Oferta laboral de las mujeres casadas

Daniel Czarnievicz & Lucía Coudet

Diciembre, 2017

1. Objetivo

El objetivo del presente trabajo es modelar la oferta de trabajo de las mujeres casadas en Uruguay para el año 2009. Para ello 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. Se cuenta entonces con un total de 19.919 observaciones.

El primer problema que presenta este tipo de estudio radica en que la variable a estudiar es una variable censurada, es decir, que se observan horas trabajadas únicamente para las mujeres que efectivamente trabajan. El resto de ellas tiene una oferta laboral, pero no es observable. Por lo tanto, estamos ante la presencia de una censura inferior en el valor 0.

En segundo lugar, 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 descriptos en los párrafos 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 exstensivo 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.

2. Los datos

2.1. 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`: experencia potencial del individuo. Se contruyó como la diferencia entre `edad` y `educ + 6`.

Un problema adicional viene dado porla variable `salmar`, la cual contiene valores faltantes. No resulta razonable pensar que esos valores faltantes fueron generados de forma aleatoria, sino que involucran un posible problema de sesgos de selección, debido a que es posible que aquellas mujeres que no quisieron declarar el salario del marido sean las mujeres que no trabajan y cuyos maridos presentan niveles de salario elevados. No obstante, en lo que sigue del trabajo se hará abstracción de éste problema, continuando solamente con las observaciones para las cuales se cuenta con la información del salario del marido.

Cuadro 1: Mujeres casadas en Uruguay en el año 2009

edad	educ	horas	hijos	sal	salmar	expot	trabaja	expot2
51	14	0	0	NA	9.45	31	0	961
45	20	56	2	5.32	11.43	19	1	361
31	10	0	1	NA	8.46	15	0	225
52	6	60	0	2.40	8.52	40	1	1600
39	15	45	2	6.46	10.55	18	1	324
41	6	48	2	5.13	9.35	29	1	841
55	6	30	0	5.62	8.41	43	1	1849
36	8	48	3	5.05	9.02	22	1	484
48	12	35	2	6.32	9.27	30	1	900
58	6	30	1	4.55	9.84	46	1	2116

2.2. Descripción

A continuación se presenta el histograma de la variable `horas`, en el cual es posible apreciar la censura mencionada en el valor `horas = 0`.

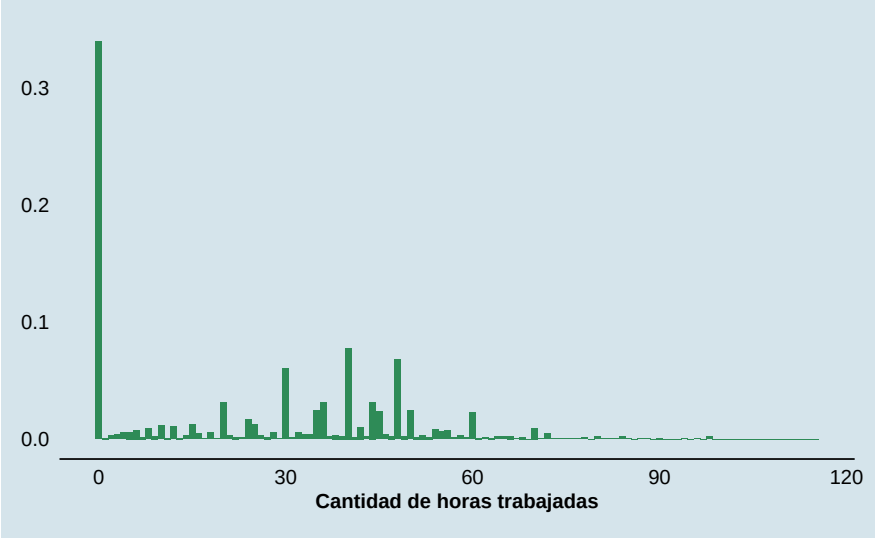

Figura 1: Histograma de la variable horas trabajadas.

En lo que respecta a la variable `hijos`, se observa que 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`. Para el resto de las mujeres, el gráfico no es concluyente.

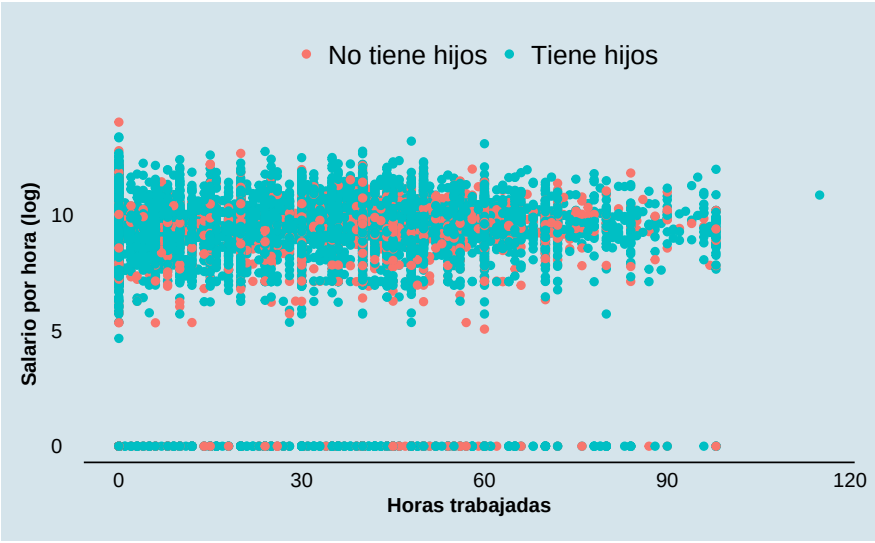

Figura 2: Scatter plot de horas vs. salario por hora de las mujeres (en log), separando a mujeres seguún tengan o no hijos.

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

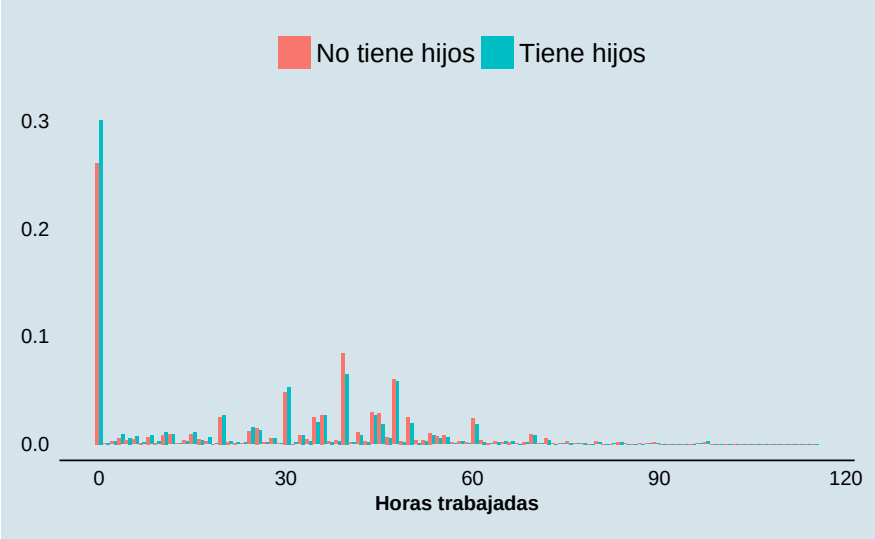

Figura 3: Histograma de las horas trabajadas separando según las mujeres tengan hijos o no.

3. 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:

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

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

$$\theta_i = \mu_2 \times \text{I}_{[\mu_1 > 0]}$$

3.1. Márgen extensivo

Se construye la variable `trabaja` como la indicatriz de que horas_i 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:[1]

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

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

3.2. Márgen intensivo

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

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

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

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

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

4. Resultados

4.1. Márgen extensivo

Debido a la complejidad computacinal de la estimación, las cadenas correpondientes al margen extensivo debieron estimarse por separado. Se realizaron cuatro cadenas utilizando los algoritmos de Hamiltonian Monte Carlo de las librerías `rstan` y `rstanarm` para `R`.

[1] 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 $u_i \sim \text{Normal}(0;\, \sigma^2)$.

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

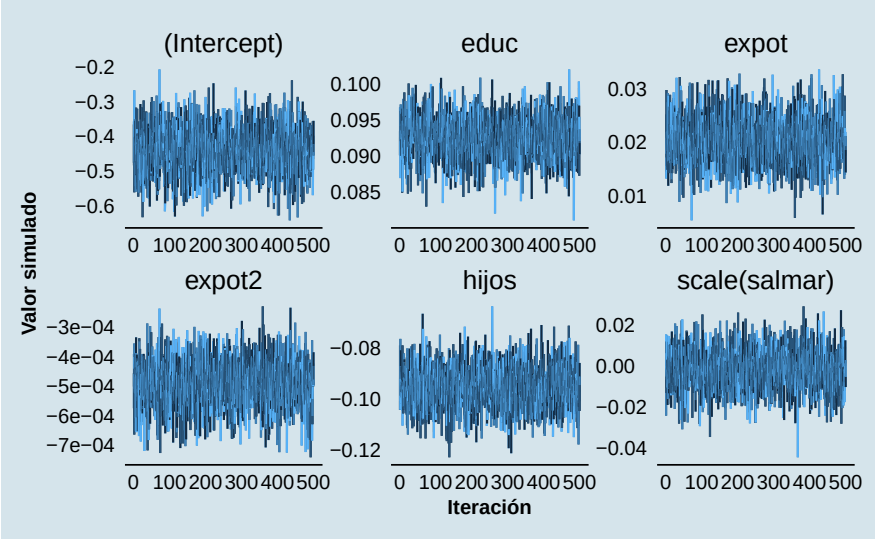

Figura 4: Cadenas simuladas para los parámetros del modelo probit para el margen extensivo. Las mismas no evidencian falta de convergencia.

Se obtienen las siguientes distribuciones de probabilidad posteriores en cada parámetro:

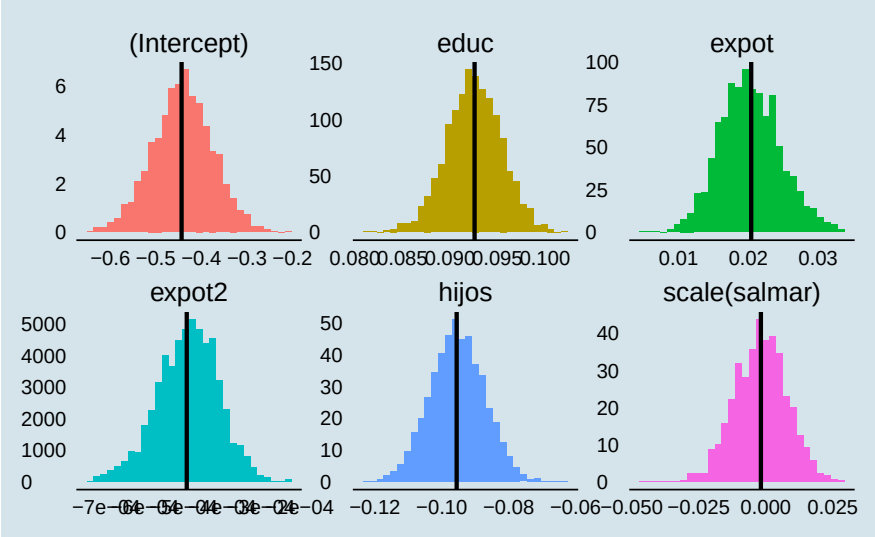

Figura 5: Histograma de las distribuciones posteriores de los coeficientes asociados a las covariables del modelo probit.

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.

Como medidas de bondad de ajuste se utilizaron el `pr0` y `pr1` donde:

- `pr0` es la probabilidad de que el modelo replique un cero para una observación que

toma valor cero

- `pr1` es la probabilidad de que el modelo replique un uno para una observación que toma valor uno

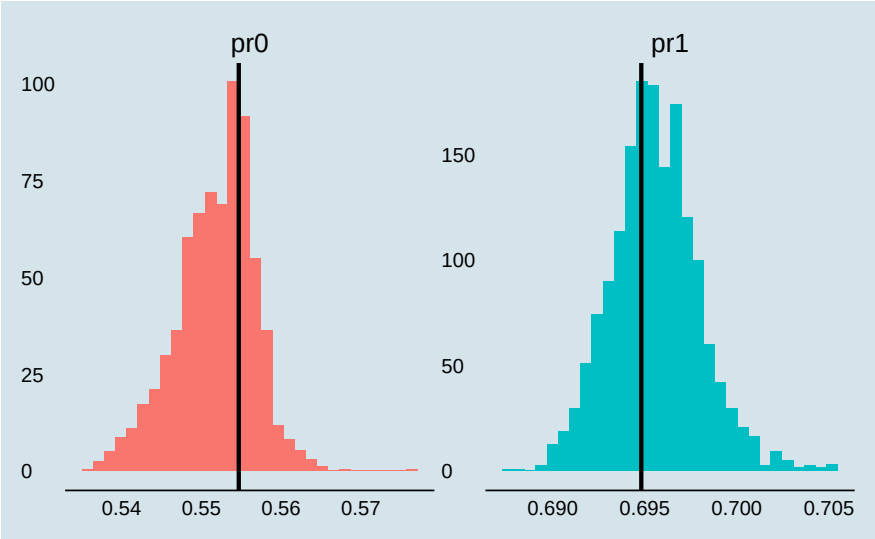

Figura 6: Histograma de las distribuciones posteriores de pr0 y pr1.

El gráfico permite observar que, para ambos casos, las distribuciones se concentran por encima del valor 0.5 lo cual valida el modelo establecido para el márgen extensivo. Las líneas verticales representan la proporcion de ceros y unos correctamente predichos por la estimación máximo verosímil, lo cual nuevamente evidencia la robustez de los resultados.

4.2. Márgen intensivo

En la `Figuara 7` se pueden observar las distribuciones de las simulaciones de las posteriores de los parámetros en el márgen intensivo.

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

```
Model Info:

 function:  stan_glm
 family:    gaussian [identity]
 formula:   horas ~ educ + expot + expot2 + hijos + lambda
```

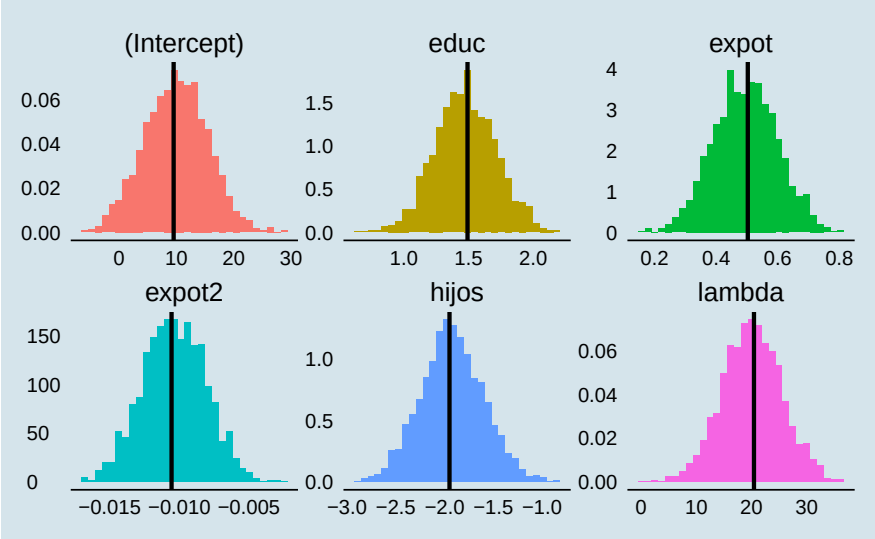

Figura 7: Histograma de las distribuciones posteriores de los parámetros asociados a las covariables del modelo para el margen intensivo

```
 algorithm: sampling
 priors:    see help('prior_summary')
 sample:    2000 (posterior sample size)
 num obs:   13167

Estimates:
                 mean     sd       2.5%     25%      50%      75%
(Intercept)        9.9      5.5     -1.0      6.2     10.1     13.7
educ               1.5      0.2      1.0      1.3      1.5      1.6
expot              0.5      0.1      0.3      0.4      0.5      0.6
expot2             0.0      0.0      0.0      0.0      0.0      0.0
hijos             -1.9      0.3     -2.5     -2.1     -1.9     -1.7
lambda            20.0      5.4      9.4     16.4     20.0     23.7
sigma             17.0      0.1     16.8     16.9     17.0     17.1
mean_PPD          37.8      0.2     37.3     37.6     37.8     37.9
log-posterior -55990.4      2.0 -55995.2 -55991.4 -55990.0 -55988.9
                 97.5%
(Intercept)       20.4
educ               2.0
expot              0.7
expot2             0.0
hijos             -1.3
lambda            30.7
sigma             17.2
mean_PPD          38.2
log-posterior -55987.7
```

```
Diagnostics:
              mcse Rhat n_eff
(Intercept)   0.3  1.0   329
educ          0.0  1.0   345
expot         0.0  1.0   373
expot2        0.0  1.0   361
hijos         0.0  1.0   396
lambda        0.3  1.0   356
sigma         0.0  1.0  2000
mean_PPD      0.0  1.0  1367
log-posterior 0.1  1.0   784

For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effe
```

5. Diagnóstico

Se obtienen réplicas para cada observación utilizando los valores posteriores θ_i. Los siguientes gráficos permiten observar que el modelo es bueno prediciendo los puntos de censura pero malo prediciendo el resto de las observacionbes. En otras palabras, el márgen extensivo ajusta bien, no así el márgen intensivo.

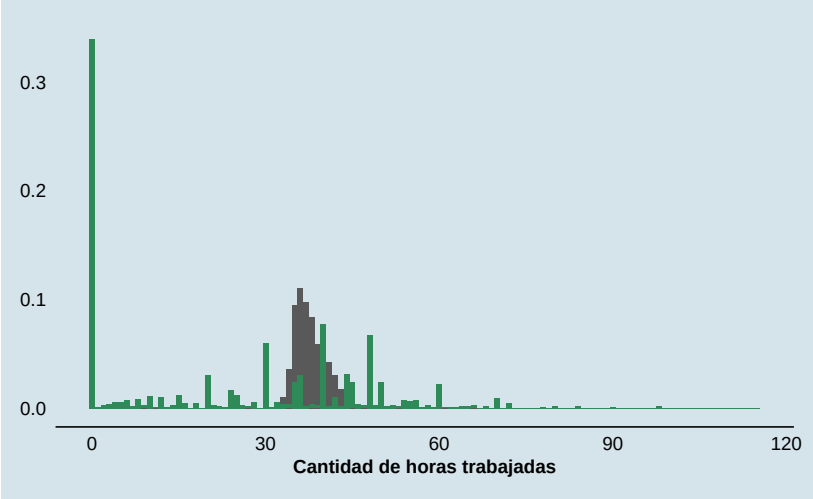

Figura 8: Histograma de la variable horas trabajadas. En verde los valores muestrales, en gris los valores simulados. Nótese que el ajuste del margen extensivo es perfecto, por lo que las barras correspondientes quedan totalmente superpuestas.

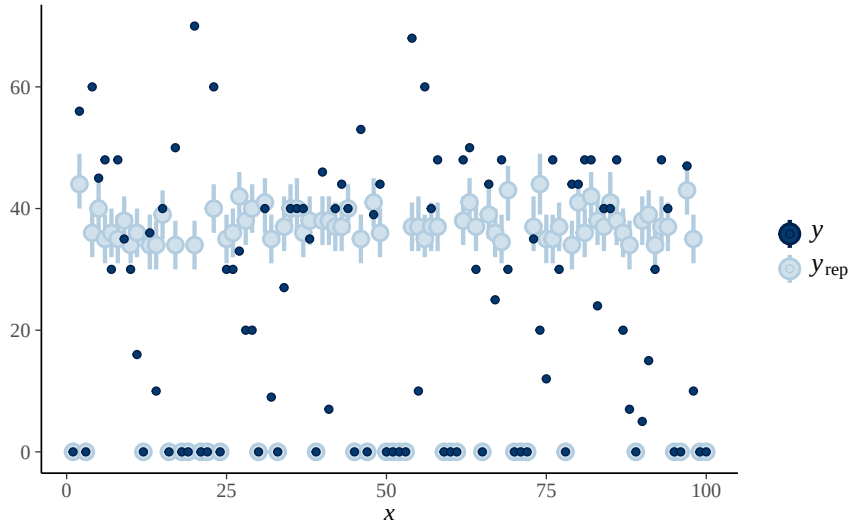

Figura 9: Intervalos de credibilidad para la predictiva posterior y los valores muestrales. Nótese que los intervalos para las observaciones censuradas son muy poco amplios. Tambien presentan buen ajuste los valores cercanos a la media de las observaciones no censuradas pero no así el rsto de las observaciones.

6. Conclusiones

La implementación bayesiana del modelo de selección muestral propuesto por Heckman presenta evidencia concluyente en cuánto a existencia de correlación entre el márgen extensivo y el márgen intensivo dado que la variable `lambda` de Heckman presenta una distribución alejada del valor 0. Como fue mencionado anteriormente, 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, como era de esperar, `hijos` tiene un efecto negativo sobre `horas` y `expot` presenta rendimientos marginales decrecientes. Por su parte, la variable `educ` tiene un efecto positivo sobre `horas`.

Tanto para la variable `lambda` 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 márgen 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}.