Econometrics **2015**, *3*, 719-732; doi:10.3390/econometrics3040719

OPEN ACCESS

econometrics

ISSN 2225-1146

www.mdpi.com/journal/econometrics

Article

Counterfactual Distributions in Bivariate Models—A Conditional Quantile Approach †

Javier Alejo [1,*] and Nicolás Badaracco [2]

[1] Center for Distributive, Labor and Social Studies (CEDLAS) and CONICET, Facultad de Ciencias Económicas, Universidad Nacional de La Plata (UNLP), La Plata 1900, Argentina

[2] Center for Distributive, Labor and Social Studies (CEDLAS), Facultad de Ciencias Económicas, Universidad Nacional de La Plata (UNLP), La Plata 1900, Argentina;
E-Mail: badaracconicolas@gmail.com

† A previous version [1] in Spanish was published in the Annals of the XLIX Annual Meeting of the AAEP.

* Author to whom correspondence should be addressed;
E-Mail: javier.alejo@depeco.econo.unlp.edu.ar; Tel.: +54-221-422-9383 (ext. 14).

Academic Editor: Gabriel Montes-Rojas

Received: 10 September 2015 / Accepted: 21 October 2015 / Published: 9 November 2015

Abstract: This paper proposes a methodology to incorporate bivariate models in numerical computations of counterfactual distributions. The proposal is to extend the works of Machado and Mata (2005) and Melly (2005) using the grid method to generate pairs of random variables. This contribution allows incorporating the effect of intra-household decision making in counterfactual decompositions of changes in income distribution. An application using data from five latin american countries shows that this approach substantially improves the goodness of fit to the empirical distribution. However, the exercise of decomposition is less conclusive about the performance of the method, which essentially depends on the sample size and the accuracy of the regression model.

Keywords: counterfactual distributions; quantile regression; numeric integration; grid method; labor market; income distribution

JEL classifications: C13; C14; C15; C31

---

1. Introduction

Most empirical studies analyze the effects of income distribution determinants through decomposition methodologies based on Oaxaca-Blinder (1973) [2,3]. Those methodologies usually focus on the wage distribution of a single individual assuming that all employment decisions are made in an isolated or independent way with respect to other household members. Notwithstanding, the literature on intra-household labor supply shows several models of the interdependence in the employment decisions within the household. Assortative mating literature provides vast evidence of interrelations of individual variables among members, such as their education levels, labor income, the choice of hours of work, *etc*. Ignoring this feature when estimating household labor earnings on decomposition exercises provides a scenario that may be biased or unrealistic.

The main component of personal earnings is labor income. Therefore, it is important to know its intra-household determinants to understand the behavior of the household incomes and their consequences on inequality. In the most traditional model, there is a sole individual responsible for making labor decisions independently of other household members. However, in the case of complete households (with head and spouse) it is usual that this decision is made by the couple. There are several models in the literature where a couple faces the problem of deciding together their labor supply according to their interests within the home (e.g., Chiappori and Pierre-Andre, 1992 [4]; Blundell *et al.*, 2005 [5]; van Klaveren *et al.*, 2008 [6], among others). The main mechanisms behind this decision are the reservation wages of each member and the bargaining power that determines the share rule of the household income.

Given the complexity involved in analyzing the joint employment decisions of all household members, the usual alternative is to focus only on the decisions made by the household head and spouse. The implicit assumption is that the rest of the household members will not change their behavior, or at least their impact on family income is small. This assumption may be too simple, but it is a starting point used in the literature to understand the complex mechanisms interacting in the labor decisions made within the household. In particular, both the reservation wages and bargaining power depend on observable and unobservable characteristics of household members such as age, education status, persuasion, *etc*. Modeling both earnings equations to analyze household income distribution while taking into account their interactions requires a methodology that generates counterfactual distributions of hypothetical changes on their determinants.

Some examples of models including employment decisions within the household are Browning *et al.* (1994) [7], Gasparini and Marchionni (2007) [8], Galiani and Weischenbaum (2012) [9], among others. Usually these studies make several parametric and/or distributional assumptions, such as normality in the unobservable income determinants. This approach could be too strict or it may not be quite representative of the actual income distribution. Another usual methodological aspect is that those papers use models focused on conditional means, relying on parametric assumptions other aspects of the distribution. Despite the progress of quantile regression literature allows exploring issues beyond the average effects, the bulk of the decompositions literature is based on counterfactual distributions of earnings equations for a single individual (*i.e.*, Machado and Mata (2005) [10], Melly (2005) [11] and

Firpo *et al.* (2009) [12]). This paper attempts to expand this literature by proposing a methodology to generate counterfactual scenarios on bivariate distributions using conditional quantiles.

The main contribution of this paper is to show that the problem of generating counterfactual income distributions for both household members is just an exercise of numerical integration involving a joint mechanism to generate a pair of random variables through their marginal distributions. Once this mechanism is established, it is possible to use the ranking association of both household members in order to get a set of replicates or realizations of the joint distribution. Nevertheless, the fact that incomes are related with observable characteristics makes necessary to introduce some structure to the conditional income distribution. Conditional quantiles are useful to model this matter for two reasons: first, they are the counterpart of the cumulative conditional distribution, and second they are easily estimable by standard methods. Quantile regressions allows an indirect way to capture the unobservable heterogeneous effects on each marginal distribution. Finally, the last step of the proposed method is to incorporate the relationship between the conditional incomes of both household members using a probabilistic association of conditional rankings.

The paper is organized as follows. In Section 2, a methodology to simulate bivariate random variable realizations based on marginal distributions is presented. Section 3 extends this idea to conditional joint distributions and its applications to counterfactual distributions. Section 4 shows an empirical application with household survey data for different countries in the Southern Cone of Latin America. Finally, Section 5 discusses the results and scope of the methodology.

2. Generating Random Variables

Generating random variables in the univariate case is relatively simple, and there are several methods available. The most widely used is the inverse cumulative function method: let U be a random variable with a uniform distribution $U(0,1)$, then the transformation $F_Y^{-1}(U)$ generates a random variable with distribution $F_Y(y)$. Thus, this procedure simply consists on taking a realization of a uniform random variable u and then computing the $u-th$ quantile $Q_Y(u) \equiv F_Y^{-1}(u)$. In the case of integer variables, the logic is quite similar to the continuous case (Devroye, 1986) [13].

The bivariate setup is more complex because the statistical relationship between two variables must be considered. A closely related problem can be found in the study of copula functions. A copula is a function that links the joint distribution to the one-dimensional marginal distributions (Nelsen, 1999) [14]. As in the univariate case, there are several methods to create a bivariate random draw. For example, the conditional distribution method allows to generate a random vector (y_1, y_2) using a vector (U_1, U_2) of independent uniform random variables. Specifically, the method of the conditional distribution requires the following two steps: (1) compute $y_2 = F_2^{-1}(U_2)$, where $F_2(.)$ is the marginal cumulative distribution of y_2; and (2) compute $y_1 = F^{-1}(u_1|y_2)$, that is, using the inverse of the cumulative distribution of y_1 conditional on y_2. The key to this process is to know the exact functional form of the conditional distribution, which can be too strict in practice.

Another strategy that allows us to adapt the univariate methods (such as the inverse cumulative function) to the bivariate problem is the grid method. Before explaining this procedure, it is appropriate to give some definitions.

Definition 1. *(encoder function). Let m_1 and m_2 be two integer variables in the domain$[1, M_1]$ and $[1, M_2]$, respectively. We define the encoder function $m = T(m_1, m_2)$ as $m = M_1 m_1 + m_2$.*

The image of this function is $Im(T) = [M_1 + 1, M_1^2 + M_2]$. The most interesting property of the encoder function is that it has a single value m for each ordered pair (m_1, m_2). Therefore, each coordinate is identified by the following decoding function:

Property 1. *(decoding functions). Let $m \in Im(T)$ be a encoder function. Then, the coordinates m_1 and m_2 can be obtained from the decoding functions: $m_1 = [m/M_1]$ and $m_2 = m - M_1 m_1$.*

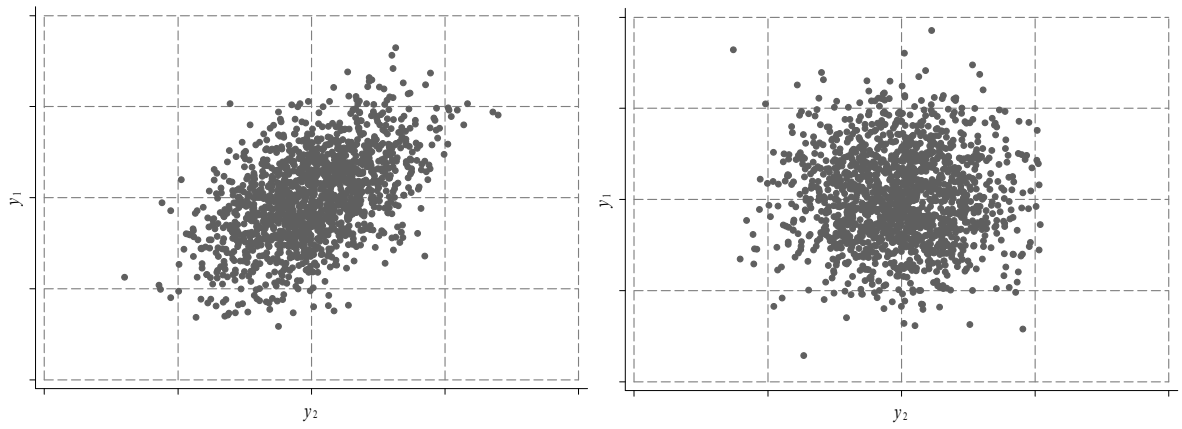
The last element that is needed is to define the set of grids of an enclosure $A \subset \mathbf{R}_+^2$.

Definition 2. *(grid C_m). Let $A \subset \mathbf{R}_+^2$ be an enclousure, a grid $C_m \subset A$ is defined as $C_m \equiv \{(y_1, y_2) \in A |\ a_{m_1-1} < y_1 < a_{m_1} \wedge b_{m_2-1} < y_2 < b_{m_2}\}$ with $m_1 = 2, \ldots, M_1$, $m_2 = 2, \ldots, M_2$, where a_{m_1} and b_{m_2} are values such that they satisfy $a_1 < a_2 < \ldots < a_{M_1}$; $b_1 < b_2 < \ldots < b_{M_2}$ and $\bigcup_{m \in Im(T)} C(m) = A$.*

Finally, consider two random variables $(y_1, y_2) \in A$ with a joint density function $f(y_1, y_2)$. To generate a realization of a vector $(\tilde{y}_1, \tilde{y}_2)$ from the population distribution $f(y_1, y_2)$ we can use the grid method by following the next steps:

1. Subdivide the enclosure A in grids C_m, where $m \in Im(T)$.
2. Calculate the probability mass of each grid $p_m = Pr[(y_1, y_2) \in C_m]$ for every m.
3. Generate a realization of an integer univariate random variable $\tilde{m}$ with probability distribution p_m, calculated in the previous step.
4. Decode $\tilde{m}$ to obtain the vector $(\tilde{m}_1, \tilde{m}_2)$.
5. Compute the realization of $(\tilde{y}_1, \tilde{y}_2)$ assigning values within the grid $C_{\tilde{m}}$.

Figure 1 presents two examples to illustrate how the method works: graph *(a)* shows the situation of two random variables with a positive relationship while *(b)* represent the independent relationship case.



(a) **(b)**

Figure 1. Random variables and correlation.

The dotted lines delimit the grids subdividing the enclosure (*i.e.*, the support of both random variables). Clearly, in the first case the probability mass (measured by the proportion of points falling

into each grid) is concentrated in the diagonal given by the bisectrix, while in the second case there is no clear pattern for the joint probability. Logically, the greater the number of grids, the better the approach of the method (Hörmann *et al.*, 2004 [15]). Therefore, the method incorporates the statistical relationship between y_1 and y_2 through the probability of each grid.

Lastly, note that the grids can be determined by their marginal quantile by defining the values $u_{m_1} \equiv F_1(a_{m_1})$ and $v_{m_2} \equiv F_2(b_{m_2})$. The validity of this equivalence is that the cumulative distribution is an increasing monotonic transformation of the random variable support. In other words, the $F_j(.)$ value represents the ranking position resulting from sorting the y_js increasingly. This establishes a one to one relationship between any value of y_j and its ranking. Then, the grid C_m can be written as:

$$\begin{aligned} C_m &= \{(y_1, y_2) \in A \mid a_{m_1-1} < y_1 < a_{m_1} \wedge b_{m_2-1} < y_2 < b_{m_2}\} \\ &= \{(y_1, y_2) \in A \mid Q_{y_1}(u_{m_1-1}) < y_1 < Q_{y_1}(u_{m_1}) \wedge Q_{y_2}(u_{m_2-1}) < y_2 < Q_{y_2}(u_{m_2})\} \end{aligned}$$

Figure 2 shows this equivalence in the definition of grids. The top graph on the right shows the point cloud in the plane (y_1, y_2), while the lower graph on the left shows its counterpart in the plane $(F_1(y_1), F_2(y_2))$. The figures in the other two quadrants show the cumulative probability of each variable viewed solely as a univariate distribution (marginal distribution). Note that although the scales of the enclosures are different, each point belonging to a grid in the upper right quadrant has a corresponding grid on the lower left quadrant.

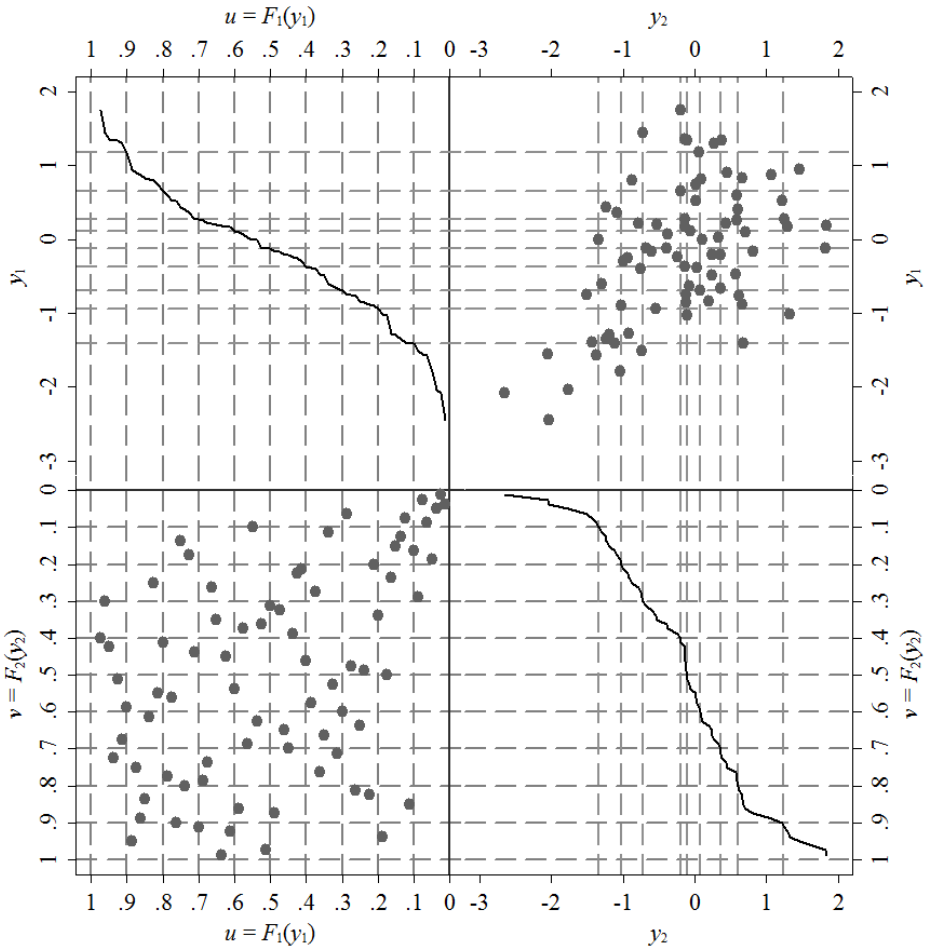


Figure 2. Equivalence of grids.

The grid definition on the marginal probability plane is equivalent to define the grid in terms of the levels of both variables. Then, looking at the grid plane makes it possible to adapt this method to the context of conditional quantiles. This is a key idea because it is precisely the estimation target of the quantile regression technique. Briefly, building the link between the probability grids and the conditional quantiles allows us to associate the marginal rankings with the univariate method of random sampling for the purpose of generating counterfactual distributions. Furthermore, this strategy requires less information than the method of the conditional distribution given that it only requires to know the probability of each grid rather than an entire functional form for the distribution of y_1 conditional on y_2.

3. Conditional Bivariate Model

3.1. Population

Consider now the distribution of (y_1, y_2) depending on a group of covariates (x_1, x_2). In particular, consider the following linear model for the pair of random variables (y_1, y_2):

$$y_1 = x_1'\beta_1 + \varepsilon_1 \quad (1)$$

$$y_2 = x_2'\beta_2 + \varepsilon_2 \quad (2)$$

where x_1 and x_2 are observable covariates vectors and the errors have a joint density $f(\varepsilon_1, \varepsilon_2|x)$, with $x \equiv (x_1, x_2)$. Using the Skorohod's representation, the same model can be formulated under the conditional quantile form:

$$Q_{y_1|x_1}(\theta_1) = x_1'\beta_1(\theta_1)$$

$$Q_{y_2|x_2}(\theta_2) = x_2'\beta_2(\theta_2)$$

where (θ_1, θ_2) are two random variables whose domain is given by $A \in [0,1] \times [0,1]$. Given the joint density $f(\varepsilon_1, \varepsilon_2|x)$, the density of this transformation becames:

$$g(\theta_1; \theta_2|x) = \frac{f[F_{Y_1}^{-1}(\theta_1), F_{Y_2}^{-1}(\theta_2)|x]}{f_{y_1}[F_{Y_1}^{-1}(\theta_1)|x] f_{y_2}[F_{Y_2}^{-1}(\theta_2)|x]}$$

Note that this function is the second derivative of the y_1 and y_2 copula conditional on x. The estimation of this object is not easy if we do not previously postulate some parametric assumptions (e.g., bivariate gaussian). While there are several available parametric forms for copulas, such as the Fréchet and Mardia families, our goal is to keep the nonparametric aspect that characterizes the quantile regression approach. However, it is unclear how the density estimation is useful for generating a sequence of random numbers to build counterfactual scenarios. [1] In this context, generating random values for a vector (y_1, y_2) conditional on x appears as a simple extension of the grids method explained in the previous section.

---

[1] Unfortunately, even though the Fréchet-Hoeffding copula bounds are useful to characterize certain particular cases such as comonotonic or countermonotonic random variables, they are too generic to be used in the context of random variables simulation.

Consider for simplicity that $M_1 = M_2 = M$ and the cutoff values in A are equally spaced—*i.e.*, $u_{m_1} = m_1/(M+1)$ and $v_{m_2} = m_2/(M+1)$. Then,

1. Subdivide the support region defined by $[0,1] \times [0,1]$ in C_m grids for $m \in Im(T)$.
2. Calculate the probability of each grid $p_m = Pr[(\theta_1, \theta_2)|x \in C_m]$ for each m.
3. Generate the realization of an integer univariate random variable $\widetilde{m}$ with probability p_m, obtained in the previous step.
4. Decode $\widetilde{m}$ at coordinates $(\widetilde{m}_1, \widetilde{m}_2)$.
5. Get realizations of the pair $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ assigning values within the grid $C_{\widetilde{m}}$.
6. Generate $(\widetilde{y}_1, \widetilde{y}_2)$ using the pair $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ and the univariate method of inverse cumulative function $\widetilde{y}_1 = Q_{\widetilde{\theta}_1}(y_1|x_1)$ and $\widetilde{y}_2 = Q_{\widetilde{\theta}_2}(y_2|x_2)$.

So far, all the elements used in each step of the process come from the population and so they are unobservable for the econometrician. Thus, an estimation strategy is required. The next section discusses about this topic when we have a random sample available instead of population data.

3.2. Sample Estimation

To generate a replicate (y_1, y_2) using a random sample we can apply the same procedure explained in the preceding paragraphs but replacing each element with its sample analogue. Specifically, we can estimate conditional quantiles $\widehat{Q}_{\theta_j}(y_j|x_j) = x_j'\widehat{\beta}_j(\theta_j)$ for $j = 1, 2$ using a certain grid of values—e.g., $\theta = 0.05, 0.10, ..., 0.90, 0.95$. The classic reference to get a consistent estimator of $\widehat{\beta}_1(\theta_1)$ and $\widehat{\beta}_2(\theta_2)$ is Koenker and Basset (1978) [16].

The grid method steps when we are working with sample estimators are:

1. Build $\widehat{C}_m$ using $x_{2i}'\widetilde{\beta}_1(a_m)$ and $x_{2i}'\widetilde{\beta}_2(b_m)$ as delimiters.
2. Estimate $\widehat{p}_m = \widehat{\Pr}(C_m) = n^{-1}\sum_{i=1}^{n} 1(i \in \widehat{C}_m)$, for every m.
3. Generate realizations of an integer random variable $\widetilde{m}$ according to the probabilities $\widehat{p}_m$.
4. Create the coordinates $(\widetilde{m}_1, \widetilde{m}_2)$ decoding $\widetilde{m}$.
5. Assign the values $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ for every grid $C_{\widetilde{m}}$.
6. Compute $(\widetilde{y}_1, \widetilde{y}_2)$ with the pair $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ using the method of the inverse function, that is $\widetilde{y}_1 = \widehat{Q}_{\widetilde{\theta}_1}(y_1|x_1)$ and $\widetilde{y}_2 = \widehat{Q}_{\widetilde{\theta}_2}(y_2|x_2)$.

All the estimates used on each of the previous steps have good asymptotic properties (consistency) under the usual exogeneity assumption (Koenker, 2005) [17]. Moreover, if the number of cells M is large enough the grid method fits better. However, the number of different quantile regressions that can be estimated with a finite sample size is limited. Portnoy (1991) [18] shows that this number is $O(n \cdot log(n))$. Nevertheless, this rate corresponds to the univariate case and to the best of our knowledge there is no study for the bivariate analysis. On the other hand, taking too many quantiles affects the consistency in the second step of the procedure because the probabilities of each grid are estimated with few observations. Therefore, there is a trade-off between the number of grids and the precision of the method. By the continuous mapping theorem, the method is expected to work well with relatively large sample sizes, provided that it allows to subdivide the enclosure into a greater number of grids.

3.3. Counterfactual Distributions

The proposed methodology can be used to generate counterfactual distributions due to a change on its determinants, as in Oaxaca-Blinder decompositions. Particularly, this proposal is in line with the literature initiated by Machado and Mata (2005) [10] and Melly (2005) [11]. Our contribution to this literature is to extend their method when there are two variables of interest (y_1, y_2) or some function of them. For example, the distribution of household per capita income is the variable of interest in the vast majority of the studies about inequality and/or poverty. If y_1 and y_2 respectively represent the head and spouse individual incomes, then the household income is the sum of them, plus the income of the other family members. After calculating the total level of income received by each household, this number is divided by the number of members in the household to obtain the household per capita income.

Assuming that incomes from other family members and those coming from non-labor sources are constant, the distribution of household per capita income depends only on the determinants of the couples income.[2] Formally, let y_t be the vector of household per capita income for all households observed in year t, then the income distribution can be represented as:

$$y_t = D(\beta_{1t}(\theta_1), \beta_{2t}(\theta_2), r_t(\theta_1, \theta_2))$$

That is, y_t is a function of the parameters of each income equation at year t as well as of the probabilistic relationship between the two conditional rankings represented by $r_t(\theta_1, \theta_2)$.

Let $I(\cdot)$ be any distributive indicator based on the vector of household incomes (e.g., Gini, Theil index, among others), then $I(y_t) - I(y_s)$ is the distributional change between the years t and s. A decomposition of this difference is an exercise of comparative statics where some income distribution determinants are changed and the others remain constant. The key is to build a set of counterfactual scenarios where some determinants are changed. The mechanism to do it is to generate replicates of the income distribution using the method explained in Section 2. For example, let's consider three counterfactual scenarios:

$$\begin{aligned} y_t^1 &= D(\beta_{1t}(\theta_1), \beta_{2t}(\theta_2), r_t(\theta_1, \theta_2)) \\ y_t^2 &= D(\beta_{1t}(\theta_1), \beta_{2t}(\theta_2), r_t(\theta_1, \theta_2)) \\ y_t^{12} &= D(\beta_{1t}(\theta_1), \beta_{2t}(\theta_2), r_t(\theta_1, \theta_2)) \end{aligned}$$

In the first equation, only the parameters of the household head have changed and this represent the first scenario. In the second, only those of the spouse have been modified, while in the third both parameter sets have changed. Then, if $I(y_t) - I(y_s)$ is the observed change in the distributive indicator, the effect of each scenario is:

$$y_t^1 - y_s = D(\beta_{1t}(\theta_1), \beta_{2s}(\theta_2), r_s(\theta_1, \theta_2)) - D(\beta_{1s}(\theta_1), \beta_{2s}(\theta_2), r_s(\theta_1, \theta_2)) \quad (3)$$

$$y_t^2 - y_s = D(\beta_{1s}(\theta_1), \beta_{2t}(\theta_2), r_s(\theta_1, \theta_2)) - D(\beta_{1s}(\theta_1), \beta_{2s}(\theta_2), r_s(\theta_1, \theta_2)) \quad (4)$$

$$y_t^{12} - y_s = D(\beta_{1t}(\theta_1), \beta_{2t}(\theta_2), r_s(\theta_1, \theta_2)) - D(\beta_{1s}(\theta_1), \beta_{2s}(\theta_2), r_s(\theta_1, \theta_2)) \quad (5)$$

---

[2] To incorporate non-labor income on a microsimulation exercise is not a simple task and depends mainly on the social policies applied in each country under analysis. See Badaracco (2014) [19] as an example for the countries in the Southern Cone of Latin America.

To ease notation, we have omitted the observable characteristics of the household head (x_1) and spouse (x_2). This obey to the fact that these determinants remain constant in our simulation exercise. Notwithstanding, our methodology admits counterfactual scenarios including isolated changes on those characteristics. In the terminology of Firpo *et al*. (2011) [20], the result of this exercise is called "characteristic effect"; while the scenarios proposed in Equations (3)–(5) are "parameter effects". The aim of this paper is to show the simulation methodology and the performance of implementing a simple exercise with different sample sizes. Therefore, to keep this analysis simple, only the counterfactual scenarios involving parameter changes were considered, separating the effect of both household members to explore the potential of the method.

4. Empirical Illustration

In this section we use real data as an application of the proposed methodology to generate counterfactual distributions of per capita household income. The model is defined by Equations (1) and (2) where y_1 and y_2 represent the labor earnings (in logs) of the household head and spouse, respectively. The vectors x_1 and x_2 are observed characteristics (age, education, gender, number of children) while ε_1 and ε_2 are terms representing the unobserved determinants of earnings. We focus our analysis on five countries in Latin America, particularly those belonging to the Southern Cone: Argentina, Brazil, Chile, Paraguay and Uruguay. The data come from household surveys collected by the statistical institutes of each country.[3]

We use three alternative methods to estimate the earning models. The first method is to use a seemingly unrelated regression model (SUR), in which the parameters in Equations (1) and (2) are estimated by OLS but allowing correlation between the error terms in both equations. The second case is the estimation of a quantile regression model (IQR), in which the assumption is that the error terms are independent. Finally, the outcomes from these methods are compared with those obtained by applying the methodology of estimating through quantile regressions but relating the model equations using the grids method (DQR).

The first exercise is to analyze the performance of the proposed methodology (DQR) relative to the other two strategies (SUR and IDR). We use an *ad-hoc* rule to choose the number of quantiles on each earning equation. This rule ensures that the number of observations on each grid will be around 40 in the case in which both equations are independent.[4] The reason behind choosing this rule is to try to get reliable estimates of each grid without losing the asymptotic properties.

Using these three methods, the model's coefficients are estimated in order to generate the joint distribution of labor earnings of the heads and spouses in a particular year. These earnings are used to build a new household per capita income and compute the Gini coefficient. Table 1 shows the results. The first panel of the table shows the Gini coefficient observed in each country, followed by the Gini coefficient of the simulated income from each method. The standard errors of each coefficient

---

[3] Table A1 on Appendix shows a brief description of the surveys used.

[4] Following this rule, the number of quantile estimates in each country are: 19 in Argentina, 39 in Brazil, 29 in Chile, 9 in Paraguay, and 15 in Uruguay.

are computed using 50 simulation replicates. Errors in the SUR model are generated 50 times from a bivariate normal distribution using all the estimated parameters. For the IQR, uniform random variables are generated independently, so that there is no conditional ranking association. Finally, under the DQR simulation, the values of the labor earnings of the head and spouse are obtained from the estimated probabilities in the grid method, namely considering the relationship between the two equations. The second panel in the table presents the mean square error (MSE) of each estimate with respect to the empirical distribution:

Table 1. Performance.

	Argentina	**Brazil**	**Chile**	**Paraguay**	**Uruguay**
Gini					
Observed	48.78	56.33	54.65	52.96	46.81
OLS	49.12	54.61	52.54	53.19	46.22
	(0.03)	(0.01)	(0.02)	(0.05)	(0.02)
IQR	48.25	55.88	54.22	52.02	46.30
	(0.01)	(0.00)	(0.02)	(0.02)	(0.01)
DQR	48.34	56.05	54.38	52.32	46.44
	(0.01)	(0.00)	(0.02)	(0.01)	(0.01)
ECM					
OLS	0.16	2.96	4.46	0.18	0.38
IQR	0.28	0.20	0.20	0.90	0.26
DQR	0.20	0.08	0.09	0.43	0.14
Obs.	16571	78188	45915	3337	10907

Note: Standard errors in parentheses. The observed Gini coefficient corresponds to the initial year (see Table A1). The number of observations corresponds to the sample of households that have both head and spouse in the initial year.

The SUR method has the lowest MSE for Argentina and Paraguay, followed closely by the DQR method. Therefore, in these two cases, using the conditional mean with an assumption of normality in errors fits relatively well to the real data. In the cases of Brazil, Chile and Uruguay the method that achieves the lowest MSE is the DQR, followed by the IQR method. As discussed above, these results suggest that the DQR method requires a certain amount of observations to achieve relatively good performance. However, in the case of Uruguay, which has a smaller sample than Argentina, DQR method has the lowest MSE. Then, this methodology may also depend on how well the model fits to the empirical distribution. However, large sample sizes should improve the approximation of the DQR method.

The next step in this section is to perform the micro-decomposition discussed in Section 3.3 in order to compare the results obtained with the three methods. As an illustration, we estimate the parameters effect in the equations of labor earnings. Table 2 shows the results of this exercise: the row Δ Obs contains the observed variation in the Gini coefficient for the entire period. The first panel in the table shows the parameters effect in the earning equation of the head, the second presents the effect corresponding to the spouse equation, and the third panel shows the decomposition of the parameters effect in both equations.

The greatest variation in the Gini coefficient corresponds to Argentina, with a fall of almost 8 points, followed by Uruguay with about 7 points. Brazil and Chile also show a reduction in inequality in terms

of the Gini coefficient with values slightly higher than 4 points, while in the case of Paraguay an increase of around a point is observed. The estimation results are interpreted as follows: the value -1.0 of the parameters effect in the equation of the spouse in Argentina under the SUR model indicates that if all that would have changed between 2004 and 2012 were the parameters governing the equation of the spouse of the household, the Gini coeficient would had been reduced by a point.

Table 2. Parameters Effects

	Argentina 2004–2012	**Brazil 2004–2012**	**Chile 2003–2011**	**Paraguay 2007–2011**	**Uruguay 2004–2012**
Δ Obs	-7.8	-4.1	-4.4	1.0	-6.8
Head					
OLS	-1.0	-1.2	1.9	0.0	-1.0
	(0.01) ***	(0.00) ***	(0.01) ***	(0.01) ***	(0.01) ***
IQR	-0.7	-0.7	-1.1	-0.5	-1.3
	(0.01) ***	(0.00) ***	(0.01) ***	(0.02) ***	(0.01) ***
DQR	-0.6	-0.6	-1.1	-0.8	-1.4
	(0.01) ***	(0.00) ***	(0.01) ***	(0.02) ***	(0.01) ***
Spouse					
OLS	-1.2	-0.7	1.0	0.3	-0.6
	(0.01) ***	(0.00) ***	(0.01) ***	(0.01) ***	(0.01) ***
IQR	-0.3	-0.1	0.4	-0.1	0.2
	(0.00) ***	(0.00) ***	(0.01) ***	(0.02) ***	(0.00) ***
DQR	-0.4	-0.1	0.3	-0.3	0.2
	(0.00) ***	(0.00) ***	(0.00) ***	(0.02) ***	(0.00) ***
Both					
OLS	-2.3	-1.7	2.9	0.3	-1.6
	(0.01) ***	(0.01) ***	(0.01) ***	(0.01) ***	(0.01) ***
IQR	-1.1	-0.4	-0.7	-0.5	-1.0
	(0.01) ***	(0.00) ***	(0.01) ***	(0.02) ***	(0.01) ***
DQR	-1.1	-0.3	-0.8	-0.9	-1.1
	(0.01) ***	(0.00) ***	(0.01) ***	(0.03) ***	(0.01) ***

Note: Standard errors in parentheses. * significative at 10%; ** significative at 5%; *** significative at 1%.

The greatest discrepancies among methodologies belong to the SUR method, while the IQR and DQR do not differ significantly from each other. The differences between the DQR and IQR are between 0 and 0.1 points (in absolute value) in the countries where the DQR achieves the lowest MSE. This result suggests a significant difference in terms of effects (between 0% and 30% in some cases). However, the economic significance of these differences is small (one tenth point of the Gini). The case of Paraguay shows a potential weakness in the DQR method when there are too few observations available. Since DQR has the lowest MSE in this country, the differences with the other methods suggest that with small samples the DQR method could present a potential bias in the estimated effects.

5. Conclusions

This paper proposes a method to incorporate the intra-household relationship between the labor incomes of the head and the spouse in decomposition studies. The paper closely follows the articles of Machado and Mata (2005) [10] and Melly (2005) [11]. We try to extend these papers by incorporating the correlation of intra-household income modeled by a simultaneous equation system. The key idea in our proposal is to associate conditional quantiles by adapting an standard method for generating random variables: the grid method.

The complexity associated with the joint employment decisions in a household leads us to focus our analysis on the behavior of the head and spouse, independently of the decisions of the rest of the family members. Furthermore, our model only analyzes the determination of labor earnings, assuming all other sources of income remain unchanged. Incorporating these other sources is an exercise that does not allow certain generalizations because non-labor income depends mainly on the social policies applied in each country (Badaracco, 2014 [19]).

An empirical application performing a simple decomposition exercise was implemented by using data from household surveys for the Southern Cone countries in Latin America. The counterfactual scenarios considered consisted on a change in the parameters in the labor earnings equations in two different moments in time. The results show that, in general, incorporating the interaction of household incomes substantially improves the goodness of fit to the empirical income distribution. Also, using quantile regression can dramatically change the results of the simulation exercise. However, although the introduction of correlation in incomes yields different results, the economic significance seems to be minor. The comparative exercise among different surveys shows that the performance of the method clearly depends on the sample size by limiting the number of grids. Moreover, given the sample size, the goodness of fit of the semiparametric method seems to be another key point.

The paper omits some important issues related to the estimation of earnings equations such as sample selection and endogeneity of covariates (e.g., education). The main reason for doing this is that our target is to propose a methodology for the generation of counterfactual distributions, showing their application using standard regression methods developed in the literature. Solving all these problems requires the use of more specific methodologies that are still under development such as those in Buchinsky (2001) [21] and Chernozhukov and Hansen (2006) [22]. Exploring the performance of the proposed method under these estimation techniques is postponed for future research.

Acknowledgments

The authors would like to thank the Center for Distributive, Labor and Social Studies (CEDLAS), at Facultad de Ciencias Económicas, Universidad Nacional de La Plata (UNLP). All errors and omissions are the sole responsibility of the authors.

Author Contributions

The authors contributed jointly to the paper.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix

Table A1. Household Surveys.

Country	Survey	Acronim	Years
Argentina	Encuesta Permanente de Hogares-Continua	EPH-C	2004–2012 (1)
Brazil	Pesquisa Nacional por Amostra de Domicilios	PNAD	2004–2012
Chile	Encuesta de Caracterización Socioeconómica Nacional	CASEN	2003–2011
Paraguay	Encuesta Permanente de Hogares	EPH	2004–2011
Uruguay	Encuesta Continua de Hogares	ECH	2004–2012

Note: (1) Second Semester.

References

1. Alejo, J.; Badaracco, N. Distribuciones contrafactuales en modelos bivariados. Un enfoque de cuantiles condicionales. In Annals of the XLIX Annual Meeting of the Argentinian Association of Political Economy, Posadas, Argentina, November 2014.
2. Oaxaca, R. Male-female wage differentials in urban labor market. *Int. Econ. Rev.* **1973**, *14*, 693–709.
3. Blinder, A. Wage discrimination: Reduced form and structural estimate. *J. Hum. Resour.* **1972**, *8*, 436–455.
4. Chiappori, P.-A. Collective Labor Supply and Welfare. In *Journal of Political Economy*; University of Chicago Press: Chicago, IL, USA, 1992; Volume 100, pp. 437–467.
5. Blundell, R.; Chiappori, P.-A.; Magnac, T.; Meghir, C. *Collective Labour Supply: Heterogeneity and Nonparticipation*; IDEI Working Papers 373; Institut d'Economie Industrielle (IDEI): Toulouse, France, 2005.
6. Klaveren, C.; van Praag, B.M.S.; van den Brink, H.M. A Public Good Version of the Collective Household Model: An Empirical Approach with an Application to British Household Data. In *Review of Economics of the Household*; Springer US: New York, NY, USA, 2008; Volume 6, pp. 169–191.
7. Browning, M.; Bourguignon, F.; Chiappori, P.; Lechene, V. Income and Outcomes: A Structural Model of Intrahousehold Allocation. *J. Political Econ.* **1994**, *102*, 1067–1096.
8. Marchionni, M.; Gasparini, L. Tracing out the effects of demographic changes on the income distribution. *J. Econ. Inequal.* **2007**, *5*, 97–114.
9. Galiani, S.; Weinschelbaum, F. Modeling informality formally: Households and firms. *Econ. Inq.* **2012**, *50*, 821–838.
10. Machado, J.A.; Mata, J. Counterfactual decomposition of changes in wage distributions using quantile regression *J. Appl. Econom.* **2005**, *20*, 445–465.

11. Melly, B. Decomposition of Differences in Distribution Using Quanitle Regressions. *Labour Econ.* **2005**, *12*, 577–590.
12. Firpo, S.; Fortin, N.; Lemieux, T. Unconditional Quantile Regressions. *Econometrica* **2009**, *77*, 953–973.
13. Devroye, L. *Non-Uniform Randor Variable Generation*; Springer-Verlag, New Inc.: New York, NY, USA, 1986; Chapter 2.
14. Nelsen, R.B. *An Introduction to Copulas*; Springer: New York, NY, USA, 1986; Chapter 2.
15. Hörmann, W.; Leydold, J.; Derflinger, G. *Automathic Nonuniform Random Variate Generation*; Springer: Berlin, Germany, 2004.
16. Koenker, R.; Bassett, G., Jr. Regression Quantiles. *Econometrica* **1978**, *46*, 33–50.
17. Koenker, R. *Quantile Regression*; Cambridge University Press: New York, NY, USA, 2005.
18. Portnoy, S. Asymptotic behavior of the number of regression quantile breakpoints. *SIAM J. Sci. Stat. Comput.* **1991**, 12, 867–883.
19. Badaracco, N. *Fecundidad y Cambios Distributivos en América Latina*. CEDLAS Working paper 173; CEDLAS, Universidad Nacional de La Plata: La Plata, Argentina, 2014.
20. Firpo, S.; Fortin, N.; Lemieux, T. Decomposition Method in Economics. In *Handbook of Labor Economics*; Elsevier: North Holland, Netherlands, 2011; Volume 4A.
21. Buchinsky, M. Quantile regression with sample selection: Estimating women's return to education in the U.S. *Empir. Econ.* **2001**, *26*, 87–113.
22. Chernozhukov, V.; Hansen, C. Instrumental quantile regression inference for structural and treatment effect models. *J. Econom.* **2006**, *132*, 491–525.