Universidad Central "Marta Abreu" de Las Villas
Facultad de Matemática, Física y Computación
Departamento de Matemática

Trabajo en opción al Título de Master en Matemática Aplicada

Análisis de Componentes Principales y Análisis de Regresión para Datos Categóricos. Aplicación en HTA.

Autor: Lic. Juan Manuel Navarro Céspedes

Tutora: Dra. Gladys Casas Cardoso

Santa Clara
2008

En la memoria de mi profesora, tutora y amiga

Gladys Cardoso Romero

Agradecimientos

No me alcanzaría el presente trabajo para reflejar los nombres de todos aquellos que han significado algo en mi vida, razón por la cual tengo algo que agradecerles. Solo que en este momento no quisiera dejar de mostrar mi agradecimiento en particular a los que de alguna manera contribuyeron a la realización del este trabajo

- A Dios por haberme dado la fe y la perseverancia para haber vencido los obstáculos y haber llegado hasta aquí.

- A mis padres y a mi hermano por toda la confianza y todo el apoyo que me brindaron. Ellos fueron y serán siempre fuente de inspiración en mi trabajo

- A Gladys Casas, por haber aceptado ser mi tutora. Un millón de gracias por todo el tiempo que ha tenido que dedicarme teniendo un montón de cosas importantes por hacer, por su paciencia, dedicación y por estar siempre disponible en los momentos en que la necesité.

- A Mayra y a Marisela por la valiosa ayuda con el inglés.

- Un agradecimiento especial, aunque quizás nunca lo sepa, a Anita J. Van der Kooij por su ayuda desinteresada siempre que tuve que acudir a ella.

- A todos los que me ayudaron con sus sugerencias y críticas.

Resumen

En la presente investigación, se presentan los métodos más importantes para el análisis de datos categóricos. Nuestra contribución está basada en la aplicación de dos nuevos métodos estadísticos categóricos: Análisis Regresión y Análisis de Componentes Principales, ambos categóricos, en un problema médico. La primera técnica aplica la metodología de escalamiento óptimo para cuantificar las variables categóricas, incluyendo la variables respuesta en el análisis de regresión, simultáneando la optimización del coeficiente de regresión múltiple. Los niveles de escalamiento que pueden ser aplicados son nominal, spline no monótono, ordinal, spline monótono o numérico.

La segunda técnica es la equivalente no lineal del Análisis de Componentes Principales (ACP). Las ventajas más importantes del no lineal sobre el ACP lineal están dadas por el hecho que incorpora variables nominales u ordinales, y además posibilita la manipulación y descripción de relaciones no lineales entre las variables.

Se presenta un problema de predicción de la hipertensión en el municipio de Santa Clara. Se obtuvo un modelo con buenos resultados con todas las variables predictoras. Se utilizó el ACP Categórico como un procedimiento exploratorio y como técnica de selección. Con las variables seleccionadas, se obtuvo un nuevo modelo de regresión categórica. Se verificaron los supuestos en todos los modelos. Finalmente, con el objetivo de resolver un problema de clasificación, se utilizó la regresión categórica como método discriminante.

Abstract

In this research, the most important methods for categorical statistical analysis are presented. Our contribution is based on the application in a medical problem of two new categorical statistical methods: Regression Analysis and Principal Component Analysis, both categorical. The first technique applies optimal scaling methodology to quantify categorical variables, including the response variable in regression analysis, simultaneously optimizing the multiple regression coefficients. The scaling levels that can be applied are nominal, nonmonotonic spline, ordinal, monotonic spline or numerical.

The second technique is the nonlinear equivalent of the standard Principal Component Analysis (PCA). The most important advantages of nonlinear over linear PCA are given by the fact that it incorporates nominal and ordinal variables, and also because is possible to handle and discover nonlinear relationships between variables.

An hypertension prediction problem in Santa Clara is presented. A model with all predictive variables was obtained with good results. Categorical PCA are used as exploratory procedure and as feature selection technique. With the selected variables, a new categorical regression model was obtained. Assumptions are verified in all models. Finally, in order to solve the classification problem, categorical regression was used as discriminant method.

Índice

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.

La tecnología informática disponible hoy en día, casi inimaginable hace solo dos décadas, ha hecho posibles avances extraordinarios en el análisis de datos psicológicos, sociológicos y de otro tipo de números referidos al comportamiento humano, incluso en otras áreas del conocimiento como la medicina, la meteorología, la bioinformática y la educación. Este impacto es más evidente en la relativa facilidad con la que los ordenadores pueden analizar enormes cantidades de datos complejos. Casi cualquier problema se puede analizar fácilmente hoy en día por un número ilimitado de programas estadísticos, incluso en ordenadores personales. Además, los efectos del progreso tecnológico han extendido aun más la capacidad de manipular datos, liberando a los investigadores de las restricciones del pasado y permitiéndoles así abordar investigaciones más sustantivas y ensayar sus modelos teóricos. Las limitaciones metodológicas no son ya un asunto crítico para el teórico empañado en la búsqueda de evidencia empírica. Gran parte de esta creciente comprensión y pericia en el análisis de datos ha venido a través del estudio y desarrollo de la estadística y de la inferencia estadística.

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.

No son pocos los métodos que introducen las denominadas variables "dummy" para trabajar con variables que no tienen propiedades numéricas reales. En estos métodos las variables categóricas son divididas en variables indicadoras de cada categoría, donde el 1 representa la presencia de la misma y el 0 la ausencia. Estas variables "dummy" son utilizadas como variables numéricas en el análisis. Tales métodos, sin embargo, suelen ser muy intensivos, especialmente cuando las variables tienen muchas categorías.

El trabajo con datos categóricos data desde 1902 con el descubrimiento más importante de Karl Pearson: el test chi cuadrado. Sobre la década de los 60 hubo una explosión, dado en gran medida por el desarrollo de la informática, de métodos de análisis estadísticos para datos categóricos [1].

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 con categorías mixtas (nominal, ordinal y numérica). En el SPSS 13 aparecen los denominadas métodos con escalamiento óptimo como el Análisis de Componentes Principales y el Análisis de Regresión. De aquí entonces que surge la idea de realizar un estudio de los mismos.

Por todo lo anteriormente expuesto se plantea lo siguiente:

Objetivo General:

Mostrar como se pueden procesar datos categóricos, con ayuda de técnicas alternativas de análisis exploratorio de datos, selección de rasgos y regresión ajustadas a este tipo de datos, para ilustrar un procedimiento típico de trabajo en análisis estadísticos donde intervienen variables con diferente nivel de medición.

El objetivo general se puede desglosar en los siguientes:

Objetivos Específicos:

- Mostrar como se pueden obtener modelos válidos de regresión para datos categóricos incluyendo el análisis de los supuestos.
- Mostrar como se puede utilizar el método de componentes principales para datos categóricos como método exploratorio de datos y como método de selección de las variables a incluir en un modelo de regresión categórica.
- Ilustrar la aplicación de los métodos anteriores en un problema médico, en particular para el pronóstico de la hipertensión arterial.

Los objetivos se pueden reformular en las siguientes:

Preguntas de Investigación:

- ¿Cómo deben aplicarse el método de componentes principales y la regresión categórica cuando existen variables con diferente nivel de medición para obtener resultados confiables y válidos?
- ¿Se obtendrán buenos resultados, al utilizar tales métodos para datos categóricos, en investigaciones médicas, específicamente relacionadas con el análisis integral de riesgos de una enfermedad como la HTA?

Después de haberse realizado el marco teórico se planteó la siguiente:

Hipótesis de Investigación:

Existen formas de aplicar satisfactoriamente los métodos de regresión y de componentes principales categóricos a problemas en los que la mayoría de las variables son discretas obteniéndose buenos y rigurosos resultados en caracterización integral de riesgos de una enfermedad como la HTA

El trabajo que se presenta a continuación está conformado de la siguiente manera:

Capítulo 1 Análisis Estadístico de Datos Categóricos

En el mismo de describen los métodos estadísticos más importantes presentes en la literatura sobre el tratamiento con datos categóricos.

Capítulo 2 Análisis de Componentes Principales y Análisis de Regresión para Datos Categóricos

En este capítulo se describen los fundamentos de ambos métodos.

Capítulo 3 Aplicación: Estudio de la Hipertensión Arterial (HTA)

En este último se muestra una aplicación de los métodos explicados en los capítulos 1 y 2, en un estudio de caracterización de la Hipertensión Arterial en cinco policlínicos del municipio de Santa Clara.

Capítulo 1. Análisis Estadístico de Datos Categóricos

En la década de los 60 comenzó una explosión en el desarrollo de métodos estadísticos para el análisis de datos categóricos [1]. Tal desarrollo continúa en nuestros días de manera acelerada. El desarrollo vertiginoso de las computadoras electrónicas constituyó sin dudas un catalizador poderoso en todo este proceso [2].

1.1 Test Chi Cuadrado. Tablas de Contingencia

En las ciencias sociales, de la salud y del comportamiento es bastante frecuente encontrarse con variables categóricas. El sexo, la raza, el padecimiento de una enfermedad o un determinado síntoma, la categoría laboral; entre otros, son ejemplos de algunas variables categóricas que se pueden encontrar. Las mismas son variables sobre las cuales únicamente se pueden obtener una medida de tipo nominal u ordinal pero con pocos valores.

Cuando se trabaja con variables categóricas, los datos suelen organizarse en tablas de doble entrada en las que cada entrada representa un criterio de clasificación (una variable categórica). Como resultado de esta clasificación, 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 [3].

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 [4].

Para identificar relaciones de dependencia entre variables no pueden utilizarse solamente las Tablas de Contingencia. Para ello se debe utilizar alguna medida de asociación, acompañada de su correspondiente prueba de significación. Un ejemplo de ello es el estadístico χ^2 (Chi-Cuadrado) propuesto por Pearson desde 1911. El mismo permite contrastar las hipótesis de que las dos variables utilizadas son independientes. Cuando dos criterios de clasificación son independientes, las frecuencias esperadas (m_{ij}) se estiman de la siguiente manera:

$$m_{ij} = \frac{(total\ de\ la\ \ fila\ i) \times (total\ de\ la\ columna\ \ j)}{N} \qquad (1.1)$$

…donde N es el total de casos analizados.

Una vez obtenidas las frecuencias esperadas se calcula el estadístico de la siguiente manera:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \qquad (1.2)$$

… donde los n_{ij} representan las frecuencias observadas.

El valor de χ^2 calculado se compara con el valor tabulado de una χ^2 para un nivel de confianza determinado y $(n-1)*(k-1)$ grados de libertad, donde *n* y *k* son el número de filas y columnas respectivamente. Si el valor calculado es mayor que el valor de una $X^2_{(n-1)(k-1)}$, significará que las diferencias entre las frecuencias observadas y las frecuencias teóricas o esperadas son muy elevadas y por tanto diremos con un determinado nivel de confianza que existe dependencia entre los factores o atributos analizados [3, 4].

El estadístico Chi-Cuadrado debe cumplir ciertas exigencias. Los datos debe provenir de muestras aleatorias con distribuciones multinomiales y las frecuencias esperadas en cada celda no deben ser excesivamente pequeñas. Tradicionalmente se ha recomendado que las frecuencias esperadas deban ser mayores o iguales a 5 aunque quizás esto sea muy exigente. Lo importante es cómo evitar en general este problema. En esencia, las tablas de contingencia no pueden tener dimensiones demasiado grandes. Si se quiere eliminar frecuencias esperadas bajas, se deben reducir las dimensiones de la tabla. Para mejorar la aproximación de la distribución en una tabla 2×2, se utiliza frecuentemente la "corrección por continuidad de Yates", (1934). Esta corrección consiste en reducir en 0.5 el valor absoluto de las diferencias $n_{ij} - m_{ij}$ del estadístico X^2 antes de elevarlas al cuadrado. Algunos autores han desarrollado una controversia sobre los méritos o insuficiencias de esta corrección, pero pese a todo, se utiliza bastante.

Por otra parte se puede usar, en lugar del Test Chi-cuadrado de Pearson, un test basado en la distribución hipergeométrica y en la hipótesis de independencia, conocido como Test exacto de Fisher (1935). El mismo ofrece la probabilidad exacta de obtener las frecuencias de hecho obtenidas o cualquier combinación alejada de la hipótesis de independencia.

Otro estadístico que se utiliza para el análisis de la relación entre las variables es el denominado Razón de Verosimilitud (Fisher, 1924; Neynam y Pearson, 1928) que se obtiene mediante la fórmula:

$$Razon\,de Verosimilitud = 2\sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{m_{ij}}\right) \qquad (1.3)$$

En este caso estamos en la presencia de un estadístico asintóticamente equivalente al Chi-Cuadrado aunque con una forma más complicada, el mismo es muy utilizado para estudiar las relaciones entre variables categóricas [3, 5] .

1.1.1 Algunas medidas de asociación entre variables aleatorias discretas nominales y ordinales

En muchas investigaciones, más que discernir la dependencia de dos variables interesa la naturaleza y fortaleza de la asociación. Los indicadores que miden esto se llaman medidas de asociación. Ninguna medida resume adecuadamente todos los tipos de asociación posibles. Las medidas se diferencian por su interpretación, y por la forma en que ellas pueden reflejar asociaciones perfectas o parciales. Una medida de asociación determinada puede tener un valor bajo para una tabla dada, pero para ello no significa que las variables objeto de estudio no estén relacionadas, sino que ellas no están relacionadas en la forma a la que es sensible dicha medida. Por ello ninguna medida individual es la mejor para todas las situaciones. Al seleccionar una, debe tenerse en cuenta el tipo de datos, la hipótesis de interés así como las propiedades de cada medida.

Las medidas nominales (se asumen apenas que las dos variables de la tabla están medidas nominalmente) pueden suministrar solamente alguna indicación sobre la estrechez de la asociación y no pueden indicar casi nada sobre la dirección o cualquier otra cosa de la naturaleza de la relación. Las medidas pueden clasificarse en dos tipos: aquellas que están basadas en el Chi-cuadrado y aquellas que se fundamentan en la lógica de reducción proporcional de error (PRE, Proportional Reduction Error, como se le dice clásicamente).

El test Chi-cuadrado en sí, no proporciona una buena medida del grado de asociación entre las dos variables; pero como está tan expandido el uso del Chi-cuadrado en la dócima de independencia, se ha estimulado la definición de medidas de asociación basadas en el Chi-cuadrado tratando de minimizar la influencia del volumen de la muestra y de los grados de libertad, así como restringir el rango de los valores de la medida al intervalo [0-1]. Estas medidas ayudan entonces a comparar los resultados del Chi-cuadrado en tablas diferentes cuando hay variación de las dimensiones y de los volúmenes de las muestras. Sin estas correcciones, es absolutamente inadmisible comparar con el Chi-cuadrado tales tablas. El llamado coeficiente Phi

modifica el Chi-cuadrado dividiéndolo por el volumen de la muestra y extrayendo la raíz cuadrada del resultado:

$$\phi = \sqrt{\frac{\chi^2}{N}} \tag{1.4}$$

Para tablas en las cuales una dimensión es mayor que 2, Phi no yace necesariamente entre 0 y 1, ya que el valor del Chi-cuadrado puede ser más grande que el volumen de la muestra. Por tanto Phi sólo queda estandarizado en el intervalo [0-1] en tablas en las cuales $R = 2$ ó $C = 2$ (*R* y *C* denotan siempre el número de filas y columnas).

Para obtener una medida que debe yacer siempre entre 0 y 1 Pearson sugirió el uso de:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \tag{1.5}$$

que se conoce como coeficiente de contingencia. El valor de este coeficiente está siempre entre 0 y 1; pero generalmente no puede llegar a ser 1. De hecho, el máximo valor de *C* depende sobre el número de filas y columnas. Por ejemplo para una tabla 4×4, el máximo valor de *C* es 0.87. Cramer introdujo la siguiente variante:

$$V = \sqrt{\frac{\chi^2}{N * (K - 1)}} \tag{1.6}$$

… donde $K = \min(\textit{número de filas}, \textit{número de columnas})$

Este estadístico, conocido como V de Cramer, puede alcanzar el máximo de 1 en tablas de cualquier dimensión y además coincide con Phi si una de las dimensiones es igual a 2.

Las medidas nominales basadas en el test Chi-cuadrado son difíciles de interpretar. Entre las tres mencionadas se recomienda la V de Cramer. Su carácter estandarizado permite al menos comparar la "estrechez de la asociación" entre tablas diferentes; pero esta estrechez o fortaleza de la asociación que estamos comparando, no responde a ningún concepto intuitivo claro de asociación [6].

La idea de la Reducción Proporcional en el error fue introducida por Goodman y Kruskal (1954). Con este tipo de medidas, el significado de la asociación resulta más claro. En esencia estas medidas indican relación entre:

- La magnitud o probabilidad del error al predecir los valores de una variable basándonos en el conocimiento sólo de esa variable.

- La magnitud o probabilidad del error al predecir los valores de esa variable basándola en el conocimiento de otra variable adicional.

El estadístico lambda de Goodman y Kruskal siempre toma valores entre 0 y 1. Un valor de 0 significa que la variable independiente no ayuda en la predicción de la variable dependiente. Un valor de 1 significa que la variable independiente permite pronosticar exactamente las categorías de la dependiente (la perfección ocurre solamente cuando en cada fila hay solamente una celda diferente de cero). Cuando dos variables son independientes, lambda es 0; pero un valor 0 para lambda, no significa que las variables tengan independencia estadística. Como todas las medidas de asociación, lambda se construye para medir un tipo específico de dependencia: la reducción del error cuando una variable es usada para predecir los valores modales de la otra. En el caso particular que no exista este tipo de asociación, lambda resultará cero. Recuérdese siempre que no existe ninguna medida sensible a todos los tipos imaginables de asociación [5].

1.2 Método Exacto y Algoritmo de Monte Carlo

Por defecto, paquetes profesionales como el SPSS calculan los niveles de significación usando el método asintótico. Esto significa que los *p* valores se estiman sobre la base del supuesto de que los datos, dado un tamaño de muestra suficientemente grande, se ajusten a una distribución particular. Sin embargo, cuando el conjunto de datos es pequeño, escaso, contiene varios lazos, están desbalanceados o están mal distribuidos, el método asintótico puede no producir resultados confiables. En esta situación, es preferible calcular el nivel de significación basado en la distribución exacta de un test estadístico. Esto se logra a través del cómputo de valores exactos de *p* para una clase amplia de prueba de hipótesis, para datos categóricos ya sean ordenados o no. La metodología estadística que sustenta estos test exactos está bien establecida en la literatura [7-12] y en esencia es una generalización del Test Exacto de Fisher para una tabla de contingencia 2×2.

Para resolver este problema se ha desarrollado el algoritmo y método de Monte Carlo. Para los pequeños conjuntos de datos, el algoritmo asegura el cómputo rápido de los valores exactos de *p*. Si un conjunto de datos es demasiado grande para el algoritmo exacto, entonces el algoritmo

Monte Carlo lo reemplaza para estimar los valores exactos de *p* a cualquier nivel exactitud deseada.

1.2.1 Método Exacto

En muchas ocaciones cuando se tienen datos pequeños, escasos, pesados o simplemente desequilibrados y la validez de la teoría de la correspondiente muestra grande está en duda, es preferible calcular el nivel de significación basado en la distribución exacta de un test estadístico. Esto permite obtener un valor exacto de *p* sin confiar en la distribución que los datos pueden no satisfacer. El cálculo exacto produce siempre un resultado confiable, sin importar el tamaño, distribución, o balance de los datos. El cálculo del resultado exacto puede resultar desde el punto de vista computacional muy intensivo, y puede en algunas ocasiones exceder el límite de la memoria de la máquina. En general, los test exactos pueden realizar los cálculos muy rápidamente para muestras con tamaño menor a 30. En la literatura puede encontrarse las condiciones para las cuales los test exactos pueden obtenerse de manera rápida [13]

Afortunadamente, dos acontecimientos han hecho los cálculos del valor *p* exacto viables en la práctica, aunque no en todos los casos, como se mencionó con anterioridad. En primer lugar, la revolución informática ha redefinido drásticamente lo que es computacionalmente factible y asequible. En segundo lugar, se han publicado en las últimas décadas, muchos algoritmos computacionales nuevos, rápidos y eficientes.

Idealmente se pudiera usar el valor *p* exacto todo el tiempo. Estos son, después de todo, el patrón de oro. Solo con la decisión de aceptar o rechazar la hipótesis nula sobre la base que un valor *p* exacto garantice no violar el error de tipo I en el nivel de significación deseado. En la práctica, sin embargo, no es posible usar el valor *p* exacto todo el tiempo. Es difícil cuantificar como un conjunto de datos puede ser resuelto por el algoritmo exacto, debido a que depende de muchos factores aparte del tamaño de la muestra. En ocasiones se puede calcular el valor *p* exacto para un conjunto de datos en los cuales el tamaño esta por encima de 20 000, y en otras ocasiones falla el cómputo del valor *p* exacto para un conjunto de datos que su tamaño es menor que 30. El tipo de test exacto deseado, el grado de desbalance en la localización de los sujetos a tratar, el número de filas y columnas en una tabla de contingencia, el número de ligas en los datos, y una variedad de otros factores interactúan de manera compleja para determinar si un conjunto de datos en particular se presta para la inferencia exacta. Es por tanto una tarea muy difícil especificar los límites superiores de viabilidad para los algoritmos exactos. Es más útil especificar el tamaño de muestra y las dimensiones de las tablas dentro de las cuales los algoritmos exactos producen respuestas rápidas, es decir, en solo unos segundos [13].

Con el fin de calcular un valor *p* exacto, deben enumerarse todos los resultados que pudieran ocurrir en algún conjunto de referencia además de los resultados que realmente se observaron. Entonces se ordenan estos resultados por alguna medida de discrepancia que refleja la desviación de la hipótesis nula. El valor *p* exacto es igual a la suma exacta de las probabilidades de esos resultados en el conjunto de referencia que son al menos tan extremos como la que se observó en realidad.

1.2.2 Método de Monte Carlo

Aunque los resultados exactos son confiables, algunos conjuntos de datos son muy grandes para que el valor exacto *p* pueda calcularse, todavía no se asume que se ajusten a la distribución necesaria para usar el método asintótico. En este caso, el método de Monte Carlo proporciona un estimador insesgado del valor exacto *p*, sin el requerimiento del método asintótico. El método de Monte Carlo es un método de muestreo repetido. En algunas tablas observadas, existen muchas, cada una con la misma dimensión, filas y columnas marginales que las tablas observadas. El método Monte Carlo muestrea reiteradamente un número específico de las posibles tablas para obtener una estimación insesgada del verdadero valor de *p* [13].

Algunos conjuntos de datos, además de ser demasiado grandes para el cálculo del valor *p* exacto, son demasiados escasos o desbalanceados para que el resultado asintótico sea fiable. Por tanto, el próximo paso es usar la opción de Monte Carlo. Esta opción puede generar un estimador del valor *p* exacto extremadamente fiel a partir de un conjunto de referencia de todas las tablas con los valores marginales observados un número grande de veces. Con la condición de cada tabla sea muestreada en proporción a su probabilidad hipergeométrica, la fracción de las tablas muestreadas que son por lo menos tan extremos como las tablas observadas proporciona un estimador insesgado del valor *p* exacto. Esto es, si *M* tablas son muestreadas a partir de un conjunto de referencia, y Q es de ellas como mínimo tan extremo como la tabla observada, el estimador Monte Carlo del valor *p* exacto es

$$\hat{p} = \frac{Q}{M} \tag{1.7}$$

La varianza de este estimador se obtiene aplicando la teoría binomial:

$$\mathrm{var}\left(\hat{p}\right) = \frac{p(1-p)}{M} \tag{1.8}$$

Por tanto, un intervalo de confianza del $100 \times (1-\gamma)\%$ para *p* es:

$$CI = \hat{p} \pm z_{\gamma/2}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{M}} \tag{1.9}$$

… donde Z_α es el α -ésimo percentil de la distribución normal estándar.

¿Qué tan buenas son las estimaciones Monte Carlo? ¿Por qué habría de usarse en vez del valor *p* asintótico? Existen varias ventajas para usar el método Monte Carlo en lugar del valor *p*:

1. El estimador Monte Carlo es insesgado.
2. La estimación de Monte Carlo se acompaña de un intervalo de confianza dentro del cual el valor *p* exacto se garantiza que esté en el nivel de confianza especificado. El valor *p* asintótico no está acompañada de ninguna de tales garantías probabilísticas.
3. La amplitud del intervalo de confianza puede ser hecha arbitrariamente pequeña, muestreando más tablas a partir del conjunto referencia.
4. En principio, se puede reducir el ancho del intervalo de confianza a tal punto que el valor *p* de Monte Carlo llega a ser indistinguible en comparación con el valor *p* exacto hasta tres cifras decimales. Para todos los propósitos prácticos, pudiera entonces afirmarse que hay valor *p* exacto. Por supuesto, esto pudiera tomar varias horas para lograrlo.
5. En la práctica, no hace falta ir tan lejos. Simplemente conociendo que el límite superior del intervalo de confianza está por debajo de 0.05, o que el límite inferior del intervalo de confianza está por encima de 0.05 es satisfactorio.

El algoritmo de Monte Carlo proporciona un estimado del valor *p* exacto, denominado *p* de Monte Carlo, el cual puede hacerse tan preciso como sea necesario para el problema que se esta resolviendo. Típicamente, su resultado es preciso al 99%, pero puede incrementarse el nivel de precisión para cualquier grado arbitrario simplemente muestreando más resultados a partir del conjunto referencia. También, ellos son garantía para resolver cualquier problema, no importa cuan largo sea el conjunto de datos. Por tanto, ellos proporcionan un robusto y fiable respaldo para las situaciones en las cuales el algoritmo de enumeración completa falla.

1.2.3 Selección entre el valor *p* Asintótico, Exacto, y de Monte Carlo

Finalmente, una guía general del uso del valor *p* asintótico, exacto, o de Monte Carlo pudiera ser la siguiente:

- Es sabio no reportar nunca un valor asintótico *p* sin chequear primero su precisión contra el correspondiente valor exacto o de Monte Carlo. No se puede predecir fácilmente a priori cuando un valor *p* asintótico será lo suficientemente exacto.
- La selección del valor exacto versus Monte Carlo es en gran medida una de las conveniencias. El tiempo requerido para el cálculo del exacto es menos predecible que los cálculos para el Monte Carlo. Generalmente, el cálculo exacto o produce una respuesta rápida, o por el contrario termina rápidamente con el mensaje que el problema es demasiado difícil para el algoritmo exacto. En ocasiones, sin embargo, los cálculos exactos pueden tomar varias horas, en dichos casos es mejor interrumpirlos y repetir el análisis con el método de Monte Carlo. Los valores *p* de Monte Carlo son para la mayoría de los propósitos prácticos tan buenos como el valor *p* exacto. El método tiene la ventaja adicional que toma una cantidad predecible de tiempo, y su respuesta está disponible en cualquier nivel de precisión deseado.

1.3 Análisis de correlación lineal

Una forma de explicar la dependencia entre dos variables aleatorias, eliminando las influencias de las dimensiones en los sistemas de medidas originalmente usados, es el coeficiente de correlación. El coeficiente de correlación establece una medida del posible nexo existente entre las variables consideradas. Stanton [14] explica que es Sir Francis Galton (1889), quien tiene el mérito de ser el primero en utilizar la correlación, aunque es su discípulo Karl Pearson (1857-1936) quien estudia con profundidad sus propiedades.

Una medida de asociación lineal especialmente apropiada para estudiar la relación entre variables de intervalo o razón es el *coeficiente de correlación de Pearson*. Pearson [15], define una medida de asociación lineal entre dos variables cuantitativas X e Y. El coeficiente de correlación entre dichas variables se escribe así:

$$\rho_{xy} = \frac{C(X,Y)}{\sqrt{V(X)V(Y)}} \qquad (1.10)$$

… donde C(X,Y) representa la covarianza y V(X) y V(Y) son las varianzas de X e Y respectivamente. Este coeficiente cumple con la siguiente propiedad:

$$-1 \leq \rho \leq 1$$

Si X e Y son variables aleatorias independientes, su coeficiente de correlación es cero. Aunque en general no es cierto que una correlación cero indique independencia, este coeficiente es una buena medida de la asociación entre las dos variables.

Cuando las dos variables son cualitativas ordinales, es posible estudiar el nexo entre ellas usando el *coeficiente de correlación de Spearman* [16]. Sea n el número de elementos de la muestra. A cada elemento de la muestra se le asignan los rangos correspondientes de las variables *X* y *Y*. Sean $x_1,\ldots,x_n$ los rangos de la primera variable y $y_1,\ldots,y_n$ los rangos de la segunda variable. Para cada elemento se calcula la diferencia $d_i = x_i - y_i$, como una indicación de la disparidad entre los dos conjuntos de rangos en esa observación. Cuanto mayor sean las d_i, menos perfecta es la asociación entre las dos variables.

El cálculo del coeficiente de correlación sería afectado por el uso directo de las d_i. Las d_i negativas cancelarían las positivas cuando se trata de determinar la magnitud de la discrepancia. Para eliminar esta dificultad se emplea d_i^2 en lugar de d_i. Mientras mayores sean los d_i, mayor será el valor de $\sum_{i=1}^{n} d_i^2$.

Usando $X_i = x_i - \bar{x}$ y $Y_i = y_i - \bar{y}$ la expresión general para un coeficiente de correlación puede escribirse como:

$$r = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i^2}} \qquad (1.11)$$

Como las variables *X* e *Y* son cualitativas, es posible usar el rango del valor original, en este caso *X* e *Y* toman como valores los rangos correspondientes. El coeficiente de correlación de las medidas originales se transforma en el coeficiente de correlación por rangos de Spearman a nivel poblacional y se escribe ρ_s

La suma de todos los valores de la variable x_i correspondiente a la suma de los n enteros $1,\ldots,n$, es:

$$\sum_{i=1}^{n} X_{(i)} = \frac{n(n+1)}{2} \tag{1.12}$$

La suma de los cuadrados:

$$\sum_{i=1}^{n} X_{(i)}^2 = \frac{n(n+1)(2n+1)}{6} \tag{1.13}$$

Así,

$$\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} (x_{(i)} - \bar{x}_{(i)})^2 = \sum_{i=1}^{n} x_{(i)}^2 - n\bar{x}_{(i)}^2 \tag{1.14}$$

$$\sum_{i=1}^{n} X_i^2 = \frac{n(n+1)(2n+1)}{6} - n\frac{(n+1)^2}{4} = \frac{n^3 - n}{12} \tag{1.15}$$

De manera análoga:

$$\sum_{i=1}^{n} Y_i^2 = \frac{n^3 - n}{12} \tag{1.16}$$

Haciendo $d_i = X_i - Y_i$ se obtiene:

$$2\sum_{i=1}^{n} X_i Y_i = \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} d_i^2 \tag{1.17}$$

Sustituyendo (1.12), (1.13) y (1.14) en (1.8) se tiene la fórmula para el coeficiente de correlaciones por ranking de Spearman:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{1.18}$$

Palmer y otros [17], usan la distribución muestral asintótica del coeficiente de correlación de Spearman:

$$r_s \xrightarrow[n\to\infty]{L} V \sim N\left[\rho_s; \frac{1}{n-1}\right] \tag{1.19}$$

Cuando de las dos variables, al menos una es cualitativa ordinal, se puede usar el *coeficiente de correlación de Kendall*, designado por r_k. El mismo es una medida de correlación conveniente para datos que se puedan ordenar [16].

En cada una de las variables se sustituye cada valor por sus respectivos rangos. Los rangos de la primera variable *X* se colocan en su orden natural:

$$x_1 = 1, x_2 = 2, \ldots, x_n = n$$

A cada $x_{(i)}$ de la secuencia anterior se le asocia el ranking de la otra variable en esa unidad poblacional. Sea $y_{[i]}$ el ranking de *Y*, cuando *X* alcanza el ranking *i*.

X	$x_{(1)} = 1$	$x_{(2)} = 2$	…	$x_{(i)} = i$	…	$x_{(n)} = n$
Y	$y_{[1]}$	$y_{[2]}$	⋯	$y_{[i]}$	⋯	$y_{[n]}$

Para $h > i$, sea α_h la cantidad de rangos $y_{(h)}$ que cumplen la propiedad: $y_{(h)} > y_{(i)}$, es decir la cantidad de concordancias en cuanto al rango. Sea γ_h la cantidad de rangos $y_{(h)}$ donde para $h > i$ se tiene $y_{(h)} < y_{(i)}$, es decir, la cantidad de discrepancias entre los rangos de ambas variables.

La cantidad total de pares de unidades poblacionales a comparar es el número de variaciones sin repetición de los *n* elementos de la muestra.

$$V_{(n,2)} = \frac{n!}{(n-2)!} = \frac{n(n-1)}{2} \tag{1.20}$$

Se define Φ de la forma siguiente:

$$\Phi = \sum_{h=1}^{n-1} (\alpha_h - \gamma_h) \tag{1.21}$$

$(\alpha_h - \gamma_h)$ es la diferencia ente concordancias y discrepancias en cuanto a los rangos, cuando se compara el *h*-ésimo ranking de Y. La suma de esas diferencias tiene la siguiente propiedad, si la suma es igual a las $V_{(n-2)}$, eso quiere decir que no hubo discrepancias entre los rangos de *X* y de *Y*, por lo tanto la concordancia es perfecta, por ello las variables *X* e *Y* están correlacionadas de forma perfecta.

Si las discrepancias son ciertas en su totalidad, Φ toma el valor $-V_{(n-2)}$. En base a estas consideraciones se define el coeficiente de correlación por rangos de Kendall:

$$r_k = \frac{\sum_{h=1}^{n-1} (\alpha_h - \gamma_h)}{\frac{1}{2} n(n-1)} \tag{1.22}$$

r_k asume valores en el intervalo $[-1,1]$. Cuando las variables *X* e *Y* no están correlacionadas, el coeficiente de correlación por ranking de Kendall r_k toma el valor cero [18].

Una medida de asociación entre variables que ha sido desarrollada la proporciona el *coeficiente de concordancia W de Kendall.* Dicho coeficiente tiene una estrecha relación con el test de Friedman, es de hecho una versión ampliada de dicho test: [13]

$$W = \frac{T_F}{N(K-1)} \tag{1.23}$$

W de Kendall puede interpretarse como el coeficiente de concordancia, que es una medida del acuerdo entre jueces. Cada caso es un juez o evaluador y cada variable es un elemento o persona siendo juzgada. Se calcula la suma de los rangos para cada variable. W de Kendall varia entre 0 (acuerdo nulo) y 1 (acuerdo absoluto) [19] .

1.3.1 Relación entre los coeficientes W de Kendall y R de Spearman

Una medida de asociación diferente es la conocida como coeficiente de correlación por rangos ordenados de Spearman. Esta medida es aplicable solamente si existen dos jueces ($N = 2$), cada uno con *K* ranking. ¿Pudiera esta medida ser extendida si $N > 2$? Una aproximación pudiera ser de la forma $\frac{N!}{(2!(N-2)!)}$ distintos pares de jueces. Entonces cada para producirá un valor para el coeficiente de correlación por rangos ordenados de Spearman. Sea $ave(R_s)$ el average para todos estos coeficientes de correlación de Spearman. Si no existen ligaduras en los datos se puede probar [20] que:

$$ave(R_S) = \frac{NW - 1}{N - 1} \tag{1.24}$$

Por lo tanto, el average del coeficiente de correlación por rangos ordenados de Spearman está linealmente relacionado con el coeficiente de concordancia W de Kendall, y se tiene una manera natural de extender el concepto de correlación a partir de una medida de asociación entre dos jueces a una asociación entre uno y varios jueces [13].

1.4 Técnicas de segmentación: CHAID

En un estudio real existen frecuentemente 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 inútil de tablas que desinforman en lugar de orientar, aún cuando se utilice la V de Cramer para ordenar la fortaleza de las asociaciones. Un estudio multivariado trata de enfocar el efecto posible de todas las variables conjuntamente incluyendo sus posibles correlaciones; pero puede ser particularmente 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. CHAID es eso: sus siglas significan Chi-squared Automatic Interaction Detector [21],

El análisis de CHAID surge realmente como una técnica de segmentación. Es muy útil en todos aquellos problemas en que se quiera subdividir una población a partir de una variable dependiente y posibles variables predictivas que cambien los valores de la variable dependiente en cada una de las subpoblaciones o segmentos. Ejemplos típicos asociados con su origen son los problemas de estudio de mercado. En estos casos la variable dependiente puede ser la aceptación o no de un producto y las variables predictivas un conjunto de características psico o socio económicas de la población que pueden influir en esta aceptación o no. La técnica de CHAID es capaz de segmentar la población en grupos de acuerdo con determinados valores de esas variables y sus interacciones que distinguen de forma óptima, diferencias esenciales en el comportamiento de la variable dependiente [22].

Desde esta formulación inicial, se concibió la posibilidad de aplicación a diversas investigaciones como pudiera ser en la salud. La más típica de ellas, es precisamente en epidemiología, en el estudio de los factores de riesgo asociados a una enfermedad. En tal caso, la variable dependiente puede ser simplemente la variable que distingue un grupo de enfermos y sanos y las variables predictivas los posibles factores de riesgo [23].

Más que segmentar la población en este caso la técnica de CHAID se usa en este caso para:

- Para conocer cuáles, entre decenas de variables (posibles factores de riesgo) pueden ser eliminadas.
- Para comprender el orden de importancia de los factores de riesgo en la caracterización de la enfermedad y en particular ayudar a detectar posibles factores confusores o modificadores de riesgo.
- Para entender cómo ciertos factores de riesgo interactúan con otros.
- Para conocer que efectos interactivos incluir en un análisis discriminante o de regresión logística de casos-controles respecto a factores de riesgo.
- Para buscar entre cientos de tablas de contingencia y seleccionar aquellas que son más significativas estadísticamente.
- Simplificar las cross tabulaciones combinando categorías de variables predictoras que no difieren significativamente.

Esto último es una de las cosas más interesantes. CHAID combina categorías de una variable predictora que no difieren significativamente. De esta forma se resuelve por ejemplo el problema de como ranquear la edad (considerada como un posible factor de riesgo) para obtener una tabla de contingencia significativa con la enfermedad. Simplemente si se ranquea en 10 o 12 categorías, CHAID se ocupará de unir las categorías consecutivas que no difieren significativamente y el resultado final mostrará muchos menos rangos de edades, evidenciando las que constituyen verdadero factor de riesgo. Lo mismo es capaz de hacer con variables incluso nominales (por ejemplo la raza o color de la piel), e incluso con variables que tienen un valor perdido, asociando éste a la categoría de la variable respecto a la cual los casos son más parecidos en su comportamiento. Este procedimiento de combinación conjuntamente con el algoritmo de ruptura o división, asegura en un mismo segmento a aquellos casos que son homogéneos respecto al criterio de segmentación.

Pero además el análisis de CHAID tiene otras aplicaciones importantes en salud. En particular permite:

- Construir una escala cuantitativa de una variable ordinal, situación típica que se presenta en los procesos por ejemplo de elaboración de tests psicométricos u otras pruebas médicas basadas en criterios cualitativos.
- Elaborar criterios diagnósticos, utilizando en este caso como variables predictoras, posibles síntomas, en lugar de factores de riesgo.
- Someter a validación resultados de ensayos clínicos [21].

Un análisis de CHAID comienza dividiendo la población en dos o más grupos distintos basado en las categorías del mejor predictor. Divide cada uno de estos grupos en pequeños subgrupos. CHAID visualiza los resultados de la segmentación en forma de un diagrama árbol cuyas ramas (nodos) corresponden a los grupos. Como cada uno de esos grupos se divide además en pequeños subgrupos, el árbol produce nuevos nodos. Entiéndase en este caso que está seleccionando sucesivamente los factores de riesgo más significativamente asociados con la enfermedad y los factores que deben ser fuentes de estratificaciones sucesivas. En cualquier punto en un análisis de CHAID, el árbol muestra el estado actual del análisis.

CHAID divide la población en grupos excluyentes y exhaustivos. Cada individuo de la población queda en uno y un solo grupo. A diferencia de otras técnicas de agrupación, como las técnicas de *Clustering*, solo CHAID utiliza una variable dependiente como criterio para la formación de subgrupos (la significación estadística entre la variable dependiente y los

predictores es lo que se maneja en el algoritmo de segmentación de CHAID). Esto es, mientras que los segmentos de CHAID se obtienen para predecir una variable dependiente, los clusters no tienen por qué ser predictivos.

1.5 Regresión Logística

La regresión logística es un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia hemos puntuado con los valores cero y uno, respectivamente) y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas (que se denominan covariables o covariadas) o categóricas. En este último caso, se requiere que sean transformadas en variables "dummy", es decir variables simuladas.

El propósito del análisis consiste en predecir la probabilidad de que a alguien le ocurra cierto "evento" (estar desempleado =1 o no estarlo = 0, ser pobre = 1 o no pobre = 0). Determinar que variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda el evento en cuestión. Esta asignación de probabilidad de ocurrencia del evento a un cierto sujeto, así como la determinación del peso que cada una de las variables dependientes en esta probabilidad, se basan en las características que presentan los sujetos a los que, efectivamente, les ocurren o no estos sucesos.

Por ejemplo, la regresión logística tomará en cuenta los valores que asumen en una serie de variables (edad, sexo, nivel educativo, posición en el hogar, origen migratorio, etc.) los sujetos que están efectivamente desocupados (=1) y los que no lo están (=0). En base a ello, predecirá a cada uno de los sujetos, independientemente de su estado real y actual, una determinada probabilidad de ser desocupado (es decir, de tener valor 1 en la variable dependiente). Si alguien es un joven no jefe de hogar, con baja educación y de sexo masculino y origen emigrante (aunque esté ocupado) el modelo le predecirá una alta probabilidad de estar desocupado (puesto que la tasa de desempleo de el grupo así definido es alta), generando una variable con esas probabilidades estimadas, y procederá a clasificarlo como desocupado en una nueva variable, que será el resultado de la predicción. Y además, analizará cuál es el peso de cada uno de estas variables independientes en el aumento o la disminución de esa probabilidad. Por ejemplo, cuando aumenta la educación disminuirá en algo la probabilidad de ser desocupados. En cambio, cuando el sexo pase de 0 = mujer a 1 = varón, aumentará en algo la probabilidad de desempleo porque la tasa de desempleo de los jóvenes de sexo masculino es mayor que la de las jóvenes mujeres. El modelo, obviamente, estima los coeficientes de tales cambios [24].

Por sus características, los modelos de regresión logística permiten dos finalidades:

- Cuantificar la importancia existente entre cada una de las covariables y la variable independiente, lo que lleva implícito también clarificar la existencia de interacción y confusión entre covariables respecto a la variable dependiente(es decir, conocer los odds ratio[1] para cada covariable)

- Clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas covariables.

La ecuación de partida en los modelos de regresión logística es como sigue:

$$P(y=1|x)=\frac{e^{b_0+\sum_{i=1}^{n} b_i x_i}}{1+e^{b_0+\sum_{i=1}^{n} b_i x_i}} \qquad (1.25)$$

… siendo $P(y=1|x)$ la probabilidad de que y tome el valor 1 (presencia de la característica estudiada), en presencia de las covariables X (aquí X es un conjunto de n covariables $x_1, x_2, \ldots, x_{n-1}, x_n$). Los componentes de esta ecuación son:

1. b_0 es la constante del modelo o término independiente
2. n el número de covariables
3. b_i los coeficientes de las covariables
4. x_i las covariables que forman parte del modelo.

Si se divide la expresión anterior por su complementario, es decir, si se construye su odds (en el ejemplo de presencia o no de enfermedad, la probabilidad de estar enfermo entre la probabilidad de estar sano), se obtiene una expresión de más fácil manejo matemático:

$$\frac{P(y-1|X)}{1-P(y=1|X)}=e^{b_0+\sum_{i=1}^{n} b_i x_i} \qquad (1.26)$$

[1] traducida al castellano con múltiples nombres como: razón de productos cruzados, razón de disparidad, razón de predominio, proporción de desigualdades, razón de oposiciones, oposición de probabilidades contrarias, cociente de probabilidades relativas, oportunidad relativa.

Esta expresión aún es de difícil interpretación. Si ahora se realiza su transformación logarítmica con el logaritmo natural, se obtiene una ecuación lineal que es lógicamente de manejo matemático aún más fácil y de mayor comprensión:

$$Log\left(\frac{P(y=1|X)}{1-P(y=1|X)}\right)=b_0+\sum_{i=1}^{n}b_i x_i \qquad (1.27)$$

En esta última expresión se observa a la izquierda de la igualdad el llamado logit, es decir, el logaritmo natural de la odds de la variable dependiente (esto es, el logaritmo de la razón de proporciones de enfermar, de fallecer, de éxito, etc.…). El término a la derecha de la igualdad es la expresión de una recta, idéntica a la del modelo general de regresión lineal:

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_n \cdot x_n \qquad (1.28)$$

Pero la regresión lineal presenta una diferencia fundamental respecto al modelo de regresión logística. En el modelo de regresión lineal se asume que los errores estándar de cada coeficiente siguen una distribución normal de media 0 y varianza constante (homoscedasticidad). En el caso del modelo de regresión logística no pueden realizarse estas asunciones pues la variable dependiente no es continua (sólo puede tomar dos valores, 0 ó 1, pero ningún valor intermedio). Si se llama ε al posible error de predicción para cada covariable x_i, se tiene que el error cometido dependerá del valor que llegue a tomar la variable dependiente y, tal como se aprecia:

$$\begin{aligned} &y = P(x) + \varepsilon \\ &\text{si} \quad y=1 \Rightarrow \varepsilon = 1 - P(x) \\ &\text{si} \quad y=0 \Rightarrow \varepsilon = -P(x) \end{aligned} \qquad (1.29)$$

Esto implica que ε sigue una distribución binomial, con media y varianza proporcionales al tamaño muestral y a $P(y=1|x_i)$ (la probabilidad de que $y=1$ dada la presencia de x_i)

Para la estimación de los coeficientes del modelo y de sus errores estándar se recurre al cálculo de estimaciones de máxima verosimilitud, es decir, estimaciones que maximicen la probabilidad de obtener los valores de la variable dependiente Y proporcionados por los datos de nuestra muestra. Estas estimaciones no son de cálculo directo, como ocurre en el caso de las estimaciones

de los coeficientes de regresión de la regresión lineal múltiple por el método de los mínimos cuadrados. Para el cálculo de estimaciones máximo-verosímiles se recurre a métodos iterativos, como el Método de Newton-Raphson. Dado que el cálculo es complejo, normalmente hay que recurrir al uso de rutinas de programación o a paquetes estadísticos. De estos métodos surgen no sólo las estimaciones de los coeficientes de regresión, sino también de sus errores estándar y de las covarianzas entre las covariables del modelo [25] .

1.5.1 Regresión Logística Multinomial. Modelos Logísticos para Respuestas Ordinales

La opción Regresión logística multinomial resulta útil en aquellas situaciones en las que desee poder clasificar a los sujetos según los valores de un conjunto de variables predictoras. Este tipo de regresión es similar a la regresión logística, pero más general, ya que la variable dependiente no está restringida a dos categorías.

En la literatura [1] se ha reportado los beneficios del uso de la ordinalidad de una variable haciendo énfasis en la inferencia en un solo parámetro. Estos beneficios se extendieron a los modelos para respuestas ordinales. Los modelos con términos que reflejan las características ordinales tales como la tendencia monótona han mejorado la parsimonia del modelo y la potencia. En este epígrafe se introduce el más popular modelo logístico para respuestas ordinales.

1.5.1.1 Logits acumulativos

Los logits acumulativos de probabilidad pueden usarse utilizando el ordenamiento categórico,

$$P(Y \le j|x) = \pi_1(x) + \ldots + \pi_j(x),\ j = 1,\ldots,J \tag{1.30}$$

Los logits acumulativos son definidos como:

$$\begin{aligned} \operatorname{logit}\left[P(Y \le j|x)\right] &= \log \frac{P(Y \le j|x)}{1 - P(Y \le j|x)} \\ &= \log \frac{\pi_1(x) + \ldots + \pi_j(x)}{\pi_{j+1}(x) + \ldots + \pi_J(x)},\ j = 1,\ldots,J-1 \end{aligned} \tag{1.31}$$

Cada logit acumulativo usa categorías con respuestas *J*.

Un modelo para el logit $\left[P(Y \le j)\right]$ solo, es un modelo logístico ordinario para una respuesta binaria en el cual las categorías desde 1 hasta j forman un resultado y las categorías entre $j+1$ y J el otro. Mejor aún, los modelos pueden usar los logits acumulativos $J-1$ en un único modelo parsimonioso.

1.5.1.2 Modelo odds proporcional

Un modelo que usa simultáneamente todos los logits acumulativos es:

$$\text{logit}\left[P\left(Y \le j|x\right)\right] = \alpha_j + \beta' x,\ j = 1,\ldots,J-1 \qquad (1.32)$$

Cada logit acumulativo tiene su propio intercepto. Los $\{\alpha_j\}$ aumentan en *j*, ya que $P\left(Y \le j|x\right)$ se incrementa en *j* para una *x* fijada, y el logit es una función creciente de esta probabilidad. Este modelo tiene el mismo efecto β para cada logit.

A un odds ratio de probabilidad acumulativa se le denomina odds radio acumulativo. Los odds que tienen respuesta menor que *j* en $x = x_1$ son $\exp\left[\beta'\left(x_1 - x_2\right)\right]$ veces el odds en $x = x_2$. El logaritmo del odds radio acumulativo es proporcional a la distancia entre x_1 y x_2. La misma constante de proporcionalidad se aplica a cada logit. Por esta propiedad se denomina a (1.32) modelo de odds proporcional [26]. Con un único predictor, el odds radio acumulativo se iguala a e^{β} siempre que $x_1 - x_2 = 1$.

El modelo (1.32) limita las curvas de respuesta $J-1$ y hace que tengan la misma forma. Por tanto, su ajuste no es el mismo que el modelo logístico para cada j. Sea $\left(y_{i1},\ldots,y_{ij}\right)$ un indicador binario de la respuesta para el sujeto i. La función de verosimilitud

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{J}\pi_j\left(x_i\right)^{y_{ij}}\right] = \prod_{i=1}^{n}\left[\prod_{j=1}^{J}\left(P\left(Y \le j|x_i\right) - P\left(Y \le j-1|x_i\right)\right)^{y_{ij}}\right] \qquad (1.33)$$

es vista como una función de $\left(\{a_j, \beta\}\right)$.

Los modelos logísticos acumulativos más complejos son formulados como en la regresión logística ordinaria. Ellos simplemente requieren un conjunto de parámetros de intersección en vez de uno solo. Los modelos en este epígrafe utilizan los odds proporcionales con el mismo efecto que los diferentes logits acumulativos. Una ventaja es que los efectos son solo para resumir e interpretar, requiriendo solamente un único parámetro para cada predictor. Los modelos se generalizan para incluir los efectos separados, sustituyendo los β en (1.32) por β_j. Esto implica un no paralelismo en las curvas para los diferentes logits. Sin embargo, las curvas para las diferentes probabilidades acumulativas se cortan para algún valor de x. Tales modelos violan el orden apropiado entre las probabilidades acumulativas.

Incluso si tales modelos ajustan mejor en el rango observado de x, por razones de parsimonia el modelo simple pudiera preferirse. Un caso es cuando los efectos $\left\{\hat{\beta}_j\right\}$ con diferentes logits no son sustancialmente diferentes en términos prácticos. Entonces la significación en un test de odds proporcionales pudiera reflejar en primer lugar un valor grande de n. Incluso para valores pequeños de n, aunque el estimador del efecto usado en el modelo simple sea sesgado, pudiera tener menor MSE que el estimador de un modelo más complejo que tiene más parámetros. Por tanto si un test de odds proporcionales tiene valores *p* pequeños, no se descarta el modelo automáticamente.

Si el modelo odds proporcional ajusta poco en términos prácticos en términos de significación estadística, existen estrategias alternativas. Pudieran ser:

- Tratar la función enlace para las cual la curva respuesta no es simétrica.
- Añadir términos adicionales, como las interacciones, al predictor lineal.
- Adicionar parámetros de dispersión
- Permitir la separación de los efectos para algún logits pero no para todos los predictores [1].

Consideraciones Finales del Capítulo

Este capítulo muestra, de manera abreviada, los métodos más importantes de análisis de datos categóricos reportados en la literatura en las últimas décadas [1].

Se muestran los fundamentos matemáticos del test chi cuadrado de independencia que se utiliza en las Tablas de Contingencia. Se explican las tres variantes que existen para el cálculo de su significación: asintótica, exacta y por métodos de Monte Carlo y se exponen algunas consideraciones y recomendaciones sobre su aplicación que pueden servir de guía a un lector no experto en el tema.

El tema de los coeficientes de correlación lineal también se aborda, haciéndose énfasis en los tests de Spearman y de Kendall. Ellos representan las alternativas no paramétricas del coeficiente de correlación de Pearson más frecuentemente usadas en presencia de variables categóricas.

A continuación se presentan dos técnicas multivariadas: los árboles de decisión basados en pruebas chi cuadrado (CHAID), como una generalización del uso de múltiples tablas de

contingencia y la regresión logística. Esta última constituye un importante método que permite realizar ajustes no lineales en presencia de variables categóricas.

Capítulo 2. Análisis de Componentes Principales y Análisis de Regresión para datos categóricos

La tecnología informática disponible hoy en día, casi inimaginable hace solo dos décadas, ha hecho posibles avances extraordinarios en el análisis de datos psicológicos [27], sociológicos y de otro tipo de datos referidos al comportamiento humano [2]. Además ha permitido enorme avances en el procesamiento de los datos en otras ramas de las ciencias como en la medicina [28]. Estos datos contienen una mixtura de diferentes tipos de variables, muchas de las cuales están medidas en categorías ordenadas y desordenadas. Variables como género, religión o profesión son medidas en categorías desordenadas. Otras como el nivel de escolaridad son medidas en categorías ordenadas. O simplemente se tienen conjuntos de variables como la edad o la longitud que son de tipo numérico. Estos diversos tipos de variables requieren distintos tipos de tratamientos que no siempre resultan tan evidentes a la hora del análisis de los datos. Además estos conjuntos de datos en ocasiones contienen variables que pueden o no estar relacionadas linealmente, lo cual debe tenerse en cuenta en el análisis. Por tanto estos conjuntos de datos presentes en varias ramas del saber no pueden siempre ser analizados de manera tan sencilla como muchos investigadores desearían [29]. En el presente capítulo se ofrece una solución a algunas de las cuestiones expuestas anteriormente haciendo uso de dos técnicas de análisis que son parte de lo que tradicionalmente se conoce como Análisis de Datos Multivariados Categóricos. Ellas son el Análisis de Componentes Principales y el Análisis de Regresión, ambas para datos categóricos. Se comenzará por la primera técnica.

2.1 Análisis de Componentes Principales Lineal

El análisis de componentes principales (ACP) se ha utilizado de manera creciente en las últimas décadas, prácticamente en todas las áreas. 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 [2].

El análisis de componentes principales realiza dos acciones fundamentales: cuantifica las variables originales y reduce la dimensionalidad de los datos. El investigador puede identificar primero las dimensiones y después determinar el grado en el que cada variable tributa a cada componente o dimensión. Si el análisis realizado fue exitoso, cada variable debe estar muy bien representada (con una correlación elevada) en una dimensión y pobremente representada (con correlaciones bajas) en las demás. Si las componentes pueden interpretarse, los datos se describen a partir de un número de conceptos mucho más reducido que las variables individuales originales [30].

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. Específicamente, los supuestos de que las variables tengan una escala de medición de intervalo y que haya relación lineal entre ellas se violan en numerosas ocasiones. Algunas veces, específicamente cuando las variables seleccionadas se miden en escala ordinal, los investigadores simplemente ignoran esta violación, y desarrollan el análisis de componentes principales independientemente a esto. Esta decisión puede o no conducir a un problema sustancial, dependiendo si estas variables tienen aproximadamente una relación lineal. Si el ACP se desarrolla sin chequear estos 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 [29].

2.2 Análisis de 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 [31]. 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 ser manipuladas más efectivamente.

El método de análisis de componentes principales categórico generaliza algunos métodos de análisis más específicos pues permite la manipulación de los datos con diferentes niveles de análisis simultáneamente. Cuando todas las variables tienen un nivel de medición numérico, ACPCat es igual al análisis de componentes principales lineal. Con todas las variables en un nivel nominal múltiple, ACPCat es equivalente al análisis de correspondencia múltiple. Mientras se desarrolla el ACP, ACPCat convierte cada categoría en un valor numérico, en concordancia con el nivel de medición de la variable, usando cuantificaciones óptimas [29].

En ACPCat, a las categorías de algunas variables le son asignados valores numéricos a través de un proceso llamado cuantificaciones óptimas. Por lo que las variables categóricas se transforman y se cuantifican de forma óptima, pudiéndose modelar relaciones no lineales entre ellas [31, 32]. Se dice que se obtiene una buena solución cuando los factores hallados son fácilmente explicables, o sea, cada variable original tiene una correlación alta en una de las dimensiones y baja en las restantes [31]. Tales valores numéricos se conocen como cuantificaciones categóricas. Las cuantificaciones óptimas reemplazan los valores de las

categorías por las cuantificaciones categóricas de tal forma que en la medida de lo posible se represente la varianza en las variables cuantificadas. Así como variables numéricas continuas, tales variables cuantificadas poseen variación en el sentido tradicional. Entonces, ACPCat logra el mismo objetivo que el análisis de componentes principales para variables categóricas cuantificadas. Si todas las variables en el ACPCat son numéricas, las soluciones del análisis de componentes principales tanto lineal como no lineal son exactamente iguales, pues en este caso no se requiere cuantificación óptima, y las variables son simplemente estandarizadas.

En ACPCat la tarea de cuantificación óptima y en el ACP la estimación del modelo se desarrollan simultáneamente, esto se logra minimizando la función de pérdida de mínimos cuadrados. En el actual análisis de componentes principales no lineal, el modelo de estimación y cuantificaciones óptimas se alternan a través del uso del uso de un algoritmo iterativo que converge a un punto estacionario donde las cuantificaciones óptimas de las categorías no cambian más. Si todas las variables se tratan numéricamente, este proceso iterativo conduce a la misma solución que el ACP.

2.2.1 Cuantificaciones Categóricas

Obviamente, cuando las variables consisten en categorías desordenadas (nominales), no tiene sentido calcular la suma o el promedio. Como las componentes principales son sumas ponderadas de las variables originales, las variables nominales no pueden ser analizadas por el método ACP. Algunas variables consisten en categorías ordenadas (ordinal). A pesar de la apariencia superficial, tales valores de la escala no son realmente valores numéricos, ya que los intervalos entre las categorías consecutivas no pueden ser considerados iguales. Aunque las variables ordinales muestran mayor estructura que las variables nominales, todavía tiene poco sentido considerar a las escalas ordinales como que tienen cualidades numéricas tradicionales. Finalmente, incluso ciertas variables numéricas pueden ser vistas como variables categóricas con c categorías, donde c indica el número de los diferentes valores observados.

ACPCat convierte las categorías en valores numéricos, debido a que la varianza solo puede ser establecida para valores numéricos. De manera similar se requiere la cuantificación, debido a que las correlaciones de Pearson se usan en la solución de ACP lineal. En ACPCat, las correlaciones no se calculan entre las variables observadas, pero sí entre las variables cuantificadas. Consecuentemente, en oposición a la matriz de correlación en ACP, la matriz de correlación en ACPCat no es fija, más bien depende del tipo de cuantificación, que se selecciona para cada una de las variables.

En contraste con la solución del ACP, la solución del ACPCat no se obtiene a partir de la matriz de correlación, sino que se calcula en forma iterativa a partir de los propios datos, utilizando el proceso de escalamiento óptimo para cuantificar las variables en relación con su nivel de medición. El objetivo del escalamiento es optimizar las propiedades de la matriz de correlación de las variables cuantificadas. Específicamente, el método maximiza los primeros p valores propios de la matriz de correlación de las variables cuantificadas, donde p indica el número de componentes que se seleccionan en el análisis. Este criterio es equivalente a la declaración previa que el objetivo de la cuantificación óptima es maximizar la varianza explicada en las variables cuantificadas.

Niveles de medición

El análisis de los datos por el ACPCat implica la toma de decisiones de manera dinámica, ya que las decisiones tomadas originalmente por el investigador pudieran necesitar ser revisadas durante el proceso de análisis; de hecho, probar varios niveles de escala y comparar sus resultados es parte de la tarea de análisis de datos. Es muy importante hacer notar que la visión del investigador, y no el nivel de medición de la variable, determina el nivel de escala de la variable [29]. En general, debe mantenerse en la mente que diferentes niveles de medición implican requerimientos diferentes. En el caso del nivel de escala nominal, el único requerimiento es que la persona que anotó la misma categoría en la variable original debe también obtener el mismo valor cuantificado. Este requerimiento es el más débil en el ACPCat. En el caso del nivel de escala ordinal, la cuantificación de las categorías deben respetar además el orden de las categorías originales: una cuantificación categórica debe siempre ser menor o igual que la cuantificación para la categoría que tiene mayor número de rango en los datos originales. Un nivel de escala numérico requiere categorías cuantificadas no solo en el orden correcto, sino también mantener el espacio relativo original de las categorías en las cuantificaciones óptimas, lo cual se logra estandarizando la variable. Si todas las variables son tratadas como numéricas, no se necesitan las cuantificaciones optimas, y simplemente las variables son estandarizadas, en este caso la relación no lineal potencial entre las variables no será explicada. Si se desea explicar la relación no lineal entre las variables numéricas, debe seleccionarse un nivel de escala no numérico.

ACPCat tiene mayor libertad en la cuantificación de la variable cuando se especifica un nivel de escala nominal, y es el más limitado cuando se especifica un nivel de escala numérico. Por lo tanto, el método obtendrá mayor explicación de la varianza cuando todas las variables son analizadas nominalmente, y menor explicación de la varianza cuando todas las variables son analizadas numéricamente [29].

Transformaciones Suaves

Los niveles nominal y ordinal descritos anteriormente usan funciones paso, las cuales pueden ser bastante irregulares. Como una alternativa, es posible usar funciones suaves (spline) para obtener una transformación no lineal. Una transformación spline monótona es menos limitada que una transformación lineal, pero más limitada que una ordinal, ya que no solo requiere que las categorías estén en el mismo orden, sino también que las transformaciones muestren una curva suave. La forma más simple de un spline es una función, por lo general un polinomio de segundo grado (función cuadrática) o un polinomio de tercer grado (función cúbica) de los datos originales, especificando el rango entero de una variable. Debido a que frecuentemente es imposible describir el rango entero de los datos con tal función simple, la separación de las funciones puede especificarse para diferentes intervalos dentro del rango de una variable. Como estas funciones son polinómicas, la suavidad de la función dentro de cada intervalo es garantizada. El punto final del intervalo donde dos funciones se interceptan es denominado nodo interior. El numero de nodos interiores y el grado de los polinomios especifican la forma del spline, y por tanto la suavidad de la transformación. Vale notar que un spline de primer grado con cero nodos interiores da lugar a una transformación lineal, y un spline de primer grado con el número de nodos interiores igual al número de categorías menos dos, da lugar a una transformación ordinal [33].

Niveles de medición y relación no lineal entre las variables

La transformación nominal y la transformación spline no monótona muestran una función creciente seguida por una decreciente describiendo la relación entre las categorías originales y las cuantificaciones. Las transformaciones ordinales y spline monótonas muestran un incremento de las cuantificaciones de las categorías, pero todas las cuantificaciones para valores altos de las categorías son ligados, ya que las cuantificaciones nominales no incrementan con los valores de las categorías como requiere el nivel de escala ordinal. La cuantificación numérica muestra (por definición) una línea recta.

Representación de las variables como vectores

Para todos los niveles de escala descritos, una forma de representar la variable cuantificada es mostrando los puntos de sus categorías en espacio de las componentes principales, donde los ejes están dados por las componentes principales. En este tipo de grafico, una variable se representa por un vector. Un vector variable es una línea recta, que va desde el origen de coordenadas hasta el punto que tiene como coordenadas las componentes de la variable. Los puntos de las categorías se colocan también en un vector variable, y sus coordenadas se hallan multiplicando las

cuantificaciones de las categorías por las correspondientes componentes en la primera (x- coordenada) y segunda (y- coordenada) componente. El orden de los puntos categóricos en el vector variable esta relacionado con las cuantificaciones: el origen representa la media de la variable cuantificada, las categorías con las cuantificaciones por encima de la media pertenecen a las vecindades del origen en el cual los puntos de las componentes están posicionadas, y las categorías con cuantificaciones por debajo de la media en la dirección opuesta, en el otro lado del origen.

La longitud total del vector variable no indica la importancia de la variable. Sin embargo, la longitud del vector variable desde el origen al punto de la componente es una indicación de la varianza total acumulada de la variable, de hecho el cuadrado de la longitud de un vector es igual a la varianza acumulada [29].

Representación de las variables como un conjunto de puntos: Nivel de escala nominal múltiple

En el proceso de cuantificación, cada categoría obtiene una sola cuantificación, o sea, una cuantificación óptima es la misma para todas las componentes. La variable cuantificada puede representarse como una línea recta a través del origen. Sin embargo, la representación de las cuantificaciones categóricas de una variable nominal (o una variable tratada como nominal) en una línea recta no siempre puede ser la más apropiada. Solamente cuando la transformación muestra una forma especifica o un orden particular, es el tipo de representación útil. En otras palabras, se tiene que ser capaz de interpretar la transformación.

Cuando la transformación es irregular, o cuando las categorías originales no pueden estar colocadas en un orden significativo, existe una manera alternativa de cuantificar las variables nominales, denominada cuantificación nominal múltiple. El objetivo de la cuantificación nominal múltiple no es representar una variable en su totalidad, sino lo suficiente para revelar de manera óptima la naturaleza de las relaciones entre las categorías de la variable y otras variables al alcance de la mano. Este objetivo se logra asignando una cuantificación por cada componente por separado.

Las cuantificaciones categóricas múltiples se obtienen promediando, por componentes, los scores de las componentes principales para todos los individuos en la misma categoría de una variable particular. Consecuentemente, tales cuantificaciones serán diferentes para cada componente, de aquí el término de cuantificación múltiple. Gráficamente, las cuantificaciones múltiples son las coordenadas de los puntos de las categorías en el espacio de las componentes principales. Como la variable categórica clasifica a los individuos en grupos mutuamente

excluyentes o clases (categorías), este punto puede ser considerado como la representación de un grupo de individuos. En contraste con las variables con otro nivel de medición, las variables nominales múltiples no obtienen componentes. El ajuste de una variable nominal múltiple en una componente se indica a través de la varianza de las cuantificaciones categórica en dicha componente. Así que, si todas las cuantificaciones están cerca del origen, la variable ajusta mal en la solución. Es importante darse cuenta que solamente se definen las cuantificaciones múltiples para variables con nivel de escala nominal. Las transformaciones spline, ordinales y numéricas siempre se obtienen por una sola cuantificación y pueden representarse como un vector [29].

Representación de los individuos como puntos

Hasta ahora, se ha descrito la representación de las variables en el espacio de las componentes principales, cada una a través de vectores o mediante un conjunto de puntos categóricos. En este epígrafe se direccionará la representación de los individuos en ACPCat. Cada individuo obtiene una puntuación de componente en cada una de las componentes principales. Estas puntuaciones de componentes son puntuaciones estándar que pueden usarse para mostrar los individuos como puntos de personas en el mismo espacio que las variables, indicando relaciones entre los individuos y las variables. En la literatura estadística a esta representación se le denomina "biplot" [34, 35]. Las variables nominales múltiples pueden representarse como un conjunto de puntos categóricos en el espacio de las componentes principales, y estos pueden combinarse con los puntos de los individuos y los vectores de las otras variables en el llamado "triplot" [36]. Cuando los individuos y los puntos de las categorías para las variables nominales múltiples se grafican juntos, un punto de categoría particular estará exactamente en el centro de los individuos que tienen resultados en esta categoría.

2.2.2 Modelo

En este epígrafe 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* personas 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 las personas 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 mostrar cómo esta función se amplía para acomodar las ponderaciones de acuerdo con los valores ausentes, ponderaciones de personas, y trasformaciones nominales múltiples.

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

$$L(Q,X,A)=n^{-1}\sum_{j=1}^{m} tr\left(q_j-Xa_j\right)'\left(q_j-Xa_j\right) \qquad (2.1)$$

…donde tr denota la función traza que suma los elementos de la diagonal de una matriz. Puede probarse que la función (2.1) es equivalente a:

$$L_2(Q, A, X) = n^{-1} \sum_{j=1}^{m} tr\left(q_j a_j' - X\right)' \left(q_j a_j' - X\right) \qquad (2.2)$$

La función de pérdida (2.2) se usa en ACPCat en lugar de (2.1), debido a que en (2.2), 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 (2.2) 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 que q_j 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 son limitados requiriendo

$$X'X = nI, \qquad (2.3)$$

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

$$1'X = 0, \qquad (2.4)$$

… donde el 1 representa la vector unidad. Las restricciones (2.3) y (2.4) 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 (2.2) 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.

La función de pérdida (2.2) se especifica por la simple situación, sin valores perdidos o la posibilidad de diferentes ponderaciones. Sin embargo, las ponderaciones por valores perdidos y las ponderaciones por personas fácilmente pueden incorporarse 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 a la persona *i*, igual a 1 para los valores no ausentes y 0 para los valores ausentes de la variable *j*. Por tanto, para las personas 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 la fila correspondiente a la persona con valores ausentes en la variable *j*. Por tanto, para la variable *j*, las personas con valores perdidos no contribuyen a la solución de ACPCat, sino que contribuyen a la solución de las variables que tienen una puntuación válida. Permitimos ponderación de la persona a través de la ponderación de error por una matriz diagonal $W_{\;n\times n}$ con elementos no negativo w_{ii}. Generalmente estas ponderaciones de personas, son todas igual a uno, donde cada persona contribuye de igual manera a la solución. Para algunos, sin embargo, puede ser conveniente para poder tener diferentes ponderaciones para diferentes personas.

Incorporando las ponderaciones de los datos ausentes M_j y las ponderaciones de las personas *W*, la función de pérdida que se minimiza en ACPCat puede expresarse como $L_3(Q,A,X) = n^{-1}\sum_{j=1}^{m}\sum_{i=1}^{n} w_{ii} m_{iij} \sum_{s=1}^{p}\left(q_{ij} a_{js} - x_{is}\right)^2$, o equivalentemente, en notación matricial como:

$$L_3\left(Q,A,X\right) = n_w^{-1}\sum_{j=1}^{m} tr\left(q_j a'_j - X\right)' M_j W\left(q_j a'_j -\right)X \qquad (2.5)$$

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 $X'M_*WX = mn_w I$.

La función de pérdida (2.5) 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 las personas

medidas 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 numero de filas de G_j es igual al número de personas, *n*, y el número de columnas de G_j es igual al número de las diferentes categorías de la variable *j*. Por cada persona, una columna de G_j contiene un 1 si la persona anotó en una categoría particular, y un cero si las persona no anotó en una categoría. Así, todas las filas de G_j contiene exactamente un 1, excepto cuando los valores ausentes son tratados pasivamente. En el caso de valores ausentes pasivos, cada fila de la matriz indicadora correspondiente a la persona 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

$$L_3(v_1, \ldots, v_m, A, X) = n^{-1} \sum_{j=1}^{m} tr\left(G_j v_j a_j' - X\right)' M_j W\left(G_j v_j a_j' - X\right) \qquad (2.6)$$

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, (2.6) es la misma que (2.5).

La ventaja de (2.6) es que la transformación nominal múltiple puede ser incorporada 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 las personas en *p* dimensiones. Entonces, la función de pérdida puede escribirse como

$$L_4(V_1,\ldots,V_m,X) = n^{-1}\sum_{j=1}^{m} tr\left(G_jV_j - X\right)' M_j W\left(G_jV_j - X\right) \qquad (2.7)$$

… 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 [29].

2.2.3 Datos ausentes en el Análisis de Componentes Principales Categórico

Una relación considerable en la literatura muestran maneras sofisticadas del manejo con datos ausentes [2, 37, 38]. ACPCat proporciona, además de otros tantos criterios conocidos, la forma de hacer frente a este problema (eliminación de valores ausentes y simple imputación), mediante dos métodos que vale la pena describir. El primero, conocido como tratamiento pasivo de los datos ausentes, garantiza que los datos con valores ausentes en una variable no contribuya a la solución de la variable, sino que contribuya a la solución de todas las otras variables. Note que este tipo de tratamiento difiere de la eliminación por pares, en que el último elimina pares de valores en "pairwise computations". El tratamiento pasivo de las perdidas es posible en ACPCat debido a que su solución no se deriva de la matriz de correlación (la cual se calcula con valores ausentes), sino a partir de los datos.

Adicionalmente, ACPCat ofrece la posibilidad del tratamiento de valores ausentes como categorías extra. Esta opción implica que las categorías ausentes obtendrán una cuantificación que es independiente al nivel de medición de la variable. La mayor ventaja de esta opción es que permite al investigador trabajar con variables que incluyen categorías numéricas u ordenadas más categorías como "no respuesta", "no se" o "no aplicable". La opción puede también ser útil si la persona omite algunas preguntas por alguna razón específica que las distingue de otras personas que sí contestaron la pregunta. Cuando la categoría ausente obtiene una cuantificación que se distingue evidentemente de otras categorías, las personas con valores ausentes difieren estructuralmente de las otras (y esto será reflejado en la puntuación de la persona). Si la categoría ausente obtiene una cuantificación cerca de la media (ponderada) de las cuantificaciones, las personas que tienen valores ausentes no pueden considerarse como un grupo homogéneo, y el tratamiento de los datos ausentes como una categoría extra proporcionará aproximadamente los mismos resultados que si se trataran como datos ausentes pasivos [29].

2.2.4 Análisis de Componentes Principales Lineal y Categórica: Semejanzas y Diferencias

El análisis de componentes principales categóricas se ha desarrollado como una alternativa del análisis de componentes principales para la manipulación de las variables categóricas y las relaciones no lineales. La comparación de ambos métodos revela semejanzas y diferencias. Comenzando por las primeras, se puede apreciar que ambos métodos proporcionan valores propios, saturaciones en componentes, puntuaciones por componentes. En ambos, los valores propios son una medida de resumen general que indican la varianza acumulada por cada componente; esto es, cada componente principal puede ser vista como una variable compuesta resumiendo las variables originales, y el valor propio indica el grado de éxito de este resumen. La suma de los valores propios de todas las componentes posibles es igual al número de variables m. Si todas las variables están altamente correlacionadas, una sola componente principal es suficiente para describir a los datos. Si las variables forman dos o más conjuntos, y las correlaciones son altas dentro de los conjuntos y baja entre ellos, una segunda o tercera componente principal se necesitará para resumir las variables. Las soluciones del ACP con más de una componente principal son referenciadas como una solución multidimensional. En dichas soluciones, las componentes principales se ordenan de acuerdo a sus valores propios. La primera componente está asociada al mayor valor propio, y acumula la mayoría de la varianza, la segunda acumula tanto como sea posible la varianza restante, etc. Esto se cumple en ambos métodos.

Las saturaciones en componentes son medidas obtenidas de las variables, y en ambos métodos, son iguales a la correlación de Pearson entre la componente principal y cada variable observada, en el caso de ACP, o a la variable cuantificada, en el caso del ACPCat. De manera similar, la suma de los cuadrados de las saturaciones en componentes sobre las componentes describe la varianza acumulada de las variables observadas (ACP) o las variables cuantificadas (ACPCat). Si existen relaciones no lineales entre las variables, y se especifican los niveles de escala nominal u ordinal, ACPCat conduce a mayor varianza acumulada que su homólogo ACP, debido a que permite transformaciones no lineales. Para ambos métodos, antes de cualquier rotación, la suma de los cuadrados de las saturaciones en componentes de todas las variables en una componente es igual al valor propio asociado a la componente.

Las componentes principales en el ACP son sumas ponderadas (combinaciones lineales) de las variables originales, mientras que en ACPCat son sumas ponderadas pero de las variables cuantificadas. En ambos las componentes son el resultado de puntuaciones estandarizadas. Resumiendo, el análisis de componentes principales y su homólogo categórico son muy similares en cuanto a objetivo, método, resultado e interpretación. La diferencia crucial radica en que el

ACP las variables medidas son directamente analizadas, mientras que el ACPCat las variables medidas son cuantificadas durante el análisis, excepto cuando todas las variables son tratados como numéricas. Otra diferencia consiste en la anidación de la solución.

Anidación de las componentes

Una manera de comprender el ACP es que este maximiza la varianza acumulada de la primera componente mediante las transformaciones lineales de las variables, y entonces maximizar la varianza acumulada de la segunda componente que es ortogonal a la primera, etc. Algunas veces a este proceso se le denomina maximización consecutiva. El proceso de maximización de la varianza acumulada se resume a través de los valores propios y sus sumas en las primeras p componentes. Los valores propios ascienden a cantidades que son iguales a los valores de la matriz de correlación. Otra manera de comprender el ACP es que este maximiza las varianza acumulada en p dimensiones simultáneamente proyectando las variables originales de un espacio m-dimensional en un espacio de componentes p-dimensional. En ACP, la maximización consecutiva de la varianza acumulada en p componentes es similar a la maximización simultánea, y se dice que las soluciones del ACP están anidadas para los diferentes valores de p.

En ACPCat, la maximización consecutiva y simultánea proporcionará diferentes resultados. En nuestra versión del ACPCat, se maximiza la varianza acumulada de las primeras p componentes simultáneamente a través de transformaciones no lineales de las variables. Los valores propios se obtienen a partir de la matriz de correlación entre las variables cuantificadas, y se maximiza la suma de los primeros p valores propios. En este caso, las diferencias entre las componentes de una solución p-dimensional y las primeras p componentes de una solución $p+1$-dimensional son frecuentemente muy pequeñas. Esto, sin embargo el cambio puede ser dramático, por ejemplo, si se trata de representar una estructura bi o tri-dimensional en una sola dimensión. Cuando uno duda si p es la dimensión más apropiada, es aconsejable buscar también la solución con $p+1$ y $p-1$ componentes.

Selección del número apropiado de componentes

En ambos tipos de análisis, el investigador tiene que decidir el número adecuado de componentes que se conservarán en la solución. Uno de los criterios más conocidos para esta decisión es el criterio de la sedimentación [39], el cual trae consigo el gráfico de sedimentación con las componentes identificadas en el eje x y su correspondiente valor propio en el eje y. esperanzadoramente, tales gráficos muestran un quebrado, o un "codo", identificando que la última componente acumula una cantidad considerable de la varianza de los datos. La

localización de dicho codo indica el número apropiado de componentes para ser incluidas en la solución [40].

Desafortunadamente, tales codos no siempre están discernibles de manera fácil en el gráfico de la sedimentación del ACP. En ACPCat, por otra parte, el hecho que la suma de los primeros *p* valores propios sea maximizado automáticamente implica que la suma de los *m*-*p* valores propios residuos sea minimizado, ya que la suma de los valores propios sobre todas las posibles componentes en el ACP permanece igual a *m* (número de variables en el análisis). Por lo tanto, el codo en el gráfico de la sedimentación del ACPCat, los cuales están basados en los valores propios de la matriz de correlación de las variables cuantificadas, puede ser más claro que en el ACP. Como las soluciones del ACPCat no están anidadas, el gráfico de la sedimentación será diferente para las distintas dimensiones, y el gráfico de la sedimentación de las soluciones *p*, *p*-1 y *p*+1-dimensionales tienen que ser comparadas. Cuando el codo está continuamente en la componente *p* o *p*+1, puede seleccionarse la solución *p*-dimensional. Existe en la literatura algunas discusiones tales como si la componente donde el codo está localizado tiene que ser incluida o no en la solución [40]. Una razón para no incluirla es que esta contribuye poco al total de la varianza acumulada. Si se selecciona un número de componentes diferente de *p*, el ACPCat tiene que ser ejecutado nuevamente con el número seleccionado de componentes, ya que las componentes no están anidadas.

Rotación

Las soluciones del ACP pueden ser rotadas libremente, sin que cambie su ajuste [40]. Un ejemplo familiar es la rotación ortogonal de la solución de manera que la carga de cada variable sea tan alta como sea posible en solo una de los dos componentes, y por tanto se simplifique la estructura (varimax). En una simple estructura, patrones similares a las saturaciones en componentes pueden ser discernidas de manera más fácil. Las variables con patrones comparables de las saturaciones en componentes pueden ser consideradas como un subconjunto, permitiendo una interpretación más sencilla. En ACPCat, la rotación ortogonal puede ser aplicada exactamente de la misma manera. Sin embargo vale notar que después de la rotación el orden de la varianza acumulada de las componentes puede perderse.

2.3 Relación del método ACPCat con otras Técnicas Estadísticas

El análisis de componentes principales se diferencia de otras técnicas multivariadas como la regresión lineal múltiple, la regresión multinomial, el análisis discriminante y los árboles de

decisión entre otras, en el sentido en que todas ellas consideran explícitamente alguna variable dependiente o de criterio, mientras que el resto son variables independientes o predictoras [2].

El método de las componentes principales es una técnicas de interdependencia que considera a todas las variables por igual: no hay dependientes e independientes, predictivas o de criterio. Las dimensiones se forman para maximizar la explicación de todas las variables y no para predecir los valores de una (la dependiente) [2] .

Existen sin embargo, otros métodos estadísticos con las que el análisis de componentes principales categóricos sí tiene relación. Resulta obvio que, si todas las variables analizadas son continuas, es decir tienen nivel de medición numérico, el análisis de componentes principales categóricos se corresponderá con el análisis de componentes principales clásico, como ya se mencionó con anterioridad. Si todas las variables tienen un nivel de escalamiento nominal múltiple, entonces el análisis de componentes principales categórico es idéntico al análisis de correspondencias múltiple [31].

En muchas ocasiones el objetivo del análisis de datos no solo tiene fines descriptivos por lo que se precisa de un análisis más profundo. En este epígrafe se pretende mostrar los fundamentos del Análisis de regresión para datos categóricos.

2.4 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 [14]. 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 [41].

En numerosas investigaciones, sobre todo en el campo médico o social [27], 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 complica notablemente.

Por otra parte, no existen dudas de que la regresión lineal múltiple es la técnica estadística más utilizada para predecir el comportamiento de una variable dependiente, a partir de los valores de varias independientes. Lo que ocurre es que no siempre tal relación es lineal. A través de los años se han reportado en la literatura numerosas contribuciones que son en esencia, generalizaciones no lineales de la regresión [42, 43]. Puede mencionarse por ejemplo, el desarrollo reciente de varios métodos de regresión no lineal en el área de minería de datos, que es una rama de la ciencia de la computación. A esas técnicas se les conoce por el nombre en inglés de "machine learning" o aprendizaje automatizado. En aras de obtener una definición más cercana al ambiente estadístico, ha surgido el término "statistical learning" (aprendizaje estadístico) para referenciar a métodos como estos [44].

Los siguientes epígrafes se centrarán en un método particular para realizar estudios de regresión lineal múltiple, para datos categóricos que se relacionan mediante transformaciones no lineales en sus categorías.

2.5 Análisis de Regresión Categórica

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 [45]. 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.

RegCat extiende la regresión lineal ordinaria, considerando simultáneamente variables continuas, ordinales y nominales. Las variables categóricas se cuantifican de manera que ellas reflejen las características de las categorías originales, utilizando transformaciones no lineales para hallar el modelo que mejor ajuste. Finalmente las variables cuantificadas se tratan de la misma forma que las variables continuas [46].

El objetivo fundamental de la regresión categórica con escalamiento óptimo consiste en describir las relaciones entre una variable respuesta y un conjunto de variables predictoras [32]. El escalamiento óptimo es un método para encontrar valores numéricos óptimos que reemplazan los valores de las categorías, por lo tanto transforma los datos categóricos en datos numéricos. En

la terminología del escalamiento óptimo, a este proceso, se le denomina "cuantificación". Las transformaciones de las variables categóricas se estiman simultáneamente con la estimación de los coeficientes de la regresión, usando una alternativa del procedimiento de los mínimos cuadrados que maximiza el cuadrado del coeficiente de regresión múltiple, para la regresión lineal en las variables transformadas. Como resultado de estos criterios de optimización, las transformaciones de escalamiento óptimo linealizan la relación entre la respuesta y los predictores. Entonces, el método RegCat resulta en variables categóricas transformadas que tienen valores con propiedades numéricas óptimas para describir la relación entre la respuesta y los predictores. Las cuantificaciones de las variables categóricas por lo general resultan una transformación no lineal, que puede ser no monótona o por la aplicación de alguna restricción, monótona o lineal. Algunas restricciones se especifican seleccionando un nivel de escalamiento óptimo. En la metodología de escalamiento óptimo, las variables numéricas se tratan como variables categóricas, con el número de categorías igual al número de los diferentes valores de la variable. Seleccionando el nivel de escalamiento numérico, para una variable numérica se obtiene una transformación lineal. Incluyendo transformaciones lineales, RegCat puede también aplicarse a datos que contienen variables numéricas. Una variable numérica puede también ser no linealmente transformada, en este caso no se respetará el espacio relativo de los valores de las categorías. Luego, el escalamiento óptimo es aplicable a ambas variables categóricas (para cuantificar) y para variables numéricas (para transformaciones no lineales) [47].

El propósito de RegCat es en esencia, el mismo que cualquier otro análisis de regresión, lo interesante es que ella puede aplicarse para aquellas variables, en las que los análisis clásicos de regresión fallan. De hecho, si se realiza un análisis de regresión sobre las variables transformadas, se obtienen los mismos resultados que con el análisis de regresión categórica.

2.5.1 Niveles de escalamiento óptimo

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.

Las propiedades de los datos que se distinguen en el enfoque de la regresión categórica son las de grupos, orden e igual espacio relativo. En dependencia del nivel de medición (nominal, ordinal o intervalo) las variables tendrán una, dos o todas estas propiedades.

- Las variables con nivel de medición nominal solamente tiene propiedades de agrupación, esto es, los valores de las categorías solamente sirven para codificar las observaciones en clases.
- Las variables ordinales tienen propiedades de agrupación y orden.
- Las variables con nivel de medición de intervalo (numéricas) tienen todas las propiedades.

Si el investigador desea preservar todas las propiedades de medición de la variable en las variables cuantificadas, el nivel de escalamiento debe seleccionarse en concordancia con el nivel de medición de la variable.

Con nivel nominal, solo se preserva la propiedad de agrupación, el nivel de escalamiento ordinal preserva la agrupación y el orden, y el nivel de escalamiento numérico preserva la agrupación, el orden e igual espacio relativo. Seleccionando el nivel de escalamiento numérico para una variable medida categóricamente implica que en el análisis los valores categóricos se tratan como valores numéricos (y cuando todas las variables se tratan numéricamente RegCat es equivalente a la regresión lineal estándar). La forma de la curva, cuando se grafican los valores cuantificados contra los valores de las categorías, está relacionada con el nivel de escalamiento: con nivel de escalamiento nominal la curva de transformación puede descender debido a que el ordenamiento de los valores cuantificados no necesitan ser el mismo que el de los valores de la categoría original. Para el nivel de escalamiento ordinal, el ordenamiento de los valores cuantificados y de los valores de la categoría original es el mismo, resultando una curva de transformación monótona. El nivel de escalamiento numérico resulta una línea recta, debido a que los intervalos entre las cuantificaciones por categorías consecutivas son proporcionales a los intervalos entre los valores de categoría.

El nivel de escalamiento, y por tanto la forma de la curva de transformación, esta 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, las cuales 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 [47].

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.

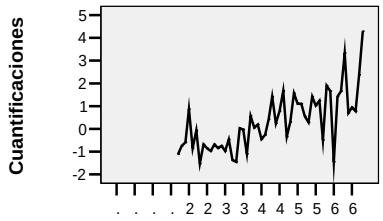
Lo expresado en el párrafo anterior se ilustra en las figuras (2.1 a 2.7) que se muestran a continuación, las cuales muestran la gráfica de transformación de la variable dependiente Diagnóstico de Expertos (DiagExp), que tiene tres categorías: (1-normotenso, 2-hiperreactivo, 3-hipertenso) y la variable independiente categórica Edad de los Pacientes (Edad). 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 2.1). En el mismo se puede apreciar que ambas variables que a medida que se incrementan alcanzan valores máximos. El R^2 que 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 2.2), mucho más si se tienen dos nodos interiores (Figura 2.3). 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 2.4). 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 2.5) y con 2 grados y 2 nodos interiores la transformación es casi lineal (Figura 2.6). Los valores de los R^2 en estos casos 0.085 y 0.078 [47]. En la figura 2.7 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.

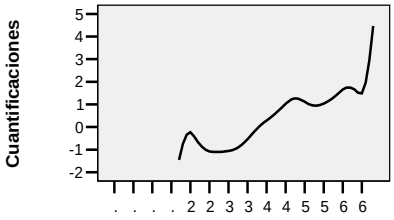
Figura 2.1 Transformacion: Edad



Categorías R Cuad = 0.128

Nominal.

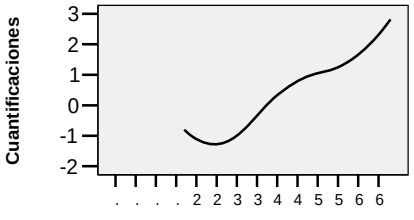
Figura 2.2 Transformacion: Edad



Categorías R Cuad = 0.088

Spline Nominal (degree 2, interior knots 10).

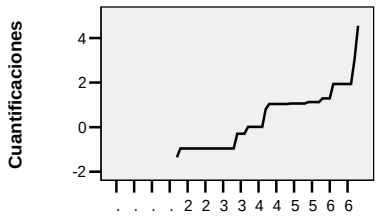
Figura 2.3 Transformacion: Edad



Categorías R Cuad = 0.081

Spline Nominal (degree 2, interior knots 2).

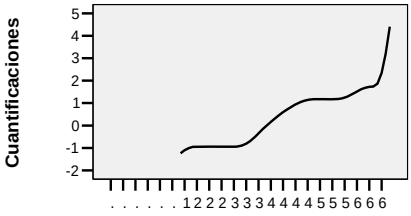
Figura 2.4 Transformacion: Edad



Categorías R Cuad = 0.094

Ordinal.

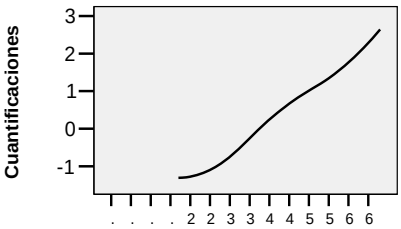
Figura 2.5 Transformacion: Edad



Categorías R Cuad = 0.085

Spline Ordinal (degree 2, interior knots 10).

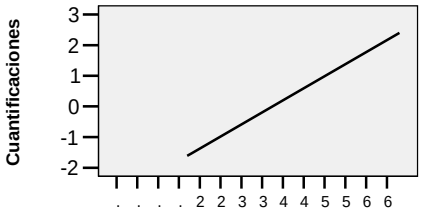
Figura 2.6 Transformacion: Edad



Categorías R Cuad = 0.078

Spline Ordinal (degree 2, interior knots 2).

Figura 2.7 Transformacion: Edad



Categorías R Cuad = 0.073

Numérico.

2.5.2 Estimación de las Transformaciones

En el método de regresión categórica, el modelo de regresión y las cuantificaciones se estiman simultáneamente en un proceso iterativo usando los *mínimos cuadrados alternantes*. El algoritmo alterna entre la estimación de la transformación de la variable respuesta y la estimación de las transformaciones y regresión ponderada de las variables predictoras. La transformación de la respuesta en una iteración se estima a partir de la combinación lineal de los predictores transformados desde las iteraciones previas.

Las cuantificaciones nominales son el punto de partida (y el punto final si el nivel de escalamiento es nominal) en la estimación de las cuantificaciones restringidas. La cuantificación nominal para una categoría es la media de los valores predictores de la categoría cuando se estima la respuesta y la media de los residuos parciales de las categorías cuando se estima el predictor. Si el nivel de escalamiento no es nominal, estas cuantificaciones se restringen según sea el nivel de escala. La restricción se impone aplicando la regresión ponderada (ponderando con las frecuencias de las categorías) de las cuantificaciones nominales, en los valores de las categorías para el nivel de escalamiento ordinal y numérica, y en I-spline base [33] para las transformaciones spline, con restricciones no negativas para los splines monótonos. Para el nivel de escalamiento ordinal, se usa la regresión monótona ponderada, la cual se reduce al promedio ponderado de las cuantificaciones nominales de las categorías que están en el orden equivocado. Con nivel de escalamiento numérico, los valores de las categorías se convierten en scores estándar, lo cual es equivalente a la regresión lineal ponderada de las cuantificaciones nominales en los valores de las categorías. Finalmente, la variable cuantificada se normaliza, y se estima el coeficiente de regresión para una variable predictora. En el método RegCat una transformación monótona es siempre creciente con los valores de las categorías. Si el nivel de escalamiento de un predictor es ordinal o spline monótono, y la relación con la respuesta (después de quitar la influencia de otros predictores) es decreciente de manera monótona, entonces el coeficiente de regresión será negativo [47].

2.5.3 Modelo

La regresión lineal múltiple es una técnica que estudia la relación lineal entre la variable respuesta y un conjunto de variables predictoras. 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:

$$\varphi_r(y) = \sum_{j=1}^{J} \beta_j \varphi_j(x_j) + e \tag{2.8}$$

con la función de pérdida:

$$L(\varphi_r, \varphi_1, \ldots, \varphi_j; \beta_1, \ldots \beta_j) = \left\| \varphi_r(y) - \sum_{j=1}^{J} \beta_j \varphi_j(x_j) \right\|^2 \tag{2.9}$$

… 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 e el vector error.

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.

2.5.4 Algoritmo

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:

$$\varphi_r(y) = G_r v_r \ \wedge \ \varphi_j(x_j) = G_j v_j \tag{2.10}$$

… donde v_r es el vector de las categorías cuantificaciones de la variable respuesta, y v_j el vector de categorías cuantificaciones para una variable predictora. Luego, el modelo de RegCat con las variables transformadas escrito en términos de matrices indicadoras y categorías cuantificadas es:

$$G_r v_r = \sum_{j=1}^{J} \beta_j G_j v_j + e \tag{2.11}$$

Con la función de pérdida mínimos cuadrados asociada:

$$L(v_r; v_1,\ldots,v_j; \beta_1,\ldots\beta_j) = \left\| G_r v_r - \sum_{j=1}^{J} \beta_j G_j v_j \right\|^2 \tag{2.12}$$

La función de pérdida (2.12) 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:

$$\tilde{v}_r = D_r^{-1} G_r' \sum_{j=1}^{J} \beta_j G_j v_j \qquad (2.13)$$

… 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:

$$v_r^+ = N^{1/2} v_r^* \left(v_r^{*\prime} D_r v_r^* \right)^{-1/2} \qquad (2.14)$$

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:

$$z = \sum_{j=1}^{J} \beta_j G_j v_j \qquad (2.15)$$

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:

$$z_j = z - \beta_j G_j v_j \qquad (2.16)$$

Las cuantificaciones no restringidas se actualizan de la manera siguiente:

$$\tilde{v}_j = sign(\beta_j) D_j^{-1} G_j' \left(G_r v_r^+ - z_j \right) \qquad (2.17)$$

Para variables con nivel de escalamiento no numérico $\tilde{v}_j$ se restringe según sea el nivel de escalamiento, y normalizada como en (2.14), 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:

$$\beta_j^+ = N^{-1}\tilde{v}_j' D_j v_j^+ \qquad (2.18)$$

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

$$z = z_j + \beta_j^+ G_j v_j^+, \qquad (2.19)$$

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 $\left\| G_r v_r^+ - z \right\|^2$. 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 [33], 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 [47].

2.6 Relación de la Regresión Categórica con otras Técnicas Estadísticas

La regresión categórica constituye una generalización de varias técnicas estadísticas, por ejemplo si la variable dependiente es continua y se tiene sólo una independiente con nivel de medición nominal, entonces la regresión categórica se convierte en un análisis de varianza clásico (ANOVA). Con un nivel de escalamiento nominal, la cuantificación para cada categoría es la media de los valores de la variable dependiente tomando los casos que pertenecen a esa categoría. La variable transformada coincide con la variable original, en la que los valores de la categoría se sustituyen por los valores de su media y el resultado se estandariza.

Si se tiene una variable dependiente nominal, la regresión categórica se convierte entonces en un análisis discriminante clásico [31].

Por otra parte, RegCat es equivalente al análisis de correlación canónica categórico mediante escalamiento óptimo (OVERALS) con dos conjuntos, uno de los cuales contiene sólo una

variable. Si se escalan todas las variables a nivel numérico, el análisis se corresponderá con el análisis de regresión múltiple típico.

2.6.1 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.

```
* 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.
```

Consideraciones Finales del Capítulo

El presente capítulo expone los fundamentos de los métodos defendidos en la tesis doctorales: *"Nonparametric inference in nonlinear principal components analysis: exploration and beyond"* y "*Prediction accuracy and stability of regression with optimal scaling transformations*". En cada uno de los casos se comienza esbozando brevemente las ideas de la técnica que ellos modifican: el análisis de componentes principales y el análisis de regresión lineal múltiple.

La importancia fundamental de ambos métodos es que permiten la modelación no lineal de las variables categóricas que utilizan. Se define un escalado óptimo para cada variable y se trabaja con esas variables transformadas. El procedimiento de escalado que ambas técnicas utilizan es esencialmente el mismo, pero debido a su importancia, se explican los detalles particulares para cada una de ellas.

Se comentan además las semejanzas y diferencias con respecto a la técnica que les dio origen, así como la relación que tienen con otras técnicas estadísticas de uso frecuente en el análisis de datos.

Capítulo 3. Aplicación: 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 [48].

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 [49]. 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, hiperreactivo (prehipertenso) o hipertenso por un comité de expertos altamente calificado. La tabla 3.1 muestra las variables originales que formaron parte de este estudio:

Tabla 3.1 Variables consideradas en el análisis

No.	Variable	Etiqueta	Valores
1.	Edad	Edad del paciente	16 – 80 años
2.	TASistB	Presión sistólica basal	Baja, Media, Alta
3.	TADiastB	Presión diastólica basal	Baja, Media, Alta
4.	TASistB1	Presión sistólica basal al primer minuto	Baja, Media, Alta
5.	TADiastB1	Presión diastólica basal al primer minuto	Baja, Media, Alta
6.	TASistB2	Presión sistólica basal al segundo minuto	Baja, Media, Alta
7.	TADiastB2	Presión diastólica basal al segundo minuto	Baja, Media, Alta
8.	TAPam	Presión arterial media	Baja, Media, Alta
9.	Col_Tot	Colesterol total	Bajo, Medio, Alto
10.	Col_Ldl	Colesterol LDL	Bajo, Medio, Alto
11.	Col_Hdl	Colesterol HDL	Bajo, Medio, Alto
12.	OImc	Índice de masa corporal	Bajo, Normal, Elevado
13.	Sexo	Sexo del paciente	Masculino, Femenino
14.	Fuma	Hábito de fumar	Sí, No
15.	Bebe	Hábito de tomar	Sí, No
16.	Diabetes	Padecimiento de Diabetes mellitus	Sí, No
17.	Dislipidemia	Padecimiento de Dislipidemia	Sí, No
18.	Raza	Raza del paciente	Blanca, Mestiza
19.	DiagExp	Diagnóstico de HTA	Normotenso, Hiperreactivo, Hipertenso

3.1 Análisis Univariado

Como primer paso en el estudio de la HTA se realizó un estudio univariado de las variables involucradas en la investigación. Se decidió comenzar analizando las posibles correlaciones entre todas las variables predictoras.

3.1.1 Análisis de Correlación

En este epígrafe se analizaron los coeficientes de correlación Pearson, Kendall y Spearman para todas las variables, pues todas tienen un nivel de medición al menos ordinal. En el cálculo de las correlaciones (Ver Anexos 1-3) se aprecia por ejemplo que la variable que representa la presión arterial media (TAPam) está altamente correlacionada con todas las variables. Se muestra además que todas las variables que miden presión están altamente correlacionadas entre ellas. De las variables que miden colesterol se aprecia que la variable Col_LDL correlaciona con Col_Tot y con Col_HDL. Estas últimas no correlacionan entre sí. La relación entre el colesterol HDL y el colesterol total suministra más información sobre el riesgo a padecer enfermedades de tipo cardiovascular [50]. Los hábitos de fumar y de beber están correlacionados entre sí. El

padecimiento de las enfermedades diabetes y dislipidemia también están correlacionadas. Para el resto de las variables no se puede establecer una relación clara de asociación. Por ejemplo, la variable Raza solamente está correlacionada con las variables Sexo y TAPam. La variable DiagExp correlaciona con todas las variables excepto con la variable Fuma.

3.1.2 Análisis de Tablas de Contingencia

En este epígrafe se realizó un estudio, univariado también, que consistió en el cálculo del estadístico chi cuadrado en tablas de contingencia. Se comparó el diagnóstico de expertos con cada una de las posibles variables predictoras que intervienen en el análisis (Ver Tabla 3.1)

Tabla 3.2 Resumen de las tablas de contingencia

	Variable	**Significación asintótica**	**Significación exacta**	**Significación de Monte Carlo**
Diagnóstico de Expertos	Edad	-	-	-
	TASistB	4.7931E-095	*	0.0000
	TADiastB	5.4909E-129	*	0.0000
	TASistB1	2.3487E-132	*	0.0000
	TADiastB1	8.6510E-117	*	0.0000
	TASistB2	2.2994E-119	*	0.0000
	TADiastB2	1.9046E-149	*	0.0000
	TAPam	3.6897E-150	*	0.0000
	Col_Tot	2.2080E-009	*	0.0000
	Col_Ldl	1.5833E-005	*	0.0000
	Col_Hdl	0.0152	*	0.0156
	OImc	1.0291E-018	*	0.0000
	Sexo	5.0764E-006	4.5306E-006	0.0000
	Fuma	0.1133	0.1160	0.1172
	Bebe	0.0027	0.0026	0.0024
	Diabetes	0.0002	0.0001	0.0001
	Dislipidemia	0.0053	0.0046	0.005
	Raza	0.0118	0.0119	0.0127

Como puede apreciarse en la tabla anterior, se realizaron los cálculos correspondientes a la significación asintótica, al test exacto y al de la simulación por el método de Monte Carlo. Debido a que la cantidad de casos es relativamente elevada (849) todos estos valores prácticamente

coinciden [13]. En el caso de la significación exacta, solo se pudo calcular dicho valor a las últimas cinco variables pues la memoria en las máquinas resultó insuficiente. La variable Edad no se incluyó en el análisis por ser una variable continua. La significación asintótica para las variables Diabetes y Dislipidemia no son confiables pues aparecen casillas con frecuencia esperada inferior a 5 (2 casillas en el caso de la Diabetes y 1 casilla en el caso de la Dislipidemia), sin embargo los cálculos de Monte Carlo y el *p* exacto muestran que ambas variables son significativas [5].

Realizando una interpretación estadística del problema puede concluirse que, desde el punto de vista univariado todas las variables, excepto la variable Fuma, son capaces de diferenciar los grupos, o sea, el test chi cuadrado arrojó resultados significativos.

Las variables predictivas pueden ordenarse por la fortaleza de su asociación [6] de acuerdo con el valor del estadístico V de Cramer. La tabla 3.3 muestra los resultados obtenidos ordenados de mayor a menor. En la misma puede apreciarse que la variable TAPam es la variable más significativa de todas las que intervinieron en el estudio de la HTA.

Tabla 3.3 Resumen de la V de Cramer

	Variable	**Significación asintótica**	**V de Cramer**
Diagnóstico de Expertos	TAPam	3.6897E-150	0.6420
	TADiastB2	1.9046E-149	0.6405
	TASistB1	2.3487E-132	0.6031
	TADiastB	5.4909E-129	0.5954
	TASistB2	2.2994E-119	0.5730
	TADiastB1	8.6510E-117	0.5669
	TASistB	4.7931E-095	0.5120
	OImc	1.0291E-018	0.2308
	Sexo	5.0764E-006	0.1694
	Col_Tot	2.2080E-009	0.1649
	Diabetes	0.0002	0.1413
	Col_Ldl	1.5833E-005	0.1272
	Bebe	0.0027	0.1178
	Dislipidemia	0.0053	0.1109
	Raza	0.0118	0.1021
	Col_Hdl	0.0152	0.0851
	Fuma	0.1133	0.0716
	Edad	-	-

La información que nos ofrece el test chi cuadrado y el estadígrafo V de Cramer es sin dudas muy valiosa, pero resulta ser demasiado abarcadora. Un estudio multivariado debería ayudar a comprender mejor las relaciones entre las variables y a simplificar la cantidad de ellas que realmente se necesitan para obtener un diagnóstico adecuado.

3.2 Análisis Multivariado

3.2.1 Técnicas de Segmentación: CHAID

En este epígrafe se aplica la técnica de segmentación CHAID tomando como variable independiente el diagnóstico de expertos (DiagExp) y como posibles variables predictoras todas las analizadas con anterioridad. Aunque el hábito de fumar no resultó significativo en las tablas de contingencia, se decidió mantener en el análisis para confirmar dicho resultado. La figura 3.1 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 muestra, 193 son prehipertensos, (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 el test chi cuadrado (Ver tabla 3.2) y con la V de Cramer (Ver Tabla 3.3)

El árbol creado tiene 7 hojas o nodos terminales. Veamos su explicación:

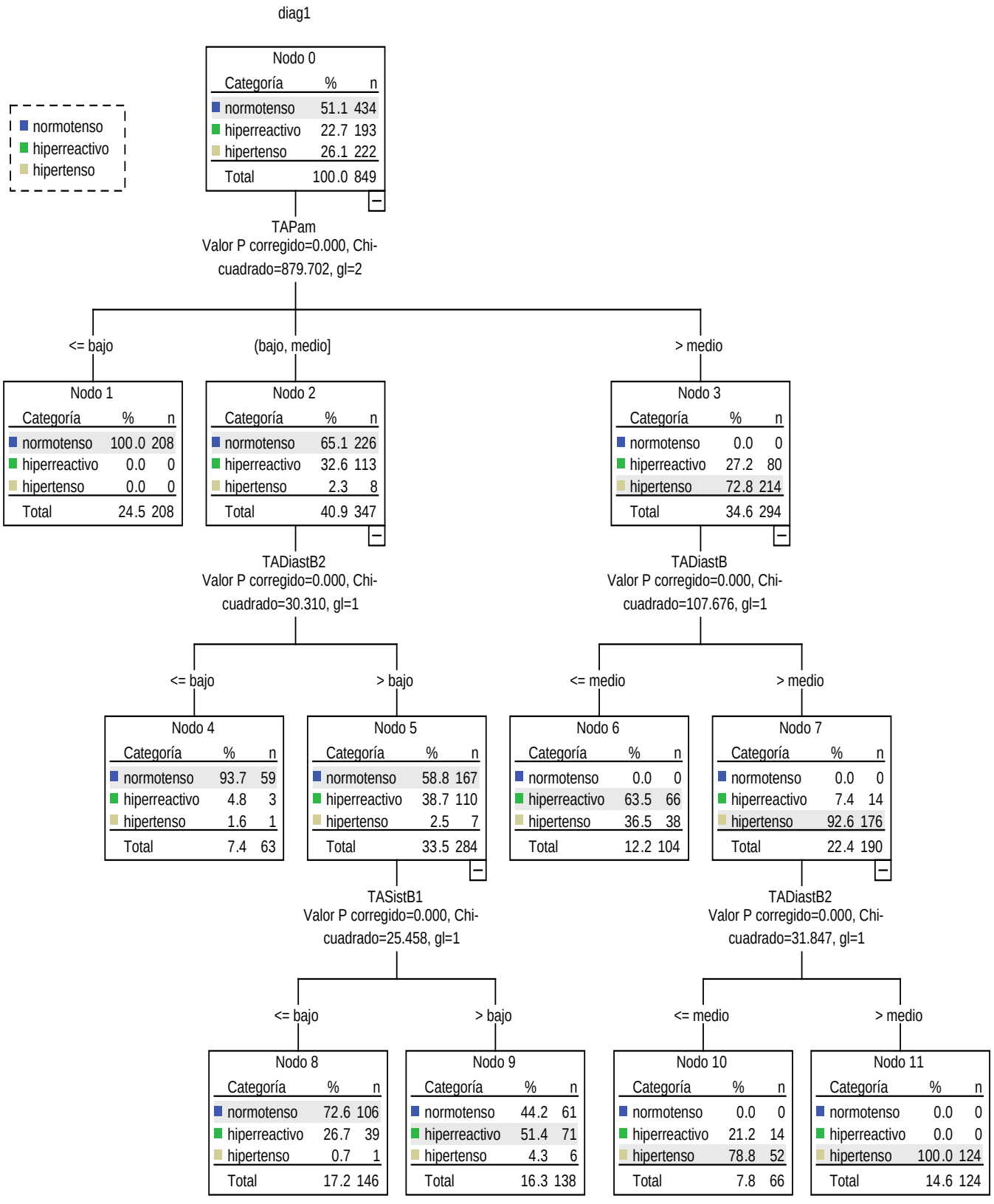
-1-. Subconjunto formado por 208 pacientes caracterizan por presentar valores bajos en la TAPam. Todos los pacientes del grupo son normotensos. Se corresponde con el Nodo 1 del árbol.

-2-. Subconjunto formado por 63 pacientes. Estos se caracterizan por tener valores de la TAPam entre bajo o medio y valores bajos de la TADiastB2. Existe predominio de normotensos (93.7%) y el resto está conformado por hiperreactivos (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 hiperreactivos (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 bajo y medio de la TAPam, valores entre medio y alto 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 hiperreactivos y uno solo de los pacientes es hipertenso. Se corresponde con el Nodo 8 del árbol.

Figura 3.1 Árbol de decisión aplicando la técnica CHAID



-5-. Subconjunto formado por 138 pacientes. Se caracterizan por tener valores entre bajo y medio de la TAPam, valores entre medio y alto en la TADiastB2 y valores entre medio y alto de la TASistB1. En este grupo predominan los hiperreactivos (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 bajo y medio en la TADiastB2. En este grupo 52 pacientes son hipertensos (78.8%) y 14 son hiperreactivos (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 3.4 muestra los resultados obtenidos:

Tabla 3.4 Clasificación

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

Método de crecimiento: CHAID
Variable dependiente: Diagnóstico de Expertos

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 hiperreactivos. 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.

Con el objetivo de corroborar los resultados anteriores se decidió aplicar otra técnica de clasificación: la regresión logística multinomial.

3.2.2 Análisis de Regresión Logística

La opción de regresión logística multinomial resulta útil en aquellas situaciones en las que se desea clasificar a los sujetos según los valores de un conjunto de variables predictoras. Este tipo de regresión es similar a la regresión logística, pero más general, ya que la variable dependiente no está restringida a dos categorías. Resulta apropiada entonces para nuestro problema, pues como ya se conoce, la variable diagnóstico tiene tres categorías.

La tabla 3.5 muestra los resultados de la clasificación utilizando la regresión logística. Observe que el porcentaje de casos correctamente clasificados ahora es de un 92.7%, muy superior al obtenido con la técnica CHAID (80.8%)

Tabla 3.5 Clasificación

	Pronosticado			
Observado	normotenso	hiperreactivo	hipertenso	Porcentaje correcto
normotenso	408	26	0	94.0%
hiperreactivo	36	157	0	81.3%
hipertenso	0	0	222	100.0%
Porcentaje global	52.3%	21.6%	26.1%	92.7%

3.2.3 Análisis cruzado CHAID – Regresión Logística

En este epígrafe se realiza un estudio cruzado para comparar los resultados de ambas técnicas, obsérvese la tabla 3.6

Tabla 3.6 Tabla de contingencia Categoría de respuesta pronosticada por CHAID * Categoría de respuesta pronosticada por Regresión Logística

Recuento

		Categoría de respuesta pronosticada por Regresión Logística			
		normotenso	hiperreactivo	hipertenso	Total
Categoría de respuesta pronosticada por CHAID	normotenso	383	32	2	417
	hiperreactivo	61	137	44	242
	hipertenso	0	14	176	190
Total		444	183	222	849

% de coincidencia: 82%

Como puede apreciarse, ambas técnicas coinciden en la clasificación de 383+137+176=696 pacientes, lo que representa aproximadamente un 82% del total de la muestra. Obsérvese que las confusiones mayores ocurren en el grupo de los hiperreactivos.

Tabla 3.7 Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	883.160[a]	4	.000
Razón de verosimilitudes	908.039	4	.000
Asociación lineal por lineal	632.306	1	.000
N de casos válidos	849		

a. 0 casillas (.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 40.95.

Se aplicó además un test chi cuadrado (Ver Tabla 3.7) y los resultados fueron significativos, lo que demuestra que existe dependencia entre ambas clasificaciones.

3.3 Análisis de Regresión Categórica

En numerosas investigaciones, sobre todo en el campo médico o social [27], 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 3.1). Se decide entonces aplicar la regresión categórica presente en el SPSS en su versión 13 [31]. En la primera corrida se consideraron todas las variables mostradas en la Tabla 3.1. 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 Tabla 3.3 y Figura 3.1) y por tanto la que mayor influencia ejerce sobre la variable dependiente (DiagExp) [31, 47]. 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.

Tabla 3.8 Resumen del modelo

R múltiple	R cuadrado	R cuadrado corregida
.910	.828	.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

El resultado del análisis de varianza resultó significativo lo que indica que el modelo es válido (Ver Anexo 4). Ahora bien el modelo que se obtiene es muy grande, o sea, está compuesto por numerosas variables predictoras (Ver Tabla 3.9) 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.

Tabla 3.9 Coeficientes

	Coeficientes tipificados				
	Beta	Error típ.	gl	F	Sig.
Edad	.020	.017	1	1.342	.247
Sexo	-.065	.017	1	15.604	.000
Raza	.025	.015	1	2.872	.090
Bebe	-.012	.016	1	.582	.446
Fuma	-.001	.015	1	.004	.947
Diabetes mellitus	-.018	.015	1	1.362	.244
Dislipidemia	-.006	.015	1	.150	.699
TASistB	.005	.024	1	.038	.845
TADiastB	.151	.026	1	34.987	.000
TASistB1	.164	.027	1	35.719	.000
TADiastB1	.088	.024	1	13.070	.000
TASistB2	.088	.028	1	10.172	.001
TADiastB2	.215	.025	1	74.572	.000
TAPam	.353	.023	2	239.095	.000
OIMC	.043	.016	1	7.761	.005
Col_Tot	-.015	.021	1	.520	.471
Col_HDL	-.011	.016	1	.532	.466
Col_LDL	-.004	.021	1	.032	.859

Variable dependiente: Diagnóstico de Expertos

Tabla 3.10 Correlaciones y Tolerancia

	Correlaciones				Tolerancia	
	Orden cero	Parcial	Semiparcial	Importancia	Después de la transformación	Antes de la transformación
Edad	.256	.040	.017	.006	.731	.730
Sexo	.169	-.136	-.057	-.013	.755	.754
Raza	.101	.059	.024	.003	.959	.959
Bebe	-.113	-.026	-.011	.002	.797	.796
Fuma	-.063	-.002	-.001	.000	.879	.878
Diabetes mellitus	-.141	-.040	-.017	.003	.918	.920
Dislipidemia	-.111	-.013	-.006	.001	.914	.913
TASistB	.646	.007	.003	.004	.365	.365
TADiastB	.729	.201	.085	.133	.317	.319
TASistB1	.758	.203	.086	.150	.276	.274
TADiastB1	.722	.125	.052	.077	.346	.339
TASistB2	.762	.110	.046	.081	.272	.304
TADiastB2	.761	.287	.124	.198	.335	.259
TAPam	.803	.473	.223	.343	.398	.192
OIMC	.319	.096	.040	.017	.856	.856
Col_Tot	.212	-.025	-.010	-.004	.482	.482
Col_HDL	-.086	-.025	-.011	.001	.840	.834
Col_LDL	.177	-.006	-.003	-.001	.480	.479

Variable dependiente: Diagnóstico de Expertos

Para interpretar la contribución de los predictores a la regresión, no es suficiente con inspeccionar los coeficientes de la regresión. Además debe inspeccionarse los valores que aparecen en la tabla 3.10. En la misma se muestra la correlación de orden cero que no es más que la correlación entre el predictor transformado y la respuesta transformada. En nuestro modelo el valor más alto corresponde a la TAPam. Se muestra además los valores de la correlación parcial, la misma representa el efecto del predictor eliminando los efectos de los otros predictores y la variable respuesta. El cuadrado de la correlación parcial explica la proporción de la varianza de la respuesta que explica el predictor, eliminando los efectos de las otras variables. En nuestro caso la variable TAPam explica un 22.37 % de variación en el diagnóstico de expertos (DiagExp). Como una alternativa de la eliminación de los efectos de las variables en la respuesta y un predictor, se puede eliminar los efectos solamente del predictor. Este indicador lo muestra la correlación semiparcial [31, 47].

Además de los coeficientes de la regresión y las correlaciones, la medida de la importancia ayuda a interpretar la contribución de los predictores a la regresión. En contraste con el coeficiente de regresión, esta medida define la importancia de los predictores aditivamente. En

nuestro ejemplo la variable más importante es la TAPam. Todas las variables que miden presión arterial acumulan un 98.6 % de importancia (Ver Tabla 3.10) [31, 47].

Valores altos en las correlaciones entre los predictores reduce la estabilidad del modelo de regresión. La tolerancia refleja en qué medida las variables independientes están linealmente relacionadas unas con las otras. Valores de la tolerancia cerca de 1 indica que la variable no puede ser bien predicha a partir de otros predictores. En contraste, una con tolerancia muy baja contribuye poca información al modelo, y puede causar problemas computacionales [31, 47]. En nuestro ejemplo todos los valores de la Tolerancia son altos indicando la ausencia de multicolinealidad (Ver Tabla 3.10).

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 (Ver Anexo 5). 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 [47] ya que la regresión categórica no realiza este tipo de análisis [31].

El estadístico de Durbin Watson obtenido fue de 1.534 (Ver Anexo 5) indicando que no hay autocorrelación y por tanto existe homogeneidad de varianza [51]. El índice de condición reafirma la ausencia de multicolinealidad (Ver Anexo 5).

En el modelo obtenido aparecen varias variables no significativas (Ver Tabla 3.9), 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.

3.4 Análisis de Componentes Principales Categóricas como método de selección de variables

El método de componentes principales ha sido utilizado de manera creciente en las últimas décadas, prácticamente en todas las áreas, es el 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 [2]. 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 (41.542 %). 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 3.11 muestra el resumen del modelo obtenido. Los autovalores pueden usarse como criterio acerca del número de dimensiones necesarias. La tabla también muestra el valor del estadístico alfa de Cronbach (0.917), que es una medida de confiabilidad que se maximiza en el procedimiento.

Tabla 3.11 Resumen del modelo

Dimensión	Alfa de Cronbach	Varianza explicada	
		Total (Autovalores)	% de la varianza
1	.872	5.670	31.499
2	.473	1.808	10.042
Total	.917[a]	7.478	41.542

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

La tabla 3.12 por su parte, muestra las saturaciones de cada variable en las dos componentes obtenidas. Obsérvese que las variables que miden presiones tienen un valor elevado (superior a 0.80) en la primera dimensión y un valor muy pequeño en la segunda. 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 la primera.

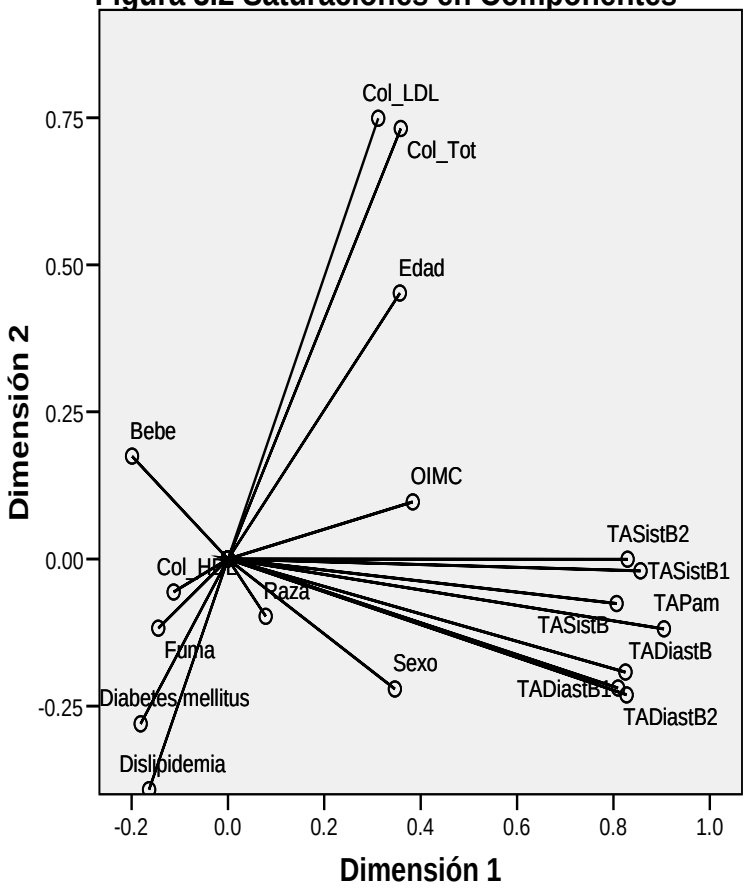
La figura 1 muestra gráficamente el mismo resultado anterior. Las variables que miden presiones están muy cercas unas de otras, luego todas esas variables están altamente correlacionadas. Todas ellas tienen valores elevados en la dimensión 1 y bajos en la 2. Dos de los colesteroles por su parte (Col_LDL y Col_Tot), se encuentran ubicados en la esquina superior izquierda de la figura, lo que muestra el efecto contrario: saturaciones pequeñas en la primera componente y elevadas en la segunda.

Tabla 3.12 Saturaciones en componentes

	Dimensión	
	1	2
Edad	.357	.452
Sexo	.346	-.221
Raza	.078	-.098
Bebe	-.199	.175
Fuma	-.144	-.118
Diabetes mellitus	-.181	-.280
Dislipidemia	-.164	-.392
TASistB	.807	-.076
TADiastB	.825	-.192
TASistB1	.856	-.020
TADiastB1	.810	-.219
TASistB2	.829	-.001
TADiastB2	.827	-.231
TAPam	.905	-.119
OIMC	.384	.097
Col_Tot	.359	.732
Col_HDL	-.113	-.056
Col_LDL	.312	.749

Normalización principal por variable.

Figura 3.2 Saturaciones en Componentes



Normalización principal por variable.

Al analizar la figura 3.2 puede sospecharse la presencia de un tercer factor. Como el objetivo de nuestro análisis es eliminar variables no significativas se decide entonces repite el análisis de componentes principales para datos categóricos pero solicitando una tercera dimensión.

La tabla 3.13 muestra el resumen del modelo obtenido. Considerando la totalidad de las variables resulta ser satisfactorio ya que el por ciento total de la varianza explicada por los factores es pequeño (49.706 %), aunque vale aclarar que supera el modelo anterior.

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.

Tabla 3.13 Resumen del modelo

Dimensión	Alfa de Cronbach	Varianza explicada	
		Total (Autovalores)	% de la varianza
1	.872	5.670	31.500
2	.473	1.807	10.041
3	.338	1.470	8.165
Total	.940[a]	8.947	49.706

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

La tabla 3.14 es bastante transparente en cuanto a las variables que intervienen en cada una de las dimensiones

Tabla 3.14 Saturaciones en componentes

	Dimensión		
	1	2	3
Edad	.357	.451	.111
Sexo	.346	-.223	.597
Raza	.078	-.097	-.192
Bebe	-.199	.177	-.707
Fuma	-.144	-.116	-.673
Diabetes mellitus	-.181	-.280	-.096
Dislipidemia	-.164	-.392	-.131
TASistB	.806	-.076	-.051
TADiastB	.825	-.192	-.087
TASistB1	.856	-.020	-.047
TADiastB1	.810	-.219	-.079
TASistB2	.830	-.001	-.073
TADiastB2	.827	-.231	-.097
TAPam	.905	-.118	-.071
OIMC	.384	.098	-.212
Col_Tot	.359	.732	.003
Col_HDL	-.113	-.056	-.015
Col_LDL	.312	.749	-.026

Normalización principal por variable.

Puede observarse que en la primera dimensión intervienen las variables que miden presión arterial. En la segunda figuran con valores elevados dos de las variables que miden colesterol (Col_Tot y Col_LDL) y en la tercera aparecen las variables Sexo, Bebe y Fuma. En consecuencia con este resultado se decide eliminar las variables que no tributan a ninguna dimensión y que además no son significativas en el modelo.

3.5 Nuevo Análisis de Regresión Categórica con las Recomendaciones del ACPCat

Con estas consideraciones se vuelve a obtener otro modelo de regresión categórica. En él se obtiene un R^2 igual a 0.827 (Ver Anexo 6). Nótese que prácticamente no disminuye, si lo

comparamos con el valor anterior, que era de 0.828. El análisis de varianza nuevamente es significativo (Ver Anexo 6).

La tabla 3.15 refleja los coeficientes del modelo encontrado. Evidentemente es un modelo más claro, sencillo y de mejor interpretación.

Tabla 3.15 Coeficientes

	Coeficientes tipificados				
	Beta	Error típ.	gl	F	Sig.
Sexo	-.064	.016	1	15.412	.000
Bebe	-.011	.016	1	.481	.488
Fuma	-.003	.015	1	.034	.854
TASistB	.000	.024	1	.000	.985
TADiastB	.153	.026	1	35.864	.000
TASistB1	.174	.027	1	41.039	.000
TADiastB1	.090	.024	1	13.707	.000
TASistB2	.088	.027	1	10.523	.001
TADiastB2	.208	.025	1	71.464	.000
TAPam	.359	.023	2	246.718	.000
OIMC	.048	.015	1	9.743	.002
Col_Tot	-.013	.020	1	.437	.509
Col_LDL	.002	.020	1	.008	.928

Variable dependiente: Diagnostico de Expertos

Se reafirma la TAPam como la variable más importante (Ver Anexo 6).

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 se comprueba que los errores están normalmente distribuidos, que existe homogeneidad de varianza y que no hay multicolinealidad, ver los resultados en el Anexo 7.

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.

3.6 Regresión Categórica como Análisis Discriminante

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 el capítulo 2, 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 el último epígrafe del capítulo 2.

En nuestro estudio y siguiendo las instrucciones anteriormente mencionadas obtuvimos un 84.57 % de pacientes bien clasificados. Los resultados se muestran en la tabla 3.16

Tabla 3.16 DiagExp * Clasificación

Count

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

Consideraciones Finales del Capítulo

En este capítulo se muestra una aplicación médica: el análisis de HTA en Santa Clara. Se comienza realizando un estudio univariado con las variables que forman parte de la investigación. Se calculan e interpretan coeficientes de correlación lineal, se realizan tablas de contingencia de todas las variables predictoras contra la variable diagnóstico de expertos (DiagExp).

A continuación se aplican dos técnicas multivariadas: se obtiene un árbol de decisión, basado en el estadístico chi cuadrado (CHAID) y se obtiene un modelo de regresión logística multinomial. Con ambos métodos se obtiene un pronóstico de clasificación y una tabla de contingencia que muestra su relación.

Posteriormente se realiza un análisis de regresión categórica utilizando todas las posibles variables predictoras. Se obtiene un modelo válido, (pues cumple con los supuestos de la regresión) pero demasiado grande, (ya que a él pertenecen las 18 variables predictivas). Con el objetivo de simplificar el modelo se decide utilizar el método de componentes principales categóricos como un criterio de eliminación de variables irrelevantes. Se muestran y explican todos los pasos seguidos, así como el modelo final. Estos pasos constituyen una guía metodológica para realizar investigaciones similares. El modelo hallado es efectivamente más sencillo y continúa siendo válido, ya que cumple con los supuestos del análisis de regresión. Como que la variable independiente es categórica, más que un análisis de regresión es útil un análisis discriminante, capaz de pronosticar la clase a la que pertenece cada uno de los 849 individuos analizados. El capítulo termina mostrando como puede utilizarse el método de regresión categórica con estos fines.

Conclusiones

Con este trabajo se arriban a las siguientes conclusiones:

1. Los datos categóricos pueden procesarse con ayuda de técnicas alternativas del análisis de datos, como son: el análisis de componentes principales categórico y el análisis de regresión categórica.
2. Utilizando un escalado adecuado de los datos en problemas donde aparecen variables continuas y discretas, es factible utilizar el análisis de componentes principales como técnica exploratoria de datos y como método de selección de las variables a tener en cuenta en un análisis de regresión.
3. Utilizando un escalado adecuado de los datos en problemas donde aparecen variables continuas y discretas, es factible utilizar el análisis de regresión categórica para hallar modelos rigurosos y válidos para el pronóstico, que cumplen con los supuestos del análisis de regresión.
4. El método de regresión categórica puede ser utilizado como un nuevo clasificador. Su importancia fundamental radica en la posibilidad que brinda de combinar variables con todos los niveles de medición.
5. Desde el punto de vista del problema de la HTA aquí presentado, puede concluirse que:
    a. La presión arterial media es la variable que más influye en el pronóstico de la variable dependiente.
    b. Se hallaron tres modelos multivariados para el pronóstico de la HTA: CHAID, regresión multinomial y regresión categórica. Analizando el porcentaje de casos correctamente clasificados, es la regresión multinomial la técnica que mejores resultados produjo.

Recomendaciones

Este trabajo sin dudas no constituye un tema terminado, más bien propicia el despertar de varias aristas en el horizonte investigativo. Se recomienda entonces:

1. Utilizar métodos estadísticos de comparación de clasificadores (curvas ROC) para estudiar las diferencias entre los resultados arrojados por los árboles de decisión, la regresión multinomial y la regresión categórica en el ejemplo de la HTA.
2. Aplicar los métodos de análisis de componentes principales para datos categóricos y análisis de regresión categórica en otros dominios diferentes al tratado en esta tesis, como pudieran ser en problemas de bioinformática, en las ciencias sociales o en las ingenierías, por solo citar algún ejemplo.
3. Continuar con el estudio de otros métodos estadísticos para el análisis de datos categóricos.

Bibliografía

1. Agresti, A., *Categorical Data Analysis*. Second ed, University of Florida. Gainesville, Florida: John Wiley & Sons, Inc., Publication.
2. Hair, J.F., et al., *Análisis multivariante. 5ed.* 1999, Madrid: Prentice Hall.
3. *SPSS 10 para Windows. Manual de usuarios. Capítulo 12*, SPSS Soft.
4. Vicéns Otero, J. and E. Medina Moral. *Análisis de datos cualitativos*. 2005 [cited 2007 22 de septiembre]; Available from: www.uam.es/personal_pdi/economicas/eva/pdf/tab_conting.pdf.
5. Grau, R., *Independencia de variables y medidas de asociación. Capítulo 3. Primera parte*
6. Press, W., et al., *NUMERICAL RECIPES in C++. The Art of Scientific Computing*. Second Edition. 2005: Cambridge University Press.
7. Agresti, A., C.R. Mehta, and N.R. Patel, *Exact inference for contingency tables with ordered categories.* Journal of the American Statistical Association, 1990. **85:410**: p. 453-458.
8. Hilton, J., C.R. Mehta, and N.R. Patel, *Exact Smirnov p values using a network algorithm.* Computational Statistics and Data Analysis, 1994. **17**: p. 4, 351-361.
9. Mehta, C.R., *An interdisciplinary approach to exact inference for contingency tables.* Statistical Sciencie, 1992. **7**: p. 167-170.
10. Mehta, C.R. and N.R. Patel, *FEXACT: A FORTRAN subroutine for Fisher´s exact test on unordered r x c contingency tables.* ACM Transactions on Mathematical Software, 1986. **12:2**: p. 154-164.
11. Mehta, C.R., N.R. Patel, and P. Senchaudhuri, *Exact stratified linear rnak test for ordered categorical and binary data.* Journal of Computational and Graphical Statistics, 1992. **1**: p. 21-40.
12. Mehta, C.R. and N.R. Patel, *A hybrid algorithm for Fisher´s exact test in unordered r x c contingency tables.* Communications in Statistics, 1986. **15:2**: p. 387-403.
13. Mehta, C.R. and N.R. Patel, *SPSS Exact Tests 7.0 for Windows*. 1996, SPSS Inc.
14. Stanton, J.M., *Galton, Pearson and the Peas: A brief history of linear regression for statistics instructors.* Journal of Statistics Education 2001. **9, Number 3**.
15. Pearson, K., *Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia.* Philosophical Transactions of the Royal Society of London 1896. **187**: p. 253-318.

16. Palmer, A., R. Jiménez, and J.J. Montaño. *Tutorial sobre el coeficiente de correlación con una o dos variables categóricas*. 2000 [cited 26 de octubre de 2007]; Available from: http://www.psiquiatria.com/psicologia/revista/50/2830
17. Palmer, A., R. Jiménez, and J.J. Montaño. *Tutorial sobre el coeficiente de correlación lineal de Pearson*. 2001 [cited 26 de octubre de 2007]; Available from: http://www.psiquiatria.com/psicologia/revista/51/2815.
18. Rojas García, G., *Muestreo para correlaciones por contingencias de Pearson*, in *Departamento de Matemática. Tesis presentada en opción al Título Académico de Máster en Matemática Aplicada.* 2007, Universidad Central de las Villas, Cuba.
19. Belmonte, M.A. *SPSS-ANOVA*. 2000 [cited 26 de octubre de 2007]; Available from: http://www.pighealth.com/Scourse/lecture/lec1071/031.htm.
20. Conover, W.J., *Practical nonparametric statistics*. 2nd ed. 1980: New York: John Wiley and Sons.
21. Grau, R., *Independencia de variables y medidas de asociación. Capítulo 3. Segunda parte*
22. *CHAID para SPSS sobre Windows. Técnicas de segmentación basadas en razones de verosimilitud Chi-cuadrado, Manual de usuario*, SPSS Soft. Inc. 1994.
23. Zamora Rodríguez, L., *Una técnica de segmentación aplicada a la epidemeología*, in *Departamento de Matemática. Tesis de Maestría en Matemática Aplicada.* 1997, Universidad Central de las Villas, Cuba.
24. Chitarroni, H. *La regresión logística*. 2002 [cited 9 /ene/ 2008]; Available from: www.salvador.edu.ar/csoc/idicso/docs/aephc1.pdf.
25. Caballero Granado, F.J. *Modelos de regresión logística incondicional (I)*. 2007 [cited 9/ene/ 2008]; Available from: http://saei.org/hemero/epidemiol/nota4.html.
26. *Regression Models for Ordinal Response Data*. [cited 18 Abr 2008; Available from: health.bsd.uchicago.edu/rathouz/HS327/lect09.pdf.
27. Aron, A. and E. Aron, *Statistics for the Behavioral and Social Sciences*. Second edition. 2002: Prentice Hall.
28. Navarro Céspedes, J.M., et al., *Estudio del riesgo cardiovascular en el municipio de Santa Clara utilizando el método de Regresión Categórica* Compumat 2007: Cuba
29. Linting, M. *Nonparametric inference in nonlinear principal components analysis : exploration and beyond.* 2007 [cited 24 Ene 2008; Doctoral thesis]. Available from: https://www.openaccess.leidenuniv.nl/dspace/handle/1887/12386.

30. Johnson, R.A. and D.W. Wichern, *Applied multivariate statistical analysis. Fifth edition*. 2002, United States of América: Pearson education international.
31. Meulman, J.J. and W.J. Heiser. *SPSS Categories 13.0*. 2004.
32. De Leeuw, J., *Multivariate analysis with optimal scaling*, in *Progress in Multivariate Analysis*. 1990: Calcutta: Indian Statistical Institute
33. Ramsay, J.O. *Monotone Regression Splines in Action*. 1988 [cited 28 Ene 2008; Statistical Science 3, 425-461]. Available from: http://www.fon.hum.uva.nl/praat/manual/spline.html.
34. Gifi, A., *Nonlinear multivariate analysis*. 1990, Chichester, England: Wiley.
35. Gower, J.C. and D.J. Hand, *Biplots*. 1996, London: Chapman&Hall.
36. Meulman, J.J., A.J. Van der Kooij, and W.J. Heiser, *Principal Components Analysis with Nonlinear Optimal Scaling Transformations for Ordinal and Nominal Data*, in *Handbook of Quantitative Methods in the Social Sciences*, D. Kaplan, Editor. 2004, Sage Publications: Newbury Park, CA. p. 49-70.
37. Allison, P.D., *Missing Data*, ed. S.U.P.S.o.Q.A.i.t.S. Sciences. 2002, Thousand Oaks,CA:Sage. 07-136.
38. McKnight, P., K. McKnight, and S. Sidani, *Missing data. A gentle introduction*. 2007, New York: The Guilford Press.
39. Perez Lopez, C., *Métodos estadísticos avanzados con SPSS*. 2005, Madrid, España: Thomson.
40. Jolliffe, I.T., *Principal component analysis*. 2002, NewYork: Springer-Verlag.
41. Draper, N.R. and H. Smith, *Applied regression analysis*. 1980: Editorial Pueblo y Educación.
42. McCullagh, P., J.A. Nelder, and . *Generalized Linear Models*. 1989, London: Chapman and Hall.
43. Berthold M, H., *Intelligent Data Analysis*. Second Edition. 2007: Springer.
44. Hastie, T.J., R.J. Tibshirani, and J.H. Friedman, *The Elements of Statistical Learning*. 2001, New York: Springer.
45. Haber, L., *Categorical regression analysis of toxicity data.* Comments on toxicology, 2001. **7 No. 5-6**: p. pp 437-452.
46. Van der Kooij, A.J. and J.J. Meulman, *MURALS: Multiple regression and optimal scoring using alternating least squares*, in *Softstat '97 Advances in Statistical Software 6*. 1997: Stuttgart: Gustav Fisher. p. pp 99-106.

47. Van der Kooij, A.J. *Prediction accuracy and stability of regression with optimal scaling transformations*. 2007 [cited 24 Jan 2008; Doctoral thesis.]. Available from: https://www.openaccess.leidenuniv.nl/dspace/handle/1887/12096.
48. *Tuotromedico: Hipertensión Arterial*. [cited 20 Mar 2008]; Available from: http://www.tuotromedico.com/temas/hipertension.htm.
49. in *Microsoft ® Encarta ® 2006*, © 1993-2005 Microsoft Corporation. Reservados todos los derechos.
50. Institute, T.H. *Colesterol*. [cited 14 Abr 2008]; Available from: http://www.texasheartinstitute.org/HIC/Topics_Esp/HSmart/cholspan.cfm.
51. Calero, A., *Estadística II*. 1998, La Habana. Cuba: Pueblo y Educación.

Anexos