Anales del VII CONGRESO NACIONAL DE ESTUDIANTES DE POSTGRADO EN ECONOMÍA (CNEPE)

DEPARTAMENTO DE ECONOMÍA
UNIVERISDAD NACIONAL DEL SUR

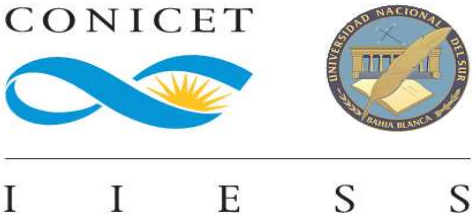
INSTITUTO DE INVESTIGACIONES ECONÓMICAS Y SOCIALES DEL SUR (IIESS)
CONICET – UNIVERSIDAD NACIONAL DEL SUR

Bahía Blanca

Mayo de 2015

ISBN: 978-987-1648-39-9

Departamento de Economía



Distribuciones contrafactuales en modelos bivariados.
Un enfoque de cuantiles condicionales.

Alejo, Javier y Badaracco, Nicolás

Distribuciones contrafactuales en modelos bivariados. Un enfoque de cuantiles condicionales

Javier Alejo
(CEDLAS-UNLP y CONICET)
javier.alejo@depeco.econo.unlp.edu.ar

Nicolás Badaracco[†]
(CEDLAS-UNLP)
nbadaracco@cedlas.org

Versión Preliminar
Agosto 2014

Abstract

Este trabajo propone un método para incorporar las relaciones intra-hogar en la determinación del ingreso del hogar siguiendo la línea iniciada por Machado y Mata (2005) y Melly (2005) usando el método de regresión por cuantiles. Mediante la estimación de modelos de regresión estándar el trabajo propone una metodología para la simulación de rankings condicionales asociados mediante una adaptación de un método de generación de variables aleatorias. Una aplicación empírica muestra que incorporar la interrelación de los ingresos de los hogares mejora sustancialmente la bondad de ajuste a la distribución empírica. Sin embargo, el ejercicio de descomposición es menos concluyente en cuanto al desempeño del método, dependiendo fundamentalmente del tamaño de la muestra y el ajuste de los modelos de regresión.

Abstract

This paper proposes a method for adding intra-household relations in the determination of household income according to the methodologies of Machado and Mata (2005) and Melly (2005) using quantile regression technique. By estimating standard regression models this paper proposes a methodology for simulating conditional rankings related by adapting a method for generating random variables. An empirical application shows that incorporating the interplay of household income substantially improves the goodness of fit to the empirical distribution. However, the decomposition exercise is less conclusive regarding the performance of the method, largely depending on the sample size and adjustment of regression models.

JEL Classification: C13, C14, C15, C31

Keywords: distribuciones contrafactuales, regresión por cuantiles, integración numérica, método de cuadrículas, mercado laboral, distribución del ingreso.

[†]Centro de Estudios Distributivos, Laborales y Sociales (CEDLAS), Facultad de Ciencias Económicas, Universidad Nacional de La Plata. Email: javier.alejo@depeco.econo.unlp.edu.ar. Todos los errores y omisiones son de exclusiva resposabilidad de los autores.

1 Introducción

La mayoría de los trabajos empíricos que analizan los efectos en los determinantes de la distribución del ingreso utilizando metodologías de descomposición al estilo Oaxaca-Blincer (1973) se focalizan en la distribución salarial de un único individuo, asumiendo que todas las decisiones laborales son tomadas en forma aislada o independiente del resto de los miembros del hogar. Sin embargo, la literatura de emparejamiento selectivo (*assortative mating*) ofrece evidencia numerosa sobre las distintas interrelaciones de variables individuales entre los miembros dentro del hogar, tales como sus niveles de educación, el ingreso laboral, la cantidad de horas dedicadas al trabajo, etc. Las repercusiones de un cambio en las decisiones o circunstancias económicas de algún miembro del hogar es probable que lleve a algún ajuste en el comportamiento de los otros. Por lo tanto, omitir este aspecto en la determinación del ingreso laboral de un hogar ofrece una visión que es potencialmente sesgada o poco realista.

Dada la gran complejidad que involucra el análisis de la decisión laboral conjunta de todos los miembros de un hogar, un supuesto usual es focalizarse solamente en las decisiones del jefe de hogar y su cónyuge, suponiendo que el resto de los miembros no cambiará su comportamiento o bien los mismos no tienen un gran impacto sobre la determinación del salario familiar. Modelar una ecuación de ingresos para ambos miembros e incorporar sus interacciones para poder estudiar los movimientos de la distribución del ingreso familiar requiere de una metodología que permita generar distribuciones contrafactuales ante cambios hipotéticos en sus determinantes, ya sea los del jefe o su cónyuge.

Incluir este aspecto en el análisis de descomposiciones no es una tarea menor. Algunos ejemplos de autores que han modelado los determinantes del ingreso que incluyen decisiones laborales al interior del hogar son Browning et al. (1994), Gasparini y Marchionni (2007), Galiani y Weischenbaum (2012), entre otros. Por lo general esta literatura realiza varios supuestos paramétricos y/o distributivos tales como normalidad en los determinantes inobservables del ingreso, los cuales pueden ser estrictos y no ser representativos de la verdadera distribución de ingresos. Otro aspecto es que por lo general utilizan modelos diseñados a estimar aspectos puntuales de la distribución, en particular la media condicional. Si bien el avance de la literatura de cuantiles permite explorar aspectos más allá de los efectos promedio, el grueso de esta literatura de descomposiciones basadas en comparaciones con distribuciones contrafactuales aún consideran ecuaciones de ingresos que modelan los ingresos laborales de un único individuo de manera independiente. Tal es el caso de los trabajos de Machado y Mata (2005), Melly (2005) y Firpo *et al.* (2009). Este trabajo intenta avanzar en ese sentido proponiendo una metodología para generar distribuciones contrafactuales en distribuciones bivariadas con un enfoque de cuantiles condicionales similar en la línea propuesta por Machado y Mata (2005) y Melly (2005).

La idea central del trabajo consiste en mostrar que el problema de generar distribuciones contrafactuales para los ingresos del jefe y el cónyuge no es otra cosa que un ejercicio de integración numérica en la cual interviene un mecanismo de generación conjunta de un par de variables aleatorias a través de sus distribuciones marginales. Una vez establecido dicho mecanismo es posible esbozar un procedimiento para obtener replicas o realizaciones de la distribución conjunta tomando como insumo la estimación de las distribuciones empíricas para cada miembro y la asociación de sus rankings. Sin embargo, el hecho de que los ingresos estén asociados a un conjunto de características observables hace que sea necesario introducir cierta estructura a la distribución condicional de los ingresos. Es aquí donde aparecen los cuantiles condicionales, que no son otra cosa que la contracara de la distribución acumulada y son fácilmente estimables mediante métodos estándar. La posibilidad

de modelar mediante una forma funcional a los cuantiles condicionales permite capturar indirectamente los efectos heterogéneos de los inobservables sobre toda la distribución. Finalmente, el paso final del método propuesto consiste en incorporar relación existente entre los ingresos de ambos miembros del hogar basado en una asociación probabilistica de los rankings condicionales.

El trabajo se ordena de la siguiente manera, en la Sección 2 se discute brevemente el método a utilizar para simular realizaciones de variables aleatorias bivariadas basadas en distribuciones marginales, mientras en la Sección 3 extiende esta idea para el caso de una distribución conjunta condicional y su aplicación al caso de distribuciones contrafactuales. La Sección 4 muestra una aplicación empírica del método usando un ejemplo sencillo con datos de distintos países del Cono Sur. Finalmente, la Sección 5 discute los resultados obtenidos y los alcances de la metodología.

2 Generación de variables aleatorias

Obtener realizaciones de variables aleatorias univariadas es relativamente sencillo, existen una variedad de métodos para ello. Uno de los más utilizados es el método de función acumulada inversa: sea U una variable aleatoria uniforme $U(0,1)$, entonces la transformación $F_Y^{-1}(U)$ genera una variable aleatoria con distribución $F_Y(y)$. Por lo tanto, el procedimiento consiste simplemente en tomar una realización de una variable aleatoria uniforme u y computar el $u-ésimo$ cuantil $Q_Y(u) \equiv F_Y^{-1}(u)$. En el caso de las variables enteras, la lógica es similar al caso continuo (Devroye, 1986).

El caso bivariado es más complejo ya que debe considerarse la relación estadística entre ambas variables. Al igual que en el caso univariado existen una diversidad de métodos; uno de ellos es el método de las cuadrículas (*grid method*) que permite en cierta forma adaptar el problema bivariado a los métodos univariados como el de la función acumulada inversa. Antes de explicar el método es conveniente dar algunas definiciones.

> **Definición** (función codificadora). Sea m_1 y m_2 dos variables enteras en el dominio $[1, M_1]$ y $[1, M_2]$, respectivamente. Se define a la función indicadora a $m = T(m_1, m_2)$ como $m = M_1 m_1 + m_2$.

La imagen de esta función es $Im(T) = [M_1 + 1, M_1^2 + M_2]$. La propiedad más interesante de la función codificadora es que a cada par ordenado (m_1, m_2) le corresponde un único valor de m, por lo tanto cada coordenada es identificada a través de la siguiente función decodificadora:

> **Propiedad** (funciones decodificadoras). Sea $m \in Im(T)$ de una función codificadora, entonces las coordenadas m_1 y m_2 pueden obtenerse mediante las funciones decodificadoras: $m_1 = [m/M_1]$ y $m_2 = m - M_1 m_1$.

Un último elemento necesario es definir al conjunto de cuadrículas de un recinto $A \subset \mathbf{R}_+^2$.

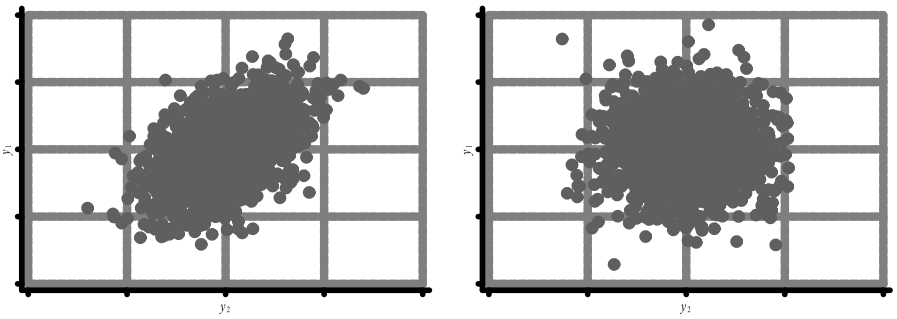
> **Definición** (cuadrículas C_m). Sea el recinto $A \subset \mathbf{R}_+^2$, se define a una cuadrícula $C_m \subset A$ como $C_m \equiv \{(y_1, y_2) \in A |\ a_{m_1-1} < y_1 < a_{m_1} \wedge b_{m_2-1} < y_2 < b_{m_2}\}$ con $m_1 = 2, \ldots, M_1$, $m_2 = 2, \ldots, M_2$ y a_{m_1} y b_{m_2} son valores elegidos de forma tal que $a_1 < a_2 < \ldots < a_{M_1}$; $b_1 < b_2 < \ldots < b_{M_2}$ y $\bigcup_{m \in Im(h)} C(m) = A$.

Finalmente, considérese dos variables aleatorias $(y_1, y_2) \in A$ con una función de densidad conjunta $f(y_1, y_2)$. Para generar una realización de un vector $(\tilde{y}_1, \tilde{y}_2)$ que respete la distribución poblacional $f(y_1, y_2)$ el método de cuadrículas sigue los siguientes pasos:

1. Subdivide a la región A en cuadrículas C_m, con $m \in Im(T)$.
2. Calcula la masa de probabilidad en cada cuadrícula $p_m = Pr[(y_1, y_2) \in C_m]$ para cada m.
3. Genera $\tilde{m}$ que es la realización de una variable aleatoria entera univariada con distribución de probabilidades p_m, calculada en el paso anterior.
4. Decodifica $\tilde{m}$ en los valores $(\tilde{m}_1, \tilde{m}_2)$.
5. Obtiene la realización de $(\tilde{y}_1, \tilde{y}_2)$ asignando valores dentro de la cuadrícula $C_{\tilde{m}}$.

La Figura 2.1 muestra dos ejemplos de cómo trabaja el método: en el caso *(a)* muestra dos variables aleatorias que tienen una relación positiva mientras que el caso *(b)* indica una relación independiente.

Figure 2.1: Variables aleatorias y correlación



(a) *(b)*

Las líneas punteadas delimitan las cuadrículas en las cuales se subdivide al recinto de en donde se mueven ambas variables aleatorias. Claramente, en el primer caso la mayor masa de probabilidad (medida por la proporción de puntos que caen en cada cuadrícula) está concentrada en la diagonal dada por la bisectriz, mientras que en el segundo caso no hay un patrón definido para la probabilidad conjunta. Lógicamente, mientras mayor sea el número de recintos, mejor será la aproximación del método (Hörmann *et al.*, 2004). La forma en la cual el método incorpora la relación estadística entre y_1 e y_2 es mediante la probabilidad de cada recinto.

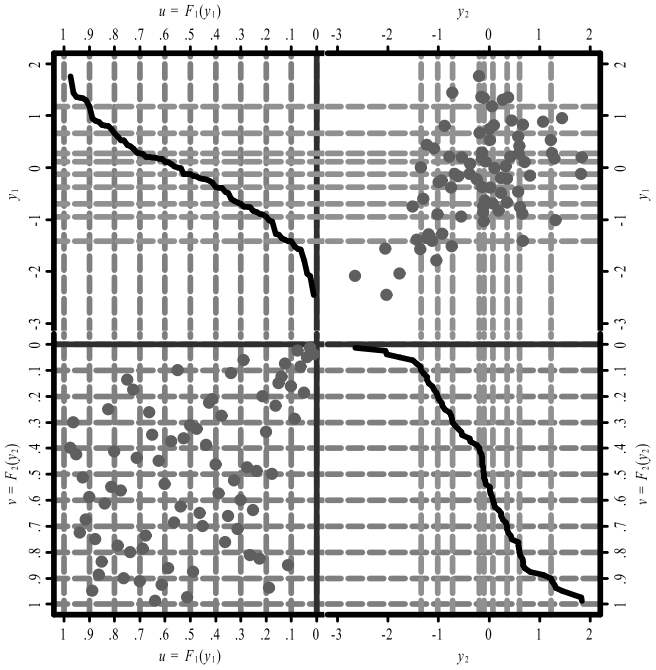
Un último aspecto es notar que las cuadrículas pueden ser definidas por medio de sus cuantiles marginales definiendo a $u_{m_1} \equiv F_1(a_{m_1})$ y $v_{m_2} \equiv F_2(b_{m_2})$. La razón de esta equivalencia es que las distribuciones acumuladas son simplemente transformaciones monótonas crecientes del soporte de la variable. En otras palabras, el valor $F_j(.)$ no es otra cosa que la posición en un ranking que surge como ordenar a y_j de manera creciente. Esto establece una relación biunívoca entre el cualquier valor de y_j y su ranking. Luego, la cuadricula C_m se puede escribir como:

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

La Figura 2.2 muestra esta equivalencia en la definición de las cuadrículas. El gráfico superior del lado derecho muestra la nube de puntos en el plano (y_1, y_2), mientras que el gráfico inferior izquierdo

muestra su contracara en el plano $(F_1(y_1), F_2(y_2))$. Los gráficos en los otros dos cuadrantes muestran la probabilidad acumulada de cada variable vista únicamente como una distribución univariada (distribución marginal). Nótese que si bien las escalas de los recintos es diferente, cada punto que pertenece a una cuadrícula en el cuadrante superior derecho tiene su correspondiente cuadrícula en el cuadrante inferior izquierdo.

Figure 2.2: *Equivalencia de cuadrículas*



Es decir, definir las cuadrículas en el plano de la probabilidad marginal es equivalente a definirlas en el plano de los niveles de ambas variables. El punto interesante de ver a las cuadrículas de esta manera es que permite adaptar este método al contexto de cuantiles condicionales, dado que es precisamente el objeto de estimación de los métodos de *quantile regression*, construyendo las correspondencias entre las cuadrículas de probabilidad y los cuantiles condicionales y en consecuencia permitiendo asociar a sus rankings marginales.

3 Modelo condicional bivariado

3.1 Población

Veamos ahora el caso en donde la distribución de (y_1, y_2) depende de un grupo de regresores (x_1, x_2). En particular, considérese el siguiente modelo lineal para el par variables aleatorias (y_1, y_2):

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

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

donde los vectores x_1 y x_2 son regresores observables y los términos de error tienen una densidad

conjunta $f(\varepsilon_1, \varepsilon_2|x)$, con $x \equiv (x_1, x_2)$. Este mismo modelo puede representarse bajo la forma de cuantiles condicionales:

$$\begin{aligned} Q_{y_1|x_1}(\theta_1) &= x_1'\beta_1(\theta_1) \\ Q_{y_2|x_2}(\theta_2) &= x_2'\beta_2(\theta_2) \end{aligned}$$

donde (θ_1, θ_2) son dos variables aleatorias en el recinto $A \in [0,1] \times [0,1]$ con una densidad conjunta dada por la cópula de $f(\varepsilon_1, \varepsilon_2|x)$.

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

Bajo este contexto, generar valores aleatorios para un vector (y_1, y_2) dado los valores de x surge como una extensión simple del método de cuadrículas explicado en la sección anterior. Por simplicidad considérese que $M_1 = M_2 = M$ y que los valores de corte en A son equiespaciados, es decir $u_{m_1} = m_1/(M+1)$ y $v_{m_2} = m_2/(M+1)$, por lo tanto:

1. Subdividir la región del soporte definida por $[0,1] \times [0,1]$ en las cuadrículas C_m para $m \in Im(T)$.
2. Calcular la probabilidad de cada cuadrícula $p_m = \Pr[(\theta_1, \theta_2)|x \in C_m]$ para cada m.
3. Generar $\widetilde{m}$ que es la realización de una variable aleatoria entera univariada con la distribución de probabilidades p_m , calculada en el paso anterior.
4. Decodificar $\widetilde{m}$ en las coordenadas $(\widetilde{m}_1, \widetilde{m}_2)$.
5. Obtener la realización $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ asignándole valores dentro de la cuadricula $C_{\widetilde{m}}$.
6. Generar $(\widetilde{y}_1, \widetilde{y}_2)$ con el par $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ recurriendo al método univariado de la función inversa $\widetilde{y}_1 = Q_{\widetilde{\theta}_1}(y_1|x_1)$ y $\widetilde{y}_2 = Q_{\widetilde{\theta}_2}(y_2|x_2)$.

Hasta aquí, todos los elementos utilizados en cada etapa del procedimiento son valores poblacionales, los cuales no son observables y debe recurrirse a estimaciones empíricas para obtenerlos. La sección siguiente profundiza sobre este aspecto cuando se cuenta con una muestra aleatoria en lugar de datos poblacionales.

3.2 Estimación muestral

Para generar una réplica de (y_1, y_2) a partir de una muestra de aleatoria se aplica el mismo procedimiento que en los párrafos anteriores pero con elementos muestrales, es decir, se debe recurrir a estimaciones de los cuantiles condicionales $\widehat{Q}_{\theta_j}(y_j|x_j) = x_j'\widehat{\beta}_j(\theta_j)$ para $j = 1, 2$; para una lista de valores, por ejemplo $\theta = 0.05, 0.10, ..., 0.90, 0.95$. La referencia clásica para encontrar una estimación consistente de $\widehat{\beta}_1(\theta_1)$ y $\widehat{\beta}_2(\theta_2)$ es Koenker y Basset (1978).

Luego, las etapas del método de cuadrículas para el caso en donde se trabaja con estimaciones muestrales son las siguientes:

1. Construir $\widehat{C}_m$ delimitando las cuadrículas con $x_{2i}'\widetilde{\beta}_1(a_m)$ y $x_{2i}'\widetilde{\beta}_2(b_m)$.

2. Estimar $\widehat{p}_m = \widehat{\Pr}(C_m) = n^{-1} \sum_{i=1}^{n} 1(i \in \widehat{C}_m)$, para cada m.

3. Generar realizaciones de una variable aleatoria entera $\widetilde{m}$, con las probabilidades $\widehat{p}_m$.

4. Decodificar $\widetilde{m}$ en las coordenadas $(\widetilde{m}_1, \widetilde{m}_2)$.

5. Asignar $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ dentro de los valores de la cuadricula $C_{\widetilde{m}}$, para cada valor realizado $\widetilde{m}$.

6. Generar $(\widetilde{y}_1, \widetilde{y}_2)$ con el par $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ recurriendo al método de la función inversa: $\widetilde{y}_1 = \widehat{Q}_{\widetilde{\theta}_1}(y_1|x_1)$ y $\widetilde{y}_2 = \widehat{Q}_{\widetilde{\theta}_2}(y_2|x_2)$.

Todos los estimadores que intervienen en el procedimiento gozan de buenas propiedades asintóticas (consistentes) bajo los supuestos de exogeneidad usuales (Koenker, 2005). Por lo tanto, con un tamaño de muestra grande el método de cuadrículas debería comportarse aproximadamente igual a su contraparte poblacional. Además, si la cantidad de recintos M también es grande el método de cuadrículas funcionará mucho mejor. Sin embargo, la cantidad de cuantiles que se pueden estimar con una muestra de tamaño n no es infinita. Portnoy (1991) obtiene una tasa a la cual crece el número de cuantiles diferentes que pueden estimarse en una muestra de tamaño n. Sin embargo, esta tasa corresponde al caso univariado y hasta donde conocemos no se ha desarrollado un análisis similar al caso bivariado. En segundo lugar, la contracara de tomar demasiados cuantiles es que queden pocas observaciones para estimar adecuadamente las probabilidades de cada cuadricula en el segundo paso del método y por lo tanto hay un *trade-off* entre el número de cuadrículas y la precisión del método. Por lo tanto se esperaría que el método funcione relativamente bien con tamaños de muestra grandes dado que permite subdividir al recinto en una mayor cantidad de cuadrículas. Este trabajo se limitará a presentar evidencia comparativa del método y por lo tanto utiliza una regla *ad-hoc* que se explicará en la sección 4.

3.3 Distribuciones contrafactuales

La metodología previa puede utilizarse para generar distribuciones contrafactuales ante un cambio en sus determinantes, tal como se hace en los ejercicios de descomposiciones del tipo Oaxaca-Blinder (1973). En particular, esta propuesta se concatena en la línea iniciada por Machado y Mata (2005) y Melly (2005). El aporte a esa literatura es extenderla al caso en el cual hay dos variables de interés (y_1, y_2) o incluso si el objetivo es alguna otra variable que es una función de las mismas. Por ejemplo, en el caso de los estudios que analizan la desigualdad y/o la pobreza, el objetivo usual es la distribución del ingreso per capita familiar. Si y_1 y y_2 son los ingresos del jefe y del cónyuge, respectivamente, el ingreso familiar surge de la suma de ambos a la que se le adiciona el ingreso generado por otros miembros además de todos los ingresos no laborales. Una vez calculado el nivel total de ingreso que percibe cada hogar, se lo divide por el número de miembros para obtener el ingreso per cápita familiar.

Si suponemos que el ingreso que aportan otros miembros así como los provenientes de fuentes no laborales permanece inalterado, la distribución del ingreso per capita familiar depende únicamente de los determinantes del ingreso del jefe y el cónyuge.[1] Formalmente, si llamamos y_t al vector de ingreso per capita familiar de todos los hogares observado en un año t, entonces:

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

[1]No es trivial cómo modelar a los ingresos no laborales en un contexto de microsimulaciones ya que dependen fuertemente de las políticas sociales de cada país analizado. Ver por ejemplo Badaracco (2014) para un ejemplo en los países del Cono Sur.

Es decir, y_t es una función de los parámetros de la ecuación de salarios del jefe y el cónyuge de ese año, así como de la relación entre ambas distribuciones condicionales.

Sea $I(\cdot)$ un indicador distributivo cualquiera basado en el vector de ingresos, como es el caso del Gini o el índice de Theil, solo para mencionar algunos. Una descomposición de la diferencia del indicador entre un año t a otro s consta de un ejercicio de estática comparada en donde se cambian algunos determinantes de la distribución del ingresos, dejando al resto en sus niveles iniciales. La forma de hacerlo es reconstruyendo escenarios contrafactuales en donde se intercambian algunos determinantes y se generan réplicas de la distribución del ingreso mediante el procedimiento explicado en la sección 2. A modo de ejemplo, vamos a considerar tres escenarios contrafactuales:

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

es decir, en la primer ecuación solo se han modificado los retornos en la ecuación de ingresos del jefe, en la segunda sólo los del cónyuge y en la tercera ambos. Luego, si $I(y_t)$–$I(y_s)$ es el cambio observado en el indicador distributivo, el efecto de cada escenario es:

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

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

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

Nótese que se ha omitido como determinantes de la distribución de ingresos a las características observables del jefe (x_1) y cónyuge (x_2). Claramente, también podría plantearse escenarios contrafactuales en los cuales lo único que cambia son dichas características. En la terminología de Firpo *et al.* (2011), este ejercicio daría como resultado lo que se conoce como "efecto características" mientras que los escenarios propuestos en las ecuaciones (3), (4) y (5) miden el "efecto parámetros". Dicha omisión corresponde al hecho de que el objetivo de este trabajo es mostrar la metodología de simulación y su desempeño en la aplicación de un ejercicio simple con distintas muestras. Por lo tanto, sólo se consideraron escenarios contrafactuales intercambiando parámetros, separando el efecto de ambos miembros del hogar para explorar el potencial del método.

4 Ilustración empírica

En está sección se lleva a cabo una aplicación del método para generar distribuciones contrafactuales para un modelo de ingresos laborales de los jefes y cónyuges de los hogares. El modelo utilizado parte de las ecuaciones (1) y (2), donde y_1 e y_2 representan los ingresos laborales del jefe y cónyuge de un hogar respectivamente, x_1 y x_2 son vectores de características observadas (edad, educación, género, número de hijos), mientras que ε_1 y ε_2 son los errores del modelo. El análisis se lleva a cabo para cinco países de América Latina, en particular para aquellos que pertenecen al Cono Sur; estos son Argentina, Brasil, Chile, Paraguay y Uruguay. Los datos utilizados provienen de las encuestas de hogares recolectadas por los Institutos de Estadística de cada país.[2]

El primer ejercicio consiste en analizar el desempeño de la metodología de realizar las estimaciones por regresiones de cuantiles relacionando las ecuaciones del modelo por el método de cuadrículas

[2]En la Tabla 5.1 se presenta una breve descripción de las encuestas utilizadas.

(DQR), comparándolo con otros que usualmente son utilizados en la literatura. Un primer caso es utilizar un *seemingly unrelated regression model* (OLS), en el cual se realizan las estimaciones de las ecuaciones (1) y (2) por mínimos cuadrados ordinarios pero permitiendo que exista correlación entre los términos de error de ambas ecuaciones. Un segundo caso es la estimación del modelo por regresiones de cuantiles (IQR), en el cual el supuesto es que los términos de error son independientes. El número de cuantiles a estimar en los últimos dos métodos se decide de la siguiente manera: dado el número de observaciones de cada país, se estiman el número de cuantiles tal que en el caso de que las ecuaciones del jefe y cónyuge no tengan ninguna relación, el número de observaciones en cada cuadricula sea al menos mayor a 40.[3]

Utilizando estos tres métodos se estiman los coeficientes del modelo para generar la distribución del ingreso laboral de los jefes y cónyuges de un año en particular. Estos ingresos laborales son utilizados para construir un nuevo ingreso per capita familiar y computar un indicador distributivo, en particular el coeficiente de Gini. En la Tabla 4.1 se presentan los resultados. En el primer panel de la tabla figura el coeficiente de Gini observado en cada país, seguido por el coeficiente de Gini a partir de los ingresos simulados bajo los diferentes métodos. Los errores estándar de cada coeficiente son computados a partir de realizar la simulación de cada método 50 veces. Esto es, en el OLS, generar los errores del modelo a partir de una función normal bivariada con los parámetros estimados de la misma. En el caso del IQR, se generan variables aleatorias uniformes en forma independiente, de forma que se genere un ranking independiente. Finalmente en el DQR, se obtienen los valores de los ingresos laborales del jefe y cónyuge a partir de las probabilidades estimadas del método de cuadrículas para considerar la relación entre ambas ecuaciones. En el segundo panel de la tabla se presenta el Error Cuadrático Medio (ECM) de cada una de las tres estimaciones, como medida para comparar la bondad del ajuste las metodologías a la distribución empírica.

Table 4.1: Desempeño

	Argentina	Brasil	Chile	Paraguay	Uruguay
Gini					
Observado	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

Nota: Desvios estándar entre paréntesis. El coeficiente de Gini observado corresponde al año inicial (ver Tabla 5.1). La cantidad de observaciones corresponde a la muestra de hogares que cuentan tanto con jefe como cónyuge en el año inicial.

El método de OLS, permitiendo correlación entre los términos de error, tiene el menor ECM en Argentina y Paraguay, seguido muy de cerca por el método de DQR. Por lo tanto, en estos dos casos, utilizar a la media condicional junto con un supuesto de normalidad en los errores se adapta relativamente bien a los datos. En los casos de Brasil, Chile y Uruguay el método que logra el menor ECM es el de DQR, seguido por el método de IQR. Tal como se discutió anteriormente, estos resultados sugieren que el método de DQR requiere cierta cantidad de observaciones para

[3]Siguiendo esta regla, el número de cuantiles estimados en cada país es: Argentina 19, Brasil 39, Chile 29, Paraguay 9 y Uruguay 15.

lograr relativamente un buen desempeño. Sin embargo, en el caso de Uruguay, que presenta un menor número de observaciones que Argentina, el método de DQR presenta el menor ECM, de forma que esta metodología posiblemente dependa también de que tan buen ajuste logra el modelo a la distribución empírica. Sin embargo, de contar con suficientes observaciones el método DQR lograría una mejor aproximación.

El siguiente paso en esta sección consiste es realizar una microdescomposición discutida en la sección anterior a fin de comparar los resultados entre los tres métodos. A modo de ilustración se descompone el efecto parámetros de las ecuaciones de ingreso laboral. En la Tabla 4.2 se presentan los resultados de este ejercicio. En fila Δ Obs figura la variación observada en el coeficiente de Gini en el período considerado. En el primer panel de la tabla se presenta el efecto parámetros de la ecuación de ingresos laborales del jefe, en el segundo el correspondiente a la ecuación del cónyuge y finalmente en el tercer panel se presenta la descomposición del efecto parámetros de ambas ecuaciones.

La mayor variación en el coeficiente de Gini corresponde a Argentina, con una reducción de casi 8 puntos, seguido por Uruguay con cerca de 7 puntos, Brasil y Chile presentan también una reducción en la desigualdad en términos de este coeficiente con valores poco más que 4 puntos. Mientras que en el caso de Paraguay se observa un incremento del indicador de un punto. Los resultados de las estimaciones se interpretan de la siguiente forma: el valor -1.0 del efecto parámetros de la ecuación del cónyuge para Argentina bajo OLS indica que si lo único que hubiera cambiado entre 2004 y 2012 fueran los parámetros que gobiernan la ecuación de los jefes del hogar, el Gini se hubiera reducido en un punto.

Table 4.2: Efectos Parámetros

	Argentina 2004-2012	Brasil 2004-2012	Chile 2003-2011	Paraguay 2007-2011	Uruguay 2004-2012
Δ Obs	-7.8	-4.1	-4.4	1.0	-6.8
Jefes					
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)***
Cónyuges					
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)***
Ambos					
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)***

Nota: Desvios estándar entre paréntesis. * significativo al 10%, ** significativo al 5%, *** significativo al 1%.

Las mayores discrepancias entre las metodologías se presentan con el método OLS, mientras que el IQR y DQR no presentan diferencias demasiado relevantes. En los países donde el DQR logra el menor ECM se tiene que las diferencias entre el DQR y el IQR son entre 0 y 0.1 (en valor absoluto). Este resultado sugiere una diferencia importante en términos de los efectos (entre 0 y 30% en algunos casos), sin embargo, la relevancia económica de esta diferencia es pequeña, dado que nos referimos a un décimo de punto del coeficiente del Gini. En el caso de Paraguay abre la posibilidad a una potencial debilidad del método en el caso de contar con pocas observaciones.

Dado que en este país el DQR no logra el menor ECM las grandes diferencias que presenta el método con los otros usualmente utilizados sugiere que en los casos que se cuente con una cantidad baja de observaciones el DQR puede presentar posibles sesgos en los efectos estimados.

5 Conclusiones

Este trabajo propone un método para incorporar las relaciones intra-hogar en la determinación del ingreso laboral del jefe y el cónyuge del hogar. El documento intenta continuar la línea iniciada por Machado y Mata (2005) y Melly (2005) intentando incorporar la correlación de ingresos intra-hogar mediante uso de técnicas semi-paramétricas estándar tales como las regresiones para cuantiles condicionales. La idea central de la propuesta consiste en tratar de asociar los rankings condicionales mediante una adaptación el método de cuadrículas estándar en el computo de generación de variables aleatorias.

La complejidad en la determinación del comportamiento conjunto de las decisiones laborales de las familias nos lleva a focalizar el análisis en el comportamiento de los dos miembros más importantes del hogar en lo que respecta al aporte de ingresos. Además, el modelo considerado solamente estudia la determinación de ingresos laborales, suponiendo que el resto de las fuentes de ingresos permanece inalterado en cada escenario contrafactual propuesto. Incorporar este último aspecto dista de ser un ejercicio que permita cierta generalización debido a que los ingresos no laborales dependen fuertemente de las políticas sociales de cada país analizado (Badaracco, 2014).

Se implementó una aplicación empírica realizando un ejercicio simple de descomposición utilizando datos provenientes de encuestas de hogares realizadas por los distintos institutos de estadísticas en los países del Cono Sur durante la última década. Los escenarios contrafactuales considerados constan en un intercambio de parámetros de las ecuaciones de ingresos en dos momentos del tiempo. Los resultados muestran que, en general, incorporar la interrelación de los ingresos de los hogares mejora sustancialmente la bondad de ajuste a la distribución empírica del ingreso. En el análisis de descomposiciones, el uso de regresión por cuantiles puede cambiar drásticamente las conclusiones del ejercicio de simulación. Sin embargo, si bien se encontraron situaciones en donde la incorporación de la correlación de ingresos arroja resultados diferentes en la descomposición, la relevancia económica de los mismos parece ser menor. Del ejercicio comparado del método entre distintas encuestas se puede concluir que el desempeño del método depende claramente del tamaño de las muestras (dado que limita el número de cuadrículas a considerar), así como el ajuste del método semi-paramétrico (dado el tamaño de la muestra).

El trabajo omite algunas cuestiones sustanciales en la estimación de ecuaciones de ingresos tales como la selección muestral o los efectos de la endogeneidad de ciertas variables explicativas como la educación. Esto corresponde a que el objetivo del mismo es proponer la metodología para generación de distribuciones contrafactuales mostrando su aplicación mediante métodos de regresión estándar en la literatura. Solucionar tales problemas requiere del uso de metodologías aún en desarrollo tales como las Buchinsky (2001) y Chernozhukov y Hansen (2006). Dado lo preliminar de este trabajo, explorar el desempeño del método bajo esas técnicas de estimación queda postergado a futuras investigaciones.

Referencias

Badaracco, N. (2014).. "Fecundidad y Cambios Distributivos en América Latina." *Tesis de Maestría. FCE-UNLP.*

Blinder, A. (1973). "Wage discrimination: reduced form and structural estimate." *Journal of Human Resources*, 8(4): 436-455.

Browning, M., Bourguignon, F., Chiappori, P. y Lechene, V. (1994). "Income and Outcomes: A Structural Model of Intrahousehold Allocation." *Journal of Political Economy*, 102(6): 1067-1096.

Buchinsky, M. (1994). "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression." *Econometrica*, 62(2), 405-458.

Buchinsky, M. (2001). "Quantile regression with sample selection: Estimating women's return to education in the U.S." *Empirical Economics*, 26: 87-113.

Chernozhukov, V., y Hansen, C. (2006). "Instrumental quantile regression inference for structural and treatment effect models" *Journal of Econometrics*, 132: 491–525.

Devroye, L. (1986). "Non-Uniform Randor Variable Generation"*Springer-Verlag, New Inc.*, Capítulo 2.

Firpo, S., N. Fortin, y T. Lemieux (2009). "Unconditional Quantile Regressions." *Econometrica*, 77(3), 953-973.

Firpo, S., N. Fortin, y T. Lemieux (2011). "Decomposition Method in Economics" *Handbook of Labor Economics*, Volume 4A.

Galiani, S. y Weinschelbaum, F. (2012). "Modeling informality formally: households and firms." *Economic Inquiry*, 50(3), 821-838.

Hörmann, W., Leydold, J. y Derflinger, G. (2004). "Automathic Nonuniform Random Variate Generation." *Springer-Berlin*

Koenker, R., y Bassett Jr, G. (1978). "Regression Quantiles." *Econometrica*, 46(1), 33-50.

Koenker, R. (2005). "Quantile Regression." *Cambridge University Press*

Machado, J. A., y Mata, J. (2005). "Counterfactual decomposition of changes in wage distributions using quantile regression" *Journal of applied Econometrics*, 20(4), 445-465.

Marchionni, M. y Gasparini, L. (2007). "Tracing out the effects of demographic changes on the income distribution." *Journal of Economic Inequality*, 5(1): 97-114.

Melly, B. (2005). "Decomposition of Differences in Distribution Using Quanitle Regressions." *Labour Economics*, 12(4), 577-90.

Oaxaca, R. (1973). "Male-female wage differentials in urban labor market." *International Economic Review* 14(3): 693-709.

Portnoy, S. (1991). "Asymtotic Behavior of the Number of Regression Quantile Breakpoints." *SIAM Journal of Scientific and Statitistical Computing*, 12, 867-883.

Table 5.1: Encuestas de Hogares

País	Encuesta	Acronimo	Años
Argentina	Encuesta Permanente de Hogares - Continua	EPH-C	2004s2-2012s2
Brasil	Pesquisa Nacional por Amostra de Domicilios	PNAD	2004-2012
Chile	Encuesta de Caracterización Socioeconómica Nacional	CASEN	2003-2011
Paraguay	Encuesta Permanente de Hogares	EPH	2007-2011
Uruguay	Encuesta Continua de Hogares	ECH	2004-2012