International Journal of Business & Economic Strategy (IJBES)
Vol.4, 2016 - pp. 13-19

How to perform a panel data analysis when we have an independent surveys? Using Rubin causal model and multiple imputation to run a dynamic analysis of poverty in the Tunisian context

Amal JMAII
Faculty of Economic Sciences and Management of Tunis
University of Tunis El Manar
BP num 94-ROMMANA 1068, Tunisia
jemaiamal@yahoo.fr

Abstract **- This paper focus on the determinants of poverty dynamics in Tunisia by performing a two stage endogenous model using amelia multiple imputation bayesian algorithm to avoid missing data. Based on this approach we show that panel data analysis can be performed through potential variables such presented by the work of Donald B. Rubin. Our contribution is rather empirical, we propose a method based on the concept of causal inference that allows us to execute a dynamic analysis of poverty with panel data using independent surveys. As we assert the dependence between poverty status over time, we choose to use a switching endogenous probit model take into consideration the form of endogeneity caused by initial condition.**

Keywords **- Amelia program, causal inference, poverty dynamics, switching endogenous model, Welfare dynamics**

I Introduction

Reducing poverty is a serious challenge for all states of the world. Until our day, we did not hear that a country has succeeded to eliminate poverty permanently. Even in the most wealthiest and powerful countries poor still exist. This proves that, policies against poverty have failed. Generally, those who are experienced poverty can not escape from his trap. Many economists have tried to find the origine of poverty persistente ([3], [2], [5]), and wondered if we are poor because we are borned poor or we are become poor over time. In fact, there are several explanations that are presented in the literature. Mainly, the failure of the economic systems, the weakness of the education programs that may limit the opportunities to have a decent job and causes a higher unemployment rates and even exogenous shocks ([1]). In the development countries, the weakness of infrastructure such as roads and communication, may limit poor people to have access to information or to the labor markets.
Tunisian government has implemented a public policies to improve economic growth and to fight against poverty. As a consequence, the real consumption per capita has increased to hit 3.31 percent at 2005 and 2.5 at 2010 (2005 as base). The INS report argue that poverty has increased face to a stable gini index(36%) but considered very high relative to other emerging countries . Despite this progress, poverty and inequlities between regions is one among the causes of the Tunisian revolution. To find the adequate policies in the fight against poverty, we must, and firstly, find a better measure of poverty and efficient use of data. A better identification of poverty leads to a better policies of fighting against poverty. Unfortunately, in Tunisia like the most developping contries, penalized data are not available. For these reasons, studies on poverty dynamics in Tunisia are scarce. Among the few searches on Tunisia, we cite the work of [4] who used pseudo-panel approach to estimate the permanent expenditures.

Proceeding from the current literatures, in our paper we assume that individuals who were poor in 2005 and remain poor in 2010, after five years, are chronically poor. Two main questions constitute the motivation of our work: what are the factors that lead individuals to stay poor? ideally to answer to this question we must conduct a panel data analysis. But as mentioned previously, panel data

are not available. From this point, we propose in this paper a new multiple approach based on multiple imputation and causal inference to compute a poverty rates that can reflect the real situation of poverty in Tunisia. In particulary, this reflection allow us to see what could be the extent of poverty in Tunisia if we use entire data, concerning the individuals, over time and without missing value. For our research work, we have chosen to apply a multiple imputation method according to theoretical and practical criteria. This method able us to impute potential variables through the logic of missing data. Following the work of [19] and based on Bayesian statistics, each variable was imputed from the observed data, allowing to take into account the uncertainty associated with each step in the imputation process. Each complete bases, thus generated, provides an estimate of the parameter of interest, and then a single estimator is obtained by calculating the mean of these estimates. Concretely, this paper study the determinants of poverty dynamics by individual socio-demographic characteristics considered to be stable over time. Using causal inference, we constructed a penalized database by imputing potential variables.

The rest of the paper is structured as follows. In section two, we expose a theoretical and conceptual vision of our proposal across the causal inference and multiple imputation appraochs. Section three deal with the methodology of the estimation. In section four, we present the source of data and the variables. The results of our model are presented in the fifth section. Finally, we concludes.

II Treatment of missing data in the causal inference models context

Missing data is an unavoidable problem in the practice of statistics. Donald Rubin has formalized a causal model that allows to identify the assumptions that are necessary to support an observational study is similar to an experimental study ([18], [13]).

A Potential variable with missing data

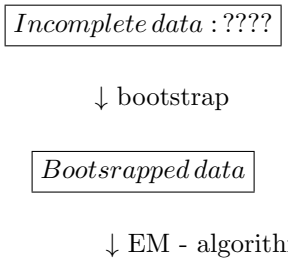
Rubin counterfactual model is based on two basic concepts: the causal statements and potential responses. Causal statements are also called treatment in the tradition of the counterfactual causal analysis. Each unit of analysis of a study should potentially be found in any of the examined causative states. In practice, this approach is close to the maching method in which researchers apply a treatment to a group in order to compare it with another group that did not receive the treatment. In fact, random assignment of an experiment ensures that each unit of analysis will be potentially found in a causal statements [14]. Conceptually, fifty percent of the information are not observable to be able to estimate the causal effect. This is a missing data problem. This inability to simultaneously observe two causal statements for the same unit of analysis was named the fundamental problem of causal inference ([11]).

According to this approach, and in the context of our analysis, panel data can be treated as a causal inference problem and a particulary case of missing data problem.

B Pretreatment step with the amelia multiple imputation algorithm

Many alternative methods of missing data have been proposed and applied for the imputation of data base. The first one is the Fully Conditionally Specified Models (FCS). This method is based on a modified version proposed by Gibbs sampler ([22]). Than a Bayesian algorithm was implemented in the package MI of the software R ([20]). However, according to [21] this approach has some theoretical weakness. Essentially, the incompatibility of the conditional densities. Hence the appearance of the AMELIA package wich is implemented in R. This algorithm represent a modified multivariate normal version which takes into consideration the specified clustering of observations ([12]).

Recently, James Honaker et al. (2011) have developped a more complete algorithm AMELIA II of multiple imputation program. The model is based on an assumption of normality $I \sim \mathcal{N}_k(\mu, \sum)$ and therefore requires some preliminary transformations of data. The EMB [1] algorithm of Amelia II combines the classic EM algorithm with a bootstrap approach. For each draw, the data are estimated by bootstrap in order to simulate the uncertainty, and then the EM algorithm is executed to find the posterior estimate $\hat{\gamma}_{MAP}$ for bootstrapping data (Figure 1).



1. The classic algorithm of maximum of likelihood ([7])

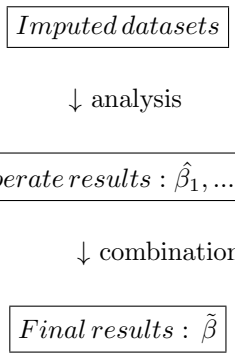


Figure 1: Shematic of multiple imputation with AMELIA II [2] program

III Methodology of the estimation

A The switching endogenous probit model: choice and modeling

The main idea of our paper is that present poverty status depends from the past poverty. However the expected correlation over time between unobservables will then cause a sample selection bias created by the called "initial conditions problem" ([10]). Among many solution to solve the problem is the use of a recursive bivariate probit model ([17])that assume the existance of a correlation between the two-probit equations distribution of poverty ([16]). Morever, an alternative solution is to use an endogenous switching model ([15], [23]) which takes into account the endogeneity of the dummy explanatory variable. In fact, sample selection problem and endogenous switching are among the most known problems in the economics and statistics literature. This problem exsit because the standard estimation techniques enable to correct the bias caused by the correlation of unobserved factors with the unobseved factors caused by the switch process [9]. Studies on poverty dynamics suggest that are more likely to be poor in the future.

In the endogenous problem, the response P_{ti} linked to the ith individual at time t is observed. We assume that the present poverty status P_{ti} depend on the past poverty $P_{t-1,i}$, which is an endogenous dummy variable, and a vector $(K \times 1)$ of other explanatory variables, X_i (we include the constant term). We define $P_{t-1,i}$ similary to P_{ti} with $S \times 1$ vector of explanatory variables, Z_i. We assume that the two vectors Z_i and X_i could containthe same elements.

We were inspired from the works of [15] to formulate our model as an equation system for two unobserved (i.e., latent) responses. Following the reasoning of [5] and using a bivariate probit model, we assume that P_i is distributed as follow:

$$P_{it}^* = X'_{it}\beta + \theta P_{t-1,i} + \epsilon_{ti}$$

$$P_{i,t} = \begin{cases} 1 & if \quad P_{it}^* > 0 \\ 0 & otherwise \end{cases} \tag{1}$$

where P_{it}^* is unobserved continous variable, β represent the parameters of the model (to be estimated), θ represents the associated coefficient of our endogenous dummy varaible $P_{t-1,i}$, and ϵ_{ti} the error term. Similary, the switching equation is specified as :

$$P_{i,t-1}^* = z'_{i,t-1}\alpha + u_{i,t-1}$$

$$P_{i,t-1} = \begin{cases} 1 & if \quad P_{i,t-1}^* > 0 \\ 0 & otherwise \end{cases} \tag{2}$$

same as equation(1), $P_{i,t-1}^*$ is unobserved continous variable, ϕ is the vector of parameters, and u_{it} represent the residual term.

We assume a bivariate normal function for ϵ_{it} and u_{it}. And to take into consideration the nature of dependence between ϵ_{ti} and $\epsilon_{i,t-1}$, we include λ which represent the random effect of the model.

$$\begin{aligned} \epsilon_{it} &= b\lambda_i + v_{it} \\ \epsilon_{i,t-1} &= \lambda_{i,t-1} + \gamma_{i,t-1} \end{aligned} \tag{3}$$

where b is free parameter [3], we define λ, v and γ as $ID \sim \mathcal{N}(0,1)$. We present the covariance matrix as :

$$\begin{aligned} cov\{(\epsilon_{it}, \epsilon_{i,t-1})'\} \equiv \sum &= \begin{bmatrix} b^2+1 & b \\ b & 2 \end{bmatrix} \\ \rho &= \frac{b}{\sqrt{2(b^2+1)}} \end{aligned} \tag{4}$$

Here $P_{i,t-1}$ is considered as exogenous (in the equation1), if and only if $\rho = 0$. In this case a sample probit regression can give a consistent estimators for the parameters of the model. But, if we assume that $\rho \neq 0$, this ordinary regression give us an inconsistent estimatrors, evidentlly, because $P_{i,t-1}$ is correlated to ϵ_{it} by λ_i (unobserved heterogeneity). The suspected endogenous of the variable let us to use the endogenous switching model to avoid the sample bias.
In fact, the model describe two status of poverty: the "basis poverty" and the "poverty transition". Then we can rewrites le model as:

The Basis poverty

$$P_r(p_{i,t-1}) = \Phi(z'_{i,t-1}\alpha) \tag{5}$$

2. for more details see Honaker et al 2014
3. we present parameter because we have information about the correlation given by ρ

The Poverty transition

$$P_r(p_{it}=1\backslash p_{i,t-1}=1)=\Phi([p_{i,t-1}\beta_1+(1-p_{i,t-1}\beta_2]X'_{it}) \quad (6)$$

Note that p_{it} and $p_{i,t-1}$ represents poverty status on t and t-1 respectively. As same as before X'_{it} and $z'_{i,t-1}$ are exogenous explanatory variables.

B State dependence

In this part, we follow the works of [5] to distinguich between *Aggregate State Dependence* (ASD) and *True State Dependence*[4] (TSD). The first one (ASD) does not take into account the individual heterogeneity. It represent the difference between the conditional probability of being poor at time t given that it was poor at t-1, and the conditional probability of being poor at time t given that it was not poor at t-1. Actually, it represent the difference between poverty persistence and poverty transitory (entry). We assume:

$$ASD=\frac{\sum_{s\in\{p_{st-1}=1\}}P_r(p_{st}=1\backslash p_{s,t-1}=1)}{\sum_s p_{s,t-1}}-\frac{\sum_{s\in\{p_{st-1}=0\}}P_r(p_{st}=1\backslash p_{s,t-1}=0)}{\sum_s(1-p_{s,t-1})} \quad (7)$$

On other hand, we measure the TSD as the rising probability of being poor on t caused by the poverty on t-1, with control to the individual heterogeneity. The choosen model assume that each composant of X'_{it} (in the transition equation) may present a various impact on the status of poverty on "t" conditionally to poverty status on "t-1".

$$GSD=N^{-1}\sum_{i=1.....N}[\hat{P}_r(p_{it}=1\backslash p_{i,t-1}=1)-\hat{P}_r(p_{it}=1\backslash p_{i,t-1}=0)] \quad (8)$$

IV Data source and variables definition

To analyze poverty dynamics in Tunisia, we chose households consumption surveys, INS (National Institute of Statistics), for 2005 and 2010 as the source of the used databases. In the two period, individuals who were born in specific yeras (for exemple, 1984-1975, or who are 21 to 30 years old in 2005), and had some observed household characteristics were selected. Within our methodology, households heads as well as their partners, their children and other dependens and relatives persons have been considered[5]. The main idea of the proposed methodology is that we impute potential variables to panelized the used database through a Bayesian algorithm. Evidently, we to impute this variables, we need some characteristics linked to the individuals.

- Generation (4): People who were born between 1984 eand 1975, between 1974 and 1965, between 1964 and 1955, or between 1954 and 1945.
- Sex (2): Man or Woman
- Education (4): Illiterate, Primary level, Secondary level and High level
- Areas (6): Urban Great Tunis, Rural Great Tunis, Urban East, Rural East, Urban West, Rural West.
- Employment[6] (4): Inactive, Salaried, Independent, Unemployed.

Taking into account this choice we run a multiple imputation using R. this propgram has therefore produced a total of obsevations
Statistic descrptives give us an interessant results. At the begining, statistics show that woman dominate the sample with a percentage of 53.8%. Those individuals who were born between 1984 and 1975 represents 31.5 % of total sample, following by those who were born between 1974 and 1965 by 27.4%, those who were born between 1964 and 1955 represents 24.9% and finally those who were born between 1954 and 1945 represents 16.1%.

Individuals who are illiterate represents 21% of total sample and those who have a primary educational level represents 38.6%. Individuals who have secondary educational level represents 29.4% and those who have a higher educational level represents 10.7%.

Our statistics highlight the large size of urban East, at the two years, compared to other regions (27.8% at 2005 and 28.9% at 2010). Followed by the great Tunis at 2005 by 18.9% but at 2010 it is the urban west by 19.5% than the other regions. This proves the migration movement between regions.

In this paper total expenditure is taken as a standard of living indicator. Indeed, consumption expenditure are further characterize by their stability over time compared to income fluctuations. They provide information about the degree of satisfaction that comes from the consumption of good and services. This approach has been advocated in recent studies by [8] [6].

4. called also as Genuine State Dependence (GSD), for more details see [5]

5. individuals without family link such us guests, family employees or ever relatives of this employees that living with the household were excluded (according to the classification of the INS

6. we did not detail the sector of activities because using a wide sector is not within the scope of our study

Table 2: Statistic descriptives of the imputed data

parameters	mean	parameters	mean
sex		*Employment2010*	
men	0.461	Inactive	0.434
Women	0.538	Salaried	0.214
Generation		Independant	0.275
born between 1984 - 1975	0.315	unemployed	0.075
born between 1974 - 1965	0.274	*Region2005*	
born between 1964 - 1955	0.249	Urban Great Tunisia	0.189
born between 1954 - 1945	0.161	Rural Great Tunisia	0.043
Education		Urban East	0.278
Illiterate	0.210	Rural East	0.161
primary level	0.386	Urban West	0.179
secondary level	0.294	Rural West	0.147
higher level	0.107	*Region2010*	
Employment2005		Urban Great Tunisia	0.184
Inactive	0.392	Rural Great Tunisia	0.039
Salaried	0.165	Urban East	0.289
Independant	0.379	Rural East	0.158
unemployed	0.062	Urban West	0.195
		Rural West	0.132

Source: Own compute based on INS data

V Results and interpretations

According to the Student test, the auto-correlation coefficient ρ is statistically significant with a p-value different of zero. This is mean that an individual how was poor in 2005 has a higher probability to be poor in 2010.

With regard to the generation variables, we found that all coefficients are statistically significant and negative for the older generation. This sign indicates that the probability of being poor is greater for younger individuals. This individuals are less likely to be initially poor than the other generation. This find might be explained by the fact that this generation was more formed in terms of education system compared to others. As a result they had more chance of having an employment than others.

Regarding to the educational level effects, we point the similar probabilities for both chronic and transient poverty for individuals whose educational level is primary or illiterate. This groups might suffer from periodical changes in their status.

Results show also that 11 per cent of poverty during the analyzed period is explained by Genuine State Dependence (GSD). This prove that observed poverty depends on the past of poverty state after controlling individuals heterogeneities. Difference between the Aggregate State Dependence (ASD)and GSD

Finally, our study shows that around 61 per cent of total observed poverty (0.49) is due to the origin of a stationary propension of poverty (ie. chronic condition) while 39 per cent coming from transient poverty. When we compare different poverty lines, we highlight similar results.

VI Conclusion

Most works on poverty dynamics, in developing country, consisted in measuring and identifying correlated variables to chronic and transient poverty. To execute such analysis panel data are recommended. However, as in most developing countries, penalized data about individual's well-being conditions does not exist. As a result, researchers does not able to conduct a *ex-post* dynamic analysis of poverty. In this paper we propose a new approach able us to panelized our data base throught the Rubin causal model. We use a bayesian algorithm included in the AMELIA II program able us to impute potential variables and then to run a switching endogenous probit model. The importance of this model is that it take into account the endogenous nature of poverty.
Based on consumer households surveys for the two year 2005 and 2010, in Tunisia, this work has yielded several results that would help to better understand the phenomenon of poverty dynamics, its contours and its determinants. Poverty in Tunisia is mostly chronic (62%). Individuals how live in rural west are the most vulnerable. Results show also that, *ceteris paribus*,a higher level of education increases the probability of escaping from poverty.
In sum, the importance of this paper is the estimate of the poverty dynamics composantes through a

Table 3: Switching model results

	initial condition		permanence in poverty		Transition to poverty	
Explanatory variables	marg effect	coef	marg effect	coef	marg effect	coef
sex: Male	-0.0222	-0.0567*	0.0002	0.0055	-0.0020	0.0045
employment(inactive as reference)						
Salaried	0.0417	0.1548*	0.0064	0.0064	0.0049	0.0050
Independent	-0.0512	-0.1575*	0.0594	0.0065*	0.0474	0.0055*
unemployed	0.0088	0.0251*	0.0402	0.0102*	0.0308	0.0076*
Education: (illiterate as reference)						
Education level: primary	-0.0556	-0.1555*	-0.0028	0.0077	-0.0068	0.0066
Education level: Secondary	-0.1234	-0.3152*	-0.0375	0.0096*	-0.0455	0.0079*
Education level: higher	-0.1742	-0.4190*	-0.1069	0.0115*	-0.1069	0.0084***
Areas: Rural Great Tunisia	0.0199	0.064	0.1207	0.0158*	0.0938	0.0128*
Areas: urban East	0.0266	0.072*	-0.0270	0.0078*	-0.0194	0.0058*
Areas: Rural East	0.0929	0.357*	0.1485	0.0097***	0.1169	0.0090*
Areas: urban West	0.0674	0.248*	0.0734	0.0089*	0.0556	0.0070*
Areas: Rural West	0.1777	0.723*	0.2462	0.0091*	0.2031	0.0101*
Second generation:	-0.0028	-0.034	0.0496	0.0071*	0.0416	0.0059*
Third generation:	0.0144	0.051**	0.0115	0.0074*	0.0089	0.0060
Fourth generation:	-0.0065	0.036	-0.0415	0.0081*	-.0398	.0062*
ρ	-0.1892	p<0.000	Number of observation: 51492			
Chronic poverty	0.3		Observed poverty		0.49	
ASD	0.1066		GSD		0.0870	

Source: Own compute based on the imputed data, significances codes: *p<0.05, **p<0.01 and ***p<0.001

new approach which allow to excute a panel data analysis even when we have independant surveys. With regard to results, it seems to us that causal inference may be a good tool of measuring poverty. All these finds are encouraging in terms of public policy since it suggests a new approach to poverty dynamics analysis in the case of non-panelized data, but they still insufficient and requires deepening.

References

[1] Aliber, M. (2003). Chronic poverty in south africa: Incidence, causes and policies. *World Development 31*(3), 473–490.

[2] Azariadis, C. (1996). The economics of poverty traps part one: complete markets. *Journal of economic growth 1*(4), 449–486.

[3] Barrett, C. B. and B. M. Swallow (2006). Fractal poverty traps. *World development 34*(1), 1–15.

[4] Ben Rejeb, J. (2008). Quantification de la pauvreté permanente sur la base de données non panélisées. *Revue économique 59*(2), 291–305.

[5] Cappellari, L. and S. P. Jenkins (2004). Modelling low income transitions. *Journal of applied econometrics 19*(5), 593–610.

[6] Chamarbagwala, R. (2010). Economic liberalization and urban–rural inequality in india: a quantile regression analysis. *Empirical Economics 39*(2), 371–394.

[7] Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

[8] Fang, Z. and C. Sakellariou (2013). Evolution of urban–rural living standards inequality in thailand: 1990–2006. *Asian Economic Journal 27*(3), 285–306.

[9] Heckman, J. J. (1978). Simple statistical models for discrete panel data developed and applied to test the hypothesis of true state dependence against the hypothesis of spurious state dependence. In *Annales de l'INSEE*, pp. 227–269. JSTOR.

[10] Heckman, J. J. (1981). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process.

[11] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association 81*(396), 945–960.

[12] Honaker, J. and G. King (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science 54*(2), 561–581.

[13] Imbens, G. W. and D. Rubin (2009). *Causal inference in statistics, and in the social and biomedical sciences*. Cambridge University Press New York.

[14] Little, R. J. and D. B. Rubin (2014). *Statistical analysis with missing data*. John Wiley & Sons.

[15] Miranda, A. and S. Rabe-Hesketh (2006). Maximum likelihood estimation of endogenous switching and sample selection models for binary, ordinal, and count variables. *Stata Journal 6*(3), 285–308.

[16] Newman, C. and S. Canagarajah (2000). Nonfarm employment, poverty, and gender linkages: evidence from ghana and uganda. *Working Draft. World Bank, Washington, DC*.

[17] Roodman, D. (2011). Fitting fully observed recursive mixed-process models with cmp. *The Stata Journal 11*(2).

[18] Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association*, 543–546.

[19] Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, Volume 81. John Wiley & Sons.

[20] Su, Y.-S., M. Yajima, A. E. Gelman, and J. Hill (2011). Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software 45*(2), 1–31.

[21] Van Buuren, S., J. P. Brand, C. Groothuis-Oudshoorn, and D. B. Rubin (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation 76*(12), 1049–1064.

[22] White, I. R., P. Royston, and A. M. Wood (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine 30*(4), 377–399.

[23] Wilde, J. (2000). Identification of multiple equation probit models with endogenous dummy regressors. *Economics letters 69*(3), 309–312.

[]