UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS MATEMÁTICAS

FACULTAD DE CIENCIAS

UNA APLICACIÓN DE MODELOS GRÁFICOS PROBABILÍSTICOS EN INVESTIGACIÓN MÉDICA

T E S I S

QUE PARA OBTENER EL GRADO ACADÉMICO DE:

MAESTRO EN CIENCIAS MATEMÁTICAS

PRESENTA:

RICARDO RAMÍREZ ALDANA

DIRECTORA DE TESIS: **DRA. GUILLERMINA ESLAVA GÓMEZ**

MÉXICO, D. F. **MAYO, 2005**

AGRADECIMIENTOS.

A mis padres. A mi madre por brindarme en todo momento su amor, comprensión y paciencia. A mi padre por su apoyo a lo largo de todos mis años de estudio.

A mi hermana, por su amistad incondicional y por siempre tener palabras de aliento cuando son necesarias.

A mis amigos, por su compañía y además lograr que uno sonría y se tranquilice cuando más difícil parece.

A mis maestros, por compartir su conocimiento. En especial a mi directora de tesis la Dra. Guillermina Eslava Gómez quien ofreció su tiempo y experiencia para lograr realizar este trabajo. A mis sinodales por sus valiosos comentarios y sugerencias. A los doctores Luis David Sánchez Velázquez y Héctor Ávila Rosas quienes ofrecieron su conocimiento en medicina y la información necesaria para poder llevar a cabo este trabajo.

Índice general

Índice de Tablas

Índice de figuras

Resumen

En este trabajo se presentan y ajustan modelos gráficos probabilísticos, en particular Redes Bayesianas, para modelar información proveniente de pacientes que ingresan a Unidades de Terapia Intensiva (UTI) en la Ciudad de México en los años 2002-2004. La base de datos con la que se trabajó está formada por 923 casos, al quitar áquellos casos con valores faltantes, se tiene un total de 861 registros, de los cuales 386 casos corresponden a pacientes que sobrevivieron tres meses o más después de su egreso de la UTI y el resto no sobrevivieron. En estos modelos las variables, continuas o discretas, se representan con círculos (u otra figura) y se llaman nodos y las relaciones de dependencia entre ellas se establecen mediante "flechas", que se denominan arcos, tal dependencia entre las variables está dada mediante probabilidades condicionales, la probabilidad condicional de un nodo dados sus nodos padres (los nodos que apuntan al primero) es la probabilidad local. En primer lugar se busca obtener una red que exprese las relaciones entre las variables a partir de una base de datos dada, en esto consiste el Aprendizaje Estructural, para ello se utiliza el programa DEAL. Se presentan los supuestos de las distribuciones de probabilidad local, se tiene que los nodos continuos están restringidos a distribuciones condicionales Gaussianas que dependen tanto de los nodos continuos como de los discretos, además debido a que no se permite que haya padres continuos en nodos discretos, las otras probabilidades locales existentes son las que únicamente involucran a nodos discretos, las cuales en principio no están restringidas. La distribución de las probabilidades locales tienen sus correspondientes parámetros los cuales se estiman con un enfoque Bayesiano bajo ciertos supuestos y restricciones sobre sus distribuciones a priori. Para encontrar un buen

modelo se define y utiliza un Puntaje, el cual forma la base de un algoritmo de búsqueda para encontrar tal modelo, que es aquel que maximiza lo más posible el puntaje. Además, en la práctica se restringe esta búsqueda para evitar que la gráfica represente relaciones imposibles entre las variables.

Una vez que se tiene una red adecuada formada por nodos, arcos y probabilidades locales, se forma un Sistema Experto. Pueden obtenerse las probabilidades marginales de cada variable a partir de la red aprendida de los datos (al compilar la red), y también puede introducirse evidencia, que consiste en proporcionar los valores que toman un subconjunto de variables y ver cómo afectan las probabilidades marginales del resto. Se explica brevemente el funcionamiento de algoritmos para llevar a cabo estos procesos, los cuales están sustentados en la Teoría de Gráficas y se utiliza para esta parte el software *Hugin 5.4*.

Posteriormente se presenta el conjunto de variables con las que se trabajó, la manera en que se eligieron, así como el proceso para seleccionar los modelos que finalmente se presentan, al final se obtuvieron los tres modelos siguientes.

Primero, un modelo donde el interés está en considerar la calidad de vida posterior a la estancia en la UTI como una variable binaria, y a las variables: edad con dos categorías, la calidad de vida inicial, la falla neurológica, la falla respiratoria y la cirugía urgente, siendo todas estas variables binarias. Se consideraron solo a los pacientes que sobrevivieron tres meses o más después del egreso de la UTI. Mediante la red obtenida, además de observar gráficamente las relaciones entre las variables, se pueden obtener independencias condicionales entre variables. Se consideró a la calidad de vida posterior como una variable respuesta y al resto como variables explicativas. Se observó que la calidad de vida posterior es independiente condicionalmente a la falla neurológica dada la falla respiratoria y que la falla neurológica era la única variable que no afectaba directamente a la respuesta. Fijando los valores de las variables

explicativas se pudo estimar la probabilidad de buena calidad de vida posterior bajo el modelo gráfico para toda combinación de valores de las variables explicativas. También se utilizó una regresión logística con solo efectos principales con la misma variable respuesta y se estimaron las probabilidades bajo este modelo, el cual solo permite inferir acerca de las relaciones entre todas las variables con una restante y no como en el gráfico, en el cual se puede ver el de todas las variables entre sí; con el modelo logístico se comprobó que cuando el individuo se encuentra en mejores condiciones en cada una de las variables explicativas es más probable que tenga buena calidad de vida posterior. Se compararon las probabilidades estimadas entre ambos modelos observando que la regresión da más importancia a la calidad de vida inicial y el gráfico a la edad. Se obtuvieron para ambos modelos tablas de clasificación para ver el porcentaje de individuos, con buena y mala calidad de vida posterior, clasificados correctamente en la base de datos original (la clasificación que se manejó en este trabajo fue utilizando un valor de corte de 0.5, i.e. si la probabilidad estimada es ≥ 0.5 el individuo se asigna a la categoría buena c.v. posterior y entonces se obtiene cuántos individuos son correctamente asignados; no se utilizó algún tipo de validación cruzada para ver, de otra forma, el poder de predicción de los modelos), resultando que los de buena calidad de vida son un poco mejor clasificados con el modelo gráfico que con el logístico, para la mala calidad de vida ocurre lo opuesto. Los médicos consultados prefieren modelos que predigan adecuadamente a individuos con mala calidad de vida posterior, entonces, en este caso prefieren quedarse con el modelo logístico. Posteriormente se ejemplificó el caso en que en el modelo gráfico la calidad de vida inicial y la posterior se consideraban como variables que no estaban fijas (como respuestas) y las demás sí.

Un segundo modelo en el que se consideran a los pacientes sobrevivientes a los tres meses o más después de su estancia en la UTI y además a los que no sobrevivieron hasta los tres meses. El interés está en la variable que identifica a los vivos de los muertos y se consideró como una variable respuesta binaria, este modelo también incluía: la edad categorizada, la calidad de vida inicial, la mortalidad predicha categorizada, la infección nosocomial, la falla respirato-

ria y la cardiaca. Resultó que la única variable que no afectaba directamente a la respuesta era la falla cardiaca y de hecho estas variables resultaron independientes condicionalmente dadas la edad, la mortalidad predicha, la infección nosocomial y la falla respiratoria. Otra vez se ajustó una regresión logística con la misma variable respuesta y las restantes variables como explicativas resultando que todas fueron significativas y también, como era de esperarse, cuando en cada una de las variables el individuo está en mejores condiciones aumenta la probabilidad de sobrevivir. Nuevamente para toda combinación de valores de las variables explicativas se obtuvieron las probabilidades estimadas de sobrevivencia con ambos modelos. Resultó que el modelo gráfico da más importancia a la calidad de vida inicial y a la edad, en cambio el modelo logístico son más importantes la mortalidad predicha y la falla respiratoria. En las tablas de clasificación resulta que el modelo gráfico clasifica mucho mejor a los sobrevivientes que el modelo logístico; sin embargo, los no sobrevivientes son mucho mejor clasificados con el modelo logístico, además en cada modelo la otra categoría no está tan bien clasificada (en el gráfico los muertos y en el logístico los sobrevivientes). Los médicos consultados prefieren las predicciones de un modelo que clasifique mejor a los no sobrevivientes, así que en este caso se preferirían las predicciones del modelo logístico, aunque por supuesto, en teoría se prefiere un modelo que clasifique bien a todos los individuos y no solo a una categoría; sin embargo, como ya se mencionó, en estos modelos en particular una categoría se clasifica mucho mejor que la otra.

Se ajustó un tercer y último modelo gráfico mixto (con variables continuas y discretas), considerando una variable tricotómica ordinal en la cual la primer categoría corresponde a los vivos con buena calidad de vida posterior, la segunda a los vivos con mala calidad de vida posterior y como última categoría a los no sobrevivientes, esta sería la variable respuesta, se consideró como si fuera continua para permitir que la variable mortalidad predicha, incluída en el modelo y continua, ingrese a ella. Se incluyeron también las variables: edad categorizada, calidad de vida inicial, mortalidad predicha, infección nosocomial, falla neurológica, falla

respiratoria y cirugía urgente. Se observó que edad y cirugía urgente son las únicas variables que no afectan directamente a la respuesta. La variable respuesta resultó independiente condicionalmente de la edad dados la calidad de vida inicial, la mortalidad predicha y la falla respiratoria, también es independiente condicionalmente a la cirugía urgente dadas la calidad de vida inicial, la infección nosocomial y la falla respiratoria. Se utilizó una regresión logística trinomial con la misma variable respuesta (pero en este caso se consideró la variable respuesta no ordenada), la categoría de referencia fue la calidad de vida posterior buena, al comparar los no sobrevivientes con la categoría de referencia para todas las variables, excepto mortalidad predicha, al estar el individuo en mejores condiciones hay mayor riesgo o probabilidad de buena calidad de vida posterior; por otro lado al comparar la mala calidad de vida posterior con la categoría de referencia resulta que la mortalidad predicha y la infección nosocomial no son significativas, para el resto de las variables explicativas cuando el individuo está en mejores condiciones es más probable tener buena calidad de vida posterior que mala. Usando la regresión para un individuo cualquiera al que se le dan valores a todas las variables explicativas se puede estimar la probabilidad en cada categoría de la variable respuesta. Para el modelo gráfico, cuando se dan valores de las variables explicativas, al ser la variable respuesta considerada como continua se obtiene una media y desviación estándar para una distribución Gaussiana (aunque esta distribución ajustada no sería muy buena, porque solo se tienen tres valores numéricos; sin embargo, por características del software se hizo así). En este modelo no se pueden listar todas las combinaciones posibles de valores de las variables explicativas, entonces se ilustra cómo se llevan a cabo las estimaciones con ambos modelos para un par de individuos y se comparan, aunque en sentido estricto los modelos ya no son comparables porque la variable dependiente está considerada de forma distinta.

Finalmente, se ofrecen conclusiones y algunos comentarios respecto a la experiencia obtenida al usar los modelos gráficos probabilísticos, al software utilizado, y al uso de la regresión logística. Entre estas conclusiones y comentarios se tiene que aunque los modelos gráficos son

una herramienta muy útil para representar la relación entre variables y sirve también como una herramienta de clasificación, todavía hay mucho trabajo teórico y técnico en desarrollo. En la parte de Aprendizaje Estructural fue muy necesaria la experiencia de los médicos para establecer cuáles relaciones de dependencia se pueden dar entre las variables, además el paquete utilizado, DEAL, tiene limitantes como el número de variables que se pueden utilizar, que no debe haber datos faltantes, que no se permiten padres continuos para nodos discretos, falta que se tenga implementado algún método de selección de variables, además como la estimación es de tipo Bayesiana sería conveniente tener un Red a priori a partir de la cual se obtengan las estimaciones iniciales para no utilizar de más la información (en este trabajo no se tenía tal red). Por otra parte, además del programa utilizado, DEAL, existen otros (por ejemplo *Hugin 6.3*) que llevan a cabo Aprendizaje Estructural, sería interesante tener acceso a estos programas para comparar los resultados; se presentaron tales comparaciones para el modelo que contiene la variable calidad de vida posterior utilizando una versión de prueba de *Hugin 6.3*. También se puede concluir que tanto las regresiones logísticas ajustadas como los modelos gráficos son dos formas de analizar un problema, cada uno con sus características particulares y limitantes, además se remarca que estos dos tipos de modelos son útiles por si solos y no compiten entre sí, sino que cada uno aporta información distinta que puede ser complementaria.

Capítulo 1

Introducción

En este trabajo se dispone de información correspondiente a pacientes que ingresaron a la Unidad de Terapia Intensiva del Centro Médico Nacional y del Centro Médico la Raza (ambos localizados en la Ciudad de México) en el período 2002-2004. Esta información fue proporcionada por los doctores Héctor Ávila Rosas y Luis David Sánchez Velázquez y corresponde a un proyecto de investigación del Posgrado de Medicina de la Facultad de Medicina de la U.N.A.M. El problema que plantearon los médicos interesados en el problema de investigación correspondía a verificar, para los sobrevivientes a Terapia Intensiva, si algunas características o variables en los individuos ya investigadas en estudios previos, ocasionaban que estos, después de su estancia, tuvieran una calidad de vida buena o mala y también averiguar cuáles otras variables podrían influir en el estado de los pacientes ya egresados de Terapia Intensiva, para así poder predecir qué ocurriría con un paciente que al ingresar a la Terapia Intensiva tuviera características específicas. Por estudios anteriores los médicos sabían que la edad del individuo, su calidad de vida previa al ingreso a la Unidad de Terapia Intensiva y alguna medida de morbilidad influían en la calidad de vida posterior de los individuos; sin embargo, también sospechaban que influía la variable sepsis, corespondiente a un tipo de infección grave. Posteriormente se observó que también era necesario incluir en el análisis

información de los individuos que ingresaron y no sobrevivieron a la Terapia Intensiva, pues los médicos suponían que como consecuencia de la infección los pacientes posiblemente no sobrevivían por lo que era conveniente incluir a estos individuos, también tuvo que recodificarse la variable sepsis y transformarla en una variable binaria con una categoría correspondiente a que el paciente tuvo una infección noscomial cualquiera y la otra correspondiente a que no se infectó o que se infectó fuera del hospital.

Al tratar de modelar un problema existe una gran diversidad de modelos estadísticos que pueden emplearse de acuerdo a las características de la información disponible. En el problema que nos ocupa los datos se modelan mediante modelos gráficos probabilísticos, en particular con las llamadas Redes Bayesianas. Con éstas se logra observar de manera gráfica las relaciones de dependencia que hay entre variables, las dependencias se visualizan mediante "flechas", así cuando hay una flecha que va de una primera variable a una segunda, se puede decir que la segunda variable está dependiendo de la primera. De hecho, también pueden observarse relaciones indirectas entre variables ya que por ejemplo pudiera ocurrir que una variable afecte a una segunda (hay una flecha entre ambas) que a su vez afecta a una tercera (hay una flecha entre ambas variables) y entonces aunque no hay una relación directa entre la primera y tercera variable se están influenciando entre sí indirectamente. Al final el modelo gráfico consiste en una figura formada por flechas y por círculos (u otra figura que representan cada una de las variables) en la que se puede ver cuáles son las relaciones de dependencia e incluso cómo están relacionadas indirectamente las variables entre sí, tal dependencia mencionada se mide mediante probabilidades.

Para construir una red se necesita saber entre cuáles variables hay dependencia y las direcciones de estas dependencias, además como la dependencia se mide mediante

probabilidades condicionales entre las variables también se necesitarían tales probabilidades. Podría darse el caso de que toda esta información fuera proporcionada por un experto ya fuera basándose en su experiencia o en base a estudios previos (Mortera et al, 2002); sin embargo, en este caso no se cuenta con tal información y entonces las relaciones de dependencia entre las variables y sus probabilidades correspondientes se aprenden o estiman a partir de la base de datos. Sin embargo, aunque a partir de los datos se puede generar la red sin ninguna restricción, también se puede restringir el modelo de tal forma que cuando las dependencias entre algunas variables sea inverosímil estas no sean incluídas en el modelo. Al final se obtendría una red que representa adecuadamente de manera gráfica las relaciones entre las variables de interés.

El objetivo de esta tesis es tratar de solucionar el problema médico mencionado arriba usando los modelos gráficos descritos para obtener experiencia de primera mano en cuanto a su uso. Paralelamente, se ajustan modelos de regresión logística, por ser estos más conocidos y usados en investigación médica por la disponibilidad de software para su ajuste.

Entonces, en este trabajo se analizan datos cuyas variables están relacionadas con la estancia de individuos en Unidades de Terapia Intensiva, se modelan tales variables con las gráficas mencionadas y se observan las interdependencias entre las variables. Además, se busca que las relaciones de dependencia obtenidas sean creíbles con lo que en la práctica médica se conoce acerca de las variables involucradas, en este paso sobre todo y en la selección de las variables con las que se va a trabajar, es donde una persona experta en la materia, en este caso médicos, dan su opinión referente al problema para poder agregar esta información al modelo que se va a generar. Una vez obtenido un modelo adecuado para las variables que se decidieron incluir, además de poder de-

terminar las dependencias e indepencias entre las variables involucradas, este modelo permite introducir valores específicos para algún grupo de variables para ver cómo se ven afectados los valores de las demás variables, el objetivo al final sería obtener una red en la que una persona cualquiera pudiera observar qué cambios ocurren en un grupo de variables (o incluso una sola variable) cuando el resto de las variables toman ciertos valores (tales valores debieran ser realistas, en el sentido de que la combinación de valores asignados sean posibles en la práctica), para así decidir qué hacer en base a los cambios obtenidos o hasta poder llevar a cabo medidas de prevención adecuadas para que la persona que se vea obligada a ingresar a una Unidad de Terapia Intensiva pueda salir de la misma con las mejores condiciones posibles y con los mínimos daños posibles

Como ya se mencionó, en el problema que nos ocupa, el interés de los médicos es por un lado obtener un modelo que permita ver qué ocurre con la calidad de vida posterior a la estancia de un individuo en una Unidad de Terapia Intensiva una vez que el paciente ha sido sometido a prácticas médicas dentro de la Unidad, entonces en un primer modelo que se obtiene se usan variables que afecten este hecho, para así ver de que forma están afectando estas variables a la calidad de vida posterior y además ver cómo se afectan entre ellas; también interesa observar qué ocurre cuando esas otras variables involucradas toman ciertos valores específicos (así que se quiere realizar tanto inferencia estadística, como inferencia predictiva). Por otro lado interesa generar también un modelo que incluya a los sobrevivientes y no sobrevivientes a la estancia en la Unidad de Terapia Intensiva y al final surgió un tercer modelo que considera a ambos grupos de pacientes y donde ahora la variable de interés toma tres valores: sobrevivientes con buena calidad de vida posterior, sobrevivientes con mala calidad de vida posterior y no sobrevivientes.

Este trabajo se dividió en cinco capítulos, incluyendo esta Introducción y las Con-

clusiones. En el Capítulo 2 se definen los conceptos necesarios para entender en que consisten los modelos gráficos, en particular las Redes Bayesianas. También se presenta y explica el Aprendizaje Estructural, el cual sirve para obtener a partir de los datos los parámetros correspondientes a las probabilidades condicionales que se tienen en una Red que se haya construído, una vez que se tienen tales parámetros se pueden estimar las probabilidades condicionales de la Red. Además mediante el Aprendizaje Estructural se obtiene a partir de los datos la interdependencia entre variables, las direcciones de tales dependencias y en general cuál es la Red que mejor representa las relaciones que se están dando en los datos, así que con el Aprendizaje Estructural se construye la mejor Red posible y se obtienen las estimaciones de las probabilidades condicionales que hay entre las variables de esta Red. En este capítulo también se introducen los Sistemas Expertos, que es la metodología con la que a partir de una Red con sus respectivas probabilidades condicionales se pueden obtener las probabilidades marginales de cada una de las variables e incluso se puede conocer cómo se ven afectados los valores de las variables cuando un grupo de variables toma valores específicos, en este apartado se explica brevemente un algoritmo eficiente que permite hacer estos procesos.

En el Capítulo 3 se presenta el problema o investigación médica, se describe la información disponible y la información con la que se trabajó, se muestra con que variables se cuenta, con cuáles se trabajó y las modificaciones que se hicieron en algunas de ellas. Se explica también de que forma la opinión experta de los médicos se consideró en los modelos obtenidos, también se hace un historial breve de cómo se fueron especificando los modelos con que finalmente se trabajó. También en este mismo capítulo se explica cómo se usa el programa DEAL necesario para la parte del Aprendizaje Estructural y también se habla brevemente del programa *Hugin* que se utilizó en la parte de Sistemas Expertos.

En el Capítulo 4 se presentan los modelos ajustados con los que finalmente se trabajó para los datos de pacientes en Unidades de Terapia Intensiva, se analizan estos modelos y se hace notar lo que se puede inferir de ellos. También se hace una comparación de los resultados que se obtienen usando los modelos gráficos con aquellos resultados que se obtienen usando regresiones logísticas ajustadas.

Finalmente en el Capítulo 5 se presentan las Conclusiones de este trabajo y se discute acerca de los beneficios y de las limitantes de los modelos gráficos en el contexto de su aplicación al problema específico.

Capítulo 2

Modelos Gráficos Probabilísticos

A lo largo de este capítulo se va a desarrollar toda la teoría indispensable para poder modelar datos con Redes Bayesianas, que son gráficas en las que se van a poder representar las relaciones de dependencia entre variables mediante flechas que unen a las variables relacionadas y cuya aplicación para unos datos específicos se lleva a cabo en los siguientes capítulos. En la primer sección se dan las definiciones, notación y los pasos a seguir para poder modelar con redes probabilísticas, en la siguiente sección se profundiza en el Aprendizaje Estructural, que consiste en poder obtener a partir de los datos la mejor red Bayesiana posible que represente las relaciones entre las variables. Finalmente en la última sección se introduce el concepto de Sistemas Expertos, que son aquellos que al cambiar en las variables de la red algunos valores o dar certeza a algún valor específico obtienen cómo se ven afectadas numéricamente las otras variables, todo esto mediante el uso de algoritmos eficientes.

2.1. Generalidades

La teoría de los modelos gráficos es una combinación entre la teoría probabilística y la teoría de gráficas. Los modelos gráficos son una herramienta para resolver problemas que ocurren en las matemáticas aplicadas e ingeniería, en particular tienen un papel importante en el diseño y análisis de algoritmos de aprendizaje en las computadoras (Inteligencia Artificial). La teoría probabilística es la que permite que todos estos elementos se fundan proporcionando maneras para poder modelar datos (Murphy, 2001).

Muchos de los modelos estudiados en campos como estadística, sistemas ingenieriles, reconocimientos de patrones, etc. son casos particulares de los modelos gráficos. Los modelos gráficos probabilísticos consisten de nodos o vértices, que representan variables aleatorias, y aristas que son líneas que unen los nodos entre sí, cuya ausencia o presencia indica la dependencia o no dependencia entre las variables.

Hay dos clases de modelos gráficos: los no dirigidos y los dirigidos, en los no dirigidos las aristas no tienen una dirección, así que solo representan asociación o correlación entre las variables, mientras que en los dirigidos sí hay una dirección, en cuyo caso se forman "flechas", que en Teoría de Gráficas se denominan arcos. Los modelos gráficos dirigidos, también se conocen como Redes Bayesianas y los no dirigidos como Redes de Markov (los modelos loglineales son un caso particular). También es posible tener un modelo con aristas no dirigidas y dirigidas (arcos) en cuyo caso se forma una Gráfica de Cadena. En este trabajo se van a manejar Redes Bayesianas, en las cuales no hay ciclos dirigidos, o sea no se forma un trayectoria que inicia en un vértice o nodo y termina en el mismo, conservando la dirección de los arcos (Bondy y Murty, 1976).[1]

[1] En este libro se definen formalmente las trayectorias, trayectorias dirigidas, ciclos, ciclos dirigidos y demás conceptos necesarios en la teoría de gráficas

Entonces se puede tomar a $D = (V, E)$, una Gráfica Acíclica Dirigida (Red Bayesiana Acíclica), donde V es un conjunto finito de nodos y E es un conjunto finito de arcos entre los nodos. A cada nodo $v \in V$ en la gráfica le corresponde una variable aleatoria X_v, así que el conjunto de variables asociadas con la gráfica dirigida D es entonces $(X_v)_{v \in V}$. Por lo regular no se distingue entre la variable aleatoria y su correspondiente nodo. Al tener un arco que va de α a β, o sea $\alpha \rightarrow \beta$, se dice que α es padre de β y se puede interpretar informalmente como un indicativo de que α causa β, el conjunto de padres de un vértice v se denomina $pa(v)$. Así, que a cada nodo v con padres $pa(v)$ se le asocia la **distribución de probabilidad local** $p(x_v|x_{pa(v)})$ obteniendo el conjunto de distribuciones de probabilidad para todas las variables P, de donde una red Bayesiana para un conjunto de variables aleatorias X es la pareja (D, P) (para ejemplificar los conceptos se tiene la figura 2.1 en la cual las variables U, V y W son padres de Z y la distribución de probabilidad local de la variable Z estaría dada por $p(Z = z|U = u, V = v, W = w)$, un ejemplo ya con valores se presenta en la sección 2.3). La posible falta de arcos en D indica independencias condicionales entre las variables aleatorias X a través de la factorización de la distribución conjunta.

$$p(x) = \prod_{v \in V} p(x_v|x_{pa(v)}). \tag{2.1}$$

Para una red probabilística cualquiera, la modelación se puede dividir en tres fases (Cowell et. al., 1999, cap. 3):

1. Definir el modelo, o sea
   - Especificar las variables relevantes
   - Especificar la dependencia estructural entre las variables

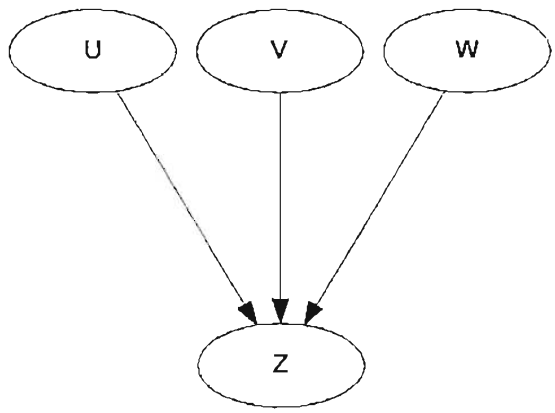

Figura 2.1: Ejemplo de Red Bayesiana

- Asignar las componentes probabilísticas del modelo

2. Construir la maquinaria de inferencia, esto se refiere a que a partir de las probabilidades en la red se construyen algoritmos que nos permiten conocer las probabilidades marginales de cada variable.

3. Usar la maquinaria de inferencia para el análisis de casos, este punto se refiere a que utilizando la maquinaria de inferencia se encuentran las distribuciones marginales de los nodos y además se puede llevar a cabo lo que se conoce como propagación de evidencia, que se refiere a que si nosostros conocemos el valor que toma una variable o varias variables, podemos ver como se modifican todas las probabilidades en cada uno de los nodos de la red.

En el paso 2 y 3 es en donde se utilizan algoritmos computacionales eficientes y es

donde se va a emplear, en nuestro caso, el programa *Hugin 5.4*. En el paso 1 al asignar las probabilidades, muchas veces se supone que un experto en la materia proporciona la estructura de la gráfica, incluyendo las probabilidades; sin embargo, otras veces tales valores no están disponibles, como en este caso, así que a partir de los datos se busca obtener un modelo apropiado, esta es la fase que se conoce como **Aprendizaje Estructural**, la cual es un área muy nueva, tratada por diversos paquetes computacionales que a partir de los datos porporcionan las probabilidades correspondientes. En esta tesis se va a utilizar el paquete DEAL (2003) para el Aprendizaje de Redes Bayesianas, el cual está disponible en lenguaje **R** y desarrollado por Susanne G. Bottcher y Claus Dethlefsen, de la Universidad de Aalborg, Dinamarca; se va a explicar la teoría detrás de este paquete, parte de la cual es exclusiva para el mismo y otra parte se aplica en general.

2.2. Aprendizaje Estructural

Se va a permitir que la red Bayesiana tenga tanto variables continuas como discretas, así que siguiendo la notación de Lauritzen (1996, p. 158), el conjunto de nodos V está dado por $V = \Delta \cup \Gamma$, donde Δ y Γ son el conjunto de nodos discretos y continuos, respectivamente. Entonces el conjunto de variables X puede ser denotado por $X = (X_v)_{v \in V} = (I, Y) = ((I_\delta)_{\delta \in \Delta}, (Y_\gamma)_{\gamma \in \Gamma})$ en donde I y Y denotan el conjunto de variables discretas y continuas, respectivamente. Para una variable discreta, δ, se tiene que I_δ, denota el conjunto de niveles.

Se va a considerar que no se permite que variables discretas tengan padres continuos, esto con la finalidad de que los metodos computacionales sean exactos (Bottcher y Dethlefsen, 2003, p.3), en cuyo caso la distribución conjunta se factoriza en una parte

discreta y una parte mixta, obteniendo:

$$p(x) = p(i, y) = \prod_{\delta \in \Delta} p(i_\delta | i_{pa(\delta)}) \prod_{\gamma \in \Gamma} p(y_\gamma | i_{pa(\gamma)}, y_{pa(\gamma)}).$$

Y en donde $i_{pa(\gamma)}$ y $y_{pa(\gamma)}$ denotan observaciones de los padres discretos y continuos, respectivamente.

Para la parte discreta de la Red, las distribuciones de probabilidad local no están restringidas. Se parametrizan de tal forma que

$$\theta_{i_\delta | i_{pa(\delta)}} = p(i_\delta | i_{pa(\delta)}, \theta_{\delta | i_{pa(\delta)}}), \tag{2.2}$$

donde $\theta_{\delta | i_{pa(\delta)}} = (\theta_{i_\delta | i_{pa(\delta)}})_{i_\delta \in I_\delta}$, o sea el conjunto de parámetros en una variable δ en todos sus posibles niveles dados sus nodos padres. Los parámetros cumplen con $\sum_{i_\delta \in I_\delta} \theta_{i_\delta | i_{pa(\delta)}} = 1$ y además $0 \leq \theta_{i_\delta | i_{pa(\delta)}} \leq 1$. Todos los parámetros asociados a un nodo δ se denotan θ_δ, así que $\theta_\delta = (\theta_{\delta | i_{pa(\delta)}})_{i_{pa(\delta)} \in I_{pa(\delta)}}$. En el programa que se va a emplear, DEAL, si no se cuenta con una Red inicial (Red bayesiana a priori) con sus respectivas probabilidades $p(i_\delta | i_{pa(\delta)})$, entonces se utiliza una distribución $p(i_\delta | i_{pa(\delta)})$ uniforme sobre los niveles de la variable para cada configuración de los padres, i.e.

$$p(i_\delta | i_{pa(\delta)}) = 1 / |I_\delta|,$$

y en donde $|I_\delta|$ es la cardinalidad del conjunto de niveles de la variable δ, o sea el número de niveles de la variable.

Además para la parte discreta de la red, la distribución conjunta estaría dada por

$$p(i) = \prod_{\delta \in \Delta} p(i_\delta | i_{pa(\delta)}).$$

Para la parte mixta de la Red se asume que las distribuciones locales son regresiones lineales Gaussianas sobre los padres continuos, cuyos parámetros dependen de la configuración de los padres discretos. Los parámetros involucrados en la distribución serían $\theta_{\gamma|i_{pa(\gamma)}} = (m_{\gamma|i_{pa(\gamma)}}, \beta_{\gamma|i_{pa(\gamma)}}, \sigma^2_{\gamma|i_{pa(\gamma)}})$, con $\theta_\gamma = (\theta_{\gamma|i_{pa(\gamma)}})_{i_{pa(\gamma)} \in I_{pa(\gamma)}}$. De donde, entonces:

$$(Y_\gamma | i_{pa(\gamma)}, y_{pa(\gamma)}, \theta_{\gamma|i_{pa(\gamma)}}) \sim N(m_{\gamma|i_{pa(\gamma)}} + y_{pa(\gamma)}\beta_{\gamma|i_{pa(\gamma)}}, \sigma^2_{\gamma|i_{pa(\gamma)}}), \tag{2.3}$$

en donde $\beta_{\gamma|i_{pa(\gamma)}}$ está formado por los coeficientes de regresión, $m_{\gamma|i_{pa(\gamma)}}$ es el término constante u ordenada al origen, y $\sigma^2_{\gamma|i_{pa(\gamma)}}$ es la varianza condicional.

Una vez definido lo anterior se puede obtener la distribución conjunta, sea $\theta = ((\theta_\delta)_{\delta \in \Delta}, (\theta_\gamma)_{\gamma \in \Gamma})$ entonces la distribución conjunta para $X = (I, Y)$ está dada por:

$$p(x|\theta) = p(i, y|\theta) = \prod_{\delta \in \Delta} p(i_\delta | i_{pa(\delta)}, \theta_{\delta|i_{pa(\delta)}}) \prod_{\gamma \in \Gamma} p(y_\gamma | i_{pa(\gamma)}, y_{pa(\gamma)}, \theta_{\gamma|i_{pa(\gamma)}}). \tag{2.4}$$

Se demuestra (Bottcher, 2004,p.17), según como se han ido definiendo las distribuciones correspondientes, que la distribución de X es una Gaussiana Condicional (CG), cuya densidad está dada por:

$$p(x|\theta) = p(i, y|\theta) = p(i)|2\pi\Sigma_i|^{-1/2} exp(-1/2(y - M_i)^t \Sigma_i^{-1}(y - M_i)). \tag{2.5}$$

Para cada i, M_i es la media no condicionada a las variables continuas y Σ_i es la matriz de covarianza para todas las variables continuas en la red. De hecho la distribución conjunta $N(M_i, \Sigma_i)$ en (2.5) puede ser obtenida para cada configuración de las variables discretas al usar un algoritmo secuencial (Bottcher y Dethhlefsen, 2003, p.6 apoyado en Shachter y Kenley, 1989) (en donde los parámetros de la Gaussiana dependen de los coeficientes de la regresión $(m_{\gamma|i_{pa(\gamma)}}, \beta_{\gamma|i_{pa(\gamma)}})$ y la respectiva varianza), de tal forma que se puede obtener la distribución conjunta correspondiente.

2.2.1. Obtención de los parámetros

Se obtendrán estimadores en las distribuciones de probabilidad locales, en el caso de DEAL el acercamiento a esta estimación es Bayesiana. Se asume que los parámetros asociados con una variable son independientes de los parámetros asociados con otras variables, a este supuesto (Spiegelhalter y Lauritzen, 1990) se le conoce como independencia parametral global. Adicionalmente, los parámetros obtenidos son independientes para cada configuración de los padres discretos, a esto se le conoce como independencia local parametral. Entonces la distribución conjunta de los parámetros puede expresarse como el siguiente producto

$$p(\theta) = \prod_{\delta \in \Delta} \prod_{i_{pa(\delta)} \in I_{pa(\delta)}} p(\theta_{\delta | i_{pa(\delta)}}) \prod_{\gamma \in \Gamma} \prod_{i_{pa(\gamma)} \in I_{pa(\gamma)}} p(\theta_{\gamma | i_{pa(\gamma)}}). \tag{2.6}$$

A lo anterior (con ambos tipos de independencia) nos referimos como **independencia parametral**. Se tiene que los parámetros son independientes dados los datos. Lo anterior significa que podemos obtener los parámetros a posteriori de un nodo independientemente de los parámetros correspondientes de otros nodos, así que se puede actualizar (obtener la distribución posteriori) la **distribución local parametral a priori** $p(\theta_{v|i_{pa(v)}})$ para cada nodo v en cada configuración de los padres discretos. De lo anterior se demuestra que si se tiene que $\theta_{\delta|i_{pa(\delta)}}$ y $\theta_{\gamma|i_{pa(\gamma)}}$ pertenecen a una familia conjugada (o sea que la distribución a priori pertenece a la misma familia que la posteriori, obtenida después de obtener una muestra d) entonces la distribución conjunta de θ pertenece a una familia conjugada. Se demuestra también, suponiendo que la muestra d es completa, o sea no hay casos faltantes, que la independencia parametral es una propiedad conjugada (propiedad que es preservada para un parámetro después de muestrar, o sea la posteriori $p(\theta|d)$ cumple (2.4) con cada una de las distribuciones condicionadas a d). Las demostraciones se encuentran en Bottcher (2004, p. 17-21).

Para las variables discretas se utiliza como distribución local parametral a priori la Dirichlet y para las variables continuas la Gaussiana para los elementos de la regresión $(m_{\gamma|i_{pa(\gamma)}}, \beta_{\gamma|i_{pa(\gamma)}}|\sigma^2_{\gamma|i_{pa(\gamma)}})$, y Gamma Inversa para la varianza $\sigma^2_{\gamma|i_{pa(\gamma)}}$, todos los parámetros correspondientes a las distribuciones mencionadas se especifican en la siguiente sección. Estas distribuciones son conjugadas a observaciones de las respectivas distribuciones, pues en el caso discreto se trata de observaciones multinomiales (proceso multinomial) cuya familia conjugada es la Dirichlet y en el continuo se demuestra que las distribuciones mencionadas son las adecuadas. El hecho de usar estas distribuciones conjugadas asegura cálculos más simples para las distribuciones posteriori. Para una descripción del proceso multinomial incluyendo las distribuciones a priori y posteriori correspondientes ver Cowell et. al. (1999, p. 269-270) y para el caso continuo ver Bottcher (marzo 2004, p. 22-24).

Como se está tratando el caso mixto, en el que hay variables discretas y continuas, la metodología a emplear es la que propone Bottcher (2001) que se llama "Procedimiento de la Priori Maestra"(*Master Priori Procedure*), que expande los resultados para cuando todas las variables son continuas o todas son discretas, método propuesto por Heckerman(1995) y Greiger y Heckerman (1994). Así que se procede a explicar de manera general tal metodología.

2.2.2. Procedimiento de la Priori Maestra

El procedimiento de la Priori Maestra es el que va a ser empleado para deducir y actualizar (obtener las distribuciones posteriori) los parámetros locales a priori de las distribuciones, tanto para nodos discretos como para continuos. En primer lugar se

necesita especificar una Red Bayesiana inicial que el usuario piensa o supone que tiene, la cual será la **Red Bayesiana a priori**, si no se tiene la información que permite construir tal red inicial o simplemente no se desea comenzar con una red específica se puede emplear una Red vacía, o sea una red que no tiene arcos que unan los nodos entre sí, que es lo que se hará en este trabajo. Entonces los pasos a seguir en el procedimiento son los siguientes.

1. Especificar la Red Bayesiana a priori y las distribuciones locales a priori, como ya se dijo estas se pueden asignar según lo que un experto considere prudente, de lo contrario en el caso discreto se va a utilizar la distribución uniforme discreta $p(i_\delta|i_{pa(\delta)}) = 1/|\mathrm{I}_\delta|$, el caso continuo se explica más adelante. Calcular la distribución conjunta a priori (ecuaciones 2.4 y 2.5).

2. De esta distribución conjunta, se puede calcular la distribución marginal de todos los parámetros en la familia que consiste de cada nodo con sus respectivos padres. A esta se le llama la distribución Priori Maestra (*Master Prior distribution*).

3. Las distribuciones locales a priori de los parámetros se pueden determinar al condicionar la distribución a Priori Maestra y de aquí se pueden obtener las distribuciones a posteriori correspondientes.

El procedimiento anterior asegura independencia en los parámetros, además de la propiedad de que si un nodo tiene el mismo conjunto de padres en dos redes diferentes, entonces el parámetro local a priori para el nodo será el mismo en ambas redes.

NODOS DISCRETOS

Heckerman et al. (1995) y Geiger y Heckerman (1994) desarrollaron el método para el caso de nodos discretos, se utilizan los resultados siguientes.

Sea A un subconjunto de Δ y sea $B = \Delta|A$, tomar a $\psi = (\psi_i)_{i\in I}$ como los parámetros de la distribución conjunta de las variables discretas $p(i)$ y en donde I son los valores que puede tomar i, o sea todas las posibles "celdas"que se pueden formar con las variables discretas. Supóngase que la distribución inicial conjunta de los parámetros es Dirichlet

$$(\psi) \sim D(\alpha),$$

con hiperparámetros $\alpha = (\alpha_i)_{i\in I}$.

Se tiene que la distribución marginal de ψ_A también es Dirichlet

$$(\psi_A) \sim D(\alpha_A),$$

con $\alpha_A = (\alpha_{i_A})_{i_A\in I_A}$, en donde $\alpha_{i_A} = \sum_{j:j_A=i_A} \alpha_j$. Finalmente la distribución condicional $(\psi_{B|i_A})$ es

$$(\psi_{B|i_A}) \sim D(\alpha_{B|i_A}),$$

con $(\alpha_{B|i_A}) = (\alpha_{i_B|i_A})_{i_B\in I_B}$ y $\alpha_{i_B|i_A} = \alpha_i = \alpha_{i_{A\cup B}}$.

Queremos obtener las α_i, para ello se utiliza la siguiente relación que se cumple en las distribuciones Dirichlet:

$$p(i) = E(\psi_i) = \frac{\alpha_i}{N}.$$

En donde $N = \sum_{i\in I} \alpha_i$. Se utilizan las probabilidades en la red a priori como estimadores de $E(\psi_i)$, es decir se obtiene $p(i)$ de la red inicial (la cual en nuestro caso sería $p(i) = \prod_{\delta\in\Delta} 1/|I_\delta|$), así que solo falta obtener N para poder calcular los parámetros α_i. Se determina la N usando la noción de una *base de datos imaginaria* (Bottcher, 2004,

p.27) que consiste en suponer que se tiene una base de datos con un cierto número de casos, con la cual partiendo de una ignorancia completa se actualiza la distribución de ψ. El tamaño de esta *base de datos imaginaria* es N y nos referimos a este valor como el *tamaño de muestra imaginario*, tal tamaño expresa cuánta confianza se tiene en las independencias o dependencias expresadas en la red a priori que hemos proporcionado (este es un valor arbitrario que uno proporciona o que el programa propone y esto significa que en la práctica no se tiene una *base de datos imaginaria*).

Se puede, entonces, determinar a partir de la distribución conjunta de ψ la distribución para la familia $A = \delta \cup pa(\delta)$, pues para ese valor de A, como ya se explicó, se tendría $(\psi_A) \sim D(\alpha_A)$, usando las definiciones dadas arriba. Esta sería la distribución a Priori Maestra en el caso discreto.

Ahora se puede proceder a obtener la **distribución local parametral a priori** a partir de la a Priori Maestra, para ello se calcula la distribución condicional de $\psi_{\delta|i_{pa(\delta)}}$ que es igual a $\theta_{\delta|i_{pa(\delta)}}$ que es la distribución local parametral a priori en el caso discreto. Usando otra vez los resultados de arriba, cuando se condicionaba ψ y en donde el lugar de B lo ocupa δ y el de A lo ocupa $pa(\delta)$ se tiene que (Bottcher, 2003, p. 8):

$$\alpha_{i_\delta|i_{pa(\delta)}} = \alpha_{i_{\delta \cup pa(\delta)}},$$

$$\alpha_{\delta|i_{pa(\delta)}} = (\alpha_{i_\delta|i_{pa(\delta)}})_{i_\delta \in I_\delta},$$

$$\theta_{\delta|i_{pa(\delta)}}|\alpha_{i_\delta|i_{pa(\delta)}} \sim D(\alpha_{\delta|i_{pa(\delta)}}).$$

Así que justamente como se había dicho en **2.2.1** la distribución local parametral a priori es una Dirichlet, entonces usando resultados para un proceso multinomial, se puede obtener la distribución parametral local posteriori correspondiente. Para ello

definir a $n_{\delta|i_{pa(\delta)}}(=(n_{i_\delta|i_{pa(\delta)}})_{i_\delta \in I_\delta})$ como el vector formado por el número de casos observados en un nivel i_δ para una particular configuración de sus padres en la base de datos con un total n de observaciones con esa configuración dada de sus padres.

Entonces la distribución local parametral posteriori $\theta_{\delta|i_{pa(\delta)}}|d$ es también Dirichlet con parámetros $\alpha'_{\delta|i_{pa(\delta)}}$ dados por:

$$\alpha'_{\delta|i_{pa(\delta)}} = \alpha_{\delta|i_{pa(\delta)}} + n_{\delta|i_{pa(\delta)}}.$$

Entonces , por definición de $\theta_{\delta|i_{pa(\delta)}}$ se obtienen las distribuciones de probabilidad local deseadas que irían en nuestra red y que fueron aprendidas usando nuestra base de datos.

NODOS CONTINUOS

Ahora, se van a encontrar las distribuciones parametrales a priori de los nodos continuos, hay que recordar que se trata de distribuciones de probabilidad Gaussianas. El caso en el que tanto los nodos, como sus respectivos padres son continuos (caso Gaussiano puro) se trata en una forma similar al caso discreto con las respectivas distribuciones Gaussianas (Bottcher, 2004, p. 28-30); sin embargo, en nuestro caso los padres de nodos continuos pueden ser tanto discretos como continuos, para el cual Bottcher (2001) desarrolló un procedimiento cuyo caso particular es cuando todos los nodos son continuos.

La solución (Bottcher, 2004, p. 31-33) es similar a la de los casos puros, se especifica una red Bayesiana a priori y se puede obtener la distribución conjunta correspondiente que como ya vimos en (2.5) es una Gaussiana condicional que es el producto de la distribución de i por una distribución Gaussiana:

$$p(i, y|H) = p(i|\psi)N(M_i, \Sigma_i),$$

en donde H es el vector de parámetros de la distribución conjunta, o sea $H = (\psi, (M_i)_{i\in I}, (\Sigma_i)_{i\in I})$. Como en el caso discreto hay que dar la distribución de cada uno de estos parámetros:

$$p(\psi) = D(\alpha),$$
$$p(M_i|\Sigma_i) = N(\mu_i, \frac{1}{\nu_i}\Sigma_i),$$
$$p(\Sigma_i) = IW(\rho_i, \Phi_i).$$

En donde IW se refiere a una distribución Wishart Inversa y en donde cada uno de los términos se obtiene de la forma siguiente. Se define una base de datos imaginaria con un tamaño de nuestra imaginario N. Como en el caso discreto , se pueden calcular los valores α_i para todas las configuraciones de i. Recordar que en el caso discreto se obtiene $\alpha_i = Np(i)$, así que α_i representa cuántas veces se ha observado $I = i$ en la base de datos imaginaria, se puede suponer que cada vez que que observamos las variables discretas I, se observan las variables continuas Y y entonces se toma $\nu_i = \rho_i = \alpha_i$. Para cada configuración de i, sea m_i la media muestral en la base de datos imaginaria y tomar a Σ_i como la varianza muestral. En el caso puro Gaussiano se demuestra que se puede tomar a $\mu_i = m_i$ y a $\Phi_i = (\nu_i - 1)\Sigma_i$, en este caso se hará lo mismo. Sin embargo; como queremos que Φ_i sea positivo, se tiene que hacer a ν_i mayor que 1 para todas las configuraciones de i y esto influye en el tamaño de N que tiene que ser al menos $N = \sum_i \nu_i$.

Se requiere obtener la distribución marginal de los parámetros en A, en particular para $A = v \cup pa(v)$ con $v = \gamma$ (en el caso discreto se obtuvó la ψ_A que era la Priori Maestra con $v = \delta$). Usando el hecho de que la distribución marginal Gaussiana Condicional (CG) de $X_A|H_A$ está dada por:

$$(X_A|H_A) \sim CG(\psi_{i_{A\cap\Delta}}, M_{A\cap\Gamma|i_{A\cap\Delta}}, \Sigma_{A\cap\Gamma|i_{A\cap\Delta}}),$$

Bottcher (2004, p. 32-33) es capaz de obtener la distribución para A en cada uno de los parámetros, la cual denomina la **distribución Priori Maestra Local**, en donde los parámetros que corresponden a distribuciones continuas (Σ y M) se toman sobre $A \cap \Gamma$ porque son los nodos que están en A y a su vez son continuos y algo similar ocurre con ψ para los nodos discretos. De hecho, como lo que nos interesa es obtener la distribución parametral local a priori para los nodos continuos, lo que en realidad queremos es la distribución que van a tener Σ y M. Tales distribuciones están dadas como (Bottcher, 2004, p. 33):

$$\Sigma_{A\cap\Gamma|i_{A\cap\Delta}} \sim IW(\rho_{i_{A\cap\Delta}}, \tilde{\Phi}_{A\cap\Gamma|i_{A\cap\Delta}}),$$

$$M_{A\cap\Gamma|i_{A\cap\Delta}}|\Sigma_{A\cap\Gamma|i_{A\cap\Delta}} \sim N(\bar{\mu}_{A\cap\Gamma|i_{A\cap\Delta}}, \frac{1}{\nu_{i_{A\cap\Delta}}}\Sigma_{A\cap\Gamma|i_{A\cap\Delta}}),$$

en donde $\rho_{i_{A\cap\Delta}} = \sum_{j:j_{A\cap\Delta}=i_{A\cap\Delta}} \rho_j$ y de forma similar para $\nu_{i_{A\cap\Delta}}$ y para $\Phi_{i_{A\cap\Delta}}$. Además:

$$\tilde{\Phi}_{A\cap\Gamma|i_{A\cap\Delta}} = \Phi_{i_{A\cap\Delta}} + \sum_{j:j_{A\cap\Delta}=i_{A\cap\Delta}} \nu_j(\mu_j - \bar{\mu}_{i_{A\cap\Delta}})(\mu_j - \bar{\mu}_{i_{A\cap\Delta}})^t,$$

$$\bar{\mu}_{i_{A\cap\Delta}} = \frac{\sum_{j:j_{A\cap\Delta}=i_{A\cap\Delta}} \mu_j\nu_j}{\nu_{i_{A\cap\Delta}}}.$$

Se puede ver que, intuitivamente la media marginal (la segunda ecuación) es un promedio ponderado de cada grupo, en donde cada grupo es una configuración de los padres discretos sobre los que se marginaliza y en donde cada peso $\nu_j = \alpha_j$ es el número de observaciones en cada grupo. En cuanto a la primera ecuación, esta corresponde a la suma de variación dentro del grupo más la variación entre los grupos. También se hace notar que $\tilde{\mu}_{A\cap\Gamma|i_{A\cap\Delta}} = \mu_{i_{A\cap\Delta}}$ y que la primer expresión solo es notación para identificar

que las μ_j en las expresiones están formados únicamente por los valores correspondientes a variables continuas que además se encuentran en el conjunto A.

Se tiene que para el conjunto A que nos interesa $A \cap \Delta = (\gamma \cup pa(\gamma)) \cap \Delta$ así que $i_{A\cap\Delta} = i_{pa(\gamma)}$, o sea los padres discretos de γ. Por otro lado, $A \cap \Gamma$ está formado por γ junto con todos sus padres continuos, podemos llamar $pa(\gamma)$ a los padres continuos, o sea a $(A \cap \Gamma)|\gamma$. También podemos asumir que $A \cap \Gamma$ está ordenado de tal manera que γ es la primera entrada. Entonces, condicionando la distribución Priori Maestra Local, como en el caso discreto, se obtiene la **distribución parametral a priori local** que sería (Bottcher, 2003, p. 9):

$$(m_{\gamma|i_{pa(\gamma)}}, \beta_{\gamma|i_{pa(\gamma)}}|\sigma^2_{\gamma|i_{pa(\gamma)}}) \sim N_{|pa(\gamma)|+1}(\mu_{\gamma|i_{pa(\gamma)}}, \sigma^2_{\gamma|i_{pa(\gamma)}}\tau^{-1}_{\gamma|i_{pa(\gamma)}}),$$

$$\sigma^2_{\gamma|i_{pa(\gamma)}} \sim I\Gamma\left(\frac{\rho_{\gamma|i_{pa(\gamma)}}}{2}, \frac{\phi_{\gamma|i_{pa(\gamma)}}}{2}\right).$$

En donde (usando las variables que ya se definieron arriba) se utiliza la siguiente partición:

$$\tilde{\Phi}_{A\cap\Gamma|i_{A\cap\Delta}} = \begin{bmatrix} \tilde{\phi}_\gamma & \tilde{\Phi}_{\gamma,pa(\gamma)} \\ \tilde{\Phi}_{pa(\gamma),\gamma} & \tilde{\Phi}_{pa(\gamma)} \end{bmatrix},$$

$$\bar{\mu}_{i_{A\cap\Delta}} = \left(\bar{\mu}_\gamma, \bar{\mu}_{pa(\gamma)}\right).$$

Además:

$$\mu_{\gamma|i_{pa(\gamma)}} = \left(\bar{\mu}_\gamma - \tilde{\Phi}_{\gamma,pa(\gamma)}\tilde{\Phi}^{-1}_{pa(\gamma)}\bar{\mu}_{pa(\gamma)}, \tilde{\Phi}_{\gamma,pa(\gamma)}\tilde{\Phi}^{-1}_{pa(\gamma)}\right),$$

$$\phi_{\gamma|i_{pa(\gamma)}} = \tilde{\phi}_\gamma - \tilde{\Phi}_{\gamma,pa(\gamma)}\tilde{\Phi}_{pa(\gamma),\gamma},$$

$$\rho_{\gamma|i_{pa(\gamma)}} = \rho_{i_{A\cap\Delta}} + |pa(\gamma)|\,,$$

$$\tau_{\gamma|i_{pa(\gamma)}} = \begin{bmatrix} 1/\nu_{i_{A\cap\Delta}} + \bar{\mu}^t_{pa(\gamma)}\tilde{\Phi}^{-1}_{pa(\gamma)}\bar{\mu}_{pa(\gamma)} & -\bar{\mu}^t_{pa(\gamma)}\tilde{\Phi}^{-1}_{pa(\gamma)} \\ -\tilde{\Phi}^{-1}_{pa(\gamma)}\bar{\mu}_{pa(\gamma)} & \tilde{\Phi}^{-1}_{pa(\gamma)} \end{bmatrix}.$$

Ahora, a partir de la distribución local parametral a priori, queremos obtener la posterior, como la distribución a priori obtenida es la conjugada, como habíamos dicho en la sección 2.2.1, entonces las **distribución local parametral posteriori** corresponde también a una Gaussiana para el caso de los regresores y una Gamma Inversa para la varianza de la siguiente forma (ver Bottcher(2004, p. 22-23) y también Bottcher y Dethlefsen (2003, p. 10)):

$$(m_{\gamma|i_{pa(\gamma)}}, \beta_{\gamma|i_{pa(\gamma)}}|\sigma^2_{\gamma|i_{pa(\gamma)}}, d) \sim N_{|pa(\gamma)|+1}(\mu'_{\gamma|i_{pa(\gamma)}}, \sigma^2_{\gamma|i_{pa(\gamma)}}\tau'^{-1}_{\gamma|i_{pa(\gamma)}}),$$

$$(\sigma^2_{\gamma|i_{pa(\gamma)}}|d) \sim I\Gamma\left(\frac{\rho'_{\gamma|i_{pa(\gamma)}}}{2}, \frac{\phi'_{\gamma|i_{pa(\gamma)}}}{2}\right).$$

En donde:

$$\tau'_{\gamma|i_{pa(\gamma)}} = \tau_{\gamma|i_{pa(\gamma)}} + (z^b_{pa(\gamma)|i_{pa(\gamma)}})^t z^b_{pa(\gamma)|i_{pa(\gamma)}},$$

$$\mu'_{\gamma|i_{pa(\gamma)}} = (\tau'_{\gamma|i_{pa(\gamma)}})^{-1}\left(\tau_{\gamma|i_{pa(\gamma)}}\mu_{\gamma|i_{pa(\gamma)}} + (z^b_{pa(\gamma)|i_{pa(\gamma)}})^t y^b_{\gamma|i_{pa(\gamma)}}\right),$$

$$\rho'_{\gamma|i_{pa(\gamma)}} = \rho_{\gamma|i_{pa(\gamma)}} + n,$$

$$\phi'_{\gamma|i_{pa(\gamma)}} = \phi_{\gamma|i_{pa(\gamma)}} + \left(y^b_{\gamma|i_{pa(\gamma)}} - z^b_{pa(\gamma)|i_{pa(\gamma)}}\mu'_{\gamma|i_{pa(\gamma)}}\right)^t y^b_{\gamma|i_{pa(\gamma)}} + (\mu_{\gamma|i_{pa(\gamma)}} - \mu'_{\gamma|i_{pa(\gamma)}})^t\tau_{\gamma|i_{pa(\gamma)}}\mu_{\gamma|i_{pa(\gamma)}}.$$

Y donde n se refiere al número de veces que en la base de datos observada, d, la variable continua y_γ tiene la misma configuración de padres discretos, o sea sería el número de observaciones y^c_γ con $c \in d$ para las que $i^c_{pa(\gamma)} = i_{pa(\gamma)}$. Además $z^b_{pa(\gamma)|i_{pa(\gamma)}}$ es una matriz formada por n renglones con una columna de unos y las otras columnas formadas por los valores observados de los padres continuos para una configuración dada de padres discretos, o sea que para cada padre continuo $l \in pa(\gamma)$ en el que $i^c_{pa(\gamma)} = i_{pa(\gamma)}$

se forma una columna con las n observaciones y_l correspondientes, así que la dimensión de esta matriz es $n \times (|pa(\gamma)| + 1)$. Por otra parte $y^b_{\gamma|i_{pa(\gamma)}}$ es un vector de observaciones del nodo γ para una configuración de padres discretos dada, así que similar a lo anterior está formado por $(y^c_\gamma)_{i^c_{pa(\gamma)}=i_{pa(\gamma)}}$ con $c \in d$.

Entonces, tenemos que a partir de los datos ya se pueden calcular todos los valores que se han descrito en este caso continuo, pues todo ya es conocido, pudiendo estimar los parámetros de las distribuciones Gaussianas para las variables (nodos) continuos. Así que, al igual que en el caso discreto, a partir de los datos se pueden calcular los parámetros y como consecuencia las probabilidades locales (probabilidades condicionales que dependen de quiénes son los padres).

Antes de proseguir, es conveniente señalar algunas de las propiedades que surgen como consecuencia del uso de metodología de la Piori Maestra descrita:

- Los parámetros H de la distribución conjunta $p(i, y|H)$ son independientes.
- Si un nodo v tiene los mismos padres en dos redes diferentes D y D^*, entonces

$$p(H_{v|pa(v)}|D) = p(H_{v|pa(v)}|D^*).$$

  A esta propiedad se le conoce como **modularidad parametral**.

- Sí la distribución conjunta $p(x)$ puede ser factorizada usando una Red Bayesiana Acíclica (DAG) D, entonces también puede ser factorizada usando otras DAG's, que representan el mismo conjunto de independencias condicionales que D, todo este conjunto de redes se dice que son equivalentes en independencia. Así que sí se tienen a D y D^* equivalentes en independencia entonces:

$$p(x|H, D) = p(x|H, D^*).$$

De hecho se demuestra que la función de verosimilitud para los datos d cumple algo similar para redes equivalentes en independencia:

$$p(d|D) = p(d|D^*).$$

Esta equivalencia se conoce como **equivalencia en verosimilitud**, e indicaría que dos redes que representan las mismas relaciones de independencia entre las variables tienen la misma función de verosimilitud.

Hasta aquí, para estas estimaciones se necesita tener identificados quiénes son los padres de las variables, pero al no conocerlos realmente bien quisiéramos encontrar una Red Bayesiana que represente las dependencias e independencias condicionales de las variables que se están manejando de la mejor manera posible, para ello se va a utilizar un "puntaje" (*score*) correspondiente a cada posible red para así elegir la mejor, esto es lo que se desarrollará en la siguiente sección y se verá que la tercera propiedad descrita arriba indicaría que dos redes equivalentes en independencia tienen el mismo puntaje por lo cual se podría usar cualquiera de las dos aunque pudieran parecer diferentes pero ambas conservan las mismas relaciones de independencia entre las variables.

2.2.3. Puntaje en las Redes

Como ya se mencionó, se necesita encontrar un puntaje que indique que tan bien una Red Bayesiana Acíclica D representa las independencias condicionales en los datos, para ello hay que medir que tan probable es D, dado que se han observado los datos d. Se elige aquella Red D con el puntaje más alto. Una medida o puntaje empleado es la probabilidad posterior de la Red, $p(D|d)$, la cual por el teorema de Bayes cumple:

$$p(D|d) \propto p(d|D)p(D),$$

donde $p(d|D)$ es la verosimilitud de D y $p(D)$ es la probabilidad a priori. Se puede ignorar la constante que normaliza la distribución y entonces obtener:

$$p(D,d) = p(d|D)p(D). \tag{2.7}$$

Ambas medidas son puntajes para la Red, en este trabajo se emplea la segunda definición y entonces el puntaje $S(D)$ en una Red está dado por $S(D) = p(d, D)$. En principio se podrían calcular los puntajes de todas las redes posibles y elegir aquella con el puntaje mayor; sin embargo, si hay muchas redes posibles puede ser computacionalmente imposible calcular todos estos puntajes para cada una, en cuyo caso es necesario encontrar una estrategia de búsqueda que permita obtener la Red con el mayor puntaje, esto se explica en la siguiente sección.

El puntaje se factoriza en una parte discreta y una parte mixta, como cuando obteníamos la probabilidad conjunta. Se tendría entonces:

$$S(D) = \prod_{\delta \in \Delta} S_\delta(D) \prod_{\gamma \in \Gamma} S_\gamma(D)$$

donde $S_\delta(D)$ es la contribución del nodo discreto δ y $S_\gamma(D)$ es la contribución del nodo continuo γ.

Según (2.7), se necesita la probabilidad a priori de D y su verosimilitud, pero se va a suponer por simplicidad que todas las redes son equiprobables en cuyo caso lo que va a interesar obtener es $p(d|D)$, para obtenerla se utiliza el hecho de que:

$$p(d|D) = \int_{\theta \in \Theta} p(d|\theta, D)p(\theta|D)d\theta.$$

Primero se obtiene la parte discreta del puntaje $S(D)$, para ello $p(d|\theta, D)$ se adquiere a partir de la parte discreta de la ecuación (2.4)

$$p(d|\theta, D) = \prod_{c \in d} p(x^c|\theta, D) = \prod_{c \in d} \prod_{\delta \in \Delta} p(i_\delta^c | i_{pa(\delta)}^c, \theta_{\delta|i_{pa(\delta)}}, D),$$

y $p(\theta|D)$ se obtiene a partir de la parte discreta de la ecuación (2.6). A partir de estos dos valores se demuestra que la parte discreta del puntaje sin considerar a $p(D)$ (o sea solo tomando la parte discreta de $p(d|D)$) está dada por la expresión siguiente (Bottcher, 2004, p. 24-25)[2]

$$\prod_{\delta \in \Delta} \prod_{i_{pa(\delta)} \in I_{pa(\delta)}} \frac{\Gamma(\alpha_{+\delta|i_{pa(\delta)}})}{\Gamma(\alpha_{+\delta|i_{pa(\delta)}} + n_{+\delta|i_{pa(\delta)}})} \prod_{i_\delta \in I_\delta} \frac{\Gamma(\alpha_{i_\delta|i_{pa(\delta)}} + n_{i\delta|i_{pa(\delta)}})}{\Gamma(\alpha_{i_\delta|i_{pa(\delta)}})}, \tag{2.8}$$

en donde $\alpha_{+\delta|i_{pa(\delta)}} = \sum_{i_\delta \in I_\delta} \alpha_{i_\delta|i_{pa(\delta)}}$ y $n_{+\delta|i_{pa(\delta)}} = \sum_{i_\delta \in I_\delta} n_{i_\delta|i_{pa(\delta)}}$.

De forma similar se puede obtener el puntaje correspondiente a la parte mixta de la red (o sea para los nodos continuos que pueden tener padres tanto discretos como continuos). Otra vez ignorando a $p(D)$, es decir tomando solo a $p(d|D)$, se obtiene la parte continua del puntaje (Bottcher, 2004, p. 25):

$$\prod_{\gamma \in \Gamma} \prod_{i_{pa(\gamma)} \in I_{pa(\gamma)}} \frac{\Gamma\left(\frac{\rho_{\gamma|i_{pa(\gamma)}} + n}{2}\right)}{\Gamma\left(\frac{\rho_{\gamma|i_{pa(\gamma)}}}{2}\right) \sqrt{det(\rho_{\gamma|i_{pa(\gamma)}} s_{\gamma|i_{pa(\gamma)}} \pi)}} \times \left[1 + \frac{1}{\rho_{\gamma|i_{pa(\gamma)}}} a_{\gamma|i_{pa(\gamma)}} s_{\gamma|i_{pa(\gamma)}}^{-1} a_{\gamma|i_{pa(\gamma)}}^t\right]^{\frac{\rho_{\gamma|i_{pa(\gamma)}} + n}{2}} \tag{2.9}$$

donde:

$$s_{\gamma|i_{pa(\gamma)}} = \frac{\phi_{\gamma|i_{pa(\gamma)}}}{\rho_{\gamma|i_{pa(\gamma)}}} \left(I + z_{pa(\gamma)|i_{pa(\gamma)}}^b \tau_{\gamma|i_{pa(\gamma)}}^{-1} (z_{pa(\gamma)|i_{pa(\gamma)}}^b)^t\right),$$

$$a_{\gamma|i_{pa(\gamma)}} = y_{\gamma|i_{pa(\gamma)}}^b - z_{pa(\gamma)|i_{pa(\gamma)}}^b \mu_{\gamma|i_{pa(\gamma)}},$$

[2] Observar que es una fórmula similar a la que se obtiene en un proceso multinomial al obtener la probabilidad marginal de los datos, *vid.* Cowell et. al. (1999, p. 270 fórmula (A.14))

y en donde los valores correspondientes necesarios ya fueron definidos en la sección anterior. Conviene observar que para los cálculos, tanto de esta parte del puntaje como de la parte discreta, se usan los valores de los parámetros correspondientes a la distribución local parametral posteriori proporcionados en la sección anterior.

Entonces, con (2.8) y (2.9) se puede obtener el puntaje total que sería el producto de ambas, multiplicado por $p(D)$, esta última probabilidad como ya se dijo se supone que tiene una distribución uniforme entre todas las redes (equiprobable).

Hay ciertas propiedades en el puntaje obtenido, una de ellas que puede verse tanto en la parte del puntaje discreto como en la parte continua, es que el producto queda factorizado en un producto que involucra sólo a un nodo y a sus padres a lo que se le llama **descomponibilidad**. Otra propiedad, como ya se mencionó anteriormente, es que dos redes equivalentes en independencia, i.e. que representan las mismas relaciones entre las variables, tienen el mismo puntaje. Finalmente conviene señalar, que el paquete DEAL que se va a emplear utiliza el logaritmo del puntaje, en lugar del puntaje simple.

2.2.4. Búsqueda del mejor modelo

Como ya se mencionó, en la búsqueda de la mejor Red Bayesiana Acíclica (DAG) que represente los datos, en teoría se podría calcular el puntaje para todas las posibles redes y elegir áquella o áquellas redes con el puntaje más alto . Robinson (1977) demostró que el número de posibles DAGs con n nodos está dado por la fórmula recursiva:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i).$$

Como las redes que estamos usando son mixtas, con nodos continuos y discretos,

no permitiéndoles a estos últimos que tengan padres continuos, entonces el número de posibles Redes estaría dado por:

$$f(|\Delta|, |\Gamma|) = f(|\Delta|) \times f(|\Gamma|) \times 2^{|\Delta||\Gamma|},$$

en donde $f(|\Delta|)$ y $f(|\Gamma|)$ representa el número de DAGs para los nodos discretos y continuos, respectivamente y en donde $2^{|\Delta||\Gamma|}$ es el número de diferentes combinaciones posibles de arcos que van de nodos discretos a continuos. Si el número de variables en la red es grande es computacionalmente imposible calcular el puntaje para cada una de ellas, es por ello que se buscan estrategias de búsqueda para poder encontrar de forma aproximada la Red con el mayor puntaje sin necesidad de calcular todos los puntajes, el método que se empleará en este caso se denomina de **"Búsqueda ávida"** (*Greedy search*) y este es el método implementado en DEAL.

Para comparar el puntaje de dos Redes Bayesianas acíclicas, D y D^* se utilizan los momios posteriores, es decir

$$\frac{p(D|d)}{p(D^*|d)} = \frac{p(D,d)}{p(D^*,d)} = \frac{p(D)}{p(D^*)} \times \frac{p(d|D)}{p(d|D^*)} = \frac{S(D)}{S(D^*)},$$

en donde $p(D)/p(D^*)$ son los *momios a priori* y $p(d|D)/p(d|D^*)$ es el *factor de Bayes*.

En el paquete DEAL, que se va a emplear, la única opción es dejar que todos las redes sean igualmente probables, así que los momios a priori son iguales a uno y entonces el *factor de Bayes* es lo que se utiliza para comparar dos redes distintas. En la sección anterior obtuvimos a $p(d|D)$ como parte del puntaje, este valor corresponde al producto de las ecuaciones (2.8) y (2.9), así que ya podríamos calcular el *factor de Bayes* para cualesquiera dos redes y compararlas.

Se tiene que para dos modelos que difieren solo en un arco, el *factor de Bayes* es, debido a la propiedad que ya se vio de descomponibilidad del puntaje, especialmente simple. Como ejemplo de la facilidad para calcular el *factor de Bayes* cuando las redes difieren en un arco vamos a suponer que tenemos dos redes, una red D y otra D^*, cuya diferencia está dada por el arco entre v y w, así que se tiene $v \leftarrow w$ en D y $v \not\leftarrow w$ en D^*. En este caso v y w pueden ser como sean, ya sea continuas o discretas, excepto que no se puede que v sea discreta y w continua. Sea ∇_v el conjunto de variables que son padres de v en D^*, así que en D los padres de v serían ∇_v y w. Para simplificar supongamos que se tiene una base de datos d con un solo caso, o sea $d = \{x\}$, como la verosimilitud es descomponible como se vió en (2.1) y además como las probabilidades que no involucran a v en el *factor de Bayes* se cancelan al tratarse de un cociente, entonces:

$$\frac{p(x|D)}{p(x|D^*)} = \frac{p(x_v|x_{w\cup\nabla_v}, D)}{p(x_v|x_{\nabla_v}, D^*)} = \frac{\int p(x_v|x_{w\cup\nabla_v}, H_{v|w\cup\nabla_v}, D)p(H_{v|w\cup\nabla_v}|D)dH_{v|w\cup\nabla_v}}{\int p(x_v|x_{\nabla_v}, H_{v|\nabla_v}, D^*)p(H_{v|\nabla_v}|D^*)dH_{v|\nabla_v}} \quad (2.10)$$

Así que para calcular el *factor de Bayes* entre D y D^*, solo se necesitan considerar los términos que involucran las distribuciones condicionales de v. Entonces en (2.8) y (2.9), por la descomponibilidad de estos factores, se considerarían solo los términos que dependen de v con padres $\nabla_v \cup w$ y los que dependen de v con padres ∇_v.

De forma similar, se puede probar que el caso en el que dos redes D y D^* difieren solamente en la dirección de un arco, es decir se tiene $v \leftarrow w$ en D y $v \rightarrow w$ en D^* (ambos nodos serían discretos o ambos continuos pues hay que recordar que se está usando el hecho de que nodos discretos no tienen padres continuos), se tiene que en el *factor de Bayes* entre D y D^* solo hay que considerar los términos que involucran distribuciones

condicionales de v y de w por lo que el cálculo también es sencillo a partir de (2.8) y (2.9).

Para simplificar más la búsqueda del mejor modelo se identifican las clases dentro de las Redes Bayesianas Acíclicas en las cuales los *factores de Bayes* para probar la existencia o no de un arco entre dos variables son los mismos. Supóngase que se tiene a D_1 y a D_1^* como dos redes que difieren solo por un arco entre v y w, es decir se cumple que $v \leftarrow w$ en D_1 y $v \not\leftarrow w$ en D_1^*. Además, sea ∇_{v1} el conjunto de variables que son padres de v tanto en D_1 como en D_1^*, i.e. v tiene como padres en D_1 a ∇_{v1} y a w y en D_1^* únicamente a ∇_{v1}. Considerar ahora, otras dos redes D_2 y D_2^* (diferentes a D_1 y D_1^*) con un conjunto ∇_{v2} que son padres tanto en D_2 como en D_2^* entre las cuales también hay un arco presente entre v y w en D_2 pero en D_2^* no. Primero se puede considerar cuando se tiene a $v \leftarrow w$ en D_2. Como en (2.10) se pueden encontrar los *factores de Bayes* correspondientes:

$$\frac{p(x|D_1)}{p(x|D_1^*)} = \frac{\int p(x_v|x_{w\cup\nabla_{v1}}, H_{v|w\cup\nabla_{v1}}, D_1)p(H_{v|w\cup\nabla_{v1}}|D_1)dH_{v|w\cup\nabla_{v1}}}{\int p(x_v|x_{\nabla_{v1}}, H_{v|\nabla_{v1}}, D_1^*)p(H_{v|\nabla_{v1}}|D_1^*)dH_{v|\nabla_{v1}}}.$$

Similarmente:

$$\frac{p(x|D_2)}{p(x|D_2^*)} = \frac{\int p(x_v|x_{w\cup\nabla_{v2}}, H_{v|w\cup\nabla_{v2}}, D_2)p(H_{v|w\cup\nabla_{v2}}|D_2)dH_{v|w\cup\nabla_{v2}}}{\int p(x_v|x_{\nabla_{v2}}, H_{v|\nabla_{v2}}, D_2^*)p(H_{v|\nabla_{v2}}|D_2^*)dH_{v|\nabla_{v2}}}.$$

Si suponemos que v cumple con $\nabla_{v1} = \nabla_{v2}$, según se vio en la sección 2.2.2 se cumple la propiedad de modularidad parametral (pues se tienen redes con los mismos padres), así que

$$p(H_{v|w\cup\nabla_{v1}}|D_1) = p(H_{v|w\cup\nabla_{v2}}|D_2)$$

y como $\nabla_{v1} = \nabla_{v2}$, también se cumple:

$$p(x_v|x_{w\cup\nabla_{v1}}, H_{v|w\cup\nabla_{v1}}, D_1) = p(x_v|x_{w\cup\nabla_{v2}}, H_{v|w\cup\nabla_{v2}}, D_2).$$

Entonces, se puede ver que el *factor de Bayes* de D_1 con D_1^* coincide con el *factor de Bayes* de D_2 con D_2^*. En conclusión el *factor de Bayes* para probar la existencia del arco $v \leftarrow w$ en D_1 es equivalente al *factor de Bayes* para probar en otra red D_2 la existencia del arco $v \leftarrow w$ cuando v en D_2 tiene los mismos padres que tiene en D_1.

Similarmente, se puede probar que el *factor de Bayes* para probar la existencia del arco $v \leftarrow w$ en D_1 es equivalente al *factor de Bayes* para probar en otra red D_2 la existencia del arco $v \rightarrow w$ cuando w en D_2 tiene los mismos padres que tiene v en D_1, con la excepción de que v es padre de w en D_2.

Las equivalencias entre redes que se acaban de plantear, se implementan en el algoritmo de Búsqueda Ávida que a continuación se presenta y lo pueden hacer más eficiente teniendo que realizar un menor número de comparaciones entre modelos (este algoritmo eficiente es el que emplea DEAL al utilizar la función *autosearch* que se verá en la sección 3.2). El algoritmo de Búsqueda Ávida es el siguiente:

1. Seleccionar una Red Bayesiana Acíclica inicial D_0 con la cual se comienza la búsqueda.

2. Calcular los factores de Bayes entre D_0 y todas las posibles redes D (o sea $p(d|D)/p(d|D_0)$), que difieren solo por un arco, es decir

   a) Se agrega un arco a D_0

   b) Se borra un arco en D_0

   c) Se cambia la dirección de un arco en D_0

3. Entre todas estas redes, seleccionar aquella que incrementa lo más posible el factor de Bayes.

4. Si el factor de Bayes no es incrementado, detener la búsqueda. En otro caso, la red elegida toma el lugar de D_0 y repetir a partir del paso 2.

Existe la posibilidad de que el máximo que se obtenga, al final del algoritmo, sea un máximo local, una manera de evitar este problema es perturbar aleatoriamente la estructura de la red inicial D_0 y repetir el algoritmo de la Búsqueda Ávida a partir de esta red. Esto puede ser repetido varias veces y entre las redes que se obtengan, se elige aquella con el puntaje más alto, a esta metodología se le llama Búsqueda Ávida con reinicios aleatorios.

2.3. Sistemas Expertos

Los sistemas expertos se pueden definir como sistemas que pretenden codificar y resumir la información de uno o más expertos en un área en una herramienta que pueda ser empleada por personas no especializadas en tal area (Cowell, 1999, p.6). Otra definición alternativa es la de que un sistema experto es un sistema informático (hardware y software) que simula a los humanos en un sistema de especialización dado (Castillo, p. 3).

Los sistemas expertos consisten de dos partes:

Sistema Experto= Conocimiento + Maquinaria de Inferencia

El **Conocimiento** incluye todo lo que se sabe acerca de un problema en un área y generalmente es proporcionado por alguien especializado en la misma. En nuestro caso

ese conocimiento está proporcionado por médicos, con ayuda del Aprendizaje Estructural del que ya se habló en las secciones anteriores. Por otra parte, la **Maquinaria de Inferencia** consiste en uno o más algoritmos para procesar el conocimiento (junto con cualquier otra información) para poder aplicarlo. Ejemplos de uso de estos sistemas expertos son variados: En el diagnóstico médico, pues bajo el supuesto de que un paciente tiene un conjunto de síntomas o características se puede predecir cuál es la enfermedad que más probablemente tiene, otro ejemplo son las aplicaciones que ha hecho Microsoft en sus sistemas de ayuda para el usuario, como en el caso del Asistente de Office que responde al usuario de acuerdo a un sistema experto. También se han usado sistemas expertos en la NASA para predecir según los datos que se tienen la posibilidad de fallas en los sistemas de propulsión de cohetes, etc.

Para los modelos que se están manejando en este trabajo la maquinaria de inferencia es aquello que va a proporcionar algoritmos para poder conocer, a partir de la red obtenida usando los datos, las probabilidades marginales de cada variable, así como permitir el hecho de poder incluir el valor que conozcamos de una o varias variables y a partir de ellos calcular las probabilidades marginales que se modifican de acuerdo a estos valores que es lo que se conoce como propagación de evidencia. De hecho, en la sección 2.1, en la que se habían especificado los pasos para obtener una red, se tenía que el hecho de construir tal maquinaria y usarla correspondían a los dos últimos pasos; en esta sección, lo que se pretende es desarrollar cada uno de estos pasos y dar una breve explicación de que es lo que hace el algoritmo que utiliza el programa *Hugin* (versión 5.4) para poder obtener las probabilidades marginales de cada variable y para propagar evidencia a través de la red.[3]

[3] Para entrar en detalle a la manera en que funcionan todos los algoritmos necesarios para obtener las probabilidades marginales y propagar evidencia, cuando todas las variables son discretas y en el caso mixto se tendrían que revisar los primeros siete capítulos de Cowell (1999).

En primer lugar, para ilustrar algunos conceptos, así como algunos de los pasos que se siguen a lo largo del algoritmo se va a emplear una red sencilla (figura 2.2) en la cual cada uno de los nodos son variables binarias, en ella se tiene la variable M que representa el evento de que el pasto se encuentre mojado y que puede ser causado por el hecho de que el rociador este encendido ($R = verdadero$) o por el hecho de que llueve ($Ll = verdadero$). A su vez la lluvia y el uso del rociador dependen del hecho de que esté o no nublado (variable N). La fuerza de la relación entre las variables está representado por una tabla de probabilidades condicionales, es decir la distribución de probabilidad local, que se presenta en la tabla 2.1. Así por ejemplo $P(M = verdadero|R = verdadero, LL = falso)$ =0.9 y entonces $P(M = falso|R = verdadero, Ll = falso)$ =1-0.9=0.1. Para el caso de la variable N al no tener padres, solo se tiene la probabilidad de que este nublado que es de 0.5.

Mediante la Gráfica y sus probabilidades se observan las siguientes relaciones: Ll (lluvia) es independiente de R (el rociador) dado su padre N (nublado). Esto se escribe como $Ll \perp R|N$. Similarmente se tiene que $M \perp N|R, Ll$, esto es porque no hay una manera ya sea directa o indirectamente en que se pueda llegar de la variable M a la N cuando se tiene fijan las variables R y Ll. Entonces al observar la red que tengamos, podemos darnos cuenta de las independencias condicionales que hay entre las variables.

La distribución conjunta de todos los nodos en la gráfica está dada por:

$$P(N, R, M, Ll) = P(N) \times P(R|N) \times P(LL|N, R) \times P(M|N, R, Ll).$$

Por las relaciones de independencia condicional dadas se tiene que la distribución con-

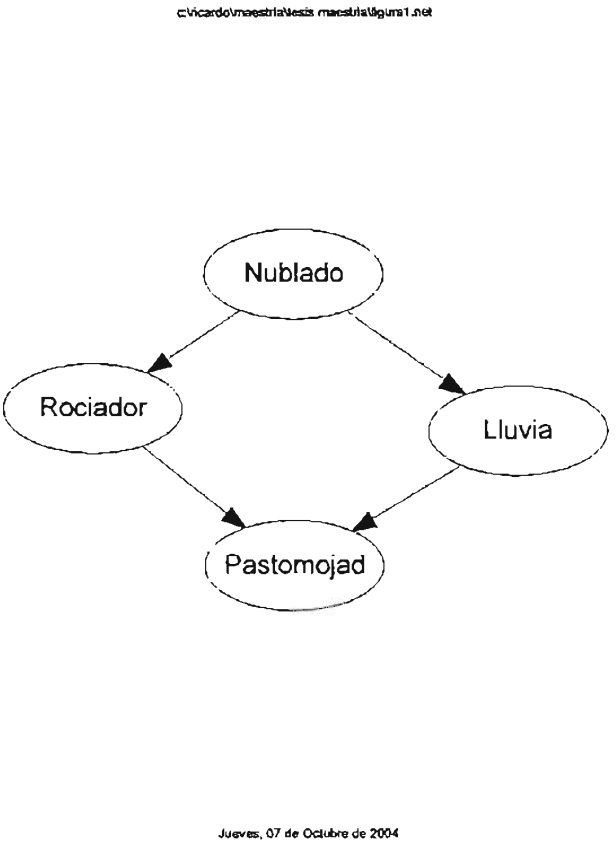

Figura 2.2: Red correspondiente al ejemplo del pasto mojado

junta se reduce a:

$$P(N, R, M, Ll) = P(N) \times P(R|N) \times P(Ll|N) \times P(M|R, Ll).$$

Y entonces, la probabilidad conjunta queda en términos de los valores que conocemos.

Vimos que uno de los primeros objetivos es el de tratar de calcular las probabilidades marginales de cada una de las variables que participan en la red, en el caso de la red en

2.2 el cálculo es sencillo, utilizando propiedades que se tiene en cualquier distribución y lo visto arriba. Así, por ejemplo para calcular la distribución marginal de la variable M

$$P(M = m) = \sum_n \sum_r \sum_{ll} P(N = n, R = r, Ll = ll, M = m).$$

De donde $P(M = m)$ sería igual a

$$\sum_n \sum_r \sum_{ll} P(N = n) \times P(R = r|N = n) \times P(Ll = ll|N = n) \times P(M = m|R = r, Ll = ll)$$

Rociador

Nublado	verdadero	falso
verdadero	0.1	0.5
falso	0.9	0.5

Lluvia

Nublado	verdadero	falso
verdadero	0.8	0.2
falso	0.2	0.8

Pastomojad

Rociador	verdadero		falso	
Lluvia	verdadero	falso	verdadero	falso
verdadero	0.989010	0.9	0.9	0.0
falso	0.010989	0.1	0.1	1.0

Nublado

verdadero	0.5
falso	0.5

Tabla 2.1: Distribución de probabilidad local para el ejemplo del pasto mojado

Al hacer los cálculos, $P(M = verdadero)$ =0.6471. La finalidad de los algoritmos es obtener estas probabilidades marginales, en este caso los cálculos son simples pero a

medida que se complique la red se necesitan algoritmos eficientes. Un primer algoritmo, todavía simple llamado de eliminación de variables consiste en colapsar las sumas tanto como sea posible, en el ejemplo se tendría que $P(M = m)$ sería igual a:

$$\sum_n P(N = n) \sum_r P(R = r|N = n) \sum_{ll} P(Ll = ll|N = n) \times P(M = m|R = r, Ll = ll).$$

Lo cual se puede reescribir como:

$$P(M = m) = \sum_n P(N = n) \sum_r P(R = r|N = n) \times T1(n, m, r),$$

donde:

$$T1(n, m, r) = \sum_{ll} P(Ll = ll|N = n) \times P(M = m|R = r, Ll = ll).$$

Y aún se puede lograr una mayor simplificación:

$$P(M = m) = \sum_n P(N = n) \times T2(n, m),$$

donde:

$$T2(n, m) = \sum_r P(R = r|N = n) \times T1(n, m, r).$$

Entonces en lugar de hacer todas las sumas de una sola vez, se hacen de forma parcial. De esta manera se pueden construir algoritmos sencillos; sin embargo, hay algoritmos más sofisticados como el que emplea *Hugin* que forman parte de lo que se conoce como **programación dinámica**, esto es, algoritmos que calculan muchas distribuciones marginales al mismo tiempo eliminando operaciones de cómputo que son redundantes. El algoritmo que se va a emplear consiste en convertir la red en un árbol (gráfica conexa y acíclica) y a partir de ese árbol que conglomera a varias variables en cada nodo del mismo, se actualizan las probabilidades hasta obtener las distribuciones marginales de cada nodo en la red original, a este algoritmo se le conoce como **algoritmo en árboles de conglomerados** (*junction tree algorithm*).

2.3.1. Algoritmo en árboles de conglomerados

Se va a a hacer una breve descripción del algoritmo de árboles en conglomerados y se va a ilustrar con el ejemplo de la figura 2.2. Los cinco pasos a seguir son los siguientes:

1. Moralizar el modelo gráfico.

   Supóngase que tenemos una Red Bayesiana acíclica K (gráfica dirigida sin ciclos dirigidos), en este paso lo que se hace es transformar la gráfica original en una gráfica no dirigida, para ello se van a ignorar las direcciones de los arcos, así que se sustituyen por aristas que unen los mismos nodos. Además se "casan los nodos padres", esto se refiere a que se unen con una arista cualesquiera dos nodos con un nodo hijo en común. El resultado es una nueva gráfica K^m llamada gráfica moral. Este proceso está ejemplificado en la figura 2.3 para la red de la figura 2.2 que ya se manejó. Una característica de está gráfica moral, es que las propiedades de independencia condicional que son satisfechas por la gráfica original también son satisfechas por la gráfica moral.

2. Triangular la gráfica moral (a esta gráfica triangulada también se le conoce como "de cuerda"(*chordal graph*)).

   Primero hay que definir lo que es una gráfica de cuerda o triangulada. Sea σ un ciclo de tamaño n, en donde el ciclo es una sucesión de n nodos $(\alpha_0, \alpha_1, ..., \alpha_n$ con $\alpha_0 = \alpha_n)$. Una *cuerda* de este ciclo es una arista (α_i, α_j) de nodos no consecutivos. Una gráfica no dirigida se llama triangulada o de cuerda cuando todos sus ciclos de longitud ≥ 4 contienen una cuerda.

   Entonces, para obtener a partir de la gráfica moral, la gráfica triangulada , hay

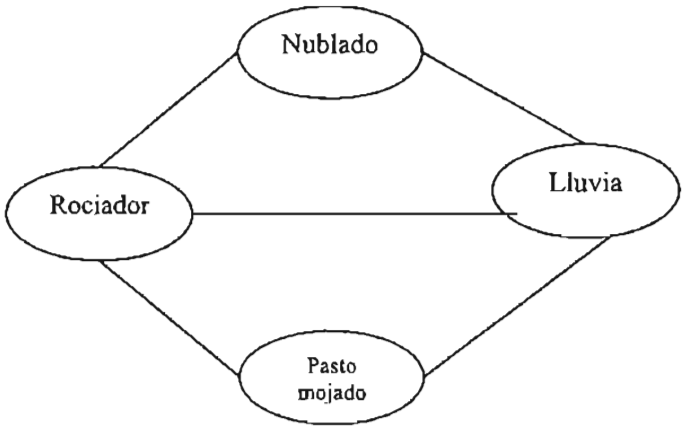

Figura 2.3: Gráfica moral de la red en el ejemplo del pasto mojado

que añadir suficientes aristas a la gráfica moral, hasta obtener una nueva gráfica $(K^m)'$ que es triangulada o de cuerda. Para obtener esta gráfica hay algoritmos (Cowell, 1999, p. 58), en el caso del ejemplo que ya se ha estado manejando, la gráfica moral ya es a su vez triangulada.

3. Encontrar los *cliques* de la gráfica moral.

   Hay que introducir el concepto de ***clique*** o **clan**, para ello hay que definir las gráficas completas, que son aquellas en las que cualesquiera par de vértices están unidos. Una subgráfica es una gráfica que está formada por un subconjunto de vértices y aristas con respecto a la gráfica original. Un *clique* es una subgráfica completa que es maximal (respecto a la contención). Un ejemplo de una gráfica completa se presenta en la figura 2.4 y en el siguiente paso se especifican los *cliques* de la gráfica de la figura 2.3.

4. Obtener un árbol de conglomerados a partir de los *cliques*.

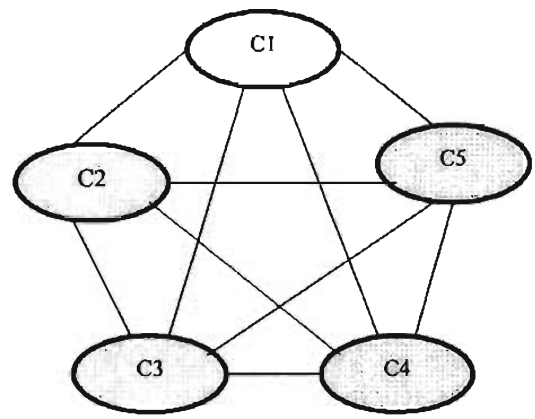

Figura 2.4: Ejemplo de gráfica completa con cinco nodos

Este paso corresponde a que una vez que se tienen los *cliques* en la gráfica moral, hay que juntar todos los nodos que forman cada *clique* y estos van a formar parte de un nuevo nodo, de tal manera que estos nuevos nodos se unen mediante una arista cuando los *cliques* comparten algunas de las variables que los conforman. Esta nueva gráfica que se obtiene es un árbol (gráfica cónexa y acíclica) y es sobre la que se va a trabajar.

Es conveniente definir lo que es un $(\alpha, \beta) - separador$ en una gráfica, este concepto se refiere a un subconjunto $C \subseteq V$ tal que todos los recorridos de α a β intersectan o pasan por C, en donde el concepto de recorrido que se está manejando es el de una sucesión de nodos de tal manera que uno se puede mover de un nodo a

otro, sin importar si se va en dirección contraria a los arcos (o sea se tiene una trayectoria en la versión no dirigida de la gráfica que se esté utilizando). Así, un subcojunto C separa A de B si es un $(\alpha, \beta) - separador$ para todo $\alpha \in A$ y $\beta \in B$. Este concepto es importante, porque resulta que en el árbol de conglomerados, si se toman a dos *cliques* adyacentes C_i y C_j y se toma al conjunto $S = C_i \cap C_j$, este conjunto es un conjunto que separa a la gráfica de cuerda o triangulada (de hecho separa a un gráfica que se llama descomponible, pero resulta que son equivalentes las gráficas de cuerda y esta última). Por otro lado, una propiedad importante de un conjunto separador S cualquiera que separa a A de B en una gráfica G es que se cumple que A es independiente de B condicionado a S, o sea $A \perp B|S$, a esta propiedad se le conoce como propiedad global de Markov[4] y la importancia de esto es que nos permite ver si dos grupos de variables A y B son independientes condicionalmente a un tercer grupo de variables S, que es justamente lo que se hizo en la gráfica 2.2 al ver que $M \perp N|R, Ll$.

En el árbol que formamos, podemos tomar los conjuntos separadores correspondientes (como se dijo están formados por los nodos en la intersección de dos *cliques* adyacentes) como otros nuevos nodos y entonces obtenemos el árbol de conglomerados, el cual no es único dada una gráfica cualquiera. En el caso del ejemplo que se ha manejado, se tienen dos *cliques* que serían $\{R, Ll, N\}$ y $\{R, Ll, M\}$, que serían adyascentes entre sí y cuyo conjunto separador sería $\{R, Ll\}$ entonces se obtendría una gráfica como en la figura 2.5, en donde el conjunto separador está representado por un rectángulo.

[4] De hecho, para que se cumpla esta propiedad se requiere que la distribución p se factorice como en la ecuación (2.11) que se presenta abajo, lo cual se está cumpliendo en nuestro caso,ver Cowell (1999, p. 67)

5. Llevar a cabo la propagación de la información de la red, es decir, calcular las probabilidades marginales (o propagar evidencia) a partir de las probabilidades locales de la red. Este paso se describe adelante.

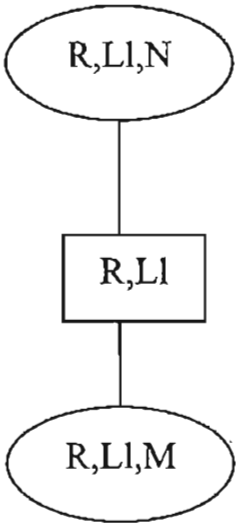

Figura 2.5: Árbol de conglomerados en el ejemplo del pasto mojado

Para todos los procesos que se han descrito hasta ahora hay algoritmos eficientes, los cuales sin importar que tan grande es la gráfica van obteniendo cada uno de los procedimientos descritos (Cowell, 1999, cap. 4). En (2.1) se vio como queda factorizada la distribución conjunta como un producto de las probabilidades locales; a partir de esta factorización resulta que en la gráfica triangulada en la que se tienen los *cliques* la distribución conjunta puede verse que también se descompone en un producto de factores, obteniendo:

$$p(x) = \prod_{c \in C} a_c(x_c), \tag{2.11}$$

así que la distribución p queda en términos de funciones a_c que dependen de los *cliques* c que se encuentran dentro de todo el conjunto de *cliques* C de la gráfica de

cliques formada a partir de la gráfica triangulada. De hecho la ecuación (2.11) se generaliza al incluir además de los *cliques* a los separadores, o sea al usar el árbol de conglomerados, de la siguiente forma:

$$p(x_V) = \frac{\prod_{c \in C} a_c(x_c)}{\prod_{s \in S} b_s(x_s)}, \tag{2.12}$$

en donde S es la familia de todos los separadores y b es función de esos separadores. Las funciones a y b se conocen como *potenciales* (ya no son necesariamente distribuciones de probabilidad, ni son funciones únicas) y cuando se cumple la ecuación (2.12) a toda la colección de funciones $(\{a_c, c \in C\}, \{b_s, s \in S\})$ se le conoce como la *representación potencial de p*. Para llevar a cabo el último paso del algoritmo en árboles de conglomerados (paso 5.) (tal paso es llamado algoritmo de propagación), en primer lugar se deben inicializar justamente estos potenciales, al comienzo se toma $b_s = 1$ y para el caso de la función a_c se inicializa usando un proceso en dos pasos: (1) en primer lugar se toman todas las funciones $a_c = 1$; y entonces (2) tomar cada factor de la densidad de probabilidad de p en la red probabilística original (o sea las distribuciones de probabilidad locales) y multiplicarlo por la función de cualquier *clique* que contenga todas las variables de ese factor como subconjunto y así redefinir los potenciales a_c de los *cliques*. Por ejemplo si se tuviera un *clique* $C1$ formado solo por dos variables $v1$ y $v2$ de tal manera que $v1$ es padre de $v2$ se puede tomar a $p(v2|v1)$ como potencial inicial y así $a_{C1} = p(v2|v1)$, en el ejemplo del pasto se tiene el *clique* $\{M, R, Ll\}$ y en este caso se podría usar a la tabla formada por $p(M|R, Ll)$ para obtener los potenciales iniciales. En conclusión, las funciones potenciales iniciales se obtienen a partir de las probabilidades originales en la red, aunque no siempre sea tan sencillo como en los dos ejemplos descritos arriba y además esta asignación no es única.

Una vez que se tienen los potenciales iniciales, la idea de los algoritmos de propagación a utilizar es la de modificar las funciones potenciales en una secuencia de pasos, lo cual se puede interpretar como ir dejando pasar un flujo a lo largo del árbol de conglomerados, modificando y actualizando cada vez los potenciales, pero de tal manera que la ecuación (2.12) siempre se cumpla. Una vez que todos los pasos se han llevado a cabo, las funciones potenciales finales, contienen la información buscada, es decir cada función potencial es una densidad de probabilidad marginal para el conjunto de variables que la conforman, de tal manera que ya todo queda en términos de probabilidades y entonces (2.12) se transforma en lo que se denomina *representación marginal de p*, obteniendo:

$$p(x_V) = \frac{\prod_{c \in C} p_c(x_c)}{\prod_{s \in S} p_s(x_s)}, \tag{2.13}$$

A partir de esa representación marginal, se pueden obtener las probabilidades marginales de cada una las variables, al hacer una suma sobre las variables de interés en los *cliques*. Por ejemplo, supongamos que se tiene una red con varios nodos o variables[5] , dentro de las cuales se tienen dos variables binarias A(con valores a y $\bar{a}$) y T (con valores t y $\bar{t}$), en la que A es padre de T y se tiene que $\{A, T\}$ forma un *clique*, como ya se dijo un potencial inicial estaría dado por las probabilidades condicionales:

	t	$\bar{t}$
a	0.05	0.95
$\bar{a}$	0.01	0.99

Tabla 2.2: Potencial Inicial

Después de aplicar el algoritmo, se tienen los valores de la Tabla 2.3 en la representación marginal. Entonces, se pueden calcular las probabilidades marginales a partir

[5] Este ejemplo está tomado de Cowell(1999, p. 109) y corresponde a un ejemplo sobre tuberculosis y en el que a partir de la red muestran cómo se aplica el algoritmo, así que los resultados y probabilidades aquí dados son reales

de estos valores, por ejemplo $P(A = a)$ =0.0005+0.0095=0.01.

	t	$\bar{t}$
a	0.0005	0.0095
$\bar{a}$	0.0099	0.9801

Tabla 2.3: Representación marginal, una vez que se ha distribuído la información

Se va a explicar brevemente cómo es que se hace pasar el flujo en los conglomerados que forman parte del árbol de conglomerados T, por simplicidad se toma el caso en que todos los nodos son discretos. Supóngase que se tienen dos *cliques* adyacentes C_1 y C_2 en T, y a S_0 el separador de estos conjuntos, suponer también que se tienen los potenciales correspondientes $\Phi = (\{\phi_c, c \in C\}, \{\phi_s, s \in S\})$ (a todo este conjunto se le conoce como *carga*), lo que se va a hacer es dejar pasar flujo desde la fuente C_1 al pozo C_2. Al hacer pasar el flujo se cambia la carga original Φ (los potenciales originales) por una nueva carga $\Phi^* = (\{\phi^*_c, c \in C\}, \{\phi^*_s, s \in S\})$, en donde los nuevos potenciales para S_0 y C_2 están dados por:

$$\phi^*_{S_0} = \sum_{C_1|S_0} \phi_{C_1},$$
$$\phi^*_{C_2} = \phi_{C_2}\lambda_{S_0},$$

en donde:

$$\lambda_{S_0} = \phi^*_{S_0}/\phi_{S_0}$$

a esta última expresión se le conoce como la *razón actualizada* que lleva el flujo desde S_0 hasta C_2. Este proceso permite que se vaya pasando flujo a través de los conglomerados.

Todo lo descrito anteriormente, permite darse una idea general de cómo funciona el algoritmo en árboles de conglomerados el cual es el proceso que está detrás de la

obtención de las probabilidades marginales de cada nodo o variable de acuerdo a la Red que se está manejando y que es parte de lo que hace el paquete *Hugin*. Sin embargo; otra parte importante es la de propagar evidencia para nuevamente obtener las probabilidades marginales de cada nodo dada esa evidencia. Esto se refiere a que si por ejemplo en el ejemplo del pasto mojado sabemos desde un principio que el pasto está mojado, podemos obtener las probabilidades marginales de cada uno de los otros nodos dado que el pasto está mojado. De forma similar podríamos tener un valor fijo para un conjunto de variables V en la red y entonces al propagar esta evidencia lo que se obtiene es la probabilidad marginal para el resto de las variables dado el conjunto V. Este proceso de propagar evidencia se realiza con el mismo algoritmo en árboles de conglomerados descrito, pero se utilizarían nuevas probabilidades: supóngase que se observa la evidencia $E : X_A = x_A^*$ se define una nueva función p^* tal que:

$$p_x^* = \begin{cases} p(x), & \text{si } x_A = x_A^* \\ 0, & \text{en otro caso} \end{cases}$$

Así, sí en el ejemplo de las variables A y T descrito anteriormente se sabe que $A = a$, entonces los potenciales iniciales al incorporar la evidencia estarían dados por:

	t	$\bar{t}$
a	0.05	0.95
$\bar{a}$	0	0

Tabla 2.4: Potencial Inicial al incorporar evidencia

Y a partir de esta tabla, se aplica el algoritmo como en el otro caso ya visto.

En conclusión, en esta sección se ha dado una breve explicación del algoritmo que emplea *Hugin* para obtener las probabilidades marginales de cada nodo ya sea que se incorpore o no evidencia. Así que a lo largo de este capítulo además de que ya se

puede construir una red a partir de una base de datos, también ya se vio la manera en que se pueden obtener las probabilidades marginales de cada variable, las cuales están dependiendo de las probabilidades locales (la probabilidad de cada variable dados sus padres) de la red construída y no sólo eso sino que también podemos incluir valores que tengamos o que damos a las variables para poder obtener las probabilidades marginales que incorporan esta información, todo usando el mismo algoritmo, el cual está implementado en *Hugin*. En los capítulos siguientes se va a utilizar todo lo que se ha visto, pero ya para un problema particular con unos datos específicos.

Capítulo 3

Metodología

En el capítulo anterior se desarrolló la teoría necesaria para poder ajustar modelos gráficos a información específica. En este capítulo se plantea el problema de aplicación que se aborda en esta tesis, así como el proceso que se siguió para poder obtener, a partir de los datos, un modelo gráfico (una Red Bayesiana) plausible.

3.1. Planteamiento del problema

Los datos corresponden a variables que se les miden a cada uno de los individuos que ingresan a la Unidad de Terapia Intensiva en el Centro Médico Nacional (Hospital Siglo XXI en la Ciudad de México) y en el Centro Médico La Raza. La base de datos correspondiente incluye pacientes elegibles en el periódo 2002-2004 hasta el 21 de octubre de ese año. Cuenta con 923 casos, en los cuales para ciertas variables hay valores no especificados o faltantes; como el paquete DEAL para el Aprendizaje Estructural no maneja datos faltantes hay que filtrarlos, al final, una vez que se eligieron variables y se filtraron los datos faltantes se tiene una base de datos definitiva con 861 registros, que es la que se usa.

Dentro de las variables que se tienen, hay variables que se refieren a fallas que presentó el paciente, como falla cardiaca (*fcard*), respiratoria (*fresp*) y neurológica (*fneur*), las cuales son binarias, pues indican si el paciente tuvo o no las fallas mencionadas. También se tiene la variable binaria cirugía urgente (*cirugia*) que indica si el paciente tuvo o no necesidad de este tipo de intervención quirúrgica. Otra variable es la calidad de vida original, es decir la calidad de vida del paciente tres meses antes de ingresar a la Unidad de Terapia Intensiva (UTI), esta variable también es binaria, indicando sí la calidad de vida original era buena o mala y esta denotada como *cvlcod*. También se maneja la variable mortalidad predicha (*mortpred*), que es una variable continua que los médicos usan como una medida para predecir la mortalidad de los pacientes, para algunas de las redes se construyó otra variable a partir de esta que es la mortalidad predicha codificada (*mortpredcod*) y que consiste en formar dos categorías: la primera cuando la mortalidad predicha toma un valor de 17 o menos y la segunda cuando la variable toma un valor mayor a 17. Por otro lado se tiene la variable infección nosocomial (*infnos*), la cual se construyó para esta tesis como una variable binaria y sirve para identificar cuando un paciente adquirió una infección dentro del mismo hospital, esta variable consta de dos categorías: en la primera se tienen a aquellos pacientes sin infección o que la adquirieron fuera del hospital (infección comunitaria) y la otra categoría corresponde a cuando la infección se adquiere en el mismo hospital. Otra variable importante a considerar es la edad de los pacientes, para fines de esta tesis se categorizó esta variable en dos categorías: menos de 60 años y 61 o más, esta categorización se basa en la opinión de los médicos y sirve para separar a las personas de edad avanzada de las demás (de hecho originalmente se usaron tres grupos: menores de 40 años, de 41 a 60 años y de 61 o más años, pero se vio que la división entre los grupos menores de 40 años y 41 a 60 años no era relevante e incluso hacía que los resultados

fueran menos significativos). Otra variable importante empleada en los modelos es la correspondiente a si el paciente vive o muere durante y al final de su estancia en la UTI (variable *vivomuer*), también fue importante considerar a otra variable que es la calidad de vida posterior a la estancia en la UTI (variable *cv2*), esta variable indica si la calidad de vida después fue buena o mala, así que estas dos categorías solo incluyen a la gente que sobrevive a la UTI, entonces se decidió ingresar una tercera categoría correspondiente a las personas que habían muerto creando una variable nueva llamada *cv2vivom* que incluye tanto a muertos como a vivos con buena calidad de vida y vivos con mala calidad de vida. En la base de datos original hay alrededor de 52 variables que se midieron para este estudio, las cuales se presentan en el Apéndice al final de esta tesis; sin embargo, aquí se han descrito las que fueron empleadas en los modelos definitivos.

Para visualizar los datos, es decir los casos con las respectivas variables de interés, en este trabajo se realiza un breve análisis de los mismos mediante componentes principales, análisis de discriminante y escalamiento multidimensional (Mardia,1979). Las variables de interés, que son las que se emplean en los modelos que más adelante se describen son: edad dicotomizada (*edadcod60*), calidad de vida inicial (*cv1cod*), falla neurológica (*fneur*), falla respiratoria (*fresp*), cirugía urgente (*qxurgent* o *cirugia*), la mortalidad predicha codificada (*mortpredcod*), infección nosocomial (*infnos*), falla cardiaca (*fcard*) y la variable que separa a la calidad de vida posterior buena de la mala y de los muertos (*cv2vivom*) (para una mejor descripción de estas variables ver la tabla 4.1); conviene aclarar que hay variables que se utilizan en los modelos pero que aquí no se introducen como la mortalidad predicha considerándola como una variable continua, la variable que especifica el estado vital o la variable que solo identifica el tipo de calidad de vida posterior, esto es porque de alguna manera estas variables ya están implícitas

en las ya variables mencionadas.

Mediante los componentes principales, que son transformaciones lineales de las variables originales con las que se pretende visualizar los datos en una menor dimensión, se obtuvó que con los tres primeros componentes principales se explica el 51.73 % de la variabilidad, al gráficar los pesos correspondientes a estos tres componentes se observa (figura 3.1) que todas las variables tienen pesos más o menos distintos (los puntos no se acercan entre sí) lo cual podría hacer pensar que no hay correlaciones grandes entre todas estas variables. Si solo se grafican los dos primeros componentes principales (figura 3.2), que explican el 38.52 % de la variabilidad, lo único un poco irregular que se observa es que la variable correspondiente a la falla neurológica se encuentra un poco más alejadas que el resto. A continuación, se calculan los puntajes correspondientes a cada observación obtenidos a partir de los dos primeros componentes principales y se grafican (figura 3.3), en esta gráfica se tienen etiquetados con 1 a los casos que corresponden a personas sobrevivientes con buena calidad de vida posterior, con 2 los casos que correspondientes a personas sobrevivientes con mala calidad de vida posterior y con 3 a los no sobrevivientes; se observan en la parte superior derecha muchos individuos que no sobrevivieron y los cuales se encuentran más o menos separados del resto, poco a poco al ir bajando en dirección a los valores negativos los valores 3 se mezclan con los valores 2, así que los individuos que no sobreviven se empiezan a mezclar con los que tuvieron mala c.v. posterior, finalmente estos últimos individuos se mezclan con los que tuvieron buena c.v. posterior y al final estos últimos individuos se separan un poco del resto.

Otra forma de visualizar los datos es mediante el uso de discriminantes, se utilizan las mismas variables pero en este caso se supone que la variable tricotómica *cv2vivom*

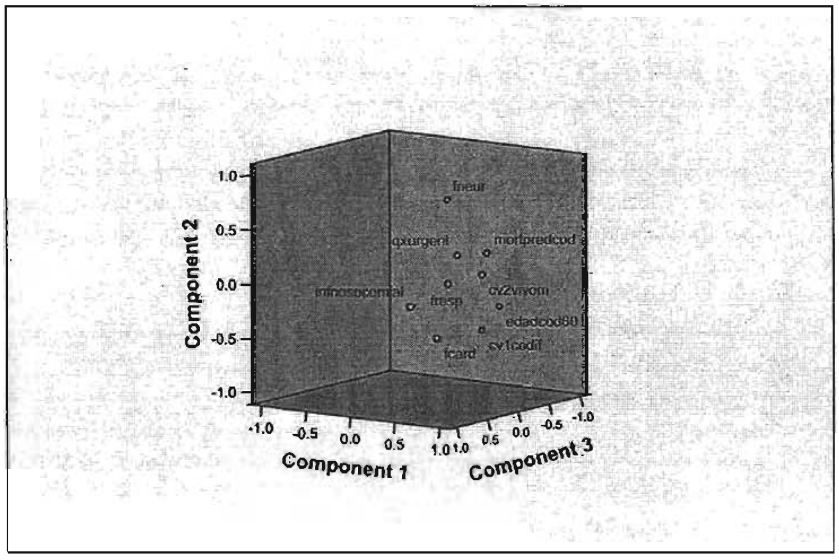

Figura 3.1: Gráfica de pesos correspondientes a los tres primeros componentes principales

separa los datos en tres grupos. Se obtuvó que las medias entre las variables son significativamente diferentes entre los grupos, también se obtienen las funciones discriminantes correspondientes, en este caso como se tienen tres grupos son dos funciones, y los puntajes de los discriminantes asociados a cada observación. Al hacer la gráfica correspondiente (figura 3.4) se observa claramente que los individuos que no sobrevivieron (marcados con M en la gráfica) se encuentran aparte de los demás en la parte superior derecha, en la parte inferior se encuentran los individuos que tuvieron mala c.v. posterior (representados como CVm) los cuales se mezclan bastante con los muertos, de hecho en la tabla de calsificación correspondiente (e incluso tambien en el modelo que se obtiene en la sección 4.3) se observa que hay confusión al clasificar y distinguir entre los sobrevivientes con mala c.v. posterior y los muertos. Finalmente las observaciones correspondientes a individuos con buena c.v. posterior (etiquetados como CVb)

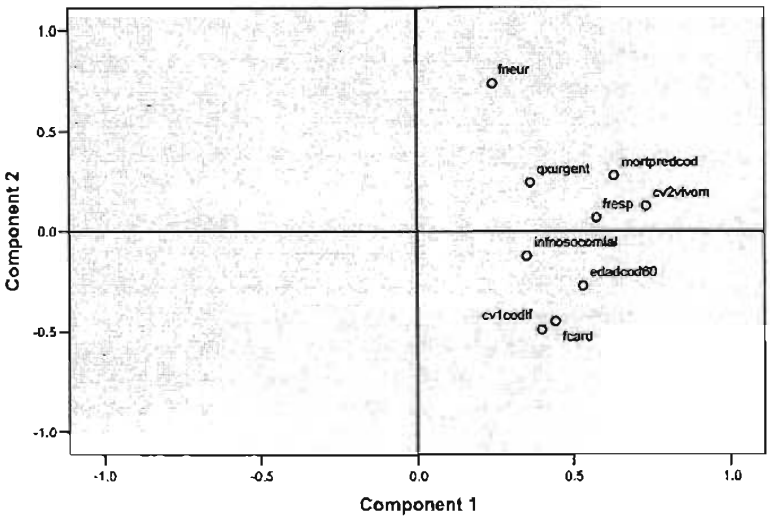

Figura 3.2: Gráfica de pesos correspondientes a los dos primeros componentes principales

se encuentran separados de los demás en la parte superior izquierda, aunque llegan a mezclarse un poco con los individuos con mala c.v. posterior. Entonces se observa que los individuos con mala calidad de vida posterior son los que se encuentran más mezclados con el resto e incluso son con los que hay más problemas en la tabla de clasificación del análisis discriminante pues son los individuos que menos se asignan correctamente a su categoría (solo el 36.2 % son correctamente asignados). También en la misma gráfica se observa que si consideramos a sobrevivientes (uniendo los individuos con etiqueta CVb y con CVm) y por otra parte a los no sobrevivientes, claramente se separan a la derecha los no sobrevientes y a la izquierda los sobrevivientes, lo mismo ocurría con los componentes principales, aunque menos evidente, al considerar el grupo formado por los individuos etiquetados con 1 y 2 y el grupo etiquetado con 3. Similarmente tanto en la figura 3.3 como en la 3.4, si se consideran solamente a los individuos con buena y mala calidad de vida posterior los puntos correspondientes se encuentran más o menos

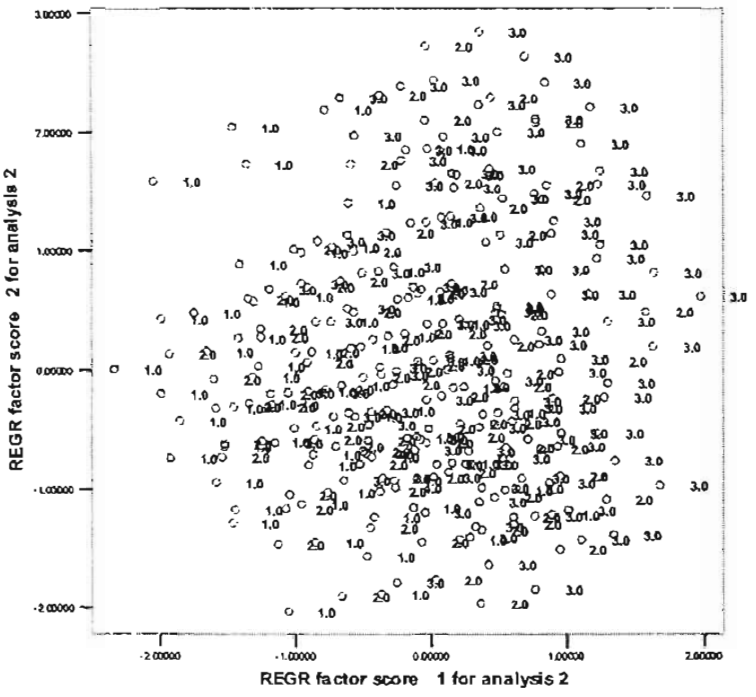

Figura 3.3: Gráfica de puntajes asociados a las dos primeros componentes principales separados entre sí.

Finalmente, otra forma de representar los datos y ver cómo se relacionan las variables es con escalamiento multidimiensional mediante el cual a partir de una matriz de disimilaridad calculada a partir de los datos se representan, en este caso en dos dimensiones, las variables. En el software que se utiliza (SPSS v.13) se utiliza la rutina Proxscal la cual permite calcular la matriz de dismilaridades para datos categóricos, todos binarios, además se utilizó la distancia Euclideana. Primero se hizo el escalamiento multidimensional con las variables ya mencionadas, excepto *cv2vivom*, la cual se sutituyó por la variable que separa individuos con buena y mala calidad de vida posterior (*cv2*), en este caso en la gráfica (figura 3.5) se observan puntos distribuídos de tal ma-

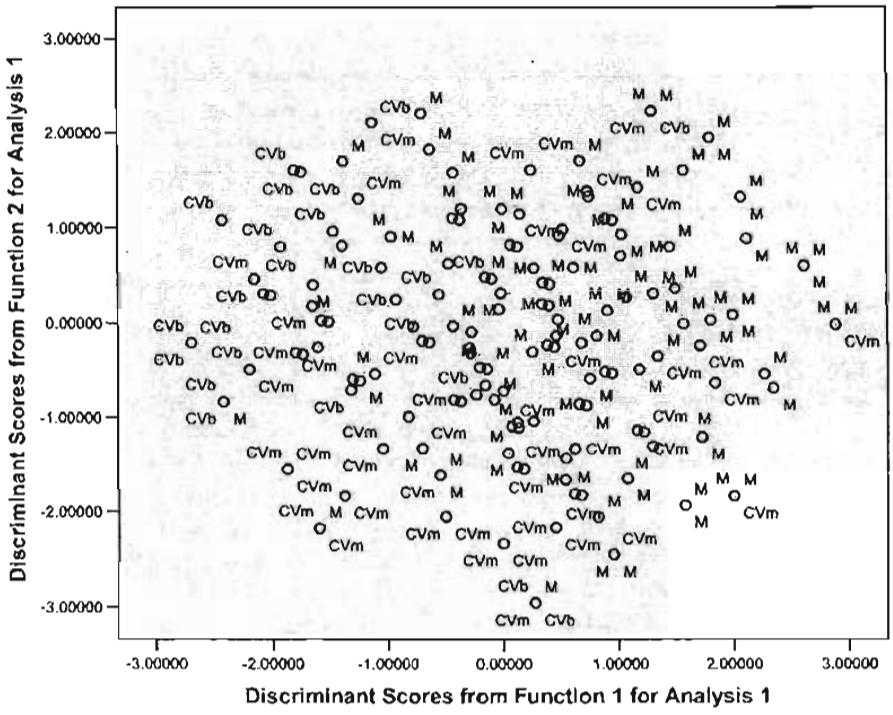

Figura 3.4: Gráfica de puntajes asociados a un análisis de discriminates

nera que no hay un patrón en el que ciertas variables se encuentren demasiado cerca entre sí, lo que se podría mencionar es que la variable relacionada con la mortalidad predicha se encuentra lejos de *cv2*, lo cual es lógico pues mortalidad predicha distingue entre vivos y muertos y no entre calidad de vida posterior buena o mala; también se encuentra un poco alejada la variable falla cardiaca, la variable falla respiratoria y se observa que la variable más cercana a *cv2* es la edad. Después se hizo un escalamiento multidimensional similar al anterior, pero sustituyendo a la variable *cv2* por la variable que distingue a sobrevivientes de no sobrevivientes (*vivomuer*) (figura 3.6), nuevamente todas las variables son binarias y se observa que en esta representación las otras variables no están demasiado cerca de esta última variable, salvo la edad y la calidad de vida inicial, lo cual indica que estas dos variables se relacionan mucho con el estado vital de

los pacientes.

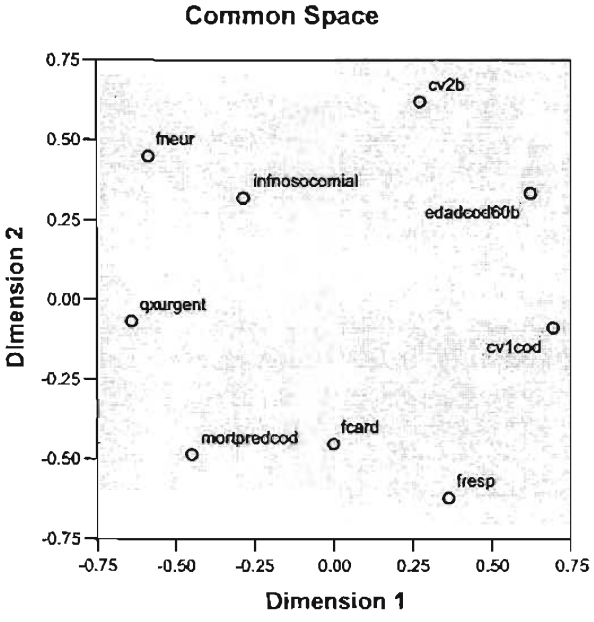

Figura 3.5: Escalamiento Multidimensional que incluye a la variable calidad de vida posterior

El objetivo que se tiene es modelar Redes Bayesianas, las cuales indiquen de manera adecuada la forma en que las variables están influyéndose entre sí, viendo como una variable depende de otra o muchas otras y midiendo este grado de dependencia con las probabilidaes condicionales llamadas distribuciones de probabilidad local ya definidas en el capítulo anterior. Para obtener estas relaciones entre variables se puede emplear el Aprendizaje Estructural, que ya se estudió en la sección 2.2; sin embargo, hay que validar que las relaciones que se obtengan sean coherentes con la realidad y que no se obtenga una Red, usando tal cual los algoritmos del Aprendizaje Estructural, en la cual surjan relaciones en las que una variable resulte ser causa de otra cuando en la realidad no es cierto. Para ello se utiliza la experiencia de los médicos en el área, los cuales indican qué relaciones son imposibles entre las variables y cuáles no, para poder

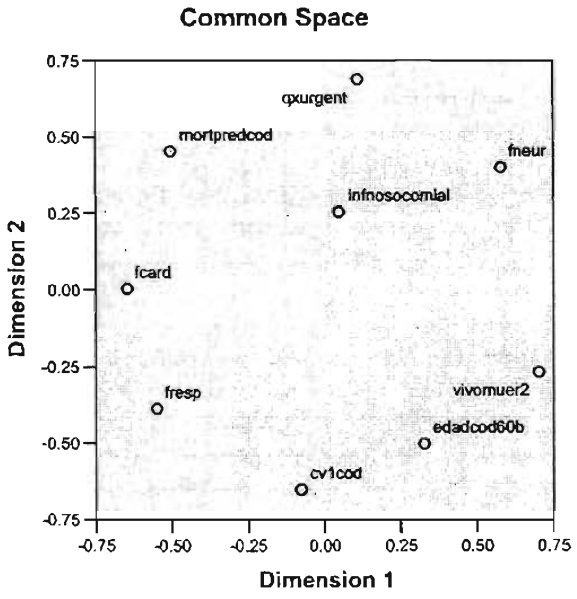

Figura 3.6: Escalamiento Multidimensional que incluye a la variable que separa sobrevivientes de no sobrevivientes

incluir esta información en la Red y que el Aprendizaje Estructural lo tome en cuenta, así, hay que restringir los arcos que puede tener la red que se va a construir, esto es posible hacerlo en DEAL y de hecho es la parte en que en este Sistema Experto en particular entra la experiencia.

Conviene recordar, que aparte de las restricciones dadas por los médicos, hay restricciones en el modelo de caracter técnico; la principal es el hecho de que los nodos o variables continuas no pueden ser padres (o ser la "causa") de nodos discretos, se debe tener muy en cuenta este hecho porque si sabemos que una variable A es causante de otra B, pero resulta que la primera es continua y la segunda discreta, aunque la relación exista en la vida real, es imposible que el modelo la incluya. Sin embargo; hay posibilidades de arreglo, una consiste en que podríamos tomar a la variable B como una variable continua, aunque sea discreta y entonces ya sería posible que hubiera un

arco que parte de A y termina en B, otra posibilidad es discretizar, cuando sea factible, a la variable continua A y entonces ambas variables serían discretas pudiendo haber un arco entre ellas. Así, por ejemplo la variable ya definida que incluye la calidad de vida posterior a la estancia en la UTI y la muerte (denotada *cv2vivom*), adquiere sus valores o es causada por muchas variables, algunas continuas y otras discretas, entonces puede convenir considerarla como una variable continua aunque sea discreta para que así cualquier variable tenga libre acceso a ella.

Tomando distintas variables potencialmente interesantes presentes en la base se probaron varios modelos gráficos, en un principio se fueron incluyendo modelos con muchas variables, algunas de las cuales se fueron eliminando porque no tenían mucha relevancia o podrían resultar repetitivas, por ejemplo las variables que consistían en si el individuo tenía enfermedad pulmonar obstructiva crónica (variable *epoc*), la variable correspondiente a cuántos días el sujeto había usado ventilación mecánica durante su estancia en la UTI (variable *vmdias*) y la variable de falla respiratoria (variable *fresp*) son variables con una alta correlación lineal (esta se observó mediante los componentes principales correspondientes) y resultaban repetitivas, por lo cual bastaba con introducir una sola de ellas, en este caso se tomó la falla respiratoria. Otro criterio que se fue empleando para ver qué variables introducir en el modelo fue la indicación por parte del médico sobre qué relaciones que se daban entre las variables resultaban más importantes y reales en su experiencia en otros estudios y a los resultados que le interesaban inferir y un último criterio fue auxiliarse por un trabajo simultáneo en el que se llevo a cabo un análisis de los mismos datos, pero usando regresiones logísticas; así que las variables que iban resultando significativas en estas regresiones en las que las variables respuesta eran las variables *cv2*, *vivomuer* y *cv2vivom*, también se iban considerando como variables que se introducían en los modelos. Con base a estos criterios se fueron refinando las

redes, hasta obtener las que satisfacieran las necesidades de los médicos, las relaciones que en la experiencia se dan y que no contuviera demasiadas variables, lo cual también podría resultar un impedimento para el programa DEAL.

Las redes que se van obteniendo, permiten ir conociendo que variable va influyendo sobre cuál otra de manera directa cuando un arco las une, pero también las relaciones indirectas que hay entre las variables, cuando por ejemplo desde una variable A hay una trayectoria dirigida que pasa por una o varias variables intermedias antes de llegar a otra variable B, entonces en este caso A y B son variables relacionadas, en la que una influye sobre otra de manera indirecta. Muchos modelos en Estadística, permiten ver cómo influyen varias variables, llamadas variables explicativas, sobre una única variable, a la cuál se le llama la variable respuesta, tal es el caso de las regresiones en cualquier modalidad, como caso particular la regresión logística en la que la variable respuesta es categórica; en estos modelos las relaciones entre varias variables explicativas pueden estar dadas mediante interacciones, pero de todos modos la finalidad de estas relaciones que se establecen entre las variables explicativas es la de explicar mejor (o predecir) la variable respuesta. Los modelos de regresión anteriores se pueden asemejar a aquellos modelos gráficos en los cuales se obliga a que los arcos solo puedan ir o no ir de todas las variables menos una a esa otra variable restante, que tomaría el lugar de la variable respuesta, mientras que todas las demás tomarían el lugar de las variables explicativas. En este sentido, los modelos gráficos pueden pensarse como un modelo más general que el de las regresiones pues se puede ver la influencia que hay entre todas las variables entre sí y no solo con una, que sería la variable respuesta y de hecho un modelo gráfico también podría ser de ayuda para ver que interacciones hay entre las variables explicativas, pues si por ejemplo se tuviera un *clique* entre un subconjunto de variables que para el modelo de regresión son explicativas y de este *clique* se llega a la variable

que se puede considerar como respuesta, entonces se podría probar la interacción entre estas variables e introducirla en la regresión correspondiente. De manera inversa, si se tiene una regresión y se obtienen variables explicativas significativas se podría pensar que en la red debiera de haber arcos de las variables que resultaron significativas en la regresión a la variable que era la respuesta en la misma regresión.

Sin embargo; las regresiones tienen ventajas respecto a los modelos gráficos en el sentido de que uno puede estimar o medir con que "fuerza" las variables influyen al poder obtener los coeficientes estimados y los niveles críticos o "p-values" de las variables y no sólo eso sino que también si damos un conjunto de variables que consideramos explican a otra variable hay métodos automatizados, como el *Forward* o el *Backward*, que auxilian en la elección de un modelo con la menor cantidad de variables que expliquen la respuesta y que además la explican mejor, en cambio, con los programas y software utilizados en este trabajo (*Hugin* y DEAL) en los modelos gráficos uno introduce las variables que considera convenientes y se obtiene la relación entre todas ellas no habiendo un método de selección que permita elegir solo algunas de las variables que serían las más importantes porque son las que se influyen más fuertemente entre sí (aunque si hay estudios acerca de métodos *Forward* y *Backward* para modelos gráficos). De hecho en los modelos gráficos se tiene que entre variables hay o no hay arcos, es decir la influencia se da o no se da y esto es en base a la experiencia que tenemos y al Aprendizaje Estructural, pero cuando alguna relación entre variables existe no sabemos que tan "fuerte" es, sino que únicamente obtenemos las probabilidades condicionales, aunque si sabemos que para que una relación entre variables se haya dado en el modelo gráfico debió de haber sido porque esta relación era importante para incrementar el puntaje de la red como se vio en la sección 2.2.4.

Hasta aquí se ha hablado de la manera en que a partir de la base de datos de UTI se pretenden construir redes que reflejen relaciones que realmente existen en la práctica en la medicina, esto se logra al elegir variables que se consideran que se están influyendo entre sí y al restringir aquellas relaciones entre las variables que se consideran incongruentes con la realidad. También se ha hablado de cómo las redes se pueden relacionar con otros modelos estadísticos y que con ellas podemos entender relaciones entre variables que puede que no sean de manera directa sino indirecta, a través de variables intermedias o a través de dependencias condicionales; tales relaciones indirectas entre variables pueden pasar desapercibidas en otros modelos, así que esta es una ventaja de estos modelos gráficos.

Más adelante, una vez que se tengan redes obtenidas a partir de los datos, que se consideren adecuadas se puede pensar en usar estas redes, en el sentido de que podemos fijar los valores de varias de las variables que conozcamos en algún paciente y ver cómo cambian las probabilidades de las otras variables, pudiendo hacer una especie de predicción de lo que ocurriría en las otras variables para ese paciente. Sin embargo; antes de pasar a esto es conveniente ver cómo funciona DEAL, qué instrucciones hay que proporcionar, cómo se introducen las restricciones en las relaciones entre variables que no son posibles, etc., recordando que todo esto se basa en la teoría que ya se explicó en la sección 2.2.

3.2. Uso de DEAL

DEAL es un programa en lenguaje R disponible de manera gratuita para su uso[1] . En primer lugar, se debe cargar el programa DEAL en R, después hay que leer la base

[1] Disponible en la dirección electrónica http://www.math.aau.dk/novo/deal

de datos, la cual debe de encontrarse guardada como un archivo tipo ASCII en el que cada columna representa una variable y cada renglón corresponde a una observación y en la que como ya se ha dicho no hay valores faltantes. Después se pueden colocar los encabezados correspondientes a los nombres o etiquetas de cada cada variable:

```
> library(deal)
> encabez<-c("edadcodif","cv1cod","cv2","fneur","fres","cirugia")
>hi<-read.table("8122004cv2buenamalaedadcv1cv2fneurfrespqxurgent.dat"
,header=TRUE,col.names=encabez)
```

Después hay que especificar el tipo de variable, ya sea que las variables sean categóricas, en cuyo caso se usa la instrucción "factor" o que se trate de variables continuas, en las cuales no habría que hacer nada pues en principio se asume que todas las variables son numéricas o continuas:

```
> hi$edadcodif<-factor(hi$edadcodif)
> hi$cv1cod<-factor(hi$cv1cod)
> hi$cv2<-factor(hi$cv2)
> hi$fneur<-factor(hi$fneur)
> hi$fres<-factor(hi$fres)
> hi$cirugia<-factor(hi$cirugia)
```

El siguiente paso, como ya se vio en el capítulo anterior, es especificar una Red Bayesiana a priori, si no tenemos alguna red que podría servir como esta red a priori, la cual sería una que se haya obtenido anteriormente en la que se tenga el conocimiento acerca de las relaciones de dependencia que existen entre las variables, entonces se puede usar la Red vacía, es decir nada más con los nodos sin arcos entre ellos, y es la alternativa que se usó.

```
> hi.nw<-network(hi)
```

A continuación se obtiene la distribución conjunta de todas las variables que se encuentran en la red, como ya se dijo en el capítulo anterior en las variables discretas se toma la distribución de probabilidad local uniforme (la distribución dados los padres) y así se obtiene la parte $p(i)$ de la probabilidad conjunta y para los nodos continuos el procedimiento es el que ya se describió en la sección 2.2. Entonces se pueden obtener los parámetros de la distribución conjunta de las variables en la red usando la función *jointprior*(), así obtendríamos los parámetros α, correspondientes a la parte discreta y los parámetros ν, ρ y Φ correspondientes a la parte continua de la red. Hay que recordar que para obtener estos valores usabamos un tamaño de muestra imaginario, el cual se tiene que el mismo programa lo especifica tomando un tamaño razonablemente pequeño, o bien uno puede agregarlo como un argumento adicional. DEAL lo especifica de la siguiente forma: recordar de la sección 2.2.1 que para calcular el tamaño de muestra imaginario se tenía $\nu_i = \alpha_i = Np(i)$ y se necesitaba que $\nu_i \geq 1$ de donde $N \geq 1/p(i)$, en este caso $p(i) = (1/2)^6 = 1/64$ ya que se tienen seis variables, todas categóricas y con dos niveles cada una, así que en este caso en particular $N \geq 64$, entonces aunque con un valor de 64 como muestra imaginaria sería suficiente para asegurar cálculos correctos, DEAL toma en principio el doble de ese valor como el tamaño de muestra imaginario, así que en este caso $N = 128$ como se observa en el segundo renglón de la siguiente instrucción.

```
> hi.prior<-jointprior(hi.nw)
Imaginary sample size: 128
```

El siguiente paso, consiste en crear un listado de arcos prohibidos en la red, por ejemplo la variable edad puede influir a todas las demás variables que se tienen como en

falla respiratoria, la calidad de vida inicial, etc.; sin embargo, no puede ocurrir al revés, que estas variables influyan sobre la edad, entonces hay que prohibir todos los ingresos de las otras variables a la variable edad. Similarmente los médicos proporcionaron la información sobre relaciones entre variables que resultan ser imposibles, todo esto se agrega en el listado de prohibiciones que se va a añadir a la red. El listado consta de dos columnas, donde cada renglón indica el arco que no está permitido. A continuación se presenta uno de esos listados y la instrucción para añadirlo a la red:

```
>banlist<-matrix(c(2,3,4,5,6,3,4,5,6,1,2,3,5,6,2,3,1,3,4,5
,1,1,1,1,1,2,2,2,2,4,4,4,4,4,5,5,6,6,6,6),ncol=2)
> banlist
      [,1] [,2]
 [1,]    2    1
 [2,]    3    1
 [3,]    4    1
 [4,]    5    1
 [5,]    6    1
 [6,]    3    2
 [7,]    4    2
 [8,]    5    2
 [9,]    6    2
[10,]    1    4
[11,]    2    4
[12,]    3    4
[13,]    5    4
[14,]    6    4
```

```
[15,]    2    5
[16,]    3    5
[17,]    1    6
[18,]    3    6
[19,]    4    6
[20,]    5    6
```

```
> hi.nw$banlist<-banlist
```

El siguiente paso es usar la instrucción *learn*(), la cual determina la Priori Maestra, las distribuciones locales parametrales a priori y las distribuciones locales parametrales posteriores, para ello en la instrucción hay que proporcionar la red que se está utilizando, la base de datos que se emplea y la distribución conjunta y parámetros que se obtuvieron con la red a priori o inicial, en este caso la red vacía:

```
> hi.nw<-learn(hi.nw,hi,hi.prior)$nw
```

A continuación, se lleva a cabo la fase del Aprendizaje Estructural en la que se busca el mejor modelo, en el sentido de que es la red que mejora el Puntaje como se vió en la sección 2.2.3 y 2.2.4, para ello se utiliza la instrucción *autosearch*() en la cual se usa todo la información ya obtenida por el programa en los pasos precedentes. El programa despliega la red que mejora el puntaje y también en cada iteración despliega el puntaje de cada una de las redes que se van creando hasta llegar a la mejor.

```
> thebest<-autosearch(hi.nw,hi,hi.prior,trace=TRUE)$nw
[Autosearch (1) -1384.663 [edadcodif][cv1cod][cv2|cv1cod][fneur][fres][cirugia]
(2) -1372.445 [edadcodif][cv1cod|edadcodif][cv2|cv1cod][fneur][fres][cirugia]
(3) -1364.549 [edadcodif][cv1cod|edadcodif][cv2|cv1cod:cirugia][fneur][fres][cirugia]
(4) -1360.323 [edadcodif][cv1cod|edadcodif][cv2|edadcodif:cv1cod:cirugia][fneur]
```

```
[fres][cirugia]
(5) -1357.459 [edadcodif][cv1cod|edadcodif][cv2|edadcodif:cv1cod:cirugia][fneur]
[fres|cirugia][cirugia]
(6) -1356.400 [edadcodif][cv1cod|edadcodif][cv2|edadcodif:cv1cod:fres:cirugia][fneur]
[fres|cirugia][cirugia]
(7) -1355.439 [edadcodif][cv1cod|edadcodif][cv2|edadcodif:cv1cod:fres:cirugia][fneur]
[fres|cirugia][cirugia|cv1cod]
(8) -1354.883 [edadcodif][cv1cod|edadcodif][cv2|edadcodif:cv1cod:fres:cirugia][fneur]
[fres|fneur:cirugia][cirugia|cv1cod]
Total 0.61 add 0.21 rem 0.05 turn 0.04 sort 0.04 choose 0 rest 0.27 ]
```

Finalmente, se puede salvar la red final y exportarla (con el nombre que se quiera), junto con todas las probabilidades obtenidas, a un formato compatible con el programa *Hugin* en el que se trabaja:

```
>savenet(thebest,file("16122004cv2buenamalaedadcv1cv2fneurfrespqxurgent
sinarcoedadacirugfrespafneurcirugiaafneurfrespcirfneurciredadfneur.net"))
```

3.3. Uso de Hugin

Se emplea el programa **Hugin 5.4**[2] , el cual es un programa de tipo comercial para Redes Bayesianas y para el cual, en la versión empleada, uno necesita proporcionar la Red Bayesiana con todo y las probabilidades locales las cuales se debieron de haber obtenido con anterioridad, ya sea que fueran proporcionadas con alguien con experiencia suficiente para dar los valores exactos de las probabilidades correspondientes (Cowell, 1999, p. 29-31) o con experiencia para saber estimarlas con los datos que se cuentan (ejemplo de estas estimaciones en Mortera et al., 2002 y en Cowell, 1999, p. 19-21, 109-111 y 143-147) o bien como en este trabajo, obteniendo estas probabilidades

[2] Hugin está disponible en la dirección electrónica http://www.hugin.dk/

con algún otro programa como DEAL y con ayuda de alguien experto en la materia.

Una vez que se tiene la red en *Hugin* con sus respectivas probabilidades locales, el paso siguiente es *compilar* la red, lo cual se refiere a que se aplican los algoritmos explicados en la sección 2.3 para obtener las probabilidades marginales de cada variable en la red. En el caso de las variables discretas, se obtiene la probabilidad de cada una de las categorías que las conforman y en el caso de las variables continuas se obtiene la media y la desviación estándar correspondientes a la distribución normal de esa variable.

El siguiente paso sería introducir la evidencia, entonces lo que se hace es cambiar las probabilidades marginales que ya se tienen según nuestros deseos, ya sea porque queremos predecir que ocurre con una nueva observación con valores específicos o simplemente porque queremos ver cómo cambian las probabilidades cuando con certeza una variable toma un valor específico. En el caso de los nodos discretos se le da probabilidad de uno a aquella categoría de la variable que forma parte de la evidencia que tenemos y en el caso de los nodos continuos hay que dar un valor para la media y el programa de manera prestablecida considera una desviación estándar de cero, aunque este valor puede ser modificado. Una vez que se ha introducido la evidencia esta se propaga a lo largo de toda la red, según ya se explicó en la sección 2.3, y entonces se obtienen nuevas probabilidades marginales dada la evidencia.

Entonces, como se ha visto, las finalidades de las redes que vamos a obtener son dos principalmente:

1. Por un lado, sirven para describir las relaciones que existen entre las variables, que a veces no son de manera directa, sino a través de variables intermedias que

las relacionan y que luego no se pueden ver en otros modelos estadísticos; sin embargo, estas relaciones deben buscarse coherentes con la realidad y experiencia de los médicos, en este caso. Además también podríamos ver qué relaciones de independencia condicional hay entre las variables involucradas. Todo esto es usando la base de datos que tenemos.

2. Una vez que hemos modelado satisfactoriamente las variables con una red, podemos saber las probabilidades marginales e introducir evidencia para conocer las probabilidades marginales modificadas, de tal manera que podemos identificar qué cambios producen valores específicos en las variables en la red o intentar predecir.

De hecho si se quisiera ir todavía más adelante, en cuestión de necesidades prácticas de los hospitales, se podría obtener al final un modelo o red validado con los datos reales, así que las predicciones para los pacientes que ya se tienen en la base sean lo más certeras posibles, de tal manera que este modelo ya se pueda aplicar para todos los pacientes incluyendo a los nuevos y entonces cuando un paciente nuevo con ciertas características ingresa al hospital ya poder saber el valor de otras variables, por ejemplo poder saber cuánto tiempo permanecerá en la UTI o bien si por ejemplo nuestra variable de interés es la calidad de vida posterior al ingreso del hospital, podríamos ser capaces de obtener cuál es la probabilidad de que su calidad de vida sea mala o que sea buena.

Hasta aquí se ha presentado cómo se hace en la práctica la modelación de los datos de la UTI usando modelos gráficos, se ha hablado de sus beneficios respecto a otros modelos y del alcance que se puede lograr con un buen modelo; sin embargo, hay que tener muy en cuenta que en la práctica hay limitantes como se mencionan a continua-

ción.

Puesto que esta es un área muy nueva de estudio en la que cada vez se van haciendo más y más avances, el software disponible es muy variado y con distintos enfoques. Para la parte del Aprendizaje Estructural no hay un software comercial que sea aplicado por la mayoría de las personas de esta área de estudio, sino que todavía está trabajándose sobre ellos, de hecho el programa *Hugin 5.4* no incluye esta parte, aunque al parecer la nueva versión, *Hugin 6.3*, ya incluye un poco más al respecto; sin embargo, la versión de prueba de *Hugin 6.3* tiene limitantes en cuanto al número de casos y variables que se pueden emplear, por otro lado la versión comercial de este paquete tiene un costo muy elevado (aproximadamente 24,000 pesos la versión para investigación y 8000 pesos la versión para estudiantes) por lo que no se sabe si incluye mejoras en la parte de Aprendizaje Estructural y si permite tener muchas ms observaciones y variables. En el programa DEAL empleado hay algunos inconvenientes como el hecho de que no permite manejar datos faltantes en la base de datos, el supuesto de que los nodos padres continuos no pueden tener hijos discretos (que de hecho también es una limitante presente en *Hugin*), también se tienen supuestos sobre las distribuciones de probabilidad local y también hay limitantes en relación al tiempo que puede tardar el algoritmo habiendo casos en los que incluso no se terminaba el proceso a pesar de que se dejaba pasar mucho tiempo y esto ocurrió cuando se introducían muchas variables, por ejemplo con 12 o 13 variables ya empezaba a haber problemas, sobre todo cuando se tenían modelos mixtos (que contiene variables continuas y discretas) con muchas variables discretas. Se supone que dentro del trabajo futuro para mejorar DEAL se están buscando resolver este tipo de problemas, además de otras mejoras y extensiones (Bottcher, 2003, p.18).

En lo que se refiere a la parte de Sistemas Expertos en la que se obtienen las proba-

bilidades marginales y se introduce evidencia, en el programa *Hugin* se requiere una red ya dada o construirla y para introducir la evidencia se realiza manualmente, entonces a uno le gustaría un procedimiento en que uno diera la evidencia de un conjunto de casos de tal manera que se obtuvieran automáticamente las probabilidades marginales dada la evidencia para todos los casos, esto con la finalidad de que se puedan hacer predicciones de ese conjunto de datos y así podríamos comparar con la misma base de datos manejada cuando es que el modelo gráfico está clasificando a un individuo de manera correcta o incorrecta de acuerdo a las probabilidades. Esta facilidad quiza ya aparece en la última versión.

Otra limitante, es el hecho de que este tema es más trabajado por investigadores del área de Inteligencia Artificial que por estadísticos, entonces falta una mayor integración entre ambas áreas para que haya más avances y programas accesibles para cualquier persona y que esta técnica pueda ser aplicada de manera general con software accesible y que integre todas las áreas necesarias para desarrollar un modelo gráfico para unos datos dados, por esto muchas veces se siguen usando otros modelos más accesibles.

3.4. Selección de los modelos

Se obtienen tres redes distintas: i) la primera utiliza la variable calidad de vida posterior a la estancia en la UTI (*cv2*) junto con otras variables, ii) esta red corresponde a cuando además de otras variables se tiene la variable que distingue vivos de muertos (*vivomuer*) y iii) en esta tercera red se incluye junto con otras variables a la variable *cv2vivom* (que ya se describió), que es tricotómica y que a pesar de ser discreta se toma continua por las limitantes que ya se han explicado respecto a que no puede

haber padres continuos de nodos discretos pues la red que se elige contiene variables continuas que pueden incidir a la variable *cv2vivom*.

A continuación se procede a presentar un historial para describir cómo se llegaron a las tres redes definitivas que fueron las que incluyen variables que en la medicina resultan relevantes en cada caso, que se aproximan más a presentar las relaciones que se dan entre las variables involucradas y que incluyen variables que resultaron importantes al utilizar regresiones logísticas como una herramienta estadística complementaria para elegir un modelo y por supuesto que son obtenidas de acuerdo al programa DEAL empleado para generar redes a partir de las bases de datos.

Se comenzó trabajando solo con los sobrevivientes tomando la variable *cv2*, con base en una regresión logística con *cv2* como variable respuesta y las demás variables de la base como explicativas se empezó con redes que tomaban las variables *edadcod1* (la edad dividida en tres categorías), *epoc* (enfermedad pulmonar crónica), *cancer*, *cv1cod*, *apacheii*, *aps1* (estas dos últimas sirven como medidas sobre la morbilidad de los individuos, por lo que están muy correlacionadas con la variable mortalidad predicha), *sepsis* (un tipo de infección adquirido en hospitales), *vmdias* (días en ventilación mecánica), *estuti* (días en la UTI) y por supuesto *cv2*. Después se experimentó agregando otras variables: falla neurológica (*fneur*), falla respiratoria (*fresp*), cirugía y diabetes.

Posteriormente se comenzo a trabajar con la base que incluye además de los sobrevivientes a los muertos así que se hicieron redes que involucran la variable *vivomuer*, esta correspondía a la nueva variable respuesta en la regresión. Se trabajan redes que incluyen de diferentes maneras las variables *edadcod1*, *epoc*, *cáncer*, *cv1cod*, *apacheii*, *sepsis*, *vm* (el individuo requiere o no de ventilación mecánica), *estuti*, la falla neu-

rológica, falla respiratoria y por supuesto *vivomuer*, se fueron agregando variables como cirugía y diabetes y se van quitando algunas otras variables, aunque en general permanecen en las redes las otras variables ya mencionadas.

Luego, se experimenta al tomar una red para la base de datos con sólo sobrevivientes que no contiene la variable *cv2*. También se toma otra red en la base de datos que incluye todos los individuos tanto los que viven como los que no pero sin la variable *vivomuer*. Posteriormente se trató de ver las correlaciones existentes entre las variables, estas correlaciones también se revisaron auxiliándose de componentes principales al ver los coeficientes que se parecían, resultó que había variables muy correlacionadas:

1. Las cuestiones respiratorias que serían los días en ventiliación mecánica, los días en falla respiratoria (o bien el tener o no falla respiratoria), el tener o no una traqueostomía y otra variable que no es relativa a estas cuestiones que es la cantidad de días en la UTI, es decir la variable *estuti*. Así que las variables correlacionadas son *vmdias* (también *vm*), *fresp*, *Traqueos* y *estuti*.

2. Variables relacionadas a cuestiones clínicas, que son medidas construídas por los médicos para predecir la mortalidad de los individuos que son el aps al ingreso (*aps1*), buselas, mortalidad predicha y por otra parte otra variable que es el nivel de creatinina. Entonces están correlacionadas las variables *aps1, Brus1, mortpred y Creat1*.

Debido a lo anterior se sugiere elegir solo una de las variables en cada rubro y de hecho en los modelos gráficos lo que ocurría era la aparición de *cliques* entre las variables correlacionadas.

Originalmente, los médicos tenían la hipótesis de que la variable binaria *sepsis* influía en *cv2* y en *vivomuer*, por eso fue constantemente incluída en los modelos, posteriormente se vio que esta variable tal como estaba no funcionaba por lo que se modificó colapsando ciertas categorías en la variable a partir de la cual se obtuvó la variable *sepsis*, así se creo una nueva variable para sustituírla que medía algo similar que fue la variable *infnos* (infección nosocomial también llamada *infnosoc*) y que como ya se ha dicho tiene dos categorías una correspondiente a cuando la infección es comunitaria (o sea adquirida fuera del hospital) o que el individuo no se infectó y otra cuando se infecta dentro del hospital. Lo que querían ver los médicos es cómo afecta infectarse en el hospital con infecciones debido al entubamiento de los pacientes, por tratamientos, etc., infectándose con patógenos más potentes y resistentes comparados con los de fuera del hospital y de hecho la variable *sepsis* medía un tipo específico de infección adquirida en el hospital, por ello las variables pueden sustituírse pues ambas representan el interés del médico de ver cómo afectan las infecciones adquiridas dentro del hospital.

Las variables falla respiratoria, falla neurológica y ventilación mecánica, en las redes que se manejaron, se pueden tomar tanto continuas como discretas. En el caso en que se toman continuas se tomaban los días con falla respiratoria, días con falla neurológica y días con ventilación mecánica y en el caso discreto las categorías eran sí o no, o sea el paciente tuvo o no falla respiratoria, tuvo o no falla neurológica, etc., así que se pudo jugar con estas variables, tomándolas continuas o discretas según se quisiera.

A continuación surgió el interés de tomar una variable que incluyera tanto muertos como vivos con sus respectivas calidad de vida después de estar hospitalizados, así surge la variable tricotómica *cv2vivom*, cuya primer categoría coresponde a sobrevivientes con buena c.v. posterior, la segunda a sobrevivientes con mala c.v. posterior y la última co-

rresponde a no sobrevivientes. Se empezó a trabajar en redes que incluían esta variable y en regresiones logísticas trinomiales con esta variable como respuesta.

Después, se empezaron a formalizar más los modelos, en el sentido de que el médico empezó a restringir más las relaciones que podían darse entre las variables para esto incluírlo en la parte del aprendizaje en redes y ver cómo se ajustaban las relaciones resultantes a las que el médico sabía se dan en la práctica. También se hicieron redes cuyos nodos son todas las fallas: nerurológica, respiratoria, cardiaca, hepática, endócrina, renal, además de *sepsis* y *cv2vivom* para ver como se están influenciando las variables entre sí, pues hasta el momento solo se habían usado las fallas respiratoria y neurológica.

Posteriormente se trabaja con la red que contiene las variables *edadcod1, epoc, cáncer, cv1cod, apacheii* (variable continua), *sepsis, estuti* (continua), *vm* (o *vmdias* siendo en el primer caso una variable discreta y el segundo una variable continua), *fneur, fresp* y *cv2vivom*, se empiezan a quitar las variables correlacionadas que son *vmdias, fresp* y *estuti* quitando una a la vez, de dos en dos, etc. se tiene la finalidad de ver si *sepsis* llega a través de algún arco a *cv2vivom*, que era la hipótesis de los médicos, no ocurre esto como ya se había mencionado y entonces la variable *sepsis* no influye sobre *cv2vivom*. Posteriormente se decide utilizar en lugar de *apacheii* a la variable mortalidad predicha (*mortpred*), que es una medida obtenida por los médicos a partir de los datos para predecir mediante un porcentaje la mortalidad de un paciente. También se deciden eliminar ya en las siguientes redes las variables correlacionadas *vmdias* y *epoc*, dejando solo a la falla respiratoria (*fresp*). Dentro de las fallas se decide agregar la falla cardiaca, además de la neurológica que ya estaba presente, esto es debido a los análisis que se hicieron con las fallas, a la importancia de estas fallas explicada por el médico y además debido a los resultados de las regresiones con las que simultánemante se está tra-

bajando, de hecho basándose en las redes se proponían interacciones que pudieran ser utilizadas en la regresión por lo que hay una retroalimentación entre ambos modelos, también se comienzan a hacer modelos gráficos con las variables solo pudiendo apuntar a la variable *cv2vivom* que es la respuesta en la regresión para establecer un paralelismo entre un modelo gráfico restringido y una regresión.

El modelo que se ha obtenido contiene las variables: *edadcod1*, *cancer*, *cv1cod*, *mortpred* (continua), *sepsis*, *estuti* (continua), *cv2vivom*, *fneur*, *fresp* y falla cardíaca (*fcard*), sigue sin ingresar *sepsis* a *cv2vivom* y aquí es donde, como se explicó arriba, se decide cambiar a *sepsis* por *infnos* (infección nososcomial) resultando que esta variable afecta directamente a *cv2vivom*, tanto en la red cuyos arcos solo pueden apuntar a *cv2vivom* como en la red no restringida, además en las regresiones resulta significativa. Se empiezan a obtener redes que están formadas únicamente por variables discretas para ello se crean las variables *mortpredcod* que ya se explicó en la primer sección y también se toma la variable *estuticod* (estancia en la UTI codificada). Se consideraron algunas variables de carácter social como alfabetismo y empleo, así como todas las fallas, obteniendo redes muy grandes, con muchas variables y que ya no resultaban prácticas, ni significativas en las regresiones, así que se retomaron los modelos más sencillos con los que se había trabajado.

Se tenían hasta el momento tres redes: una para la variable *cv2*(calidad de vida posterior buena o mala), otra para *vivomuer* y la última para *cv2vivom*. Cada una contaba además con las variables *edadcod1*, *cáncer*, *cv1cod*, *mortpred* (o *mortpredcod*), *infnos*, *estuti*, *fneur*, *fresp* y *fcard*. Posteriormente se decide quitar a la variable cáncer pues hay pocos casos con esta enfermedad y a *estuti*, pues en el trabajo ya no resultó de interés esta variable, además de que esta aparece correlacionada con falla respiratoria.

Se empieza a trabajar entonces independientemente cada uno de los tres modelos.

Para el caso del modelo con *cv2* se tienen las variables *edadcod1, cv1cod, mortpredcod, infnos, cv2, fneur, fresp y fcard*, se decide quitar la falla cardiaca y sustituirla por la variable cirugía urgente (denotada *qxurgent* o también para este trabajo como *cirugia*), posteriormente *infnos* se puede eliminar pues no resulta importante en la regresión correspondiente, además *mortpredcod* también puede eliminarse puesto que esta medida se crea para identificar a los individuos que viven o mueren y no para ver si un individuo va a tener buena o mala calidad de vida que es en lo que consiste la variable *cv2*, así que al final el modelo que resultó más plausible, después de experimentar quitando y poniendo algunas variables es aquel que contiene las variables: *edadcod60* (esta variable es la edad codificada en dos grupos los menores de 60 años y los mayores de 61), *cv1cod, fneur,fresp* y cirugía urgente (*qxurgent* o *cirugia*), una vez obtenido el modelo final, ya sea este o los otros dos, no se ha concluído pues falta restringir de manera más exigente los arcos y ocurre que al cambiar alguna restricción aparecen arcos que tampoco son muy deseables, así que hay que refinar el modelo hasta llegar a algo coherente y deseado. En este caso, se obtiene un modelo gráfico formado únicamente por variables discretas cuyo modelo ya restringiendo los arcos a aquellas relaciones no plausibles en la práctica, la correspondiente regresión logística, así como la red que solo apunta a *cv2* se presentan en el siguiente capítulo.

Para el caso del modelo que incluye *cv2vivom*, de manera similar que en el modelo anterior para llegar a algo definitivo se trabajó sobre las mismas variables mencionadas, se quita falla cardiaca y se incluye cirugía urgente, después de experimentar diversas variaciones de modelos incluyendo o no incluyendo algunas de las variables mencionadas queda el modelo con las variables *edadcod60, cv1cod, mortpred, infnos, cv2vivom, fneur,*

fresp y cirugía urgente (*qxurgent* o *cirugia*), por consejo del médico en este modelo gráfico entra la variable mortalidad predicha pues era una variable de su interés, a pesar de que en la regresión correspondiente no es siempre significativa, pero hay que recordar que la regresión solo sirve para hacer una comparación y guiarse y no necesariamente lo que indique la regresión tiene que determinar qué variables tomar. En este caso la variable *cv2vivom* se toma como continua, pues *mortpred* es continua y se quiere permitir que haya posibilidad de un arco entre estas variables, además que *cv2vivom* es tricotómica y se observaron algunos problemas al exportar las probabilidades de DEAL a *Hugin* al usar variables tricotómicas discretas, por ello se prefirió tomar a *cv2vivom* como variable continua, aunque no lo sea.

En el caso del modelo que tiene la variable *vivomuer* ocurre algo similar que en los casos anteriores y al trabajar en los modelos incluyendo y quitando las variables se terminó con uno formado por las variables *edadcod60, cv1cod, mortpredcod, infnos, vivom, fresp* y *fcard*, se tiene que este modelo está formado únicamente por variables discretas.

Con el historial anterior se pretende dar una idea de cómo fue el proceso de ir eligiendo o quitando variables, se trata de explicar cómo se trabajó para obtener los modelos finales y cómo a partir de la información dada por los mismos, el médico y las regresiones se llegaron a modelos aceptables. También se puede ver como en el proceso interviene de manera activa la opinión del experto, así como el hacer un estudio concienzudo de las variables involucradas. En el siguiente capítulo se presentan los tres modelos definitivos y lo que se puede inferir de ellos.

Capítulo 4

Resultados

En este capítulo se presenta cada uno de los tres modelos gráficos con los que se trabajó: i) un modelo gráfico para la calidad de vida posterior a la estancia en la UTI, ii) un modelo para la variable que identifica a vivos de muertos, y iii) un último modelo para la variable que identifica a vivos con buena calidad de vida posterior, vivos con mala calidad de vida posterior y muertos. También se presentan las regresiones logísticas correspondientes en las que las variables ya mencionadas son las variables respuestas y en el que las variables explicativas son las variables restantes para cada uno de los tres modelos gráficos. Posteriormente se hace una comparación para cada uno de los tres modelos entre los resultados que se obtienen utilizando la regresión logística y el modelo gráfico.

Un resumen de las variables que fueron utilizadas a lo largo de este capítulo, sus nombres y en caso de que la variable sea categórica el significado de cada una de sus categorías se presentan en la tabla 4.1.

Etiqueta	Variable	Categorías
edadcod60 (o *edad*)	Edad dicotomizada	1=60 o menos años de edad 2=61 o más años
cv1cod	Calidad de vida del paciente dos meses previos a la hospitalización dicotomizada	1=Buena calidad de vida inicial 2=Mala calidad de vida inicial
cv2	Calidad de vida del paciente a los tres meses del alta hospitalaria dicotomizada	1=Buena calidad de vida 2=Mala calidad de vida
fneur	Falla neurológica	0=No hay falla; 1=Sí hay falla
fresp	Falla respiratoria	0=No hay falla; 1=Sí hay falla
cirugia (o *qxurgent*)	Cirugía urgente	0=No; 1=Sí
mortpredcod	Mortalidad predicha (expresada en porcentaje) dicotomizada	0=17 o menos 1=Más de 17
infnos	Infección nosocomial (en el hospital)	0=Sin infección o infección comunitaria(no hay infección nosocomial) 1=Sí hay infección nosocomial
vivomuer	Estado vital	1=Vivo; 2=Muerto
fcard	Falla cardiaca	0=No hay falla; 1=Sí hay falla
mortpred	Mortalidad predicha (expresada en porcentaje)	
cv2vivom	Variable que identifica vivos con buena calidad de vida, vivos con mala calidad de vida y muertos	1=Vivo con buena calidad de vida a los tres meses de alta hospitalaria 2=Vivo con mala calidad de vida a los tres meses de alta hospitalaria 3=Muerto

Tabla 4.1: Variables que se utilizan en los modelos del capítulo 4

4.1. Modelo para la calidad de vida posterior a la estancia en UTI

En este caso se obtuvo una red formada únicamente por seis variables discretas: la variable edad dicotomizada (*edadcod60* o *edad*), la calidad de vida inicial (*cv1cod*), la calidad de vida posterior a la estancia en la UTI (*cv2*), la falla neurológica (*fneur*), la falla respiratoria (*fresp*) y la variable correspondiente a cirugía urgente (*cirugia*), todas

las variables son binarias como se observa en la tabla 4.1. El aprendizaje de la red o estimación de probabilidades condicionales en la misma se llevó a cabo en el programa DEAL basándose en las restricciones, proporcionadas por el médico de acuerdo a su experiencia y también de acuerdo a otras restricciones que el médico no había percibido pero que iban surgiendo al ir refinando el modelo hasta llegar a un modelo plausible. Las instrucciones que se muestran en la sección 3.2 para llevar a cabo la obtención de una red son las que corresponden justamente a este modelo. Al final se obtiene la red que se muestra en la figura 4.1.

Basándose en la red se pueden hacer las siguientes observaciones: 1) A la variable calidad de vida posterior a la estancia en la UTI (*cv2*) ingresan de manera directa las variables: edad dicotomizada (*edad*), calidad de vida anterior a la estancia en la UTI (*cv1cod*), falla respiratoria (*fresp*) y cirugía urgente (*cirugia*), la única variable que no ingresa a *cv2* de manera directa es la falla neurológica (*fneur*); sin embargo, esta variable ingresa indirectamente a través de la falla respiratoria, ya que en la red se sugiere que la falla respiratoria depende de la falla neurológica (hay un arco entre ellas). 2) Al tomar en cuenta el concepto de conjuntos separadores como se definieron en la sección 2.3 se puede observar que falla neurológica es condicionalmente independiente de *cv2* dada la falla respiratoria, esto es porque al considerar la versión no dirigida de la figura 4.1 (es decir la gráfica en la que no se toman en cuenta la dirección de los arcos) resulta que la variable *fresp* separa a la variable *cv2* de la variable *fneur*. La independencia condicional anterior se denota como $fneur \perp cv2|fresp$. 3) La calidad de vida inicial depende de la edad, 4) El hecho de tener que hacerse una cirugía urgente depende a su vez de la calidad de vida inicial y 5) El tener o no falla respiratoria puede pensarse como una consecuencia de la cirugía urgente.

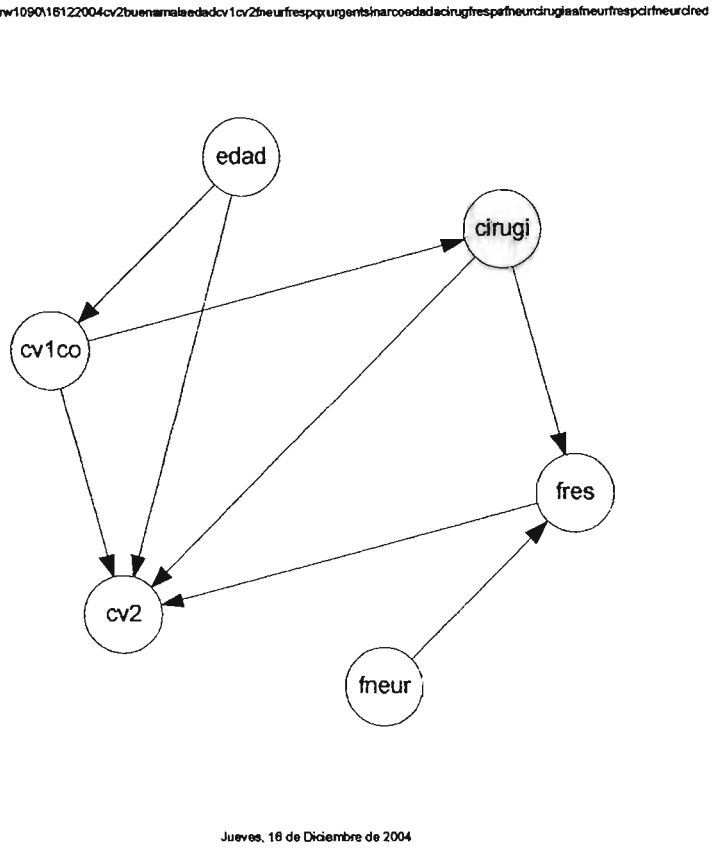

Figura 4.1: Modelo gráfico que involucra la variable calidad de vida posterior (*cv2*)

Con base en la gráfica de la figura 4.1 y al concepto de conjuntos separadores en **2.3** se puede observar la presencia de independencias condicionales entre algunas variables de manera similar a aquella que se obtuvo en el párrafo anterior. Las independencias condicionales obtenidas se presentan a continuación. Se hace notar que solo se listan las independencias condicionales por pares, esto es no se listan las independencias entre conjuntos de variables.

1. La falla neurológica es condicionalmente independiente de la calidad de vida pos-

terior dada la falla respiratoria, o sea $fneur \perp cv2|fresp$.

2. La edad codificada es condicionalmente independiente de la cirugía urgente dada la calidad de vida inicial y la calidad de vida posterior, $edad \perp cirugia|cv1cod, cv2$.

3. La falla respiratoria es condicionalmente independiente de la calidad de vida inicial dada la variable cirugía urgente y la calidad de vida posterior, $fresp \perp cv1cod|cirugia, cv2$; de hecho $fresp \perp cv1cod|cirugia, cv2, edad$.

4. La edad codificada es condicionalmente independiente de la falla neurológica dada la falla respiratoria, $edad \perp fneur|fresp$; también la edad es condicionalmente independiente de la falla neurológica dada la cirugía y la calidad de vida posterior, $edad \perp fneur|cirugia, cv2$, es mas $edad \perp fneur|cirugia, cv1cod, fresp, cv2$.

5. La edad codificada es condicionalmente independiente de la falla respiratoria dada la calidad de vida inicial y la calidad de vida posterior, $edad \perp fresp|cv1cod, cv2$; también las mismas variables son condicionalmente independientes dadas la cirugía urgente y la calidad de vida posterior, $edad \perp fresp|cirugia, cv2$, de hecho $edad \perp fresp|cirugia, cv2, cv1cod$.

6. La calidad de vida inicial es condicionalmente independiente de la falla neurológica dada la falla respiratoria, $cv1cod \perp fneur|fresp$, también las mismas variables son condicionalmente independientes dadas la cirugía urgente y la calidad de vida posterior, $cv1cod \perp fneur|cirugia, cv2$ y dadas la falla respiratoria y la calidad de vida posterior, $cv1cod \perp fneur|fresp, cv2$; de hecho $cv1cod \perp fneur|cirugia, fresp, cv2$.

7. La falla neurológica es condicionalmente independiente de la cirugía urgente dada la falla respiratoria, es decir $fneur \perp cirugia|fresp$.

Dada la naturaleza del problema o pregunta de investigación médica donde se tienen un conjunto de variables que pueden verse como explicativas y una variable binaria que puede verse como respuesta se puede llevar a cabo una regresión logística (Agresti, 2002, cap. 5) con la variable binaria calidad de vida posterior (*cv2*) como variable respuesta, cuyo éxito (o sea el valor uno en la variable binaria correspondiente) sería cuando la calidad de vida posterior es buena. El resto de las variables se consideran como variables explicativas, usando el paquete SPSS v. 10 se obtiene el ajuste de la regresión con los valores mostrados en la tabla 4.2, en la cual se observan los coeficientes correspondientes a cada variable explicativa, y resulta que de acuerdo al "p-value"todas las variables son estadísticamente significativas.

Al usar solo variables categóricas y observar que las exponenciales de cada uno de los coeficientes son valores mayores que uno (equivalentemente todos los coeficientes son positivos) significa que la probabilidad (o riesgo) de que la calidad de vida sea buena se incrementa cuando cada una de las variables se encuentra en la categoría codificada distinta de cero, que en este caso corresponde a cuando el individuo se encuentra en mejores condiciones; por ejemplo, en el caso de la variable *edad* la categoría diferente de cero corresponde a cuando el individuo tiene 60 o menos años, la exponencial del coeficente es de 1.8666, es decir que cuando el individuo tiene 60 o menos años aumenta el riesgo de que su calidad de vida posterior a su estancia en la UTI sea buena manteniendo fijo el valor de las otras variables explicativas, de manera similar al tener una calidad de vida inicial buena aumenta la probabilidad de que su calidad de vida posterior sea también buena, cuando el individuo no tiene falla neurológica (o respiratoria) incrementa la probabilidad de que la variable *cv2* sea buena, y finalmente al no tener cirugía urgente también aumenta la probabilidad de tener posteriormente buena calidad de vida. Así que como era de esperarse al tener el individuo mejores condiciones

tiene más probabilidad de una buena calidad de vida posterior.

Para ver que es aceptable la calidad del ajuste del modelo logístico mencionado, el cual solo tiene efectos principales y no incluye interacciones entre las variables explicativas, se puede analizar la devianza residual, que es una medida de la diferencia entre el modelo ajustado y el modelo saturado (el modelo que incluye efectos principales e interacciones de todos los ordenes), se busca que este valor sea "chico" pues esto indicaría que el modelo saturado no es mejor que el ajustado. Entonces, se obtiene una prueba de hipótesis, cuya hipótesis nula es que la diferencia entre el modelo saturado y el ajustado es pequeña, en la cual la estadística de prueba es la devianza residual, la cual se compara con el cuantil de una ji cuadrada con tantos grados como la diferencia de parámetros estimados entre el modelo ajustado y el saturado. En este caso la devianza residual es de 23.864 (subtabla 4 tabla 4.2), los grados de libertad son 21 y el nivel crítico o "p-value" de la prueba correspondiente es 0.30 indicando que en general, por ejemplo para un nivel de significancia de 0.05, no se rechaza la hipótesis nula correspondiente y entonces la diferencia entre el modelo ajustado y el saturado es pequeña y como consecuencia el modelo ajustado es aceptable y entonces para mejorar el modelo no habría necesidad de incluir interacciones de algún orden.

En la tabla 4.2 obtenida usando SPSS también se muestra la estadística de Pearson la cual también sirve para ver si el ajuste es adecuado, en este caso el nivel crítico es de 0.379 lo cual indica como en la prueba anterior, que el modelo ajustado es adecuado. También, para ver que tan bueno es el ajuste que se obtienen con el modelo logístico elegido, se ajusta a los datos un modelo saturado (las probabilidades estimadas a partir del modelo logístico saturado coinciden con las probabilidades observadas) y se comparan las tablas de clasificación de este modelo con el utilizado. En el modelo saturado

resultó que 73.4 % de los individuos con mala calidad de vida posterior son clasificados correctamente y 70.70 % de los individuos con buena calidad de vida posterior son clasificados adecuadamente, resultados similares a los que se tienen en la regresión con solo efectos principales (66.5 % de los individuos con mala c.v. posterior son clasificados correctamente y 70.70 % de los individuos con buena c.v. posterior son clasificados adecuadamente) entonces es preferible quedarse con el modelo elegido (con solo efectos principales), por el principio de parsimonia. Sin embargo, los resultados mencionados para el modelo saturado no son muy confiables porque en el software utilizado hubo problemas numéricos que afectan las estimaciones.

Classification Table[a]

			Predicted		
			BUENMALA		Percentage Correct
	Observed		mala	buena	
Step 1	BUENMALA	mala	125	63	66.5
		buena	51	147	74.2
	Overall Percentage				70.5

a. The cut value is .500

Variables in the Equation

								95.0% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	EDADCOD60(1)	.635	.244	6.774	1	.009	1.886	1.170	3.042
	CV1COD(1)	1.768	.312	32.007	1	.000	5.858	3.175	10.807
	FNEUR(1)	.959	.428	5.024	1	.025	2.609	1.128	6.036
	FRESP(1)	.742	.261	8.051	1	.005	2.100	1.258	3.505
	QXURGENT(1)	.746	.258	8.384	1	.004	2.108	1.273	3.493
	Constant	-3.384	.552	37.511	1	.000	.034		

a. Variable(s) entered on step 1: EDADCOD60, CV1COD, FNEUR, FRESP, QXURGENT.

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	161.194	165.150	159.194			
Final	83.516	107.251	71.516	87.678	5	.000

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	22.364	21	.379
Deviance	23.864	21	.300

Pseudo R-Square

Cox and Snell	.203
Nagelkerke	.271
McFadden	.164

Tabla 4.2: Coeficientes ajustados, tabla de clasificación y tablas para ver la calidad de ajuste de la regresión con variable respuesta binaria calidad de vida posterior (*cv2*)

Además, se obtiene una prueba de razones de verosimilitud (subtabla 3 tabla 4.2) entre el modelo ajustado y el el modelo "nulo" (aquel cuyos coeficientes para todos los parámetros son cero) cuya hipótesis nula es que el modelo ajustado no supera al modelo nulo, así que la estadística es menos dos veces la diferencia entre los logaritmos de las verosimilitudes correspondientes la cual se compara con el cuantil de una ji-cuadrada, en este caso el nivel crítico obtenido es cero indicando que el modelo ajustado supera al nulo. También se obtienen pseudo R^2's que comparan la devianza nula D_0 (la devianza residual correspondiente al modelo que incluye solo al término constante) con la devianza residual del modelo ajustado D_k, obteniendo la reducción relativa en la devianza nula debido al modelo $((D_0 - D_k)/D_0)$, lo cual significaría que porcentaje de la variabilidad es explicada por el modelo, en subtabla 5 de la tabla 4.2 aparecen estos

valores calculados de distintas formas y se observa que están entre 0.164 y 0.271 entonces de manera aproximada entre el 16.4 % y el 27.1 % de la variabilidad es explicada por el modelo.

Finalmente, otro punto de interés al ajustar una regresión logística es el hecho de determinar cuáles son las variables que influyen más sobre la variable respuesta, para ello se puede utilizar el coeficiente de Wald asociado a cada variable explicativa (ver tabla 4.2), el cual se calcula al dividir el coeficiente estimado entre la desviación estándar estimada del coeficiente estimado y elevando toda esta expresión al cuadrado. De esta forma, resulta que la variable *cv1cod* sería la variable más influyente sobre la respuesta, pues el coeficente asociado a la misma es mayor que la los otros coeficentes (32.007). Sin embargo; el coeficiente anterior no es realmente un coeficiente estandarizado, un coeficiente estandarizado propuesto (Agresti, 2002, p. 191) es $\hat{\beta}_{jest} = \hat{\beta}_j s_{X_j}$, donde $\hat{\beta}_j$ es el coeficiente ajustado para la variable explicativa X_j y s_{X_j} es la desviación estándar en la base de datos para la variable X_j, cuando la variable es binaria $s_{X_j} = \sqrt{p_j(1-p_j)}$, donde p_j es la proporción de individuos en la base de datos que para la variable explicativa X_j tienen la característica de interés, así por ejemplo para la variable *edadcod60*, la propoprción de individuos en el rango inferior de edad para la base de datos que incluye únicamente a los sobrevivientes (que es la base de datos utilizada en esta regresión logística) es $p_{edadcod60}$=0.63 y entonces $\hat{\beta}_{edadcod60est}$=0.635*$\sqrt{0.63(0.37)}$=0.306. De manera similar se calculan todos los demás coeficientes estandarizados y nuevamente resulta que la variable de más relevancia es la calidad de vida inicial, con un coeficiente relativo de 0.741 que es mucho mayor que los coeficientes relativos de las otras variables (ver el Apéndice, tabla B.1).

Para ver la relación que existe entre la regresión anterior con un modelo gráfico que

represente algo similar, se ajusta una gráfica en la cual se elimina la posibilidad de cualquier arco existente entre las variables, exceptuando aquellos arcos que van de cada una de las variables consideradas como explicativas a la variable respuesta *cv2*, de tal manera que al llevar a cabo el Aprendizaje Estructural en DEAL, se obtiene una gráfica en la que hay arcos que apuntan o no apuntan de cada una de las variables "explicativas": calidad de vida inicial (*cv1cod*), la edad categorizada (*edad*), la falla neurológica (*fneur*), la falla respiratoria (*fresp*) y la cirugía urgente (*cirugia*) hacia la variable respuesta calidad de vida posterior a la estancia, o sea a *cv2*. Esta gráfica se muestra en la figura 4.2, se observa en ella que hay arcos desde todas las variables, excepto de la variable falla neurológica, hacia la variable *cv2*; así que el mejor modelo gráfico con las características mencionadas no indica que la falla neurológica sea causante directa de una mala o buena calidad de vida posterior a la estancia en la UTI.

Al observar los "p-values" o valores críticos obtenidos en la regresión para cada una de las variables se observa que para la variable falla neurológica se tiene el "p-value" más grande de 0.025, lo que indicaría que *fneur* es un variable que explica la variable respuesta calidad de vida posterior (*cv2*); sin embargo, como ya se dijo en el caso del modelo gráfico correspondiente, falla neurológica no aparece relacionado con *cv2*.

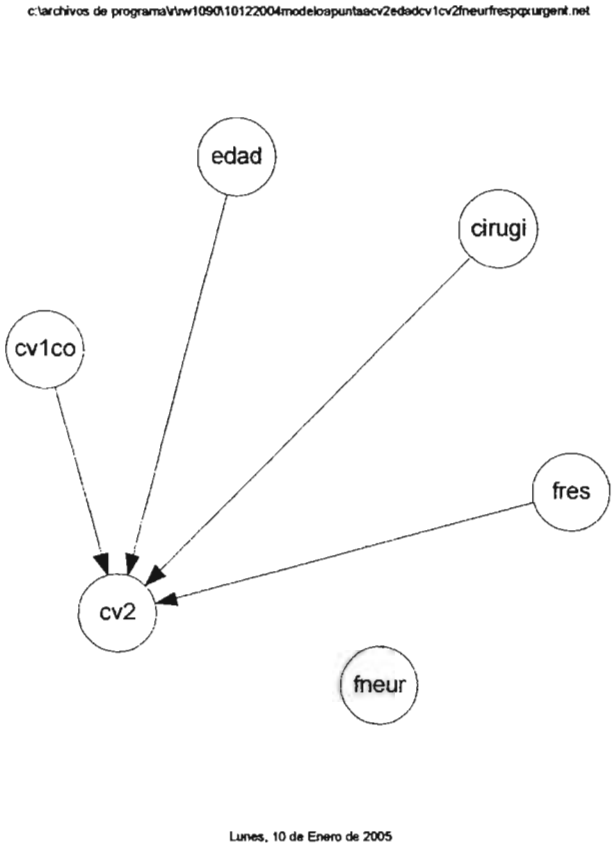

Figura 4.2: Modelo gráfico que solo permite arcos que inciden a calidad de vida posterior (*cv2*)

El siguiente paso corresponde a obtener probabilidades ajustadas, tanto para el modelo gráfico como para el logístico y hacer una comparación entre ellas. En el caso de la regresión logística para obtener las probabilidades ajustadas bajo el modelo simplemente hay que utilizar los coeficientes ajustados y sustituirlos según los valores de las variables, esto se logra al recordar que el modelo empleado en una regresión logística es:

$$\log(\frac{p}{1-p}) = \alpha + \beta'\mathbf{x}, \tag{4.1}$$

donde $\mathbf{x}$ es un vector formado por todas las variables explicativas, α es el término constante, β es el vector formado por los coeficientes de la regresión y son los valores que se estiman, finalmente se tiene que $p = P(Y = 1; \mathbf{x})$ significa la probabilidad de que la variable respuesta tome el valor uno cuando las variables explicativas toman el valor $\mathbf{x}$ o bien puede pensarse como la probabilidad de éxito. Entonces, según los valores de la tabla 4.2 la regresión logística ajustada en este caso está dada por:

$\log\left(\frac{\hat{P}(cv2=buena)}{\hat{P}(cv2=mala)}\right)$ =-3.38+ 0.63 $(edadcod60 \leq 60)$+ 1.76 $(cv1cod = buena)$+ 0.95 $(fneur = no)$+ 0.74 $(fresp = no)$+ 0.74 $(cirugia = no)$.

Además a partir de la ecuación (4.1), una vez que se han estimado los coeficientes se tiene:

$$\hat{p} = \frac{\exp(\hat{\alpha} + \hat{\beta}'\mathbf{x})}{1 + \exp(\hat{\alpha} + \hat{\beta}'\mathbf{x})}.$$

Entonces, se pueden obtener las probabilidades estimadas para valores específicos de las variables explicativas. Por ejemplo si se tiene un individuo con menos de 60 años, con buena calidad de vida inicial, sin falla neurológica ni respiratoria y sin que se le

haya practicado una cirugía urgente, la probabilidad estimada bajo el modelo logístico de que su calidad de vida posterior a la estancia en la UTI sea buena, $\hat{P}(cv2 = buena)$, está dada por:

$$\frac{\exp(-3.384 + 0.635 + 1.768 + 0.959 + 0.742 + 0.746)}{1 + \exp(-3.384 + 0.635 + 1.768 + 0.959 + 0.742 + 0.746)} = 0.8124$$

De forma similar se pueden obtener las probabilidades estimadas bajo el modelo logístico para todas las combinaciones posibles o patrones que pueden tomar las variables explicativas binarias.

En un siguiente paso se estiman las mismas probabilidades de que un individuo tenga una buena calidad de vida posterior a la UTI cuando las variables explicativas toman distintos valores, pero ahora usando el modelo gráfico correspondiente a la figura 4.1 y para ello se utiliza el programa *Hugin 5.4* ya mencionado. Lo que se hace es tomar la red original y *compilarla* (sección 3.3), obteniendo las probabilidades marginales de cada variable obtenidas a partir de las probabilidades locales aprendidas usando DEAL, en la figura 4.3 se observa como se presentan los resultados obtenidos en *Hugin*. Conviene recordar que cada variable es binaria y que el valor que toman las categorías en cada variable se encuentran en la tabla 4.1 al inicio de este capítulo, así por ejemplo, en el caso de la variable edad la categoría etiquetada como 1 representa a un individuo que tiene 60 o menos años de edad y la categoría etiquetada 2 representa cuando tiene 61 o más años.

Una vez que se ha *compilado* el modelo gráfico, el siguiente paso corresponde a incorporar la evidencia. Para ejemplificar este paso se utiliza al mismo tipo de individuo que se uso para mostrar cómo se obtienen las probabilidades en la regresión logística,

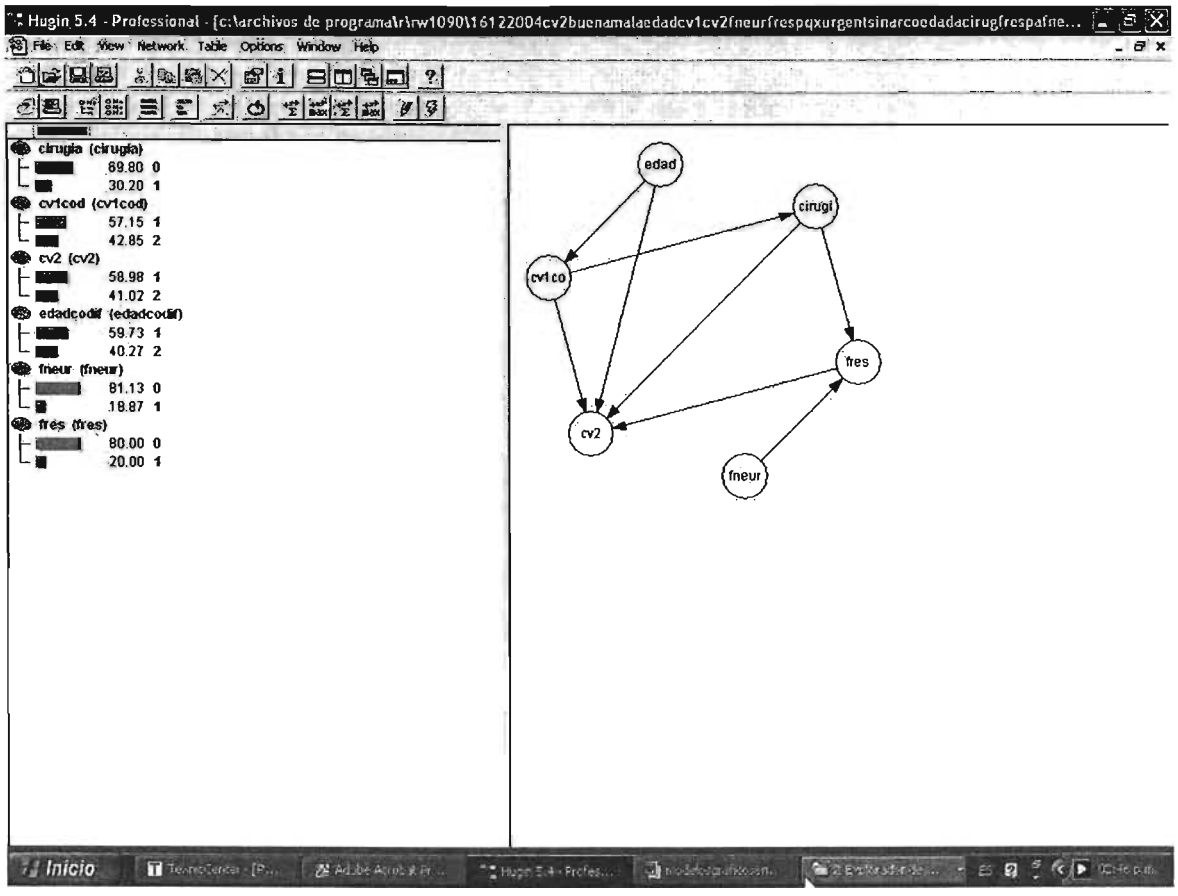

Figura 4.3: Probabilidades marginales una vez compilado el modelo gráfico en *Hugin*

es decir un individuo con menos de 60 años, con buena calidad de vida inicial, sin falla neurológica ni respiratoria y sin que se le haya practicado una cirugía urgente, a cada categoría en la que el individuo se encuentra para cada una de las variables se le da probabilidad uno, así por ejemplo en la variable edad a la categoría que corresponde a individuos menores de 60 años se le da probabilidad uno y de forma similar para las otras variables. Una vez que se han introducido estos valores se propaga la evidencia, es decir se actualizan la probabilidades para cada categoría en las variables restantes, en este caso solo en la variable *cv2* y entonces se puede obtener la probabilidad bajo el modelo gráfico de que el individuo con las características mencionadas tenga una buena

calidad de vida posterior a su estancia en la UTI y que en este caso corresponde a un valor de 0.7467 (figura 4.4).

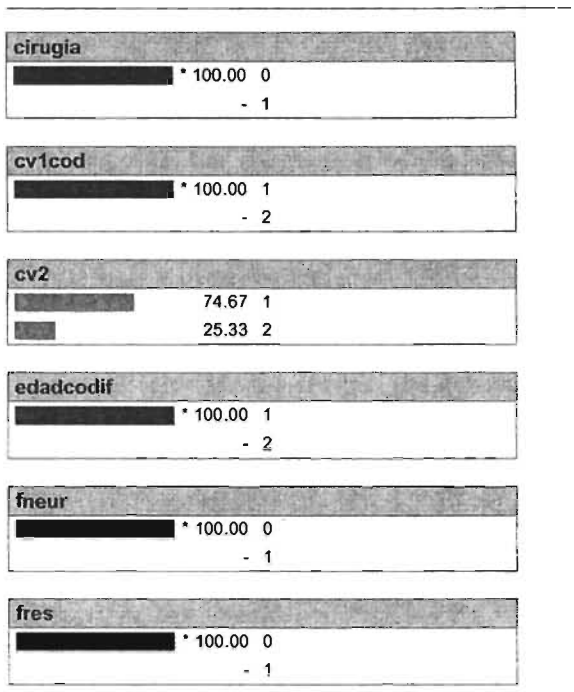

Figura 4.4: Probabilidades marginales para *cv2* una vez introducida la evidencia

El siguiente paso es estimar las probabilidades de tener una buena calidad de vida posterior a la estancia en la UTI para toda posible combinación de los valores o categorías que pueden tomar las demás variables. Como se tienen cinco variables explicativas y todas son binarias, entonces hay un total de 32 posibles combinaciones de valores que pueden tomar los individuos en sus variables explicativas. Entonces, tal como se ha descrito arriba, se estiman las probabilidades para cada una de estas combi-

naciones, tanto con el modelo de regresión logística como con el modelo gráfico. Estas estimaciones se presentan en la tabla 4.3 y posteriormente se grafican, primero con una gráfica en la que se comparan las probabilidades entre ambos modelos para cada una de los 32 posibles valores que pueden tomar los individuos (figura 4.5) y luego en una gráfica en la que en un eje se tiene la probabilidad estimada bajo la regresión y en el otro eje la probabilidad estimada bajo el modelo gráfico propuesto (figura 4.6).

En la tabla 4.3 se observa que bajo el modelo gráfico los valores de las probabilidades se encuentran entre 0.333 y 0.795, en cambio, bajo el modelo logístico los valores van desde 0.032 hasta 0.812, así que en ninguna combinación de categorías del modelo gráfico hay valores tan pequeños como los que se alcanzan en el logístico, de tal forma que los valores que toman las probabilidades se encuentran en un intervalo de menor tamaño.

En la misma tabla y en la figura 4.5, se observa como hasta la celda 16 o renglón 16 de la tabla (a partir de la cual la edad cambia de la categoría uno a dos) las tendencias entre ambas estimaciones de las probabilidades son parecidas, es decir cuando un valor sube para el modelo gráfico también sube para el logístico y lo mismo cuando el valor baja. Después de esta observación ya no se observan estas tendencias similares.

Se observa que las diferencias en valor absoluto de las probabilidades entre ambos modelos van desde 0.008 hasta 0.496 (celda 12) y el promedio de estas diferencias en las 32 combinaciones de valores es de 0.249 unidades. Observando la celda 16, que corresponde a una persona joven pero con todo mal, bajo el modelo logístico la probabilidad de buena calidad de vida posterior es de 0.060 y en la siguiente celda (17), correspondiente a una persona mayor pero con todo lo demás indicando que la persona estaba

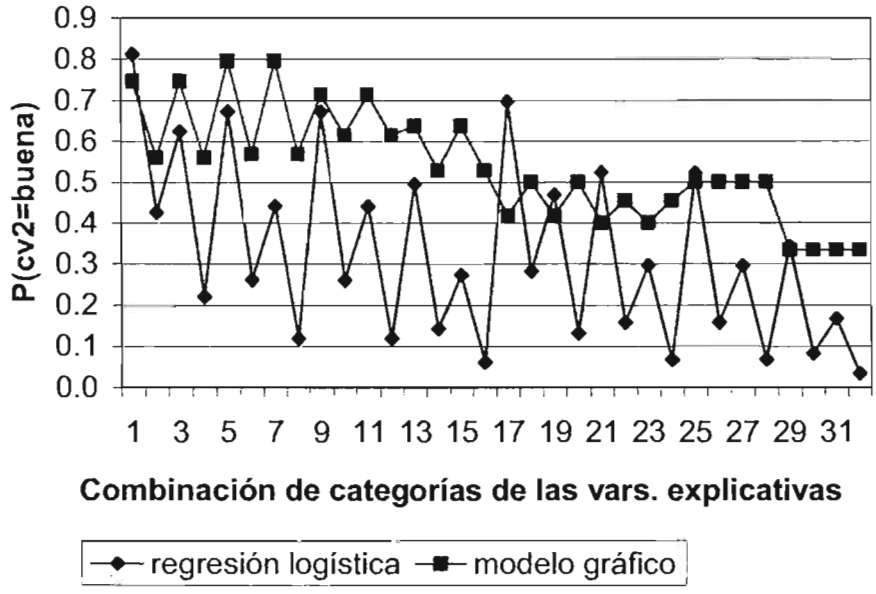

Figura 4.5: Probabilidades estimadas, $\hat{P}(cv2 = buena)$, bajo una regresión logística y bajo un modelo gráfico (orden 1)

bien, la probabilidad sube hasta 0.696; en cambio, en el modelo gráfico al pasar de la celda 16 a la 17 la probabilidad de tener una buena calidad de vida posterior disminuye un poco (de 0.528 a 0.416) y de hecho las probabilidades de las siguientes celdas (que corresponden a personas mayores de 61 años) ya no suben por encima del valor que toma la celda 16, así que en el modelo gráfico el hecho de pasar al grupo de edad con personas mayores implica una disminución en la probabilidad de tener una buena calidad de vida posterior.

Como ya se observó cuando el individuo tiene todo favorable bajo el modelo logístico se tiene una probabilidad de tener buena calidad de vida posterior de 0.812, mientras

celda	edad cod.	cirugía urg.	falla resp.	falla neur.	cv1 cod.	$\hat{P}_{reg.}$	$\hat{P}_{graf.}$	frec. rel. buena c.v.	obs.	$\lvert\hat{P}_{reg.} - \hat{P}_{graf.}\rvert$	$\frac{\lvert\hat{P}_{reg.} - \hat{P}_{graf.}\rvert}{\hat{P}_{graf.}}$
1	1	0	0	0	1	0.8124	0.7467	0.8333	60	0.0657	0.0881
2	1	0	0	0	2	0.4251	0.5600	0.1429	7	0.1349	0.2409
3	1	0	0	1	1	0.6241	0.7467	1.0000	2	0.1226	0.1642
4	1	0	0	1	2	0.2208	0.5600	0.0000	0	0.3392	0.6057
5	1	0	1	0	1	0.6735	0.7955	0.7143	84	0.1220	0.1534
6	1	0	1	0	2	0.2604	0.5690	0.1176	17	0.3086	0.5424
7	1	0	1	1	1	0.4415	0.7955	0.5455	11	0.3540	0.4450
8	1	0	1	1	2	0.1189	0.5690	0.0000	1	0.4501	0.7911
9	1	1	0	0	1	0.6726	0.7143	0.6000	10	0.0417	0.0584
10	1	1	0	0	2	0.2596	0.6154	0.3333	3	0.3558	0.5781
11	1	1	0	1	1	0.4405	0.7143	0.0000	0	0.2738	0.3833
12	1	1	0	1	2	0.1185	0.6154	0.0000	0	0.4969	0.8075
13	1	1	1	0	1	0.4945	0.6389	0.4595	37	0.1444	0.2260
14	1	1	1	0	2	0.1431	0.5283	0.0000	4	0.3852	0.7292
15	1	1	1	1	1	0.2727	0.6389	0.3333	6	0.3662	0.5732
16	1	1	1	1	2	0.0601	0.5283	0.0000	1	0.4682	0.8862
17	2	0	0	0	1	0.6966	0.4167	0.7143	21	0.2799	0.6716
18	2	0	0	0	2	0.2815	0.5000	0.3333	9	0.2185	0.4370
19	2	0	0	1	1	0.4680	0.4167	0.0000	1	0.0513	0.1232
20	2	0	0	1	2	0.1306	0.5000	0.0000	0	0.3694	0.7389
21	2	0	1	0	1	0.5222	0.4000	0.4375	32	0.1222	0.3056
22	2	0	1	0	2	0.1572	0.4545	0.2083	24	0.2973	0.6541
23	2	0	1	1	1	0.2953	0.4000	0.0000	3	0.1047	0.2619
24	2	0	1	1	2	0.0667	0.4545	0.0000	1	0.3878	0.8532
25	2	1	0	0	1	0.5212	0.5000	0.0000	1	0.0212	0.0425
26	2	1	0	0	2	0.1567	0.5000	0.5000	2	0.3433	0.6866
27	2	1	0	1	1	0.2944	0.5000	0.0000	0	0.2056	0.4112
28	2	1	0	1	2	0.0665	0.5000	0.0000	1	0.4335	0.8670
29	2	1	1	0	1	0.3414	0.3333	0.3333	27	0.0081	0.0243
30	2	1	1	0	2	0.0813	0.3333	0.2667	15	0.2520	0.7561
31	2	1	1	1	1	0.1658	0.3333	0.0000	3	0.1675	0.5027
32	2	1	1	1	2	0.0328	0.3333	0.0000	3	0.3005	0.9016

Categorías o celdas ordenadas por *edadcod*, *cirugía*, *fresp*, *fneur* y *cv1cod*.

Tabla 4.3: Tabla de probabilidades estimadas, $\hat{P}(cv2 = buena)$, bajo regresión logística y bajo el modelo gráfico (orden 1)

tanto en el modelo gráfico la probabilidad es de 0.746, observando que los puntos se acercan entre sí; sin embargo, cuando todo esta mal en el individuo (es grande, con cirugía, con fallas y mala calidad de vida inicial) en el modelo gráfico la probabilidad es de 0.33 mientras que en el logístico es de 0.03, o sea hay una diferencia de 0.3 unidades (diferencia por arriba de la diferencia promedio mencionada), de hecho, como ya se dijo, en el modelo gráfico los valores no bajan de 0.33.

También se observa que en las probabilidades estimadas con el modelo gráfico hay valores que se repiten. Estas repeticiones reflejan la presencia de independencias condicionales entre variables. Por ejemplo si se observa la celda 1 y la 3 de la tabla 4.3 se tiene que sus probabilidades estimadas con el modelo gráfico son las mismas (0.746). La celda 1 corresponde a individuos con las mismas características que los correspondientes a la celda 3 (son personas con 60 o menos años, sin cirugía urgente, sin falla respiratoria y con buena calidad de vida inicial) a excepción de que los de la celda 1 no presentan falla neurológica y los de la 3 sí, pero como bajo el modelo considerado la calidad de vida posterior es condicionalmente independiente de la falla neurológica dada la falla respiratoria, entonces como ambas celdas tienen los mismos valores en la falla respiratoria y en todas las demás variables salvo en falla neurológica y debido a la independencia condicional entre la calidad de vida posterior y la falla neurológica mencionada, aunque se cambien los valores que toma la variable falla neurológica las probabilidades de la calidad de vida posterior son las mismas en ambas celdas. De forma similar se explica que la celda 2 y 4 tengan las mimas probabilidades estimadas, la celda 5 y la 7, la 6 y la 8 y así sucesivamente para el resto de la tabla.

Si las probabilidades estimadas por ambos modelos fueran los mismos o muy parecidos, en la figura 4.6 los puntos estarían muy cercanos a una recta imaginaria de 45°;

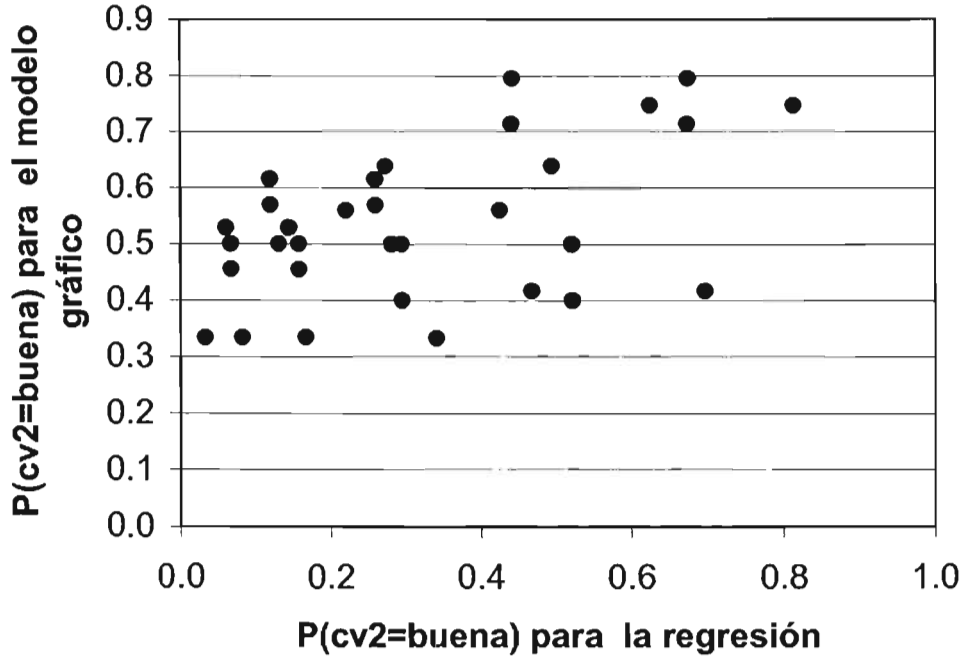

Figura 4.6: Probabilidades estimadas $\hat{P}_{reg.}(cv2 = buena)$ vs $\hat{P}_{graf.}(cv2 = buena)$

sin embargo, en este caso la mayoría de los puntos están cargados hacia arriba a la izquierda, indicando que en el modelo gráfico se estiman probabilidades mayores respecto a las que se estiman utilizando la regresión, lo cual también se observaba en la figura 4.5. Los puntos más extremos son los siguientes: un punto en la parte inferior izquierda el cual corresponde a una persona mayor con todo mal, como ya se mencionó su probabilidad estimada bajo el modelo logístico es de 0.032, mientras que en el modelo gráfico estima un valor de 0.333, un segundo punto se encuentra en la parte izquieda superior y corresponde a una persona en el rango de edad inferior, pero con todas las demás variables indicando que estaba mal y cuyas probabilidades son de 0.528 en el modelo gráfico y de 0.060 en el logístico. Un tercer punto extremo es el que se encuentra en la parte inferior derecha y que corresponde a una persona mayor pero bien

en lo demás, cuyas probabilidades estimadas son de 0.416 para el modelo gráfico y de 0.696 para el logístico, también se tiene un cuarto punto en la parte superior derecha y que corresponde a una persona en el rango de edad inferior con todo lo demás bien y cuyos valores entre ambos modelos se parecen (0.746 en el modelo gráfico y 0.812 en el logístico). Finalmente, otro punto extremo también se encuentra en la parte superior izquierda (arriba del segundo punto) y corresponde a aquel en el que la diferencia de las probabilidades entre ambos modelos era la mayor; se trata de una persona en el rango de edad inferior, con cirugía urgente, sin falla respiratoria pero con falla neurológica y mala calidad de vida inicial; en este caso la regresión logística estima una probabilidad de 0.118 y el modelo gráfico estima el valor 0.615.

En la antepenúltima columna de la tabla 4.3 se observa que la combinación de categorías con la mayor frecuencia de observaciones, 84, corresponde al renglón o celda 5 que incluye a personas con 60 o menos años, sin cirugía urgente, con falla respiratoria, sin falla neurológica y con buena calidad de vida inicial en este caso la diferencia de probabilidades estimadas entre ambos modelos es de 0.122 (el modelo logístico estima una valor de 0.673 y el gráfico de 0.795) y en las combinaciones o celdas donde no hay ninguna observación son: la celda 4 (o renglón 4 de la tabla 4.3) correspondiente a personas en el rango de edad inferior, sin cirugía urgente, sin falla respiratoria, con falla neurológica y con mala calidad de vida inicial y cuya diferencia entre las estimaciones es de 0.339; la celda 11 correspondiente a personas en el rango de edad menor, con cirugía urgente, sin falla respiratoria, con falla neurológica y con buena calidad de vida inicial con una diferencia entre las estimaciones de 0.2738; la celda 12 que tiene las mismas características de la observación anterior, salvo que en este caso la calidad de vida inicial era mala, para este caso la diferencia entre estimaciones es de 0.496, que como ya se dijo anteriormente es la mayor diferencia que se dio entre las estimaciones; también

se tiene a la celda 20 correspondiente a individuos ya mayores, sin cirugía urgente, sin falla respiratoria, con falla neurológica y con mala calidad de vida inicial y finalmente se tiene a la celda 27 que corresponde a personas mayores, con cirugía urgente, sin falla respiratoria, con falla neurológica y buena calidad de vida inicial, en este caso la diferencia entre las probabilidades estimadas es de 0.205.

Se puede realizar para cada patrón de covariables una comparación entre las probabilidades estimadas de tener buena calidad de vida posterior con ambos modelos y la frecuencia relativa observada de individuos con buena calidad de vida posterior (novena columna de la tabla 4.3), tal frecuencia mencionada se refiere a la proporción de individuos que en la base de datos tuvieron buena calidad de vida para un patrón de covariables específico, por ejemplo para la combinación de valores correspondientes a la celda 1 de la tabla 4.3 un 83.33% de individuos tuvieron buena calidad de vida y obviamente el restante 16.67% tuvieron mala calidad de vida. A continuación se realiza el análisis comparativo para aquellas celdas que tuvieron 15 o más observaciones.

1) Para la celda 1, que consta de 60 individuos, se observa que tanto la probabilidad de buena calidad de vida posterior estimada con el modelo gráfico como la estimada con el logístico se acercan a la frecuencia relativa observada, por lo que ambos modelos al tener una probabilidad estimada de 0.5 o más, asignan a un individuo tipo la celda 1 a la categoría buena calidad de vida posterior y efectivamente en la base de datos la mayoría de los individuos con los valores correspondientes a la celda 1 tuvieron buena calidad de vida posterior. 2) Para la celda 5 en la que se encuentran la mayor cantidad de observaciones, 84, nuevamente las estimaciones y la frecuencia relativa son similares (alrededor de 0.7), por lo que ambos modelos asignan a los individuos con la combinación de valores correspondientes a esta celda a la categoría buena calidad de

vida posterior, en la base de datos ocurre que la mayoría de estos individuos tuvieron buena calidad de vida posterior (71.43%). 3) Para la celda 6, con 17 individuos, el modelo gráfico estima un valor de 0.569, en cambio el logístico de 0.260, la frecuencia relativa correspondiente es de 0.117, entonces aunque ambos modelos estiman valores por encima de la frecuencia relativa la diferencia es mayor para el modelo gráfico, el cual asignaría a los individuos a buena calidad de vida posterior aunque en la base la mayoría (88.24%) tienen mala calidad de vida posterior, para este caso la edad se encuentra el rango de edad inferior y la calidad de vida inicial es mala, como la edad es la variable de más peso para el modelo gráfico (esto se vio al obtener las probabilidades estimadas para todo patrón de covariables y comparar con cuáles valores de las variables explicativas aumenta o disminuye mucho la probabilidad) entonces al encontrarse el individuo en la categoría de edad inferior no disminuye mucho la probabilidad de buena calidad de vida posterior mientras que para el logístico la calidad de vida inicial es la variable de más peso (esto se vio con los coeficientes estandarizados y los de Wald) y al tener mala calidad inicial disminuye la probabilidad estimada. 4) Para la celda 13, con 37 observaciones, las estimaciones y la frecuencia relativa están alrededor de 0.5; en el modelo gráfico la estimación es mayor a esta frecuencia y también a 0.5 (0.638) por lo que a los individuos se les asignaría a tener buena calidad de vida posterior, aunque en realidad un poco menos de la mitad (45.95%) de los individuos con las carcterísticas de la celda 13 tuvieron buena calidad de vida posterior, en cambio la regresión logística estima un valor apenas menor a 0.5 (0.494) por lo que asigna a los individuos a tener mala calidad de vida posterior. Nuevamente ocurre que el individuo se encuentra en el rango inferior de edad y esto influye en que en el modelo gráfico la probabilidad de buena calidad de vida posterior sea mayor. 5) Para la celda 17, con 21 individuos, ocurre que la probabilidad estimada bajo el modelo logístico se aproxima a la frecuencia relativa correspondiente (ambos valores alrededor de 0.7), en cambio el

modelo gráfico estima una probabilidad mas pequeña (0.416), esto es porque en este caso los individuos se encuentran en el rango de edad superior y esto ocasiona que para este modelo disminuya la probabilidad de buena calidad de vida posterior, entonces el modelo gráfico asignaría a los individuos a la categoría de mala calidad de vida posterior; sin embargo, la mayoría de los individuos en la base (71.43 %) tienen buena calidad de vida posterior. 6) Para la celda 21, con 32 observaciones, la probabilidad estimada bajo el modelo gráfico y la frecuencia relativa correspondiente se aproximan, con valores alrededor de 0.4; sin embargo, el modelo logístico estima una probabilidad mayor a 0.5, 0.522, por lo que el modelo logístico asignaría a los individuos con las características de esta celda a tener buena calidad de vida posterior aunque en realidad tan solo cerca del 40 % de individuos tienen buena calidad de vida posterior, el hecho de que la estimación bajo el modelo logístico sea mayor se debe a que los individuos de esta celda tiene buena calidad de vida inicial y bajo regresión en estos casos se estiman mayores probabilidades. 7) Para la celda 22, con 24 observaciones, las estimaciones con ambos modelos son menores a 0.5, por lo que asignarían a los individuos a la categoría de mala calidad de vida posterior y en la base de datos efectivamente la mayoría, cerca del 80 %, de los individuos tienen mala calidad de vida posterior. 8) Para la celda 29, con 27 observaciones, tanto las estimaciones como la frecuencia se aproximan bastante entre sí, con un valor alrededor de 0.3, de hecho la probabilidad estimada bajo el modelo gráfico es exactamente la frecuencia correspondiente, así que con ambos modelos se asignan a los individuos a la categoría calidad de vida posterior mala y efectivamente dos terceras partes de los individuos con las características de esta celda tienen mala calidad de vida posterior. 9) Para la celda 30, con 15 individuos, nuevamente las estimaciones bajo ambos modelos son menores a 0.5 al igual que la frecuencia, se observa que otra vez la estimación bajo el modelo gráfico se acerca más al valor de la frecuencia y que bajo ambos modelos se asignan a los individuo a la categoría mala calidad de vida

y efectivamente la mayoría, 73.33 % de los individuos en la base tienen mala calidad de vida posterior.

Para seguir comparando ambos modelos con la frecuencia relativa, se puede contar en cuántas celdas bajo cada modelo se asigna al individuo a la categoría con la frecuencia relativa observada adecuada, es decir se cuentan las celdas en las que por ejemplo si bajo la regresión se estiman valores de 0.5 o más la frecuencia relativa también es 0.5 o más, en estas celdas ocurriría que bajo regresión se asignan los individuos a tener buena calidad de vida posterior y en la base de datos la mayoría de estos individuos tienen buena calidad de vida posterior, de forma similar también se cuentan las celdas en las que si la estimación es inferior a 0.5 la frecuencia también es inferior a 0.5. En el caso de la regresión logística 28 de las 32 celdas se asignan adecuadamente, en cambio en el modelo gráfico 16 de las 32 celdas se asignan adecuadamente, por supuesto que hay celdas que no presentan observaciones y que no sería tan importante considerarlas y hay otras que cuentan con mayor cantidad de observaciones y que fueron las que se analizaron arriba.

Otra forma para ver qué tan bien se ajustan los modelos a los datos es mediante pruebas de bondad de ajuste (se utilizó la prueba ji- cuadrada). Como en el análisis comparativo de arriba, las pruebas de bondad de ajuste se llevaron a cabo solo para aquellas celdas con 15 o más observaciones. En primer lugar se obtuvieron el número de observaciones con buena c.v. posterior bajo regresión, para ello se multiplicó el número de observaciones de cada celda por la probabilidad estimada de buena c.v. posterior para esa misma celda bajo regresión, este valor sería el número de observaciones con buena c.v. posterior bajo regresión, la probabilidad p_i de que una observación se encuentre en cada una de las celdas bajo regresión resulta de dividir el número de observaciones

con buena c.v. posterior bajo regresión en cada una de las celdas entre la suma sobre todas las celdas del número de observaciones con buena c.v. posterior bajo regresión. Finalmente el número esperado de observaciones con buena c.v. posterior bajo regresión para cada celda, E_i, resulta de multiplicar el número total de observaciones con buena c.v. posterior que se obtiene de la base de datos (176 en este caso) por la p_i correspondiente. El número de observaciones con buena c.v. posterior para la celda i en la base de datos se denota como O_i y entonces se obtiene el estadístico de prueba $T = \sum_{j=1}^{9}(O_j - E_j)^2/E_j$=8.702 (Apéndice, tabla B.2), tal estadístico se compara con $\chi^{c-1}_{1-\alpha}$ el cuantil de una distribución ji cuadrada con $c-1$ grados de libertad que acumula una probabilidad de $1-\alpha$, con α un nivel de significancia y donde c es el número de celdas que se utilizan para la prueba (en este caso 9). Si el valor de T es mayor que el cuantil mencionado se rechaza la hipótesis nula H_0 de que la distribución desconocida de los datos es la misma distribución que uno proporciona, en este caso la distribución que se genera bajo la regresión logística. Tomando un α=0.05 se tiene que $\chi^8_{0.95}$=15.507 y entonces no se rechaza la hipótesis mencionada, lo cual quiere decir que la distribución ajustada bajo regresión es similar a la distribución real de los datos. De manera similar se puede calcula T bajo el modelo gráfico (Apéndice, tabla B.3), obteniendo T=17.386 que al compararlo con el cuantil correspondiente para el mismo valor de α utilizado arriba indicaría rechazar la hipótesis nula de que la distribución bajo el modelo gráfico coincide con la distribución real; sin embargo, para α=0.025, se tiene que $\chi^8_{0.975}$=17.534, lo cual indicaría no rechazar la hipótesis nula. En conclusión, la distribución bajo ambos modelos es similar a la distribución real, bajo el modelo logístico esto es cierto con el α usual y en el otro caso no es así.

En la figura 4.5 y tabla 4.3 las categorías están ordenadas primero de acuerdo a la variable edad, luego respecto a cirugía urgente, falla respiratoria, falla neurológica y

calidad de vida inicial, de tal manera que cada variable va de mejor a peor, por ejemplo para la variable edad primero van los del rango de edad inferior (categoría 1) y luego los mayores (categoría 2), luego dentro de cada grupo de edad se tiene que en la variable cirugía primero va el valor correspondiente a cuando el individuo no necesito cirugía (0) y luego cuando sí (1) y así sucesivamente para las otras variables. Ahora, lo que se pretende es cambiar este orden colocando en primer lugar la variable correspondiente a la calidad de vida inicial debido a que en la regresión logística es la variable de más peso (el valor estimado de su coeficiente es 1.768, el relativo de 32.007 y el estandarizado de 0.741, valores mucho mayores que los de las otras variables) y luego se tendrían las variables edad, cirugía, falla respiratoria y neurológica, obteniendo la tabla 4.4 y la figura 4.7, lo que se observa es que las estimaciones de las probabilidades para la regresión logística tienen una tendencia a la baja, en cambio en el modelo gráfico no hay tal tendencia, de hecho la probabilidad sube de la celda 16, cuando el individuo tiene buena calidad de vida inicial pero todo lo demás mal, a la 17, cuando el individuo tiene mala calidad de vida inicial pero todo lo demás bien, de un valor de 0.333 a 0.560 y a partir de esta observación todas las probabilidades se estabilizan (entre 0.333 y 0.615), así que en el modelo gráfico la calidad de vida inicial no está influyendo tanto de tal forma que al tener mala calidad de vida inicial no se observan necesariamente las probabilidades más bajas de toda la serie; en cambio en la regresión logistica esto sí ocurre. De hecho, para el modelo gráfico, las probabilidades más bajas son la celda 15 (individuo con buena calidad de vida inicial, pero todo mal a excepción de que no tiene falla neurológica), 16 (individuo con buena calidad de vida inicial, pero todo lo demás mal), 31 (individuo con mala calidad de vida inicial y todo mal excepto que no tiene falla neurológica) y 32 (individuo con todo mal). También se observa que las diferencias más grandes entre las probabilidades estimadas con ambos modelos se dan sobre todo cuando la calidad de vida inicial fue mala ($cv1cod$=2).

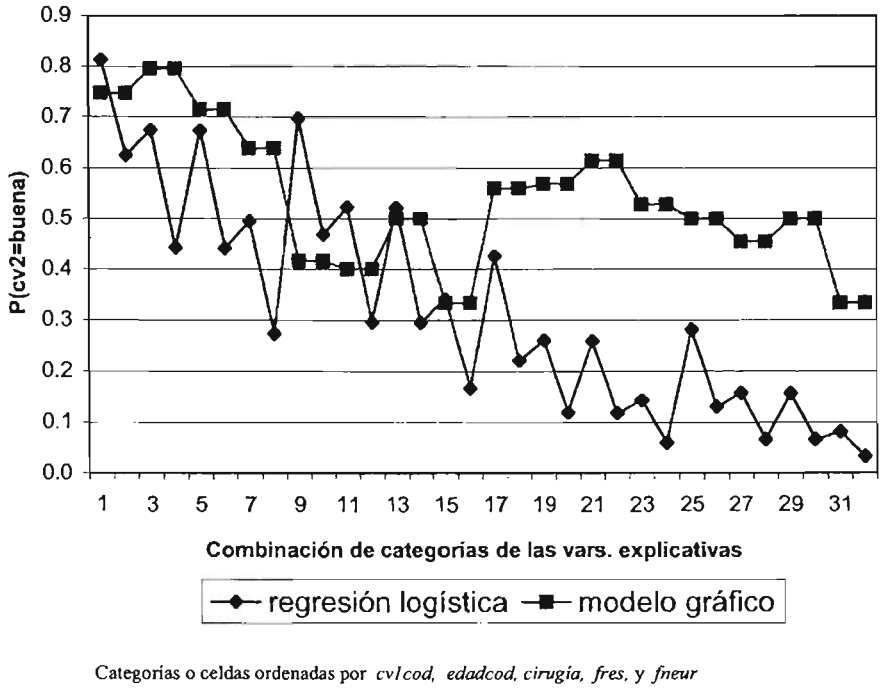

Figura 4.7: Probabilidades estimadas, $\hat{P}(cv2 = buena)$, bajo una regresión logística y bajo un modelo gráfico (orden 2)

Finalmente, se puede hacer una tabla de clasificación para ver con que porcentaje el modelo gráfico está clasificando correctamente las observaciones de la base de datos original, es decir, hay que comparar la calidad de vida posterior a la estancia en la UTI del individuo con la que predice el modelo. Para esto, a cada individuo en la base de datos se le asigna la probabilidad estimada de que su calidad de vida posterior sea buena según la combinación de valores que toman las variables explicativas, una vez hecho esto se utilizó como punto de corte el valor 0.5, de tal manera que si el individuo tiene una probabilidad estimada de 0.5 o más se le asigna a la categoría de que su calidad de vida posterior es buena y en caso contrario a que es mala, este valor de 0.5 se tomó de manera

arbitraria porque es el valor que el software empleado para llevar a cabo el ajuste de la regresión logística utiliza de manera prestablecida, entonces el mismo punto de corte se utilizó tanto en la regresión logística como en el modelo gráfico; sin embargo, los puntos de corte se puede asignar en base a otros métodos, como por ejemplo realizando un análisis de las curvas ROC. Una vez hecho esto para cada uno de los individuos en la base de datos original (en este caso son solo los sobrevivientes a la UTI que son 386 individuos), se obtiene la tabla 4.5 en la cual se observa que el 76.26 % de los individuos que al salir de la UTI tuvieron buena calidad de vida posterior son clasificados de manera correcta usando el modelo gráfico, es decir, el modelo predice que van a tener buena calidad de vida cuando realmente la tuvieron; de manera similar el 44.14 % de los individuos con mala calidad de vida posterior fueron clasificados correctamente y además se tiene que el 60.62 % del total de individuos están bien clasificados. Al usar la regresión logística se puede observar en la primer subtabla de la tabla 4.2 que el 74.20 % de los individuos con buena calidad de vida posterior fueron clasificados correctamente, así que hay un porcentaje un poco mayor de individuos bien clasificados con el modelo gráfico, por otro lado, se observa que el 66.5 % de los individuos con mala calidad de vida posterior son clasificados correctamente, así que en el modelo logístico la clasificación de los individuos con mala calidad es más certera.

celda	edad cod.	cirugía urg.	falla resp.	falla neur.	cv1 cod.	$\hat{P}_{reg.}$	$\hat{P}_{graf.}$	frec. rel. buena c.v.	obs.	$\lvert\hat{P}_{reg.} - \hat{P}_{graf.}\rvert$
1	1	0	0	0	1	0.8124	0.7467	0.8333	60	0.0657
2	1	0	0	1	1	0.6241	0.7467	1.0000	2	0.1226
3	1	0	1	0	1	0.6735	0.7955	0.7143	84	0.1220
4	1	0	1	1	1	0.4415	0.7955	0.5455	11	0.3540
5	1	1	0	0	1	0.6726	0.7143	0.6000	10	0.0417
6	1	1	0	1	1	0.4405	0.7143	0.0000	0	0.2738
7	1	1	1	0	1	0.4945	0.6389	0.4595	37	0.1444
8	1	1	1	1	1	0.2727	0.6389	0.3333	6	0.3662
9	2	0	0	0	1	0.6966	0.4167	0.7143	21	0.2799
10	2	0	0	1	1	0.4680	0.4167	0.0000	1	0.0513
11	2	0	1	0	1	0.5222	0.4000	0.4375	32	0.1222
12	2	0	1	1	1	0.2953	0.4000	0.0000	3	0.1047
13	2	1	0	0	1	0.5212	0.5000	0.0000	1	0.0212
14	2	1	0	1	1	0.2944	0.5000	0.0000	0	0.2056
15	2	1	1	0	1	0.3414	0.3333	0.3333	27	0.0081
16	2	1	1	1	1	0.1658	0.3333	0.0000	3	0.1675
17	1	0	0	0	2	0.4251	0.5600	0.1429	7	0.1349
18	1	0	0	1	2	0.2208	0.5600	0.0000	0	0.3392
19	1	0	1	0	2	0.2604	0.5690	0.1176	17	0.3086
20	1	0	1	1	2	0.1189	0.5690	0.0000	1	0.4501
21	1	1	0	0	2	0.2596	0.6154	0.3333	3	0.3558
22	1	1	0	1	2	0.1185	0.6154	0.0000	0	0.4969
23	1	1	1	0	2	0.1431	0.5283	0.0000	4	0.3852
24	1	1	1	1	2	0.0601	0.5283	0.0000	1	0.4682
25	2	0	0	0	2	0.2815	0.5000	0.3333	9	0.2185
26	2	0	0	1	2	0.1306	0.5000	0.0000	0	0.3694
27	2	0	1	0	2	0.1572	0.4545	0.2083	24	0.2973
28	2	0	1	1	2	0.0667	0.4545	0.0000	1	0.3878
29	2	1	0	0	2	0.1567	0.5000	0.5000	2	0.3433
30	2	1	0	1	2	0.0665	0.5000	0.0000	1	0.4335
31	2	1	1	0	2	0.0813	0.3333	0.2667	15	0.2520
32	2	1	1	1	2	0.0328	0.3333	0.0000	3	0.3005

Categorías o celdas ordenadas por *cv1cod*, *edadcod*, *cirugía*,*fresp* y *fneur*.

Tabla 4.4: Tabla de probabilidades estimadas, $\hat{P}(cv2 = buena)$, bajo regresión logística y bajo el modelo gráfico (orden 2)

Calidad de vida posterior				
	Predicha			
		mala	buena	Porcentaje correcto
Observada	mala	83	105	0.4414
	buena	47	151	0.7626
	total			0.6062

Tabla 4.5: Tabla de clasificación para la calidad de vida posterior, *cv2*, usando el modelo gráfico

A continuación surge la pregunta de cómo serían las probabilidades estimadas usando el modelo de la figura 4.2 (el cual solo permitía que hubiera o no arcos de cada una de las otras variables a la variable calidad de vida posterior *cv2*) cuando se fijan valores para las variables explicativas, las cuales son las restantes variables excepto falla neurológica pues esta variable no incide sobre *cv2* por lo que ya no se tomaría dentro del conjunto de variables explicativas (i.e. las variables explicativas de este modelo son calidad de vida inicial, edad codificada, cirugía urgente y falla respiratoria). Los valores estimados para toda combinación posible de valores de las variables explicativas para el modelo de la figura 4.2 se presentan en la tabla 4.6.

Hay que hacer algunas observaciones, resulta que en el modelo de la figura 4.2 coincide que las variables que apuntan a la variable *cv2* son las mismas que apuntan a la misma variable *cv2* en la figura 4.1, es decir que los nodos padres para *cv2* en ambas gráficas son los mismos (calidad de vida inicial, edad codificada, cirugía y falla respiratoria), también resulta que las probabilidades locales de *cv2* (las probabilidades condicionales de un nodo dado sus padres) aprendidas a partir de DEAL son las mismas en ambas gráficas, ahora bien, resulta que para predecir una observación se fijan todas las variables excepto a la variable *cv2* por lo cual en la figura 4.1 no se está permitiendo que en el momento de propagar la evidencia se consideren en los cálculos las

celda	edad cod.	cirugía urg.	falla resp.	cv1 cod.	$\hat{P}_{reg.}$	$\hat{P}_{graf.}$	obs.	$\lvert\hat{P}_{reg.} - \hat{P}_{graf.}\rvert$
1	1	0	0	1	0.8053	0.7467	62	0.0586
2	1	0	0	2	0.4226	0.5600	7	0.1374
3	1	0	1	1	0.6491	0.7955	95	0.1464
4	1	0	1	2	0.2466	0.5690	18	0.3224
5	1	1	0	1	0.6604	0.7143	10	0.0539
6	1	1	0	2	0.2560	0.6154	3	0.3594
7	1	1	1	1	0.4651	0.6389	43	0.1738
8	1	1	1	2	0.1333	0.5283	5	0.3950
9	2	0	0	1	0.6915	0.4167	22	0.2748
10	2	0	0	2	0.2839	0.5000	9	0.2161
11	2	0	1	1	0.5005	0.4000	35	0.1005
12	2	0	1	2	0.1506	0.4545	25	0.3039
13	2	1	0	1	0.5130	0.5000	1	0.0130
14	2	1	0	2	0.1571	0.5000	3	0.3429
15	2	1	1	1	0.3202	0.3333	30	0.0131
16	2	1	1	2	0.0769	0.3333	18	0.2564

Categorías o celdas ordenadas por *edadcod*, *cirugía*, *fresp* y *cv1cod*.

Tabla 4.6: Tabla de probabilidades estimadas, $\hat{P}(cv2 = buena)$, bajo regresión logística y bajo el modelo gráfico de la figura 4.2

probabilidades condicionales que hay entre las variables explicativas, como consecuencia de todo lo anterior al fijar todos los valores de las variables explicativas y dejar libre solo a la variable respuesta *cv2* y propagar la evidencia, las probabilidades estimadas de tener buena calidad de vida posterior sin considerar a la falla neurológica son las mismas para ambos modelos, pues aunque en el modelo de la figura 4.1 se considera a la falla neurológica al ser ésta condicionalmente independiente a *cv2* dada la falla respiratoria entonces una vez dado el valor de la falla respiratoria no importa qué valor tenga la falla neurológica, esto no afecta a las probabilidades estimadas, de hecho ya se mencionó antes que es por esta razón que hay valores repetidos con un cierto esquema de repetición en la tabla 4.3. Entonces, si no se toma en cuenta a la variable falla neurológica la tabla de probabilidades estimadas para el modelo de la figura 4.2 tiene los mismos valores que la tabla para la gráfica de la figura 4.1 (tabla 4.3), es decir se colapsa la tabla, y ya no hay valores estimados con el patrón de repetición que se observaba en la tabla 4.3, el cual como ya se dijo se debía a la independencia condicional.

El modelo gráfico de la figura 4.2 puede compararse con una regresión logística con la calidad de vida posterior como variable respuesta y la calidad de vida inicial, edad codificada, cirugía y falla respiratoria como variables explicativas; los coeficientes estimados se presentan en la tabla 4.7 en la cual se observa que la variable de más peso sigue siendo la calidad de vida inicial (nuevamente los coeficientes estimados, coeficientes de Wald y los coeficientes estandarizados, Apéndice Tabla B.1, son mayores que los correspondientes a los de las otras variables explicativas). A partir de la tabla 4.7 se obtuvieron las estimaciones de la probabilidad de tener buena calidad de vida bajo el modelo logístico. Nuevamente se podría hacer una gráfica en la que se comparan las probabilidades estimadas con ambos modelos para las 16 posibles combinaciones de valores de las variables explicativas (figura 4.8) y otra gráfica en la que en un eje

se tiene la probabilidad estimada bajo el modelo gráfico y en el otro bajo regresión logística (figura 4.9); sin embargo, las conclusiones que se tenían antes no cambian: en el modelo gráfico la edad es la variable de mayor peso, al ser la que más influye en el aumento o disminución de la probabilidad de tener buena calidad de vida posterior, en el modelo logístico, como ya se dijo, la variable de más peso es la calidad de vida inicial, además cuando la celda corresponde a observaciones cuya edad se encuentra en el rango inferior las tendencias de las estimaciones son similares entre ambos modelos, otra vez en general las estimaciones para el modelo gráfico son mayores que para el modelo logístico, nuevamente la celda con mayor número de observaciones en la base de datos (95) corresponde a individuos en el rango inferior de edad, sin cirugía urgente, con falla respiratoria y buena calidad de vida inicial y con una diferencia entre las estimaciones de 0.146 unidades; sin embargo, ahora las diferencias entre las estimaciones son menores, ahora la diferencia mayor es de 0.395 (antes era 0.496) y corresponde a la celda 8 en la que el rango de edad es inferior, con cirugía urgente, con falla respiratoria y con mala calidad de vida inicial.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	EDADCOD60(1)	.613	.242	6.417	1	.011	1.847	1.149	2.968
	CV1COD(1)	1.732	.311	31.081	1	.000	5.649	3.073	10.384
	FRESP(1)	.805	.259	9.688	1	.002	2.237	1.347	3.713
	QXURGENT(1)	.755	.255	8.768	1	.003	2.127	1.291	3.505
	Constant	-2.485	.361	47.441	1	.000	.083		

a. Variable(s) entered on step 1: EDADCOD60, CV1COD, FRESP, QXURGENT.

Tabla 4.7: Coeficientes estimados para la regresión con variable respuesta calidad de vida posterior (*cv2*) sin *fneur*

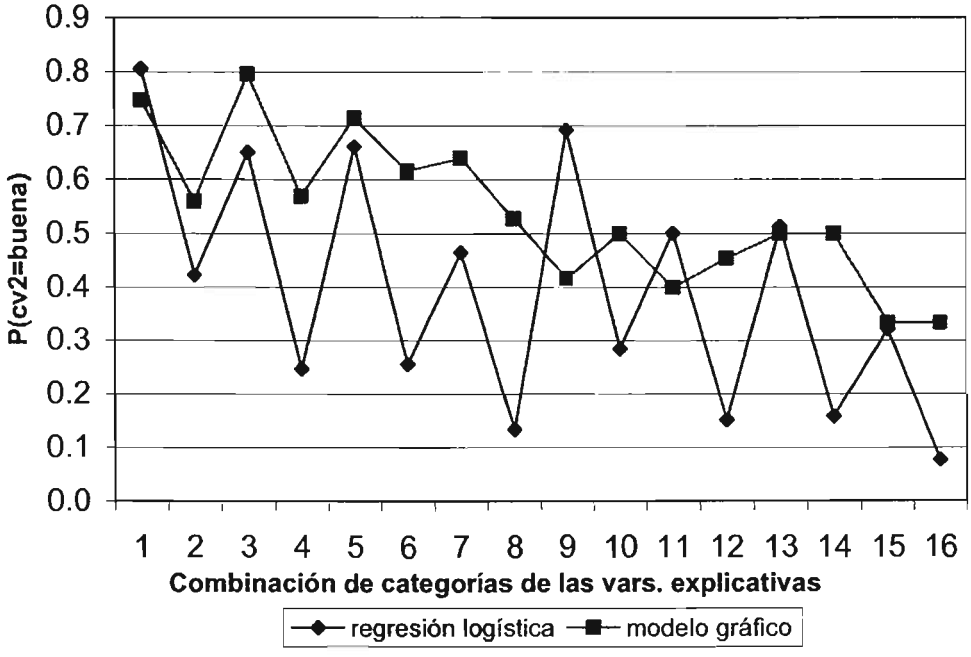

Figura 4.8: Probabilidades estimadas, $\hat{P}(cv2 = buena)$, bajo una regresión logística y bajo el modelo gráfico de la figura 4.2

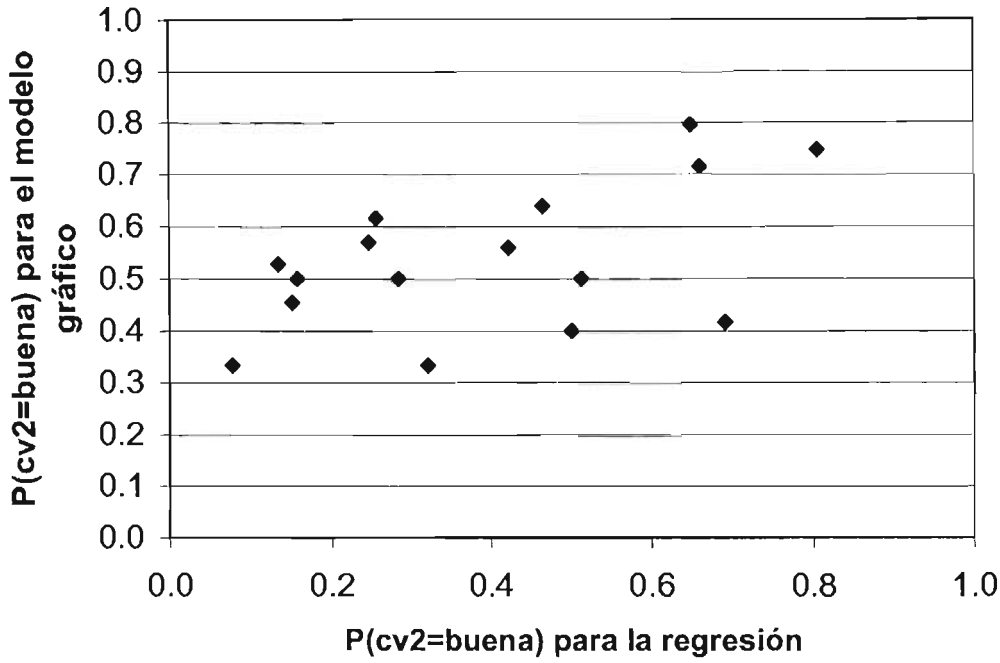

Figura 4.9: Probabilidades estimadas $\hat{P}_{reg.}(cv2 = buena)$ vs $\hat{P}_{graf.}(cv2 = buena)$ usando el modelo gráfico de la figura 4.2

Como ya se mencionó anteriormente una de las causas de que las estimaciones sean las mismas con los dos modelos gráficos, el de la figura 4.1 y el de la 4.2, era que a todas las variables excepto la calidad de vida posterior, *cv2*, se les daba un valor específico y entonces al propagar evidencia no se tomaban en cuenta las probabilidades condicionales entre las variables que no fueran las que involucraban a *cv2*. Entonces surge la pregunta de qué pasaría si dejáramos un par de variables libres, en el sentido que no les asignamos valores específicos, mientras que al resto de las variables sí y entonces la idea es ver que ocurre con este par de variables en ambos modelos al propagar la evidencia. Por ejemplo supongamos que se tienen valores fijos para todas las variables excepto la calidad de vida inicial (*cv1cod*) y la calidad de vida posterior (*cv2*) así que

queremos ver las probabilidades en cada categoría de estas variables una vez que se ha propagado la evidencia en la red o sea una vez que se han actualizado las probabilidades. Este ejemplo correspondería al caso en que se saben los valores de las variables correspondientes a cuando el paciente está en la terapia intensiva (se sabe si tiene fallas, necesitó cirugía urgente y la edad del paciente), pero no se sabe en que estado llegó el paciente, o sea su calidad de vida inicial, ni tampoco se sabe en que estado va a salir una vez que estuvó en la UTI, o sea su calidad de vida posterior. Supógase que se tiene un individuo sin cirugía urgente, en el grupo de edad inferior, sin falla neurológica y sin falla respiratoria, entonces al propagar la evidencia la probabilidad estimada bajo el modelo de la gráfica en la figura 4.1 (en el que las variables pueden apuntarse entre sí) de que el individuo tuviera buena calidad de vida inicial sería de 0.692 y de que tenga buena calidad de vida posterior es de 0.689 (ver figura 4.10), en cambio la probabilidad estimada bajo el modelo de la gráfica en la figura 4.2 (en que las variables solo apuntan a *cv2*) de que el individuo tuviera buena calidad de vida inicial sería de 0.704 y de que tenga buena calidad de vida posterior es de 0.6915 (ver figura 4.11), entonces se observa que en este caso las estimaciones sí son diferentes puesto que se están tomando en cuenta otras probabilidades condicionales además de las del nodo correspondiente a *cv2* y de hecho ocurre que las probabilidades estimadas en el modelo de la figura 4.2 son mayores que las del modelo de la figura 4.1.

De manera similar, se pueden asignar valores fijos a cualquier subconjunto de variables de interés y observar cómo se afectan las probabilidades de que otras variables tomen valores específicos.

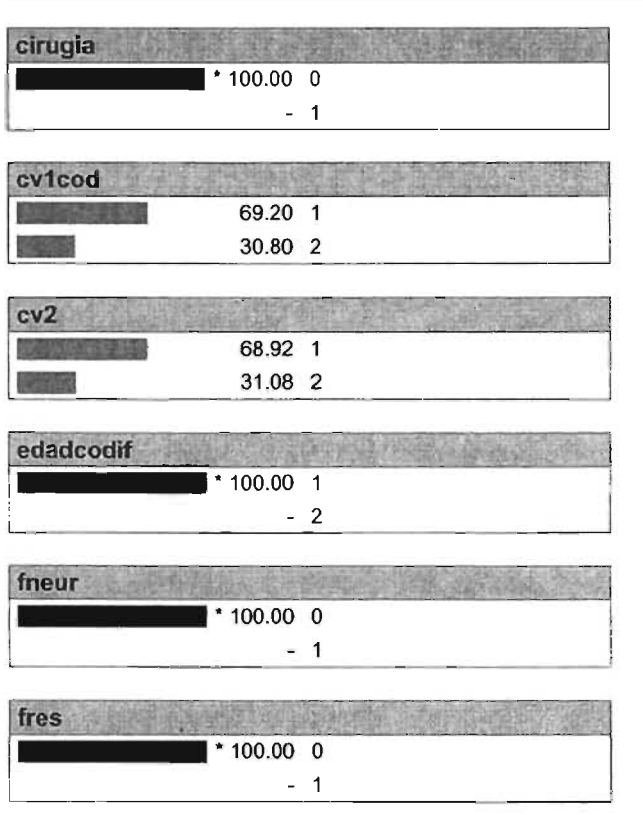

Figura 4.10: Probabilidades marginales para *cv1cod* y *cv2* una vez introducida la evidencia para la gráfica de la figura 4.1

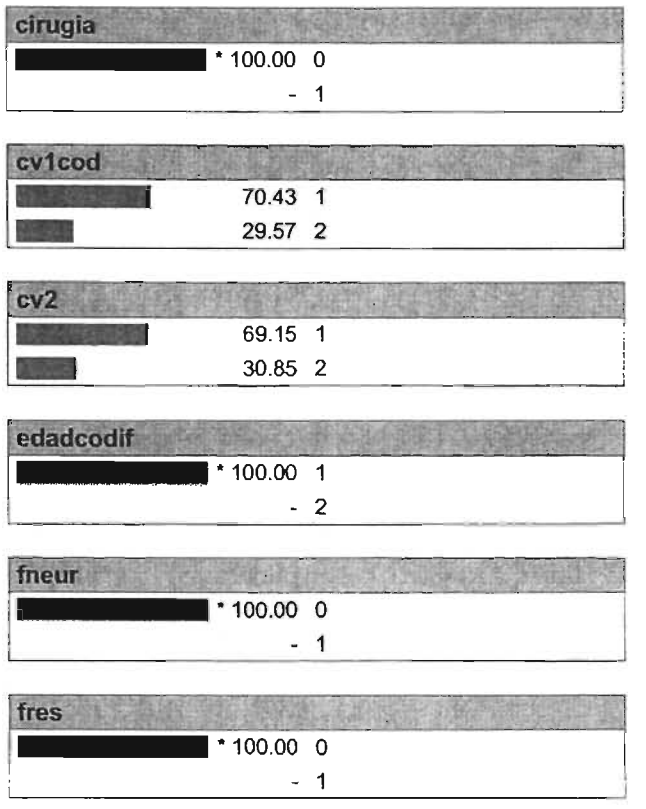

Figura 4.11: Probabilidades marginales para *cv1cod* y *cv2* una vez introducida la evidencia para la gráfica de la figura 4.2

4.2. Modelo para la variable que identifica vivos y muertos

En este apartado consideramos el caso donde el modelo no solo incluye los sobrevivientes sino que también incluye a los muertos, entonces se utiliza la variable binaria *vivomuer* con categorías vivo y muerto, como ya se dijo en el capítulo anterior la gráfica definitiva a la que se llegó incluye las variables: edad dicotomizada (*edadcod60*), la

calidad de vida inicial (*cv1cod*), la mortalidad predicha categorizada (*mortpredcod*), la infección noscomial (*infnos*), la variable que identifica y separa a los vivos de los muertos (*vivomuer*), la falla respiratoria (*fresp*) y la falla cardiaca (*fcard*).

Una vez que se han hecho las restricciones pertinentes en los arcos para obtener una gráfica con relaciones coherentes con la realidad se obtiene la gráfica de la figura 4.12, se puede observar: 1) Todas las variables inciden a la variable que distingue vivos de muertos (*vivomuer*) a excepción de la variable relacionada con falla cardiaca (*fcard*); sin embargo, esta variable afecta indirectamente a la variable *vivomuer*, puesto que la falla cardiaca (*fcard*) afecta a la mortalidad predicha codificada que a su vez afecta a la variable que distingue vivos de muertos (*vivomuer*), de forma similar la falla cardiaca incide en la variable relativa a tener o no infección nosocomial que incide a su vez en la variable *vivomuer*. 2) Se observa que la variable relativa a la edad se relaciona directamente con todas las variables al haber arcos que unen edad con cada una de las variables, por otra parte la variable mortalidad predicha se puede ver que aparte de depender de la variable edad también depende de las fallas orgánicas, en este caso falla cardiaca y falla respiratoria. 3) El tener o no infección nosocomial depende de la edad y además de las fallas, lo cual es lógico pues al tener fallas orgánicas el paciente es más vulnerable a infecciones al requerir de tratamientos como respiración artificial, entubamientos, etc. con los cuales podría infectarse en el hospital. 4) La variable falla respiratoria parece como un factor que incide en la falla cardiaca, además de que afecta la mortalidad predicha, infección nosocomial y a la variable que separa a vivos de muertos (*vivomuer*) como ya se dijo.

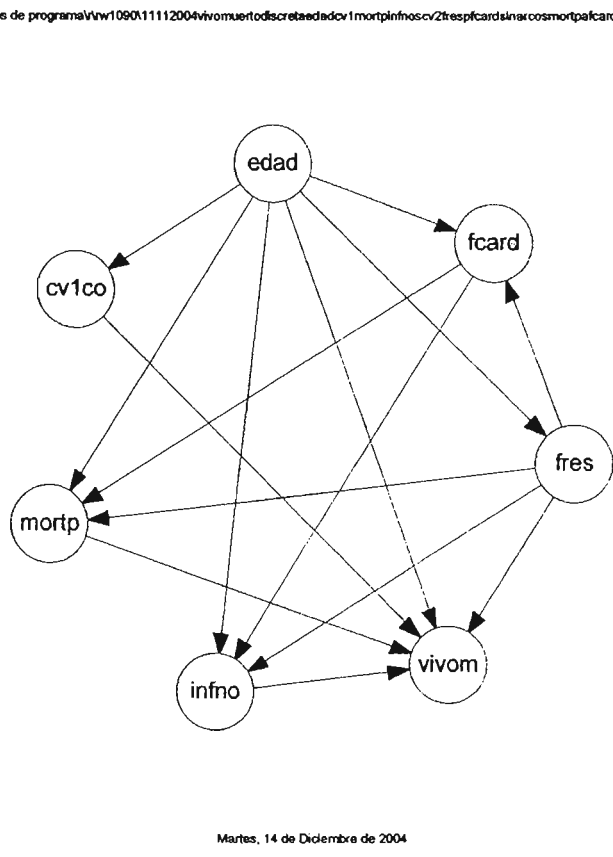

Figura 4.12: Modelo gráfico que involucra la variable que separa a vivos de muertos (*vivomuer*)

Se podrían ver algunas relaciones de independencia condicional entre variables o conjunto de variables, nuevamente tomando en cuenta los conjuntos separadores como se definieron en la sección 2.3, así se observan las siguientes independencias condicionales.

1. La falla cardiaca es condicionalmente independiente de la variable que identifica a los vivos de los muertos dada la edad codificada, la mortalidad predicha categorizada, la infección nosocomial y la falla respiratoria, es decir $fcard \perp vivomuer$

$|edad, mortpredcod, infnos, fresp.$

2. La calidad de vida inicial es condicionalmente independiente de la mortalidad predicha categorizada dada la edad y la variable que separa a vivos de muertos, $cv1cod \perp mortpredcod|edad, vivomuer$.

3. Las fallas y la mortalidad predicha son condicionalmente independientes de la calidad de vida inicial dadas las variables edad categorizada y la variable que identifica a los vivos de los muertos, $fcard, fresp, mortpredcod \perp cv1cod|edad, vivomuer$.

4. La calidad de vida inicial es independiente condicionalmente de la infección nosocomial dada la edad categorizada y la variable que representa a vivos y muertos, $cv1cod \perp infnos|edad, vivomuer$.

5. La calidad de vida inicial es condicionalmente independiente de la falla respiratoria dada la edad categorizada y la variable que separa a vivos de muertos, $cv1cod \perp fresp|edad, vivomuer$.

6. La calidad de vida inicial es condicionalmente independiente de la falla cardiaca dada la falla respiratoria, la infección nosocomial, la mortalidad predicha y la edad categorizadas, o sea $cv1cod \perp fcard|\ fresp, infnos, mortpredcod, edad$, también ocurre que ambas variables son condicionalmente independientes dadas la edad categorizada y la variable que separa a vivos de muertos, $cv1cod \perp fcard|\ edad, vivomuer$.

7. La mortalidad predicha categorizada es condicionalmente independiente de la infección nosocomial dadas la edad, las dos fallas y la variable que identifica a vivos de muertos, así que $mortpredcod \perp infnos|edad, fresp, fcard, vivomuer$.

De manera similar que en la sección anterior se ajusta una regresión logística, los coeficientes estimados se presentan en la tabla 4.8. En el modelo se considera como variable respuesta a la variable *vivomuer* y como explicativas a las restantes, de acuerdo a los valores de los "p-values" asociados a cada variable se observa que todas las variables son estadísticamente significativas, se observa también que todos los coeficentes son positivos y la exponencial de ellos son números mayores que uno, lo cual indica que cuando un individuo se encuentra en mejores condiciones se incrementa la probabilidad o riesgo de que la persona esté viva después de su estancia en la UTI. Así, cuando el individuo tiene una edad inferior a 60 años aumenta la probabilidad de que esté vivo, de manera similar cuando su calidad de vida inicial es buena, cuando su mortalidad predicha se encuentra en la categoría con valores inferiores, cuando no hay una infección nosocomial y cuando no presenta falla cardiaca o falla respiratoria aumenta la probabilidad de sobrevivir tal y como era de esperarse.

Nuevamente, se puede analizar la calidad de ajuste del modelo. Se tiene que la devianza residual (subtabla 3 tabla 4.8) es de 53.812 que es pequeña comparada con un cuantil de la ji cuadrada con 51 grados de libertad para un α de por ejemplo 0.05, de hecho el nivel crítico correspondiente es 0.367 lo cual indica que de manera significativa no se rechaza que la diferencia entre el modelo saturado y el ajustado es poca, similarmente con la estadística de Pearson se tiene un nivel crítico de 0.644 lo que indica, como en la prueba anterior, que el modelo se ajusta bien a los datos y que no hay necesidad de incluir interacciones entre las variables. Otra vez, la tabla de clasificación del modelo ajustado puede compararse la del modelo saturado correspondiente; sin embargo, como en el caso del modelo que incluía la calidad de vida posterior hay problemas numéricos en el software utilizado que afectan las estimaciones, entonces aunque la tabla de calsificación es similar a la de la regresión ajustada (en el modelo saturado 77.9% de

los muertos están bien clasificados en el ajustado 81.5 %; en el modelo saturado 65.8 % de los vivos están bien clasificados en el ajustado 57.8 %) los resultados obtenidos no son confiables. Por otra parte, el nivel crítico para la estadística de prueba con la que se compara el modelo ajustado con el nulo es aproximadamente cero (subtabla 4 tabla 4.8), indicando que en general se rechaza la hipótesis nula de que el modelo ajustado no supera al nulo y entonces las variables explicativas que se están agregando al modelo son importantes para explicar a la variable respuesta. Finalmente se tienen valores de las pseudo R^2 (subtabla 5 tabla 4.8) entre 0.159 y 0.263, así que aproximadamente entre el 15.90 % y el 26.30 % de la variabilidad está siendo explicada por el modelo.

Classification Table[a]

Observed			Predicted		
			VIVOMUER		Percentage Correct
			muerto	vivo	
Step 1	VIVOMUER	muerto	387	88	81.5
		vivo	163	223	57.8
	Overall Percentage				70.8

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	EDADCOD60(1)	.392	.166	5.596	1	.018	1.480	1.070	2.049
	CV1COD(1)	.458	.177	6.693	1	.010	1.580	1.117	2.235
	MORTPR_A(1)	.968	.161	35.959	1	.000	2.632	1.918	3.611
	INFNOSOC(1)	.696	.170	16.775	1	.000	2.006	1.438	2.799
	FRESP(1)	1.359	.227	35.873	1	.000	3.894	2.496	6.075
	FCARD(1)	.418	.162	6.682	1	.010	1.519	1.106	2.087
	Constant	-1.981	.205	93.559	1	.000	.138		

a. Variable(s) entered on step 1: EDADCOD6, CV1CODIF, MORTPR_A, INFNOSOC, FRESP, FCARD.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	46.725	51	.644
Deviance	53.812	51	.367

Model Fitting Information

	Model Fitting Criteria			Likelihood Ratio Tests		
Model	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	365.355	370.113	363.355			
Final	189.226	222.533	175.226	188.129	6	.000

Pseudo R-Square

Cox and Snell	.196
Nagelkerke	.263
McFadden	.159

Tabla 4.8: Coeficientes ajustados, tabla de clasificación y tablas para ver la calidad de ajuste para la regresión con variable respuesta *vivomuer*, variable que separa a vivos de muertos

Por otra parte, otra vez se pueden analizar los coeficientes de Wald (Tabla 4.8) y sobre todo los coeficientes estandarizados de este modelo (Apéndice, tabla B.1) para determinar cuáles son las variables que influyen más sobre la variable respuesta. se puede observar que los coeficientes más grandes corresponden a las variables *mortpredcod* y a *fresp*, así que estas dos variables son las que más influyen sobre la variable respuesta, de tal modo que al cambiar los valores de estas variables aumenta o disminuye en mayor medida la probabilidad estimada.

Para analizar la relación de la regresión anterior con otro modelo gráfico, se ajusta

un modelo que solo permite que haya o no arcos de cada una de las variables explicativas a la variable que *vivomuer*, se observa en la figura 4.13 que, a diferencia de la regresión logística, no hay relación de dependencia (no hay un arco) entre la variable falla cardiaca y la variable *vivomuer*, así que en el modelo gráfico no se considera que la falla cardiaca afecta directamente al hecho de que una persona viva o muera.

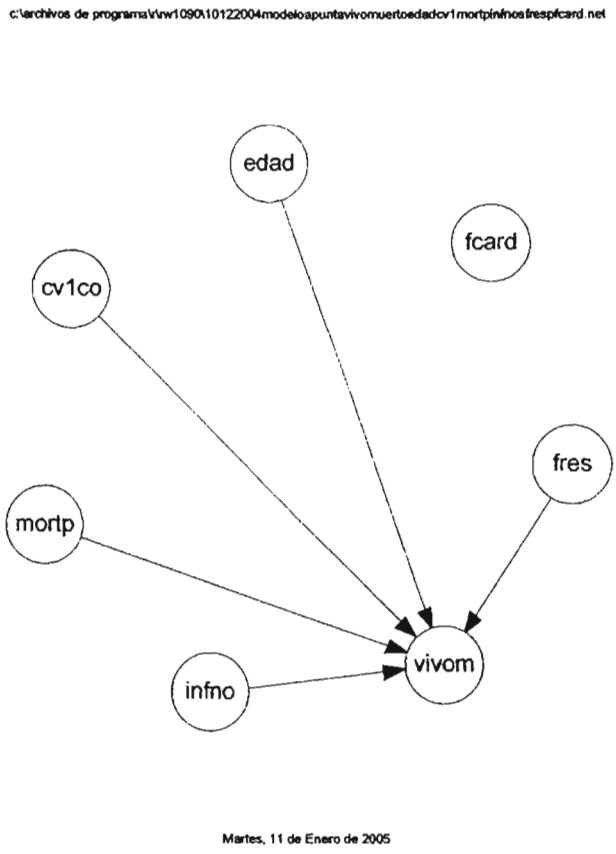

Figura 4.13: Modelo gráfico que solo permite arcos que inciden a *vivomuer*

Al igual que en la sección 4.1, el siguiente paso es estimar probabilidades usando

tanto el modelo gráfico de la figura 4.12 como el logístico, en este caso la probabilidad p que se está estimando corresponde a la probabilidad de que el paciente sobreviva. Para estimar las probabilidades bajo el modelo logístico se utilizan los coeficientes estimados presentados en la tabla 4.8, correspondientes al modelo siguiente

$\log\left(\frac{\hat{P}(vivomuer=vivo)}{\hat{P}(vivomuer=muerto)}\right)$ =-1.98+ 0.39 $(edadcod60 \leq 60)$+ 0.45 $(cv1cod = buena)$+ 0.96 $(mortpredcod \leq 17)$+ 0.69 $(infnos = no)$+ 1.35 $(fresp = no)$+ 0.41 $(fcard = no)$.

Entonces, según los valores de las variables explicativas, puede estimarse p como en la sección anterior; así, si por ejemplo se tiene un individuo de 60 o menos años de edad, con una buena calidad de vida inicial, con niveles de mortalidad predicha bajos, sin infección noscomial y sin falla cardíaca, ni respiratoria entonces la probabilidad estimada de que el individuo sobreviva es:

$$\hat{p} = \frac{\exp(-1.981 + 0.392 + 1.458 + 0.968 + 0.696 + 1.359 + 0.418)}{1 + \exp(-1.981 + 0.392 + 1.458 + 0.968 + 0.696 + 1.359 + 0.418)} = 0.9097$$

Usando el modelo gráfico presentado en la figura 4.12 puede estimarse para el mismo individuo la probabilidad de que sobreviva. Nuevamente todas las variables son binarias y los valores o etiquetas de cada una de las categorías de las variables se encuentran en la tabla 4.1, por ejemplo para la variable calidad de vida inicial la primer categoría etiquetada con 1 representa cuando la calidad inicial era buena y el valor 2 cuando era mala. Entonces, para obtener la probabilidad estimada, otra vez se compila la red y después se ingresan los valores que toman las variables para el individuo (figura 4.14), obteniendo al final que la probabilidad estimada es 0.7612.

Finalmente, como en la sección anterior, se pueden estimar las probabilidades de

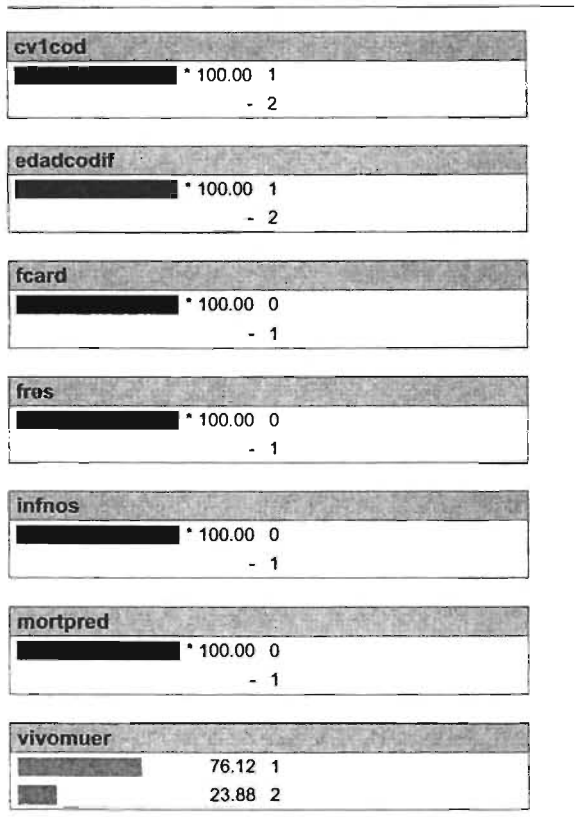

Figura 4.14: Probabilidades marginales para *vivom* una vez introducida la evidencia

sobrevivencia con ambos modelos para cada posible combinación de valores que toman las variables explicativas o patrones de covariables, en este caso como se tienen seis variables explicativas el total de valores posibles que pueden tomar en conjunto las variables explicativas es $2^6 = 64$. La tabla 4.9 presenta las probabilidades estimadas bajo ambos modelos, la gráfica en la figura 4.15 permite comparar las probabilidades entre ambos modelos para cada una de los 64 posibles valores y finalmente en la figura 4.16 se grafica la probabilidad estimada bajo la regresión contra la probabilidad estimada bajo el modelo gráfico para cada una de las 64 celdas.

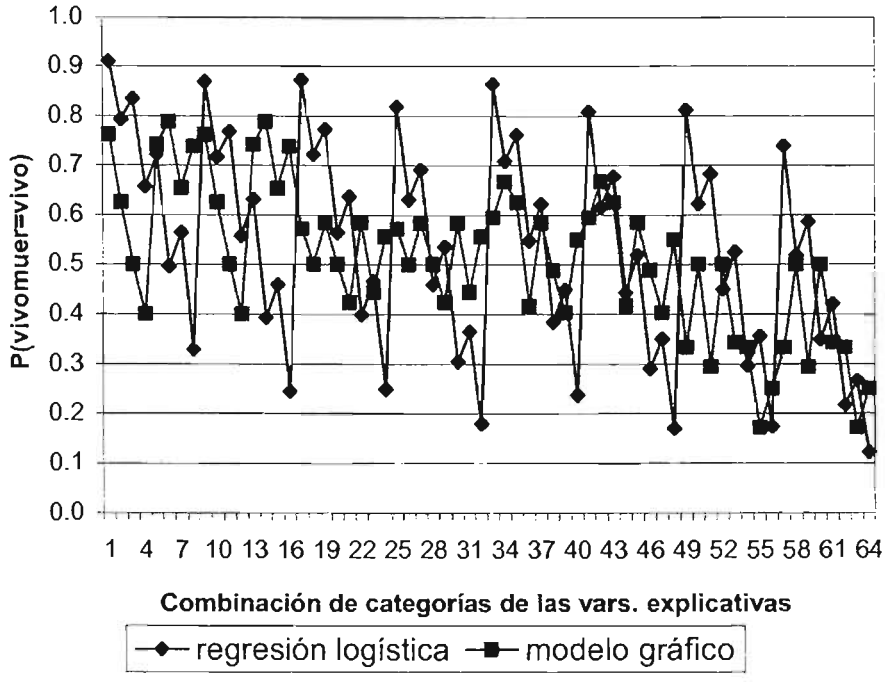

Figura 4.15: Probabilidades estimadas, $\hat{P}(vivomuer = vivo)$, bajo una regresión logística y bajo un modelo gráfico (orden 1)

Se observa que en el modelo gráfico las probabilidades estimadas se encuentran entre 0.171 y 0.787, en el logístico entre 0.121 y 0.909, así que las estimaciones para el modelo gráfico toman un rango amplio de valores, sobre todo al compararlo con el modelo de la sección anterior; de hecho este rango de valores para el modelo gráfico es parecido al que toman las estimaciones bajo la regresión aunque un poco menor. También se observa que no hay tendencias similares en las estimaciones entre ambos modelos, puesto que a veces en una combinación de valores la estimación para el modelo gráfico sube, mientras que para el logístico baja y viceversa.

Al comparar con el modelo de la sección anterior (para la calidad de vida posterior

o *cv2*) se tiene que las predicciones se encuentran más cercanas entre sí, en promedio la diferencia en valor absoluto entre las estimaciones (64 estimaciones) usando ambos modelos es de 0.177, en cambio en el modelo para *cv2* de la sección anterior era 0.249. El valor mínimo de esta diferencia es de 0.018 y el máximo de 0.494. La diferencia más grande corresponde a la celda 16 que es un individuo con buena calidad de vida inicial, en el rango de edad menor, con falla cardiaca y respiratoria, infección nosocomial y en el rango mayor de mortalidad predicha, para este individuo el modelo logístico considera que tiene poca probabilidad de vivir (0.244) en cambio el gráfico considera que su probabilidad de vivir es alta (0.738). De la celda 16 (tipo de persona con buena calidad de vida inicial, más joven y con todo lo demás mal) a la 17 (tipo de persona con buena calidad de vida inicial, en el rango de edad mayor y todo lo demás bien) la probabilidad estimada en el modelo logístico sube hasta 0.871, en cambio en el gráfico baja a 0.571; así que para el modelo gráfico la edad es importante pues para una persona con buena calidad de vida inicial, en el rango de edad menor y todo lo demás mal la probabilidad de vivir es más alta que para una persona que también tiene buena calidad de vida inicial y todo lo demás a su favor, salvo que es una persona ya mayor (algo similar ocurre de la celda 48 a la 49 solo que en este caso la persona tiene mala calidad de vida inicial), entonces nuevamente edad es una variable de peso en el modelo gráfico.

Otra celda donde la diferencia entre las probabilidades estimadas es importante (0.376) es la 32 que corresponde a una persona con buena calidad de vida inicial y todo lo demás mal (personas mayores, con fallas, infección nosocomial y alta mortalidad predicha) pues el modelo logístico proporciona un valor de 0.179 y el gráfico de 0.555, de hecho ocurre que para personas con buena calidad de vida inicial el modelo gráfico siempre proporciona estimaciones por arriba de 0.4, en cambio en la regresión hay valores que son menores (hasta 0.179). En realidad, hasta que la calidad de vida inicial

es mala (a partir de la celda 33) el modelo gráfico estima los valores más pequeños (hasta 0.171), así que la calidad de vida inicial también es una variable importante en el modelo gráfico que hace disminuir o aumentar mucho las probabilidades estimadas.

Para comprobar como la calidad de vida inicial y la edad están afectando al modelo gráfico se pueden observar las siguientes celdas: la celda 33 que corresponde a una persona con mala calidad de vida inicial, joven y todo lo demás bien, en este caso el modelo gráfico estima una probabilidad de vivir de 0.593 y la regresión logística de 0.864, así que el hecho de tener mala calidad de vida inicial está afectando más al modelo gráfico que al logístico; al tomar la celda 49 correspondiente a una persona con mala calidad de vida inicial, en un rango de edad mayor y lo demás bien el modelo logístico estima una probabilidad de 0.811, mientras que el gráfico de 0.333, lo cual verifica como la calidad de vida inicial y la edad afectan la probabilidad de vivir en el modelo gráfico y no tanto al logístico. De hecho cuando la calidad de vida inicial es mala y la persona es mayor (a partir de la celda 49) en el modelo gráfico todas las probabilidades estimadas son menores o iguales a 0.5, en el modelo logístico no ocurre esto. Finalmente, otra combinación de valores interesante de considerar es aquella en la que todo está mal en un inividuo, en este caso el modelo gráfico estima una probabilidad de 0.25, mientras que el logístico de 0.12.

celda	cv1 cod.	edad cod.	falla card.	falla resp.	inf. nos.	mort. pred.	$\hat{P}_{reg.}$	$\hat{P}_{graf.}$	frec. rel. vivos	obs.	$\lvert\hat{P}_{reg.} - \hat{P}_{graf.}\rvert$	$\frac{\lvert\hat{P}_{reg.} - \hat{P}_{graf.}\rvert}{\hat{P}_{graf.}}$
1	1	1	0	0	0	0	0.9097	0.7612	0.9697	33	0.1485	0.1951
2	1	1	0	0	0	1	0.7928	0.6250	0.7333	15	0.1678	0.2685
3	1	1	0	0	1	0	0.8340	0.5000	1.0000	5	0.3340	0.6679
4	1	1	0	0	1	1	0.6561	0.4000	0.6667	3	0.2561	0.6403
5	1	1	0	1	0	0	0.7213	0.7416	0.7609	46	0.0203	0.0274
6	1	1	0	1	0	1	0.4958	0.7879	0.5476	42	0.2921	0.3708
7	1	1	0	1	1	0	0.5634	0.6532	0.6250	16	0.0898	0.1375
8	1	1	0	1	1	1	0.3289	0.7381	0.1875	16	0.4092	0.5543
9	1	1	1	0	0	0	0.8690	0.7612	0.9375	16	0.1078	0.1416
10	1	1	1	0	0	1	0.7159	0.6250	0.6667	6	0.0909	0.1454
11	1	1	1	0	1	0	0.7678	0.5000	1.0000	1	0.2678	0.5356
12	1	1	1	0	1	1	0.5568	0.4000	1.0000	2	0.1568	0.3919
13	1	1	1	1	0	0	0.6302	0.7416	0.6136	44	0.1114	0.1502
14	1	1	1	1	0	1	0.3929	0.7879	0.3333	45	0.3950	0.5013
15	1	1	1	1	1	0	0.4593	0.6531	0.3636	33	0.1938	0.2967
16	1	1	1	1	1	1	0.2440	0.7381	0.2826	46	0.4941	0.6695
17	1	2	0	0	0	0	0.8719	0.5714	0.8333	6	0.3005	0.5259
18	1	2	0	0	0	1	0.7211	0.5000	0.7500	8	0.2211	0.4422
19	1	2	0	0	1	0	0.7724	0.5833	0.5000	4	0.1891	0.3242
20	1	2	0	0	1	1	0.5632	0.5000	0.0000	1	0.0632	0.1263
21	1	2	0	1	0	0	0.6362	0.4242	0.6923	13	0.2120	0.4998
22	1	2	0	1	0	1	0.3992	0.5833	0.2667	30	0.1841	0.3157
23	1	2	0	1	1	0	0.4658	0.4444	0.3333	3	0.0214	0.0482
24	1	2	0	1	1	1	0.2488	0.5556	0.3636	11	0.3068	0.5522
25	1	2	1	0	0	0	0.8176	0.5714	0.8750	8	0.2462	0.4308
26	1	2	1	0	0	1	0.6299	0.5000	1.0000	3	0.1299	0.2599
27	1	2	1	0	1	0	0.6908	0.5833	0.0000	0	0.1075	0.1843
28	1	2	1	0	1	1	0.4591	0.5000	0.0000	2	0.0409	0.0818
29	1	2	1	1	0	0	0.5352	0.4242	0.5263	19	0.1110	0.2616
30	1	2	1	1	0	1	0.3043	0.5833	0.2609	69	0.2790	0.4783
31	1	2	1	1	1	0	0.3647	0.4444	0.2857	7	0.0797	0.1793
32	1	2	1	1	1	1	0.1790	0.5556	0.3023	43	0.3766	0.6778
33	2	1	0	0	0	0	0.8644	0.5938	0.7778	9	0.2706	0.4556
34	2	1	0	0	0	1	0.7077	0.6667	0.0000	1	0.0410	0.0614
35	2	1	0	0	1	0	0.7606	0.6250	0.0000	0	0.1356	0.2170
36	2	1	0	0	1	1	0.5469	0.4167	0.0000	0	0.1302	0.3124
37	2	1	0	1	0	0	0.6208	0.5833	0.7143	7	0.0375	0.0643
38	2	1	0	1	0	1	0.3834	0.4878	0.6667	6	0.1044	0.2140
39	2	1	0	1	1	0	0.4494	0.4031	0.6667	3	0.0463	0.1149
40	2	1	0	1	1	1	0.2367	0.5495	0.0000	1	0.3128	0.5693
41	2	1	1	0	0	0	0.8075	0.5938	0.5000	2	0.2137	0.3599
42	2	1	1	0	0	1	0.6144	0.6667	1.0000	1	0.0523	0.0784
43	2	1	1	0	1	0	0.6766	0.6250	0.0000	0	0.0516	0.0825
44	2	1	1	0	1	1	0.4428	0.4167	0.5000	2	0.0261	0.0625
45	2	1	1	1	0	0	0.5187	0.5833	0.4545	11	0.0646	0.1107
46	2	1	1	1	0	1	0.2905	0.4878	0.3636	11	0.1973	0.4045
47	2	1	1	1	1	0	0.3496	0.4031	0.0909	11	0.0535	0.1328
48	2	1	1	1	1	1	0.1695	0.5495	0.1538	13	0.3800	0.6915
49	2	2	0	0	0	0	0.8115	0.3333	1.0000	2	0.4782	1.4348
50	2	2	0	0	0	1	0.6206	0.5000	0.5000	6	0.1206	0.2412
51	2	2	0	0	1	0	0.6822	0.2941	0.0000	0	0.3881	1.3197
52	2	2	0	0	1	1	0.4492	0.5000	0.0000	1	0.0508	0.1016
53	2	2	0	1	0	0	0.5252	0.3429	0.6364	11	0.1823	0.5317
54	2	2	0	1	0	1	0.2959	0.3333	0.2500	24	0.0374	0.1123
55	2	2	0	1	1	0	0.3555	0.1711	0.0000	1	0.1844	1.0777
56	2	2	0	1	1	1	0.1732	0.2500	0.0000	5	0.0768	0.3071
57	2	2	1	0	0	0	0.7392	0.3333	0.7500	4	0.4059	1.2179
58	2	2	1	0	0	1	0.5185	0.5000	0.3750	8	0.0185	0.0370
59	2	2	1	0	1	0	0.5856	0.2941	0.0000	0	0.2915	0.9913
60	2	2	1	0	1	1	0.3493	0.5000	1.0000	1	0.1507	0.3013
61	2	2	1	1	0	0	0.4214	0.3429	0.5333	15	0.0785	0.2290
62	2	2	1	1	0	1	0.2167	0.3333	0.2407	54	0.1166	0.3498
63	2	2	1	1	1	0	0.2664	0.1711	0.1250	8	0.0953	0.5569
64	2	2	1	1	1	1	0.1212	0.2500	0.1702	47	0.1288	0.5152

Categorías o celdas ordenadas por *cv1cod*, *edadcod60*, *fcard,fresp*, *infnos* y *mortpredcod*.

Tabla 4.9: Tabla de probabilidades estimadas, $\hat{P}(vivomuer = vivo)$, bajo regresión logística y bajo el modelo gráfico (orden 1)

Se tiene que la combinación de categorías con mayor número de observaciones en la base de datos es la correspondiente a la celda 30 (69 observaciones) que son personas con buena calidad de vida inicial, mayores, con fallas, sin infección nosocomial y alta mortalidad predicha, en este caso la regresión estima un valor de 0.304 y el modelo gráfico de 0.583. También hay 54 individuos con mala calidad de vida inicial, mayores, con fallas, sin infección y alta mortalidad predicha, en este caso el modelo gráfico estima un valor de 0.333, mientras que el logístico de 0.216.

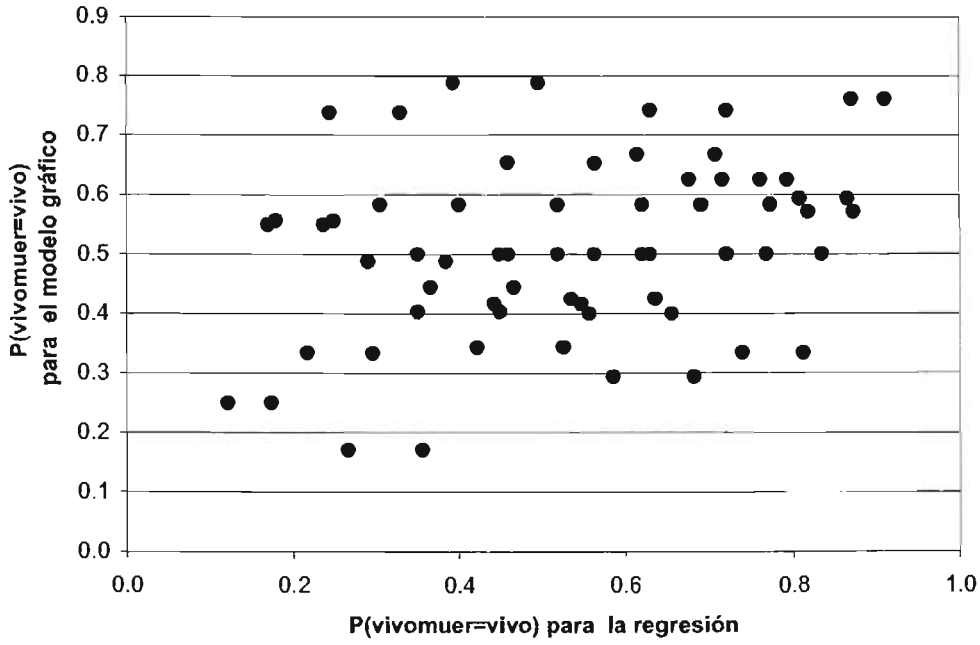

Figura 4.16: Probabilidades estimadas $\hat{P}_{reg.}(vivomuer = vivo)$ vs $\hat{P}_{graf.}(vivomuer = vivo)$

Como en la sección anterior, puede observarse en la tabla 4.9 que hay probabilidades estimadas en el modelo gráfico que se repiten, por ejemplo se tiene que la serie

de valores de la celda 1 a la 8 se repite de la celda 9 a la celda 16, esto es debido a que $fcard \perp vivomuer \mid edad, mortpredcod, infnos, fresp$, esto es la falla cardiaca es independiente condicionalmente a la variable que separa a vivos de muertos dada la edad categorizada, la mortalidad predicha categorizada, la infección noscomial y la falla respiratoria. Se observa que la celda 1 tiene los mismos valores en casi todas las variables que la celda 9, excepto en la falla cardiaca pues en la celda 1 no hay falla cardiaca mientras que en la 8 sí la hay, y las probabilidades estimadas son las mismas, esto es precisamente porque la falla cardiaca es independiente condicionalmente a la variable que separa vivos de muertos, entonces no importa que valor tome la falla cardiaca mientras que el resto de las variables tomen el mismo valor (también la calidad de vida inicial, pues la calidad inicial afecta directamente a la variable *vivomuer*). De forma similar la probabilidad estimada de la celda 2 coincide con la de la celda 10 y así sucesivamente hasta ver que todo el bloque de probabilidades estimadas de la celda 1 a la 8 se repite de la celda 9 a la 16. Similarmente se puede ver que el bloque de la celda 17 a la 24 se repite de la celda 25 a la celda 32, el bloque de la celda 33 a la 40 se repite de la celda 41 a la 48 y el bloque de la celda 49 a la 56 se repite en las últimas 8 celdas, de la 57 a la 64. Así que la presencia de independencias condicionales explica el hecho de la repetición de valores en las probabilidades estimadas.

En la figura 4.16 los puntos se encuentran alrededor de la recta de 45°, sobre todo si se compara con la misma gráfica de la sección anterior (figura 4.6), lo cual indica que las estimaciones entre ambos modelos se acercan entre sí. Hay algunos puntos alejados en la parte superior izquierda que son: la celda 16 (buena c.v. inicial, rango menor de edad, con fallas, con infección nosocomial y alta mortalidad predicha), la 8 (buena c.v. inicial, rango menor de edad, sin falla card., con falla resp., con infección nosocomial y alta mortalidad predicha), la celda 14 (buena c.v. inicial, rango menor de edad, con

fallas, sin infección nosocomial y alta mortalidad predicha), la 6 (buena c.v. inicial, rango menor de edad, sin falla card., con falla resp., sin infección nosocomial y alta mortalidad predicha), la celda 48 (mala c.v. inicial, rango menor de edad, con fallas, con infección nosocomial y alta mortalidad predicha) y la celda 32 (mala c.v. inicial, personas mayores, con fallas, con infección nosocomial y alta mortalidad predicha), todos estos puntos se caracterizan porque la estimación de la probabilidad de vivir para el modelo gráfico es mucho mayor que para la regresión logística, lo que tienen en común es la presencia de falla respiratoria y alta mortalidad predicha; de hecho la diferencia en las estimaciones es debido a que estás variables son las que tienen más peso en la regresión (son las variables con los coeficentes y coeficientes estandarizados más grandes) y entonces al tener falla respiratoria y alta mortalidad al momento de estimar las probabilidades los correspondientes coeficientes toman el valor de cero y como consecuencia disminuye la probabilidad de vivir. Otro grupo de puntos alejados y que se encuentran en la parte inferior derecha son: el correspondiente a la celda 51 (mala c.v. inicial, personas mayores, sin fallas, con infección nosocomial y baja mortalidad predicha), 57 (mala c.v. inicial, personas mayores, con falla card., sin falla resp., sin infección nosocomial y baja mortalidad predicha) y a la celda 49 (mala c.v. inicial, personas mayores, sin fallas, sin infección nosocomial y baja mortalidad predicha), en estos puntos la estimación para la regresión es mayor respecto a la del modelo gráfico, lo que tienen en común es la mala calidad de vida inicial, que son personas mayores, con baja mortalidad predicha y sin falla respiratoria, entonces se comprueba el peso que da el modelo gráfico a una mala calidad de vida inicial y a que la persona sea mayor, pues la probabilidad de vivir disminuye respecto a las probabilidades que se obtienen con la regresión.

Como en la sección anterior, para cada patrón de covariables se realiza una com-

paración entre las probabilidades estimadas de sobrevivencia usando ambos modelos y la frecuencia relativa observada de individuos sobrevivientes. Se realiza el análisis comparativo para aquellas celdas que tuvieron más de 20 individuos.

1) Para la celda 1, con 33 observaciones, se tiene que tanto las estimaciones como la frecuencia relativa son mayores a 0.5, así que en ambos modelos se está asignando a los individuos con los valores en las variables correspondientes a la celda 1 a la categoría vivo y en la base de datos el 96.97 % de los individuos con tales características sobreviven. 2) Para la celda 5, con 46 observaciones, tanto las probabilidades estimadas como la frecuencia relativa se aproximan entre sí con valores alrededor de 0.7, así que en ambos modelos se asigna a los individuos a la categoría vivos que es en la que en la base hay una mayor proporción de individuos. 3) Para la celda 6, con 42 observaciones, la probabilidad estimada bajo el modelo gráfico es mayor a 0.5 (0.787), de igual forma la frecuencia relativa observada es mayor a 0.5 (0.5476); sin embargo, el modelo logístico estima una probabilidad inferior a 0.5, 0.495, así que el modelo gráfico asigna a los individuos a la categoría vivo en la cual se encuentran más de la mitad de los individuos que en la base de datos tienen la combinación de valores de la celda 6 , en cambio la regresión logística asigna a estos individuos a la categoría muerto, esto ocurre porque como ya se dijo las variables de más peso en la regresión son la mortalidad predicha y la falla respiratoria , en esta celda el individuo tiene alta mortalidad predicha y falla respiratoria lo cual ocasiona que la probabilidad estimada de sobrevivir disminuya. 4) Para la celda 13, con 44 observaciones, se tiene que tanto las estimaciones como la frecuencia relativa son valores cercanos entre sí y mayores de 0.5, por lo cual ambos modelos asignan a los individuos con el patrón de covariables de la celda 13 a la categoría vivo, en la cual en la base de datos se encuentran la mayor parte de los individuos (más del 60 %). 5) Para la celda 14, con 45 observaciones, la probabilidad estimada bajo la regresión se aproxima a

la frecuencia relativa correspondiente, la cual indica que una tercera parte de los individuos con este patrón de covariables sobreviven. En cambio el modelo gráfico estima una probabilidad de sobrevivencia mucho más alta, 0.7879, entonces los individuos bajo este modelo se estarían asignando a la categoría vivo cuando en realidad, como se dijo, solo una tercera parte de los individuos con estas carcterísticas sobreviven, el hecho de que la probabilidad de sobrevivencia estimada bajo el modelo gráfico sea tan alta se debe a que los individuos con el patrón de covariables de esta celda corresponden a individuos con buena calidad de vida inicial y en el rango inferior de edad, y como ya se vio antes, en el modelo gráfico estas variables son las de más importancia y entonces al encontrarse bien el individuo en ellas aumenta mucho la probabilidad estimada de sobrevivir. 6) Para la celda 15, con 33 observaciones, la situación es prácticamente la misma que en el caso anterior. 7) Para la celda 16, con 46 observaciones, ocurre prácticamente lo mismo que para la celda anterior descrita (la 14), incluso los valores estimados y la frecuencia relativa son similares a los de la celda ya descrita y nuevamente el hecho de que la probabilidad estimada bajo el modelo gráfico sea alta se debe a que este tipo de individuo tiene buena calidad de vida inicial y se encuentra en el rango de edad inferior. 8) Para la celda 22, con 30 observaciones, se tiene que la calidad de vida inicial es buena y aunque el rango de edad es el superior, de todos modos la probabilidad de sobrevivencia estimada bajo el modelo gráfico no es baja (0.583), al ser mayor a 0.5 a estas observaciones se les asignaría la categoría vivo; sin embargo, menos de la tercera parte de los individuos en la base de datos sobreviven, 26.67 %, valor al cual se acerca más la probabilidad estimada bajo el modelo logístico. 9) La celda 30, con 69 observaciones, tiene las mismas carcterísticas que la celda anterior salvo que se tiene presencia de falla cardiaca, pero ya se vio que bajo el modelo gráfico esta variable no afecta los resultados, así que la probabilidad estimada es la misma (0.583). Nuevamente bajo este modelo se asignan a los individuos a la categoría vivo; sin embargo, solo el 26.09 % de

los individuos con las características de esta celda sobreviven por lo que bajo el modelo gráfico muchos individuos están mal clasificados. Bajo el modelo logístico se tiene una probabilidad estimada más parecida a la frecuencia relativa por lo que los individuos se asignan a la categoría muerto que es donde la mayoría de los individuos se encuentran (el 73.91 %). 10) En la celda 32, con 43 observaciones, la frecuencia relativa observada de sobrevivientes fue de un 30.23 %, la probabilidad estimada bajo el modelo logístico resulta baja comparada con este valor, 0.179, esto se debe a que los individuos tipo la celda 32 tienen alta mortalidad predicha y falla respiratoria, lo cual disminuye mucho la probabilidad estimada al ser las variables de más peso para este modelo; en cambio, bajo el modelo gráfico la probabilidad estimada, 0.555, es mayor que la frecuencia relativa, debido a que otra vez la calidad de vida inicial es buena, así que bajo este modelo se asignarían los individuos a la categoría vivos cuando en realidad aproximadamente 30 % de los individuos en la base pertenecen a esta categoía, por otra parte en el modelo logístico los individuos se asignan a la categoría muerto donde se encuentran la mayoría de los individuos de la base. 11) Para la celda 54, con 24 individuos, ocurre que tanto las probabilidades estimadas como la frecuencia relativa tienen valores similares, las estimaciones son cercanas a 0.3 por lo que los individuos se asignarían a la categoría muertos y esta es la categoría, con la combinación de valores de esta celda, en la que en la base de datos se encuentran el 75 % de los individuos, así que con ambos modelos los individuos se están clasificando lo mejor posible. 12) La celda 62, con 54 observaciones, es un caso parecido al anterior pues los valores estimados son similares a la frecuencia relativa correspondiente, con valores cercanos a 0.25, así que nuevamente se asignan los individuos a la categoría de muertos en la cual se encuentran la mayoría de los individuos de la base para este patrón de covariables. 13) En la celda 64, con 47 observaciones, se observa que la probabilidad estimada bajo el modelo gráfico, 0.250, es un poco mayor que la frecuencia relativa observada correspondiente, 0.170, y también

se observa que la probabilidad estimada bajo el otro modelo es un poco menor, 0.121; sin embargo, ambos modelos asignan a los individuos a la categoría de muertos en la cual, para esta celda, se encuentran aproximadamente el 83 % de los individuos; así que bajo los modelos se está asignando a los individuos a la categoría con mayor cantidad de individuos en la base, muerto, esto es lógico porque esta celda corresponde al caso en el que el individuo se encuentra en las peores condiciones posibles de acuerdo a las variables que se han utilizado.

En conclusión, para aquellas celdas que tuvieron más de 20 individuos (13 celdas de este tipo), el modelo logístico asigna en la mayoría de las celdas, 11, a las observaciones a la categoría en la que en base de datos hubo mayor número o frecuencia relativa de observaciones; por otra parte, el modelo gráfico asigna en menos celdas, 7, a las observaciones a la categoría en la que en la base de datos hubo mayor frecuencia relativa de observaciones. También se puede considerar todas las celdas y contar en cuántas, bajo cada modelo, se asigna al individuo a la categoría con la frecuencia relativa observada adecuada, así que por ejemplo bajo la regresión se cuentan las celdas en las que si se estiman valores de 0.5 o más la frecuencia relativa también es 0.5 o más y en estas celdas ocurriría que bajo regresión se asignan los individuos a la categoría vivos y en la base de datos la mayoría de estos individuos efectivamente sobreviven, también se cuentan las celdas en las que si la estimación es inferior a 0.5 la frecuencia también es inferior a 0.5. En el caso de la regresión logística 47 de las 64 celdas se asignan adecuadamente, en cambio en el modelo gráfico 34 de las 64 celdas se asignan adecuadamente; sin embargo, hay que tomar en cuenta que ciertas celdas no tienen observaciones y que no sería tan importante considerarlas y hay otras que cuentan con mayor cantidad de observaciones que fueron las que se analizaron arriba. Por otro lado, nuevamente como en la sección anterior, se pueden realizar pruebas ji cuadrada de bondad de ajuste, las

pruebas se hicieron para aquellas celdas con más de 20 observaciones. Bajo la regresión logística la estadística de prueba T=8.702 (Apéndice, Tabla B.4), este valor se compara con $\chi^{12}_{0,95}$=21.026 (utilizando α=0.05), entonces como T no es mayor a este valor no se rechaza la hipótesis nula de que la distribución que tienen los datos es la distribución que se obtiene bajo regresión logística; en cambio, bajo el modelo gráfico T=34.301 (Apéndice, Tabla B.5), entonces usando el mismo nivel de significancia (e incluso niveles mucho más pequeños) se observa que la estadística de prueba es mayor que el cuantil correspondiente y se rechaza la hipótesis nula de que la distribución bajo el modelo gráfico sea siempre similar a la distribución que tienen los datos.

Debido a que en la regresión la variable de más peso es la falla respiratoria (*fresp*) (coeficiente estimado 1.359, coeficiente de Wald 35.873 y coeficiente relativo 0.515) seguida de la mortalidad predicha (*mortpredcod*) (coeficiente estimado 0.968, coeficiente de Wald 35.959 y coefiente estandarizado 0.4141), lo cual de hecho ya surgió en análisis anteriores, se puede ordenar la tabla considerando en primer lugar la falla respiratoria (*fresp*), luego la mortalidad predicha (*mortpredcod*), la calidad de vida inicial (*cv1cod*), la edad (*edadcod*), la falla cardiaca (*fcard*) y la infección nosocomial (*infnos*), se presentan las probabildades estimadas (tabla 4.10) y la gráfica correspondiente de las probabildades estimadas (figura 4.17). Claramente, en el modelo logístico hay una tendencia a la baja en las probabilidades, lo cual no se observa en el modelo gráfico. En las primeras 16 celdas, que es cuando la mortalidad predicha es baja y no hay falla respiratoria, las brechas entre las estimaciones con ambos modelos son grandes y lo mismo ocurre con las últimas 16 celdas que es cuando la mortalidad predicha es alta y hay falla respiratoria; en el primer caso las probabilidades estimadas con el modelo gráfico están por debajo de las obtenidas con el modelo logístico y en el otro caso están por arriba, el resto de los puntos presentan estimaciones similares entre sí. Entonces,

otra vez se comprueba como mortalidad predicha y falla respiratoria afectan bastante las estimaciones en la regresión y no tanto las del modelo gráfico, pues cuando la falla esta ausente y simultáneamente hay baja mortalidad predicha las probabilidades para la regresión aumentan considerablemente respecto a las del modelo gráfico y cuando hay falla y hay alta mortalidad predicha la probabilidad de vivir disminuye considerablemente respecto a los valores estimados con el modelo gráfico. Sin embargo, en el modelo gráfico las variables que influyen más son la calidad de vida inicial y la edad, como se puede observar por ejemplo en las celdas 13 a la 16 que son individuos mayores y con mala calidad de vida inicial cuyas probabilidades disminuyen bastante.

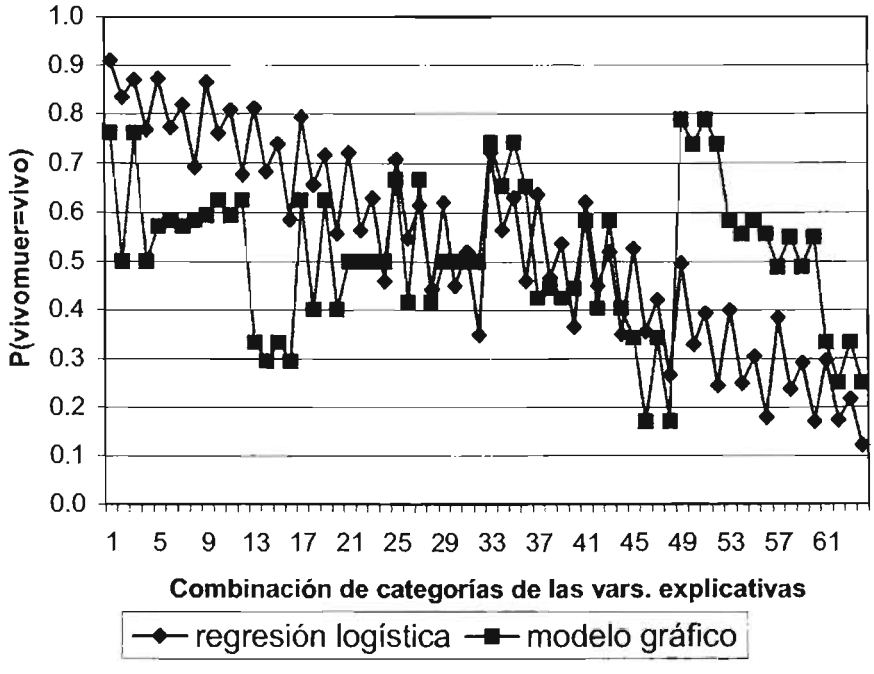

Figura 4.17: Probabilidades estimadas, $\hat{P}(vivomuer = vivo)$, bajo una regresión logística y bajo un modelo gráfico (orden 2)

Finalmente, como en la sección anterior se puede hacer una tabla de clasificación para el modelo gráfico (tabla 4.11), resulta que el 72.27 % de las personas que sobrevivieron a la terapia intensiva (en la base de datos) son clasificados correctamente de acuerdo al modelo, por otra parte, el 45.68 % de los individuos que murieron están bien clasificados, además, se tiene que un 57.60 % del total de individuos se encuentran correctamente asignados. Al comparar con los resultados que se obtienen con la regresión logística (subtabla 1 de la tabla 4.8) resulta que en la regresión logística los vivos no están tan bien clasificados como con el modelo gráfico, puesto que con la regresión el 57.80 % de los individuos son clasificados correctamente; sin embargo, los muertos en la regresión están mejor clasificados que con el modelo gráfico pues resulta que un 81.50 % de los individuos son asignados correctamente. Entonces se observa que en el caso del modelo logístico la asignación a muertos es más certera mientras que para el modelo gráfico la asignación más certera corresponde a los vivos.

celda	cv1 cod.	edad cod.	falla card.	falla resp.	inf. nos.	mort. pred.	$\hat{P}_{reg.}$	$\hat{P}_{graf.}$	frec. rel. vivos	obs.	$\lvert\hat{P}_{reg.} - \hat{P}_{graf.}\rvert$
1	1	1	0	0	0	0	0.9097	0.7612	0.9697	33	0.1485
2	1	1	0	0	1	0	0.8340	0.5000	1.0000	5	0.3340
3	1	1	1	0	0	0	0.8690	0.7612	0.9375	16	0.1078
4	1	1	1	0	1	0	0.7678	0.5000	1.0000	1	0.2678
5	1	2	0	0	0	0	0.8719	0.5714	0.8333	6	0.3005
6	1	2	0	0	1	0	0.7724	0.5833	0.5000	4	0.1891
7	1	2	1	0	0	0	0.8176	0.5714	0.8750	8	0.2462
8	1	2	1	0	1	0	0.6908	0.5833	0.0000	0	0.1075
9	2	1	0	0	0	0	0.8644	0.5938	0.7778	9	0.2706
10	2	1	0	0	1	0	0.7606	0.6250	0.0000	0	0.1356
11	2	1	1	0	0	0	0.8075	0.5938	0.5000	2	0.2137
12	2	1	1	0	1	0	0.6766	0.6250	0.0000	0	0.0516
13	2	2	0	0	0	0	0.8115	0.3333	1.0000	2	0.4782
14	2	2	0	0	1	0	0.6822	0.2941	0.0000	0	0.3881
15	2	2	1	0	0	0	0.7392	0.3333	0.7500	4	0.4059
16	2	2	1	0	1	0	0.5856	0.2941	0.0000	0	0.2915
17	1	1	0	0	0	1	0.7928	0.6250	0.7333	15	0.1678
18	1	1	0	0	1	1	0.6561	0.4000	0.6667	3	0.2561
19	1	1	1	0	0	1	0.7159	0.6250	0.6667	6	0.0909
20	1	1	1	0	1	1	0.5568	0.4000	1.0000	2	0.1568
21	1	2	0	0	0	1	0.7211	0.5000	0.7500	8	0.2211
22	1	2	0	0	1	1	0.5632	0.5000	0.0000	1	0.0632
23	1	2	1	0	0	1	0.6299	0.5000	1.0000	3	0.1299
24	1	2	1	0	1	1	0.4591	0.5000	0.0000	2	0.0409
25	2	1	0	0	0	1	0.7077	0.6667	0.0000	1	0.0410
26	2	1	0	0	1	1	0.5469	0.4167	0.0000	0	0.1302
27	2	1	1	0	0	1	0.6144	0.6667	1.0000	1	0.0523
28	2	1	1	0	1	1	0.4428	0.4167	0.5000	2	0.0261
29	2	2	0	0	0	1	0.6206	0.5000	0.5000	6	0.1206
30	2	2	0	0	1	1	0.4492	0.5000	0.0000	1	0.0508
31	2	2	1	0	0	1	0.5185	0.5000	0.3750	8	0.0185
32	2	2	1	0	1	1	0.3493	0.5000	1.0000	1	0.1507
33	1	1	0	1	0	0	0.7213	0.7416	0.7609	46	0.0203
34	1	1	0	1	1	0	0.5634	0.6532	0.6250	16	0.0898
35	1	1	1	1	0	0	0.6302	0.7416	0.6136	44	0.1114
36	1	1	1	1	1	0	0.4593	0.6531	0.3636	33	0.1938
37	1	2	0	1	0	0	0.6362	0.4242	0.6923	13	0.2120
38	1	2	0	1	1	0	0.4658	0.4444	0.3333	3	0.0214
39	1	2	1	1	0	0	0.5352	0.4242	0.5263	19	0.1110
40	1	2	1	1	1	0	0.3647	0.4444	0.2857	7	0.0797
41	2	1	0	1	0	0	0.6208	0.5833	0.7143	7	0.0375
42	2	1	0	1	1	0	0.4494	0.4031	0.6667	3	0.0463
43	2	1	1	1	0	0	0.5187	0.5833	0.4545	11	0.0646
44	2	1	1	1	1	0	0.3496	0.4031	0.0909	11	0.0535
45	2	2	0	1	0	0	0.5252	0.3429	0.6364	11	0.1823
46	2	2	0	1	1	0	0.3555	0.1711	0.0000	1	0.1844
47	2	2	1	1	0	0	0.4214	0.3429	0.5333	15	0.0785
48	2	2	1	1	1	0	0.2664	0.1711	0.1250	8	0.0953
49	1	1	0	1	0	1	0.4958	0.7879	0.5476	42	0.2921
50	1	1	0	1	1	1	0.3289	0.7381	0.1875	16	0.4092
51	1	1	1	1	0	1	0.3929	0.7879	0.3333	45	0.3950
52	1	1	1	1	1	1	0.2440	0.7381	0.2826	46	0.4941
53	1	2	0	1	0	1	0.3992	0.5833	0.2667	30	0.1841
54	1	2	0	1	1	1	0.2488	0.5556	0.3636	11	0.3068
55	1	2	1	1	0	1	0.3043	0.5833	0.2609	69	0.2790
56	1	2	1	1	1	1	0.1790	0.5556	0.3023	43	0.3766
57	2	1	0	1	0	1	0.3834	0.4878	0.6667	6	0.1044
58	2	1	0	1	1	1	0.2367	0.5495	0.0000	1	0.3128
59	2	1	1	1	0	1	0.2905	0.4878	0.3636	11	0.1973
60	2	1	1	1	1	1	0.1695	0.5495	0.1538	13	0.3800
61	2	2	0	1	0	1	0.2959	0.3333	0.2500	24	0.0374
62	2	2	0	1	1	1	0.1732	0.2500	0.0000	5	0.0768
63	2	2	1	1	0	1	0.2167	0.3333	0.2407	54	0.1166
64	2	2	1	1	1	1	0.1212	0.2500	0.1702	47	0.1288

Categorías o celdas ordenadas por *fresp, mortpredcod, cv1cod, edadcod, fcard* e *infnos*.

Tabla 4.10: Tabla de probabilidades estimadas, $\hat{P}(vivomuer = vivo)$, bajo regresión logística y bajo el modelo gráfico (orden 2)

Calidad de vida posterior				
		predicha		
		muerto	vivo	Porcentaje correcto
observada	muerto	217	258	0.4568
	vivo	107	279	0.7227
	total			0.5760

Tabla 4.11: Tabla de clasificación para el estado vital, *vivomuer*, usando el modelo gráfico

4.3. Modelo para la variable que identifica vivos con buena calidad de vida, vivos con mala calidad de vida y muertos

Como ya se mencionó en el capítulo anterior la red definitiva con la que se trabaja contiene las variables: edad categorizada en dos grupos (*edadcod60*), calidad de vida inicial (*cv1cod*), mortalidad predicha (*mortpred*), infección nosocomial (*infnos*), la variable que identifica sobrevivientes con buena c.v., con mala c.v. y los muertos (*cv2vivom*), falla neurológica (*fneur*), falla respiratoria (*fresp*) y cirugía urgente (*cirugia* o *qxurgent*), dentro de estas variables mortalidad predicha es una variable continua y a la variable *cv2vivom* se le va a considerar también como una variable continua, aunque en realidad como ya se ha dicho es una variable tricotómica que debería tomarse como tal pero que por razones técnicas se considera como continua. Así que este es un ejemplo de una red mixta, pues contiene tanto nodos continuos como discretos. Nuevamente, como en los modelos de las secciones anteriores, con ayuda de los médicos se trabajaron las restricciones en los arcos entre las variables hasta llegar a un modelo definitivo, que se muestra en la figura 4.18.

A partir de la red obtenida se tienen los siguientes resultados: a la variable relacionada con la buena o mala calidad de vida en los sobrevivientes y a los muertos (es decir a *cv2vivom*) ingresan de forma directa todas las otras variables a excepción de las variables correspondientes a cirugía urgente (*cirugia*) y grupo de edad (*edad*); sin embargo, la variable relativa a la edad afecta de manera indirecta a *cv2vivom* a través de las variables *cv1cod, mortpred* y *fres* las cuales ingresan de manera directa a la variable *cv2vivom*, por otro lado la variable correspondiente a cirugía urgente (*cirugia*)

ingresa indirectamente a *cv2vivom* a través de la presencia o no de infección nosocomial (*infnos*) y a la falla respiratoria (*fres*). Además, se observa que la variable calidad de vida inicial (*cv1cod*) incide de manera directa sobre la variable (*cirugia*). A la variable correspondiente a la mortalidad predicha (*mortpred*) ingresan las variables *edad, fresp, fneur* e *infnos* (es decir, la edad, la falla respiratoria, la falla neurológica y la infección nosocomial respectivamente), a esta última variable ingresan a su vez las variables *fneur, fresp* y *cirugia* (falla neurológica, falla respiratoria y cirugía urgente), también se tiene que no ingresa ninguna variable a la falla neurológica (*fneur*) sino que más bien solo salen arcos (hacia *mortpred, infnos* y *cv2vivom* como ya se mencionó), por otra parte, a la variable falla respiratoria (*fresp*) ingresan las variables cirugía urgente (*cirugia*) y la edad categorizada (*edad*), a la variable *cirugia* solo ingresa la variable calidad de vida inicial (*cv1cod*) y finalmente *edad* afecta únicamente a las variables *cv1cod, mortpred* y *fresp* (calidad de vida inicial, mortalidad predicha y falla respiratoria) sin que ninguna variable pueda ingresar a la misma.

También a través de la red se observan las independencias condicionales siguientes.

1. Para la variable de interés (*cv2vivom*) se obtiene que la edad categorizada (*edad*) es condicionalmente independiente de *cv2vivom* dadas la calidad de vida inicial (*cv1cod*), la mortalidad predicha (*mortpred*) y la falla respiratoria (*fresp*), es decir, $edad \perp cv2vivom | cv1cod, mortpred, fresp$.

2. También se tiene que la cirugía urgente es condicionalmente independiente a la variable que separa vivos con buena y mala c.v. de los muertos dadas la calidad de vida inicial, la infección nosocomial y la falla respiratoria, $cirugia \perp cv2vivom | cv1cod, infnos, fresp$.

3. La edad categorizada es condicionalmente independiente de la cirugía urgente

dadas la calidad de vida inicial, la falla respiratoria y la infección nosocomial, *edad*⊥ *cirugia*| *cv1cod, fresp, infnos.* Las mismas variables también son condicionalmente independientes dadas la calidad de vida inicial, la mortalidad predicha y la falla respiratoria, *edad*⊥ *cirugia*| *cv1cod, mortpred, fresp.*

4. La edad categorizada es condicionalmente independiente de la falla neurológica dadas la calidad de vida inicial, la mortalidad predicha y la falla respiratoria, *edad*⊥ *fneur*| *cv1cod,mortpred,fresp.* También ambas variables son condicionalmente independientes dadas la mortalidad predicha, la infección nosocomial y la variable que separa vivos con buena y mala c.v. de los muertos, *edad*⊥ *fneur*| *mortpred, infnos, cv2vivom.*

5. La variable edad categorizada es condicionalmente independiente de la infección nosocomial dadas la calidad de vida inicial, la mortalidad predicha y la falla respiratoria, *edad*⊥ *infnos*| *cv1cod,mortpred,fresp.*

6. La calidad de vida inicial es condicionalmente independiente de la mortalidad predicha dadas la cirugía urgente, la variable que separa vivos con buena y mala c.v. de los muertos y la edad dicotomizada, *cv1cod*⊥ *mortpred*| *cirugia,cv2vivom, edad.*

7. La calidad de vida inicial es condicionalmente independiente de la infección nosocomial dadas la cirugía urgente, la variable que separa vivos con buena y mala c.v. de los muertos y la edad categorizada, *cv1cod*⊥ *infnos*| *cirugia,cv2vivom,edad.*

8. La calidad de vida inicial es condicionalmente independiente de la falla neurológica dadas la cirugía urgente, la variable que separa vivos con buena y mala c.v. de los muertos y la edad dicotomizada, *cv1cod*⊥ *fneur*| *cirugia,cv2vivom,edad.* También son condicionalmente independientes dadas la edad, la cirugía urgente y la variable

que separa vivos con buena y mala c.v. de los muertos, *cv1cod*⊥ *fneur*| *edad, cirugia, cv2vivom.*

9. La calidad de vida inicial es condicionalmente independiente de la falla respiratoria dadas la cirugía urgente, la variable que separa vivos con buena y mala c.v. de los muertos y la edad categorizada, *cv1cod*⊥ *fresp*| *cirugia, cv2vivom, edad.*

10. La mortalidad predicha es condicionalmente independiente de la cirugía urgente dadas la falla respiratoria, la calidad de vida inicial y la infección nosocomial, *mortpred*⊥ *cirugia*| *fresp, cv1cod, infnos.*

11. La falla neurológica es condicionalmente independiente de la falla respiratoria dadas la mortalidad predicha, la infección nosocomial y la variable que separa vivos con buena y mala c.v. de los muertos, *fneur*⊥ *fresp*| *mortpred, infnos, cv2vivom.*

12. Finalmente, la falla neurológica es condicionalmente independiente de la cirugía urgente dadas la infección nosocomial, la falla respiratoria y la calidad de vida inicial, *fneur*⊥ *cirugia*| *infnos, fresp, cv1cod.* También son condicionalmente independientes dadas la mortalidad predicha, la infección nosocomial y la variable que separa vivos con buena y mala c.v. de los muertos, *fneur*⊥ *cirugia*| *mortpred, infnos, cv2vivom.*

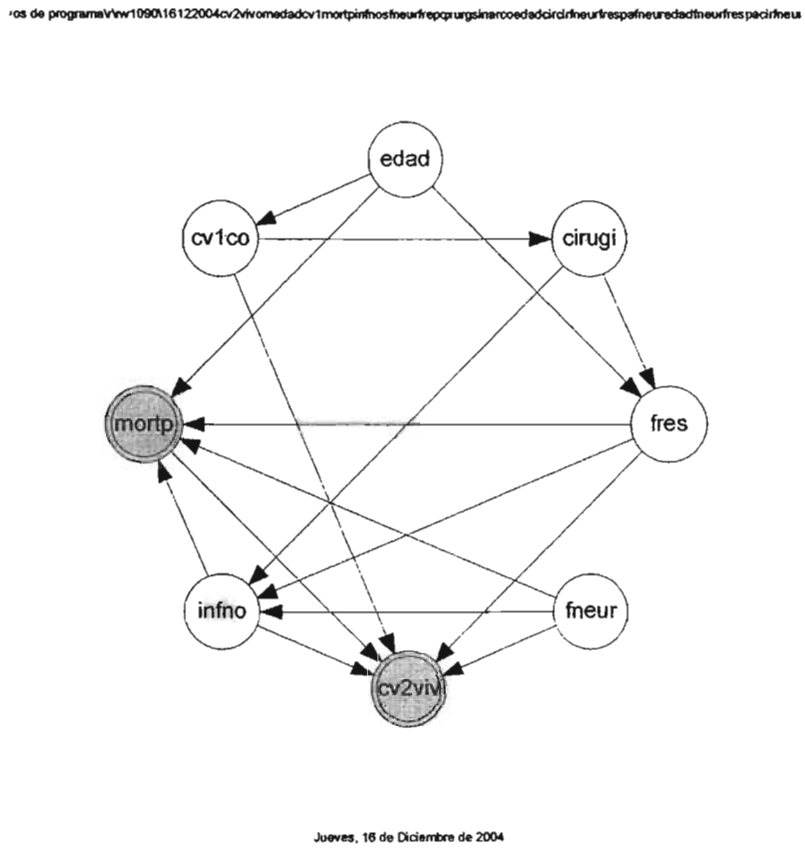

Figura 4.18: Modelo gráfico que involucra la variable que separa vivos con buena y mala calidad de vida de los muertos (*cv2vivom*)

Se ajustó una regresión logística trinomial, cuya variable respuesta es *cv2vivom* (se consideró un modelo logístico con variable respuesta no ordenada), y con categoría de referencia a las personas vivas con buena calidad de vida, los resultados obtenidos se presentan en la tabla 4.12. Al ver la prueba para determinar si es importante introducir cada una de las variables (primer tabulado) se observa como cada una de las variables es significativa rechazando la hipótesis de que los parámetros estimados para cada una de las variables sea cero.

En el siguiente rubro de la tabla 4.12 se presenta cada uno de los coeficientes; para cuando se comparan los muertos contra la categoría de vivos con buena calidad de vida posterior, se tiene que todos los coeficientes son estadísticamente distintos de cero. Para la mortalidad predicha el coeficiente es positivo pero muy pequeño y de hecho su exponencial es casi igual a uno indicando que al incrementar la mortalidad predicha una unidad aumenta ligeramente la probabilidad o riesgo de pasar de la categoría de buena c.v. a muerto o que de hecho hay independencia y el riesgo de tener buena c.v. o estar muerto es similar a pesar de incrementar el valor de la mortalidad predicha. Las otras variables son categóricas y toman el valor de cero para la categoría más grande que corresponde en todos los casos a cuando el individuo está peor, así para la calidad de vida inicial se tiene un coeficiente de -1.707, cuya exponencial es 0.181 y que indicaría que cuando la calidad de vida inicial del individuo es buena disminuye el riesgo de pasar de la categoría de buena calidad de vida posterior a la categoría de muerto, o sea que es más probable que el individuo tenga buena calidad de vida posterior. De manera similar las exponenciales para todas las otras variables son menores a uno (incluso en los intervalos de confianza correspondientes), entonces se obtiene que cuando el individuo no tiene infección nosocomial hay menor riesgo de morir que de tener buena calidad de vida, de forma parecida cuando no hay falla neurológica, cuando no hay falla

respiratoria, cuando no hay cirugía urgente y cuando el individuo pertenece al grupo de edad más joven hay menor riesgo de morir que de tener buena calidad de vida. Para cuando se comparan la categoría de vivos con mala calidad de vida posterior contra los vivos con buena calidad de vida posterior resulta que las variables mortalidad predicha e infección nosocomial no son significativas; en el caso de mortalidad predicha es lógico pues la mortalidad predicha es un índicador que distingue vivos de muertos y no buena de mala calidad de vida; en cuanto a las otras variables todas ellas son significativas, con coeficientes negativos y exponenciales menores a uno lo cual significa que cuando el individuo tiene buena c.v. inicial, no tiene falla respiratoria, no tiene falla neurológica, no requirió de cirugía urgente o su rango de edad es inferior hay menor riesgo de tener mala calidad de vida posterior que de tener buena, o sea que como la lógica lo indica hay más posibilidad de tener una buena calidad de vida.

Para analizar la calidad del ajuste del modelo logístico, nuevamente se analiza la devianza residual, en este caso es de 1316.88 (tabla 4.12 subtabla 4), el cual es un valor pequeño comparado con un cuantil para una ji cuadrada con 1572 grados de libertad para un nivel de significancia de por ejemplo 0.05; de hecho el nivel crítico correspondiente es de prácticamente 1, por lo que no se rechaza la hipótesis nula de que la diferencia entre el modelo ajustado y el saturado es pequeña y entonces este modelo ajusta bien a los datos. De hecho, lo anterior se corrabora con la estadística de Pearson con un nivel crítico de 0.374 que también indica que el modelo ajustado es adecuado, entonces estadísticamente se observa que no hay necesidad de incluir interacciones en el modelo y que se tiene un buen modelo. Posteriormente (tabla 4.12 subtabla 5), se observa que el nivel crítico de la prueba que compara el modelo nulo con el ajustado es de aproximadamente cero por lo que en general se rechaza la hipótesis de que las variables incluidas en el modelo ajustado no ayudan a explicar a la variable respuesta.

Finalmente, se observa que las pseudo R cuadradas estimadas (tabla 4.12 subtabla 6) se encuentran entre 0.181 y 0.351, entonces entre el 18.10 % y el 35.10 % de la variabilidad es explicada por el modelo ajustado. En conclusión, el modelo ajustado es adecuado para los datos.

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	1361.721	.000	0	.
MORTPRED	1390.240	28.518	2	.000
CV1CODIF	1408.443	46.722	2	.000
INFNOSOC	1385.741	24.020	2	.000
DIASFRES	1409.895	48.173	2	.000
DIASFNEU	1389.469	27.747	2	.000
QXURGENT	1370.678	8.957	2	.011
EDADCOD6	1378.426	16.705	2	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Parameter Estimates

CV2CODBI		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
muerto vs bueno	Intercept	5.143	.607	71.867	1	.000			
	MORTPRED	2.257E-02	.006	12.990	1	.000	1.023	1.010	1.035
	[CV1COD=1]	-1.707	.299	32.652	1	.000	.181	.101	.326
	[CV1COD=2]	0[a]	.	.	0	.	.	.	.
	[INFNOS=.00]	-1.006	.234	18.484	1	.000	.366	.231	.579
	[INFNOS=1.00]	0[a]	.	.	0	.	.	.	.
	[FRESP=0]	-1.771	.267	43.863	1	.000	.170	.101	.287
	[FRESP=1]	0[a]	.	.	0	.	.	.	.
	[FNEUR=0]	-1.736	.381	20.720	1	.000	.176	8.340E-02	.372
	[FNEUR=1]	0[a]	.	.	0	.	.	.	.
	[QXURGENT=0]	-.548	.230	5.658	1	.017	.578	.368	.908
	[QXURGENT=1]	0[a]	.	.	0	.	.	.	.
	[EDADCOD60=1]	-.889	.222	15.985	1	.000	.411	.266	.636
	[EDADCOD60=2]	0[a]	.	.	0	.	.	.	.
malo vs bueno	Intercept	3.878	.660	34.515	1	.000			
	MORTPRED	-1.75E-03	.007	.059	1	.809	.998	.984	1.012
	[CV1COD=1]	-1.752	.309	32.148	1	.000	.173	9.462E-02	.318
	[CV1COD=2]	0[a]	.	.	0	.	.	.	.
	[INFNOS=.00]	-.353	.264	1.787	1	.181	.703	.419	1.179
	[INFNOS=1.00]	0[a]	.	.	0	.	.	.	.
	[FRESP=0]	-.809	.256	9.971	1	.002	.445	.269	.736
	[FRESP=1]	0[a]	.	.	0	.	.	.	.
	[FNEUR=0]	-1.126	.422	7.122	1	.008	.324	.142	.741
	[FNEUR=1]	0[a]	.	.	0	.	.	.	.
	[QXURGENT=0]	-.721	.248	8.425	1	.004	.486	.299	.791
	[QXURGENT=1]	0[a]	.	.	0	.	.	.	.
	[EDADCOD60=1]	-.703	.242	8.442	1	.004	.495	.308	.795
	[EDADCOD60=2]	0[a]	.	.	0	.	.	.	.

a. This parameter is set to zero because it is redundant.

Classification

Observed	Predicted			
	muerto	malo	bueno	Percent Correct
muerto	424	9	42	89.3%
malo	135	13	40	6.9%
bueno	81	2	115	58.1%
Overall Percentage	74.3%	2.8%	22.9%	64.1%

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	1589.385	1572	.374
Deviance	1316.889	1572	1.000

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1677.250	1686.767	1673.250			
Final	1393.721	1469.851	1361.721	311.529	14	.000

Pseudo R-Square

Cox and Snell	.304
Nagelkerke	.351
McFadden	.181

Tabla 4.12: Coeficientes estimados, tabla de clasificación y tablas para ver la calidad del ajuste para la regresión con variable respuesta *cv2vivomuer*, que separa vivos con buena y mala c.v. de los muertos

Nuevamente, se compara la regresión logística con un modelo gráfico que solo permite la existencia de arcos a la variable que separa a los sobrevivientes con buena calidad de vida, de aquellos con mala calidad de vida y de los muertos (*cv2vivom*) (figura 4.19) y en el cual se observa que a diferencia de la regresión logística las variables que no afectan de manera directa a la respuesta son las variables correspondientes

a cirugía urgente y a la edad codificada en dos grupos, pero que ya se vio en la red anterior que bajo otras restricciones sí afectan la respuesta aunque de manera indirecta.

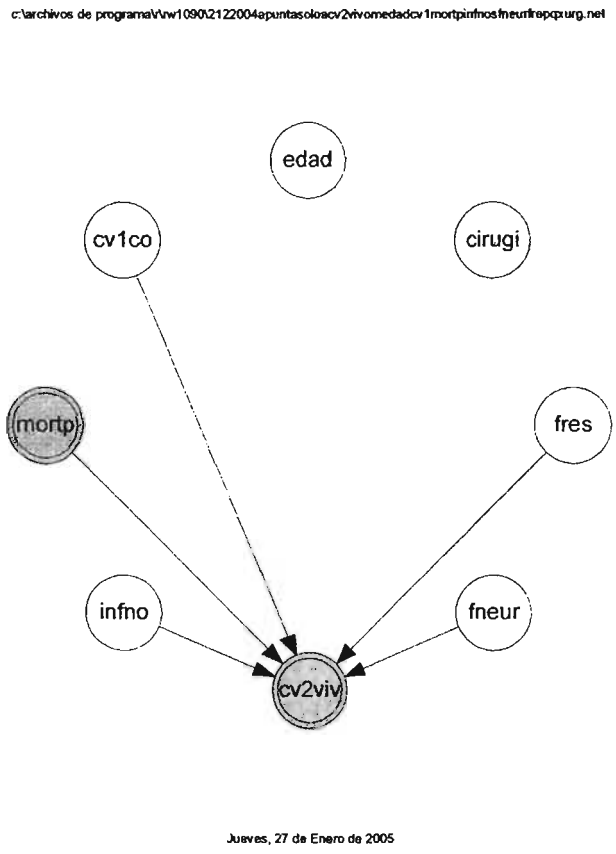

Figura 4.19: Modelo gráfico que solo permite arcos que inciden a *cv2vivomuer*

En capítulos anteriores, se mencionó que el hecho de que haya un grupo de variables correlacionadas entre sí se ve reflejado en las redes mediante la formación de *cliques* entre las variables correlacionadas, para ilustrar esto se toma una red formada por las mismas variables que ya se han manejado en esta sección pero agregando las variables

calificación fisiológica aguda en el primer día (*aps1*) y la calificación de la gravedad de la enfermedad (*apacheii*), que como ya se dijo anteriormente son variables continuas y que están correlacionadas entre ellas y a su vez con las variable mortalidad predicha. Además, a cada una de estas variables agregadas (*aps1* y *apacheii*) se les aplican las mismas restricciones que se emplearon en la gráfica de la figura 4.18 para la variable *mortpred*, como resultado se obtiene una nueva red (figura 4.20) en la cual se observa, como era de esperarse, que las tres variables correlacionadas (*aps1*, *apacheii* y *mortpred*) se encuentran unidas entre sí, de tal manera que ignorando las direcciones de los arcos las aristas resultantes forman un "triángulo"que une los tres nodos entre sí y esto es justamente un *clique* con tres nodos.

En las secciones anteriores se estimó la probabilidad de éxito (que correspondía a la probabilidad de vivir con buena calidad de vida en **4.1** y la probabilidad de vivir en **4.2**) para cada una de las combinaciones de valores que podían tomar las variables explicativas y después se comparaban los resultados entre sí; en este caso no se puede hacer lo mismo por varias razones: en primer lugar la variable mortalidad predicha (*mortpred*) es continua por lo cual no toma un rango finito de valores así que no se pueden obtener todas las estimaciones para todas las combinaciones de valores posibles de las variables explicativas, pues en teoría sería un número infinito de posibilidades (en la práctica sería un número finito muy grande); en segundo lugar, las estimaciones que se obtienen con la regresión logística son probabilidades que se asignan a cada una de las categorías de la variable respuesta que es aquella que separa a los sobrevivientes con buena calidad de vida, de aquellos con mala calidad de vida y de los muertos (*cv2vivom*) mientras que para el modelo gráfico al considerar a *cv2vivom* como variable continua (recordar que se le está considerando continua principalmente por la restricción de que no puede haber padres continuos de nodos discretos y como *mortpred* es continua no podría haber arcos

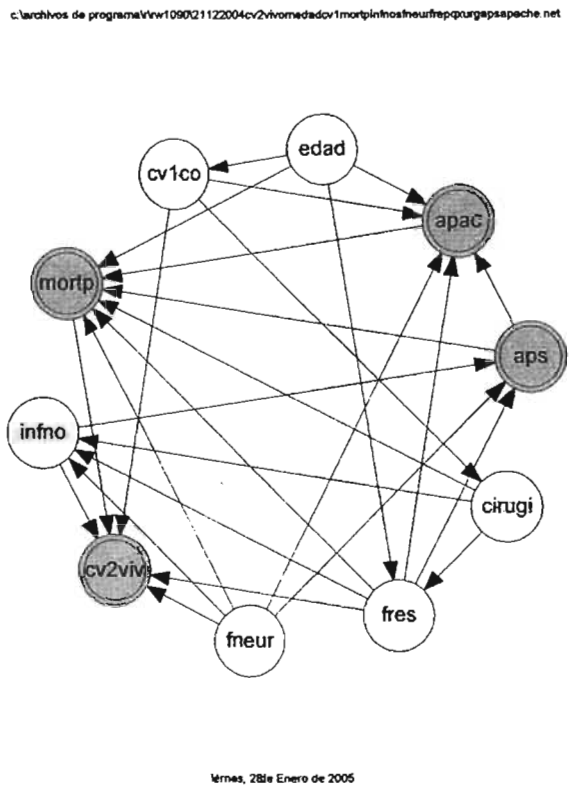

Figura 4.20: Modelo gráfico que ilustra la presencia de un *clique* formado por las variables *aps1*, *apacheii* y *mortpred*

entre estas variables) lo que se obtienen no son probabilidades, sino un valor continuo estimado que de hecho puede considerarse como un intervalo pues *Hugin* proporciona la desviación estándar de la estimación correspondiente (en realidad hay que recordar que para nodos continuos lo que se obtiene es la media y la desviación estándar estimadas de una distribución Gaussiana). Por lo tanto la comparación entre los modelos no se puede llevar a cabo como antes; sin embargo, se va a presentar cómo estima *Hugin* los valores para *cv2vivom* para algunas cantidades que toman las variables explicativas.

Las categorías de cada una de las variables discretas en la red son las mismas que ya se manejaron en las dos secciones anteriores y las cuales se pueden obtener en la tabla 4.1, así que al proporcionar las categorías a las que pertenece el individuo ya se sabe que valor numérico toman la variables. Para la variable respuesta *cv2vivom* se tienen tres categorías: la categoría que toma el valor 1 corresponde a vivos con buena calidad de vida posterior, la 2 corresponde a vivos con mala calidad de vida posterior y la 3 corresponde a los muertos. Supongamos que se tienen a un individuo sin cirugia urgente, con buena calidad de vida inicial, en el rango de edad correspondiente a personas de 60 o menos años, sin falla neuronal ni falla respiratoria y con una mortalidad predicha de 5 (5%), en primer lugar se *compila* la red, posteriormente se incorpora la evidencia de tal forma que, a la categoría en la que el individuo se encuentra en las variables discretas se le da probabilidad de uno y en el caso de la variable continua, *mortpred*, se introduce el valor deseado (en este caso 5) como la media de esa variable y en *Hugin* se asignó el valor cero a la desviación estándar. A continuación se propaga la evidencia y se obtienen los resultados de la figura 4.21, en la cual la media estimada para *cv2vivom* es de 1.378 y la desviación estándar estimada es 0.647, se observa que el valor estimado para *cv2vivom* se encuentra más próximo a uno pues la media estimada se acerca más a este valor, aunque al considerar una desviación estándar los valores estimados estarían entre 0.731 y 2.025 alcanzando al valor dos, pero apenas lo alcanza; en cualquier caso, para este individuo hay más posibilidad de que se encuentre en la categoría 1 de la variable *cv2vivom* la cual corresponde a que su calidad de vida posterior a su estancia en la UTI sea buena. En la figura 4.21, también se observa que *Hugin* permite ver para los nodos continuos la distribución Gaussiana estimada para un nodo continuo de interés (en este caso para *cv2vivom*), hay que notar que *Hugin* permite ver de manera prestablecida la distribución estimada con dos desviaciones estándar.

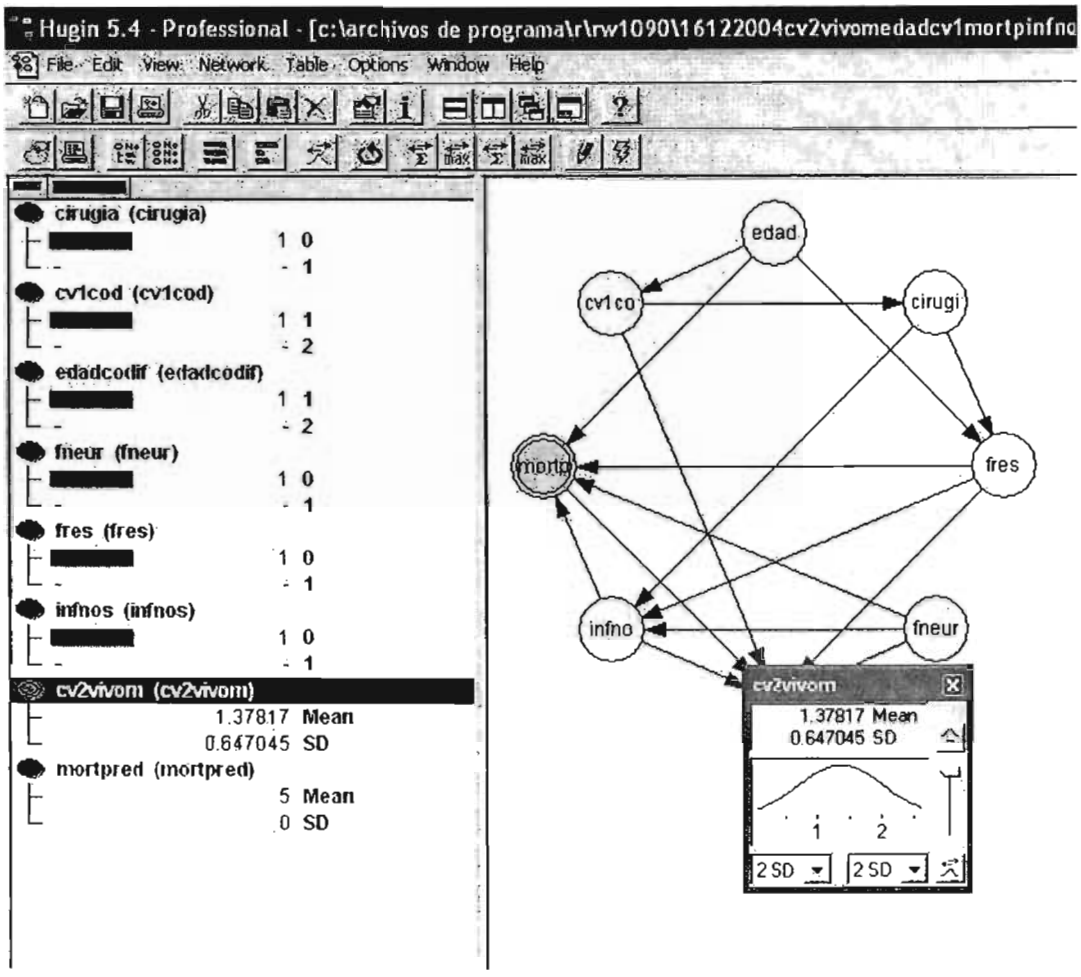

Figura 4.21: Probabilidades marginales para *cv2vivom* una vez introducida la evidencia

Usando la regresión logística descrita, se pueden estimar las probabilidades para el mismo individuo en cada una de las categorías de la variable respuesta *cv2vivom*, como se consideró un modelo logístico con la variable respuesta no ordenada, su forma es la siguiente (Agresti, 2002, cap. 7):

$$\log\left(\frac{P(Y=j)}{P(Y=0)}\right) = \alpha_j + \beta_j'\mathbf{x}, j = 1\ldots k. \tag{4.2}$$

donde la categoría en la que Y toma el valor cero corresponde a la categoría de referencia (en este caso la calidad de vida posterior buena) y las categorías 1 hasta k son las categorías restantes (en este caso solo son dos: calidad de vida posterior mala y muertos) y $\mathbf{x}$ es un vector formado por los valores que toman las variables explicativas. Entonces el modelo ajustado (basándose en la tabla 4.12) estaría dado por las ecuaciones siguientes:

$\log\left(\frac{\hat{P}(cv2vivom=muerto)}{\hat{P}(cv2vivom=buenac.v.)}\right)$ =5.14+ 0.02 ($mortpred$) -1.70 ($cv1cod = buena$) -1 ($infnos = no$) -1.77 ($fresp = no$) -1.73 ($fneur = no$) -0.54 ($cirugia = no$) -0.88 ($edadcod60 \leq 60$),

$\log\left(\frac{\hat{P}(cv2vivom=malac.v.)}{\hat{P}(cv2vivom=buenac.v.)}\right)$ =3.87+ 0.001 ($mortpred$) -1.75 ($cv1cod = buena$) -0.35 ($infnos = no$) -0.80 ($fresp = no$) -1.12 ($fneur = no$) -0.72 ($cirugia = no$) -0.70 ($edadcod60 \leq 60$).

Basándose en la ecuación 4.2 y considerando que se debe cumplir que $\sum_{i=0}^{k} P(Y = i) = 1$, se obtienen las estimaciones para cada una de las categorías en la variable respuesta:

$$\hat{P}(Y = 0) = \frac{1}{1 + \sum_{i=1}^{k} \exp(\hat{\alpha}_i + \hat{\beta}_i'\mathbf{x})},$$

$$\hat{P}(Y = j) = \frac{\exp(\hat{\alpha}_j + \hat{\beta}_j'\mathbf{x})}{1 + \sum_{i=1}^{k} \exp(\hat{\alpha}_i + \hat{\beta}_i'\mathbf{x})}, j = 1 \ldots k.$$

Para el individuo tipo que se ha estado manejando (individuo que no necesitó cirugia urgente, con buena calidad de vida inicial, en el rango de edad correspondiente a personas de 60 o menos años, sin falla neurológica ni falla respiratoria y con una mortalidad predicha de 5) se tiene: $\exp(\hat{\alpha}_{muerto} + \hat{\beta}'_{muerto}\mathbf{x})$ = exp(5.143+0.02257(5)-1.707-1.006-1.771-1.736-0.548-0.889)=0.0906 y $\exp(\hat{\alpha}_{mala} + \hat{\beta}'_{mala}\mathbf{x})$ = exp(3.878-0.001749(5)-1.752-0.353-0.809-1.126-0.721-0.703)=0.2031, de donde para la categoría base:

$$\hat{P}(cv2vivom = buena) = 0.773$$

Y para el resto de las categorías:

$$\hat{P}(cv2vivom = muerto) = 0.070$$

$$\hat{P}(cv2vivom = mala) = 0.157$$

Entonces se observa que en la regresión logística el individuo también tiene más probabilidad de tener buena calidad de vida posterior a su estancia en la UTI, luego de tener mala y finalmente tiene menos probabilidad de morir.

Para ejemplificar nuevamente los resultados que se obtienen con el modelo gráfico, se utiliza otro individuo con sus correspondientes valores en cada una de sus variables, este individuo corresponde a una persona que tiene todas las variables en su contra; es decir, requirió de cirugía urgente, su calidad de vida inicial era mala, su edad corresponde al nivel más alto, tiene falla respiratoria y neurológica, además tuvo infección nosocomial y una mortalidad predicha de 60. En este caso nuevamente se incorpora la evidencia, justo como en el caso anterior, y se propaga obteniendo una media estimada para la variable *cv2vivom* de 2.796 con una desviación estándar de 0.454, así que la media se encuentra muy cercana al valor tres y además si se considera una sola desviación estándar el valor para *cv2vivom* se encontraría entre 2.342 y 3.250, así que podríamos decir que lo más probable es que el valor de la variable *cv2vivom* sea tres y este valor corresponde a que la persona muere. Ahora bien, si se hace el análisis con la regresión trinomial para el mismo individuo hay que calcular, como en el caso anterior, $\exp(\hat{\alpha}_{muerto} + \hat{\beta}'_{muerto}\mathbf{x})$ = exp(5.143+0.02257(60))=663.281 y $\exp(\hat{\alpha}_{mala} + \hat{\beta}'_{mala}\mathbf{x})$ = exp(3.878-0.001749(60))=43.513 y a partir de estos valores se obtiene:

$$\hat{P}(cv2vivom = buena) = 0.0014$$

$$\hat{P}(cv2vivom = muerto) = 0.9371$$

$$\hat{P}(cv2vivom = mala) = 0.0614$$

Así que se observa que la regresión logística también indica que la categoría en donde con mayor probabilidad se encuentra el individuo es en la de morir.

Finalmente conviene recalcar, como ya se mencionó arriba, que para poder hacer comparaciones entre los modelos habría que dar los valores específicos para cada una de las variables explicativas, de tal modo que la comparación se haría de manera particular cada vez que se tiene una observación o individuo con distintos valores en sus variables y el proceso a seguir sería el mismo que se ha hecho para los dos individuos anteriores.

Capítulo 5

Comentarios y Conclusiones

En este capítulo se presentan las conclusiones obtenidas a partir de la realización de este trabajo, tales conclusiones son desde los puntos de vista metodológico, de la investigación médica y desde el punto de vista técnico práctico. Además se presentan comentarios respecto a la experiencia que se obtuvo al usar los modelos gráficos probabilísticos y además se discute acerca de los beneficios y limitantes de los mismos.

En primer lugar se formó una red en la que la variable de interés era la calidad de vida posterior a la estancia en la Unidad de Terapia Intensiva, la cual incluía a las variables: edad categorizada en dos grupos de edad, la calidad de vida inicial, la falla neurológica, la falla respiratoria y la cirugía urgente. Se observó que en la red la falla neurológica no tenía una relación directa con la calidad de vida posterior aunque sí habia una relación indirecta, de hecho se observó que la falla neurológica era independiente condicionalmente a la calidad de vida posterior dada la falla respiratoria, con la misma gráfica se listaron otras independencias condicionales entre las demás variables.

Las mismas variables fueron analizadas ajustando una regresión logística con solo efectos principales, cuya variable respuesta era la calidad de vida posterior a la estancia

en la Unidad de Terapia Intensiva y el resto de las variables eran las explicativas, todas las variables explicativas resultaron estadísticamente significativas, indicando que todas las variables realmente afectan a la variable respuesta. En la regresión ajustada resultó, como era de esperarse, que cuando las variables tomaban los valores en los que el individuo estaba en mejores condiciones, por ejemplo con buena calidad de vida inicial, aumentaba el riesgo de que la calidad de vida posterior fuera buena. Además, como las variables explicativas eran binarias y tan solo eran cinco se pudieron obtener las probabilidades estimadas de que la calidad de vida posterior fuera buena para toda combinación de valores posibles de estas variables explicativas o patrón de covariables. Se observó, al analizar los coeficientes estimados, coeficientes estimados relativos y estandarizados para esta regresión, que la variable a la que se le da más peso es la calidad de vida inicial de tal modo que la calidad de vida inicial determina en mayor medida la probabilidad estimada de que una persona tenga buena o mala calidad de vida posterior. La misma estimación de las probabilidades para toda posible combinación de valores de las variables explicativas se llevó a cabo para el modelo gráfico, así que se pudieron comparar las probabilidades estimadas entre ambos modelos. Por medio de estas comparaciones se observó que la variable relativa a la edad era de más peso en el modelo gráfico que las otras variables, pues cuando la categoría de edad asignada era la correspondiente a personas mayores de 60 años (61 o más años) las probabilidades estimadas de tener buena calidad de vida posterior eran más bajas que para la otra categoría. Posteriormente, se observo que la calidad de vida inicial no afecta tanto en el modelo gráfico la probabilidad estimada de que la calidad de vida posterior sea buena comparada con la regresión. Se observó también que en general las probabilidades estimadas de tener buena calidad de vida posterior en el modelo gráfico son mayores que las del modelo logístico, también se observó que en las probabilidades estimadas mediante el modelo gráfico se ven reflejadas las independencias condicionales mediante la presen-

cia de valores repetidos, en este caso se refleja que la falla neurológica es independiente condicionalmente a la calidad de vida posterior dada la falla respiratoria. Finalmente se utilizó el modelo gráfico para predecir a qué categoría de la calidad de vida posterior (buena o mala) pertenecería cada individuo de la base de datos original, logrando hacer tablas de clasificación, de tal modo que resultó que el 76.26 % de los individuos con buena calidad de vida posterior son clasificados correctamente y el 44.14 % de los individuos con mala calidad de vida posterior son clasificados correctamente, de forma similar en el caso de la regresión ajustada resultó que un porcentaje menor, el 74.20 %, de los individuos con buena calidad de vida posterior son clasificados correctamente; sin embargo, el 66.5 % de los individuos con mala calidad de vida posterior son clasificados correctamente.

Como ya se mencionó, se utilizó una regresión logística con efectos principales únicamente, entonces surge el cuestionamiento de qué ocurre al incluir interacciones entre las variables. Se explicó en esta tesis que se trabajó simultáneamente con las regresiones y con el modelo gráfico, de tal manera que cuando en el modelo gráfico se observaban relaciones entre las variables que sugieren interacciones entre las mismas se incluían; por otra parte, variables relevantes en las regresiones se incluían en los modelos gráficos; sin embargo, al realizar los ajustes en las regresiones con las interacciones correspondientes se observó que las interacciones que se iban incluyendo no resultaban significativas, por ello y para simplificar el modelo se utilizaron únicamente efectos principales. Además, el ajuste del modelo con solo efectos principales para las variables mencionadas resultó bastante bueno, para ello se analizó la devianza residual y otras pruebas de hipótesis para ver la calidad del ajuste. También se ajustó un modelo saturado (con interacciones de todas los ordenes) y se compararon los resultados de este modelo con el utilizado, en el modelo saturado resultó que 73.4 % de los individuos con mala calidad de vida posterior

son clasificados correctamente y 70.70 % de los individuos con buena calidad de vida posterior son clasificados adecuadamente, resultados similares a los que se tienen en la regresión con solo efectos principales que indicarían que es preferible quedarse con el modelo elegido (con solo efectos principales), por el principio de parsimonia; sin embargo, los resultados anteriores no son muy confiables porque en el software utilizado hubo problemas numéricos que afectan las estimaciones, esto es debido a la colinearidad al tener interacciones de alto orden. En conclusión el modelo ajustado es bueno, no es necesario agregar interacciones para tener un modelo satisfactorio, aunque si se quisiera obtener un mejor modelo logístico para la calidad de vida posterior se podría hacer un análisis exhaustivo de todas las interacciones: incluir todas las de segundo orden, luego las de tercer orden, etc. y comparar estos modelos con el que aquí se eligió.

En el caso del modelo gráfico, no hay un modelo saturado; sin embargo, un modelo con el que se puede comparar el modelo elegido es con áquel que no tiene restricciones, en el sentido de que cualesquiera variables pueden relacionarse entre sí; sin embargo, al ajustar este modelo gráfico se observa en primer lugar la existencia de dependencias incongruentes con la realidad, como por ejemplo la edad depende de variables como la calidad de vida inicial o la calidad de vida posterior, en segundo lugar al obtener la tabla de clasificación, aunque 96.46 % de los individuos con buena calidad de vida posterior son clasificados adecuadamente, tan solo 23.40 % de los individuos con mala calidad de vida posterior son clasificados correctamente, así que los individuos con mala calidad de vida posterior son muy mal clasificados comparados con el modelo gráfico con restricciones que se utilizó. Entonces, se observa que restringir el modelo gráfico es importante para obtener modelos con mayor capacidad de predicción. En este caso, el hecho de que los individuos de buena calidad de vida posterior con el modelo gráfico sin restricciones sean tan bien clasificados se debe a que la mayoría de las probabilidades

estimadas son mayores a 0.5 por lo cual prácticamente todos los individuos se asignan a calidad de vida posterior buena, por lo cual es lógico que los individuos que efectivamente tienen buena calidad de vida posterior sean clasificados correctamente, así resulta que este modelo a los individuos con buena calidad de vida posterior los clasifica bien pero a cambio a los de mala calidad de vida posterior los clasifica muy mal, incluso las probabilidades estimadas prácticamente no se aproximan a las frecuencias observadas.

En conclusión, desde el punto de vista médico, independientemente del modelo utilizado (logístico con efectos principales o el gráfico restringido) se observó que la calidad de vida posterior depende de las otras variables consideradas: edad dicotomizada, calidad de vida inicial, cirugía urgente, falla respiratoria y falla neurológica. Al final se obtienen dos modelos distintos, siendo decisión del médico cuál de los modelos se prefiere utilizar puesto que en cada uno de ellos las variables involucradas tienen distinta importancia, en cuanto a su grado de influencia, al momento de estimar la probabilidad de que un individuo tenga buena o mala calidad de vida posterior y entonces los modelos muchas veces clasifican de manera distinta a que categoría de calidad de vida posterior, buena o mala, pertenecen los individuos. De hecho, según la opinión de médicos consultados, por lo general se prefieren las estimaciones de un modelo que asigne y prediga adecuadamente a los individuos que pertenecen a la categoría mala calidad de vida posterior, en este caso se preferirían las estimaciones o predicciones que se obtienen usando la regresión logística pues con este modelo un mayor porcentaje de individuos con mala calidad de vida posterior son clasificados correctamente que con el modelo gráfico; además, aunque con el modelo gráfico los individuos con buena calidad de vida posterior son mejor clasificados que con el modelo logístico, la diferencia es mínima; entonces, desde el punto de vista médico y predictivo se eligiría el modelo correspondiente a la regresión logística.

Posteriormente, para las mismas variables, se obtuvó otra red que solo permitía que entraran o no arcos a la variable respuesta (calidad de vida posterior), para esta red se observó que la variable respuesta no dependía directamente de la falla neurológica. Se obtuvieron las probabilidades estimadas de buena calidad de vida posterior para esta red, fijando los valores que tomaban el resto de las variables (explicativas) y se observó que eran las mismas que en la otra red (ignorando la falla neurológica) por varias razones: los nodos padres de calidad de vida posterior eran los mismos en ambas redes, las probabilidades condicionales de esta variable dados sus padres también coincidían, en ambos casos se estaban fijando todas las variables explicativas por lo que no se estaban tomando en cuenta en la primer red las probabilidades condicionales entre las variables explicativas y también debido a la independencia condicional de la calidad de vida posterior con la falla neurológica dada la respiratoria. Se decidió comparar los resultados con ambas redes, pero ahora permitiendo que la calidad de vida inicial y la posterior fueran las variables "respuesta", es decir solo se fijaron los valores de las otras variables y así se obtuvieron resultados diferentes entre ambos modelos, además de que se ilustró cómo con un modelo gráfico no solo se puede ver cómo es afectada una variable, sino que se pueden dar valores de un subconjunto cualquiera de variables y ver cómo se ven afectadas las restantes.

A continuación, se analizó una segunda red en la que la variable de interés fue la variable binaria que identifica a los vivos de los muertos, esta red estaba conformada además por las variables: edad categorizada, calidad de vida inicial, mortalidad predicha codificada, infección nosocomial, falla respiratoria y falla cardiaca. Se observó que todas las variables afectaban directamente a la variable que separa a vivos de muertos, excepto la falla cardiaca, aunque la afecta indirectamente. También se observaron un

conjunto de independencias condicionales, entre ellas, para la variable que separa a vivos de muertos se observó que era independiente condicionalmente a la falla cardiaca dadas la edad, la mortalidad predicha, la infección nosocomial y la falla respiratoria.

Nuevamente se ajustó una regresión logística con solo efectos principales, cuya variable respuesta era la variable que separa a vivos de muertos y el resto de las variables eran explicativas. Todas las variables explicativas resultaron significativas por lo que estadísticamente todas influyen en la variable respuesta, se observó que al pertenecer un individuo al rango de edad inferior, al tener buena calidad de vida inicial, al tener baja mortalidad predicha, al no tener infección nosocomial, al no tener infección respiratoria y al no tener falla cardiaca aumentaba el riesgo de que la persona sobreviviera. Posteriormente se estimaron las probabilidades de sobrevivencia para toda posible combinación de las variables explicativas, tanto con el modelo logístico como con el gráfico. Se observaron que las predicciones entre ambos modelos son más parecidas entre sí que en la primer red para la calidad de vida posterior. Mediante estas estimaciones se observó que el modelo gráfico proporciona un mayor peso a las variables calidad de vida inicial y a la edad categorizada puesto que cuando un individuo tiene buena calidad de vida inicial o cuando pertenece al grupo inferior de edad (o ambas cosas) la probabilidad estimada de sobrevivencia aumenta en gran medida (teniendo menos importancia los valores de las otras variables) sobre todo al comparar con el modelo logístico, de manera similar al tener mala calidad de vida inicial o pertenecer al grupo de edad superior (o ambas cosas) disminuye la probabilidad de sobrevivir. En el modelo logístico se observó; al analizar los coeficientes, coeficientes relativos y coeficientes estandarizados ajustados y al comparar las estimaciones obtenidas, que las variables de más peso son la mortalidad predicha y la falla respiratoria, de tal modo que al no tener falla respiratoria o baja mortalidad predicha (o ambas) aumenta mucho

la probabilidad estimada de sobrevivir, afectando menos los valores de las variables restantes, y al contrario al tener falla respiratoria o alta mortalidad predicha (o ambas) disminuye mucho la probabilidad de sobrevivir. Así que las mayores diferencias entre ambos modelos se dan debido a que hay distintas variables que tienen más peso o importancia en cada modelo y estas hacen que se incremente mucho o diminuyan mucho las probabilidades estimadas de un modelo respecto a las del otro. Nuevamente se hizo una tabla de clasificación para la variable que separa vivos de muertos con la finalidad de ver el porcentaje de individuos que el modelo gráfico está asignado correctamente a cada categoría en la base de datos de la que se partió, se observa que bajo el modelo gráfico el 72.27 % de los sobrevivientes están clasificados correctamente y el 45.68 % de los muertos están bien clasificados, al comparar con la tabla de clasificación correspondiente para el modelo logístico se observa que en este la situación es al revés puesto que están mejor clasificados los muertos, con un 81.50 %; sin embargo, solo el 57.80 % de los vivos se están clasificando adecuadamente, así que son dos modelos en el que distintas categorías están mejor clasificadas en uno y en otro. Otra vez, este modelo logístico con solo efectos principales resultó ser un buen modelo (analizando la devianza residual y otras pruebas), pueden compararse nuevamente la tabla de clasificación de este modelo con la del modelo saturado correspondiente; sin embargo, como en el caso del modelo que incluía la calidad de vida posterior hay problemas numéricos en el software utilizado que afectan las estimaciones y aunque la tabla de clasificación es similar a la de la regresión utilizada, lo cual indicaría que no hay mucha ganancia en utilizar un modelo complejo cuando uno simple es suficiente, los resultados obtenidos no son muy confiables. Si quisiera mejorarse el modelo logístico, se haría incluyendo interacciones y eligiendo aquellas significativas (algunas interacciones ya se analizaron para realizar este trabajo) hasta llegar a algun modelo adecuado que mejore la clasificación de los individuos; sin embargo, esto ya no sería necesario puesto que el modelo con solo efectos

principales es adecuado.

Se puede concluir para estas otras variables, que en ambos modelos (gráfico y logístico) el estado vital (sobrevivir o no sobrevivir) depende de las otras variables involucradas (en el modelo gráfico la dependencia a veces es indirecta) que son la edad dicotomizada, la calidad de vida inicial, la mortalidad predicha, la infección nosocomial, la falla respiratoria y la falla cardiaca. Entonces se observa, que el sobrevivir o no, está en parte determinado por la calidad de vida inicial, la edad, la mortalidad predicha y por el hecho de infectarse o no en el hospital, y esto era precisamente la hipótesis de los médicos. Otra vez, al tener dos modelos distintos, depende de los médicos decidir con cual prefieren trabajar de acuerdo a si prefieren el que clasifica mejor a los sobrevivientes (modelo gráfico) o a los no sobrevientes (modelo logístico) y también de acuerdo a su experiencia en cuestión de qué variables influyen más sobre el estado vital, pues cada modelo le da importancia o peso distinto a la influencia que algunas variables explicativas tienen sobre las probabilidades estimadas de la variable respuesta. De hecho, desde el punto de vista de los médicos consultados se prefieren las predicciones que se obtienen con un modelo que clasifica mejor a los no sobrevivientes, que para este caso como ya se dijo, corresponden a las obtenidas a partir del modelo logístico.

Conviene aclarar que en las tablas de clasificación de los modelos descritos arriba, para la calidad de vida posterior y para el estado vital, se utilizó como punto de corte el valor 0.5, de tal modo que si la probabilidad estimada de que el individuo tuviera buena calidad de vida posterior era mayor o igual a 0.5 el individuo se asignaba a la categoría buena c.v. posterior, similarmente si la probabilidad de sobrevivencia era mayor o igual a 0.5 el individuo fue asignado a la categoría vivo. Sin embargo, el punto de corte no es un valor fijo y para cada modelo se puede asignar un punto de corte distinto, para ello

se analizan las curvas ROC, que son gráficas en las que para distintos puntos de corte se tienen parejas ordenadas formadas por la sensitividad (que es la probabilidad de que bajo el modelo se asigne como éxito a lo que en realidad es éxito) y por 1-especificidad (donde especificidad consiste en la probabilidad de asignar bajo el modelo como fracaso a lo que realmente fue un fracaso), se prefieren modelos cuya sensitividad sea grande y 1-especificidad chica, lo cual se cumple cuando el área bajo la curva sea grande. Entonces, para tomar un punto de corte adecuado en un cierto modelo, se puede utilizar aquel valor en el que veamos que la senstividad aumenta lo más posible y 1-especificidad disminuye lo más posible, así que si la probabilidad estimada bajo el modelo es mayor o igual a este valor se asigna al individuo a la categoría correspondiente a éxito, en caso contrario a la categoría fracaso. Se ajustaron las curvas ROC, utilizando SPSS v13, para cada uno de los modelos (ejemplo en la figura B.1 y tabla B.6) y se observó: Para el modelo que contiene a la variable calidad de vida posterior (cuya categoría éxito sería buena c.v. posterior) y que resulta de ajustar una regresión logística se observó que un punto de corte adecuado es aproximadamente 0.5 y el área bajo la curva ROC es 0.767, para el modelo gráfico que contiene a la variable calidad de vida posterior se observó que un punto de corte adecuado es aproximadamente 0.6 y bajo la curva ROC se tiene un área de 0.689, para el modelo que contiene a la variable que separa a vivos de muertos (cuya categoría éxito sería vivo) y que resulta de ajustar una regresión logística un buen punto de corte es aproximadamente 0.4 con un área bajo la curva de 0.753 y finalmente en el modelo gráfico que también contiene a la variable que separa a vivos de muertos un punto de corte adecuado es aproximadamente 0.5, con un área bajo la curva de 0.630. En conclusión, analizando las curvas ROC cada modelo tendría un punto de corte distinto; sin embargo, se puede observar que los puntos de corte adecuados en los modelos son muy cercanos o incluso coinciden con 0.5, que es el valor que se utilizó en este estudio, porque es el valor prestablecido en el software y además de esta forma

todos los modelos tuvieron el mismo punto de corte.

Finalmente se manejó un modelo mixto, formado por variables continuas y discretas, en el cual hay dos variables continuas: la mortalidad predicha y una variable tricotómica que distingue a los sobrevivientes con buena calidad de vida posterior de los sobrevivientes con mala calidad de vida posterior y de los muertos. En este caso se manejó a esta última variable como continua debido a que hay una limitante técnica en el software utilizado que no permite que variables discretas dependan de variable continuas (no se permiten padres continuos de nodos discretos) y al tomar a la variable mencionada como continua entonces cualquier otra variable, continua o discreta, puede ingresar a ella, también se decidió considerarla como continua porque en el programa DEAL se observaron problemas al exportar a *Hugin* resultados para variables tricotómicas y además para ilustrar el caso mixto. Las variables de este modelo son la edad categorizada, la calidad de vida inicial, como ya se mencionó la mortalidad predicha, la infección nosocomial, la falla neurológica, la falla respiratoria y la cirugía urgente. En la red se observó que las variables que afectan directamente a la variable que separa a vivos con buena calidad de vida posterior, de los de mala calidad de vida posterior y de los muertos son todas excepto la edad y la cirugía urgente; las cuales de todos modos la afectan indirectamente. Para esta misma variable se observó que es independiente condicionalmente de la edad dados la calidad de vida inicial, la mortalidad predicha y la falla respiratoria, además también es independiente condicionalmente a la cirugía urgente dadas la calidad de vida inicial, la infección nosocomial y la falla respiratoria.

Otra vez se ajustó una regresión logística, en este caso trinomial, con variable respuesta la variable que separa a vivos con buena calidad de vida posterior, de vivos

con mala calidad de vida posterior y de muertos, cuyas variables explicativas son las restantes. Se tomó como categoría de referencia a la categoría de buena calidad de vida posterior y se observó que el riesgo de tener buena calidad de vida (c.v.) o morir es similar a pesar de incrementar el valor de la mortalidad predicha, también se observó que cuando un individuo tiene buena calidad de vida inicial, no tiene infección nosocomial, no tiene falla neurológica, no tiene falla respiratoria, no necesita cirugía urgente y cuando el rango de edad es inferior disminuye el riesgo de pasar de la categoría de buena c.v. posterior a muerto, o sea es más probable tener una buena c.v. posterior. Por otra parte, cuando se compara la mala c.v. posterior con la buena se observó que la mortalidad predicha y la infección nosocomial no eran significativas y para el resto de las variables se observó que cuando el individuo está en mejores condiciones hay más posibilidad de tener buena calidad de vida posterior que mala. La calidad del ajuste de esta regresión logística es buena, lo cual se observa mediante la devianza residual y otras pruebas de hipótesis, entonces no es necesario agregar interacciones para tener un modelo que se ajusta adecuadamente a los datos . En base a los coeficientes de las variables explicativas estimados en la regresión se pueden obtener las probabilidades estimadas para cada una de las categorías de la variable respuesta; similarmente para el modelo gráfico se pueden obtener valores estimados para la variable que se considera como respuesta en la regresión; sin embargo, en este caso como la variable es continua se obtendría como valor estimado la media y la desviación estándar de una distribución Gaussiana, entonces aunque el valor obtenido sea continuo se podrían obtener ciertos intervalos de valores que toma la variable y ver en que categoría es más posible que se encuentre el individuo: mala calidad de vida posterior, buena o si murió. Se observó que la comparación entre las estimaciones obtenidas mediante la red con las obtenidas con la regresión se haría individuo por individuo y no como en las redes anteriores en que se podían listar todas las combinaciones de valores de las variables explicativas, además

la comparación no es tan fácil pues en un modelo se tienen probabilidades y en otro intervalos de valores; sin embargo, en este trabajo se compararon con ambos modelos un par de individuos con distintos valores en sus variables explicativas obteniendo resultados similares.

Entonces, desde el punto de vista médico, se observa que bajo el modelo gráfico las variables consideradas explicativas en el último modelo: edad dicotomizada, calidad de vida inicial, mortalidad predicha, infección nosocomial, falla neurológica, falla respiratoria y cirugía urgente; influyen, directa o indirectamente, en la variable tricotómica, considerada respuesta, que separa a los sobrevivientes con buena calidad de vida posterior, mala y los no sobrevivientes. Bajo el modelo logístico resultó que ni la mortalidad predicha, ni la infección nosocomial sirven para distinguir a individuos que tienen mala o buena calidad de vida posterior; en el caso de la mortalidad predicha esto era de esperarse pues esta variable nada más sirve para separar a vivos de muertos; en el caso de la infección nosocomial, se observó en el modelo anterior, que esta variable es importante para distinguir si una persona sobrevive o no, pero no influye en la calidad de vida posterior. Entonces, la infección nosocomial y la mortalidad predicha son variables cuya inclusión es adecuada en el contexto de modelos que tratan de distinguir entre sobrevivientes y no sobrevivientes pero no para distinguir entre buena y mala calidad de vida posterior. En el caso del modelo gráfico, resulta como ya se dijo, que todas las variables incluso infección nosocomial y mortalidad predicha afectan a la variable respuesta, lo cual es congruente pues en este modelo no se comparan las categorías de la variable respuesta con una categoría de referencia como en el modelo logístico, así que el modelo no distingue si la influencia sobre la variable respuesta es al separar a vivos de muertos o a individuos con buena o mala calidad de vida posterior y entonces como sobrevivir o no son categorías de la variable respuesta que dependen de la infección

nosocomial y de la mortalidad predicha es lógico que en la gráfica haya dependencias entre estas variables y la respuesta. Entonces, con el modelo gráfico, se verifica que las variables que sospechaban los médicos que influían sobre la respuesta: edad, calidad de vida inicial, mortalidad predicha e infección nosocomial realmente están influyendo.

En conclusión, por medio de los modelos gráficos, en particular las Redes Bayesianas, se puede llevar a cabo un análisis de datos, en este caso fue para pacientes que ingresan en Unidades de Terapia Intensiva, este análisis gráfico en primer lugar permite modelar datos de una manera sencilla, pues se puede ver la relación de dependencia entre todas las variables para una base de datos dada mediante figuras formadas por flechas y círculos (o cualquier figura geométrica que represente a las variables) que ilustran estas relaciones, las cuales pueden darse entre todas las variables y no únicamente hacia una sola variable como ocurre en una regresión logística con una variable respuesta, en la que aunque haya interacciones el objetivo de incluirlas es explicar a la respuesta, lo cual en términos de un modelo gráfico se traduce en que aunque se permiten aristas (líneas sin dirección) que unen variables, los arcos solo pueden apuntar a una variable respuesta. Por otra parte, al obtener una representación gráfica de las variables estos modelos facilitan la comprensión a cualquier persona, ajena o no a la estadística, de las relaciones entre variables e incluso con ellos uno se puede dar cuenta de independencias condicionales entre variables. Para llevar a cabo la modelación se requirió del uso de un programa, DEAL, el cual sirve para obtener a partir de una base de datos una Red Bayesiana que represente adecuadamente las relaciones entre las variables e incluso restringir relaciones imposibles entre variables.

Los modelos gráficos además de representar las relaciones entre variables también pueden servir como un modelo para predicción, en el sentido de que uno proporciona

los valores de un grupo de variables y se ve cómo se modifican las variables restantes a través de las probabilidades marginales correspondientes; en este trabajo, para el proceso descrito se utilizó el software *Hugin 5.4*. Para que las predicciones mencionadas sean acertadas se necesita que la Red generada represente y simplifique de la mejor forma la manera en que en la realidad están relacionadas las variables entre sí y entonces entre mejor esté representada la realidad mediante una Red mejores resultados se obtendrían. En este sentido hay dos áreas que se están desarrollando en los modelos gráficos: por un lado el generar una Red que represente de la mejor forma posible la relación entre variables partiendo de una base de datos cualquiera (es decir aprender de los datos para obtener una Red) y por otro lado el hecho de, que una vez que están representados los datos mediante una Red específica, obtener algoritmos eficientes para poder inferir acerca de lo que ocurre en las variables en conjunto (en sus probabilidades marginales) y también ver cómo afectan los valores de un subconjunto de variables de interés en las variables restantes.

El objetivo o meta a largo plazo de los modelos gráficos sería llegar a un punto en el que dando una base de datos sin ninguna limitación en cuanto al número de variables, las cuales pudieran ser continuas o discretas, se obtuviera una red que representara de la mejor manera posible la relación que en la realidad existe entre las variables, que un programa de forma más o menos automática propusiera a partir de un conjunto de variables un subconjunto de variables que fueran las más importantes para el problema debido a que son las más relacionadas entres sí (es decir que el programa permita la selección de variables para así obtener modelos parsimoniosos) y que una vez obtenida esta Red se pudiera implementar en la práctica, de tal modo que cualquier persona pudiera dar los valores que conoce de algunas variables y predecir lo más acertado posible lo que ocurre con las restantes. Por ejemplo, en el caso de los pacientes en la Unidad

de Terapia Intensiva se podría obtener una Red en la que al llegar un nuevo paciente (o incluso una base de datos de pacientes nuevos) a la Unidad y medirle algunos valores de ciertas variables con las que llega y otros durante su estancia en la UTI los médicos podrían predecir simultáneamente cuánto tiempo va a permanecer en la UTI, qué calidad de vida es más probable que tenga, si es más probable que viva o muera, incluso podría hacerse una análisis en el que sabiendo que su calidad de vida posterior es mala y midiendo otras variables se supiera con que probabilidad se tenía o no un factor de riesgo, en fin, se podría llegar a un momento en que con una Red se pudiera predecir o también encontrar o confirmar la existencia de factores de riesgo. Además con las redes se podría saber que variables se están influenciando entre sí e incluso darse cuenta de relaciones indirectas entre variables que pudieran pasar desapercibidas anteriormente.

Sin embargo; aunque mucho de lo anterior puede ser realizado actualmente, en la realidad los modelos gráficos son todavía un área de estudio muy reciente, que como ya se ha dicho involucra en gran medida a la Inteligencia Artificial y al área de Computación, por lo mismo falta un mayor apoyo e interacción con el área Estadística, también falta mayor difusión de los mismos. Además los modelos gráficos obtenidos y analizados con los programas utilizados en este trabajo todavía cuentan con muchas limitantes técnicas, entre ellas el hecho de que no se permiten padres continuos de nodos discretos, también que debido a los algoritmos el número de variables todavía es limitado; además, todavía no se permiten datos faltantes en la base de datos que se maneja, por otra parte uno proporciona las variables a utilizar y el programa no las sugiere así que podría ocurrir que no se están incluyendo las más importantes para lo que interesa estudiar o que al contrario se estén incluyendo demasiadas variables, además para introducir evidencia (valores de algunas variables) en *Hugin 5.4* esto se hace individuo por individuo, lo cual sería muy laborioso para una base de datos muy grande cuyas

variables tomaran muchos valores y no como en este caso que muchas de las variables eran discretas. Por otro lado, en el caso del programa DEAL que se utilizó para llevar a cabo el Aprendizaje Estructural, es decir el hecho de obtener un modelo gráfico a partir de los datos, una crítica fuerte es que si no se tiene una Red Bayesiana a priori, como fue el caso de este estudio, las distribuciones locales parametrales a priori, en el caso continuo sobre todo, se obtienen a partir de los datos (por ejemplo en el caso continuo se utiliza una matriz de varianza muestral Σ_i y la media muestral m_i) y luego las distribuciones locales parametrales posteriori y como consecuencia las estimaciones nuevamente se obtienen a partir de los datos, así que pareciera que se usa de manera doble la información, lo cual no es muy correcto desde el punto de vista Bayesiano, entonces sería muy conveniente obtener de alguna manera una Red Bayesiana a priori para que así las distribuciones locales parametrales a priori no dependan de los valores de la base de datos y que no se utilice de más la información con la que se cuenta.

Por otro lado existen otros programas que permiten manejar módelos gráficos, con diferentes enfoques, supuestos y entonces es cuestión personal el que se vaya a usar. Como ya se mencionó, en este estudio se utilizó para la parte de Aprendizaje Estructural el programa DEAL; sin embargo, existen otros programas que pudieran emplearse como por ejemplo nuevas versiones de *Hugin* (6.3); sin embargo, como este programa tiene un costo muy elevado solo se tuvo acceso a la versión de prueba correspondiente, la cual incluye limitantes en cuanto al número de variables y observaciones que se pueden utilizar; de hecho, con este programa solo se pudo llevar a cabo el Aprendizaje Estructural para el modelo que incluye la variable calidad de vida posterior (la base de datos correspondiente incluye 386 casos y 6 variables). *Hugin 6.3* utiliza otros algoritmos para obtener la estructura de las redes (i.e. para obtener los arcos que relacionan las variables entre sí) a partir de los datos, que son el algoritmo NPC (*Necessary Path Condition*)

o el PC (*Path Condition*), una vez que se tiene la estructura de la red Bayesiana las probabilidades condicionales de la red (probabilidades locales) se estiman usando el algoritmo llamado EM y entonces se obtendría una red con sus respectivas probabilidades locales asociadas de forma similar a como se hizo usando DEAL. Sería objeto de otra tesis u otro estudio el analizar estos otros algoritmos, su funcionamiento, limitantes, etc. Nuevamente se puede obtener, como con DEAL, una red en la que no se restringen las direcciones en los arcos entre las variables, pero nuevamente se obtuvieron relaciones incongruentes con la realidad como que la c.v. inicial dependía de la c.v. posterior; sin embargo, como con DEAL se pueden prohibir las direcciones de los arcos o incluso la relación entre variables (cuando estas no se relacionan en ninguna dirección, o sea cuando dos variables no dependen entre sí) y hasta se pueden forzar las direcciones de los arcos entre algunas variables, incluso se podría forzar a que la red que se obtenga en *Hugin 6.3* sea la misma que se obtuvo utilizando DEAL (la gráfica de la figura 4.1) pero estimando las probabilidades locales utilizando *Hugin*. Para esta red (obtenida con *Hugin*) otra vez se obtienen las probabilidades estimadas de tener buena c.v. posterior para toda posible combinación de valores de las variables explicativas, la gráfica que muestra estas probabilidades estimadas comparándolas con las obtenidas con el modelo logístico y con el otro modelo gráfico se muestra en el Apéndice (tabla B.7, figura B.2 y B.3). Se observa que las probabilidades estimadas se parecen en algunas celdas (combinación de valores de las variables explicativas) a las estimadas bajo el modelo gráfico de la figura 4.1, en otras a las estimadas bajo la regresión logística correspondiente y en otras celdas discrepa con ambos modelos, los valores estimados se encuentran entre 0 y 0.8387, además se observa que la variable edad no es de tanta importancia como en el otro modelo gráfico, pues hay celdas cuyas probabilidades estimadas son pequeñas independientemente de que la variable edad se encuentra en el rango inferior o en el rango superior, en cambio en el modelo gráfico obtenido usando DEAL las probabili-

dades estimadas en las celdas en las cuales la variable edad se encontraba en el rango superior eran las más pequeñas y las probabilidades estimadas en las celdas en las cuales la variable edad se encontraba en el rango inferior eran las mayores. Sin embargo, la variable calidad de vida inicial influye mucho en las probabilidades estimadas, como en el caso de la regresión logística, pues en aquellas celdas en las que las c.v. inicial es mala las probabilidades estimadas está por debajo de 0.3333, en cambio en las celdas en las que las c.v. inicial es buena las probabilidades estimadas están por arriba de 0.3 (aunque hay un par de celdas con valores cero, pero estas corresponden a celdas en las que no hay prácticamente ninguna observación). Una última observación es que bajo el modelo gráfico aprendido usando *Hugin* las probabilidades estimadas de buena c.v. posterior se aproximan mucho a las frecuencias relativas observadas correspondientes a cada combinación de valores de las variables explicativas. Sin embargo, un inconveniente del modelo anterior es, como ya se mencionó, que forzamos a que las relaciones de dependencia fueran las mismas que las del modelo gráfico obtenido en DEAL, si no se hace así, sino que se restringieran las direcciones de los arcos entre las variables y después se obtiene la red y se estiman las probabilidades condicionales correspondientes, resulta un gráfica en la que la cirugía urgente y la falla respiratoria no se relacionan con el resto de las variables (incluyendo, obviamente, la c.v. posterior) pero tanto la regresión logística como el modelo gráfico aprendido usando DEAL indican que esta relación existe. Entonces, para llevar a cabo el Aprendizaje Estructural con *Hugin* se debe llevar a cabo nuevamente un análisis de las variables, de las relaciones de dependencia , etc. con ayuda de los médicos, también se debe indagar el funcionamiento de los algoritmos y entonces se tienen otros modelos gráficos cuya obtención se basa en otra metodología distinta a la de DEAL; sin embargo, para que se pueda llevar a cabo el análisis sin limitantes, en cuanto a número de observaciones y variables, sería necesario tener acceso a la versión comercial más reciente de *Hugin*. Finalmente, con todo

lo anterior se puede ver que hay distintos programas para analizar modelos gráficos; aunque cada programa tienen su sintaxis particular y una metodogía y teoría detrás del mismo distinta, por lo que aprender cada uno implica tiempo y algunos no son tan simples o amigables para el usuario.

En conclusión, los modelo gráficos en la actualidad son un área naciente y poco explotada, son una buena herramienta más para analizar datos, todavía tienen muchas limitantes técnicas, además para obtener algún modelo para ciertos datos se requiere mucho apoyo de expertos en el área que se esté estudiando (en el caso de este trabajo, se necesitó una interdisciplina adecuada entre medicina y estadística), lo cual limita más su aplicación pues al no tener cuidado se podría llegar a modelos no compatibles con la realidad, pero a pesar de sus defectos, como otra herramienta más en estadística merecen atención, sobre todo por el potencial que pueden llegar a tener si es que se van resolviendo las dificultades técnicas que se han señalado.

Apéndice A

Variables en la base de datos

Antes de hospitalizarse		
Título	**Significado**	**Códigos**
Hospital	Hospital	1=CMN; 2=CMLR
Caso	Caso	
Sexo	Sexo	1=Mujer; 2=Hombre
edadcont	Edad	
edadcod1	Edad codificada	1≤40; 2=41-60; 3≥ 61 años
Procede	Sitio de procedencia	1=Urgencias; 2=Piso; 3=Quirófano
epoc	Enfermedad pulmonar	0=No; 1=Sí
diabetes	Diabetes mellitus	0=No; 1=Sí
cáncer	Cáncer codificado	0=No; 1=Sí
Charlson	Gravedad de las enfermedades antes de hospitalizarse	
Charlso	Gravedad codif. de las enfermedades antes de hospitalizarse	0=Ninguna; 1=1 ó más
cv1	Calidad de vida 2 meses previos a la hospitalización	
cv1cod (cv1codif)	Calidad de vida 2 meses previos a la hospitalización codificada	1=Buena; 2=Mala

Durante la hospitalización, antes de ingresar a la terapia intensiva		
Servicio	Servicio.	1=Medicina; 2=Cirugía
Durante la estancia en la terapia intensiva		
apacheii	APACHE II (Calificación de la gravedad de la enfermedad)	
Aii	APACHE II codificado	0=15 ó menos; 1=16 ó más
qxurgent (cirugia)	Cirugía urgente	0=Cirugía electiva; 1=Cirugía urgente
mortpred	Mortalidad predicha (Calculada por el modelo APACHE II) Expresada en porcentaje	
Altah	Motivo de alta hospitalaria	0=Mejoría; 1=Muerto
Altauti	Motivo de alta de la terapia intensiva	1=Mejoría; 3=Muerte; 5=Máximo beneficio
Adquirid	Sitio de adquisición de la infección	0=Sin infección; 1=En la comunidad; 2=En el hospital
aps1	APS1. Calificación fisiológica aguda en el primer día de estancia en la terapia	
Brus1	Bruselas de ingreso (Gravedad del enfermo)	
sepsis	Sepsis grave	0=No; 1=Sí
fcard	Falla cardiaca	0=No; 1=Sí
fendócri	Falla endócrina	0=No; 1=Sí
frenal	Falla renal	0=No; 1=Sí
fresp	Falla respiratoria	0=No; 1=Sí
fneur	Falla neurológica.	0=No; 1=Sí
fhepátic	Falla hepática	0=No; 1=Sí
Díasfcar	Días en falla cardiaca	
Diasfen	Días en falla endócrina	
Díasfren	Días en falla renal	
Díasfres	Días en falla respiratoria	
Díasfneu	Días en falla neurológica.	
Díasfhep	Días en falla hepática	
vm	Ventilación asistida mecánica	0=No; 1=Sí
vmdias	Días en ventilación mecánica asistida	
npt	Nutrición artificial	0=No; 1=Sí
Traqueos	Traqueostomía	0=No; 1=Sí
Glasg1	Glasgow día 1	0=No; 1=Sí
Ingr1	Ingresos hídricos del primer día	
Bh1d	Balance hídrico del primer día	
Gluc1	Glucosa del primer día	
Creat1	Creatinina día 1	
ik1	Índice de Kirby respiratorio día 1	
esthosp	Estancia hospitalaria	
estuti	Estancia en UTI (días)	
estuticod	Estancia en UTI codificada	1=5 ó menos; 2=6 ó más
estposut	Estancia post-UTI	

A los 3 meses de egreso del hospital		
cv2	Calidad de vida a los 3 meses del alta hospitalaria	
cv2codif	Calidad de vida a los 3 meses del alta hospitalaria codificada	0=Buena CV; 1=Mala CV
calsep	Calidad de vida a los 3 meses del alta y sepsis	0=Buena CV sin sepsis; 1=Buena CV con sepsis; 2=Mala CV sin sepsis; 3=Mala CV con sepsis

Apéndice B

Otras tablas y gráficas

Variable	$\hat{\beta}_{jest}$ reg. tabla 4.2	Variable	$\hat{\beta}_{jest}$ reg. tabla 4.7	Variable	$\hat{\beta}_{jest}$ reg. tabla 4.8
edadcod60	0.3060	edadcod60	0.2959	edadcod60	0.1825
cv1cod	0.7410	cv1cod	0.7266	cv1cod	0.2114
fneur	0.2674	fresp	0.3699	mortpredcod	0.4416
fresp	0.3409	cirugia	0.3436	infnos	0.3277
cirugia	0.3395			fresp	0.5152
				fcard	0.2029

Tabla B.1: Coeficientes estandarizados ajustados para las regresiones logísticas de la sección 4.1 y 4.2

edad cod.	cirugía urg.	falla resp.	falla neur.	cv1 cod.	$\hat{P}_{reg.}$	obs.	obs. buena c.v. post. O_i	obs. buena c.v. post. bajo reg.	p_i	E_i
1	0	0	0	1	0.8124	60	50	48.7469	0.2808	49.4228
1	0	1	0	1	0.6735	84	60	56.5729	0.3259	57.3573
1	0	1	0	2	0.2604	17	2	4.4264	0.0255	4.4878
1	1	1	0	1	0.4945	37	17	18.2965	0.1054	18.5502
2	0	0	0	1	0.6966	21	15	14.6279	0.0843	14.8307
2	0	1	0	1	0.5222	32	14	16.7115	0.0963	16.9432
2	0	1	0	2	0.1572	24	5	3.7735	0.0217	3.8258
2	1	1	0	1	0.3414	27	9	9.2182	0.0531	9.3460
2	1	1	0	2	0.0813	15	4	1.2193	0.0070	1.2362
Totales						317	176	173.5932	est. T	8.7027

Tabla B.2: Tabla para calcular la prueba ji cuadrada de bondad de ajuste en la regresión logística para *cv2* de la sección 4.1 (celdas con 15 o más observaciones)

edad cod.	cirugía urg.	falla resp.	falla neur.	cv1 cod.	$\hat{P}_{graf.}$	obs.	obs. buena c.v. post. O_i	obs. buena c.v. post. bajo graf.	p_i	E_i
1	0	0	0	1	0.8124	60	50	44.8020	0.2341	41.1986
1	0	1	0	1	0.6735	84	60	66.8220	0.3491	61.4476
1	0	1	0	2	0.2604	17	2	9.6730	0.0505	8.8950
1	1	1	0	1	0.4945	37	17	23.6393	0.1235	21.7380
2	0	0	0	1	0.6966	21	15	8.7507	0.0457	8.0469
2	0	1	0	1	0.5222	32	14	12.8000	0.0669	11.7705
2	0	1	0	2	0.1572	24	5	10.9080	0.0570	10.0307
2	1	1	0	1	0.3414	27	9	8.9991	0.0470	8.2753
2	1	1	0	2	0.0813	15	4	4.9995	0.0261	4.5974
Totales						317	176	191.3936	est. T	17.3862

Tabla B.3: Tabla para calcular la prueba ji cuadrada de bondad de ajuste en el modelo gráfico para *cv2* de la figura 4.1 (celdas con 15 o más observaciones)

cv1 cod.	edad cod.	falla card.	falla resp.	inf. nos.	mort. pred.	$\hat{P}_{reg.}$	obs.	obs. vivos O_i	obs. vivos bajo reg.	p_i	E_i
1	1	0	0	0	0	0.9097	33	32	30.0202	0.1358	30.2944
1	1	0	1	0	0	0.7213	46	35	33.1805	0.1502	33.4837
1	1	0	1	0	1	0.4958	42	23	20.8215	0.0942	21.0117
1	1	1	1	0	0	0.6302	44	27	27.7280	0.1255	27.9814
1	1	1	1	0	1	0.3929	45	15	17.6820	0.0800	17.8435
1	1	1	1	1	0	0.4593	33	12	15.1582	0.0686	15.2967
1	1	1	1	1	1	0.2440	46	13	11.2229	0.0508	11.3255
1	2	0	1	0	1	0.3992	30	8	11.9746	0.0542	12.0840
1	2	1	1	0	1	0.3043	69	18	20.9953	0.0950	21.1871
1	2	1	1	1	1	0.1790	43	13	7.6979	0.0348	7.7682
2	2	0	1	0	1	0.2959	24	6	7.1011	0.0321	7.1660
2	2	1	1	0	1	0.2167	54	13	11.7018	0.0530	11.8087
2	2	1	1	1	1	0.1212	47	8	5.6970	0.0258	5.7490
Totales							556	223	220.9810	est. T	8.3730

Tabla B.4: Tabla para calcular la prueba ji cuadrada de bondad de ajuste en el modelo logístico para *vivomuer* de la sección 4.2 (celdas con 20 o más observaciones)

cvl cod.	edad cod.	falla card.	falla resp.	inf. nos.	mort. pred.	$\hat{P}_{reg.}$	obs.	obs. vivos O_i	obs. vivos bajo reg.	p_i	E_i
1	1	0	0	0	0	0.9097	33	32	25.1196	0.0749	16.7064
1	1	0	1	0	0	0.7213	46	35	34.1136	0.1017	22.6881
1	1	0	1	0	1	0.4958	42	23	33.0918	0.0987	22.0085
1	1	1	1	0	0	0.6302	44	27	32.6304	0.0973	21.7017
1	1	1	1	0	1	0.3929	45	15	35.4555	0.1057	23.5806
1	1	1	1	1	0	0.4593	33	12	21.5523	0.0643	14.3339
1	1	1	1	1	1	0.2440	46	13	33.9526	0.1013	22.5810
1	2	0	1	0	1	0.3992	30	8	17.4990	0.0522	11.6381
1	2	1	1	0	1	0.3043	69	18	40.2477	0.1200	26.7677
1	2	1	1	1	1	0.1790	43	13	23.8908	0.0713	15.8892
2	2	0	1	0	1	0.2959	24	6	7.9992	0.0239	5.3201
2	2	1	1	0	1	0.2167	54	13	17.9982	0.0537	11.9701
2	2	1	1	1	1	0.1212	47	8	11.7500	0.0350	7.8146
Totales							556	223	335.3007	est. T	34.3016

Tabla B.5: Tabla para calcular la prueba ji cuadrada de bondad de ajuste en el modelo gráfico para *vivomuer* de la figura 4.12 (celdas con 20 o más observaciones)

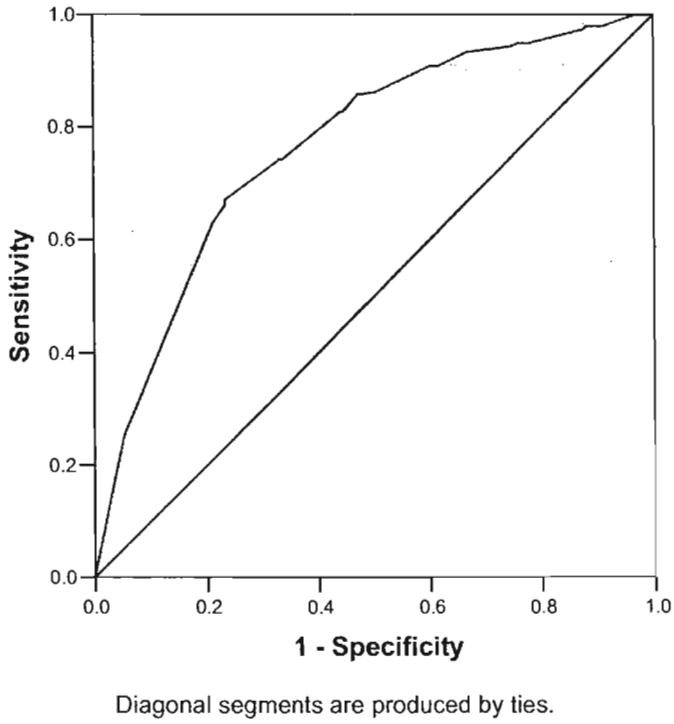

Figura B.1: Curva ROC para el modelo en que se ajusta una regresión logística con variable respuesta *cv2*

Sintaxis para generar curvas ROC en SPSS v 13.0

```
ROC
  probestcv2buena  BY cv2 (1)
  /PLOT = CURVE
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .
```

Puntos de corte	sensitividad	1-especificidad
0.000	1.000	1.000
0.046	1.000	0.984
0.063	1.000	0.979
0.067	1.000	0.973
0.074	1.000	0.968
0.100	0.980	0.910
0.131	0.980	0.904
0.150	0.980	0.883
0.157	0.975	0.878
0.161	0.949	0.777
0.213	0.949	0.761
0.260	0.944	0.750
0.266	0.934	0.670
0.277	0.924	0.649
0.288	0.909	0.617
0.318	0.909	0.601
0.383	0.864	0.505
0.433	0.859	0.473
0.455	0.828	0.447
0.481	0.828	0.441
0.508	0.742	0.335
0.522	0.742	0.330
0.573	0.672	0.234
0.648	0.662	0.234
0.673	0.631	0.213
0.685	0.328	0.085
0.754	0.253	0.053
1.000	0.000	0.000

Tabla B.6: Coordenadas de la curva ROC para distintos puntos de corte en el modelo en que se ajusta una regresión logística con variable respuesta *cv2*

celda	edad cod.	cirugía urg.	falla resp.	falla neur.	cv1 cod.	$\hat{P}_{reg.}$	$\hat{P}_{graf.}$	$\hat{P}_{Hugin}$	frec. rel. buena c.v.	obs.
1	1	0	0	0	1	0.8124	0.7467	0.8387	0.8333	60
2	1	0	0	0	2	0.4251	0.5600	0.1429	0.1429	7
3	1	0	0	1	1	0.6241	0.7467	0.8387	1.0000	2
4	1	0	0	1	2	0.2208	0.5600	0.1429	0.0000	0
5	1	0	1	0	1	0.6735	0.7955	0.6947	0.7143	84
6	1	0	1	0	2	0.2604	0.5690	0.1111	0.1176	17
7	1	0	1	1	1	0.4415	0.7955	0.6947	0.5455	11
8	1	0	1	1	2	0.1189	0.5690	0.1111	0.0000	1
9	1	1	0	0	1	0.6726	0.7143	0.6000	0.6000	10
10	1	1	0	0	2	0.2596	0.6154	0.3333	0.3333	3
11	1	1	0	1	1	0.4405	0.7143	0.6000	0.0000	0
12	1	1	0	1	2	0.1185	0.6154	0.3333	0.0000	0
13	1	1	1	0	1	0.4945	0.6389	0.4419	0.4595	37
14	1	1	1	0	2	0.1431	0.5283	0.0000	0.0000	4
15	1	1	1	1	1	0.2727	0.6389	0.4419	0.3333	6
16	1	1	1	1	2	0.0601	0.5283	0.0000	0.0000	1
17	2	0	0	0	1	0.6966	0.4167	0.6818	0.7143	21
18	2	0	0	0	2	0.2815	0.5000	0.3333	0.3333	9
19	2	0	0	1	1	0.4680	0.4167	0.6818	0.0000	1
20	2	0	0	1	2	0.1306	0.5000	0.3333	0.0000	0
21	2	0	1	0	1	0.5222	0.4000	0.4000	0.4375	32
22	2	0	1	0	2	0.1572	0.4545	0.2000	0.2083	24
23	2	0	1	1	1	0.2953	0.4000	0.4000	0.0000	3
24	2	0	1	1	2	0.0667	0.4545	0.2000	0.0000	1
25	2	1	0	0	1	0.5212	0.5000	0.0000	0.0000	1
26	2	1	0	0	2	0.1567	0.5000	0.3333	0.5000	2
27	2	1	0	1	1	0.2944	0.5000	0.0000	0.0000	0
28	2	1	0	1	2	0.0665	0.5000	0.3333	0.0000	1
29	2	1	1	0	1	0.3414	0.3333	0.3000	0.3333	27
30	2	1	1	0	2	0.0813	0.3333	0.2222	0.2667	15
31	2	1	1	1	1	0.1658	0.3333	0.3000	0.0000	3
32	2	1	1	1	2	0.0328	0.3333	0.2222	0.0000	3

Categorías o celdas ordenadas por *cv1cod*, *edadcod*, *cirugía,fresp* y *fneur*.

Tabla B.7: Tabla de probabilidades estimadas, $\hat{P}(cv2 = buena)$, bajo regresión logística, bajo el modelo gráfico de la fig. 4.1 y bajo el mismo modelo gráfico usando *Hugin 6.3*

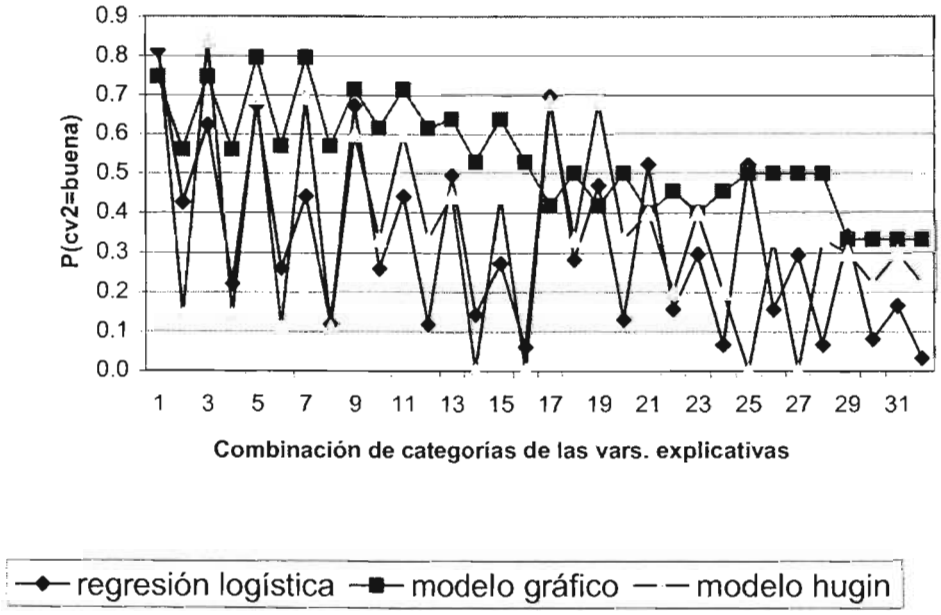

Figura B.2: Probabilidades estimadas, $\hat{P}(cv2 = buena)$, bajo una regresión logística, bajo el modelo gráfico de la fig. 4.1 y bajo el mismo modelo gráfico usando *Hugin 6.3* (orden 1)

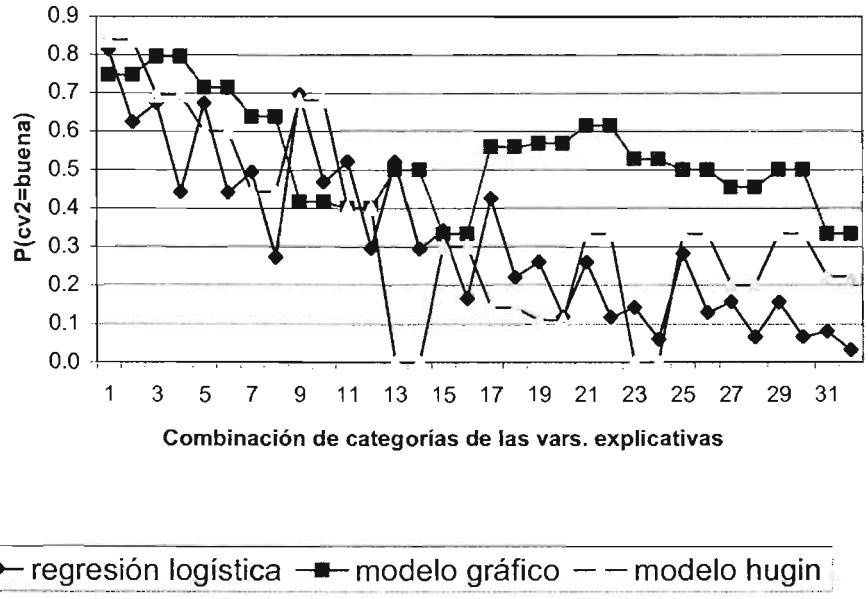

Figura B.3: Probabilidades estimadas, $\hat{P}(cv2 = buena)$, bajo una regresión logística, bajo el modelo gráfico de la fig. 4.1 y bajo el mismo modelo gráfico usando *Hugin 6.3* (orden 2)

Bibliografía

Agresti, A. (2002), "Categorical Data Analysis", 2a ed., John Wiley and Sons, Nueva York.

Bondy J.A. y Murty, U.S.R. (1976) "Graph Theory with Applications", MacMillan Press, Londres.

Bottcher, S. (marzo 2004), "Learning Bayesian Networks with Mixed Variables", Ph. D. Thesis, Universidad de Aalborg, Dinamarca.

Bottcher, S. G. y Dethhlefsen, C. (2003) "DEAL: A Package for Learning Bayesian Networks", Aalborg University, disponible en www.math.auc.dk/research/ resports/reports.htm.

Castillo, E., Gutiérrez, J. M. y Hadi A.S. (1998) "Sistemas Expertos y Modelos de Redes Probabilísticas", Monografías de la Academia Española de Ingeniería, disponible en http://personales.unican.es/gutierjm/BookCGH.html

Cowell, R. G., Dawid, A. P., Lauritzen, S. y Spiegelhalter, D. J. (1999), "Probabilistic Networks and Expert Systems", Springer-Verlarg, Nueva York.

Cox D.R. y Wermuth, N. (1996),"Multivariate Dependencies models, analysis and interpretation", Chapman & Hall, Inglaterra.

Davies, S. y Moore, A. (abril 2000),"Mix-nets: Factored Mixtures of Gaussian in Bayesian Networks with Mixed Continuous and Discrete Variables", Schoool of Computer Science, Carnegie Mellon University.

Geiger, D. y Heckerman D. (1994) "Learning Gaussian networks", reporte técnico MSR-TR-94-10, Microsoft Research.

Harary, F. (1969), "Graph Theory", Adisson-Wesley.

Heckerman, D., Geiger, D. y Chickering, D. (1995), "Learning Bayesian Networks: The combination of knowledge and statistical data", Machine Learning, 20, 197–243.

Hosmer, D. W. (2000),"Applied Logistic Regression", 2a ed., John Wiley and Sons, Nueva York.

Lauritzen, S. L.(1996), "Graphical Models", Oxford University Press.

Mardia, K. V., Kent, J. T., Bibby, J. M. (1979),"Multivariate Analysis", Academic Press.

Mortera, J., Dawid, A. P., Lauritzen, S. L. (2002),"Probabilistic expert systems for DNA mixture profiling", Theoretical Population Biology, 63, 191–205.

Murphy, K. (mayo 2001), "An Introduction to graphical models", Reporte Técnico, Intel Research Technical Report. disponible en http://www.cs.ubc.ca/~murphyk/Papers/intro_gm.pdf

Shachter, R.D. y Kenley, C.R. (1989), "Gaussian influence Diagrams". Management Science, 35, 527-550.

Whittaker, J. (1990), "Graphical Models and Multivariate Statistics", John Wiley and Sons.

Páginas electrónicas

`http://www.math.aau.dk/novo/deal`

`http://www.hugin.dk`

`http://www.cs.ubc.ca/~murphyk/papers.html`

`http://personales.unican.es/castie/#Index`