Práctica de técnicas multivariantes II

Autores:

Melina Peressini Álvarez Guillermo Calvo Arenaza

Objetivo de la práctica

El objetivo de este trabajo es el de poder predecir mediante una regresión logística, cuánto ganará una persona en función de una serie de variables, siendo el resultado final si gana o no más de 50.000 dólares al año (0 si es < de 50.000$ ó 1 si es >50.000$).

Base de datos

Para este trabajo se ha decidido trabajar con una base de datos extraída de la web de Kaggle, siendo la base de datos seleccionada:

https://www.kaggle.com/lodetomasi1995/income-classification

Nota sobre la reducción de la base de datos
Debido a la gran cantidad de datos, se ha procedido a recategorizar la variable que hace referencia a la nacionalidad de la persona, de modo que tomase 0 cuando no es americano y 1 cuando es americano. Tras esto se han seleccionado de forma aleatoria y balanceada los datos en función de si es americano o no, dando como resultado una composición de 50% americanos y 50% no americanos.

Estructura de la base de datos

Una vez categorizado y modificado las variables (eliminado y creando variables nuevas) no queda la siguiente estructura de la base de datos.

Nota sobre las variables
A continuación, se presentará una tabla con un total de X variables, de las cuales solo las que tengan un asterisco (*) al lado del nombre, se emplearán para el modelo de regresión logística para así poder determinar medidas de asociación con el fin de comprender la naturaleza de los datos y evitar problemas de multicolinealidad entre las variables del modelo.

Variables de la base de datos y naturaleza:

Nombre de la variable	En el modelo	Tipo	Nº categorías
Age	Si	Continua	-
Age_cat	No	Categórica ordinal	5
Hoursperweek	Si	Continua	-
hoursperweek_cat	No	Categórica ordinal	3
Sexo_cat	Si	Dicotómica	2
wordclass_cat	No	Categórica nominal	9
education_cat	No	Categórica ordinal	16
maritalstatus_cat	No	Categórica nominal	7
occupation_cat	No	Categórica nominal	15
race_cat	Si	Categórica nominal	5
income_cat	Respuesta	Categórica ordinal	2
Americano	No	Dicotómica	2
pierde_en_inversion	No	Dicotómica	2
gana_en_inversion	No	Dicotómica	2
diferencia	No	Continua	-

Una vez vista la estructura de la base de datos, procederemos a calcular los estadísticos básicos de las variables separadas por tipo (continua, categórica y dicotómica) con el fin de tener una visión más amplia de la realidad de los datos.

Nota sobre las variables age_cat y hoursperweek_cat

Para poder realizar algunas de las medidas de asociación se ha procedido a categorizar de forma ordinal dichas variables, siendo el criterio de categorización para las variables:

Edad:

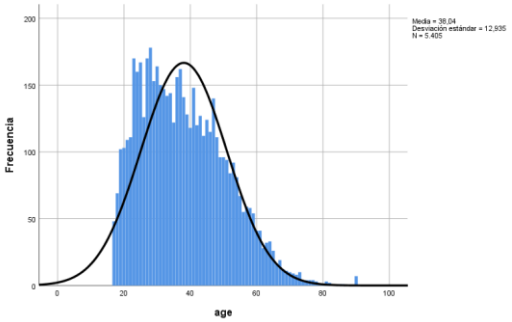
$$\{1\ si\ edad\ \in [17,24]\ 2\ si\ edad\ \in [25,40]\ 3\ si\ edad\ \in [41,64]\ 4\ si\ edad\ \in [65,90]$$

Horas de trabajo:

$$\{1\ si\ horas\ \in [0,19]\ 2\ si\ horas\ \in [20,40]\ 3\ si\ horas\ \in [41,99]$$

Análisis descriptivo básico

Age: Edad

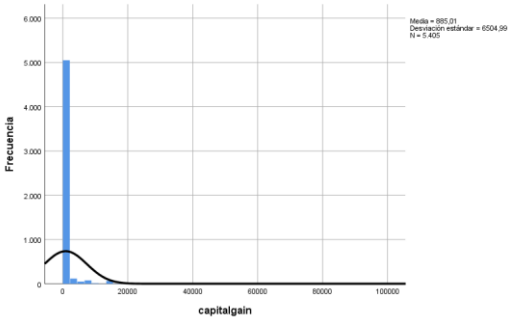


Como se puede observar, la variable edad está truncada de forma natural, debido a que la edad de las personas está comprendida en el intervalo de 17 a 90 años aproximadamente. De forma que la media de la edad de las personas es de 38 años con una desviación estándar de 13 años, por lo que estamos ante una muestra de la población joven.

Estadísticos descriptivos

	N	Rango	Mínimo	Máximo	Media	Desv. Desviación
age	5405	73	17	90	38,04	12,935
N válido (por lista)	5405					

Capitalgain (Capital ganado)



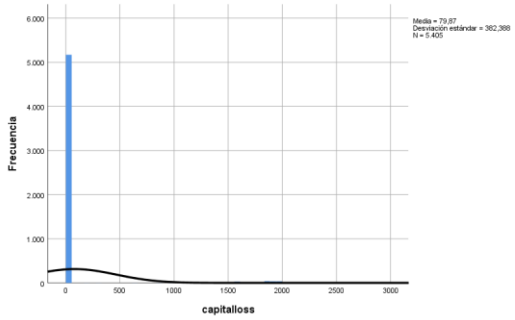
Se puede observar que el capital ganado de las personas que componen esta base de datos es muy dispar, puesto que claramente se observa que mucho de ellos toman valores muy bajos (e incluso siendo valores de 0 por lo que no se podría transformar en una variable logarítmica y por ende poder arreglar el problema de outliers), mientras que existen pocas observaciones outliers que toman valores muy extremos. Esto además se puede observar en la media que es muy baja en comparación con la desviación estándar

Estadísticos descriptivos

	N	Rango	Mínimo	Máximo	Media	Desv. Desviación
capitalgain	5405	99999	0	99999	885,01	6504,990
N válido (por lista)	5405					

Se puede observar que el capital ganado de las personas que componen esta base de datos es muy dispar, puesto que claramente se observa que mucho de ellos toman valores muy bajos (e incluso siendo valores de 0 por lo que no se podría transformar en una variable logarítmica y por ende poder arreglar el problema de outliers), mientras que existen pocas observaciones outliers que toman valores muy extremos. Esto además se puede observar en la media que es muy baja en comparación con la desviación estándar.

Capitalloss (capital perdido)



De nuevo nos encontramos ante la misma situación que en la variable de capitalgain (capital ganado), donde podemos observar que existen muchos valores que toman valores cercanos a 0 (o incluso que tomen el valor 0) y algunos casos donde toman valores muy extremos. De nuevo, debido a la presencia de valores 0, no se puede plantear una transformación logarítmica.

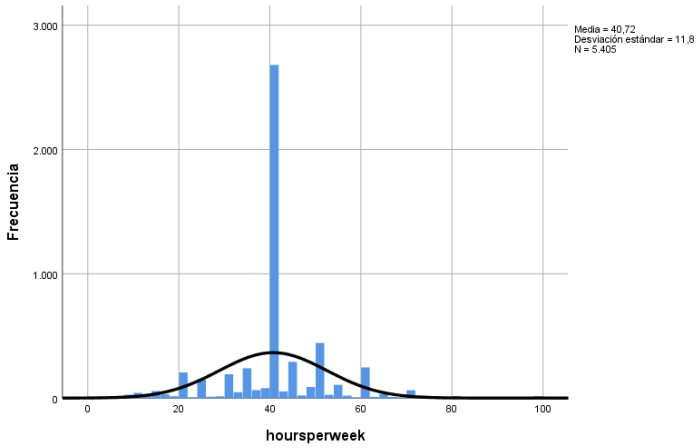
Estadísticos descriptivos

	N	Rango	Mínimo	Máximo	Media	Desv. Desviación
capitalloss	5405	2603	0	2603	79,87	382,388
N válido (por lista)	5405					

Nota sobre las variables capitalgain y capitalloss

Como se ha podido observar, las dos variables toman valores cercanos o iguales a 0 por lo que puede que estas no aporten información significativa al modelo es por ello que nos podremos plantear el modelo con o sin estas variables. Peor esto ya se verá más adelante en este trabajo.

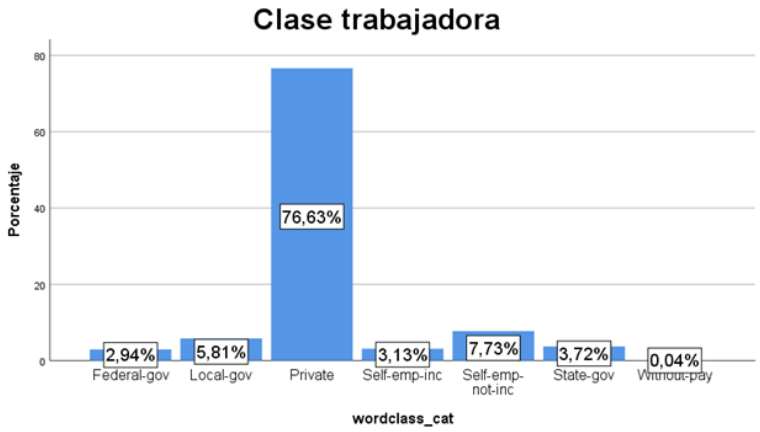
Hoursperweek (horas de trabajo a la semana)



Estadísticos descriptivos

	N	Rango	Mínimo	Máximo	Media	Desv. Desviación
hoursperweek	5405	98	1	99	40,72	11,800
N válido (por lista)	5405					

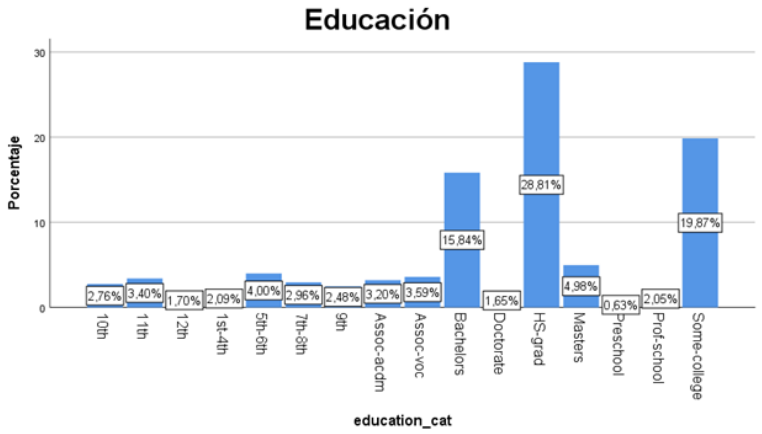
Wordclass_cat (categoría de trabajo)



wordclass_cat

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Federal-gov	159	2,9	2,9	2,9
	Local-gov	314	5,8	5,8	8,8
	Private	4142	76,6	76,6	85,4
	Self-emp-inc	169	3,1	3,1	88,5
	Self-emp-not-inc	418	7,7	7,7	96,2
	State-gov	201	3,7	3,7	100,0
	Without-pay	2	,0	,0	100,0
	Total	5405	100,0	100,0	

Education_cat (estudios que se tienen categorizados)



education_cat

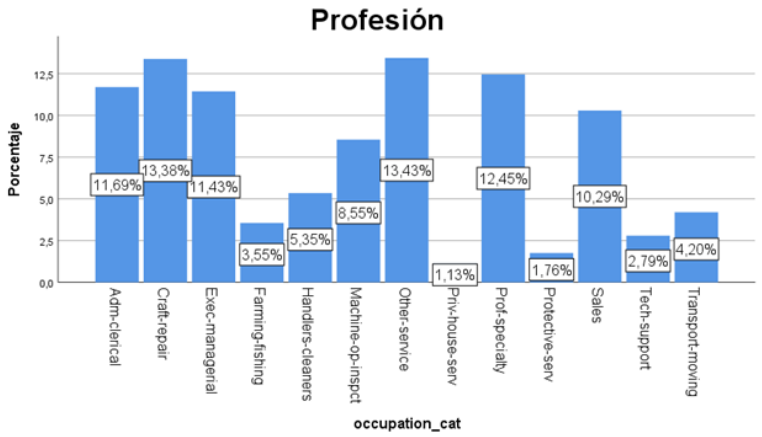
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	10th	149	2,8	2,8	2,8
	11th	184	3,4	3,4	6,2
	12th	92	1,7	1,7	7,9
	1st-4th	113	2,1	2,1	10,0
	5th-6th	216	4,0	4,0	14,0
	7th-8th	160	3,0	3,0	16,9
	9th	134	2,5	2,5	19,4
	Assoc-acdm	173	3,2	3,2	22,6
	Assoc-voc	194	3,6	3,6	26,2
	Bachelors	856	15,8	15,8	42,0
	Doctorate	89	1,6	1,6	43,7
	HS-grad	1557	28,8	28,8	72,5
	Masters	269	5,0	5,0	77,4
	Preschool	34	,6	,6	78,1
	Prof-school	111	2,1	2,1	80,1
	Some-college	1074	19,9	19,9	100,0
	Total	5405	100,0	100,0	

Maritalstatus_cat (estado civil categorizado)

maritalstatus_cat

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Divorced	597	11,0	11,0	11,0
	Married-civ-spouse	2543	47,0	47,0	58,1
	Married-spouse-absent	170	3,1	3,1	61,2
	Never-married	1762	32,6	32,6	93,8
	Separated	183	3,4	3,4	97,2
	Widowed	150	2,8	2,8	100,0
	Total	5405	100,0	100,0	

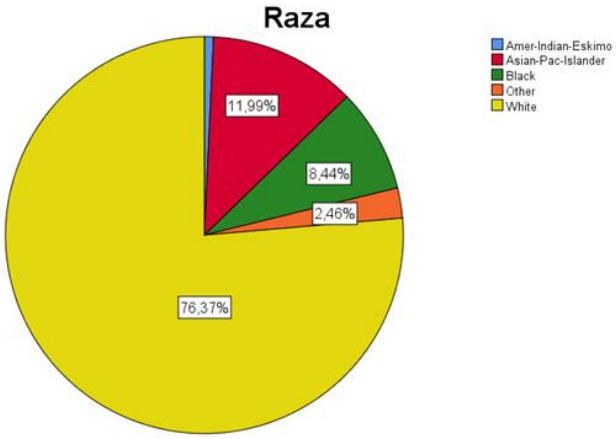
occupation_cat (profesion)



occupation_cat

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Adm-clerical	632	11,7	11,7	11,7
	Craft-repair	723	13,4	13,4	25,1
	Exec-managerial	618	11,4	11,4	36,5
	Farming-fishing	192	3,6	3,6	40,1
	Handlers-cleaners	289	5,3	5,3	45,4
	Machine-op-inspct	462	8,5	8,5	54,0
	Other-service	726	13,4	13,4	67,4
	Priv-house-serv	61	1,1	1,1	68,5
	Prof-specialty	673	12,5	12,5	81,0
	Protective-serv	95	1,8	1,8	82,7
	Sales	556	10,3	10,3	93,0
	Tech-support	151	2,8	2,8	95,8
	Transport-moving	227	4,2	4,2	100,0
	Total	5405	100,0	100,0	

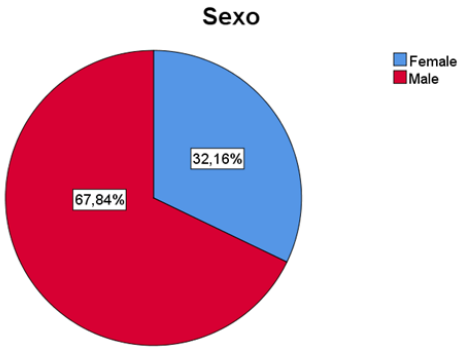
Race (raza)



race_cat

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Amer-Indian-Eskimo	40	,7	,7	,7
	Asian-Pac-Islander	648	12,0	12,0	12,7
	Black	456	8,4	8,4	21,2
	Other	133	2,5	2,5	23,6
	White	4128	76,4	76,4	100,0
	Total	5405	100,0	100,0	

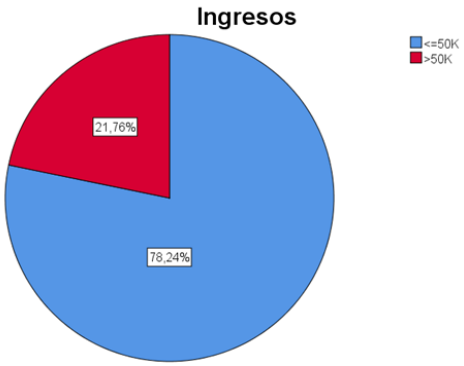
Sexo_cat (sexo categorizado por 0 y 1)



sexo_cat

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Female	1738	32,2	32,2	32,2
	Male	3667	67,8	67,8	100,0
	Total	5405	100,0	100,0	

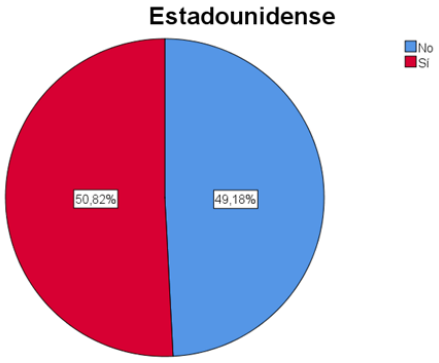
income_cat (Ingresos)



income_cat

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	<=50K	4229	78,2	78,2	78,2
	>50K	1176	21,8	21,8	100,0
	Total	5405	100,0	100,0	

Estadounidense



Americano

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	0	2658	49,2	49,2	49,2
	1	2747	50,8	50,8	100,0
	Total	5405	100,0	100,0	

Medidas de asociación

Una vez visto el análisis descriptivo básico de las variables que componen la base de datos, trabajaremos con las medidas asociación, distinguiendo en función de la naturaleza de las variables:

- Entre dos variables nominales dicotómicas
- Entre dos variables nominales sin restricción en el número de categorías
- Entre variables categóricas ordinales
- Entre una variable cuantitativa y una variable explicativa categórica
- Entre dos variables cuantitativas

De esta forma, estudiaremos la relación entre diferentes pares de variables de interés.

Entre dos variables dicotómicas: ingresos anuales y sexo

Resulta de interés analizar si existe una relación entre el sexo y los ingresos anuales que perciba un individuo; es decir, si existen diferencias en los ingresos anuales percibidos en función de si la persona es hombre o mujer. En tanto que nuestra variable relativa a los ingresos anuales presenta dos categorías (ingresos menores o iguales a 50.000$ anuales e ingresos superiores a 50.000$ anuales), estamos tratando con dos variables categóricas que son, ambas, dicotómicas (presentan únicamente dos categorías cada una de ellas).

income_cat*sexo_cat tabulación cruzada

Recuento

		sexo_cat		
		Female	Male	Total
income_cat	<=50K	1561	2668	4229
	>50K	177	999	1176
Total		1738	3667	5405

1. **Coeficiente Phi**

El coeficiente Phi es una medida de asociación basada en el estadístico Chi-cuadrado de Pearson para el caso particular de un par de variables categóricas dicotómicas (es decir, con únicamente dos categorías cada una de ellas). Se obtiene como la raíz cuadrada del cociente del estadístico Chi-cuadrado entre el número total de observaciones (N). El coeficiente Phi puede tomar valores comprendidos en el intervalo [0, 1]: valores próximos a cero indican ausencia de asociación entre las variables consideradas y valores próximos a uno indican una asociación fuerte entre ellas.

Por tanto, el coeficiente Phi nos puede permitir estudiar la relación entre los ingresos anuales y el sexo. Toma un valor de **0,193**, cercano a 0,2 y por tanto más próximo a cero que a uno. Este valor del coeficiente Phi indica, por tanto, que existe una asociación entre los ingresos anuales percibidos y el sexo, pero que esta asociación es de grado débil.

Medidas simétricas

		Valor	Aprox. Sig.
Nominal por Nominal	Phi	,193	,000
	V de Cramer	,193	,000
N de casos válidos		5405	

2. **Riesgo relativo**

Otra medida de asociación que nos permite estudiar la relación entre estas dos variables es el riesgo relativo. En este caso, la variable correspondiente a los ingresos anuales juega el papel de variable respuesta y la variable correspondiente al sexo juega el rol de variable explicativa.

Estimación de riesgo

	Valor	Intervalo de confianza de 95 %	
		Inferior	Superior
Odds ratio para sexo_cat (Female / Male)	3,302	2,782	3,920
Para cohorte income_cat = <=50K	1,234	1,204	1,266
Para cohorte income_cat = >50K	,374	,322	,434
N de casos válidos	5405		

La proporción de mujeres que ganan como máximo 50.000$ anuales es 1561 / 1738 = 0,898 (es decir, prácticamente un 90%).

El el caso de los hombres, la proporción de aquellos que ganan 50.000$ al año o menos también es mayor a aquella correspondiente a los que ganan más de 50.000$ anuales, suponiendo un 0,728 (2668 / 3667).

La estimación del **riesgo relativo** se obtiene como el cociente de estas proporciones y toma, en este caso, el valor **1,234**. Ello nos indica que la probabilidad de percibir ingresos inferiores o iguales a 50.000$ anuales siendo mujer es 1,234 veces la probabilidad de percibir estos ingresos anuales siendo hombre. El intervalo al 95% de confianza es de (1,204 , 1,266); no contiene al uno, por lo que puede concluirse que la relación entre variables es estadísticamente significativa.

Por el contrario, la proporción de mujeres que ganan más de 50.000$ anuales es 177 / 1738 = 0,102; y la de los hombres que ganan más de 50.000$ anuales es 999 / 3667 = 0,272 (es decir, cerca de un 10% y cerca de un 3%, respectivamente).

Por tanto, el **riesgo relativo** correspondiente a los ingresos superiores a 50.000$ toma el valor **0,374**. Ello implica que la probabilidad de percibir ingresos superiores a 50.000$ siendo mujer es 0,374 veces la probabilidad de percibir ingresos superiores a 50.000$ siendo hombre (por tanto, más pequeña). El intervalo al 95% de confianza es de (0,322 , 0,434). Tampoco en este caso el intervalo de confianza contiene al uno, de manera que la relación entre variables se considera estadísticamente significativa.

3. Razón de probabilidades y ods ratio

Para las mujeres, la **razón de probabilidades** toma el valor **8,819**: es decir, la proporción de aquellas que ganan como máximo 50.000$ al año es 8,819 veces la proporción de aquellas que ganan más de 50.000$ anuales (1561 / 177 = 8,819).

Para los hombres, el valor la **razón de probabilidades** es **2,671**: es decir, la proporción de aquellos que ganan como máximo 50.000$ al año es 2,671 veces la proporción de aquellos que ganan más de 50.000$ anuales (2668 / 999 = 2,671).

En ambos casos, por tanto, encontramos proporciones mayores en el intervalo correspondiente a los menores ingresos anuales.

El **ods ratio** se obtiene como resultado del cociente de estas dos razones de probabilidades, es decir (1561 / 177) / (2668 / 999), o bien 8,819 / 2,671. Esta medida de asociación toma el valor **3,302** e indica que la probabilidad estimada de ganar como máximo 50.000$ al año frente a ganar más de 50.000$ al año es, en las mujeres, 3,302 veces la probabilidad estimada para los hombres. El intervalo al 95% de confianza es (2,782 , 2,671); no contiene al uno, por lo que la relación entre ambas variables es estadísticamente significativa.

Por tanto, estas medidas de asociación (**riesgo relativo** y **ods ratio**) nos permiten complementar la información otorgada por el **coeficiente Phi** de manera que podamos ver en qué consiste la relación dada entre ambas variables. De forma general, tanto mujeres como para hombres, son más frecuentes los ingresos de 50.000$ o inferiores, mientras que percibir unos ingresos superiores a 50.000$ resulta menor habitual. Sin embargo, este desbalance es más acusado para las mujeres que para los hombres; es decir, es menos probable que una mujer gane más de 50.000$ a que lo haga un hombre.

Entre dos variables categóricas sin restricción en el número de categorías

En las siguientes tablas encontraremos información sobre la asociación entre dos variables categóricas con diferentes niveles siendo dichas variables la correspondiente a los ingresos anuales (con dos categorías) y la raza (con cinco categorías). En primer lugar, la tabla cruzada correspondiente es la siguiente:

income_cat*race_cat tabulación cruzada

Recuento

		race_cat					
		Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White	Total
income_cat	<=50K	37	464	401	122	3205	4229
	>50K	3	184	55	11	923	1176
Total		40	648	456	133	4128	5405

A continuación, se presentan los valores obtenidos para las diferentes medidas de asociación que nos permiten estudiar la relación entre la raza y los ingresos anuales.

Medidas direccionales

			Valor	Error estándar asintótico[a]	Aprox. S[d]	Aprox. Sig.
Nominal por Nominal	Lambda	Simétrico	,000	,000	.[b]	.[b]
		income_cat dependiente	,000	,000	.[b]	.[b]
		race_cat dependiente	,000	,000	.[b]	.[b]
	Tau Goodman y Kruskal	income_cat dependiente	,011	,002		,000[c]
		race_cat dependiente	,002	,001		,000[c]
	Coeficiente de incertidumbre	Simétrico	,010	,002	4,456	,000[e]
		income_cat dependiente	,012	,003	4,456	,000[e]
		race_cat dependiente	,008	,002	4,456	,000[e]

Medidas simétricas

		Valor	Aprox. Sig.
Nominal por Nominal	Phi	,107	,000
	V de Cramer	,107	,000
	Coeficiente de contingencia	,106	,000
N de casos válidos		5405	

El **coeficiente de contingencia** se basa en el estadístico Chi-cuadrado de Pearson. Se trata de un coeficiente que toma valores comprendidos entre 0 y 1, si bien la cota superior depende del número de categorías de cada una de las dos variables consideradas. Esta cota se calcula de la siguiente manera.

$$C_{\max} = \sqrt{\frac{\min(I-1, J-1)}{1+\min(I-1, J-1)}}$$

Valores próximos a 0, por tanto, indican ausencia de asociación entre la raza y los ingresos anuales percibidos; mientras que valores próximos a la cota superior indicarán una fuerte asociación. En este caso, la cota superior toma un valor de **0,707** y el coeficiente de contingencia toma un valor de **0,106**, indicando la existencia de relación muy débil entre la raza y los ingresos anuales.

La **V de Cramer** se obtiene, al igual que el coeficiente de contingencia, a partir del estadístico Chi-cuadrado de Pearson. Toma, con independencia del número de categorías de las variables consideradas, valores entre 0 y 1. Valores próximos a 0 indican ausencia de asociación, mientras que valores próximos a 1 indican asociación fuerte. En nuestro caso, toma el valor **0.107**, cercano a 0, si bien debe tenerse en cuenta que esta medida de asociación tiende a subestimar el grado de asociación entre las variables. Nos indica, nuevamente, una relación muy débil entre la raza y los ingresos anuales percibidos.

Respecto a la medida de asociación **Lambda**, nos interesa particularmente su expresión asimétrica, de manera que nos indica en qué proporción se reduce el error al predecir los ingresos anuales cuando se considera la información proporcionada por la variable correspondiente a la raza. En este caso, se obtiene un valor de Lambda igual a **0**. Sabemos que puede ocurrir que esta medida de asociación toma el valor 0 en casos en los que la independencia entre las dos variables no sea total. Sin embargo, de este resultado podemos deducir en cualquier caso que, de existir asociación, ésta es de grado muy débil.

En el caso del **coeficiente de incertidumbre**, de similar interpretación a Lambda, nos interesa nuevamente su expresión asimétrica en la que los ingresos anuales juegan el papel de variable respuesta y la raza el de variable explicativa. En este caso, el coeficiente de incertidumbre toma el valor **0,012**, que se interpreta como la proporción de incertidumbre reducida al predecir los ingresos anuales a partir de la raza. Como se obtiene un valor próximo a 0, deducimos una asociación muy débil entre la raza y los ingresos anuales.

El **coeficiente Tau-y de Goodman y Kruskal**, al igual que lambda o el coeficiente de incertidumbre es un estadístico basado en el error. Toma valores entre 0 y 1 y se interpreta como la proporción de mejora obtenida en la predicción de los valores de una de las variables a partir de los valores de la otra, frente a la predicción cuando únicamente se considera la información de la propia variable. Si suponemos que los ingresos anuales juegan el papel de variable respuesta y que la raza es variable explicativa, obtenemos un valor para este coeficiente de **0,11**; es decir, la predicción sobre los

ingresos anuales mejora en un 11% cuando se considera la información proporcionada por la variable relativa a la raza.

Con toda la información anterior, podemos sacar en claro que los ingresos de las personas no dependen de la raza a la cual pertenezcan siendo otros factores los que implican mayor o menor ingresos para el individuo.

La medida de asociación **Kappa** es apropiada cuando ambas variables toman los mismos valores (lo que implica, necesariamente pero no sólo, que tengan el mismo número de categorías). Mide el grado de acuerdo entre los valores de las dos variables tomando valores en el intervalo [-1,1]. Valores próximos a 1 indican que las variables están relacionadas y se presenta una total concordancia en las respuestas a las dos variables. Por el contrario, valores próximos a -1 indican que, aunque las variables están relacionadas, existe una total discordancia en las respuestas. Por último, valores próximos a cero indican que las variables son independientes.

Se calcula a partir de la proporción de acuerdos observados (suma de los casos que presentan el **mismo valor** en ambas variables, correspondientes a la diagonal principal de la tabla de contingencia) y la proporción de acuerdos esperados de forma azarosa.

Dado que no contábamos en nuestra base de datos con dos variables que tomasen los mismos valores, hemos hecho algunas modificaciones para conseguirlo y poder realizar un ejemplo de la medida de asociación Kappa. Partiendo de las variables continuas correspondientes al capital ganado y al capital perdido hemos, primero, forzado a que ambas tengan el mismo rango (seleccionando únicamente las observaciones en las que ambas tuvieran valores iguales o inferiores a 2500 €). En segundo lugar, las hemos categorizado por intervalos: [0, 1000] y [1001, 2500]. La tabla cruzada de contingencia que se obtiene es la siguiente.

capitalgain*capitalloss tabulación cruzada

Recuento

		capitalloss		
		[0, 1000]	[1001, 2500]	Total
capitalgain	[0, 1000]	4783	224	5007
	[1001, 2500]	49	0	49
Total		4832	224	5056

Medidas simétricas

		Valor	Error estándar asintótico[a]	Aprox. S[b]	Aprox. Sig.
MEdida de acuerdo	Kappa	-,016	,002	-1,515	,130
N de casos válidos		5056			

a. No se supone la hipótesis nula.

b. Utilización del error estándar asintótico que asume la hipótesis nula.

El valor que se obtiene para la medida de asociación Kappa es de **-0,016**. Por tanto, ello indica independencia entre las dos variables.

Entre variables categóricas ordinales: edad y número de horas trabajadas por semana

Ambas variables se han generado a partir de la categorización de las variables de naturaleza continua con que contaba la base de datos. Tras esta categorización, se obtiene la siguiente tabla cruzada de contingencia.

hoursperweek (Agrupada)*age (Agrupada) tabulación cruzada

Recuento

		age (Agrupada)				
		[17, 24] años	[25, 40] años	[41, 64] años	[65, 90] años	Total
hoursperweek (Agrupada)	20 horas o menos	188	82	82	60	412
	De 21 a 40 horas	578	1560	1276	76	3490
	Más de 40 horas	106	726	650	21	1503
Total		872	2368	2008	157	5405

Las medidas de asociación para escala ordinal permiten estudiar el signo de la asociación entre ambas variables. Es decir, si existe una asociación positiva: valores más altos en una de las variables se asocian a valores más altos en la otra (con concentración de frecuencias en la diagonal principal de la tabla cruzada de contingencia); o bien una asociación negativa: valores más altos en una de las variables se asocian a valores más bajos en la otra (con concentración de frecuencias de la diagonal secundaria de la tabla cruzada de contingencia).

Las medidas de asociación **coeficiente Gamma**, **Tau-b y Tau-c de Kendall** y **D de Sommers** se construyen a partir del número de **observaciones concordantes** (P) y el número de **observaciones discordantes** (Q).

$$P = \sum_{i=1}^{I-1}\sum_{j=1}^{J-1} n_{ij} \sum_{i'>i}\sum_{j'>j} n_{i'j'} \qquad Q = \sum_{i=1}^{I-1}\sum_{j=2}^{J} n_{ij} \sum_{i'>i}\sum_{j'<j} n_{i'j'}$$

En este sentido, la prevalencia de observaciones concordantes sobre discordantes corresponde a una asociación positiva; y lo contrario a una asociación negativa entre las variables.

Medidas direccionales

			Valor	Error estándar asintótico[a]	Aprox. S[b]	Aprox. Sig.
Nominal por Nominal	Lambda	Simétrico	,021	,003	6,476	,000
		age (agrupado) dependiente	,035	,005	6,476	,000
		hoursperweek (Agrupada) dependiente	,000	,000	.[c]	.[c]
	Tau Goodman y Kruskal	age (agrupado) dependiente	,029	,003		,000[d]
		hoursperweek (Agrupada) dependiente	,027	,003		,000[d]
Ordinal por ordinal	d de Somers	Simétrico	,112	,013	8,463	,000
		age (agrupado) dependiente	,128	,015	8,463	,000
		hoursperweek (Agrupada) dependiente	,099	,012	8,463	,000

Medidas simétricas

		Valor	Error estándar asintótico [a]	Aprox. S [b]	Aprox. Sig.
Ordinal por ordinal	Tau-b de Kendall	,113	,013	8,463	,000
	Tau-c de Kendall	,096	,011	8,463	,000
	Gamma	,194	,023	8,463	,000
N de casos válidos		5405			

El **coeficiente Gamma** es la más sencilla de estas medidas. Se calcula como:

$$\gamma = \frac{P - Q}{P + Q} \quad -1 \leq \gamma \leq 1$$

Valores próximos a 1 indican fuerte asociación positiva, mientras que valores próximos a -1 indican fuerte asociación negativa. Valores próximos a cero indican que no se da una relación positiva ni negativa entre las variables. Puede existir, no obstante, asociación de otro tipo entre las variables. En nuestro caso, el coeficiente Gamma toma un valor de **0.194**.

La **Tau-b de Kendall** se interpreta de manera similar al coeficiente Gamma. Debe tenerse en cuenta que, para esta medida de asociación, los valores -1 y 1 sólo pueden alcanzarse cuando ambas variables presentan el mismo número de categorías; y sólo se alcanzan cuando, además, ambas variables se encuentran en situación de total asociación (negativa o positiva, respectivamente).

En nuestro caso, las variables consideradas no tienen el mismo número de categorías, por lo que sabemos que los valores -1 y 1 no podrán ser alcanzados. Obtenemos un valor para la tau-b de Kendall de **0,113**.

La **Tau-c de Kendall** supone una corrección de la Tau-b de Kendall, que presenta la ventaja de que sí permite que los valores -1 y 1 sean alcanzados cuando las variables estudiadas presentan diferente número de categorías. Como desventaja, tiende a subestimar el grado de asociación entre las variables. Obtenemos para la Tau-c de Kendall un valor de **0,096**.

La **D de Sommers** presenta una versión asimétrica (en el que una de las dos variables se considera como variable respuesta y la otra como variable explicativa) y una versión simétrica. En ambos casos, toma valores entre -1 y 1 y la interpretación es similar a la de las medidas presentadas anteriormente. Presenta el inconveniente de que los valores -1 y 1 pueden ser alcanzados en situaciones en las que la relación entre las variables (negativa o positiva, respectivamente) no es total. El valor de la D de Sommers simétrica siempre está comprendido entre los valores obtenidos para la D de Sommers asimétrica tomando cada una de las variables como variable respuesta.

Los valores que obtenemos para la D de Sommers son **0,099** (cuando la variable de **horas trabajadas por semana** juega el rol de variable dependiente), **0,128** (cuando la variable de **edad** hace de variable dependiente) y **0,112** en la versión simétrica.

Todas estas medidas de asociación toman valores positivos, pero próximos a cero. Por tanto, nos indican práctica ausencia de relación positiva o negativa entre ambas variables.

Entre una variable cuantitativa y una variable explicativa categórica: edad y estado civil

El **coeficiente Eta** permite estudiar la relación entre una variable cuantitativa y una variable categórica. En este marco, la variable cuantitativa juega el papel de variable respuesta y el cuadrado del coeficiente Eta corresponde a la proporción de su variabilidad que es explicada por la variable categórica.

Para esta medida de asociación, trabajaremos con la edad (variable cuantitativa, sin categorizar) y el estado civil de las personas (variable categórica). El coeficiente Eta nos proporciona información para dar respuesta a la pregunta: ¿se observan diferencias en la edad de las personas en función de su estado civil?

Debido la gran dimensión de la tabla de valores cruzadas entre la edad y el estado matrimonial de las personas, se ha optado por no mostrar la tabla. En lo que a la medida de asociación eta de dicha tabla cruzada obtenemos los siguiente resultados.

Medidas direccionales

			Valor
Nominal por intervalo	Eta	age dependiente	,544
		maritalstatus_cat dependiente	,413

El valor del coeficiente Eta que nos interesa es el que corresponde a la variable edad (cuantitativa) como variable respuesta o dependiente; es decir, **0,544**. Por tanto, **0,296** (su cuadrado) es la proporción de variabilidad de la edad que es explicada por las diferencias dadas en el estado civil.

Entre dos variables cuantitativas: edad y número de horas trabajadas por semana

Los coeficientes de correlación de Pearson y de Spearman son medidas del grado de correlación lineal entre dos variables cuantitativas. Ambos toman valores comprendidos en [1, 1], con similar interpretación: valores próximos a 1 indican fuerte correlación lineal positiva, valores próximos a -1 indican fuerte correlación lineal negativa y valores próximos a 0 indican ausencia o debilidad en la correlación lineal.

Las siguientes expresiones de ambos permiten obtenerlos cuando no se dispone de los datos desagregados, sino de forma resumida mediante las frecuencias correspondientes a cada valor. Para el coeficiente de correlación de Pearson, la expresión es:

$$r = \frac{\sum_{j=1}^{J}\sum_{i=i}^{I} \mathrm{n}_{ij}\cdot(x_i-\bar{X})\cdot(\mathrm{y}_j-\bar{Y})}{\sqrt{\sum_{i=i}^{I}\mathrm{n}_{i+}\cdot(x_i-\bar{X})^2}\cdot\sqrt{\sum_{j=i}^{J}\mathrm{n}_{+j}\cdot(\mathrm{y}_j-\bar{Y})^2}}$$

Para el coeficiente de correlación de Spearman, la expresión es (donde *rxi* y *ryj* son los rangos asignados a los valores ordenados):

$$r = \frac{\sum_{j=1}^{J}\sum_{i=i}^{I} \mathrm{n}_{ij}\cdot(\mathrm{r}\,x_i-\bar{X})\cdot(\mathrm{ry}_j-\bar{Y})}{\sqrt{\sum_{i=i}^{I}\mathrm{n}_{i+}\cdot(\mathrm{r}\,x_i-\bar{X})^2}\cdot\sqrt{\sum_{j=i}^{J}\mathrm{n}_{+j}\cdot(\mathrm{ry}_j-\bar{Y})^2}}$$

Utilizamos estos coeficientes para estudiar la relación entre la variable correspondiente a la edad y la variable correspondiente a la jornada laboral semanal (en horas).

Para ambos coeficientes obtenemos resultados positivos, si bien cercanos al 0: el coeficiente de correlación de Pearson vale **0,087** y el coeficiente de correlación de Spearman toma el valor **0,157**. Nos indican, por tanto, una relación muy débil entre la edad y el número de horas trabajadas a la semana. Esta relación, a pesar de su debilidad, es de orden positivo, es decir: a mayor edad, mayor número de horas trabajadas a la semana.

Medidas simétricas

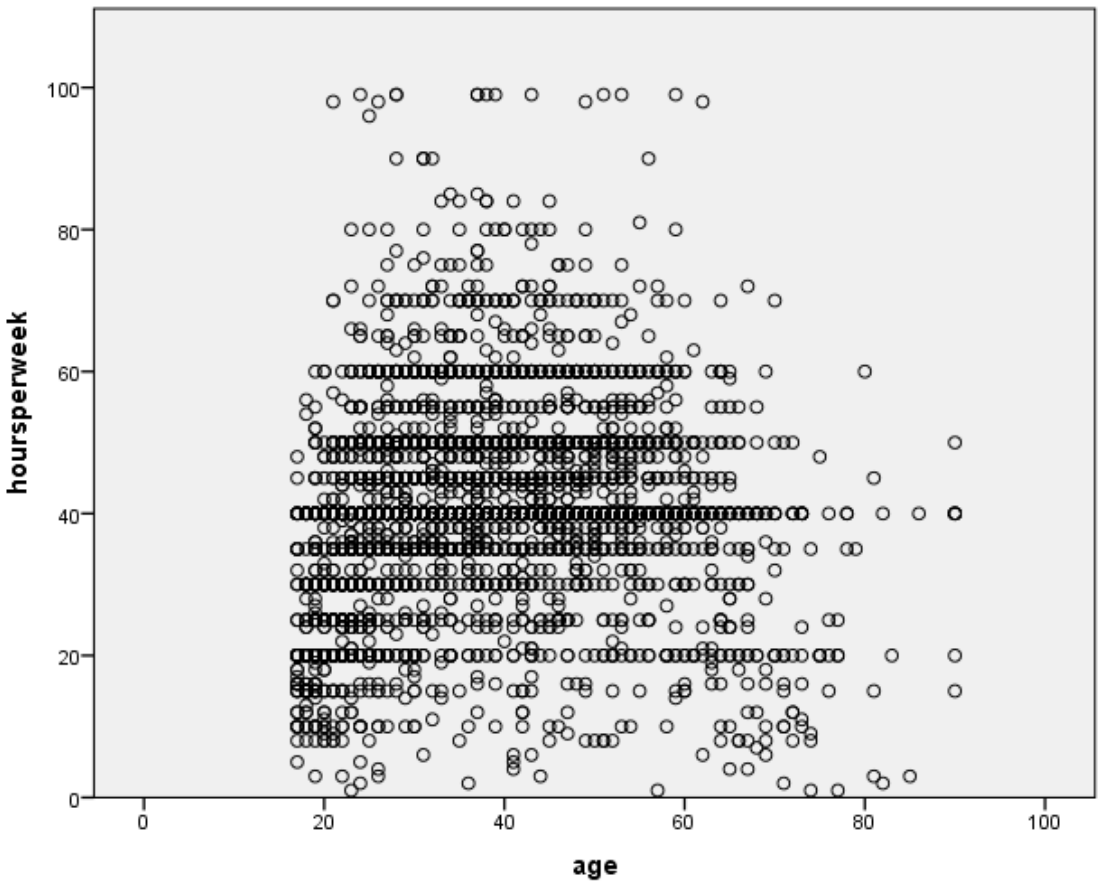
		Valor	Error estándar asintótico[a]	Aprox. S[b]	Aprox. Sig.
Intervalo por intervalo	R de persona	,087	,016	6,421	,000[c]
Ordinal por ordinal	Correlación de Spearman	,157	,014	11,676	,000[c]
N de casos válidos		5405			

a. No se supone la hipótesis nula.

b. Utilización del error estándar asintótico que asume la hipótesis nula.

c. Se basa en aproximación normal.

Si realizamos, de manera complementaria, un gráfico de dispersión para ambas variables, podemos apreciar visualmente la debilidad de la relación entre las variables.



Se aprecia que, para las personas más jóvenes (en torno a los 20 años), los puntos se reparten con similar densidad para jornadas completas (de 40 horas semanales) y también parciales (de menos de 40 horas semanales). Sin embargo, conforme aumenta la edad, las jornadas parciales pasan a ser menos habituales. Esto explica el carácter positivo de los coeficientes de correlación obtenidos. Sin embargo, no se observa una relación lineal clara entre ambas variables (lo que corresponde a que ambos coeficientes toman valores cercanos al 0).

Modelo loglineal

Plantearemos un modelo loglineal con la finalidad de estudiar las relaciones existentes entre las variables categóricas correspondientes a los ingresos anuales, el sexo y la raza. El modelo loglineal es un modelo lineal para los logaritmos de las frecuencias de la tabla de contingencia cruzada. La tabla cruzada de contingencia para las tres variables consideradas es la siguiente.

sexo_cat*race_cat*income_cat tabulación cruzada

Recuento

income_cat			race_cat					Total
			Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White	
<=50K	sexo	Female	16	172	217	46	1110	1561
		Male	21	292	184	76	2095	2668
	Total		37	464	401	122	3205	4229
>50K	sexo	Female	2	32	10	1	132	177
		Male	1	152	45	10	791	999
	Total		3	184	55	11	923	1176
Total	sexo	Female	18	204	227	47	1242	1738
		Male	22	444	229	86	2886	3667
	Total		40	648	456	133	4128	5405

Modelo saturado general

El modelo saturado para las tres variables consideradas (es decir, aquel que incluye todos los efectos) y que explica perfectamente las frecuencias observadas en la tabla cruzada de contingencia es el siguiente.

$$\ln \mu_{ijk} = \lambda + \lambda_i^X + \lambda_i^Y + \lambda_k^Z + \lambda_{ij}^{X,Y} + \lambda_{ik}^{X,Z} + \lambda_{jk}^{Y,Z} + \lambda_{ijk}^{X,Y,Z}$$

Donde:

λ_i^X es el efecto debido a la categoría i de la variable correspondiente a la sexo

λ_i^Y es el efecto debido a la categoría j de la variable correspondiente a los ingresos anuales

λ_k^Z es el efecto debido a la categoría k de la variable correspondiente a la raza

$\lambda_{ij}^{X,Y}$ es el efecto debido a la interacción de las categorías *i* y *j* de las variables sexo e ingresos anuales, respectivamente

$\lambda_{ik}^{X,Z}$ es el efecto debido a la interacción de las categorías *i* y *k* de las variables sexo y raza, respectivamente

$\lambda_{jk}^{Y,Z}$ es el efecto debido a la interacción de las categorías *j* y *k* de las variables ingresos anuales y raza, respectivamente

$\lambda_{ijk}^{X,Y,Z}$ es el efecto debido a la interacción de las categorías *i*, *j* y *k* de las variables sexo, ingresos anuales y raza, respectivamente

λ es el término independiente

Las estimaciones de los parámetros se presentan a continuación.

Estimaciones de parámetro

Efecto	Parámetro	Estimación	Error estándar	Z	Sig.	Intervalo de confianza de 95 %	
						Límite inferior	Límite superior
race_cat*income_cat*sexo_cat	1	-,418	,222	-1,879	,060	-,853	,018
	2	,031	,084	,365	,715	-,135	,196
	3	,184	,101	1,826	,068	-,013	,382
	4	,138	,188	,737	,461	-,230	,506
race_cat*income_cat	1	,146	,222	,659	,510	-,289	,582
	2	-,411	,084	-4,866	,000	-,576	-,245
	3	,117	,101	1,158	,247	-,081	,314
	4	,364	,188	1,938	,053	-,004	,732
race_cat*sexo_cat	1	,461	,222	2,076	,038	,026	,897
	2	-,119	,084	-1,406	,160	-,284	,047
	3	,074	,101	,738	,461	-,123	,272
	4	-,211	,188	-1,125	,261	-,579	,157
income_cat*sexo_cat	1	,224	,073	3,048	,002	,080	,368
race_cat	1	-2,224	,222	-10,008	,000	-2,660	-1,789
	2	,812	,084	9,611	,000	,646	,977
	3	,170	,101	1,682	,092	-,028	,367
	4	-1,289	,188	-6,867	,000	-1,657	-,921
income_cat	1	,991	,073	13,504	,000	,847	1,135
sexo_cat	1	-,400	,073	-5,448	,000	-,544	-,256

A partir de las estimaciones presentadas en la tabla anterior, pueden calcularse la totalidad de ellos.

Mijk	□	□x	□y	□z	□xy	□xz	□yz	□xyz
111	3.9602	-0.4	0.991	-2.224	0.224	0.461	0.146	-0.418
112	3.9602	-0.4	0.991	0.812	0.224	-0.119	-0.411	0.031
113	3.9602	-0.4	0.991	0.17	0.224	0.074	0.117	0.184
114	3.9602	-0.4	0.991	-1.289	0.224	-0.211	0.364	0.138
115	3.9602	-0.4	0.991	2.531	0.224	-0.205	-0.216	0.065
121	3.9602	-0.4	-0.991	-2.224	-0.224	0.461	-0.146	0.418
...	...	...	...	...	...	...	...	...

La interpretación que podemos hacer de las estimaciones de los **efectos principales** es la siguiente:

- Un valor estrictamente positivo indica que la categoría de la variable correspondiente presenta una frecuencia superior a la esperada bajo equiprobabilidad entre todas las categorías
- Por el contrario, un valor estrictamente negativo indica que la categoría correspondiente presenta una frecuencia inferior a la esperada bajo equiprobabilidad entre todas las categorías de la variable
- Un valor igual a 0 indica que la frecuencia observada para la categoría correspondiente es igual a la esperada bajo equiprobabilidad entre las categorías de la variable

Por otra parte, la interpretación que podemos hacer de las estimaciones de los **efectos relativos a interacciones entre variables** es:

- Un valor estrictamente positivo indica que el cruce de categorías correspondiente presenta una frecuencia superior a la esperada bajo independencia entre las variables
- Por el contrario, un valor estrictamente negativo indica que el cruce de categorías correspondiente presenta una frecuencia inferior a la esperada bajo independencia entre las variables
- Un valor igual a 0 indica que la frecuencia observada para el cruce de categorías correspondiente es igual a la esperada bajo independencia entre las variables

Método backward

El modelo saturado, por tratarse de un modelo jerárquico, puede simplificarse a través del método backward. Mediante este método, se contrasta, paso a paso, la significación de los efectos de mayor orden que se mantienen en el modelo. El estadístico del contraste es la **razón de verosimilitud chi-cuadrado**. De los efectos de mayor orden, será candidato a ser eliminado del modelo aquél para el

que el contraste de hipótesis presente un **menor p-valor.** Esto equivale a un mayor valor para la razón de verosimilitud chi-cuadrado, siempre que la distribución de estos estadísticos bajo la hipótesis nula sea idéntica: es decir, una Chi-cuadrado con los mismos grados de libertad. Este efecto será eliminado del modelo únicamente en caso de que el contraste resulte **no significativo** (es decir, presente un **p-valor superior al nivel de significación fijado: 0,05**).

Los resultados correspondientes a la aplicación del método backward sobre el modelo saturado se presentan a continuación.

Resumen de los pasos

Escalón[a]			Effects	Chi-cuadrado[c]	gl	Sig.	Número de iteraciones
0	Clase generadora[b]		sexo_cat*race_cat*income_cat	,000	0	.	
	Efecto suprimido	1	sexo_cat*race_cat*income_cat	5,906	4	,206	3
1	Clase generadora[b]		sexo_cat*race_cat, sexo_cat*income_cat, race_cat*income_cat	5,906	4	,206	
	Efecto suprimido	1	sexo_cat*race_cat	56,834	4	,000	2
		2	sexo_cat*income_cat	207,702	1	,000	2
		3	race_cat*income_cat	53,304	4	,000	2
2	Clase generadora[b]		sexo_cat*race_cat, sexo_cat*income_cat, race_cat*income_cat	5,906	4	,206	

a. En cada paso, se suprime el efecto con el nivel de significación más grande para el cambio de la razón de verosimilitud, siempre que el nivel de significación sea mayor que ,050.

b. Las estadísticas se visualizan para el mejor modelo en cada paso después del paso 0.

c. Para 'Efecto suprimido', este es el cambio en el chi-cuadrado después de que se suprima el efecto del modelo.

En el **paso 0**, se contrasta si el efecto de orden superior en el modelo saturado, si es el correspondiente a la triple interacción entre variables, es nulo. El estadístico del contraste (la razón de verosimilitud chi-cuadrado) toma el valor 5,906 y, bajo esta hipótesis nula, se distribuye según una Chi-cuadrado de cuatro grados de libertad. El p-valor asociado es de 0,206, por lo que no puede rechazarse que el efecto sea nulo (la interacción triple **no** es significativa). Por tanto, este efecto es eliminado del modelo.

Una vez eliminado este efecto del modelo, los efectos de mayor orden son los que corresponden a las interacciones dobles (es decir, por pares entre las tres variables). En el **paso 1**, se contrasta individualmente si cada uno de estos efectos es nulo. Los tres contrastes arrojan **p-valores de 0,000**. Por tanto, las tres interacciones dobles o por pares entre las tres variables **son significativas** y se mantienen en el modelo.

En consecuencia, quedamos ante el siguiente modelo:

$$\ln \hat{\mu}_{ijk} = \lambda + \lambda_i^X + \lambda_i^Y + \lambda_k^Z + \lambda_{ij}^{X,Y} + \lambda_{ik}^{X,Z} + \lambda_{jk}^{Y,Z}$$

Bondad de ajuste

Para el modelo no saturado (es decir, tras la eliminación del efecto correspondiente a la interacción triple), debe evaluarse la bondad del ajuste. El modelo será adecuado cuando las frecuencias observadas y las estimadas mediante el modelo no puedan ser consideradas significativas. Para contrastar esta hipótesis, pueden emplearse dos estadísticos: **chi-cuadrado** y **razón de verosimilitud**.

Pruebas de bondad de ajuste

	Chi-cuadrado	gl	Sig.
Razón de verosimilitud	5,906	4	,206
Pearson	6,515	4	,164

El estadístico chi-cuadrado toma el valor 5,906 y el estadístico razón de verosimilitud toma el valor 6,515. Ambos se distribuyen bajo la hipótesis nula (correspondiente a la adecuación del modelo) según una Chi-cuadrado de cuatro grados de libertad. Los p-valores asociados en ambos casos son superiores a 0,05 (0,206 y 0,164, respectivamente). Por tanto, no puede rechazarse la hipótesis nula y no se rechaza tampoco la validez del modelo.

Las diferencias entre las frecuencias observadas y las estimadas mediante el modelo loglineal se presentan en la siguiente tabla. En ella podemos observar que los residuos estandarizados toman, en general, valores pequeños. Los mayores (en valor absoluto) toman valores de 1,679 y -1,144 (estando ambos por debajo de 2).

Recuentos de casilla y residuos

			Observado		Esperado			Residuos
sexo_cat	race_cat	income_cat	Recuento	%	Recuento	%	Residuos	estándar
Female	Amer-Indian-Eskimo	<=50K	16,000	0,3%	17,351	0,3%	-1,351	-,324
		>50K	2,000	0,0%	,648	0,0%	1,352	1,679
	Asian-Pac-Islander	<=50K	172,000	3,2%	174,798	3,2%	-2,798	-,212
		>50K	32,000	0,6%	29,200	0,5%	2,800	,518
	Black	<=50K	217,000	4,0%	212,670	3,9%	4,330	,297
		>50K	10,000	0,2%	14,332	0,3%	-4,332	-1,144
	Other	<=50K	46,000	0,9%	45,286	0,8%	,714	,106
		>50K	1,000	0,0%	1,711	0,0%	-,711	-,544
	White	<=50K	1110,000	20,5%	1110,899	20,6%	-,899	-,027
		>50K	132,000	2,4%	131,106	2,4%	,894	,078
Male	Amer-Indian-Eskimo	<=50K	21,000	0,4%	19,649	0,4%	1,351	,305
		>50K	1,000	0,0%	2,352	0,0%	-1,352	-,881
	Asian-Pac-Islander	<=50K	292,000	5,4%	289,202	5,4%	2,798	,165
		>50K	152,000	2,8%	154,800	2,9%	-2,800	-,225
	Black	<=50K	184,000	3,4%	188,330	3,5%	-4,330	-,316
		>50K	45,000	0,8%	40,668	0,8%	4,332	,679
	Other	<=50K	76,000	1,4%	76,714	1,4%	-,714	-,081
		>50K	10,000	0,2%	9,289	0,2%	,711	,233
	White	<=50K	2095,000	38,8%	2094,101	38,7%	,899	,020
		>50K	791,000	14,6%	791,894	14,7%	-,894	-,032

Prueba de los k efectos

La prueba de los k efectos contrasta si todos los efectos de orden k son nulos. Para ellos, se emplean los estadísticos chi-cuadrado y razón de verosimilitud. Si la hipótesis nula se rechaza, ello significa que **al menos uno** (uno o más) de los efectos de orden k es significativamente distinto de cero. No puede concluirse, sin embargo, que todos ellos sean significativamente distintos de cero.

Efectos de K y de orden superior

			Razón de verosimilitud		Pearson		Número de
	K	gl	Chi-cuadrado	Sig.	Chi-cuadrado	Sig.	iteraciones
Efectos K y de orden superior[a]	1	19	11679,967	,000	18628,195	,000	0
	2	13	354,958	,000	346,482	,000	2
	3	4	5,906	,206	6,515	,164	3
Efectos K[b]	1	6	11325,008	,000	18281,713	,000	0
	2	9	349,052	,000	339,967	,000	0
	3	4	5,906	,206	6,515	,164	0

a. Prueba que los efectos K y de orden superior son cero.

b. Prueba que los efectos K son cero.

Para el efecto de orden tres (correspondiente a la interacción triple de las variables), la prueba de los k efectos arroja un p-valor de 0,206 asociado al estadístico chi-cuadrado y un p-valor de 0,164 asociado a la razón de verosimilitud. En ambos casos, se trata de p-valores elevados (superiores a los niveles de significación habitualmente empleados: 0,05 o 0,01). Por lo tanto, no hay evidencia estadística para rechazar que el efecto de orden tres sea nulo.

Tanto para los efectos de orden dos (correspondientes a las tres interacciones dobles que se dan, por pares, entre las tres variables consideradas en el modelo), como para los efectos de orden uno (o efectos principales, correspondientes a las categorías de cada una de las variables), se obtienen en la prueba de los k efectos p-valores de 0,000 asociados al estadístico chi-cuadrado y también p-valores de 0,000 asociados al estadístico razón de verosimilitud. Ello implica, por una parte, que los efectos de orden dos son conjuntamente significativos; y, por otra, que los efectos de orden uno son conjuntamente significativos.

Mediante la prueba de los efectos de orden k y superior se contrasta si todos los efectos de orden k y superior son nulos. Se emplean, nuevamente, los estadísticos chi-cuadrado y razón de verosimilitud; y la interpretación del rechazo de la hipótesis nula es equivalente.

Prueba de asociación parcial

Para contrastar, de manera individual, si cada uno de los efectos del modelo es significativo, se emplea la prueba de asociación parcial. En ella, la hipótesis nula es que el efecto considerado es nulo. El estadístico empleado es el estadístico chi-cuadrado.

Asociaciones parciales

Efecto	gl	Chi-cuadrado parcial	Sig.	Número de iteraciones
sexo_cat*race_cat	4	56,834	,000	2
sexo_cat*income_cat	1	207,702	,000	2
race_cat*income_cat	4	53,304	,000	2
sexo_cat	1	703,859	,000	2
race_cat	4	8790,758	,000	2
income_cat	1	1830,392	,000	2

Evaluado los p-valores asociados al valor del estadístico chi-cuadrado en cada una de las pruebas de asociación parcial, podemos concluir que todos los efectos de orden dos son significativos. Es decir, todas las interacciones dos a dos o entre pares son significativas.

Por otra parte, también todos los efectos principales son significativos. Ello indica que, en cada una de las tres variables (correspondientes al sexo, a los ingresos anuales y a la raza), se da una distribución de frecuencias no uniforme entre sus categorías (es decir, una distribución de frecuencias distinta a la esperada si todas las categorías de la variable fuesen equiprobables).

Modelo probit

Para la realización del modelo probit, consideramos la variable dicotómica correspondiente a los ingresos anuales, la variable correspondiente a la edad (que jugará el papel de variable independiente) y la variable sexo (que jugará el papel de variable que segmenta, generando grupos).

Las dos categorías de la variable correspondiente a los ingresos anuales, como ya se ha comentado anteriormente, son:

1. Ingresos anuales de hasta 50.000€
2. Ingresos anuales superiores a 50.000€

En este modelo, la variable independiente debe ser de tipo cuantitativo, pero encontrarse categorizada. La variable correspondiente a la edad es la que jugará este papel en nuestro modelo. Por este motivo, a partir de la variable original que presenta nuestra base de datos de trabajo, construimos la siguiente variable categórica, con los niveles presentados a continuación:

1. [17, 24] años
2. [25, 40] años
3. [41, 64] años
4. [65, 90] años

A través del modelo probit, podremos estimar la probabilidad de presentar ingresos anuales superiores a 50.000€ en función de la edad y del sexo.

A continuación, se presenta la tabla cruzada de contingencia que considera estas tres variables: ingresos anuales, edad y sexo.

age (Agrupada)*income_cat*sexo_cat tabulación cruzada

Recuento

sexo_cat			income_cat <=50K	income_cat >50K	Total
Female	age (Agrupada)	[17, 24] años	346	3	349
		[25, 40] años	638	87	725
		[41, 64] años	524	82	606
		[65, 90] años	53	5	58
	Total		1561	177	1738
Male	age (Agrupada)	[17, 24] años	515	8	523
		[25, 40] años	1259	384	1643
		[41, 64] años	827	575	1402
		[65, 90] años	67	32	99
	Total		2668	999	3667
Total	age (Agrupada)	[17, 24] años	861	11	872
		[25, 40] años	1897	471	2368
		[41, 64] años	1351	657	2008
		[65, 90] años	120	37	157
	Total		4229	1176	5405

Dado que el modelo probit requiere que los datos se presenten de forma agregada o resumida, generamos un nuevo fichero con estas características, en el que las variables son las siguientes:

1. **Sexo**
2. **Edad ~** variable categórica numérica y ordinal, que tiene por niveles las marcas de clase correspondientes a los intervalos anteriormente construidos:

Intervalo	Marca de clase
[17, 24] años	20,5 años
[25, 40] años	32,5 años
[41, 64] años	52,5 años
[65, 90] años	77,5 años

3. **Individuos ~** número de individuos en la muestra para cada combinación de niveles de las variables Sexo y Edad
4. **IngresosAltos ~** número de individuos en la muestra, para cada combinación de niveles de las variables Sexo y Edad, que presentan ingresos anuales superiores a los 50.000€

El fichero de datos construido se presenta a continuación:

	Sexo	Edad	Individuos	IngresosAltos
1	Mujer	20,50	349	3
2	Mujer	32,50	725	87
3	Mujer	52,50	606	82
4	Mujer	77,50	58	5
5	Hombre	20,50	523	8
6	Hombre	32,50	1643	384
7	Hombre	52,50	1402	575
8	Hombre	77,50	99	32

Modelo matemático

Dado que la variable Sexo presenta dos niveles y, por tanto, genera dos grupos, tendremos dos modelos probit a ajustar. La expresión matemática de estos modelos es la siguiente:

$$T(p) = \beta_0^k + \beta_1 Edad$$

Donde:

- *p* es la probabilidad de ganar más de 50.000€ y T(p) es el valor correspondiente de la distribución Normal estándar o N(0,1) que acumula, en la cola izquierda, dicha probabilidad. Es decir, la variable dependiente del modelo sigue una distribución normal estándar o N(0,1)

- *k* = 1, 2 (correspondiendo a los dos niveles de la variable Sexo: 1 = "mujer", 2 = "hombre"). Es decir, cada uno de los modelos probit presenta su propio término independiente. Ambos modelos difieren únicamente en estos parámetros, mientras que el parámetroβ_1 es común a ambos.

Estimación de los parámetros del modelo

Los parámetros a estimar son, por tanto,β_{01}, β_{02} y β_{01}. La estimación se realiza mediante el método de máxima verosimilitud. Los estimadores obtenidos se presentan a continuación.

Estimaciones de parámetro

	Parámetro		Estimación	Error estándar	Z	Sig.	Intervalo de confianza de 95 %	
							Límite inferior	Límite superior
PROBIT[a]	Edad		,024	,001	15,911	,000	,021	,027
	Interceptación[b]	Mujer	-2,241	,076	-29,462	,000	-2,317	-2,165
		Hombre	-1,576	,066	-23,873	,000	-1,642	-1,510

a. Modelo PROBIT: PROBIT(p) = Interceptación + BX

b. Se corresponde a la variable de agrupación Sexo.

Los modelos ajustados obtenidos son, por tanto, los siguientes:

$$T(\hat{p})_{MUJERES} = -2'241 + 0'024Edad$$

$$T(\hat{p})_{HOMBRES} = -1'567 + 0'024Edad$$

El signo de la estimación del parámetro β_1 indica una relación positiva entre la edad y la variable dependiente T(p); y, por tanto, una relación positiva entre la edad y la probabilidad de respuesta. Es decir: a mayor edad, se estima una mayor probabilidad de presentar unos ingresos anuales superiores a 50.000€.

La estimación del parámetro β_{02}, mayor a la estimación del parámetro B01, se corresponden con una mayor probabilidad de presentar ingresos anuales de más de 50.000€ para los hombres (fijada la edad).

Las probabilidades estimadas mediante el modelo ajustado para la presencia de respuesta (es decir, para presentar ingresos anuales superiores a 50.000€), se pueden visualizar en la siguiente tabla. En ellas podemos observar las relaciones anteriormente comentadas.

Recuentos de casilla y residuos

	Número	Sexo	Edad	Número de sujetos	Respuestas observadas	Respuestas esperadas	Residuo	Probabilidad
PROBIT	1	1	20,500	349	3	13,792	-10,792	,040
	2	1	32,500	725	87	51,055	35,945	,070
	3	1	52,500	606	82	96,155	-14,155	,159
	4	1	77,500	58	5	19,793	-14,793	,341
	5	2	20,500	523	8	71,932	-63,932	,138
	6	2	32,500	1643	384	344,371	39,629	,210
	7	2	52,500	1402	575	517,032	57,968	,369
	8	2	77,500	99	32	59,495	-27,495	,601

Validación de los modelos

Como se observa en la expresión matemática, se trata de modelos **lineales** (la variable dependiente *T(p)* es combinación lineal de la variable independiente Edad) y que guardan una relación de **paralelismo** (el parámetro a estimar β_1 es común a ambos modelos; es decir: se trata de dos rectas con misma pendiente y, por tanto, paralelas). Éstas son las hipótesis del modelo que se deben validar.

La **bondad de ajuste** del modelo se determina mediante el **contraste de Chi-cuadrado**, en el que se consideran las respuestas observadas en el muestra frente a las respuestas esperadas bajo la hipótesis nula de bondad de ajuste del modelo. Estas últimas se obtienen, para cada combinación de niveles de las variables Edad y Sexo, mediante el producto de las **probabilidades estimadas** mediante el modelo probit por el número de individuos correspondiente en la muestra.

En la siguiente tabla pueden visualizarse, para cada combinación de niveles de las variables Edad y Sexo, las respuestas observadas, las respuestas esperadas bajo la hipótesis de bondad de ajuste del modelo, y los residuos correspondientes (construidos como la diferencia entre ellas).

Recuentos de casilla y residuos

	Número	Sexo	Edad	Número de sujetos	Respuestas observadas	Respuestas esperadas	Residuo	Probabilidad
PROBIT	1	1	20,500	349	3	13,792	-10,792	,040
	2	1	32,500	725	87	51,055	35,945	,070
	3	1	52,500	606	82	96,155	-14,155	,159
	4	1	77,500	58	5	19,793	-14,793	,341
	5	2	20,500	523	8	71,932	-63,932	,138
	6	2	32,500	1643	384	344,371	39,629	,210
	7	2	52,500	1402	575	517,032	57,968	,369
	8	2	77,500	99	32	59,495	-27,495	,601

Pruebas de chi-cuadrado

		Chi-cuadrado	gl[b]	Sig.
PROBIT	Prueba de bondad de ajuste de Pearson	169,068	5	,000[a]
	Prueba de paralelismo	16,484	1	,000

a. Puesto que el nivel de significación es menor que ,150, se utiliza un factor de heterogeneidad en el cálculo de los límites de confianza.

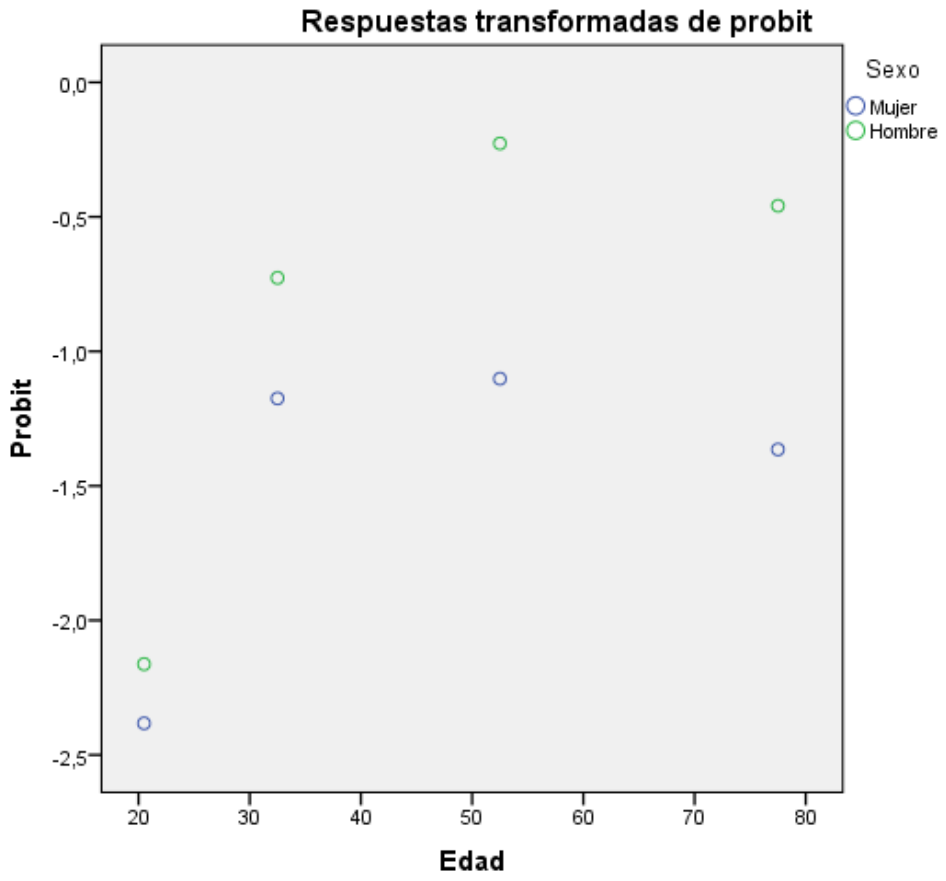
b. Las estadísticas basadas en casos individuales difieren de las estadísticas basadas en casos agregados.

El estadístico del contraste toma el valor **169,068** y, bajo la hipótesis nula de bondad de ajuste, se distribuye según una Chi-cuadrado de cinco grados de libertad. El p-valor asociado al contraste vale 0,000. Por tanto, el contraste es significativo: se rechaza la bondad de ajuste del modelo.

Por otra parte, el estadístico correspondiente a la prueba de paralelismo toma el valor 16,484. La hipótesis nula, en este caso, es la de paralelismo entre las dos rectas, correspondientes a cada uno de los dos modelos probit. Bajo esta hipótesis, el estadístico del contraste sigue una distribución Chi-cuadrado de un grado de libertad. El p-valor asociado toma el valor 0,000 y lleva a rechazar la hipótesis de paralelismo.

Por tanto, se rechaza la validez de los modelos probit; es decir, no pueden considerarse adecuadas las probabilidades estimadas mediante ellos.

La ausencia de linealidad y de paralelismo se puede visualizar mediante el gráfico de dispersión de las respuestas observadas transformadas en función de la Edad y el Sexo. Para cada combinación de niveles de las variables Edad y Sexo, no se representa la proporción muestral correspondiente de individuos con ingresos de más de 50.000€, sino el valor de la distribución Normal estándar que acumula, en la cola izquierda, una probabilidad del valor de dicha proporción.



La disposición curva de los puntos (para cada uno de los grupos generados por la variable sexo) denota la ausencia de linealidad. Por otra parte, se puede observar que las diferencias entre las respuestas transformadas de ambos grupos (hombres y mujeres) no son constantes, sino que varían en función de la edad. Esto indica falta de paralelismo. La prueba de paralelismo arroja, en concordancia, un **p-valor** de **0,000** y, por tanto, lleva a rechazar la hipótesis nula de paralelismo.

Regresión logística binaria

La regresión logística binaria nos permite clasificar a un individuo en una variable dicotómica a partir de la construcción de una función lineal de una serie de variables independientes, que pueden ser tanto cuantitativas como cualitativas.

En nuestro caso, la variable de interés y sobre la que queremos realizar la clasificación de un individuo son los ingresos anuales (ingresos de hasta 50.000€ al año o bien ingresos superiores a 50.000€ al año).

La clasificación se realiza mediante la estimación de la probabilidad de pertenencia a una de las dos subpoblaciones de la variable dicotómica. La variable dependiente del modelo de regresión logística binaria no es, no obstante, esta probabilidad, sino la función:

$$e^Z = \frac{p}{1-p} \rightarrow Z = ln\left(\frac{p}{1-p}\right)$$

Selección de variables

En primer lugar, seleccionamos las variables a ser consideradas como independientes en el modelo de regresión logística binaria mediante el método por pasos. El conjunto de variables sobre el que realizamos la selección son:

1. Age: edad (variable cuantitativa continua)
2. Hoursperweek: jornada laboral semanal en horas (variable cuantitativa continua)
3. Sexo_cat: sexo (variable categórica dicotómica)
4. Race_cat: raza (variable categórica con más de dos categorías)

Resumen de procesamiento de casos

Casos sin ponderar[a]		N	Porcentaje
Casos seleccionados	Incluido en el análisis	5405	100,0
	Casos perdidos	0	,0
	Total	5405	100,0
Casos no seleccionados		0	,0
Total		5405	100,0

Contamos con 5405 casos, ninguno de ellos con valores perdidos.

El modelo de regresión logística binaria supone las siguientes codificaciones de las variables categóricas. En primer lugar, la variable objetivo de la siguiente manera.

Codificación de variable dependiente

Valor original	Valor interno
<=50K	0
>50K	1

Por tanto, la probabilidad *p* estimada corresponderá (por el software utilizado) a la probabilidad de pertenencia al segundo grupo. Es decir, se estimará la probabilidad de presentar ingresos anuales superiores a 50.000€ .

Las variables categóricas se codifican de la siguiente manera, mediante variables *dummy* que indican, para cada una de las categorías de la variable (a excepción de la última de ellas), la pertenencia o no del individuo a dicha categoría. La pertenencia a la última categoría de la variable corresponde a la ausencia de pertenencia a todas las anteriores.

Codificaciones de variables categóricas

			Codificación de parámetro			
		Frecuencia	(1)	(2)	(3)	(4)
race_cat	Amer-Indian-Eskimo	40	1,000	,000	,000	,000
	Asian-Pac-Islander	648	,000	1,000	,000	,000
	Black	456	,000	,000	1,000	,000
	Other	133	,000	,000	,000	1,000
	White	4128	,000	,000	,000	,000
sexo_cat	Female	1738	1,000			
	Male	3667	,000			

Por tanto, se generan, a partir de la variable race_cat, las variables:

1. race_cat(1), correspondiente a la pertenencia (1) o no (0) a la categoría Amer-Indian-Eskimo
2. race_cat(2), correspondiente a la pertenencia (1) o no (0) a la categoría Asian-Pac-Islander
3. race_cat(3), correspondiente a la pertenencia (1) o no (0) a la categoría Black
4. race_cat(4), correspondiente a la pertenencia (1) o no (0) a la categoría Other

La pertenencia a la categoría White vendrá dada por la no pertenencia a ninguna de las categorías anteriores (es decir, por valores 0 en las cuatro variables *dummy* anteriormente indicadas).

A partir de la variable sexo_cat, se genera la variable:

1. sexo_cat(1), correspondiente a la pertenencia (1) o no (0) a la categoría Female (mujeres)

La pertenencia a la categoría Male (hombres) vendrá dada por la no pertenencia a la categoría Female (es decir, por un valor 0 en la variable *dummy* sexo_cat(1)).

Las variables no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	sexo_cat(1)	201,563	1	,000
		race_cat	61,819	4	,000
		race_cat(1)	4,812	1	,028
		race_cat(2)	19,054	1	,000
		race_cat(3)	27,504	1	,000
		race_cat(4)	14,570	1	,000
		age	309,657	1	,000
		hoursperweek	250,915	1	,000
	Estadísticos globales		669,669	7	,000

En el paso 0, ninguna de las variables está incluida en el modelo, que corresponde únicamente al término independiente. Para determinar la variable a entrar en el modelo, se contrasta, para cada uno de los coeficientes correspondientes, la hipótesis nula que el valor del coeficiente sea cero. El estadístico del contraste es la **puntuación eficiente de Rao**.

$$H_0: \beta_j = 0$$

La interpretación de dicha hipótesis es que la variable Xj, siendo incluida en el modelo en el siguiente paso, no aporta información significativa sobre la variable dependiente (que es la función Z). Por tanto, para que se considere la entrada de una variable en el modelo, el contraste correspondiente debe ser **significativo** (con un **p-valor asociado inferior a 0,05**). Entrará en el modelo aquella con menor p-valor (siempre que sea inferior a 0,05). Se consideran los resultados obtenidos para las variables originales y no las dicotomizadas (si bien éstas serán las que aparezcan en el modelo finalmente).

Entrará, por tanto, variable Age, para la que la puntuación eficiente de Rao es de 309,657 y el p-valor es menor (siendo inferior a 0,05).

En cada paso, tras la inclusión de una nueva variable en el modelo, se contrasta la aportación de información de cada una de las variables sobre la variable dependiente mediante el mediante el **estadístico de Wald.** Es decir, la hipótesis nula del contraste es:

$$H_0: \beta_j = 0$$

Para que una variable salga del modelo, el contraste debe ser **no significativo** (con un **p-valor asociado mayor a 0,1**). Saldrá del modelo aquella variable para la que el contraste de hipótesis arroje un p-valor mayor, siempre que éste sea mayor a 0,1. Se consideran los resultados obtenidos para las variables originales y no las dicotomizadas (si bien éstas serán las que aparezcan en el modelo finalmente).

Las variables no están en la ecuación

			Puntuación	gl	Sig.
Paso 1	Variables	hoursperweek	248,575	1	,000
		sexo_cat(1)	197,743	1	,000
		race_cat	58,479	4	,000
		race_cat(1)	3,699	1	,054
		race_cat(2)	21,677	1	,000
		race_cat(3)	27,136	1	,000
		race_cat(4)	10,677	1	,001
	Estadísticos globales		412,248	6	,000
Paso 2	Variables	sexo_cat(1)	141,279	1	,000
		race_cat	49,118	4	,000
		race_cat(1)	2,959	1	,085
		race_cat(2)	20,725	1	,000
		race_cat(3)	18,093	1	,000
		race_cat(4)	11,354	1	,001
	Estadísticos globales		179,619	5	,000
Paso 3	Variables	race_cat	40,569	4	,000
		race_cat(1)	2,367	1	,124
		race_cat(2)	20,205	1	,000
		race_cat(3)	9,839	1	,002
		race_cat(4)	11,481	1	,001
	Estadísticos globales		40,569	4	,000

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	age	,044	,003	287,186	1	,000	1,045
	Constante	-3,037	,113	716,463	1	,000	,048
Paso 2[b]	age	,047	,003	283,431	1	,000	1,048
	hoursperweek	,046	,003	225,408	1	,000	1,047
	Constante	-5,116	,191	719,036	1	,000	,006
Paso 3[c]	sexo_cat(1)	-1,056	,091	133,345	1	,000	,348
	age	,047	,003	280,088	1	,000	1,048
	hoursperweek	,041	,003	169,753	1	,000	1,042
	Constante	-4,646	,195	568,691	1	,000	,010
Paso 4[d]	sexo_cat(1)	-1,032	,092	125,673	1	,000	,356
	race_cat			39,065	4	,000	
	race_cat(1)	-,923	,621	2,211	1	,137	,397
	race_cat(2)	,398	,103	14,993	1	,000	1,488
	race_cat(3)	-,451	,156	8,316	1	,004	,637
	race_cat(4)	-1,056	,331	10,203	1	,001	,348
	age	,047	,003	274,164	1	,000	1,048
	hoursperweek	,040	,003	165,429	1	,000	1,041
	Constante	-4,624	,197	550,266	1	,000	,010

En lo que a la interpretación de las tablas anteriores encontramos:

1. En el **paso 1**, entra en el modelo la variable **Age** (correspondiente a la edad), por ser la variable con mínimo p-valor asociado a la **puntuación eficiente de Rao** (e inferior a 0,05)
2. En el **paso 2**, entra en el modelo la variable **Hoursperweek** (correspondiente a la jornada laboral semanal, en horas), por ser la variable con mínimo p-valor asociado a la **puntuación eficiente de Rao** (e inferior a 0,05)
3. En el **paso 3**, entra en el modelo la variable **Sex_cat(1)** (correspondiente al sexo), por ser la variable con mínimo p-valor asociado a la **puntuación eficiente de Rao** (e inferior a 0,05). Tras entrar en el modelo Sex_cat(1), el p-valor asociado al **estadístico de Wald** es, para los contrastes de hipótesis correspondientes a todas las variables, inferior a 0,1. Por tanto, ninguna de las variables sale del modelo.
4. En el **paso 4**, entra en el modelo la variable **Race_cat** (las variables en el modelo serán race_cat(1), race_cat(2), race_cat(3), race_cat(4)), por ser la variable con mínimo p-valor asociado a la **puntuación eficiente de Rao** (e inferior a 0,05). Tras su entrada, el p-valor asociado al **estadístico de Wald** es, para los contrastes de hipótesis correspondientes a todas las variables, inferior a 0,1. Por tanto, ninguna de las variables sale del modelo.

Modelo teórico a ajustar

La selección por pasos no ha prescindido de ninguna variable para el modelo: todas ellas aportan información significativa sobre la variable dependiente (función Z). Por tanto, si consideramos:

Variable	Nombre
X1	Age
X2	Hoursperweek
X3	Sex_cat(1)
X4	Race_cat(1)
X5	Race_cat(2)
X6	Race_cat(3)
X7	Race_cat(4)

El modelo a ajustar es:

$$\frac{p}{q} = e^{\beta_0} * e^{1X_1} * e^{\beta_2 X_2} * \ldots * e^{\beta_7 X_7}$$

Modelo ajustado e interpretación de coeficientes

Y el modelo ajustado, queda:

$$\frac{\hat{p}}{\hat{q}} = 0'1 * 1'048^{x_{i1}} * 1'041^{x_{i2}} * 0'356^{x_{i3}} * 0'397^{x_{i4}} * 1'488^{x_{i5}} * 0'637^{x_{i6}} 0'348^{x_{i7}}$$

Los coeficientes estimados permiten interpretar la influencia de cada una de las variables independientes del modelo, de la siguiente manera:

- **Age**: es el *ods ratio* correspondiente a cualquier valor de edad frente a una unidad inferior. Por tanto, esto implica que, por cada año más de edad, es 1,048 veces más probable tener ingresos anuales superiores a 50.000€

$$e^{\beta_1} = 1{,}048$$

- **Hoursperweek**: el *ods ratio* correspondiente a cualquier valor de horas trabajadas a la semana frente a una unidad inferior. Por tanto, esto implica que, por cada hora más de jornada laboral semanal, es 1,041 veces más probable tener ingresos anuales superiores a 50.000€

$$e^{\beta_2} = 1{,}041$$

- **Sex_cat(1)**: es el *ods ratio* correspondiente a ser mujer frente a ser hombre. Eso implica que la probabilidad de tener ingresos anuales superiores a 50.000€ siendo mujer es 0,356 veces la probabilidad de tenerlos siendo hombre. Es decir, ser mujer disminuye la probabilidad de tener ingresos superiores a 50.000€.

$$e^{\beta_3} = 0{,}356$$

- **Race_cat(1)**, **Race_cat(2)**, **Race_cat(3)** y **Race_cat(4)**: son los *ods ratios* correspondientes a ser de raza Amer-Indian-Eskimo, Asian-Pac-Islander, Black y Other frente a no serlo, respectivamente. Su interpretación indica que ser de raza Asian-Pac-Islander aumenta la probabilidad de tener ingresos superiores a 50.000€; concretamente, la probabilidad de tener ingresos superiores a 50.000€ es 1,488 veces mayor si se es de esta raza. Sin embargo, la probabilidad de tener ingresos superiores a 50.000€ disminuye siendo de alguna de las otras razas.

Bondad de ajuste

Estadístico G

Se contraste si todas las variables independientes del modelo de regresión logística binaria aportan, de manera conjunta, información significa sobre la variable dependiente. La hipótesis nula de este contraste es:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_J = 0$$

El estadístico G es el estadístico del contraste y, bajo la hipótesis nula se distribuye bajo una Chi-cuadrado de j grados de libertad (siendo j el número de variables independientes incluidas en el modelo). Su valor se obtiene a partir del estadístico D, de manera que en el paso *m* el estadístico G se calcula:

G(*m*) = D(modelo correspondiente al paso *m*) - D(modelo correspondiente al paso 0)

El valor del estadístico D aparece en la tabla siguiente (columna *Logaritmo de la verosimilitud -2)* para cada uno de los pasos a partir del paso 1.

Resumen del modelo

Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	5360,994[a]	,054	,084
2	5114,406[a]	,096	,149
3	4961,667[a]	,122	,187
4	4918,872[a]	,129	,198

a. La estimación ha terminado en el número de iteración 5 porque las estimaciones de parámetro han cambiado en menos de ,001.

Pruebas ómnibus de coeficientes de modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Escalón	301,535	1	,000
	Bloque	301,535	1	,000
	Modelo	301,535	1	,000
Paso 2	Escalón	246,588	1	,000
	Bloque	548,123	2	,000
	Modelo	548,123	2	,000
Paso 3	Escalón	152,739	1	,000
	Bloque	700,862	3	,000
	Modelo	700,862	3	,000
Paso 4	Escalón	42,795	4	,000
	Bloque	743,657	7	,000
	Modelo	743,657	7	,000

El estadístico D toma los siguientes valores, en cada paso del método de selección de variables por pasos:

- Paso 1 - G(1) = 301,535, que se distribuye bajo la hipótesis nula según una Chi-cuadrado de un grado de libertad
- Paso 2 - G(2) = 548,123, que se distribuye bajo la hipótesis nula según una Chi-cuadrado de 2 grados de libertad
- Paso 3 - G(3) = 700,862, que se distribuye bajo la hipótesis nula según una Chi-cuadrado de 3 grados de libertad
- Paso 4 - G(4) = 743,657, que se distribuye bajo la hipótesis nula según una Chi-cuadrado de 4 grados de libertad

El contraste de hipótesis resulta significativo para cada uno de los cuatro modelos, con asociados p-valores de valor 0,000. Se observa, además, que la inclusión de variables en cada paso da lugar a un incremento en el estadístico G (y, por tanto, aunque no se aprecia en la tabla, a una disminución del p-valor).

Por tanto, para el modelo con el que hemos trabajo (que corresponde al paso 4), contamos con evidencia estadística para rechazar la hipótesis del contraste. Es decir, las variables Age, Hoursperweek, Sex_cat(1), Race_cat(1), Race_cat(2), Race_cat(3) y Race_cat(4) aportan conjuntamente información significativa sobre la variable dependiente (función Z) a través del modelo. La misma conclusión puede extraerse para los modelos correspondientes a los pasos 1, 2 y 3.

R cuadrado de Cox y Snell y R cuadrado de Nagelkerke

Ambos pretenden cuantificar, mediante un valor comprendido entre cero y uno, la bondad del ajuste del modelo de regresión binaria. La R cuadrado de Cox y Snell presenta el inconveniente de que no alcanza la cota superior del uno. La R cuadrado de Nagelkerke supone una corrección de la R cuadrado de Cox y Snell, de manera que esta cota pueda ser alcanzada.

Resumen del modelo

Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	5360,994[a]	,054	,084
2	5114,406[a]	,096	,149
3	4961,667[a]	,122	,187
4	4918,872[a]	,129	,198

a. La estimación ha terminado en el número de iteración 5 porque las estimaciones de parámetro han cambiado en menos de ,001.

El valor de la R cuadrado de Nagelkerke (0,198) nos indica que la calidad del modelo es de un 19,8%, lo que supone una calidad pobre del modelo.

Análisis de los residuos

Otra forma de evaluar la bondad del ajuste es mediante el estudio de los residuos (**diferencia entre la probabilidad observada y la estimada mediante el modelo de regresión logística**). Cuanto menores sean los residuos, mejor será la calidad del modelo. Los residuos tipificados corregidos, que se distribuyen según una Normal estándar o N(0,1), permiten detectar individuos anómalos (aquellos para los que el residuo tipificado sea superior, en valor absoluto, a 2).

En la siguiente tabla, se presentan las observaciones con residuos tipificados corregidos superior, en valor absoluto, a 2,5. Se puede apreciar que son numerosos los casos en los que estos residuos toman valores elevados.

Lista por casos[b]

Caso	Estado seleccionado[a]	Observado income_cat	Pronosticado	Grupo pronosticado	Variable temporal Resid	Variable temporal ZResid
40	S	2**	,073	1	,927	3,561
41	S	2**	,038	1	,962	5,053
78	S	2**	,050	1	,950	4,376
122	S	2**	,087	1	,913	3,243
181	S	2**	,125	1	,875	2,650
182	S	2**	,076	1	,924	3,491
283	S	2**	,087	1	,913	3,243
311	S	2**	,122	1	,878	2,688
348	S	2**	,082	1	,918	3,338
374	S	2**	,130	1	,870	2,588
...	...	...	...	...	...	...
5090	S	2**	,127	1	,873	2,620
5098	S	2**	,129	1	,871	2,600
5143	S	2**	,080	1	,920	3,398
5209	S	1**	,897	2	-,897	-2,949
5220	S	2**	,048	1	,952	4,470
5266	S	2**	,100	1	,900	3,003
5277	S	2**	,119	1	,881	2,721
5295	S	2**	,097	1	,903	3,050
5321	S	2**	,105	1	,895	2,925
5394	S	2**	,103	1	,897	2,952

a. S = Seleccionado, U = casos sin seleccionar, y ** = casos clasificados incorrectamente.

b. Se listan los casos con residuos estudentizados mayores que 2,000.

Nota: debido a la extensión de la tabla se ha decidido mostrar las 10 primeras y últimas observaciones de la tabla.

Prueba de Hosmer y Lemeshow

La prueba de Hosmer y Lemeshow valora la capacidad explicativa del modelo y se recomienda cuando el mismo incluye al menos una variable cuantitativa, como es nuestro caso.

El estadístico correspondiente se calcula mediante la **comparación del número de individuos pertenecientes al segundo grupo** (en nuestro caso, número de individuos con ingresos anuales superiores a 50.000€) **y el número esperado en función de las probabilidades estimadas mediante el modelo**, tras realizar una partición de los datos en función de los deciles correspondientes a las probabilidades estimadas para cada individuo de la muestra.

La **hipótesis nula** correspondiente a esta prueba es la de **adecuación del modelo** (es decir, no hay diferencias significativas entre las frecuencias observadas y las esperadas en función de las probabilidad estimadas). Bajo esta hipótesis nula, el estadístico de la prueba sigue una distribución Chi-cuadrado.

No se puede emplear cuando se obtienen frecuencias esperadas muy bajas (inferiores a 5). En nuestro caso, y como se puede observar en la siguiente tabla, esto no ocurre:

Tabla de contingencia para la prueba de Hosmer y Lemeshow

		income_cat = <=50K		income_cat = >50K		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	539	485,994	3	56,006	542
	2	480	436,835	17	60,165	497
	3	436	409,357	38	64,643	474
	4	508	518,965	106	95,035	614
	5	422	462,786	142	101,214	564
	6	373	436,072	176	112,928	549
	7	356	387,285	151	119,715	507
	8	389	425,665	197	160,335	586
	9	320	347,724	194	166,276	514
	10	406	318,318	152	239,682	558
Paso 2	1	537	518,862	9	27,138	546
	2	488	454,923	15	48,077	503
	3	507	476,753	34	64,247	541
	4	451	458,984	84	76,016	535
	5	430	462,764	128	95,236	558
	6	408	445,573	152	114,427	560
	7	375	402,174	156	128,826	531
	8	365	385,875	177	156,125	542
	9	344	347,393	193	189,607	537
	10	324	275,699	228	276,301	552
Paso 3	1	537	528,365	11	19,635	548
	2	494	504,738	47	36,262	541
	3	479	476,403	52	54,597	531
	4	497	468,644	45	73,356	542
	5	473	455,762	74	91,238	547
	6	423	431,827	119	110,173	542
	7	385	406,259	156	134,741	541
	8	358	385,527	199	171,473	557
	9	313	333,410	232	211,590	545
	10	270	238,065	241	272,935	511
Paso 4	1	536	528,179	10	17,821	546
	2	509	506,391	32	34,609	541
	3	487	488,663	54	52,337	541
	4	479	469,463	61	70,537	540
	5	469	452,039	73	89,961	542
	6	429	429,742	111	110,258	540
	7	392	405,923	149	135,077	541
	8	345	374,009	196	166,991	541
	9	312	330,898	230	211,102	542
	10	271	243,693	260	287,307	531

Prueba de Hosmer y Lemeshow

Escalón	Chi-cuadrado	gl	Sig.
1	255,080	8	,000
2	112,640	8	,000
3	47,047	8	,000
4	26,769	8	,001

Para el modelo con el que hemos trabajado, que es el correspondiente al paso 4, el estadístico de la prueba toma el valor 26,769. Bajo la hipótesis nula, se distribuye según una Chi-cuadrado de 8 grados de libertad. El **p-valor** asociado al estadístico de la prueba es de **0,001**. Esto nos lleva a rechazar la hipótesis nula de adecuación del modelo.

Clasificación de individuos

En cuanto a la calidad de predicción del modelo, encontramos la siguiente matriz de confusión. El valor de corte es 0,50. Esto significa que cuando la estimación de la probabilidad *p* de tener ingresos superiores a 50.000€ para un individuo sea igual o superior a 0,50, este individuo se clasificará en este grupo. Por el contrario, cuando la estimación de esta probabilidad sea inferior a 0,50, se clasificará en el grupo de ingresos iguales o inferiores a 50.000€.

Tabla de clasificación[a]

			Pronosticado		
			income_cat		Corrección de porcentaje
	Observado		<=50K	>50K	
Paso 1	income_cat	<=50K	4169	60	98,6
		>50K	1163	13	1,1
	Porcentaje global				77,4
Paso 2	income_cat	<=50K	4120	109	97,4
		>50K	1090	86	7,3
	Porcentaje global				77,8
Paso 3	income_cat	<=50K	4085	144	96,6
		>50K	1037	139	11,8
	Porcentaje global				78,1
Paso 4	income_cat	<=50K	4071	158	96,3
		>50K	1015	161	13,7
	Porcentaje global				78,3

a. El valor de corte es ,500

De forma que el modelo final es capaz de predecir de forma correcta un **78,30%** de las observaciones, a nivel global.

El porcentaje de clasificación correcta para los individuos con ingresos iguales o inferiores a 50.000€ (es decir, la **especificidad** del modelo) es del **96,3%** (sobre un total de 4229 individuos). Sin embargo, el porcentaje de clasificación correcta para los individuos con ingresos superiores a 50.000€ (es decir, la **sensibilidad** del modelo) es de tan sólo un **13,7%** (sobre un total de 1176 individuos).

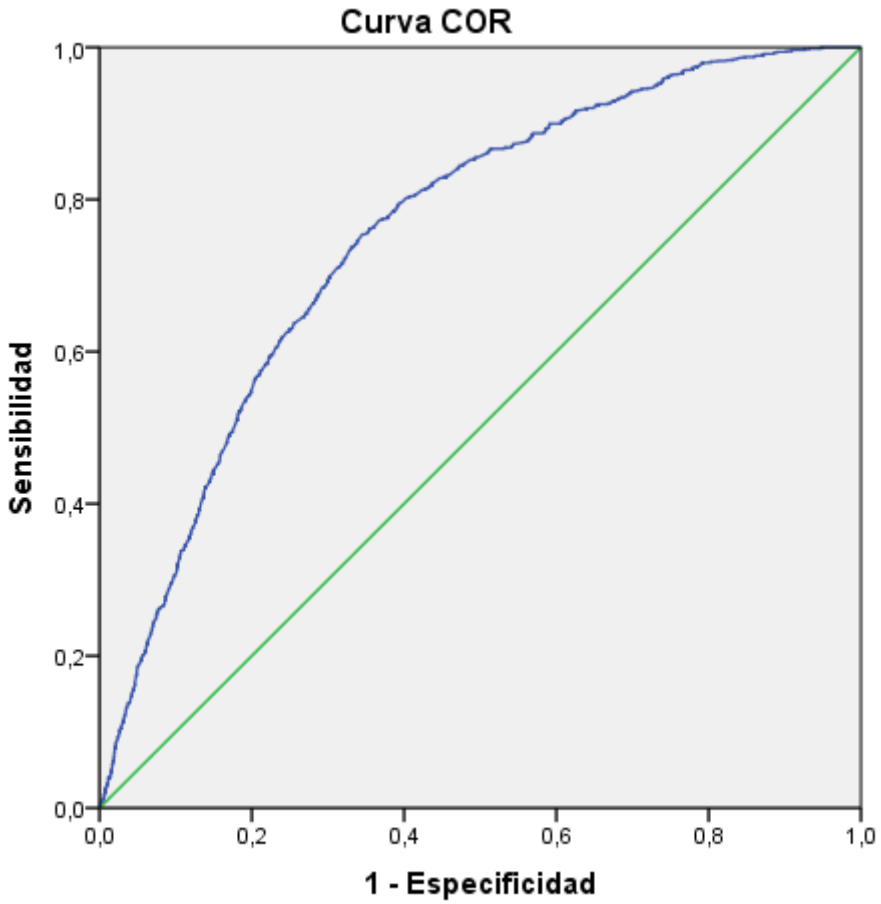
Por tanto, la calidad de clasificación es alta para los individuos de ingresos iguales o inferiores a 50.000€, pero muy baja para aquellos con ingresos más elevados. El porcentaje global de

clasificación correcta está muy influido por el elevado porcentaje de clasificación de los individuos con ingresos de hasta 50.000€ y por el desbalance entre el número de individuos correspondiente a cada uno de los dos grupos en la muestra (son mucho menos numerosos los individuos con ingresos superiores a 50.000€).

La incorporación de variables en el modelo, en cada paso, supone un aumento del porcentaje de clasificación global del modelo. En el paso 1, que corresponde al modelo que incluye únicamente a la variable **Age** como variable independiente, el porcentaje global de clasificación es del **77,4%**. Al incluir en el modelo, también, a la variable **Hourspeerweek**, este porcentaje aumenta a un **77,8%**. La inclusión de la variable **Sex_cat(1)** en el modelo da lugar a un porcentaje global de clasificación del **78,1%**; y por último, el modelo final (que incluye, también, las variables dicotomizadas correspondientes a la variable categórica **Race_cat**) se obtiene el **78,3%** anteriormente mencionado.

De igual forma, se produce en cada paso un incremento de la **sensibilidad** del modelo, a costa de perder **especificidad**. En el modelo que incluye únicamente a la variable **Age**, la sensibilidad es de apenas un **1,1%**. Con la incorporación de **Hourspeerweek** en el modelo, aumenta al **7,3%**. Al considerar también en el modelo la variable **Sexo_cat(1)**, se consigue una sensibilidad del **11,8%**; y, en el modelo correspondiente al paso 4, se alcanza el **13,7%** comentado inicialmente. La disminución en la especificidad del modelo es menor, bajando progresivamente en cada paso desde un **98,6%** hasta un **96,3%**.

La sensibilidad y la especificidad del modelo dependen del punto de corte empleado para la clasificación. La curva ROC permite analizar los resultados correspondientes a diferentes puntos de corte, representando para ellos **sensibilidad** y el **complementario de la especificidad** (**1 - especificidad**).



Los segmentos de diagonal se generan mediante empates.

El área comprendida bajo la curva ROC, comprendida siempre entre 0,5 y 1, puede emplearse como medida de la capacidad de clasificación del modelo. En este sentido, se considera una capacidad de clasificación aceptable si el área bajo la curva ROC está comprendida entre 0,7 y 0,8 y muy buena por encima de este último valor. En nuestro caso, el área correspondiente toma un valor de **0,756**.

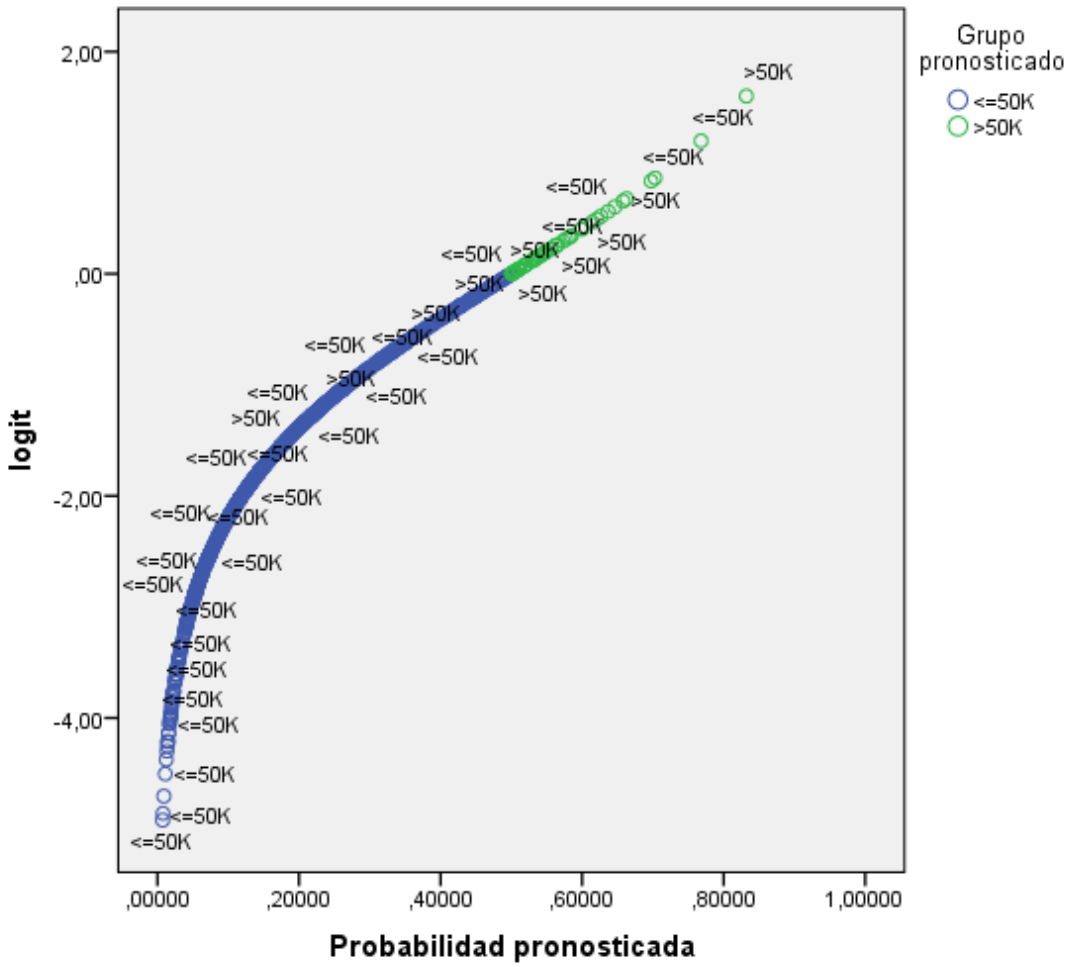
Área bajo la curva

Variable(s) de re:

Área
,756

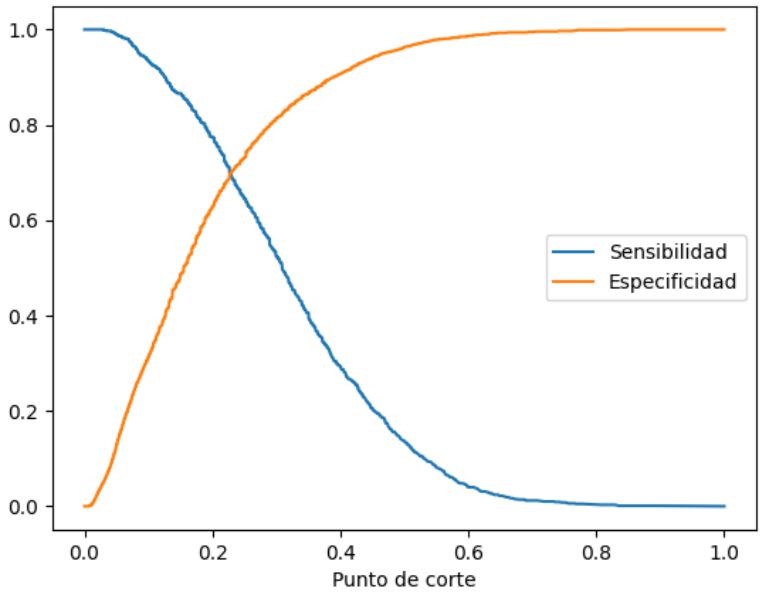
La(s) variable (s) de resultado de prueba: Probabilidad pronosticada tiene, como mínimo, un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Las estadísticas podrían estar sesgadas.

En el gráfico presentado a continuación, se representa la función logit (o función Z) en función de la estimación de la probabilidad *p* obtenida mediante el modelo. En él, podemos observar el grupo pronosticado (para valores estimados de *p* inferiores a 0,5, se pronostican ingresos anuales iguales o inferiores a 50.000€; para valores estimados de *p* iguales o superiores a 0,5 se pronostican ingresos anuales superiores a 50.000€) frente al grupo observado, indicado mediante el etiquetado de los puntos.



Un punto de corte superior a 0,5 hará aumentar la especificidad del modelo, a la vez que disminuirá la sensibilidad. Por el contrario, un punto de corte inferior a 0,5, hará aumentar la sensibilidad del modelo, a precio de disminuir la especificidad.

En nuestro caso, con el punto de corte igual a 0,5, hemos obtenido resultados muy desequilibrados para la especificidad y la sensibilidad del modelo. Se obtiene una elevada especificidad (**96,3%**), pero una muy baja sensibilidad (**13,7%**). Por tanto, buscar un mayor equilibrio entre ellas supondrá **utilizar un punto de corte de valor inferior a 0,5**.



La representación de la sensibilidad y la especificidad del modelo en función del punto de corte utilizado nos permite ver que el punto de corte que equilibra ambas es algo superior a 0,2. Es muy próximo a **0,23**, según podemos ver en la siguiente tabla.

Punto de corte	Sensibilidad	1 - Especificidad	Especificidad	Diferencia
0,2270255	0,699	0,302	0,698	0,001
0,2272727	0,699	0,302	0,698	0,001
0,2273626	0,699	0,302	0,698	0,001
0,2276142	0,699	0,302	0,698	0,001
0,2279341	0,698	0,302	0,698	0
0,2280683	0,698	0,301	0,699	-0,001
0,2283877	0,696	0,301	0,699	-0,003
0,2287568	0,696	0,301	0,699	-0,003
0,228964	0,696	0,301	0,699	-0,003

Los resultados de clasificación si se utiliza este nuevo punto de corte (0,23) en lugar de 0,5 se presentan en la siguiente tabla.

Tabla de clasificación[a]

			Pronosticado		
			income_cat		Corrección de porcentaje
	Observado		<=50K	>50K	
Paso 1	income_cat	<=50K	2870	1359	67,9
		>50K	518	658	56,0
	Porcentaje global				65,3
Paso 2	income_cat	<=50K	2911	1318	68,8
		>50K	454	722	61,4
	Porcentaje global				67,2
Paso 3	income_cat	<=50K	2972	1257	70,3
		>50K	371	805	68,5
	Porcentaje global				69,9
Paso 4	income_cat	<=50K	2963	1266	70,1
		>50K	362	814	69,2
	Porcentaje global				69,9

a. El valor de corte es ,230

Tomando como punto de corte 0,23, se obtiene una **especificidad** del **70,1%** y una **sensibilidad** del **69,2%**. La disminución de la especificidad implica, por el desbalance existente entre el número de individuos pertenecientes a cada uno de los grupos en la muestra, que el porcentaje global de individuos bien clasificados disminuya a un **69,9%**.