SEMINARIO DE DOCTORADO

METODOLOGÍA DE INVESTIGACIÓN SOCIAL

Agustín Salvia
Santiago Poy

APLICACIÓN

4 REGRESIÓN LOGÍSTICA

Análisis de regresión logística

- En las ciencias sociales, muchas veces tratamos con variables que no son cuantitativas sino **cualitativas** (nominales, ordinales).

- El análisis de regresión logística es, al igual que el análisis de regresión lineal, apropiado para el estudio de relaciones de *dependencia* entre variables. A partir de una serie de variables *independientes* queremos explicar el comportamiento de una variable *dependiente*. La especificidad del análisis de regresión logística es que la **variable dependiente es cualitativa**.

- El análisis de regresión logística combina dos tradiciones de estudio en ciencias sociales: el análisis de **tablas de contingencia** y el **análisis de regresión lineal**. A diferencia del primero, permite introducir variables cuantitativas como independientes; a diferencia del segundo, es más flexible en cuanto a los supuestos que utiliza.

- Existen dos modalidades principales de regresión logística: la **binaria** (con una variable dependiente dicotómica) y la **multinomial** (con una variable dependiente que tiene más de dos categorías). Aquí nos enfocamos en la binaria.

Análisis de regresión logística

- Recordemos que, en el modelo lineal clásico, la relación entre variables puede escribirse del siguiente modo:

$$y = a + bx$$

- La recta de regresión se extiende entre $-\infty$ y $+\infty$, aun cuando los valores de x se interpretan en el rango de valores observados en la muestra. El problema que surge es cuando los valores pronosticados son imposibles: ello sucede cuando la variable dependiente es dicotómica y sólo asume valores 0 y 1. Por lo tanto, cualquier valor intermedio resulta imposible, así como valores que superen 0 y 1.

- La regresión logística soluciona este problema usando una función no lineal como la función logística. La formulación matemática de la función logística tiene la siguiente forma:

$$y = \Pr(y = 1 \mid x) = \frac{1}{1 + e^{-(a+bx)}}$$

Análisis de regresión logística

- La interpretación de los coeficientes de regresión logística difiere con respecto a los de la regresión lineal. El **coeficiente** es la **medida de cuánto varía el logaritmo neperiano del cociente de probabilidades de dos sucesos**. Esto es lo que se conoce como ***transformación logit***.

- La transformación logit surge de considerar la relación entre dos sucesos, es decir, la **razón** de ocurrencia (*odds*). Así, por ejemplo, si consideramos el evento "concluir la secundaria", podemos plantear:

$$odds = \frac{P}{1-P} = \frac{0{,}6}{1-0{,}6} = 1{,}5$$

- Así, la probabilidad de terminar la secundaria frente a no terminarla arroja un *odds* de 1,5. Si a esta expresión se le aplica la transformación logit, tenemos que:

$$log\left(\frac{P}{1-P}\right) = a + bx$$

- Esta transformación permite identificar el modelo en forma lineal y aditiva.

Análisis de regresión logística

- El proceso de análisis involucra una serie de pasos (López-Roldán y Fachelli, 2015):

  - ✓ Selección de las **variables del modelo**:

    - ✓ Definición de un tipo de modelo que requiera análisis de dependencia.
    - ✓ Evaluar la existencia de asociación entre las distintas variables independientes y la dependiente.
    - ✓ Descartar colinealidad entre variables independientes.
    - ✓ Evaluar la importancia de introducir interacciones relevantes.
    - ✓ Relación tamaño/variables: alrededor de 15 casos por cada variable.

  - ✓ Estimación de los coeficientes: método de **máxima verosimilitud**.

  - ✓ Evaluación del modelo:

    - ✓ Examinar la **bondad de ajuste**: R2 de Nagelkerke, R2 de Cox y Snell (SPSS) o pseudo R2 de McFadden (Stata).
    - ✓ Complementar con la prueba de Hosmer y Lemeshow (evalúa capacidad predictiva por subgrupos). Sólo para regresión múltiple (y sensible a N).
    - ✓ Examinar la **capacidad clasificatoria** del modelo (*overall*), mediante la tabla de clasificación.

Ejercicio 1: *análisis de regresión logística simple*

Análisis de regresión logística binaria simple

- **Efecto de <u>la condición de registro (formalidad)</u> sobre la <u>pobreza</u> entre trabajadores/as asalariados/as.**

			asal_noreg Condicion de registro en la seguridad social para asalariados		
			,00 Registrados	1,00 No registrados	Total
pobreza Poblacion bajo la linea de pobreza	,00 No pobre	Recuento	761	204	965
		% dentro de asal_noreg Condicion de registro en la seguridad social para asalariados	86,0%	55,1%	76,9%
	1,00 Pobre	Recuento	124	166	290
		% dentro de asal_noreg Condicion de registro en la seguridad social para asalariados	14,0%	44,9%	23,1%
Total		Recuento	885	370	1255
		% dentro de asal_noreg Condicion de registro en la seguridad social para asalariados	100,0%	100,0%	100,0%

La razón de probabilidades **global** de ser pobre frente a no serlo es:
0,231/0,769=0,30

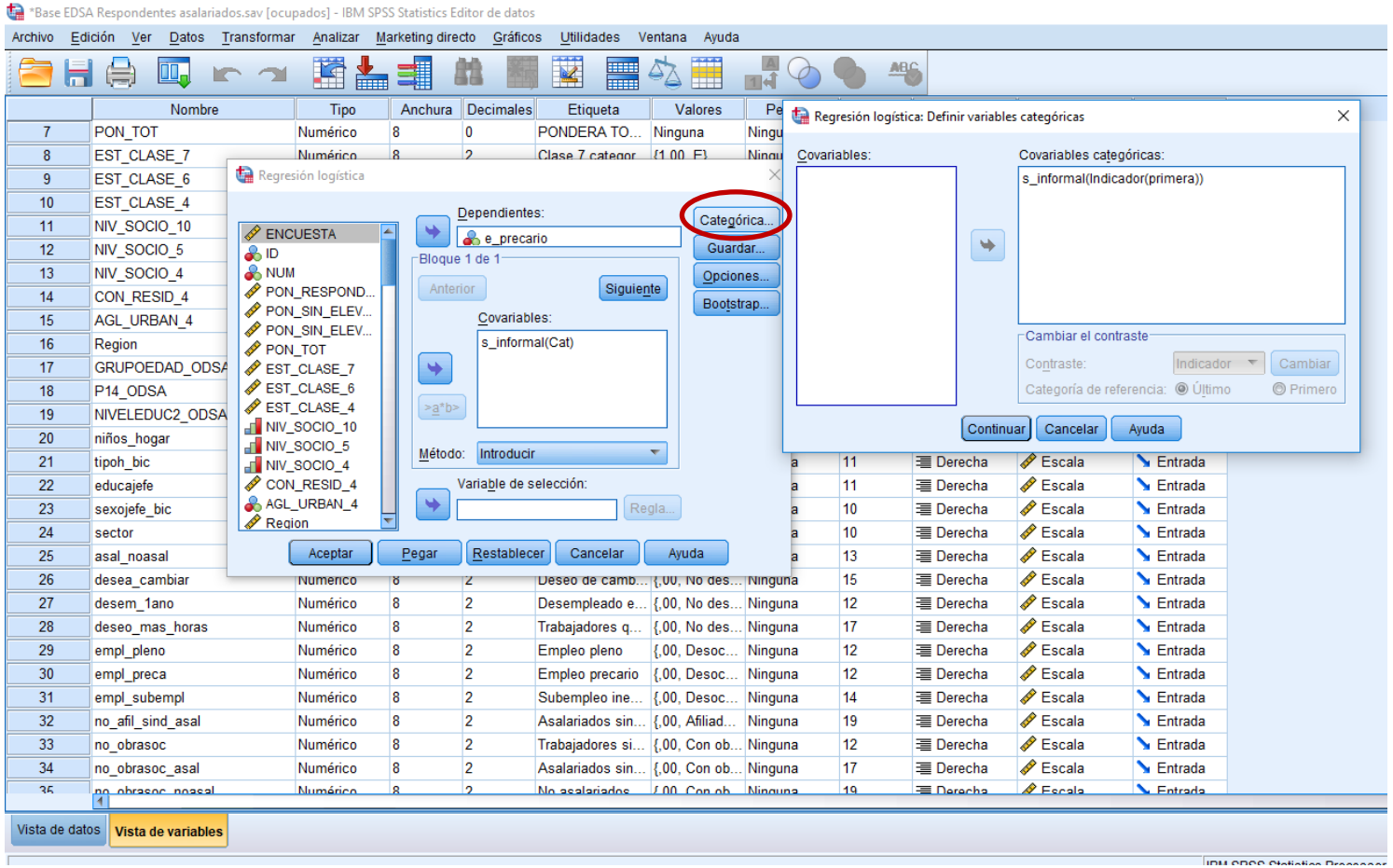
Si tiene trabajo **registrado**:
0,14/0,86=0,162

Si tiene trabajo **no registrado:**
0,449/0,551=0,815

Cuando la variable independiente varía en una unidad (pasa de registrado a no registrado), **la razón de ser pobre frente a no serlo** se incrementa en **0,815/0,162=4,98**

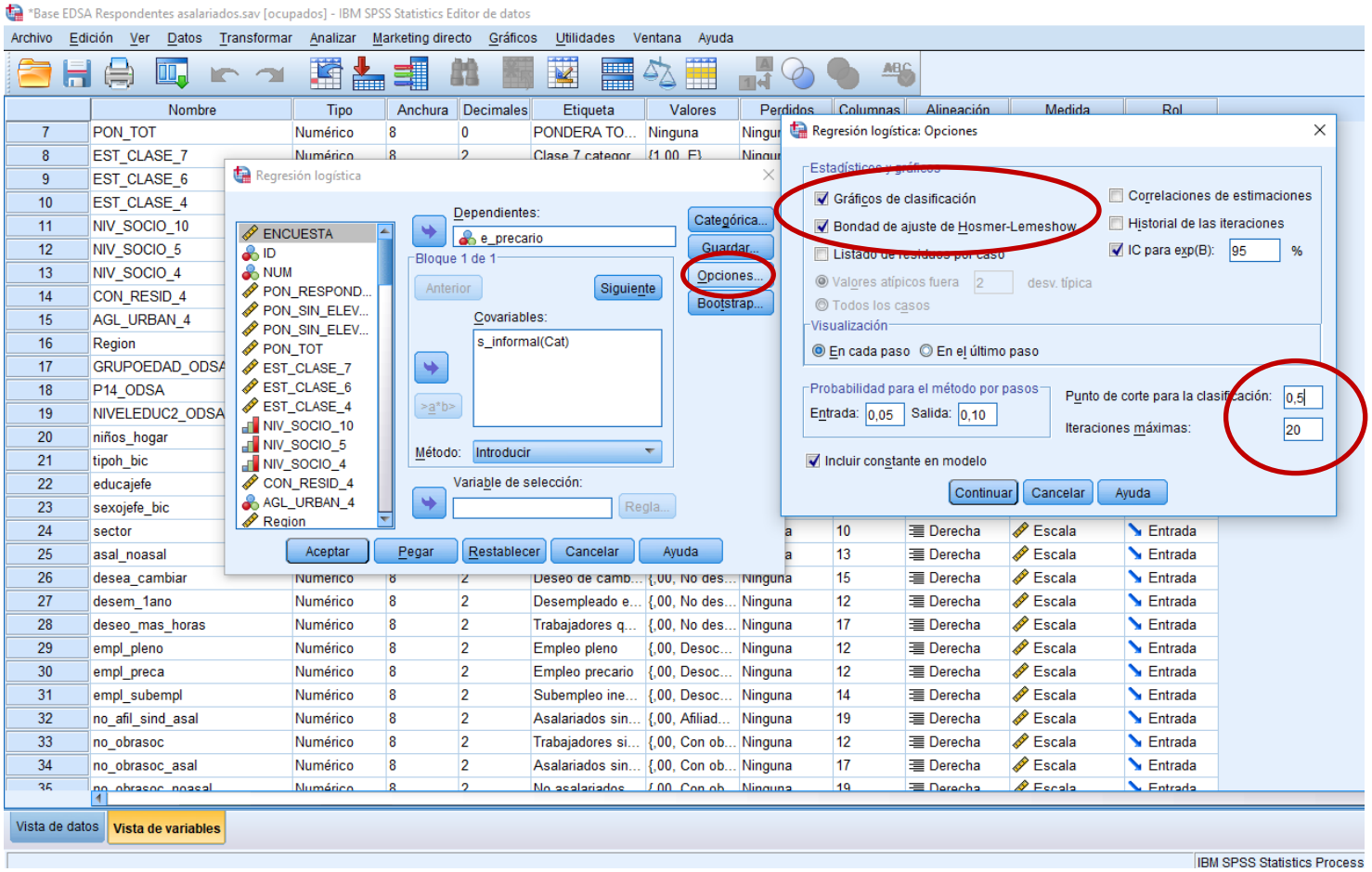
Análisis de regresión logística binaria simple

- **Comandos en SPSS:**



Análisis de regresión logística binaria simple

- **Comandos en SPSS:**



Análisis de regresión logística binaria simple

- **Efecto de <u>la condición de registro (formalidad)</u> sobre la <u>pobreza</u> entre trabajadores/as asalariados/as.**

Resumen de procesamiento de casos

Casos sin ponderar[a]		N	Porcentaje
Casos seleccionados	Incluido en el análisis	1256	100,0
	Casos perdidos	0	,0
	Total	1256	100,0
Casos no seleccionados		0	,0
Total		1256	100,0

a. Si la ponderación está en vigor, consulte la tabla de clasificación para el número total de casos.

El resumen de procesamiento nos permite identificar si tenemos casos perdidos. Habitualmente, los casos perdidos se originan en que alguna variable independiente tiene *missing.*

Codificación de variable dependiente

Valor original	Valor interno
.00 No pobre	0
1.00 Pobre	1

Codificaciones de variables categóricas

		Frecuencia	Codificación de parámetro (1)
asal_noreg Condicion de registro en la seguridad social para asalariados	.00 Registrados	914	,000
	1.00 No registrados	342	1,000

La **codificación** de parámetros nos permite entender los resultados.

Aquí lo que nos dice es que el modelo va a evaluar qué sucede al pasar del valor 0 a 1 de "asal_no_reg", es decir, cuánto se incrementa la razón de <u>ser no pobre a serlo</u> cuando ***x*** pasa de <u>empleo registrado a no registrado</u>.

Análisis de regresión logística binaria simple

- **Efecto de <u>la condición de registro (formalidad)</u> sobre la <u>pobreza</u> entre trabajadores/as asalariados/as.**

Tabla de clasificación[a,b]

			Pronosticado		
			pobreza Poblacion bajo la linea de pobreza		Porcentaje correcto
	Observado		.00 No pobre	1.00 Pobre	
Paso 0	pobreza Poblacion bajo la linea de pobreza	.00 No pobre	5821	0	100,0
		1.00 Pobre	1753	0	,0
	Porcentaje global				76,9

a. La constante se incluye en el modelo.

b. El valor de corte es ,500

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-1,200	,027	1940,876	1	,000	,301

Las variables no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	Condicion de registro en la seguridad social para asalariados(1)	842,791	1	,000
	Estadísticos globales		842,791	1	,000

La primera salida que genera es **sin introducir** ninguna variable independiente. Nos permite comparar qué pasa con el modelo cuando incorporamos una variable independiente.

Análisis de regresión logística binaria simple

- **Efecto de <u>la condición de registro (formalidad)</u> sobre la <u>pobreza</u> entre trabajadores/as asalariados/as.**

Pruebas ómnibus de coeficientes de modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	788,040	1	,000
	Bloque	788,040	1	,000
	Modelo	788,040	1	,000

Cuando tenemos un modelo con varios pasos (porque introducimos variables en pasos o *steps*) esperamos que mejore Chi Cuadrado.

Resumen del modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	7406,322[a]	,099	,149

a. La estimación ha terminado en el número de iteración 4 porque las estimaciones de parámetro han cambiado en menos de ,001.

Cuando introducimos distintas variables en pasos, el -2 Log de verosimilitud debe ir reduciéndose. Ello significa que el modelo ajusta mejor a los datos.

Pueden entenderse como el grado en que se reduce la probabilidad de no explicar nada con las variables que se agregan. Es habitual que sus valores oscilen entre .200 y .300.

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	,000	0	.

Es otra medida de bondad de ajuste. Sólo tiene sentido cuando hay más de 1 VI. *En este test esperamos que el p-valor sea mayor a .05, porque buscamos aceptar la hipótesis nula.*

Análisis de regresión logística binaria simple

- **Efecto de <u>la condición de registro (formalidad)</u> sobre la <u>pobreza</u> entre trabajadores/as asalariados/as.**

Tabla de clasificación[a]

Observado			Pronosticado		
			pobreza Poblacion bajo la linea de pobreza		Porcentaje correcto
			.00 No pobre	1.00 Pobre	
Paso 1	pobreza Poblacion bajo la linea de pobreza	.00 No pobre	5821	0	100,0
		1.00 Pobre	1753	0	,0
	Porcentaje global				76,9

a. El valor de corte es ,500

Capacidad clasificatoria del modelo a partir de la ecuación de regresión especificada. De acuerdo con esto, el 100% de los que no son pobres son clasificados correctamente, y nadie de los que son pobres es asignado correctamente. Podemos mejorar la capacidad clasificatoria alterando el punto de corte (por *default* en .50), ya que la capacidad clasificatoria es sensible a la incidencia del evento.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Condicion de registro en la seguridad social para asalariados(1)	1,607	,058	768,163	1	,000	4,987
	Constante	-1,812	,039	2115,491	1	,000	,163

a. Variables especificadas en el paso 1: Condición de registro en la seguridad social para asalariados.

Esta es la tabla que más nos interesa en términos sustantivos. Nos dice qué efectos tienen las variables independientes sobre la dependiente.

Los coeficientes B son los logaritmos de los *odds ratios* y pueden entenderse en términos de **jerarquía e intensidad** explicativa

El **error típico** de los coeficientes **B** no debería superar el 50% del valor del coeficiente

La significancia de los coeficientes (a partir de los Wald) se interpreta como en la regresión lineal.

Exp(B) es el ***odds ratio*** antes visto. Entre quienes tienen empleos **no registrados**, las chances de ser pobre (frente a no serlo) son casi **5 veces** las que enfrentan quienes tienen empleos registrados.

Análisis de regresión logística binaria simple

- La reconstrucción de la ecuación de regresión nos permite predecir la probabilidad de ocurrencia del evento:

$$\boldsymbol{Pr(y=1)=\frac{1}{1+e^{-(a+bx)}}}$$

- A partir de lo anterior, tenemos que, para un trabajador no registrado la probabilidad de ser pobre:

$$\boldsymbol{Pr(y=1)=\frac{1}{1+e^{-(-1{,}807+1{,}607*1)}}}$$

$$\boldsymbol{Pr(y=1)=\frac{1}{2,22}=0,450=45\%}$$

- Para un trabajador no registrado, la probabilidad de ser pobre:

$$\boldsymbol{Pr(y=1)=\frac{1}{1+e^{-(-1{,}807+1{,}607*0)}}=0,141=14,1\%}$$

Análisis de regresión logística binaria simple

- Estos son los valores predichos por el modelo de regresión:

$$Pr(y = 1) = \frac{1}{2,22} = 0,450 = 45\%$$

$$Pr(y = 1) = \frac{1}{1 + e^{-(-1,807 + 1,607 * 0)}} = 0,141 = 14,1\%$$

PRE_1	var
,44884	
,14038	
,14038	
,44884	
,44884	
,14038	
,14038	
,14038	
.	
,14038	
,14038	
,44884	
,14038	
.	
,14038	

Ejercicio 2: *análisis de regresión logística múltiple*

Análisis de regresión logística binaria múltiple

Introducimos más covariables en el modelo de regresión logística.

```
OUTPUT CLOSE ALL.

USE ALL.
COMPUTE filter_$=(cat_ocup=3).
FILTER BY filter_$.
EXECUTE.

weight by PONDERA_SIN_ELEVAR.

LOGISTIC REGRESSION VARIABLES pobreza
  /METHOD=ENTER asal_noreg d_mujer CH06 n_educ d_ninos d_gba
  /CONTRAST (asal_noreg)=Indicator(1)
  /CONTRAST (d_mujer)=Indicator(1)
  /CONTRAST (n_educ)=Indicator
  /CONTRAST (d_ninos)=Indicator(1)
  /CONTRAST (d_gba)=Indicator(1)
  /SAVE=ZRESID
  /CLASSPLOT
  /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

Análisis de regresión logística binaria múltiple

Antes de ello, hacemos una prueba de correlaciones:

Correlaciones

		CH06 ¿Cuántos años cumplidos tiene?	asal_noreg Condicion de registro en la seguridad social para asalariados	n_educ Maximo nivel educativo alcanzado	d_mujer Mujer	d_ninos Presencia de niños en el hogar	d_gba Partidos del Conurbano
CH06 ¿Cuántos años cumplidos tiene?	Correlación de Pearson	1	-,084**	-,022	,060*	-,205**	-,014
	Sig. (bilateral)		,003	,426	,034	,000	,618
	N	1256	1256	1256	1256	1256	1256
asal_noreg Condicion de registro en la seguridad social para asalariados	Correlación de Pearson	-,084**	1	-,206**	,026	,038	,092**
	Sig. (bilateral)	,003		,000	,357	,174	,001
	N	1256	1256	1256	1256	1256	1256
n_educ Maximo nivel educativo alcanzado	Correlación de Pearson	-,022	-,206**	1	,210**	-,120**	-,308**
	Sig. (bilateral)	,426	,000		,000	,000	,000
	N	1256	1256	1256	1256	1256	1256
d_mujer Mujer	Correlación de Pearson	,060*	,026	,210**	1	,018	-,048
	Sig. (bilateral)	,034	,357	,000		,531	,092
	N	1256	1256	1256	1256	1256	1256
d_ninos Presencia de niños en el hogar	Correlación de Pearson	-,205**	,038	-,120**	,018	1	,116**
	Sig. (bilateral)	,000	,174	,000	,531		,000
	N	1256	1256	1256	1256	1256	1256
d_gba Partidos del Conurbano	Correlación de Pearson	-,014	,092**	-,308**	-,048	,116**	1
	Sig. (bilateral)	,618	,001	,000	,092	,000	
	N	1256	1256	1256	1256	1256	1256

**. La correlación es significativa en el nivel 0,01 (bilateral).

*. La correlación es significativa en el nivel 0,05 (bilateral).

Análisis de regresión logística binaria múltiple

Codificaciones de variables categóricas

		Frecuencia	Codificación de parámetro (1)	(2)	(3)
n_educ Maximo nivel educativo alcanzado	1.00 Hasta secundaria incompleta	329	1,000	,000	,000
	2.00 Secundaria completa	302	,000	1,000	,000
	3.00 Terc/Univ incompleto	234	,000	,000	1,000
	4.00 Terc/Univ completo	391	,000	,000	,000
d_mujer Mujer	,00	684	,000		
	1,00	572	1,000		
d_gba Partidos del Conurbano	,00	357	,000		
	1,00	899	1,000		
d_ninos Presencia de niños en el hogar	.00 Sin niños	616	,000		
	1.00 Con niños	640	1,000		
asal_noreg Condicion de registro en la seguridad social para asalariados	.00 Registrados	914	,000		
	1.00 No registrados	342	1,000		

Es fundamental evaluar cómo están codificadas las variables categóricas. En este caso, al indicar que la categoría de referencia es la **última** vemos que es el valor=0.

De igual modo, vemos que la categoría=1 es "secundaria incompleta" y la categoría=3 es universitario incompleto.

Por lo tanto, vamos a interpretar el coeficiente como **el pasaje de la variable independiente en 1, 2 o 3 unidades** *(p. ej., cuando se pasa de universitario completo a secundaria incompleta, la razón de momios aumenta en…)*

Análisis de regresión logística binaria múltiple

Resumen del modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	946,110[a]	,194	,312

a. La estimación ha terminado en el número de iteración 6 porque las estimaciones de parámetro han cambiado en menos de ,001.

Apreciamos la **mejora** en el coeficiente R2 de Nagelkerke con respecto al modelo simple.

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	10,970	8	,203

La prueba de **Hosmer y Lemeshow** indica bondad de ajuste cuando la significancia es > 0.05

Aquí buscamos "quedarnos" con la hipótesis nula de que no hay diferencias entre la distribución observada y la pronosticada por el modelo especificado.

En este caso, vemos que la prueba es un indicador de **buen ajuste**.

Tabla de clasificación[a]

Observado			Pronosticado		
			pobreza Poblacion bajo la linea de pobreza		Porcentaje correcto
			.00 No pobre	1.00 Pobre	
Paso 1	pobreza Poblacion bajo la linea de pobreza	.00 No pobre	966	53	94,8
		1.00 Pobre	165	72	30,4
	Porcentaje global				82,6

a. El valor de corte es ,500

La tabla de clasificación nos muestra que el **82,6%** de los casos están bien clasificados.

Sin embargo, clasificó correctamente al **94,8%** de los no pobres (**Especificidad o "verdaderos negativos"**) y sólo al **30,4%** de los pobres (**Sensibilidad o "verdaderos positivos"**).

Análisis de regresión logística binaria múltiple

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Condicion de registro en la seguridad social para asalariados(1)	1,309	,171	58,860	1	,000	3,701
	Mujer(1)	,051	,171	,090	1	,764	1,053
	¿Cuántos años cumplidos tiene?	-,006	,007	,731	1	,393	,994
	Maximo nivel educativo alcanzado			57,795	3	,000	
	Maximo nivel educativo alcanzado(1)	1,884	,279	45,502	1	,000	6,579
	Maximo nivel educativo alcanzado(2)	1,198	,288	17,351	1	,000	3,313
	Maximo nivel educativo alcanzado(3)	,550	,325	2,869	1	,090	1,733
	Presencia de niños en el hogar(1)	1,346	,187	52,036	1	,000	3,842
	Partidos del Conurbano (1)	,551	,231	5,705	1	,017	1,735
	Constante	-4,071	,465	76,625	1	,000	,017

a. Variables especificadas en el paso 1: Condicion de registro en la seguridad social para asalariados, Mujer, ¿Cuántos años cumplidos tiene?, Maximo nivel educativo alcanzado, Presencia de niños en el hogar, Partidos del Conurbano.

Entre quienes tienen empleos **no registrados**, las chances de ser pobre (frente a no serlo) son **3,7 veces** las que enfrentan quienes tienen empleos registrados, *ceteris paribus.*

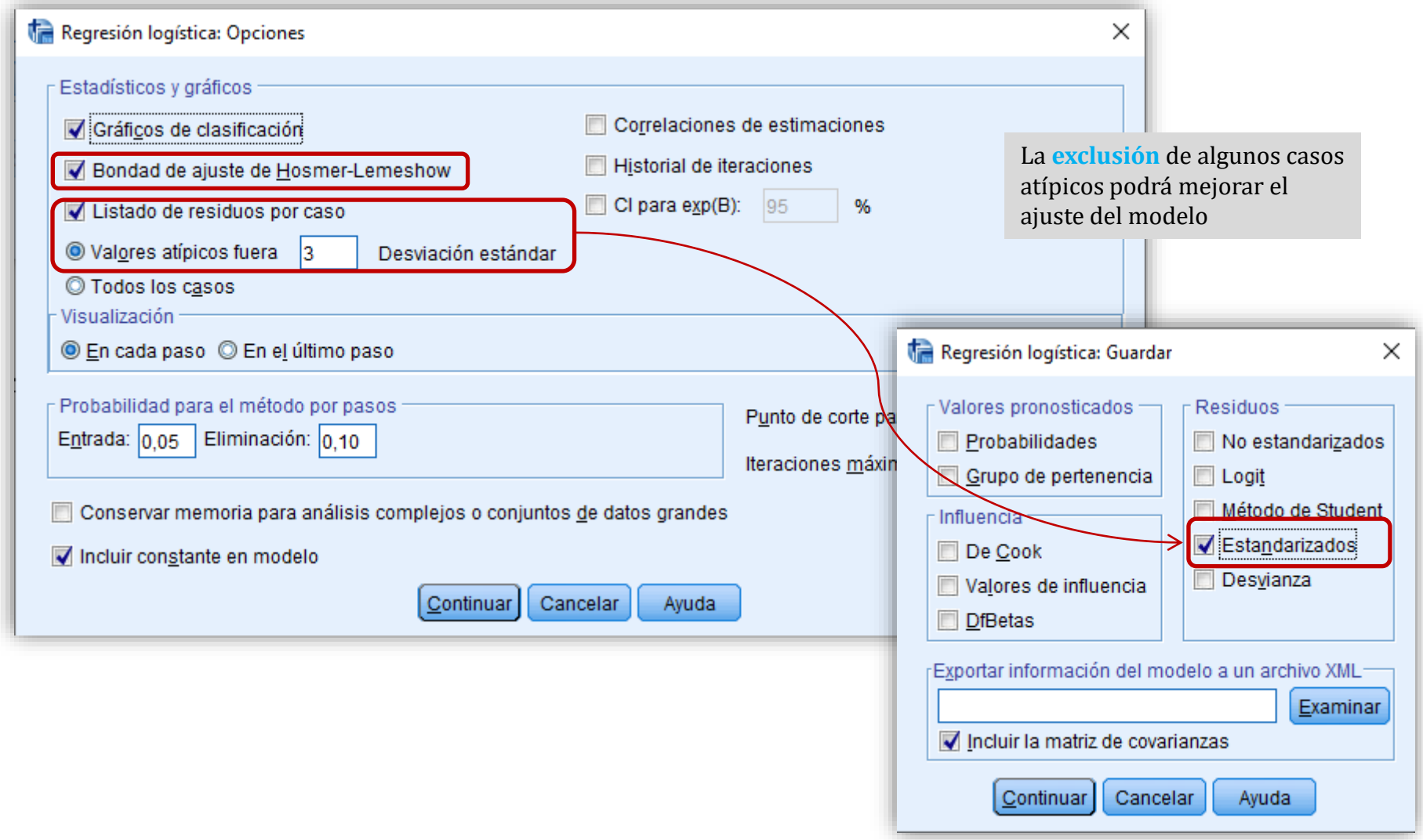
Por cada año cumplido, la razón de ser pobre frente a no serlo se reduce **1-0.994=0.006%**

Los coeficientes de sexo y edad **no resultan** estadísticamente significativos

Ejercicio 3: *mejorando el ajuste del modelo*

Estadísticos de bondad de ajuste

- 1) Excluyendo valores con residuos altos:



Mejorando el ajuste del modelo

- 1) Excluyendo valores con residuos altos:

```
OUTPUT CLOSE ALL.

USE ALL.
COMPUTE filter_$=(cat_ocup=3 & (ZRE_1>-3 & ZRE_1<3)).
FILTER BY filter_$.
EXECUTE.

weight off.

LOGISTIC REGRESSION VARIABLES pobreza
  /METHOD=ENTER asal_noreg d_mujer CH06 n_educ d_ninos d_gba
  /CONTRAST (asal_noreg)=Indicator(1)
  /CONTRAST (d_mujer)=Indicator(1)
  /CONTRAST (n_educ)=Indicator
  /CONTRAST (d_ninos)=Indicator(1)
  /CONTRAST (d_gba)=Indicator(1)
  /CLASSPLOT
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

Estadísticos de bondad de ajuste

- 1) Excluyendo valores con residuos altos:

Resumen del modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	797,025[a]	,240	,398

a. La estimación ha terminado en el número de iteración 7 porque las estimaciones de parámetro han cambiado en menos de ,001.

Apreciamos la **mejora** en el coeficiente R2 de Nagelkerke con respecto al modelo anterior

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	14,853	8	,062

La prueba de Hosmer & Lemeshow sigue indicando buen ajuste

Tabla de clasificación[a]

Observado			Pronosticado		
			pobreza Poblacion bajo la linea de pobreza		Porcentaje correcto
			.00 No pobre	1.00 Pobre	
Paso 1	pobreza Poblacion bajo la linea de pobreza	.00 No pobre	966	53	94,8
		1.00 Pobre	141	72	33,8
	Porcentaje global				84,3

a. El valor de corte es ,500

No se producen cambios importantes con respecto a la clasificación

Estadísticos de bondad de ajuste

- 2) Mejorando la tabla de clasificación: ***Índice de Youden***

Tabla de clasificación[a]

			Pronosticado		
			pobreza Poblacion bajo la linea de pobreza		Porcentaje correcto
	Observado		.00 No pobre	1.00 Pobre	
Paso 1	pobreza Poblacion bajo la linea de pobreza	.00 No pobre	966	53	94,8
		1.00 Pobre	165	72	30,4
	Porcentaje global				82,6

a. El valor de corte es ,500

A partir de la tabla de clasificación original

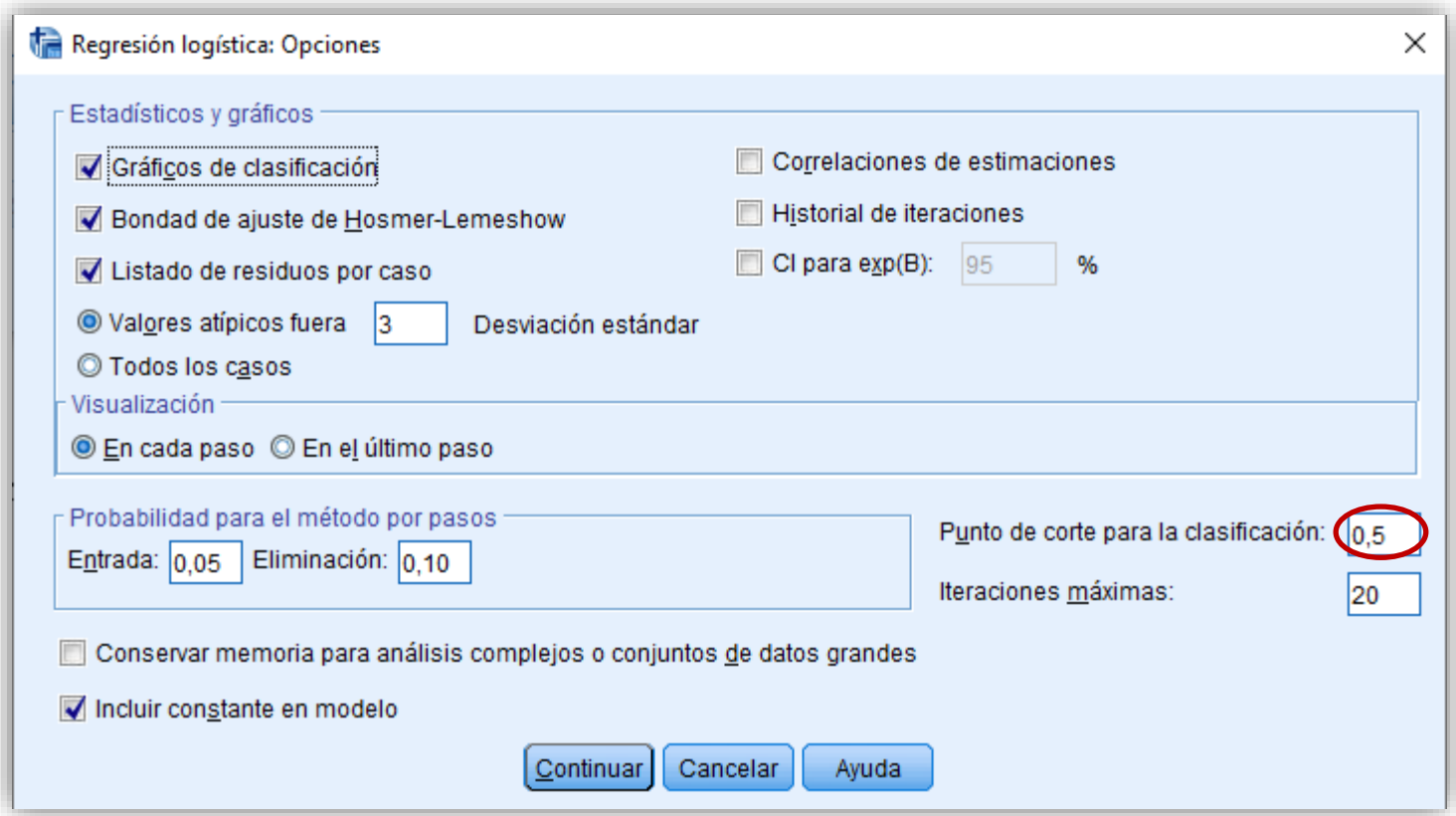
Especificidad: $P(\hat{y} = 0 \mid y = 0) = 0.948$
Sensibilidad: $P(\hat{y} = 1 \mid y = 1) = 0.304$

Índice de Youden: $Sensibilidad + Especificidad - 1$

$$IY = 0.304 + 0.948 - 1 = \mathbf{0.252}$$

Estadísticos de bondad de ajuste

- 2) Mejorando la tabla de clasificación:



Estadísticos de bondad de ajuste

- 2) Mejorando la tabla de clasificación

Tabla de clasificación[a]

			Pronosticado		
			pobreza Poblacion bajo la linea de pobreza		Porcentaje correcto
	Observado		.00 No pobre	1.00 Pobre	
Paso 1	pobreza Poblacion bajo la linea de pobreza	.00 No pobre	837	182	82,1
		1.00 Pobre	86	151	63,7
	Porcentaje global				78,7

a. El valor de corte es ,252

Apreciamos una **mejora importante** en la clasificación

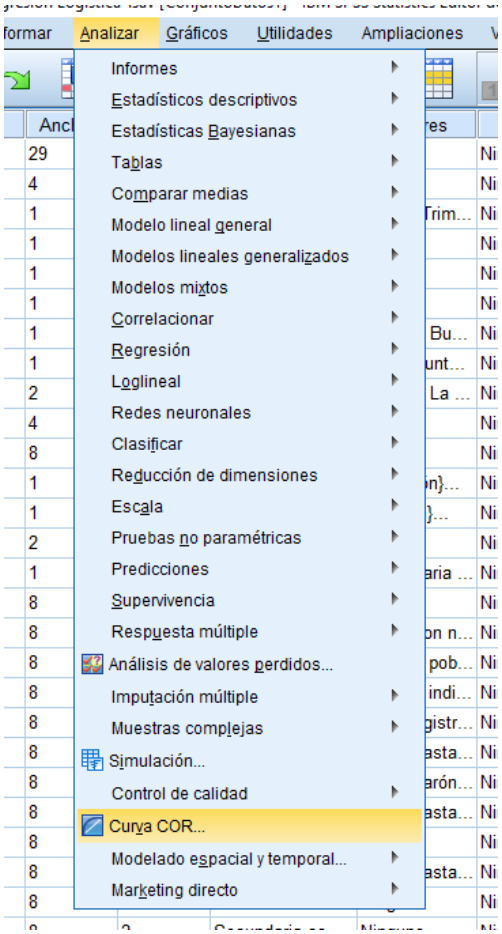
Estadísticos de bondad de ajuste

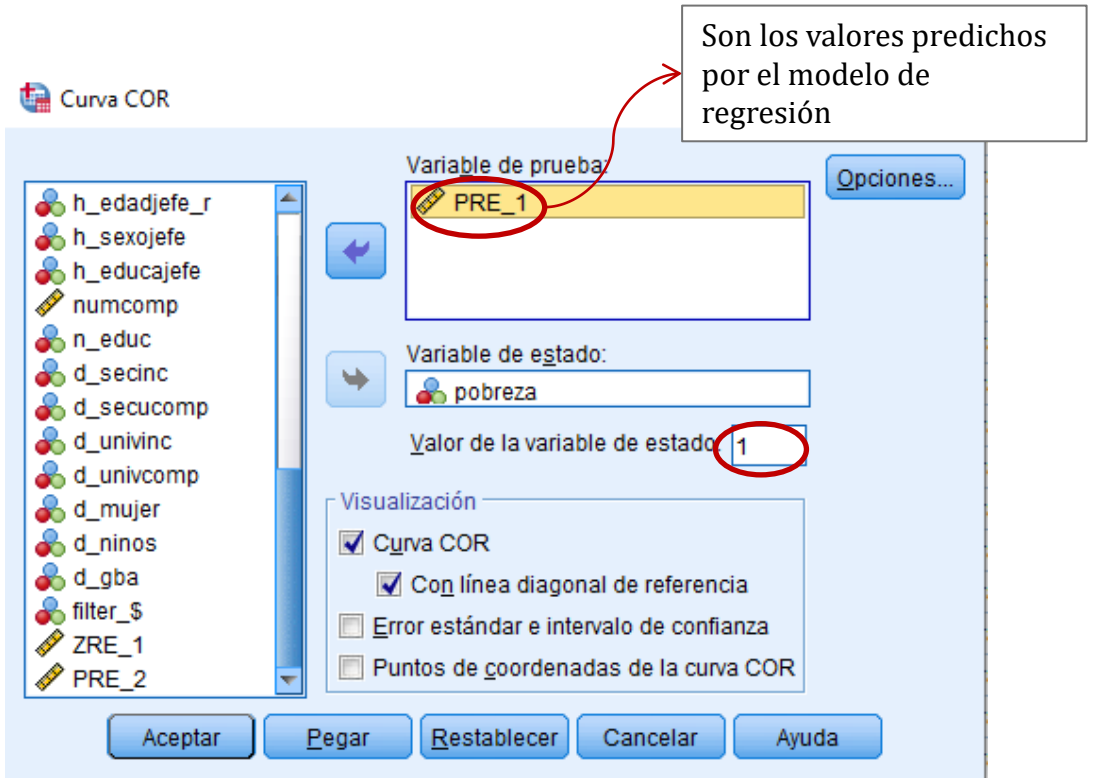
- 2) Mejorando la tabla de clasificación

```
                  Step number: 1

                  Observed Groups and Predicted Probabilities

     160 +                                                                                                    +
         I                                                                                                    I
         I  P                                                                                                 I
F        I  N                                                                                                 I
R    120 +  N                                                                                                 +
E        I  N                                                                                                 I
Q        I  N                                                                                                 I
U        I NN N   P                                                                                           I
E     80 + NN N  PN                                                                                           +
N        I NN N  NN                                                                                           I
C        I NN N  NN                                                                                           I
Y        I NN N  NN   PN                                                                                      I
      40 + NN NN NN   NNP       PPP           P                                                               +
         I NNNNNNNN   NNN       PNN          PPPP                                                             I
         I NNNNNNNNN NNNN       NNNN        PNNNPP               P              P                             I
         I NNNNNNNNN NNNNN    NNNNNNP      NNNNNNNN           PPNNP         PNNNNP                            I
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0        .1        .2        .3        .4        .5        .6        .7        .8        .9         1
  Group:  NNNNNNNNNNNNNNNNNNNNNNNNNPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP

          Predicted Probability is of Membership for Pobre
          The Cut Value is .25
          Symbols: N - No pobre
                   P - Pobre
          Each Symbol Represents 10 Cases.
```

Estadísticos de bondad de ajuste

- 3) Curva ***COR (Característica de Operador Receptor)***: representa la **sensibilidad** (eje de las *y, ordenadas*) en función de **1- especificidad** (eje de las *x, abscisas)*

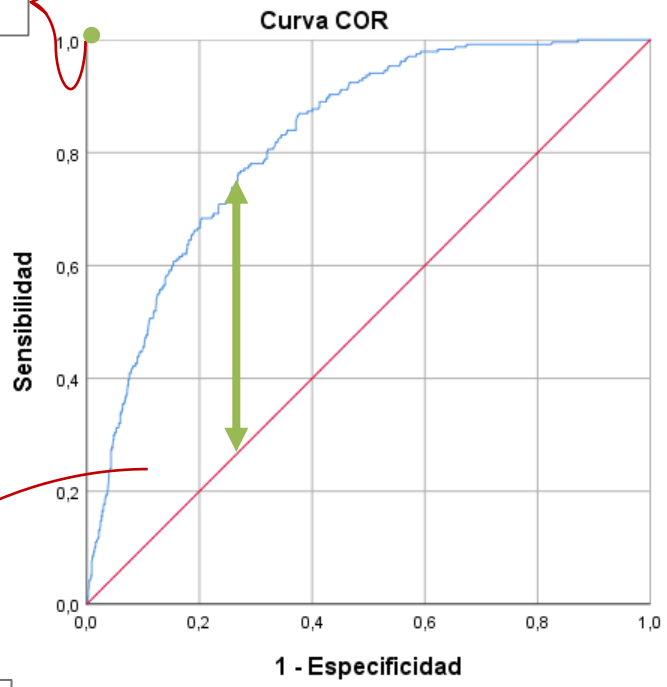




Estadísticos de bondad de ajuste

- 3) Curva ***COR (Característica de Operador Receptor)***: representa la **sensibilidad** (eje de las *y, ordenadas*) en función de **1- especificidad** (eje de las *x, abscisas)*

Este punto sería el de clasificación óptima



Los segmentos de diagonal se generan mediante empates.

La distancia vertical máxima desde la curva ROC a la diagonal principal es el **índice de Youden** y representa el valor óptimo de corte para la clasificación

Cuanto **mayor sea el área bajo la curva,** mejor es la capacidad clasificatoria del modelo

Ejercicio 4: *pasos para incorporar las covariables*

Método Hacia adelante condicional (Forward Step)

Este método utiliza un criterio estadístico para ir incorporando variables en la medida que supongan una mejora en el ajuste del modelo.

El principal criterio es la ***Puntuación de Rao.*** Entrará al modelo aquella cuyo *p-valor* en la Puntuación de Rao sea más pequeña. El supuesto es que su incorporación al modelo sería significativa.

```
OUTPUT CLOSE ALL.

USE ALL.
COMPUTE filter_$=(cat_ocup=3).
FILTER BY filter_$.
EXECUTE.

weight off.


LOGISTIC REGRESSION VARIABLES pobreza
  /METHOD=FSTEP(COND) asal_noreg d_mujer CH06 n_educ d_ninos d_gba
  /CONTRAST (asal_noreg)=Indicator(1)
  /CONTRAST (d_mujer)=Indicator(1)
  /CONTRAST (n_educ)=Indicator
  /CONTRAST (d_ninos)=Indicator(1)
  /CONTRAST (d_gba)=Indicator(1)
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

Método Hacia adelante condicional (Forward Step)

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-1,459	,072	409,031	1	,000	,233

Las variables no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	Condicion de registro en la seguridad social para asalariados(1)	109,079	1	,000
		Mujer(1)	1,320	1	,251
		¿Cuántos años cumplidos tiene?	9,587	1	,002
		Maximo nivel educativo alcanzado	123,593	3	,000
		Maximo nivel educativo alcanzado(1)	87,155	1	,000
		Maximo nivel educativo alcanzado(2)	4,823	1	,028
		Maximo nivel educativo alcanzado(3)	8,953	1	,003
		Presencia de niños en el hogar(1)	78,036	1	,000
		Partidos del Conurbano (1)	35,688	1	,000
	Estadísticos globales		255,855	8	,000

Puntuación de Rao

Significancia del estadístico

Método Hacia adelante condicional (Forward Step)

Bloque 1: Método = Avanzar por pasos (Condicional)

Pruebas ómnibus de coeficientes de modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	129,730	3	,000
	Bloque	129,730	3	,000
	Modelo	129,730	3	,000
Paso 2	Paso	65,868	1	,000
	Bloque	195,599	4	,000
	Modelo	195,599	4	,000
Paso 3	Paso	67,953	1	,000
	Bloque	263,552	5	,000
	Modelo	263,552	5	,000
Paso 4	Paso	6,181	1	,013
	Bloque	269,733	6	,000
	Modelo	269,733	6	,000

Nos informa la mejora que se produce por las variables que se incorporan en cada paso

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	,000	2	1,000
2	6,610	6	,358
3	9,767	8	,282
4	10,868	8	,209

Resumen del modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	1086,892[a]	,098	,158
2	1021,023[a]	,144	,232
3	953,070[a]	,189	,305
4	946,890[a]	,193	,312

a. La estimación ha terminado en el número de iteración 6 porque las estimaciones de parámetro han cambiado en menos de ,001.

Vemos cómo mejora la bondad de ajuste a medida que incorporamos covariables

El -2LL debería reducirse en cada paso

Método Hacia adelante condicional (Forward Step)

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Maximo nivel educativo alcanzado			102,019	3	,000	
	Maximo nivel educativo alcanzado(1)	2,352	,257	84,029	1	,000	10,512
	Maximo nivel educativo alcanzado(2)	1,722	,267	41,605	1	,000	5,597
	Maximo nivel educativo alcanzado(3)	,925	,305	9,170	1	,002	2,521
	Constante	-2,920	,230	161,857	1	,000	,054
Paso 2[b]	Condicion de registro en la seguridad social para asalariados(1)	1,289	,159	65,796	1	,000	3,628
	Maximo nivel educativo alcanzado			79,660	3	,000	
	Maximo nivel educativo alcanzado(1)	2,115	,262	65,248	1	,000	8,292
	Maximo nivel educativo alcanzado(2)	1,559	,272	32,850	1	,000	4,756
	Maximo nivel educativo alcanzado(3)	,793	,311	6,510	1	,011	2,209
	Constante	-3,241	,239	184,301	1	,000	,039
Paso 3[c]	Condicion de registro en la seguridad social para asalariados(1)	1,350	,167	65,456	1	,000	3,857
	Maximo nivel educativo alcanzado			71,518	3	,000	
	Maximo nivel educativo alcanzado(1)	2,014	,267	56,976	1	,000	7,495
	Maximo nivel educativo alcanzado(2)	1,351	,278	23,623	1	,000	3,863
	Maximo nivel educativo alcanzado(3)	,651	,318	4,191	1	,041	1,918
	Presencia de niños en el hogar(1)	1,402	,181	59,946	1	,000	4,065
	Constante	-4,015	,272	217,344	1	,000	,018
Paso 4[d]	Condicion de registro en la seguridad social para asalariados(1)	1,336	,167	63,646	1	,000	3,804
	Maximo nivel educativo alcanzado			58,893	3	,000	
	Maximo nivel educativo alcanzado(1)	1,859	,273	46,341	1	,000	6,416
	Maximo nivel educativo alcanzado(2)	1,216	,283	18,451	1	,000	3,373
	Maximo nivel educativo alcanzado(3)	,574	,321	3,195	1	,074	1,775
	Presencia de niños en el hogar(1)	1,385	,182	57,967	1	,000	3,993
	Partidos del Conurbano (1)	,557	,231	5,825	1	,016	1,745
	Constante	-4,335	,311	194,131	1	,000	,013

En el primer paso, introdujo la **educación** que tenía la mayor puntuación de Rao.
En el segundo paso, introdujo la **condición de registro**

En el tercer paso introdujo la **presencia de niños**
En el último paso, introdujo la **región**

Método Hacia adelante condicional (Forward Step)

Las variables no están en la ecuación

			Puntuación	gl	Sig.
Paso 1	Variables	Condicion de registro en la seguridad social para asalariados(1)	69,433	1	,000
		Mujer(1)	1,428	1	,232
		¿Cuántos años cumplidos tiene?	15,450	1	,000
		Presencia de niños en el hogar(1)	64,536	1	,000
		Partidos del Conurbano (1)	10,041	1	,002
	Estadísticos globales		135,155	5	,000
Paso 2	Variables	Mujer(1)	,214	1	,643
		¿Cuántos años cumplidos tiene?	7,684	1	,006
		Presencia de niños en el hogar(1)	65,106	1	,000
		Partidos del Conurbano (1)	8,167	1	,004
	Estadísticos globales		71,211	4	,000
Paso 3	Variables	Mujer(1)	,055	1	,814
		¿Cuántos años cumplidos tiene?	,810	1	,368
		Partidos del Conurbano (1)	5,924	1	,015
	Estadísticos globales		6,691	3	,082
Paso 4	Variables	Mujer(1)	,048	1	,827
		¿Cuántos años cumplidos tiene?	,689	1	,406
	Estadísticos globales		,779	2	,678

Confirmamos que, mediante un método por pasos, se han eliminado variables que no aportaban información relevante a nuestro modelo de análisis

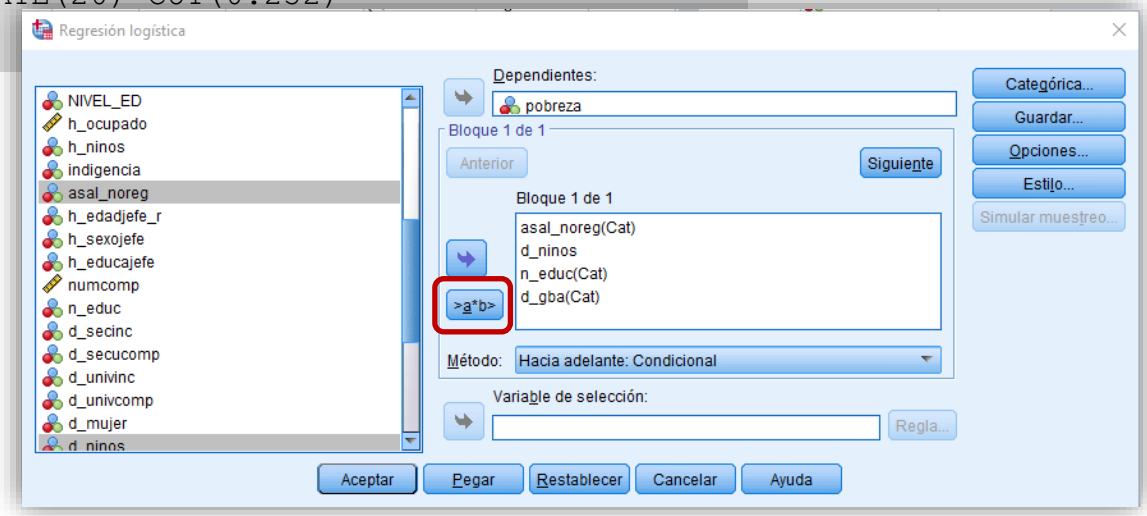
Ejercicio 5: *análisis con interacciones*

Interacciones en el modelo de regresión logística

Introducimos una interacción entre la condición de registro a la seguridad social y la presencia de niños/as.

En el caso del modelo logit, la interacción se elabora sencillamente incluyendo una multiplicación en la sintaxis o en ela ventana de comandos:

```
LOGISTIC REGRESSION VARIABLES pobreza
  /METHOD=ENTER asal_noreg d_ninos asal_noreg*d_ninos n_educ d_gba
  /CONTRAST (asal_noreg)=Indicator(1)
  /CONTRAST (n_educ)=Indicator
  /CONTRAST (d_ninos)=Indicator(1)
  /CONTRAST (d_gba)=Indicator(1)
   /PRINT=GOODFIT
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(0.252)
  /SAVE=PRED.
```



Interacciones en el modelo de regresión logística

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Condicion de registro en la seguridad social para asalariados(1)	1,806	,315	32,925	1	,000	6,086
	Presencia de niños en el hogar(1)	1,725	,273	39,787	1	,000	5,611
	Condicion de registro en la seguridad social para asalariados(1) by Presencia de niños en el hogar(1)	-,669	,372	3,230	1	,072	,512
	Maximo nivel educativo alcanzado			58,481	3	,000	
	Maximo nivel educativo alcanzado(1)	1,865	,273	46,612	1	,000	6,455
	Maximo nivel educativo alcanzado(2)	1,238	,283	19,123	1	,000	3,447
	Maximo nivel educativo alcanzado(3)	,607	,321	3,577	1	,059	1,834
	Partidos del Conurbano (1)	,555	,230	5,830	1	,016	1,743
	Constante	-4,611	,357	166,609	1	,000	,010

a. Variables especificadas en el paso 1: Condicion de registro en la seguridad social para asalariados, Presencia de niños en el hogar, Condicion de registro en la seguridad social para asalariados * Presencia de niños en el hogar , Maximo nivel educativo alcanzado, Partidos del Conurbano.

Este coeficiente ahora es cuánto se incrementa la razón de chances de ser pobre *entre los **informales**, cuando **no** se tienen niños, con respecto a los formales sin niños.*

Este coeficiente ahora es cuánto se incrementa la razón de chances de ser pobre *cuando **sí** se tienen niños entre los **formales,** con respecto a los formales que no tienen niños.*

Este coeficiente captura la interacción. Nos indica que la penalidad asociada *a tener niños sobre la pobreza es menor entre los informales que entre los trabajadores formales.*

Interacciones en el modelo de regresión logística

Tablas personalizadas

		asal_noreg Condicion de registro en la seguridad social para asalariados	
		.00 Registrados	1.00 No registrados
		PRE_1 Probabilidad pronosticada	PRE_1 Probabilidad pronosticada
		Media	Media
d_ninos Presencia de niños en el hogar	.00 Sin niños	,03922	,23567
	1.00 Con niños	,19780	,49730

Podemos calcular **brechas de probabilidad**: Entre los formales, la probabilidad de ser pobre pasa de **0,39** cuando no se tienen niños a **0,198** cuando se tienen (**5 veces**)

Entre los informales, la probabilidad de ser pobre pasa de **0,236** cuando no se tienen niños en el hogar a **0,497** cuando se tienen (**2,1 veces**)

El término de interacción en la regresión justamente nos está diciendo que la "penalidad" relativa de tener niños en el hogar sobre la probabilidad de ser pobre cuando se es **informal** **es alrededor de la mitad** de la que tienen **los formales**.

Interacciones en el modelo de regresión logística

Otra manera de ver las interacciones:

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	d_noreg_ninos(1)	2,861	,290	97,374	1	,000	17,487
	d_noreg_noninos(1)	1,806	,315	32,925	1	,000	6,086
	d_reg_ninos(1)	1,725	,273	39,787	1	,000	5,611
	Maximo nivel educativo alcanzado			58,481	3	,000	
	Maximo nivel educativo alcanzado(1)	1,865	,273	46,612	1	,000	6,455
	Maximo nivel educativo alcanzado(2)	1,238	,283	19,123	1	,000	3,447
	Maximo nivel educativo alcanzado(3)	,607	,321	3,577	1	,059	1,834
	Partidos del Conurbano (1)	,555	,230	5,830	1	,016	1,743
	Constante	-4,611	,357	166,609	1	,000	,010

a. Variables especificadas en el paso 1: d_noreg_ninos, d_noreg_noninos, d_reg_ninos, Maximo nivel educativo alcanzado, Partidos del Conurbano.

$$Interacción = \frac{\frac{17{,}487}{6.086}}{\frac{5.611}{1.000}} = \frac{2.87}{5.61} = 0.512$$

Análisis de regresión logística binaria con interacciones

- La reconstrucción de la ecuación de regresión nos facilita interpretar las interacciones.

$$\boldsymbol{Pr(y=1)=\frac{1}{1+e^{-z}}, donde\ z=a+bx+\cdots}$$

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	d_noreg_ninos(1)	2,861	,290	97,374	1	,000	17,487
	d_noreg_noninos(1)	1,806	,315	32,925	1	,000	6,086
	d_reg_ninos(1)	1,725	,273	39,787	1	,000	5,611
	Maximo nivel educativo alcanzado			58,481	3	,000	
	Maximo nivel educativo alcanzado(1)	1,865	,273	46,612	1	,000	6,455
	Maximo nivel educativo alcanzado(2)	1,238	,283	19,123	1	,000	3,447
	Maximo nivel educativo alcanzado(3)	,607	,321	3,577	1	,059	1,834
	Partidos del Conurbano (1)	,555	,230	5,830	1	,016	1,743
	Constante	-4,611	,357	166,609	1	,000	,010

a. Variables especificadas en el paso 1: d_noreg_ninos, d_noreg_noninos, d_reg_ninos, Maximo nivel educativo alcanzado, Partidos del Conurbano.

	No registrados con niños	No registrados sin niños	Registrados con niños	Registrados sin niños
D*X	2.861	1.806	1.725	0.000
Educación	1.238	1.238	1.238	1.238
Partidos	0.555	0.555	0.555	0.555
Constante	-4.611	-4.611	-4.611	-4.611
-z=-(a+bx)	**-0.044**	**1.012**	**1.093**	**2.818**
1 / 1 + e $^{-(z)}$	0.957	2.750	2.984	16.740
Pr (Y)	**51.1%**	**26.7%**	**25.1%**	**5.6%**

Nota: es necesario “fijar” los valores de los regresores en alguna categoría (por ejemplo: vive en Partidos del GBA).