UNIVERSITÀ DEGLI STUDI FIRENZE
DISIA
DIPARTIMENTO DI STATISTICA, INFORMATICA, APPLICAZIONI "GIUSEPPE PARENTI"

Self-Selection and Direct Estimation of Across-Regime Correlation Parameter

Giorgio Calzolari, Antonino Di Pino

DISIA WORKING PAPER
2014/04

Self-Selection and Direct Estimation of Across-Regime Correlation Parameter

Giorgio Calzolari and Antonino Di Pino

Abstract A direct Maximum Likelihood (ML) procedure to estimate the "generally unidentified" across-regime correlation parameter in a two-regime endogenous switching model is here provided. The results of a Monte Carlo experiment confirm consistency of our direct ML procedure, and its relative efficiency over widely applied models and methods. As an empirical application, we estimate a Two-Regime simultaneous equation model of domestic work of Italian married women in which the two regimes are given by their working status (employed or unemployed).

Key words: Endogenous switching model, Across-Regime correlation parameter

JEL classifications: C31, C34, J22

1 Introduction

We consider a simultaneous Two-Equation model (Roy model with two regimes), in which a non null correlation between the error terms may occur as a consequence of the joint influence of latent factors on the outcome gained by the subject in the chosen regime. However, this correlation (or covariance) across regimes is not empirically identifiable as a result of the selection criterion, involving that both dependent variables cannot be jointly observed. Nevertheless, "some knowledge" of this parameter is considered relevant to provide information about the agents' behaviour in a two-regime switching model (Vijverberg,1993) and to obtain predicted out-of-sample distribution of the outcome gains (Poirier and Tobias, 2003).

The aim of this study is to suggest a simple selection criterion in a two-equation switching model that permits identification and "direct *ML* estimation also" of the across-regime correlation. The model simply specifies two outcome regression equations plus the straightforward selection rule that the larger outcome is chosen and observed, while the smaller is latent. It is therefore a sort of

Giorgio Calzolari
Università di Firenze, DISIA "G. Parenti", Viale Morgagni, 59, Firenze, email: calzolar@disia.unifi.it

Antonino Di Pino
Università di Messina, Dipartimento S.E.A.M. Via T. Cannizzaro, 278, Messina, email: dipino@unime.it
(We thank the financial support of the project MIUR PRIN MISURA - Multivariate Models for Risk Assessment)

"two simultaneous censored equations" with endogenous censoring, endogeneity being due to the across-equation correlation. For each individual, the contribution to the likelihood is given by the probability density of the observed outcome (the larger) and by the (conditional) probability that the alternative outcome has a smaller value: besides coefficients and variances, the (Gaussian) likelihood includes therefore also the across-equation correlation. Model and estimation method will be called *Two-Equation ML* in the following.

Other widely used approaches add a stochastic selection equation to the two outcome equations. Two of them are compared in this paper with the proposed *Two-Equation ML*: the Maximum Likelihood described in Poirier and Ruud (1981) and Maddala (1983 and 1986), hereafter called *Three-Equation ML*; the Two Stage approach (Heckman, 1976 and 1990; Lee, 1978) hereafter called *TS Heckman.* In these cases, estimates of the correlation parameter are possible only "indirectly", after coefficients and variances have been computed, applying the relationships among the errors' second-order moments (Maddala, 1983 pp. 223-228).

Other approaches would also be available, but will not be considered explicitly in this paper. For instance, assuming that across-correlation is determined by a latent factor (e.g. the unobserved individual ability) common to both outcome equations and to a third selection equation, the estimation procedure would apply factor analysis methods. However, since individuals cannot be observed jointly in both regimes, a distribution of "counterfactuals" should be preliminarily provided (e.g., Carneiro et al., 2003; Aakvik et al. 2005).

Using simulated data, we evidence the better performance of our *Two-Equation ML* estimator over the traditional methods that include a third selection equation, where the across regime correlation cannot be estimated directly: the *Three-Equation ML* and the *TS Heckman* method. Performances are first compared by Monte Carlo in a "correct specification context", where simulated data are produced from the normal distribution, coherently with the Gaussian likelihoods adopted. Then, comparison is also performed in a "misspecification context", where simulated data are produced from a "heavy-tails" Student-*t*.

We provide also an empirical application to compare the methods. We estimate the time devoted to domestic work by the Italian (married or cohabiting) women under two different regimes, given by their working status: employed or unemployed. The source of data is given by the cross sectional ISTAT (Italian National Institute of Statistics) Survey on Time Use in Italy in the years 2002-2003. Estimation results reveal that a strong positive correlation across the two regimes of employed and unemployed women occurs. This result can be interpreted in the sense that the ability in doing housework is identical for employed and unemployed women.

The paper is organized as follows: In the next section, we discuss about the specification of two-regime model and we explain the rationale of our methodology. In Sect. 3, the specification of our

Two-Equation ML estimator is provided. In Sect. 4, the results of simulations, based on the use of *Three-Equation ML* (Poirier and Ruud, 1981), *TS Heckman* and our *Two-Equation ML* procedures, respectively, are discussed. In Sect. 5 the results of the empirical application are reported. In Sect. 6 we conclude with final remarks.

2 Methodological issues

We consider a simultaneous Two-Equation Roy model with two regimes (Roy, 1951). This model is specified by two regression equations whose dependent variables (outcomes) are excluding each other in a cross-sectional framework, and where selection is simply based on the choice of the larger outcome[2]:

$y_{1i} = \mathbf{x'}_{1i}\,\boldsymbol{\beta}_1 + u_{1i}$ $y_{2i} = \mathbf{x'}_{2i}\,\boldsymbol{\beta}_2 + u_{2i}$ 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.	(1)

Specific target of this study is to estimate the correlation between u_{1i} and u_{2i} (ρ_{12}). A non-zero correlation between the error terms may occur as a consequence of the joint influence of latent factors. This correlation across regimes is not "empirically identifiable", as both dependent variables cannot be jointly observed.

model (1) is widely discussed by analysts (e.g. Maddala, 1983 and 1986; Heckman and Honoré, 1990, Vella and Verbeek, 1999). Surprisingly, however, it is not usual in the literature to directly tackle estimation of the model exactly as it is specified in the above model's (1) equations, whose parameters are coefficients, variances, "and" the across regimes correlation. Rather, estimation is usually performed after transformation of the two equations into a Three-Equation model, with the inclusion of a third equation to select between the two regimes (e.g. Lee, 1978; Heckman, 1990)[3].

[2] A further assumption is that observations can be classified in those belonging to the first regime and to the second regime, as in a "sample selection known" framework.

[3] A relevant exception is represented by the *ML* approach to estimate simultaneously demand and supply equations in a disequilibrium model, where the observations belong, respectively, to a demand or to a supply function (see, among others, Maddala and Nelson, 1974). The choice of the regime does not depend on a third selection equation, but the likelihood function specified in this model, including coefficients and variances, does not include the across-regime covariance.

$$
\begin{aligned}
& y_{1i} = \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + u_{1i} \qquad \text{if } L_i = 1;\ \text{otherwise latent} \\
& y_{2i} = \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + u_{2i} \qquad \text{if } L_i = 0;\ \text{otherwise latent} \\
& L_i^{*} = \mathbf{z}'_i\boldsymbol{\gamma} + \eta_i \qquad (2) \\
& \begin{cases} L_i = 1 & \text{if } L_i^{*} > 0 \\ L_i = 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

The selection rule is no longer "deterministic" (as it is in model 1, where the larger is chosen and observed, the smaller is latent), but becomes stochastic. As in model (1), the error terms u_{1i} and u_{2i} are normally distributed with zero mean and variances equal to σ^2_1 and σ^2_2. The disturbance term of the selection equation η_i is assumed $N(0,1)$, while the covariances $\sigma_{1\eta}$ and $\sigma_{2\eta}$ with the disturbances of both outcome equations can be different from zero. Differently from model (1), ρ_{12} does not appear in the (Gaussian) Likelihood.

In principle, model (2) has some important advantages over model (1). First of all, it has a high degree of generality in the choice of the variables that "explain" selection; they can be the same, or can be different from the explanatory variables included in $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$. If the explanatory variables of the selection equation include all the variables in $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ (without duplications), then estimates of model's (2) parameters can be consistent even if data have been generated as assumed in model (1); thus it is in some sense legitimate to ignore that one is facing a problem of misspecification. Second, the variables in $\mathbf{x}_{1i}$ need to be observed only when y_{1i} is observed, and not for all individuals i (analogously for $\mathbf{x}_{2i}$). Third, estimation of model (2) is surely easier; also, Heckman's Two Stage method and Maximum Likelihood are implemented in well known and widely used software packages[4] .

As anticipated, however, there is an important disadvantage in estimating model (2) when the data generating process is assumed to be as in model (1): the (Gaussian) Likelihood of model (2) does not include the correlation (or covariance) between the error terms of the two outcome equations, u_{1i} and u_{2i}. Estimates of this parameter can be obtained only indirectly, applying the relationships among the errors' second-order moments, after model (2) has been estimated (by *ML* or *Two-Stage* procedures, see Appendix 2 for details).

Assuming that outcome can always be observed in one of the two regimes, model (1) can be theoretically specified as a switching regression model with "sample separation known" (cf. Maddala, 1986 for a survey). The agent is assumed to compare the outcomes of the two equations, and to choose the larger (or the smaller, depending on the problem); thus, the larger between the two dependent variables is observed, the smaller is latent, but its value is "upper bounded" by the observed one. The model is therefore a sort of "two simultaneous censored equations" with endogenous censoring. For each individual, the contribution to the likelihood is given by the

[4] e.g. Stata routine "Movestay" provided by Lokshin and Sajaia (2004)

probability density of the observed variable (the larger) and by the (conditional) probability that the other variable has a smaller value: besides coefficients and variances, the (Gaussian) likelihood includes therefore also the across-equation correlation. With respect to other *ML* or Two-Stage (*TS*) methods (such as the *TS Heckman* or Control Function), there is no additional stochastic equation to select between the two regimes.

3 *Two-Equation ML* Estimator

With respect to the traditional Roy model, based on the assumption that the same regressors are included in each regime, in model (1) the vectors $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ can include different regressors as well as the same regressors in each regime. However, as a consequence of the specification of the likelihood function, a regressor included in only one of the two regimes must also be "observable" in the other regime (where the corresponding y_i is latent). The error terms u_{1i} and u_{2i} are jointly normally distributed with zero means, variances σ^2_1 and σ^2_2, and across-regime covariance of errors, σ_{12}, that may be different from zero. The specification of our likelihood function is based on the probability of a subject to gain the outcome of the two regimes; it is the probability density of the observed variable, multiplied by the conditional probability that the other variable (latent) is smaller than the observed variable.

Censoring rule in model (1) implies that:

$$\begin{aligned} y_{1i}\, observed \quad &\Rightarrow y_{2i} < y_{1i} \Rightarrow \mathbf{x'}_{2i}\boldsymbol{\beta}_2 + u_{2i} < \mathbf{x'}_{i1}\boldsymbol{\beta}_1 + u_{1i} \\ y_{2i}\, observed \quad &\Rightarrow y_{1i} \le y_{2i} \Rightarrow \mathbf{x'}_{1i}\boldsymbol{\beta}_1 + u_{1i} \le \mathbf{x'}_{2i}\boldsymbol{\beta}_2 + u_{2i} \end{aligned} \tag{3}$$

Hence:

$$\begin{aligned} \phi(y_{1i})P(y_{2i} < y_{1i}) &= \phi(u_{1i})P\left(u_{2i} < \mathbf{x'}_{1i}\boldsymbol{\beta}_1 - \mathbf{x'}_{2i}\boldsymbol{\beta}_2 + u_{1i} \middle| y_{1i} observed\right) \\ \phi(y_{2i})P(y_{1i} \le y_{2i}) &= \phi(u_{2i})P\left(u_{1i} \le \mathbf{x'}_{2i}\boldsymbol{\beta}_2 - \mathbf{x'}_{1i}\boldsymbol{\beta}_1 + u_{2i} \middle| y_{2i} observed\right) \end{aligned} \tag{4}$$

where $\phi()$ is a normal probability density function. y_{2i} and y_{1i} result censored, respectively, if:

$$u_{2i} < \mathbf{x'}_{1i}\boldsymbol{\beta}_1 - \mathbf{x'}_{2i}\boldsymbol{\beta}_2 + u_{1i} \middle| y_{1i} observed$$

and (5)

$$u_{1i} \le \mathbf{x'}_{2i}\boldsymbol{\beta}_2 - \mathbf{x'}_{1i}\boldsymbol{\beta}_1 + u_{2i} \middle| y_{2i} observed$$

Then, the Likelihood function is given by:

$$\mathcal{L}\left(\sigma_1^2,\sigma_2^2,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\sigma_{12}\right)= \prod_{\substack{y_2\\ observed}} \phi(u_2)\Phi\left(u_{1i} \le \mathbf{x'}_{2i}\,\boldsymbol{\beta}_2 - \mathbf{x'}_{1i}\,\boldsymbol{\beta}_1 + u_{2i} | y_{2i} observed\right)$$
$$\cdot \prod_{\substack{y_1\\ observed}} \phi(u_{1i})\Phi\left(u_{2i} < \mathbf{x'}_{1i}\,\boldsymbol{\beta}_1 - \mathbf{x'}_{2i}\,\boldsymbol{\beta}_2 + u_{1i} | y_{1i} observed\right) \tag{6}$$

Considering the bivariate normal distribution of the error terms, $u_{2i} = y_{2i} - \mathbf{x'}_{2i}\,\boldsymbol{\beta}_2$ and $u_{1i} = y_{1i} - \mathbf{x'}_{1i}\,\boldsymbol{\beta}_1$, we obtain the following log-likelihood function:

$$ln\,\mathbf{L} = -\frac{(\,y_{1i} - \mathbf{x'}_{1i}\,\boldsymbol{\beta}_1\,)^2}{2\sigma_1^2} - \frac{1}{2} ln\,\sigma_1^2 + ln\,\Phi\left(\frac{(\,y_{i1} - \mathbf{x'}_{2i}\,\boldsymbol{\beta}_2\,) - \dfrac{\sigma_{12}}{\sigma_1^2}(\,y_{1i} - \mathbf{x'}_{1i}\,\boldsymbol{\beta}_1\,)}{\sqrt{\sigma_2^2 - \sigma_{12}^2 \big/ \sigma_1^2}}\right)$$

$$-\frac{(\,y_{2i} - \mathbf{x'}_{2i}\,\boldsymbol{\beta}_2\,)^2}{2\sigma_2^2} - \frac{1}{2} ln\,\sigma_2^2 + ln\,\Phi\left(\frac{(\,y_{2i} - \mathbf{x'}_{1i}\,\boldsymbol{\beta}_1\,) - \dfrac{\sigma_{12}}{\sigma_2^2}(\,y_{2i} - \mathbf{x'}_{2i}\,\boldsymbol{\beta}_2\,)}{\sqrt{\sigma_1^2 - \sigma_{12}^2 \big/ \sigma_2^2}}\right) \tag{7}$$

where Φ() is the standard normal cumulative distribution function used to specify the contribution to the likelihood of censoring y_{1i} or y_{2i}.

Note that, using this *Two-Equation ML* procedure, no limitation in model specification should be adopted if we do not include some regressors in the specification of one of the two equations, provided that these regressors should be observable under both regimes to ensure the identification of the probability of censoring in both regimes (see, above, Eqs. (4) and (5))

Applying this *Two-Equation ML* procedure, the parameter σ_{12} (or ρ_{12}) can be directly estimated under the assumption of endogenous selection. This parameter, in particular, measures the correlation in unobserved productivity (ability) between the two regimes (or sectors) (e.g. Heckman and Honoré, 1990).

4 Monte Carlo Results

A Monte Carlo experiment allows to compare inferential properties of our *Two-Equation ML* estimator of the across-regime correlation, with the estimators obtained "indirectly", after applying Poirier-Ruud (1981) *Three–Equation ML* or the *TS Heckman* methods.

Experiments are conducted under Normal and Student-*t* (dof: 5) distributional assumptions, respectively. The application to a Student-*t* error distribution is particularly interesting because, among various departures from normality occurring in practice, one of the most common is when the distribution of the data has heavier tails than the normal distribution.

The data generating process here considered is represented as model (1), characterized by the inclusion in the right hand sides of a single regressor variable (plus the constant). Slope and intercept coefficients are the same for both equations. Then, in order to simulate the presence of a large cross correlation, we set the across-regime correlation, alternatively, with positive (ρ_{12}= 0.90) and negative sign (ρ_{12}= -0.90). We simulate also estimators performance setting absence of across regime correlation (ρ_{12}= 0). To estimate with *Three–Equation ML* and *TS Heckman* methods, a third *Probit* selection equation is included, where the explanatory variables are the variables of the two equations of model (1) (without duplications: thus the two exogenous variables plus the constant). Consequently, taking into account the meaning of the relationships between the errors' second-order moments of outcome and selection equations (cf. Heckman and Honoré, 1990; and Vijverberg, 1993), $\sigma^2_1 = \sigma^2_2$ implies null correlation between outcome and selection equation (absence of "endogenous selection"). Instead, when σ^2_1 and σ^2_2 are different (our results are obtained setting $\sigma^2_1 = 4\sigma^2_2$) a nonzero correlation between the outcome and the selection equation occurs (thus we may talk of "endogenous selection"). We use mean bias and root mean-square error (*RMSE*) to compare the performance of the estimators. The percentage of cases observed in each regime on the total of cases is symmetrically equal to 50%.

We summarize in Tables 1 and 2 the results of the estimated correlation parameters only (detailed estimation results on all parameters are reported in Appendix 1). With respect to other methods, the *Two-Equation ML* procedure generally provides much more efficient estimates for both regression coefficients and across-regime correlation parameter. In addition, the "direct" estimate of the across-regime correlation, provided by the *Two-Equation ML* method, appears to be always consistent, while *TS-Heckman* method fails when ρ_{12}= 0.90 and absence of endogeneous selection is imposed. If the error distribution is normal, we can observe (see Tables A1 and A2 in Appendix 1) how the bias in the estimated coefficients of both regression equations is substantially negligible for all the three estimation methods, but our *Two-Equation ML* procedure performs better than *Three-Equation ML* (*Poirier-Ruud*) and *TS Heckman* methods in terms of relative efficiency.

If we consider error terms distributed like a Student-*t* (5 d.o.f.), the bias in estimates are negligible adopting *Two-Equation ML* when positive and negative values of across-regime correlation are set. The *Two-Equation ML* procedure performs better than other methods as, for example, the *TS-Heckman* method. In particular, the latter produces non-negligible bias in across-regime correlation estimates (Tables A3 and A4 in Appendix 1). However, if across-regime correlation is imposed as null, estimates of ρ_{12} results generally biased using all the procedures, even if the estimated values of ρ_{12} result to be non significantly different from zero.

Table 1: Across-Regime correlation parameter (ρ_{12}) estimation: Simulation results - **Errors Distribution: Bivariate Normal**; sample: 10000; Monte Carlo Replications: 10000

	Two-Equation ML		***TS Heckman***		***Three-Equation ML (Poirier-Ruud)***	
ρ_{12} = 0.90 (positive)	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$\sigma^2_1 = \sigma^2_2$ *Absence of endogenous selection*	0.9000	*0.0048*	0.7998	*0.0070*	0.8999	*0.0049*
$\sigma^2_1 = 4\sigma^2_2$ *Endogenous selection*	0.9000	*0.00002*	0.8993	*0.0003*	0.9007	*0.0002*
ρ_{12}= -0.90 (negative)	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$\sigma^2_1 = \sigma^2_2$ *Absence of endogenous selection*	-0.9002	*0.0068*	-0.8993	*0.0597*	-0.8985	*0.0392*
$\sigma^2_1 = 4\sigma^2_2$ *Endogenous selection*	-0.9003	*0.00005*	-0.9016	*0.0050*	-0.8990	*0.0020*
ρ_{12}= 0	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$\sigma^2_1 = \sigma^2_2$ *Absence of endogenous selection*	-0.0001	*0.0010*	-0.0004	*0.0015*	-0.0002	*0.0013*
$\sigma^2_1 = 4\sigma^2_2$ *Endogenous selection*	0.0003	*0.0001*	-0.0047	*0.0005*	0.0005	*0.0002*

Table 2: Across-Regime correlation parameter (ρ_{12}) estimation: Simulation results - **Errors Distribution: Student-*t*** (dof: 5); sample: 10000; Monte Carlo Replications: 10000

	Two-Equation ML		***TS Heckman***		***Three-Equation ML (Poirier-Ruud)***	
ρ_{12} = 0.90 (positive)	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$\sigma^2_1 = \sigma^2_2$ *Absence of endogenous selection*	0.9001	*0.0063*	0.7989	*0.1015*	0.9004	*0.0064*
$\sigma^2_1 = 4\sigma^2_2$ *Endogenous selection*	0.9051	*0.0082*	0.9216	*0.0318*	0.8955	*0.0198*
ρ_{12}= -0.90 (negative)	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$\sigma^2_1 = \sigma^2_2$ *Absence of endogenous selection*	-0.9189	*0.0205*	-0.4780	*0.4252*	-0.860	*0.0637*
$\sigma^2_1 = 4\sigma^2_2$ *Endogenous selection*	-0.9239	*0.0250*	-0.4350	*0.4692*	-0.872	*0.0670*
ρ_{12}= 0	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$\sigma^2_1 = \sigma^2_2$ *Absence of endogenous selection*	-0.1520	*0.1630*	0.1225	*0.1286*	-0.0360	*0.0561*
$\sigma^2_1 = 4\sigma^2_2$ *Endogenous selection*	-0.1591	*0.1710*	0.1974	*0.2053*	-0.0243	*0.0580*

5 Empirical Application

We specify a simultaneous equation model to estimate the Italian women's supply of domestic work under two regimes: Employed women are included in the first regime (regime 1), while unemployed women are included in the second regime (regime 2), respectively. We specify an outcome equation in each of the two regimes; the dependent variable of both outcome equations is given by the logarithm of the time spent in domestic work by a married (or cohabiting) Italian woman. The selection rule for the *Two-Equation ML estimator* is: the lower time is selected and observed, the higher time is latent. For the *TS Heckman* and *Three-Equation ML* methods, a third selection equation (*probit*) is included, where the binary dependent *L* is 1 if the woman is employed, and 0 if the woman is unemployed. Besides the obvious interesting information deriving from coefficients and variances, the correlation parameter ρ_{12} deserves some interest. When $\rho_{12} > 0$ we have a "hierarchical structure", in which the two regime-specific earnings-abilities or skills are positively correlated. This implies that employed women working at home more than average would usually work at home more than average even if not employed. Alternatively, when $\rho_{12} < 0$

the two-regime earnings-abilities are negatively correlated, in the sense that employed and unemployed subjects may have different skills. This implies that employed women working at home less than average would usually work at home more than average if not employed. In this case, we have a "comparative advantage" structure of the model in which, on average, those who work more at home if he/she would be unemployed, would work less indoor if he/she would be employed, with respect to the other employed subjects.

The source is given by the cross-sectional dataset of the ISTAT Survey on Time Use in Italy in the years 2002-2003. In this survey the interviewed subjects provide detailed information on their own time use through the compilation of a diary in which they register all of their daily activities (measured in minutes). The selected sample is the same used in a recent study by Campolo and Di Pino (2012), and it is composed of 5698 Italian women, aged 18-60, living with their partners, and equitably distributed by area of residence and employment status (subjects who have retired by work are excluded). In particular, the employed women (Regime 1) are 3091, while the unemployed women (Regime 2) are 2607.

Explanatory variables of the model are described in the Table 3. The variables *Partime, Wage and Woman's work*, are surveyed under the Regime 1 only (employed women). For this reason, these variables are included as regressors only in the first outcome equation, but not in the second outcome equation (and not in the selection equation as well).

The results reported in Table 4 (first equation) and in Table 5 (second equation) show that, in general, marked differences do not emerge from the estimated coefficients obtained applying the three methods. Estimation results, common to all the three estimation procedures, show that the less educated women (both employed and unemployed) work more in the house. Moreover, the time devoted to domestic chores is greater for more aged women, especially if they are younger than their partner (variable *D_age*). We can observe also that the woman's domestic work decreases if she works in the market, and especially if her time spent working in the market increases with respect to the time spent in the market by her partner (variable *Work_gender gap*). A positive sign is reported by the coefficients of the regressors *Chcare_partner* and *Housework_partner*, in both regression equations. This indicates that the domestic chores and the activity of care generally lead to an increases of the commitment of both partners. Applying the *TS Heckman* method, a negative sign is obtained in the coefficient of the dummy-variable *Area* in the equation of employed women. This would be surprising (but the coefficient is not significantly different from zero): employed women would work less in the house in the (less-developed) southern regions. However, an opposite result (significantly non-zero) is obtained adopting both *Two-Equation ML* and *Three-Equation ML* estimators.

In general, the regressors' coefficients estimated by *Two-Equation ML* appear to be more significant than the estimates obtained by using the other methods[5]. The across-regime correlation, directly estimated using *Two-Equation ML* is equal to 0.96 and it results highly significant. Performing both *TS Heckman* and *Three-Equation ML* "indirect" estimation of the across regime correlation, we obtain, respectively, 0.71 and 0.46. The positive sign of this coefficient signals that common latent factors positively influence domestic work supply of both employed and unemployed women. As a consequence, if we assume that a relevant latent factor is given by the individual ability in domestic work, estimation results lead to conclude that employed and unemployed women have not different skill regarding to their commitment in domestic and care activity. This implies that Italian married women who are unemployed have not a comparative advantage in domestic activity with respect to employed women. However, the high (absolute) value of the across-regime correlation, ρ_{12}, directly estimated using *Two-Equation ML*, means that important explanatory variables are omitted by the specification of the regressors set of both equations. On the other hand, the negative sign of the *Lambda* coefficients ($\sigma_{1\eta}$ and $\sigma_{2\eta}$), estimated by *TS Heckman*, implies that participation in the labour market and the domestic work supply are negatively correlated as a consequence of the endogenous selection. This means that the commitment in housework and childcare discourages the woman to participate in the labour market.

6 Final Remarks

Our *Two-Equation ML* approach allows us to identify the across-regime correlation parameter, and to obtain a reliable point-estimation without introducing a further selection equation in the specification of the two-regime model.

Simulation results show that the *Two-Equation ML* procedure provides consistent estimates of across-regime correlation parameter, even in presence of endogenous selection. In general, the Monte Carlo experiments show that *Two-Equation ML* procedure performs better with respects to other methods, no matter if the assumption of normality of errors is introduced or relaxed.

As a result of the estimation of the domestic work supply of the Italian women, we obtain a large positive value of the across-regime correlation, that reveals how the attendance to housework and childcare of, respectively, employed and unemployed women is not affected by a different skill. This conclusion supports the thesis that employed women generally seek ways to maximize time

[5] Among the alternative ways of computing ML standard errors, the results here reported were obtained from the Hessian matrix (e.g. Calzolari and Panattoni, 1988)

devoted to children and domestic chores, as well as unemployed women (a similar result has been found by Bianchi, 2000, for US).

Table 3: *Italian Women's domestic work estimation: Description of variables*

Dependent variable:	
Logdom	Log of time spent for childcare and housework by a woman in a day
Regressors:	
Age	Age of woman
Children	No of children living in the household 0-13 years
Dummy_Weekend	Dummy Weekend: Reference day activities of the diary: Weekend=0; Mon.-Fri. =1.
Area	Dummy: Area of residence (Southern Regions=1; North-Centre=0)
Education	Woman education (years of schooling)
D_age	Age of woman - age of partner (man)
Chcare_partner	Log of time spent for childcare by the partner in a day
Housework_partner	Log of time spent for domestic work by the partner in a day
Table 3 - *continued*	
Help	Help received (paid) for domestic work and childcare
Work_gender_ gap	Gender gap in paid work: log of difference between woman's paid work and man's paid work (minutes in a day)
Partime	Categorical: If she works partime =1; fulltime =2; unemployed = 0
Wage	Log of yearly labour income[6]
Woman's work	Log of woman's working time in a day.
Lambda	Lambda Heckman ($\sigma_{1\eta}$ and $\sigma_{2\eta}$)

[6] ISTAT Survey on Time Use in Italy does not contain information on income (wage); consequently, a matching procedure has been performed in order to "import" income data from another source: the Bank of Italy 2004 Survey on Household Income and Wealth data collected in the year 2002. Details on matching procedure (balancing score and sensitivity test) are reported in the article of Campolo and Di Pino (2012), in which the same dataset of this study is used.

Table 4: Estimation results of the domestic work of Italian women in the years 2002-2003 (Eq. 1 : Employed women). Subsample size: no. 3091

Dependent variable: Log of daily domestic work	**Employed women (Regime 1)**					
	Two-Eqs. ML		*TS Heckman*		*Three-Eqs. ML*	
Regressors:	coef	*SE*	coef	*SE*	coef	*SE*
Intercept	7.8344	*0.1980*	6.5769	*0.4618*	6.9795	*0.3183*
Age	0.0169	*0.0018*	0.0158	*0.0020*	0.0154	*0.0020*
Children	0.2175	*0.0216*	0.1600	*0.0252*	0.1824	*0.0245*
Dummy_Weekend	-0.0605	*0.0285*	-0.0766	*0.0304*	-0.0914	*0.0309*
Area	0.1204	*0.0448*	-0.0814	*0.0642*	0.1147	*0.0476*
Education	-0.0151	*0.0038*	-0.0186	*0.0068*	-0.0108	*0.0043*
D_age	-0.0095	*0.0036*	-0.0110	*0.0040*	-0.0108	*0.0041*
Chcare_partner	0.0280	*0.0041*	0.0262	*0.0047*	0.0248	*0.0047*
Housework_partner	0.0172	*0.0035*	0.0225	*0.0040*	0.0254	*0.0039*
Help	0.0000	*0.0002*	0.0001	*0.0002*	0.0001	*0.0002*
Work_gender_ gap	-0.0029	*0.0078*	-0.0122	*0.0096*	-0.0320	*0.0087*
Partime	0.0497	*0.0366*	-0.0475	*0.0448*	-0.0892	*0.0738*
Wage	-0.0301	*0.0187*	0.1132	*0.0232*	0.0766	*0.0191*
Woman's work	-0.3586	*0.0264*	-0.3416	*0.0491*	-0.3125	*0.0354*
Lambda ($\sigma_{1\eta}$)			-0.2285	*0.2268*		
σ^2_1	0.6086		0.6491		0.7457	
ρ_{12}	**0.9594**	*0.0089*	**0.7103**		**0.4629**	

Table 5: Estimation results of the domestic work of Italian women (Eq. 2 : Unemployed women). Subsample size: no. 2607

Dependent variable: Log of daily domestic work	**Unemployed women (Regime 2)**					
	Two-Eqs. ML		*TS Heckman*		*Three-Eqs. ML*	
Regressors:	coef	*SE*	coef	*SE*	coef	*SE*
Intercept	5.9127	*0.0792*	5.4314	*0.2073*	5.6203	*0.1092*
Age	0.0070	*0.0013*	0.0061	*0.0014*	0.0062	*0.0014*
Children	0.1250	*0.0128*	0.1163	*0.0132*	0.1161	*0.0132*
Dummy_Weekend	0.2678	*0.0208*	0.2946	*0.0201*	0.2949	*0.0201*
Area	0.0172	*0.0189*	0.0665	*0.0219*	0.0560	*0.0197*
Education	-0.0024	*0.0030*	-0.0073	*0.0037*	-0.0053	*0.0033*
D_age	-0.0044	*0.0025*	-0.0037	*0.0026*	-0.0035	*0.0026*
Chcare_partner	0.0130	*0.0030*	0.0123	*0.0031*	0.0122	*0.0031*
Housework_partner	0.0072	*0.0023*	0.0075	*0.0024*	0.0069	*0.0024*
Help	-0.0001	*0.0002*	-0.0012	*0.0004*	-0.0011	*0.0004*
Work_gender_ gap	0.0339	*0.0036*	-0.0322	*0.0250*	-0.0072	*0.0095*
Lambda ($\sigma_{2\eta}$)			-0.1202	*0.0979*		
σ^2_2	0.2682		0.2412		0.2367	
ρ_{12}	**0.9594**	*0.0089*	**0.7103**		**0.4629**	

Table 6: Estimation results of Selection Equation using *TS Heckman* and *Three-Eqs. ML*

Dependent variable(dummy): L = 1 if the woman is employed		*TS Heckman first stage (Probit)*	*Three-Eqs. ML*	
Regressors:	coef	*SE*	coef	*SE*
Intercept	0.7476	*0.2377*	1.3571	*0.1865*
Age	0.0003	*0.0044*	-0.0113	*0.0034*
Children	-0.0798	*0.0471*	-0.1750	*0.0368*
Dummy_Weekend	0.0751	*0.0668*	0.0803	*0.0535*
Area	-0.7810	*0.0689*	-0.4374	*0.0544*
Education	0.1188	*0.0102*	0.0640	*0.0081*
D_age	0.0059	*0.0084*	0.0162	*0.0065*
Chcare_partner	0.0016	*0.0103*	-0.0145	*0.0081*
Housework_partner	-0.0300	*0.0084*	-0.0313	*0.0063*
Help	0.0019	*0.0010*	0.0012	*0.0007*
Work_gender_ gap	0.4453	*0.0093*	0.3968	*0.0085*

References

(1) Aakvik, A., Heckman J. J., Vytlacil E. J.: Estimating treatment effects for discrete outcomes when responses to treatment vary: An application to Norwegian vocational rehabilitation programs. J. Econometrics 125, 15–51 (2005)

(2) Bianchi, S. M.: Maternal employment and time with children: Dramatic change or surprising continuity? Demography, 37(4), 401–414 (2000)

(3) Calzolari, G., and L. Panattoni: Alternative estimators of FIML covariance matrix: A Monte Carlo study. Econometrica, 56, 701-714 (1988)

(4) Campolo, M.G., Di Pino, A.: An empirical analysis of women's working time, and an estimation of female labour supply in Italy. Statistica. 72(2), 173-193 (2012)

(5) Carneiro, P., Hansen, K. T., Heckman, J. J.: Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. Int. Econ. Rev. 44 (2), 361-422 (2003)

(6) Chen, H., Fan, Y., Liu, R.: Inference for the Correlation Coefficient between Potential Outcomes in the Gaussian Switching Regime Model. Department of Economics Vanderbilt University, Working Paper, August (2012)

(7) Heckman, J. J.: "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Model," Annals of Economic and Social Measurement, 15, 475-492. (1976)

(8) Heckman, J. J.: Varieties of Selection Bias. The American Economic Review 80(2), 313-318 (1990)

(9) Heckman, J. J., Honoré, B. E.: The Empirical content of the Roy Model. Econometrica 58(5), 1121-1149 (1990)

(10) ISTAT - Italian National Institute of Statistics, (2007), The Use of Time – Multi-Purpose Survey on Italian Families in 2002-2003.

(11) Lee, L. F.: Unionism and wage rates: A simultaneous equation model with qualitative and limited dependent variables. International Economic Review **19**(2), 415-433 (1978)

(12) Lokshin, M., Sajaia, Z.,: Maximum likelihood estimation of endogenous switching regression models. The Stata Journal. 4(3), 282–289 (2004)
(13) Maddala, G.S.: Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press. Cambridge (UK) (1983)
(14) Maddala, G.S.: Disequilibrium, Self-Selection, and Switching Models. In Griliches, Z., Intriligator, M.D. (eds.) Handbook of Econometrics, Vol. III, pp. 1633-1688. Elsevier Science, North-Holland (1986)
(15) Maddala, G.S., Nelson, F.: Maximum likelihood methods for markets in disequilibrium. Econometrica 42, pp. 1013-1030 (1974)
(16) Poirier, D.J. and P.A. Ruud: On the appropriateness of endogenous switching. J. Econometrics 16, 249-256 (1981)
(17) Poirier D. J., Tobias J. L.: On the predictive distributions of outcome gains in the presence of an unidentified parameter. J Bus Econ Stat. 24(2) 258-268 (2003)
(18) Roy, A.: Some thoughts on the distribution of earnings. Oxford Economic Papers, 133, 145-146 (1951)
(19) Vella, F., Verbeek, M.: Estimating and interpreting models with endogenous treatment effects. J Bus Econ Stat. 17(4), 473-478 (1999)
(20) Vijverberg W. P .M.: Measuring the unidentified parameter of the extended Roy Model of selectivity. J Econometrics 57**,** 69-89 (1993)

Appendix 1 – Monte Carlo Experiments: Estimated Coefficients

Table A1: Simulation results - **Errors distribution: Normal**; sample: 10000; Monte Carlo Replications: 10000. ***Absence of endogenous selection*** ($\sigma^2_2 = \sigma^2_1$)

ρ_{12} = 0.90 (**positive**)		***Two-Equation ML***		***TS Heckman***		***Three-Equation ML***	
	coefs set	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$const_1$	***10***	10.0060	*0.0580*	9.3040	*0.2120*	9.9940	*0.1300*
β_1	***1***	0.9990	*0.0000*	1.0250	*0.0000*	1.0000	*0.0000*
$const_2$	***10***	10.0010	*0.0550*	10.1340	*0.1230*	10.0060	*0.1280*
β_2	***1***	0.9997	*0.0002*	0.9870	*0.0003*	0.9990	*0.0003*
σ^2_1	***100***	99.98	*2.52*	100.63	*4.08*	99.99	*3.89*
σ^2_2	***100***	100.01	*2.67*	100.02	*4.21*	100.01	*4.20*
ρ_{12}	***0.9***	0.9000	*0.0048*	0.7998	*0.0070*	0.8999	*0.0049*

ρ_{12} = -0. 90 (**negative**)		***Two-Equation ML***		***TS Heckman***		***Three-Equation ML***	
	coefs set	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$const_1$	***10***	9.9910	*0.0420*	9.9760	*0.2130*	9.9930	*0.0470*
β_1	***1***	1.0010	*0.0000*	1.0010	*0.0000*	1.0010	*0.0000*
$const_2$	***10***	10.0030	*0.0460*	10.0090	*0.2180*	10.0020	*0.0510*
β_2	***1***	0.9996	*0.0001*	1.0000	*0.0003*	0.9997	*0.0001*
σ^2_1	***100***	100.11	*5.37*	100.04	*21.18*	100.10	*5.90*
σ^2_2	***100***	99.98	*5.04*	100.29	*21.04*	100.00	*5.54*
ρ_{12}	***-0.9***	-0.9002	*0.0068*	-0.8993	*0.0597*	-0.8985	*0.0392*

ρ_{12} = 0		***Two-Equation ML***		***TS Heckman***		***Three-Equation ML***	
	coefs set	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$const_1$	***10***	9.9970	*0.0800*	9.9900	*0.2170*	9.9910	*0.1420*
β_1	***1***	1.0002	*0.0002*	1.0005	*0.0003*	1.0004	*0.0003*
$const_2$	***10***	10.0030	*0.0830*	10.0120	*0.2010*	10.0100	*0.1420*
β_2	***1***	0.9997	*0.0002*	0.9990	*0.0003*	0.9990	*0.0003*
σ^2_1	***100***	99.95	*4.73*	99.87	*10.51*	100.00	*7.76*
σ^2_2	***100***	99.98	*4.63*	100.15	*9.79*	99.96	*7.29*
ρ_{12}	***0***	-0.0001	*0.0010*	-0.0004	*0.0015*	-0.0002	*0.0013*

Table A2: Simulation results - **Errors distribution: Normal**; sample: 10000; Monte Carlo Replications: 10000. **Endogenous selection** ($\sigma^2_1 = 4*\sigma^2_2$)

ρ_{12} = 0.90 (**positive**)		***Two-Equation ML***		***TS Heckman***		***Three-Equation ML***	
	coefs set	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$const_1$	***10***	10.0022	*0.3064*	9.9990	*0.5431*	9.9933	*0.3939*
β_1	***1***	0.9996	*0.0179*	0.9996	*0.0253*	0.9998	*0.0216*
$const_2$	***10***	10.0012	*0.1704*	10.0055	*0.2820*	10.0053	*0.2404*
β_2	***1***	0.9998	*0.0093*	0.9996	*0.0132*	0.9995	*0.0120*
σ^2_1	***200***	199.93	*3.50*	199.63	*5.10*	200.09	*4.32*
σ^2_2	***50***	50.03	*0.91*	50.08	*1.20*	50.01	*1.13*
ρ_{12}	***0.9***	0.9000	*0.00002*	0.8993	*0.0003*	0.9007	*0.0002*

ρ_{12} = -0. 90 (**negative**)		***Two-Equation ML***		***TS Heckman***		***Three-Equation ML***	
	coefs set	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$const_1$	***10***	9.9850	*0.2751*	9.9670	*0.6663*	9.9880	*0.2872*
β_1	***1***	1.0010	*0.0155*	1.0020	*0.0234*	1.0010	*0.0159*
$const_2$	***10***	10.0020	*0.1591*	10.0060	*0.3419*	10.0000	*0.1685*
β_2	***1***	0.9997	*0.0080*	1.0000	*0.0119*	0.9998	*0.0083*
σ^2_1	***200***	200.25	*4.54*	200.10	*9.77*	200.21	*4.66*
σ^2_2	***50***	50.00	*1.21*	50.14	*2.40*	50.01	*1.27*
ρ_{12}	***-0.9***	-0.9003	*0.00005*	-0.9016	*0.0050*	-0.8990	*0.0020*

ρ_{12} = 0		***Two-Equation ML***		***TS Heckman***		***Three-Equation ML***	
	coefs set	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$const_1$	***10***	10.0060	*0.0810*	9.9870	*0.6527*	9.9930	*0.1430*
β_1	***1***	1.0020	*0.0010*	1.0007	*0.0252*	1.0000	*0.0000*
$const_2$	***10***	10.0060	*0.0830*	10.0140	*0.3541*	10.0190	*0.1430*
β_2	***1***	1.0020	*0.0010*	0.9990	*0.0138*	1.0050	*0.0000*
σ^2_1	***200***	200.12	*4.73*	199.57	*7.83*	200.45	*7.76*
σ^2_2	***50***	50.09	*4.63*	50.01	*1.34*	50.26	*7.29*
ρ_{12}	***0***	0.0003	*0.0001*	-0.0047	*0.0005*	0.0005	*0.0002*

Table A3: Simulation results. **Errors distribution: Student-*t*** (dof: 5) - sample: 10000; Monte Carlo Replications: 10000. **Absence of endogenous selection** ($\sigma^2_1 = \sigma^2_2$)

ρ_{12} = 0.90 (**positive**)		***Two Equations ML***		***TS Heckman***		***Three-Equation ML***	
	coefs set	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$const_1$	***10***	9.9773	*0.2432*	9.2121	*0.9170*	9.9588	*0.3630*
β_1	***1***	1.0008	*0.0140*	1.0279	*0.0354*	1.0013	*0.0189*
$const_2$	***10***	9.9680	*0.2441*	10.1310	*0.3844*	10.0310	*0.3686*
β_2	***1***	1.0012	*0.0140*	0.9880	*0.0216*	0.9994	*0.0185*
σ^2_1	***100***	99.81	*2.67*	100.62	*3.64*	99.94	*3.56*
σ^2_2	***100***	100.11	*2.90*	99.93	*3.79*	99.89	*3.77*
ρ_{12}	***0.9***	0.9001	*0.0063*	0.7989	*0.1015*	0.9004	*0.0064*

ρ_{12} = -0.90 (**negative**)		***Two Equations ML***		***TS Heckman***		***Three-Equation ML***	
	coefs set	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$const_1$	***10***	9.7070	*0.3701*	9.8480	*0.4925*	9.7380	*0.3867*
β_1	***1***	0.9960	*0.0134*	1.0040	*0.0181*	0.9940	*0.0162*
$const_2$	***10***	9.3300	*0.7130*	9.7640	*0.5462*	9.3190	*0.7441*
β_2	***1***	1.0178	*0.0217*	1.0060	*0.0188*	1.0176	*0.0238*
σ^2_1	***100***	104.98	*6.35*	100.93	*5.20*	105.07	*7.20*
σ^2_2	***100***	104.54	*6.23*	102.47	*5.97*	104.71	*6.89*
ρ_{12}	***-0.9***	-0.9189	*0.0205*	-0.4780	*0.4252*	-0.8600	*0.0637*

ρ_{12} = 0		***Two Equations ML***		***TS Heckman***		***Three-Equation ML***	
	coefs set	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$const_1$	***10***	9.0240	*1.0790*	9.8600	*0.4915*	8.7790	*1.3280*
β_1	***1***	1.0300	*0.0360*	1.0038	*0.0194*	1.0310	*0.0370*
$const_2$	***10***	8.9940	*1.0790*	9.8380	*0.5007*	8.7550	*1.3280*
β_2	***1***	1.0320	*0.0360*	1.0050	*0.0194*	1.0310	*0.0370*
σ^2_1	***100***	103.69	*9.22*	100.72	*7.23*	107.69	*11.93*
σ^2_2	***100***	103.76	*6.20*	101.20	*4.69*	107.85	*9.68*
ρ_{12}	***0***	-0.1520	*0.1630*	0.1225	*0.1286*	-0.0360	*0.0561*

Table A4: Simulation results. **Errors distribution: Student-*t*** (dof: 5) - sample: 10000; Monte Carlo Replications: 10000. **Endogenous selection** ($\sigma^2_1 = 4*\sigma^2_2$)

ρ_{12} = 0.90 (**positive**)		***Two Equations ML***		***TS Heckman***		***Three-Equation ML***	
	coefs set	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$const_1$	***10***	9.9679	*0.3276*	9.6882	*0.6260*	9.5597	*0.6054*
β_1	***1***	0.9990	*0.0201*	1.0105	*0.0280*	1.0135	*0.0269*
$const_2$	***10***	10.0430	*0.1863*	10.0970	*0.3081*	10.2130	*0.3340*
β_2	***1***	0.9986	*0.0103*	0.9971	*0.0136*	0.9944	*0.0138*
σ^2_1	***200***	200.57	*5.80*	201.63	*8.04*	202.46	*7.94*
σ^2_2	***50***	49.82	*1.64*	50.29	*2.15*	50.65	*2.28*
ρ_{12}	***0.9***	0.9051	*0.0082*	0.9216	*0.0318*	0.8955	*0.0198*

ρ_{12} = -0.90 (**negative**)		***Two Equations ML***		***TS Heckman***		***Three-Equation ML***	
	coefs set	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$const_1$	***10***	9.6840	*0.4398*	9.8080	*0.7018*	9.9390	*1.1792*
β_1	***1***	0.9890	*0.0204*	1.0050	*0.0252*	0.9800	*0.0486*
$const_2$	***10***	9.4110	*0.6175*	9.8460	*0.3913*	9.3550	*0.7468*
β_2	***1***	1.0152	*0.0177*	1.0040	*0.0132*	1.0174	*0.0251*
σ^2_1	***200***	208.49	*11.45*	201.74	*10.81*	206.09	*14.62*
σ^2_2	***50***	53.28	*3.98*	51.17	*3.03*	53.66	*4.60*
ρ_{12}	***-0.9***	-0.9239	*0.0250*	-0.4350	*0.4692*	-0.8720	*0.0670*

ρ_{12} = 0		***Two Equations ML***		***TS Heckman***		***Three-Equation ML***	
	coefs set	coef	*RMSE*	coef	*RMSE*	coef	*RMSE*
$const_1$	***10***	8.7180	*1.3850*	9.6850	*0.7353*	8.4930	*1.6130*
β_1	***1***	1.0440	*0.0490*	1.0088	*0.0273*	1.0420	*0.0490*
$const_2$	***10***	9.1760	*0.8950*	9.9860	*0.3620*	9.0610	*1.0660*
β_2	***1***	1.0230	*0.0260*	1.0010	*0.0139*	1.0220	*0.0280*
σ^2_1	***200***	206.34	*17.31*	202.80	*15.19*	213.17	*21.36*
σ^2_2	***50***	52.19	*3.36*	50.12	*2.13*	53.55	*4.83*
ρ_{12}	***0***	-0.1590	*0.1710*	0.1974	*0.2053*	-0.024	*0.058*

Appendix 2 – Estimation and Partial Identification of Across Regime Covariance Using Second-Order Moments Relationships.

In a two-regime Roy model with a selection equation the covariances between the outcome equations and the selection equation (model 2) are allowed to be different from zero. The error terms u_{1i} and u_{2i} are normally distributed with zero mean and variances equal to σ^2_1 and σ^2_2. From the censoring rule imposed to both outcome equations we derive that y_{2i} and y_{1i} can be, respectively, observed only if : $v_i = u_{1i} - u_{2i} > -(\mathbf{x'}_{1i}\boldsymbol{\beta}_1 - \mathbf{x'}_{2i}\boldsymbol{\beta}_2)$, or $v_i = u_{1i} - u_{2i} \geq -(\mathbf{x'}_{1i}\boldsymbol{\beta}_1 - \mathbf{x'}_{2i}\boldsymbol{\beta}_2)$, where the random variable $v_i = u_{2i} - u_{1i}$ is normally distributed with zero mean and variance σ^2_v.

The covariance σ_{12} can be indirectly estimated using one among the methods based on the second order moments' relationships (e. g. Maddala, 1983 and 1986). The procedure here adopted uses the predicted values of the selection equation and of both outcome equations to estimate preliminarily σ^2_v. In doing this, we first consider the sample composition: $n = n_1 + n_2$ with n_1 observations under the Regime 1 and n_2 observations under the Regime 2. Then, given n_1 row vectors $\mathbf{x'}_{1i}$ in the regressors matrix of the Regime 1, n_2 row vectors $\mathbf{x'}_{2i}$ in the regressors matrix of the Regime 2, and n row vectors $\mathbf{z'}$ in the regressors matrix of the selection equation, we have:

$$\left(\mathbf{x}'_i\,\hat{\boldsymbol{\beta}}_1 - \mathbf{x}'_i\,\hat{\boldsymbol{\beta}}_2\right)/\hat{\sigma}_v = \mathbf{z}'_i\,\hat{\boldsymbol{\gamma}} \qquad \text{where:} \quad \mathbf{x}'_i = \left[\mathbf{x'}_{1i} \quad \mathbf{x'}_{1i}\right] \tag{8}$$

and:

$$\hat{\sigma}^2_v = \sum_{i=1}^{n}\left(\mathbf{x}'_i\,\hat{\boldsymbol{\beta}}_1 - \mathbf{x}'_i\,\hat{\boldsymbol{\beta}}_2\right)^2 \Big/ \sum_{i=1}^{n}\left(\mathbf{z}'_i\,\hat{\boldsymbol{\gamma}}\right)^2 \tag{9}$$

Hence, estimating $\hat{\sigma}^2_1$ and $\hat{\sigma}^2_2$ by the outcome equations and computing $\hat{\sigma}^2_v$ by the Eq. (9), the moment relationship $\sigma^2_v = \sigma^2_1 + \sigma^2_2 - 2\sigma_{12}$ allows us to obtain an estimate of the covariance σ_{12}.

Alternatively, second order moments relationships allow us also to obtain a partial identification of ρ_{12}, by the computation of a lower and an upper bound of the across-regime correlation parameter, as Vijverberg (1993) demonstrated. Assume that the random variable $v_i / \sigma_v = \eta_i$ is the $N(0,1)$ disturbance term of selection equation. Therefore, in model (2) error terms u_{1i}, u_{2i}, and η_i can be distributed as a trivariate normal[7], where the covariances $\sigma_{1\eta}$ and $\sigma_{2\eta}$ may be different from zero (Vijverberg, 1993; Chen et al., 2012). Given the semi-positive definiteness of errors covariance matrix of model's (2), the correlation parameter $\rho_{12}=\sigma_{12}/\sigma_1\sigma_2$ is included in the following interval:

$$\rho_{2\eta}\rho_{1\eta} - \left[\left(1-\rho^2_{2\eta}\right)\left(1-\rho^2_{1\eta}\right)\right]^{1/2} \leq \rho_{12} \leq \rho_{2\eta}\rho_{1\eta} + \left[\left(1-\rho^2_{2\eta}\right)\left(1-\rho^2_{1\eta}\right)\right]^{1/2} \tag{10}$$

[7] unlike in model (1) where error terms u_{1i}, u_{2i} are distributed as a bivariate normal

The estimation of $\rho_{1\eta}=\sigma_{1\eta}/\sigma_1$ and $\rho_{2\eta}=\sigma_{2\eta}/\sigma_2$ are provided by both *TS Heckman* or *Three-Equation ML* estimation procedures. We can observe by the inequality (9) that ρ_{12} is point identified only if one among $\rho^2_{1\eta}$ and $\rho^2_{2\eta}$ is equal to one at either $\rho_{1\eta}\,\text{sign}\,(\rho_{2\eta})$ or $\rho_{2\eta}\,\text{sign}\,(\rho_{1\eta})$.