Identificación de potencial sesgo por no respuesta en encuestas de hogares del nordeste argentino sobre el ingreso familiar y el nivel de pobreza

Maestría en Estadística Aplicada
Universidad Nacional de Córdoba

Tesista	Lic. Celine Iliana Cabás
Directora	Dra. Patricia Caro
Co-Director	Dr. Carlos Matías Hisgen

Octubre 2024

1. Introducción

En el campo estadístico del muestreo, la no respuesta es una de las principales problemáticas a abordar debido a que si no se comporta de manera aleatoria puede introducir sesgos en las distribuciones de las variables que nos interesan medir y, por lo tanto, en las conclusiones que saquemos de testear hipótesis a partir de ellas.

> "Si la decisión de responder depende estadísticamente de las variables bajo investigación, entonces la submuestra de encuestados no reflejará con precisión la distribución real de las variables de interés en la población y esto, a su vez, dará como resultado inferencias basadas en muestras sistemáticamente sesgadas". [1]

Con respecto a esta problemática, un producto estadístico muy estudiado son las encuestas de hogares que relevan los países para medir diversos indicadores económicos y sociales. Entre sus finalidades, una de ellas es poder conocer la distribución del ingreso de las familias e individuos de cierto entorno geográfico y poder realizar inferencia sobre ellas.

A su vez, dentro de las medidas más relevantes calculadas a partir del ingreso, se encuentra el nivel de pobreza monetaria para las distintas unidades geográficas que conforman la muestra. Este indicador constituye una de las medidas centrales a tener en cuenta cuando se trata de implementación de políticas públicas.

El presente trabajo tiene como objetivo identificar si existe presencia de sesgo por no respuesta unitaria en el ingreso familiar medido por encuestas de hogares relevadas en nordeste argentino. El estudio se centra en los aglomerados urbanos de Gran Resistencia, Corrientes, Formosa y Posadas para los años 2018-2022 con periodicidad trimestral.

De los distintos enfoques que existen para abordar el problema de la no respuesta, este proyecto pretende estudiar el caso particular de la no respuesta 'unitaria' que hace referencia a la pérdida de observaciones por entrevistas no realizadas. Se trabajará con modelos que determinen la potencial presencia de sesgo por no respuesta y, en caso de existir, se propondrán métodos de reponderación de los datos para minimizar el sesgo.

En primera instancia, se pretende identificar si la cantidad total de entrevistas realizadas por un mismo hogar a lo largo del período depende estadísticamente o no del ingreso per cápita familiar y de otras variables que reflejen su condición socioeconómica. Se plantea un modelo logístico multinomial con variable de respuesta ordinal para estudiar si la predisposición a responder depende significativamente de las variables bajo estudio. [2] [1]

En el caso de probar la presencia potencial de sesgo, en segunda instancia se pretende trabajar con métodos de reponderación de los datos que corrijan por no respuesta. Principalmente, aquellos que utilicen la probabilidad predicha de respuesta de los hogares.

En este trabajo, se plantea esta problemática en el marco de la Encuesta Permanente de Hogares (EPH) para los aglomerados urbanos del nordeste argentino (NEA). Específicamente en lo que refiere al ingreso per cápita familiar y, por ende, al nivel de pobreza de los hogares. La selección de la región NEA viene justificada por el interés de estudiar si las marcadas diferencias entre el nivel de pobreza monetaria de Gran Resistencia y los demás aglomerados urbanos de la región puede verse justificada por diferencias observables en sus estructuras de no respuesta.

2. Antecedentes

En la literatura se han propuesto y estudiado procedimientos para minimizar el sesgo por no respuesta en las distintas etapas de una encuesta. Ya sea respecto a cómo tratar la no respuesta antes y durante la recolección de datos, previendo cuestiones vinculadas al contacto con la persona encuestada, como también métodos que evalúan sesgos potenciales posterior a la recolección y maneras de corregirlos [3]. En esta investigación se trabajará en el segundo aspecto, ya que se emplearán datos secundarios relevados por programas nacionales.

En estudios basados en muestras, puede darse que la no respuesta no se comporte de manera aleatoria. Es decir, cuando las personas que sí respondieron a la encuesta difieren significativamente de aquellas que no lo hicieron, debe tenerse cuidado respecto a la validez de la inferencia porque se estaría trabajando con distribuciones sesgadas [4].

En esta línea, el problema de la no respuesta puede entenderse principalmente de dos maneras. Como "no respuesta al ítem", cuando el individuo no responde a un ítem en particular pero sí participa de la encuesta, o como "no respuesta unitaria", donde el individuo directamente no realiza la entrevista [1]. Ambas problemáticas pueden ser abordadas mediante técnicas de imputación, o bien mediante técnicas de reponderación de los datos que corrijan sesgos de selección.

Entonces, ¿cómo detectar el sesgo por no respuesta? Existen diversas alternativas como el test de Little para determinar si el patrón de la no respuesta sigue un comportamiento completamente aleatorio [5]. Otra alternativa son los test chi-cuadrado de contraste de independencia, o bien modelos de regresión logísticos de respuesta binaria o politómica que permiten testear si existe relación entre ciertas variables de interés y la propensión a responder de los hogares.

Una vez detectado que la no respuesta no sigue un comportamiento aleatorio en la muestra, la siguiente cuestión a resolver es cómo lidiar con la no respuesta ¿de qué manera corregir el sesgo? Existen técnicas de imputación para sustituir individuos no encuestados, por ejemplo dividir la muestra en subclases e imputar un individuo similar a la misma subclase. Además, se desarrollaron métodos basados en variables auxiliares que son ampliamente utilizados, conocidos como métodos de calibración. Estas variables deben estar disponible para los respondentes y al mismo tiempo conocer algún tipo de información sobre la distribución de esta variable en la población [2] [6].

Por otro lado, una alternativa muy utilizada son los modelos basados en la probabilidad de respuesta, ampliamente conocidos como Propensity Score Adjustment (PSA). Estos métodos parten desde modelos logísticos de respuesta binaria o modelos multinomiales de respuesta politómica hasta modelos más modernos vinculados a la literatura de Machine Learning (CART, RF, XGBoost) como en los trabajo de [7], [8], [9] que predicen la probabilidad de respuesta para incluirla en la corrección del ponderador . El objetivo de este método es mejorar la representatividad en la muestra de aquellos hogares con baja propensión a responder respecto a aquellos con alta propensión a responder, y de esta manera equilibrar la estructura de la muestra para corregir posibles sesgos [2].

Una vez calculada la probabilidad, el método más usual es el de la ponderación inversa de la probabilidad que recalibra los pesos de los hogares con baja probabilidad de respuesta mejorando su representatividad en la muestra [7].

En línea con lo anterior, dependiendo la flexibilidad del modelo implementado para la predicción de la probabilidad de respuesta, tendremos mayor o menor varianza en las estimaciones basados en los factores de expansión corregidos. Por este motivo, otro aspecto importante a tener en cuenta es el efecto del sobreajuste con potenciales costos en el incremento de la variabilidad de las estimaciones de nuestra variable de interés. Se han desarrollado y estudiado la eficacia de métodos conocidos como de Weight Smoothing para reducir la varianza de los estimadores mediante el modelado de los ponderadores condicionados a las variables de interés [10].

En relación a antecedentes empíricos, la investigación [1] trabaja con la encuesta Current Population Survey (CPS) de U.S. Census Bureau para analizar la sensibilidad de la distribución acumulada del ingreso frente a ajustes en los factores de expansión basados en la no respuesta unitaria. Para ello, plantea una especificación de la propensión a responder como función del ingreso y utiliza estas predicciones en la corrección de la matriz de varianzas y covarianzas del ingreso mediante GMM (Generalized method of moments). Por otro lado, el artículo de Pkewis y Shlomo [11] estudia estrategias utilizadas para mantener la representatividad, en este caso, en muestras longitudinales en vistas a mejorar la asignación de recursos en el relevamiento de este tipo de encuestas. Trabaja con la encuesta UK Millennium Cohort Study (MCS) definiendo subgrupos en base al menor o mayor trabajo de mantenimiento o intervención que requieran según su propensión a responder. Aplica los modelos de PSA con indicadores R de representatividad y curvas ROC para definir los grupos que requieren seguimiento. Otro ejemplo de estudios en muestras longitudinales es el de König y Sakshaug [12] sobre la encuesta alemana IAB Establishment Panel (IAB-EP) donde analizan tendencias en la tasa de respuesta en un período de 17 años y describen las dificultades que existen para mantener que la motivación de los establecimientos para seguir participando en el panel de entrevistas y cómo esta motivación disminuye luego de la primera entrevista.

3. Formulación del problema y objetivos

La **pregunta de investigación** planteada en este trabajo es la siguiente:
¿Qué efectos tiene la presencia de sesgo por no respuesta en encuestas de hogares del nordeste argentino sobre la estimación del ingreso familiar y el nivel de pobreza?

El **objetivo general** es identificar la potencial presencia de sesgo por no respuesta en encuestas de hogares del nordeste argentino y sus efectos sobre el ingreso familiar y el nivel de pobreza durante el período 2018-2022.

Los **objetivos específicos** son:

- Comparar las estructuras de respuesta de la encuesta de hogares entre los aglomerados urbanos del nordeste argentino para el período 2018-2022.
- Comprobar si la predisposición a responder por parte de los hogares depende significativamente del ingreso familiar u otras variables en los distintos aglomerados urbanos del nordeste argentino.
- Comparar modelos para predecir la probabilidad de los hogares de responder de manera completa el esquema de entrevistas de la encuesta.
- Plantear una corrección del sesgo por no respuesta basada en la probabilidad predicha de responder mediante técnicas de reponderación de los datos.
- Contrastar las distribuciones de ingresos y el nivel de pobreza estimado antes y después de la corrección por no respuesta.

4. Fuentes de información

La principal fuente de información de este trabajo es la Encuesta Permanente de Hogares (EPH) relevada por el Instituto Nacional de Estadísticas y Censos (INDEC) de la República Argentina, que se constituye como un programa nacional de relevamiento de información sobre trabajo e ingresos de los hogares y es empleada para el cálculo de la pobreza. Se trabajará con las bases de datos por hogar e individuo de periodicidad trimestral para el período que comprende el primer trimestre 2018 al cuarto trimestre 2022. El listado preliminar de variables a utilizar se presenta en el Cuadro 1.

Cuadro 1: Descripción de variables a utilizar en EPH individual y hogar.

Variable	Descripción
Identificación	
CODUSU	Código de identificación de la vivienda
NRO_HOGAR	Código de identificación del hogar
REGION	Código de región geográfica
AGLOMERADO	Código de aglomerado urbano
ANO4	Año de relevamiento
TRIMESTRE	Trimestre de relevamiento
Base individual	
CH03	Relación de parentesco (Jefe de hogar=1)
CH04	Sexo
CH06	Edad
NIVEL_ED	Nivel educativo
ESTADO	Condición de actividad
CAT_OCUP	Categoría ocupacional
Base hogar	
IV1	Tipo de vivienda
IX_TOT	Cantidad de miembros del hogar
ITF	Ingreso total familiar
IPCF	Ingreso per cápita familiar
Variable de respuesta	
NRO_REP	Número de entrevistas realizadas en el período 2018-2022
Ponderador	
PONDIH	Ponderador del ITF y del IPCF

Se complementa su uso con el Índice de Precios al Consumidor (IPC) para deflactar los ingresos y poder estudiar los efectos constantes durante el período. Además, se empleará la valorización de la canasta básica total (CBT) para la determinación de la condición de pobreza o no del hogar por período. Ambas variables definidas para la región nordeste de Argentina.

La muestra sigue un esquema de rotación trimestral de las viviendas por área muestral. Es decir que, idealmente, una misma vivienda debe ser encuestada dos trimestres consecutivos, descansar los dos trimestres subsiguientes y volver a ser encuestada dos trimestres consecutivos más para garantizar la estructura de panel de datos que caracteriza a la encuesta. Sin embargo, debido a la no respuesta no todos los hogares completan el esquema. Por lo cual, el número de entrevistas realizadas (del 1 al 4) nos sirve como variable proxy para medir la menor o mayor tendencia a responder que tienen los hogares que son encuestados.

Dado que los datos públicos de la encuesta incluyen únicamente los casos con entrevista realizada, para poder estudiar la probabilidad de respuesta de los hogares trabajaremos con la variable que cuenta el número de entrevistas realizadas en todo el período considerado. Es decir, los hogares que sólo contestaron un trimestre de los cuatro que corresponden tienen baja propensión a responder mientras que los que contestaron las cuatro veces tiene alta propensión a responder. De esta manera, podremos evaluar si existen diferencias sistemáticas

en la probabilidad de responder más o menos veces según el nivel de ingreso y otras variables socioeconómicas.

En los datos se observa que el grupo mayoritario de casos corresponde al de hogares que respondieron 4 veces, es decir, aquellos que completaron el esquema de la encuesta. Dado que además este es el escenario ideal de respuesta al que buscan llegar los paneles de encuestas, se trabajará con una indicadora de si completó o no el total de entrevistas para plantear el modelo de probabilidad. Este modelo tendrá el objetivo final de equilibrar las estructuras de respuesta de los aglomerados y corregir el posible sesgo por no respuesta mediante una reponderación de los datos que mejore la representatividad de los casos con baja propensión a responder.

5. Metodología

En una primera etapa, se realizará un análisis descriptivo de la estructura de la muestra para los cuatro aglomerados urbanos que representan al nordeste argentino en la encuesta (Gran Resistencia, Corrientes, Formosa y Posadas). Esto permitirá observar, a priori, si existen diferencias importantes en el número de entrevistas realizadas por hogar que pueda causar sesgos por no respuesta.

En una segunda etapa, se trabajará un modelo que explique el ingreso per cápita a precios constantes de los hogares en base a distintas características socioeconómicas con el objetivo de testear si, controlando por estas variables, el número de repeticiones tiene una relación significativa con el ingreso. Dada la estructura de panel de los datos se implementará un modelo mixto con ordenada aleatoria para controlar la heterogeneidad de los hogares.

Siguiendo a [1] y [2], en una tercera etapa se ajustará un modelo lineal generalizado multinomial de respuesta politómica ordinal que explique el número de entrevistas realizadas por hogar (del 1 al 4) para testear la aleatoriedad teórica que esta variable debería tener respecto a características socioeconómicas de interés. En caso de identificar la presencia de sesgo porque la decisión de responder depende de las variables de interés, pasamos a la siguiente etapa.

Con el objetivo de predecir la probabilidad de respuesta de los hogares basado en la propensión que tengan a completar el esquema de la encuesta, se compararán modelos alternativos para PSA como ser el logístico, DecisionTree, Random Forest y XGBoost seleccionando el de mejor desempeño para esta aplicación en particular [7]. Las probabilidades predichas por estos modelos se implementarán en la corrección de los factores de expansión de la encuesta.

Por último, para analizar el efecto de la no respuesta sobre el ingreso per cápita familiar y el nivel de pobreza por aglomerado urbano, se compararán las distribuciones de ingresos pre y post corrección. Lo cual implicará el análisis principalmente de las colas de la distribución que representan los percentiles de más bajos y mayores ingresos [1]. En este último punto del trabajo, se estimarán con las distribuciones acumuladas empíricas de ingresos familiares modificando los factores de expansión.

6. Resultados esperados

Se espera encontrar que la propensión a responder por parte de los hogares no se comporte de manera aleatoria respecto al ingreso y otras variables socioeconómicas. Es decir, que el modelo multinomial con las categorías basadas en el conteo de entrevistas realizadas ajuste una relación significativa con el ingreso per cápita familiar. Esto confirmaría la presencia de potencial sesgo por no respuesta debido a que la decisión de responder más o menos veces depende sistemáticamente de nuestra variable de interés.

Habiendo determinado variables influyentes sobre la propensión a responder, se espera lograr un modelo de buen desempeño que prediga la probabilidad de que el hogar complete el esquema de la encuesta. Esta probabilidad predicha permitirá equilibrar la representatividad de los hogares en la muestra período a período de manera que los de baja propensión a responder estén mejor representados en los datos.

En los microdatos de la encuesta se observa que el caso de Gran Resistencia se encuentra notablemente desbalanceado, con minoría de casos con esquema completo respecto a los demás aglomerados. Dadas estas estructuras de respuesta, se espera que la implementación del método de reponderación modifique en mayor medida la distribución acumulada del ingreso per cápita familiar de este aglomerado y en menor medida las distribuciones de Corrientes, Posadas y Formosa. Por el mismo motivo, se espera que el cambio en el nivel de pobreza monetaria registrado para Gran Resistencia sea en mayor magnitud que en este indicador en los demás aglomerados del NEA post ajuste del factor de expansión.

Este trabajo busca contribuir al estudio de la no respuesta unitaria en encuestas longitudinales de hogares mediante un trabajo empírico de aplicación sobre los hogares del Nordeste Argentino, teniendo en cuenta algunos de los indicadores socioeconómicos que resultan de este relevamiento.

Referencias

[1] A. Korinek, J. A. Mistiaen y M. Ravallion. "An econometric method of correcting for unit nonresponse bias in surveys". En: *Journal of Econometrics* 136 (2007), págs. 213-235. ISSN: 03044076. DOI: 10.1016/j.jeconom.2006.03.001.

[2] J. Bethlehem, F. Cobben y B. Schouten. *Handbook of nonresponse in household surveys*. Wiley, 2011, pág. 474. ISBN: 9780470542798.

[3] T. Krenzke, W. Van de Kerckhove y L. Mohadjer Westat. "Identifying and Reducing Nonresponse Bias throughout the Survey Process". En: *Survey Research Methods* (2005).

[4] F. Butar Butar y C. Chang. "Weighting Methods in Survey Sampling". En: *Survey Research Methods* (2012), págs. 4768-4782.

[5] L. González Allendes. "Propuesta de tratamiento de la no respuesta parcial para la medición de la Pobreza Multidimensional en Chile". Tesis de mtría. Universidad de Chile, 2019.

[6] R. Ferri-García y M. Del Mar Rueda. "Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys". En: *SORT* 42 (2018), págs. 159-182. ISSN: 20138830. DOI: 10.2436/20.8080.02.73.

[7] R. Ferri-García et al. "Estimating response propensities in nonprobability surveys using machine learning weighted models". En: *Mathematics and Computers in Simulation* 225 (2024), págs. 779-793. ISSN: 03784754. DOI: 10.1016/j.matcom.2024.06.012.

[8] B. K. Lee, J. Lessler y E. A. Stuart. "Improving propensity score weighting using machine learning". En: *Statistics in Medicine* 29 (2010), págs. 337-346. ISSN: 02776715. DOI: 10.1002/sim.3782.

[9] D. Westreich, J. Lessler y M. Jonsson Funk. "Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression". En: *Journal of Clinical Epidemiology* 63 (8 2010), págs. 826-833. ISSN: 0895-4356. DOI: 10.1016/J.JCLINEPI.2009.11.020.

[10] R. Ferri-García et al. "Weight smoothing for nonprobability surveys". En: *Test* 31 (2022), págs. 619-643. ISSN: 18638260. DOI: 10.1007/s11749-021-00795-7.

[11] I. Plewis y N. Shlomo. "Using response propensity models to improve the quality of response data in longitudinal studies". En: *Journal of Official Statistics* 33 (3 2017), págs. 753-779. ISSN: 20017367. DOI: 10.1515/JOS-2017-0035.

[12] C. König y J. W. Sakshaug. "Nonresponse trends in establishment panel surveys: findings from the 2001–2017 IAB establishment panel". En: *Journal for Labour Market Research* 57 (1 dic. de 2023). ISSN: 25105027. DOI: 10.1186/s12651-023-00349-4.