Maestría en Estadística Aplicada
Universidad Nacional de Córdoba

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

Documento de trabajo
Avance de resultados

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

Agosto 2024

Índice general

Índice de figuras

Capítulo 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 el 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. Siguiendo la literatura de Propensity Score Adjustment

(PSA) 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 utilicen la probabilidad predicha de respuesta de los hogares planteada como la predisposición a completar el esquema de panel de la encuesta. Posteriormente se pretende analizar las modificaciones sufridas por las distribuciones acumuladas de ingresos frente a los cambios en los factores de expansión de la encuesta.

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. Se pretende extender el ajuste del factor de expansión por probabilidad de respuesta implementado por el Instituto Nacional de Estadística y Censos (INDEC), haciendo especial énfasis en los aglomerados urbanos mencionados y en la propensión a responder de manera completa el esquema de la encuesta como principal diferencia observada entre ellos.

Las consecuencias de las discontinuidades en los paneles de observaciones sobre la medición de indicadores socioeconómicos en aglomerados urbanos pequeños constituye aún una temática poco estudiada a nivel nacional, más allá de los avances metodológicos publicados por el INDEC [3]. Resaltando que aún no se han realizado investigaciones de este tipo a nivel regional por lo cual este trabajo contribuye al área metodológica de medición del ingreso familiar y la pobreza monetaria medido por encuestas, para unidades geográficas pequeñas como es el caso de los aglomerados del NEA.

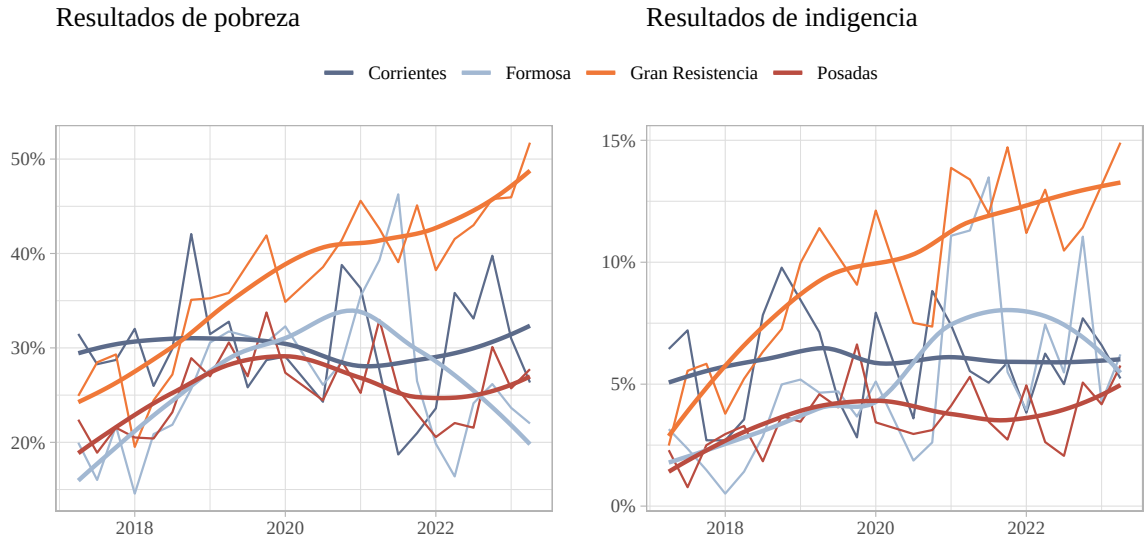


Figura 1.1: Resultados de pobreza e indigencia de los aglomerados urbanos del Nordeste Argentino, período 2017-I Trim a 2023-II Trim.

Capítulo 2

Antecedentes

2.1. Marco teórico

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 [4]. 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 [5].

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 [6]. 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 ella ¿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] [7].

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 los trabajos de [8], [9], [10] que predicen esta probabilidad para incluirla en el factor de expansión de la encuesta. 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 su estructura y 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 [8]. En esta instancia, es importante trabajar con variables de grupos o áreas muestrales que suavicen los ajustes de los pesos de diseño.

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 [11].

2.2. 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 [12] estudia estrategias 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 [13] sobre la encuesta alemana IAB Establishment Panel (IAB-EP) donde analizan tendencias en la tasa de respuesta en un período de 17 años. Describen las dificultades que existen para mantener la motivación de los establecimientos en seguir participando en el panel de entrevistas y cómo esta motivación disminuye luego de la primera entrevista.

En relación a antecedentes empíricos en Argentina, el trabajo de Comari y Hoszowski [14] estudia el efecto de la no respuesta en el panel rotativo de la EPH entre 2005 y 2011 sobre la estimación de la tasa de desempleo a nivel nacional. Encuentra que existe relación entre el número de participaciones (entrevistas realizadas por hogar) y el nivel de la tasa de desempleo. Los hogares con mayor número de participaciones tienen mayor probabilidad de menores tasas de desempleo. Trabajando con indicadores

agregados a nivel nacional, concluyen que la incidencia de sesgo por desvíos en el esquema rotativo es débil aunque puede deberse a mejoras en las participaciones.

Sin embargo, el impacto de esta problemática sobre los resultados de indicadores para aglomerados urbanos pequeños aún es poco estudiado en Argentina más allá de los trabajos realizados por el instituto nacional. El INDEC ha incorporado actualizaciones metodológicas con ajustes por probabilidad de respuesta en su factor de expansión basado en variables como nivel educativo, edad, cantidad de ocupados y desocupados, régimen de tenencia de la vivienda, entre otras variables demográficas y socioeconómicas del hogar [3]. Con respecto a la variable ingreso, la no respuesta fue abordada desde el enfoque de "no respuesta al ítem" con correcciones a niveles de ingresos mediante ajustes a los pesos de diseño asignando a los no respondentes el comportamiento de los respondentes por estratos [15]. En el presente proyecto de investigación se plantea que los efectos sobre el ingreso deberían ser abordados también desde la no respuesta unitaria.

Para el caso argentino, los trabajos de investigación en su mayoría abordan la problemática de la discontinuidad en los paneles de observaciones al momento de evaluar el impacto y resultados de políticas públicas sobre variables socioeconómicas que son medidas en el tiempo para un mismo individuo u hogar. Algunos ejemplos de ellos son los trabajos [16] y [17] que incorporan técnicas de Propensity Score Matching (PSM). En el primer caso, para evaluar la presencia de sesgo de informalidad laboral de un programa social de transferencias a jefes de hogares desempleados y en el segundo caso, para analizar la eficacia del prgrama social Asignación Universal por Hijo (AUH) como amortiguador de la inestabilidad de ingresos de los hogares económicamente vulnerables en Argentina.

Capítulo 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.

Capítulo 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.

Se complementa su uso con el Índice de Precios al Consumidor (IPC) para deflactar los ingresos y 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.

Cuadro 4.1: Variables preliminares 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

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.

Capítulo 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.

Además, 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. Este modelo y el análisis descriptivo anterior constituirán una primera exploración de la relación subyacente entre ingresos familiares y la propensión a responder a la encuesta por parte de los hogares.

Siguiendo a ~ [1] y [2], en una segunda 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 distintas características socioeconómicas. En caso de identificar la presencia de sesgo porque la decisión de responder depende de las variables de interés, pasamos a la siguiente etapa. Se estudiarán distintas especificaciones del modelo respecto al ingreso familiar como nuestra variable de interés principal.

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 [8]. Este modelo tendrá el objetivo final de equilibrar las estructuras de respuesta de los aglomerados y corregir el posible sesgo mediante una reponderación de los datos que mejore la representatividad de los casos con baja propensión a responder.

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 las distribuciones acumuladas empíricas de ingresos familiares modificando los factores de expansión basado en PSA.

5.1. Modelo lineal generalizado multinomial: respuesta categórica ordinal

Siguiendo a Agresti [18] y . . . , los modelos de regresión logística con respuesta categórica binaria pueden extenderse a modelos de regresión con variables de respuesta categórica múltiple, con más de dos categorías. Dependiendo de la característica nominal u ordinal de la variable respuesta, se puede trabajar con: (1) modelos logísticos con una categoría de referencia (baseline-category logit models), o bien (2) modelos logísticos acumulativos. En este trabajo se aplicará el segundo caso para la identificación de potencial sesgo por no respuesta basado.

El modelo multinomial que tiene en cuenta el orden de las categorías es llamado modelo logit acumulativo para respuesta categórica ordinal. Estos modelos estiman los coeficientes de manera tal que se contrastan respecto a las escalas superiores vs inferiores de la variable ordinal, a diferencia del otro tipo de modelo ordinal que compara respecto a una categoría de referencia. Por lo tanto, ajustan una cantidad menor de parámetros y son relativamente más simples de interpretar. Como este

$$P(Y \leq j) = \pi_1 + ... + \pi_j, j = 1, ..., c$$

$$P(Y \leq 1) \leq P(Y \leq 2) \leq ... \leq P(Y \leq c) = 1$$

Los modelos ordinales acumulativos consideran el orden de las categorías modelando las probabilidades acumuladas

$$P(Y \leq r|\mathbf{x}) = \sum_{j=1}^{r} p_j(\mathbf{x})$$

En términos generales, estos modelos asumen la existencia de una variable subyacente continua no observable U tal que:

$$Y = c \iff \theta_{c-1} < U \leq \theta_r,$$

donde $c = 1, ..., J$ y $-\infty = \theta_0 < \theta_1 < ... < \theta_j = \infty$.

Dado el vector de variables explicativas $\mathbf{x}$, se asume que podemos expresar a esta variable de la forma $U = -\mathbf{x}'\boldsymbol{\beta} + \varepsilon$ donde $\boldsymbol{\beta}$ es un vector de parámetros desconocidos y ε es una variable aleatoria con función de distribución acumulada F. Teniendo

Dado que $Y = c \iff \theta_{c-1} < U \leq \theta_c$, $U = -\mathbf{x}'\boldsymbol{\beta} + \varepsilon$ y $\varepsilon \sim F$, resulta:

$$P(Y \leq c|\mathbf{x}) = P(U \leq \theta_c) = P(\varepsilon \leq \theta_c + \mathbf{x}'\boldsymbol{\beta}) = F(\theta_c + \mathbf{x}'\boldsymbol{\beta})$$

y siendo

$$F^{-1}(P(Y \leq c|\mathbf{x})) = \theta_c + \mathbf{x}\boldsymbol{\beta}$$

5.2. Bootstrap

Empleamos bootstrap en la estimación de los errores estándar de

Bootstrap no requiere del cumplimiento de los supuestos del modelo para computar los errores estándar de las estimaciones, por lo cual es más probable de arrojar una estimación más precisa de los errores estándar de los coeficientes del modelo.

5.3. Stepwise y validación cruzada

5.4. Árboles de decisión

5.5. Random Forest

5.6. XGBoost

Capítulo 6

Resultados

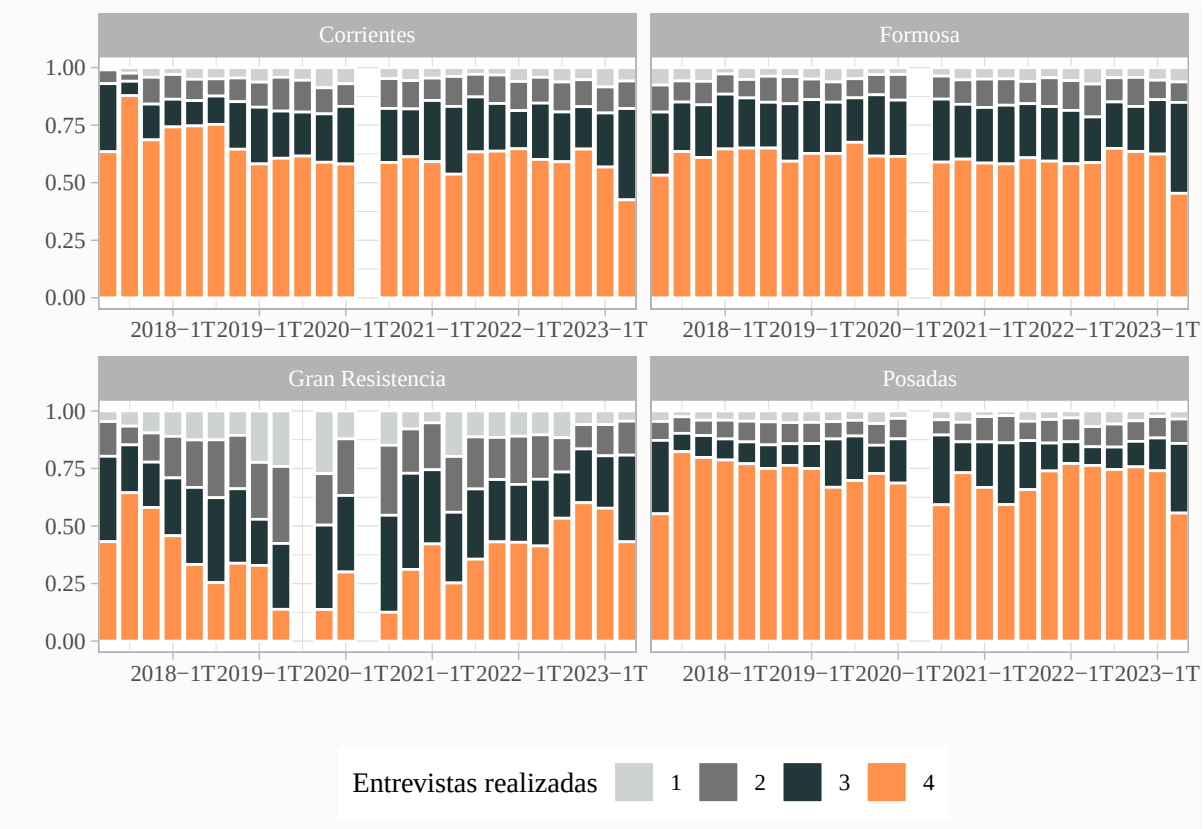
6.1. Caracterización de estructuras de respuesta

Trabajando con la variable de cantidad total de entrevistas realizadas por el hogar en el período, podemos tener una aproximación de cuál es la propensión de los hogares a responder. Analizamos esta propensión por aglomerado urbano teniendoe en cuenta la proporción de hogares en la muestra según el número de entrevistas realizadas (Tabla 1).

Vemos entonces que alrededor del 5 % de los hogares contesta sólo una vez en Corrientes, Posadas y Formosa mientras que para Gran Resistencia este porcentaje sube a casi 14 %. Además, aproximadamente el 10 % de los hogares responde dos veces en estas mismas localidades mientras que en Gran Resistencia tenemos que el 22 % responde sólo dos veces. En el caso de tres encuestas realizadas, las proporciones de hogares en cada categoría son bastantes variadas entre aglomerados. Sin embargo cuando se trata de completar el esquema de la encuesta, es decir, contestar los cuatro trimestres del esquema de rotación de hogares, Gran Resistencia se ubica en aproximadamente 32 % mientras que los demás aglomerados superan el 60 %.

Cuadro 6.1: Proporción de hogares según cantidad total de entrevistas realizadas por aglomerado urbano

	Entrevistas realizadas			
	1	2	3	4
Corrientes	0.1798	0.2099	0.2271	0.2578
Formosa	0.1940	0.2151	0.2729	0.2733
Gran Resistencia	0.4714	0.4012	0.3298	0.1630
Posadas	0.1548	0.1739	0.1702	0.3060



Si pasamos a analizar esta estructura de respuesta en el tiempo, vemos que los aglomerados de Corrientes, Formosa y Posadas mantienen un porcentaje de hogares con esquema completo superior al 60 % en casi todos los casos mientras que Gran Resistencia presenta una estructura muy distinta. La distribución de proporciones de hogares por categoría es bastante cambiante además de que en la mayoría de los casos los hogares con esquema de respuesta completo no superan el 50 %.

6.2. Identificación de sesgo por no respuesta

En primera instancia, planteamos test bivariados de asociación para evaluar si existe relación entre el número de entrevistas realizadas como variable categórica y algunos factores de interés. De las pruebas resulta que se rechaza la hipótesis nula de independencia entre el número de entrevistas realizadas y que el hogar sea o no pobre, el nivel educativo del jefe de hogar, que el jefe sea casado o unido, el estado de ocupación y si la vivienda es casa o departamento. Este resultado nos da un primer indicio de que la propensión a responder no se comporta de manera aleatoria en relación a estas variables medidas por la encuesta.

Cuadro 6.2: Resultados de pruebas chi-cuadrado de asociación de variables.

Variables	Estadístico	df	p.value
NRO_REP ~ hogar_pobre	22.06904	3	6.31e-05
NRO_REP ~ NIVEL_ED	750.32527	12	0.00e+00
NRO_REP ~ casadounido	94.11168	3	0.00e+00
NRO_REP ~ ESTADO	152.65589	6	0.00e+00
NRO_REP ~ casadpto	178.58000	3	0.00e+00

En segunda instancia y con el objetivo de detectar el sesgo por no respuesta, se plantea un modelo lineal generalizado multinomial de respuesta politómica ordinal que busca testear relaciones estadísticamente significativas entre el número de entrevistas realizadas (del 1 al 4) y el ingreso además de otras variables de interés socioeconómico.

Del modelo ajustado resulta que la chance de contestar pocas veces en vez de varias veces o completar el esquema de entrevistas aumenta en un 0.265 % a medida que el ingreso per cápita familiar en términos reales aumenta en un 1 % en Gran Resistencia. En Corrientes, Formosa y Posadas este efecto es 0.188 %, 0.273 % y 0.262 % respectivamente. Por otro lado, la chance de contestar pocas veces en vez de varias o bien completar el esquema aumenta en 0.074 % a medida que el hogar promedio suma un miembro adicional.

Cuadro 6.3: Modelos para cantidad de entrevistas realizadas por aglomerado urbano.

	Gran Rcia		Corrientes		Formosa		Posadas	
	OR	p	OR	p	OR	p	OR	p
(Intercept) × 1	0.525	0.134	0.395*	0.010	0.143***	<0.001	0.000	0.971
(Intercept) × 2	1.856	0.152	1.532	0.235	0.519+	0.074	0.000	0.974
(Intercept) × 3	6.637***	<0.001	5.240***	<0.001	1.972+	0.065	0.000	0.976
logIPCF_d	**1.302*** **	**<0.001**	**1.178*** **	**<0.001**	**1.270*** **	**<0.001**	**1.264*** **	**<0.001**
CH06	0.944***	<0.001	0.960***	<0.001	0.969***	<0.001	0.885***	<0.001
I(CH06^2)	1.000***	<0.001	1.000	0.110	1.000	0.503	1.001***	<0.001
IX_TOT	0.949***	<0.001	0.852***	<0.001	0.872***	<0.001	0.861***	<0.001
casadpto	0.266***	<0.001	0.293***	<0.001	0.322***	<0.001	146657.850	0.973
casadounido	0.887**	0.005	0.876**	0.010	0.862**	0.002	0.755***	<0.001
Num.Obs.	8777		8349		9082		8652	
AIC	22658.0		16331.5		18125.3		14305.8	
BIC	22721.7		16394.8		18189.3		14369.4	
RMSE	2.52		3.47		3.54		4.13	

Nota:
+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Resultados bootstrap de los coeficientes del modelo anterior

En tercera instancia, como proceso exploratorio del sesgo por no respuesta, ajustamos modelos para explicar el ingreso per cápita familiar en términos reales (deflactado). Dado que el ingreso, y posteriormente el nivel de pobreza, son nuestras principales variables de interés en este estudio analizamos si controlando esta variable por factores de influencia puede captarse una dependencia de las entrevistas realizadas.

Se presentan nuevamente cuatro modelos, uno para cada aglomerado, y se incluyen como variables de control a las horas trabajadas, el nivel educativo del jefe de hogar, su categoría y calificación ocupacional, otros ingresos no laborales, una binaria de sexo (mujer=1, hombre=0) y una categórica del año para controlar por posibles dinámicas en el tiempo. Por último, se incorpora la variable NRO_REP que mide el número de entrevistas realizadas.

Vemos que el número total de entrevistas realizadas en el período analizado tiene un coeficiente significativo sobre el nivel de ingreso, que nos indica que a medida que aumentan las entrevistas realizadas por hogar se reduce el nivel de ingreso. (El coeficiente indica relación negativa controlando por lo demás, pero ver si es correcto presentar esto sino quedarnos con el mlg multinomial justificado en la teoría dada la relación de causalidad. Las entrevistas realizadas no influyen en el nivel de ingreso sino viceversa).

Cuadro 6.4: Resultados de simulación por bootstrap para coeficientes de los modelos multinomiales, por aglomerado urbano.

	Original	Bootstrap	IC 95 % (percentiles)
Gran Resistencia			
logIPCF_d	1.302	1.343	[1.223, 1.385]
CH06	0.944	0.950	[0.930, 0.957]
I(CH06^2)	1.000	1.000	[1.000, 1.000]
IX_TOT	0.949	0.962	[0.924, 0.973]
casadpto	0.266	0.336	[0.157, 0.400]
casadounido	0.887	0.925	[0.817, 0.968]
Corrientes			
logIPCF_d	1.178	1.220	[1.103, 1.261]
CH06	0.960	0.968	[0.943, 0.976]
I(CH06^2)	1.000	1.000	[1.000, 1.000]
IX_TOT	0.852	0.867	[0.822, 0.883]
casadpto	0.293	0.348	[0.207, 0.410]
casadounido	0.876	0.924	[0.789, 0.974]
Formosa			
logIPCF_d	1.270	1.314	[1.191, 1.364]
CH06	0.969	0.978	[0.954, 0.986]
I(CH06^2)	1.000	1.000	[1.000, 1.000]
IX_TOT	0.872	0.887	[0.843, 0.902]
casadpto	0.322	0.385	[0.226, 0.455]
casadounido	0.862	0.903	[0.786, 0.943]
Posadas			
logIPCF_d	1.264	1.317	[1.167, 1.372]
CH06	0.885	0.892	[0.869, 0.899]
I(CH06^2)	1.001	1.001	[1.001, 1.001]
IX_TOT	0.861	0.880	[0.825, 0.899]
casadounido	0.755	0.800	[0.669, 0.850]

Cuadro 6.5: Modelos para logaritmo del ingreso per cápita familiar deflactado por aglomerado urbano.

	Gran Rcia	Corrientes	Formosa	Posadas
(Intercept)	7.767***	7.664***	8.304***	7.907***
	[7.513, 8.020]	[7.413, 7.916]	[8.038, 8.569]	[7.676, 8.137]
horas_trab	0.003***	0.003***	0.003***	0.002***
	[0.002, 0.004]	[0.003, 0.004]	[0.002, 0.005]	[0.001, 0.003]
NIVEL_EDSecundario completo	0.206***	0.217***	0.140***	0.275***
	[0.142, 0.270]	[0.153, 0.281]	[0.080, 0.201]	[0.217, 0.334]
NIVEL_EDSecundario incompleto	-0.001	0.066+	-0.059+	0.105***
	[-0.070, 0.069]	[-0.001, 0.133]	[-0.125, 0.006]	[0.047, 0.164]
NIVEL_EDUniversitario completo	0.495***	0.465***	0.444***	0.653***
	[0.414, 0.576]	[0.388, 0.542]	[0.362, 0.527]	[0.582, 0.724]
NIVEL_EDUniversitario incompleto	0.366***	0.380***	0.211***	0.414***
	[0.280, 0.453]	[0.300, 0.459]	[0.121, 0.301]	[0.342, 0.485]
CAT_OCUP	0.187***	0.114***	0.092***	0.149***
	[0.140, 0.234]	[0.076, 0.152]	[0.053, 0.131]	[0.113, 0.186]
CALIFICACIONNo calificados	-0.271**	-0.050	-0.259*	-0.273**
	[-0.474, -0.068]	[-0.250, 0.150]	[-0.475, -0.042]	[-0.460, -0.086]
CALIFICACIONOperativos	-0.027	0.097	-0.066	-0.096
	[-0.225, 0.171]	[-0.097, 0.291]	[-0.279, 0.146]	[-0.277, 0.086]
CALIFICACIONProfesionales	0.431***	0.495***	0.403***	0.268**
	[0.211, 0.652]	[0.292, 0.698]	[0.174, 0.633]	[0.074, 0.461]
CALIFICACIONTécnicos	0.107	0.224*	0.052	0.037
	[-0.095, 0.310]	[0.028, 0.421]	[-0.164, 0.269]	[-0.148, 0.222]
otros_ing_nolab	0.000***	0.000***	0.000***	0.000***
	[0.000, 0.000]	[0.000, 0.000]	[0.000, 0.000]	[0.000, 0.000]
mujer	-0.126***	-0.119***	-0.132***	-0.103***
	[-0.179, -0.073]	[-0.174, -0.064]	[-0.187, -0.077]	[-0.155, -0.052]

6.3. Métodos de reponderación de la muestra

Según lo hallado en la sección anterior, la distribución del ingreso per cápita familiar puede verse sesgada y no reflejar la distribución real. Por lo tanto, para poder analizar el nivel de ingreso de los hogares y, por ende, el nivel de pobreza monetaria por aglomerado, debemos formular alguna corrección del sesgo por no respuesta. La idea es implementar un método de reponderación de la muestra basado en la probabilidad de respuesta de los hogares, corrigiendo el factor de expansión de la encuesta PONDIH.

Actualmente y siguiendo la literatura de *response propensity modeling* que proponen modelos alternativos para predecir la probabilidad de respuesta, se está trabajando con un modelo logístico, un árbol de decisión y un bosque aleatorio. La idea sería comparar medidas de la clasificación para quedarnos con el mejor modelo que prediga la probabilidad de que el hogar complete el esquema de la encuesta. Se busca predecir una binaria indicadora de la categoría 4 en número de entrevistas realizadas, dado que son las diferencias más marcadas que observamos entre aglomerados y refleja la propensión a responder que tienen los hogares.

Una vez seleccionado el mejor modelo, se corrigen los factores de expansión. Por el momento se está implementando la metodología más tradicional de dividir el ponderador por la probabilidad de respuesta, y reescalando para mantener la representación de viviendas por área muestral.

Corregidos los ponderadores, se pretende comparar las distribuciones del ingreso per cápita familiar acumulado para ver posibles cambios pre y post corrección por no respuesta y balanceo de la muestra. Pasando posteriormente a analizar los cambios en el nivel de pobreza monetaria que depende del ingreso.

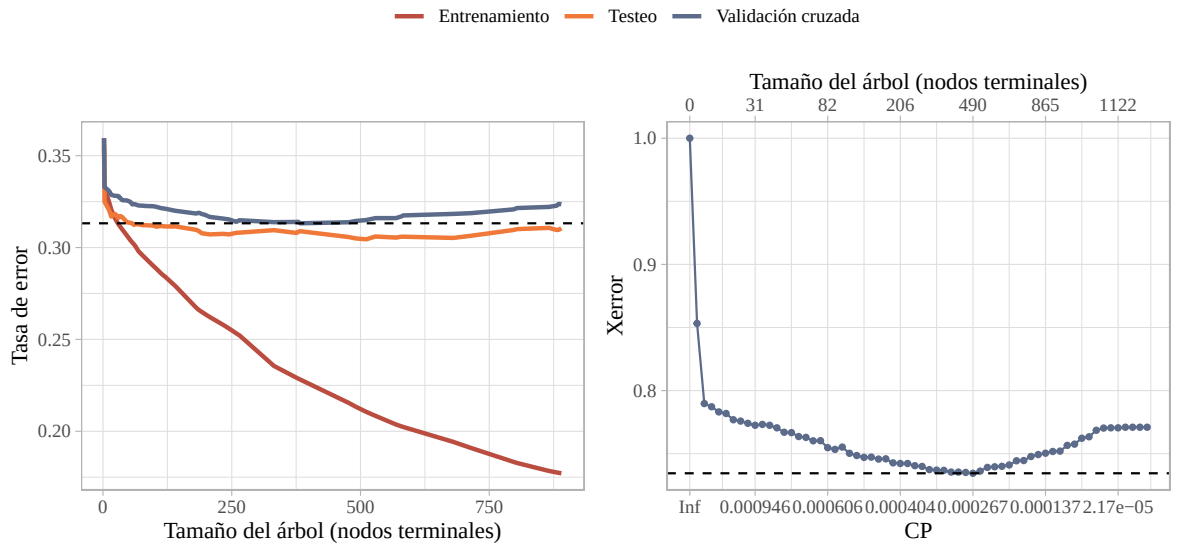


Figura 6.1: Tasa de error de testeo, de entrenamiento y por validación cruzada según cantidad de nodos terminales (izquierda) junto con tasa de error relativa por validación cruzada según relación costo-complejidad del árbol.

6.3.1. Modelos con suavizado de ponderadores

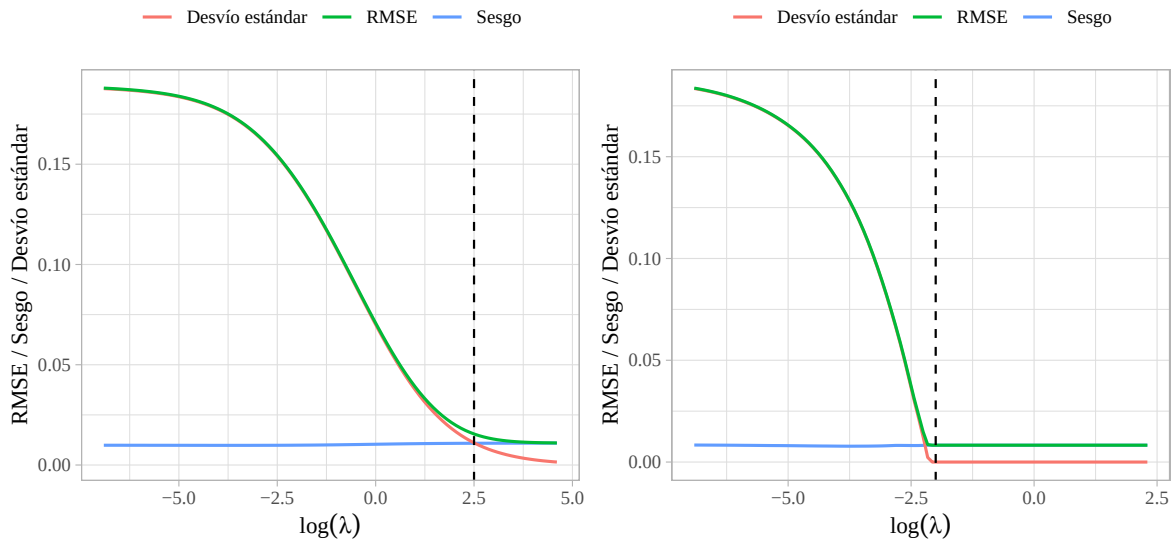


Figura 6.2: Sesgo, desvío estándar y raíz del error cuadrático medio para regresión Ridge (izquierda) y lasso (derecha) estimados mediante validación cruzada.

Capítulo 7

Tareas pendientes/dudas por resolver

- Probar algún método como stepwise que funcione con los modelos multinomiales para intentar mejorar la bondad de ajuste del modelo. Igualmente este quedaría como modelo más bien descriptivo para la interpretación de los OR.
- Como se está trabajando con una muestra, ver alguna manera de robustecer los resultados con boostrap por ejemplo (si es apropiado). Para el modelo multinomial principalmente, vi algunos papers que lo hacen así.
- En principio probar la corrección del PONDIH con el método propuesto (dividiéndolo por la probabilidad de que el hogar complete el esquema de la encuesta y reescalando para mantener los pesos de las áreas muestrales), pero hay otros cálculos alternativos que usan la probabilidad de respuesta también. Investigar un poco esto.
- Ver métodos de optimización de hiperparámetros y validación cruzada para los modelos de predicción de la probabilidad de respuesta completa. Incluir esto previo al análisis de medidas de bondad de ajuste de la clasificación para mejorar los modelos.
- La idea previamente a ver resultados sobre pobreza es ver cómo se modifica la distribución del ingreso per cápita familiar. Principalmente la distribución acumulada para analizar las colas (percentiles vs ingreso acumulado), hasta ahora va quedando que la distribución corregida queda por debajo de la original lo que indicaría que los percentiles de más bajos ingresos (influyentes para el nivel de pobreza) acumulan más ingreso. Vengo usando la función wtd.Ecdf() que calcula la distribución empírica ponderada pero no sé si es muy apropiada porque las curvas no quedan suavizadas, al contrario parecen sobreajustadas. Probé usando GMM (Método generalizado de momentos) para una gamma o normal inversa pero no estoy segura si ese método admite ponderaciones.

Bibliografía

[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] Instituto Nacional de Estadística y Censos. *Encuesta Permanente de Hogares: Consideraciones metodológicas sobre el tratamiento de la información del segundo trimestre de 2020*. 2020.

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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 2023). ISSN: 25105027. DOI: 10.1186/s12651-023-00349-4.

[14] C. Comari y A. E. Hoszowski. "Non response in rotating panel surveys: analysis on Argentina's labor force survey". En: *Joint Statistical Meetings*. 2014. ISBN: 2.1.1579.0960. DOI: 10.13140/RG.2.1.1579.0960.

[15] Instituto Nacional de Estadística y Censos. *Encuesta Permanente de Hogares: Diseño de registro y estructura para las bases preliminares Hogar y Personas*. 2024.

[16] L. Gasparini, F. Haimovich y S. Olivieri. "Labor informality bias of a poverty-alleviation program in Argentina". En: *Journal of Applied Economics* 12 (2 2009), págs. 181-205. ISSN: 1514-0326. DOI: 10.1016/S1514-0326(09)60012-X.

[17] Sebastien Carrere. "How effective are cash transfer programs in mitigating income instability? Evidence from the AUH in Argentina". 2024. URL: https://hal.science/hal-04525248v3.

[18] Alan Agresti. *Wiley Series in Probability and Statistics*. Ed. por John Wiley y Sons. 3.ª ed. 2019. ISBN: 9781119405283. URL: http://www.wiley.com/go/wsps.