UNIVERSIDAD DE CIENFUEGOS ”CARLOS RAFAEL RODRÍGUEZ”
FACULTAD DE INGENIERÍA INFORMÁTICA
MATEMÁTICA BÁSICA Y APLICADA

MODELO DE PRONÓSTICO ALTERNATIVO PARA EL ESTADO AL EGRESO DE PACIENTES CON ENFERMEDADES CEREBRO-VASCULARES EN EL HOSPITAL PROVINCIAL DE CIENFUEGOS

Tesis presentada en opción al grado científico
de Master en Matemática Aplicada

Autor: Ing. ALEJANDRO GONZÁLEZ DELGADO
Tutores: Prof. Aux., ROBERTO SUÁREZ SURÍ, MsC.

Cienfuegos
2008

Las matemáticas poseen no solo la verdad, sino cierta belleza suprema. Una belleza fría y austera, como la de una escultura.

Bertran Russell

A todos los que compartieron mis preocupaciones.

A los que sufrieron mis tropiezos y festejaron mis alegrías.

A Mailyn que tuvo siempre listo su hombro para apoyarme después de la fatiga.

No pocas han sido las personas que me han aprovechado para mostrar su buena voluntad para conmigo a lo largo de estos meses de trabajo.
A todos ellos gracias.
No obstante, siempre se está en la obligación de mencionar a los que más cerca han estado y, por consiguiente, más posibilidad tuvieron de mostrarse en el momento preciso.
A mi tutor Roberto que apareció cuando la barca estaba a punto de zozobrar.
A Ridelio que ha sido un padre científico y un buen amigo
A Rubén por su comprensión y buen juicio.
A todos los profesores del departamento que brindaron su apoyo hasta donde pudieron.
A Bello, Migue y Pacheco porque nunca me dieron la espalda cuando acudí a ellos.
A todos mis colegas de la maestría por el equipo de trabajo que formamos juntos.
A Lily por la disposición siempre de ayudarme.
A mi nueva familia y amigos.
A Day, porque gracias a ella todo fue mucho más fácil.
Y por último a esas dos personas especiales que no conocen el límite de la generosidad, Mirtha y Mailyn.

RESUMEN

El presente trabajo ha tenido como propósito la búsqueda de un modelo de pronóstico alternativo para el estado al egreso de pacientes con enfermedades cerebro-vasculares en el Hospital Provincial de Cienfuegos, sobre la base de un conjunto de variables clínicas al momento del ingreso, con el objetivo de explorar la posibilidad de incrementar la calidad del pronóstico y/o reducir los predictores. Para ello se utilizó una muestra en el periodo 1-7-2000 al 31-1-2003. Las técnicas empleadas son el detector automático de interacciones Chi-cuadrado, la regresión logística multinomial, tablas de contingencia y correlaciones bivariadas basadas en el coeficiente de Spearman y se introduce la correlación canónica con escalamiento óptimo (OVERALS) en lugar del análisis de componentes principales para datos categóricos (PRINCALS).

Como resultado se obtuvieron dos modelos comparables: el primero a través del uso de OVERALS y el segundo mediante el empleo de PRINCALS. Se pudo comprobar que el primero es ligeramente superior al segundo a partir de los criterios de comparación considerados con este propósito, esto es, porcentaje de clasificación correcta, coeficiente Kappa, errores estándares en los coeficientes, sensibilidad y especificidad.

Por último se estableció una comparación del modelo propuesto aplicando la técnica de OVERALS con el obtenido por Bembibre-Suárez, lográndose como resultado más destacable el ligero incremento en el porcentaje de clasificación correcta y la disminución de la cantidad de variables predictoras a emplear.

ÍNDICE

Pág.

INTRODUCCIÓN

La necesidad del hombre de tener conocimientos los más certeros posibles sobre lo que ha de acontecer, siempre ha sido una constante en la historia. Desde los tiempos más remotos se han recurrido a diferentes fuentes con estos fines. Se pueden citar, por ejemplo, el famoso Oráculo de Delphos al cual apelaron, incluso en circunstancias históricas críticas, no pocos hombres ilustres de la antigüedad. El interés por conocer el futuro radica precisamente en el hecho de posibilita trazar planes que permitan afrontar de forma más efectiva lo que pueda acontecer. En la actualidad, se han desarrollados herramientas científicas que posibilitan satisfacer precisamente está necesidad de manera más o menos satisfactoria. Dentro del campo de las estadísticas destacan las técnicas regresión, el análisis de series de tiempo, etc.

El campo de la salud tampoco escapa a esta necesidad. Se debe a Hipócrates el mérito de haber sido el primer médico que se ocupó del pronóstico de los enfermos que atendía. Consideraba él, que la capacidad de preveer en un médico era un elemento primordial en el éxito de su profesión. Decía Hipócrates que ”el médico puede predecir la evolución de una enfermedad mediante la observación de un número suficiente de casos”(Novas & Machado, 2002).

No obstante, el interés por el pronóstico ha sido menor que por el diagnóstico de los pacientes. En la actualidad se incrementa el interés por el pronóstico en el campo de la salud al tener en cuenta que del resultado de este depende la actitud que pueden asumir médicos, familiares y hasta el propio paciente a la hora de tomar decisiones relacionadas, por ejemplo, con someterse a una intervención quirúrgica, afrontar cierto tratamiento en fase experimental, etc.

En las últimas décadas el desarrollo de la informática ha producido un incremento considerable de los modelos de pronóstico. Se han utilizado por ejemplo, con el propósito de medir

desigualdades sociales en diferentes áreas geográficas sobre la base del pronóstico de la mortalidad infantil (Schneider $et\ al.$, 2002). Otro reporte relacionado con el embarazo está dado por el análisis de las complicaciones de las mujeres en este estado sometidas a violencia física (Helena & Aguirre, 1996). En ambos casos se emplearon la regresión lineal múltiple y la regresión logística como técnica para el pronóstico. También se reporta una aplicación del análisis discriminante para el estudio de la relación entre la aparición de casos de dengue a partir del grado de vulnerabilidad a esta enfermedad, de las regiones urbanas analizadas (Martínez, Rojas, Valdés, & Remond, 2003). Entre las técnicas más empleadas con estos fines se encuentran la regresión tanto lineal como logística, el análisis discriminante y otras técnicas multivariadas auxiliares como el análisis de componentes principales.

En Cuba el proceso de pronóstico relativos a enfermedades cerebro-vasculares (ECV) se ha realizado fundamentalmente a partir de la experiencia de los médicos relacionados con el paciente y, en muchos casos, sobre la base de escalas diagnósticas. En el Hospital Provincial ”Gustavo Aldereguía Lima”de Cienfuegos, se desarrolló un modelo de pronóstico a partir de la combinación de las técnicas de Análisis de Componentes Principales no lineal (PRINCALS) y Regresión Logística Multinomial, para el estado al egreso de pacientes con ECV. En este estudio se obtuvieron modelos con información al ingreso, a las 24 horas y a las 72 horas (Taboada $et\ al.$, 2003).

A pesar de que el modelo obtenido con la información al momento del ingreso presenta un 88 por ciento de buena clasificación, aceptable en la práctica médica, es realmente significativo el hecho de que requieren un número considerable de variables, lo cual es un elemento que afecta su introducción en este campo. Esta cuestión impone la necesidad de trabajar en la búsqueda de modelos alternativos que mantengan o mejoren su poder predictivo y reduzcan el número de variables independientes.

Teniendo en cuenta todo lo anteriormente planteado, se define como **Problema Científico** de la presente investigación: la necesidad de mejora del modelo establecido para el pronóstico del estado al egreso de los pacientes con ECV, sobre la base de la información proporcionada por

14 variables clínicas en el momento del ingreso, en cuanto a su capacidad de predictiva y/o la reducción de la cantidad de variables predictoras (Taboada $et\ al.$, 2003).

Como **pregunta de investigación se plantea:**

¿La aplicación de la correlación canónica con escalamiento óptimo (OVERALS) o de otras técnicas estadísticas, permite establecer modelos de pronóstico que mejoren la capacidad predictiva y/o la parsimonia con respecto al establecido por Bembibre-Suárez al momento del ingreso?

La presente investigación tiene como **hipótesis** de que si se elabora un modelo de pronóstico a partir del empleo de la técnica de OVERALS u otras técnicas estadísticas, es posible mejorar el modelo de pronóstico al ingreso establecido (Taboada $et\ al.$, 2003) en cuanto a su capacidad predictiva y/o la reducción de las variables predictoras empleadas.

Objeto de Investigación es el pronóstico de la evolución de los pacientes hospitalizados con ECV.

El **Campo de acción** son los modelos estadístico para la ayuda pronóstica de la evolución de los pacientes con ECV.

Objetivo General consiste en establecer al momento del ingreso un modelo para el pronóstico del estado al egreso de los pacientes con ECV, que eleve la capacidad predictiva con respecto al modelo propuesto por Bembibre-Suarez y/o reduzca la cantidad de variables predictoras en el modelo.

Objetivos específicos

- Establecer una variante del procedimiento estadístico empleado por Bembibre-Suárez con el empleo de OVERALS u otras técnicas estadísticas.
- Establecer y validar el modelo de pronóstico para el momento del ingreso.
- Comparar los resultados de los modelos obtenidos por ambos procedimientos en cuanto a porcentaje de casos clasificados correctamente, coeficiente Kappa, sensibilidad y especificidad de los modelos, errores estándar de sus coeficientes, estabilidad ante el

cambio de la muestra y cantidad de variables predictoras.

Las **tareas de la investigación** se pueden resumir como sigue:

- Revisión bibliográfica sobre el pronóstico en el campo de la salud y resumir las soluciones dadas en diferentes contextos a esta problemática.
- Adecuación de la base de datos de acuerdo al propósito de la investigación.
- Formular y validar variantes de modelos de pronósticos para el estado al egreso de pacientes con ECV que mejoren la capacidad predictiva y/o la parsimonia con respecto al obtenido por Bembibre-Suárez.
- Comparar resultados con el modelo obtenido por Bembibre-Suárez.
- Validar el modelo obtenido.
- Redactar el informe final.

El trabajo posee **novedad científica metodológica y práctica.**

El **aporte metodológico** consiste en introducir una variante en el procedimiento empleado para establecer un modelo de pronóstico del estado al egreso de pacientes con ECV, con la introducción de la técnica OVERALS, lo cual enriquece el estudio de problemas con esta naturaleza.

Desde el punto de vista **práctico** se logra el aumento de la calidad del pronóstico y el aumento de la facilidad de utilización del modelo obtenido a partir de la disminución del número de variables predictoras a emplear.

La **justificación del la investigación** está dada por la necesidad de encontrar un modelo que permitan elevar el porcentaje de pronóstico correcto del estado al egreso de pacientes con ECV y/o disminuir el número de variables predictoras, que ofrezcan una mayor certeza en la toma de decisiones y a su vez sea menos engorrosa su manipulación tendiendo en cuenta la disminución de la información necesaria para su utilización. Lo cual se revierte en una práctica médica de

mayor calidad.

La tesis está estructurada en Resumen, Introducción, tres Capítulos, Conclusiones, Recomendaciones, Referencias Bibliográficas y Anexos.

Capítulo 1: Modelos de pronósticos de uso frecuente en la salud. En este Capítulo se exponen las principales técnicas de pronóstico empleadas fundamentalmente en el campo de la salud. Se realiza una breve descripción de cada una de ellas y se referencian algunas aplicaciones de estas a problemas concretos en el campo de la salud. Por último se ofrece un conjunto de acciones que se puede seguir para establecer modelos de pronóstico.

Capítulo 2: Procedimiento para establecer el modelo de pronóstico. Una variante. El Capítulo 2 se dedica a la descripción por pasos del procedimiento empleado por Bembibre-Suárez para obtener el modelo de pronóstico, que resultó de la combinación de las técnicas de PRINCALS y la regresión logística multinomial. En un segundo momento se presenta la variante a introducir en este modelo a partir del empleo de la técnica OVERALS en lugar de PRINCALS. Se tienen en cuenta además los criterios de comparación para ambos modelos. Por último se ofrecen las conclusiones del capítulo.

Capítulo 3: Implementación del procedimiento propuesto para la obtención del modelo de pronóstico. Siguiendo el procedimiento descrito en el capítulo anterior, se obtiene el modelo de pronóstico deseado y luego de su validación se realizan las comparaciones, primeramente con un modelo obtenido mediante la técnica de PRINCALS, en el que se emplearon las mismas variables e interacciones de manera tal que pudieran compararse y por último, con el propuesto con Bembibre-Suárez. Se ofrecen además las conclusiones parciales a las que se arribó luego del análisis de los resultados alcanzados en ambos casos.

CONCLUSIONES Y RECOMENDACIONES. Las Conclusiones ofrecen los resultados obtenidos y se exponen los resultados de las comparaciones realizadas, enfatizando en el cumplimiento de los objetivos propuestos al inicio de la investigación. Las Recomendaciones se centran en potenciales mejoras al modelo obtenido y la posible aplicación en otros momentos posteriores al ingreso.

Capítulo 1

Modelos de pronóstico de uso frecuente en salud

En este capítulo se exponen de forma resumida algunas de las técnicas de pronóstico utilizadas con mayor frecuencia en la investigación, específicamente en el campo de la salud. Profundizándose en aquellas que van a ser posteriormente aplicadas para darle cumplimiento a los objetivos propuestos. Se presentan además reportes de aplicaciones, todos vinculadas a la salud.

1.1. Introducción

El desarrollo de modelos de pronóstico para predecir un fenómeno en particular no es una tarea fácil. Obtener modelos de pronóstico que presenten altos niveles de correcta clasificación conlleva a realizar un profundo estudio de problema concreto que se esté tratando. En dependencia de la propia naturaleza de este, será factible la utilización de unos u otros modelos de pronóstico. En los epígrafes siguientes se exponen algunas de estas técnicas estadísticas para la estimación de modelos de pronóstico.

1.2. Técnicas estadísticas para el pronóstico

En este epígrafe se va a abordar fundamentalmente la regresión lineal múltiple, la regresión logística y el análisis discriminante. Se expondrán los fundamentos teóricos de cada una de ellas, profundizando en la regresión logística por ser una de las herramientas a emplear en la presente investigación. Además de hacer referencia a algunas aplicaciones encontradas en la salud.

1.2.1. Regresión lineal múltiple

Los modelos de regresión permiten modelar matemáticamente la relación entre variables numéricas. Estos modelos posibilitan obtener medidas cuantitativas de relación mutua entre el conjunto de variables predictoras y la variable independiente. La variable dependiente es aleatoria y las k variables independientes o predictores son controladas. Se pueden obtener determinados modelos matemáticos para describir estas relaciones. Unos de los más populares son los lineales.

El modelo general de regresión lineal múltiple se puede expresar de la forma siguiente:

$$Y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \varepsilon_i \tag{1.2.1}$$

.

los β_j son los parámetros del modelo, que se estiman a partir de los datos de una muestra de tamaño n empleando el método de mínimos cuadrados ordinarios. La ecuación anterior expresa en forma abreviada el siguiente sistema de ecuaciones:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + ... + \beta_k x_{k1} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_{12} + ... + \beta_k x_{k2} + \varepsilon_2 \\ &..... \\ Y_n &= \beta_0 + \beta_1 x_{1n} + ... + \beta_k x_{kn} + \varepsilon_n \end{aligned} \tag{1.2.2}$$

En notación matricial compacta queda de la siguiente forma:

$$Y = X\beta + \varepsilon \tag{1.2.3}$$

donde Y es vector formado por las n observaciones de la variable dependiente, el vector β de orden $k+1$ está conformado por los coeficientes de regresión, X es la matriz de datos de orden $n \times (k+1)$, y ε es el vector de las perturbaciones aleatorias de n filas.

Los supuestos básicos del análisis de regresión múltiple son:

1. $E(\varepsilon/X_i) = 0$; con $i = \overline{1,n}$,
2. $Cov(\varepsilon_i, \varepsilon_j) = 0; \ \ \forall i \neq j$, es decir que, no existe autocorrelación entre las componentes aleatorias,
3. $\sigma_i^2 = \sigma^2; \ \ \forall i = \overline{1,n}$, supuesto de igualdad de varianza,
4. Las variables independientes son no aleatorias sino controladas,
5. No existe multicolinealidad entre las variables independientes, o sea, ninguna de las variables independientes es combinación lineal de las restantes,
6. $\varepsilon \sim N(0, \sigma^2 I)$; esto es, que el vector de las perturbaciones aleatorias sigue distribución normal multivariada con media cero y matriz de varianza $\sigma^2 I$ y
7. No existen errores de especificación.

El objetivo es encontrar el vector de los estimadores mínimos cuadrados ordinarios de los parámetros del modelo (1.2.1). Con este propósito se procede a minimizar la siguiente función:

$$min \ \ L = \sum_{i=1}^{n} (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \cdots + \widehat{\beta}_k x_{ki}))^2 \tag{1.2.4}$$

Los estimadores mínimos cuadráticos $\widehat{\beta}_0$, $\widehat{\beta}_1$,..., $\widehat{\beta}_k$ se obtienen de derivar (1.2.4) e igualarla a cero. Esto se expresa como sigue:

$$\left(\frac{\partial L}{\partial \beta}\right)\Bigg|_{(\widehat{\beta}_1, \cdots, \widehat{\beta}_k)} = 0 \tag{1.2.5}$$

Los estimadores obtenidos por (1.2.5) se obtienen a partir de la expresión:

$$\widehat{\beta} = (X^T X)^{-1} X^T Y \tag{1.2.6}$$

Si X es una matriz de rango completo, entonces el vector $\widehat{\beta}$ es único y minimiza la suma de los residuos al cuadrado (Greene, 2003).

Análisis de los residuos del modelo

La medida básica del error de predicción de un modelo de regresión es el residuo (F. Hair, Anderson, Tatham, & C. Black, 1999). El método básico para el análisis de incumplimiento de supuestos son precisamente los gráficos de los residuos. Estos pueden graficarse contra los valores pronosticados de la variable dependiente y/o las variables independientes por separado. La idea consiste en determinar si existen en estos gráficos determinados patrones de comportamiento que permitan identificar el incumplimiento de algún o algunos supuestos. También existen pruebas de hipótesis para el análisis de algunos de estos supuestos: prueba de Durbin-Watson para la autocorrelación; prueba de Glejser para la homocedasticidad ; prueba de Kolmogorov-Smirnov para la normalidad, etc.

Algunos reportes de aplicaciones de la regresión lineal múltiple

Existen algunos reportes de aplicaciones de la regresión múltiple en el campo de la salud, por ejemplo, Schneider obtiene un modelo para medir desigualdades en diferentes áreas geográficas tomando como variable dependiente la tasa de mortalidad infatil y como Independiente el Producto nacional Bruto (Schneider $et\ al.$, 2002). Otra aplicación consistió en su uso para el pronóstico del peso al nacer de los hijos de mujeres embarazadas víctimas de violencia doméstica (Helena & Aguirre, 1996). En Murcia se empleó también, pero en este caso con el fin de pronosticar el gasto farmacéutico en esa región (Córdoba, 2007). Un ejemplo, aunque no es

propiamente una aplicación lo presenta Salina , donde la presión arterial sistólica se determina a partir de variables como el peso y la edad (Salinas & Silva, 2007). No obstante, existen pocas referencias del empleo de esta técnica en la actividad clínica.

1.2.2. Regresión logística

La regresión logística, al igual que la regresión lineal, es una herramienta que permite estudiar la dependencia entre una variable dependiente y un conjunto de variables independientes o predictoras. Un elemento que las diferencia es que en la primera se tiene que la variable dependiente es categórica y las variables predictoras pueden ser númericas o categóricas. Este hecho es precisamente el que hace de la regresión logística una herramienta de mayor aplicación en determinados campos de la ciencia como el de la salud. Los supuestos básicos de la regresión logística son:

- Independencia entre las observaciones sucesivas y
- Existencia de una relación lineal entre $ln(\frac{\pi_i}{1-\pi_i})$ y las variables predictoras.

Donde $\pi(y = l/X_i)$ es la probabilidad de que cierto individuo pertenezca a la categoría l de la variable dependiente (con $l = \overline{1,c}$), dada una configuración determinada de las variables predictoras.
El modelo es el siguiente:

$$\pi(y = l/X_i) = \frac{e^{\beta_l+\beta_{1l}x_{1i}+\cdots+\beta_{kl}x_{ki}}}{1+\sum_{m=1}^{c-1} e^{\beta_m+\beta_{1m}x_{1i}+\cdots+\beta_{km}x_{ki}}} \tag{1.2.7}$$

los β_j son los coeficientes de las k variables independientes (factores) y X es el vector de los valores observados de las variables independientes asociado a cierto individuo i (con $i = \overline{1,n}$). El procedimiento seguido para determinar los coeficientes del modelo no puede ser el de mínimos cuadrado ordinarios como en la regresión múltiple debido a tres razones fundamentales (Pérez, Pliego, Lorenzo, & Tomé, 1995):

- Los residuos no se distribuyen normalmente,
- No cumplen con el supuesto de homocedasticidad y
- Las estimaciones por ese método podría conducir a valores fuera del intervalo [0,1].

En este caso los parámetros del modelo se pueden estimar por dos método diferentes, el de mínimos cuadrados generalizados y el método de máxima verosimilitud. El primero es válido solamente cuando se tienen varios valores de la variable de respuesta para una misma configuración de las variables independientes. En caso contrario debe emplearse para la estimación de los coeficientes el segundo método, que es el caso que se aborda en la presente investigación. No obstante, Hosmer emplea el método de maxima verosimilitud en ambos casos conduciendo a resultados asintóticamente coincidentes (Pérez $et\ al.$, 1995). Es método es precisamente el que se procede a exponer brevemente.

La función de verosimilitud se obtiene como sigue:

$$\vartheta(X_i) = K\prod_{i=1}^{n_1}\pi(y=1/X_i)\prod_{i=1}^{n_2}\pi(y=2/X_i)\cdots\prod_{i=1}^{n_c}\pi(y=c/X_i) \tag{1.2.8}$$

Considérese que $n_1+n_2+\cdots+n_c=n$, donde n_l indica la cantidad de individuos que pertenecen a la categoría l de la variable dependiente. La idea del método de máxima verosimilitud es que los valores estimados de los β_j son los que hacen máxima a (1.2.8). Sustituyendo en la función anterior las probabilidades por sus respectivas expresiones en función del modelo (1.2.7), y realizando una transformación logarítmica con el propósito de facilitar el procedimiento matemático de minimización queda como sigue:

$$\begin{aligned} L(\vartheta(X_i)) = lnK + n_1\beta_1 + n_2\beta_2 + \cdots + n_c\beta_c + \textstyle\sum_{i=1}^{n_1}\sum_{j=1}^{k}\beta_{1j}x_{ji} + \sum_{i=1}^{n_2}\sum_{j=1}^{k}\beta_{2j}x_{ji} + \cdots \\ + \textstyle\sum_{i=1}^{n_c}\sum_{j=1}^{k}\beta_{cj}x_{ji} + \sum_{i=1}^{n} ln(1 + e^{\beta_1+\beta_{11}x_{1i}+\cdots+\beta_{1k}x_{ki}} + \cdots + e^{\beta_c+\beta_{1k}x_{1i}+\cdots+\beta_{ck}x_{ki}}) \end{aligned} \tag{1.2.9}$$

Para determinar los valores de β_j que maximizan $L(\beta)$ se deriva (1.2.9) con respecto a cada uno de los β_j y se igualan cada una de estas ecuaciones resultantes a cero. A partir de este resultado se obtiene un sistema no lineal de c(k+1) ecuaciones con la misma cantidad de incógnitas cuya

solución puede obtenerse siguiendo algún método iterativo.

Valoración del ajuste del modelo

Una vez obtenido el modelo se debe verificar cuan buena o mala es la calidad del ajuste obtenido. Este paso es de vital importancia, debido a que si no se realizan los análisis pertinentes con este fin, los modelos obtenidos pueden conducir a pronósticos aberrantes. Para esto se han desarrollado un conjunto de criterios de evaluación. Algunos de estos criterios se exponen a continuación:

- Análisis de residuos mediante los estadísticos Chi-cuadrado de Pearson y Deviance,
- Significación de los coeficientes del modelo y
- Tablas clasificación correcta,

A continuación se exponen las particularidades de cada uno de estos criterios.

- **Chi-cuadrado de Pearson y Deviance**

Estas son medidas que miden de alguna forma la diferencia entre los valores observados y los pronósticados por el modelo. Un elemento importante que debe emplearse con este fin es lo que se conoce en la literatura como Deviance (D). Esta medida es equivalente a la suma de cuadrado de los residuos. Antes de exponer su expresión de cálculo se define que es un modelo saturado. Por modelo saturado se entiende aquel que tiene tantos parámetros como observaciones. La expresión de la Deviance es la siguiente:

$$D = -2\ln[\frac{VMA}{VMS}] \tag{1.2.10}$$

En el que VMA es la verosimilitud del modelo actual y VMS es la verosimilitud del modelo saturado. A la razón entre ambas es lo que se conoce como razón de verosimilitud.

Usando la expresión (1.2.9) la expresión anterior se convierte en:

$$D = 2\sum_{i=1}^{n_1} y_i ln\frac{y_i}{\widehat{\pi}(y_i = 1/X_i)} + 2\sum_{i=1}^{n-n_1}[(1-y_i)ln\frac{1-y_i}{1-\widehat{\pi}(y_i = 1/X_i)}] \tag{1.2.11}$$

donde n_1 representa el número de opciones de respuesta de la variable dependiente codificada con 1.

El estadístico de resumen basado en el χ^2 de Pearson tiene como expresión:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\pi}_i)^2}{\widehat{\pi}_i(1-\widehat{\pi}_i)} \tag{1.2.12}$$

La distribución de ambos estadígrafos, teniendo como hipótesis nula que el modelo ajustado es correcto, se puede ajustar a una distribución χ^2 con $n-k-1$ grados de libertad para ambos criterios. En este caso la regla de decisión está dada por $p(D > \chi^2_\alpha(n-k-1)) < \alpha$. De manera análoga se plantearía para el estadístico de Pearson.

- **Significación de los coeficientes del modelo**

Una vez que se esté seguro sobre la calidad del ajuste del modelo, se debe proceder a contrastar la significación de los coeficientes del mismo. Para ello es útil determinar la matriz de varianza-covarianza de los coeficientes estimados. Para ello se procede primeramente a calcular la matriz $I(\beta)$, conocida como matriz de información. Hosmer propone con este fin, estimar esta matriz a partir de la siguiente expresión (Hosmer, 1989):

$$\widehat{I}(\widehat{\beta}) = X^T V X \tag{1.2.13}$$

donde X es una matriz de orden $nx(k+1)$ que contiene la información para cada individuo y V es una matriz diagonal con el elemento general en su diagonal de la forma $\widehat{\pi}_i(1-\widehat{\pi}_i)$, esto es:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{1k} \\ & & \vdots & \\ 1 & x_{n1} & \cdots & x_{1k} \end{pmatrix}$$

y la matriz V es:

$$V = \begin{pmatrix} \widehat{\pi}_1(1-\widehat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \widehat{\pi}_2(1-\widehat{\pi}_2) & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \widehat{\pi}_n(1-\widehat{\pi}_n) \end{pmatrix}$$

En lo adelante se seguirá hablando de matriz de información, pero debe saberse que verdaderamente se trabaja con su estimación. La manera de determinarla se puede consultar en la propia fuente anteriormente referida. Para determinar la matriz de varianza covarianza se invierte la matriz de información y se denota por $\Sigma(\beta) = I^{-1}(\beta)$. Al j-ésimo elemento de la diagonal de la matriz anterior se le va a denotar como $\sigma^2(\beta_j)$, el cual es la varianza de $\widehat{\beta}_j$, y $\sigma(\beta_j, \beta_m)$, con $j \neq m$ para denotar a los elementos fuera de su diagonal, este valor es la covarianza entre β_j y β_m.

Los estimadores de la varianzas y covarianzas se denotan por $\widehat{\Sigma}(\widehat{\beta})$, se obtienen evaluando $\Sigma(\beta)$ en $\widehat{\beta}$. Se van a emplear las notaciones $\widehat{\sigma}^2(\widehat{\beta}_j)$ y $\widehat{\sigma}(\widehat{\beta}_j, \widehat{\beta}_m)$ para denotar los valores de dicha matriz. El error estandar de los coeficientes estimados van a estar dados por la siguiente expresión:

$$\widehat{EE}(\widehat{\beta}_j) = \sqrt{\widehat{\sigma}^2(\widehat{\beta}_j)} \tag{1.2.14}$$

Se va a emplear la notación anterior para las prueba de significación de los coeficientes.

Exiten varios criterios para verificar la significación estadística de los coeficientes del modelo (Abraira & Vargas, 1996; Hosmer, 1989). Para ello se puede proceder de la siguiente manera:

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0$

Se toma como estadígrafo de prueba:

$$Z = \frac{\widehat{\beta}_j}{\widehat{EE}(\widehat{\beta}_j)} \tag{1.2.15}$$

La región crítica asociada a esta prueba está dada por $|Z| \geq Z_{\frac{\alpha}{2}}$.

De forma equivalente se puede determinar:

$$W = \frac{\widehat{\beta}_j}{\widehat{\sigma^2}(\widehat{\beta}_j)} \tag{1.2.16}$$

Cuando la muestra es grande, el estadígrafo W tiene distribución chi-cuadrado con $n-k$ grados de libertad. La región crítica para este contraste es $W > \chi^2_\alpha$. A estos contrastes se les denominan contrastes de Wald.

La prueba multivariada de Wald propone que todos los coeficientes del modelo son iguales a cero, exceptuando al término independiente. El estadígrafo de prueba en este caso es:

$$W = \widehat{\beta}^T[\widehat{\Sigma}(\widehat{\beta})]^{-1}\widehat{\beta} = \widehat{\beta}^T[X^T V X]\widehat{\beta} \tag{1.2.17}$$

el cual tiene también una distribución chi-cuadrado con $k+1$ grados de libertad. Para determinar la significación global de los coeficientes del modelo se procede de la manera siguiente:

$$G = D(Modelo\ \ sin\ \ variables\ \ predictoras) - D(Modelo\ \ con\ \ variables\ \ predictoras) \tag{1.2.18}$$

La regla de decisión para rechazar la hipótesis nula para esta prueba es $P[\chi^2(k) > G] < \alpha$.

La prueba de Wald falla con frecuencia cuando los coeficientes son significativos. De ahí que se recomiende el uso de la razón de verosimilitud (Hosmer, 1989).

- **Tablas de clasificación correcta**

Este es un recurso que no está basada en las diferencias entre los valores observados y estimado, sino que constituyen más bien una idea intuitiva para resumir los resultados del modelo estimado. Estas tablas ofrecen los resultados cruzados de las variables de salida para los valores observados por las filas y los estimados por las columnas (o viceversa). Para determinar que a que valores de salida pertenece un individuo se establece un punto de corte, que se le puede llamar c. Para el caso de una regresión binomial, si el valor probabilístico pronósticado excede a c, entonces se considera el valor uno de la variable de salida, en caso contrario se asocia al valor cero. Generalmente este valor de corte está ubicado en $0{,}5$. En una regresión multinomial se comparan entre sí las probabilidades estimadas de pertenencia a los c grupos o categorías.
Supóngase que se tiene una variable de respuesta con tres categoría ($l = \overline{1,3}$). La probabilidad de pertenecer a la l-ésima categoría es π_l, las cuales quedarían definidas en un modelo de regresión logística multinomial de la manera siguiente:

$$\pi_1 = \frac{e^{f_1}}{1 - e^{f_1} - e^{f_2}}$$
$$\pi_2 = \frac{e^{f_2}}{1 - e^{f_1} - e^{f_2}}$$
$$\pi_3 = 1 - p_1 - p_2$$

donde f_1 y f_2 representan las funciones de clasificación para los grupos 1 y 2 respectivamente. La regla de clasificación de un individuo i en uno de los tres grupos a partir de estos modelos está dada por:

$$\pi_1 = max\{\pi_1, \pi_2, \pi_3\} \;\; indivi \in categ \;\; 1$$
$$\pi_2 = max\{\pi_1, \pi_2, \pi_3\} \;\; indivi \in categ \;\; 2$$
$$\pi_3 = max\{\pi_1, \pi_2, \pi_3\} \;\; indivi \in categ \;\; 3$$

A partir de este enfoque, se emplean las probabilidades estimadas para predecir la clasificación de los grupos. De acuerdo con esto, si el modelo es capaz de lograr porcentaje significativo

de clasificación correcta tanto global como por grupo, entonces se tiene una sólida evidencia de la calidad del ajuste del modelo. Si embargo, esto no es suficiente dado el hecho que para grupos que difieren mucho en tamaño, se va a ver favorecidos aquellos que contengan la mayor cantidad de individuos. De ahí que se pudiera analizar el porcentaje de clasificación correcta más allá del azar. En resumen, la tabla de clasificación correcta da una idea de la calidad del ajuste global de modelo, pero no debe considerarse aisladamente.

Reportes de aplicación de la regresión logística

El análisis de regresión logística ha encontrado aplicaciones en el campo de la salud con múltiples propósitos. En Cuernavaca se obtuvo un modelo logístico para pronosticar las complicaciones de las mujeres embarazadas sometidas a violencia física (Helena & Aguirre, 1996). En Ciudad de la Habana se empleó la regresión logística para buscar la relación entre la aparición de casos de dengue en manzanas identificadas por cierto grado de vulnerabilidad a esta pandemia (Martínez $et\ al.$, 2003). Se ha empleado también para pronosticar el estado al egreso de pacientes con ECV en el hospital de Cienfuegos, que es este precisamente el modelo que se pretende mejorar con esta investigación (Taboada $et\ al.$, 2003).

1.2.3. Análisis discriminante

El análisis discriminante (AD), al igual que la regresión logística, es una técnica que permite diferenciar entre dos o más grupos preestablecidos de individuos a partir de un conjunto de características o variables predictoras. A este conjunto de variable se le conoce como variables discriminantes. De manera formal se puede decir que el AD tiene dos objetivos fundamentales (Johnson & Wichern, 2002):

- Describir las diferencias existentes entre los grupos, sobre la base de los valores que tomen las variables discriminantes y,

- Construir una regla de decisión que asigne a los nuevos individuos a uno de los grupos contemplados en el análisis.

Como restricciones del AD se tiene que las variables discriminantes deben ser numéricas o permitir ser cuantificadas con un significado práctico. Esta es una diferencia importante con respecto a la regresión logística, que si permite el análisis con variables categóricas. Otro elemento que las diferencia es la forma de clasificación. Mientras que la regresión logística ofrece un valor de probabilidad de pertenencia a cierta categoría, el AD clasifica directamente a los individuos en un grupo específico. Los supuestos del AD son:

- Normalidad multivariante de las variables independientes,
- Igualdad de matrices de covarianza y dispersión de cada uno de los grupos definidos previamente para el análisis y
- No multicolinealidad entre las variables discriminantes.

Cuando no se cumple el supuesto de normalidad multivariante Hair recomienda el uso de la regresión logística (F. Hair $et\ al.$, 1999). No obstante existen otros enfoque del AD que no necesitan de este supuesto para obtener las funciones discriminantes como el de Fisher (Johnson & Wichern, 2002). Puede encontrarse además otro enfoque como el de Anderson que no parte del supuesto de igualdad de matrices de varianza-covarianza, pero si necesita del cumplimiento de la normalidad multivariada (Johnson & Wichern, 2002).

El AD tiene como punto de partida una matriz de datos que tiene en las columnas las k variables que va a emplearse para discriminar y por las filas aparecen los individuos que ofrecen las puntuaciones a cada una de las variables. El problema matemático del AD es encontrar una medida unidimensional de las variables discriminantes a partir de obtener una combinación lineal de estas. A esta combinación lineal se le conoce como función discriminante (FD) y se puede escribir como sigue:

$$D = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k \tag{1.2.19}$$

El resultado de evaluar las k variables discriminantes en la ecuación (1.2.19) es un valor que se conoce como puntuación discriminante y permite determinar a que grupo pertenece un nuevo individuo. Se van a obtener m FD, donde m=min{g-1,k}. El problema a considerar ahora es como determinar los coeficientes α_j (con $j = \overline{0,k}$) de las funciones discriminantes que mejor diferencien entre los individuos. Este aspecto se representa matemáticamente de la forma siguiente:

$$\begin{aligned} D_1 &= \alpha_{10} + \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1k}x_k \\ D_2 &= \alpha_{20} + \alpha_{21}x_1 + \alpha_{22}x_2 + \cdots + \alpha_{2k}x_k \\ &\vdots \\ D_m &= \alpha_{m0} + \alpha_{m1}x_1 + \alpha_{m2}x_2 + \cdots + \alpha_{mk}x_k \end{aligned} \tag{1.2.20}$$

Estas funciones deben discriminar al máximo los grupos. Esto es, maximizar la varianza entre los grupos y minimizar la varianza dentro de los grupos. Para ello se parte del siguiente punto. Si T_{kxk} representa la matriz de varianza-covarianza total de las variables discriminates, B_{kxk} la matriz de varianza-covarianza entre grupos y I_{kxk} la matriz de varianza-covarianza dentro de los grupos, se puede plantear que $T = B + I$.

Para obtener las FD se busca una función lineal de $(x_1, x_2, \cdots, x_k)$ $D = \alpha^T X$, de tal forma que:

$$Var(D) = \alpha^T T \alpha = \alpha^T B \alpha + \alpha^T I \alpha \tag{1.2.21}$$

esto es, la variabilidad entre los grupos más la variabilidad dentro de los grupos.

Se quiere maximizar la variabilidad entre los grupos para discriminarlos mejor. Esto es:

$$max \frac{\alpha^T B \alpha}{\alpha^T T \alpha} \tag{1.2.22}$$

o sea, maximizar la varianza entre los grupos en relación al total de la varianza. Esto es equivalente a calcular:

$$max(\alpha^T B \alpha) \qquad sujeto\ a: \ \ \alpha^T T \alpha = 1 \tag{1.2.23}$$

Luego, aplicando el método de los multiplicadores de Lagrange a la ecuación (1.2.23) se tiene que:

$$L = \alpha^T B \alpha - \lambda(\alpha^T T \alpha - 1) \tag{1.2.24}$$

si se deriva (1.2.24) con respecto a α y se iguala a cero se obtiene:

$$\frac{\partial L}{\partial \alpha} = 2B\alpha - 2\lambda T\alpha = 0 \tag{1.2.25}$$

o lo que es lo mismo:

$$B\alpha = \lambda T\alpha \tag{1.2.26}$$

La ecuación (1.2.26) implica que $(T^{-1}B)\alpha = \lambda\alpha$. Por tanto, el vector propio asociado a la primera función discriminante lo es de la matriz $(T^{-1}B)$.

Debido a que $B\alpha = \lambda T\alpha$,

$$\alpha^T B\alpha = \lambda\alpha^T T\alpha = \lambda \tag{1.2.27}$$

Entonces, si se toma el vector propio asociado el máximo valor propio, como resultado se va a obtener la función que tiene el mayor poder discriminante. El valor propio asociado a la FD indica la proporción de varianza total explicada por las m funciones discriminantes que recoge la variable D_i. Para obtener más restantes FD se buscan los vectores propios de la matriz $(T^{-1}B)$ asociados a los valores propios en orden decreciente, hasta obtener el número de FD permisibles. Debe resaltarse el hecho de que estos vectores propios son linealmente independientes y originan funciones que no están correlacionadas entre sí.

Una vez obtenidas las FD, se recomienda determinar si la separación entre las medias muestrales de los grupos difieren de manera estadísticamente significativa, ya que de otra forma puede ser infructuosa la búsqueda de una regla de clasificación (Johnson & Wichern, 2002).

Antes de presentar una regla de clasificación de individuos se definen algunos elementos que se van a emplear. Supóngase que se tienen n_1 observaciones de las k variables discriminantes $X^{(T)} = (x_1, x_2, \cdots, x_k)$ del grupo I y de igual forma n_2 observaciones del grupo II, las matrices de datos respectivos son:

$$X_1 = \begin{pmatrix} x_{11}^{(T)} \\ x_{12}^{(T)} \\ \vdots \\ x_{1n_1}^{(T)} \end{pmatrix} \qquad \text{y} \qquad X_2 = \begin{pmatrix} x_{21}^{(T)} \\ x_{22}^{(T)} \\ \vdots \\ x_{2n_2}^{(T)} \end{pmatrix}$$

Luego se determinar los vectores de media respectivos para cada uno de estos grupos $\overline{X}_{1(kx1)}$ y $\overline{X}_{2(kx1)}$. Los elementos de cada uno de estos vectores se calcula de la siguiente manera:

$$\begin{aligned} \overline{x}_{1j} &= \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij} \qquad para \;\; j = \overline{1,k} \\ \overline{x}_{2j} &= \frac{1}{n_2}\sum_{i=1}^{n_2} x_{ij} \qquad para \;\; j = \overline{1,k} \end{aligned} \tag{1.2.28}$$

En las expresiones anteriores $\overline{x}_{1j}$ y $\overline{x}_{2j}$ representan los j elementos de cada uno de los vectores de media.

De acuerdo con (1.2.19), se obtienen valores $(D_{11}, D_{12}, \cdots, D_{1n_1})$ para las observaciones del primer grupo y $(D_{21}, D_{22}, \cdots, D_{2n_2})$ para las observaciones del segundo. Se obtienen los centroides para cada uno de estos a partir de:

$$\begin{aligned} \overline{D}_1 &= \alpha^{(T)}(\overline{X}_1) \\ \overline{D}_2 &= \alpha^{(T)}(\overline{X}_2) \end{aligned} \tag{1.2.29}$$

Para el caso de dos grupos de clasificación, el punto de corte discriminante es:

$$C = \frac{\overline{D}_1 + \overline{D}_2}{2} \tag{1.2.30}$$

El criterio para clasificar el individuo i es:

$$\begin{aligned} &Si \;\; D_l < C \quad el \;\; individuo \;\; se \;\; clasifica \;\; en \;\; el \;\; grupo \;\; I, \\ &Si \;\; D_l > C \quad el \;\; individuo \;\; se \;\; clasifica \;\; en \;\; el \;\; grupo \;\; II \end{aligned}$$

Por lo general, lo que se hace es restar el valor C a la función discriminante. En este caso lo que se hace es clasificar al individuo en el grupo I si $D > C$ y en el grupo dos en caso contrario (Carvajal, Trejos, & Mejías, 2004).

Significación de las funciones discriminantes

Existen varios criterios para evaluar la significación de las FD. Uno de los más empleados es el conocido como λ de Wilks. Este se utiliza para medir de forma secuencial, el poder discriminatorio de cada una de las FD que se van obteniendo. Se comienza siempre por la de mayor poder discriminatorio y se prosigue de esa misma forma hasta la última FD que se obtuvo en el análisis. En cada etapa se plantea si se debe construir una nueva FD. Si existen todavía diferencias significativas entre los grupos se construye una nueva FD (la función obtenida discrimina bien), en caso contrario se dice que la separación obtenida ya es suficiente y, por tanto, no se necesitan más FD.

Un valor pequeño de λ implica que la discriminación de la función es buena. Esta interpretación es preferible complementarla con una prueba de significación de la FD. Para ello se emplea es estadístico V de Barlett. Si se han extraído r FD se tienen como hipótesis:

H_0: Existen diferencias significativas entre los grupos.

H_1: La separación conseguida ya es suficiente.

Es estadígrafo es:

$$V = -\{(n-1) - \frac{k+g}{2}\} \ln \lambda_r \tag{1.2.31}$$

Donde V $\sim \chi^2$ con k-r y g-r-1 grados de libertad. La expresión de cálculo de la λ de Wilks es:

$$\lambda_r = \frac{SCB}{SCT}$$

En la que SCB significa la suma de cuadrados intragrupo y SCT la suma de cuadrado total.

Reportes de aplicación del AD

Algunas de las aplicaciones del AD se ofrecen a continuación. En Ciudad de la Habana se realizó un estudio sobre el dengue y luego de identificar cluster de manzanas en cuanto a vulnerabilidad (poco vulnerable, medianamente vulnerable y muy vulnerable) a partir de condiciones ambientales y sociales, se aplicó el AD para realizar los agrupamientos resultantes (Martínez $et\ al.$, 2003). Se reporta una aplicación con el propósito de diferenciar a los individuos aparentemente sanos de acuerdo a diferentes criterios, detectándose a partir de su aplicación, que el mejor discriminate era la edad (Rodríguez, Sánchez, Blanco, Romero, & Majen, 2004). La incorporación de grasas de origen animal o vegetal a productos lácteos se considera una adulteración del producto. Pinto desarrollo una metodología para detectar adulteraciones a este tipo de productos. Aunque no es una aplicación propiamente dicha en la salud, si está relacionada directamente con ella, teniendo en cuenta los resultados perjudiciales que este tipo de prácticas puede acarrear al consumidor (Pinto $et\ al.$, 2002). En el campo de la pedagogía, sirvió para determinar que el examen de ingreso al ICFES [1] no discriminaba correctamente al ahora de pronosticar el éxito de los estudiantes en la asignatura algebra lineal (Carvajal $et\ al.$, 2004).

1.3. Algunas técnicas multivariadas auxiliares en la búsqueda de modelos de pronóstico

En este epígrafe se discutirán algunas técnicas de la estadística multivariada que pueden combinarse con las técnicas de pronóstico comentadas en epígrafes anteriores para garantizar cualidades deseables en los modelos obtenidos. La discusión se centrará en:

- Análisis de componentes principales lineal (ACP),
- Análisis de correlación canónica lineal (ACC),

[1] Instituto Colombiano de Fomento para la Educación Superior

- Análisis de componentes principales para datos categóricos (PRINCALS) y
- Análisis de correlación canónica con escalamiento óptimo (OVERALS)

A continuación se procede a discutir los elementos básicos de cada una de ella.

1.3.1. Análisis de componentes principales lineal

El ACP es una técnica que permite reducir el número de variables a un número menor o igual de componentes con la menor pérdida posible de información. La nuevas variables que resulten de la aplicación de esta técnica serán combinación lineal de las originales, y además, independientes entre sí. Estas nuevas variables ficticias (componentes o factores) son capaces de explicar gran parte de la variabilidad total de las variables originales. Para poder reducir la dimensión de un conjunto de variables, estas deben poseer cierto grado de relación lineal entre ellas, ya que de otra forma no es posible lograrlo. Este es precisamente el único supuesto que se necesita para aplicar un ACP lineal, además de trabajar con variables numéricas. El problema matemático consiste en obtener a partir de k variables iniciales un conjunto de p variables menores o iguales k. El sistema de los componentes es:

$$C_j = \alpha_j^T X \qquad donde \;\; j = \overline{1,p} \tag{1.3.1}$$

El problema radica en determinar los vectores α_j (de orden k) que permitan obtener las p componentes como combinaciones lineales de las variables originales. La matriz X de orden nxk contiene la información de las k variables independientes estandarizadas para los n individuos. Dado el supuesto que las variables originales se supone tienen información redundante, se tiene el propósito de que las p componentes que se obtengan estén incorrelacionadas entre sí. El procedimiento que se sigue es secuencial, donde se obtiene primeramente el vector α_1 correspondiente a la primera componente principal, como combinación lineal de las variables originales con máxima varianza y así sucesivamente se obtienen los restantes vectores α_j. El

interés en maximizar la varianza se debe al hecho de que la mayor información está relacionado con la mayor variabilidad. Esto es, cuanto mayor variabilidad tengan los datos, se considera que existe mayor información.

Para garantizar la unicidad de la solución, los α_i están sujetos a la restricción $\alpha_1^T\alpha_1 = 1$, para $j = \overline{1,p}$. Esta problemática se resuelve con el siguiente planteamiento:

$$\underset{\alpha_{(1)}}{max} V(C_1) \qquad sujeto\ \ a\ \ \alpha_1^T\alpha_1 = 1 \tag{1.3.2}$$

Teniendo en cuenta que $Var(C_1) = \alpha_1^T\Sigma\alpha_1$, donde Σ es la matriz de varianza-covarianza, luego se resuelve (1.3.2) mediante el método de los multiplicadores de Lagrange como sigue:

$$\underset{\alpha_{(1)}}{max} L = \alpha_1^T\Sigma\alpha_1 - \lambda(\alpha_1^T\alpha_1 - 1) \tag{1.3.3}$$

si se deriva (1.3.3) con respecto a α_1 y se iguala a cero:

$$2\Sigma\alpha_1 - 2\lambda\alpha_1 = 0 \tag{1.3.4}$$

sacando como factor común a α_1 se tiene:

$$(\Sigma - \lambda I)\alpha_1 = 0 \tag{1.3.5}$$

Esto es un sistema de ecuaciones lineales. Para que el sistema tenga una solución distinta de cero la matriz $(\Sigma - \lambda I)$ tiene que ser singular (Varela, Gandolff, García, & Peña, 2003). Esto implica que el determinante tenga que ser igual a cero, o sea:

$$|\Sigma - \lambda I| = 0 \tag{1.3.6}$$

y de este modo λ es un autovalor de Σ. Esta matriz es de orden p y si además es definida positiva, tendrá p autovalores diferentes $\lambda_1, \lambda_2, \cdots, \lambda_p$ de tal forma que, por ejemplo, $\lambda_1 > \lambda_2 > \cdots, > \lambda_p$. Desarrollando (1.3.5) se tiene que:

$$\begin{gathered}(\Sigma - \lambda I)\alpha_1 = 0 \\ \Sigma\alpha_1 - \lambda I\alpha_1 = 0 \\ \Sigma\alpha_1 = \lambda I\alpha_1\end{gathered} \tag{1.3.7}$$

entonces,

$$Var(C_1) = \alpha_1^T \Sigma \alpha_1 = \alpha_1^T \lambda I \alpha_1 = \lambda \alpha_1^T \alpha_1 = \lambda \tag{1.3.8}$$

Entonces, si lo que se quiere es maximizar la varianza de C_1 lo que tiene es que tomar el mayor autovalor λ_1 y el correspondiente autovector α_1, que es el vector de los coeficientes de las variables originales. Se puede probar que α_2 es el vector propio asociado al segundo mayor valor propio de Σ y así sucesivamente (Tusell, 2005).

En las aplicaciones de esta técnica estadística frecuentemente se trabaja con un grupo de componentes que expliquen un porcentaje considerable de la variabilidad del conjunto de datos y por estar incorrelacionadas son útiles como predictoras en modelos de regresión lineal múltiple, regresión logística, análisis discriminante, correlación canónica, etc.

1.3.2. Análisis de correlación canónica lineal

El ACC es una técnica de la estadística multivariada que facilita el estudio de las interrelaciones entre múltiples variables predictoras y múltiples variables criterio. El propósito de esta técnica está en encontrar pares de variables que sean combinación lineal de las variables de cada conjunto, de tal forma que la correlación entre ambas sea máxima. Este es un elemento importante de esta técnica, teniendo en cuenta que técnicas tan poderosas como el análisis de regresión no se puede aplicar cuando se tiene más de una variable dependiente. Para poder aplicar un ACC los datos deben ser numéricos. En caso de no serlo se realizan determinadas transformaciones. Sobre este aspecto se volverá en el próximo epígrafe.

Los supuestos básicos del ACC son:

- Linealidad de las variables involucradas en el análisis,
- Normalidad multivariante,
- Homocedasticidad y
- La no multicolinealidad entre las variables.

A continuación se ofrecen los elementos básicos de esta técnica multivariada.

La matriz X, de orden nxp representa el primer conjunto de p variables aleatorias, observadas en n individuos,

La matriz Y, de orden nxq representa el segundo conjunto de q variables aleatorias, observadas en n individuos

O sea, X e Y representan el conjunto de variables predictoras y variables criterios respectivamente. Sea el primer par de ecuaciones canónicas

$$\begin{aligned} U_k &= \alpha_{k1}x_1 + \alpha_{k2}x_2 + \cdots + \alpha_{kp}x_p \\ V_k &= \beta_{k1}y_1 + \beta_{k2}y_2 + \cdots + \beta_{kq}y_q \\ con\ \ k &= \overline{1, min(p,q)} \end{aligned} \tag{1.3.9}$$

El objetivo es estimar $\alpha_{k1}, \alpha_{k2}, \cdots, \alpha_{kp}$ y $\beta_{k1}, \beta_{k2}, \cdots, \beta_{kq}$ tal que la correlación entre U_k y V_k sea máxima (correlación canónica). Donde U_k y V_k se llaman variables canónicas. Para este propósito se establece el siguiente procedimiento.

Se determinan las siguientes matrices:

Σ_{xx} matriz de varianza covarianza de X

Σ_{yy} matriz de varianza covarianza de Y

Σ_{xy} matriz de varianza covarianza entre X e Y

La correlación entre dos variables canónicas se expresa como sigue:

$$\rho(U,V) = \frac{\alpha^T \Sigma_{xy} \alpha}{\sqrt{\alpha^T \Sigma_{xx} \alpha}\sqrt{\beta^T \Sigma_{yy} \beta}} \tag{1.3.10}$$

El propósito es maximizar la correlación entre las variables canónicas U y V con respecto a α y β, lo que equivale a maximizar ρ. También se exigen las condiciones de varianza unitaria para las variables canónicas U y V respectivamente:

$$Var(U) = \alpha^T \Sigma_{xx} \alpha = 1 \qquad y \qquad Var(V) = \beta^T \Sigma_{yy} \beta = 1 \tag{1.3.11}$$

Se utiliza el método de los multiplicadores de Lagrange teniendo en cuenta que se trata de un problema de optimización con restricciones de igualdad. Se tiene como función a maximizar:

$$L = \alpha^T \Sigma_{xy} \beta - \lambda(\alpha^T \Sigma_{xx} \alpha - 1) - \mu(\beta^T \Sigma_{yy} \beta) \tag{1.3.12}$$

Derivando (1.3.12) con respecto a α y β se obtiene como resultado:

$$\Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{xx}^{-1} \Sigma_{yx} = \lambda^2 \alpha \tag{1.3.13}$$

Esto implica que λ^2 es el valor propio asociado al vector propio de la matriz:

$$A_{pxp} = \Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{xx}^{-1} \Sigma_{yx} \tag{1.3.14}$$

De la misma forma se obtiene β como el valor propio asociado al vector propio μ^2 de la matriz:

$$A_{qxq} = \Sigma_{yy}^{-1} \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \tag{1.3.15}$$

Se tiene que λ^2=μ^2=ρ^2 que es el cuadrado del coeficiente de correlación entre las variables canónicas U y V. El vector asociado al mayor valor propio (que es el mismo para ambas matrices) proporciona el primer par de variables canónicas (primera función canónica). Solo es necesario conocer los vectores propios de una de las matrices ya que a partir de α se puede conocer β y viceversa.

Para determinar las siguientes variables canónicas se trabajan con los mayores valores propios que le siguen y así se procede de manera recursiva hasta obtener todos los pares restantes de variables canónicas que coinciden con el min (p,q). Cuando se ha realizado el procedimiento anteriormente descrito, tomando en consideración los vectores propios de las matrices simétricas son ortogonales, se obtienen las variables canónicas que cumplen lo siguiente:

- C_m es máximo,
- $Cor(V_j, V_k) = 0$ para $\forall$ j $\neq$ k,
- $Cor(U_j, U_k) = 0$ para $\forall$ j $\neq$ k,
- $Cor(U_j, V_k) = 0$ para $\forall$ j $\neq$ k,

1.3.3. Las técnicas de PRINCALS y OVERALS

Para aplicar un ACP o un ACC se parte de la restricción de que las variables involucradas en el análisis (en ambos casos) deben ser numéricas. Cuando algunas de las variables son nominales u ordinales, la aplicación directa de estas técnicas paramétricas no es posible. La idea en este caso sería crear cuantificaciones de las categorías de las variables sin violentar sus características de medida y buscando que el ajuste de las observaciones originales al modelo sea máximo. Para el caso de variables nominales, las transformaciones deben realizarse de forma tal que se conserve la pertenencia de las observaciones en cada categoría. Si las variables fueran ordinales, deben realizarse transformaciones monótonas para conservar las propiedades de orden de estas categorías. De esta forma, representando cada una de estas categorías como parámetros, se realiza el proceso de optimización cuyo resultado conduce a los denominados parámetros de escalado óptimo (Portillot, Mar, & Martínez, 2007).

El procedimiento de cuantificación de las categorías de las variables originales que tiene en cuenta estos elementos, se conoce en la literatura como escalamiento óptimo y está basado en el método de mínimos cuadrados alternantes (Vinacua, 1998; Gifi, 1985). Las técnicas ACP y ACC en las que primeramente se requiera realizar este tipo de transformaciones se les denomina PRINCALS y OVERALS respectivamente.

En ambos casos el criterio para el escalamiento que se sigue es que las puntuaciones de los individuos de la muestra en las dimensiones a obtener en las solución, tengan una relación lo más alta posible con cada una de las variables del análisis y que ya han sido previamente cuantificadas. Una vez cuantificadas las categorías de las variables originales, se obtienen combinaciones lineales de estas óptimamente escaladas (cuantificadas), esto es, variables que garanticen un mejor ajuste del modelo.

Estas técnicas van a ser utilizadas mediante el empleo de paquetes de programas estadísticos como el SPSS, que permite obtener las cuantificaciones de las categorías de las variables y las dimensiones pedidas; lo cual posibilita el uso de estos resultados en otros análisis estadísticos.

1.3.4. Segmentación jerárquica. El algoritmo CHAID

La segmentación jerárquica es un método que utiliza un algoritmo basado en criterios que identifican grupos homogéneos de una población. El proceso que emplea es de división secuencial, iterativo y descendente, mediante la definición de una variable dependiente discreta. La idea es ir construyendo una estructura de representación de la información similar a un árbol. La raíz del árbol es el conjunto de la totalidad de los datos, los subconjuntos conforman las ramas del árbol. Un conjunto en el que se hace una partición se denomina nodo. El árbol de decisión se construye al dividir el conjunto de datos en dos o más subconjuntos de observaciones a partir de los valores que toman las variables predictoras. Cada uno de estos conjuntos vuelve después a ser particionados mediante el mismo algoritmo. Este proceso continúa hasta que no se encuentren diferencias significativas en la influencia de las variables de predicción de uno de estos grupos hacia el valor de la variable de respuesta.

La segmentación jerárquica, y más particularmente el algoritmo CHAID, puede utilizar variables nominales, ordinales y continuas y actúa por disgregación politómica de las categorías o clases. Esto significa que es capaz de en un mismo nodo de obtener más de una rama (Santín, 2006). Este se diferencia de otros algoritmos similares como el Detector Automático de interacción (AID) que solo es capaz de lograr una segmentación binaria (Mangin, Alonso, & Clavel., 2002). De manera general, la segmentación jerárquica permite descomponer una base de datos en varios grupos con base en el mejor predictor de la variable dependiente (en función de la prueba de la más significativa). El resultado es un ”valor-p”. Este valor representa la probabilidad de que la relación sea estadísticamente significativa. A continuación, los valores-p para cada tabulación cruzada de todas las variables independientes se clasifican, y si el mejor (el valor más pequeño) se encuentra bajo un nivel determinado, se realiza una ramificación del nodo raíz en esa ubicación. La segmentación permite determinar la mejor partición de cada nodo y aparejar clases o categorías de las variables independientes predictoras. Se repite el proceso hasta que no queden más parejas significativas. Seguidamente se vuelve a seleccionar el nuevo

mejor predictor y el grupo (nodo) será dividido en dos o más clases. Se repetirá el proceso hasta que se llegue a la regla de fin.

Cuanto más se alarguen las ramas menos variables independientes habrá disponibles ya que el alargamiento se lleva a cabo precisamente con las variables. La ramificación llega a su fin cuando el mejor valor-p ya no es menor que el α prefijado. Los nodos, hoja del árbol, son aquéllos que no han sufrido ramificaciones.

Lo anteriormente planteado de manera formal es (Santín, 2006):

Sea una variable dependiente d con al menos dos categorías ($d \geq 2$) y sea una variable explicativa x con $c \geq 2$ categorías. El primer objetivo es reducir la tabla de contingencia cxd actual a la tabla más significativa jxd a partir de la combinación de categorías de x. Para ello se calcula el conjunto $T_j^* = max_i T_j^i T_j^{(i)}$ formado por los estadísticos de los i métodos posibles de configurar una tabla de contingencia $j \times d = \overline{2,c}$. Finalmente se escoge $T_j^* = max_i T_j^i$ que representa el estadístico χ^2 más elevado para la tabla de contingencia $j \times d$. La búsqueda de T_j^* cuando se dispone de varias variables explicativas se lleva a cabo de manera secuencial aplicando el siguiente algoritmo.

1. Para cada variable x se calculan las tablas de contingencia a partir de las categorías de x y de la variable dependiente d.

2. Buscar el par de categorías de x cuya tabulación cruzada $2 \times d$ es menos significativa. Si esta significación no alcanza un determinado valor crítico, por ejemplo que sea estadísticamente significativa al 95 por ciento, ambas categorías se funden y son consideradas como una nueva categoría compuesta independiente.

3. Para cada categoría compuesta formada por 3 o más de las categorías originales, se busca la división binaria más significativa. Si alcanza un valor crítico determinado se parte la categoría y se pasa al punto 2.

4. Calcular la significación de todas las variables explicativas compuestas por sus categorías

óptimas y elegir la más significativa. Si este valor es mayor que un valor criterio se subdividen los datos de acuerdo a las categorías de esta variable.

5. Para cada partición de los datos generada que no ha sido analizada, comenzar de nuevo con el paso 1. Este nuevo análisis termina al excluir particiones que no superan un número mínimo de observaciones definidas por el analista o un p-valor crítico de división.

1.4. Acciones para establecer un modelo de pronóstico

Cuando se tiene como propósito estimar un modelo, cuya finalidad sea emplearlo con fines predictivos, se pueden sugerir un grupos de acciones válidas independientemente de la naturaleza del problema que se esté abordando. En el siguiente subepígrafe se resumen las acciones fundamentales del proceso de obtención y validación de los modelos.

1.4.1. Secuencia de acciones a seguir

A continuación se exponen los elementos de cada una de estas acciones:

1. Determinación del problema, objetivos, población, muestra, variables a emplear y su nivel de medición,
2. Determinación de los procedimientos estadísticos a emplear,
3. Estimación del modelo de pronóstico, valoración del ajuste y análisis del cumplimiento de los supuestos,
4. Validación e interpretación de los resultados y
5. Aplicación del modelo obtenido con fines pronósticos.

Ante todo, se debe tener claro el problema a enfrentar y los objetivos propuestos previamente, así como las variables, la población estudiada y la muestra analizada, de forma tal que no

exista ambig$\ddot{u}$edad durante el transcurso de la investigación y la valoración del cumplimiento de los objetivos propuestos sea posible. Con estos elementos claramente especificado, se puede seleccionar la técnica más adecuada a emplear para obtener el modelo de pronóstico.

Luego se pasa a considerar los procedimientos estadísticos a aplicar partiendo de la naturaleza del problema a tratar y el tipo de información que se disponga. En función de estos elementos se opta por uno u otro procedimiento para obtener el modelo de pronóstico deseado.

Una vez estimado el modelo, debe valorarse su significación estadística. Para ello se toman como referencia las herramientas que se adecúen a la técnica en particular con la que se esté tratando. La idea consiste es determinar el grado de ajuste de forma tal que se pueda realizar pronósticos confiable a partir del modelo obtenido.

Otro paso primordial consiste en verificar si cumple con los supuestos establecidos para su correcta aplicación. En el caso de no cumplirse, valorar el posible impacto de incumplimiento de algún supuesto en particular, así como alguna posible medida remedial a aplicar en caso de ser posible. En este punto debe considerarse además si se cumplen determinados requerimientos basados en la experiencia de investigaciones precedentes.

Toda vez que se ha considerado aceptable el ajuste del modelo y restantes elementos, se deben validar para determinar su estabilidad y, de esta forma, poder garantizar su generalización con fines predictivos. Para esto se puede proceder a seleccionar una muestra adicional si es posible o, por el contrario, dividir la muestra si su tamaño lo permite y comparar si existen diferencias significativas entre uno y otro. Los criterios de comparación dependen de la técnica y particular y de los propósitos de la investigación. Otro aspecto a tener en cuenta es la significación práctica (y no solo estadística) de modelos. Esto se refiere a determinar, por ejemplo, si todas las variables incluidas en el modelo son realmente significativas, su significación está adecuadamente reflejada en las ponderaciones asignadas a estas por el modelo. Pudiera considerase además la inclusión de alguna variable que estadísticamente no sea significativa pero desde el punto de vista práctico no debe ser marginada.

Como último paso se aplica el modelo obtenido con fines predictivos y de esta forma se le da

cumplimiento a los objetivos trazados el inicio del proceso investigativo. Es importante conocer las limitaciones del modelo obtenido, de manera que no se utilicen indiscriminadamente. Aquí debe considerarse el contexto en que se realizó la investigación, ya que si las condiciones condiciones bajo las que se obtuvieron, la utilidad de los mismos puede verse seriamente afectada. En este caso habría que recalcularlo con la nueva información que se disponga o, posiblemente buscar nuevos modelos alternativos.

1.4.2. Algunos recursos empleados en la validación de modelos de pronóstico

Como recurso adicional para validar la calidad del ajuste de un modelo de regresión logística pueden emplearse además de los propios desarrollados para este modelo, el coeficiente Kappa y los análisis de sensibilidad y especificidad. A continuación se discuten cada uno de estos recursos

- **Coeficiente Kappa**

Teniendo en cuenta el análisis realizado sobre la clasificación correcta cuando la composición de los grupos es significativamente diferente, es posible determinar un coeficiente que corrige el porcentaje de clasificación correcta que pueda deberse al azar. Si la variable de respuesta tiene c categorías. El coeficiente Kappa se determina como:

$$K = \frac{p_o - p_e}{1 - p_e} \tag{1.4.1}$$

siendo

$$p_o = \frac{CA}{CA + CD} \tag{1.4.2}$$

y

$$p_e = \sum_{i=1}^{c} p_{oi} p_{ei} \tag{1.4.3}$$

donde:

CA: cantidad de acuerdos en la clasificación,

CD: cantidad de desacuerdos en la clasificación,

i: número de la categoría de la variables dependiente ($i = \overline{1, c}$),

p_{oi}: proporción de ocurrencia observada en la categoría i de la variable dependiente,

p_{ei}: proporción de ocurrencia esperada en la categoría i de la variable dependiente.

Un criterio para valorar el valor del coeficiente Kappa es el siguiente (Latour, Abraira, Cabello, & Sánchez, 1997):

Kappa	Grado de concordancia
0.81-1.00	Excelente
0.61-0.80	Buena
0.41-0.60	Moderada
0.21-0.40	Ligera
0.00-0.20	Mala

- **Análisis de sensibilidad y especificidad**

Los indicadores de sensibilidad y especificidad del modelo permiten reforzar el análisis de sobre la capacidad de pronóstico del modelo y, por ende, su validez. Aunque estos indicadores son empleados para valorar la calidad de pruebas diagnósticas, pueden extrapolarse para el caso del pronóstico. Para el caso del diagnóstico, se entiende por sensibilidad la probabilidad de la prueba empleada de detectar los enfermos, mientras que la especificidad se relaciona con la probabilidad de detectar a los individuos sanos (Altman & Bland, 1994).

De manera general, si se tienen en cuenta dos categorías de la variable de salida (V y F), los resultados permiten clasificar a los sujetos en cuatro grupos según la tabla siguiente:

	Pronóstico verdadero	
Resultados del pronóstico	V	F
V	VV	VF
F	FV	FF

La sensibilidad puede interpretarse, por ejemplo, como la probabilidad de clasificar correctamente a un individuo en F. Esta se puede entender como la capacidad del modelo de pronosticar correctamente la pertenencia de un individuo a la categoría F. La fórmula de cálculo sería la siguiente:

$$Sensibilidad = \frac{FF}{FF + VF} \tag{1.4.4}$$

La especificidad sería la probabilidad de clasificar a un individuo en V, esto es, la capacidad del modelo para pronosticar los individuos que pertenecen a la categoría V. Se calcularía como:

$$Especificidad = \frac{VV}{VV + FV} \tag{1.4.5}$$

De acuerdo a las categorías de salida de la variable se puede definir para cada contexto en específico la sensibilidad y especificidad del modelo.

1.5. Conclusiones del capítulo

1. El desarrollo vertiginoso de la informática y por ende, la disponibilidad de softwares especializados cada vez más potentes y de fácil uso, hace que técnicas de pronóstico con algún grado de complejidad sean empleados frecuentemente en las investigaciones. De ahí que se reporten múltiples aplicaciones en tan disímiles áreas de investigación.

2. La regresión logística es un modelo de especial interés debido al hecho de que permite trabajar con variable dependientes categóricas y de ahí su amplio uso en el campo de la salud.

3. El análisis de correlación canónica no lineal puede ser aplicado con el propósito de obtener nuevas variables que sean combinaciones lineales de las originales, y estás nuevas variables (variables canónicas del primer conjunto), pueden ser utilizadas como variables predictoras en un modelo de regresión logística.

4. La combinación de las técnicas de análisis de correlación canónica no lineal y la regresión logística; de la cual no hay reportes de su aplicación para el pronóstico, constituye un recurso potencial para pronosticar el estado al egreso de pacientes con ECV, teniendo en cuenta que las primeras combinaciones de las variables predictoras tienen correlación máxima con el conjunto de variables criterios.

Capítulo 2

Procedimiento para establecer el modelo de pronóstico. Una variante

2.1. Introducción

Este capítulo se dedica a la descripción del procedimiento seguido por Bembibre-Suárez para establecer los modelos para pronosticar la perspectiva de la evolución de los pacientes hospitalizados con ECV, teniendo en cuenta las variables clínicas al ingreso. Se exponen los pasos del procedimiento empleado y se describe además la muestra de casos utilizada para la obtención de tales modelos, los criterios para la selección de las variables predictoras, incluidas las obtenidas como interacciones de las variables originales; el uso de la regresión logística multinomial sobre los componentes obtenidos por un análisis de componentes principales categórico y los criterios seguidos para validar el modelo obtenido. Por último se presenta la variante propuesta al procedimiento para la obtención del modelo de pronóstico y los criterios a seguir para la comparación de los modelos obtenidos por ambos procedimientos.

2.2. Descripción del procedimiento propuesto por Bembibre-Suárez

A continuación se exponen los elementos considerados por Bembibre-Suárez desglosado en cada una de las etapas mencionadas anteriormente.

2.2.1. Selección de la muestra y variables participantes

La muestra empleada para obtener el modelo de pronóstico del estado de los pacientes en una primera fase de la investigación fue la totalidad de los pacientes hospitalizados en cuidados intensivos en el hospital provincial ¨Gustavo Aldereguía Lima¨ de Cienfuegos (1318 pacientes), en el período del 1-7-2000 al 31-1-2003.

En una primera etapa de la investigación se trabajó con aquellas variables clínicas que mostraban relación con la variable dependiente ¨estado al egreso¨. Las variables escogidas en esta fase fueron: conciencia, parálisis facial, fuerzas motoras de los miembros derechos e izquierdos, mirada (desviación conjugada de los ojos), sensibilidad, lenguaje, tono, reflejos osteotendinosos y babinski. Posteriormente se incluyen las variables ataxia y nistagmo en interacciones. A continuación se exponen las escalas de medición para cada una de ellas:

- conciencia (1-vigilia, 2- somnolencia, 3-estupor, 4-coma),
- parálisis facial (1-no parálisis, 2- parálisis incompleta, 3- parálisis completa, 4-no procede)
- fuerzas motoras (5- normal; 4- movimiento activo, disminución de la fuerza, mueve las articulaciones con una fuerza mayor que contra la gravedad, por sí solo; 3- movimiento activo, sólo mueve las articulaciones contra la gravedad; 2-movimiento activo, no vence la fuerza de gravedad; 1-contracción muscular visible o pequeño movimiento; 0- ausencia de movimiento visible o contracción en cualquier músculo),
- mirada (1- no desviación, 2- si desviación, 3-no procede),

- sensibilidad (1-normal, 2-pérdida parcial, 3-pérdida densa, 4- no procede),
- lenguaje (1-articula normal, 2-disartria ligera, 3-disartria severa, 4-disfasia, 5-afasia, 6-no procede),
- tono (1-normal, 2-flacidez, 3-hipertonía, 4-no procede),
- reflejos osteotendinosos (1-normal, 2-hiporreflexia, 3-hiperreflexia, 4-ausentes, 5-no procede),
- babinski (1-reflejo no presente, 2-si presente, 3-no procede),
- ataxia (1-si, 2-no y 3-no procede) y
- nistagmo (1-si, 2-no y 3-no procede)

El modelo de regresión logística estimado directamente a partir de estas variables no resultó satisfactorio en cuanto a su poder predictivo y, además, muestra coeficientes con grandes errores estándar, razones suficientes para buscar variantes más adecuadas. El problema de los grandes errores estándar de los estimadores de los parámetros de los modelos de regresión logística multinomial, se resolvió definitivamente con el empleo de un análisis de componentes principales categóricos, en correspondencia con la naturaleza cualitativa de los datos registrados y el problema del bajo poder predictivo se resuelve parcialmente. En este caso particular, con el uso de otras variables derivadas de las originales y obtenidas por dos vías: en un primer caso se construyen variables dicotómicas con algunas categorías de algunas variables clínicas originales, como es el caso de las fuerzas motoras en las categorías 4 o 5, que muestra una relación fuerte con el estado de los pacientes al egreso, cuestión que no se aprecia cuando se consideran estas variables con todas sus categorías; y en un segundo caso, se construyen interacciones de ordenes superiores con el empleo del detector automático de interacciones Chi-cuadrado (CHAID). En esencia, primero se aplica un análisis de componentes principales categóricos sobre todas las variables (originales seleccionadas o derivadas de las originales) y sobre los componentes

obtenidos se aplica la regresión logística multinomial. El procedimiento en detalle se expone a continuación.

Procedimiento estadístico

El procedimiento que definitivamente se aplicó para cumplimentar el objetivo propuesto es el siguiente:

1. Determinación de las variables mejores relacionadas con el estado al egreso y derivación de nuevas variables determinadas por las interacciones significativas, por la combinación de categorías influyentes de las variables originales.

2. Realización de un análisis de componentes principales categóricos a partir de las variables obtenidas en el primer paso.

3. Determinación del modelo de regresión logística multinomial para pronosticar el estado del paciente al egreso en las categorías: vivo sin discapacidad (1), vivo con discapacidad (2) y fallecidos (3), a partir de las variables independientes constituidas por las dimensiones obtenidas en el escalamiento anterior y empleando como variable dependiente el estado del paciente al egreso en sus tres categorías: vivos sin discapacidad (vsd), vivo con discapacidad (vcd) y fallecidos. Este modelo se calcula para toda la población considerada en el estudio.

4. Validación interna de los modelos obtenidos mediante el recálculo de los modelos con una muestra aleatoria del 50 % de los casos y el cálculo del coeficiente Kappa para establecer la estabilidad del modelo y su capacidad de pronóstico.

5. Validación de los resultados obtenidos con un conjunto de datos no participantes en la búsqueda de los modelos, empleando para ello el coeficiente Kappa entre la clasificación lograda por los modelos obtenidos en tres categorías (vsd, vcd y fallecidos) con respecto al estado real de los pacientes al egreso; además de realizar un estudio de correlaciones con otras escalas existentes.

2.2.2. Búsqueda de interacciones significativas

Una vez determinadas las variables clínicas relevantes con fines pronósticos, se procedió a establecer algunas variables derivadas de las variables originales:

Inicialmente se incluyen 4 variables referentes a fuerzas motoras: Fuerza motora superior derecha (FMSD), Fuerza motora superior izquierda (FMSI), Fuerza motora inferior derecha (FMID) y Fuerza motora inferior izquierda (FMII). Estas variables son de suma importancia en la evolución de estos pacientes y sin embargo algunas de ellas apenas correlacionan con la variable ¨estado al egreso". Después de algunos análisis exploratorios, de estas 4 variables de derivaron las siguientes 4 (Surí, 2004):

- fmd 45:Todas las fuerzas musculares derechas en las categorías 4 o 5,
- fmi 45:Todas las fuerzas musculares izquierdas en las categorías 4 o 5,
- fm45:Todas las fuerzas musculares en las categorías 4 o 5 y
- fm01:Alguna fuerza muscular en las categorías 0 o 1.

Búsqueda de interacciones con el CHAID

Se aplicó el procedimiento CHAID sobre las variables originales seleccionadas y del árbol obtenido se tomaron aquellas interacciones que tenían una fuerte relación con la variable estado al egreso, estas interacciones se someten al criterio de un grupo de expertos para establecer su significación en la actividad clínica, además de la significación estadística ya probada. Con las interacciones que reciben la aprobación de los expertos se definen variables dicotómicas que asumen valor uno si satisfacen la proposición lógica correspondientes y cero en caso contrario.

Las interacciones consideradas como significativas fueron:

- Inti1: Estado normal en: conciencia, mirada y tono, con alguna de las fuerzas musculares en categoría menor o igual a 4, pero las fuerzas izquierdas son siempre 4 o 5. (Mayoritariamente son vivos sin discapacidad (vsd).

- Inti2: Estado normal en: conciencia y mirada, con $tono \geq 2$. (Mayoritariamente son vivos con discapacidad (vcd)).
- Inti3: Conciencia con categoría 3 o superior. (Mayoritariamente son fallecidos (fall).
- Inti4: Conciencia=2 y $mirada \geq 2$. (Mayoritariamente son vcd)).
- Inti5: Mirada =2 y $sensibilidad \geq 2$ (Mayoritariamente son fall).
- Inti6: Mirada =2 y sensibilidad =1 (Mayoritariamente son vcd).
- Inti7: Estado normal en: mirada, tono, babinski, sensibilidad y parálisis facial, con lenguaje =2, nistagmo presente (1) y ataxia ausente. (Mayoritariamente son vsd).
- Inti9: Mirada=1 y tono=1 y (todas las fuerzas al ingreso en 5) y sensibilidad=1 (v.s.d).

2.2.3. Aplicación de PRINCALS sobre las variables predictoras al ingreso

En los análisis de componentes principales realizados se empleó el método de normalización principal por variables y la cantidad de dimensiones se decidió por el criterio del máximo número de componentes posibles con valores propios mayores que 1 y buscando la mayor cantidad posible de varianza explicada.

2.2.4. Empleo de la regresión logística multinomial para el cálculo del modelo de pronóstico

La justificación del empleo de la regresión logística para determinar el modelo de pronóstico se debe a la naturaleza de la variable dependiente (estado al egreso). El objetivo consiste en clasificar a los pacientes en uno de los grupos: 1- Vivo sin discapacidad (vsd) ; 2- vivo con discapacidad (vcd); 3-Fallecido (fall). Como variables independientes se emplearon las 10 primeras componentes resultado de la aplicación de PRINCALS, comentado en el epígrafe

anterior. Del modelo obtenido, en una primera corrida, se eliminaron las dimensiones que no resultaron estadísticamente significativas.

2.2.5. Validación de los resultados

Con el objetivo de validar la estabilidad del modelo se procedió al recálculo de este con la mitad de los datos tomados aleatoriamente, para lo cual se segmentó el archivo de datos mediante una muestra aleatoria del 50 % de los casos. Las comparaciones realizadas entre el modelo obtenido con la mitad de las observaciones y con toda la muestra se realiza teniendo en cuenta: porcentaje de clasificación correcta, coeficiente Kappa, magnitud de los coeficientes estimados y sus errores estándar. Tales comparaciones permitieron establecer la estabilidad de los mismos y dar paso a la fase final de validación con la información de 444 pacientes, registrados en el período 1-2-2003 al 31-8-2003 que no participaron en los análisis para establecer los modelos de pronóstico. De ahí que se consideren estable el modelo y se acepte su utilización en el pronóstico (Surí, 2004). Los resultados del modelo obtenido con toda la muestra se ofrecen en el Anexo A.

2.3. Variante introducida en el procedimiento

Es este epígrafe se van a exponer los aspectos relacionados con la nueva variante propuesta, para la búsqueda modelo para el pronóstico de la evolución de pacientes con ECV. Primeramente se expone el uso de la técnica de OVERALS para determinar las variables predictoras a emplear en la regresión logística y posteriormente la discusión de una prueba analítica para la estabilidad de los coeficientes del modelo.

2.3.1. Empleo de la técnica de OVERALS

En correspondencia con los objetivos propuestos en esta investigación, se pretende aplicar una variante en el procedimiento, consistente en la aplicación de OVERALS en sustitución del

PRINCALS, utilizando como segundo grupo de variables las variables al egreso, las cuales no fueron utilizadas en el procedimiento anterior. Esta variante está basada en el hecho probado que las variables del segundo grupo (estado de las variables clínicas al momento del egreso) correlacionan más fuertemente con el estado al egreso que las variables al momento del ingreso y son portadoras de información que no es recogida totalmente por la variable estado al egreso. Las dimensiones obtenidas en mediante el empleo OVERALS son combinaciones lineales de las variables del primer grupo que maximizan las correlaciones respectivamente con las dimensiones del segundo grupo, absorbiendo mejor de esta forma las correlaciones descritas anteriormente, con su consiguiente influencia en la capacidad predictiva de los modelos de regresión logística estimados a partir de tales dimensiones. Se espera también que el establecimiento de nuevas interacciones preferentemente que, preferentemente discriminen a los pacientes del grupo vsd, contribuyan al aumento de la capacidad predictiva de los modelos estimados o la disminución del número de variables independientes a emplear.

El empleo de OVERALS se realizará a partir del conjunto de variables clínicas que se consideren para la estimación del modelo, luego de realizar algunos análisis exploratorios. Para decidir sobre las variables que participarán inicialmente de forma directa en el proceso de búsqueda de los modelos, se analizan los coeficientes de correlación de Spearman y las correspondientes tablas de contingencia entre las diferentes variables clínicas (tanto al ingreso como al egreso) con respecto a la variable estado el egreso, se considerará además; el aporte de las variables a las dimensiones canónicas obtenidas. El conjunto de variables seleccionadas puede ser enriquecido con variables que aunque no correlacionan bien, generan interacciones significativas.

Dentro del primer conjunto (compuesto por las variables clínicas al ingreso y sus interacciones) se van a considerar solamente aquellas que se emplearán definitivamente en el modelo de pronóstico, mientras que en el segundo conjunto se pueden incluir incluso aquellas variables que no necesariamente se van a emplear con fines predictivos. Una vez obtenidas las dimensiones del primer conjunto mediante OVERALS, se procede a realizar la regresión logística multinomial.

Estimado ya el modelo se pasa a validarlo. Para ello se va a proceder de manera análoga a

como anteriormente se hizo, con la diferencia de que en este caso solamente se va a realizar la validación cruzada con aproximadamente la mitad de las observaciones y se aplicará una prueba analítica para la estabilidad del modelo estimado respecto a la magnitud de los parámetros estimados, con el empleo del enfoque de la variable dicotómica, la cual en este caso, clasifica los individuos en cada una de las mitades de la muestra (Gujarati, 2005). Para ello se va a seleccionar una muestra aleatoria de la mitad de las observaciones y se valorará la estabilidad de los coeficientes.

2.3.2. Prueba de estabilidad mediante el enfoque de la variable dicotómica

El planteamiento de la prueba de estabilidad con el empleo de una variable indicadora para el caso de la RLM es el siguiente (Gujarati, 2005): suponga se tienen k variables independientes: $X_1, X_2, \cdots, X_k$ y la variable indicadora D, resultando el modelo:

$$y_i = \alpha_1 + \alpha_2 D_i + \beta_1 X_{1i} + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} D_i X_{1i} + \beta_{k+2} D_i X_{2i} + \cdots + \beta_{2k} D_i X_{ki} + \mu_i \tag{2.3.1}$$

con $i = \overline{1, n}$.

El primer grupo de casos en la muestra tiene n_1 casos y el segundo n_2 casos, con $n_1 + n_2 = n$, así la variable dicotómica D toma el valor 1 para los casos del primer grupo y el valor cero para los casos del segundo grupo. Luego si en el modelo obtenido según (2.3.1) se considera el caso $D = 0$, se obtiene el modelo para el segundo grupo, el cual es:

$$(y_i / D = 0, x_j) = \alpha_1 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i \tag{2.3.2}$$

si se considera el caso $D = 1$, se obtiene el modelo para el primer grupo, a saber:

$$(y_i / D = 1, x_j) = \alpha_1 + \alpha_2 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \beta_{k+1} X_{1i} + \beta_{k+2} X_{2i} + \cdots + \beta_{2k} X_{ki} + \mu_i \tag{2.3.3}$$

el cual se puede expresar como:

$$(y_i/D = 1, x_j) = (\alpha_1 + \alpha_2) + (\beta_1 + \beta_{k+1})X_{1i} + (\beta_2 + \beta_{k+2})X_{2i} + \cdots + (\beta_k + \beta_{2k})X_{ki} \quad (2.3.4)$$

En este modelo el coeficiente α_2 indica la intersección diferencial, es decir el incremento con respecto a la intersección del modelo del segundo grupo, y los coeficientes β_{n+j}, con $j = \overline{1,k}$ indican la pendiente diferencial, es decir, en que medida difieren los coeficientes de las variables independientes en ambos modelos. Luego si hay coeficientes del modelo (2.3.1) estadísticamente significativos, se puede afirmar que hay cambios en los modelos de ambos grupos de datos y no se puede aceptar la estabilidad de los modelos.

Para extender esta prueba a la regresión logística, se trabajará con los modelos de regresión logística linealizados. Se crea una variable dicotómica de selección (D) donde los casos seleccionados como muestra aleatoria del 50 % se le asigna el valor 1 y el resto el valor cero. Se corre el modelo de regresión logística multinomial sobre las variables independientes $(dim_1; dim_2; \cdots ; dim_8)$, sobre los productos de cada variable independiente y la variable dicotómica de selección D: $(dim_1.D_i; dim_2.D_i; \cdots ; dim_8.D_i)$ y sobre la variable D_i. Si algunos de los coeficientes de este último grupo de variables en los modelos obtenidos son estadísticamente significativos , se asume que no existe estabilidad de los coeficientes en los modelos de ambos grupos de casos.

2.4. Criterios para la comparación de modelos

Una vez realizado el proceso de validación, se va a realizar un análisis comparativo entre los modelos en estudio. Los criterios a emplear en este análisis va a ser, fundamentalmente los siguientes:

- por ciento de clasificación correcta,
- número de variables predictoras,

- errores estándares en los coeficientes de lo modelos,
- índice Kappa,
- estabilidad y comparación descriptiva de los coeficientes y
- sensibilidad y especificidad de ambos modelos.

La idea de comparar ambos modelos a partir de una serie de criterios posibilita poder realizar un análisis comparativo más exhaustivo y, de esta forma, conocer las bondades de cada uno de estos. El primer indicador es un elemento primordial a la hora de considerar un modelo para el pronóstico. Es evidente que mientras mayor sea este valor, pues más atractivo es el modelo. No obstante el análisis de los restantes no se debe descartar. Si se obtienen modelo con similares porcentajes de pronóstico correcto, es mucho más deseado aquel que emplee menor cantidad de variables predictoras. Un modelo más parsimonioso hace que este sea mejor aceptado en la práctica dado el hecho de que, en muchas ocasiones, se hace realmente engorrosa la medición de un gran número de variables.

2.5. Conclusiones del capítulo

1. Las variantes introducidas al procedimiento para establecer los modelos de pronóstico constituye un intento en la búsqueda de mejores modelos de pronóstico para el estado de pacientes con ECV, en cuanto a sus capacidad de pronóstico y la parsimonia de los mismos, aplicables en otros tipos de enfermedades.

2. La exploración de nuevas interacciones pueden mejorar el porcentaje de pronóstico correcto para el primer grupo de pacientes (vsd).

3. Se introduce la aplicación de la prueba de estabilidad con el empleo de una variable indicadora lo cual constituye un elemento que valida la extrapolación de los modelos obtenidos a otros grupos de pacientes.

Capítulo 3

Implementación del procedimiento propuesto para la obtención del modelo de pronóstico

En este capítulo se van a presentar todos los resultados relacionados con el proceso de búsqueda y obtención del modelo de pronóstico para el estado al egreso de pacientes con ECV, sobre la base de un conjunto de variables clínicas al momento del ingreso, siguiendo el procedimiento expuesto en el capítulo anterior. Una vez determinado el modelo que se considere adecuado de acuerdo a los objetivos propuesto, se procede a su validación mediante el enfoque de validación cruzada. Luego se pasa a realizar la comparación entre el modelo propuesto sobre la base del conjunto de criterios previamente especificados con uno obtenido a partir de las mismas variables e interacciones consideradas pero aplicando PRINCALS, de tal forma que puedan realizarse comparaciones (teniendo en cuenta que ambos son modelos equivalentes) y por último, con los resultados alcanzados por el modelo propuesto por Bembibre-Suárez.

3.1. Análisis de las variables participantes en el modelo

Antes de realizar cualquier análisis de los anteriormente expuestos, se realiza una revisión de la base de datos con el propósito de disminuir las incidencias que pueden tener algunas observaciones particulares en la calidad del modelo obtenido, específicamente mediante la unión de algunas categorías de las variables clínicas en el momento del ingreso. Las categorías que se

unieron fueron aquellas que tenían muy pocos casos registrados. Este análisis se realizó a partir de la observaciones en tablas de contingencia de cada una de estas con respecto a la variable egreso. Los resultados de estas observaciones son los siguientes: se eliminó la categoría 3 de la variable orientación por no presentar ningún caso con este valor; se unieron las categorías 3 y 4 de parálisis facial, sensibilidad y tono y las 4 y 5 de la variable lenguaje.

Se modifican además en las variables al egreso, los valores de estas relacionados con pacientes que no presentan mediciones bien porque se fueron de alta antes que se les realizara estas mediciones, o porque fallecieron en sala. La idea fue la siguiente, a los individuos que salieron de alta sin discapacidad se les consideró en la mejor categoría en todas las variables. En el caso de los que fallecieron en sala, se les incorporó una categoría adicional que se interpreta como que no procede la medición. En las variables ordinales, esta nueva categoría aparece como la peor de todas.

Por último, se van a excluir del análisis las variables nistagmo, ataxia y babinski; dada la escasa relación entre estas variables y la variable egreso (Ver Anexo B). La variable reflejos osteotendinosos se elimina también por no aportar de manera significativa a los modelos obtenidos en el proceso de búsqueda del modelo definitivo. Como resultado del mismo proceso se decide incluir la variable orientación.

3.2. Búsqueda de interacciones significativas

Con el fin de mejorar la capacidad predictiva global del modelo, se procede a buscar nuevas interacciones con ayuda del CHAID de las variables al ingreso (considerando algunas de las ya existentes), que mejoren la discriminación preferentemente en el primer grupo de pacientes (vsd); dado que este es el grupo peor clasificado en el modelo propuesto por Bembibre-Suárez (Ver Anexo A). Los resultados que se obtuvieron en este proceso son los que se ofrecen (Ver Anexo C):

- Intin1: Mirada=1, conciencia=1, tono=1, fmi5=0 y $fmii \geq 4$ (vsd),

- Intin2: $Conciencia = 1$, $mirada = 1$ y $tono \geqslant 2$ (vcd),
- Intin3: $Conciencia \geq 3$ (f),
- Intin4: $Mirada \geqslant 2$ y conciencia=2 (vcd),
- Intin5: Sensibilidad=2 y mirada=1,
- Intin6: Sensibilidad=3 y P facial en 2 o 3,
- Intin7: Sensibilidad=1 y $tono \geqslant 2$,
- Intin8 Mirada=1 y tono=1,
- Intin9 Tono=1 y sensibilidad=1,
- Intin10: Sensibilidad=2 y $mirada \geqslant 2$,
- Intin11: Mirada=1 y $tono \geqslant 2$,
- Intin12: Mirada=2 y Sensibilidad=2,
- Intin13: Mirada=2, sensibilidad=2, $P\ \ facial \leq 2$ y $lenguaje \geqslant 3$,
- Intin14: Mirada=1, $tono \geqslant 2$ y $lenguaje \leq 4$,
- Intin15: $Coni \geq 3$ y $mirada \geq 2$,
- Intin16: $P\ \ facial \geq 3$ y $mirada \geq 2$,
- Intin17: P facial=2, mirada=1, tono=1 y sensibilidad=1,
- Intin18: P facial=2, mirada=1 y tono=1,
- Intin19: P facial=1, $orient \leq 2$ y $tono \geqslant 2$,
- Intin20: Intin14=2 y Intin8=2

- Intin21: Intin14=1, Intin5=2 y Intin19=1,
- Intin22: Mirada=1, Intin14=2, Intin8=1, Intin9=1, Intin17=1 y lenguaje=1 y
- Intin23: Mirada=1, Intin14=2, Intin8=2

De igual manera se buscaron interacciones significativas entre las variables al egreso y la variable egreso. El resultado de este proceso de búsqueda es el siguientes (Ver Anexo D):

- Inte1: Lenguaje=1 y alguna muscular en la categorías 0 o 1 (Mayoritariamente son vsd),
- Inte2: Lenguaje=2 y todas las fuerzas musculares derechas en las categorías 4 o 5 (Mayoritariamente son vcd) y
- Inte3: Lenguaje=1, ninguna fuerza muscular en las categorías 0 o 1 y tono al $egreso \geqslant 2$ (vcd).

3.3. Aplicación de la técnica OVERALS sobre los conjuntos de variables analizados

Una vez determinadas las interacciones significativas entre ambos conjuntos de variables, se procede a la aplicación de la correlación canónica con escalamiento óptimo. En este análisis se incluyeron directamente, luego de una serie de análisis basados fundamentalmente en el método de prueba y error; las variables clínicas conciencia y lenguaje además de las interacciones buscadas. Las variables al egreso (segundo conjunto) que fueron consideradas en el análisis son: la propia variable egreso,conciencia, parálisis facial, mirada, sensibilidad, lenguaje, tono, las fuerzas musculares (fmde45, fmie45, fme45 y fme01) y las interacciones entre las variables de este conjunto. La primera dimensión canónica obtenida correspondiente al primer grupo de variables se ofrece a continuación.

$Dim1 = -0{,}077*(coni)+0{,}012*(lengi)-0{,}003*(intin1)-0{,}011*(intin2)-0{,}109*(intin3)+0{,}001*(intin4)+0{,}034*(intin5)+0{,}047*(intin6)-0{,}018*(intin7)-0{,}332*(intin8)-0{,}037*$

$(intin9)+0{,}156*(Intin10)-0{,}361*(Intin11)-0{,}086*(Intin12)-0{,}003*(Intin13)-0{,}012*(Intin14)+0{,}033*(Intin15)-0{,}003*(Intin16)-0{,}004*(Intin17)+0{,}009*(Intin18)+0{,}001*(Intin19)+0{,}580*(Intin20)-0{,}006*(Intin21)-0{,}006*(Intin22)-0{,}151*(Intin23)$

Estas dimensiones se determinan, como se puede apreciar, a partir de la combinación lineal de las variables e interacciones consideradas en el conjunto predictor, toda vez que estas han sido recodificadas según las cuantificaciones correspondientes. Las restantes dimensiones se pueden obtener de igual forma a partir de las ponderaciones que se ofrecen en el Anexo E.

3.4. Estimación del modelo de pronóstico con regresión logística multinomial

Sobre las dimensiones obtenidas en el primer conjunto de variables mediante OVERALS, se estima el modelo de regresión logística multinomial con la variable dependiente estado al egreso de los pacientes.

Antes de considerar un modelo con fines predictivos se verifica la calidad de ajuste del mismo. La prueba de la Deviance sobre la bondad del ajuste del modelo permite asumir la validez del mismo con una significación estadística de 0.01, teniendo en cuenta el valor de 0.074 de la significación de la misma. Esta tiene como hipótesis nula que el modelo ajusta adecuadamente a los datos. Este resultado se muestra en la Figura 3.1.

Bondad de ajuste

	Chi-cuadrado	gl	Sig.
Pearson	299,089	108	,000
Desviación	129,986	108	,074

Figura 3.1: Ajuste del modelo.

En la Figura 3.2 se ofrecen los resultados de las pruebas para la significación del modelo.

La prueba sobre la significación global de las dimensiones incluidas como variables predictoras siguiendo el criterio de la razón de verosimilitud (primera tabla), sugiere que son estadísticamente significativa. No obstante, la dimensión cuatro no muestra una alta significación global. La prueba de Wald para la significación individual de los coeficientes en cada una de las funciones de clasificación confirma precisamente este hecho, a pesar de ello no se elimina del modelo debido a que disminuiría el porcentaje de clasificación correcta (segunda tabla). Los valores de estos coeficientes en cada una de las ecuaciones se pueden ver en la Figura 3.3.

Información del ajuste del modelo

Modelo	-2 log verosimilitud	Chi-cuadrado	gl	Sig.
Sólo la intersec	1919,250			
Final	185,538	1733,712	14	,000

Contrastes de la razón de verosimilitud

Efecto	-2 log verosimilitud del modelo reducido	Chi-cuadrado	gl	Sig.
Intersección	520,367	334,829	2	,000
DIM1	1034,744	849,206	2	,000
DIM2	279,826	94,288	2	,000
DIM3	286,807	101,269	2	,000
DIM4	188,659	3,121	2	,210
DIM5	369,081	183,543	2	,000
DIM6	287,126	101,588	2	,000
DIM8	236,918	51,380	2	,000

Figura 3.2: Significación de los coeficientes del modelo.

Estimaciones de los parámetros

EGRES		B	Error típ.	Wald	gl	Sig.
1	Intersecci	-,239	,912	,069	1	,794
	DIM1	7,430	,882	70,891	1	,000
	DIM2	3,197	,534	35,797	1	,000
	DIM3	3,677	,750	24,038	1	,000
	DIM4	-,627	,501	1,564	1	,211
	DIM5	-1,250	,352	12,601	1	,000
	DIM6	-5,902	1,308	20,373	1	,000
	DIM8	2,213	,345	41,273	1	,000
2	Intersecci	4,438	,659	45,345	1	,000
	DIM1	4,556	,493	85,303	1	,000
	DIM2	1,274	,419	9,238	1	,002
	DIM3	2,301	,702	10,748	1	,001
	DIM4	-,766	,476	2,592	1	,107
	DIM5	-1,569	,162	94,086	1	,000
	DIM6	-5,199	1,303	15,910	1	,000
	DIM8	,317	,124	6,478	1	,011

Figura 3.3: Significación de los coeficientes en las ecuaciones de regresión.

3.4.1. Validación del modelo

Considerado el adecuado ajuste del modelo se procede a validarlo con el propósito de garantizar su generalización a otros conjunto de datos no participantes en la investigación. Primeramente se utiliza el enfoque de validación cruzada. Para ello se toma una submuestra aleatoria del 50 % del total de las observaciones y se realizan el análisis de dos formas. Una vez seleccionada ambas muestras y obtenido un modelo para cada una de estas se comparan de forma descriptiva los coeficientes de los modelos para cada una. Esta comparación permite apreciar que estos no presentan diferencias sustanciales en sus coeficientes, existiendo una coincidencia casi total en la significación de las variables participantes. Se obtienen además, porcentajes similares de clasificación correcta (89.1 y 89.4 respectivamente,Ver Anexo F).

La otra idea consistió en la introducción de la prueba de estabilidad mediante el enfoque de la

variable dicotómica discutida en el Capítulo 2. Los resultados de la misma permiten plantear que los coeficientes diferenciales (intercepto y pendiente) no son significativos; de donde se concluye que el modelo es estable ante cambios en la muestra (Ver Figura 3.4). Comprobada

Contrastes de la razón de verosimilitud

Efecto	-2 log verosimilitud del modelo reducido	Chi-cuadrado	gl	Sig.
Intersección	400,788	187,642	2	,000
D	214,578	1,432	2	,489
DIM1	687,154	474,008	2	,000
DIM2	219,967	6,821	2	,033
DIM3	252,923	39,778	2	,000
DIM4	213,983	,837	2	,658
DIM5	294,497	81,351	2	,000
DIM6	256,065	42,919	2	,000
DIM8	244,798	31,652	2	,000
DDIM1	215,049	1,903	2	,386
DDIM2	214,123	,977	2	,613
DDIM3	213,462	,316	2	,854
DDIM4	213,626	,480	2	,787
DDIM5	213,705	,559	2	,756
DDIM6	214,343	1,197	2	,550
DDIM8	215,595	2,450	2	,294

Figura 3.4: Estabilidad de los coeficientes del modelo.

la calidad del ajuste del modelo y la validez del mismo, se está en condiciones de realizar las comparaciones comentadas al inicio del capítulo.

Las ecuaciones para el cálculo de las probabilidades de pertenencia de los pacientes a cada grupo considerando los resultados que se muestran en la Figura 3.3 son:

$$\begin{aligned} p_1 &= \frac{e^{f_1}}{1 + e^{f_1} + e^{f_2}} \\ p_2 &= \frac{e^{f_2}}{1 + e^{f_1} + e^{f_2}} \\ p_3 &= 1 - p_1 - p_2 \end{aligned} \tag{3.4.1}$$

donde f_1 y f_2 son las funciones de clasificación para los grupos 1 y 2 respectivamente. O sea:

$$f_1 = -0{,}239 + 7{,}43Dim1 + 3{,}197Dim2 + 3{,}677Dim3 - 0{,}627Dim4 - 1{,}25Dim5 \\ -5{,}902Dim6 + 2{,}213Dim8$$

$$f_2 = 4{,}438 + 4{,}556Dim1 + 1{,}274Dim2 + 2{,}301Dim3 - 0{,}766Dim4 - 1{,}569Dim5 \\ -5{,}199Dim6 + 0{,}317Dim8$$

De acuerdo a las diferentes tentativas de búsqueda de modelos para el pronóstico mediante el empleo de OVERALS, se pudo observar que pequeñas variaciones en el conjunto de variables al egreso (inclusión o exclusión de una variable), provocaron cierta repercusión en los resultados de la capacidad predictiva, específicamente en el porcentaje de clasificación correcta.

3.5. Análisis comparativo de modelos

En este epígrafe se va a realizar la comparación de los modelos obtenidos tomando como referencia los criterios expuestos en el Capítulo 2. Estos se van a especificar previamente antes de cada comparación. A modo de convenio se va a hacer referencia al modelo obtenido mediante el empleo de OVERALS como Variante 1 y como Variante 2 al modelo obtenido a partir de la utilización de PRINCALS.

3.5.1. Comparación entre los modelos obtenidos con OVERALS y PRINCALS

Para la comparación de estos modelos se va a emplear los siguientes criterios:

- por ciento de clasificación correcta,
- errores estándares en los coeficientes de lo modelos,
- índice Kappa,
- estabilidad mediante la comparación descriptiva de las magnitudes de los coeficientes y

- sensibilidad y especificidad de ambos modelos.

Se va a omitir del análisis el número de variables por las razones ya expuestas.

Teniendo en cuenta el porcentaje de clasificación correcta, se puede apreciar que la Variante 1 supera en aproximadamente un 1 % a la Variante 2; se consigue además mejor clasificación correcta en el primer grupo, que es el de más difícil clasificación (Ver Figuras 3.5 y 3.6).

Clasificación

	Pronosticado			
Observado	1	2	3	Porcentaje correcto
1	146	48	0	75,3%
2	40	704	20	92,1%
3	2	33	325	90,3%
Porcentaje globa	14,3%	59,6%	26,2%	89,2%

Figura 3.5: Tabla de clasific correcta del modelo con OVERALS.

Classification

	Predicted			
Observed	1	2	3	Percent Correct
1	118	76	0	60,8%
2	24	702	38	91,9%
3	0	18	342	95,0%
Overall Percentage	10,8%	60,4%	28,8%	88,2%

Figura 3.6: Tabla de clasific del modelo obtenido con PRINCALS.

El análisis de los errores estándares de los coeficientes muestran que mediante la Variante 1 esto son menores. Este elemento se puede confirmar a través del cálculo del módulo de la razón $\frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}$, con $j = \overline{1,p}$, donde p es el número de dimensiones empleadas en cada caso (7 para Variante 1 y 8 para Variante 2). Realizando una comparación dimensión a dimensión para cada ecuación de los modelo comparados, se aprecia que los coeficientes del modelo con la Variante 1 tienen menor variabilidad que los coeficientes del modelo con la Variante 2. Esto se justifica atendiendo al hecho de que en el 75 % comparaciones realizadas, se obtienen razones con valores superiores favorables para la Variante 1. En el Anexo G se muestran los valores de las razones para cada una de las variantes. Otros de los resultados obtenidos para la Variante 2 se pueden ver en el Anexo H.

Los índices Kappa de los modelos que se comparan son de 0.807 y 0.787 para las Variantes 1 y 2 respectivamente. Estos valores indican una diferencia de dos unidades porcentuales a favorable a la Variante 1 (Ver Figuras 3.7 y 3.8). A pesar de que no es una diferencia sustancial, según el criterio de Latour (Latour $et\ al.$, 1997), este indicador correspondiente a la Variante 1 se clasifica como excelente.

En la prueba de estabilidad mediante el enfoque de la variable dicotómica, se verifica que el

Medidas simétricas

	Valor	Error típ. asint.[a]	aproximada[b]	Sig. aproximada
Medida de acuerdo	Kappa ,807	,015	38,907	,000
N de casos válidos	1318			

Figura 3.7: Coeficiente Kappa para el modelo con OVERALS.

modelo obtenido mediante la Variante 2 también cumple con este requisito (Ver Anexo I).

La comparación de las variantes de modelos sobre la base de los análisis de sensibilidad, se va a realizar a partir de las categoría vivos y fallecidos de la variable estado al egreso. Para ello se realiza una transformación de las tablas de clasificación correcta de estos de tal manera que

Symmetric Measures

	Value	Asymp. Std. Error[a]	Approx.[b]	Approx. Sig
Measure of Agr Kappa	,787	,016	37,407	,000
N of Valid Cases	1318			

Figura 3.8: Coeficiente Kappa para el modelo con PRINCALS.

queden agrupados en estas dos categorías. Estas transformaciones se ofrecen en el Anexo J.

A los efectos de la investigación se va a entender por sensibilidad la probabilidad que tiene el modelo de pronóstico de clasificar correctamente a los individuos fallecidos. De manera análoga se define la especificidad como la probabilidad de clasificar a los individuos vivos acertadamente. Aplicando las expresiones (1.4.4) y (1.4.5) para el cálculo de la sensibilidad y especificidad respetivamente se tiene que: la Variante 1 es más sensible que la Variante 2 (0.942 y 0.90); sin embargo esta última presente un mayor valor para la sensibilidad (0.98 y 0.964 en ese orden).

De las comparaciones realizadas se puede concluir que ambos modelos muestran un comportamiento similar en cuanto a los indicadores analizados con una ligera diferencia a favor del modelo obtenido empleando la técnica de OVERALS. En la siguiente tabla se presenta un resumen de las comparaciones realizadas entre ambos modelos siguiendo los diferentes criterios establecidos con este propósito.

Modelo	Clasif Correcta (%)	Variab coef	Kappa	Estab	Sensib	Especif
Variante 1	89.2	Menor	0.807	Estable	0.942	0.964
Variante 2	88.2	Mayor	0.787	Estable	0.90	0.98

3.5.2. Comparación entre el modelo propuesto y el obtenido por Bembibre-Suárez

En este caso los criterios a considerar son:

- por ciento de clasificación correcta,
- número de variables predictoras,
- errores estándares en los coeficientes de lo modelos,
- índice Kappa y
- sensibilidad y especificidad de ambos.

En cuanto al porcentaje de clasificación correcta se puede afirmar que el modelo obtenido en la presente investigación es también ligeramente superior en este aspecto. El obtenido por Bembibre-Suárez es de 88.4 % mientras que el valor alcanzado por el modelo calculado es de 89,2 %. En el Anexo A se muestran todos los resultados de este.

Ahora bien, considerando el número de variables predictoras, en este caso se obtuvo un modelo con tres variable menos (en total 11), lo que hace que sea más parsimonioso y, por ende, más fácilmente introducible como modelo para el pronóstico.

A manera descriptivas, los errores estándares de los coeficientes para ambos modelos no tienen diferencias realmente sustanciales. En ambos casos se consideran pequeños.

De acuerdo al coeficiente Kappa, se aprecian valores muy próximos en ambos casos. El valor alcanzado por este último es de 0.794 (Ver Anexo A), mientras que el del modelo que se propone es de 0.807 (Figura 3.7).

A partir del análisis de los resultados de la comparación de ambos modelos, se observa, como elemento a destacar, la reducción de la cantidad de variables predictivas a emplear y el ligero aumento del porcentaje de clasificación correcta, alcanzada por el modelo propuesto en esta investigación.

3.6. Conclusiones del capítulo

1. La inclusión de nuevas interacciones posibilitó mejorar el porcentaje de pronóstico acertado en los pacientes pertenecientes a la categoría vsd, que en modelo anterior resultó ser el grupo de peor clasificación.

2. El análisis comparativo entre los modelos obtenidos mediante el empleo de OVERALS y PRINCALS arroja resultados ligeramente favorables en el primero sobre la base de los elementos de comparación tomados en cuenta,

3. Basado en los diferentes criterios sobre la calidad del ajuste y la prueba de estabilidad, el modelo propuesto es adecuado y posibilita realizar pronósticos con un aceptable porcentaje de aciertos, y

4. La alternativa propuesta como variante del procedimiento implementado por Bembibre-Suárez es válida y muestra resultados levemente superiores.

CONCLUSIONES

Como resultado del proceso de búsqueda de un modelo de pronóstico para el estado al egreso de pacientes con ECV, a partir de un conjunto de variable clínicas al momento del ingreso y haciendo uso de las técnicas de OVERALS, CHAID y regresión logística multinomial; se pueden establecer las siguientes conclusiones:

1. El uso de la técnica OVERALS en el procedimiento para establecer un modelo para el pronóstico, constituye una variante que garantiza una capacidad predictiva adecuada.

2. El empleo de la nueva variante eleva moderadamente el poder predictivo para el conjunto de datos participantes en la investigación.

3. La búsqueda de nuevas interacciones posibilitó mejorar el porcentaje de pronóstico correcto tanto en el grupo de más difícil clasificación (vsd) como de manera global.

4. La introducción de la prueba analítica de estabilidad de los modelos mediante el enfoque de la variable dicotómica en el proceso de validación de un modelo candidato, incrementa la seguridad de las conclusiones sobre esta cualidad deseable en los modelos de pronóstico.

5. Como resultado de la búsqueda del modelo en la presente investigación, se logró reducir la cantidad de variables independientes, lo cual facilita la introducción del mismo en la práctica médica.

RECOMENDACIONES

1. Aplicar ambos procedimientos en nuevas investigaciones de esta naturaleza con vista a establecer posibles regularidades en los resultados.

2. Estudiar el efecto que provoca en la capacidad predictiva, la inclusión de variables no clínicas en el conjunto de variables al egreso para la correlación canónica .

3. Extender la variante introducida en el procedimiento para obtener los modelos de pronóstico al ingreso, para las 24 y 72 horas.

4. Implementar un software para la aplicación práctica del modelo obtenido como recurso para la ayuda pronóstica en la actividad clínica.

REFERENCIAS BIBLIOGRÁFICAS

Abraira, V. & Vargas, A. de. (1996). *Métodos multivariantes en bioestadística.* Centro de Estudios Ramón Areces.

Altman, D. & Bland, J. (1994). Statistic notes: Diagnostic test 1: sensitivity and specificity. *BMJ*.

Carvajal, P., Trejos, A., & Mejías, J. S. (2004). Aplicación del análisis discriminate para explorar la relación entre el examen de icfes y el redimiento en álgebre lineal de los estudiantes de ingeniería de la upt en el periodo 2001-2003. *Scientia et Technica*(25).

Códoba, J. A. G. (2007). Aplicaciones de la matemática y la estadística al ambito sanitario. *Jornadas sobre nuevas tecnologías al servicio de la salud y el bienestar social*.

F. Hair j., Anderson, J., Tatham, R. L., & C. Black w. (1999). *Análisis multivariante* (fifth ed.). Madrid: Prentice Hall Iberia.

Gifi, A. (1985). Princals. En *Internal report – UG-85-02*.

Greene, W. H. (2003). *Econometric analysis* (Fifth ed.). New Jersey: Prentice Hall.

Gujarati, D. (2005). *Econometría.* La Habana: Felix Varela.

Helena, R. V. S. L. & Aguirre, S. (1996). La violencia doméstica durante el embarazo y su relación con el peso al nacer. *Salud pública Méx*, *38*(5).

Hosmer, D. W. (1989). *Applied logistic regression.* John Wiley and Sons.

Johnson, R. A. & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (fifth ed.). N.J.: Prentice-Hall, Inc.

Latour, J., Abraira, V., Cabello, J. B., & Sánchez, J. L. (1997). Métodos de investigación en cardiología: validez y errores de medición. *Rev Esp de Card*, 117-128.

Mangin, J.-P. L., Alonso, M. A. S., & Clavel., J. S. (2002). La predicción y la clasificación de datos en marketing. un análisis comparativo mediante técnicas multivariantes, árboles jerárquicos y redes neuronales. *Ciencia Ergo Sum*, *9*(1), 21 a 30.

Martínez, T. T. P., Rojas, L. Íñiguez, Valdés, L. S., & Remond, R. (2003). Vulnerabilidad espacial al dengue. una aplicación de los sistemas de información geográfica en el municipio playa de ciudad de la habana. *Rev Cubana Salud Pública*, *29*(4).

Novas, J. D. & Machado, B. R. G. (2002). El pronóstico. *Revista cubana de medicina general e integral*, *20*(2).

Pinto, M., Contreras, M., Carrasco, E., Brito, C., Molina, L. H., Ah-Hen, K. S., & León, S. V. y. (2002). Determinación de la autenticidad de grasas lácteas. análisis discriminante lineal de triacilglicéridos. *Agro Sur*, *30*(1).

Portillot, F., Mar, C., & Martínez, T. (2007). Métodos no lineales de escalado óptimo: una aplicación al análisis del empleo en la compañía ferroviaria mza.

Pérez, L. R.-M., Pliego, F. J. M., Lorenzo, J. M. M., & Tomé, P. U. (1995). *Análisis estadísitco de encuestas: datos cualitativos.* Madrid, España: AC.

Rodríguez, E. R., Sánchez, P. H., Blanco, F. L., Romero, C. D., & Majen, L. S. (2004). Aplicación de análisis multivariado para la diferenciación de individuos sanos según su contenido sérico de minerales. *Nutrición Hospitalaria*, *XIX*(5), 263-268.

Salinas, M. & Silva, C. (2007). Modelos de regresión y correlación ii. regresión lineal múltiple. *Ciencia and Trabajo*(23), 39-41.

Santín, D. (2006). La medición de la eficiencia de las escuelas:una revisión crítica. *Revista de Economía Pública*, 57-82.

Schneider, M. C., Castillo-Salgado, C., Bacallao orge, Loyola, E., Mujica, O. J., Vidaurre, M., & Roca, A. (2002). Métodos de medición de las desigualdades de salud. *Revista Panamericana de Salud Pública*, *12*(6).

Surí, R. S. (2004). *Procedimiento estadístico para la búsqueda de un modelo que permita pronosticar la perspectiva de evolución de los pacientes hospitalizados con enfermedades cerebrovasculares en base de la información estadística recopilada y la experiencia médica.*

Taboada, D. R. M. B., Suri, R. S., Morales, E. C., Gómez, J. C., Brito, A. E., Lafonte, R. E., & Peraza, M. V. (2003). Creación y validación de un instrumento para el seguimiento de pacientes con enfermedad cerebrovascular. *Revista Cubana de Medicina*, *42*(1).

Tusell, F. (2005). *Análisis multivariante.*

Varela, M. V., Gandolff, L. S., García, M. C., & Peña, G. B. de la. (2003). *Algebra lineal* (Segunda ed.). Ciudad de la Habana: Félix Varela.

Vinacua, B. V. (1998). *Análisis estadístico con spss para windows. estadística multivariante.* Mc Grau Hill.

ANEXOS

Anexo A. Resultados del modelo obtenido por Bembibre-Suárez.

Bondad de ajuste

	Chi-cuadrado	gl	Sig.
Pearson	810,503	262	,078
Desviación	230,448	262	,921

Información del ajuste del modelo

Modelo	-2 log verosimilitud	Chi-cuadrado	gl	Sig.
Sólo la intersec	1938,765			
Final	287,361	1651,404	10	,000

Contrastes de la razón de verosimilitud

Efecto	-2 log verosimilitud del modelo reducido	Chi-cuadrado	gl	Sig.
Intersección	451,127	163,766	2	,000
OBSCO1_3	1115,728	828,367	2	,000
OBSCO2_3	358,218	70,857	2	,000
OBSCO3_3	342,686	55,325	2	,000
OBSCO4_3	314,978	27,617	2	,000
OBSCO6_3	341,057	53,696	2	,000

Estimaciones de los parámetros

EGRESO		B	Error típ.	Wald	gl	Sig.
1	Intersección	-2,259	,587	14,805	1	,000
	OBSCO1_3	-7,561	1,155	42,827	1	,000
	OBSCO2_3	1,339	,623	4,618	1	,032
	OBSCO3_3	,705	,623	1,280	1	,258
	OBSCO4_3	-,832	,181	21,042	1	,000
	OBSCO6_3	,528	,290	3,312	1	,069
2	Intersección	1,578	,193	66,883	1	,000
	OBSCO1_3	-4,193	,421	99,049	1	,000
	OBSCO2_3	1,668	,243	47,085	1	,000
	OBSCO3_3	1,215	,165	54,364	1	,000
	OBSCO4_3	-,443	,166	7,094	1	,008
	OBSCO6_3	-,612	,130	22,275	1	,000

Clasificación

	Pronosticado			
Observado	1	2	3	Porcentaje correcto
1	135	58	1	69,6%
2	35	695	34	91,0%
3	0	25	335	93,1%
Porcentaje global	12,9%	59,0%	28,1%	88,4%

Tabla de contingencia EGRESO * Categoría de respuesta pronosticada

Recuento

		Categoría de respuesta pronosticada			
		1	2	3	Total
EGRESO	1	135	58	1	194
	2	35	695	34	764
	3		25	335	360
Total		170	778	370	1318

Medidas simétricas

		Valor	Error típ. asint.[a]	Sig. aproximada
Medida de acuerdo	Kappa	,794	,016	,000
N de casos válidos		1318		

a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

Anexo B. Correlación de variables clínicas eliminadas del análisis con la variable egreso.

Correlaciones

			EGRESO	ATAXI
Rho de Spearman	EGRESO	Coeficiente de correlación	1,000	,050
		Sig. (bilateral)	.	,071
		N	1318	1318
	ATAXI	Coeficiente de correlación	,050	1,000
		Sig. (bilateral)	,071	.
		N	1318	1318

Correlaciones

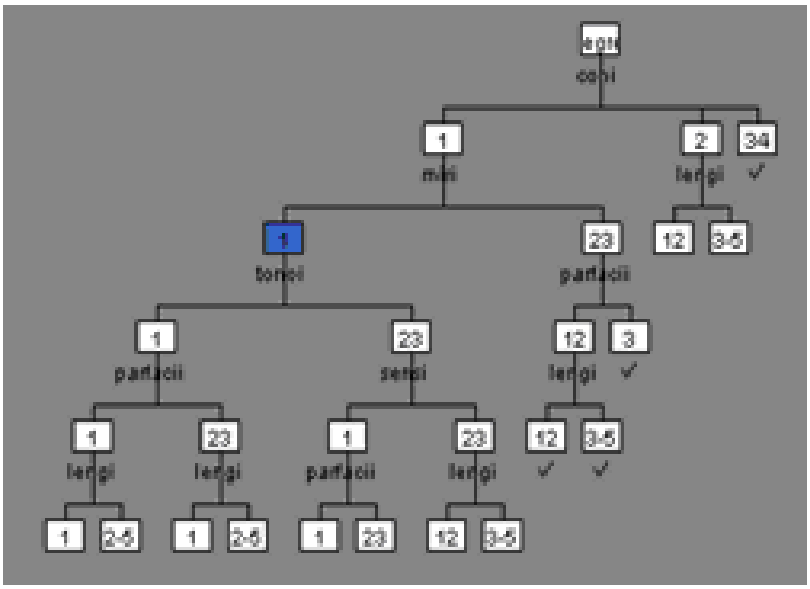
			EGRESO	NISTI
Rho de Spearman	EGRESO	Coeficiente de correlación	1,000	,165**
		Sig. (bilateral)	.	,000
		N	1318	1318
	NISTI	Coeficiente de correlación	,165**	1,000
		Sig. (bilateral)	,000	.
		N	1318	1318

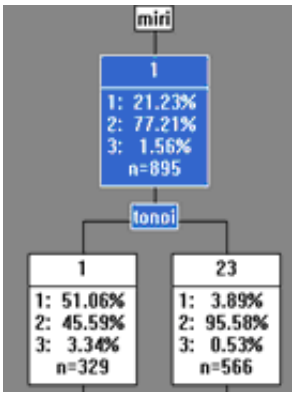
Correlaciones

			EGRESO	BABII
Rho de Spearman	EGRESO	Coeficiente de correlación	1,000	,175**
		Sig. (bilateral)	.	,000
		N	1318	1318
	BABII	Coeficiente de correlación	,175**	1,000
		Sig. (bilateral)	,000	.
		N	1318	1318

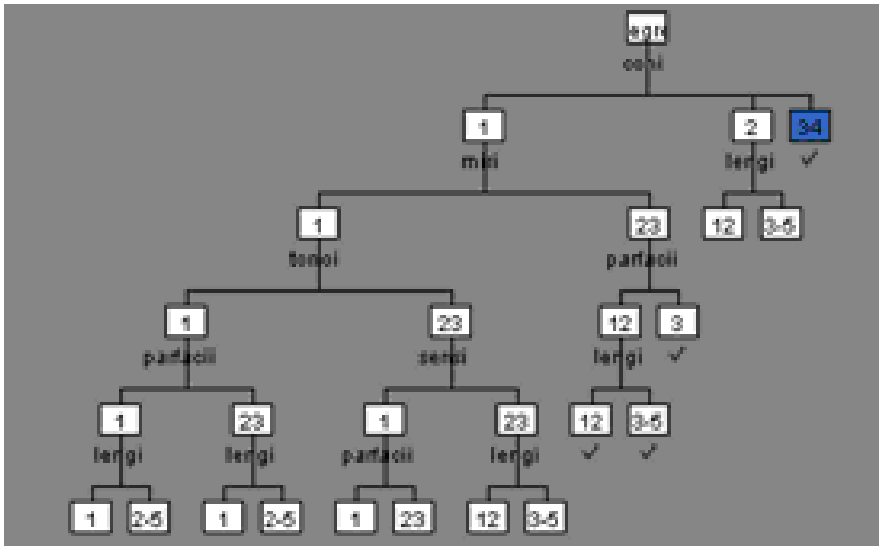
Anexo C. Algunas nuevas interacciones de las variables al ingreso.

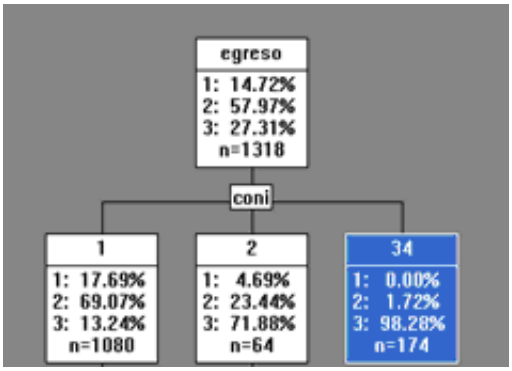
Intin2 : coni=1 y miri=1 y $tonoi \geq 2$ (Vcd)



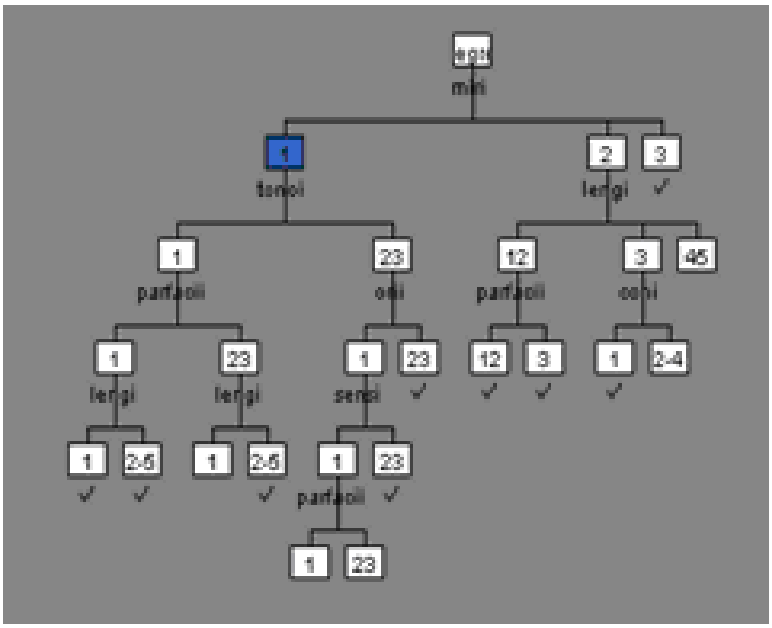


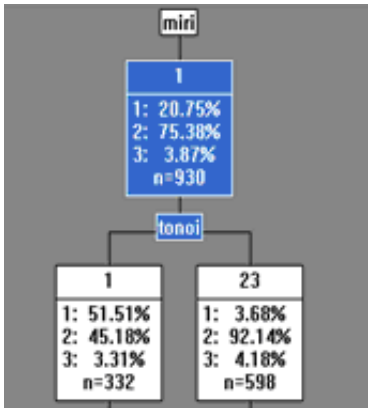
Intin3: $coni \geq 3$ (fallecidos)





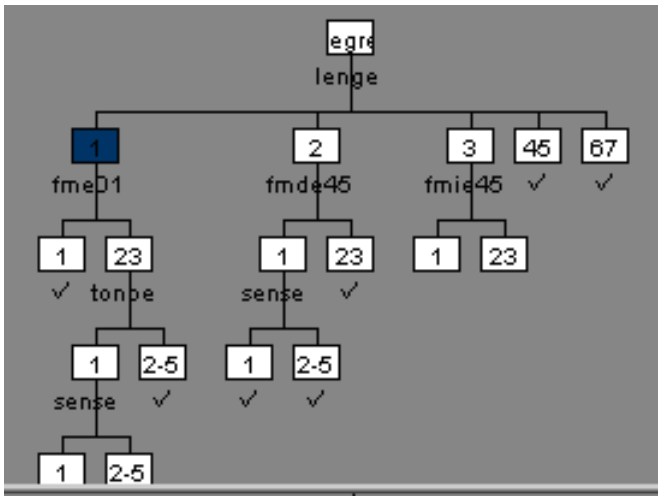
Intin11: miri1 y $tonoi \geq 2$ (mayoritariamente vcd)

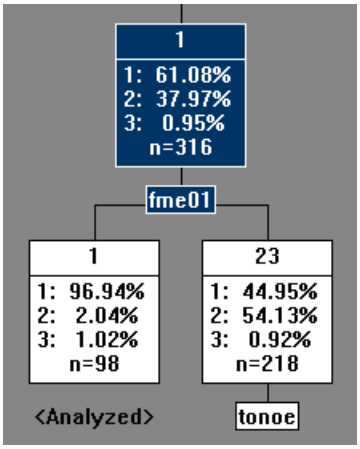




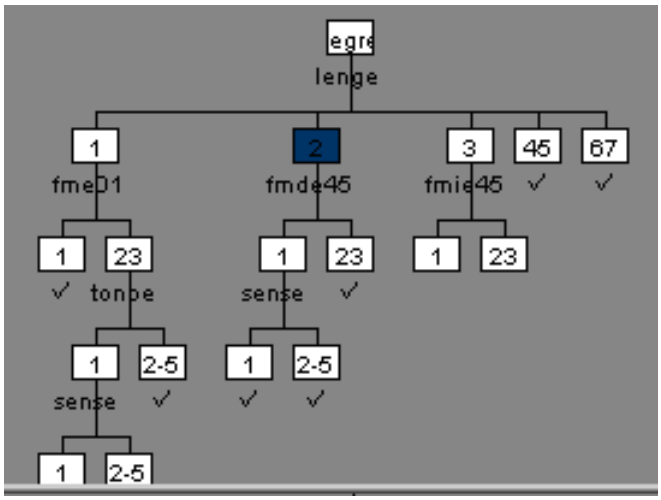
Anexo D. Algunas nuevas interacciones de las variables al egreso.

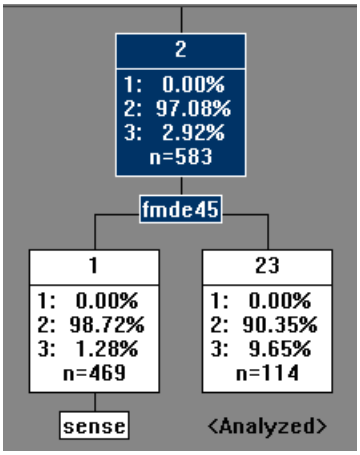
Inte1: Lenge=1 y alguna fme en la categorías 0 o 1 (mayoritariamente vsd)



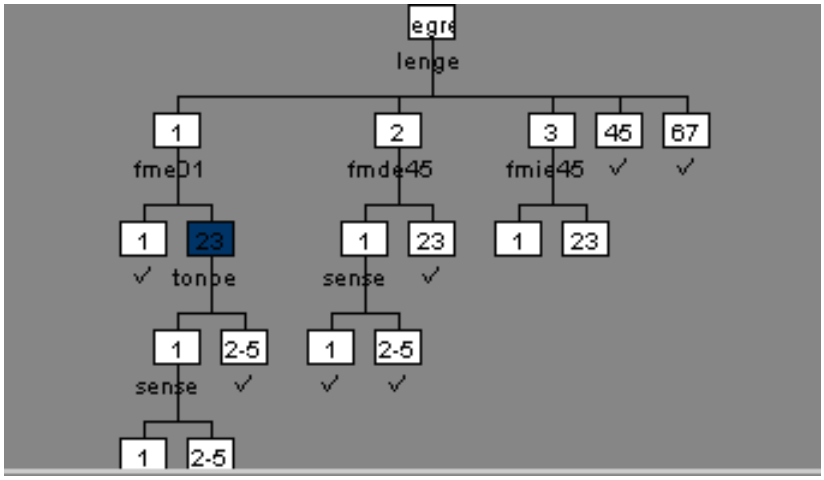


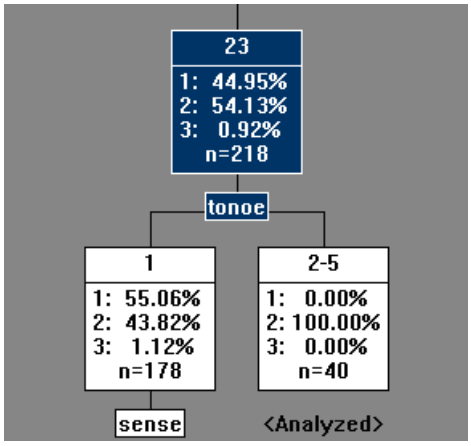
Inte2: Lenge=2 y fmde en 4 o 5 (Mayoritariamente son vcd)





Inte3: Lenge=1, ninguna fme en 0 o 1 y $tonoe \geqslant 2$ (vcd)





Anexo E. Pesos de las variables en las dimensiones canónicas.

Weights

Set		Dimension 1	2	3	4	5	6	7	8
1	coni	-,077	-,061	,086	,145	,523	-,210	-,505	-,221
	lengi	,012	,007	,005	,106	-,153	-,020	1,008	,297
	miri=1+coni=1+tonoi=1+fmi5=0+fmii>=4 (vsd)	-,003	-,015	-,022	,072	-,011	,924	,177	-,065
	coni=1+miri=1+tonoi>=2 (vcd)	-,011	,004	,000	,066	,206	-,034	-,075	-,336
	coni>=3 (muere)	-,109	-,061	,120	-,053	-,572	-,135	-,492	-,300
	(miri >= 2) & (coni = 2) vcd	,001	,005	,003	,002	-,133	-,003	,033	-,291
	(sensi= 2) &((miri= 1)	,034	-,080	,100	-,285	,112	,144	-,251	,021
	(sensi = 3) & ((parfacii = 2)or(parfacii= 3))	,047	-,076	-,046	-,320	-,414	,053	,035	,031
	(sensi= 1) & ((tonoi= 2)or(tonoi= 3))	-,018	-1,045	,397	-,755	-,158	,042	-,040	,187
	(tonoi= 1) & (miri= 1)	-,332	,150	-,208	,334	-,163	-,369	-,202	-,271
	(tonoi= 1) & (sensi= 1)	-,037	-,969	,103	-,661	-,244	,048	-,493	,342
	(sensi= 2) &((miri= 2)or(miri = 3))	,156	-,100	,058	-,081	,678	-,202	,889	1,289
	(miri= 1) &((tonoi= 2)or(tonoi= 3))	-,361	,348	,024	-,647	-,627	,264	,201	,760
	(miri= 2) &(sensi = 2)	-,086	,007	-,042	-,247	-1,089	,258	-,913	-,590
	(miri= 2) &(sensi = 2) & ((parfacii = 1)or(parfacii= 2)) & ((lengi = 3)or(lengi = 4)or(lengi = 5))	-,003	-,004	-,001	,023	-,432	-,004	-,051	,268
	(miri= 1) & ((tonoi= 2)or(tonoi= 3)) & ((lengi = 1)or(lengi = 2)or(lengi = 3)or(lengi = 4))	-,012	-,008	-,024	-,128	,069	,044	,061	-,043
	((coni= 3) or(coni = 4))& ((miri = 2) or(miri= 3))	,033	,002	,009	,220	,549	-,071	-,080	-,344
	((parfacii= 3) or(parfacii= 4))& ((miri = 2) or(miri= 3))	-,003	,002	-,243	-,287	-,487	-,038	,190	,686
	(parfacii= 2) &(miri = 1) & (tonoi= 1) & (sensi = 1)	-,004	,295	-,153	,128	,306	,294	-,119	,058
	(parfacii= 2) &(miri = 1) & (tonoi= 1)	,009	-,296	-,159	-,417	-,144	-,393	,132	,477
	(parfacii= 1) &((orii= 1) or(orii = 2))& ((tonoi= 2) or(tonoi = 3)or(tonoi = 4))	,001	,033	,504	,675	,014	,051	-,216	-,005
	((inter11 = 2) & (inter5 = 2))	,580	-,346	,289	-,161	,896	,316	-,154	-,307
	((inter11 = 1) & (inter1 = 2) & (inter18 = 1))	-,006	-,033	,034	-,033	,069	,032	-,194	-,066
	((miri = 1) & (inter11 = 2) & (inter5 = 1) & (inter6 = 1) & (inter15 = 1) & (lengi) = 1)	-,006	-,007	,064	,084	,003	,408	-,305	-,218
	((miri = 1) & (inter11 = 2) & (inter5 = 2))	-,151	,083	,124	,375	-,190	-,113	-,201	,375

Anexo F. Comparación de los coeficientes de los modelos en la validación cruzada.

Muestra aleatoria
Primera mitad

Parameter Estimates

egreso[a]		B	Std. Error	Wald	df	Sig.
1	Intercept	,301	1,282	,055	1	,815
	dim1	6,701	1,139	34,597	1	,000
	dim2	3,817	,877	18,937	1	,000
	dim3	3,893	1,131	11,856	1	,001
	dim4	-,886	,764	1,345	1	,246
	dim5	-1,171	,505	5,377	1	,020
	dim6	-6,287	1,915	10,775	1	,001
	dim8	1,621	,470	11,905	1	,001
2	Intercept	4,434	,899	24,311	1	,000
	dim1	4,604	,691	44,436	1	,000
	dim2	1,611	,665	5,878	1	,015
	dim3	2,309	1,052	4,815	1	,028
	dim4	-1,128	,739	2,331	1	,127
	dim5	-1,707	,237	51,799	1	,000
	dim6	-5,423	1,907	8,089	1	,004
	dim8	,252	,197	1,646	1	,199

a. The reference category is: 3.

Clasificación

Observado	Pronosticado			
	1	2	3	Porcentaje correcto
1	70	20	0	77,8%
2	23	327	17	89,1%
3	1	9	174	94,6%
Porcentaje global	14,7%	55,5%	29,8%	89,1%

Segunda mitad

Parameter Estimates

egreso[a]		B	Std. Error	Wald	df	Sig.
1	Intercept	-2,620	8,462	,096	1	,757
	dim1	11,248	13,167	,730	1	,393
	dim2	2,663	1,053	6,391	1	,011
	dim3	3,738	1,095	11,665	1	,001
	dim4	-,515	,712	,524	1	,469
	dim5	-1,375	,572	5,786	1	,016
	dim6	-5,938	1,904	9,726	1	,002
	dim8	2,784	,594	21,968	1	,000
2	Intercept	4,539	1,004	20,449	1	,000
	dim1	4,662	,746	39,018	1	,000
	dim2	1,067	,561	3,617	1	,057
	dim3	2,458	1,006	5,966	1	,015
	dim4	-,579	,651	,793	1	,373
	dim5	-1,522	,241	40,003	1	,000
	dim6	-5,305	1,899	7,801	1	,005
	dim8	,327	,167	3,827	1	,050

a. The reference category is: 3.

Clasificación

Observado	Pronosticado			
	1	2	3	Porcentaje correcto
1	75	29	0	72,1%
2	17	370	10	93,2%
3	1	15	160	90,9%
Porcentaje global	13,7%	61,2%	25,1%	89,4%

Anexo G. Razón entre los coeficientes del modelo y sus errores típicos.

Dimensión	Variante 1	Variante 2
Intercep	0.262	4.82
Dim 1	8.424	0.817
Dim 2	5.987	5.545
Dim 3	4.903	1.406
Dim 4	1.252	4.8
Dim 5	3.55	3.393
Dim 6	4.512	0.03
Dim 8	6.141	0.792
Intercept	6.734	3.01
Dim 1	9.241	6.026
Dim 2	3.04	2.974
Dim 3	3.297	2.545
Dim 4	1.609	6.237
Dim 5	9.685	1.667
Dim 6	3.99	2.697
Dim 8	2.556	3.827

Anexo H. Resultado del modelo obtenido mediante el empleo de PRINCALS (Variante 2).

Bondad de ajuste

	Chi-cuadrado	gl	Sig.
Pearson	141908,417	162	,000
Desviación	172,878	162	,265

Información del ajuste del modelo

Modelo	-2 log verosimilitud	Chi-cuadrado	gl	Sig.
Sólo la intersecció	1951,376			
Final	232,638	1718,738	16	,000

Contrastes de la razón de verosimilitud

Efecto	-2 log verosimilitud del modelo reducido	Chi-cuadrado	gl	Sig.
Intersección	350,866	118,228	2	,000
OBSCO1_1	1252,578	1019,940	2	,000
OBSCO2_1	362,935	130,297	2	,000
OBSCO3_1	241,147	8,509	2	,014
OBSCO4_1	274,732	42,094	2	,000
OBSCO5_1	249,307	16,669	2	,000
OBSCO6_1	244,547	11,909	2	,003
OBSCO7_1	295,804	63,166	2	,000
OBSCO8_1	251,911	19,273	2	,000

Estimaciones de los parámetros

EGRESO		B	Error típ.	Wald	gl	Sig.
1	Intersecció	-38,200	7,925	23,237	1	,000
	OBSCO1_1	1,706	5,373	,101	1	,751
	OBSCO2_1	-28,104	5,068	30,753	1	,000
	OBSCO3_1	11,344	8,064	1,979	1	,159
	OBSCO4_1	-55,435	11,544	23,058	1	,000
	OBSCO5_1	11,187	3,297	11,510	1	,001
	OBSCO6_1	,137	1,669	,007	1	,935
	OBSCO7_1	31,787	6,065	27,466	1	,000
	OBSCO8_1	1,626	2,052	,628	1	,428
2	Intersecció	1,734	,576	9,054	1	,003
	OBSCO1_1	8,695	1,443	36,287	1	,000
	OBSCO2_1	-3,045	1,024	8,848	1	,003
	OBSCO3_1	-1,815	,713	6,490	1	,011
	OBSCO4_1	,116	,489	,056	1	,812
	OBSCO5_1	1,035	,621	2,781	1	,095
	OBSCO6_1	2,139	,793	7,276	1	,007
	OBSCO7_1	1,883	,464	16,452	1	,000
	OBSCO8_1	,773	,202	14,681	1	,000

Anexo I. Resultado de la prueba de estabilidad para la Variante 2.

Jtilizando la misma muestra aleatoria que para la Variante 1

Contrastes de la razón de verosimilitud

Efecto	-2 log verosimilitud del modelo reducido	Chi-cuadrado	gl	Sig.
Intersección	346,687	83,871	2	,000
MUESTRA	267,123	4,308	2	,116
OBSCO1_1	774,480	511,665	2	,000
OBSCO2_1	348,003	85,188	2	,000
OBSCO3_1	270,598	7,782	2	,020
OBSCO4_1	297,252	34,437	2	,000
OBSCO5_1	281,970	19,154	2	,000
OBSCO6_1	272,697	9,882	2	,007
OBSCO7_1	307,798	44,982	2	,000
OBSCO8_1	274,484	11,668	2	,003
DOBS1	267,068	4,253	2	,119
DOBS2	264,672	1,856	2	,395
DOBS3	269,112	6,297	2	,043
DOBS4	266,300	3,485	2	,175
DOBS5	267,026	4,211	2	,122
DOBS6	264,543	1,728	2	,422
DOBS7	267,264	4,448	2	,108
DOBS8	264,137	1,321	2	,516

Anexo J. Tablas de categorías reagrupadas para análisis de sensibilidad y especificidad.

Variante 1. Modelo obtenido a partir de OVERALS

		Categoría pronosticada	
		V	F
Categoría	V	938	20
observada	F	35	325

Variante 2. Modelo obtenido a partir de PRINCALS

		Categoría pronosticada	
		V	F
Categoría	V	920	38
observada	F	18	342

Modelo propuesto por Bembibre-Suárez

		Categoría pronosticada	
		V	F
Categoría	V	923	35
observada	F	25	335