Rodríguez-Oreggia, Eduardo; López Videla, Bruno

Working Paper

Imputación de ingresos laborales: Una aplicación con encuestas de empleo en México

Working Papers, No. 2014-21

Provided in Cooperation with:
Bank of Mexico, Mexico City

Suggested Citation: Rodríguez-Oreggia, Eduardo; López Videla, Bruno (2014) : Imputación de ingresos laborales: Una aplicación con encuestas de empleo en México, Working Papers, No. 2014-21, Banco de México, Ciudad de México

This Version is available at:
https://hdl.handle.net/10419/129960

Banco de México

Documentos de Investigación

Banco de México

Working Papers

N° 2014-21

Imputación de ingresos laborales: una aplicación con encuestas de empleo en México

Eduardo Rodríguez-Oreggia
EGAP, ITESM

Bruno López Videla
Banco de México

Septiembre 2014

Documento de Investigación
2014-21

Working Paper
2014-21

Imputación de ingresos laborales: una aplicación con encuestas de empleo en México*

Eduardo Rodríguez-Oreggia[y]
EGAP, ITESM

Bruno López Videla[z]
Banco de México

Resumen: El objetivo de este artículo es el de realizar una imputación de ingresos a observaciones con ingresos no reportados en la Encuesta Nacional de Ocupaciones y Empleo (ENOE). Se presenta imputaciones por dos métodos y una corrección de estimaciones por re-muestreo para las observaciones con ingreso reportado. Posteriormente, se analiza el posible sesgo en estimaciones de ecuaciones mincerianas y de pobreza laboral derivado de ignorar a las observaciones con ingreso no reportado. Los resultados señalan que cuando no se considera el sub-reporte de ingresos en las estimaciones existen diferencias en los parámetros que definen la relación entre el nivel de capital humano y el ingreso, así como en los que describen los determinantes de pobreza laboral, en comparación con los resultados obtenidos de estimaciones que consideran la muestra total. Las diferencias son más significativas cuando se analiza pobreza laboral.
Palabras Clave: imputación, ingreso, capital humano, pobreza laboral, matching

Abstract: The aim of this paper is to make imputations of earnings to observations with missing earnings in the Encuesta Nacional de Ocupaciones y Empleo (ENOE). We present imputations by two methods and also a correction of estimations by reweighting observations with reported earnings. Then, we analyze the possible bias in estimations of mincer equations and labor poverty derived from ignoring observations with missing earnings. The results show that when missing earnings are not considered in estimations, there are differences in the parameters that define the relationship between the human capital level and earnings, as well as those that describe the determinants of labor poverty, compared to the results obtained from estimations that consider all observations. Differences are more significant when we analyze labor poverty.
Keywords: imputations, earnings, human capital, labor poverty, matching
JEL Classification: C18, C81, D10, J24

*Los autores agradecen la asistencia de investigación de Javier Sinco. También agradecen los comentarios recibidos por Nelly Aguilera, Virginia Abrin, Gerardo Leyva Parra, y de los participantes en el Coloquio de Investigación del INEGI en octubre de 2013 y en el Congreso de Investigación del ITESM 2014.

[y] EGAP, Instituto Tecnológico y de Estudios Superiores de Monterrey. Correo electrónico: eduardo.oreggia@gmail.com.

[z] Dirección General de Investigación Económica, Banco de México. Correo electrónico: bruno.lopez@banxico.org.mx.

1 Introducción

Una preocupación constante en el análisis de encuestas que involucran la declaración de ingresos es el sub-reporte que puede haber por parte de individuos y de hogares, los cuales pueden afectar tanto a diversas estimaciones utilizando estos datos, como a indicadores de incidencia de política pública basados en ellos. En la literatura se pueden encontrar estudios que muestran que no considerar los salarios no reportados en investigaciones sobre salarios e ingreso puede generar sesgos en diversas estimaciones, bien estos estudios usualmente son para países desarrollados que cuentan con mejores bases de datos.

Por ejemplo, para el caso de Estados Unidos, utilizando la Encuesta de Población Actual (Current Population Survey, o CPS por su siglas en inglés), Lillard, Smith y Welch (1986) encuentran un sesgo potencial en las estimaciones con la presencia de observaciones con salario no reportado, aunque subestimaciones en ocupaciones muy detalladas.[1] Rubin (1996) sugiere que irrespectivamente de la aleatoriadad de los faltantes, es importante proveer de imputaciones, ya que la falta de ello crea estudios derivados de una muestra más pequeña. En general, si bien no hay un consenso hacia el mejor método de imputación, es una cuestión que académicamente al menos debe considerarse para su discusión en cuanto a las implicaciones que puede representar.

En algunos países, las oficinas de estadística reportan encuestas con ingresos directamente imputados. Por ejemplo, la CPS, levantada por el Bureau of Labor Statistics (BLS), y que es una de las más utilizadas en Estados Unidos para realizar diversos indicadores de políticas públicas, así como para realizar investigaciones académicas, realiza una imputación de ingresos tomando observaciones donantes del mismo período en una primera fase, y posteriormente en periodos previos, con el fin de llenar las celdas de ingresos faltante. La tasa de ingresos faltantes en esta encuesta llega hasta el 30%. Bollinger y Hirsch (2006) analizaron si esta imputación causa algún sesgo en estimaciones de capital humano que incluyen las imputaciones, sugiriendo algunas alternativas de correcciones, y dejando al debate académico su aplicación o no. No obstante, estos autores no analizan los posibles

[1] Para estudios similares ver Horowitz y Manski (1998, 2000).

sesgos en indicadores de política pública derivados de los ingresos faltantes en ciertas observaciones (e.g. pobreza, desigualdad. etc).

En este sentido, México se ha vuelto un caso interesante de analizar, ya que la Encuesta Nacional de Ocupaciones y Empleo (ENOE), que trimestralmente captura tanto ingresos como otros indicadores laborales, ha experimentado un amplio crecimiento en observaciones con ingreso faltante, pasando de alrededor del 13% en 2005 a 24% en 2012; aunque las demás características sí se encuentran reportadas de forma normal en la encuesta. Ya previamente dos estudios derivados de forma similar han analizado esta cuestión.

En el primero, Rodríguez-Oreggia *et al.* (2012) utilizan un proceso de matching para imputar ingresos faltantes en la Encuesta Nacional de Ocupación y Empleo (ENOE). En el segundo, Vázquez-Campos (2013) analiza la muestra que no reporta ingresos en la ENOE y encuentra que las tendencias de no reporte no son aleatorias. Además, el autor realiza una revisión de las metodologías de imputación y aplica cuatro diferentes métodos de imputación (i.e. pareamiento por puntaje, Hot-Deck, imputación en la mediana de un grupo con ruido, y pareamiento por promedios predictivos). En ambos casos, los autores encuentran que no considerar los ingresos faltantes puede conllevar a sobreestimaciones en el indicador de pobreza laboral, calculado por el Consejo Nacional de Evaluación de la Política Social (CONEVAL, 2013). En términos generales, cuando se realiza esta estimación no se toman en cuenta los salarios faltantes y se le suma cero al ingreso del hogar para aquellos individuos que no reportan un salario pese a que en realidad sí lo perciben, por lo que existen hogares que pueden estar considerados dentro de pobreza laboral, cuando en realidad no necesariamente lo están.

Derivado de estos estudios se deduce la necesidad no solo de analizar la incidencia de los ingresos faltantes sobre indicadores como pobreza laboral, sino más allá de eso, de estudiar cuál es el papel que tiene la imputación de ingresos o el uso solo de ingresos reportados sobre estimaciones de capital humano, donde factores como la educación juegan un papel relevante. Como se señaló anteriormente, los ingresos faltantes no ocurren de forma aleatoria en la encuesta, por lo que es posible que existan sesgos en la medición de factores relacionados con ingreso y pobreza laboral. Por ello, el objetivo de este artículo es el de analizar el posible sesgo que existe de ignorar los ingresos faltantes en la ENOE al medir los

efectos de diversas variables sobre ingreso laboral y pobreza laboral, una vez imputados los ingresos mediante el método de Hot-Deck. Para realizar esta contribución a la literatura, se utilizan algunos métodos de imputación y posteriormente se realizan estimaciones de capital humano considerando ingresos y pobreza laboral para medir los posibles sesgos al no considerar a las observaciones con ingreso faltante.

El artículo se estructura como sigue. En una primera parte, se presentan los mecanismos por los cuales pueden existir ingresos faltantes en las encuestas. Posteriormente, se analizan los faltantes en la ENOE y sus características comparadas con los reportados. A continuación, se presentan los supuestos de imputación y el método de estimación. Después se analizan los resultados y finalmente, se delinean conclusiones y algunas recomendaciones.

2 Mecanismos de respuestas faltantes

El análisis de encuestas que involucran la declaración de ingresos se ha enfrentado constantemente a un sub-reporte en esta pregunta. Las razones por las cuales los individuos y hogares pueden no reportar su ingreso son muy diversas. Groves, Singer y Corning (1999) recuentan estas razones en una serie de múltiples factores que actúan en forma simultánea, desde los temas que va cubriendo la encuesta, la presencia de otras personas, el tema de confidencialidad, hasta factores socioeconómicos y físicos.

La estadística ha afrontado el problema de ingresos faltantes en encuestas mediante la imputación de salarios a aquellos individuos que no reportan un salario de diversas maneras. Primero, para realizar una imputación de salarios es importante considerar tres mecanismos de respuesta resumidos en Rubin (1987) y Little y Rubin (2002): salarios faltantes completamente de forma aleatoria (MCAR por sus siglas en inglés), salarios faltantes de forma aleatoria (MAR por sus siglas en inglés), y salarios faltantes de forma no aleatoria (MNAR por sus siglas en inglés). La estrategia a seguir para tratar los salarios no reportados va a depender de cuál sea el mecanismo de respuesta de los individuos que no reportan un salario.

Un dato faltante completamente de forma aleatoria implica que la probabilidad de no reportar es independiente de la variable de análisis y de las variables del modelo explicativo; en este caso, no reportar un salario sería independiente del salario mismo y de las variables que lo

explican. Si se diera este caso, entonces no existiría un sesgo en las estimaciones al no considerar a las observaciones que no reportan un salario. Cabe señalar que encontrar este mecanismo en datos faltante es algo muy raro empíricamente.

Por otro lado, un dato faltante de forma aleatoria implica que la probabilidad de no reportar es independiente del verdadero valor de la variable en cuestión pero es dependiente al menos de una de las variables explicativas del modelo (Treiman, 2009). En este sentido, $P(No\ reportar|y_i, x_i) = P(No\ reportar|x_i)$. Si se diera este caso, entonces una simple estimación del salario sin considerar el problema de las observaciones con salario no reportado pudiera estar sesgada. Para evitar un sesgo en las estimaciones es importante realizar una imputación a la variable de salarios para aquellos individuos que no reportan un salario.

Finalmente, se puede presentar el caso de datos faltantes de forma no aleatoria. En este escenario, la probabilidad de no reportar es dependiente de la variable en cuestión y de las variables explicativas del modelo, de tal forma que se tiene un problema de selección en variables no observadas. Por ejemplo, con este mecanismo de respuesta los individuos con mayores salarios pudieran tener una mayor probabilidad de no reportar comparado con los individuos con salarios medios, de tal manera que no reportar está correlacionado con el nivel del salario. Esto lleva a un problema adicional de no observables en la imputación, lo cual todavía no está resuelto en la literatura estadística.

3 Mecanismo de respuesta de las observaciones con salario no reportado en la ENOE

En México el Instituto Nacional de Estadística y Geografía (INEGI) levanta la encuesta ENOE trimestralmente, a hogares e individuos de forma rotativa por cinco trimestres consecutivos, preguntando acerca de características socio económicas, así como detalles de características laborales como si tiene una ocupación o no, en su caso ingreso laboral, horas trabajadas, tipo de ocupación, sector de actividad, prestaciones, etc., para todos los integrantes de un hogar de 14 años en adelante. Es una encuesta utilizada para dar seguimiento a indicadores agregados como el desempleo, actividades formales, entre otros usos.

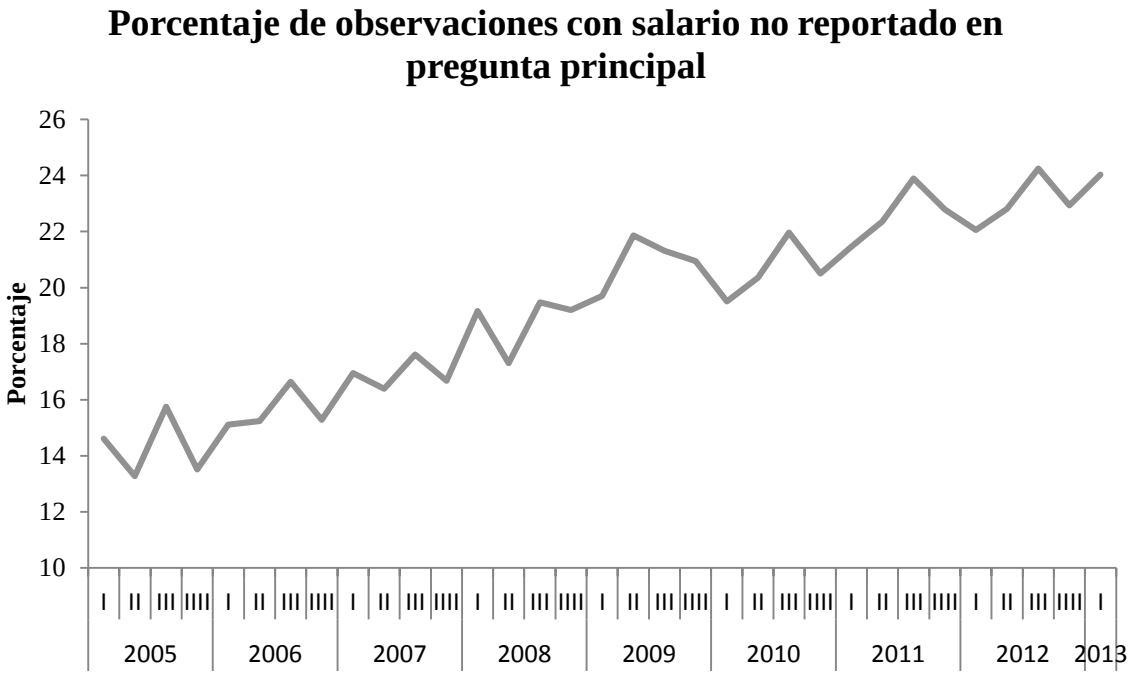
La ENOE captura el ingreso laboral a través de dos preguntas en secuencia. En la primera, se pregunta directamente a las personas ocupadas que deben recibir un pago ¿Cuánto ganó o en cuánto calcula sus ingresos? La pregunta incluye la periodicidad, y en las bases de microdatos del INEGI ya se reporta el monto estandarizado mensual.[2] Si los individuos contestan esta pregunta, entonces se pasa a la parte de prestaciones laborales. Sin embargo, si no se responde esta pregunta, el cuestionario pasa a una segunda pregunta para tratar de determinar el ingreso laboral: Actualmente el salario mínimo es de (cantidad) ¿la cantidad que obtiene al mes por su trabajo es…? Y se detallan algunos rangos acotados en niveles de salarios mínimos. Si se contesta o no se quiere contestar esta segunda pregunta igualmente el cuestionario avanza a la parte de prestaciones laborales, dejando una parte de individuos sin respuesta en ambas preguntas sobre ingresos.

La muestra que analizamos a continuación abarca a individuos de 14 años o más, que forman parte de la Población Económicamente Activa (PEA) y están ocupados en categorías que implican un pago. Esto es, se excluye de la muestra a individuos que caen en la categoría de ocupados sin pago. El periodo de análisis va desde el primer trimestre de 2005 al primer trimestre de 2013.

Como se puede observar en la gráfica 1, existe una tendencia creciente en el porcentaje de individuos que no reportan un salario en la ENOE tomando en cuenta solamente la pregunta principal de en cuánto calcula su ingreso laboral. En el primer trimestre de 2005, el 14.61% de los ocupados de 14 años o más que deberían reportar un salario no lo reportan; para el primer trimestre de 2013 este porcentaje aumenta a 24.03%.[3] El porcentaje máximo de observaciones que no reportan un salario se observa en el tercer trimestre de 2012, alcanzando el 24.25%. De la gráfica se puede deducir que a lo largo del tiempo el porcentaje de observaciones que no reportan un salario ha aumentado de forma constante. Lo anterior reafirma que el mecanismo de respuesta de las observaciones que no reportan un salario no es completamente aleatorio y que pueden existir algunas características observables, y no observables, correlacionadas con el mismo.

[2] Esta pregunta es la única que considera CONEVAL para sus cálculos de pobreza laboral.
[3] Considerando la pregunta alternativa indirecta por niveles de salarios mínimos se reduciría la tasa de no respuesta del 24 al 12%.

Gráfica 1.



Notas: Se incluye solo la pregunta de ingreso y no la indirecta por salarios mínimos. La muestra incluye a observaciones con 14 años o más que forman parte de la PEA, ocupados y que no caen dentro de la categoría de ocupados sin pago. Datos de la ENOE.

En este sentido, es importante identificar si los factores correlacionados con la probabilidad de no reportar son aquellas variables explicativas del salario (Treiman, 2009). Para ello, se realizan pruebas de media con las variables explicativas del salario entre los individuos que reportan un salario y los individuos que no lo hacen. Si existe una diferencia en las medias del vector de características, entonces la probabilidad de no reportar un salario está correlacionada con las mismas.

En el Cuadro 1 se presentan los resultados de las pruebas de media para el segundo trimestre de 2005 de un vector de variables explicativas del nivel del salario.[4] Como se puede observar, existen diferencias significativas a lo largo del tiempo en las medias de las variables entre las observaciones que reportan un salario y las observaciones que no lo hacen. En este sentido,

[4] En general, los estudios laborales utilizan las mismas variables o muy similares para explicar el salario.

no reportar un salario no se da de forma completamente aleatoria; la probabilidad de no reportar un salario está correlacionada con el vector de variables explicativas del mismo.

Cuadro 1. Pruebas de diferencia en medias.
Ho: Existe diferencia en las medias entre grupos.

	No reporta un salario	N	Media	t-estadístico
Hombre	0	144,988	0.6265484	-23.2977
	1	10,527	0.7396219	
Cohorte de edad	0	144,988	7.21619	41.7451
	1	10,527	6.07267	
Casado	0	144,988	0.6255276	-8.4159
	1	10,527	0.6665717	
Nivel educativo	0	144,988	3.019815	-0.4723
	1	10,527	3.025363	
Tipo de ocupación	0	144,988	1.475591	-55.8764
	1	10,527	1.944429	
Formal	0	144,988	0.4330427	29.7167
	1	10,527	0.2852665	
Sector de actividad	0	144,988	2.562702	64.9936
	1	10,527	2.117222	
Trabajo de tiempo completo	0	144,988	0.72935	-4.1182
	1	10,527	0.7477914	
Urbano	0	144,988	0.8676235	32.7141
	1	10,527	0.753396	
Entidad federativa	0	144,988	16.49731	-0.2215
	1	10,527	16.51753	

Notas: La muestra incluye a observaciones con 14 años o más que forman parte de la PEA, ocupados y que no caen dentro de la categoría de ocupados sin pago. 1 = no reporta un salario, 0 el caso contrario. Salario reportado pertenece a la pregunta principal de cuánto calcula su ingreso laboral.

Una forma adicional de probar la robustez de este resultado es estimando un modelo probit en donde la variable dependiente es una dummy, donde 1 son aquellas observaciones que no reportan un salario y 0 el caso contrario. Los resultados se muestran en el Anexo 1. Por simplificación, las estimaciones se presentan para los segundos trimestres de 2005 a 2012. Los resultados muestran una fuerte correlación entre el vector de variables explicativas del salario y la probabilidad de no reportar el mismo. Adicionalmente, en los resultados se puede observar que los hombres tienen una mayor probabilidad de no reportar un salario con respecto a las mujeres. De igual forma, los individuos con estudios profesionales tienen

mayor probabilidad de no reportar un salario, al igual que los trabajadores formales y los trabajadores de tiempo completo. Por lo tanto, dado que existen factores observables correlacionados con la probabilidad de no reportar un salario, se descarta la posibilidad de que el mecanismo de respuesta de las observaciones con salario faltante en la ENOE sea faltante de forma completamente aleatoria.

No existe una prueba formal para mostrar que la estructura de las observaciones con salario no reportado es faltante de forma no aleatoria (MNAR). Es difícil probar que la probabilidad de no reportar un salario está correlacionada con el nivel del salario en cuestión, al menos de forma directa. David, Little, Samuel y Triest (1986), utilizando la Encuesta de Población Actual, encuentran para Estados Unidos que aunque existe evidencia de que el mecanismo de respuesta sea no aleatorio, en la práctica no es cualitativamente importante. En este sentido, en el presente artículo se asume que la probabilidad de no reportar un salario no tiene una correlación directa con el nivel del mismo.[5] Por lo tanto, de las pruebas anteriores se concluye que la estructura de las observaciones que no reportan un salario en la ENOE es del tipo MAR, es decir, existe una correlación entre la probabilidad de no reportar un salario y las variables explicativas del mismo, pero en las categorías del vector de estas variables se da de forma aleatoria.

En este sentido, dada la existencia de observaciones con salario no reportado en la ENOE, probaremos los posibles sesgos en estimaciones y mediciones de indicadores al considerar solo la variable de ingreso reportado, con aquellas derivadas de añadir ingresos imputados. En la siguiente sección se presentan los supuestos necesarios para realizar una imputación de salarios y la metodología a utilizar.

4 Supuestos de la imputación

La metodología a utilizar está basada en un aparejamiento o matching entre características observables tanto demográficas de cada individuo y de su hogar, así como laborales, de forma

[5] Hirsch y Schumacher (2004) toman en cuenta el problema de selección en la Encuesta de Población Actual. Identifican la probabilidad de no responder utilizando como variable instrumental si la encuesta fue respondida por el individuo o por algún otro miembro del hogar. Aunque es posible considerar problemas de selección en la probabilidad de no responder utilizando variables instrumentales, este no es tema del presente artículo.

que se busca un donante que tenga las mismas características del individuo que no reporta un salario.

Una vez que se prueba que la estructura de las observaciones que no reportan un salario es del tipo MAR, es importante definir los supuestos bajo los cuales se realiza la imputación de salarios entre donantes y receptores. En el presente artículo, y_i es la variable de interés para la observación i; x_i es el vector de variables explicativas de y_i para la observación i, y z_i es el vector de variables para la observación i bajo las cuales se realiza la imputación. Se debe cumplir que $\boldsymbol{z} \subseteq \boldsymbol{x}$. Bollinger y Hirsch (2006) establecen cinco supuestos que se deben cumplir para una imputación:

SUPUESTO 1: Existen datos faltantes únicamente en algunas observaciones de la variable y_i.

SUPUESTO 2: $E_O[y_i|x_i, z_i] = E_M[y_i|x_i, z_i] = E[y_i|x_i, z_i]$.

SUPUESTO 3: $z_i = h(x_i)$, donde $h(\cdot)$ es una función determinística conocida.

SUPUESTO 4: $E[y_i|x_i, z_i] = E[y_i|x_i] = \alpha + x_i'\beta$

SUPUESTO 5: Los valores imputados de y_i son seleccionados aleatoriamente de la función de distribución $f_O(y_i|z_i)$.

El primer supuesto parece un poco redundante. Sin embargo, es importante notar que únicamente se podrá hacer las imputaciones si la única variable con datos faltantes es la variable *y*. Las observaciones con datos faltantes en alguna de las variables del vector $\boldsymbol{z}$ no serán consideradas para la imputación.

El segundo supuesto es muy importante y está relacionado con la condición MAR. $E_O[\cdot]$ es la esperanza condicional de y_i para las observaciones con salario reportado; $E_M[\cdot]$ es la esperanza condicional de y_i para las observaciones con salario no reportado y $E[\cdot]$ es la esperanza condicional de y_i para la población. El supuesto implica que la media condicionada de la variable y_i es igual entre donantes y receptores, de tal forma que no existe un problema de selección entre donantes y receptores. Este supuesto permite considerar datos faltantes de forma aleatoria dado que la distribución de (x_i, z_i) puede variar entre donantes y receptores, pero se garantiza que en promedio el salario es el mismo dentro de cada celda. Bollinger y

Hirsch (2006) llaman a esta condición como media condicional faltante de forma aleatoria (CMMAR por sus siglas en inglés).

El tercer supuesto establece que el vector de variables con las que se realiza la imputación es un subconjunto del vector de variables explicativas de análisis. Adicionalmente, el vector ***z*** puede agregar al vector ***x*** en categorías más amplias. Por ejemplo, considérese los años de educación como una variable explicativa del salario. Una forma de agregar estos años es mediante categorías: nivel de instrucción primaria, secundaria, preparatoria o profesional. De esta manera se tiene que si se conocen los años de educación, entonces se puede conocer su agregación, pero lo contrario no necesariamente se cumple.

El cuarto supuesto implica que la relación entre el vector de variables explicativas y la variable dependiente es lineal y que no existe una relación entre el vector ***z*** y la variable dependiente más allá de la que se contiene en el vector ***x***.

Finalmente, el quinto supuesto establece que los valores imputados de la variable dependiente son independientes de cualquier variable no observada o no incluida en el vector ***z***. Esto permite que los valores imputados sean seleccionados de manera aleatoria para evitar subestimaciones en la varianza de la variable dependiente.

5 Método de imputación

Los métodos de imputación han sido ampliamente estudiados en la literatura. En general, lo que tratan de hacer es asignar un valor conocido o estimado a aquellas observaciones con datos faltantes condicionado a un vector de características sociodemográficas.[6] Esta sección revisa brevemente algunos métodos, ya que otros estudios contienen descripciones más detalladas (e.g. Frick and Grabka, 2003). Utilizando la ENOE, Campos-Vázquez (2013) compara cuatro diferentes métodos de imputación sobre el cálculo del indicador de pobreza laboral y concluye que todos arrojan resultados similares, sugiriendo que se utilice el que represente menos complicaciones para estimar.

[6] Little y Rubin (2002) describen ampliamente los métodos de imputación y el análisis estadístico en presencia de observaciones con datos faltantes. Adicionalmente, el BLS cuenta con un gran número de artículos en donde se detallan y comparan los métodos de imputación más utilizados (e.g. West, *et al.*,1990 ; Montaquila y Ponikowski, 1995).

Uno de los métodos más utilizados es el método no paramétrico de celdas de Hot-Deck. Este método permite identificar individuos similares en celdas construidas con el vector de variables sociodemográficas, para posteriormente imputar un salario a las observaciones con salario faltante de manera aleatoria o con base en una función de distancia.

En el presente artículo el método utilizado para la imputación de salarios será el de Hot-Deck con dos variantes. Primero, se realizará la imputación mediante una asignación aleatoria. Posteriormente, se realizará una imputación mediante una función de distancia de Mahalanobis. Ambos métodos son utilizados por el BLS para la imputación de salarios en la Encuesta de Población Actual y se ha demostrado que son robustos a otros métodos paramétricos de imputación (West, *et al.*,1990). Adicionalmente, la ventaja de estos métodos es que mantienen la misma distribución de características para cada celda k (Kalton y Kasprzyk, 1982).

Hot-Deck con imputación aleatoria

Este método consiste en identificar las observaciones en celdas construidas con base en un vector de variables sociodemográficas, y asignar de manera aleatoria el salario de individuos donantes a individuos receptores para cada celda. De tal manera, se tiene que:

$$y_{M,k,i} = y_{O,k,j}$$

en donde y_M es el salario no reportado, y y_O es el salario reportado. Existen k celdas, en donde i son los individuos que no reportan un salario y j son los individuos seleccionados aleatoriamente (con reemplazo) que reportan un salario para cada celda k.

Es importante considerar que de acuerdo a Little y Rubin (2002), una imputación simple puede generar sesgos en la varianza de la variable de interés debido a que se realiza una imputación de manera aleatoria en cada celda k. No obstante, Rubin (1987) establece que estos problemas se pueden solucionar mediante una imputación múltiple, en la que se simulan de manera aleatoria m imputaciones, en donde $m > 1$. Después de las m simulaciones, se promedia el valor imputado para cada simulación y se le asigna este valor a las observaciones con valores no reportados. Esto evita sesgos en la estimación de la varianza de la variable de

interés por lo que puede realizarse inferencia. En general, $m = 5$ es suficiente para tener una buena imputación. Por lo tanto, en el presente artículo se realizan cinco simulaciones.

Hot-Deck con función de distancia

El método de Hot-Deck considerando una función de distancia es muy similar al anterior, con la particularidad de que la imputación en cada celda k se hace mediante la minimización de una función de distancia, en este caso, una función de distancia de Mahalanobis. Algunos estudios utilizan una función de distancia euclidiana (e.g. West, *et al.*, 1990), sin embargo, en el presente artículo se considera la función de distancia de Mahalanobis debido a que pondera por la matriz de varianzas y covarianzas del vector de covariables utilizadas para realizar la imputación. De tal manera, se tiene que:

$$y_{M,k,i} = y_{O,k,j}$$

En donde las observaciones i (receptor) y j (donante) cumplen con la minimización de la siguiente función de distancia:

$$D(i,j) = (z_i - z_j)' V^{-1} (z_i - z_j)$$

$\forall\, i, j$ en cada celda k.

Donde V^{-1} es la matriz inversa de varianzas y covarianzas de las covariables utilizadas para la imputación.

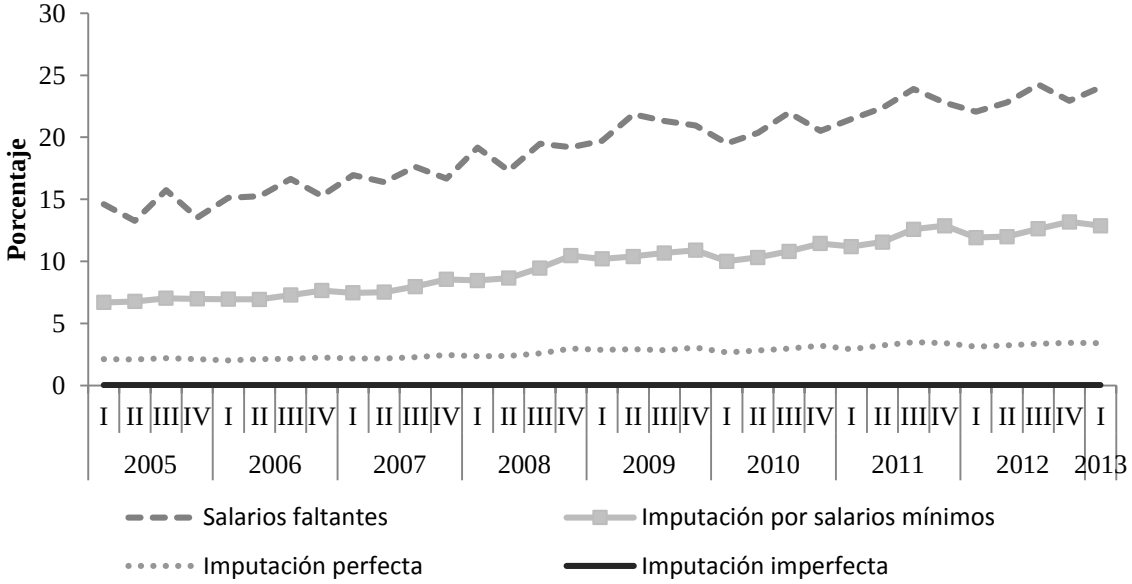
6 Resultados

En esta sección se presentan los resultados de la imputación considerando los criterios descritos anteriormente, y se demuestra el sesgo potencial en las estimaciones con salario no reportado. Bollinger y Hirsch (2006) utilizan la Encuesta de Población Actual (CPS) de Estados Unidos para estudiar el sesgo potencial en estimaciones de capital humano que consideran únicamente el salario reportado comparando con las que imputan los ingresos. En este sentido, en el presente artículo se quiere identificar el sesgo potencial y su dirección en los parámetros comúnmente estimados en estudios laborales en México que utilizan la ENOE, y en parte seguiremos el artículo mencionado.

En la gráfica 2 se puede observar el porcentaje de imputaciones realizadas en cada etapa. La muestra abarca individuos de 14 años o más, que forman parte de la PEA, ocupados, y que no caen dentro de la categoría de ocupados sin pago. En una primera etapa, a aquellas observaciones que no reportan la pregunta principal, pero que sí la de salarios mínimos, se les asigna un salario estimado en el rango que se manifiesta. A lo largo del periodo, estas observaciones representan en promedio alrededor del 50.1% de las observaciones totales que no reportan un salario en la ENOE (gráfica 1). A las observaciones restantes se les realiza la imputación con las variantes del método Hot-Deck. Primero, se realiza una imputación perfecta condicionada al siguiente vector de variables sociodemográficas (***z***): género del individuo, cohorte de edad (12 cohortes), estado civil (soltero o casado), nivel de instrucción completo (sin instrucción, primaria, secundaria, preparatoria y profesional), tipo de ocupación (asalariado, cuenta propia, patrón), trabajador formal, sector de ocupación, trabajo de tiempo completo, área urbana y entidad federativa. La descripción de las variables se presenta en el Anexo 2.

Gráfica 2.

Porcentaje de observaciones con salarios no reportados después de la imputación



Notas: La muestra incluye a observaciones con 14 años o más que forman parte de la PEA, ocupados y que no caen dentro de la categoría de ocupados sin pago. Los datos reportados son para la imputación utilizando el método de Hot-Deck con una función de distancia de Mahalanobis; los porcentajes son similares para el método de Hot-Deck con imputación aleatoria.

En promedio, con este proceso se logra imputar un salario a 86% de las observaciones que no lo reportan sin incluir a las que se le asignó un ingreso laboral con múltiplos de salarios mínimos. El 14% restante se debe a que existen celdas en las que no existe un donante o un receptor para realizar la imputación. Para solucionar este problema, siguiendo la metodología del BLS (Bollinger y Hirsch, 2006), en una segunda etapa se realiza una imputación imperfecta a las observaciones que no se les pudo asignar un salario en la primera etapa. Esta etapa consiste en seguir la misma metodología propuesta de imputación pero ampliando las categorías de ciertas variables del vector de variables sociodemográficas. En este caso, se agregan los niveles de instrucción en tres categorías: nivel de instrucción bajo, medio y alto. Adicionalmente, se agregan las entidades federativas en tres regiones: norte, centro y sur. Al finalizar este proceso, en promedio el 99.8% de las observaciones totales de la muestran cuentan con un salario asignado.

6.1 Sesgo en las estimaciones

Una vez que se realiza la imputación es importante identificar los sesgos potenciales en estimaciones que no toman en cuenta las observaciones con salario no reportado. Para ello, se estima una ecuación minceriana (Mincer, 1974) en donde la variable dependiente es el logaritmo del salario mensual real; se especifica el modelo con un vector de variables sociodemográficas $\boldsymbol{x}$, en donde $\boldsymbol{z} \subseteq \boldsymbol{x}$. Posteriormente, se estima un modelo probit en donde la variable dependiente es una variable dummy donde 1 son todos aquellos individuos que caen en pobreza laboral (de acuerdo a la línea mínima de bienestar determinada por el Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL)), y 0 el caso contrario; se especifica el modelo utilizando las variables sociodemográficas incluidas en la ecuación minceriana.[7]

Es importante notar que para estimar los errores estándar se asume que las observaciones son independientes entre sí. Sin embargo, como se realiza una imputación de salarios de un donante a un receptor, este supuesto no necesariamente se mantiene. Este problema puede

[7] De acuerdo a CONEVAL (2013), las personas en pobreza laboral son todas aquellas personas que no pueden cubrir con su ingreso laboral la canasta básica alimentaria. Para calcular los individuos en pobreza laboral, se suma el ingreso laboral del hogar y se lo divide entre el número de miembros por hogar; si el ingreso por miembros del hogar no logra cubrir la canasta básica alimentaria, entonces todos los miembros de ese hogar son pobres laborales (CONEVAL, 2013).

ser corregido considerando re-muestreo a través de bootstrap. Este proceso se realizará para la muestra de imputados y el total de la muestra sumando informantes más imputados. Para la muestra que comprende solo a los ingresos reportados, estimamos errores estándar robustos que nos permiten ponderar por los factores de expansión de las observaciones reportados en las encuestas; de tal manera, la matriz de varianzas y covarianzas estimada es $A'VA$, donde A es una matriz que corrige el sesgo.[8]

De forma adicional y siguiendo a Bollinger y Hirsch (2006), se puede sospechar que los coeficientes varían entre trabajadores con diferentes características en el caso de considerar solo a los que sí reportan un ingreso, dado que al no ser faltantes aleatorios, la muestra que incluye únicamente observaciones que reportan un salario puede diferir de la muestra total, afectando las estimaciones. Estos autores proponen una corrección muestral basada en un re-muestreo de los ponderadores considerando el peso de las observaciones con ingreso faltante y como alternativa a la imputación. Para ello, primero se realiza un modelo probit para determinar si contesta la pregunta principal de ingreso o no, sujeta a una serie de características (en este caso las características utilizadas para la imputación de ingresos) y posteriormente, se calcula el inverso de la probabilidad de esta estimación, la cual se utiliza como ponderador en las estimaciones. En nuestro caso sería realizarlo invirtiendo la variable dependiente de las estimaciones en el Anexo 1. Las estimaciones con este ponderador para aquellos que reportan ingreso se presentan en el cuadro 2 en la columna 2.

El cuadro 2 muestra los resultados obtenidos de la estimación de la ecuación minceriana para el tercer trimestre de 2012. La columna (1) muestra los resultados obtenidos únicamente con las observaciones que reportan un salario y ponderando con el factor de expansión usual de la ENOE, es decir, como se realizaría de forma común. La columna dos muestra los resultados para los reportados corregidos con los nuevos ponderadores mencionados anteriormente por ingresos faltantes. La columna (3) muestra los resultados obtenidos solo para aquellas observaciones con imputaciones, y la columna (4) con toda la muestra, en ambos casos utilizando bootstrap con 500 replicaciones de re-muestreo para los errores estándar. En cada estimación se presentan los coeficientes, errores estándar en paréntesis y

[8] Otros trabajos también ignoran la variación en la muestra debido a las imputaciones (e.g. Little y Rubin, 2002).

entre corchetes a los intervalos de confianza al 95%. Las dos últimas columnas presentan razones entre los estimadores obtenidos para determinar los posibles sesgos.

Cuadro 2. Estimaciones de la ecuación minceriana. III de 2012. Variable dependiente: logaritmo del salario mensual real

	Muestra					
	Reportados (factor de expansión)	**Reportados (ponderado por faltantes)**	**Muestra solo imputados (bootstrap)**	**Total reportados más imputados (bootstrap)**		
	(1)	(2)	(3)	(4)	(1)/(4)	(1)/(2)
Hombre	0.287***	0.273***	0.183***	0.261***	1.1	1.05
	(0.00639)	(0.00420)	(0.00700)	(0.00361)		
	[0.287-0.299]	[0.264-0.280]	[0.169-0.197]	[0.254-0.268]		
Edad	0.0396***	0.0386***	0.0253***	0.0358***	1.11	1.03
	(0.00119)	(0.00087)	(0.00129)	(0.00072)		
	[.0372-.0419]	[0.0237-0.040]	[0.023-0.028]	[0.034-0.037]		
Edad al cuadrado	-0.000391***	-0.000374***	-0.000224***	-0.000345***	1.13	1.05
	(0.00001)	(0.00001)	(0.00001)	(0.00001)		
	[-0.00042- -0.00036]	[-0.00039- -0.00035]	[-0.00025- -0.00019]	[-0.00036- -0.00033]		
Casado	0.0540***	0.0632***	0.0330***	0.0573***	0.94	0.85
	(0.00655)	(0.00441)	(0.00730)	(0.00371)		
	[0.041-0.067]	[0.055-0.719]	[0.019-0.047]	[0.050-0.064]		
Años de escolaridad	0.0578***	0.0614***	0.0531***	0.0567***	1.02	0.94
	(0.00081)	(0.00056)	(0.00082)	(0.00041)		
	[0.056-0.059]	[0.060-0.062]	[0.051-0.055]	[0.056-0.057]		
Patrón	0.213***	0.208***	0.157***	0.188***	1.13	1.02
	(0.00944)	(0.00648)	(0.01070)	(0.00530)		
	[0.195-0.232]	[0.195-0.220]	[0.136-0.178]	[0.177-0.198]		
Ocupado por cuenta propia	0.642***	0.656***	0.614***	0.641***	1	0.98
	(0.01810)	(0.01250)	(0.01470)	(0.00910)		
	[0.606-0.677]	[0.637-0.680]	[0.585-0.642]	[0.624-0.659]		
Formal	0.224***	0.231***	0.253***	0.244***	0.92	0.97
	(0.00643)	(0.00430)	(0.00810)	(0.00358)		
	[0.211-0.237]	[0.222-0.239]	[0.237-0.269]	[0.237-0.251]		
Industria	0.0695	0.0466	0.282***	0.135***	0.51	1.49
	(0.05530)	(0.04150)	(0.03340)	(0.02300)		
	[-0.039-0.178]	[-0.035-0.128]	[0.216-0.348]	[0.088-0.178]		
Servicios	0.357***	0.366***	0.262***	0.321***	1.11	0.98
	(0.01270)	(0.01020)	(0.01860)	(0.00820)		
	[0.332-0.382]	[0.346-0.386]	[0.225-0.298]	[0.305-0.337]		
Otro sector	0.368***	0.352***	0.242***	0.304***	1.21	1.05
	(0.01260)	(0.01010)	(0.01780)	(0.00800)		
	[0.343-0.392]	[0.332-0.371]	[0.208-0.277]	[0.229-0.320]		
Trabaja tiempo completo	0.451***	0.419***	0.216***	0.384***	1.17	1.08
	(0.00745)	(0.00508)	(0.01022)	(0.00420)		
	[0.436-0.466]	[0.409-0.429]	[0.196-0.236]	[0.376-0.393]		
Urbano	0.121***	0.139***	0.207***	0.156***	0.78	0.87
	(0.00812)	(0.00649)	(0.01370)	(0.00571)		
	[0.105-0.137]	[0.127-0.152]	[0.180-0.234]	[0.145-0.167]		
Constante	5.507***	5.503***	6.216***	5.700***	0.97	1
	(0.02820)	(0.02170)	(0.03440)	(0.01650)		
	[5.451-5.562]	[5.460-5.545]	[6.148-6.283]	[5.668-5.732]		
Efectos fijos a nivel estado	Sí	Sí	Sí	Sí		
N	120,111	120,111	38,389	158,500		
R-cuadrado	0.459	0.451	0.415	0.436		

Notas: Errores estándar en paréntesis; intervalos de confianza al 95% en corchetes. En (1) y (2) errores estándar robustos; en (1) ponderados por el factor de expansión reportado en la ENOE; en (2) ponderados por el inverso de la probabilidad de reportar ingreso; en (3) y (4) con bootstrap con 500 replicaciones de re-muestreo. *** p<0.01. Categorías base: Asalariados, Agricultura. Las columnas (3) y (4) consideran las imputaciones mediante el método de Hot-Deck con función de distancia de Mahalanobis; los resultados son robustos a las estimaciones considerando el método de Hot-Deck con pareamiento aleatorio.

Para conocer la dirección del sesgo en las estimaciones que no consideran a las observaciones con salario no reportado, en la quinta columna se muestra una razón de los parámetros obtenidos entre las observaciones que reportan un salario y la muestra total. Si la razón es mayor a 1, existe un sesgo hacia la derecha, y si la razón es menor a 1, entonces existe un sesgo hacia la izquierda.[9] En general, como se puede observar en el cuadro, existe un sesgo hacia la derecha cuando no se consideran las observaciones que no reportan un salario. La segunda columna muestra los parámetros obtenidos con la ponderación por re-muestreo con base en la muestra que reporta un salario; en general presentan coeficientes dentro del rango de las estimaciones obtenidas con la muestra que reporta un salario y la muestra total que incluye las observaciones con salario imputado. Esto es similar a los resultados obtenidos en Bollinger y Hirsch (2006), aunque en algunos casos, los resultados son mayores a ese rango, como es en el caso del coeficiente de educación.

Ahora bien, el cuadro anterior muestra solo un período como ejemplo por cuestión de espacio, pero es importante identificar el sesgo en las estimaciones a lo largo del tiempo. Para ello, en la siguientes subsecciones, se analizan dos variables como ejemplo de interés ampliamente estudiadas para el caso de México: retornos a la educación y premio a la formalidad, y sus efectos sobre salarios y sobre pobreza laboral.

6.2 Retornos a la educación

Los retornos a la educación han sido ampliamente estudiados en México (e.g. Psacharopoulos *et al*., 1996; Bracho y Zamudio, 1994; López-Acevedo, 2001; Rodríguez-Oreggia, 2014), sin embargo, ninguno de estos estudios considera el sesgo potencial en las estimaciones por no incluir a las observaciones con salario no reportado. En la gráfica 3 se muestran los retornos a la educación por trimestres en México de 2005 a 2013, medidos como el porcentaje de retorno promedio por cada año adicional de educación. Se incluyen los resultados obtenidos

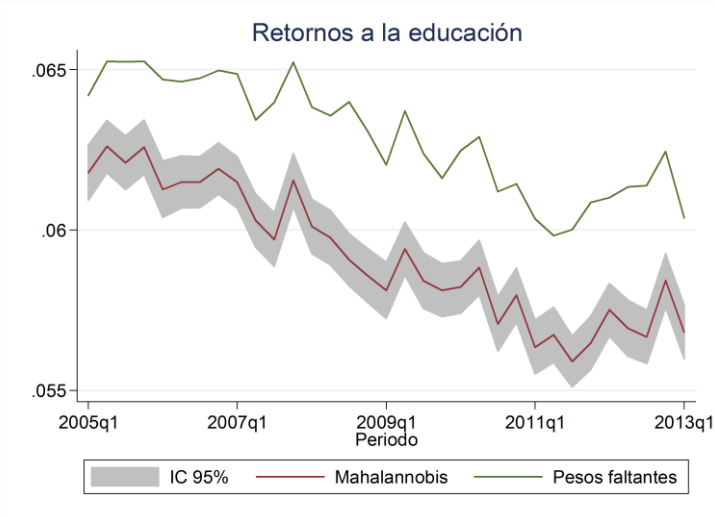
[9] Bajo este criterio, una razón mayor a uno implicaría un sesgo hacia la derecha mientras que una razón menor a uno implicaría un sesgo hacia la izquierda. Sin embargo, es importante notar que si los coeficientes son negativos, una razón mayor a uno estaría indicando un sesgo hacia la derecha cuando en realidad el sesgo es hacia la izquierda. Por lo tanto, para evitar este problema, de ahora en adelante, un sesgo hacia la derecha implica una sobre-estimación de la magnitud del coeficiente, mientras que un sesgo hacia la izquierda implica una subestimación de la magnitud del coeficiente.

con la muestra que reporta un salario (con ambas ponderaciones), y con la muestra total.[10] Para que los resultados sean robustos a la imputación, se incluyen los resultados obtenidos con los dos criterios de imputación Hot-Deck propuestos, además se calculan los intervalos de confianza al 95% en cada caso.

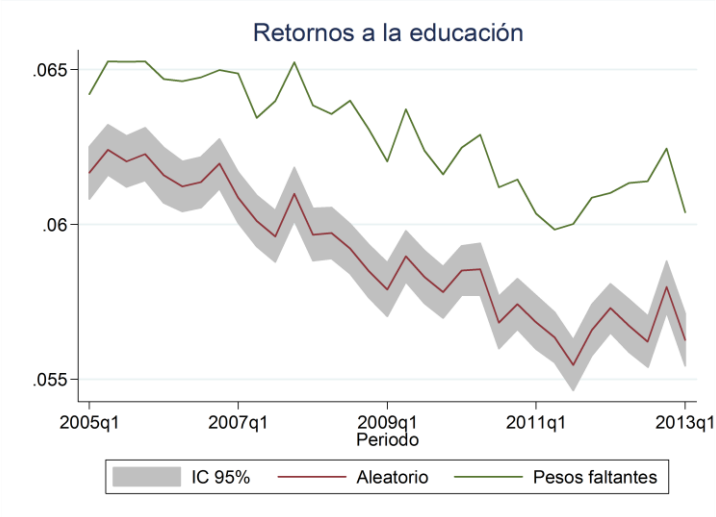
[10] La muestra total son todos aquellos individuos de 14 años o más, ocupados y que no caen dentro de la categoría de ocupados sin pago. Las observaciones a las que no se les pudo imputar un salario después de las dos etapas de imputación no son consideradas en la muestra total y corresponden al 0.2% de la muestra.

Gráfica 3. Retornos a la educación

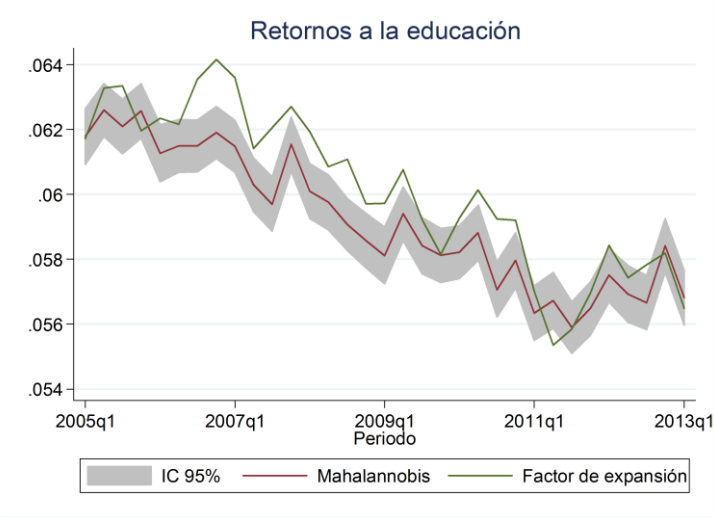
A.



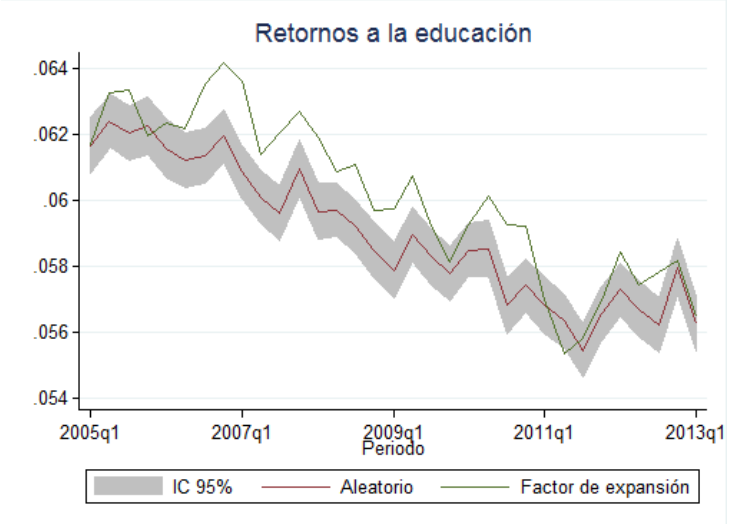
B.



C.



D.



Notas: Intervalos de confianza al 95%. Se incluyen los resultados obtenidos con los dos criterios de imputación (HD-A: Hot-Deck con imputación aleatoria; HD-M: Hot-Deck con función de distancia de Mahalanobis), en ambos casos los errores estándar se calcularon utilizando bootstrap con 500 replicaciones de re-muestreo; así como la muestra solo reportada ponderada por ingresos faltantes y la reportada ponderada con factor de expansión de la ENOE. Se controló por género, experiencia laboral y su cuadrado, estado civil, tipo de ocupación, formalidad, sector, trabajo de tiempo completo, área urbana y entidad federativa. En los paneles A y B se compara las estimaciones de la muestra total considerando ambos métodos de imputación con las estimaciones utilizando los pesos faltantes. En los paneles C y D se compara las estimaciones de la muestra total considerando ambos métodos de imputación con las estimaciones utilizando el factor de expansión reportado en la ENOE.

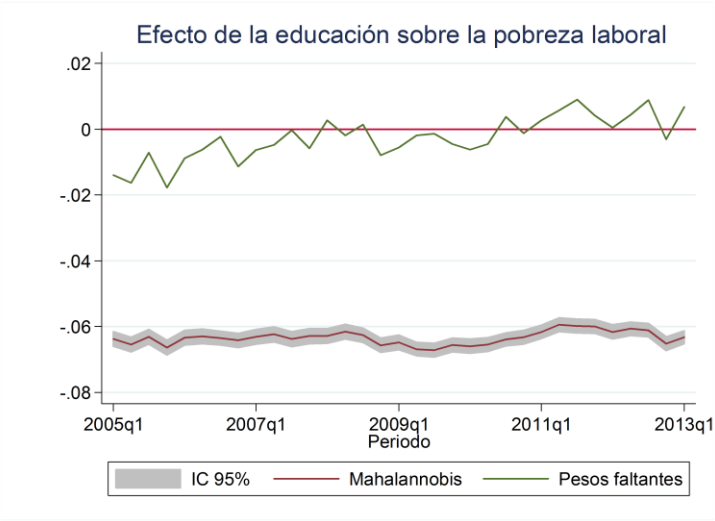
Como se puede observar en la gráfica 3, en promedio, para todos los periodos analizados a mayor educación mayor salario, aunque existe una caída de los retornos en el tiempo. Es importante notar que las estimaciones de los retornos a la educación tienen un sesgo cuando no se incluyen a las observaciones que no reportan un salario, aunque para el caso de los paneles C y D (cuando se compara las estimaciones con imputación por ambos métodos con las estimaciones considerando factores de expansión) el sesgo no es necesariamente significativo para todos los periodos. No obstante, en general se puede decir que no incluir a las observaciones con salario no reportado pudiera generar sobre-estimaciones en los retornos a la educación. Para este caso, la diferencia es más notoria cuando se consideran los pesos faltantes.

En la gráfica 4 se muestran los efectos promedio de la educación sobre la pobreza laboral para cada trimestre de 2005 a 2013.[11] Como se puede observar en los cuatro paneles de la gráfica (paneles A, B, C y D), considerar únicamente a las observaciones con salario reportado genera una subestimación en la magnitud del coeficiente de la educación sobre la pobreza laboral. El efecto promedio de la educación sobre la pobreza laboral es sustancialmente más bajo cuando no se considera la muestra total, aún con ambos tipos de ponderadores; inclusive para ciertos periodos es no significativo. En cambio, para la muestra completa con ingresos imputados, la dirección del efecto de educación sobre la pobreza laboral claramente va de acuerdo a la teoría para todos los periodos, es decir, mayor educación reduce la probabilidad de caer en pobreza laboral; los resultados son robustos a ambos criterios de imputación.

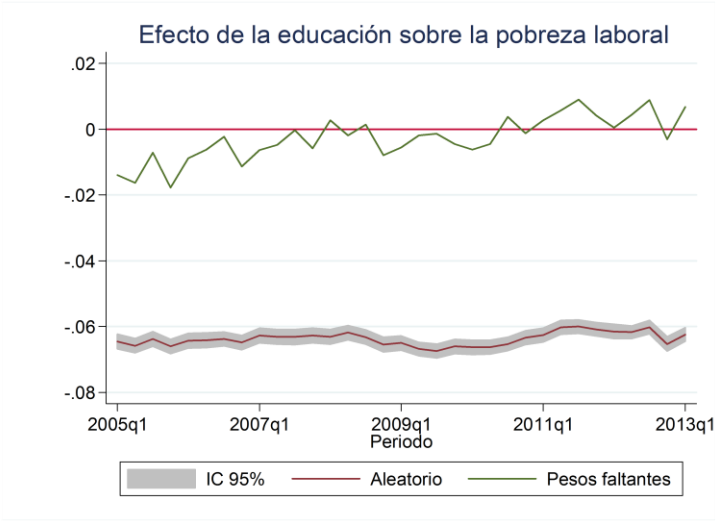
[11] Se realizaron tres estimaciones de pobreza laboral. La primera únicamente incluye a las observaciones con salario reportado. Las otras dos son estimaciones utilizando la muestra total con cada uno de los criterios de imputación propuestos. Ver Anexo 3 para una descripción más amplia sobre la evolución de la pobreza laboral en México considerando los salarios no reportados. En Campos-Vázquez (2013) se hace una descripción más detallada por diversos grupos poblacionales.

Gráfica 4. Efecto de la educación sobre probabilidad de caer en pobreza laboral

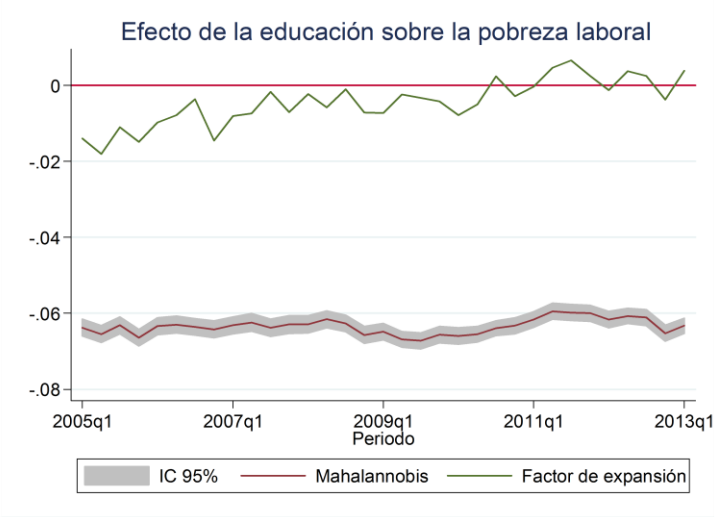
A.



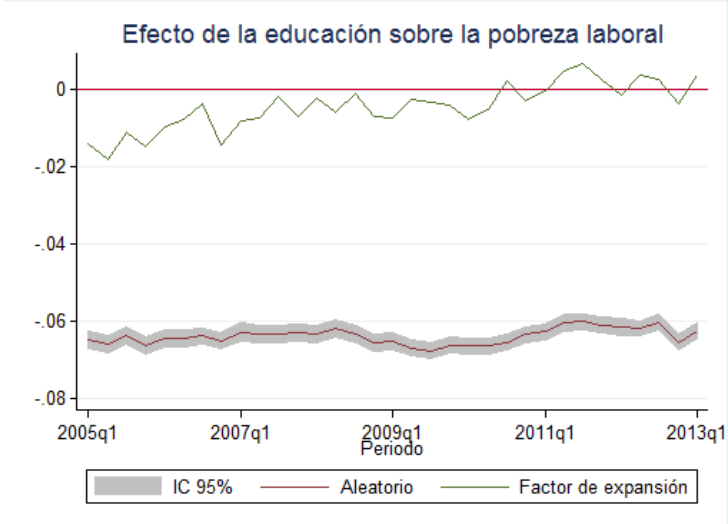
B.



C.



D.



Notas: Intervalos de confianza al 95%. Se incluyen los resultados obtenidos con los dos criterios de imputación (HD-A: Hot-Deck con imputación aleatoria; HD-M: Hot-Deck con función de distancia de Mahalanobis), en ambos casos los errores estándar se calcularon utilizando bootstrap con 500 replicaciones de re-muestreo; así como la muestra solo reportada ponderada por ingresos faltantes y la reportada ponderada con factor de expansión de la ENOE. Se controló por género, experiencia laboral y su cuadrado, estado civil, tipo de ocupación, formalidad, sector, trabajo de tiempo completo, área urbana y entidad federativa. En los paneles A y B se compara las estimaciones de la muestra total considerando ambos métodos de imputación con las estimaciones utilizando los pesos faltantes. En los paneles C y D se compara las estimaciones de la muestra total considerando ambos métodos de imputación con las estimaciones utilizando el factor de expansión reportado en la ENOE.

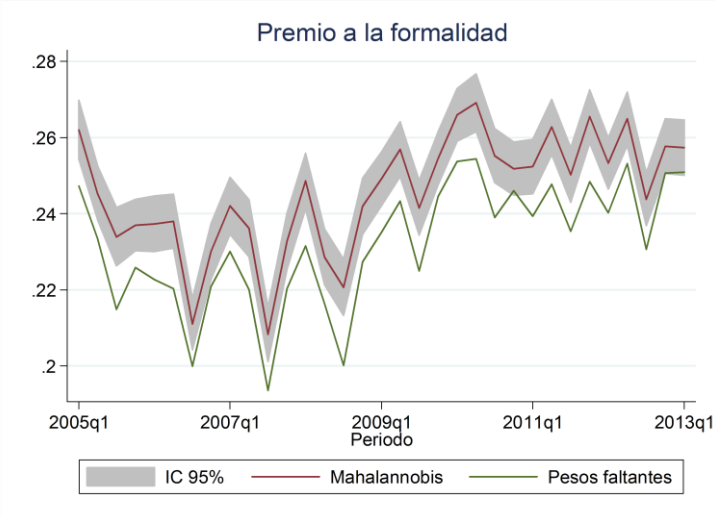
De los resultados se concluye que en general, a mayor educación menor probabilidad de caer en pobreza laboral, considerando que los resultados son más consistentes utilizando la muestra completa con imputaciones. Esto es, la probabilidad estimada es más consistente en términos teóricos e intuitivos si se considera a la muestra total que incluye las observaciones imputadas por salario no reportado.

6.3 Formalidad

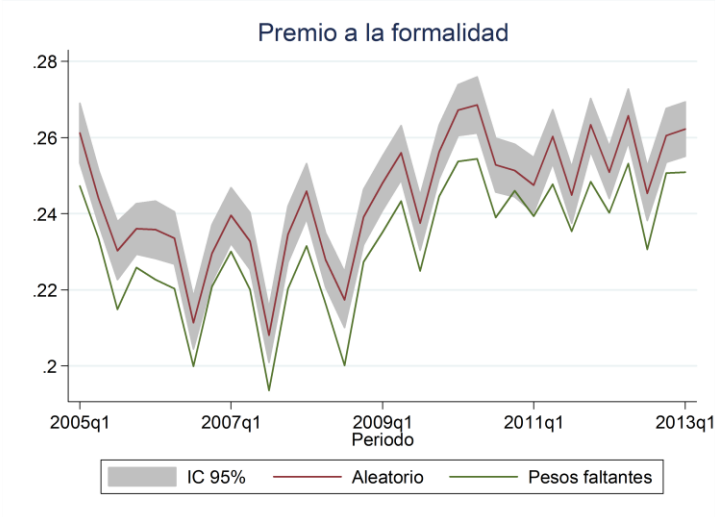
La formalidad laboral es un tema amplio de estudio, sobre todo por su impacto en la productividad laboral y por sus implicaciones en la provisión de beneficios por seguridad social. En México, alrededor de dos tercios de los trabajos se encuentran en el sector informal de la economía y tienen un impacto muy bajo en la productividad de los trabajadores (Rodríguez-Oreggia, 2007 y 2010). En este sentido, para el caso de México es importante identificar el efecto de la formalidad en los trabajos sobre los salarios o sobre la probreza laboral. En la gráfica 5 se pueden observar los premios salariales a la formalidad laboral, medida como acceso a seguridad social por el trabajo, del primer trimestre de 2005 al primer trimestre de 2013, reportados en la ENOE. Como se puede observar en la gráfica, los premios salariales a la formalidad son positivos y elevados. En promedio, un trabajador formal gana 24.1% más que un trabajador informal.

Gráfica 5. Premio salarial a la formalidad

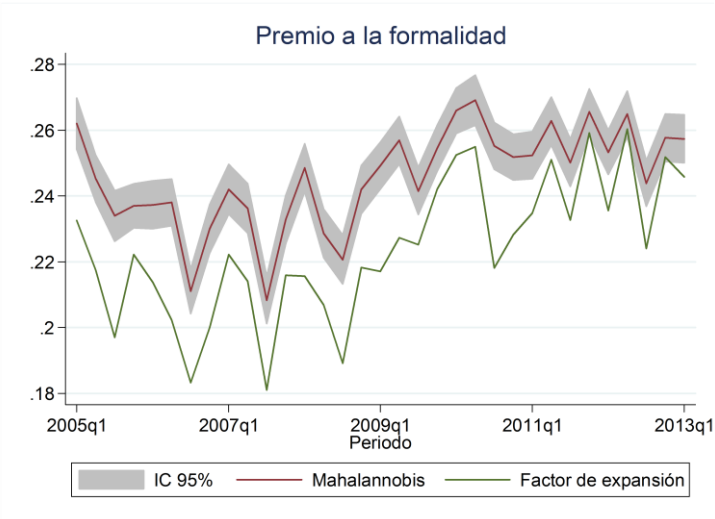
A.



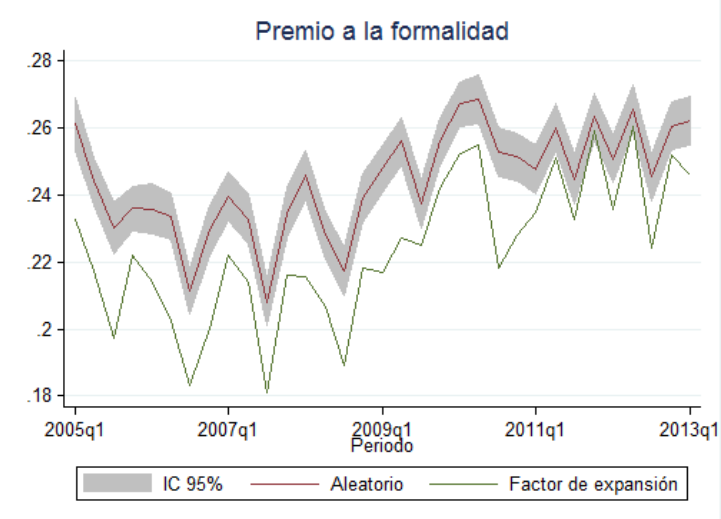
B.



C.



D.



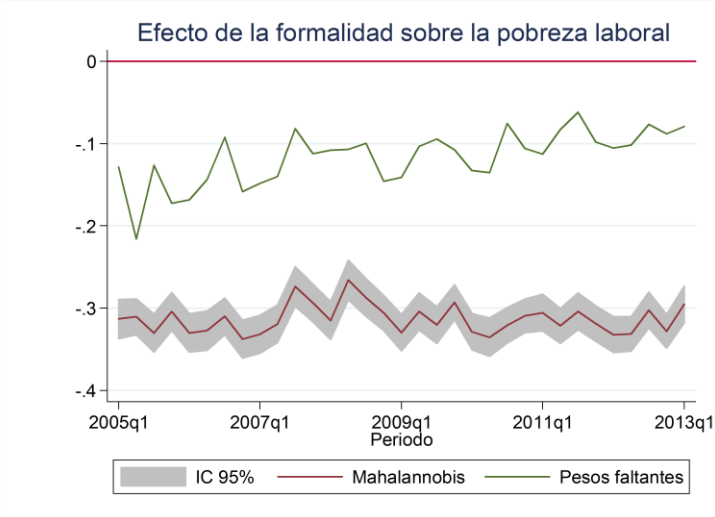
Notas: Intervalos de confianza al 95%. Se incluyen los resultados obtenidos con los dos criterios de imputación (HD-A: Hot-Deck con imputación aleatoria; HD-M: Hot-Deck con función de distancia de Mahalanobis), en ambos casos los errores estándar se calcularon utilizando bootstrap con 500 replicaciones de re-muestreo; así como la muestra solo reportada ponderada por ingresos faltantes y la reportada ponderada con factor de expansión de la ENOE. Se controló por género, experiencia laboral y su cuadrado, estado civil, tipo de ocupación, formalidad, sector, trabajo de tiempo completo, área urbana y entidad federativa. En los paneles A y B se compara las estimaciones de la muestra total considerando ambos métodos de imputación con las estimaciones utilizando los pesos faltantes. En los paneles C y D se compara las estimaciones de la muestra total considerando ambos métodos de imputación con las estimaciones utilizando el factor de expansión reportado en la ENOE.

Sin embargo, es importante identificar el sesgo de los premios salariales a la formalidad al no considerar las observaciones con salario faltante. En la gráfica 5 se pueden observar los premios salariales a la formalidad considerando las observaciones con salario reportado y el total de observaciones, incluyendo aquellas con salario imputado. Como se puede observar en la gráfica 5 (paneles A, B, C y D) existe una subestimación consistente en el tiempo del premio salarial cuando no se considera a las observaciones con salario no reportado comparado con las estimaciones considerando pesos faltantes y factores de expansión. No obstante, es importante notar que la subestimación es menor cuando se consideran los pesos faltantes (paneles A y B). Los resultados son robustos a ambos criterios de imputación.

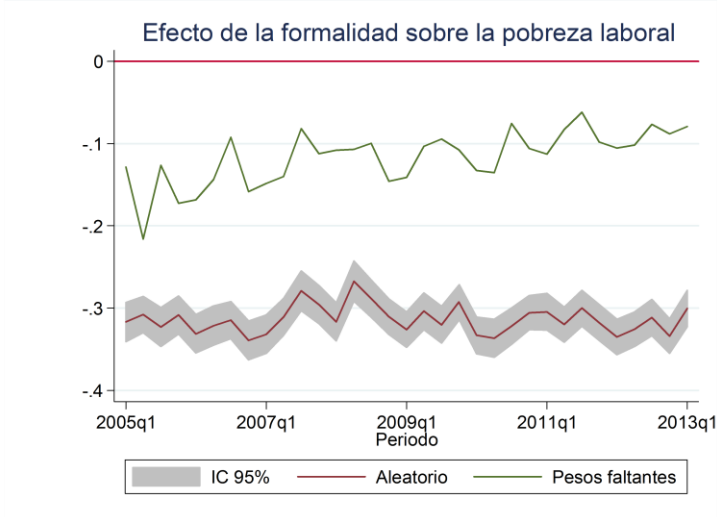
Finalmente, en la gráfica 6 se muestran los coeficientes de la formalidad laboral sobre la pobreza laboral. Como se puede observar en la gráfica, la formalidad laboral tiene un efecto promedio negativo sobre la pobreza laboral. Es decir, los trabajadores formales tienen menor probabilidad de caer en pobreza laboral comparado con los trabajadores informales.

Gráfica 6. Efecto de la formalidad sobre probabilidad de caer en pobreza laboral

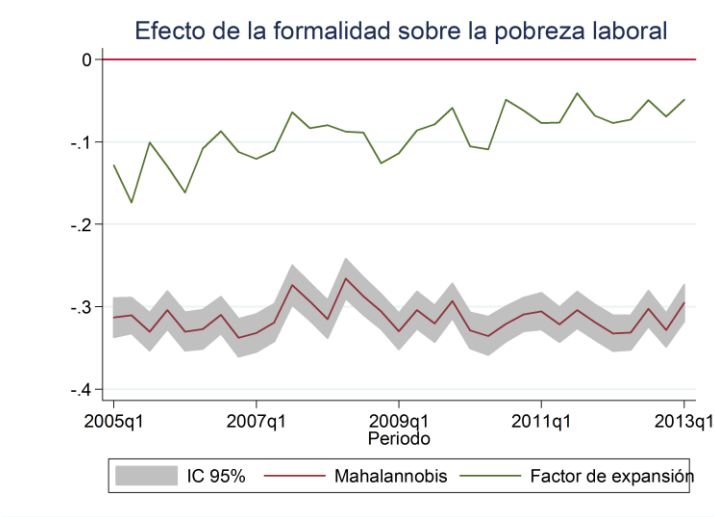
A.



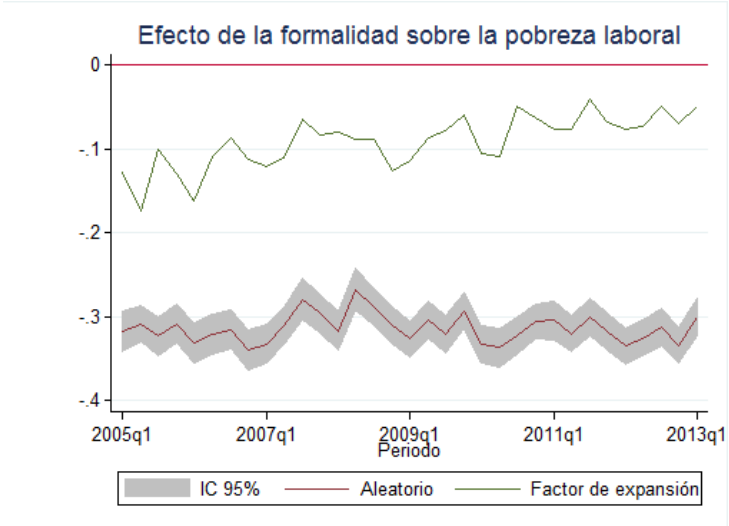
B.



C.



D.



Notas: Intervalos de confianza al 95%. Se incluyen los resultados obtenidos con los dos criterios de imputación (HD-A: Hot-Deck con imputación aleatoria; HD-M: Hot-Deck con función de distancia de Mahalanobis), en ambos casos los errores estándar se calcularon utilizando bootstrap con 500 replicaciones de re-muestreo; así como la muestra solo reportada ponderada por ingresos faltantes y la reportada ponderada con factor de expansión de la ENOE. Se controló por género, experiencia laboral y su cuadrado, estado civil, tipo de ocupación, formalidad, sector, trabajo de tiempo completo, área urbana y entidad federativa. En los paneles A y B se compara las estimaciones de la muestra total considerando ambos métodos de imputación con las estimaciones utilizando los pesos faltantes. En los paneles C y D se compara las estimaciones de la muestra total considerando ambos métodos de imputación con las estimaciones utilizando el factor de expansión reportado en la ENOE.

De los resultados se puede observar claramente una subestimación del efecto promedio de la formalidad sobre la probabilidad de caer en pobreza laboral cuando no se considera la muestra total; la probabilidad de caer en pobreza laboral es aún menor cuando las estimaciones se realizan con la muestra total considerando las observaciones con salario imputado. Los resultados también son robustos a ambos métodos de imputación.

En general, se puede afirmar que existe una amplia diferencia entre resultados tanto en niveles de pobreza laboral como en estimaciones derivadas del análisis laboral involucrando solo ingresos reportados, comparado con aquellas estimaciones que incluyen a las observaciones con ingresos imputados. Las implicaciones parecen ser claras respecto al uso de la información de ingreso derivado de la ENOE: pareciera que el uso de imputación de ingresos puede ayudar a identificar mejor los efectos de determinadas variables sobre ingreso y pobreza laboral, y por ende a focalizar mejor a grupos vulnerables (ver por ejemplo Rodríguez-Oreggia, López-Videla y Prudencio, 2013). Sin embargo, la selección del modelo con correcciones siempre depende de los datos que se tienen, así como del objetivo de cada investigación, pero siempre considerando algún método de corrección (para una discusión ver Bollinger y Hirsch, 2006). Lo que sí queda claro es que el ignorar los ingresos no reportados conlleva a un sobre reporte en el cálculo de los que caen en pobreza laboral, lo cual tiene implicaciones en estimaciones que involucran a esta variable.

7 Conclusiones

Las respuestas faltantes, sobre todo en ingresos, han sido materia de análisis en diversos países, especialmente en Estados Unidos. No obstante, en México ya se empieza a considerar el efecto potencial que puede tener este problema sobre indicadores de política pública, en especial sobre mediciones de la pobreza laboral (ver Rodríguez-Oreggia *et al*., 2012, y Campos-Vázquez, 2013). En este artículo se ha buscado analizar el método de imputación de ingresos para la ENOE bajo dos variantes y sus efectos sobre estimaciones de capital humano, así como indicadores basados en estos datos, tales como pobreza laboral. En una primera etapa, se presentan los métodos de imputación con base en la secuencia de preguntas de ingreso de la ENOE. En una segunda parte, se analizan los resultados de la imputación sobre ecuaciones mincerianas y sobre la probabilidad de caer en pobreza laboral comparándolos

con los resultados ponderados que incluyen en la muestra solo a observaciones con ingreso reportado.

En general, se muestra que al no considerar las observaciones con ingresos faltantes en la ENOE, el nivel de las estimaciones de pobreza laboral está sobre estimado. Adicionalmente, al comparar las estimaciones que incluyen a la muestra con ingresos no reportados con las estimaciones que incluyen únicamente a la muestra con ingresos reportados, se detectaron sesgos en las estimaciones de capital humano y de pobreza laboral. Los retornos a la educación serían un poco más bajos al considerar la muestra imputada, alrededor de medio punto porcentual, pero mayor si se corrige muestralmente, tomando en cuenta que para el primer caso el sesgo no necesariamente es significativo para todos los periodos. El uso de la muestra imputada permite también obtener cálculos más consistentes sobre el efecto de la educación en la reducción de la pobreza laboral. También hay un efecto mayor de la formalidad en evitar caer en pobreza laboral al utilizar esta muestra. Además, se presentaron los resultados que utilizan una corrección muestral basada en ponderadores de peso por los que sí reportan un ingreso laboral en la encuesta.

Las implicaciones en términos de políticas públicas se deriva del hecho de identificar los efectos de variables asociadas ya sea al ingreso laboral o a la pobreza laboral, y que permitiría una mejor focalización y un mayor impacto al establecer programas públicos enfocados en reducir la vulnerabilidad y/o aumentar ingresos. Los resultados sugieren que el uso de imputaciones en el ingreso para el caso de la ENOE merece una discusión académica seria y un replanteamiento en el cálculo de indicadores que utilicen esta variable por parte de organismos gubernamentales.

8 Referencias

Bollinger, C. y Hirsch, B. (2006), "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching". Journal of Labor Economics, 24, pp. 483-519.

Campos-Vázquez, R. (2013), "Efecto de los Ingresos No Reportados en el Nivel y Tendencia de la Pobreza Laboral en México", Ensayos, 32, 2, pp. 23-54.

CONEVAL (2013), "Tendencias Económicas y Sociales de Corto Plazo. Resultados Nacionales. Mayo 2013" (http://www.coneval.gob.mx/Informes/ITLP/PRIMER%20TRIMESTRE%202013/ITLP%20NACIONAL%20mayo%202013.pdf) 5 de agosto.

Crawford, S. (1990), "Internal Memoranda", Bureau of Labor Statistics, 1989.

David, M., Little, R., Samuel, M. y Triest, R. (1986), "Alternative Methods for CPS Income Imputation", Journal of the American Statistical Association, 77, pp. 251-261.

Frick, J.R., y Grabka, M.M. (2003), "Missing Income Data in the German SOEP: Incidence, Imputation, and its Impact on the Income Distribution", DIW Discussion Paper 376. DIW, Berlín.

Groves, R., Singer, E., y Corning, A.D. (1999), "Decision Making in Survey Participation: Theory and a Test", University of Michigan, mimeo.

Hirsch, B. y Schumacher, E. (2004), "Match Bias in Wage Gap Estimates due to Earnings Imputation", Journal of Labor Economics, 22, pp. 689-722.

Horowitz, J., y Manski C. (1998), "Censoring of Outcomes and Regressors Due to Survey Non-Response: Identification and Estimation Using Weights and Imputations", Journal of Econometrics, 84, pp. 37-58.

Horowitz, J., y Manski C. (2000), "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data", Journal of the American Statistical Association, 95, pp. 77-84.

Kalton, G. y Kaspryzk, D. (1982), "Imputing for Missing Survey Responses", Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 22-31.

Lillard, L., Smith, J. y Welch, F. (1986), "What do We Really Know About Wages? The Importance of Nonreporting and Census Imputation", Journal of Political Economy, 94, pp. 489-506.

Little, R. y Rubin, D. (2002), Statistical Analysis with Missing Data (2da ed), Nueva York, John Wiley and Sons.

Lopez-Acevedo, G. (2001), "Evolution of Earnings and Rates of Return to Education in Mexico", Policy Research Working Paper 2691, Washington, World Bank.

Mincer, J. (1974), Schooling, Experience and Earnings, Nueva York, National Bureau of Economic Research.

Montaquila, J. y Ponikowski, C. (1995), "An Evaluation of Alternative Imputation Methods", Proceedings of the Section on Survey Research Methods, American Statistical Association.

Psacharopoulos, G., Velez, E., Panagides, A., y Yang, H. (1996), "Returns to Education During Economic Boom and Recession: Mexico 1984, 1989 and 1992", Education Economics, 4, pp. 219-230.

Rodríguez-Oreggia, E. (2007), "The Informal Sector in Mexico: Characteristics and Dynamics, Social Perspectives, 9, 1, pp. 89-156.

Rodíguez-Oreggia, E. (2014), "Instituciones, Geografía y Retornos a la Educación en México". En R. de la Torre, E. Rodríguez-Oreggia, e I. Soloaga (eds), Temas de Política Social en México. CIDE, en imprenta.

Rodríguez-Oreggia, E. (2010), "Informalidad y Políticas Públicas: El caso de México", en Adenauer (ed) Informalidad y Políticas Públicas en América Latina, Río de Janeiro, Adenauer-KAS.

Rodríguez-Oreggia, E., López-Videla, B. y Prudencio, D. (2012), "Índice de la Tendencia Laboral de la pobreza: Consideraciones Sobre Adiciones al Grupo de Trabajadores con Ingreso Reportado", EGAP, mimeo.

Rodríguez-Oreggia, E., López-Videla, B. y Prudencio, D. (2013) "Labor Vulnerability and the Evolution of the Working Poor in Mexico", artículo presentado en el congreso de la Society for the Study of Economic Inequality, Bari, Italia.

Rubin, D. (1987), "Multiple Imputation for Nonresponse in Surveys", Nueva York, Wiley.

Treiman, D. (2009), Quantitative Data Analysis, San Francisco, Jossey-Bass.

West, S.A. Butani, S. y Witt, M. (1990), "Alternative Imputation Methods for Wage Data", Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 254-259.

ANEXO 1 Cuadro A1. Probabilidad de no reportar un salario
Efectos marginales dy/dx

	2005'II	2006'II	2007'II	2008'II	2009'II	2010'II	2011'II	2012'II
Hombre	0.0130***	0.0128***	0.0119***	0.0126***	0.0196***	0.0197***	0.0163***	0.0199***
	(0.00184)	(0.00168)	(0.00182)	(0.00206)	(0.00232)	(0.00227)	(0.00245)	(0.00244)
Casado	-0.0161***	-0.0163***	-0.0206***	-0.0173***	-0.0208***	-0.0220***	-0.0226***	-0.0276***
	(0.00201)	(0.00189)	(0.00206)	(0.00228)	(0.00256)	(0.00246)	(0.00265)	(0.00266)
Primaria	-0.00139	-0.00399	-0.000838	0.00595*	0.0000427	-0.00165	-0.00683	-0.00727
	(0.00262)	(0.00248)	(0.00288)	(0.00349)	(0.00407)	(0.00400)	(0.00449)	(0.00450)
Secundaria	0.00748**	0.00426	0.00548	0.0182***	0.0159***	0.0166***	0.0110**	0.00705
	(0.00354)	(0.00313)	(0.00352)	(0.00435)	(0.00492)	(0.00481)	(0.00521)	(0.00519)
Preparatoria	0.0323***	0.0257***	0.0275***	0.0551***	0.0552***	0.0582***	0.0510***	0.0536***
	(0.00478)	(0.00412)	(0.00459)	(0.00584)	(0.00630)	(0.00628)	(0.00660)	(0.00656)
Profesional	0.0740***	0.0547***	0.0639***	0.102***	0.110***	0.127***	0.127***	0.149***
	(0.00676)	(0.00581)	(0.00636)	(0.00793)	(0.00832)	(0.00870)	(0.00904)	(0.00961)
Patrón	0.0620***	0.0682***	0.0816***	0.0933***	0.0844***	0.106***	0.114***	0.123***
	(0.00714)	(0.00660)	(0.00713)	(0.00778)	(0.00787)	(0.00855)	(0.00872)	(0.00938)
Ocupado por cuenta propia	0.0723***	0.0734***	0.0809***	0.0831***	0.0918***	0.0955***	0.107***	0.116***
	(0.00340)	(0.00337)	(0.00373)	(0.00374)	(0.00411)	(0.00415)	(0.00440)	(0.00440)
Formal	0.00903***	0.0177***	0.0172***	0.0235***	0.0250***	0.0339***	0.0401***	0.0496***
	(0.00260)	(0.00243)	(0.00266)	(0.00287)	(0.00328)	(0.00335)	(0.00355)	(0.00358)
Industria	-0.0661***	-0.0580***	-0.0729***	-0.0802***	-0.0924***	-0.0906***	-0.103***	-0.109***
	(0.00173)	(0.00162)	(0.00180)	(0.00194)	(0.00218)	(0.00211)	(0.00235)	(0.00223)
Servicios	-0.108***	-0.0855***	-0.111***	-0.122***	-0.129***	-0.129***	-0.136***	-0.161***
	(0.00382)	(0.00346)	(0.00400)	(0.00430)	(0.00468)	(0.00451)	(0.00481)	(0.00491)
Otro sector	0.602***	0.476***	0.436***	0.391***	0.414***	0.476***	0.439***	0.289***
	(0.03330)	(0.02910)	(0.03100)	(0.02910)	(0.03000)	(0.03220)	(0.03890)	(0.02450)
Trabajo de tiempo completo	0.0133***	0.00959***	0.0136***	0.0150***	0.0216***	0.0187***	0.0220***	0.0223***
	(0.00163)	(0.00157)	(0.00166)	(0.00190)	(0.00215)	(0.00216)	(0.00232)	(0.00233)
Urbano	-0.00831***	-0.0119***	-0.00865***	0.000091	-0.00315	-0.00568**	0.00425	0.00192
	(0.00222)	(0.00234)	(0.00241)	(0.00247)	(0.00307)	(0.00289)	(0.00300)	(0.00287)
Efectos fijos por cohorte de edad	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Efectos fijos por estado	Sí	Sí	Sí	Sí	Sí	Sí	Sí	Sí
N	155,515	161,245	162,462	162,737	154,822	158,646	155,513	159,679

Notas: Errores estándar en paréntesis. * p<0.1, ** p <0.05, *** p<0.01. Categorías base: Sin instrucción, Asalariados, Sector Primario.

Anexo 2. Variables consideradas para el matching entre individuos que reportan y no reportan ingreso

Cuadro A2. Celdas de acuerdo al criterio de imputación

Criterio de imputación	Número de Celdas	Categorías
Género	2	Hombre, Mujer
Cohorte de edad	12	14-17, 18-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, 61-65, 66+
Estado civil	2	Soltero, Casado
Educación		
Imputación perfecta	5	Sin instrucción, Primaria, Secundaria, Preparatoria, Profesional
Imputación imperfecta	3	Nivel de instrucción bajo, medio y alto
Ocupación	3	Asalariado, Cuenta Propia, Patrón
Formalidad	2	Formal, Informal
Sector de Ocupación	4	Primario, Secundario, Terciario, No especificado
Trabajo de tiempo completo	2	Trabaja tiempo completo, No trabaja tiempo completo
Área urbana	2	Urbano, Rural
Estado		
Imputación perfecta	32	32 entidades federativas
Imputación imperfecta	3	Norte, Centro, Sur
Celdas Posibles		
Imputación perfecta	737,280	
Imputación imperfecta	41,472	

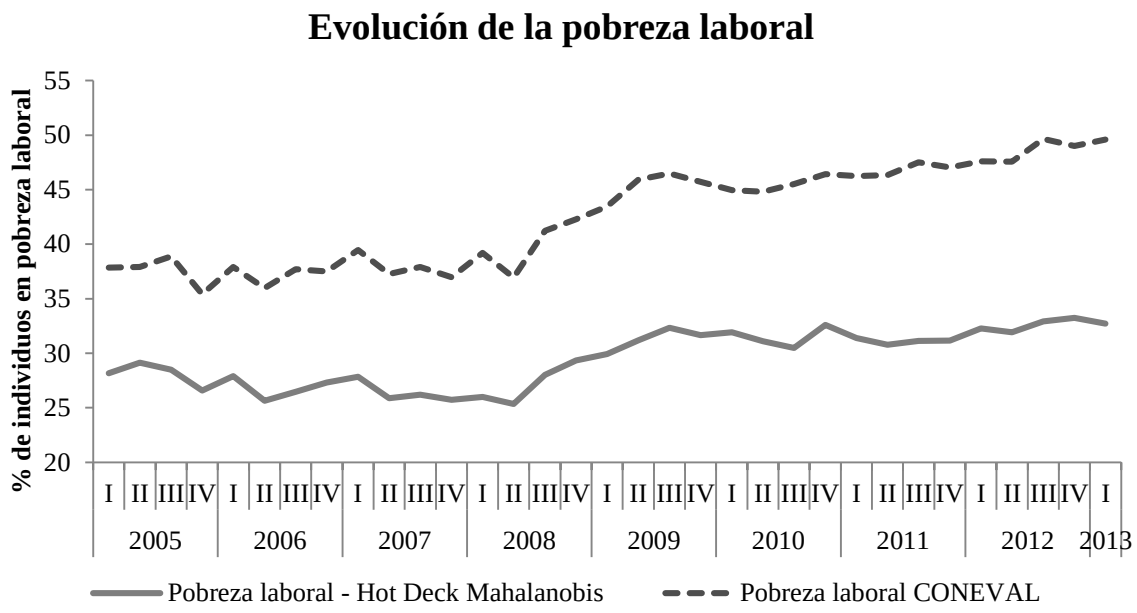
Anexo 3. Evolución de la pobreza laboral en México

De acuerdo a CONEVAL (2013), los individuos en pobreza laboral son todas aquellas personas que no pueden cubrir una canasta alimentaria básica con el salario laboral.[12] Para realizar el cálculo se suma el ingreso laboral y se lo divide entre los miembros por hogar. Si el salario por miembro del hogar cae por debajo de la línea de bienestar mínimo, entonces todos los individuos del hogar caen dentro de la categoría de pobres laborales. Si el salario está por encima de la línea de bienestar mínimo, entonces los miembros de ese hogar son no pobres laborales.

En la gráfica A.1 se puede observar la evolución de la pobreza laboral estimada con el criterio de CONEVAL desde el primer trimestre de 2005 al primer trimestre de 2013. Adicionalmente, en la gráfica se muestra la evolución de la pobreza laboral considerando las observaciones con salario imputado de acuerdo al criterio de Hot-Deck con función de distancia de Mahalanobis, para el mismo periodo.[13] Como se observa claramente en la gráfica, con ambos criterios la pobreza laboral ha aumentado significativamente a partir de la mitad de año de 2008 y todavía no se han podido disminuir los niveles de pobreza laboral a los periodos pre crisis. Además, es importante resaltar la sobre estimación de los individuos que caen en pobreza laboral cuando no se consideran las observaciones con salario imputado. Aunque en los últimos años la pobreza laboral ha aumentado, los niveles no son tan altos como los estimados de acuerdo al criterio de CONEVAL. En este sentido, estos datos deben leerse cuidadosamente tomando en cuenta la importancia de las observaciones que trabajan pero que no reportan un salario.

[12] CONEVAL publica de forma recurrente las líneas de pobreza laboral en donde todos aquellos individuos que se encuentren por debajo de ellas son considerados como pobres laborales.
[13] Los resultados son muy similares a la imputación mediante Hot-Deck de forma aleatoria.

Gráfica A.1.



Notas: En la muestra se incluyen a todos los hogares de la muestra de la ENOE. Los resultados con las observaciones con salario imputado mediante el método de Hot-Deck de forma aleatoria son similares a los reportados con imputación mediante el método de Hot-Deck con función de distancia de Mahalanobis.