Revista de Matemática: Teoría y Aplicaciones 2010 **17**(2) : 199–230
cimpa – ucr issn: 1409-2433

ANÁLISIS DE COMPONENTES PRINCIPALES Y ANÁLISIS DE REGRESIÓN PARA DATOS CATEGÓRICOS. APLICACIÓN EN LA HIPERTENSIÓN ARTERIAL

PRINCIPAL COMPONENT AND REGRESSION ANALYSIS FOR CATEGORICAL DATA. APPLICATION TO ARTERIAL HYPERTENSION

Juan M. Navarro Céspedes *
Gladys M. Casas Cardoso†
Emilio González Rodríguez‡

Received: 11 Nov 2008; Revised: 16 Oct 2009; Accepted: 6 Apr 2010

*Grupo de Estadística, Facultad de Matemática, Física y Computación, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba. E-Mail: juanma@uclv.edu.cu

†Laboratorio de Bioinformática, Facultad de Matemática, Física y Computación, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba. E-Mail: gcasas@uclv.edu.cu

‡Centro de Desarrollo de la Electrónica, Universidad Central "Marta Abreu" de Las Villas, La Habana, Cuba. E-Mail: eglez@uclv.edu.cu

Resumen

El presente trabajo aborda el tema relacionado con el procesamiento estadístico de variables categóricas. Se explican los fundamentos matemáticos del análisis de Componentes Principales y del análisis de Regresión para datos categóricos. La unión de estas técnicas puede utilizarse para resolver problemas de clasificación. Debido a que estos son métodos relativamente nuevos, se decide utilizar otra técnica más conocida (árboles de clasificación siguiendo criterios chi cuadrado) para realizar comparaciones de sus los resultados, con ayuda de la teoría de las curvas ROC. En la aplicación desarrollada se estudiaron pacientes supuestamente sanos del municipio de Santa Clara, Cuba, diagnosticados como hipertensos, pre hipertensos y normotensos por un Comité de Expertos Médicos altamente calificados. La regresión categórica unida al análisis de Componentes Principales como método de selección de variables, resultó ser la mejor técnica ara resolver el problema de clasificación.

Palabras clave: regresión categórica, hipertensión arterial, clasificación, curvas ROC.

Abstract

The present work is about the statistical processing of categorical data. The mathematical details of the Categorical Principal Components and the Categorical Regression Analysis are explained. The combination of both techniques can be used to solve classification problems. Because these techniques are relatively new, we decided to use another technique (classification trees following the chi squared criteria) to make a comparison of their results, with the help of the theory of ROC curves.

In the application, supposedly healthy patients of Santa Clara, Cuba, were diagnosed as hypertensive, pre hypertensive and no hypertensive by a Committee of Medical Experts. Categorical Component Analysis and Categorical Regression Analysis were applied in order to successfully solve the classification problem.

Keywords: categorical regression, arterial hypertension, classifiers, ROC curves.

Mathematics Subject Classification: 62P10.

1 Introducción

El cambiante mundo moderno está sustentado por un conjunto de ciencias empleadas por el hombre para, entre otras cosas, controlar y perfeccionar

los procesos; tal es el caso de la Estadística. En los últimos años se han desarrollado varios métodos que se ocupan de los modelos matemáticos en general, métodos que han sido automatizados gracias al desarrollo de la informática, por lo que resultan de gran utilidad práctica para solucionar problemas presentes en la sociedad.

En las investigaciones de corte social, fundamentalmente intervienen conjuntos de datos que reflejan alguna cualidad o categoría. A estos datos se les conoce como datos categóricos. Dichos datos pueden contener una mezcla de diferentes tipos de variables, muchas de las cuales están medidas en categorías ordenadas o desordenadas. Variables como las estaciones del año, los tipos de determinado producto en el mercado, o el hecho que un estudiante apruebe o no un examen, son ejemplos de variables con categorías desordenadas. Variables como el nivel de educación o la frecuencia con que se desarrolla cierta actividad, (poca, regular o mucha) son ejemplos de variables con categorías ordenadas. Las variables continuas pueden considerarse variables categóricas, coincidiendo cada categoría o cualidad con su valor. Estos tipos de variables requieren diferentes tratamientos en el proceso de análisis de datos, los cuales no siempre son tan evidentes como pudieran parecer. En adición a esto, muchas de estos conjuntos pueden contener variables que pueden o no estar relacionados linealmente, lo cual también tendrá que ser reflejado en el resultado del análisis. Por tanto, el análisis de datos categóricos no siempre se realizará tan fácilmente como el investigador desearía.

El método de Componentes Principales ha sido una herramienta estadística ampliamente utilizada en diversas áreas del conocimiento, sobre todo en aquellas donde se tienen un volumen considerable de datos y por tanto aumenta la necesidad de conocer la estructura de los mismos y sus interrelaciones. En muchos casos los supuestos del método no se satisfacen especialmente los relacionados con el nivel de medición de las variables y la relación lineal entre ellas. El Análisis de Regresión Lineal, por su parte ha sido una de las herramientas estadísticas más utilizada para predecir una variable respuesta o dependiente a partir de una combinación lineal de variables predictoras o independientes. El modelo de regresión se realiza bajo la suposición que la variable respuesta esté linealmente relacionada con el conjunto de variables predictoras. En investigaciones donde intervienen variables categóricas no pueden aplicarse dichos métodos precisamente por violar los supuestos de los mismos.

Alternativamente se han desarrollados varios métodos para el análisis de datos categóricos. En recientes versiones del paquete estadístico SPSS

aparecen los denominadas métodos con escalamiento óptimo como el Análisis de Componentes Principales y el Análisis de Regresión.

2 Análisis estadístico de datos categóricos

Numerosas son las pruebas estadísticas que se utilizan en la actualidad para procesar datos categóricos [1]. En la medida en la que la sociedad progresa, van apareciendo y desarrollándose otras técnicas nuevas. Es por ello que surge la necesidad de establecer semejanzas y diferencias entres las técnicas existentes para determinar su superioridad o para establecer sus limitaciones y poder saber cuál método es correcto aplicar ante una nueva situación.

2.1 Tablas de contingencia

Cuando se trabaja con variables categóricas, los datos suelen organizarse en tablas de doble entrada en las que cada una representa un criterio de clasificación (una variable categórica). Como resultado, las frecuencias aparecen organizadas en casillas que contienen información sobre la relación existente entre ambos criterios. A estas tablas de frecuencias se les denomina Tablas de Contingencia [2].

Las Tablas de Contingencia tienen dos objetivos fundamentales: organizar la información contenida en un experimento cuando esta es de carácter bidimensional, o sea cuando está referida a dos variables categóricas y analizar si existe alguna relación de dependencia e independencia entre los niveles de las variables objeto de estudio [3]. La significación puede calcularse de manera asintótica usando el test chi cuadrado de Pearson, de manera exacta o a través del método de simulación de Monte Carlo.

2.2 Árboles de decisión: CHAID

En un estudio real existen con frecuencia múltiples variables (predictivas o independientes) que pueden tener asociación con una variable dependiente. La presentación de muchas tablas de contingencia, no siempre refleja las asociaciones esenciales, y usualmente se convierte en un listado enorme de tablas que desinforman en lugar de orientar. Un estudio multivariado trata de enfocar el efecto posible de todas las variables conjuntamente incluyendo sus posibles correlaciones; pero puede ser interesante si se considera además las posibilidades de la interacción entre las variables predictivas sobre la variable dependiente. Cuando el número de

variables crece, el conjunto de las posibles interacciones crece en demasía, resulta entonces prácticamente imposible analizarlas y por ello adquiere especial interés una técnica de detección automática de interacciones fundamentales que construya un árbol de decisión. CHAID es eso: sus siglas significan Chi-squared Automatic Interaction Detector [4].

3 Análisis de componentes principales y análisis de regresión para datos categóricos

3.1 Componentes principales

El análisis de componentes principales (ACP) se ha utilizado de manera creciente en las últimas décadas, prácticamente en todas las áreas [5]. El análisis de componentes principales realiza dos acciones fundamentales: cuantifica las variables originales y reduce la dimensionalidad de los datos. Si el análisis realizado es exitoso, cada variable debe estar muy bien representada (con una correlación elevada) en una dimensión y pobremente representada (con correlaciones bajas) en las demás [6].

En muchos casos, el análisis de componentes principales constituye el objeto de estudio, pero los supuestos del método no se cumplen para los datos observados. Si el ACP se desarrolla sin chequear los supuestos, nunca se podrá estar totalmente seguro de que los resultados serán dignos de confianza. En esta situación, el ACP no lineal o categórico con cuantificaciones óptimas es una alternativa útil [7].

3.2 Componentes principales para datos categóricos

El método de componentes principales categóricos (ACPCat), al igual que su homólogo para variables continuas, puede considerarse como una técnica exploratoria de reducción de las dimensiones de una base de datos incorporando variables nominales y ordinales de la misma manera que las numéricas. El método pone al descubierto relaciones existentes entre las variables originales, entre los casos y entre ambos: variables y casos [8]. Puede además analizar variables con su nivel de medición. Cuando existe relación no lineal entre las variables, pueden especificarse también otros niveles de análisis, de manera que estas relaciones pueden manipularse de manera más efectiva.

En este apartado se describe matemáticamente el análisis de componentes principales categórico. Se supone que se tiene una matriz de datos $H_{n\times m}$, la cual consiste en las puntuaciones observadas de n casos en m

variables. Cada variable puede ser denotada como la $j-ésima$ columna de H; h_j como un vector $n \times 1$, con $j = 1, \ldots, m$. Si las variables h_j no tienen nivel de medición numérico, o se espera que la relación entre ellas no sea lineal, se aplica una transformación no lineal. Durante el proceso de transformación, cada categoría obtiene un valor escalado óptimo, denominado cuantificación categórica. ACPCat puede ser desarrollado minimizando la función de perdida mínima cuadrática en la que la matriz de datos observados H es reemplazada por una matriz $Q_{n\times m}$, que contiene las variables transformadas $q_j = \phi_j(h_j)$. En la matriz Q, las puntuaciones observadas de los casos se reemplazan por las cuantificaciones categóricas. El modelo ACPCat es igual al modelo del ACP, capturando las posibles no linealidades de las relaciones entre las variables en las transformaciones de las variables. Se comenzará explicando como el objetivo del ACP se alcanza por el ACPCat minimizando la función de pérdida, y por tanto se mostrará cómo esta función se amplía para acomodar las ponderaciones de acuerdo con los valores ausentes, ponderaciones por casos, y trasformaciones nominales múltiples.

A las puntuaciones de los casos en las componentes principales obtenidas a partir del ACP se le denominan puntuaciones de las componentes (puntuaciones de los objetos en ACPCat). ACP intenta mantener la información en las variables tanto como sea posible en las puntuaciones de las componentes. A las puntuaciones de las componentes, multiplicadas por un conjunto de ponderaciones óptimas, se les denominan saturaciones en componentes, y tienen que aproximar los datos originales tan cerca como sea posible. Usualmente en ACP, las puntuaciones de las componentes y las saturaciones en componentes se obtienen de una descomposición en valor singular de la matriz de datos estandarizada, o de una descomposición en valores propios de la matriz de correlación. Sin embargo, el mismo resultado puede obtenerse a través de un proceso iterativo en el que se minimiza la función de pérdida mínima cuadrática. La pérdida que se minimiza es la pérdida de la información debido a la representación de las variables por un número pequeño de componentes: en otras palabras, la diferencia entre las variables y las puntuaciones de las componentes ponderadas a través de las saturaciones en componentes. Si $X_{n\times p}$ se considera la matriz de las puntuaciones de las componentes, siendo p el número de las componentes, y si $A_{m\times p}$ es la matriz de las saturaciones en componentes, siendo su $j-ésim$a fila indicada por a_j, la función de perdida que se usa en el ACP para la minimización de la diferencia entre los datos originales y las componentes principales puede ser expresada como

$L(Q, X, A) = .n^{-1} \sum_j \sum_n (q_{ij} - \sum_s x_{is} a_{js})^2$ En notación matricial, esta función puede escribirse como:

$$L(Q, X, A) = n^{-1} \sum_{j=1}^{m} tr\,(q_j - X a_j)'\,(q_j - X a_j) \tag{1}$$

donde tr denota la función traza que suma los elementos de la diagonal de una matriz. Puede probarse que la función (1) es equivalente a:

$$L_2(Q, A, X) = n^{-1} \sum_{j=1}^{m} tr\left(q_j a_j' - X\right)'\left(q_j a_j' - X\right). \tag{2}$$

La función de pérdida (2) se usa en ACPCat en lugar de (1), debido a que en (2), la representación vectorial de las variables así como la representación de las categorías como un conjunto de puntos agrupados puede ser incorporada, comos será mostrada dentro de poco.

La función de pérdida (2) está sujeta a un número de restricciones. Primero, las variables transformadas son estandarizadas, a fin de que $q_j' q_j = n$. Tal restricción se necesita para resolver la indeterminación entre q_j y a_j en el producto escalar $q_j a_j'$. Esta normalización implica q_j que contenga z-scores y garantice que las saturaciones en componentes en a_j estén correlacionadas entre las variables y las componentes. Para evitar la solución trivial $A = 0$ y $X = 0$, las puntuaciones de los objetos se limitan y se requiere que:

$$X'X = nI \tag{3}$$

donde I es la matriz identidad. Se necesita también que las puntuaciones de los objetos estén centradas, por lo tanto:

$$1'X = 0 \tag{4}$$

donde el 1 representa el vector unidad. Las restricciones (3) y (4) implican que las columnas de X (componentes) son z-scores ortonormales: su media es cero, su desviación estándar es uno, y están incorrelacionadas. Para el nivel de escala numérica, $q_j = \phi_j(h_j)$ implica una transformación lineal, o sea, la variable observada h_j es simplemente transformada en z-scores. Para los niveles no lineales (nominal, ordinal, spline), $q_j = \phi_j(h_j)$ denotan una transformación acorde con el nivel de medición seleccionado para la variable j.

La función de pérdida (2) se minimiza aplicando los mínimos cuadrados alternantes, actualizando cíclicamente uno de los parámetros X, Q y A,

mientras que los otros dos se mantienen constantes. Este proceso iterativo se continúa hasta que la mejora en los valores perdidos posteriores esté por debajo de algún valor pequeño especificado por el usuario. En ACPCat, los valores de partida de X son aleatorios.

Las ponderaciones por valores perdidos y las ponderaciones por casos pueden incorporarse fácilmente a la función de pérdida. Para acomodar el tratamiento pasivo de los valores, se introduce una matriz diagonal $M_{j_{n\times n}}$, con la $i-ésima$ diagonal principal de entrada ii, correspondiente al caso i, igual a 1 para los valores no ausentes y 0 para los valores ausentes de la variable j. Por tanto, para los casos con valores perdidos en la variable j, los elementos de la diagonal correspondiente en M_j son ceros, así que la matriz error premultiplicada por M_j, $M_j\left(q_j a_j' - X\right)$, contiene ceros en las filas correspondientes a los casos con valores ausentes en la variable j. Por tanto, para la variable j, los casos con valores perdidos no contribuyen a la solución de ACPCat, sino que contribuyen a la solución de las variables que tienen una puntuación válida. Por otra parte, se permite la ponderación de los casos a través de la ponderación del error por una matriz diagonal $W_{n\times n}$ con elementos no negativos w_{ii}. Generalmente estas ponderaciones son todas igual a uno, pues cada caso contribuye de igual manera a la solución. Para algunos, sin embargo, puede ser conveniente tener diferentes ponderaciones para diferentes casos.

Incorporando las ponderaciones de los datos ausentes M_j y las ponderaciones de los casos W, la función de pérdida que se minimiza en ACPCat puede expresarse como:

$$L_3\left(Q,A,X\right) = n^{-1}\sum_{j=1}^{m}\sum_{i=1}^{n} w_{ii}m_{iij}\sum_{s=1}^{p}\left(q_{ij}a_{js}-x_{is}\right)^2,$$

o equivalentemente, en notación matricial como:

$$L\left(Q,A,X\right) = n_w^{-1}\sum_{j=1}^{m} tr\left(q_j a_j' - X\right)' M_j W\left(q_j a_j' - X\right). \qquad (5)$$

Entonces, la restricción centrada se torna en $1'M_*WX = 0$, donde $M_* = \sum_{j=1}^{m} M_j$, y la restricción de estandarización en $X'M_*WX = mn_wI$.

La función de pérdida (5) puede ser usada para las transformaciones nominales, ordinales y spline, donde los puntos de las categorías se restringen para estar en una línea recta (vector). Si las categorías de una variable están representadas como un grupo de puntos (utilizando el nivel de escala nominal múltiple), con el grupo de puntos en el centro de los puntos de los casos medidos en una categoría particular, las categorías no estarán en una

línea recta, sino que cada categoría obtendrá cuantificaciones múltiples, una de las cuales es la componente principal. En contraste, si la representación del vector se usa en lugar de la representación de los puntos de las categorías, cada categoría obtiene una sola cuantificación categórica, y la variable obtiene diferentes saturaciones en componentes por cada componente. Para incorporar las cuantificaciones múltiples en la función de pérdida, se expresa $L_3(Q, A, X)$ de manera conveniente para introducir las variables nominales múltiples. Considerando para cada variable una matriz indicadora G_j. El número de filas de G_j es igual al número de casos, n, y el número de columnas de G_j es igual al número de las diferentes categorías de la variable j. Por cada caso, una columna de G_j contiene un 1 si el caso tiene una categoría particular, y un cero si no la tiene. Así, todas las filas de G_j contiene exactamente un 1, excepto cuando los valores ausentes son tratados pasivamente. Si se estuviera en presencia de valores ausentes pasivos, cada fila de la matriz indicadora correspondiente a la observación con valores ausentes contiene solamente ceros. En la función de pérdida, las variables cuantificadas q_j pueden ahora ser escritas como $G_j v_j$, con v_j representando las cuantificaciones de las categorías de la variable j. Entonces, la función de pérdida se torna en:

$$L_3(v_1, \ldots, v_m, A, X) = n^{-1} \sum_{j=1}^{m} tr\left(G_j v_j a_j' - X\right)' M_j W \left(G_j v_j a_j' - X\right). \tag{6}$$

La matriz $v_j a_j'$ contiene coordenadas p-dimensionales que representan las categorías en una línea recta a través del origen, en la dirección dada por las saturaciones en componentes a_j. Como $q_j = G_j v_j$ para todas las variables que no son nominales múltiples, (6) es la misma que (5).

La ventaja de (6) es que la transformación nominal múltiple puede incorporarse directamente. Si se especifica el nivel de escala nominal múltiple, con las categorías representadas como puntos de grupos, $v_j a_j'$ se reemplaza por V_j, conteniendo los puntos de grupos, los centroides de los objetos de puntos para los casos en p dimensiones. Entonces, la función de pérdida puede escribirse como:

$$L_4(V_1, \ldots, V_m, X) = n^{-1} \sum_{j=1}^{m} tr\left(G_j V_j - X\right)' M_j W \left(G_j V_j - X\right) \tag{7}$$

donde V_j contiene las coordenadas de los centroides para las variables dadas con nivel de medición nominal múltiple, y $V_j = v_j a_j'$ contiene las coordenadas de los puntos categóricos localizados en un vector para otros niveles de medición [7].

3.3 Análisis de regresión lineal

El análisis de regresión lineal estándar es una técnica estadística ampliamente utilizada desde la segunda mitad del siglo XIX, cuando el científico británico Francis Galton introdujo dicho término [9]. El análisis de regresión lineal clásico minimiza las diferencias de la suma de los cuadrados entre una variable de respuesta (dependiente) y una combinación ponderada de las variables predictoras (independientes). Las variables son normalmente cuantitativas, con los datos categóricos (nominales) recodificados como variables binarias. Los coeficientes estimados reflejan cómo los cambios en las variables predictoras afectan a la respuesta. Puede obtenerse un pronóstico de la respuesta para cualquier combinación de los valores predictores [10].

3.4 Análisis de regresión para datos categóricos

El análisis de regresión categórica es un método a través del cual la regresión se aplica a los datos de la respuesta en forma de categorías con el propósito de predecir la probabilidad de ocurrencia de una categoría particular de la respuesta como función de una o más variables independientes [11]. La regresión categórica (RegCat) se ha desarrollado como un método de regresión lineal para variables categóricas. La regresión categórica cuantifica los datos categóricos mediante la asignación de valores numéricos a las categorías, obteniéndose una ecuación de regresión lineal óptima para las variables transformadas.

3.4.1 Cuantificaciones categóricas

En el proceso de cuantificación ciertas propiedades de los datos se preservan en la transformación. Las propiedades que se seleccionan para ser preservadas se especifican seleccionando un nivel de escalamiento óptimo para las variables. Es importante para realizarlo, que el nivel de escalamiento óptimo es el nivel en el que una variable se analiza, el que no necesariamente coincide con el nivel de medición de la variable.

El nivel de escalamiento, y por tanto la forma de la curva de transformación, está también relacionado con el número de grados de libertad de la transformación y por tanto al ajuste del modelo. Las transformaciones con más libertad resultan transformaciones menos suaves y ajustan mejor, mientras que transformaciones más restrictivas son más suaves pero los resultados ajustan menos. De manera que, existe un equilibrio entre las propiedades de preservación de los datos y la preservación de

la información relacional en los datos: restringiendo las transformaciones, preservando más propiedades de los datos, se alcanza un costo de ajuste y se pierde información relacional. La transformación con el máximo de libertad es el resultado a partir del nivel de escalamiento nominal, donde el número de grado de libertad es igual al número de categorías menos uno. El nivel de escalamiento ordinal requiere una restricción de orden sobre las cuantificaciones categóricas, resultando el número de grado de libertad igual al número de categorías con diferentes valores cuantificados menos uno. El escalamiento numérico impone una restricción de intervalo adicional a la restricción de orden y tiene un grado de libertad.

El nivel de escalamiento nominal y el ordinal dan lugar a transformaciones que son funciones paso, la cuales son adecuadas para variables con un número pequeños de categorías. Para variables con un número más grande de categorías, las funciones spline son más apropiadas, entre estas distinguimos splines no monótonos para transformaciones no ordenadas y splines monótonos para transformaciones ordenadas. Las funciones spline son funciones polinomiales por trozos, ellas son más restrictivas que las funciones paso, dando lugar a curvas de transformación más suaves, pero con un ajuste menor. Para obtener una transformación spline, el rango de la variable se divide en un número de intervalos, igual al número de nodos especificado menos uno. Los nodos son los puntos extremos de los intervalos. Entonces las funciones polinomiales de un grado específico se ajustan en cada intervalo y se empatan en cada nodo. La suavidad y el numero de grados de libertad de una curva de transformación spline depende del número de nodos y del grado de las funciones polinomiales [12].

En términos de restricciones, o sea, de suavidad de la curva de transformación y ajuste, la transformación spline no monótona está entre una nominal y una transformación lineal. Con número de nodos interiores igual al número de categorías menos dos y usando un polinomio de primer grado, la transformación spline es la misma que la transformación nominal. Con el número de nodos interiores igual a cero y con un polinomio de primer grado, la transformación spline es la misma que la transformación lineal. De la misma manera, una transformación spline monótona está entre una ordinal y una transformación lineal.

Lo expresado en el párrafo anterior se ilustra en la figura 1 que se muestran a continuación, la que muestra la gráfica de transformación de la variable dependiente Diagnóstico de Expertos (DiagExp), que tiene tres categorías: (1-normotenso, 2-pre hipertenso, 3-hipertenso) y cierta variable independiente categórica (X1). A la variable dependiente se le

fijó el nivel de medición ordinal mientras que a la independiente se le variaron los niveles de medición.

Con el nivel de medición nominal aplicada a la variable independiente se obtiene una curva bastante dentada (figura 1.1). En el mismo se puede apreciar que ambas variables que a medida que se incrementan alcanzan valores máximos. El R^2que se obtiene es igual a 0.128. Al aplicar una transformación spline no monótona (2do grado con 10 nodos interiores) las irregularidades son más suaves (figura 1.2), mucho más si se tienen dos nodos interiores (figura 1.3). Los R^2 para estos casos son 0.088 y 0.081 respectivamente. Obsérvese que el R^2 disminuye en la medida en que el nivel de escalado utilizado conserva más propiedades.

Como las transformaciones ordinales se obtienen mediante el average de las cuantificaciones nominales que están en el orden equivocado, la aplicación de niveles de escala ordinales da lugar a transformaciones que restringen todos los valores cuantificados en forma de mesetas (figura 1.4). El R^2 que se obtiene en esta transformación es 0.094. Cuando se aplica una transformación monótona (2 grados con 10 nodos interiores) muchas de las mesetas desaparecen (figura 1.5) y con 2 grados y 2 nodos interiores la transformación es casi lineal (figura 1.6).

Los valores de los R^2 en estos casos son 0.085 y 0.078 [12]. En la figura 1.7 se muestra la transformación con nivel de escalado numérico. El R^2 que se obtiene es 0.073. En todas estas gráficas de observa que a medida que se gana en suavidad se pierde en ajuste.

La regresión categórica múltiple es una técnica no lineal, donde la no linealidad radica en las transformaciones de las variables. El modelo de la regresión categórica es el modelo de la regresión lineal clásica, aplicado a las variables transformadas:

$$\varphi_r(y) = \sum_{j=1}^{J} \beta_j \varphi_j(x_j) + e \tag{8}$$

con la función de pérdida:

$$L\left(\varphi_r, \varphi_1, \ldots, \varphi_j; \beta_1, \ldots, \beta_j\right) = \left\| \varphi_r(y) - \sum_{j=1}^{J} \beta_j \varphi_j(x_j) \right\|^2 \tag{9}$$

donde J es el número de variables predictoras, y representa la variable respuesta observada o discretizada, x_j representa las variables predictoras observadas o discretizadas, β_j los coeficientes de regresión, φ_r las transformaciones de la variable respuesta, φ_j las transformaciones de las variables predictoras y el vector error.

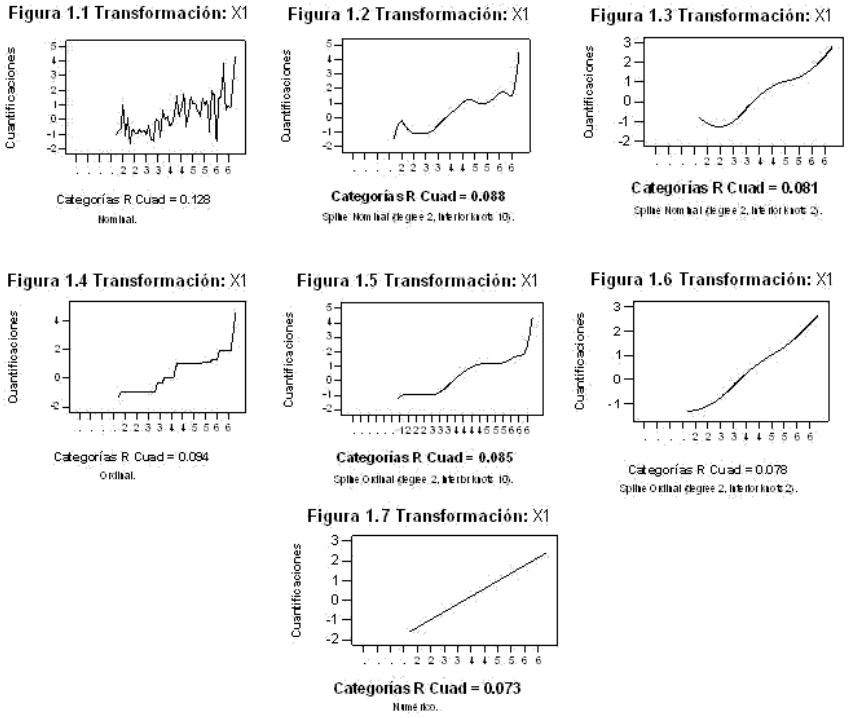

Figura 1: Transformación de la variable X1.

Todas las variables son centradas y normalizadas para obtener la suma de los cuadrados igual a N, y $\|\cdot\|^2$ representa el cuadrado de la norma euclideana.

La forma de las transformaciones depende del nivel de escalamiento óptimo, el cual puede seleccionarse para cada variable por separado y es independiente del nivel de medición. El nivel de escalamiento define que parte de la información que está en la variable observada o discretizada (según sea el nivel de medición) se retiene en la transformación de la variable. Con nivel de escalamiento numérico, los valores de la categoría de una variable se tratan como cuantitativos. Entonces toda la información se retiene y la única transformación aplicada es la estandarización, resultando una transformación lineal. Luego, cuando para todas las variables se aplica el nivel de escalamiento numérico, el resultado de la RegCat es igual al resultado de la regresión lineal múltiple con las variables estandarizadas.

Con niveles de escalamiento no numérico, los valores de las categorías se tratan como cualitativos, y se transforman en valores cuantitativos.

En este caso, alguna parte de la información en la variable observada o discretizada se pierde.

Con nivel ordinal o spline monótono, la información de intervalo se pierde y solamente la información de grupo y orden se retienen, así se posibilita una transformación monótona.

Con nivel nominal y spline no monótono solo la información de agrupación tiene que conservarse, dando lugar a una transformación no monótona.

Aplicando niveles de escalamiento no lineales, las relaciones no lineales entre la variable respuesta y las variables predictoras se linealizan, por lo tanto el modelo de regresión lineal del término es todavía aplicable.

En RegCat las variables observadas o discretizadas se codifican en una matriz indicadora $G_m$de tamaño $N \times C_m$, donde N es el número de observaciones y C_m representa el número de categorías de la variable m, $m = 1, \ldots, M$, donde M es el número total de variables.

Una entrada $g_{ic(m)}$ de G_m , donde $c = 1, \ldots, C_m$, es 1 si la observación i está en la categoría c de la variable m y 0 en otro caso. Entonces las variables transformadas pueden escribirse como el producto de la matriz indicador G_m y el $C_m - vector$ de las cuantificaciones categóricas v_m:

$$\varphi_r(y) = G_r v_r \wedge \varphi_j(x_j) = G_j v_j \tag{10}$$

donde v_r es el vector de las categorías cuantificaciones de la variable respuesta, y v_j el vector de categoríascuantificaciones para una variable pre-

dictora. Luego, el modelo de RegCat con las variables transformadas escrito en términos de matrices indicadoras y categorías cuantificadas es:

$$G_r v_r = \sum_{j=1}^{J} \beta_j G_j v_j + e. \tag{11}$$

Con la función de pérdida mínimos cuadrados asociada:

$$L\left(v_r; v_1, \ldots, v_j; \beta_1, \ldots, \beta_j\right) = \left\| G_r v_r - \sum_{j=1}^{J} \beta_j G_j v_j \right\|^2. \tag{12}$$

La función de pérdida (12) se minimiza por el algoritmo de mínimos cuadrados alternantes, que alterna entre la cuantificación de la variable respuesta por un lado, y la cuantificación de las variables predictoras y estimación de los coeficientes de regresión por el otro.

Primero se inicializan las cuantificaciones y los coeficientes de regresión. RegCat tiene dos formas de inicialización: aleatoria y numérica. Una inicialización aleatoria usa valores aleatorios estandarizados para las cuantificaciones iniciales, y los coeficientes de regresión iniciales son las correlaciones de orden cero de la variable respuesta cuantificada aleatoriamente con las variables predictoras cuantificadas de manera aleatoria. Los valores iniciales con una inicialización numérica se obtienen a partir de un análisis con nivel de escalamiento numérico para todas las variables.

En el primer paso del algoritmo, las cuantificaciones de las variables predictoras y los coeficientes de regresión se mantienen fijos. Con nivel de escalamiento numérico las cuantificaciones $v_r$de la variable respuesta son los valores de las categorías de la variable observada o discretizada centrada y normalizada. Con nivel de escalamiento no numérico las cuantificaciones son actualizadas en la siguiente forma:

$$\tilde{v}_r = D_r^{-1} G_r' \sum_{j=1}^{J} \beta_j G_j v_j \tag{13}$$

donde $D_r = G_r' G_r$. Las cuantificaciones $\tilde{v}_r$ son las cuantificaciones no estandarizadas para el nivel de escalamiento nominal. Para los niveles ordinal, no monótono o spline monótono, se aplica una restricción para $\tilde{v}_r$, en relación con el nivel de escalamiento, produciendo v_r^*. Por tanto, $v_r^* = \tilde{v}_r$ para el nivel de escalamiento nominal, y $v_r^* = \tilde{v}_r$ (restringida) para los niveles ordinales y spline. Entonces v_r^* se estandariza:

$$v_r^+ = N^{1/2} v_r^* (v_r^{*\prime} D_r v_r^*)^{-1/2}. \tag{14}$$

En el segundo paso del algoritmo, las cuantificaciones de la variable respuesta mantienen fijas, y las cuantificaciones v_j de las variables predictoras con nivel de escalamiento no numérico, y los coeficientes de regresión se actualizan para cada variable al mismo tiempo. El enfoque trabaja como sigue. Primero se calcula el $N - vector$ de los valores predictores:

$$z = \sum_{j=1}^{J} \beta_j G_j v_j. \tag{15}$$

Para actualizar las cuantificaciones de la variable j, la contribución de la variable j a la predicción (la combinación lineal ponderada de los predictores transformados) se sustrae de z:

$$z_j = z - \sum_{j=1}^{J} \beta_j G_j v_j. \tag{16}$$

Las cuantificaciones no restringidas se actualizan de la manera siguiente:

$$\tilde{v}_j = sign(\beta_j) D_j^{-1} G_j' (G_j v_j^+ - z_j). \tag{17}$$

Para variables con nivel de escalamiento no numérico $\tilde{v}_j$ se restringe según sea el nivel de escalamiento, y normalizada como en (14), produciendo v_j^+. Para variables con nivel de escalamiento numérico, v_j^+ contiene los valores de las categorías de los datos observados o discretizados centrados y estandarizados. Luego los coeficientes de regresión β_j se actualizan:

$$\beta_j^+ = N^{-1} \tilde{v}_j' D_j v_j^+. \tag{18}$$

Entonces, la contribución actualizada de la variable j para la predicción se adiciona a z_j:

$$z = z_j + \beta_j^+ G_j v_j^+. \tag{19}$$

y el algoritmo continua con la actualización de la cuantificación para la próxima variable predictora, hasta que todos los predictores sean actualizados.

Los valores perdidos se calculan como $\left\| G_j v_j^+ - z \right\|^2$. Estos dos pasos se repiten hasta que se alcance el criterio de convergencia especificado por el usuario.

Para el nivel de escalamiento ordinal, se usa la regresión monótona ponderada de las cuantificaciones nominales en la variable observada o discretizada. Para la restricción en relación con los niveles de escalamiento spline se usa la regresión ponderada de las cuantificaciones nominales en

un I-spline base [13], con restricciones no negativas adicionales para el nivel de escalamiento spline monótono. En este punto, pudiera ocurrir una complicación adicional. Una restricción creciente de manera monótona puede a veces dar lugar a una variable transformada con valores constantes. Por ejemplo, cuando los valores de $\tilde{v}$ son decrecientes de manera monótona, excepto para el primer y el ultimo valor, las cuantificaciones restringidas son la media de $\tilde{v}$ para todas las categorías. En este caso, la transformación en una constante puede evitarse dando lugar a una función monótona decreciente [12].

3.4.2 Relación con el Análisis de Discriminante

El método de regulación RegCat puede fácilmente extenderse al Análisis de Discriminante tanto lineal como no lineal regularizado para clasificar los casos en los grupos. La RegCat con escalamiento nominal aplicado a una variable categórica dependiente y con transformaciones lineales a los predictores continuos es equivalente a un Análisis Discriminante lineal (unidimensional; solamente resultará una función discriminante). Al seleccionar una transformación no lineal, se logrará un Análisis Discriminante no lineal. La adaptación de RegCat en el Análisis Discriminante Categórico no es asunto del algoritmo, sino solamente el resultado: coeficientes de regresión tienen que ser convertidos en coeficientes discriminantes, lo cual es sencillo debido a que son proporcionales entre ellos, y el resultado específico hacia el Análisis Discriminante necesitan ser suministrado.

La pertenencia final de cada caso a una de las clases no puede realizarse a nivel de menú en el SPSS, por lo que se necesita auxiliarse de una ventana de sintaxis. A continuación se muestran los conjuntos de pasos necesarios para convertir los valores de la variable dependiente en valores de una clase.

Pasos necesarios para convertir los valores de la variable dependiente en valores de una clase.

```
* x = 1 cuantificación categórica de la variable dependiente
* y = 2 cuantificación categórica de la variable dependiente
* z = 3 cuantificación categórica de la variable dependiente
compute dist1= (pre_1 - x )**2.
compute dist2= (pre_1 - y)**2.
compute dist3= (pre_1 - z)**2.
compute mindist = MIN(dist1, dist2, dist3).
compute class1 = (mindist = dist1).
```

```
compute class2 = (mindist = dist2).
recode class2 (1 = 2).
compute class3 = (mindist = dist3).
recode class3 (1 = 3).
compute class = class1 + class2 + class3. exe.
CROSSTABS
    /TABLES= depvar BY class.
```

4 Estudio de la hipertensión arterial (HTA)

La hipertensión arterial (HTA) es la elevación de la presión arterial por encima de un límite que se considera normal (140/90 mmHg). Es la principal enfermedad crónica degenerativa y la más común causa de muerte, afecta aproximadamente al 20% de la población mundial. La elevación anormal de la presión constituye un importante factor de riesgo coronario y de padecer accidentes vasculares cerebrales [14].

Se cree que tanto los factores ambientales como los genéticos son causas de la hipertensión. La tensión arterial tiende a elevarse con la edad. Es también más frecuente que aparezca si la persona es obesa, tiene una dieta rica en sal y pobre en potasio, bebe elevadas cantidades de alcohol, no tiene actividad física y sufre de un elevado estrés psicológico. Aunque está claro que la tendencia a la hipertensión puede ser heredada, se desconocen en gran medida los factores genéticos responsables de la misma [15]. El conocimiento actual de éste problema de salud pública a nivel mundial, obliga a buscar estrategias certeras de detección, control y tratamiento.

En este trabajo se presenta un estudio realizado con los 849 individuos de cinco policlínicos de la ciudad de Santa Clara. Cada caso fue inicialmente clasificado como normotenso, pre hipertenso o hipertenso por un comité de expertos altamente calificado. La tabla 1 muestra las variables originales que formaron parte de este estudio.

4.1 Árboles de decisión: CHAID

En este epígrafe se aplica la técnica de segmentación CHAID tomando como variable dependiente el diagnóstico de expertos (DiagExp) y como posibles variables predictoras el resto de las variables que aparecen en la tabla 1. La figura 2 muestra un esquema que resume el primer árbol obtenido.

En el nodo raíz del árbol se encuentran los 849 casos estudiados. De ellos, 434 personas son normotensas, lo que representa un 51.1% de la mues-

Variable	Etiqueta	Valores
Edad	Edad del paciente	16-80 años
TASistB	Presión Sistólica Basal	Baja, Media, Alta
TADiastB	Presión Diastólica Basal	Baja, Media, Alta
TASistB1	Presión Sistólica al minuto 1	Baja, Media, Alta
TADiastB1	Presión Diastólica al minuto 1	Baja, Media, Alta
TASistB2	Presión Sistólica al minuto 2	Baja, Media, Alta
TADiastB2	Presión Diastólica al minuto 2	Baja, Media, Alta
TAPam	Presión arterial media	Baja, Media, Alta
Col_Tot	Colesterol total	Bajo, Medio, Alto
Col_Ldl	Colesterol LDL	Bajo, Medio, Alto
Col_Hdl	Colesterol HDL	Bajo, Medio, Alto
OImc	Índice de masa corporal	Bajo, Normal, Elevado.
Sexo	Sexo del paciente	Masculino Femenino
Fuma	Hábito de fumar	Sí, No
Bebe	Ingestión de bebidas alcohólicas	Sí, No
Diabetes	Padecimiento de Diabetes Mellitus	Sí, No
Dislipidemia	Padecimiento de dislipidemia	Sí, No
Raza	Raza del paciente	Blanca, Mestiza
DiagExp	Diagnóstico de HTA	Normotenso, Pre hipertenso, Hipertenso.

Tabla 1: Variables consideradas en el análisis.

tra, 193 son pre hipertensos (22.7%) y 222 casos son hipertensos (26.1%). La variable que mejor ayuda a diferenciar los grupos es la TAPam, esta es la más significativa, acorde con lo reportado por los especialistas [14][15].

El árbol creado tiene 7 hojas o nodos terminales, veamos su explicación:

1. Subconjunto formado por 208 pacientes caracterizan por presentar valores bajos en la TAPam. Todos los pacientes del grupo son normotensos. Se corresponde con el Nodo 1 del árbol.

2. Subconjunto formado por 63 pacientes. Estos se caracterizan por

tener valores de la TAPam entre baja o media y valores bajos de la TADiastB2. Existe predominio de normotensos (93.7%) y el resto está conformado por pre hipertensos (4.8%) e hipertensos (1.6 %). Se corresponde con el Nodo 4 del árbol.

3. Subconjunto formado por 104 pacientes. Se caracterizan por tener valores altos en la TAPam y valores bajos en la TADiastB. Es un grupo donde predominar los pre hipertensos (63.5%) sobre los hipertensos (36.5%). Se corresponde con el Nodo 6 del árbol.

4. Subconjunto formado por 146 pacientes. Se caracterizan por tener valores entre baja y media de la TAPam, valores entre media y alta en la TADiastB2 y valores bajos de TASistB1. Es un grupo donde predominan los normotensos (72.6%). El 26.7% de los pacientes del grupo son pre hipertensos y uno solo de los pacientes es hipertenso. Se corresponde con el Nodo 8 del árbol.

5. Subconjunto formado por 138 pacientes. Se caracterizan por tener valores entre baja y media de la TAPam, valores entre media y alta en la TADiastB2 y valores entre media y alta de la TASistB1. En este grupo predominan los pre hipertensos (51.4%). Los normotensos representan un 44.2% del total del grupo mientras que los hipertensos solo representan 4.3%. Se corresponde con el Nodo 9 del árbol.

6. Subconjunto formado por 66 pacientes. Es característica de este grupo presentar valores altos en la TAPam, valores altos en la TADiastB y valores entre baja y media en la TADiastB2. En este grupo 52 pacientes son hipertensos (78.8%) y 14 son pre hipertensos (21.2%). Es válido destacar la ausencia de pacientes normotensos en el grupo. Se corresponde con el Nodo 10 del árbol.

7. Subconjunto formado por 124 pacientes que se caracterizan por tener valores altos en la TAPam, valores altos en la TADiastB y también valores altos en la TADiastB2. Es un grupo donde los 124 pacientes que lo conforman son hipertensos (100%). Se corresponde con el Nodo 11 del árbol.

El árbol de decisión obtenido, además de segmentar la población, crea reglas de clasificación. La tabla 2 muestra los resultados obtenidos:

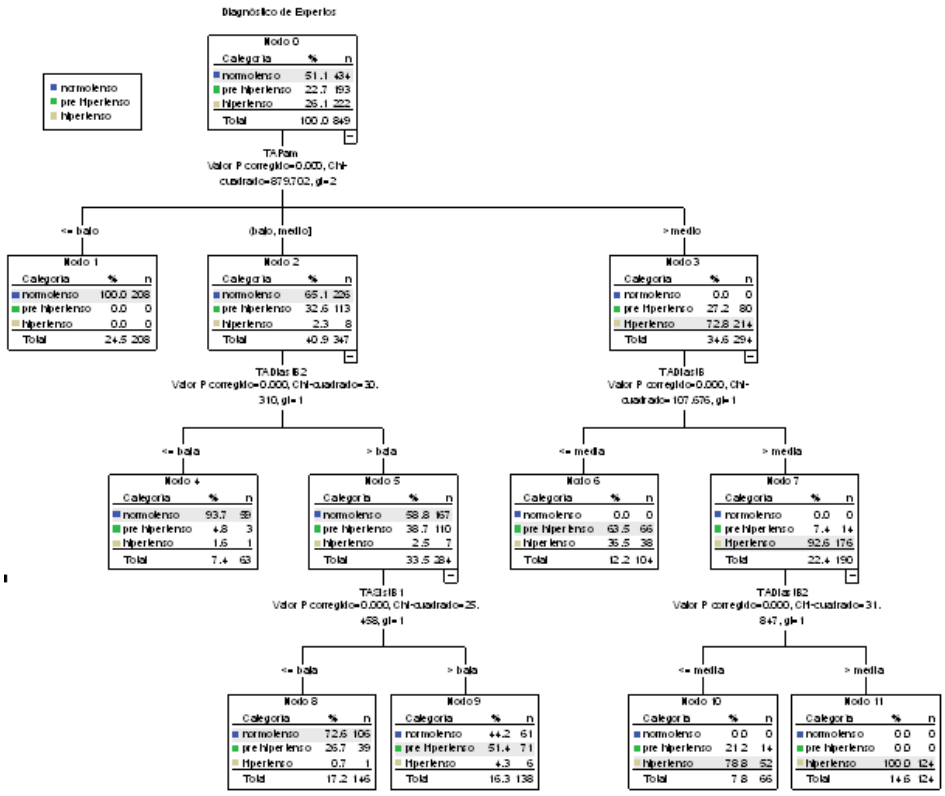

Figura 2: Árbol de decisión aplicando la técnica CHAID.

Se clasifican adecuadamente un 80.8% de la totalidad de los casos. Debe señalarse que los resultados más interesantes se encuentran en el hecho de que el árbol casi no se equivoca entre pacientes normotensos e hipertensos. Ningún normotenso fue clasificado como hipertenso y sólo dos hipertensos fueron clasificados como normotenso. Las dudas aparecen en el grupo de los pre hipertensos. Esto se corresponde plenamente con el criterio de los expertos, pues este grupo se considera dudoso. A él pertenecen aquellas personas que no son hipertensas, pero que tienen una probabilidad elevada de serlo en un futuro no muy lejano.

Observado	Pronosticado			
	normotenso	pre hipertenso	hipertenso	% correcto
normotenso	373	61	0	85.9%
pre hipertenso	42	137	14	71.0%
hipertenso	2	44	176	79.3%
% global	49.1%	28.5%	22.4%	80.8%

Método de crecimiento: CHAID
Variable dependiente: Diagnóstico de Expertos

Tabla 2: Clasificación.

4.2 Regresión categórica con componentes principales como método de selección de variables

En numerosas investigaciones, sobre todo en el campo médico o social [16], se tienen variables predictoras categóricas. Algunas tienen un orden entre sus valores, otras son simplemente nominales. En estos casos pudiera pensarse en realizar una regresión de la respuesta con respecto a los propios valores predictores categóricos. Como consecuencia, se estima un coeficiente para cada variable. Sin embargo, para las variables discretas, los valores categóricos son arbitrarios. La codificación de las categorías de diferentes maneras proporciona diferentes coeficientes, dificultando las comparaciones entre los análisis de las mismas variables. De manera general, la aplicación de las técnicas clásicas de regresión se dificulta notablemente. Para subsanar estas deficiencias surge la regresión categórica.

En este epígrafe se pretende encontrar un modelo de regresión que permita caracterizar el padecimiento de la HTA en pacientes de cinco policlínicos del municipio de Santa Clara. El problema que se presenta en este trabajo no puede tratarse adecuadamente por una regresión lineal múltiple, pues la variable dependiente (DiagExp) es ordinal y todas las predictoras son categóricas (ver tabla 1). Se decide entonces aplicar la regresión categórica presente en el SPSS en su versión 13 [8]. En la primera corrida se consideraron todas las variables mostradas en la tabla 1. A la variable presión arterial media (TAPam) se le aplicó el nivel de escalamiento nominal con el objetivo que tuviera mayor grado de libertad y por tanto lograr así un mejor ajuste en el modelo, ya que de todas las variables predictoras ésta es la más importante o significativa (ver figura

2) y por tanto la que mayor influencia ejerce sobre la variable dependiente (DiagExp) [8][12]. El valor del coeficiente de determinación R^2 obtenido fue igual a 0.828, lo cual indica que el 82.8% de la variable diagnóstico está explicado en el modelo.

R Múltiple	R Cuadrado	R Cuadrado Ajustado
0.910	0.828	0.824

Variable dependiente: Diagnóstico de Expertos
Predictores: Edad Sexo Raza Bebe Fuma Diabetes mellitus Dislipidemia TASistB TADiastB TASistB1 TADiastB1 TASistB2 TADiastB2 TAPam OIMC Col_Tot Col_HDL Col_LDL

Tabla 3: Resumen del modelo.

El resultado del análisis de varianza resultó significativo lo que indica que el modelo es válido [17]. Ahora bien el modelo que se obtiene es muy grande, o sea, está compuesto por numerosas variables predictoras (ver tabla 4) y algunas de ellas son no significativas. El método de regresión categórica no tiene implementado aún ningún método de selección de variables y por consiguiente todas las variables independientes consideradas pasaron a formar parte de la ecuación.

Para analizar los supuestos de la regresión se utilizó el test de Kolmogorov Smirnov para comprobar si los residuos estaban normalmente distribuidos. La significación fue 0.161 indicando la normalidad [17]. Para verificar la homogeneidad de la varianza y comprobar la ausencia de multicolinealidad se realizó una regresión lineal tomado como datos los valores de las variables transformadas [12] ya que la regresión categórica no realiza este tipo de análisis [8].

El estadístico de Durbin Watson obtenido fue de 1.534 indicando que no hay autocorrelación y por tanto existe homogeneidad de varianza [18]. El índice de condición reafirma la ausencia de multicolinealidad [17].

En el modelo obtenido aparecen varias variables no significativas (ver tabla 4), además que son muchas por lo que el modelo pudiera no ser sencillo y por tanto de difícil interpretación. Para realizar la selección de las variables se decidió utilizar el método de componentes principales para variables categóricas precisamente por la naturaleza de las variables que intervienen en el estudio.

El método de componentes principales ha sido utilizado de manera creciente en las últimas décadas, prácticamente en todas las áreas, es el

	Coeficientes Estandarizados	
	Beta	Significación
Edad	0.020	0.247
Sexo	-0.065	0.000
Raza	0.025	0.090
Bebe	-0.012	0.446
Fuma	-0.001	0.947
Diabetes mellitus	-0.018	0.244
Dislipidemia	-0.006	0.699
TASistB	0.005	0.845
TADiastB	0.151	0.000
TASistB1	0.164	0.000
TADiastB1	0.088	0.000
TASistB2	0.088	0.001
TADiastB2	0.215	0.000
TAPam	0.353	0.000
OIMC	0.043	0.005
Col_Tot	-0.015	0.471
Col_HDL	-0.011	0.466
Col_LDL	-0.004	0.859

Variable Dependiente: Diagnóstico de Expertos

Tabla 4: Coeficientes.

análisis de componentes principales. En la medida en que aumenta el número de las variables a considerar en una investigación dada, aumenta la necesidad de conocer en profundidad su estructura y sus interrelaciones [5]. Las investigaciones sobre la HTA no constituyen una excepción.

El nivel de escalamiento aplicado a las variables fue el mismo que el que se utilizó en el análisis de regresión categórica. El modelo que se obtiene considerando la totalidad de las variables resulta ser poco satisfactorio ya que el por ciento total de la varianza explicada por los factores es pequeño. Ello puede deberse a que a la mayoría de las variables consideradas se le asignó un escalado numérico, que es de todos, el más restrictivo.

La tabla 5 muestra el resumen del modelo obtenido. Como puede apreciarse el porcentaje total de la varianza explicada por los factores es pequeño (49.706%), pero en nuestro caso este hecho no es tan importante,

debido a que no se van a sustituir las variables originales por los factores hallados.

		Varianza explicada	
Dimensión	Alfa de Cronbach	Total (Autovalores)	% de la varianza
1	0.872	5.670	31.500
2	0.473	1.807	10.041
3	0.338	1.470	8.165
Total	0.940	8.947	49.706

a. El Alfa de Cronbach Total está basado en los autovalores totales

Tabla 5: Resumen del modelo.

La tabla también muestra el valor del estadístico alfa de Cronbach (0.940), que es una medida de confiabilidad que se maximiza en el procedimiento.

La tabla 6 muestra las variables que intervienen en cada una de las dimensiones. Obsérvese que las variables que miden presiones tienen un valor elevado (superior a 0.800) en la primera dimensión y valores pequeños en las demás. El efecto contrario ocurre con dos de las variables que miden colesteroles, pues ellas tienen un valor muy elevado en la segunda componente y pequeño en las otras. La tercera componente por su parte, se describe fundamentalmente por factores de riesgo: hábito de fumar (Fuma) y consumo de bebidas alcohólicas (Bebe).

Realizando un análisis detallado de estos resultados, se decide eliminar las variables que no tributan a ninguna dimensión y que además no son significativas en el modelo de regresión.

Con estas consideraciones se vuelve a obtener otro modelo de regresión categórica. En él se obtiene un R^2 igual a 0.827 [17]. Nótese que prácticamente el R^2 no disminuye, si lo comparamos con el valor anterior, que era de 0.828. El análisis de varianza nuevamente es significativo.

La tabla 7 refleja los coeficientes del modelo encontrado. Evidentemente es un modelo más claro, sencillo y de mejor interpretación. Además se reafirma la TAPam como la variable más importante.

Para tener certeza de que este modelo es válido se estudia nuevamente en detalle el cumplimiento de los supuestos en el nuevo modelo encontrado siguiendo la misma metodología que en el primer modelo. Nuevamente

	Dimensión		
	1	2	3
Edad	0.357	0.451	0.111
Sexo	0.346	-0.223	0.597
Raza	0.078	-0.097	-0.192
Bebe	-0.199	0.177	-0.707
Fuma	-0.144	-0.116	-0.673
Diabetes mellitus	-0.181	-0.280	-0.096
Dislipidemia	-0.164	-0.392	-0.131
TASistB	0.806	-0.076	-0.051
TADiastB	0.825	-0.192	-0.087
TASistB1	0.856	-0.020	-0.047
TADiastB1	0.810	-0.219	-0.079
TASistB2	0.830	-0.001	-0.073
TADiastB2	0.827	-0.231	-0.097
TAPam	0.905	-0.118	-0.071
OIMC	0.384	0.098	-0.212
Col_Tot	0.359	0.732	0.003
Col_HDL	-0.113	-0.056	-0.015
Col_LDL	0.312	0.749	-0.026

Normalización principal variable

Tabla 6: Saturaciones en componentes.

se comprueba que los errores están normalmente distribuidos, que existe homogeneidad de varianza y que no hay presencia de multicolinealidad.

Hasta aquí estamos satisfechos porque se ha encontrado un modelo de regresión categórico sencillo y que cumple con los supuestos del análisis de regresión. Pero no debe olvidarse que la variable dependiente, o sea, el diagnóstico de expertos (DiagExp) es una variable categórica, luego estamos en presencia de un problema de clasificación.

La regresión categórica nos proporciona un valor predicho de la variable dependiente, sin embargo, lo que realmente se necesita es el pronóstico predicho de la clase a la que cada uno de los pacientes pertenece, según el modelo hallado.

Como se explicó en uno de los epígrafes anteriores, a nivel de menú del SPSS no aparecen opciones que brinden estas facilidades, ellos debe hacerse a nivel de sintaxis siguiendo las orientaciones que aparecen en

	Coeficientes Estandarizados	
	Beta	Significación
Sexo	-0.064	0.000
Bebe	-0.011	0.488
Fuma	-0.003	0.854
TASistB	0.000	0.985
TADiastB	0.153	0.000
TASistB1	0.174	0.000
TADiastB1	0.090	0.000
TASistB2	0.088	0.001
TADiastB2	0.208	0.000
TAPam	0.359	0.000
OIMC	0.048	0.002
Col_Tot	-0.013	0.509
Col_LDL	0.002	0.928

Variable Dependiente: Diagnóstico de Expertos

Tabla 7: Coeficientes.

dicho epígrafe.

En nuestro estudio y siguiendo las instrucciones anteriormente mencionadas obtuvimos un 84.57% de pacientes bien clasificados. Los resultados se muestran en la tabla 8 .

DiagExp	Clasificación			
	normotenso	pre hipertenso	hipertenso	Total
normotenso	397	37	0	434
pre hipertenso	51	123	19	193
hipertenso	0	24	198	222
Total	448	184	217	849

Tabla 8: Recuento DiagExp*Clasificación.

4.3 Comparación de métodos a través de las curvas ROC

Las diferentes alternativas de clasificación de la hipertensión arterial pueden ser validadas por diferentes vías, siguiendo los criterios de evaluación de los clasificadores clásicos de la Teoría Estadística y de la Inteligencia Artificial, en particular los gráficos ROC [19]. Spackman [20] demostró el valor de las curvas ROC en la evaluación y comparación de algoritmos.

Las curvas ROC constituyen otra manera de examinar el desempeño de un clasificador. Una curva ROC es un gráfico con la Razón de Falsos Positivos ($FP = 1 - Sp$) en el eje X y la Razón de Verdaderos Positivos (TP) en el eje Y. Las curvas quedan en el cuadrado $[0,1]\times[0,1]$. El vértice superior izquierdo de este cuadrado: $(0,1)$ representa al clasificador perfecto porque clasifica todos los casos positivos y todos los casos negativos correctamente porque $FP = 0$ y $TP = 1$. El vértice inferior izquierdo $(0,0)$ representa un clasificador que predice todos los casos como negativos, mientras que el vértice superior derecho $(1,1)$ corresponde a un clasificador que predice todos los casos como positivos. El punto $(1,0)$ es un clasificador pésimo que resulta incorrecto en todas las clasificaciones.

En muchos casos, un clasificador tiene un parámetro que puede ser ajustado para incrementar TP al costo de incrementar FP o decrecer FP al costo de decrecer TP. Cada parámetro puede suministrar un par (FP, TP) o lo que es lo mismo un punto sobre este cuadrado y una serie de tales puntos pueden utilizarse para plotear la curva ROC. Un clasificador que no dependa de parámetros, se representa por un punto simple, correspondiente a su par (FP, TP).

El área bajo la curva ROC puede ser usada como una medida de la exactitud en muchas aplicaciones. Si se comparan dos clasificadores, a través de sendas curvas ROC podemos decidir en general que la de mayor área bajo ella identifica al mejor clasificador, o más precisamente, el clasificador para el cual se pueda obtener un punto más alto en el eje Y (mayor ordenada) con una punto más bajo en el eje X (menor abscisa). Para un clasificador no paramétrico, e identificado por un punto ROC, la eficiencia puede medirse en términos de la distancia del punto (FP, TP) correspondiente al punto $(0,1)$. En ambos criterios, pueden introducirse pesos en términos de la importancia relativa de los FP o los TP [21].

A continuación se mostrará la comparación de la técnica de segmentación CHAID y la regresión categórica, ambos utilizados en el problema de la clasificación de la hipertensión arterial. Para la realización de la comparación se salvaron las probabilidades de pertenencia a cada clase (1-normotenso, 2-pre hipertenso, 3-normotenso) en el caso del CHAID. Como

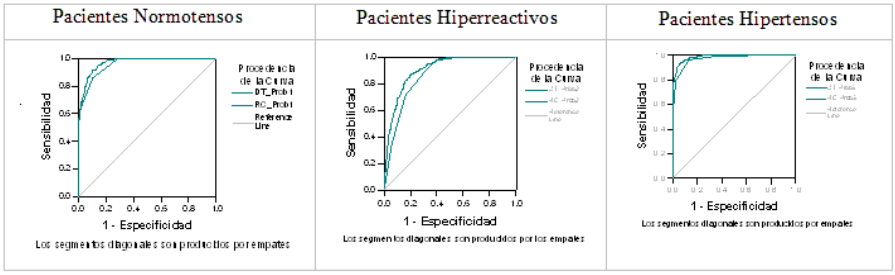

Figura 3: Resultados de las curvas ROC.

la regresión categórica no es exactamente una técnica de clasificación, en el SPSS 13 no aparece implementado las probabilidades de pertenencia a las distintas clases. Entonces se decidió calcularlas utilizando la siguiente sintaxis:

```
compute dist1 = ABS(YPRONOST + 0.89).
compute dist2 = ABS(YPRONOST - 0.3).
compute dist3 = ABS(YPRONOST - 1.48).
compute RC_Prob1 = (dist2 + dist3) / (2*(dist1 + dist2 + dist3)).
compute RC_Prob2 = (dist1 + dist3) / (2*(dist1 + dist2 + dist3)).
compute RC_Prob3 = (dist1 + dist2) / (2*(dist1 + dist2 + dist3)). exe.
```

La figura 3 y la tabla 9 muestran los resultados.

	Pacientes Normotensos	Pacientes Pre Hipertensos	Pacientes Hipertensos
Árbol de decisión	0.957	0.868	0.974
Regresión Categórica	0.969	0.904	0.988

Tabla 9: Resultado del área bajo la curva ROC

Como puede apreciarse en todos los casos la regresión categórica proporciona mejores resultados en cuanto a la clasificación de la hipertensión arterial si la comparamos con los árboles de decisión.

5 Conclusiones

El análisis de regresión categórica resulta ser una buena opción cuando nos enfrentamos a problemas en los que la mayoría de las variables analizadas son del tipo categóricas. Aplicando esta técnica se puede realizar un estudio para descubrir los modelos que relacionen las variables que intervienen en el análisis y poder realizar predicciones sobre los datos que intervienen en el análisis. Además se muestra como el Análisis de Componentes Principales para datos categóricos puede emplearse como método de selección de variables. En el ejemplo que se desarrolla se obtiene un modelo cuyo coeficiente de determinación indica que el 82.7% de la variable respuesta es explicado por las predictoras, lo cual indica que existen varias variables que influyen en el riego de padecer hipertensión arterial.

Se realizó una comparación con otro método de clasificación: los árboles CHAID. Aplicando la teoría de las curvas ROC se corroboró que la regresión categórica ofrece mejores resultados en cuanto a la clasificación de la Hipertensión Arterial.

Referencias

[1] Agresti, A. (2002) *Categorical Data Analysis*, Second ed.. John Wiley & Sons, New York.

[2] SPSS 10 para Windows. Manual de usuarios. Capítulo 12, SPSS Soft.

[3] Vicéns Otero, J.; Medina Moral, E. (2005) "Análisis de datos cualitativos", en: www.uam.es/personal_pdi/economicas/eva/pdf/tab_conting.pdf, consultado el 22-Sep-2007, 9:30 a.m.

[4] Grau, R. (2000) "Independencia de variables y medidas de asociación", Capítulo 3. Segunda parte. Preprint, Universidad Central de las Villas, Cuba.

[5] Hair, J.F. et al. (1999) *Análisis Multivariante*, 5a ed. Prentice Hall, Madrid.

[6] Johnson, R.A.; Wichern, D.W. (2002) *Applied Multivariate Statistical Analysis*, Fifth edition. Pearson Education International, United States of America.

[7] Linting, M. (2007) *Nonparametric Inference in Nonlinear Principal Components Analysis: exploration and beyond*. Doctoral Thesis, Leiden University.

[8] Meulman, J.J.; Heiser, W.J. (2004) SPSS Categories 13.0.

[9] Stanton, J.M., et al. (2001) "A brief history of linear regression for statistics instructors", *Journal of Statistics Education* **9**(3).

[10] Draper, N.R.; Smith, H. (1980) *Applied Regression Analysis*. Editorial Pueblo y Educación.

[11] Haber, L. (2001) "Categorical regression analysis of toxicity data", *Comments on toxicology* **7**(5-6): 437–452.

[12] Van der Kooij, A.J. (2007) *Prediction Accuracy and Stability of Regression with Optimal Scaling Transformations*. Doctoral Thesis, Leiden University.

[13] Ramsay, J.O.; Monotone, Wichern. (1988) "Regression splines in action", en:
`http://www.fon.hum.uva.nl/praat/manual/spline.html`, consultada 28-Ene-2008, 10:15 a.m.

[14] "Tuotromedico: Hipertensión Arterial", en:
`http://www.tuotromedico.com/temas/hipertension`, consultada 20-Mar-2008, 1:18 p.m.

[15] Microsoft ® Encarta ® 2006, © 1993-2005 Microsoft Corporation. Reservados todos los derechos.

[16] Aron, A.; Aron, E. (2002) *Statistics for the Behavioral and Social Sciences*, Second edition. Prentice Hall.

[17] Navarro Céspedes, J.M. (2008) *Análisis de Componentes Principales y Análisis de Regresión para datos categóricos. Aplicación en HTA*. Tesis de Maestría, Universidad Central de las Villas, Santa Clara, Cuba.

[18] Calero, A. (1998) *Estadística II*. Pueblo y Educación, La Habana, Cuba.

[19] Swets, A.J. (1988) "Measuring the accuracy of diagnostic systems", *Science* 240: 1285–1293.

[20] Spackman, K.A. (1989) "Signal detection theory: Valuable tools for evaluating inductive learning", *Sixth International Workshop on Machine Learning*, San Mateo, CA.

[21] Fawcett, T. (2004) "ROC graphs: notes and practical considerations for researchers, en:
home.comcast.net/˜tom.fawcett/public_html/papers/ROC101.pdf,
consultado 5-May-2008, 3:58 p.m.