

Oferta laboral de las mujeres casadas

Daniel Czarniewicz & Lucía Coudet

December 8, 2017

Objetivo

- ▶ El objetivo del presente trabajo es modelar la oferta laboral de las mujeres casadas en Uruguay para el año 2009.

Dos problemas que presentan este tipo de estudios:

- ▶ La variable a estudiar es una variable censurada. Se observan horas trabajadas únicamente para las mujeres que efectivamente trabajan. El resto de ellas tiene una oferta laboral, pero no es observable. Censura inferior en el valor 0.
- ▶ Existe un potencial problema de autoselección de la unidades del marco. Esto se traduce en que las observaciones que se acumulan en el punto de censura no lo hacen de forma aleatoria, sino que existen factores adicionales que determinan este comportamiento.

- ▶ Para solucionar estos problemas realizaremos una implementación bayesiana del método de Heckman.
- ▶ Este método busca incorporar en el modelo para la variable latente los problemas anteriores (variable censurada y autoselección muestral).
- ▶ Para ello Heckman propone un procedimiento de dos etapas de la siguiente forma:
 1. Estimar un modelo PROBIT para el margen extensivo usando toda la muestra.
 2. Estimar mediante MCO el margen intensivo usando únicamente las unidades que no se encuentran en el punto de censura.
- ▶ Lo que Heckman busca testear es la existencia de una correlación entre los márgens.
- ▶ Si dicha correlación no existiera, entonces no estaríamos frente a un problema de autoselección muestral, y la censura podría considerarse aleatoria.
- ▶ Por lo tanto, la segunda estimación deberá contemplar esta característica de los datos.

Los datos

- ▶ Se trabaja con un extracto de la ECH 2009 en el cual se seleccionan a todas las mujeres casadas en edad de trabajar.
- ▶ Para determinar dicha edad, se toman únicamente a las mujeres mayores de 25 años y menores a 60 años.
- ▶ Total de observaciones 19.919.

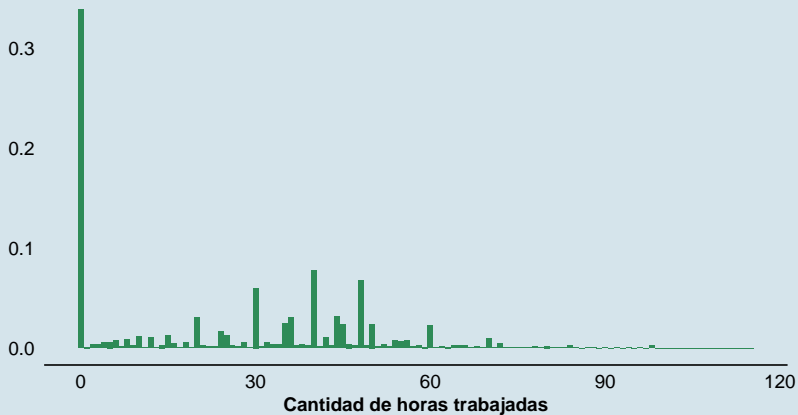
Variables utilizadas

La base de datos contiene información respecto de las siguientes variables:

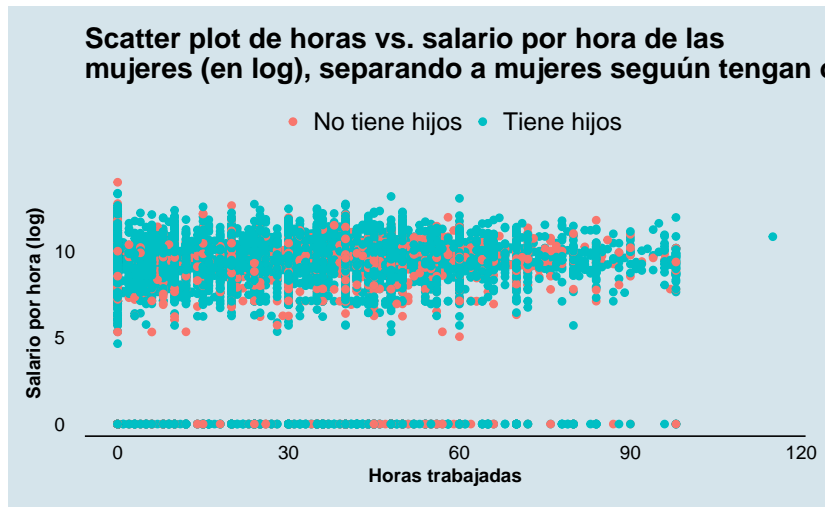
- ▶ horas: cantidad de horas semanales trabajadas.
- ▶ sal: logaritmo del salario por hora percibido.
- ▶ edad: años cumplidos.
- ▶ educ: años de educación completados.
- ▶ hijos: cantidad de hijos.
- ▶ salmar: logaritmo del salario del marido.
- ▶ expot: experiencia potencial del individuo. Se contruyó como la diferencia entre edad y educ + 6.

Descripción

Histograma de la variable horas trabajadas



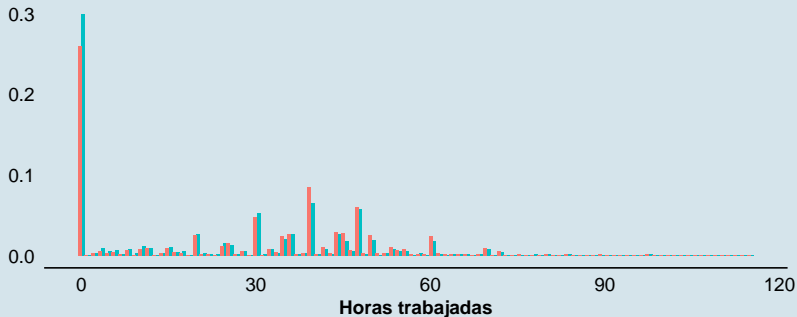
En el punto de censura las mujeres que tienen hijos predominan sobre las que no tienen hijos para todos los niveles de la variable sal.



El histograma de horas según la variable hijos muestra una tendencia a que las mujeres que tienen hijos trabajan menos horas.

Histograma de las horas trabajadas separando según las mujeres tengan hijos o no

■ No tiene hijos ■ Tiene hijos



El modelo

Dado que la variable de interés, horas, es un conteo de horas semanales trabajadas, se propone modelarla por una distribución Poisson de tasa θ_i :

$$\text{horas}_i \sim \text{Poisson}(\theta_i) I_{[\theta_i > 0]}$$

Se considera al parámetro θ_i como el producto entre la indicatriz de que la observación no esté en el punto de censura, $I_{[\mu_1 > 0]}$, y el margen intensivo μ_2 :

$$\theta_i = \mu_2 \times I_{[\mu_1 > 0]}$$

Márgen extensivo

Se construye la variable trabaja como la indicatriz de que horas; tome valores estrictamente positivos. Con esa variable como dependiente se estima el siguiente modelo probit:

$$\text{Probit}(\mu_{1i}) = \mathbf{z}'_i \boldsymbol{\gamma}$$

donde el vector de covariables \mathbf{z}_i incluye a las variables educ, expot, expot2, hijos y salmar.

Para el vector de parámetros asociados a las covariables, $\boldsymbol{\gamma}$, se selecciona una distribución previa Normal:

$$\gamma_j \sim \text{Normal}(0; 1, 6^2) \quad \forall j$$

Se construye la variable lambda de Heckman como la inversa del ratio de Mills:

$$\hat{\lambda}_i^k = \frac{\phi(\mathbf{z}'_i \gamma^k)}{\Phi(\mathbf{z}'_i \gamma^k)} \quad \forall i = 1; \dots; n \quad \forall k = 1; \dots; S$$

$$\hat{\lambda}_i = \frac{1}{S} \sum_{k=0}^S \hat{\lambda}_i^k \quad \forall i = 1; \dots; n$$

El motivo por el cual el lambda de Heckman es relevante en esta clase de problemas esta vinculado a la esperanza de un modelo Tobit. El resultado fundamental es que:

$$E(y_i|y_i^* > 0; \mathbf{x}_i) = \mathbf{x}_i' \boldsymbol{\beta} + \sigma \frac{\phi\left(\frac{\mathbf{x}_i' \boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{x}_i' \boldsymbol{\beta}}{\sigma}\right)} = \mathbf{x}_i' \boldsymbol{\beta} + \sigma \lambda_i$$

donde y_i^* es la variable latente tal que $y_i^* = \mathbf{x}_i' \boldsymbol{\beta} + u_i$ donde $u_i \sim \text{Normal}(0; \sigma^2)$.¹

¹El detalle puede encontrarse en Wooldridge (2010) - Econometric analysis of cross section and panel data - 2nd edition.

Márgen intensivo

Utilizando solamente las observaciones no censuradas, se modelan las horas efectivamente trabajadas de la siguiente manera:

$$\mu_{2i} = \mathbf{x}'_i \boldsymbol{\beta} + \hat{\lambda}_i \sigma_{12}$$

donde el vector de covariables \mathbf{x}_i incluye las variables educ, expot, expot2, hijos, y lambda.

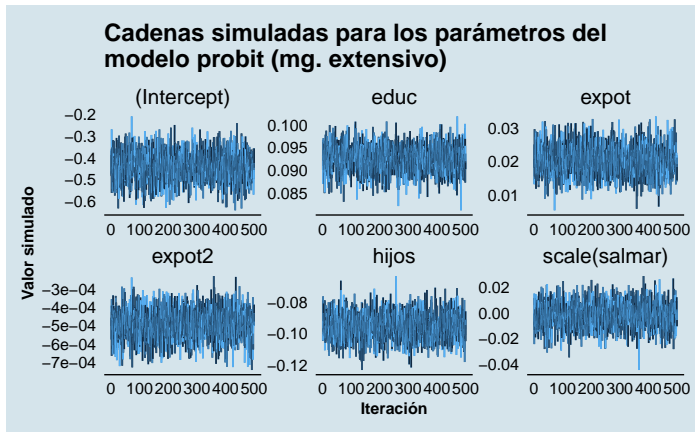
Para el vector de parámetros asociados a las covariables, $\boldsymbol{\beta}$, se selecciona una distribución previa Normal, así como también para el coeficiente asociado al λ de Heckman:

$$\beta_j \sim \text{Normal}(0; 2, 5^2) \quad \forall j$$

$$\sigma_{12} \sim \text{Normal}(0; 2, 5^2)$$

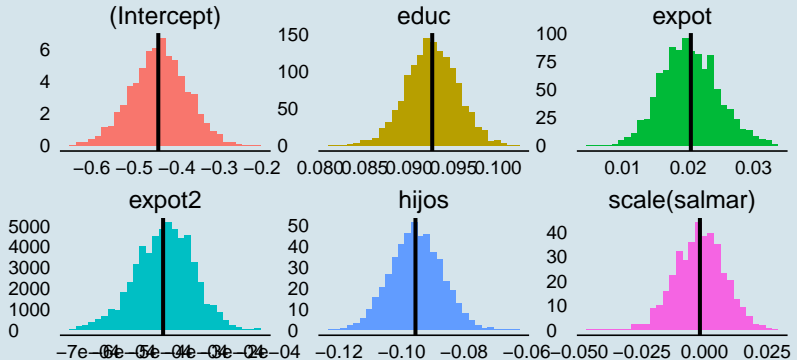
Resultados - Márgen extensivo

Debido a la complejidad computacional de la estimación, las cadenas correspondientes al margen extensivo debieron estimarse por separado.



No se evidencia falta de convergencia.

Histograma de las distribuciones posteriores de los coeficientes asociados a las covariables del modelo probit (mg extensivo)



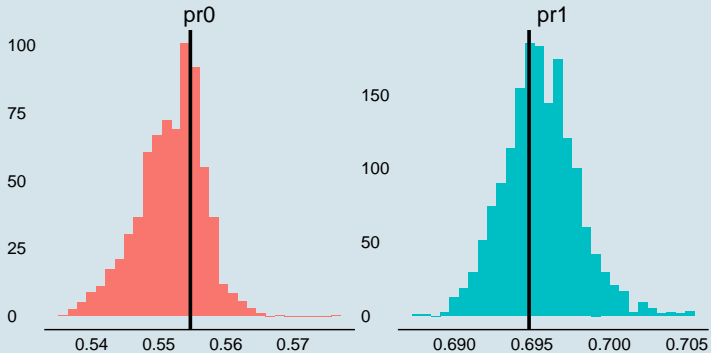
Las líneas verticales corresponden a estimaciones máximo verosímiles de los coeficientes. Se observa entonces que los resultados son robustos al método de estimación que se utilice.

Diagnóstico - margen extensivo

Como medidas de bondad de ajuste se utilizaron el pr_0 y pr_1 donde:

- ▶ pr_0 es la probabilidad de que el modelo replique un cero para una observación que toma valor cero
- ▶ pr_1 es la probabilidad de que el modelo replique un uno para una observación que toma valor uno

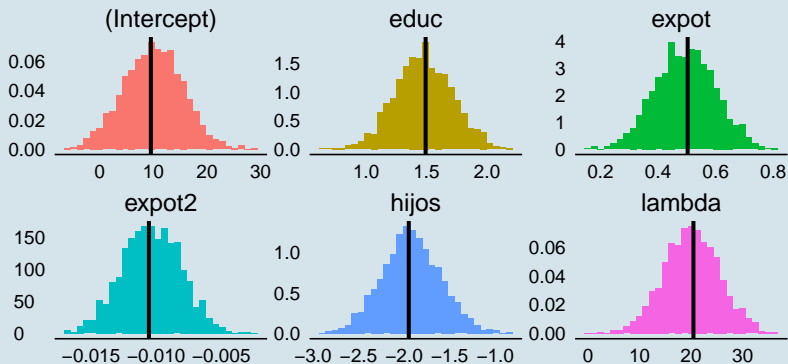
Histograma de las distribuciones posteriores de $pr0$ y $pr1$



- ▶ Las distribuciones se concentran por encima del valor 0.5 lo cual valida el modelo establecido para el margen extensivo.
- ▶ Las líneas verticales representan la proporción de ceros y unos correctamente predichos por la estimación máximo verosímil, lo cual nuevamente evidencia la robustez de los resultados.

Márgen intensivo

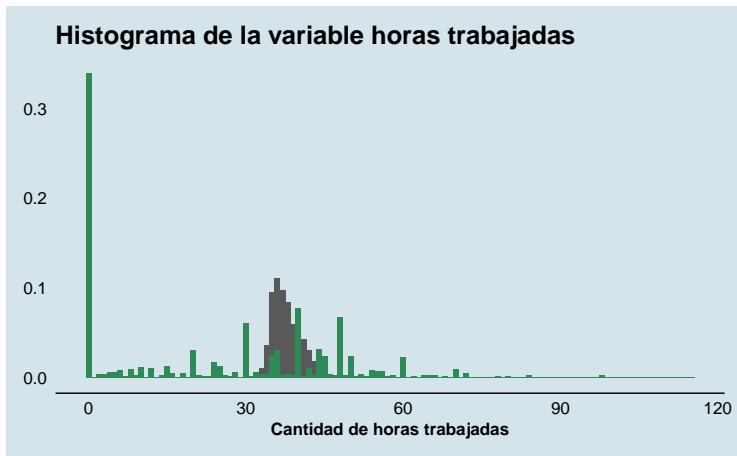
Histograma de las distribuciones posteriores de los parámetros asociados a las covariables del modelo para el margen intensivo



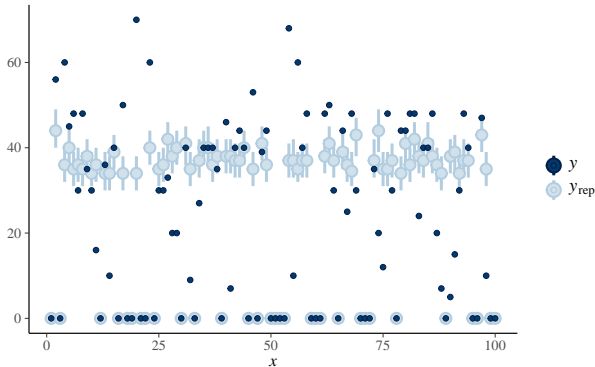
- ▶ La línea negra corresponde a estimaciones máximo verosímiles de cada parámetro, las cuales se sitúan cerca del centro de las distribuciones posteriores para todos los parámetros.
- ▶ Se observa además que la distribución posterior del coeficiente asociado a la variable λ tiene una esperanza alejada del valor 0, lo cual evidencia correlación entre los márgenes y por lo tanto de autoselección muestral.

Diagnóstico

- ▶ Se obtienen réplicas para cada observación utilizando los valores posteriores θ_i .
- ▶ El modelo es bueno prediciendo los puntos de censura pero malo prediciendo el resto de las observaciones.



Intervalos de credibilidad para la predictiva posterior y los valores muestrales



- ▶ Nótese que los intervalos para las observaciones censuradas son muy poco amplios.
- ▶ También presentan buen ajuste los valores cercanos a la media de las observaciones no censuradas pero no así el resto de las observaciones.

Conclusiones

- ▶ Existe evidencia concluyente en cuanto a la presencia de correlación entre el margen extensivo y el margen intensivo dado que la variable λ de Heckman presenta una distribución alejada del valor 0.
- ▶ Esto implica autoselección de unidades en la muestra y por lo tanto justifica la utilización del presente modelo.
- ▶ En lo que respecta a las variables seleccionadas para el análisis:
 - ▶ hijos tiene un efecto negativo sobre horas
 - ▶ expot presenta rendimientos marginales decrecientes
 - ▶ educ tiene un efecto positivo sobre horas.

- ▶ Tanto para la variable λ de Heckman como para el resto de las variables, los resultados son consistentes con los obtenidos en la implementación clásica del modelo de Heckman.
- ▶ El mal ajuste del margen intensivo puede deberse a la distribución seleccionada para los datos. El histograma de la variable `horas` sugiere que sería razonable proponer una distribución con varios modos. Una posible alternativa sería una mixtura de distribuciones Poisson cada una con diferentes parámetros θ_{ji} .