Aplicación de modelos categóricos para la estimación del nivel de escolaridad de los jefes de hogar en las regiones de Bío-Bío y Metropolitana usando la encuesta CASEN 2022

IECD 423: Análisis de datos categóricos

Bastián Barraza - Javiera Contador - Pablo Estay - Hilda Núñez - Emilia Sepúlveda

Universidad de Valparaíso

11 de Dicicembre

Índice

Introducción

Problemática

- Los problemas en la escolaridad en Chile afectan la calidad de la educación y el futuro de las nuevas generaciones.
- Según la encuesta CASEN 2017, aproximadamente 5 millones de adultos no habían completado sus 12 años de escolaridad, con 500,000 analfabetos.
- La deserción escolar se ha agravado tras la pandemia y está vinculada a la violencia y el delito en las instituciones educativas.

Introducción

Características de la investigación

- La Encuesta CASEN, realizada por el Ministerio de Desarrollo Social y Familia de Chile, busca conocer la situación de hogares y población, especialmente aquellos en pobreza y grupos prioritarios.
- La investigación propone aplicar modelos de regresión binomial, multinomial ordinal, multinomial nominal, regresión Poisson y regresión binomial negativa para entender las variables que influyen en la escolaridad de jefes de hogar en el Bío-Bío y la Metropolitana.
- Los datos provienen de la Encuesta CASEN 2022 y los resultados se obtienen con el software Stata 18.

Índice

Objetivos

Objetivo General: Aplicar modelos de regresión para la estimación del nivel de escolaridad de los jefes de hogar en las regiones de Bío-Bío y Metropolitana usando la encuesta CASEN 2022.

Objetivos específicos:

- Realizar un análisis descriptivo de las variables de interés.
- Desarrollar modelos de regresión, precisamente regresión binomial, regresión multinomial ordinal y nominal, regresión Poisson y regresión binomial negativa.
- Describir los resultados obtenidos de cada modelo de regresión aplicados a la Encuesta Casen.

Índice

Regresión Logística

Definición de regresión Logística

El modelo de regresión logística utiliza la función logística (también conocida como la función sigmoide) para transformar una combinación lineal de variables predictoras en una probabilidad que se encuentra en el rango de 0 a 1.

Modelo de regresión

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)}}$$

Donde:

- $P(Y = 1)$ es la probabilidad de que la variable dependiente Y sea igual a 1.
- e es la base del logaritmo natural.
- $\beta_0, \beta_1, \ldots, \beta_k$ son los coeficientes del modelo.
- $X_1, X_2, \ldots, X_k$ son las variables predictoras.

Regresión Logística

Pregunta de investigación

¿Cómo influyen el nivel socioeconómico, el sexo, la región y la edad en la probabilidad de que los jefes de hogar tengan más de 10 años de escolaridad?

Hipótesis

El nivel socioeconómico, el sexo, la región y la edad desempeñan roles significativos en la determinación de la duración de la escolaridad de los jefes de hogar en las regiones de Biobío y Santiago, Chile.

Regresión Logística

Variable	**Odds Ratio**	**P**$>\|t\|$
Region		
Región del Biobío	1 (base)	
Región Metropolitana de Santiago	1.266	0.000
Edad	0.938	0.000
Nivel Socio Económico		
Bajo	1 (base)	
Medio	2.421	0.000
Alto	12.359	0.000
Bajo-medio	2.060	0.000
Bajo-alto	0.396	0.165
Bajo-medio-alto	3.125	0.000
Medio-alto	5.379	0.000
Sexo		
Hombre	1 (base)	
Mujer	0.854	0.000
Constante	25.597	0.000

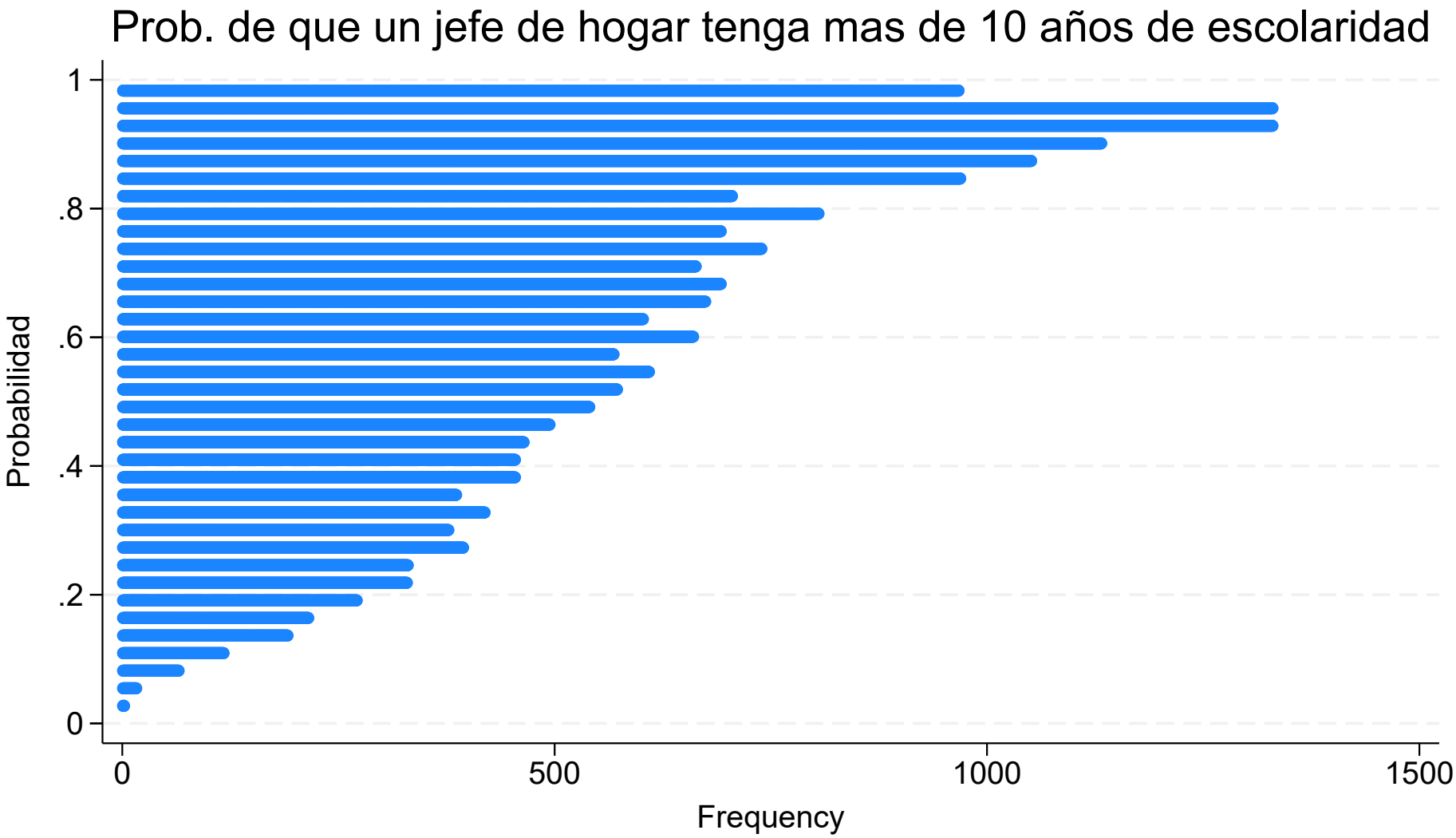


Figura: Diagrama de frecuencias de las estimaciones del modelo logístico.

Regresión Logística

Análisis de residuos

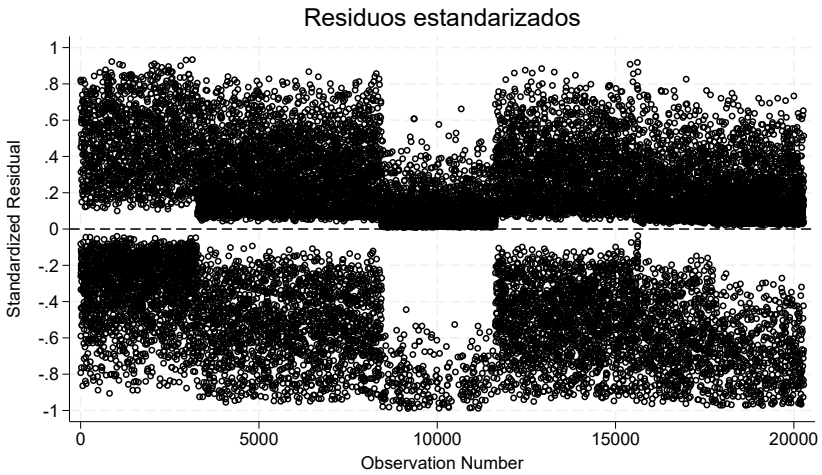


Figura: Diagrama de dispersión de residuos.

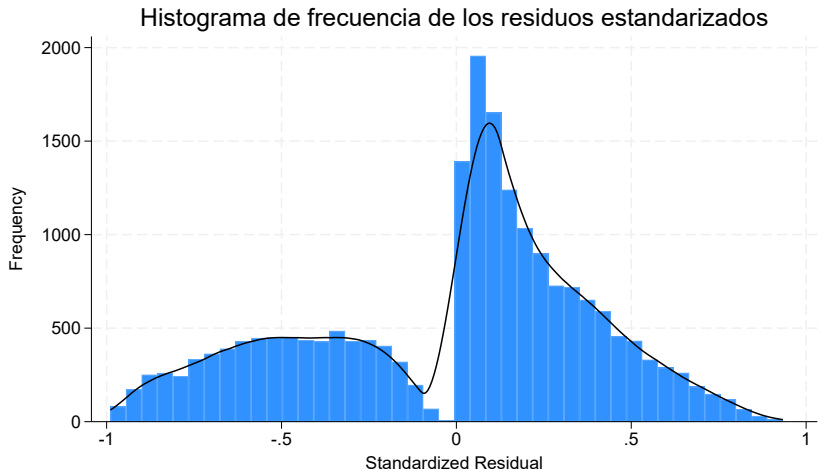


Figura: Densidad de kernel de residuos.

Proporciones marginales estimadas de la variable Región

Variable	Margin	P$>\|t\|$
Región		
Región del Biobío	0.710	0.000
Región Metropolitana de Santiago	0.745	0.000

Proporciones marginales estimadas de la variable Sexo

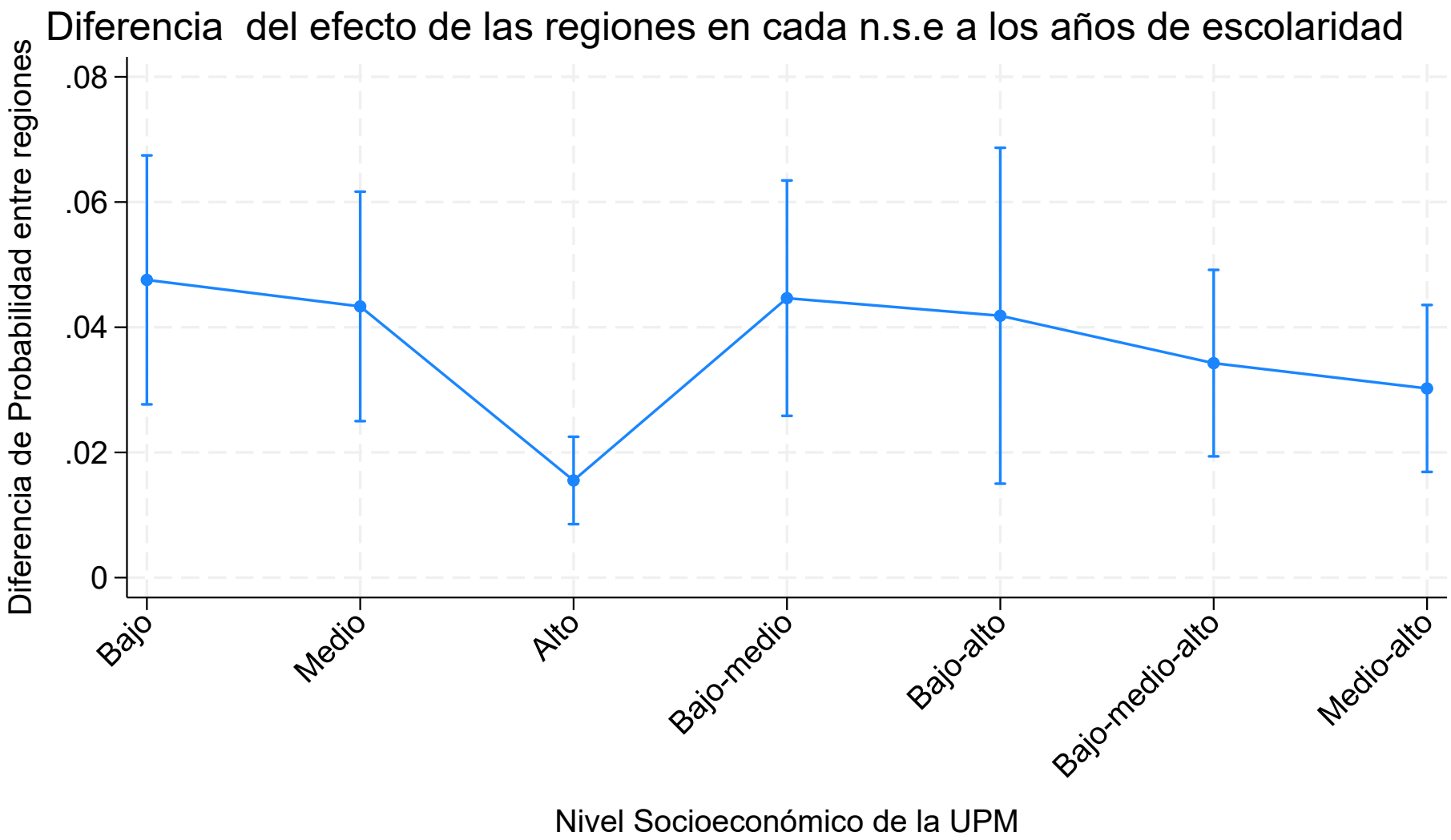
Variable	Margin	P$>\|t\|$
Sexo		
Hombre	0.749	0.000
Mujer	0.727	0.000

Proporciones marginales estimadas de la variable Sexo

Variable	**Margin**	**P**$>\|t\|$
Nivel Socio-Económico		
Bajo	0.535	0.000
Medio	0.701	0.000
Alto	0.904	0.000
Bajo-medio	0.673	0.000
Bajo-alto	0.351	0.005
Bajo-medio-alto	0.743	0.000
Medio-alto	0.820	0.000

Regresión Logística

Diferencia de la probabilidad de tener mas de 10 años de escolaridad entre regiones de entre regiones de cada nivel socioeconómico



Regresión multinomial ordinal

El modelo multinomial ordinal es una extensión del modelo de regresión logística ordinal, que se utiliza cuando la variable dependiente es ordinal, es decir, tiene un orden inherente pero las distancias entre las categorías no son necesariamente iguales. Este modelo es adecuado cuando se trata de predecir una variable categórica con tres o más niveles ordenados.

Este modelo acumula las probabilidades de las categorías anteriores y no utiliza la ultima categoría como referencia pues la probabilidad acumulada es igual a 1.

Regresión multinomial ordinal

Pregunta de investigación

¿Como es la variación las probabilidades de pertenencia dado los niveles de las variable explicativa nse,sexo, región y edad dado la variable respuesta esc(años de escolaridad), para los jefes de hogar que residen en la región del Bio-Bío y Metropolitana?

Hipótesis

Existe variación importante en las probabilidades de pertenencia dado diferentes años de escolaridad para valores específicos de las variables explicativas.

Regresión multinomial ordinal

Supongamos que tenemos J categorías ordenadas (o niveles) en nuestra variable dependiente, y denotamos las probabilidades acumulativas de pertenecer a la categoría j o inferior como $P(Y \leq j)$ para $j = 1, 2, \ldots, J-1$. La función logística ordinal está dada por la siguiente expresión:

$$P(Y \leq j) = F(\alpha_j - \beta_1 x_1 - \beta_2 x_2 - \ldots - \beta_p x_p)$$

donde:

- $F()$ es la función logística acumulativa.
- α_j es un parámetro de umbral
- asociado con la categoría j. $\beta_1, \beta_2, \ldots, \beta_p$ son los coeficientes del modelo asociados con las variables predictoras $x_1, x_2, \ldots, x_p$.

Regresión multinomial ordinal

Esc	Coefficient	$P > \|t\|$
region		
Región Metropolitana	.0534362	0.275
edad	-.0478688	0.000
yoprcor	$1{,}02e-06$	0.000
nse		
Medio	.7665288	0.000
Alto	1.849287	0.000
Bajo-medio	.6022216	0.000
Bajo-alto	-.7877184	0.350
Bajo-medio-alto	1.281207	0.000
Medio-alto	1.566959	0.000
sexo		
2. Mujer	.2926723	0.000

muestra la probabilidad prevista para cada uno de los valores de la variable especificada

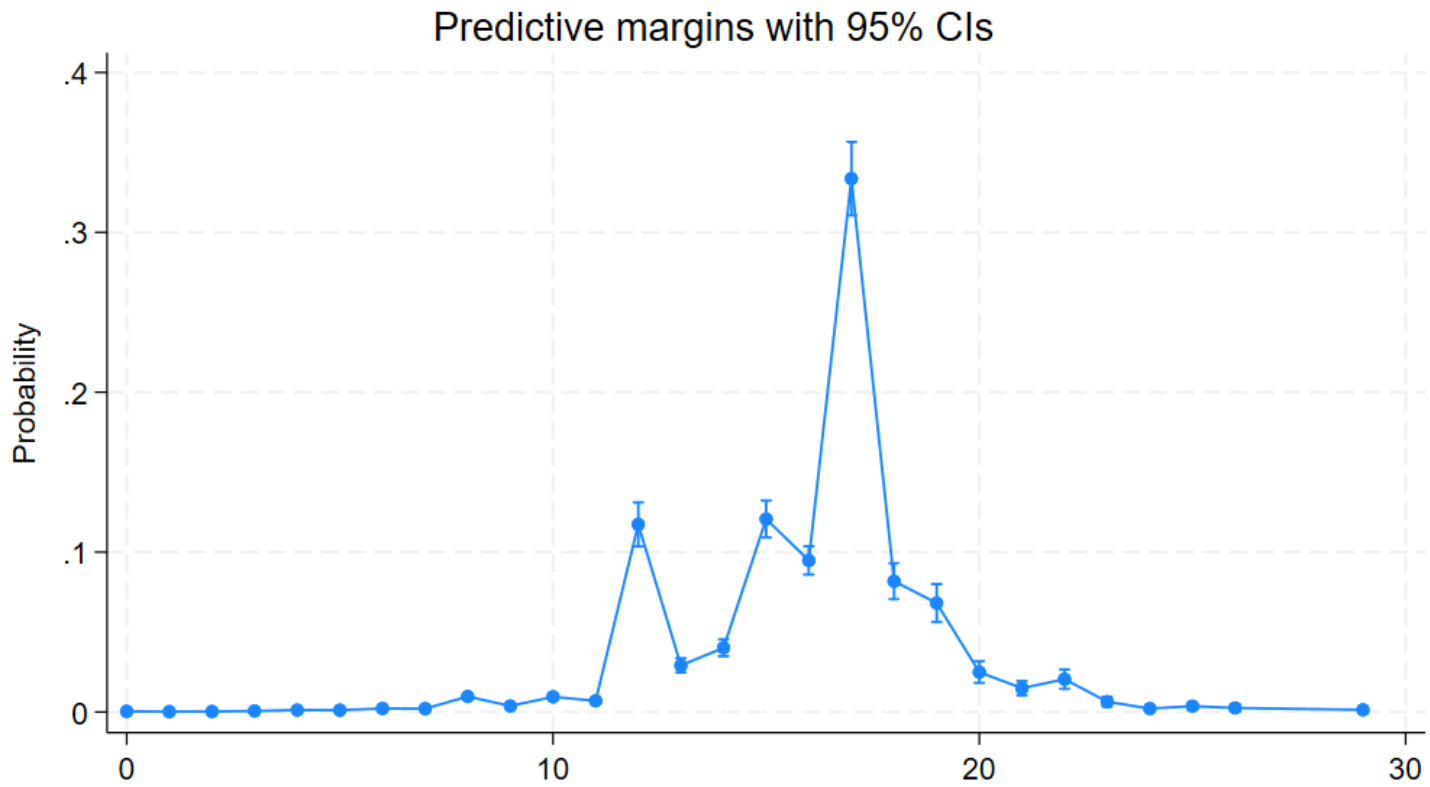


Figura: Probabilidades para el caso de mujeres de la región del Bio-bio de un nivel socio económico Alto con 30 años de edad Jefas de hogar para lo años de escolaridad

Regresión multinomial ordinal

Resultados de estimación caso 2

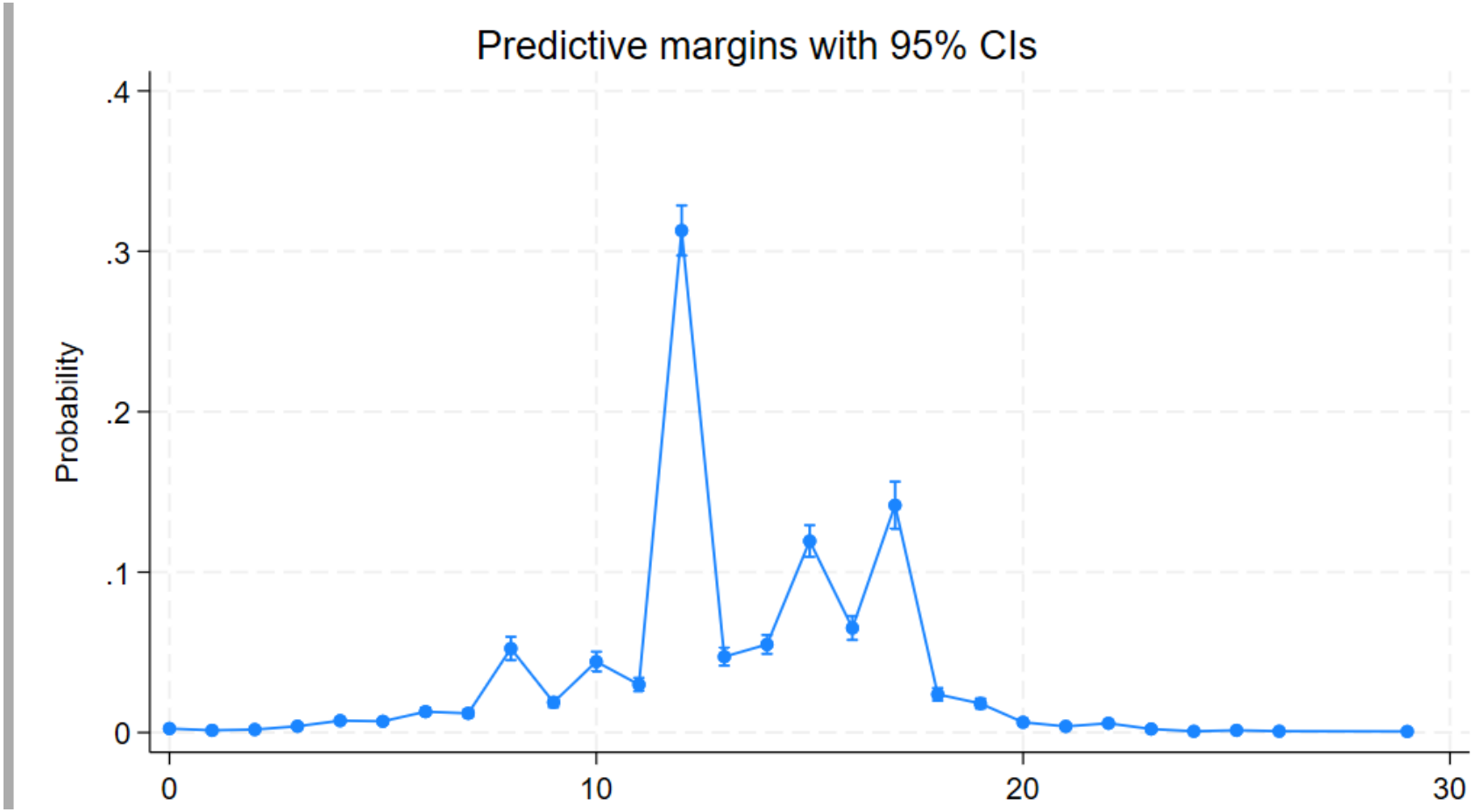


Figura: Probabilidades para el caso de mujeres de la región del Bio-bio de un nivel socio económico Bajo con 30 años de edad Jefas de hogar para lo años de escolaridad.

Regresión multinomial ordinal

Residuos

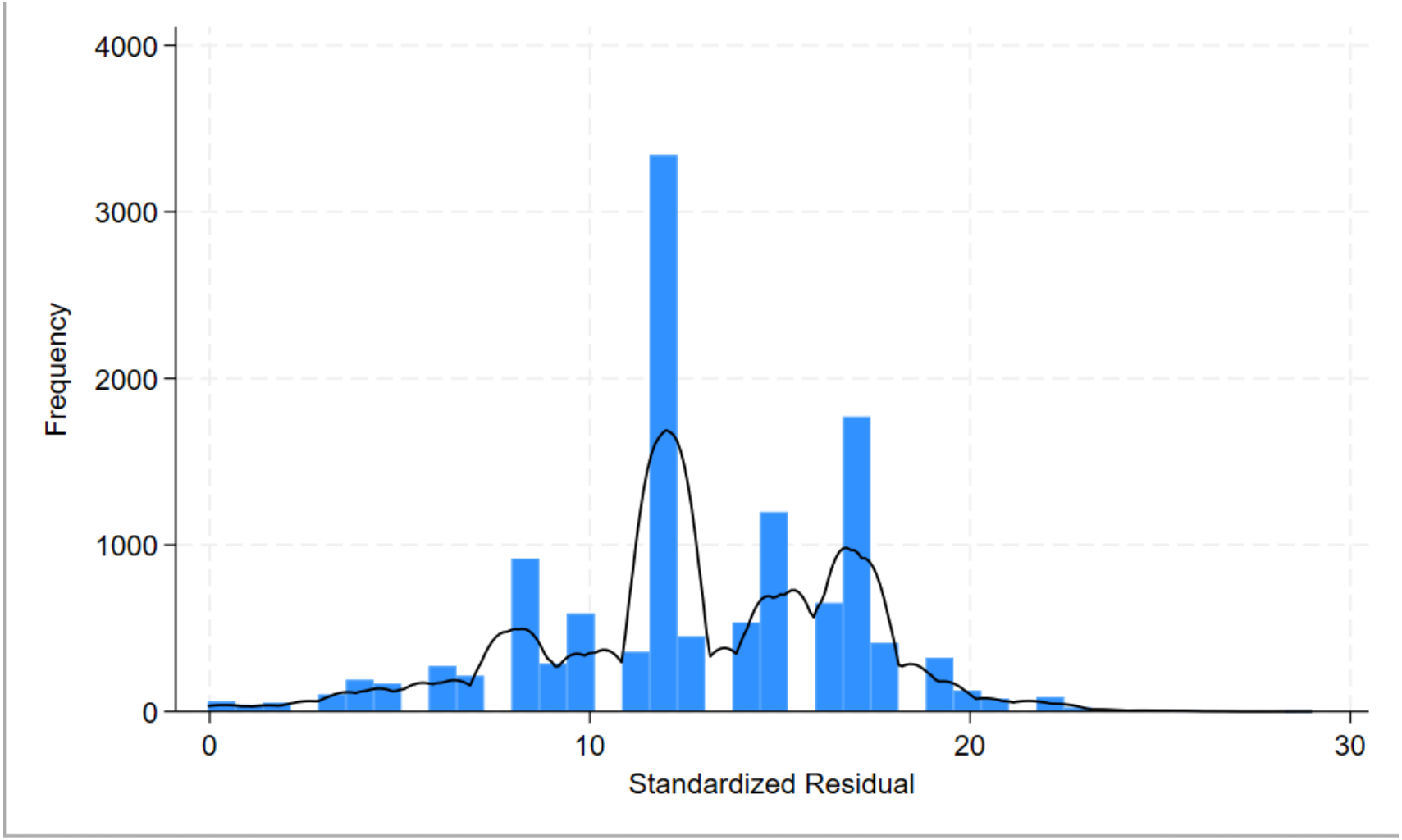


Figura: Histogramas Residuos para modelo multinomial ordinal de los años de escolaridad de los jefes de hogar en la región del Bio-bío y Metropolitana.

Regresión multinomial ordinal

Regresión paralela o supuesto de probabilidades proporcionales

Uno de los supuestos que subyacen a la regresión logística ordenada es que la relación entre cada par de grupos de resultados es la misma. En otras palabras, la regresión logística ordenada supone que los coeficientes que describen la relación entre, digamos, las categorías más bajas versus todas las categorías superiores de la variable de respuesta son los mismos que los que describen la relación entre la siguiente categoría más baja y todas las categorías superiores

- La hipótesis nula es que no hay diferencia en los coeficientes entre modelos.
- La hipótesis alternativa es que existe diferencia entre los coeficientes entre los modelos.

Regresión multinomial ordinal

Prueba de razón de verosimilitud aproximada de proporcionalidad de probabilidades entre categorías de respuesta

$$\begin{aligned} \text{chi}\,2(130) &= 1318{,}75 \\ \text{Prob} > \text{chi2} &= 0{,}0000 \end{aligned}$$

Índice

Regresión Multinomial Nominal

La regresión logística multinomial se usa para modelos con una variable dependiente nominal de más de dos categorías (multiclases) y es una extensión multivariante de la regresión logística binaria clásica.

$$P(Y = j|X) = \frac{e^{\beta_{0j}+\beta_{1j}X_1+\cdots+\beta_{pj}X_p}}{1 + \sum_{k=1}^{J-1} e^{\beta_{0k}+\beta_{1k}X_1+\cdots+\beta_{pk}X_p}}$$

Dónde:

- (P(Y=j—X)) es la probabilidad de la j-ésima categoría de la variable dependiente dadas las variables independientes.
- $(e^{\beta_{0j}+\beta_{1j}X_1+\cdots+\beta_{pj}X_p})$ es la combinación lineal exponencial de los predictores para la j-ésima categoría.
- (J) es el número total de categorías de la variable dependiente.
- $(\beta_{0j}, \beta_{1j}, \ldots, \beta_{pj})$ son los coeficientes de regresión para la j-ésima categoría.

Pregunta de investigación

¿Existe relación significativa entre la variable respuesta esc (años de escolaridad) para los jefes de hogar que residen en la región Metropolitana y la del Biobío, según los niveles de las variables explicativas nse, sexo y edad?

Hipótesis nula: No existe relación significativa entre la variable respuesta esc (años de escolaridad) para los jefes de hogar que residen en la región Metropolitana y la del Biobío, según los niveles de las variables explicativas nse, sexo y edad.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

Hipótesis alternativa: Existe relación significativa entre la variable respuesta esc (años de escolaridad) para los jefes de hogar que residen en la región Metropolitana y la del Biobío, según los niveles de las variables explicativas nse, sexo y edad.

$$H_1 : \beta_i \neq 0$$

Aplicación

Survey: Multinomial logistic regression

Number of strata = 204	Number of obs	12,271
Number of PSUs = 3,653	Population size	2,373,194
	Design df	3,449
	$F(20, 3430)$	= 54,31
	Prob > F	= 0,0000

esc_multicat	Coefficient	Linearized std. err.	t	$P > \|t\|$	[95 % conf.	interval]
1						
edad	0.0720844	0.0032464	22.20	0.000	0.0657193	0.0784496
sexo						
Mujer	-0.7279051	0.0926635	-7.86	0.000	-0.909586	-0.5462243
yoprcor	-3.25e-06	4.55e-07	-7.13	0.000	-4.14e-06	-2.35e-06
region						
Región Metropolitana de Santiago	-0.1644654	0.0869549	-1.89	0.059	-0.3349538	0.006023
nse						
Medio	-1.136914	0.1276845	-8.90	0.000	-1.387259	-0.8865692
Alto	-2.667539	0.1677573	-15.90	0.000	-2.996453	-2.338626
Bajo-Medio	-0.9225988	0.1431167	-6.45	0.000	-1.203201	-0.6419968
Bajo-Alto	0.6170455	0.6645032	0.93	0.353	-0.6858141	1.919905
Bajo-Medio-Alto	-1.665747	0.1441566	-11.56	0.000	-1.948387	-1.383106
Medio-Alto	-1.819277	0.1540891	-11.81	0.000	-2.121392	-1.517162
$_cons$	-0.6246321	0.2716396	-2.30	0.022	-1.157223	-0.0920412

Figura: 1

Aplicación

2						
edad	0.0275966	0.0026739	10.32	0.000	0.022354	0.0328393
sexo						
Mujer	-0.4169939	0.0663444	-6.29	0.000	-0.5470721	-0.2869157
yoprcor	-1.87e-06	1.92e-07	-9.76	0.000	-2.25e-06	-1.50e-06
region						
Región Metropolitana de Santiago	-0.0924241	0.0687721	-1.34	0.179	-0.2272622	0.042414
nse						
Medio	-0.5883586	0.1245093	-4.73	0.000	-0.832478	-0.3442393
Alto	-1.595242	0.1385153	-11.52	0.000	-1.866822	-1.323662
Bajo-Medio	-0.5255517	0.1377088	-3.82	0.000	-0.7955508	-0.2555526
Bajo-Alto	-1.859659	0.8265996	-2.25	0.025	-3.480333	-0.2389845
Bajo-Medio-Alto	-1.081162	0.1509599	-7.16	0.000	-1.377142	-0.7851821
Medio-Alto	-1.154935	0.1440045	-8.02	0.000	-1.437278	-0.8725923
$_cons$	0.7236811	0.1940188	3.73	0.000	0.3432778	1.104084
3 (base outcome)						

Figura: Tabla de los resultados obtenidos aplicando la regresión logística multinomial

Interpretaciones

Podemos ver que las variables edad, sexo, yoprcor y nse son todas estadísticamente significativas, ya que todos sus valores p son menores a 0.05. Esto nos quiere decir que tienen un impacto estadísticamente significativo en la probabilidad de pertenecer a una categoría específica de años de escolaridad en comparación con la categoría base.

Residuos

A continuación se mostrará un histograma de los residuos:

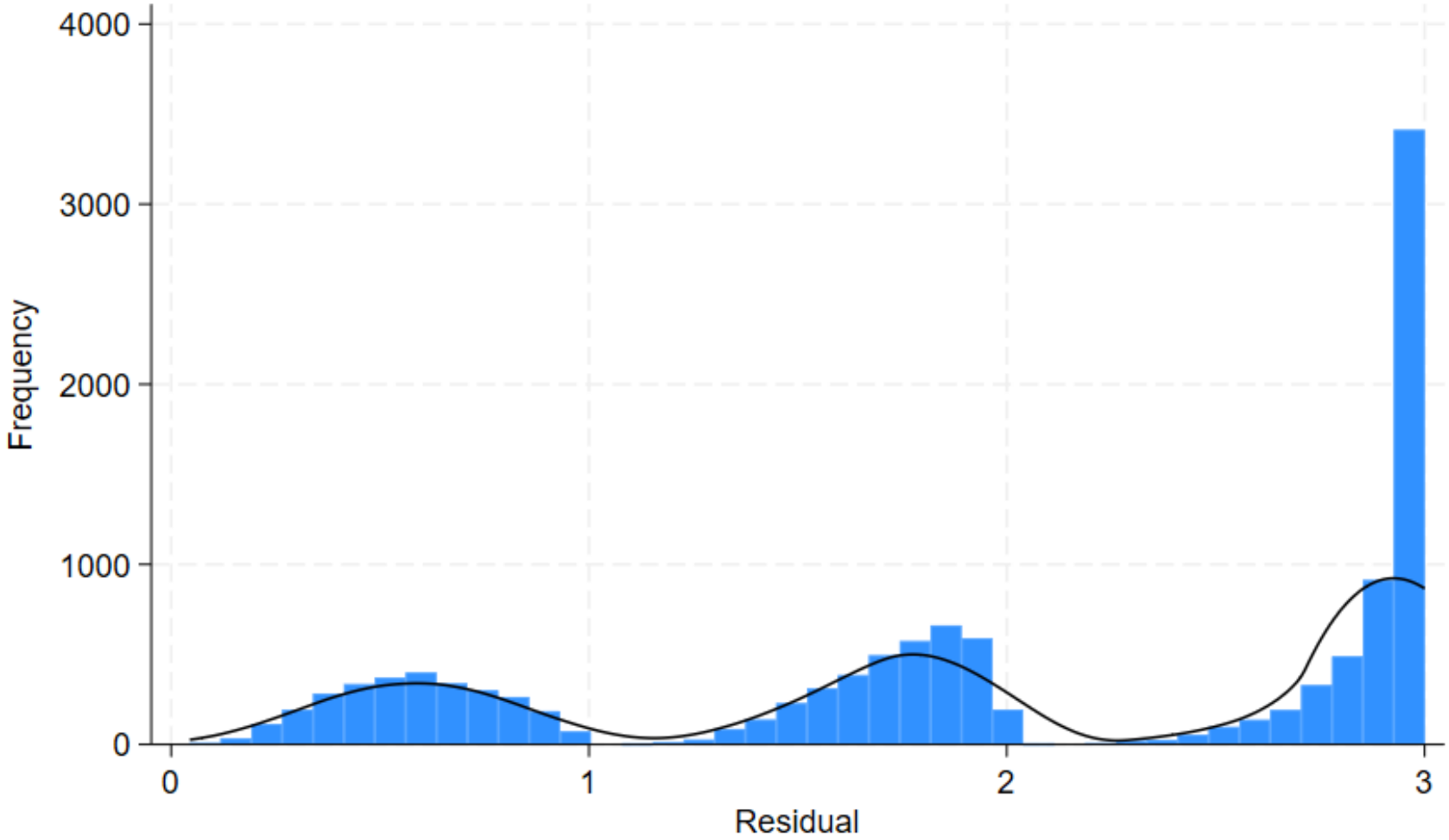


Figura: Histograma residuos

Al ver el histograma, podría indicar que los residuos están distribuidos de manera diferente para cada categoría. Podría ser por las siguientes razones:

- Existen factores adicionales que están influyendo en la cantidad de años de escolaridad de los jefes de hogar, pero que no están siendo considerados por el modelo.
- El modelo está capturando demasiados factores de confusión.
- El modelo es correcto, pero los residuos se están separando por las categorías de la variable respuesta.

Comentarios

summarize residuos

Variable	Obs	Mean	Std. dev.	Min	Max
residuos	**12,271**	2.002561	.9464557	.0439225	3

Figura: Resumen de los residuos

En lugar de evaluar la normalidad de los residuos, este modelo se centra en otras consideraciones como la bondad de ajuste del modelo. La bondad de ajuste puede evaluarse mediante estadísticas específicas del modelo, como la devianza y la prueba de razón de verosimilitudes. Estas pruebas evalúan qué tan bien se ajustan los datos observados a las predicciones del modelo.

Índice

Regresión Poisson

Definición

El modelo de regresión Poisson es un modelo el cual se distingue por ser utilizado en estudios de variable de recuento, ideal para poder modelar valores enteros no negativos que cuenten.

Modelo de regresión

$$\log(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... \beta_n X_n$$

donde,

i) y es la variable respuesta,

ii) β_i, con i=0,1,...,n, son los coeficientes númericos, β_0 es la intersección.

iii) x es la variable predictora.

Regresión Poisson

Pregunta de investigación

¿Existe una diferencia significativa en la media de años de escolaridad de los jefes de hogar en las regiones de Biobío y Metropolitana entre hombres y mujeres?

Hipótesis

No hay diferencia significativa en la media de años de escolaridad de los jefes de hogar en las regiones de Biobío y Metropolitana entre hombres y mujeres.

Regresión Poisson

Survey: Mean estimation

Number of strata	$= 204$	Number of obs	$= 20,169$
Number of PSUs	$= 3,737$	Population size	$= 3,439,776$
		Design df	$= \mathbf{3,533}$

	Mean	Linearized std. err.	[95 % conf.	interval]
esc	12.60054	.0547548	12.49319	12.7079

Cuadro: Estimación de la media.

Regresión Poisson

Aplicación del modelo

Survey: Poisson regression

esc	Coefficient	$P > \|t\|$
region		
RM	.0197	0.008
edad	-.0065	0.000
yoprcor	$6{,}43e-08$	0.000
nse		
Medio	.1380	0.000
Alto	.2932	0.000
Bajo-medio	.1145	0.000
Bajo-alto	-.0811	0.687
Bajo-medio-alto	.2016	0.000
Medio-alto	.2573	0.000
sexo		
2. Mujer	.0218	0.000
cons	2.619	0.000

Regresión Poisson

Análisis residual

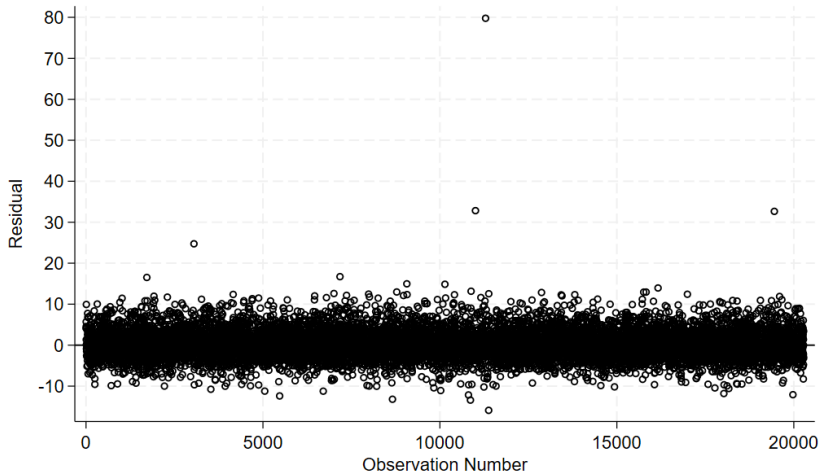


Figura: Diagrama de dispersión de residuos.

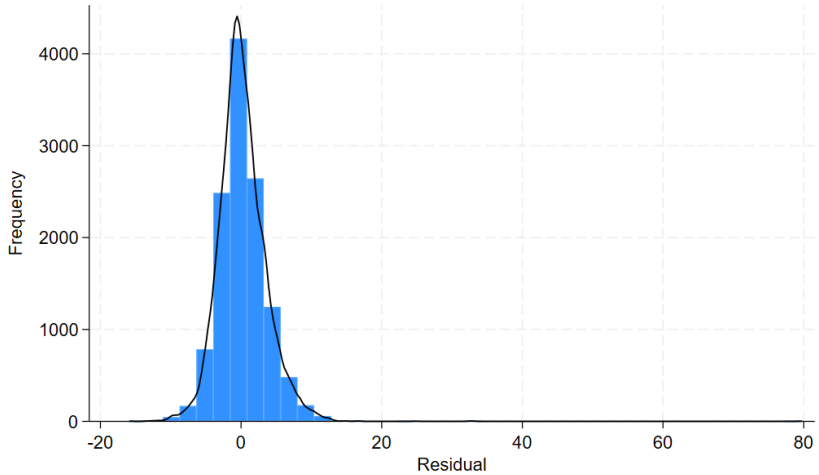


Figura: Densidad de los residuos.

Regresión Poisson

Survey: Mean estimation

Number of strata	$= 204$	Number of obs	$= 20,169$
Number of PSUs	$= 3,737$	Population size	$= 3,439,776$
		Design df	$= 3,533$

	Mean	Linearized std. err.	[95 % conf.	interval]
c.esc@sexo				
1. Hombre	12.91868	.0651949	12.79085	13.0465
2. Mujer	12.25202	.0687661	12.11719	12.38684

Cuadro: Estimación de la media de los años de escolaridad por sexo con los valores observados.

Regresión Poisson

Survey: Mean estimation

Number of strata = 204 Number of obs = 12,329
Number of PSUs = 3,653 Population size = 2,383,757
Design df = 3,449

	Mean	Linearized std. err.	[95 % conf.	interval]
c.pred_esc@sexo				
1. Hombre	13.60845	.0503696	13.5097	13.70721
2. Mujer	13.69826	.0398347	13.62016	13.77636

Cuadro: Estimación de la media de los años de escolaridad por sexo con los valores predichos obtenidos mediante la regresión de Poisson.

Índice

Regresión Binomial Negativa

Definición

El modelo de regresión Binomial Negativa surge como una alternativa al modelo Poisson, ya que este último sufre de problemas cuando existe una gran sobredispersión en los datos o hay varianza no constante.

Modelo de regresión

$$y_i = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ...\beta_n X_{in} + \epsilon_i)$$

donde,

i) y_i es la i-ésima observación de la variable respuesta,
ii) β_i, con i=0,1,...,n, son los coeficientes númericos, β_0 es la intersección.
iii) x_{ij} es la i-esima observación de la j-ésima variable predictora.
iv) ϵ_i es el error, se asume que está incorrelacionado con los x_i

Regresión Binomial Negativa

Pregunta de investigación

¿Existe una diferencia significativa en el promedio de años de escolaridad entre hombres y mujeres en las regiones de Biobío y Metropolitana?

Hipótesis

No hay diferencia significativa en el promedio de años de escolaridad entre hombres y mujeres en las regiones de Biobío y Metropolitana.

Regresión Binomial Negativa

Aplicación del modelo

```
. svy: nbreg esc i.region edad yoprcor i.nse /*modelo regresión binomial negativa*/
(running nbreg on estimation sample)
convergence not achieved

Survey: Negative binomial regression

Number of strata   =       204                  Number of obs      =     24,934
Number of PSUs     =     3,714                  Population size    =  4,711,360
                                                Design df          =      3,510
                                                F(   9,    3502)   =     282.69
Dispersion         = mean                       Prob > F           =     0.0000
```

esc	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
region						
Región Metropolitana de Santiago	.0087801	.0057975	1.51	0.130	-.0025868	.020147
edad	-.0056652	.000192	-29.51	0.000	-.0060417	-.0052888
yoprcor	6.21e-08	1.25e-08	4.98	0.000	3.76e-08	8.65e-08
nse						
Medio	.1032798	.0093761	11.02	0.000	.0848966	.1216631
Alto	.2529465	.0113115	22.36	0.000	.2307688	.2751243
Bajo-medio	.0931688	.0101527	9.18	0.000	.0732629	.1130746
Bajo-alto	-.0848279	.1187133	-0.71	0.475	-.3175819	.1479261
Bajo-medio-alto	.1611487	.0104957	15.35	0.000	.1405704	.181727
Medio-alto	.2180995	.0153242	14.23	0.000	.1880542	.2481449
_cons	2.624556	.0106367	246.75	0.000	2.603701	2.64541
/lnalpha	-37.49322	.			.	.
alpha	5.21e-17	.			.	.

Regresión Binomial Negativa

Análisis residual

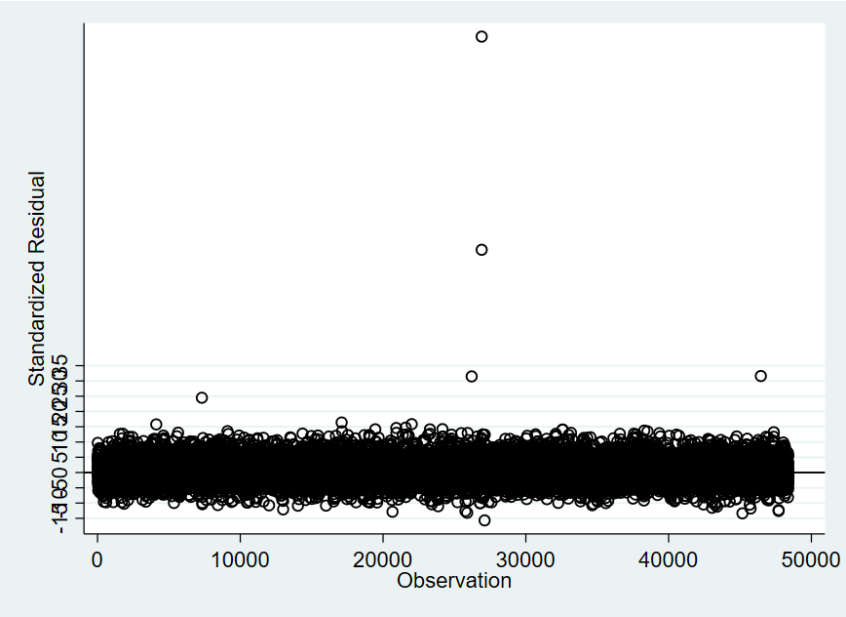


Figura: Diagrama de dispersión de residuos.

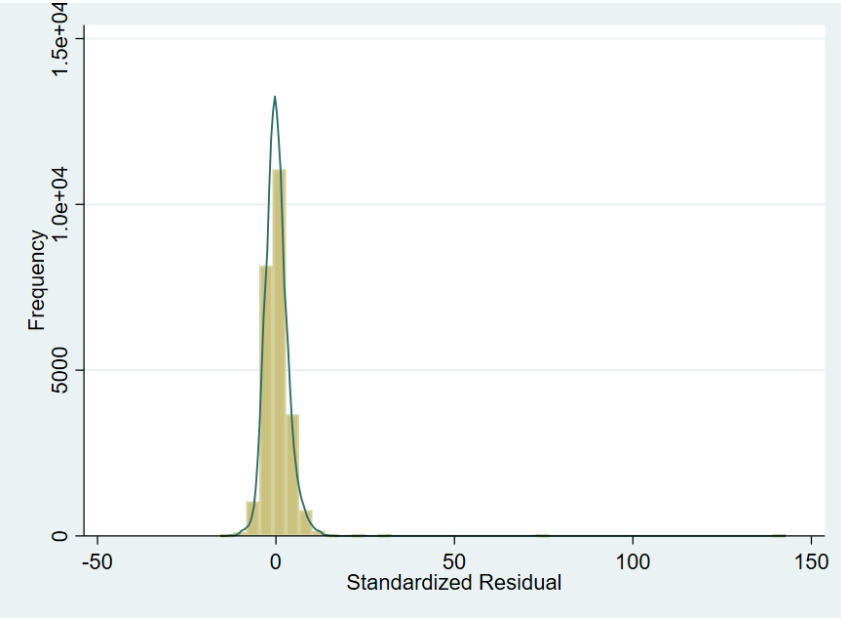


Figura: Densidad de los residuos.

Índice