

ANÁLISIS DEL COMPORTAMIENTO ECONÓMICO DE LOS HOGARES ESPAÑOLES RESPECTO AL GRADO DE FACILIDAD PARA LLEGAR A FIN DE MES, MEDIANTE LA APLICACIÓN DE MODELOS DE PREDICCIÓN CATEGÓRICOS.

Miguel Ángel Fajardo Caldera

Departamento de Economía Aplicada y Organización de Empresas

Universidad de Extremadura

e-mail: fajardo@unex.es

Jesús Pérez Mayo

Departamento de Economía Aplicada y Organización de Empresas

Universidad de Extremadura

e-mail: jperez@unex.es

Resumen

El Panel de Hogares de la Unión Europea (PHOGUE) incluye algunas preguntas cuya finalidad es conocer la opinión de los hogares sobre determinados aspectos relacionados con su situación económica. En concreto, una de estas preguntas se refiere al grado de facilidad para llegar a fin de mes, pregunta subjetiva y categórica, ya que los hogares responden según su propia valoración de estos aspectos, teniendo en cuenta sus ingresos mensuales del que disponen y categórica, porque la encuesta ofrece a los hogares seis alternativas de respuesta, tres opciones que indican un mayor o menor grado de dificultad y otras tres referidas a menor o mayor grado de facilidad.,

Los autores realizan hipótesis sobre el comportamiento de los hogares en el grado de facilidad para llegar a fin de mes, basados en los datos obtenidos del Panel de Hogares de la Unión Europea (PHOGUE) para los hogares españoles, durante los años 1994, 1995 y 1996. Estas hipótesis serán contrastadas a través de modelos categóricos para el Análisis de Predicción.

Descriptores: Panel de Hogares de la Unión Europea (PHOGUE); modelos log-lineales; Análisis de predicción en modelos categóricos.

1. Introducción

Uno de los objetivos de análisis del Gobierno como de los investigadores económicos es conocer la situación económica de los hogares españoles, con el fin de mejorar aquellas situaciones económicas más desfavorables para ellos, mediante la toma de medidas políticas y económicas que puedan resolverlas.

Actualmente, es de todos conocidos la preocupación de organismos nacionales e internacionales sobre el excesivo endeudamiento de los hogares españoles en la adquisición de una vivienda propia, lo que ha propiciado una serie de advertencias, artículos y opiniones sobre las posibles consecuencias que pueden ocurrir si las condiciones económicas iniciales que han permitido este endeudamiento de los hogares cambiasen a otras condiciones menos favorables, especialmente para aquellos hogares que parten de una situación económica más desfavorable.

Los autores han querido analizar en este trabajo cuál ha sido el comportamiento de los hogares españoles ante la dificultad de llegar a fin de mes, teniendo en cuenta sus ingresos mensuales netos, contrastando los posibles cambios que hayan podido darse en la dificultad de llegar a fin de mes de esos hogares, lo que nos podría servir de análisis de la capacidad de cambio de los hogares españoles ante situaciones cambiantes.

Para poder realizar este análisis, vamos a utilizar el Panel de Hogares de la Unión Europea (PHOGUE), que es una encuesta que se realiza a diversos países de la Unión Europea desde 1994. Esta incluye algunas preguntas cuya finalidad es conocer la opinión de los hogares sobre determinados aspectos relacionados con su situación económica. En concreto, una de estas preguntas se refiere al grado de facilidad para llegar a fin de mes, siendo esta pregunta subjetiva, porque los hogares responden según su propia valoración de estos aspectos, teniendo en cuenta sus ingresos mensuales netos del que disponen y categórica, porque la encuesta ofrece a los hogares seis alternativas de respuesta, tres opciones que indican un mayor o menor grado de dificultad (con mucha dificultad, con dificultad, con cierta dificultad) y otras tres referidas a menor o mayor grado de facilidad (con cierta facilidad, con facilidad y con mucha facilidad).

Para realizar este análisis es necesario establecer primeramente una relación funcional (global) entre las variables criterios y las variables predictoras y posteriormente

establecer aquellos valores de las variables predictoras que son de interés para obtener predicciones de las variables criterios.

Cuando analizamos variables continuas, frecuentemente estamos interesados en realizar predicciones utilizando métodos tales como el análisis de regresión. Tales métodos permiten predecir un valor de una variable dependiente (variable criterio) a partir de un valor establecido para cada una de las variables independientes (variables predictoras).

En el análisis de predicción en tablas de n-vías de clasificación cruzada para variables categóricas, el procedimiento de realizar predicciones sigue de una manera formal las mismas pautas que en el análisis de regresión analizado anteriormente. En primer lugar establecemos una relación logística o proposicional entre las variables predictoras y criterios, análoga a la relación funcional del análisis de regresión, de tal forma que esta relación global no sea, mediante el cálculo de proposiciones, ni contradictoria ni tautológica; y en segundo lugar, descomponer esta relación global en otras relaciones particulares entre las categorías de las variables predictoras y criterios que nos permitirán analizar la hipótesis global.

En el caso del tipo de datos con los que vamos a trabajar, datos longitudinales, uno de los objetivos es predecir los comportamientos de los hogares, en cuanto a los posibles cambios que pueden producirse en el grado de facilidad para llegar a fin de mes y en que sentido son estos cambios. Una de las técnicas estadísticas que nos permiten realizar este análisis es a través de los modelos categóricos de Análisis de Predicción.

El Análisis de Predicción (AP), desarrollado por Hildebrand, Laing y Rosenthal (1977) y extendido por von Eye y Brandtstädter (1988a; von Eye, Brandtstädter, & Rovine, 1993), es una herramienta estadística creada para predecir sucesos empíricos usando variables categóricas. El Análisis de Predicción es un método para testar hipótesis específicas que están recogidas en las llamadas celdas de aciertos y errores. Las celdas aciertos contienen casos que son consistentes con la predicción o hipótesis bajo investigación, mientras que las celdas error contienen casos que lógicamente contradicen la hipótesis. Las hipótesis son testadas comparando la distribución de frecuencias observadas con la distribución de frecuencias esperadas.

2. Fases de un análisis de predicción en clasificaciones cruzadas de variables categóricas

Para realizar un Análisis de Predicción entre variables categóricas de clasificación cruzada es necesario seguir las siguientes fases:

1. Formulación de hipótesis de predicción global y descomposición de la hipótesis global en hipótesis parciales.
2. Identificación de celdas que contienen casos que son consistentes con las hipótesis (celdas aciertos) y aquellas celdas que contienen casos que no son consistentes con las hipótesis (celdas errores).
3. Estimación de frecuencias esperadas.
4. Evaluación estadística de la hipótesis global y de las hipótesis parciales.

2.1. Formulación de hipótesis de predicción

Los elementos básicos de las predicciones son las proposiciones, es decir, sentencias de las cuales tiene sentido catalogarlas como ciertas (T) o falsas (F). La combinación de dos o más proposiciones simples las denominaremos proposiciones compuestas. Las proposiciones se combinan a partir de los siguientes operadores lógicos: conjunción (AND, \wedge), disyunción (OR, \vee), implicación (IF-THEN, \rightarrow) y negación (NOT, \neg).

Una vía de determinar la verdad de las proposiciones compuestas es a través de las llamadas tablas de verdad, de las cuales exponemos un ejemplo a continuación:

TABLAS DE VERDAD

Prop. Elementales		Proposiciones compuestas			
a	b	$a \wedge b$	$a \vee b$	$a \rightarrow b$	$\neg a$
T	T	T	T	T	F
T	F	F	T	F	F
F	T	F	T	T	T
F	F	F	F	T	T

Se dice que una proposición compuesta es contradictoria si ella siempre es falsa, sean cuales sean los valores que tomen las proposiciones elementales. Por el contrario, decimos que es tautológica si es siempre cierta, sean cuales sean los valores de las proposiciones elementales. Por ejemplo, en la siguiente tabla se especifican dos proposiciones compuestas, de las cuales (1) es una contradicción y (2) tautología (2):

Proposiciones elementales		Proposiciones compuestas	
a	$\neg a$	(1) $a \wedge (\neg a)$	(2) $a \vee (\neg a)$
T	F	F	T
F	T	F	T

Las predicciones en el análisis de predicción se basan, en general, en las relaciones IF-THEN. Los modelos se formulan entre predictores y criterios mediante la relación lógica de implicación; y las categorías de predictores y categorías de criterios con conjunción y disyunción.

Sean por ejemplo tres variables categóricas A, B y C con a, b y c categorías respectivamente. Consideremos A y B como variables predictoras y C como variable criterio. El modelo de predicción global entre predictores [P] y criterio [C] viene dado por:

[P] \rightarrow [C] viene dado por: $A \wedge B \rightarrow C$. Hipótesis Global.

Dentro de este modelo de predicción global especificamos un conjunto de predicciones parciales entre las categorías, por ejemplo, si (a₁, a₂) son las categorías de la variable A, (b₁, b₂) las de B y (c₁, c₂, c₃) las de C, podemos establecer las siguientes hipótesis parciales:

$$H: a_1 \wedge b_1 \rightarrow c_2$$

$$a_1 \wedge b_2 \rightarrow c_1 \text{ Predicciones parciales}$$

$$a_2 \rightarrow (c_2 \vee c_3)$$

Cuando se establecen relaciones entre diseños (combinaciones de categorías) de la misma variable (predictor o criterio) usualmente utilizamos el operador OR y cuando son diseños de distintas variables usaremos el operador AND.

Es fundamental que la predicción global, formulada a través de predicciones parciales sea no-tautológica y no-contradictoria.

2.2. Identificación de celdas con respecto a las hipótesis.

Cuando realizamos hipótesis de predicción para un conjunto de variables categóricas, la matriz de datos asociada a la distribución conjunta de frecuencias, contiene las celdas que son confirmatorias de las hipótesis, que denominaremos celdas ACIERTOS y aquellas que no son confirmatorias con las hipótesis enunciadas y que denominaremos celdas ERRORES. Por tanto, cada hipótesis establecida contendrá celdas de ACIERTOS y celdas de ERRORES.

Así, en el caso anteriormente expuesto de tres variables categóricas, tendríamos la siguiente tabla de clasificación cruzada:

TABLA DE CLASIFICACIÓN CRUZADA PARA LAS VARIABLES CATEGÓRICAS A,B,C.	
INDICES DE A,B,C	ESTATUS
$a_1b_1c_1$	ERROR
$a_1b_1c_2$	ACIERTO
$a_1b_1c_3$	ERROR
$a_1b_2c_1$	ACIERTO
$a_1b_2c_2$	ERROR
$a_1b_2c_3$	ERROR
$a_2b_1c_1$	ERROR
$a_2b_1c_2$	ACIERTO
$a_2b_1c_3$	ACIERTO
$a_2b_2c_1$	ERROR
$a_2b_2c_2$	ACIERTO
$a_2b_2c_3$	ACIERTO

2.3. Estimación de frecuencias esperadas.

Diversas aproximaciones se han realizado para abordar este problema. La primera de ellas era propuesta por Hildebrand, Laing y Rosenthal (1994, 1997). Para el cálculo de las estimaciones de las frecuencias esperadas proponen un modelo log-lineal estándar jerárquico de la siguiente forma:

Sean [P] y [C], donde [P] es el conjunto de predictores y [C] el conjunto criterio. El modelo presenta las siguientes características:

1. Es saturado en los predictores (efectos principales e interacciones).

2. Es saturado en los criterios.
3. Se asume la independencia entre predictores y criterios.

El modelo se acepta si existen desviaciones respecto de la independencia, es decir, existen relaciones entre predictores y criterios. Su medida de las desviaciones del supuesto de independencia se hace a través de medidas de Reducción Proporcional del error (PRE). Una medida que utiliza es la medida ∇ , relacionada con la medida de asociación asintótica de Goodman & Kruskal (1954).

Como alternativa a esta primera aproximación Goodman & Kruskal (1974) proponen utilizar modelos log-lineal de cuasi-independencia. Para ello, declara a las celdas de error como ceros estructurales y de este modo las excluye de la estimación de las frecuencias esperadas. El problema es la falta de grados de libertad para medir el ajuste del modelo.

La tercera aproximación corresponde a Hurbert (1979) que utiliza modelos de emparejamientos.

La cuarta aproximación se debe a Von Eye (1993) que utiliza de nuevo los modelos log-lineal jerárquicos no estándar a través de la formulación siguiente:

$$\text{Modelo log- lineal: } \log \text{ Freq} = X_k \lambda_k + X_l \lambda_l$$

donde el índice k indica los vectores que pertenecen al primer grupo (modelo base, relaciones entre predictores y criterios) y l indica los vectores que pertenecen al segundo grupo (vectores de las hipótesis de predicción).

La matriz de diseño del primer grupo de vectores X_k esta dividida en dos partes, la primera especifica los efectos relativos a los predictores y la segunda parte especifica los efectos relativos a los criterios. Estas relaciones van a constituir lo que llamaremos el **modelo log-lineal base** para las hipótesis de predicción (PA).

La matriz del segundo grupo X_l , refleja las relaciones realizadas relativas a las hipótesis de predicción. Los parámetros correspondientes a las restricciones e interacciones de este grupo, suelen definirse de la siguiente forma:

$$\lambda_{ij} = \frac{\gamma_i}{h_i}, \text{ si la celda } ij \text{ es una celda acierto.}$$

y

$$\lambda_{ij} = \frac{\gamma_i}{e_i}, \text{ si la celda } ij \text{ es una celda error.}$$

siendo h_i el número de celdas hit en la fila i , y e_i el número de celdas error en la fila i . El parámetro γ_i corresponde a la hipótesis de predicción establecida en la fila i .

Por tanto, bajo estas restricciones, el modelo anterior queda de la forma:

$$\log F = X_i \lambda_i + \gamma_i \left[\frac{1}{h_i} \delta_{ij} - \frac{1}{e_i} (1 - \delta_{ij}) \right]$$

donde $\delta_{ij} = 1$ si la celda es acierto y $\delta_{ij}=0$ si la celda es error.

Una alternativa a este diseño para las hipótesis de predicción viene dada por:

$$X_i = \begin{cases} 1 & \text{si la celda } i \text{ confirma la predicción elemental } j \\ -1 & \text{si la celda } i \text{ no confirma la predicción elemental } j \\ 0 & \text{en otro caso} \end{cases}$$

Los autores de este trabajo utilizaran este último diseño junto a la aproximación debida a Von Eye, para la contrastación de las hipótesis de predicción establecidas.

3. Matriz de datos

En este trabajo utilizamos los datos correspondientes a 1994, 1995 y 1996, para los hogares españoles del PHOGUE. Para centrarnos en el problema que queremos estudiar, respecto a la pregunta sobre el grado de dificultad para llegar a fin de mes de acuerdo con los ingresos netos mensuales percibidos por el hogar; hemos tomado una muestra de 5.714 hogares, donde hemos construido las variables M_i , $i = 1994, 1995$ y 1996 , las cuales son del tipo cualitativas dicotómicas formuladas de la forma siguiente:

$$M_i = \begin{cases} 1 & \text{si el hogar ha tenido dificultad para llegar a fin de mes en el año } i \\ -1 & \text{si el hogar ha tenido facilidad para llegar a fin de mes en el año } i \end{cases}$$

con $i = 94, 95, 96$.

Los datos observados se recogen en la siguiente tabla:

MATRIZ DE DATOS

M94	M95	M96	Frecuencia
1	1	1	2.687
1	1	2	445
1	2	1	411
1	2	2	377
2	1	1	370
2	1	2	235
2	2	1	274
2	2	2	915

4. Hipótesis

Los autores se plantean las siguientes hipótesis:

1. Los hogares que en el año 1994 señalaron que tenían dificultad para llegar a fin de mes, presentan una mayor probabilidad de volver a encontrarse en esa misma situación en años sucesivos, que en otras situaciones alternativas (celdas 112, 121, 122).
2. Los hogares que en el año 1994 señalaron que tenían facilidad para llegar a fin de mes, presentan una mayor probabilidad de volver a encontrarse en esta misma situación en años sucesivos, que en otras situaciones alternativas (celdas 211, 212, 221).

Por tanto, la hipótesis global viene definida por:

$$M94 \Rightarrow M95 \wedge M96.$$

siendo las hipótesis parciales correspondientes las dadas por:

Hipótesis parcial 1: $M94 (1) \Rightarrow M95 (1) \wedge M96 (1)$ celda 111

Hipótesis parcial 2: $M94 (2) \Rightarrow M95 (2) \wedge M96 (2)$ celda 222

Atendiendo a las hipótesis, tenemos que las celdas 111 y 222 son las celdas aciertos o celdas que son consistentes con la hipótesis parciales 1 y 2 respectivamente mientras que las celdas 112,121 y 122 son celdas errores o celdas no consistentes con la predicción parcial 1 y las celdas 211, 212 y 221 son celdas errores para la hipótesis parcial 2.

Puede probarse que este conjunto de hipótesis es no contradictorio y no tautológico.

5. Planteamiento del modelo

Basándonos en las hipótesis formuladas en el apartado anterior, consideramos el modelo log-lineal que envuelve una variable predictor (grado de dificultad en 1994) y dos variables criterios (grados de dificultad en 1995 y 1996 respectivamente).

El modelo base viene dado por el conjunto predictor [P]= [M94] y el conjunto criterio [C]= [M95, M96], cuya expresión formal viene dada por:

$$\text{Log Frecuencia} = \text{intercepto} + M94 + M95 + M96 + M95:M96$$

Este modelo nos indica la independencia estadística entre el predictor y el conjunto criterio. La estimación del modelo nos lleva a ser rechazado, pues $LR-X^2 = 1496,109$ con $g.l = 3$ y $p=0,00$; lo que nos indica que existe una relación entre el predictor y el conjunto criterio.

Si añadimos los vectores correspondientes a la formulación de las hipótesis de predicción parcial formuladas, obtenemos la siguiente tabla o matriz de diseño:

Índices	Vectores de efectos principales e interacción				Vectores específicos para la predicción de hipótesis	
	M94	M95	M96	M95*M96	V1	V2
111	1	1	1	1	1	0
112	1	1	-1	-1	-1	0
121	1	-1	1	-1	-1	0
122	1	-1	-1	1	-1	0
211	-1	1	1	1	0	-1
212	-1	1	-1	-1	0	-1
221	-1	-1	1	-1	0	-1

222	-1	-1	-1	1	0	1
-----	----	----	----	---	---	---

Si estimamos el modelo completo formado por el modelo base más los dos vectores de correspondientes a las hipótesis parciales tenemos:

$$\text{Modelo: log frecuencia} = \text{modelo base} + V1 + V2$$

Siendo la estimación de sus coeficientes dadas por:

Coefficientes:

(Intercepto)	M94	M95	M96	M95:M96	V1	V2
6.57838	0.27399	0.12344	0.12710	0.06183	0.73143	0.70325

Con una verosimilitud dada por LR-X2=4.32365, con 1 grado de libertad y para $p < 0,03758623$ se rechazaría la hipótesis, lo que nos indica que al 1% del nivel de significación podemos señalar que los vectores correspondientes a las hipótesis parciales explican gran parte de la independencia entre los predictores y criterios.

Si atendemos a la significatividad de los coeficientes, tenemos que:

Coefficientes:

Valor	Error Std.	t valor	
(Intercepto)	6.57837882	0.01695205	388.058050
M94	0.27399968	0.02064625	13.271161
M95	0.12343853	0.02671436	4.620681
M96	0.12710155	0.02671436	4.757799
M95:M96	0.06183439	0.02303274	2.684630
V1	0.73142762	0.03939454	18.566727
V2	0.70325062	0.04146767	16.959011

Podemos observar en la tabla anterior que todos los coeficientes son altamente significativos, lo que nos indica que todos los coeficientes son diferentes de cero y que las hipótesis parciales no son rechazadas por los datos.

En cuanto a la correlación de los coeficientes obtenemos que:

Coefficientes de correlación:

	(Intercepto)	M94	M95	M96	M95:M96	V1
M94	-0.1287039					
M95	-0.1084312	-0.6068364				
M96	-0.1113940	-0.6068364	0.4867217			
M95:M96	-0.5014847	-0.0065955	0.0849025	0.0870831		
V1	0.2187287	0.4725621	-0.6931248	-0.6931248	-0.4820430	
V2	0.0527216	-0.5468388	0.6528103	0.6528103	-0.2429179	-0.4794402

Es interesante observar las correlaciones negativas entre los coeficientes de la hipótesis 1 (V2) y las variables M94 y M95; y las positivas de la hipótesis 1 (V2) y las variables M94 y M95.

En la tabla de frecuencias observadas y esperadas, así como los residuos estandarizados se presentan en la siguiente tabla:

Índices de celdas	Frecuencias		Residuos estandarizados
	Observadas	Esperadas	
M94M95M96			
111	2.687	2.687,00	0,00
112	445	426,43	0.89
121	411	429,57	-0.90
122	377	377,00	0,00
211	370	370.00	-0,00
212	235	253,57	-1.18
221	274	255,43	1.15
222	915	915,00	0,00

Los residuos estandarizados nos indica un buen ajuste del modelo global a los datos observados.

6. Conclusiones:

1. A través de la estimación de los modelos log-lineal del modelo de relación de independencia entre predictor (M94) y conjunto criterio (M95, M96) ha sido rechazada, por lo tanto se pueden establecer relaciones de predicción entre ellos.

2. Las hipótesis H1 y H2 parciales establecidas son altamente significativas, lo que nos indica que ante cambios económicos desfavorables que pudieran producirse, los hogares que tiene una situación económica más desfavorables (medida a través de llegar a fin de mes con dificultades) y que, en general, tienen una mayor probabilidad de persistir en ella, podrían agravar en mayor medida su situación económica que a aquellos que tiene una situación económica más favorable.

3. Las hipótesis estadísticas parciales de predicción H1 y H2, capturan gran parte de la dependencia entre las variables predictoras t variables criterios.

Bibliografía

- Agresti, A. (1984). Analysis of ordinal categorical data. New York: Wiley.

- Agresti, A. (1996). An introduction to categorical data analysis. New York: Wiley.

- Christensen, R. (1997) Log-linear models and logistic regression (2ª ed.) New York: Springer.

- Clogg, C.C., Eliason, S.R., & Grego, J. (1990). Models for the analysis of change in discrete variables. In A. von Eye (Ed.), Statistical methods for longitudinal research (Vol. 2, pp. 409-441). New York: Academic Press.

- Hildebrand D.K., Laing, J.D., & Rosenthal, H. (1977) Prediction analysis of cross-classifications. New York: Wiley.

- von Eye, A. (1990). Introduction to Configural Frequency Analysis: The search for types and antitypes in cross-classifications. Cambridge: Cambridge University Press.

- von Eye, A. (Ed.). (1990). Statistical methods in longitudinal research, Vol. 1: Principles and structuring change. New York: Academic Press.

- von Eye, A. (Ed.). (1990). Statistical methods in longitudinal research, Vol. 2: Time series and categorical longitudinal data. New York: Academic Press.

- von Eye, A. (2002). *Configural Frequency Analysis - Methods, Models, and Applications*. Mahwah, NJ: Lawrence Erlbaum.
- .von Eye, A., & Brandtstädter (1988). Formulating and testing developmental hypotheses using statement calculus and nonparametric statistics. In P.B.Balttes, D. Featherman, & R.M. Lerner (Eds.).*Life-span development and behaviour* (Vol 8, pp. 61-97). Hillsdale, NJ: Lawrence Erlbaum Associates.
- von Eye, A., Brandtstädter, J., & Rovine, M.J. 1993). Models for prediction analysis. *Journal of mathematical Sociology*, 22, 235-371.
- .von Eye, A., & Clogg, C.C. (Eds.). (1994). *Latent variables analysis - Applications for developmental research*. Newbury Park, CA: Sage.
- von Eye, A., & Clogg, C.C. (Eds.) (1996). *Analysis of categorical variables in developmental research*. San Diego, CA: Academic Press.
- von Eye, A., & Schuster, C. (1998). *Regression analysis for social sciences - models and applications*. San Diego: Academic Press.