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
Primer avance de resultados

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

Agosto 2024

Índice

1. Introducción

2. Antecedentes

3. Problema y objetivos

4. Metodología

5. Resultados

Se presentan resultados preliminares del trabajo.

5.1. 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 1: Proporción de hogares según cantidad total de entrevistas realizadas por aglomerado urbano

	Entrevistas realizadas			
	1	2	3	4
Corrientes	0.0501	0.1153	0.2183	0.6163
Formosa	0.0479	0.1064	0.2400	0.6056
Gran Resistencia	0.1386	0.2205	0.3153	0.3256
Posadas	0.0423	0.0894	0.1554	0.7128

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 %.

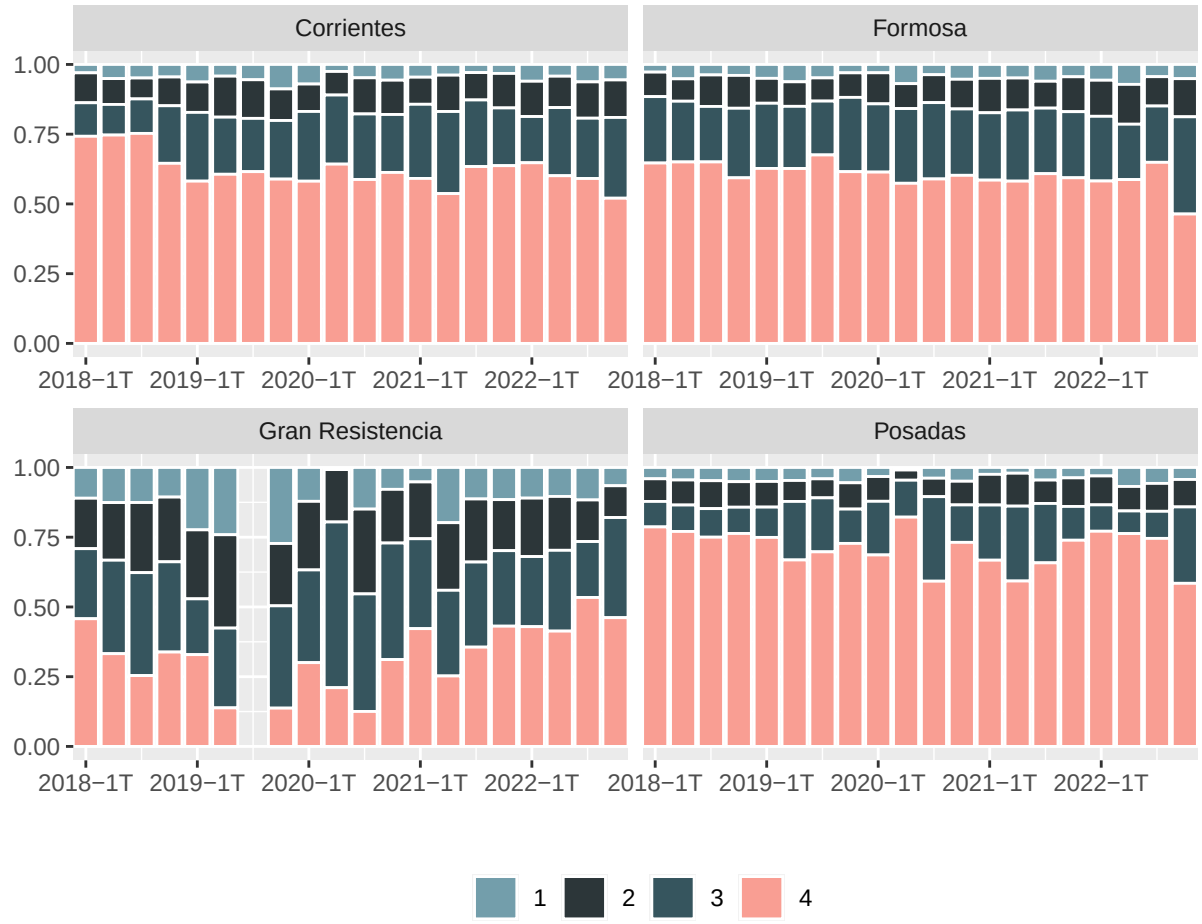

Figura 1: Estructura de respuesta por aglomerado urbano del NEA, período 2018-2022.

5.2. 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 2: Resultados de pruebas chi-cuadrado de asociación de variables.

Variables	Estadístico	df	p.value
NRO_REP ~ hogar_pobre	22.71314	3	4.63e-05
NRO_REP ~ NIVEL_ED	609.38911	12	0.00e+00
NRO_REP ~ casadounido	79.67789	3	0.00e+00
NRO_REP ~ ESTADO	149.61655	6	0.00e+00
NRO_REP ~ casadpto	128.48034	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 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	0.674	0.420	0.400*	0.019	0.128***	<0.001	2.468	0.626
(Intercept) × 2	2.451+	0.066	1.603	0.225	0.484+	0.070	9.253	0.230
(Intercept) × 3	9.463***	<0.001	5.491***	<0.001	1.851	0.124	26.991+	0.075
logIPCF_d	**1.265*** **	**<0.001**	**1.188*** **	**<0.001**	**1.273*** **	**<0.001**	**1.262*** **	**<0.001**
CH06	0.942***	<0.001	0.954***	<0.001	0.973**	0.002	0.885***	<0.001
I(CH06^2)	1.000***	<0.001	1.000*	0.037	1.000	0.645	1.001***	<0.001
IX_TOT	0.931***	<0.001	0.835***	<0.001	0.848***	<0.001	0.819***	<0.001
casadpto	0.312**	0.003	0.321***	<0.001	0.295***	<0.001	0.157	0.307
Num.Obs.	7129		7083		7536		7013	
AIC	18744.5		13962.1		15050.3		11641.3	
BIC	18799.5		14017.0		15105.7		11696.1	
RMSE	2.48		3.52		3.62		4.10	

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

	Gran Rcia	Corrientes	Formosa	Posadas
(Intercept)	8.031***	7.755***	8.397***	8.106***
	[7.800, 8.263]	[7.504, 8.006]	[8.135, 8.659]	[7.876, 8.336]
horas_trab	0.004***	0.003***	0.004***	0.002***
	[0.003, 0.005]	[0.002, 0.004]	[0.002, 0.005]	[0.001, 0.003]
NIVEL_EDSecundario completo	0.186***	0.234***	0.139***	0.268***
	[0.125, 0.247]	[0.173, 0.296]	[0.082, 0.197]	[0.214, 0.323]
NIVEL_EDSecundario incompleto	-0.022	0.043	-0.038	0.078**
	[-0.090, 0.047]	[-0.025, 0.112]	[-0.104, 0.027]	[0.020, 0.137]
NIVEL_EDUniversitario completo	0.503***	0.538***	0.460***	0.664***
	[0.424, 0.582]	[0.463, 0.612]	[0.381, 0.538]	[0.597, 0.730]
NIVEL_EDUniversitario incompleto	0.365***	0.431***	0.213***	0.400***
	[0.280, 0.449]	[0.354, 0.508]	[0.126, 0.299]	[0.333, 0.468]
CAT_OCUPObrero o empleado	0.310***	0.188***	0.170***	0.306***
	[0.258, 0.361]	[0.140, 0.237]	[0.119, 0.221]	[0.262, 0.350]
CAT_OCUPPatrón	0.236**	0.098*	0.077	0.091+
	[0.081, 0.390]	[0.003, 0.194]	[-0.024, 0.177]	[-0.005, 0.187]
CAT_OCUPTrabajo fliar sin rem	-0.083	0.034	0.290	-0.162
	[-0.521, 0.355]	[-0.447, 0.514]	[-0.917, 1.497]	[-0.712, 0.388]
CALIFICACIONNo calificados	-0.329**	-0.043	-0.297*	-0.351**
	[-0.544, -0.114]	[-0.266, 0.181]	[-0.534, -0.061]	[-0.564, -0.137]
CALIFICACIONOperativos	-0.019	0.142	-0.038	-0.066
	[-0.229, 0.192]	[-0.076, 0.360]	[-0.271, 0.195]	[-0.275, 0.142]
CALIFICACIONProfesionales	0.403***	0.566***	0.426***	0.338**
	[0.171, 0.635]	[0.340, 0.792]	[0.179, 0.673]	[0.119, 0.557]
CALIFICACIONTécnicos	0.063	0.219+	0.049	0.033
	[-0.153, 0.280]	[-0.001, 0.440]	[-0.187, 0.285]	[-0.179, 0.245]
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	-0.097***	-0.100***	-0.096***	-0.091***
	[-0.144, -0.049]	[-0.148, -0.053]	[-0.144, -0.047]	[-0.137, -0.046]
anio2019	-0.333***	0.105***	-0.205***	-0.156***
	[-0.392, -0.273]	[0.045, 0.164]	[-0.260, -0.150]	[-0.203, -0.110]
anio2020	-0.340***	0.033	-0.272***	-0.165***
	[-0.405, -0.275]	[-0.032, 0.097]	[-0.332, -0.211]	[-0.219, -0.112]
anio2021	-0.457***	0.045	-0.432***	-0.175***
	[-0.522, -0.392]	[-0.019, 0.110]	[-0.494, -0.369]	[-0.229, -0.121]
anio2022	-0.419***	-0.022	-0.317***	-0.201***
	[-0.482, -0.355]	[-0.089, 0.045]	[-0.381, -0.253]	[-0.257, -0.145]
nro_rep	**-0.049*** **	**-0.069*** **	**-0.073*** **	**-0.046*** **
	[-0.070, -0.028]	**[-0.093, -0.046]**	**[-0.097, -0.050]**	**[-0.068, -0.023]**
Num.Obs.	4158	4359	3601	4869
R2	0.713	0.755	0.794	0.793
R2 Adj.	0.711	0.754	0.793	0.792
AIC	5952.3	5095.6	3488.2	4943.7
BIC	6085.3	5229.5	3618.2	5080.0
RMSE	0.49	0.43	0.39	0.40

5.3. 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.

6. 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.