1. EXECUTIVE SUMMARY

El objetivo de este informe es realizar un modelo de regresión logística mediante el cual podamos tener un modelo para la variable explicada "Hogar en riesgo de pobreza", la cual es cualitativa. A través de los modelos de elección discreta podemos modelizar variables cualitativas en donde del conjunto total de observaciones, fragmentamos la muestra en dos partes, el "tranning· y el "trst" con el 60% y 40% de los datos, respectivamente, realziando a continuación el test ANOVA en donde veremos que las variables estad´siticamente significativas son AyudaFamilias (*), VacacionesOutdoor (***), CapacidadAfrontar (***), LlegarFinMes (***), Miembros (*), HogaresSemanales (***) y finalmente ActMayor (*). Seguidamente realizaremos el test R^2 de McFadden el cual arroja un resultado de 0.3789. Para terminar estableceremos como umbral el 68% para nuestra predicción, de tal forma que valores superiores a este umbral tomarán el valor 1 y 0 en caso contrario. Una vez establecido el umbral generaremos nuestra matriz de confusión, la cual nos servirá para calcular la precisión de nuestro modelo "accuracy", el cual realizado en R nos arroja como resultado una precisión del 74,34%.

2. INTRODUCCIÓN

En primer lugar tendremos que trabajar con la librería "readxl" fundamental para poder abrir el archivo en formato xlsx, "tibble" y "CaTools".

De los datos contenidos en el fichero tenemos 477 observaciones y 18 variables. Lo primero que tenemos que conocer es el significado que aporta cada una de las variables y para ello sustituiremos cada uno de los índices de las columnas por su definición contenida en el fichero Word.

En segundo lugar, seleccionamos las variables que nos interesan para nuestro estudio. Dicho de otro modo, las variables que eliminamos son "Hogar", "TVcolor", "RentaTotalAnterior", "Region", "RentaMenos16" y "SexoMayor", por lo que hemos pasado de trabajar con 18 a 12 variables.

En tercer lugar, todas nuestras variables están en formato numérico, algunas de ellas nos interesa convertirlas en factor, por lo que realizamos la conversión de las variables "VacacionesOutdoor", "CapacidadAfrontar", "Ordenador", "LlegarFinMes", "RegimenTenencia", "SexoMayor" y "ActMayor" y construimos nuestro data frame.

3. MODELO DE REGRESIÓN LOGÍSTICA

Nos creamos una semilla llamada 123 para que a través de las muestras aleatorias nos den los mismos resultados si los deseamos repetir y dividimos la muestra en dos partes. Por una parte, está el "training.set", el cual contiene el 60% del total, esto es, 286 observaciones, y por otra parte el "test" que contiene el 40% restante, es decir, 191 obsercaciones.

A continuación realizamos nuestro modelo de regresión glm binomial sobre nuestra variable explicativa "HogarPobreza" para el 60% de las observaciones y realizamos el test ANOVA con la Chi-cuadrado:

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			285	374.94		
AyudaFamilias	1	4.474	284	370.46	0.0344175	*
VacacionesOutdoor	1	45.008	283	325.45	1.962e-11	***
CapacidadAfrontar	1	23.414	282	302.04	1.306e-06	***
Ordenador	2	0.875	280	301.16	0.6455544	
LlegarFinMes	5	21.940	275	279.22	0.0005376	***
RegimenTenencia	4	4.782	271	274.44	0.3103709	
Miembros	1	6.523	270	267.92	0.0106505	*
EdadMayor	1	1.456	269	266.46	0.2275871	
HorasSemanales	1	16.048	268	250.41	6.175e-05	***
Mayores16	1	0.210	267	250.21	0.6468497	
ActMayor	8	17.337	259	232.87	0.0267840	*

Nuestras variables estadísticamente significativas:

- AyudaFamilias (*)
- VacacionesOutdoor (***)
- CapacidadAfrontar (***)
- LlegarFinMes (***)
- Miembros (*)
- HogaresSemanales (***)
- ActMayor (*)

Una vez establecidas cuáles son nuestras variables estadísticamente signficativas tendremos que comprobar la bondad del ajuste. A través del comando "pR2".

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu

```
> pR2(modelo01, 4)
         llh      llhNull           G2     McFadden         r2ML         r2CU
-116.4339686 -187.4677873  142.0676375    0.3789121    0.3914888    0.5359632
```

El pseudo R^2 de McFadden representa lo que se reduce la "desviance", la cual es poroporcional al aumento de verosimilitud del modelo. Nos arroja un resultado de 0.3789. Este estadístico calculado cuanto más se acerque a 1 indica que mejor ajustado estará el modelo.

A continuación, por medio del comando fitted.results y a través de nuestro "test" establecemos como umbral el 68% para nuestra predicción, de tal forma que valores superiores a este umbral tomarán el valor 1 y 0 en caso contrario.

El modelo logit tiene la ventaja sobre los modelos de probabilidad, que las probabilidades calculadas siempre están comprendidas entre 0 y 1, con lo cual se evita el tener que hacer aproximaciones a 0,01 cuando las probabilidades son negativas, o a 0,99 cuando son mayores a 1.

Aunque el modelo logit es lineal en las variables explicativas, las probabilidades en sí mismas no lo son, lo cual contrasta con el modelo lineal de probabilidad en donde las probabilidades aumentan linealmente con las variables independientes.

Con todos estos cálculos nuestra matriz de confusión, que no es más que una tabla bidimensional donde se analiza la relación entre los valores observados y los predichos del cual obtenemos lo siguiente:

```
> logit.perf
      Predicted
Actual   0   1
     0 102   7
     1  42  40
```

Fijandonos en nuestra matriz observamos que (0,0) y (1,1) cuentan con la mayoría del total de las observaciones, esto es, 102 y 40, respectivamente. En cambio, para los valores (1,0) y (0,1) contamos solo con 42 y 7, respectivamente, siendo estas las observaciones fuera de nuestra predicción.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu

Tomado todo ello en su conjunto podemos calcular la precisión de nuestro modelo "accuracy", la cual es la proporción entre las predicciones correctas que ha hechogenerado nuestro modelo y el total de las predicciones. Dicho de otro modo, esta precisión será equivalente a restar el ratio de error de la unidad: 1- ratio de error.

Explicado de otro modo, el Accuracy se refiere a la dispersión del conjunto de valores obtenidos a partir de las mediciones repetidas, de tal forma que cuanto menor es la dispersión mayor es la precisión y se representa por la proporción entre el número de predicciones correctas (tanto positicas como negativas) y el total de prediciones, el cual realizado en R nos arroja como resultado una precisión del 74,34%.

4. CONCLUSIONES

Los modelos de regresión logística son un tipo de análisis de regresión que nos ha servido para predecir el resultado de una variable categórica "Hogar riesgo pobreza" en función de las variables independientes o predictoras, de las cuales hemos trabajado finalmente con 12 de ellas. Realizando el test ANOVA, el test R^2 de McFadden y la matriz de confusión nos sale una bondad del ajuste de nuestro modelo de un 74,34%.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu