The Stata Journal (2021)
21, Number 2, pp. 430–461 DOI: 10.1177/1536867X211025834

Maximum likelihood estimation of an across-regime correlation parameter

Giorgio Calzolari
University of Firenze
Firenze, Italy
giorgio.calzolari@unifi.it

Maria Gabriella Campolo
University of Messina
Messina, Italy
mgcampolo@unime.it

Antonino Di Pino
University of Messina
Messina, Italy
dipino@unime.it

Laura Magazzini
Sant'Anna School of Advanced Studies
Pisa, Italy
laura.magazzini@santannapisa.it

Abstract. In this article, we describe the `mlcar` command, which implements a maximum likelihood method to simultaneously estimate the regression coefficients of a two-regime endogenous switching model and the coefficient measuring the correlation of outcomes between the two regimes. This coefficient, known as the "across-regime" correlation parameter, is generally unidentified in the traditional estimation procedures.

Keywords: st0642, mlcar, mlcartestn, Roy model, endogenous switching, maximum likelihood, across-regime correlation

1 Introduction

The two-regime switching regression models have been widely used in applied economic analysis, such as in the estimation of the earnings equations for unionized and nonunionized workers or in the estimation of wage equations of subjects employed in the private sector and in the public sector (Lee 1978; Lee and Trost 1978). Researchers have adopted several estimation methods to obtain estimates of the coefficients of the outcome equations in both regimes. The model is usually extended, and a further selection equation is included. Within this framework, maximum likelihood (ML) methods (Poirier and Ruud 1981; Maddala 1983) and two-stage procedures (Heckman 1976, 1990; Lee 1978) provided estimated coefficients of the outcome equations and of the selection equation, including variances of the error terms and covariances between the errors of the outcome equations and the selection equation.[1] In such models, the selection equation allows one to identify the choice of the regime (the decision of the agent of belonging to regime 1 or to regime 2) supporting the two outcome equations. The estimation of the outcome equations in both regimes accounts for the endogenous effect of the selection by introducing, in the respective regressors set, a correction term obtained by the "generalized residuals" of the selection equation, estimated at a first stage.

1. For example, the command `movestay`, provided by Lokshin and Sajaia (2004), implemented the ML procedure to estimate simultaneously both outcome equations and the selection equation of an endogenous switching model.

In general, the two-stage method was recognized as consistent and computationally feasible. The ML approach also considers the same three-equation set, simultaneously estimating all parameters.

However, these methods did not provide the estimation of the parameter measuring the correlation between the error terms of the two outcome equations, the so-called across-regime correlation (or covariance). The reason is that this parameter is not empirically identifiable because of the selection rule specifying a two-regime switching model, in which the dependent variable referred to an observation cannot be jointly observed in both regimes.

Despite the difficulty in identification, some "knowledge" about this parameter was considered relevant in terms of interpretation of the agent's behavior in an endogenous switching model (see Heckman and Honoré [1990] and Vijverberg [1993]). The across-regime correlation measures the correlation in unobserved productivity (ability) of the subject in both regimes (or sectors). The traditional estimation methods allow estimating the cross-correlation parameter only "indirectly", based on the estimate of coefficients and variances, and applying the relationships among the errors' second-order moments as in Maddala (1983, 223–228) and in French and Taber (2010).

Differently from these approaches, which provide an "indirect" estimation of the across-regime correlation parameter, Calzolari and Di Pino (2017) suggested that identification and direct ML estimation of the across-regime correlation parameter are possible if the model specification is closer to the traditional Roy model rather than its more widely used generalized versions. The model is specified as "two equation", implying a sort of "rational" behavior of the agent, who simply chooses the regime with the higher outcome. For each individual, the contribution to the likelihood is given by the probability density of the observed (larger) outcome and by the (conditional) probability that the alternative (censored) outcome has a smaller value.

This approach allows us to obtain a reliable simultaneous point estimation of both the outcome equations without introducing a further selection equation explaining the choice of the subject, such as in the specification of the "generalized" Roy model (for example, Carneiro, Hansen, and Heckman [2003]). This allows us to obtain more efficient estimates than those provided applying two-stage estimation methods.[2]

In this article, we describe the `mlcar` command, which implements the two-equation ML method of Calzolari and Di Pino (2017) to estimate simultaneously the coefficients of an (endogenous switching) two-equation model including the across-regime correlation coefficient. This full-information approach relies on the assumption of joint normality of the error terms of each of the two outcome equations in the respective regime.

In the next section, we briefly discuss the properties of the across-regime correlation coefficient and its relevance for economic analysis. In section 3, we provide a brief description of the methodology and model specification.

2. Calzolari and Di Pino (2017) checked the relative efficiency of the two-equation ML estimates, performing several Monte Carlo experiments. In some experiments, efficiency was confirmed also by considering distributions different from the normal.

Because our full-information approach relies on the assumption of normality of the error terms in each regime, we also provide a postestimation command to verify the hypothesis of normality of the error terms in both regimes (`mlcartestn`). This testing procedure is an extension to the two-regime endogenous switching models of the conditional moment (CM) test, which verifies the normality assumption in the censored regression model (tobit; see, for example, Newey [1985]; Tauchen [1985]; Skeels and Vella [1999]). We report a brief description of this procedure in section 3.1.

In section 4, we describe the `mlcar` command and its options followed by general examples of application. In section 5, we report the results of some empirical applications of the `mlcar` command.

To provide a comparison with the `mlcar` results, in the appendix, we consider the procedure that should be applied for the indirect estimation of the cross-correlation coefficient if the endogenous switching model is estimated in one of the traditional ways. Appendix A briefly describes how to obtain the indirect estimate of the across-regime correlation parameter via the two-stage Heckman procedure, and in appendix B, we consider the same empirical applications of section 5 and report the indirect estimation of this parameter. Finally, in appendix C, we report the results of several Monte Carlo experiments, checking the performance of the CM test statistics by simulating data with different distributions of the error terms.

2 Relevance and empirical content of the across-regime correlation coefficient

In many cases, the two-regime switching models extend the Roy model of self-selection to include the decision rule adopted for selecting into different regimes. For example, the two-regime wage's model of self-selection aims to explain the workers' occupational choice and its consequences for the distribution of earnings when individuals differ in their endowments of specific skills (see Heckman and Honoré [1990]; Vijverberg [1993]). In doing this, one should obtain information about the joint distribution of the potential (counterfactual) outcomes. A relevant parameter of such a distribution is the across-correlation coefficient, ρ_{12}.[3]

Heckman and Honoré (1990) proved that the identification of the joint distribution of potential outcomes is essential to the empirical content of this model. As shown by these authors, if the ρ_{12} coefficient is identified, one can, by adopting a two-regime specification as in a Roy model, estimate the population distribution of potential outcomes knowing only the outcomes of subjects observed into one of the two regimes.

The sign of the across-regime correlation, in particular, allows us to know more in detail what criterion the agents follow to select the regime. Considering a wage model in a public or private sector choice, for example, a positive sign of ρ_{12} signals that the agents, supported by their own skills, manage to gain a higher-than-average level of

3. The subscripts 1 and 2 indicate the two different regimes.

outcome in both regimes. Thus, one of the two sectors (public sector) absorbs most of the above-average productive workers.

At the opposite, a negative sign of ρ_{12} means that the agent has different skills in each regime, and he or she chooses the regime in which he or she is more productive. In this case, the workers are absorbed by the sector in which they gain a comparative advantage in terms of productivity. This condition generally increases the segmentation of the labor market.

An example on the use of ρ_{12} to obtain information about the skills of the agents is provided by Calzolari and Di Pino (2017), who estimated the time devoted to domestic work by employed and unemployed women in Italy. In this case, a positive sign of ρ_{12} indicated that common latent factors positively influence the domestic work supply of women in both regimes. This result led to the conclusion that employed and unemployed women do not have different skills regarding their commitment in domestic work.

Some studies showed that a knowledge of the ρ_{12} coefficient supports methods for obtaining the predictive distributions of outcomes and, consequently, an estimation of the treatment parameters (average treatment effect, average treatment effect on the treated) measuring outcome gains from program participation. Poirier and Tobias (2003), in particular, showed how the entire distribution associated with these gains can be obtained in certain situations if the ρ_{12} coefficient is, at least in part, identified.

Along this line, Fan and Wu (2010) provide sharp bounds to obtain a partial identification of the correlation coefficient of the potential outcomes, their joint distribution, and the distribution of treatment effects.

The aforementioned studies on the use of two-regime switching models adopt partial information on the ρ_{12} coefficient to derive predictive distributions. Instead, an important result achieved by applying the estimator implemented by the `mlcar` procedure consists in obtaining a direct point estimation of the ρ_{12} parameter, supported by the typical inferential properties of the ML estimators.

3 Methodological issues

Calzolari and Di Pino (2017) specified an endogenous switching model with two regression equations whose dependent variables (outcomes) are mutually exclusive in a cross-sectional framework and where selection is simply based on the choice of the larger outcome.

$$\begin{aligned} y_{1i} &= \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + u_{1i} \quad \text{if observed in regime 1; otherwise latent} \\ y_{2i} &= \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + u_{2i} \quad \text{if observed in regime 2; otherwise latent} \end{aligned}$$

The agent is assumed to behave rationally; thus, if $y_{1i} > y_{2i}$, then y_{1i} is observed and y_{2i} is latent; otherwise, y_{2i} is observed and y_{1i} is latent.

A relevant characteristic of this model is that the two dependent variables, y_{1i} and y_{2i}, are explicitly factors in the choice of the regime. For each individual, $y_{1i} - y_{2i}$ represents the net gain (or net loss) from the choice between two options.

The error terms u_{1i} and u_{2i}, given by $u_{1i} = y_{1i} - \mathbf{x}'_{1i}\boldsymbol{\beta}_1$ and $u_{2i} = y_{2i} - \mathbf{x}'_{2i}\boldsymbol{\beta}_2$, are assumed to be normally distributed with zero mean and variances σ_1^2 and σ_2^2. Identification and estimation of the across-regime covariance, σ_{12}, becomes possible by considering (as in a tobit model) the probability density of the observed outcome, multiplied by the conditional probability that the other outcome (latent) is smaller than the observed. More in detail, the censoring rule in the model implies that

$$
\begin{aligned}
y_{1i} \text{ observed} &\Rightarrow y_{2i} < y_{1i} \Rightarrow \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + u_{2i} < y_{1i} \\
y_{2i} \text{ observed} &\Rightarrow y_{1i} \leq y_{2i} \Rightarrow \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + u_{1i} \leq y_{2i}
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\phi(y_{1i})P(y_{2i} < y_{1i}) &= \phi(u_{1i})P(u_{2i} < y_{1i} - \mathbf{x}'_{2i}\boldsymbol{\beta}_2 | y_{1i} \text{ observed}) \\
\phi(y_{2i})P(y_{1i} \leq y_{2i}) &= \phi(u_{2i})P(u_{1i} \leq y_{2i} - \mathbf{x}'_{1i}\boldsymbol{\beta}_1 | y_{2i} \text{ observed})
\end{aligned}
$$

where $\phi(\cdot)$ is a normal probability density function.

We consider also the CMs of the error terms; namely, $E(u_{1i}|u_{2i}) = (\sigma_{12}/\sigma_2^2)u_{2i} = (\sigma_{12}/\sigma_2^2)(y_{2i} - \mathbf{x}'_{2i}\boldsymbol{\beta}_2)$ and $\text{Var}(u_{1i}|u_{2i}) = \sigma_1^2 - (\sigma_{12}^2/\sigma_2^2)$ are, respectively, the conditional mean and variance of u_{1i} given u_{2i}. Analogously, $E(u_{2i}|u_{1i}) = (\sigma_{12}/\sigma_1^2)u_{1i} = (\sigma_{12}/\sigma_1^2)(y_{1i} - \mathbf{x}'_{1i}\boldsymbol{\beta}_1$ and $\text{Var}(u_{2i}|u_{1i}) = \sigma_2^2 - (\sigma_{12}^2/\sigma_1^2)$ are, respectively, the conditional mean and variance of u_{2i} given u_{1i}. Hence, σ_{12} is the covariance between the error terms of both regimes, known as the across-regime covariance.

Therefore, in (1) we have the probability that an agent does not belong to regime 2, under the condition that he or she chooses regime 1:

$$
P(u_{2i} \leq y_{1i} - \mathbf{x}'_{2i}\boldsymbol{\beta}_2 | y_{1i}\text{observed}) = \Phi\left\{ \frac{(y_{1i} - \mathbf{x}'_{2i}\boldsymbol{\beta}_2) - \frac{\sigma_{12}}{\sigma_1^2}(y_{1i} - \mathbf{x}'_{1i}\boldsymbol{\beta}_1)}{\sqrt{\sigma_2^2 - \sigma_{12}^2/\sigma_1^2}} \right\} \tag{1}
$$

Analogously, in (2) we have the probability that an agent does not belong to regime 1, under the condition that he or she chooses regime 2:

$$
P(u_{1i} \leq y_{2i} - \mathbf{x}'_{1i}\boldsymbol{\beta}_1 | y_{2i}\text{observed}) = \Phi\left\{ \frac{(y_{2i} - \mathbf{x}'_{1i}\boldsymbol{\beta}_1) - \frac{\sigma_{12}}{\sigma_2^2}(y_{2i} - \mathbf{x}'_{2i}\boldsymbol{\beta}_2)}{\sqrt{\sigma_1^2 - \sigma_{12}^2/\sigma_2^2}} \right\} \tag{2}
$$

$\Phi(\cdot)$ is the standard normal cumulative distribution function used to specify, in both (1) and (2), the contribution to the likelihood of censoring, respectively, y_{2i} and y_{1i}.

Therefore, given the conditional probabilities (1) and (2), we finally obtain the following contribution of the ith observation to the log likelihood,

$$\begin{aligned}\ln L(\boldsymbol{\theta})_i = R_i & \left[-\frac{(y_{1i} - \mathbf{x}_{1i}\boldsymbol{\beta}_1)^2}{2\sigma_1^2} - \frac{1}{2}\ln\sigma_1^2 + \ln\Phi \left\{ \frac{(y_{1i} - \mathbf{x}'_{2i}\boldsymbol{\beta}_2) - \frac{\sigma_{12}}{\sigma_1^2}(y_{1i} - \mathbf{x}'_{1i}\boldsymbol{\beta}_1)}{\sqrt{\sigma_2^2 - \sigma_{12}^2/\sigma_1^2}} \right\} \right] \\ + (1 - R_i) & \left[-\frac{(y_{2i} - \mathbf{x}'_{2i}\boldsymbol{\beta}_2)^2}{2\sigma_2^2} - \frac{1}{2}\ln\sigma_2^2 + \ln\Phi \left\{ \frac{(y_{2i} - \mathbf{x}'_{1i}\boldsymbol{\beta}_1) - \frac{\sigma_{12}}{\sigma_2^2}(y_{2i} - \mathbf{x}'_{2i}\boldsymbol{\beta}_2)}{\sqrt{\sigma_1^2 - \sigma_{12}^2/\sigma_2^2}} \right\} \right] \end{aligned} \quad (3)$$

with $\boldsymbol{\theta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \sigma_1^2, \sigma_2^2, \sigma_{12})'$, while R_i is a dummy-indicator variable equal to 1 if y_{1i} is observed (regime 1) and equal to 0 if y_{2i} is observed (regime 2). Applying this two-equation ML procedure, we can directly estimate the parameter σ_{12} (or ρ_{12}) under the assumption of endogenous selection.

3.1 A CM test of normality for a two-regime switching model

The ML estimator critically relies on the assumption of normality of the error terms of both equations. As a complement to the estimation procedure, we implement a CM test to verify the normality assumption. The proposed test procedure extends, to the two-equation case, the CM test available in the literature to verify the normality assumption in the context of the tobit model (for example, Skeels and Vella [1999]). In particular, the test is based on the comparison of the third and fourth moments of u_{1i} and u_{2i} with the theoretical values implied under the assumption of normally distributed error terms. Absent censoring, we could write

$$\begin{aligned} E(u_{1i}^3) &= 0 & E(u_{2i}^3) &= 0 \\ E(u_{1i}^4 - 3\sigma_1^4) &= 0 & E(u_{2i}^4 - 3\sigma_2^4) &= 0 \end{aligned}$$

However, these equalities cannot be satisfied on the "observed" part of each regime, because of censoring.

The CM test is built by considering the following observed residuals:

$$\begin{aligned} v_{3i} &= R_i\{u_{1i}^3 - E(u_{1i}^3|y_{1i}\text{observed})\} + (1 - R_i)\{u_{2i}^3 - E(u_{2i}^3|y_{2i}\text{observed})\} \\ v_{4i} &= R_i\{u_{1i}^4 - E(u_{1i}^4|y_{1i}\text{observed})\} + (1 - R_i)\{u_{2i}^4 - E(u_{2i}^4|y_{2i}\text{observed})\} \end{aligned}$$

The moment conditions that we exploit to verify the normality assumption can therefore be written as

$$\begin{aligned} E(v_{3i}) &= 0 \\ E(v_{4i}) &= 0 \end{aligned}$$

with v_{3i} and v_{4i} including powers of the observed residuals in regime 1 and regime 2 as defined before.

For observations in regime 1, we can write

$$\widehat{u}_{1i}^3 = \left(y_{1i} - \mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1\right)^3 \quad \text{and} \quad \widehat{u}_{1i}^4 = \left(y_{1i} - \mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1\right)^4$$

Analogous formulas hold for observations in regime 2:

$$\widehat{u}_{2i}^3 = \left(y_{2i} - \mathbf{x}_{2i}'\widehat{\boldsymbol{\beta}}_2\right)^3 \quad \text{and} \quad \widehat{u}_{2i}^4 = \left(y_{2i} - \mathbf{x}_{2i}'\widehat{\boldsymbol{\beta}}_2\right)^4$$

To perform the computations related to the testing procedure, we also need to evaluate the following CMs:

$$\begin{array}{ll} E\left(u_{1i}^3 | y_{1i}\text{observed}\right) & E\left(u_{2i}^3 | y_{2i}\text{observed}\right) \\ E\left(u_{1i}^4 | y_{1i}\text{observed}\right) & E\left(u_{2i}^4 | y_{2i}\text{observed}\right) \end{array}$$

Focus on the computation related to u_{1i}; an analogous formula applies for u_{2i}.

Under the assumption of joint normality of u_{1i} and u_{2i}, we note that the difference $\delta_i = u_{1i} - u_{2i}$ is also normally distributed. Thus, u_{1i} can be written as a linear function of δ_i plus an independent error term,

$$u_{1i} = \tau_1 \delta_i + \epsilon_{1i}$$

with ϵ_{1i} normally distributed, independent of δ_i, and $\tau_1 = \text{cov}(\delta_i, u_{1i})/\text{var}(\delta_i)$. It holds that $E(\epsilon_{1i}) = 0$, $E(\epsilon_{1i}^2) = \sigma_\epsilon^2$, $E(\epsilon_{1i}^3) = 0$, and $E(\epsilon_{1i}^4) = 3\sigma_\epsilon^4$. We therefore can write

$$\begin{aligned} E(u_{1i}^3 | y_{1i}\text{observed}) &= E\{(\tau_1\delta_i + \epsilon_{1i})^3 | \delta_i \leq \mathbf{x}_{1i}'\boldsymbol{\beta}_1 - \mathbf{x}_{2i}'\boldsymbol{\beta}_2\} \\ E(u_{1i}^4 | y_{1i}\text{observed}) &= E\{(\tau_1\delta_i + \epsilon_{1i})^4 | \delta_i \leq \mathbf{x}_{1i}'\boldsymbol{\beta}_1 - \mathbf{x}_{2i}'\boldsymbol{\beta}_2\} \end{aligned}$$

The two expected values can be computed by exploiting the recursive formula that characterizes the moments of a truncated normal distribution (see, for example, Chesher and Irish [1987, 40]) and exploiting the independence of ϵ_{1i} and δ_i (see also Pfaffermayr [2014]).

The computation in `mlcartestn` is based on the outer-product-gradient formula: consider the vector $\mathbf{w}_i$, which includes the gradient of the log likelihood function (3) and the residuals,

$$\mathbf{w}_i = \left(\frac{\partial \text{ln} L_i}{\partial \boldsymbol{\theta}'}, \widehat{v}_{3i}, \widehat{v}_{4i}\right)$$

with $\boldsymbol{\theta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \sigma_1^2, \sigma_2^2, \sigma_{12})'$. Build the matrix $\mathbf{W}$ with rows $\mathbf{w}_i$. The test is obtained as

$$\text{CM} = \boldsymbol{\iota}'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\iota}$$

with $\boldsymbol{\iota}$ a column vector of ones. The test corresponds to nR^2 with the uncentered coefficient of determination of the regression of $\boldsymbol{\iota}$ on $\mathbf{w}_i$. Computed in this form, the test is known to have small-sample problems in finite samples (for example, Drukker [2002]); it is oversized in finite samples. To address this issue, we also provide a simulated version of the CM test as in Orme (1995).

4 The mlcar command

4.1 Syntax

`mlcar` fits a two-equation endogenous switching model using the procedure described in Calzolari and Di Pino (2017). The dependent variable (*depvar*) is recorded across two regimes, as identified by the selection variable specified in the (required) option `regime(`*varname*`)`. The generic syntax for the command is as follows:

`mlcar` *depvar* [*if*] [*in*] [*weight*] `, regime(`*varname*`) x1(`*varlist*`)` [`x2(`*varlist*`)` `accuracy(0|1|2)` `olsinit` `level(`*#*`)` *maximize_options*]

`fweights`, `aweights`, `iweights`, and `pweights` are allowed; see [U] **11.1.6 weight**.

The dependent variable *depvar* is recorded across two regimes, as identified by the variable specified in the (required) option `regime(`*varname*`)`:

$$y_1 = \textit{depvar} \text{ if } \textit{varname} = 1$$
$$y_2 = \textit{depvar} \text{ if } \textit{varname} = 0$$

It is assumed that the individual chooses the regime with the highest outcome; that is,

$$y_2 \geq y_1 \text{ if } \textit{varname} = 0$$
$$y_2 < y_1 \text{ if } \textit{varname} = 1$$

The variances of the error terms of the outcome equations are $\sigma_1^2 = \text{s11}$ and $\sigma_2^2 = \text{s22}$, and the covariance between the two error terms is $\sigma_{12} = \text{s12}$. The across-regime correlation can be computed as $\rho_{12} = \text{r12} = \text{s12}/\text{sqrt}(\text{s11} \times \text{s22})$.[4]

4.2 Options

`regime()` identifies the variable that specifies the two regimes, one coded as 0 (y_2 is recorded in *depvar*) and the other as equal to 1 (y_1 is recorded in *depvar*). `regime()` is required.

`x1(`*varlist*`)` and, optionally, `x2(`*varlist*`)` specify the list of variables. When the same set of regressors `$XLIST` is specified in both outcome equations, these can be specified in the (required) option `x1()` as `x($XLIST)`. However, the set of regressors in `x1()` and `x2()` need not be the same: a different list of variables can be specified in `x1()` and in `x2()` to be used as independent variables for the outcome equation of regime 1 and 2, respectively. `x1()` is required.

4. Besides the coefficients of the equations that characterize the two regimes, the likelihood function is written in terms of θ_1, θ_2, and θ_3, where $\text{s11} = \exp(\theta_1)$, $\text{s22} = \exp(\theta_2)$ (as to guarantee that the variances are positive), and $\text{r12} = \tanh(\theta_3)$ [to bound the correlation parameter in the interval $(-1, 1)$].

`accuracy()` defines how the gradient vector and the Hessian matrix are computed:

If `accuracy(0)`, both gradient and Hessian are obtained in a numeric way (`method(lf0)` is used with the `ml` command).

If `accuracy(1)`, the gradient vector is computed using the analytic formula (`method(lf1)` is used with the `ml` command; the Hessian is still computed using numeric approximation).

If `accuracy(2)` (the default), both gradient and Hessian are computed using the analytic formula (`method(lf2)` is used with the `ml` command).

`olsinit` specifies to use the ordinary least-squares estimates as initial values for the `ml` estimation (in this case, the starting value of r12 is set equal to 0). Alternatively, the user can specify different initial values using the option `init(`*ml_init_args*`)`, available with the `ml` command. If no initial value is specified, `mlcar` lets the `ml` command search for initial values.

`level(#)` specifies the confidence level. By default, the value in macro `S_level` is considered. The default is `level(95)`.

maximize_options specifies the options of the Stata command `ml model`; see [R] **ml** for details.

4.3 Postestimation

The postestimation command `predict` can be used after `mlcar`. The syntax is

`predict` *newvar* [`, xb1 xb2 pnb12 pnb21`]

The following options are allowed to compute these conditional and unconditional expectations:

`xb1` calculates the linear prediction in regime 1 for observations in regime 1 and in regime 2 (the default):

$$\widehat{y}_{1i} = \mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1$$

`xb2` calculates the linear prediction in regime 2 for observations in regime 2 and in regime 1:

$$\widehat{y}_{2i} = \mathbf{x}_{2i}'\widehat{\boldsymbol{\beta}}_2$$

`pnb12` calculates the probability of not being in regime 1, for units deciding to belong to regime 2:

$$\Phi\left\{\frac{\left(y_{2i} - \mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1\right) - \frac{\widehat{\sigma}_{12}}{\widehat{\sigma}_2^2}\left(y_{2i} - \mathbf{x}_{2i}'\widehat{\boldsymbol{\beta}}_2\right)}{\sqrt{\widehat{\sigma}_1^2 - \widehat{\sigma}_{12}^2/\widehat{\sigma}_2^2}}\right\}$$

`pnb21` calculates the probability of not being in regime 2, for units deciding to belong to regime 1:[5]

$$\Phi\left\{\frac{\left(y_{1i}-\mathbf{x}_{2i}'\widehat{\boldsymbol{\beta}}_2\right)-\frac{\widehat{\sigma}_{12}}{\widehat{\sigma}_1^2}\left(y_{1i}-\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1\right)}{\sqrt{\widehat{\sigma}_2^2-\widehat{\sigma}_{12}^2/\widehat{\sigma}_1^2}}\right\}$$

After `mlcar`, `mlcartestn` performs the CM test for joint normality of the error terms. The default computation of the test statistics uses the outer-product-gradient form (Skeels and Vella 1999). The syntax is

`mlcartestn` [, `sim(`#`)`]

`sim(`#`)` permits one to compute the simulated version of the CM test as in Orme (1995).

5 Examples

We illustrate the use of the `mlcar` command with four examples. The first two datasets used are available from Wooldridge (2010) and readable within Stata (https://www.stata.com/texts/eacsap/); the third dataset is used by Hamermesh and Biddle (1994), and it can be downloaded from http://fmwww.bc.edu/ec-p/data/wooldridge/beauty.dta.

5. Calzolari and Di Pino apologize for some typos in their article of 2017. First of all, on the right hand side of (10), (11), and (12), the symbol $\boldsymbol{\Phi}()$ is correctly used to indicate the cumulative distribution function of the standard normal; but in all the other places between page 5 and page 7, it would have been more appropriate to replace "$\boldsymbol{\Phi}(u\ldots$" with "$P(u\ldots$". Also, the explanations of the "cumulative normal" that follow four lines after (8) have been erroneously interchanged.

Still on page 6, three lines before the end or the page, in the expression of the conditional variance, σ_{12} should be squared.

At the top of page 7, after (10), the lines 2 and 3 should be written as "Analogously, the probability of a subject not belonging to regime 2 under the condition that he or she chooses regime 1 is given by [equation (11) follows]."

In (10), (11), and (12) parentheses have been incorrectly applied to the denominators, that should be, respectively, $\sqrt{(\sigma_1^2-\sigma_{12}^2/\sigma_2^2)}$ and $\sqrt{(\sigma_2^2-\sigma_{12}^2/\sigma_1^2)}$ in place of $\sqrt{(\sigma_1^2-\sigma_{12}^2)/\sigma_2^2}$ and $\sqrt{(\sigma_2^2-\sigma_{12}^2)/\sigma_1^2}$.

In appendix A, lines 5 and 6 should be rewritten as : "$\ldots v_i = u_{1i}-u_{2i} > -(\mathbf{x}_{1i}'\boldsymbol{\beta}_1-\mathbf{x}_{2i}'\boldsymbol{\beta}_2)$, or $v_i = u_{1i}-u_{2i} \leq -(\mathbf{x}_{1i}'\boldsymbol{\beta}_1-\mathbf{x}_{2i}'\boldsymbol{\beta}_2)$, where the random variable $v_i = u_{1i}-u_{2i}$ is normally distributed with zero mean and variance $\sigma_{v\ldots}^2$".

Finally, between the two lines of (12), there was the sentence "if y_{1i} is observed (regime 1); otherwise it is", but the entire sentence was erroneously canceled.

5.1 Example 1

In the first example, we use `fringe.dta`, a dataset reporting wages, hourly benefits and demographic information on 616 workers. The dataset includes information about the individual earning, the years of work experience, the years at school, and about the membership of single workers to a union. This dataset allows us to estimate the individual wage in a two-regime union or nonunion model. We start by loading the dataset and providing some descriptive statistics:

```
. use https://www.stata.com/data/jwooldridge/eacsap/fringe
. generate lannhrs_noff = lannhrs
. replace lannhrs_noff = 0 if office == 1
(298 real changes made)
. label variable lannhrs_noff "log(annual hours worked) if no office worker"
. generate lannhrs_off = lannhrs
. replace lannhrs_off = 0 if office == 0
(318 real changes made)
. label variable lannhrs_off "log(annual hours worked) if office worker"
. describe lannearn lannhrs_off lannhrs_noff lannhrs exper expersq male annbens
> office educ union

              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
lannearn        float   %9.0g                 log(annearn)
lannhrs_off     float   %9.0g                 log(annual hours worked) if office
                                                worker
lannhrs_noff    float   %9.0g                 log(annual hours worked) if no
                                                office worker
lannhrs         float   %9.0g                 log(annhrs)
exper           byte    %9.0g                 years work experience
expersq         int     %9.0g                 exper^2
male            byte    %9.0g                 =1 if male
annbens         float   %9.0g                 vacdays+sicklve+insur+pension
office          byte    %9.0g                 =1 if office worker
educ            byte    %9.0g                 years schooling
union           byte    %9.0g                 =1 if union member
. by union, sort: summarize lannearn lannhrs_off lannhrs_noff lannhrs exper male
> annbens office educ
-------------------------------------------------------------------------------
-> union = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    lannearn |        420    9.216772    .6507578   6.575912   11.68688
 lannhrs_off |        420    4.584254    3.734404          0   8.451054
lannhrs_noff |        420    3.032933    3.740712          0   8.313852
     lannhrs |        420    7.617186    .2546784   6.436151   8.451054
       exper |        420    17.47381    12.24343          0         60
-------------+---------------------------------------------------------
        male |        420    .5857143    .4931858          0          1
     annbens |        420    1613.699    1299.026          0    4780.01
      office |        420     .602381    .4899896          0          1
        educ |        420    12.86429    2.660264          6         18
-------------------------------------------------------------------------------
```

```
-> union = 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    lannearn |        196    9.458707     .411354   7.867565   10.30895
 lannhrs_off |        196    1.729401    3.177236          0   7.917172
lannhrs_noff |        196    5.876007    3.219449          0     8.2623
     lannhrs |        196    7.605408    .1775188   7.090077     8.2623
       exper |        196    21.06633    12.18451          1         50
-------------+---------------------------------------------------------
        male |        196    .7602041    .4280522          0          1
     annbens |        196    2495.977    1276.506          0    5129.13
      office |        196    .2295918    .4216474          0          1
        educ |        196    11.76531    2.743014          6         18
```

The outcome of interest is `lannearn`, the logarithmic of the annual earnings, while the variable that identifies the regime is `union`, a dummy variable that assumes a value equal to 1 if workers have established any form of workers' representation at the workplace. The set of covariates, in the output above, includes the years of experience and its square, the level of education measured in years of schooling and its square, a dummy variable equal to 1 if the subject is a male, a dummy variable equal to 1 if the subject is an office worker (equal to 0 if the subject performs manual work), the annual hours worked, and the level of the annual benefits. The basic syntax for `mlcar` is the following:

```
. mlcar lannearn, regime(union)
> x1(lannhrs_off lannhrs_noff exper expersq male annbens)
> x2(lannhrs exper expersq office educ male)
initial:       log likelihood = -26712.266
alternative:   log likelihood = -14666.355
rescale:       log likelihood = -1812.6506
rescale eq:    log likelihood =  -1176.497
Iteration 0:   log likelihood =  -1176.497  (not concave)
Iteration 1:   log likelihood = -891.12989  (not concave)
Iteration 2:   log likelihood = -652.68786  (not concave)
Iteration 3:   log likelihood = -430.03486  (not concave)
Iteration 4:   log likelihood = -204.32947  (not concave)
Iteration 5:   log likelihood = -117.10794
Iteration 6:   log likelihood = -22.215082
Iteration 7:   log likelihood = -6.4281359
Iteration 8:   log likelihood =  16.555086
Iteration 9:   log likelihood =  18.706246
Iteration 10:  log likelihood =  18.719583
Iteration 11:  log likelihood =  18.719584
                                                  Number of obs   =        616
                                                  Wald chi2(6)    =     462.73
Log likelihood =  18.719584                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
    lannearn |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Y1           |
 lannhrs_off |   .3796578   .1143129     3.32   0.001     .1556086     .603707
lannhrs_noff |   .4202827   .1131525     3.71   0.000     .1985078    .6420576
       exper |   .0200713   .0080979     2.48   0.013     .0041997    .0359428
     expersq |  -.0003675   .0001712    -2.15   0.032    -.0007031    -.000032
        male |   .1943418   .0664524     2.92   0.003     .0640974    .3245862
     annbens |   .0003687   .0000275    13.41   0.000     .0003148    .0004226
       _cons |    4.72104   .8388029     5.63   0.000     3.077016    6.365063
-------------+----------------------------------------------------------------
Y2           |
     lannhrs |   1.141442   .0920309    12.40   0.000     .9610652     1.32182
       exper |   .0166853   .0058733     2.84   0.004     .0051739    .0281967
     expersq |   -.000239   .0001266    -1.89   0.059    -.0004871    9.12e-06
      office |   .3521486   .0474501     7.42   0.000     .2591481    .4451491
        educ |   .0564712   .0085692     6.59   0.000     .0396759    .0732665
        male |   .3736418   .0467833     7.99   0.000     .2819482    .4653353
       _cons |  -.9233093    .696087    -1.33   0.185    -2.287615    .4409961
-------------+----------------------------------------------------------------
lnsigma11    |
       _cons |  -1.562106   .1373886   -11.37   0.000    -1.831383   -1.292829
-------------+----------------------------------------------------------------
lnsigma22    |
       _cons |  -1.607133   .0813521   -19.76   0.000     -1.76658   -1.447686
-------------+----------------------------------------------------------------
tanhrho      |
       _cons |  -.3751207    .148885    -2.52   0.012    -.6669299   -.0833115
-------------+----------------------------------------------------------------
     sigma11 |    .209694   .0288096                      .1601919    .2744931
     sigma22 |   .2004615    .016308                      .1709165    .2351137
       rho12 |  -.3584626   .1297539                     -.5829568   -.0831193
------------------------------------------------------------------------------
  Y1 corresponds to regime!=0
  Y2 corresponds to regime==0
------------------------------------------------------------------------------
```

```
. mlcartestn, sim(100)
Conditional moment test for normality of residuals after mlcar
     chi(2) =      863 - p-value = 0.0000
```

Null hypothesis of normality of the errors is rejected. In this application, the set of regressors is not the same for both regimes, so we specify both the option x1(*varlist*) and x2(*varlist*).

The option regime() identifies the variable (union) that specified the two regimes (unionized or nonunionized workers). The variable *depvar* includes observations on both y_1 and y_2. Observations corresponding to union that are equal to 0 identify y_2 in *depvar*; when union is coded as 1 (or any value different from 0), y_1 is recorded in *depvar*.

The first panel of the output of mlcar provides the estimated coefficients of the equation under regime 1 (unionized workers). The second panel provides estimated coefficients of the equation under regime 2 (nonunionized workers). In the last part of the output, the value of the across-regime correlation is reported.

sigma11 and sigma22 are the variances of the residuals of the regression part of the model, and lnsigma11 and lnsigma22 are their log.

The estimation results show that the impact of the yearly worked hours on earned income is generally positive and stronger for nonunionized workers than unionized workers. Among the latter, the effect of worked hours is strongest for those who do not perform office work. Education exerts a positive influence on labor income of nonunionized workers. Finally, in both union and nonunion regimes, work experience exerts a positive influence on labor income, albeit with decreasing rates of growth.

The across-regime correlation, rho12, is equal to -0.358, while the covariance s12 is equal to -0.0734. The negative sign of rho12 signals how less skilled workers, who usually gain less than average if nonunionized, have a "comparative advantage" in terms of perceived earnings if they join the union.

We obtain a cross-correlation parameter with a negative sign ($\rho_{12} = -0.18$) even if we apply the indirect procedure of the two-step Heckman estimation (appendix A). The model's estimation results after the two-step Heckman estimation are reported in appendix B.

5.2 Example 2

In the second example, we use 401ksubs.dta, a cross-sectional survey on eligibility for participation of 9,275 individuals in the U.S. 401k pension plan, including their income data and other demographic information. We adopt the family financial assets as a dependent variable, while we include household per capita income, age, participation in another pension plan (individual retirement account [IRA]), and family status as explanatory variables in the model. A subject belongs to regime 1 if he or she participates in the 401k plan, while he or she belongs to regime 2 if not associated with the 401k pension plan. In the following table, we report the descriptive statistics relative to the variables used in our analysis:

```
. use http://www.stata.com/data/jwooldridge/eacsap/401ksubs, clear
. generate inc_percap = inc/fsize
. label variable inc_percap "=inc/fsize"
. generate marr_pira=marr*pira
. label variable marr_pira "=married*IRA"
. generate nonmarr_pira = (1-marr)*pira
. label variable nonmarr_pira "=(1-married)*IRA"
. describe nettfa p401k inc_percap age agesq marr_pira nonmarr_pira
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
nettfa          float   %9.0g                 net total fin. assets, $1000
p401k           byte    %9.0g                 =1 if participate in 401(k)
inc_percap      float   %9.0g                 =inc/fsize
age             byte    %9.0g                 age^2
agesq           int     %9.0g                 age^2
marr_pira       float   %9.0g                 =married*IRA
nonmarr_pira    float   %9.0g                 =(1-married)*IRA
. by p401k, sort: summarize nettfa inc_percap age agesq marr_pira nonmarr_pira
--------------------------------------------------------------------------------
-> p401k = 0
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      nettfa |      6,713    11.66722    55.28923   -502.302   1462.115
  inc_percap |      6,713    15.95789    12.69433       1.02    143.067
         age |      6,713    40.91494    10.53225         25         64
       agesq |      6,713    1784.944    916.4837        625       4096
   marr_pira |      6,713    .1479219     .355049          0          1
-------------+---------------------------------------------------------
nonmarr_pira |      6,713    .0652465    .2469788          0          1
--------------------------------------------------------------------------------
-> p401k = 1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      nettfa |      2,562    38.47296    79.27108   -283.356   1536.798
  inc_percap |      2,562    21.45778    15.25077   1.640625    102.396
         age |      2,562    41.51327    9.651726         25         64
       agesq |      2,562    1816.471    838.3487        625       4096
   marr_pira |      2,562    .2802498    .4492089          0          1
-------------+---------------------------------------------------------
nonmarr_pira |      2,562    .0819672    .2743683          0          1
```

The outcome of interest is `nettfa`, the net family financial assets in thousands of dollars, and the variable that identifies the regime is `p401k`, which assumes value equal to 1 if the individual is associated with the 401k pension plan (0 otherwise). The set of covariates, in the output above, includes the income per capita, the age of the individual and its square, and two interaction dummy variables signaling whether the subject is both married and associated with the IRA or whether he or she is not married and associated with the IRA.

In this second example, we used the same covariates for both regimes. Thus, the list of variables is specified only in `x1()`.

As for the results of the estimates, we can observe that married people who are also associated with an IRA pension plan are generally more willing to participate in the 401k plan. In addition, the results show that income availability and married condition jointly affect the propensity to set aside financial assets and participate in the 401k plan. The availability of financial assets is positively correlated with age for those who choose to join the 401k plan; the opposite occurs for those who do not join the 401k, whose financial assets decrease with increasing age.

```
. mlcar nettfa, regime(p401k) x1(inc_percap age agesq marr_pira nonmarr_pira)
initial:       log likelihood =     -<inf>  (could not be evaluated)
feasible:      log likelihood = -1679277.6
rescale:       log likelihood =  -53682.56
rescale eq:    log likelihood = -49287.352
Iteration 0:   log likelihood = -49287.352
Iteration 1:   log likelihood = -46651.285
Iteration 2:   log likelihood = -45617.634
Iteration 3:   log likelihood = -45515.138
Iteration 4:   log likelihood = -45512.746
Iteration 5:   log likelihood = -45512.745
                                                Number of obs     =      9,275
                                                Wald chi2(5)      =     477.54
Log likelihood = -45512.745                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
      nettfa |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Y1           |
  inc_percap |   1.303126   .1016447    12.82   0.000     1.103906    1.502346
         age |    7.94995   1.166818     6.81   0.000     5.663029    10.23687
       agesq |  -.0882846   .0133711    -6.60   0.000    -.1144915   -.0620777
   marr_pira |   48.70116   3.527254    13.81   0.000     41.78787    55.61445
nonmarr_pira |    13.6085   5.603566     2.43   0.015     2.625714    24.59129
       _cons |  -272.7242   24.69672   -11.04   0.000    -321.1289   -224.3196
-------------+----------------------------------------------------------------
Y2           |
  inc_percap |   .1554649    .051188     3.04   0.002     .0551383    .2557915
         age |  -3.408809   .5127297    -6.65   0.000    -4.413741   -2.403877
       agesq |    .046012   .0058888     7.81   0.000       .03447    .0575539
   marr_pira |   21.45378   1.804081    11.89   0.000     17.91785    24.98971
nonmarr_pira |   18.68462   2.695435     6.93   0.000     13.40166    23.96757
       _cons |   43.77803   10.69928     4.09   0.000     22.80782    64.74823
-------------+----------------------------------------------------------------
lnsigma11    |
       _cons |   9.319316   .0324298   287.37   0.000     9.255754    9.382877
-------------+----------------------------------------------------------------
lnsigma22    |
       _cons |   8.125216   .0189222   429.40   0.000     8.088129    8.162303
-------------+----------------------------------------------------------------
tanhrho      |
       _cons |  -1.056451   .0184238   -57.34   0.000    -1.092561   -1.020341
-------------+----------------------------------------------------------------
     sigma11 |   11151.35   361.6364                      10464.61    11883.15
     sigma22 |   3378.598   63.93062                      3255.591    3506.252
       rho12 |  -.7843016   .0070908                     -.7978108   -.7700052
------------------------------------------------------------------------------
  Y1 corresponds to regime!=0
  Y2 corresponds to regime==0
------------------------------------------------------------------------------
```

```
. mlcartestn, sim(100)
Conditional moment test for normality of residuals after mlcar
     chi(2) =  3.0e+06 - p-value = 0.0000
```

The null hypothesis of normality of the errors is rejected. The estimated across-regime correlation, `rho12`, is equal to −0.78, while the covariance, s12, is equal to −4814.1. In this case, the high level of the coefficient `rho12` denotes that relevant latent factors, not specified in the model as covariates, influence the choice of the regime. The negative sign of this coefficient signals that workers with net family financial assets (`nettfa`) lower than average and not participating in pension plans would have a comparative advantage in `nettfa` by joining a 401k pension plan. If we fit the model by performing a two-stage Heckman procedure (estimation results are reported in appendix B), the application of the indirect estimation of `rho12` gives an absurd value of −98.75, thus being absolutely inconsistent as a measure of correlation.

5.3 Example 3

In this example, we use `beauty.dta`. It is a dataset reporting hourly wages and demographic characteristics on 1,260 U.S. workers. The dataset can be downloaded from http://fmwww.bc.edu/ec-p/data/wooldridge/beauty.dta, and it includes information about the individual wage, the years of workforce experience, the years at school, gender and race, and whether the subject works in the service industry. We start by loading the dataset, and we provide some descriptive statistics after trimming some observations with outliers in the dependent variable.

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/beauty, clear
. generate lwage2 = lwage
. summarize lwage2, det
  (output omitted)
. replace lwage2 = . if lwage<=r(p1)
(15 real changes made, 15 to missing)
. replace lwage2 = . if lwage>r(p99)
(12 real changes made, 12 to missing)
. generate collgrad =  educ>=12
. drop if lwage2 == .
(27 observations deleted)
. describe lwage2 exper expersq collgrad female black service

              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------

lwage2          float   %9.0g
exper           byte    %8.0g                 years of workforce experience
expersq         int     %8.0g                 exper^2
collgrad        float   %9.0g
female          byte    %8.0g                 =1 if female
black           byte    %8.0g                 =1 if black
service         byte    %8.0g                 =1 if service industry
```

```
. by service, sort: summarize lwage2 exper expersq collgrad female black

-> service = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      lwage2 |        897    1.703497    .5304814   .2468601   3.208017
       exper |        897    18.65998    12.30306          0         48
     expersq |        897    499.3913    555.8697          0       2304
    collgrad |        897    .7781494     .415723          0          1
      female |        897    .2653289    .4417545          0          1
-------------+---------------------------------------------------------
       black |        897    .0691193    .2537984          0          1

-> service = 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      lwage2 |        336    1.541193    .5800109   .2468601   3.149311
       exper |        336     17.1994    10.93231          0         46
     expersq |        336    414.9792    474.6492          0       2116
    collgrad |        336    .8928571    .3097561          0          1
      female |        336    .5505952    .4981754          0          1
-------------+---------------------------------------------------------
       black |        336    .0803571    .2722507          0          1
```

In this example, the outcome of interest is `lwage2`, the logarithm of the hourly wage, while the variable that identifies the regime is `service`, a dummy variable that assumes value equal to 1 if the subject works in the service industry. The set of covariates, in the output above, includes the years of experience and its square, a dummy variable equal to 1 if the years of schooling are greater or equal to 12, a dummy variable equal to 1 if the subject is a female, and a dummy variable equal to 1 if the subject is black.

The basic syntax for `mlcar` is the following:

```
. mlcar lwage2, r(service) x1(exper expersq collgrad female black)
initial:       log likelihood = -1981.0794
alternative:   log likelihood = -1369.8356
rescale:       log likelihood = -1369.8356
rescale eq:    log likelihood = -731.70649
Iteration 0:   log likelihood = -731.70649  (not concave)
Iteration 1:   log likelihood = -494.77166
Iteration 2:   log likelihood = -357.24137
Iteration 3:   log likelihood = -337.93524
Iteration 4:   log likelihood = -319.37493
Iteration 5:   log likelihood = -315.19668
Iteration 6:   log likelihood = -315.11101
Iteration 7:   log likelihood = -315.08604
Iteration 8:   log likelihood = -315.08432
Iteration 9:   log likelihood =  -315.0843
                                                Number of obs     =      1,233
                                                Wald chi2(5)      =     277.84
Log likelihood =  -315.0843                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
      lwage2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Y1           |
       exper |   .0484381   .0074527     6.50   0.000      .033831    .0630451
     expersq |  -.0008431   .0001696    -4.97   0.000    -.0011755   -.0005108
    collgrad |   .4065657   .1018238     3.99   0.000     .2069947    .6061367
      female |  -.2420134   .1264511    -1.91   0.056    -.4898531    .0058262
       black |  -.0483121   .0643132    -0.75   0.453    -.1743637    .0777394
       _cons |   .6306671   .3611614     1.75   0.081    -.0771962     1.33853
-------------+----------------------------------------------------------------
Y2           |
       exper |   .0359758   .0049783     7.23   0.000     .0262184    .0457332
     expersq |  -.0005748   .0001092    -5.26   0.000    -.0007888   -.0003607
    collgrad |   .2234244   .0439926     5.08   0.000     .1372006    .3096483
      female |  -.5172188   .0618616    -8.36   0.000    -.6384654   -.3959723
       black |  -.0719072   .0543878    -1.32   0.186    -.1785053    .0346909
       _cons |     1.2124   .0559033    21.69   0.000     1.102831    1.321968
-------------+----------------------------------------------------------------
lnsigma11    |
       _cons |  -1.443819   .2134236    -6.77   0.000    -1.862122   -1.025517
-------------+----------------------------------------------------------------
lnsigma22    |
       _cons |  -1.509647   .0619937   -24.35   0.000    -1.631152   -1.388141
-------------+----------------------------------------------------------------
tanhrho      |
       _cons |   .9010903   .6939439     1.30   0.194    -.4590147    2.261195
-------------+----------------------------------------------------------------
     sigma11 |   .2360246   .0503732                      .1553427    .3586112
     sigma22 |    .220988   .0136999                      .1957039    .2495387
       rho12 |   .7168284   .3373657                     -.4292808    .9785074
------------------------------------------------------------------------------
  Y1 corresponds to regime!=0
  Y2 corresponds to regime==0
------------------------------------------------------------------------------

. mlcartestn
Conditional moment test for normality of residuals after mlcar
     chi(2) =    5.586 - p-value = 0.0612
```

```
. mlcartestn, sim(100)
Conditional moment test for normality of residuals after mlcar
     chi(2) =    9.143 - p-value = 0.0103
```

The null Hypothesis of normality of the errors is not rejected if we consider a nominal test size of 0.05 when the asymptotic formula is considered and a nominal size of 0.01 when the simulated version of the test is computed.

As for the estimation results, note in particular that women's wage is lower than that of men in both regimes, especially if the women work outside the service industry. We did not obtain analogous results by performing a two-stage Heckman procedure (see appendix B).

The estimated across-regime correlation, `rho12`, is equal to 0.72. The positive sign of this coefficient signals how workers gaining more in the service industry would have gained more also working in the other sectors. However, this parameter is not statistically different from zero.

If we fit the model by performing a two-stage Heckman procedure, the value of `rho12` is equal to -110.74228, absolutely inconsistent as a measure of correlation.

6 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 21-2
. net install st0642      (to install program files, if available)
. net get st0642          (to install ancillary files, if available)
```

7 References

Azzalini, A. 1985. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* 12: 171–178.

Calzolari, G., and A. Di Pino. 2017. Self-selection and direct estimation of across-regime correlation parameter. *Journal of Applied Statistics* 44: 2142–2160. https://doi.org/10.1080/02664763.2016.1247789.

Carneiro, P., K. T. Hansen, and J. J. Heckman. 2003. 2001 Lawrence R. Klein lecture: Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review* 44: 361–422. https://doi.org/10.1111/1468-2354.t01-1-00074.

Chesher, A., and M. Irish. 1987. Residual analysis in the grouped and censored normal linear model. *Journal of Econometrics* 34: 33–61. https://doi.org/10.1016/0304-4076(87)90066-2.

Drukker, D. M. 2002. Bootstrapping a conditional moments test for normality after tobit estimation. *Stata Journal* 2: 125–139. https://doi.org/10.1177/1536867X0200200202.

Fan, Y., and J. Wu. 2010. Partial identification of the distribution of treatment effects in switching regime models and its confidence sets. *Review of Economic Studies* 77: 1002–1041. https://doi.org/10.1111/j.1467-937X.2009.00593.x.

French, E., and C. Taber. 2010. Identification of models of the labor market. In *Handbook of Labor Economics, vol. 4A*, ed. O. Ashenfelter and D. Card, 537–617. Amsterdam: Elsevier. https://doi.org/10.1016/S0169-7218(11)00412-6.

Hamermesh, D. S., and J. E. Biddle. 1994. Beauty and the labor market. *American Economic Review* 84: 1174–1194.

Heckman, J. J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5: 475–492.

———. 1990. Varieties of selection bias. *American Economic Review* 80: 313–318.

Heckman, J. J., and B. E. Honoré. 1990. The empirical content of the Roy model. *Econometrica* 58: 1121–1149. https://doi.org/10.2307/2938303.

Lee, L. 1978. Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *International Economic Review* 19: 415–433. https://doi.org/10.2307/2526310.

Lee, L.-F., and R. P. Trost. 1978. Estimation of some limited dependent variable models with application to housing demand. *Journal of Econometrics* 8: 357–382. https://doi.org/10.1016/0304-4076(78)90052-0.

Lokshin, M., and Z. Sajaia. 2004. Maximum likelihood estimation of endogenous switching regression models. *Stata Journal* 4: 282–289. https://doi.org/10.1177/1536867X0400400306.

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

Newey, W. K. 1985. Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53: 1047–1070. https://doi.org/10.2307/1911011.

Orme, C. 1995. Simulated conditional moment tests. *Economics Letters* 49: 239–245. https://doi.org/10.1016/0165-1765(95)00679-A.

Pfaffermayr, M. 2014. A GMM-based test for normal disturbances of the Heckman sample selection model. *Econometrics* 2: 151–168. https://doi.org/10.3390/econometrics2040151.

Poirier, D. J., and P. A. Ruud. 1981. On the appropriateness of endogenous switching. *Journal of Econometrics* 16: 249–256. https://doi.org/10.1016/0304-4076(81)90111-1.

Poirier, D. J., and J. L. Tobias. 2003. On the predictive distributions of outcome gains in the presence of an unidentified parameter. *Journal of Business & Economic Statistics* 21: 258–268. https://doi.org/10.1198/073500103288618945.

Skeels, C. L., and F. Vella. 1999. A Monte Carlo investigation of the sampling behavior of conditional moment tests in tobit and probit models. *Journal of Econometrics* 92: 275–294. https://doi.org/10.1016/S0304-4076(98)00092-X.

Tauchen, G. 1985. Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics* 30: 415–443. https://doi.org/10.1016/0304-4076(85)90149-6.

Vijverberg, W. P. M. 1993. Measuring the unidentified parameter of the extended Roy model of selectivity. *Journal of Econometrics* 57: 69–89. https://doi.org/10.1016/0304-4076(93)90059-E.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

About the authors

Giorgio Calzolari is professor of econometrics at the University of Firenze (Italy), Department of Statistics, Computer Science, Applications. His main previous appointment was at the IBM Scientific Center in Pisa (Italy), where he was a research staff member of the econometric team for twenty years. In 1989 he chaired the Econometrics Programme Committee of the European Meeting of the Econometric Society (ESEM'89, Muenchen). In 2012 he was President of the "Societa' Italiana di Econometria" (SIdE-IEA, Italian Econometric Association).

Maria Gabriella Campolo is associate professor in social statistics at the University of Messina, Department of Economics. Her research interests are mainly focused on the study and application of models for estimating the relationships between socio-economic and demographic phenomena, including the relationship between fertility, women labor supply, time-use and well-being.

Antonino Di Pino is professor in social statistics at the Department of Economics of the University of Messina (Italy). His research interests in statistical methodology are focused on causal inference problems and on models with censoring. He carried out applied researches on opportunity cost of children, on the economic evaluation of housework, and on the influence of demographic variables on the partners' participation to the labor market.

Laura Magazzini is associate professor of econometrics at the Institute of Economics and EMbeDS (Economics and Management in the era of Data Science), Sant'Anna School of Advanced Studies, Pisa (since August 2020). She previously was at the Department of Economics, University of Verona (Italy). Her research interests are centered around microeconometrics, industrial economics, the economics of innovation, competition policy and econometric methods, with particular reference to panel data analysis.

A Indirect identification of across-regime covariance in a two-regime switching model

As shown above in section 3, adopting the two-equation ML method, the across-regime covariance is identified and estimated simultaneously with the regression coefficients and errors variances. Unlike this approach, that of previous two-regime switching models with a selection equation, generally following a two-stage procedure (Heckman 1976, 1990; Lee 1978), provided only an indirect identification (and a "gross" estimation) of the across-regime covariance. In the applications proposed in section 4, we compare the estimates applying both our two-equation ML method (`mlcar` command) and the traditional two-stage estimation, which requires a selection equation. In the second case, the estimation of the across-regime correlation is obtained indirectly as in Lee and Trost (1978) and Vijverberg (1993).

In a two-regime switching model, the error terms u_{1i} and u_{2i} are assumed to be normally distributed with zero mean and variances equal to σ_1^2 and σ_2^2. From the censoring rule imposed on both outcome equations, we derive that y_{1i} and y_{2i} can be, respectively, observed if $v_i = u_{1i} - u_{2i} > -(\mathbf{x}'_{1i}\boldsymbol{\beta}_1 - \mathbf{x}'_{2i}\boldsymbol{\beta}_2)$ or $v_i = u_{2i} - u_{1i} \geq -(\mathbf{x}'_{2i}\boldsymbol{\beta}_2 - \mathbf{x}'_{1i}\boldsymbol{\beta}_1)$, where the random variable v_i is normally distributed with zero mean and variance σ_v^2.

Then, the random variable v_i/σ_v is distributed as a standard normal. In this way, reparameterizing as $(\mathbf{x}'_{1i}\boldsymbol{\beta}_1 - \mathbf{x}'_{2i}\boldsymbol{\beta}_2)/\sigma_v = \mathbf{z}'_i\boldsymbol{\gamma}$, we obtain the linear predictions $\mathbf{z}'_i\widehat{\boldsymbol{\gamma}}$ of the choice of the regime (according to the censoring rule) by running a probit regression on the selection equation.

Hence, we can obtain an indirect estimation of the covariance σ_{12} estimating preliminarily σ_v^2. In doing this, we use the predicted values of the selection equation, $\mathbf{z}'_i\widehat{\boldsymbol{\gamma}}$, and of both outcome equations, $\mathbf{x}'_{1i}\widehat{\boldsymbol{\beta}}_1$ and $\mathbf{x}'_{2i}\widehat{\boldsymbol{\beta}}_2$.

To estimate σ_v^2, we first consider the sample composition $n = n_1 + n_2$ with n_1 observations under regime 1 and n_2 observations under regime 2. Then, given n_1 row vectors $\mathbf{x}'_{1i}$ in the regressors matrix of regime 1, n_2 row vectors $\mathbf{x}'_{2i}$ in the regressors matrix of regime 2, and n row vectors $\mathbf{z}'$ in the regressors matrix of the selection equation, we have

$$\left(\mathbf{x}'_i\widehat{\boldsymbol{\beta}}_1 - \mathbf{x}'_i\widehat{\boldsymbol{\beta}}_2\right)/\widehat{\sigma}_v = \mathbf{z}'_i\widehat{\boldsymbol{\gamma}} \quad \text{where } \mathbf{x}'_i = [\mathbf{x}'_{1i} \quad \mathbf{x}'_{2i}]$$

and

$$\widehat{\sigma}_v^2 = \sum_{i=1}^{n}\left(\mathbf{x}'_i\widehat{\boldsymbol{\beta}}_1 - \mathbf{x}'_i\widehat{\boldsymbol{\beta}}_2\right)^2 / \sum_{i=1}^{n}(\mathbf{z}'_i\widehat{\boldsymbol{\gamma}})^2 \tag{4}$$

Then, estimating $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ by the outcome equations and computing $\widehat{\sigma}_v^2$ by (4), we obtain, through the well-known moment relationship $\sigma_v^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$, an estimate of the cross-covariance $\widehat{\sigma}_{12}$ and of the cross-correlation parameter, $\widehat{\rho}_{12} = \widehat{\sigma}_{12}/(\widehat{\sigma}_1\widehat{\sigma}_2)$.

B Heckman two-stage estimation results

We show below the results of the Heckman two-stage estimation applied to the three examples of two-regime models exposed in section 4.4. In doing this, we describe more in detail the procedure, using the Stata command, to obtain the indirect `rho12` estimation as explained in appendix A.

B.1 Example 1

```
. use https://www.stata.com/data/jwooldridge/eacsap/fringe, clear
. generate lannhrs_noff = lannhrs
. replace lannhrs_noff = 0 if office == 1
(298 real changes made)
. label variable lannhrs_noff "log(annual hours worked) if no office worker"
. generate lannhrs_off = lannhrs
. replace lannhrs_off = 0 if office == 0
(318 real changes made)
. label variable lannhrs_off "log(annual hours worked) if office worker"
. probit union lannhrs exper expersq office educ male annbens
Iteration 0:   log likelihood = -385.30268
Iteration 1:   log likelihood = -286.00449
Iteration 2:   log likelihood = -285.21203
Iteration 3:   log likelihood = -285.21175
Iteration 4:   log likelihood = -285.21175
Probit regression                               Number of obs     =        616
                                                LR chi2(7)        =     200.18
                                                Prob > chi2       =     0.0000
Log likelihood = -285.21175                     Pseudo R2         =     0.2598
```

union	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lannhrs	-1.273315	.2873008	-4.43	0.000	-1.836414	-.7102159
exper	.0175563	.019202	0.91	0.361	-.0200789	.0551914
expersq	-.0002715	.0004088	-0.66	0.507	-.0010728	.0005298
office	-1.29563	.1505457	-8.61	0.000	-1.590694	-1.000566
educ	-.060494	.0256906	-2.35	0.019	-.1108466	-.0101414
male	-.1288234	.1471372	-0.88	0.381	-.4172069	.1595602
annbens	.0005211	.0000575	9.06	0.000	.0004083	.0006339
_cons	9.37351	2.185413	4.29	0.000	5.09018	13.65684

```
. predict lin_pred, xb
. generate lin_predsq = lin_pred^2
. generate mills1 = normalden(-lin_pred) / (1 - normal(-lin_pred))
. generate mills2 = -normalden(-lin_pred) / (normal(-lin_pred))
```

```
. regress lannearn lannhrs_off lannhrs_noff exper expersq male annbens mills1 if
> union == 1
      Source |       SS       df       MS              Number of obs =     196
-------------+------------------------------           F(7, 188)     =   40.88
       Model |  19.9142816     7  2.84489738           Prob > F      =  0.0000
    Residual |  13.0820799   188  .069585532           R-squared     =  0.6035
-------------+------------------------------           Adj R-squared =  0.5888
       Total |  32.9963616   195  .169212111           Root MSE      =  .26379

------------------------------------------------------------------------------
    lannearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 lannhrs_off |   .3293442   .1711405     1.92   0.056    -.0082582    .6669467
lannhrs_noff |    .322118   .1565905     2.06   0.041     .0132178    .6310182
       exper |   .0187162   .0069879     2.68   0.008     .0049313     .032501
     expersq |  -.0003416   .0001414    -2.42   0.017    -.0006205   -.0000627
        male |   .2702128    .052182     5.18   0.000     .1672754    .3731503
     annbens |   .0002021   .0000482     4.19   0.000      .000107    .0002972
      mills1 |   .1178812    .154788     0.76   0.447    -.1874633    .4232257
       _cons |   5.998155   1.043655     5.75   0.000     3.939376    8.056935
------------------------------------------------------------------------------

. matrix b = e(b)

. generate predict1 = b[1,1]*lannhrs_off + b[1,2]*lannhrs_noff + b[1,3]*exper +
> b[1,4]*expersq + b[1,5]*male + b[1,6]*annbens + b[1,8]

. generate sigma11 = e(rss)/e(df_r)

. regress lannearn lannhrs exper expersq office educ male mills2 if union == 0
      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(7, 412)     =  106.70
       Model |  114.357123     7  16.3367319           Prob > F      =  0.0000
    Residual |  63.0833638   412  .153114961           R-squared     =  0.6445
-------------+------------------------------           Adj R-squared =  0.6384
       Total |  177.440487   419   .42348565           Root MSE      =   .3913

------------------------------------------------------------------------------
    lannearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lannhrs |   1.289579   .0848471    15.20   0.000     1.122792    1.456366
       exper |   .0052416   .0056803     0.92   0.357    -.0059245    .0164076
     expersq |  -.0000341   .0001204    -0.28   0.777    -.0002707    .0002026
      office |   .5116217   .0526985     9.71   0.000     .4080302    .6152132
        educ |   .0581582   .0082158     7.08   0.000     .0420081    .0743083
        male |   .3288602   .0442127     7.44   0.000     .2419496    .4157707
      mills2 |  -.8607312   .0895085    -9.62   0.000    -1.036682   -.6847808
       _cons |  -2.258669   .6512697    -3.47   0.001    -3.538895   -.9784428
------------------------------------------------------------------------------

. matrix b = e(b)

. generate predict2 = b[1,1]*lannhrs + b[1,2]*exper + b[1,3]*expersq +
> b[1,4]*office + b[1,5]*educ + b[1,6]*male + b[1,8]

. generate sigma22 = e(rss)/e(df_r)

. generate diff_pred = (predict1 - predict2)^2
```

```
. total diff_pred lin_predsq
Total estimation                  Number of obs   =       616

             |      Total   Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------
   diff_pred |   186.4592   9.075711       168.636    204.2823
  lin_predsq |   714.9428   34.21549      647.7495    782.1362

. matrix b = e(b)
. generate sigma_diff = b[1,1] / b[1,2]
. generate sigma12 = ((sigma11 + sigma22) - sigma_diff) /2
. generate rho12 = sigma12/ sqrt(sigma11 * sigma22)
. display rho12
-.18456715
```

B.2 Example 2

```
. use http://www.stata.com/data/jwooldridge/eacsap/401ksubs, clear
. generate inc_percap = inc/fsize
. label variable inc_percap "=inc/fsize"
. generate marr_pira = marr*pira
. label variable marr_pira "=married*IRA"
. generate nonmarr_pira = (1-marr)*pira
. label variable nonmarr_pira "=(1-married)*IRA"
. describe nettfa p401k inc_percap age agesq marr_pira nonmarr_pira
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
nettfa          float   %9.0g                 net total fin. assets, $1000
p401k           byte    %9.0g                 =1 if participate in 401(k)
inc_percap      float   %9.0g                 =inc/fsize
age             byte    %9.0g                 age^2
agesq           int     %9.0g                 age^2
marr_pira       float   %9.0g                 =married*IRA
nonmarr_pira    float   %9.0g                 =(1-married)*IRA
```

```
. probit p401k inc_percap age agesq nonmarr_pira marr_pira
Iteration 0:   log likelihood = -5466.2574
Iteration 1:   log likelihood = -5212.3587
Iteration 2:   log likelihood = -5211.9516
Iteration 3:   log likelihood = -5211.9516
Probit regression                               Number of obs     =      9,275
                                                LR chi2(5)        =     508.61
                                                Prob > chi2       =     0.0000
Log likelihood = -5211.9516                     Pseudo R2         =     0.0465

------------------------------------------------------------------------------
       p401k |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  inc_percap |   .0155672   .0010476    14.86   0.000      .013514    .0176205
         age |   .0946554   .0117114     8.08   0.000     .0717015    .1176094
       agesq |  -.0011147   .0001347    -8.27   0.000    -.0013788   -.0008507
nonmarr_pira |   .0195348   .0576982     0.34   0.735    -.0935516    .1326212
   marr_pira |   .4367102   .0364846    11.97   0.000     .3652017    .5082188
       _cons |  -2.862799   .2453548   -11.67   0.000    -3.343686   -2.381913
------------------------------------------------------------------------------

. predict lin_pred, xb
. generate lin_predsq = lin_pred^2
. generate mills1 = normalden(-lin_pred) / (1 - normal(-lin_pred))
. generate mills2 = -normalden(-lin_pred) / (normal(-lin_pred))
. regress nettfa inc_percap age agesq nonmarr_pira marr_pira  mills1 if p401k == 1
      Source |       SS           df       MS      Number of obs   =     2,562
-------------+----------------------------------   F(6, 2555)      =    106.93
       Model |  3229982.69         6  538330.449   Prob > F        =    0.0000
    Residual |  12863094.1     2,555  5034.47909   R-squared       =    0.2007
-------------+----------------------------------   Adj R-squared   =    0.1988
       Total |  16093076.8     2,561  6283.90346   Root MSE        =    70.954

------------------------------------------------------------------------------
      nettfa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  inc_percap |   7.150479   1.115339     6.41   0.000     4.963418     9.33754
         age |   41.48615   7.255142     5.72   0.000     27.25959    55.71271
       agesq |  -.4718318   .0853272    -5.53   0.000    -.6391493   -.3045144
nonmarr_pira |   24.76967    5.81559     4.26   0.000     13.36592    36.17341
   marr_pira |    219.646   32.53806     6.75   0.000     155.8424    283.4497
      mills1 |    553.734   103.5124     5.35   0.000     350.7573    756.7107
       _cons |  -1681.013   297.3186    -5.65   0.000    -2264.023   -1098.003
------------------------------------------------------------------------------

. matrix b = e(b)
. generate predict1 = b[1,1]*inc_percap + b[1,2]*age + b[1,3]*agesq +
> b[1,4]*nonmarr_pira + b[1,5]*marr_pira + b[1,7]
. generate sigma11 = e(rss)/e(df_r)
```

```
. regress nettfa inc_percap age agesq nonmarr_pira marr_pira  mills2 if p401k == 0
      Source |       SS           df       MS      Number of obs   =     6,713
-------------+----------------------------------   F(6, 6706)      =    204.04
       Model |  3167500.94         6  527916.823   Prob > F        =    0.0000
    Residual |  17350402.8     6,706  2587.29537   R-squared       =    0.1544
-------------+----------------------------------   Adj R-squared   =    0.1536
       Total |  20517903.7     6,712  3056.89864   Root MSE        =    50.865

------------------------------------------------------------------------------
      nettfa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  inc_percap |  -1.130697   .3157658    -3.58   0.000    -1.749698   -.5116952
         age |  -11.31436   1.715672    -6.59   0.000    -14.67762   -7.951095
       agesq |    .140205   .0201568     6.96   0.000     .1006911    .1797188
nonmarr_pira |   18.78395   2.656777     7.07   0.000     13.57582    23.99208
   marr_pira |  -10.83976   8.637384    -1.25   0.210    -27.77178    6.092255
      mills2 |  -228.4327   37.94426    -6.02   0.000    -302.8155   -154.0499
       _cons |   142.4178   24.51337     5.81   0.000      94.3638    190.4718
------------------------------------------------------------------------------

. matrix b = e(b)

. generate predict2 = b[1,1]*inc_percap + b[1,2]*age + b[1,3]*agesq +
> b[1,4]*nonmarr_pira + b[1,5]*marr_pira + b[1,7]

. generate sigma22 = e(rss)/e(df_r)

. generate diff_pred = (predict1 - predict2)^2

. total diff_pred lin_predsq

Total estimation                  Number of obs   =      9,275

--------------------------------------------------------------
             |      Total   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
   diff_pred |   3.21e+09   1.59e+07      3.18e+09    3.24e+09
  lin_predsq |    4460.02   30.90237      4399.444    4520.595
--------------------------------------------------------------

. matrix b = e(b)

. generate sigma_diff = b[1,1] / b[1,2]

. generate sigma12 = ((sigma11 + sigma22) - sigma_diff) /2

. generate rho12 = sigma12/ sqrt(sigma11 * sigma22)

. display rho12
-98.753319
```

B.3 Example 3

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/beauty, clear

. generate lwage2 = lwage

. summarize lwage2, det
  (output omitted)

. replace lwage2 = . if lwage<=r(p1)
(15 real changes made, 15 to missing)

. replace lwage2 = . if lwage>r(p99)
(12 real changes made, 12 to missing)

. generate collgrad =  educ>=12
```

```
. drop if lwage2 == .
(27 observations deleted)

. probit service exper expersq collgrad female black

Iteration 0:   log likelihood = -722.21193
Iteration 1:   log likelihood = -664.80736
Iteration 2:   log likelihood = -664.34963
Iteration 3:   log likelihood = -664.34934
Iteration 4:   log likelihood = -664.34934

Probit regression                               Number of obs     =      1,233
                                                LR chi2(5)        =     115.73
                                                Prob > chi2       =     0.0000
Log likelihood = -664.34934                     Pseudo R2         =     0.0801

------------------------------------------------------------------------------
     service |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |    .037829   .0132769     2.85   0.004     .0118068    .0638512
     expersq |  -.0007867   .0002999    -2.62   0.009    -.0013745    -.000199
    collgrad |   .5120305   .1138821     4.50   0.000     .2888257    .7352353
      female |   .7676342   .0841543     9.12   0.000     .6026948    .9325736
       black |   .0628345   .1556089     0.40   0.686    -.2421533    .3678223
       _cons |  -1.655462   .1683095    -9.84   0.000    -1.985343   -1.325581
------------------------------------------------------------------------------

. predict lin_pred, xb

. generate lin_predsq = lin_pred^2

. generate mills1 = normalden(-lin_pred) / (1 - normal(-lin_pred))

. generate mills2 = -normalden(-lin_pred) / (normal(-lin_pred))

. regress  lwage2 exper expersq collgrad female black mills1 if service == 1

      Source |       SS           df       MS      Number of obs   =       336
-------------+----------------------------------   F(6, 329)       =     14.45
       Model |  23.5045356         6  3.9174226   Prob > F        =    0.0000
    Residual |   89.193708       329  .271105496   R-squared       =    0.2086
-------------+----------------------------------   Adj R-squared   =    0.1941
       Total |  112.698244       335  .336412668   Root MSE        =    .52068

------------------------------------------------------------------------------
      lwage2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |     .14808   .0670659     2.21   0.028      .016148    .2800121
     expersq |  -.0029665   .0014009    -2.12   0.035    -.0057224   -.0002105
    collgrad |   1.833999    .941139     1.95   0.052    -.0174104    3.685408
      female |   1.905758   1.357603     1.40   0.161    -.7649205    4.576436
       black |    .109329   .1554572     0.70   0.482    -.1964864    .4151444
      mills1 |   4.024335   2.440734     1.65   0.100    -.7770794    8.825749
       _cons |  -6.962463   4.885576    -1.43   0.155    -16.57337    2.648445
------------------------------------------------------------------------------

. matrix b = e(b)

. generate predict1 = b[1,1]*exper + b[1,2]*expersq + b[1,3]*collgrad +
> b[1,4]*female + b[1,5]*black + b[1,7]

. generate sigma11 = e(rss)/e(df_r)
```

```
. regress lwage2 exper expersq collgrad female black mills2 if service == 0
      Source │       SS           df       MS      Number of obs   =       897
─────────────┼──────────────────────────────────   F(6, 890)       =     76.72
       Model │  85.9561851         6  14.3260308   Prob > F        =    0.0000
    Residual │   166.18766       890  .186727707   R-squared       =    0.3409
─────────────┼──────────────────────────────────   Adj R-squared   =    0.3365
       Total │  252.143845       896  .281410541   Root MSE        =    .43212

─────────────┬────────────────────────────────────────────────────────────────
      lwage2 │      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
─────────────┼────────────────────────────────────────────────────────────────
       exper │   .0486707   .0111441     4.37   0.000     .0267988    .0705425
     expersq │  -.0008191   .0002315    -3.54   0.000    -.0012735   -.0003648
    collgrad │   .3733219   .1352818     2.76   0.006     .1078133    .6388305
      female │  -.2798039   .2388327    -1.17   0.242    -.7485448    .1889369
       black │  -.0461786   .0605579    -0.76   0.446    -.1650315    .0726743
      mills2 │   .4667468   .6280203     0.74   0.458    -.7658266     1.69932
       _cons │   1.186469   .0571013    20.78   0.000       1.0744    1.298538
─────────────┴────────────────────────────────────────────────────────────────

. matrix b = e(b)

. generate predict2 = b[1,1]*exper + b[1,2]*expersq + b[1,3]*collgrad +
> b[1,4]*female + b[1,5]*black + b[1,7]

. generate sigma22 = e(rss)/e(df_r)

. generate diff_pred = (predict1 - predict2)^2

. total diff_pred lin_predsq

Total estimation                  Number of obs   =      1,233

─────────────┬────────────────────────────────────────────
             │      Total   Std. Err.     [95% Conf. Interval]
─────────────┼────────────────────────────────────────────
   diff_pred │   37957.01    458.185       37058.1    38855.92
  lin_predsq │   754.7499   19.83436       715.837    793.6627
─────────────┴────────────────────────────────────────────

. matrix b = e(b)

. generate sigma_diff = b[1,1] / b[1,2]

. generate sigma12 = ((sigma11 + sigma22) - sigma_diff) /2

. generate rho12 = sigma12/ sqrt(sigma11 * sigma22)

. display rho12
-110.74233
```

C Monte Carlo experiments on the mlcartestn procedure to test normality

Monte Carlo simulations allow us to evaluate the performance, in finite samples, of the proposed testing procedure (see section 3.1), implemented by the `mlcartestn` command. We based the experiments on a design similar to that previously used by Calzolari and Di Pino (2017) to check the properties of the two-equation ML estimator.

The simulated two-regime model is specified as follows:

$$y_{1i} = 2 + x_{1i} + u_{1i} \tag{5}$$

$$y_{2i} = 10 + 0.8x_{2i} + u_{2i} \tag{6}$$

The explanatory variables, x_{1i} and x_{2i}, are both generated from a normal distribution with mean 50 and variance 100. The error terms, u_{1i} and u_{2i}, are random variables with zero mean and variance, respectively, $\sigma_1^2 = 100$ and $\sigma_2^2 = 10$. The percentage of cases observed in each regime on the total of cases is symmetrically equal to 50%.

Then, to simulate the presence of a large cross-correlation, we set the across-regime correlation alternatively with positive ($\rho_{12} = 0.90$ and $\sigma_{12} = 28.4605$) and negative signs ($\rho_{12} = -0.90$ and $\sigma_{12} = -28.4605$). We also simulated estimation and testing performance by setting absence of across-regime correlation ($\rho_{12} = 0$).

We checked the performance of the testing procedure assuming normally distributed errors and, alternatively, accounting for some cases of misspecification given by the violation of the assumption of normality. To this end, we simulated error terms that deviate from the normal distribution in terms of higher kurtosis following Student t distributions with 9, 30, and 100 degrees of freedom, although the errors distributed as a Student t (100) reproduce the case in which the kurtosis is closer to the normality condition.

We also simulated the model whose error terms deviate from normality because of the presence of asymmetry. To this purpose, we generate error terms following a Skew Normal distribution (for example, Azzalini [1985]) with the Shape parameter, α, equal to 5 (generally involving a level of skewness close to 0.8–0.9).

Summing up, we simulate several data-generating processes (DGPs) based on (5) and (6) under different distributive assumptions on the errors, accounting for, respectively, positive, negative, and null cross-correlation between the errors of the two equations:

Covariance matrix under positive cross-correlation: ($\rho_{12} = 0.90$):

$$\mathbf{\Sigma}_{(u_{1i};u_{2i})} = \begin{pmatrix} 100 & 28.4605 \\ 28.4605 & 10 \end{pmatrix}$$

Covariance matrix under negative cross-correlation: ($\rho_{12} = -0.90$):

$$\mathbf{\Sigma}_{(u_{1i};u_{2i})} = \begin{pmatrix} 100 & -28.4605 \\ -28.4605 & 10 \end{pmatrix}$$

Covariance matrix in absence of cross-correlation: ($\rho_{12} = 0$):

$$\mathbf{\Sigma}_{(u_{1i};u_{2i})} = \begin{pmatrix} 100 & 0 \\ 0 & 10 \end{pmatrix}$$

In the following table 1, we report the simulation results, given by the means of the empirical test sizes obtained setting several DGPs, under different assumptions of the errors distribution.

Table 1. Empirical test size* of CM test of normality (`mlcartestn` command)

	Positive cross-correlation ($\rho_{12} = 0.9$)			
Sample:	$n = 500$	$n = 1000$	$n = 1500$	$n = 5000$
DGP1_Normal	0.0506	0.0502	0.0495	0.0398
DGP2_t(9)	0.3570	0.5666	0.6613	0.9889
DGP3_t(30)	0.0872	0.1154	0.1152	0.2000
DGP4_t(100)	0.0446	0.0444	0.0401	0.0370
DGP5_Sk_Norm($\alpha = 5$)	0.7614	0.9776	0.9969	1.0000
	Negative cross-correlation ($\rho_{12} = -0.9$)			
Sample:	$n = 500$	$n = 1000$	$n = 1500$	$n = 5000$
DGP1_Normal	0.0481	0.0635	0.0678	0.0655
DGP2_t(9)	0.3396	0.5252	0.6289	0.9830
DGP3_t(30)	0.0951	0.1189	0.1770	0.2856
DGP4_t(100)	0.0439	0.0594	0.0403	0.0630
DGP5_Sk_Norm($\alpha = 5$)	0.0738	0.1479	0.2735	0.8658
	Absence of cross-correlation ($\rho_{12} = 0$)			
Sample:	$n = 500$	$n = 1000$	$n = 1500$	$n = 5000$
DGP1_Normal	0.0625	0.0532	0.0544	0.0475
DGP2_t(9)	0.4976	0.7305	0.9007	1.0000
DGP3_t(30)	0.1357	0.1648	0.1772	0.4585
DGP4_t(100)	0.0638	0.0748	0.0594	0.0691
DGP5_Sk_Norm($\alpha = 5$)	0.6255	0.9406	0.9879	1.0000

NOTES: Nominal test size: 5%. No of reps = 1000
* Proportion of cases in which the null hypothesis of normality is rejected.

The results reported in table 1 show that the CM test, implemented with the command `mlcartestn`, with the `sim(100)` option, allows us to detect misspecification given by the departure from the normality assumption because of an excess of kurtosis or skewness. Note that as in the cases in which the null hypothesis is expected to be rejected because of misspecification [being the errors distributed as Student $t(9)$, Student $t(30)$, and skew-normal($\alpha = 5$)], the share of rejections approaches 100% as the sample dimension increases. Note also that the empirical test size performs better in the cases in which DGPs are simulated assuming positive or null cross-correlation between the errors.

If we simulate DGPs following normal or Student $t(100)$ distributions, the results of empirical test size are consistent to the nominal size fixed for the rejection of the null hypothesis of normality.