Estimación de la tasa de pobreza en Cantabria mediante matching estadístico

Francisco Parra Rodríguez
Pablo Lobete López
Instituto Cántabro de Estadística (ICANE)

Instituto Cántabro de ESTADÍSTICA

Índice

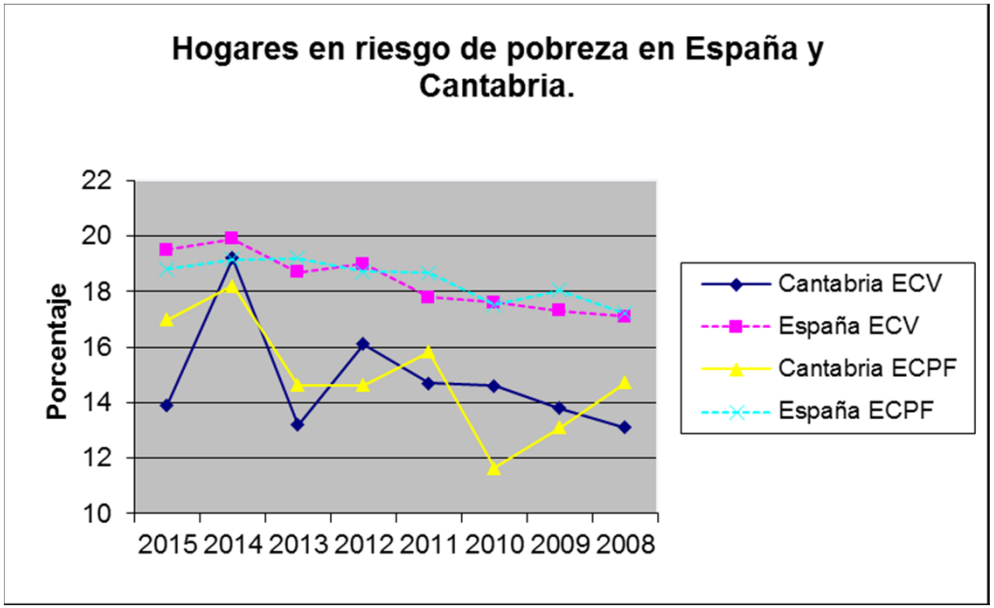
- Matching estadístico
- Fuentes estadísticas: Recodificación de variables
- Modelos estadísticos utilizados
- Resultados obtenidos

GOBIERNO de CANTABRIA

Cantabria Camino Lebaniego

¿Por qué un Matching con la Encuesta Social de Cantabria (ESOC)?

- La tasa de pobreza en las encuestas de hogares de Cantabria es muy variable.
- Las encestas tienen una muestra pequeña para este propósito:
  - Encuesta de Condiciones de Vida: 480 hogares
  - Encuesta de Presupuestos familiares: 762 hogares
- Dos posibilidades:
  - Hacer una encuesta nueva
  - Establecer una metodología para el cálculo de la tasa de riesgo de pobreza en Cantabria a partir de la Encuesta Social de Cantabria (ESOC).
- Solución: Matching estadístico.



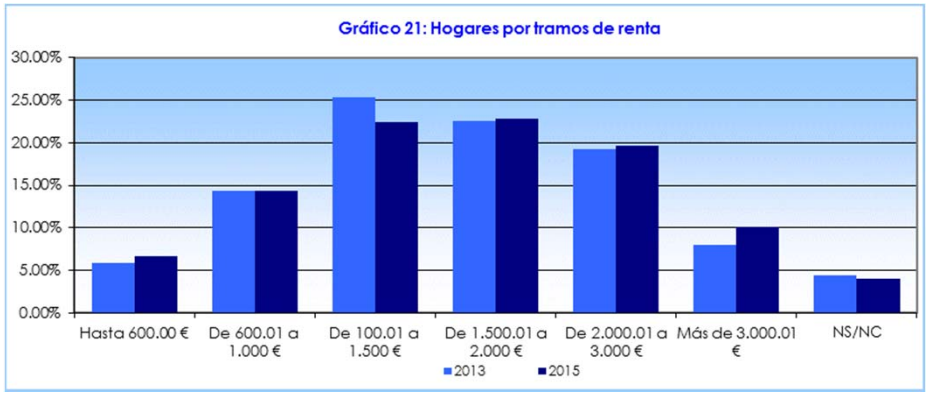
El matching estadístico:

- Integración de dos bases de datos referidas a una misma población objetivo.
- Poner en relación dos bases de datos a través de una serie de variables comunes, correlacionadas con una variable objetivo Z, que solo está presente en uno de los conjuntos de datos.
- Buscar un modelo explicativo de la variable Z en el conjunto de datos donante, compuesto por esas variables comunes, que pueda permitir la estimación en dicha variable en el conjunto de datos receptor.
- Bibliografía:
  - Leulescu, A. and Agafitei, M. (2013) *Statistical matching: a model based approach for data integration*. Eurostat methodologies and working paper, Eurostat.
  - Serafino, P. and Tonkin, R. (2013) *Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey.* Statistical Working Papers. Eurostat

Fases del Matching estadístico con la ESOC:

- Estudio de las bases de datos e identificación de variables comunes a ambos conjuntos de datos
- Recodificación y armonización de las variables comunes cuando ha sido necesario.
- Evaluación individual de dichas variables, comprobando que su distribución es similar.
- Formulación de modelos explicativos o funciones de enlace de la variable objetivo Z en la base de datos de la EPF (2016). Evaluación de la capacidad explicativa de cada modelo y elección del más adecuado.
- Cálculo de la tasa de riesgo de pobreza en la Encuesta Social de Cantabria (2015).

Ingresos en la ECPF y la ESOC



4.1.2 Indique de cuál de las siguientes fuentes proviene la renta mensual total de su ho

	2015	
	Hogares	%
Ingresos por trabajo	170,367	71.96%
Prestaciones contributivas (pensiones por jubilación,...)	86,428	36.51%
Prestaciones no contributivas	12,135	5.13%
Prestaciones de subsidio por desempleo	20,134	8.50%
Ingreso Mínimo de Inserción	5,464	2.31%
Ayudas públicas por familia	2,642	1.12%
Ayudas públicas por hijos	9,071	3.83%
Ayuda para vivienda (subvenciones para compra o alquiler)	2,254	0.95%
Ingresos por rentas	13,171	5.56%
Transferencias periódicas monetarias percibidas entre hogares	1,839	0.78%
Otros ingresos	6,781	2.86%

- Los ingresos en la EPF son los percibidos regularmente por el hogar y sus miembros, cualquiera que sea su origen, una vez descontadas las cotizaciones a la Seguridad Social y otros pagos asimilados, así como las cantidades satisfechas en concepto de impuestos.
- Rentas procedentes del trabajo por cuenta ajena, por cuenta propia, subsidios, prestaciones, pensiones, rentas del capital y de la propiedad, así como otros ingresos regulares (donaciones, transferencias de otros hogares, remesas de emigrantes, etc...).

Recodificación de variables en EPF y ESOC

Tabla 1: Estructura de las variables comunes (EPF-ESOC)

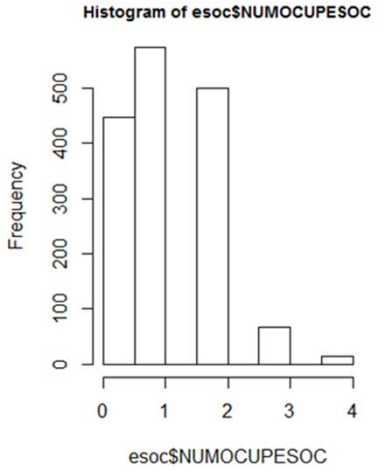
Variables en EPF 2016 y ESOC 2015	Estructura
Régimen de tenencia de la vivienda	Recodificada para armonizarla con la ESOC 2015. Finalmente contiene tres niveles: 1 Propiedad 2 Alquiler 3 Cesión
Disposición o no de otras viviendas	1 Sí 2 No
Tamaño del municipio	Recodificada en EPF, pasado de 6 estratos 5 (por no tener el 6º valores para Cantabria). A su vez, se ha generado la variable "Tamaño del municipio" en la ESOC 2015, a partir de código de municipio del hogar. Finalmente, cuenta con 5 niveles: 1 Menos de 10.000 hab 2 De 10.000 a 20.000 hab 3 De 20.000 a 50.000 hab 4 De 50.000 a 100.000 hab 5 100.000 hab o más
Sexo del/la sustentador/a principal	1 Hombre 0 Mujer
Número de ocupados en el hogar	
Número de personas que reciben ingresos en el hogar	
Ocupación del sustentador principal	1 Ocupado 2 No ocupado
Número de habitaciones de la vivienda	Finalmente excluida por contener un alto número de casos perdidos en la ESOC 2015
Número de mayores de 65 en el hogar	Esta variable se crea en la EPF a partir de las edades individuales de los miembros del hogar

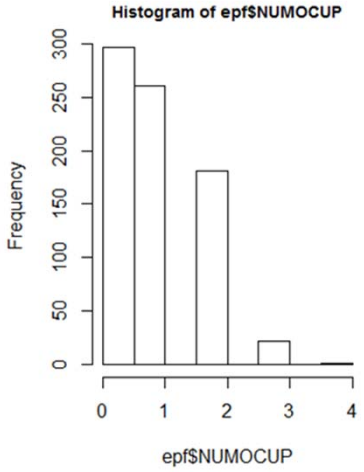
Evaluamos la similitud de la distribución de las variables en cada uno a través de un análisis gráfico, así como de la medida de la Distancia de Hellinger (HD), considerándose adecuado un resultado inferior al 5%, de acuerdo con Leulescu & Agafitei (2013). Resultado que se obtuvo en todos los casos.

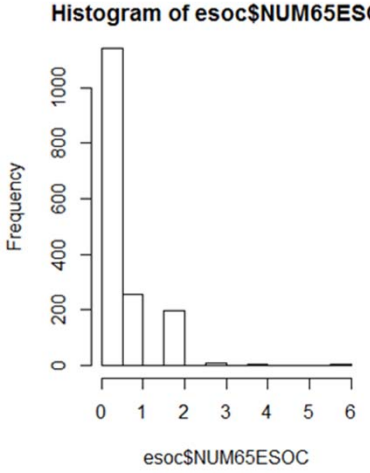
$$HD(V,V') = \sqrt{\frac{1}{2}\cdot\sum_{i=1}^{K}\left(\sqrt{p(V=i)}-\sqrt{p(V'=i)}\right)^2} = \sqrt{\frac{1}{2}\cdot\sum_{i=1}^{K}\left(\sqrt{\frac{n_{Di}}{N_D}}-\sqrt{\frac{n_{Ri}}{N_R}}\right)^2}$$

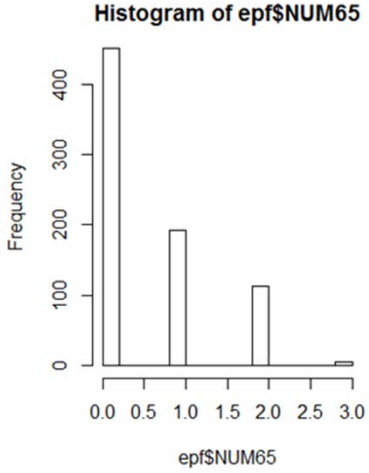
Donde K es el número total de celdas de la tabla de contingencia, n_{Di} es la frecuencia de la celda "i" de los datos del donante D, n_{Ri} es la frecuencia de la celda "i" de los datos del recipiente R y N el tamaño total de la tabla de contingencia específica.

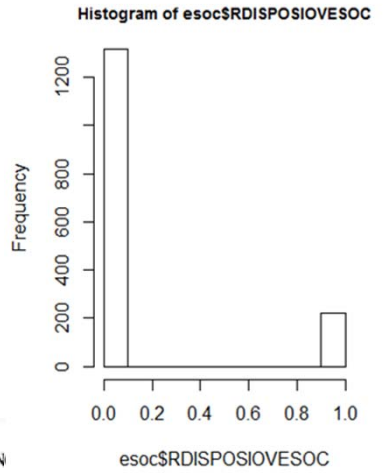
Resultados de la Recodificación I

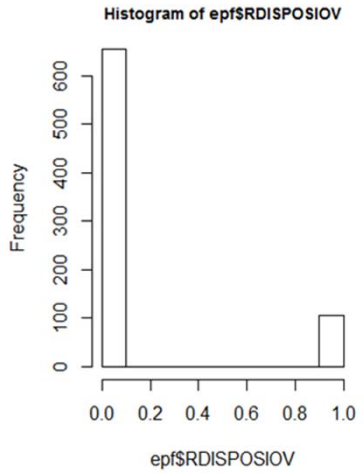


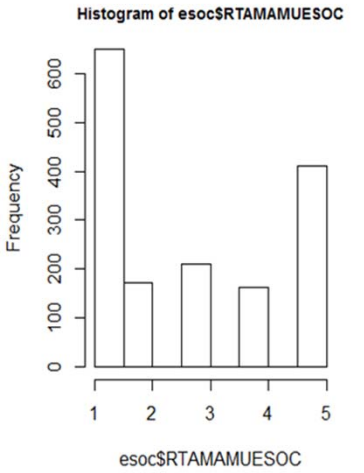


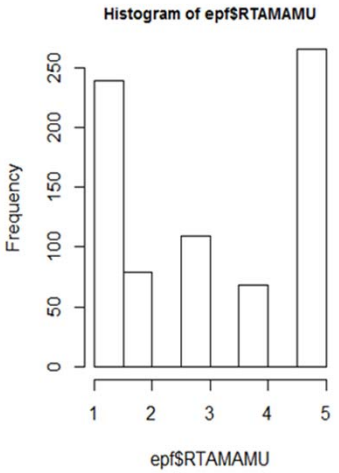




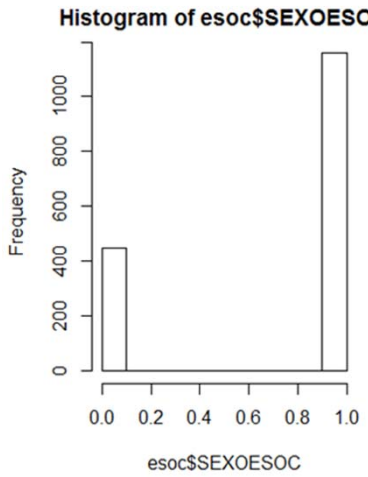


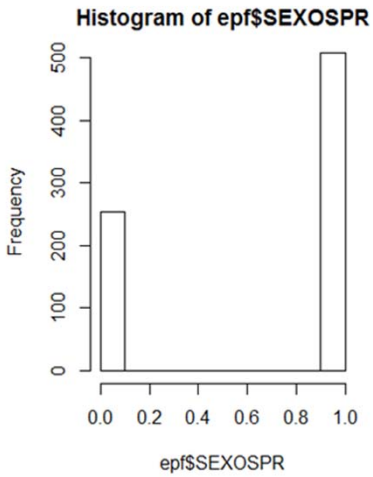


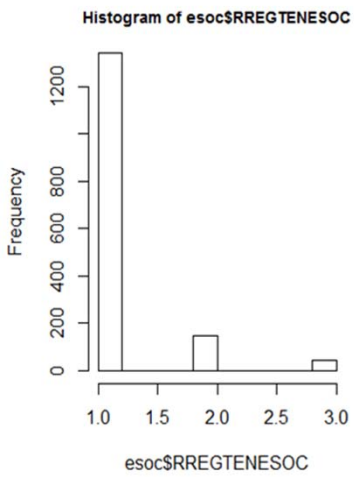


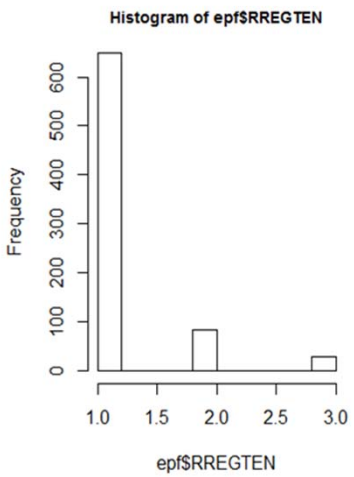


Resultados de la Recodificación II







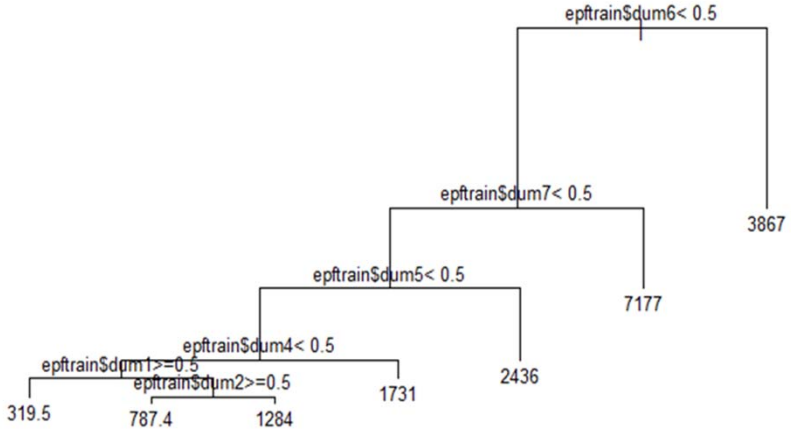


Función de Enlace

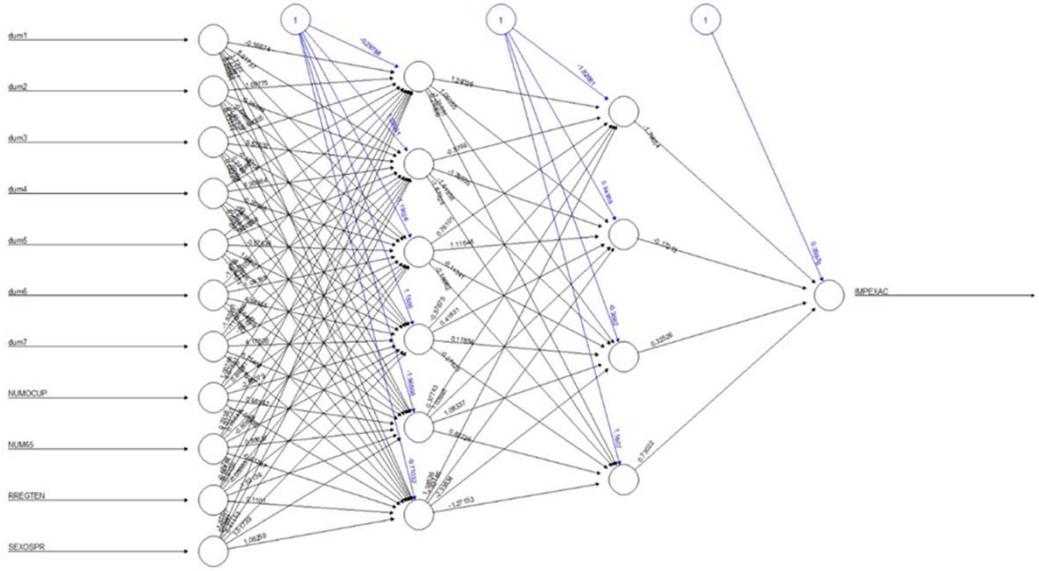
- Utiliza como regresores: los intervalos o tramos de ingresos del hogar manifestados por la persona entrevistada, nº de ocupados, nº de mayores de 65 años, régimen de tenencia de la vivienda, sexo del sustentador principal del hogar.
- Métodos no paramétricos:
  - Árbol de regresión
  - Red neuronal
- Métodos paramétricos:
  - Regresión lineal
  - Regresión gamma
  - Regresión logarítmica

Métodos no paramétricos

Un árbol de regresión. Se trata de un método no paramétrico, basado en la segmentación jerárquica. Se ha utilizado el package "rpart", *Recursive Partitioning and Regression Trees.*



Una red neuronal. Para la aplicación y entrenamiento de la red neuronal, utilizamos el package de R "neuralnet". Se ha segmentado la muestra en dos submuestras, una de test y otra de entrenamiento. La muestra de entrenamiento recoge el 70% de los casos. El objetivo era obtener la definición de red con mejor capacidad explicativa (nº de nodos ocultos).



Resultados Modelos I

Figura 9: Puntuaciones de la variable ingresos en la EPF 2016, estimación de ingresos de la red neuronal y estimación de ingresos del árbol de regresión

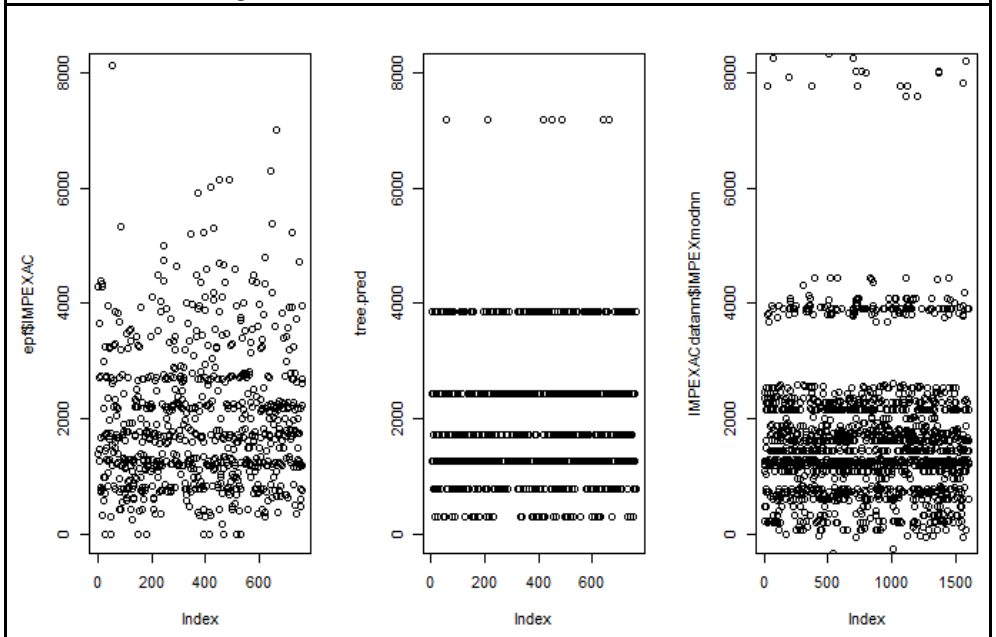
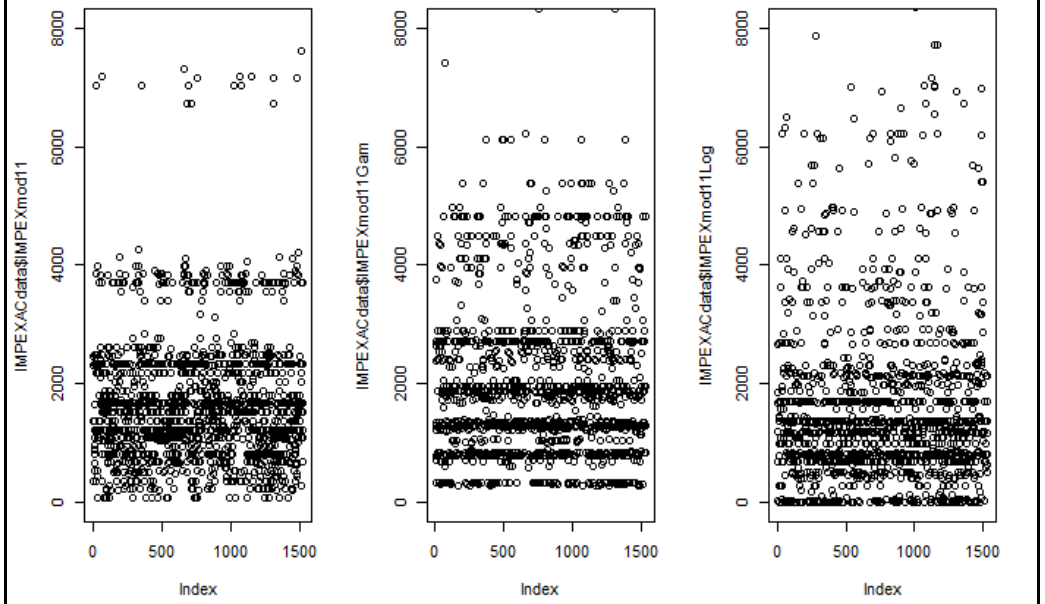


Figura 10: Estimaciones del modelo lineal, modelo gamma y modelo logarítmico.



Resultados Modelos II

Tabla 2: AIC, suma del error cuadrático medio y % de acierto de cada modelo

Modelo	AIC	Suma del error cuadrático medio	% de acierto[1]
Regresión lineal	11325.44	164867.67	87,54%
Regresión gamma	12384.02	494310.80	90,47%
Regresión logarítmica	2300.67[2]	7399601	85,29%
Red neuronal	-	209242.85	79,04%
Árbol de regresión	-	114144.31	94,14%

[1] El % de acierto o tasa de acierto se calcula a partir de la matriz de confusión entre la estimación realizada por el modelo y el intervalo de ingresos de la ESOC para cada hogar. El cociente de los hogares clasificados correctamente (verdaderos positivos), entre todos los elementos clasificados en esa clase (verdaderos positivos + falsos positivos).

[2] No es posible comparar el valor de AIC de la distribución logarítmica con el AIC de las otras distribuciones.

Modelo Elegido para el Matching:

- ### **Modelo Regresión Lineal**

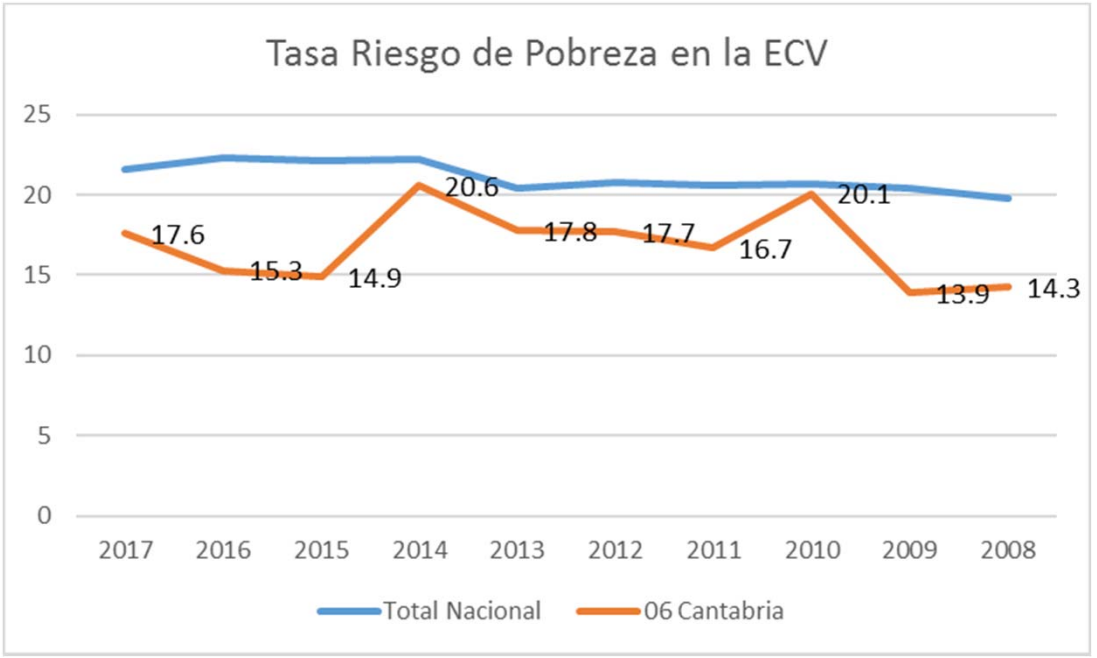
- ```
  attach(epf)
  mod11<- lm(formula = epf$IMPEXAC ~ 0 + dum1 + dum2 + dum3 + dum4 + dum5 + dum6 + dum7 + NUMOCUP + NUM65 + RREGTEN + SEXOSPR, data = epf)
  summary(mod11)
  ```

- ```
  ##
  ## Call:
  ## lm(formula = epf$IMPEXAC ~ 0 + dum1 + dum2 + dum3 + dum4 + dum5 +
  ##     dum6 + dum7 + NUMOCUP + NUM65 + RREGTEN + SEXOSPR, data = epf)
  ##
  ## Residuals:
  ##     Min      1Q  Median      3Q     Max
  ## -1147.7  -171.8    13.8   166.5  3264.7
  ##
  ## Coefficients:
  ##         Estimate Std. Error t value Pr(>|t|)
  ## dum1     -133.05      75.51  -1.762   0.0785 .
  ## dum2      319.02      58.09   5.492 5.44e-08 ***
  ## dum3      726.05      53.71  13.518  < 2e-16 ***
  ## dum4     1163.43      58.70  19.819  < 2e-16 ***
  ## dum5     1826.77      58.86  31.036  < 2e-16 ***
  ## dum6     3193.10      68.89  46.348  < 2e-16 ***
  ## dum7     6518.09     163.93  39.761  < 2e-16 ***
  ## NUMOCUP   146.52      25.40   5.769 1.17e-08 ***
  ## NUM65     138.92      24.12   5.759 1.24e-08 ***
  ## RREGTEN   215.13      29.39   7.320 6.42e-13 ***
  ## SEXOSPR   145.39      31.95   4.551 6.23e-06 ***
  ## ---
  ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  ##
  ## Residual standard error: 409 on 750 degrees of freedom
  ## Multiple R-squared:  0.9686, Adjusted R-squared:  0.9681
  ## F-statistic:  2101 on 11 and 750 DF,  p-value: < 2.2e-16
  ```

Estimación de la tasa de riesgo de pobreza ESOC 2015

Tasa de Riesgo de Pobreza en la ESOC: 0.1928104575

```
#Boostrap para calcular IC
library(boot)
x=ifelse(reg<=tmen,1,0)
u=rep(1,length(x))
datos=data.frame(x,u)
porcentaje <- sum(datos$x)/sum(datos$u)
porcentaje

## [1] 0.1928104575

porcentaje.fun <- function(data,w) {sum(data$x*w)/sum(data$u*w)}
prcent.boot <- boot(datos,porcentaje.fun,R=1000,stype="w",sim="ordinar
y")
boot.ci(prcent.boot,conf=c(0.90,0.95))

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates

CALL :

boot.ci(boot.out = prcent.boot, conf = c(0.9, 0.95))

Intervals :

Level      Normal              Basic
90%   ( 0.1769,  0.2095 )   ( 0.1771,  0.2091 )
95%   ( 0.1738,  0.2126 )   ( 0.1739,  0.2118 )

Level      Percentile          BCa
90%   ( 0.1765,  0.2085 )   ( 0.1771,  0.2093 )
95%   ( 0.1739,  0.2118 )   ( 0.1745,  0.2118 )
```



Conclusiones

- La estadística oficial que recoge información sobre la tasa de pobreza monetaria presenta dificultades muestrales en Cantabria.
- Dificultades metodológicas para la aplicación del matching: diferencias de diseño de las preguntas y casos perdidos.
- La función de enlace utiliza como regresores los intervalos o tramos de ingresos del hogar manifestados por la persona entrevistada, teniendo el resto de variables del modelo (nº de ocupados, nº de mayores de 65 años, régimen de tenencia de la vivienda, sexo del sustentador principal del hogar) un peso relativamente bajo en el modelo, actuando como graduadores del nivel de ingresos entre los intervalos.
- Los modelos no paramétricos son los que ofrecen los peores resultados para nuestros propósitos.
- Se obtuvo un resultado que resulta coherente en relación a las tasas de pobreza conocidas mediante otras fuentes estadísticas hasta la fecha, pero con menor variabilidad estadística.