"EVALUACIÓN DE MÉTODOS DE CORRECCIÓN POR DATOS OMITIDOS EN EL CONTEXTO DE DESIGUALDAD DE OPORTUNIDADES"

TESIS PARA OPTAR AL GRADO DE

Magíster en Economía

Alumno: Daniela Edith Luengo Aravena

Profesor Guía: Esteban Puentes – Dante Contreras

Santiago, 15 de mayo de 2013

Evaluación de Métodos de Corrección por Datos Omitidos en el Contexto de Desigualdad de Oportunidades

Alumna: Daniela Luengo

Profesores Guía: Dante Contreras - Esteban Puentes

15 de mayo de 2013

Resumen

El presente trabajo evalúa el desempeño de procedimientos simples -ampliamente utilizados en la literatura para corregir problemas de pérdida de datos- cuando la distribución de la omisión es no aleatoria. Esto, dentro del contexto de la medición de la desigualdad y la desigualdad de oportunidades en el mercado laboral. Para ello se utiliza un experimento en donde se generan observaciones omitidas desde un vector de datos completo, siguiendo un proceso estocástico conocido y luego, se aplica a dicha muestra con datos faltantes, algunos métodos de corrección, asumiendo que no se está al tanto de la distribución que sigue la pérdida de datos. Se observa que para tasas de no respuesta bajas y niveles de severidad del sesgo de selección moderados, la mayoría de los procedimientos revisados evidencia un buen desempeño al estimar la media del ingreso e indicadores de desigualdad. Además, se concluye que no existe un mejor método pues cada situación es particular y un análisis exhaustivo de la tasa de no respuesta, las características de los datos y la distribución de la omisión será necesario. Los métodos aquí revisados, si bien no corrigen el problema del sesgo de selección, nos entregan un set de herramientas muy útiles y simples para evaluar la robustez de los indicadores de desigualdad y desigualdad de oportunidades en el mercado del trabajo y otros contextos similares.

1. Introducción

Un desarrollo clave en el pensamiento moderno ha sido la incorporación de la responsabilidad personal en la definición de justicia social. A partir de los trabajos de Rawls (1971) y Sen (1980), los cientistas sociales han comenzado a preguntarse si toda inequidad es injusta y cuál sería la dimensión adecuada en la que se debiera promover la igualdad.

La distinción entre la inequidad debida a esfuerzo personal, la que puede considerarse como éticamente aceptable, y aquella debida a factores exógenos al individuo, la cual podría clasificarse como injusta, genera consenso en la literatura académica, lo que a su vez, re-direcciona la política pública cuya tarea pasa a enfocarse en proveer igualdad de oportunidades, o en otras palabras, un campo de juego nivelado entre quienes poseen circunstancias externas favorables y quienes no. En esta misma línea, la desigualdad producida por diferencias en los niveles de esfuerzo, no se consideraría dañina y hasta podría suponerse como necesaria por los incentivos positivos que genera (Bourguignon et al., 2007b; Marrero, 2009).

Debido al interés académico y social que suscita el concepto desigualdad de oportunidades y a su naturaleza un tanto abstracta, se han desarrollado diversas metodologías que buscan medirlo. Siguiendo las definiciones formales entregadas en Roemer (1993, 1998) y Van de Gaer (1993), trabajos como Bourguignon et al. (2007a), Lefranc et al. (2008), Checchi and Peragine (2010) y Ferreira and Gignoux (2011b) han propuesto técnicas para cuantificarlo en la práctica. Por otra parte, investigaciones de Contreras et al. (2009a,b, 2012), han estimado la desigualdad de oportunidades en el contexto chileno. Si bien, las metodologías descritas en estos trabajos difieren entre sí, todas ellas se basan en datos a nivel de hogar o individuo que contengan dos tipos de variables: las que miden bienestar (típicamente el ingreso, la riqueza, indicadores de logro educacional como notas promedio, entre otros) y las que miden circunstancias, es decir, variables que pueden influenciar a las de bienestar, pero sobre las cuales el individuo no tiene control alguno (por ejemplo, sexo, raza o *background* familiar).

Sin embargo, al intentar medir este tipo de desigualdad, es común encontrar datos perdidos en las variables de interés, y en la mayor parte de las ocasiones, esa omisión de datos no se produce al azar. De esta forma, estaremos en presencia de algún tipo de sesgo de selección. Ferreira and Gignoux (2011a), por ejemplo, indagan acerca de la desigualdad de oportunidades en educación, utilizando como variable de bienestar el puntaje de la prueba PISA. Sin embargo, no todos los individuos asisten a la escuela, por lo que se cuenta con medida de resultado solo para aquéllos que están inscritos en el sistema escolar, dejando fuera, probablemente, a

los niños más desaventajados, sesgando las estimaciones de desigualdad. Lo mismo ocurre, por ejemplo, si se quisiera medir desigualdad de ingresos pues solo se contaría con datos de salario para los sujetos que se encuentran insertos en el mercado laboral, dejando fuera, posiblemente, a los trabajadores menos calificados de la fuerza de trabajo.

Estos problemas de pérdida de datos de manera no aleatoria, que surgen inevitablemente al intentar medir desigualdad, no son simples de corregir. Los modelos de selección y los pattern mixture models se han encargado de esta tarea a través de la modelación de la distribución de la omisión de datos, sin embargo, necesariamente recurren a ciertos supuestos no testeables, para hacerlo (Schafer and Graham, 2002).

Algunos trabajos proponen métodos más simples y directos, que si bien no están diseñados para corregir la omisión de datos no aleatoria, permitirían evaluar la sensibilidad de las medidas de desigualdad a posibles sesgos de selección. Ferreira and Gignoux (2011a), por ejemplo, en su investigación acerca de la desigualdad de oportunidades en educación, utilizan el enfoque de reponderación para analizar si sus indicadores de desigualdad están sujetos a cambios al agregar al análisis a los individuos que se encuentran fuera del sistema escolar. Los mismos autores proponen un procedimiento del tipo colddeck que permite medir desigualdad de oportunidades en países que no disponen de datos de ingresos y de background socioeconómico en una misma encuesta (Ferreira et al., 2011).

En este sentido, el presente trabajo pretende evaluar el desempeño de estos procedimientos más simples -ampliamente utilizados en la literatura para corregir problemas de pérdida de datos- cuando la distribución de la omisión es no aleatoria (*missing not at random*, MNAR). Esto, dentro del contexto de la medición de la desigualdad y la desigualdad de oportunidades en el mercado laboral.

La presente investigación se organiza como sigue. La sección 2 presenta un breve resumen acerca de la caracterización que la literatura académica ha dado a los patrones y distribución de la omisión de datos. La sección 3 expone los principales enfoques y métodos que se han utilizado para corregir la no respuesta. La sección 4 desarrolla un mecanismo de simulación de datos que permite estudiar el desempeño de diversos procedimientos para corregir la pérdida de datos, cuando esta pérdida sigue un proceso no aleatorio. Posteriormente se aplican dichos mecanismos y se evalúan sus resultados. Por último se concluye en la sección 5.

2. Caracterización de la pérdida u omisión de datos

A continuación se analizarán los inconvenientes que puede ocasionar la presencia de datos perdidos u omitidos en las estimaciones, las principales razones detrás de la no respuesta y los diferentes patrones y distribuciones que ésta puede seguir.

2.1. Datos perdidos

En la mayoría de las bases de datos con las que trabajan investigadores y tomadores de decisiones se presenta el problema de datos perdidos, en una o más variables. La tasa de esta pérdida puede variar desde menos del 1 % hasta más del 40 % según la base de datos y la variable en análisis. No obstante, por más pequeña que sea dicha magnitud, puede causar inconvenientes en términos de sesgo e ineficiencia en los parámetros estimados (Schafer and Graham, 2002).

Las causas de la pérdida de datos son diversas. El azar o bien el diseño y/o tipo de estudio pueden tener alguna influencia. Por ejemplo, el cansancio en encuestas muy extensas puede llevar a la no respuesta al final del cuestionario, o bien, en ocasiones, los entrevistados pueden negarse a responder acerca de temas muy sensibles. Otras causas, pueden estar relacionadas a desconocimiento por parte del encuestado de la información que se le solicita, cuestionarios autoaplicados en que el entrevistado podría saltarse preguntas por descuido, entre otras.

Un caso especial de pérdida de datos lo constituyen los estudios longitudinales, en los cuales es probable que se pierdan observaciones de una ola de encuestas a otra debido a que no es posible volver a contactar al entrevistado por fallecimiento, cambio de dirección u otro.

Por último, el tipo de datos omitidos en el que nos enfocaremos en el presente documento ocurre cuando una pregunta no aplica para el entrevistado en cuestión. Por ejemplo, la pregunta "¿Cuál fue su salario del mes pasado?" generará un dato perdido para toda aquella persona que no trabajó durante ese mes, sin importar si pertenece a la fuerza laboral o no, y por lo tanto los análisis que se obtengan a partir de esta variable deben tener en consideración a la población a la que se estará haciendo referencia. Lo mismo sucede si se quisiera trabajar, por ejemplo, con medidas de desempeño escolar: solo se contaría con el promedio de notas de alumnos insertos en el sistema educacional y por lo tanto, si no se realizan ajustes posteriores, las conclusiones de los análisis deben referirse a aquella población.

2.2. Patrón de los datos perdidos

La elección de un mecanismo de corrección por omisión de datos debe considerar el patrón que dicha omisión sigue. Little and Rubin (1987) los clasifica en univariados, multivariados, monótonos, aleatorios y de variable latente. El patrón univariado hace referencia a la omisión concentrada en solo una variable de la base de datos; el patrón multivariado, por su parte, se refiere a la no respuesta para un conjunto de preguntas de un subgrupo de entrevistados; el patrón de omisión de datos del tipo monótono ocurre particularmente en estudios longitudinales, cuando en cada ronda de la encuesta se van perdiendo observaciones (atrición); el tipo aleatorio consiste en pérdidas dispuestas en cualquier variable e individuo, sin seguir un patrón determinado; y por último, el patrón del tipo variable latente se refiere a la pérdida subjetiva que ocurre al no contar con datos no observables (muchas veces por tratarse de variables que son difícilmente cuantificables, como habilidad, inteligencia, etc.).

2.3. Distribución de los datos perdidos

La pérdida o ausencia de datos es entendida como un proceso estocástico y por lo tanto, se le puede categorizar según su función de probabilidad. La clasificación más utilizada en la literatura académica se origina en el trabajo de Rubin (1976) y agrupa a los procesos en completamente aleatorios (*missing completely at random*, MCAR), aleatorios (*missing at random*, MAR) y no aleatorios (*missing not at random*, MNAR).

2.3.1. Procesos Completamente Aleatorios (MCAR)

Se dice que un proceso es MCAR cuando la distribución de los valores omitidos no depende de los valores de ninguna variable.

Suponer que la distribución de la omisión de datos es MCAR es equivalente a afirmar que la pérdida no se encuentra relacionada con ningún factor, ni observable ni tampoco no observable.

Si se asume este tipo de distribución para todas las variables en una base de datos, no existe la necesidad de corregir por omisión, puesto que la muestra observada puede ser considerada como una submuestra aleatoria simple de la base de datos original (Allison, 1999).

2.3.2. Procesos Aleatorios (MAR)

Un proceso de datos omitidos se considera MAR si la distribución de la ausencia de datos en una variable Y no es aleatoria, sino que depende de otra variable X, pero para cada valor

de X, los valores observados de Y sí representan una muestra aleatoria de Y. Así, por ejemplo, si X corresponde al género del encuestado e Y a su renta, se estaría en presencia de un proceso MAR si existen más valores ausentes de Y en hombres que en mujeres y, sin embargo, tanto dentro del grupo de los hombres como dentro del de las mujeres, el patrón de datos omitidos es completamente aleatorio. En otras palabras, sostener que la distribución de datos es MAR es equivalente a aseverar que la distribución de la omisión depende solo de variables que podemos observar en la base de datos. Como se puede concluir de las definiciones anteriores MCAR es un caso particular de procesos MAR.

2.3.3. Procesos No Aleatorios (MNAR)

Cuando el proceso de datos faltantes no es MCAR ni MAR, se clasifica como MNAR. En este tipo de procesos, la probabilidad de que un dato esté omitido en la variable Y depende del valor de la propia variable omitida Y, aún después de controlar por otras variables.

Esta es la distribución de omisión de datos con la que probablemente nos encontraríamos al buscar medir desigualdad de ingresos, pues solo contaríamos con datos de salario para los sujetos que se encuentran insertos en el mercado laboral, dejando fuera, posiblemente, a los trabajadores menos calificados, es decir, los que obtendrían un menor salario si estuviesen trabajando.

Estos problemas de pérdida de datos de manera no aleatoria, que surgen inevitablemente al intentar medir desigualdad, no son simples de corregir. Los modelos de selección y los pattern mixture models se han encargado de esta tarea a través de la modelación de la distribución de la omisión de datos, sin embargo, necesariamente recurren a ciertos supuestos, muchas veces no testeables, para hacerlo (Schafer and Graham, 2002).

2.3.4. Plausibilidad de Procesos MAR

En general, para cualquier tipo de problema de datos omitidos, es difícil asumir una distribución MAR, pues aún si contáramos con una serie de variables que nos permitieran controlar por posibles razones de la omisión de datos, nunca se podría asegurar que se están utilizando todas las variables que efectivamente participan de la pérdida. Quizá se estén dejando algunas de lado o se estén agregando otras sin importancia. En otras palabras, cuando la omisión de los datos va más allá del control del investigador, la distribución de ésta será desconocida y que sea MAR será solo un supuesto impuesto por el investigador. En general, no hay forma de aseverar si en un determinado set de datos se cumple la condición MAR, excepto obteniendo respuestas

posteriores de los individuos que no contestan (Glynn et al., 1993; Graham and Donaldson, 1993; Little and Rubin, 1987).

Así, en la mayor parte de los casos se debe tener presente que, probablemente, se esté trabajando con datos MNAR. Y Si bien, es posible formular y estimar modelos para este tipo de omisión de datos, dichos modelos son complejos y no-testeables (Allison, 2001).

3. Principales enfoques para el manejo de la pérdida de datos

Algunos trabajos proponen métodos más simples y directos, que si bien no están diseñados para corregir la omisión de datos no aleatoria, permitirían evaluar la sensibilidad de las medidas de desigualdad a posibles sesgos de selección. Estos se basan en enfoques ampliamente utilizados en la literatura para corregir problemas de pérdida de datos. Para analizar sus objetivos, ventajas y desventajas, es de utilidad clasificarlos en 3 grandes categorías que serán descritas a continuación (Little and Rubin, 2002).

3.1. Enfoque de datos completos

Bajo esta manera de proceder, se encuentran dos métodos denominados listwise y pairwise.

El **listwise** excluye todas las observaciones que poseen al menos un dato perdido en cualquiera de las variables que serán utilizadas, es decir, trabaja solo con las observaciones que disponen de información completa para todas las variables en análisis. Es el método utilizado por defecto en la mayoría de los paquetes estadísticos y supone que los datos omitidos siguen un patrón MCAR.

Su principal ventaja es ser un mecanismo simple y fácil de ejecutar. No genera sesgo e ineficiencia en los parámetros estimados cuando el tamaño muestral es grande y el patrón de datos perdidos es MCAR. Sin embargo, estos supuestos son difíciles de cumplir en la práctica y más aún en el contexto del cálculo de desigualdad en el mercado laboral.

Cuando la proporción de los datos perdidos es grande relativa al tamaño de muestra y el patrón de omisión es MAR o MNAR, este procedimiento impactará, primeramente disminuyendo el tamaño muestral, y luego haciendo que las estimaciones derivadas de dicha muestra sean sesgadas e ineficientes (Kalton and Kasprzyk, 1982).

El **pairwise**, por su parte, utiliza todos los pares de observaciones que no poseen valores perdidos en las variables usadas para un indicador en particular y se utiliza frecuentemente en estimaciones de matrices de varianzas y covarianzas. Por ejemplo, si suponemos una base de datos como la de la tabla 1, y se quisiera medir la correlación entre salario y escolaridad, se haría uso de las observaciones 1, 4 y 6, mientras que para medir la correlación entre salario y

ocupación se utilizarían los datos del individuo 1, 3, 4 y 6.

Tabla 1: Patrones de datos omitidos

Observación	Sexo	Edad	Escolaridad	Ingresos	Ocupación
1	0	41	16	45000	3
2	1	36	15	·	1
3	0	64	·	12000	1
4	1	22	12	22000	4
5	1	26	·	·	3
6	0	37	15	18000	2

Fuente: elaboración propia

Este método recupera algo de la información que se pierde completamente al recurrir al método listwise, y además, bajo el supuesto de que la distribución de la omisión es MCAR, los estimadores de covarianza y correlación serán consistentes. Sin embargo, debido a la diferencia en tamaño muestral que implica este procedimiento, no todos los análisis serán completamente comparables entre sí, lo que en la práctica podría dificultar el obtener conclusiones acerca de los resultados.

3.2. Enfoque de reponderación

Esta técnica pondera cada observación con información completa (sin dato omitido en ninguna de las variables de interés) para considerar a las observaciones con algún dato perdido. Es un mecanismo similar al de post-estratificación, con la diferencia de que para ponderar las observaciones se utiliza información obtenida desde dentro de la misma muestra, en tanto que en la post-estratificación se recurre a datos exógenos provenientes de otras encuestas, censos o registros administrativos (Galván and Medina, 2007).

Existen diversas formas para calcular el ponderador de cada observación, aunque la mayoría de las técnicas divide a la muestra en estratos, en base a variables observables, y luego calcula un ponderador para cada estrato, considerando que la distribución final de la muestra ponderada se asemeje a las características observables de la muestra completa.

Esta técnica permite corregir problemas de sesgo en los parámetros cuando el patrón de omisión de datos es MAR, asumiendo que los estratos están correctamente definidos.

3.3. Enfoque basado en la imputación

Bajo este enfoque se reemplaza cada valor ausente de la base de datos por otro, siguiendo distintos métodos. Cuando el procedimiento se aplica una sola vez, es decir, se genera un solo valor imputado para cada dato perdido, se le conoce como imputación simple. Por el contrario, si un valor omitido es imputado dos veces o más, permitiendo calcular un error debido a la imputación, entonces se le llama múltiple. A su vez, si la imputación se realiza en base a un modelo estadístico expresamente definido se le conoce como enfoque explícito, mientras que si la imputación se realiza en base a algoritmos que siguen tácitamente un modelo, se dice que se sigue un enfoque implícito.

Entre los enfoques explícitos se encuentran: el **método de imputación de la media incondicional**, que reemplaza el valor omitido por la media de la variable en la muestra observada; el **método de imputación de la media condicional**, en donde se definen estratos y cada valor perdido es sustituido por la media del estrato correspondiente, el método de **imputación por regresión**, que reemplaza el valor omitido por el predicho utilizando información de variables observadas; el **método de imputación por regresión estocástica**, que es similar al anterior pero además de sustituir el valor perdido por el predicho, se agrega un residuo para dar cuenta de incertidumbre en dicho valor; por último el **método de imputación propensity score**, que ordena a los individuos según su probabilidad de no respuesta y reemplaza el valor omitido con el de alguno de los entrevistados con similar probabilidad de no respuesta.

Por otro lado, entre los enfoques implícitos se puede encontrar: el **método de imputación hotdeck**, que reemplaza el valor perdido por uno de alguna otra observación que posea características similares en ciertas variables y generalmente, se aplica dividiendo a la muestra en estratos, desde donde a cada observación sin datos se le asigna aleatoriamente un valor a imputar, dentro de cada estrato; el **método de imputación colddeck**, que utiliza fuentes externas, como censos u otras encuestas, para asignar el valor en las celdas sin dato; y por último el **método de imputación compuesto** que combina dos o más métodos.

4. Análisis del desempeño de métodos de corrección por omisión con datos MNAR

En general, los métodos descritos en la sección anterior aunque poseen la ventaja de ser fáciles de ejecutar, se ocupan de corregir la omisión de datos del tipo MAR. Por otro lado, en la literatura académica que busca medir desigualdad se trabaja, la mayor parte del tiempo, con variables del tipo MNAR.

El presente trabajo pretende estudiar el desempeño de procedimientos simples del tipo MAR -ampliamente utilizados en la literatura para corregir problemas de pérdida de datos- cuando la distribución de la omisión es no aleatoria (*missing not at random*, MNAR). Esto, dentro del contexto de la medición de la desigualdad y la desigualdad de oportunidades en el mercado laboral.

Para ello, se seguirán los siguientes pasos. Primero, desde una muestra de datos ficticia se crearán observaciones perdidas en la variable de ingresos, siguiendo un proceso conocido y luego, se aplicarán a dicha muestra con datos omitidos, algunos de los métodos de corrección revisados en la sección 3, asumiendo que no se está al tanto de la distribución que sigue la pérdida de datos. Finalmente, se evaluarán y compararán los distintos procedimientos, observando su impacto en la media del ingreso e indicadores de desigualdad y desigualdad de oportunidades.

4.1. Experimento para la creación de datos omitidos

En la práctica, un investigador que trabaja con datos probablemente encontrará observaciones perdidas en las variables a analizar. Sin embargo, al no poseer el vector de datos completos, le será imposible evaluar de forma certera la calidad del procedimiento por el cual corrigió la pérdida de dichos datos. El experimento que pasaremos a describir a continuación, tiene por objetivo crear una pérdida de datos artificial, desde un vector de datos completos, para luego proceder a comparar y evaluar los resultados de varios tipos de métodos de corrección por omisión de datos.

El primer paso consiste en generar un vector de ingresos, Y, con ciertas características que permitan reflejar la realidad del contexto del mercado laboral. Primero, es deseable que variables observables jueguen un rol importante en su predicción y segundo, debe estar compuesto

por una parte atribuible a la "suerte".

Siguiendo lo expuesto en Contreras et al. (2012), se considerarán las variables edad, escolaridad, escolaridad del padre, escolaridad de la madre, ocupación del padre (dummy igual a "1" si el padre fue trabajador de cuello blanco, "0" en otro caso), número de hermanos y región donde vivió la niñez (dummy para norte y sur), para generar el dato de ingreso mensual del individuo i[1]:

$$\begin{aligned} ln(Y_i) &= \beta_0 + \beta_1 edad_i + \beta_2 esc_i + \beta_3 esc\_p_i + \beta_4 esc\_m_i + \beta_5 ocup\_p_i + \beta_6 hnos_i \\ &\quad + \beta_7 norte_i + \beta_8 sur_i + \varepsilon_{1,i} \end{aligned} \tag{1}$$

donde $\beta_0 = 9{,}8$, $\beta_1 = 0{,}01$, $\beta_2 = 0{,}15$, $\beta_3 = 0{,}03$, $\beta_4 = 0{,}01$, $\beta_5 = 0{,}1$, $\beta_6 = -0{,}001$, $\beta_7 = 0{,}1$, $\beta_8 = -0{,}08$ y $\varepsilon_{1,i} \sim N(0, 0{,}6)$.

Así, la variable generada a través del proceso 1, Y, corresponderá el vector de ingresos mensuales con datos completos que nos permitirá evaluar las propiedades de los distintos métodos. Este ingreso fue generado asumiendo que son una cierta cantidad de variables y un shock aleatorio los factores que afectan el ingreso de cada individuo. La escolaridad de éste, junto con la de su padre, se asume que son las variables que afectan en mayor magnitud el salario del entrevistado[2].

A continuación, crearemos artificialmente, una pérdida de datos en el vector de ingresos, Y, que partirá del 5 % hasta el 35 % de la muestra. Este paso se realiza del siguiente modo. Se genera una nueva variable de ingresos a la que llamaremos *ingreso de corte* (Yc), a partir del mismo proceso generador de datos de la ecuación 1, pero con otro error, ε_2 , distribuido de igual manera que ε_1:

$$\begin{aligned} ln(Yc_i) &= \beta_0 + \beta_1 edad_i + \beta_2 esc_i + \beta_3 esc\_p_i + \beta_4 esc\_m_i + \beta_5 ocup\_p_i + \beta_6 hnos_i \\ &\quad + \beta_7 norte_i + \beta_8 sur_i + \varepsilon_{2,i} \end{aligned} \tag{2}$$

donde $\varepsilon_{2,i} \sim N(0, 0{,}6)$, $i = 1; \ldots; 3,165$.

[1] Estos datos se obtuvieron al seleccionar aleatoriamente, un subgrupo de 3,165 entrevistados hombres de la Encuesta de Protección Social 2009 (EPS 2009).

[2] El valor de los coeficientes y del error estándar, se asumieron similares a los calculados a través de una regresión OLS entre el ingreso y las características que los individuos seleccionados reportaron en la EPS 2009.

La variable Yc, entonces, solo se utilizará con el objetivo de decidir qué observaciones de la variable Y quedarán sin dato. Para esto, ordenamos a los individuos de la muestra en base a la variable Yc, y dejamos sin dato de ingreso mensual (Y) al 5, 15, 25 y 35 % de la muestra con los valores más bajos de la variable Yc.

En otras palabras, la ecuación 2 representa el proceso generador de la omisión de los datos o la *regla* que sigue dicha omisión. En este caso, se trabajó bajo el supuesto de que la regla depende de las mismas covariables que el proceso que genera los datos de ingresos, y que la "suerte" se distribuye de igual manera en ambos procesos.

De esta forma, contaremos con 5 vectores de ingresos: uno con datos completos (que nos servirá para evaluar los métodos de corrección por omisión de datos) y 4 con datos omitidos (ver anexo B). La tabla 2 presenta la distribución de los datos perdidos creados de forma artificial de acuerdo a la regla descrita en la ecuación 2, según decil de ingresos. Mientras más bajo sea el ingreso del individuo, mayor es la probabilidad de que el dato se encuentre omitido, tal como fue modelado en 1 y 2. Esto pretende simular el sesgo de selección presente en el mercado del trabajo, que se refiere al hecho de que solo se cuenta con datos de salario para los sujetos que se encuentran insertos en el mercado laboral, dejando fuera, posiblemente, a los trabajadores menos calificados.

Tabla 2: Distribución de la omisión de datos según tasa de no respuesta

Decil de Ingresos	Omisión			
	5 %	15 %	25 %	35 %
1	25.6 %	51.1 %	68.1 %	78.9 %
2	9.2 %	31.3 %	47.2 %	62.0 %
3	5.4 %	20.8 %	36.6 %	53.9 %
4	5.1 %	16.1 %	31.6 %	42.7 %
5	1.9 %	9.8 %	18.3 %	31.9 %
6	1.6 %	8.5 %	18.4 %	28.2 %
7	0.3 %	4.4 %	11.0 %	19.9 %
8	0.9 %	5.1 %	11.4 %	17.7 %
9	0.3 %	2.2 %	5.7 %	10.1 %
10	0.0 %	0.6 %	1.9 %	4.7 %

En la tabla 3, se muestra que el promedio del ingreso va en aumento mientras se acrecienta

la cantidad de datos omitidos, al igual que la desviación estándar (aunque va disminuyendo si la pensamos como proporción de la media). Por último, las medidas de desigualdad presentadas en la tabla 4 revelan que ésta disminuye, mientras aumenta la pérdida de datos.

Tabla 3: Estadística descriptiva del ingreso según porcentaje de omisión

Omisión en Y	Media	Desv. Est.	Mínimo	Máximo
0 %	233,623	243,441	6,898	2,872,599
5 %	242,468	246,233	8,014	2,872,599
15 %	258,939	253,568	8,014	2,872,599
25 %	276,355	262,476	8,014	2,872,599
35 %	295,135	271,406	15,535	2,872,599

Tabla 4: Desigualdad de ingresos según porcentaje de omisión

Omisión en Y	Gini	MLD GE(0)	Theil GE(1)
0 %	0.47	0.41	0.39
5 %	0.46	0.38	0.37
15 %	0.45	0.36	0.35
25 %	0.44	0.34	0.33
35 %	0.43	0.32	0.32

De esta manera, el patrón de omisión que se ha generado es univariado (concentrado exclusivamente en la variable de ingresos) y dependiente de la variable Yc, la que a su vez depende de las covariables escolaridad, escolaridad del padre y la madre, ocupación del padre, región de residencia en la niñez y número de hermanos. Según esto, la distribución de la omisión de datos podría ser considerada MAR, pues si se corrigiese por cada una de las covariables anteriormente descritas, entonces no habría diferencias entre las observaciones de Y con y sin dato observado (ver definición 2.3.2). Sin embargo, apoyándonos en un supuesto más realista, en el presente trabajo se asumirá que el proceso que genera la pérdida de los datos no es conocido, y por lo tanto no incluiremos el set completo de variables que forman parte del proceso generador de la pérdida de datos al aplicar los métodos de corrección por datos omitidos.

Así, asumiremos que el investigador no conoce ni el proceso ni tampoco todas las variables que explican la pérdida de datos, pero supone que son la escolaridad del individuo y la de su

padre las covariables que determinan dicha omisión, y por lo tanto, corregirá la no respuesta bajo este supuesto. Si este es el caso, entonces, la probabilidad de que un dato esté omitido en la variable Y quedará dependiendo del valor de la propia variable con datos omitidos Y, dado que no se controló por todas las variables que correspondían, lo que implicará un patrón de pérdida de datos MNAR. Esto es justamente lo que sucede en la práctica: cuando la omisión de los datos va más allá del control del investigador, la distribución de ésta será desconocida y asumir un criterio MAR será solo un supuesto (Schafer and Graham, 2002).

No obstante, aunque no se esté seguro de trabajar con datos MAR o MNAR, recientes investigaciones han demostrado que en casos realistas, el asumir un criterio MAR, puede mejorar las estimaciones logrando solo un impacto menor sobre parámetros estimados y los errores estándar, haciendo uso de variables que están correlacionadas con la variable que genera la pérdida de los datos (Collins et al., 2001; Schafer and Graham, 2002).

4.2. Aplicación de métodos de corrección por omisión de datos

El siguiente paso será corregir los datos omitidos que se acaban de generar asumiendo que no se conoce el proceso de omisión de los datos. Se aplicarán 6 de los métodos revisados en la sección 3: listwise, medias condicionales, reponderación, hotdeck, regresión lineal estocástica y propensity score estocástico.

Siguiendo a Efron (1994) y utilizando la técnica bootstrap, se obtienen 2,001 muestras independientes a partir de la muestra original con datos omitidos, y a cada una se le aplican los 6 métodos de imputación. Finalmente, se calculan los parámetros de interés (ingreso promedio e indicadores de desigualdad) y sus respectivos intervalos de confianza[3].

4.2.1. Listwise

Esta es la técnica más simple. Se trabajó utilizando solo las observaciones que poseían dato de ingreso.

4.2.2. Medias condicionales

Para imputar la media condicional, primeramente, se debieron definir los estratos. Éstos se eligieron realizando una estimación MCO con los datos disponibles, lo cual reveló que los mejores predictores del ingreso eran la escolaridad del individuo y la de su padre. Con esta

[3]Se utilizó el método de percentiles de Efron con 2,001 simulaciones para el cálculo de los intervalos de confianza.

información se categorizaron dichas variables en 4 grupos: “sin escolaridad”, “de 1 a 8 años”, “de 9 a 12 años” y “más de 12 años de estudios”. Como cada variable se constituye de cuatro ítems, al cruzar la variable escolaridad del individuo (4 grupos) con escolaridad del padre (4 grupos más), se contabilizan finalmente 16 estratos.

De esta manera, es posible calcular el ingreso promedio dentro de cada estrato (tomando en consideración solamente a los individuos con dato) e imputárselo a los demás sujetos que están en dicho estrato, pero que no tienen dato de ingreso.

Si bien este método es fácil de aplicar, se deben tener algunas consideraciones. No será posible elegir una cantidad muy grande de categorías para formar los estratos pues, cada estrato debe tener al menos una observación con dato para poder imputárselo a las demás. Igualmente, si las categorías definidas son muy amplias, se terminará con muchas observaciones con igual dato lo que afectará directamente a la varianza del ingreso y por consiguiente a los indicadores de desigualdad.

4.2.3. Reponderación

Esta técnica también requiere de la definición de estratos. Éstos fueron determinados de igual forma que en el método anterior. La ponderación que se da a cada estrato se realizó de manera tal que la distribución por estratos en la muestra con dato omitido lograra igualarse a la distribución por estratos de la muestra sin datos perdidos. Esto se consigue, ponderando cada observación por el siguiente factor:

$$Ponderador_e = \frac{n\_total_e}{n\_con\_dato_e}, \quad e = 1, 2, \ldots, 16. \tag{3}$$

donde e indica el estrato, n_total_e el número total de observaciones en el estrato e (sin importar si tiene dato de ingreso o no) y $n\_con\_dato_e$ corresponde al número de observaciones en el estrato e que posee dato de ingreso.

Este método presenta el mismo tipo de dificultad que el de medias condicionadas con respecto a la elección de los estratos. Además, dada la forma de la ponderación, el promedio de ingresos por estrato no variará. Sin embargo, al no imputar un mismo valor para todos los datos omitidos dentro de un estrato, dará lugar a una mayor varianza en el ingreso con respecto al procedimiento anterior.

4.2.4. Hotdeck

Para la utilización de este método se definieron los mismos estratos mencionados anteriormente. Esta técnica reemplaza los datos perdidos con algún dato de ingreso aleatoriamente determinado desde dentro de cada estrato. Todas las observaciones con dato de ingreso, tienen la misma probabilidad de ser elegidas para imputar ese valor a alguna observación sin dato dentro de cada estrato.

Se tienen las mismas dificultades que los dos métodos anteriores con respecto a la determinación de los estratos, pero al igual que el método de ponderación, la variable de ingresos imputada presentará una mayor varianza que en el método de medias, pues no se imputa un mismo valor a todos los individuos con dato faltante dentro del estrato.

4.2.5. Regresión Lineal Estocástica

Este procedimiento imputa los valores omitidos basándose en una regresión lineal entre el ingreso y variables explicativas. Sea Y_{obs} la parte de la variable ingreso con dato observado, e Y_{miss} el componente con dato omitido. Sea X_{obs} los predictores para los n_{obs} individuos con dato de ingreso observado, y X_{miss} el complemento de los n_{miss} casos sin dato de ingreso. Sea r el número de predictores utilizados (escolaridad y escolaridad del padre en este caso particular). El algoritmo que se utiliza para este procedimiento sigue a Van Buuren et al. (1999) y consiste en los siguientes pasos:

a) Calcular $W = (X'_{obs}X_{obs})^{-1}$, $\hat{\beta} = WX'_{obs}Y_{obs}$ y $\hat{Y}_{obs} = X_{obs}\hat{\beta}$.

b) Obtener una variable aleatoria g desde una distribución ji-cuadrado con $n_{obs} - r$ grados de libertad.

c) Calcular $\sigma^2_* = (Y_{obs} - \hat{Y}_{obs})(Y_{obs} - \hat{Y}_{obs})/g$.

d) Obtener un vector de dimensión r desde una distribución normal $D \sim N(0, I_r)$, donde I_r es la matriz identidad de orden r.

e) Calcular $\hat{\beta}_* = \hat{\beta} + \sigma_* W^{1/2}$ es la raíz cuadrada triangular de W obtenida a través de la descomposición de Cholesky.

f) Calcular los valores predichos $\hat{Y}_{miss} = X_{miss}\hat{\beta}_*$.

g) 7) Para cada valor perdido $i = 1, \ldots, n_{miss}$ encontrar la observación cuyo $\hat{Y}_{obs}$ es el más cercano a $\hat{Y}_{miss,i}$ y tomar Y_{obs} de este individuo como el valor a imputar en la observación i.

El paso 1 obtiene $\hat{\beta}$ e $\hat{Y}_{obs}$ de los datos observados a través de una regresión lineal. Los pasos 2 a 5 intentan generar coeficientes aleatorios, pero que provengan desde la misma distribución de β. Finalmente, la idea de los pasos 6 y 7 es tomar prestados para la imputación los datos de individuos muy similares, pero que poseen dato de ingreso observado.

4.2.6. Propensity Score Estocástico

Esta técnica utiliza una estimación de la probabilidad de que un dato en particular se encuentre omitido. El supuesto detrás es que la no respuesta puede ser explicada a través de un set de covariables. En este caso se utilizó un modelo de regresión probit donde la variable dependiente es igual a "1" si el dato se encuentra en la base de datos y "0" si está perdido. Las covariables utilizadas fueron la escolaridad y escolaridad del padre.

Una vez estimada la probabilidad de que la variable se encuentre omitida, se ordenaron los datos según este propensity score y se procedió a dividir la muestra en deciles. Dentro de cada decil, se imputó al azar el ingreso de alguno de los sujetos con dato a aquellos que no lo tenían.

4.3. Resultados

A continuación se presentarán los resultados para la estimación del ingreso y medidas de desigualdad en términos de la media y su intervalo de confianza, el sesgo promedio y la tasa de cobertura para cada método y nivel de omisión. La tasa de cobertura (llamada también tasa de cobertura real para enfatizar la diferencia con la nominal) corresponde a la proporción de veces que el parámetro estimado está contenido en el verdadero intervalo de confianza (aquel calculado con la muestra sin datos omitidos), mientras que la tasa de cobertura nominal está dada por el nivel de confianza del intervalo (95 % en nuestro caso). Si el método de corrección por omisión de datos es eficaz en estimar el valor verdadero del dato perdido, entonces el valor nominal de la tasa de cobertura debería ser muy similar al real.

4.3.1. Ingreso Promedio

Los resultados para la estimación del ingreso promedio se muestran en la tabla 5. Esta indica que para cualquier nivel de omisión de los datos, es el método listwise el que más sesgo posee. Los métodos de reponderación y media condicional, como es de esperarse dado que seleccionamos estratos idénticos, reportan igual ingreso promedio, seguido muy de cerca por el método hotdeck. Los métodos de regresión lineal y propensity score estocástico son los que reportan el menor sesgo. El sesgo promedio de todos los métodos aumenta con el aumento del

porcentaje de omisión en la muestra.

Cuando el porcentaje de omisión de los datos alcanza el 5 % de la muestra, la tasa de cobertura real (de los intervalos de confianza al 95 %) del método de regresión lineal y propensity score estocástico corresponden al valor nominal de cobertura. Los métodos de media condicional, reponderación y propensity score, por su parte, se encuentran cercanos a este valor.

El método de propensity score estocástico, alcanza tasas de cobertura reales cercanas a las nominales hasta que la omisión de datos se encuentra en un 15 %. Por su parte, el método de regresión lineal, hasta una omisión de datos de 25 %. Los métodos de media condicional, reponderación y hotdeck, alcanzan coberturas reales de alrededor de un 88 % para una omisión del 15 %, de 70 % para una omisión del 25 %, y del 50 % para una omisión del 35 %.

Por otro lado, se aprecia que si no se corrigiese la falta de datos (listwise), los niveles de cobertura no alcanzarían siquiera el 50 % para una omisión de datos del 5 %. Para niveles mayores de omisión, no hay estimaciones del ingreso promedio bajo este método que se encuentren contenidos en el intervalo de confianza real.

Tabla 5: Resultados de la corrección por datos omitidos para el ingreso promedio

Procedimiento	Media	95 % CI		Sesgo	Tasa de
		lím. inf.	lím. sup.	Promedio	Cobertura***
0 % omisión					
Muestra Completa	**233,665**	**225,415**	**242,138**	–	–
5 % omisión					
Listwise	242,520	233,901	251,271	8,855	46.8 %
Media Condicional	235,361	227,152	243,876	1,696	93.2 %
Reponderación	235,361	227,152	243,876	1,696	93.2 %
Hotdeck	235,359	227,050	243,803	1,694	93.0 %
Regresión Lineal*	233,705	225,415	242,330	40	94.6 %
Regresión Lineal**	233,753	225,491	242,286	88	94.8 %
Propensity Score*	233,967	225,766	242,566	302	94.9 %
Propensity Score**	233,994	225,903	242,603	329	94.7 %
15 % omisión					
Listwise	258,938	249,657	268,857	25,273	0.0 %
Media Condicional	237,142	228,732	245,829	3,477	87.8 %
Reponderación	237,142	228,732	245,829	3,477	87.8 %
Hotdeck	237,137	228,620	246,030	3,472	87.4 %
Regresión Lineal*	233,770	225,380	242,872	105	94.3 %
Regresión Lineal**	233,878	225,079	242,873	213	93.6 %
Propensity Score*	235,241	226,689	244,112	1,576	93.4 %
Propensity Score**	235,727	227,354	244,580	2,062	92.0 %
25 % omisión					
Listwise	276,337	266,056	286,972	42,671	0.0 %
Media Condicional	240,024	231,392	248,820	6,359	69.1 %
Reponderación	240,024	231,392	248,820	6,359	69.1 %
Hotdeck	239,990	231,312	248,872	6,325	69.2 %
Regresión Lineal*	235,083	226,251	244,281	1,418	92.2 %
Regresión Lineal**	235,063	225,944	244,616	1,398	90.6 %
Propensity Score*	237,029	228,396	246,022	3,364	86.6 %
Propensity Score**	237,914	229,518	247,026	4,249	82.2 %
35 % omisión					
Listwise	295,144	283,534	307,208	61,479	0.0 %
Media Condicional	241,859	233,039	251,049	8,194	52.2 %
Reponderación	241,859	233,039	251,049	8,194	52.2 %
Hotdeck	241,833	232,525	251,453	8,168	51.7 %
Regresión Lineal*	235,265	225,485	245,027	1,600	89.2 %
Regresión Lineal**	234,218	223,837	244,626	553	88.1 %
Propensity Score*	240,165	231,166	250,182	6,500	66.7 %
Propensity Score**	240,855	231,627	250,422	7,190	60.9 %

*Método considerando las variables escolaridad, escolaridad del padre y edad.

**Método considerando las variables escolaridad y escolaridad del padre. Mismas utilizadas para formar estratos en los demás métodos.

***Proporción de veces que el parámetro estimado está contenido en el verdadero intervalo de confianza.

4.3.2. Medidas de Desigualdad

La tabla 5 expone los resultados de la aplicación de los distintos procedimientos para el cálculo del coeficiente de Theil. Tal como en la estimación del ingreso promedio, los métodos de regresión lineal y propensity score, son los que mejor desempeño evidencian. En este indicador, a diferencia de los resultados anteriores, el método de media condicional baja su desempeño en comparación a los procedimientos de reponderación y hotdeck. Esto porque el coeficiente de Theil (y en general cualquier medida de desigualdad) dependerá no solo del promedio de los ingresos, sino también de su distribución y como se mencionó anteriormente, el método de medias condicionales tiende a disminuir la desviación estándar de la variable al imputar los mismos valores a todos los individuos sin dato dentro un mismo estrato.

El método de regresión lineal presenta una tasa de cobertura real similar a la nominal, hasta que la omisión en el vector de ingresos alcanza el 15 %. El desempeño del procedimiento propensiy score parece dar buenos resultados hasta una tasa de omisión del 5 %. Los demás métodos no alcanzan niveles aceptables de cobertura, sin embargo, los métodos hotdeck y reponderación se muestran muy superiores al de medias condicionales para todo nivel de omisión en los datos.

Para las estimaciones del nivel de desigualdad de oportunidades (ver tabla 6) todos los métodos, presentan un nivel de cobertura real cercano al nominal cuando la tasa de no respuesta es de un 5 %. Para tasas mayores de omisión, todos los métodos parecen mostrar un desempeño aceptable a excepción del listwise. Esto se debe a que la forma de estimación del nivel de desigualdad de oportunidades utiliza todas las covariables que afectan la omisión de datos (ver anexo A) y a que la estimación del ingreso es, en general, aceptable para todos los métodos, a excepción del listwise.

La tabla 7, presenta los resultados de la proporción de desigualdad que se debe a desigualdad de oportunidades, esta se define como el cociente entre el nivel de desigualdad de oportunidades y la desigualdad total (ver anexo A). Los métodos propensity score y regresión lineal logran niveles aceptables de desempeño para todos los niveles de omisión, pues consiguen estimadores adecuados para la desigualdad y la desigualdad de oportunidades. Como los indicadores hotdeck y reponderación tienden a subestimar los niveles de la desigualdad total y de

la desigualdad de oportunidades, los indicadores de la relación entre estos se encuentran muy cercanos a los verdaderos valores. Por su parte, como el método de medias condicionales hace un buen trabajo estimando la desigualdad de oportunidades pero subestima de gran manera la desigualdad total, tiende a sobreestimar la proporción de la desigualdad de oportunidades. Finalmente, el método listwise subestima la desigualdad total y la desigualdad de oportunidades, resultando la relación entre ambas en una estimación aceptable para hasta un 15 % de omisión de datos y mostrando un mejor desempeño en términos de cobertura que el método de medias condicionales para cualquier tasa de no respuesta.

Tabla 6: Resultados de la corrección por datos omitidos para medidas de desigualdad Coeficiente de Theil.

Procedimiento	Media	95 % CI		Sesgo	Tasa de
		lím. inf.	lím. sup.	Promedio	Cobertura***
0 % omisión					
Muestra Completa	**0.39**	**0.37**	**0.41**	–	–
5 % omisión					
Listwise	0.37	0.35	0.39	-0.02	62.6 %
Media Condicional	0.37	0.35	0.40	-0.01	72.5 %
Reponderación	0.38	0.36	0.40	-0.01	84.9 %
Hotdeck	0.38	0.36	0.40	-0.01	84.4 %
Regresión Lineal*	0.39	0.36	0.41	0.00	94.2 %
Regresión Lineal**	0.39	0.36	0.41	0.00	94.3 %
Propensity Score*	0.39	0.36	0.41	0.00	93.1 %
Propensity Score**	0.38	0.36	0.41	0.00	93.2 %
15 % omisión					
Listwise	0.35	0.33	0.37	-0.04	7.6 %
Media Condicional	0.35	0.33	0.38	-0.03	19.3 %
Reponderación	0.37	0.35	0.39	-0.02	61.6 %
Hotdeck	0.37	0.35	0.39	-0.02	61.6 %
Regresión Lineal*	0.39	0.36	0.41	0.00	93.7 %
Regresión Lineal**	0.39	0.36	0.41	0.00	91.7 %
Propensity Score*	0.38	0.35	0.40	-0.01	83.1 %
Propensity Score**	0.38	0.35	0.40	-0.01	78.8 %
25 % omisión					
Listwise	0.33	0.31	0.36	-0.06	0.4 %
Media Condicional	0.34	0.31	0.36	-0.05	1.4 %
Reponderación	0.36	0.34	0.39	-0.03	40.0 %
Hotdeck	0.36	0.34	0.39	-0.03	39.8 %
Regresión Lineal*	0.39	0.36	0.41	0.00	90.9 %
Regresión Lineal**	0.38	0.36	0.41	0.00	85.5 %
Propensity Score*	0.37	0.35	0.40	-0.01	74.1 %
Propensity Score**	0.37	0.35	0.40	-0.02	64.0 %
35 % omisión					
Listwise	0.32	0.29	0.34	-0.07	0.0 %
Media Condicional	0.31	0.29	0.34	-0.07	0.0 %
Reponderación	0.36	0.33	0.38	-0.03	22.0 %
Hotdeck	0.36	0.33	0.39	-0.03	23.9 %
Regresión Lineal*	0.38	0.35	0.42	0.00	84.1 %
Regresión Lineal**	0.38	0.35	0.42	0.00	81.2 %
Propensity Score*	0.36	0.33	0.39	-0.03	32.4 %
Propensity Score**	0.36	0.33	0.39	-0.03	28.7 %

*Método considerando las variables escolaridad, escolaridad del padre y edad.

**Método considerando las variables escolaridad y escolaridad del padre. Mismas utilizadas para formar estratos en los demás métodos.

***Proporción de veces que el parámetro estimado está contenido en el verdadero intervalo de confianza.

Tabla 7: Resultados de la corrección por datos omitidos para medidas de desigualdad Nivel de Desigualdad de Oportunidades.

Procedimiento	Media	95 % CI		Sesgo	Tasa de
		lím. inf.	lím. sup.	Promedio	Cobertura***
0 % omisión					
Muestra Completa	**0.10**	**0.09**	**0.12**	–	–
5 % omisión					
Listwise	0.10	0.08	0.12	-0.01	92.4 %
Media Condicional	0.11	0.09	0.13	0.00	94.3 %
Reponderación	0.10	0.09	0.12	0.00	95.6 %
Hotdeck	0.10	0.09	0.12	0.00	95.6 %
Regresión Lineal*	0.11	0.09	0.13	0.00	94.7 %
Regresión Lineal**	0.11	0.09	0.13	0.00	95.0 %
Propensity Score*	0.11	0.09	0.13	0.00	94.9 %
Propensity Score**	0.11	0.09	0.13	0.00	94.7 %
15 % omisión					
Listwise	0.09	0.07	0.11	-0.01	70.8 %
Media Condicional	0.11	0.09	0.13	0.01	90.9 %
Reponderación	0.10	0.08	0.12	0.00	95.7 %
Hotdeck	0.10	0.09	0.12	0.00	96.4 %
Regresión Lineal*	0.11	0.09	0.13	0.00	94.2 %
Regresión Lineal**	0.11	0.08	0.13	0.00	92.9 %
Propensity Score*	0.11	0.09	0.13	0.00	95.0 %
Propensity Score**	0.11	0.09	0.13	0.00	95.5 %
25 % omisión					
Listwise	0.08	0.06	0.10	-0.02	38.2 %
Media Condicional	0.11	0.09	0.13	0.01	93.1 %
Reponderación	0.10	0.08	0.12	0.00	94.3 %
Hotdeck	0.10	0.08	0.12	-0.01	92.0 %
Regresión Lineal*	0.11	0.08	0.13	0.00	93.0 %
Regresión Lineal**	0.11	0.08	0.13	0.00	90.2 %
Propensity Score*	0.10	0.09	0.12	0.00	95.4 %
Propensity Score**	0.10	0.09	0.12	0.00	95.6 %
35 % omisión					
Listwise	0.07	0.06	0.09	-0.03	7.9 %
Media Condicional	0.11	0.09	0.13	0.00	96.8 %
Reponderación	0.10	0.08	0.12	-0.01	87.1 %
Hotdeck	0.09	0.07	0.11	-0.01	80.0 %
Regresión Lineal*	0.10	0.08	0.12	0.00	90.4 %
Regresión Lineal**	0.10	0.08	0.13	0.00	90.1 %
Propensity Score*	0.10	0.08	0.12	-0.01	87.9 %
Propensity Score**	0.10	0.08	0.12	-0.01	88.3 %

*Método considerando las variables escolaridad, escolaridad del padre y edad.

**Método considerando las variables escolaridad y escolaridad del padre. Mismas utilizadas para formar estratos en los demás métodos.

***Proporción de veces que el parámetro estimado está contenido en el verdadero intervalo de confianza.

Tabla 8: Resultados de la corrección por datos omitidos para medidas de desigualdad Proporción de Desigualdad de Oportunidades.

Procedimiento	Media	95 % CI		Sesgo	Tasa de
		lím. inf.	lím. sup.	Promedio	Cobertura***
0 % omisión					
Muestra Completa	**27.0 %**	**22.6 %**	**31.2 %**	–	–
5 % omisión					
Listwise	26.8 %	22.4 %	31.1 %	-0.15	95.1 %
Media Condicional	29.0 %	24.7 %	33.2 %	2.08	83.3 %
Reponderación	27.5 %	23.2 %	31.8 %	0.58	94.1 %
Hotdeck	27.8 %	23.5 %	31.9 %	0.81	93.7 %
Regresión Lineal*	27.3 %	23.0 %	31.5 %	0.33	94.4 %
Regresión Lineal**	27.3 %	22.9 %	31.5 %	0.36	94.7 %
Propensity Score*	27.6 %	23.3 %	31.9 %	0.60	94.0 %
Propensity Score**	27.6 %	23.3 %	31.8 %	0.62	93.8 %
15 % omisión					
Listwise	25.9 %	21.2 %	30.3 %	-1.07	92.2 %
Media Condicional	31.5 %	27.1 %	35.6 %	4.55	43.3 %
Reponderación	27.9 %	23.4 %	32.2 %	0.98	91.9 %
Hotdeck	28.0 %	23.6 %	32.2 %	1.07	91.7 %
Regresión Lineal*	27.5 %	22.8 %	31.7 %	0.51	93.2 %
Regresión Lineal**	27.5 %	22.3 %	32.0 %	0.53	90.9 %
Propensity Score*	28.1 %	23.7 %	32.3 %	1.15	91.5 %
Propensity Score**	28.2 %	23.8 %	32.4 %	1.29	90.6 %
25 % omisión					
Listwise	24.8 %	20.1 %	29.4 %	-2.19	81.0 %
Media Condicional	33.0 %	28.6 %	37.1 %	6.06	20.5 %
Reponderación	28.0 %	23.4 %	32.5 %	1.05	90.4 %
Hotdeck	27.3 %	22.8 %	31.7 %	0.36	93.5 %
Regresión Lineal*	27.3 %	21.8 %	32.0 %	0.32	91.0 %
Regresión Lineal**	27.6 %	21.7 %	32.5 %	0.63	86.4 %
Propensity Score*	27.8 %	23.2 %	32.2 %	0.87	91.9 %
Propensity Score**	28.0 %	23.5 %	32.3 %	1.05	90.8 %
35 % omisión					
Listwise	23.0 %	18.1 %	27.7 %	-4.00	57.1 %
Media Condicional	34.1 %	29.6 %	38.3 %	7.15	9.3 %
Reponderación	27.1 %	22.3 %	31.9 %	0.19	92.0 %
Hotdeck	26.3 %	21.4 %	30.8 %	-0.67	91.8 %
Regresión Lineal*	26.2 %	20.8 %	31.1 %	-0.73	88.9 %
Regresión Lineal**	27.0 %	21.4 %	31.9 %	0.01	89.2 %
Propensity Score*	26.9 %	22.0 %	31.4 %	-0.08	93.2 %
Propensity Score**	27.0 %	22.4 %	31.4 %	0.08	93.1 %

*Método considerando las variables escolaridad, escolaridad del padre y edad.

**Método considerando las variables escolaridad y escolaridad del padre. Mismas utilizadas para formar estratos en los demás métodos.

***Proporción de veces que el parámetro estimado está contenido en el verdadero intervalo de confianza.

4.3.3. Robustez

La calidad de los métodos de corrección por datos perdidos dependerá de tres factores principales: i) el porcentaje de omisión presente en la muestra, ii) la distribución del vector aleatorio que describe los datos y iii) la distribución de los datos faltantes. En el apartado anterior se ha revisado cómo afecta el primer factor en el desempeño de los distintos procedimientos de corrección. El segundo factor, por su parte, queda también zanjado por el proceso descrito en 1. Esto porque estamos intentando replicar las circunstancias y condiciones que se pueden observar en el contexto del mercado del trabajo. Con respecto al tercer factor, hasta ahora, se ha asumido que el proceso generador de la pérdida de datos está dado por la ecuación 2. No obstante, el haber modelado la omisión bajo esta regla, entrega la posibilidad de variar algunos parámetros de ella y observar el desempeño de los distintos métodos de corrección, bajo estos nuevos supuestos.

El variar el valor de la desviación estándar de ε_2 (σ_2) en la ecuación 2, permitirá cambiar el grado de severidad del sesgo de selección, es decir, podremos decidir cuan alejados pretendemos que estén los datos de distribuirse bajo el supuesto MAR y cómo cambia el desempeño de los distintos métodos de imputación dado esta variación.

Esto porque, al variar la desviación estándar del proceso generador de la pérdida de datos, se está variando la proporción de la regla que se debe a la suerte versus la proporción debida al poder explicativo de las covariables. En otras palabras, el aumentar la desviación estándar de ε_2, implicará que las variables explicativas presentes en la regla tengan un menor peso a la hora de decidir qué individuos quedarán sin dato de ingresos y será la suerte la que lo decida, haciendo la omisión de datos más aleatoria. Por el contrario, mientras menor sea la desviación estándar del error en la ecuación 2, la regla será más clara, pues el peso relativo de las variables explicativas versus el de la suerte será mayor, haciendo que el azar tenga poco que ver con la decisión de qué individuos quedarán sin dato de ingresos, y por lo tanto, la omisión de datos será menos aleatoria.

Una forma de medir qué proporción de la decisión de quienes se quedan sin dato de ingresos se debe a variables explicativas (y no al azar), es la bondad del ajuste (R^2) del modelo en la

ecuación 2. Específicamente, una mayor desviación estándar corresponderá a un bajo R^2 o a una regla en donde la suerte se encarga de la decisión de quien se queda sin dato, mientras que una desviación estándar más baja, es equivalente a un alto R^2 o a una regla clara, en donde son ciertas variables explicativas las que resuelven el patrón de la omisión de los datos.

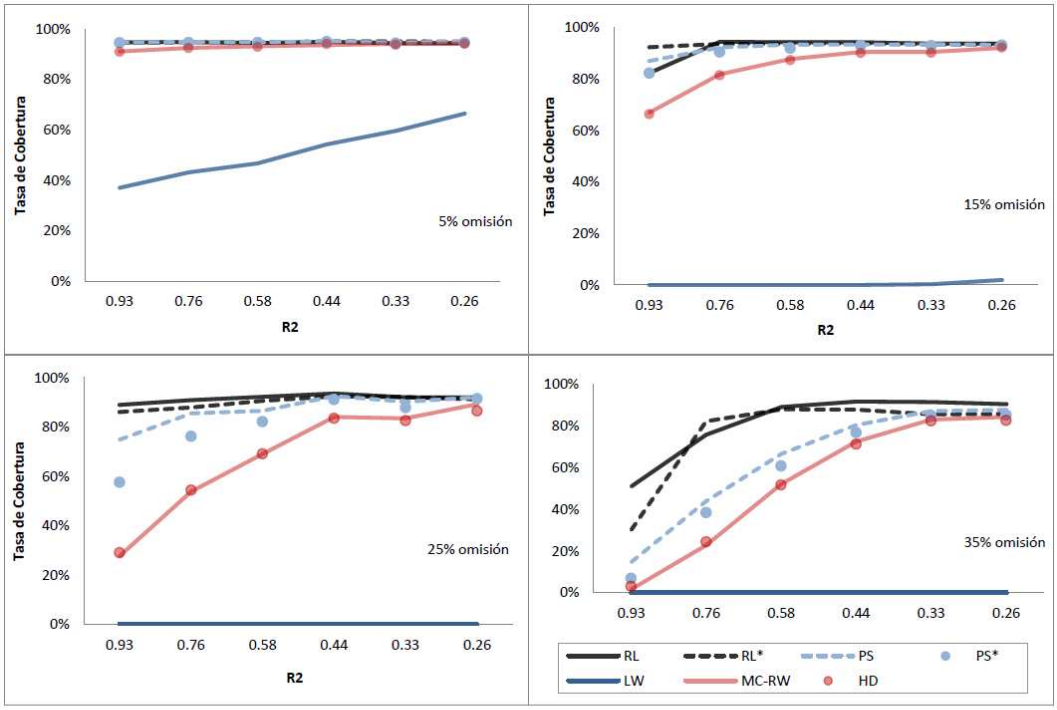
En la figura 1, se observa la tasa de cobertura real con intervalos de confianza del 95 % para la estimación del ingreso promedio, según el nivel de severidad del sesgo de selección, para una omisión del 5, 15, 25, y 35 % de la muestra. Se demuestra que a mayor acercamiento al supuesto MAR (menor R^2), todos los métodos mejoran su desempeño. Además, mientras mayor sea la cantidad de datos omitidos, peor es el funcionamiento de cada método. En el caso más difícil: omisión del 35 % de los datos y $R^2 = 0{,}93$, es donde se aprecia de mejor forma la superioridad del método de regresión lineal para estimar el ingreso promedio. Esto porque sigue muy de cerca el verdadero proceso que genera la omisión en los datos. Aun así, su porcentaje de aciertos alcanza solo un 50 %. En otras palabras, cuando nos alejamos del supuesto MAR y el porcentaje de datos perdidos es alto, difícilmente se tendrán buenas estimaciones de los parámetros a través de este tipo de procedimientos.

La figura 2 presenta el porcentaje de aciertos en la estimación del coeficiente de Theil para cada método. Destaca el bajo desempeño del método de media condicional, que para altos porcentajes de omisión, es aún más bajo que el no corregir (listwise), sobre todo para niveles bajos de R^2. Los resultados de los métodos de reponderación y hotdeck resultan ser similares, para cualquier nivel de severidad del sesgo de selección. Las técnicas del propensity score y regresión lineal estocásticos son superiores al resto de los métodos, para cualquier R^2 y entre estos el de mejor desempeño es el de regresión lineal.

La figura 3 indica la tasa de acierto en el cálculo del nivel de desigualdad de oportunidades. Se aprecia que el nivel de severidad del sesgo no afecta demasiado el desempeño de los diferentes métodos (excepto listwise) para niveles de omisión del 5 y 15 %. Cuando la omisión y el sesgo de selección son más altos, son los métodos de regresión lineal y propensity score los que presentan un mejor desempeño.

Figura 1: Porcentaje de aciertos según severidad del sesgo de selección.

Ingreso Promedio.



RL = Regresión lineal estocástica considerando edad, escolaridad y escolaridad del padre.

RL* = Regresión lineal estocástica considerando escolaridad y escolaridad del padre.

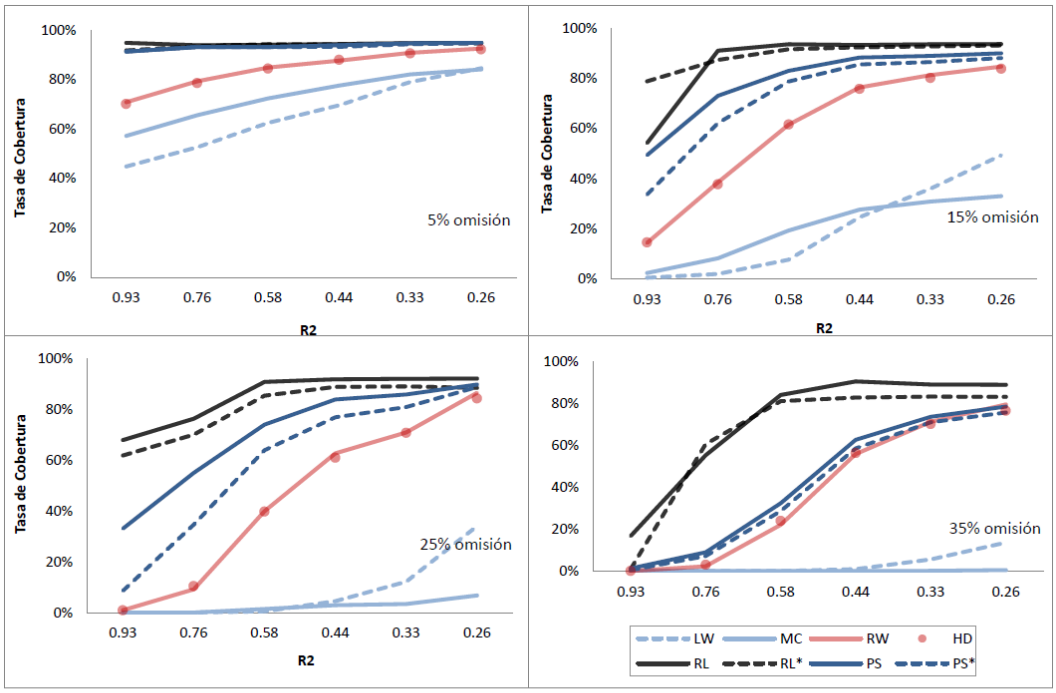
PS = Propensity score considerando edad, escolaridad y escolaridad del padre.

PS* = Propensity score considerando escolaridad y escolaridad del padre.

LW = Listwise. MC = Media condicional. RW = Reponderación. HD = Hotdeck.

Figura 2: Porcentaje de aciertos según severidad del sesgo de selección.

Ingreso Promedio.



RL = Regresión lineal estocástica considerando edad, escolaridad y escolaridad del padre.

RL* = Regresión lineal estocástica considerando escolaridad y escolaridad del padre.

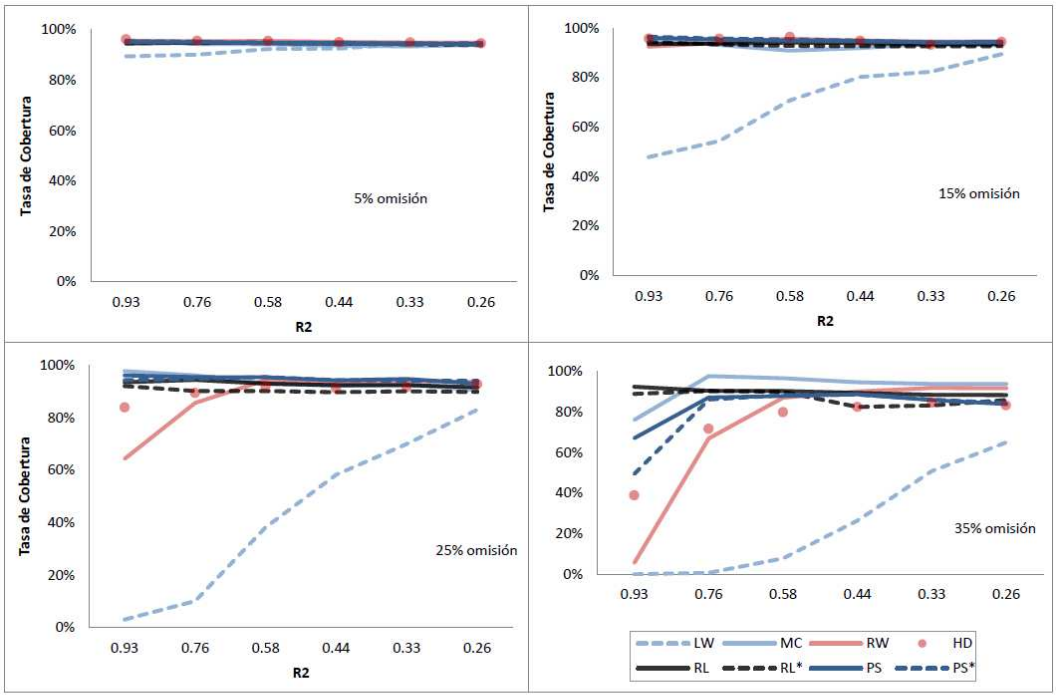
PS = Propensity score considerando edad, escolaridad y escolaridad del padre.

PS* = Propensity score considerando escolaridad y escolaridad del padre.

LW = Listwise. MC = Media condicional. RW = Reponderación. HD = Hotdeck.

Figura 3: Porcentaje de aciertos según severidad del sesgo de selección.

Ingreso Promedio.



RL = Regresión lineal estocástica considerando edad, escolaridad y escolaridad del padre.

RL* = Regresión lineal estocástica considerando escolaridad y escolaridad del padre.

PS = Propensity score considerando edad, escolaridad y escolaridad del padre.

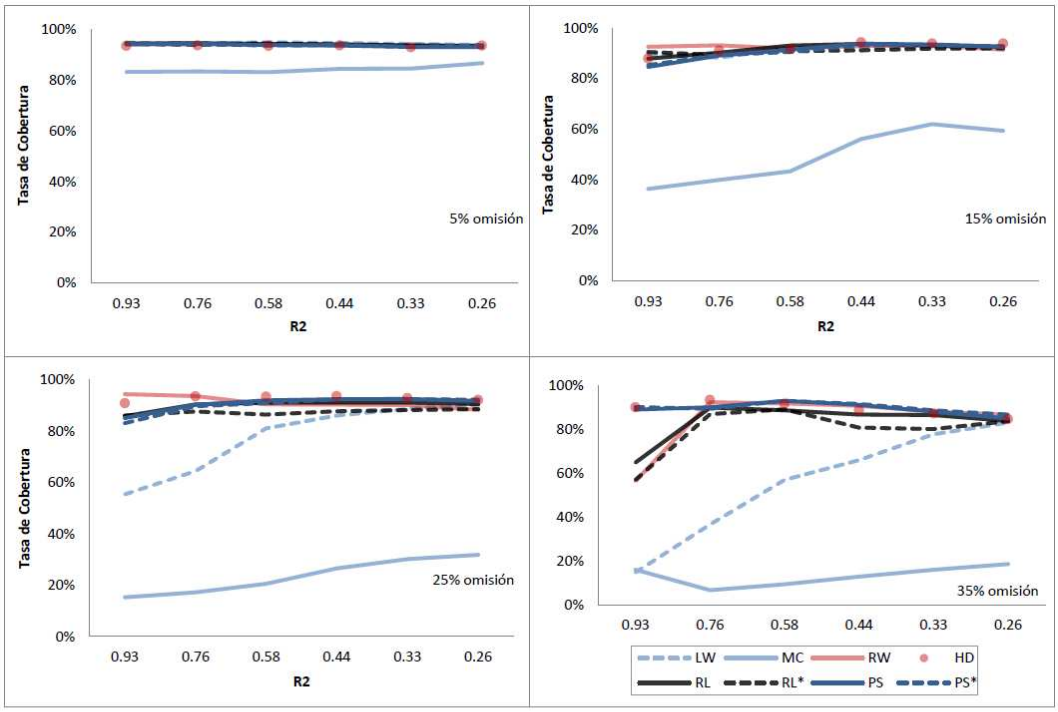
PS* = Propensity score considerando escolaridad y escolaridad del padre.

LW = Listwise. MC = Media condicional. RW = Reponderación. HD = Hotdeck.

Como la proporción de la desigualdad total que se le atribuye a la desigualdad de oportunidades se define como el cociente entre dos indicadores, ésta dependerá netamente de cuan bien estimados se encuentren la desigualdad total y el nivel de desigualdad de oportunidades (figura 2 y 3, respectivamente). Como el método de media condicional subestima de gran manera la desigualdad total pero por otro lado, hace un buen trabajo al calcular el nivel de desigualdad de oportunidades, entonces, sobreestima la proporción de la desigualdad de oportunidades y su desempeño en este indicador es más bajo que el resto de los métodos, para todo nivel de severidad del sesgo de selección (ver figura 4. El resto de los procedimientos tienden a tener subestimar (algunos más que otros) tanto la desigualdad total como el nivel de la inequidad de oportunidades, y por lo tanto, el cociente se mantiene cercano al verdadero valor, para todo nivel de severidad del sesgo de selección en niveles bajos de omisión (5 y 15 %). Para niveles de pérdida de datos mayores al 15 %, las tasas de coberturas del procedimiento listwise se encuentran cercanas al 80 % para niveles bajos de R^2 (menores o iguales a 0.43 para una omisión del 25 % y mayores o iguales a 0.33 para una omisión de datos del 35 %).

Figura 4: Porcentaje de aciertos según severidad del sesgo de selección.

Ingreso Promedio.



RL = Regresión lineal estocástica considerando edad, escolaridad y escolaridad del padre.

RL* = Regresión lineal estocástica considerando escolaridad y escolaridad del padre.

PS = Propensity score considerando edad, escolaridad y escolaridad del padre.

PS* = Propensity score considerando escolaridad y escolaridad del padre.

LW = Listwise. MC = Media condicional. RW = Reponderación. HD = Hotdeck.

5. Conclusiones

En el presente trabajo se ha evaluado y comparado el desempeño de distintos procedimientos, generalmente utilizados en la literatura para corregir problemas de datos faltantes, cuando la distribución de la omisión es no aleatoria. Esto, enmarcado en el contexto de la medición de la desigualdad y la desigualdad de oportunidades en el mercado laboral.

Se ha comprobado que todos los métodos analizados poseen limitaciones y su correcta aplicación depende principalmente de la magnitud de la omisión de datos en la muestra y la forma en que se comporta la distribución de los datos faltantes. Para un mismo nivel de severidad del sesgo de selección, en la medida que aumenta la tasa de omisión de datos, la eficacia de todos los procedimientos se debilita. De la misma forma, para una tasa de pérdida de datos determinada, mientras menos aleatorio sea el patrón de los datos faltantes, más bajo será el desempeño de todos los procedimientos estudiados.

Si bien es complicado dar órdenes de magnitud, pues cada situación será particularmente diferente y dependiente de la variable en estudio, del porcentaje de datos faltantes, del tipo de encuesta que se analice y del uso que se hará de la información imputada, es posible sacar algunas conclusiones generales para contextos similares al aquí analizado.

Primero, cuando se trabaja con tasas de no respuesta que alcanzan magnitudes superiores al 25 %, y los niveles de severidad del sesgo de selección son altos (por ejemplo, estudios en que se pudiera esperar un proceso generador de datos faltantes con un mayor a 0.4), se recomienda ser extremadamente cuidadoso y no basar decisiones a partir de estos indicadores pues todos los procedimientos analizados entregan estimaciones débiles.

Segundo, al trabajar con bajas tasas de no respuesta (15 % o menos) y niveles de severidad del sesgo de selección moderados (por ejemplo, estudios en que se pudiera esperar un proceso generador de datos faltantes con un menor a 0.3), se espera un buen rendimiento por parte de los distintos procedimientos analizados. Para estimar la media del ingreso, los métodos propensity score y regresión lineal estocástica muestran mejor desempeño que los demás, y su ventaja radica en que no requieren la definición de estratos, los que disminuyen la cantidad de

variables relevantes que se pueden considerar en el proceso de corrección de datos faltantes. Para efectos de la estimación de indicadores de desigualdad, por otra parte, no se recomienda bajo ningún término la utilización del método de imputación de medias condicionales, pues distorsiona la distribución de la variable imputada (ingresos en la mayoría de los casos), que es justamente lo que miden los indicadores de inequidad.

Tercero, los métodos hotdeck y reponderación son una buena alternativa a los métodos más sofisticados cuando la información de las covariables es limitada. Por ejemplo, si se tuviesen datos de la variable de resultado en una encuesta, e información acerca de la distribución de los individuos con respecto a su background en otra, se podrían combinar ambas creando estratos con variables relevantes que se encuentren en las dos encuestas y evaluar la robustez de los indicadores de desigualdad y desigualdad de oportunidades a través de estos métodos.

Finalmente se concluye que no existe un mejor método de imputación, cada situación será particular y un análisis exhaustivo de la tasa de no respuesta, las características de los datos y la distribución de la omisión será necesario. Los métodos aquí revisados, ampliamente utilizados en la literatura de pérdida de datos, si bien no corrigen el problema del sesgo de selección, nos entregan un set de herramientas muy útiles y simples para evaluar la robustez de los indicadores de desigualdad y desigualdad de oportunidades en el mercado del trabajo y otros contextos similares.

Referencias

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1).

Allison, P. D. (1999). Multiple imputation for missing data: A cautionary tale. *SAGE Quantitative Research Methods*, page 259.

Allison, P. D. (2001). *Missing data*. Number 136. SAGE Publications, Incorporated.

Bourguignon, F., Ferreira, F. H., and Menendez, M. (2007a). Inequality of opportunity in Brazil. *Review of Income and Wealth*, 53(4):585–618.

Bourguignon, F., Ferreira, F. H., and Walton, M. (2007b). Equity, efficiency and inequality traps: A research agenda. *The Journal of Economic Inequality*, 5(2):235–256.

Checchi, D. and Peragine, V. (2010). Inequality of opportunity in italy. *The Journal of Economic Inequality*, 8(4):429–450.

Collins, L. M., Schafer, J. L., and Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures.

Contreras, D., Larrañaga, O., Puentes, E., and Rau, T. (2009a). Evidence for inequality of opportunities. a cohort analysis for Chile.

Contreras, D., Larrañaga, O., Puentes, E., and Rau, T. (2009b). The evolution of opportunities for children in chile 1990-2006.

Contreras, D., Larrañaga, O., Puentes, E., and Rau, T. (2012). Inequality of opportunities and long term earnings measures: Evidence for chile.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89(426):463–475.

Ferreira, F. H. and Gignoux, J. (2011a). The measurement of educational inequality: Achievement and opportunity.

Ferreira, F. H. and Gignoux, J. (2011b). The measurement of inequality of opportunity: Theory and an application to latin america. *Review of Income and Wealth*, 57(4):622–657.

Ferreira, F. H., Gignoux, J., and Aran, M. (2011). Measuring inequality of opportunity with imperfect data: the case of turkey. *The Journal of Economic Inequality*, 9(4):651–680.

Galván, M. and Medina, F. (2007). *Imputación de Datos: Teoría y Práctica*. UN.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association*, 88(423):984–993.

Graham, J. W. and Donaldson, S. I. (1993). Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology*, 78(1):119.

Kalton, G. and Kasprzyk, D. (1982). Imputing for missing survey responses. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, volume 22.

Lefranc, A., Pistolesi, N., and Trannoy, A. (2008). Inequality of opportunities vs. inequality of outcomes: Are western societies all alike? *Review of Income and Wealth*, 54(4):513–546.

Little, R. J. and Rubin, D. B. (1987). *Statistical analysis with missing data*, volume 539. Wiley New York.

Little, R. J. and Rubin, D. B. (2002). Statistical analysis with missing data.

Marrero, G. A. (2009). Inequality of opportunity and growth. *Documentos de trabajo (FEDEA)*, (24):1–40.

Rawls, J. (1971). A theory of justice. *Harvard University Press*.

Roemer, J. E. (1993). A pragmatic theory of responsibility for the egalitarian planner. *Philosophy & Public Affairs*, 22(2):146–166.

Roemer, J. E. (1998). *Equality of opportunity*. Harvard University Press.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods*, 7(2):147–177.

Sen, A. (1980). Equality of what? *The Tanner lectures on human values*, 1:353–369.

Van Buuren, S., Boshuizen, H. C., Knook, D. L., et al. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18(6):681–694.

Van de Gaer, D. (1993). *Equality of opportunity and investment in human capital*. Katholieke Universiteit Leuven, Faculteit der Economische en Toegepaste Economische Wetenschappen.

Anexo

A. Metodología para el cálculo de la desigualdad de oportunidades

Ésta se basa en los trabajos de Bourguignon et al. (2007a) y Ferreira and Gignoux (2011a).

Roemer (1998) postula que las variables de bienestar, como los ingresos, son consecuencia de dos tipos de variables: las circunstancias y el esfuerzo. Las circunstancias son todas aquellas variables que el individuo no puede controlar, por ejemplo, el sexo o el nivel educacional de sus padres. Por otra parte, el individuo sí controla su nivel de esfuerzo. Esta idea se puede modelar como sigue:

$$w = g(C, E, u)$$

donde w es la variable de interés o de resultado; C el vector de variables de circunstancias que se asumen exógenos; E el vector de variables de esfuerzo; y u se refiere a la suerte o factores aleatorios.

Siguiendo a Bourguignon et al. (2007a), el modelo puede ser aproximado al siguiente sistema lineal:

$$lnw = C\alpha + E\beta + u$$

$$E = C\lambda + v$$

Donde la ecuación estructural relaciona ingreso con variables de circunstancias y esfuerzo log-linealmente. Además, contamos con una ecuación lineal para las variables de esfuerzo que responde a las circunstancias y a no observables.

La forma reducida de este sistema será:

$$lnw = C(\alpha + \beta\lambda) + v\beta + u$$

La cual puede ser estimada por OLS dada la exogeneidad de las circunstancias:

$$lnw = C\theta + \varepsilon$$

Bajo estos supuestos, una distribución paramétricamente estandarizada $\{\bar{y}_i\}$ se estima como sigue:

$$\bar{y}_i = exp(\bar{C}\theta_{ols} + \varepsilon_{i,ols})$$

donde $\bar{C}$ corresponde al promedio poblacional de las circunstancias, θ_{ols} al vector de parámetros obtenidos a través de OLS y $\varepsilon_{i,ols}$ el residuo para cada individuo i.

Esta distribución paramétricamente estandarizada es una distribución contrafactual del ingreso, donde todas las diferencias en las circunstancias han sido eliminadas y solo la suerte y el esfuerzo permanecen. La medida de desigualdad de oportunidad que se obtiene de esta estimación viene dada por:

$$S = 1 - [D(\bar{y}_i)/D(y_i)]$$

Donde $D()$ es cualquier índice de desigualdad calculado sobre la distribución de ingresos. S, entonces, mide la proporción de la desigualdad de oportunidades en la desigualdad total. Como se menciona en Ferreira and Gignoux (2011a), S es una cota inferior de la medida de desigualdad de oportunidades, debido a que en el caso de incluir más variables de circunstancias en la ecuación, la parte del total de la distribución del ingreso explicada por todas las circunstancias nunca disminuirá. Luego, si existiesen variables de circunstancias omitidas, el resultado sería un S mayor.

También es posible medir el nivel de desigualdad de oportunidades, como sigue:

$$D(\bar{y}_i) - D(y_i)$$

Si bien existen otros métodos para medir desigualdad de oportunidades que no imponen una forma funcional a la relación entre las variables, estas estimaciones no paramétricas restringen el número de circunstancias a ser incluidas y a su vez, el número de categorías dentro de cada variable de circunstancia que se pueden definir, dado que si estos son muy altos, la

varianza muestral se vuelve demasiado grande, haciendo de los indicadores de desigualdad de oportunidades, estimaciones muy imprecisas (Contreras et al., 2012).

B. Distribución de la omisión de datos según severidad del sesgo de selección para distintas tasas de no respuesta

Tabla 9: Distribución de la omisión de datos según severidad del sesgo de selección
Tasa de no respuesta del 5 %

Decil de Ingresos	R^2					
	0.93	0.76	0.58	0.44	0.33	0.26
1	31.2 %	27.8 %	26.2 %	22.1 %	18.9 %	16.1 %
2	11.1 %	9.8 %	9.2 %	9.5 %	8.5 %	8.2 %
3	4.1 %	5.7 %	5.4 %	5.7 %	6.3 %	6.6 %
4	2.8 %	4.1 %	4.4 %	4.7 %	5.1 %	5.7 %
5	0.3 %	1.3 %	1.9 %	3.2 %	3.8 %	4.4 %
6	0.3 %	0.9 %	1.6 %	2.2 %	2.8 %	3.2 %
7	0.3 %	0.3 %	0.3 %	0.9 %	1.3 %	1.3 %
8	0.0 %	0.3 %	0.9 %	1.3 %	1.9 %	2.5 %
9	0.0 %	0.0 %	0.3 %	0.6 %	0.9 %	1.3 %
10	0.0 %	0.0 %	0.0 %	0.0 %	0.6 %	0.9 %
Total	5.0 %	5.0 %	5.0 %	5.0 %	5.0 %	5.0 %

Tabla 10: Distribución de la omisión de datos según severidad del sesgo de selección
Tasa de no respuesta del 15 %

Decil de Ingresos	R^2					
	0.93	0.76	0.58	0.44	0.33	0.26
1	65.9 %	59.6 %	52.1 %	44.8 %	41.3 %	38.2 %
2	34.2 %	32.6 %	31.3 %	28.5 %	26.9 %	25.3 %
3	21.5 %	20.8 %	20.5 %	20.5 %	20.5 %	19.2 %
4	13.3 %	14.2 %	16.1 %	17.1 %	17.1 %	17.7 %
5	6.9 %	8.2 %	9.1 %	9.5 %	10.1 %	10.7 %
6	3.8 %	7.0 %	8.5 %	8.9 %	11.1 %	11.4 %
7	2.2 %	3.2 %	4.1 %	7.3 %	7.6 %	9.1 %
8	1.6 %	3.2 %	5.1 %	8.2 %	9.2 %	9.5 %
9	0.6 %	0.9 %	2.5 %	4.1 %	4.4 %	5.4 %
10	0.0 %	0.3 %	0.6 %	1.3 %	1.9 %	3.5 %
Total	15.0 %	15.0 %	15.0 %	15.0 %	15.0 %	15.0 %

Tabla 11: Distribución de la omisión de datos según severidad del sesgo de selección Tasa de no respuesta del 25 %

Decil de Ingresos	R^2					
	0.93	0.76	0.58	0.44	0.33	0.26
1	83.0 %	77.0 %	69.1 %	58.7 %	55.5 %	49.2 %
2	57.3 %	53.2 %	47.2 %	44.3 %	40.8 %	38.6 %
3	41.0 %	38.8 %	36.0 %	36.6 %	36.3 %	35.6 %
4	29.4 %	29.1 %	32.0 %	31.3 %	30.7 %	30.4 %
5	16.4 %	18.6 %	17.7 %	19.2 %	20.8 %	21.5 %
6	10.8 %	14.2 %	18.4 %	21.2 %	20.9 %	22.2 %
7	5.7 %	8.2 %	10.7 %	12.3 %	14.2 %	16.1 %
8	4.4 %	7.3 %	11.4 %	14.2 %	15.5 %	17.1 %
9	1.9 %	3.2 %	6.0 %	8.5 %	10.4 %	12.3 %
10	0.3 %	0.6 %	1.9 %	3.8 %	5.1 %	7.3 %
Total	25.0 %	25.0 %	25.0 %	25.0 %	25.0 %	25.0 %

Tabla 12: Distribución de la omisión de datos según severidad del sesgo de selección Tasa de no respuesta del 35 %

Decil de Ingresos	R^2					
	0.93	0.76	0.58	0.44	0.33	0.26
1	92.1 %	88.0 %	79.8 %	71.9 %	67.2 %	62.5 %
2	74.7 %	66.5 %	62.0 %	56.0 %	52.8 %	51.3 %
3	59.3 %	54.9 %	53.6 %	49.5 %	46.4 %	46.1 %
4	43.0 %	44.3 %	42.4 %	43.0 %	41.5 %	42.4 %
5	30.3 %	30.6 %	31.5 %	31.9 %	32.2 %	31.5 %
6	21.8 %	25.3 %	28.2 %	30.7 %	32.3 %	32.9 %
7	14.2 %	16.4 %	19.6 %	23.0 %	24.9 %	26.8 %
8	9.5 %	14.2 %	17.7 %	20.9 %	21.8 %	23.1 %
9	3.8 %	7.3 %	10.4 %	14.8 %	19.2 %	20.2 %
10	1.3 %	2.5 %	4.7 %	8.2 %	11.7 %	13.3 %
Total	35.0 %	35.0 %	35.0 %	35.0 %	35.0 %	35.0 %