UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

"ALGUNAS APLICACIONES DE LOS MODELOS DE REGRESIÓN LOGÍSTICA"

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I A

P R E S E N T A :

ADRIANA RAMÍREZ VELAZQUEZ

DIRECTORA DE TESIS: DRA. GUILLERMINA ESLAVA GÓMEZ

2005

DAD NACIONAL
ÓNOMA DE
MÉXICO

ACT. MAURICIO AGUILAR GONZÁLEZ
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito: "Algunas Aplicaciones de los Modelos de Regresión Logística"

realizado por Adriana Ramírez Velázquez

con número de cuenta 09515661-7 , quien cubrió los créditos de la carrera de: Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Dra. Guillermina Eslava Gómez	
Propietario	Dra. Rebeca Aguirre Hernández	
Propietario	Mat. Margarita Elvira Chávez Cano	
Suplente	Act. Francisco Sánchez Villarreal	
Suplente	Act. Víctor Manuel Solís Nájera	

Consejo Departamental de Matemáticas

Act. Jaime Vázquez Alamilla

La elaboración de este trabajo demanda tiempo, trabajo y dedicación; así como la cooperación de algunas personas, en especial tuve la suerte de contar con ellos y aquí expreso mi agradecimiento:

Agradezco al M.C. Luis David Sánchez Velázquez y al Dr. Héctor Ávila Rosas profesor del postgrado en medicina de la U.N.A.M., el apoyo brindado al facilitarme la base de datos para llevar a cabo este trabajo.

A la Dra. Guillermina Eslava por dedicarme tiempo y apoyarme en todo momento

A la Dra. Rebeca Aguirre porque me enseñó todo lo relacionado a la Regresión Logística

A la Mat. Margarita Chávez, al Act. Víctor Manuel Solís y al Act. Francisco Sánchez por haber invertido parte de su tiempo en leer este trabajo

Hay muchas personas a quien me gustaría dedicarles este trabajo pero en especial te lo dedico a ti, María Noé Saavedra que siempre supiste ser la mejor abuela de todas y que sé que siempre vas a estar a mi lado para celebrar todos y cada uno de mis logros, muchas gracias por ser mi abuelita y por estar orgullosa de mi, te prometo que siempre voy a tratar de superarme y a seguir adelante.

Quiero aprovechar para agradecer a mis padres por su comprensión y apoyo y que sepan que sin ellos no hubiera podido llegar tan lejos, a mis hermanas Karla y Andrea por estar siempre conmigo, a mi abuelito y a mi tía Lupe porque ellos siempre me explicaban cuando no le entendía a algo y a mis tíos Luis, Leticia, Miriam y Raúl por darme consejos y brindarme su amistad, a mis primos: Luis, Arturo, José, Rodrigo y Alfonso

Finalmente quiero darle gracias a los amigos que siempre han estado conmigo no sólo en los buenos momentos sino también en los malos, Mónica, Ramón, Carlos, Pedro, Paty, Brenda, Sergio, Gustavo, Raúl, Gaby, Guadalupe, Marilú, los del museo, los del inglés, los matemáticos, etc., y a todas aquellas personas que han estado a mi lado y que sé que puedo contar con ellas.

INDICE

Resumen

El objetivo de esta tesis es ajustar un modelo de regresión logística tricotómico y algunos binarios a información proveniente de un estudio médico, con el fin de ayudar a explicar algunos de los factores de riesgo asociados a la calidad de vida (buena, mala o murió) de pacientes egresados de terapia intensiva.

En la primera parte de este trabajo se expone la teoría necesaria y suficiente concerniente al modelo de regresión logística tanto binaria como multinomial, para aplicarla en un estudio clínico con el fin de evaluar cuales son los factores que permiten pronosticar el nivel de estado vital en enfermos con sepsis grave hospitalizados en terapia intensiva. Dicho estudio se llevó a cabo en los hospitales: Centro Médico Nacional Siglo XXI y Centro Médico Nacional La Raza del 1° de Mayo del 2002 al 31 de Septiembre del 2004.

En la segunda parte, se explica la metodología médica (criterios de exclusión e inclusión, población en estudio, descripción de las variables clínicas, etc.) y se muestran los modelos ajustados de regresión logística binaria y multinomial, también se muestran los modelos ajustados de regresión lineal múltiple y regresión poisson ya que se decidió analizar que factores podían estar relacionados a los días de estancia en terapia intensiva y esta es una variable que puede considerarse continua, aunque también puede tratarse como una variable discreta.

Cabe señalar que este problema ha sido tratado en la tesis de Medrano Ortiz (2004) para el caso de pacientes sobrevivientes a la terapia intensiva; en el presente trabajo contemplamos a pacientes sobrevivientes y no sobrevivientes.

Se hizo un análisis exploratorio de las variables continuas por medio de gráficas de dispersión mientras que para las variables categóricas se usaron tablas cruzadas. Se calcularon las frecuencias de las variables para eliminar a aquellas que tuvieran un porcentaje muy bajo de casos válidos. Para saber cuales de las variables estaban relacionadas entre sí, se hizo un análisis de componentes principales y a partir de la matriz de correlación se identificó cuales eran las variables continuas que estaban altamente correlacionadas.

Después de lo anterior, se procedió a ajustar varios modelos de regresión logística usando los paquetes estadísticos SPSS y STATA.
Algunos de los métodos de selección de variables que se usaron para auxiliarse en la elección del modelo adecuado son: el de forward, backward y la estadística de bondad de ajuste de Hosmer y Lemeshow. Este trabajo no se limita solamente a la estimación de los parámetros para el modelo, también se realiza la validación de supuestos, la evaluación de la bondad de ajuste del modelo, análisis de los residuos y finalmente se evalúa la capacidad predictiva de los modelos propuestos para el caso binario, tricotómico, regresión lineal múltiple y regresión poisson.

Finalmente se propone un modelo plausible que ayuda a explicar los factores que influyen en el grado de la calidad de vida de los pacientes después de ser sometidos a terapia intensiva. El modelo sin embargo no tiene un poder predictivo bueno juzgado por la tabla de clasificación de las observaciones con dicho modelo.

Introducción

Casi todos los campos de la investigación científica se pueden beneficiar del análisis estadístico, en la investigación de mercados, la estadística representa una ayuda inestimable para determinar si es probable que un nuevo producto pueda tener éxito. Incluso el investigador médico, preocupado por la eficacia de un nuevo medicamento, si un sujeto operado se infecta o no durante cierto lapso postoperatorio o si un paciente hospitalizado muere o no antes del alta.

En situaciones como las mencionadas, suele interesar a los investigadores la evaluación del efecto de uno o más antecedentes sobre el hecho de que el acontecimiento se produzca.

El propósito de este trabajo es mostrar algunas de las aplicaciones de los modelos de regresión logística por medio de datos reales de manera tal, que pueda ser tomado en cuenta como una herramienta útil y se ilustre el concepto de la regresión logística de forma clara y concisa.

Los métodos de regresión se han convertido en un componente integral de cualquier análisis de datos enfocado a describir la relación entre una variable de respuesta y una o más variables explicativas.

Es muy frecuente el caso en que la variable de respuesta es discreta, tomando dos o más valores posibles. En la última década el modelo de regresión logística se ha convertido, en muchos campos, en un método estándar del análisis.

En el presente trabajo se expone, en el primer capítulo, la teoría referente al modelo de regresión logística. En la primera sección se presenta el modelo y se explica el método de máxima verosimilitud que sirve para estimar los parámetros, se mencionan los métodos para la selección de variables y se muestran algunos métodos de bondad de ajuste. Finalmente se expone la teoría referente al modelo de regresión logística multinomial (respuesta ordinal y respuesta nominal).

En el segundo capítulo se presentan las aplicaciones de los modelos logísticos: binario y tricotómico así como la explicación teórica del modelo de regresión múltiple y su aplicación.

En la última parte se encuentran los anexos, en el Anexo I se encuentran algunas gráficas de dispersión para las variables continuas y categóricas, histogramas y tablas cruzadas. En el Anexo II se encuentran los comandos que fueron utilizados para ajustar algunos de los modelos de regresión logística en SPSS y en STATA.

Capítulo 1. Regresión Logística

La regresión logística es parte de una familia de modelos estadísticos llamados Modelos Lineales Generalizados (MLG). Esta amplia familia de modelos incluye regresión ordinaria y Análisis de la Varianza (ANOVA), así como modelos loglineales. Lo que distingue un modelo de regresión logística de un modelo de regresión lineal es que la variable de respuesta en regresión logística es binaria o multinomial.

La regresión logística es un método de análisis adecuado cuando se requiere modelar una variable de respuesta binaria o multinomial, por ejemplo, del tipo presencia o ausencia de enfermedad, y permite el uso de un conjunto de covariables de tipo categórico y continuo, permitiendo la interpretación a sus parámetros (Hosmer & Lemeshow, 1989).

En ella se suele simbolizar con Y a la variable respuesta, del tipo presencia ($Y = 1$) o ausencia ($Y = 0$) de enfermedad y con $\pi(X)$ a la siguiente probabilidad: $P(Y = 1 \mid \mathbf{X})$, donde $\mathbf{X}$ es un vector de k covariables.

Antes de empezar un estudio de regresión logística es importante entender que el objetivo en un análisis usando este método es el mismo que en cualquier otra técnica usada en estadística: encontrar el mejor ajuste y el modelo más parsimonioso que describa la relación entre una variable de respuesta o dependiente y un grupo de variables explicativas o independientes. Estas variables independientes son llamadas con frecuencia covariables.

Llamemos Y a la variable dependiente, que refleja la ocurrencia o no del suceso. Puesto que Y es dicotómica, puede asumir los dos valores siguientes.

Y = 1 si el hecho ocurre
Y = 0 si el hecho no ocurre

De estas consideraciones se desprende una advertencia clara. Al interpretar los coeficientes de las variables es imprescindible tener en cuenta cómo se ha definido a la variable de respuesta: un coeficiente con signo positivo indica que (P[Y=1]) crece cuando lo hace

la variable, pero el sentido cualitativo de este hecho depende, desde luego, de lo que representen tanto la variable en cuestión como el suceso Y=1.

Lo que se procura mediante la regresión logística es, en principio, expresar la probabilidad de que ocurra el hecho en cuestión como función de ciertas variables (supongamos que son *k*) que se presumen relevantes o influyentes en dicho suceso.

La forma analítica en que esa probabilidad se vincula con las variables explicativas se expone a continuación.

El caso más simple es aquel en que se incluye una sola variable independiente:

$$P(Y=1) = \frac{1}{1+e^{(-\beta_0-\beta_1 X)}} \qquad [1.1]$$

Equivalentemente

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X$$

El caso más general es el siguiente:

$$P(Y=1) = \frac{1}{1+e^{(-\beta_0-\beta_1 X_1-\beta_2 X_2-.....-\beta_k X_k)}} \qquad [1.2]$$

Equivalentemente

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Donde β_0, β_1, β_2,..... , β_k son los parámetros del modelo, y donde *exp* denota la función exponencial. La expresión [1.2] es lo que se conoce como la función logística y [1.1] es su versión univariada. Al construir el modelo de regresión logística, las variables explicativas pueden ser de cualquier naturaleza: dicotómicas, ordinales, continuas o nominales.

1.1 Modelos Lineales Generalizados

Como dijimos anteriormente, el modelo logístico pertenece a la familia de los Modelos Lineales Generalizados. Todos ellos se caracterizan por ser lineales a alguna transformación de la esperanza de la variable respuesta cuya distribución pertenece a la familia exponencial.

Están determinados por tres componentes:

1.- Aleatoria. Identifica a la variable respuesta y a su distribución de probabilidad. Supóngase que $y_1, y_2, ..., y_n$ es una muestra de n variables independientes con función de densidad de la familia exponencial, cuya forma es la siguiente:

$$f(y_i ; \theta_i) = a(\theta_i)\, b(y_i) \exp[y_i\, Q(\theta_i)] \qquad [1.3]$$

Donde $Q(\theta_i)$ es el parámetro natural.

2.- Sistemática. También se le conoce como predictor lineal de las variables explicativas. Las variables explicativas $X_1, X_2, ..., X_k$ pueden ser cualitativas o cuantitativas. El vector de parámetros β_k es desconocido.

$$\eta_i = \beta_0 + \sum_{j=1}^{k} \beta_{ij} X_{ij} \; i = 1,2,....,n \qquad [1.4]$$

3.- Función Liga. Relaciona la componente aleatoria, E (y_i) = μ_i, con la sistemática, η_i.

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} \qquad [1.5]$$

g (μ_i) es una transformación de la esperanza expresada como una función lineal de los parámetros $\beta_0, \beta_1, \ldots, \beta_k$ y además es una función monótona diferenciable.

A la función liga g (μ_i) = μ_i se le llama liga identidad. Esta es la función liga para la regresión lineal con distribución Normal en la variable respuesta. La función liga que transforma la media en el parámetro natural se le llama función canónica g (μ_i) =Q (θ_i). (Agresti, 2000. p. 116).

Algunos ejemplos de los modelos lineales generalizados son: regresión lineal, loglineal, poisson y regresión logística (binario y multinomial).

1.2 El Modelo de Regresión Logística Múltiple

El modelo de regresión logística múltiple está dado por: $g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$

donde:

$$g(\mu) = \ln \frac{P[Y=1 \mid X]}{1 - P[Y=1 \mid X]}$$

A cada individuo se le mide una variable de respuesta Y y una o más variables explicativas X_k. El objetivo es describir y/o predecir el valor de Y a partir de las variables explicativas. La variable de respuesta es aleatoria y por lo tanto tiene una distribución de probabilidad. Las variables explicativas no se consideran aleatorias, es decir, su valor está fijo.

Se tienen los siguientes casos:

* Si Y es continua generalmente se supone $Y_i \sim$ Normal (μ_i, σ^2) y el modelo que describe la relación entre la variable de respuesta y las variables explicativas es el modelo de regresión lineal.

* Si Y es discreta se supone que $Y_i \sim$ Poisson (λ_i) y en este caso, la relación entre la variable de respuesta y las variables explicativas se describe mediante el modelo de regresión Poisson.

* Si Y es binaria o dicotómica entonces $Y_i \sim$ Bernoulli (P_i) dónde $P_i = P(Y_i = 1; X_i)$ y un modelo que describe la relación entre la variable de respuesta y la variable explicativa es el modelo de regresión logística binaria.

* Si Y es multinomial entonces $Y_i \sim$ Multinomial (n, p) y el modelo que describe la relación entre la variable de respuesta y las variables explicativas es el modelo de regresión logística multinomial.

En el proceso de selección del modelo se tienen en cuenta los siguientes aspectos:

a) análisis univariado de las variables;
b) selección de escala apropiada para las variables continuas, para elegir la expresión funcional más lógica para el modelo; y
c) consideración de términos de interacción, basada tanto en criterios estadísticos como del problema en cuestión.

1.2.1 Método de Máxima Verosimilitud

Para el ajuste del modelo [1.1] y la estimación de los parámetros β_0 y β_1 no puede seguirse el procedimiento de mínimos cuadrados como en el caso de la regresión lineal. Una alternativa de uso general para la estimación de los parámetros consiste en utilizar el procedimiento de estimación por máxima verosimilitud. En síntesis este método proporciona valores estimados ($\hat{\beta}_0$ y $\hat{\beta}_1$) para los parámetros desconocidos β_0 y β_1 que maximizan la probabilidad de que con ellos se obtengan los valores observados. Para aplicar este método, se precisa construir, en primer lugar, la denominada función de verosimilitud (L) que expresa la probabilidad de los datos observados como una función de parámetros desconocidos. Los valores que maximizan la función L serán los estimadores maximoverosímiles de dichos parámetros. Así pues, una vez ajustado el modelo y obtenidos los estimadores maximoverosímiles $\hat{\beta}_0$ y $\hat{\beta}_1$, la estimación de la probabilidad es inmediata:

$$\hat{P}(Y=i \mid X)=\frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_i}}$$

$$(1-\hat{P}(Y=i \mid X))=\frac{1}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_i}}$$

La muestra es aleatoria y las observaciones son independientes entre sí, la probabilidad de que un sujeto de la muestra experimente el suceso es independiente de lo que le ocurra a cualquier otro, por lo que la probabilidad conjunta se calcula como el producto de las probabilidades individuales y de esa forma obtenemos la función de verosimilitud, que tiene en cuenta todos los datos de forma global, y será función únicamente de los coeficientes.

Se calcula la derivada de esa función, se iguala a cero y se obtienen los valores de los coeficientes que maximizan esa función. Aunque esto que se dice fácil, al menos en el modelo logístico, es algo más complicado de efectuar ya que las ecuaciones verosímiles no son lineales a los parámetros por lo que se resuelven por métodos numéricos iterativos que por lo general están incluidos en los paquetes estadísticos.

Para terminar esta visión introductoria al modelo de regresión logística es necesario definir dos conceptos básicos relacionados con el mismo y que serán de suma utilidad para su completa comprensión.

El primero son los "odds" o momios. Se define como la razón entre la probabilidad de que ocurra un suceso y su probabilidad complementaria, esto es, de que no ocurra:

$$momios \quad = odds \quad = \frac{\mathrm{P}(Y = i)}{1 - \mathrm{P}(Y = i)}$$

Donde

$$\frac{\mathrm{P}(Y = i)}{1 - \mathrm{P}(Y = i)} = e^{\beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}}$$

e indica la preferencia de elegir la opción 1 de la variable de respuesta frente a la opción 0. También se le conoce como momio de la probabilidad de éxito.

La probabilidad de éxito cuando la variable independiente toma el valor de x es:

$$P(Y = 1 \mid X = x)$$

Entonces los momios de presentar cierta característica para una persona con X=1 es:

$$\frac{\mathrm{P}(Y=1 \mid X=1)}{1-\mathrm{P}(Y=1 \mid X=1)} \qquad [1.6]$$

Mientras que los momios de presentar cierta característica para una persona con X=0 es:

$$\frac{\mathrm{P}(Y=1 \mid X=0)}{1-\mathrm{P}(Y=1 \mid X=0)} \qquad [1.7]$$

El modelo de regresión logística se interpreta en términos de razón de momios.

A continuación se presentan la Razón de Momios que se define como el cociente de los momios [1.6] y [1.7] y la tabla que muestra la probabilidad condicional de Y para cada valor de X.

$$RM = \frac{Momios(Y=1 \mid X=1)}{Momios(Y=1 \mid X=0)} = \frac{\dfrac{\mathrm{P}(Y=1 \mid X=1)}{1-\mathrm{P}(Y=1 \mid X=1)}}{\dfrac{\mathrm{P}(Y=1 \mid X=0)}{1-\mathrm{P}(Y=1 \mid X=0)}} \qquad [1.8]$$

Tabla 1. Probabilidades Condicionales Calculadas a Partir del Modelo Logístico cuando la variable explicativa es binaria

	X = 1	X = 0
Y = 1	$P(Y=1 \mid X=1)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$P(Y=1 \mid X=0)=\frac{e^{\beta_0}}{1+e^{\beta_0}}$
Y = 0	$1-P(Y=1 \mid X=1)=\frac{1}{1+e^{\beta_0+\beta_1}}$	$1-P(Y=1 \mid X=0)=\frac{1}{1+e^{\beta_0}}$

Sustituyendo los valores de la tabla anterior en la ecuación [1.8] se obtiene lo siguiente:

$$RM=\frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\div\frac{1}{1+e^{\beta_0+\beta_1}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}}\div\frac{1}{1+e^{\beta_0}}}=\frac{\frac{(e^{\beta_0+\beta_1})(1+e^{\beta_0+\beta_1})}{1+e^{\beta_0+\beta_1}}}{\frac{(e^{\beta_0})(1+e^{\beta_0})}{1+e^{\beta_0}}}=\frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}=e^{\beta_1}$$

Por definición los momios son no negativos pero lo interesante de sus valores es saber si son mayores o menores a uno.

Si los momios son mayores a 1, indica que la probabilidad de éxito es mayor a la de fracaso. Sin embargo, si los momios son menores a 1 esto implica que la probabilidad de fracaso es mayor a la del éxito.

Como $RM=e^{\beta_1}$ entonces la relación que existe entre la razón de momios y el coeficiente de regresión, es la exponencial del coeficiente. Para obtener más información del valor del parámetro se recomienda usar intervalos de confianza para la razón de momios.

Los puntos extremos del intervalo de confianza se calculan de acuerdo a la siguiente expresión:

$$\exp\left\{\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}} S\hat{E}\left(\hat{\beta}_j\right)\right\}$$

El segundo concepto es el de la transformación logística, definida como el logaritmo de la probabilidad o preferencia de la opción 1 frente a la opción 0.

$$\ln\left(\frac{\mathrm{P}(Y=1)}{1-\mathrm{P}(Y=1)}\right) = \beta_0 + \beta_1 X_1$$

Hay que destacar que mientras que la probabilidad se expresa a través de un modelo no lineal (logístico), el logaritmo de la razón de las probabilidades sí lo es, lo cual facilita la interpretación del modelo.

1.3 Selección de Variables

Uno de los objetivos principales de cualquier método es elegir aquellas variables que resulten en el mejor modelo dentro del contexto del problema. El criterio para incluir una variable en un modelo puede variar de un problema a otro y de una disciplina a otra. El método tradicional para construir un modelo estadístico involucra la búsqueda del modelo más parsimonioso que siga explicando la relación entre las variables en base a los datos. Entre más variables se incluyan en un modelo, habrá más parámetros que estimar.

Hay ciertos pasos que podemos seguir para facilitar la selección de variables al construir un modelo de regresión logística.

El proceso de selección debe empezar con un cuidadoso análisis univariado de cada variable. La selección de las variables a incluir en el modelo como predictoras del fenómeno que se desea estudiar ha de realizarse siguiendo dos criterios que no tienen por qué conducir a resultados coincidentes: modelización estadística y modelización sustantiva. Según el criterio estadístico tan sólo se incluirán en el modelo aquellas variables que tienen una capacidad de predicción estadísticamente significativa, es decir, que contribuyan a la mejora de la bondad del ajuste del modelo. En el criterio sustantivo, el investigador decide qué variables debe incluir en función de la base teórica en la que se apoya la hipótesis de investigación que se pretende verificar. (Luque Martínez, Pág.458).

Una vez completado el análisis univariado, elegimos variables para el análisis multivariado. Cualquier variable cuya prueba univariada tenga un p-valor < 0.25 deberá ser considerada como candidata para el modelo multivariado junto con todas las variables de importancia biológica conocida. Una vez que las variables han sido identificadas, empezamos con un modelo conteniendo a todas las variables elegidas.

Siguiendo con el ajuste del modelo multivariado, la importancia de cada variable incluida en el modelo debe ser verificada con una prueba de la estadística de Wald para cada variable y una comparación de cada coeficiente estimado con el coeficiente del modelo univariado que contiene solamente a esa variable.

1.3.1 Las Variables Cualitativas en el Modelo Logístico

Cuando algunas de las variables explicativas son de índole nominal, de más de 2 categorías (politómicas), para incluirlas en el modelo hay que darles un tratamiento especial.

La asignación de un número a cada categoría no resuelve el problema ya que si tenemos, por ejemplo, la variable ejercicio físico con tres posibles respuestas: sedentario, realiza ejercicio esporádicamente, realiza ejercicio frecuentemente, y le asignamos los valores 0, 1, 2, significa a efectos del modelo, que efectuar ejercicio físico frecuentemente es dos veces mayor que solo hacerlo esporádicamente, lo cual no tiene ningún sentido. Más absurdo sería si se trata, a diferencia de ésta, de una variable nominal, sin ninguna relación de orden entre las respuestas, como puede ser el estado civil.

La solución a este problema es crear tantas variables dicotómicas como número de categorías menos *1*. Estas nuevas variables, artificialmente creadas, reciben en la literatura anglosajona el nombre de "*dummy*", traduciéndose en español con diferentes denominaciones como pueden ser variables indicadoras, o binarias.

Así por ejemplo si la variable en cuestión recoge datos de tabaquismo con las siguientes respuestas: *Nunca fumó*, *Ex-fumador*, *Actualmente fuma menos de 10 cigarrillos diarios*, *Actualmente fuma 10 o más cigarrillos diarios*, tenemos 4 posibles respuestas por lo que construiremos 3 variables internas dicotómicas (valores 0,1), existiendo diferentes posibilidades de codificación, que conducen a diferentes interpretaciones, y siendo la más habitual la siguiente:

	I1	**I2**	**I3**
Nunca fumó	0	0	0
Ex- fumador	1	0	0
Menos de 10 cigarrillos diarios	0	1	0
10 o más cigarrillos diarios	0	0	1

En este tipo de codificación el coeficiente de la ecuación de regresión para cada variable binaria (siempre transformado con la función exponencial), corresponde a la razón de momios de esa categoría con respecto al nivel de referencia (la primera respuesta), en nuestro ejemplo cuantifica cómo cambia el riesgo respecto a no haber fumado nunca.

Existen otras posibilidades entre las que se destaca con un ejemplo para una variable cualitativa de tres respuestas:

	I1	I2
Respuesta 1	0	0
Respuesta 2	1	0
Respuesta 3	1	1

Con esta codificación cada coeficiente se interpreta como una media del cambio del riesgo al pasar de una categoría a la siguiente.

En el caso una categoría que NO pueda ser considerada de forma natural como nivel de referencia, como por ejemplo el grupo sanguíneo, un posible sistema de clasificación es:

	I1	I2
Respuesta 1	-1	-1
Respuesta 2	1	0
Respuesta 3	0	1

Dónde los coeficientes de los contrastes I1 e I2 tienen una interpretación directa como cambio en el riesgo con respecto a la media de las tres respuestas.

1.3.2 Variables Ordinales

En el caso de las variables ordinales se puede asumir que la escala funciona aproximadamente a un nivel cuantitativo, desde luego, tal maniobra presupone que se considere que la "distancia" entre categorías contiguas es la misma. En el caso contrario, las variables ordinales pueden manejarse del mismo modo que se ha explicado para las nominales, o sea como variables binarias o como contrastes.

En el ejemplo que se describe en el Capítulo 2, algunas de las variables, tanto nominales como ordinales, fueron introducidas en el modelo como variables categóricas, con la consiguiente formación de variables binarias.

1.3.3 Interacciones

En ocasiones se piensa que la influencia de una de las variables sobre la probabilidad de que ocurra el hecho se modifica en función del valor de otra de las variables y es necesario incluir en el modelo una tercera que sea el producto de las anteriores. Estos son los conocidos como términos de interacción que pueden incluir 2 o más variables.

Introducimos términos de interacción cuando tenemos razones para suponer que la influencia de una de las variables sobre la probabilidad de éxito varía en función del valor que asume otra de las variables incluidas en el modelo; o sea, si la influencia de X_1 sobre la probabilidad de éxito varía en función del valor que toma X_2, incluimos en el modelo un término que represente la interacción de X_1 y X_2.

1.4 Selección de modelos

Al estar hablando de modelos de regresión múltiple, un aspecto de interés es cómo seleccionar el mejor conjunto de variables independientes a incluir en el modelo.

La definición de mejor modelo depende del tipo y el objetivo del estudio. En un modelo con finalidad predictiva se considerará como mejor modelo aquél que produce predicciones más fiables, mientras que en un modelo que pretende estimar la relación entre dos variables

(corrigiendo el efecto de otras, como se vio anteriormente), se considerará mejor aquél con el que se consigue una estimación más precisa del coeficiente de la variable de interés. Esto se olvida a menudo y sin embargo conduce a estrategias de modelado completamente diferentes. Así en el segundo caso una covariable con coeficiente estadísticamente significativo pero cuya inclusión en la ecuación no modifica el valor del coeficiente de la variable explicativa de mayor interés, será excluida de la ecuación, ya que no se trata de un factor de confusión: la relación entre la variable explicativa de interés y la probabilidad no se modifica si se tiene en cuenta esa covariable.

Otra consideración que hay que hacer siempre que se analizan datos es distinguir entre diferencias numéricas, diferencias estadísticamente significativas y diferencias clínicamente relevantes. No siempre coinciden los tres conceptos.

Lo primero que se habrá de plantear es el modelo máximo, o lo que es lo mismo el número máximo de variables dependientes que pueden ser incluidas en la ecuación, considerando también las interacciones si fuera conveniente.

Aunque existen diferentes procedimientos para escoger el modelo sólo hay tres mecanismos básicos para ello: empezar con una sola variable dependiente e ir añadiendo nuevas variables según un criterio prefijado (procedimiento hacia adelante) *forward*, o bien empezar con el modelo máximo e ir eliminando de él variables según un criterio prefijado (procedimiento hacia atrás) *backward*. El tercer método, denominado en la literatura *stepwise*, combina los dos anteriores y en cada paso se puede tanto añadir una variable como eliminar otra que ya estaba en la ecuación.

En el caso de la regresión logística el criterio para decidir en cada paso si escogemos un nuevo modelo frente al actual viene dado por el logaritmo del cociente de verosimilitudes de los modelos.

La función de verosimilitud de un modelo es una medida de cuán compatible es éste con los datos observados. Si al añadir una nueva variable al modelo no mejora la verosimilitud de forma apreciable, en sentido estadístico, ésta variable no se incluye en la ecuación. Dicha función nos permite comparar modelos, por ejemplo dos modelos uno de ellos incluye una variable adicional con respecto al primer modelo.

Las diferencias en la función de verosimilitud se alteran arbitrariamente con la escala de medida, por lo que la forma adecuada de compararlas es mediante cocientes. De ahí que cuando se comparan modelos que han sido estimados mediante este procedimiento se hable de cociente de verosimilitudes.

También en las salidas de los programas suele aparecer el término *likelihood ratio* o cociente de verosimilitudes para un modelo, sin que se especifique que se esté contrastando con otro diferente. En estos casos el contraste es frente al modelo que sólo incluye el término constante y por tanto no se consideran las variables X o los factores de riesgo, y se compara con el modelo que sí incluye las variables.

Dicho cociente de verosimilitudes se distribuye asintóticamente según una χ^2 con grados de libertad igual al número de variables incluidas en el modelo, que es la diferencia frente al modelo con solo la constante. Al igual que antes, si el contraste resulta no significativo pensamos que incluir a las variables X no mejora significativamente la verosimilitud del modelo y por lo tanto se trata de un modelo sin utilidad.

Para evaluar la significancia estadística de una variable concreta dentro del modelo, nos fijaremos en el valor de χ^2 (estadística de Wald) correspondiente al coeficiente de la variable y en su desviación estándar.

1.4.1 Estadística de Wald

Esta estadística juega el mismo rol que la estadística T en el análisis de regresión lineal múltiple. Permite contrastar las hipótesis de que los parámetros del modelo son iguales a cero.

Para cualquier variable independiente X_j seleccionada, si β_j es el parámetro asociado a X_j en la ecuación de regresión logística, la estadística de Wald permite contrastar la siguiente hipótesis nula:

H_0: $\beta_j = 0$

Vs

H_1: $\beta_j \neq 0$

La estadística de prueba está dada por: $$W = \frac{\hat{\beta}_j}{S\hat{E}(\hat{\beta}_j)}$$

Bajo la hipótesis nula, la estadística de Wald tiene una distribución asintótica χ^2 con un grado de libertad.

Donde $S\hat{E}$ es la desviación estándar asintótica estimada de $\hat{\beta}_j$.

La interpretación de dicha hipótesis es que la información que se perderá al eliminar o no incluir a la variable X_j en el siguiente paso, no es significativa. La variable a ser eliminada o ignorada será la que presente un mayor p-valor.

1.4.2 Algoritmos de selección de modelos

Añadiendo más términos, más variables, a un modelo la función de verosimilitud aumentará de valor y si la muestra es grande será difícil distinguir mediante el contraste del cociente de verosimilitud entre una mejora "real" y una aportación trivial. El modelo perfecto no existe, puesto que todos constituyen simplificaciones de la realidad y siempre son preferibles modelos con menos variables, puesto que además de ser más sencillos, son más estables y menos sometidos a sesgo.

Existen, esencialmente, dos clases de algoritmos automatizados para llevar a cabo la selección:

Los métodos de selección hacia delante y los de eliminación hacia atrás.

1.4.3 Algoritmos de selección de variables hacia adelante (forward)

Comienzan con el modelo que tiene como única variable explicativa el término constante β_0. En cada paso del algoritmo entra en la ecuación del modelo aquella variable con el menor p-valor calculado utilizando uno de los tres criterios de selección de variables expuestos anteriormente y así sucesivamente hasta que entran todas las variables significativas al modelo.

1.4.4 Algoritmos de eliminación de variables hacia atrás (backward)

Comienza con el modelo que tiene todas las variables en la ecuación de regresión. En cada paso elimina de la ecuación del modelo aquella variable con un coeficiente β_i que no es significativamente distinto de 0 utilizando alguno de los criterios expuestos anteriormente.

1.4.5 Bondad de ajuste del modelo

Por medidas de bondad de ajuste hemos de entender aquellas pruebas que evalúen el grado de efectividad del modelo considerado en cuanto a la descripción de la variable dependiente, es decir, cuán cerca están los valores estimados $\hat{y}_i$ de los realmente observados. Analizaremos tres grupos de medidas de bondad de ajuste: las basadas en pruebas estadísticas de contraste de hipótesis, las derivadas de la comparación directa entre los valores estimados y observados de la variable de respuesta y, por último, las que son análogas al coeficiente de determinación múltiple (R^2) de la regresión lineal. (Luque Martínez, 2000).

1.4.5.1 Bondad del Ajuste: Contraste de Hipótesis.

Este tipo de medidas de bondad de ajuste se basa en contrastar la hipótesis nula H_0 de que el modelo seleccionado ajusta bien los datos por medio de un estadístico con una distribución conocida.

El ajuste del modelo es bueno si:

- Las distancias entre los valores de la variable de respuesta observada con respecto a la ajustada son pequeñas.

a) Devianza

El estadístico devianza *(D)* se define como una función del logaritmo natural del cociente de la función de verosimilitud del modelo seleccionado y la del modelo saturado. Un modelo saturado es aquel que contiene tantos parámetros como datos y que predice perfectamente los valores observados. La devianza tiene la siguiente expresión:

$$D = -2\ln\left\{\frac{verosimilitud\ del\ modelo\ selecciona do}{verosimilitud\ del\ modelo\ saturado}\right\}$$

La cantidad entre corchetes se denomina razón de verosimilitudes y el propio estadístico *D* es también llamado prueba de razón de verosimilitudes. En algunas pruebas y paquetes estadísticos se suele utilizar *-2 ln L(β)* o *-2 ln likelihood* para referirse a la devianza de un determinado modelo.

Contrastes de Bondad de Ajuste

Se utilizan, entre otros, el estadístico de la devianza y el contraste de bondad de ajuste de Hosmer-Lemeshow.

A través de estas estadísticas se contrastan las hipótesis:

H_0: El modelo describe adecuadamente a los datos

$$logit\ (p_i) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$$

Vs

H_1: El modelo no describe adecuadamente a los datos

$$logit(p_i) \neq \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$$

La estadística de bondad de ajuste basada en los residuos de la devianza es:

$$D = 2\sum_{i=1}^{n} \left\{ y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right\}$$

Dónde $\hat{y}_i = n_i \hat{p}_i$

Bajo el supuesto de H_0 cierta y de n_i grandes en un número fijo de niveles o categorías en las variables explicativas se distribuye asintóticamente como una χ^2 con grados de libertad igual a la diferencia del número de patrones de covariables (número de combinaciones de valores observados de las variables explicativas) y el número de parámetros estimados. (Tomas P. Ryan. 1997. Pág.267). Cabe señalar que esta estadística se puede usar cuando el número de patrones de covariables es pequeño comparado con el número total de

individuos, lo cual generalmente sucede cuando todas las variables explicativas son cualitativas. En el caso de que el número total de patrones de covariables es aproximadamente igual al número total de individuos estudiados, se propone agrupar los datos ajustados y construir la estadística $\hat{C}$, (Hosmer y Lemeshow, 2000. p.147).

D es el origen de la prueba *G* de contraste de hipótesis para la significancia conjunta de todas las variables incluidas en el modelo. Este último estadístico no hace sino recoger el cambio en *D* debido a incluir las variables independientes con respecto al modelo que sólo contiene el término constante. Así:

G = D (modelo sin las variables, sólo con la constante) – D (modelo con las variables)
Puesto que la verosimilitud del modelo saturado es la misma en ambos valores de D, la diferencia puede expresarse como:

$$G = -2\ln\left\{\frac{verosimilitud\ del\ \text{mod}elo\ \text{sin}\ \text{var}iables\ (sólo\ con\ la\ cons\,\text{tan}\,te)}{verosimilitud\ del\ \text{mod}elo\ con\ las\ \text{var}iables}\right\}$$

b) Prueba de la χ^2

Tanto esta prueba como la siguiente son medidas de bondad de ajuste que se basan en comparar los valores observados y los estimados por el modelo que se desea evaluar (valores esperados), todo ello, una vez más, bajo la hipótesis H_0 de que dicho modelo ajusta bien a los datos observados.

Esta prueba se basa en la obtención de un estadístico χ^2 que mide el nivel de discordancia que puede existir al comparar, para cada uno de los diferentes patrones de covariables existentes, el número de respuestas (afirmativas) observadas con el número de éxitos estimados por el modelo. Por patrón de covariables se entiende cada una de las diferentes combinaciones de valores que pueden adoptar las variables independientes incluidas en el modelo. Por ejemplo, las variables SEXO (1 = hombre; 0 = mujer) y ESTADO CIVIL (1 = soltero; 0 = casado) determinarían cuatro patrones de covariables, puesto que cada uno

de los individuos que componen la muestra pueden clasificarse en uno de los siguientes grupos: hombre-soltero, hombre-casado, mujer-soltera y mujer-casada. La estadística χ^2 de Pearson generalizada para el modelo logístico es:

$$\chi^2 = \sum_{i=1}^{n} r_i^2$$

Donde $$r_i = \frac{y_i - n_i \hat{p}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}}$$

n_i = número de individuos en el patrón de covariables i.

$Y = y_1, y_2, \ldots y_k$

Donde y_i es el número de éxitos observados en el patrón de covariables i y $\hat{P}_i$ es la probabilidad estimada por el modelo para el patrón de covariables i.

Dónde $\hat{P}_i = \hat{P}(Y = y_i)$

c) Prueba de Hosmer – Lemeshow

Esta prueba es especialmente adecuada para evaluar la bondad de ajuste de aquellos modelos que incluyan una o más variables independientes de tipo continuo y que cuenten con un número de patrones de covariables prácticamente igual al número de casos observados ($M \approx N$).

Estos autores proponen ordenar de menor a mayor las N probabilidades estimadas (una para cada caso observado) y a continuación agruparlas en g grupos de tal modo que en el primero de ellos se encuentren los $n_i = N/g$ sujetos con las probabilidades estimadas más pequeñas. Todos los grupos deben tener aproximadamente el mismo número de individuos. Generalmente se trabaja con $g = 10$ grupos (el mínimo requerido es $g = 6$ grupos). Estos grupos son conocidos como deciles de riesgo.

El estadístico de bondad de ajuste de Hosmer – Lemeshow $\hat{C}$, se obtiene calculando el estadístico ji-cuadrado de Pearson de una tabla de $2xg$ referida a las frecuencias observadas y estimadas para cada uno de los g grupos.

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n_k \hat{P}_k)^2}{n_k \hat{P}_k (1 - \hat{P}_k)}$$

Dónde:

n_k = Número de individuos del grupo k. $k = 1,2,3,\ldots..., g$

$\hat{P}k$ = Promedio de las probabilidades estimadas de los individuos del grupo k.

O_k = Número observado de éxitos en el grupo k.

N = Casos observados.

M = Número de patrones de covariables.

Se rechaza H_0 si $\hat{C} > \chi^2_{(g-2)}$. Cuando H_0 es cierta, entonces $\hat{C}$ se aproxima a una distribución χ^2 con (g-2) grados de libertad.

d) Eficiencia Predictiva

Otro modo de evaluar la bondad del ajuste del modelo seleccionado consiste en comparar las predicciones del mismo con los datos muestrales observados, siendo la tabla de clasificación - y una serie de medidas derivadas de la misma -, el procedimiento más utilizado para este fin.

La tabla de clasificación es una tabla de doble entrada donde se clasifican los casos que componen la muestra según los valores observados de la variable de respuesta y los valores pronosticados por el modelo estimado, de tal modo que, dado un valor de corte (generalmente 0.5), todos los casos cuya probabilidad estimada sea igual o mayor que este valor serán clasificados en el grupo que denota la presencia de la característica representada por la variable dependiente. Mientras que aquellas observaciones que obtengan una probabilidad menor a 0.5 serán clasificadas en el grupo que implica la ausencia de dicha característica.

Una vez construida la tabla es conveniente examinar algunas medidas que actúan como índices de la eficacia predictiva del modelo. La tabla de clasificación obtenida adoptará la siguiente forma:

Tabla 1.1 Tabla de Clasificación

Observados	Pronosticados		Totales
	Negativo	Positivo	
Negativo	A	B	(A+B)
Positivo	C	D	(C+D)
Totales	(A+C)	(B+D)	N

Donde A y D son los casos clasificados correctamente por el modelo y B y D son los casos clasificados incorrectamente. De este modo se pueden definir los siguientes índices:

- Tasa de aciertos: (A+D)/N
- Tasa de errores: (B+C)/N
- Especificidad: proporción entre la frecuencia de negativos correctos y el total de resultados negativos observados (A/(A+B))
- Sensibilidad: razón entre los positivos correctos y el total de positivos observados (D/(C+D))
- Tasa de falsos negativos: C/(C+D)
- Tasa de falsos positivos: B/(B+A)

Aunque la interpretación de estos resultados, y en especial de la tasa de aciertos, puede conducir a afirmar que el modelo goza de una alta eficacia predictiva, cabe preguntarse hasta que punto son "buenas" estas tasas de clasificación.

El modelo de regresión logística ajustado usando SPSS permite obtener una imagen adicional sobre la eficacia predictiva del modelo estimado por medio del denominado *histograma de probabilidades estimadas*. Si el modelo estimado distingue acertadamente los dos grupos, los casos para los que se ha observado que ocurre el fenómeno a estudiar (Y=1) deben estar situados a la derecha del punto de corte elegido (0.5), mientras que aquellos casos para los que se ha observado la ausencia del evento (Y=0) deben de situarse a la izquierda de 0.5. Cuanto más agrupados estén ambos grupos en sus respectivos extremos mayor será la eficacia predictiva del modelo. (Luque Martínez, 2000) (Ver Anexo I).

1.4.5.2 Comparación de Modelos no Anidados

Ya hemos visto algunos métodos para seleccionar el modelo que mejor ajuste a los datos cuando uno de los modelos está anidado en el otro, es decir, cuando un modelo es un caso particular de otro modelo, pero también podemos encontrarnos con que los modelos no tienen nada en común, en este caso se han propuesto otras medidas de contraste entre modelos una de ellas es el criterio de información de Akaike o AIC.

El criterio de información de Akaike pondera la función logaritmo de máxima verosimilitud, usando la varianza residual, y el número de parámetros en el modelo. En principio el criterio de selección será escoger modelos con valores más bajos de AIC, el mejor modelo es el que tiene un AIC más pequeño. (Ver e.g. Agresti, 2002, Pág.216).

$$AIC=-2(\ln \text{verosimilitud} - n^{o} \text{ parámetros del modelo ajustado})$$

1.5.1 Diagnósticos del modelo

A la hora de examinar la plausibilidad del modelo de regresión logística seleccionado es importante valorar la posible presencia de valores extremos (outliers) que puedan alterar el ajuste de los datos. Para ello existe un conjunto de métodos de diagnóstico basados en estadísticos o indicadores que examinan la relación existente entre los valores observados y los estimados por el modelo para cada patrón de covariable o sujeto. En general, estas medidas pueden sintetizarse en dos grupos: valores residuales y medidas de influencia. Las primeras se apoyan en diferentes análisis de los residuos (diferencia entre los valores observados de la variable de respuesta y los valores derivados del modelo ajustado) para cada observación a fin de detectar aquellos casos para los cuales el modelo no ajusta bien.

1.5.2 Residuos del modelo

Los residuos más utilizados son los siguientes:

- **Residuos de Pearson:** es el residuo dividido por una estimación de su desviación típica. Para muestras grandes se distribuye como una Normal (0,1).

$$r_p = \frac{y_i - \hat{P}_i}{\sqrt{\hat{P}_i(1-\hat{P}_i)}} \; i = 1,....,n$$

- **Residuos de Pearson Estandarizados:** es el residuo de Pearson dividido entre $(1\text{-}h_i)$ dónde $\hat{V}_i = n_i \hat{P}_i(1-\hat{P}_i)$ y h_i es el i-ésimo elemento de la diagonal de la matriz $n \; x \; n$ $H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$. En esta expresión para H, W es la matriz diagonal $n \; x \; n$ de los casos usados al ajustar el modelo y X es la matriz diseño $n \; x \; p$, dónde p es el número de parámetros desconocidos en el modelo.

$$Z_i = \frac{y_i - n_i\hat{P}_i}{\sqrt{\hat{V}_i(1-h_i)}} \; i = 1,....,n$$

- **Residuos desviación o de la devianza:** comparan la probabilidad estimada de que el caso en cuestión sea un éxito con el valor observado.

$$d_i = signo(y_i - \hat{y}_i)\left[2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right]^{1/2}$$

Dónde $signo(y_i - \hat{y}_i)$ es la función que hace a d_i positivo cuando $y_i \geq \hat{y}_i$ y negativo cuando $y_i < \hat{y}_i$

1.5.3 Medidas de Influencia

Las medidas de influencia identifican aquellos sujetos que ejercen una notable influencia en las estimaciones derivadas del modelo.

- **El valor de influencia (leverage):** se utiliza para detectar aquellos casos que tienen gran impacto sobre el ajuste del modelo.

- **La distancia de Cook:** es una medida que cuantifica el cambio en el vector de coeficientes de regresión cuando una determinada observación es excluida del cálculo de los coeficientes de regresión.

- **El cambio en los coeficientes del modelo:** cuando se excluye un caso concreto. Este estadístico mide el cambio para cada coeficiente incluido el término independiente. Obviamente, la presencia de valores de cambio altos identifica observaciones que deberían ser examinadas con más detalle.

1.6 Modelo Logístico Multinomial

En este caso se supone que la variable dependiente Y tiene más de dos categorías y utiliza como distribución subyacente la distribución multinomial.

Supongamos que las categorías de la variable de respuesta, Y, son codificadas como 0, 1 y 2. En el modelo de regresión logística binaria, la variable de respuesta es parametrizada en términos del logito de Y=1 contra Y=0. En el modelo tricotómico tenemos dos funciones logísticas: una para Y=1 contra Y=0, y otra para Y=2 contra Y=0. En teoría podemos usar cualquiera de las dos parejas pero la extensión obvia del caso binario es usar el logito de Y=2 contra Y=0 para la segunda función. De esta manera, el grupo codificado como Y=0 servirá como referencia. Los efectos varían de acuerdo a la categoría de referencia.

Sea **X** el vector de covariables de dimensión p + 1 con $x_0 = 1$ para explicar el término constante. Denotaremos las dos funciones logísticas como:

$$g_1(X) = \ln\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right]$$

$$= \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \ldots\ldots + \beta_{1p}x_p$$

$$= (1, \mathbf{X})' \beta_1$$

y

$$g_2(X) = \ln\left[\frac{P(Y=2|X)}{P(Y=0|X)}\right]$$

$$= \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \ldots\ldots + \beta_{2p}x_p$$

$$= (1, \mathbf{X})' \beta_2$$

Las tres probabilidades condicionales para cada categoría dado el vector de covariables son:

$$P(Y=0 \mid x) = \frac{1}{1+e^{g_1(x)}+e^{g_2(x)}}$$

$$P(Y=1 \mid x) = \frac{e^{g_1(x)}}{1+e^{g_1(x)}+e^{g_2(x)}}$$

$$P(Y=2 \mid x) = \frac{e^{g_2(x)}}{1+e^{g_1(x)}+e^{g_2(x)}}$$

Sea $\pi_j = \mathrm{P}(Y=j \mid X)$ para j = 0, 1, 2. La expresión general para la probabilidad condicional en un modelo logístico multinomial con tres categorías de respuesta es:

$$\mathrm{P}(Y=j \mid X) = \frac{e^{g_j(x)}}{\sum_{k=0}^{2} e^{g_k(x)}}$$

Donde $$g_0(X) = \ln\left[\frac{\mathrm{P}(Y=0 \mid X)}{\mathrm{P}(Y=0 \mid X)}\right] = 0$$

La regresión logística multinomial puede ser dividida en dos casos:

1) Regresión Logística con Respuesta Ordinal. La variable dependiente es ordinal, por ejemplo, no importante, importante y muy importante.

2) Regresión Logística con Respuesta Nominal. La variable dependiente es nominal, por ejemplo, demócratas, republicanos e independientes.

1.6.1 Regresión Logística Multinomial con Respuesta Ordinal.

Cuando la variable de respuesta es ordinal, los logitos acumulativos pueden ser modelados con los modelos de momios proporcionales. Los modelos de momios proporcionales asumen que los logitos acumulativos pueden ser representados como funciones lineales paralelas de variables independientes, es decir, para cada logito acumulativo los parámetros de los modelos son los mismos, excepto por el término constante.

Por ejemplo, supongamos que la variable dependiente *Y* toma los siguientes valores, 0 = muy importante, 1 = importante y 2 = no importante y sean $P_0 = P(Y=0)$, $P_1 = P(Y=1)$ y $P_2 = P(Y=2)$.

La regresión logística ordinal modela la relación entre los logitos acumulativos de *Y*, es decir,

$$\log \frac{P_0}{1-P_0} = \log \frac{P_0}{P_1 + P_2}$$

Y

$$\log \frac{P_0 + P_1}{1-(P_0 + P_1)} = \log \frac{P_0 + P_1}{P_2}$$

y de las variables independientes. El modelo asume una relación lineal para cada logito y líneas de regresión paralelas.

$$\log (P_0/(1-P_0)) = b_1 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k,$$

$$\log ((P_0+P_1)/P_2) = b_2 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k,$$

Esto es, las intersecciones b_1 y b_2 son diferentes pero los parámetros de regresión restantes son los mismos.

Es fácil ver que las razones $\frac{P_0}{1 - P_0}$ y $\frac{P_0 + P_1}{P_2}$ son proporcionales.

$$(P_0 /(1-P_0)) = e^{b_1} e^{\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}$$

$$((P_0 + P_1)/P_2) = e^{b_2} e^{\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k} = const\,(P_0 /(1-P_0))$$

Dónde, $const = e^{b_2} / e^{b_1}$, de aquí el nombre de modelo de momios proporcionales.
Esto implica que las razones de probabilidades para Y son las mismas.

El método de máxima verosimilitud es usado para obtener las estimaciones de los parámetros del modelo.

Para obtener las probabilidades esperadas se usan las siguientes fórmulas derivadas de las ecuaciones de arriba.

$$P_0 = e^{b_1+\beta_1 X_1+\beta_2 X_2+.....+\beta_k X_K} /(1+e^{b_1+\beta_1 X_1+\beta_2 X_2+.....+\beta_k X_k})$$

$$P_0 + P_1 = e^{b_2+\beta_1 X_1+\beta_2 X_2+.....+\beta_k X_K} /(1+e^{b_2+\beta_1 X_1+\beta_2 X_2+.....+\beta_k X_k})$$

$$P_2 = 1-(P_0+P_1)$$

Si el parámetro β_i es positivo, entonces para P_0 las probabilidades esperadas de (Y=0), así como la probabilidad acumulativa de Y=0 o Y=1, el valor de P_0+P_1 es grande cuando X_i toma valores grandes. Por otro lado, si *bi* es negativo entonces los valores que toman P_0 y P_0+P_1 son pequeños cuando X_i toma valores grandes.

1.6.2 Regresión Logística Multinomial con Respuesta Nominal.

Si la hipótesis del modelo de probabilidades proporcionales no es satisfactoria, entonces la aproximación generalizada de los logitos puede ser usada para modelar la relación entre la variable de respuesta y las variables independientes.

Los modelos generalizados de regresión logística también son usados cuando la variable de respuesta es nominal.

Para una variable categórica, los logitos generalizados son definidos como el logaritmo natural de la probabilidad de cada categoría por encima de la probabilidad de la última categoría de respuesta. Estos logitos generalizados son modelados como funciones lineales de las variables independientes con diferentes parámetros de regresión para cada logito.

Por ejemplo, supongamos que la variable dependiente Y toma los siguientes valores 0= democrático, 1= republicano y 2= independiente y sean $P_0 = P\ (Y=0)$, $P_1 = P\ (Y=1)$ y $P_2 = P\ (Y=2)$.

Los logitos generalizados modelan la relación entre $\log \frac{P_0}{P_2}$ y $\log \frac{P_1}{P_2}$ y las variables independientes.

El modelo supone una relación lineal distinta para cada logito.

$$\log (P_0/P_2) = b_1 + \beta_{11}X_1 + \beta_{12}X_2 + \ldots + \beta_{1k}X_k,$$

$$\log (P_1/P_2) = b_2 + \beta_{21}X_1 + \beta_{22}X_2 + \ldots + \beta_{2k}X_k,$$

Es decir, todos los parámetros de regresión son diferentes.

Al igual que en el modelo de regresión logística multinomial con respuesta ordinal, el método de máxima verosimilitud es usado para obtener las estimaciones de los parámetros del modelo.

Para obtener las probabilidades esperadas se usan las siguientes fórmulas derivadas de las ecuaciones de arriba.

$$P_0 = e^{u1} / (1+ e^{u1} + e^{u2}),$$
$$P_1 = e^{u2} / (1+ e^{u1} + e^{u2}),$$
$$P_2 = 1- (P_0+P_1),$$

Donde,

$$u_1= b_1+ \beta_{11}X_1 + \beta_{12}X_2 +... + \beta_{1k}X_k$$

$$u_2 = b_2+ \beta_{21}X_1 + \beta_{22}X_2 +... + \beta_{2k}X_k.$$

1.6.3 Inferencia en un Modelo Logístico Multinomial

Para construir la función de verosimilitud, es conveniente formular tres variables binarias codificadas como *0* y *1* para indicar el grupo de miembros en una observación. Las variables se codifican de la siguiente manera:

Si Y = 0	Si Y = 1	Si Y = 2
$Y_0 = 1$	$Y_0 = 0$	$Y_0 = 0$
$Y_1 = 0$	$Y_1 = 1$	$Y_1 = 0$
$Y_2 = 0$	$Y_2 = 0$	$Y_2 = 1$

La función condicional de verosimilitud para una muestra de n observaciones independientes es:

$$L(\beta) = \prod_{i=1}^{n} \left[\mathrm{P}_0(x_i)^{y_{0i}} \mathrm{P}_1(x_i)^{y_{1i}} \mathrm{P}_2(x_i)^{y_{2i}} \right]$$

Aplicando logaritmo y tomando en cuenta que $\sum y_{ji} = 1$ para cada i, entonces la función de verosimilitud es:

$$L(\beta) = \sum_{i=1}^{n} y_{1i} g_1(x_i) + y_{2i} g_2(x_i) - \ln(1 + e^{g_1(x_i)} + e^{g_2(x_i)})$$

Las ecuaciones de verosimilitud se obtienen a partir de las primeras derivadas parciales de $L(\beta)$ respecto a cada uno de los parámetros desconocidos. La solución requiere el mismo tipo de métodos iterativos usados en el caso binario.

La estimación de los parámetros se lleva a cabo por máxima verosimilitud y la solución se encuentra de forma numérica mientras que la selección y el estudio bondad de ajuste del modelo se llevan a cabo de forma similar al modelo de regresión logística binaria.

Capítulo 2. Aplicación del Modelo de Regresión Logística Binaria
Factores pronósticos del estado vital en enfermos hospitalizados en la Unidad de Terapia Intensiva.

Introducción

No cabe ninguna duda que la regresión logística es una herramienta estadística con capacidad para el análisis de datos en investigación clínica y epidemiología, debido a que es muy útil para identificar y estimar posibles factores de riesgo en el desarrollo de algunas enfermedades.

El propósito de este trabajo es mostrar la aplicación de modelos logísticos en un estudio realizado en dos hospitales del Distrito Federal: el Centro Médico Nacional Siglo XXI y el Centro Médico Nacional La Raza con un total de 917 casos. A fin de indagar cuáles son los factores de riesgo que pronostican diferentes resultados funcionales (vivo, muerto o vivo con mala calidad de vida).

Se utilizaron modelos de regresión logística con la intención de valorar la importancia de los atributos referidos a los antecedentes ligados a la mala calidad de vida así como de otros factores en relación con la muerte de los pacientes. Para ahondar en las asociaciones entre ciertas variables explicativas, se completó el análisis a partir de una matriz de correlación. El recurso analítico empleado permitió no solo hallar las asociaciones más relevantes entre los factores considerados y la sepsis, sino también lograr una mejor comprensión acerca de la trama de relaciones existentes entre ellos.

El MC. Luis David Sánchez Velázquez estudiante del postgrado en medicina de la UNAM., encargado del diseño y actualización del proyecto fue quien nos proporcionó la base de datos sobre la cual trabajamos para obtener el modelo de regresión logística, binaria y tricotómica.

2.1 Metodología de la Investigación Médica.

Objetivo. Evaluar qué factores permiten pronosticar el nivel de estado vital en enfermos hospitalizados en la UTI durante el período de estudio.

Diseño del Estudio. Se trata de un estudio observacional, longitudinal, homodémico, de cohorte y con recolección prospectiva de la información (estudio de pronóstico).

Lugares del Estudio. Unidades de terapia intensiva (UTI) polivalentes de:

- Centro Médico Nacional Siglo XXI
- Centro Médico Nacional La Raza

Período de Estudio. 1° de Mayo del 2002 al 31 de Septiembre del 2004.

Población en Estudio. Enfermos hospitalizados en la UTI durante el período de estudio.

Criterios de Inclusión:

Enfermos hospitalizados en la UTI durante el período de estudio.
Ambos géneros.
Edad de 18 años o mayores.
Residencia en el Distrito Federal o en el área metropolitana.

Criterios de Exclusión:

Enfermos que pasan a la UTI después de cirugía sólo para vigilancia post-operatoria.
Enfermos con diagnóstico de sepsis procedentes de otro hospital.
Enfermos y/o familiar que rechacen participar en el estudio.
Enfermos que reingresan a la UTI y de quienes ya se tenga el cuestionario de CV.

Criterios de Eliminación:

Enfermos con información incompleta. Carencia de la CFA (Calificación Fisiológica Aguda), de los cuestionarios de CV (Calidad de Vida) previa a la hospitalización y a los 3 meses de egreso hospitalario.

2.2 Planteamiento del Problema.

Los modelos predictivos se han convertido en una herramienta eficaz, siempre que predomine el sentido común y la experiencia en su uso, para el médico dedicado al cuidado del enfermo crítico, respaldo de acciones éticas y legales, han sido elementos importantes a tener en cuenta a la hora de privar de una terapéutica o proceder a algún enfermo. Como sucede con todas las decisiones que tienen un impacto sobre los cuidados del paciente, debe ponderarse del conjunto de conocimientos médicos disponibles, los deseos de los pacientes, familiares y médicos, y la probabilidad de que los cuidados intensivos beneficien al paciente o no. A veces estas decisiones incluirán solo un juicio médico; otras veces, la elección reflejará una perspectiva ética, legal o filosófica.

Las infecciones adquiridas en ámbito hospitalario, también llamadas infecciones nosocomiales representan un serio problema desde distintos puntos de análisis, entre ellos se incluyen sus consecuencias humanas, individuales, sociales y económicas. Entre las infecciones nosocomiales más frecuentemente reportadas se encuentran la de los tractos respiratorio y urinario, bacteriemias y heridas quirúrgicas.

La sepsis es la patología que se observa con mayor frecuencia en las unidades de terapia intensiva (UTI) y que conlleva a disfunción o falla orgánica múltiple e incremento en la mortalidad y aumento en los costos de atención médica.

Los factores que contribuyen a la elevada incidencia actual de la sepsis son:

- Una mayor sobrevida de enfermos con predisposición a la sepsis, tales como neonatos, ancianos y aquéllos con diabetes, cáncer y/o granulocitopenia.
- Un mayor uso de accesos invasivos en la terapéutica de los enfermos, tales como prótesis, equipo de inhaloterapia, catéteres venosos y sondas urinarias.
- El uso indiscriminado de antibióticos que crean condiciones favorables para el sobrecrecimiento, colonización e infección subsecuente por microorganismos más virulentos y multirresistentes a los antibióticos. Tratamientos con quimioterapia y radioterapia.

- Un extenso uso de terapias inmunosupresoras y de esteroides para el transplante de órganos y/o enfermedades inflamatorias.

La evaluación del resultado de la atención médica en la UTI se ha enfocado tradicionalmente sólo en una variable objetiva, la mortalidad. Sin embargo, algunos autores consideran un mal resultado de la intervención de la UTI no sólo a la muerte, sino también a una condición clínica peor a la que tenía el enfermo o a un estado vegetativo al egreso de la UTI, por lo que la medición de la calidad de vida (CV) es una variable de interés cada vez mayor.

Los enfermos que sobreviven a la sepsis pueden tener función orgánica gravemente comprometida, la cual resulta en síntomas persistentes (disnea, fatiga, depresión), estado funcional comprometido (función física, social y emocional) y reducción en la CV relacionada con la salud.

2.3 Análisis Estadístico Exploratorio

Los principales paquetes estadísticos (GLIM, BMDP, SAS, STATA, SPSS, etc.) pueden ajustar modelos de regresión logística. En concreto, utilizaremos el procedimiento de regresión logística del programa SPSS y el del programa STATA.

Antes de comenzar a ajustar los modelos de regresión logística, se hizo un análisis exploratorio de la base de datos. En primer lugar se recodificó a las variables binarias con valores 0 y 1 y se identificaron cuales variables tenían más casos faltantes.

Las variables de interés se clasifican en:

I. **Variables Demográficas:** género, edad, escolaridad (en años), estado civil, estado laboral, procedencia, servicio, CV, fecha de ingreso al hospital, fecha de ingreso a la UTI, fecha de egreso de la UTI, fecha de egreso hospitalario, motivos de egreso de la UTI y del hospital.

II. **Variables Clínicas:** diagnóstico principal, comorbilidad (Calificación de Charlson) modificada, peso, talla, superficie corporal, índice de masa corporal, falla orgánica (Calificación de Bruselas), escala de gravedad de la enfermedad (APACHE II), Calificación Fisiológica Aguda (CFA), Calificación de Coma de Glasgow, balance hídrico.

III. **Variables Paraclínicas:** biometría hemática, química sanguínea, gasometrías arteriales, cultivos, microorganismos y antibiogramas.

IV. **Variables Terapéuticas:** asistencia mecánica ventilatoria, vasopresor, inotrópico, empleo de insulina, sedación, relajación, diálisis, nutrición artificial, traqueotomía y cánula endotraqueal.

V. **Variables de Morbilidad:** choque, paro cardiorrespiratorio, síndrome de insuficiencia respiratoria aguda, dependencia ventilatoria, coagulación intravascular diseminada, reintervención quirúrgica y coma.

VI. **Variables correspondientes a las Enfermedades:** Enf. Pulmonar obstructiva crónica, Demencia, Enf. vascular periférica, Hepatopatía, Diabetes mellitus, Diabetes mellitus complicada, Úlcera péptica, Leucemia, Insuficiencia renal crónica, Linfoma, Cáncer, Cáncer metastásico, SIDA, Insuficiencia cardiaca congestiva, Esteroide y cirrosis.

Tabla 2.1 Descripción de las variables explicativas agrupadas por temporalidad.

Antes de Hospitalizarse		
Variable	**Significado**	**Código**
Hospital	Hospital	1=CMN , 2=CMLR
Caso	Caso	
Sexo	Sexo	1=Mujer , 2=Hombre
Edad	**Edad**	
Edadcod	Edad codificada	1=<40 ; 2=41-60 ; 3=>60años
Procede	Sitio de procedencia	1=Urgencias; 2 Piso; 3=Quirófano
Epoc	Enfermedad pulmonar	0=No; 1=Si
Diabetes	Diabetes mellitus	0=No; 1=Si
Cáncerto	**Cáncer codificado**	**0=No; 1=Si**
Charlson	Gravedad de las enfermedades antes de hospitalizarse	
Charlso	Gravedad codif. de las enfermedades antes de	0=Ninguna; 1=1 ó más
Cv1	hospitalizarse	
Cv1codif	Calidad de vida 2 meses previos a la hospitalización **Calidad de vida codif. 2 meses previos a la hospitalización**	**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)	0=15 ó menos; 1=16 ó más
Aii	APACHE II codificado	**0=Cirugía electiva1= Cirugía**
Qxurgent	**Cirugía urgente**	**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	0=Mejoría;1=Muerte;2=MáximoBe
Adquirid	**Sitio de adquisición de la infección**	**0=sin infección; 1=comunitaria; 2=nosocomial**
Aps1	APS1 Cal. fisiológica aguda en el 1er día de estancia en la terapia	
Brus1	Bruselas de ingreso (Gravedad del enfermo)	
Sepsisgr	**Sepsis grave**	**0=No; 1=Si**
Fcardiac	**Falla cardiaca**	**0=No; 1=Si**
Fendócri	**Falla endocrina**	**0=No; 1=Si**
Frenal	Falla renal	0=No; 1=Si
Díasfren	**Días en falla renal**	
Frespira	**Falla respiratoria**	**0=No; 1=Si**
Díasfcar	Días en falla cardiaca	
Díasfen	**Días en falla endocrina**	
Díasfres	Días en falla respiratoria	
Vm	Ventilación asistida mecánica	0=No; 1=Si
Vmdías	**Días en ventilación mecánica asistida**	

Tabla 2.1 Descripción de las variables explicativas agrupadas por temporalidad (Continuación)

Durante la estancia en la Terapia Intensiva		
Npt	**Nutrición artificial**	**0=No; 1=Si**
Traqueos	Traqueotomía	0=No; 1=Si
Glasl	Glasgow día 1	
Ingrl	Ingresos hídricos del 1er día	
Bhld	Balance hídrico del 1er día	
Glucl	Glucosa del 1er día	
Creatl	Creatinina día 1	
Ikl	Indice de Kirby respiratorio día 1	
Esthosp	Estancia hospitalaria	
Estuti	**Estancia UTI**	
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; 3=Mala CV con sepsis

NOTA: Las variables que se encuentran en negritas son todas aquellas variables que se tomaron en cuenta para ajustar los modelos de regresión: binario, multinomial y lineal.

2.4 Modelación Estadística para el caso Binario.

Con el propósito de valorar la importancia de los atributos referidos a los antecedentes ligados a la sobrevivencia y de los condicionantes clínicos en relación con la condición de muerte de los pacientes hospitalizados en terapia intensiva, se realizó una serie de análisis estadísticos para saber cuales de estos factores son los que más influyen en el fallecimiento de los pacientes o en que los pacientes sobrevivan. En este caso, la variable independiente es *vivomuer* y está codificada de la siguiente manera:

$$vivomuer = \begin{cases} 1 \text{ Muerto} \\ 0 \text{ Vivo} \end{cases}$$

Se comenzó por hacer gráficas de dispersión y gráficas de barras para las variables continuas mientras que para las variables binarias se hicieron tablas cruzadas respecto a la variable *sepsis* ya que está variable es un factor muy importante que va ligado a la mortalidad y se cree que es la más importante. Algunas de estas gráficas se presentan en el Anexo I.

Del total de pacientes, 588 pertenecen al Centro Médico Nacional Siglo XXI y 329 al Centro Médico Nacional La Raza. En la siguiente tabla se ve la proporción de pacientes infectados con sepsis que mueren o sobrevien.

Tabla 2.2 Proporción de pacientes que mueren infectados con sepsis

SEPSISGR Sepsis grave * VIVMUERT Variable blanco Crosstabulation

Count

		VIVMUERT Variable blanco		
		0 Vivo	1 Muerto	Total
SEPSISGR Sepsis grave	0 No	185	161	346
	1 Si	178	322	500
Total		363	483	846

En general, el 64% de los pacientes que tienen sepsis ya sean hombres o mujeres mueren mientras que solamente el 36% de estos pacientes sobrevive. Al hacer la gráfica cruzada de *edad* vs *vivomuer* se notó que en el grupo de 61 años o más mueren más personas lo cual significa que la edad es un factor importante que debemos tomar en cuenta al ajustar el modelo.

Después de esto, se ajustó un modelo de regresión logística binario con todas las variables que se cree explican el hecho de que una persona viva o muera usando como variable de respuesta a *vivomuer* pero debido a que había muchos casos faltantes y además había problemas de multicolinealidad se hizo un análisis de componentes principales a partir de la matriz de correlación para identificar a las variables que estaban altamente correlacionadas entre si. Ya que la presencia de una multicolinealidad elevada puede producir unos coeficientes estimados sesgados y unos errores estándar elevados, lo que alteraría tanto los valores estimados de las probabilidades como el resultado de las pruebas de Wald.

En este análisis se encontró que *apacheii* está relacionada con *aps1*, *aps2*, *bruselas* y *mortpred*. A continuación se muestran las tablas correspondientes a la matriz de correlación de las variables continuas y binarias 0,1.

Tabla 2.3 Matriz de correlación de variables continuas

Correlation Matrix

		APACHEII APACHE II	APS1 Calif. Fisiológica aguda Ingreso	APS2 Calif. Fisiológica aguda sepsis grave	BRUS2 Bruselas sepsis grave	CREAT1 Creatinina 1	MORTPRED Mortalidad predicha
Correlation	APACHEII APACHE II	1.000	.860	.679	.503	.404	.855
	APS1 Calif. Fisiológica aguda Ingreso	.860	1.000	.803	.519	.413	.723
	APS2 Calif. Fisiológica aguda sepsis grave	.679	.803	1.000	.548	.331	.597
	BRUS2 Bruselas sepsis grave	.503	.519	.548	1.000	.396	.473
	CREAT1 Creatinina 1	.404	.413	.331	.396	1.000	.333
	MORTPRED Mortalidad predicha	.855	.723	.597	.473	.333	1.000

La siguiente tabla muestra que existe una alta correlación entre *infnos01* y *sepsis* mientras que por otro lado, la correlación de *calsep* con *sepsis* no es tan alta debido a que dicha variable presenta un valor de 0 .417 y no es lo suficientemente cercano a 1 para poder decir que hay una alta correlación entre dichas variables.

Tabla 2.4 Matriz de correlación de variables binarias 0,1

Correlation Matrix

		SEPSISGR Sepsis grave	CALSEP calidad de vida 2 y sepsis	INFNOS01 Nueva Variable	ANTIBIÓT Antibióticos
Correlation	SEPSISGR Sepsis grave	1.000	.417	.538	.486
	CALSEP calidad de vida 2 y sepsis	.417	1.000	.330	.272
	INFNOS01 Nueva Variable	.538	.330	1.000	.348
	ANTIBIÓT Antibióticos	.486	.272	.348	1.000

Por medio de este análisis también se encontró que la variable *diasfres* está correlacionada a *vmdias*, *diasseda* y *diasrela* y a su vez *vmdias* está correlacionada a *diasseda* y *diasrela*. La matriz de correlación se muestra en el Anexo I.

Se decidió ajustar un modelo para cada grupo de variables y después tomar a las variables que resultaran ser significativas para ajustar otro modelo usando esas variables.

Se ajustó un modelo usando únicamente a las variables correspondientes a las enfermedades y a la edad, cabe mencionar que no se incluyeron las siguientes enfermedades: demencia, Enf. vascular periférica, hepatopatía, leucemia, linfoma, cáncer metastásico, sida y cirrosis debido a que sus frecuencias son muy pocas (menores a 20) en este caso, las variables más significativas fueron epoc, irc, esteroide y edad.

Las variables que resultaron ser significativas de cada grupo son: **EDAD, APACHEII, EPOC, IRC, ESTEROID, DIASFNEU, DIASRES Y VMDIAS.**

Pero estas variables explicativas no son suficientes ya que clínicamente hay algunas variables que influyen mucho en que una persona viva o muera. Así que se ajustó un modelo binario incluyendo a las variables que se creía eran importantes clínicamente.

Las variables clínicas que son importantes y que se consideraron al ajustar el modelo son: falla respiratoria codificada, falla neurológica codificada, falla cardiaca codificada, edad, cáncer, estancia en Unidad de Terapia Intensiva (estuti), apacheii, calidad de vida 1 codificada (cv1) y sepsis grave.

El modelo resultante tomando en cuenta *edad, cáncer, estuti, sepsis, apacheii, fcard, cv1 codif, fneurol* y *frespirc*, es:

MODELO 1

$$\ln\left(\frac{P(Y=muerto)}{1-P(Y=muerto)}\right) = -3.598 + 0.021 edad + 0.035 apacheii + 0.398 cv1codif(mala) + 0.526 sepsisgr(si) + 0.538 fcard(si) + 1.284 fneurol(si) + 1.367 frespirc(si)$$

Tabla 2.5 Coeficientes Estimados en el Intervalo de Confianza Modelo 1

	B	S.E	Wald	df	Sig.	Exp (B)	IC 95% para B Inferior	Superior
EDAD	.021	.005	**17.753**	1	.000	1.021	0.011	0.031
APACHEII	.035	.014	6.410	1	.011	1.035	0.008	0.062
CV1CODIF(1)	.398	.185	4.630	1	.031	1.489	0.035	0.761
SEPSISGR(1)	.526	.172	9.338	1	.002	1.691	0.188	0.862
FCARD(1)	.538	.172	9.796	1	.002	1.713	0.201	0.875
FNEUROL(1)	1.284	.247	**26.994**	1	.000	3.610	0.799	1.768
FRESPIRC(1)	1.367	.236	**33.474**	1	.000	3.922	0.904	1.173
CONSTANTE	-3.598	.370	94.461	1	.000	.027		

En este caso las variables más significativas son: *edad, fneurol(1)* y *frespirc(1)* ya que son las que tienen un valor alto en la estadística de Wald. Las otras variables también son significativas aunque a un nivel más bajo.

Estos resultados se comentaron con el investigador encargado del estudio y se analizó nuevamente la base de datos para encontrar otras variables explicativas que pudieran aportar más información y de esta forma aprovechar lo más posible los datos que tenemos disponibles.

En este nuevo análisis, se encontró que la variable *adquirid* nos proporciona más información que *sepsis* sobre que tipo de infección tiene el paciente ya que está codificada con los siguientes valores: 0 si no tiene infección, 1 si la infección es comunitaria y 2 si la infección es nosocomial, también se pensó que la variable correspondiente a la mortalidad predicha (mortpred) podía ofrecer más información que *apacheii*. De ahora en adelante, para ajustar los demás modelos vamos a utilizar las siguientes variables: *falla respiratoria codificada, falla neurológica codificada, falla cardiaca codificada, edad, cáncer, estuti, mortpred, cvl codificada* y *adquirid*. Como dijimos anteriormente, *cáncer* no resultó ser significativa cuando ajustamos el modelo usando solamente enfermedades, sin embargo el investigador cree que *cáncer* es una variable muy importante y sí la debemos tomar en cuenta.

El modelo que resulta usando *falla respiratoria codificada, falla neurológica codificada, falla cardiaca codificada, edad, cáncer, estuti, mortpred, cvl codificada* y *adquirid* es:

MODELO 2

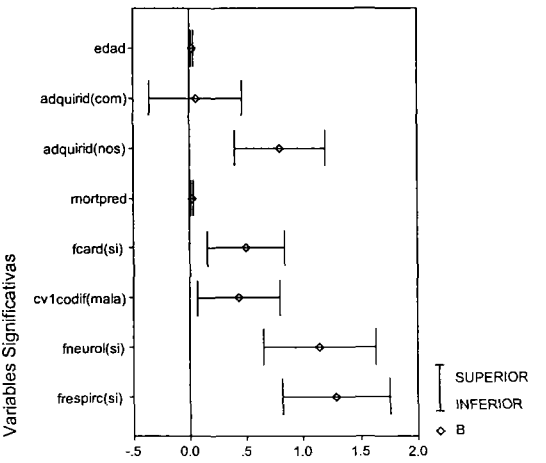
$$\ln\left(\frac{P(Y = muerto)}{1 - P(Y = muerto)}\right) = -3.337 + 0.020edad + 0.055adquirid(com) + 0.799adquirid(nos) + 0.019mortpred + 0.501fcard(si) + 0.430cvlcodif(si) + 1.147fneurol(si) + 1.290frespirc(si)$$

Tabla 2.6 Coeficientes Estimados en el Intervalo de Confianza Modelo 2

	B	S.E	Wald	df	Sig.	Exp (B)	IC 95% para B Inferior	Superior
EDAD	.020	.005	**17.030**	1	.000	1.020	0.011	0.030
ADQUIRID(com)	.055	.203	**15.551**	1	**.794**	.450	-0.355	0.465
ADQUIRID(nos)	.799	.213	**12.188**	1	.000	.475	0.402	1.196
MORTPRED	.019	.005	**15.825**	1	.000	1.019	0.010	0.028
FCARD(1)	.501	.175	8.226	1	.004	.606	0.159	0.843
CV1CODIF(1)	.430	.187	5.276	1	.022	.651	0.063	0.797
FNEUROL(1)	1.147	.249	**21.177**	1	.000	.318	0.659	1.635
FRESPIRC(1)	1.290	.237	**29.564**	1	.000	.275	0.825	1.754
CONSTANTE	-3.337	.355	88.286	1	.000	.036		

Las variables más significativas en el modelo 2 son: *edad, adquirid (nos), mortpred, fneurol (si)* y *frespirc (si)*. Como se dijo anteriormente, las otras variables también lo son pero a un nivel más bajo. Solamente la variable *adquirid (com)* no es significativa.

Gráfica 1. Intervalo de Confianza de los Coeficientes Estimados para las Variables Significativas del modelo 2



Por otro lado, se creó una nueva variable a la que se le llamó *infnos01* con los siguientes valores: 0 si adquirid = sin infección o comunitaria y 1 si adquirid = infección nosocomial para poder seguir ajustando modelos binarios y seguir ajustando modelos de manera tal que se tengan más alternativas para poder escoger el mejor modelo.

El modelo resultante es:

MODELO 3

$$\ln\left(\frac{P(Y=muerto)}{1-P(Y=muerto)}\right) = -3.321 + 0.020edad + 0.775\inf nos01(nos) + 0.019mortpred + 0.509fcard(si) + 0.431cvlcodif(mala) + 1.134fneurol(si) + 1.293frespirc(si)$$

2.4.1 Tablas de Clasificación.

Como se mencionó anteriormente, una manera de evaluar el ajuste del modelo es a través de una tabla de clasificación que consiste en construir una tabla cruzada para la variable de respuesta, *Y*, con una variable dicotómica cuyos valores se derivan de las probabilidades estimadas por el modelo logístico ajustado.

Tabla 2.7 Tabla de Clasificación Modelo 2

Classification Table [a]

			Predicted		
			Variable blanco		
	Observed		Vivo	Muerto	Percentage Correct
Step 1	Variable blanco	Vivo	225	138	62.0
		Muerto	92	381	80.5
	Overall Percentage				72.5

a. The cut value is .500

El porcentaje total de los casos que fueron clasificados correctamente por el modelo 2 es de 72.5% = [(225+381)/836]*100, el porcentaje correcto para los que viven es de 62.0% = [225/(225+138)]*100 mientras que el porcentaje correspondiente a los que mueren es de 80.5% = [381/(381+92)]*100.

Con base en los resultados anteriores se dice que el índice de sensibilidad del modelo es de 80.5% y el índice de especificidad es de 62.0%. Por otro lado, la tasa de aciertos es de 72.5% y la tasa de errores es de 27.5%.

Tabla 2.8 Tabla de Clasificación Modelo 3

Classification Table [a]

Observed			Predicted		
			Variable blanco		
			Vivo	Muerto	Percentage Correct
Step 1	Variable blanco	Vivo	224	139	61.7
		Muerto	91	382	80.8
	Overall Percentage				72.5

a. The cut value is .500

Por otro lado, para el modelo 3 tenemos los siguientes resultados:

El porcentaje total de los casos que fueron clasificados correctamente por el modelo 3 es de 72.5%, el porcentaje correcto para los que viven es de 61.7% mientras que el porcentaje correspondiente a los que mueren es de 80.8%.

El índice de sensibilidad del modelo 3 es de 80.8% y el índice de especificidad es de 61.7%. Por otro lado, la tasa de aciertos es de 72.5% y la tasa de errores es de 27.5%.

De acuerdo con los resultados anteriores, podemos decir que ambos modelos son equivalentes ya que su tasa de aciertos es igual.

2.4.2 Criterio de Información de Akaike

De acuerdo con el criterio de información de Akaike, el mejor modelo es el que tenga el valor más pequeño.

AIC=-2(ln verosimilitud - nº parámetros del modelo ajustado)

Modelo	Devianza	Número de Parámetros
2	914.199	9
3	914.267	8

$$AIC_{modelo\ 4} = 932.199$$

$$AIC_{modelo\ 5} = 930.267$$

Ambos modelos tienen un AIC muy parecido, aunque podríamos decir que el mejor modelo es el modelo 3 ya que es el que tiene el valor más bajo.

2.4.3 Significancia Estadística de las Variables.

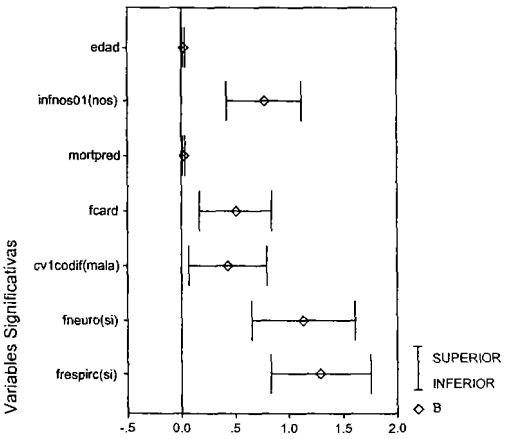
Tabla 2.9 Coeficientes Estimados en el Intervalo de Confianza Modelo 3

	B	S.E	Wald	df	Sig.	Exp (B)	IC 95% para B Inferior	Superior
EDAD	.020	.005	**16.979**	1	.000	1.020	0.011	0.030
INFNOS01(1)	.775	.179	**18.630**	1	.000	2.170	0.423	1.126
MORTPRED	.019	.005	**16.898**	1	.000	1.019	0.010	0.028
FCARD(1)	.509	.172	8.769	1	.003	1.663	0.172	0.846
CV1CODIF(1)	.431	.187	5.303	1	.021	1.538	0.064	0.797
FNEUROL(1)	1.134	.244	**21.541**	1	.000	3.108	0.655	1.613
FRESPIRC(1)	1.293	.237	**29.868**	1	.000	3.645	0.829	1.757
CONSTANTE	-3.321	.350	90.243	1	.000	.036		

A pesar de que las pruebas globales son significativas sobre el ajuste del modelo se tiene que revisar cuidadosamente la significancia de cada variable. En la tabla 2.9 se muestra la estadística de Wald, según dicha estadística todas las variables son significativas ya que presentan valores altos y el valor de p es prácticamente cero, las únicas variables que tienen un valor bajo son *fcard* y *cv1codif* pero de igual forma son significativas.

Gráficamente puede verse que ninguno de los intervalos de confianza calculados para los coeficientes contiene al cero por lo que se puede afirmar con un 95% de confianza que las variables: edad, mortpred, fneurol, fcard, frespirc, *infnos01* y *cv1codif* tienen un efecto significativo en la probabilidad de que el paciente muera.

Gráfica 2. Intervalo de Confianza de los Coeficientes Estimados para las Variables Significativas del modelo 3



2.4.4 Estadística de Hosmer y Lemeshow

Los modelos propuestos están formados por variables continuas por lo que una prueba alternativa a las estadísticas χ^2 y D es la estadística de Hosmer – Lemeshow.

En este caso las hipótesis que usamos para probar la bondad de ajuste del modelo son:

H_0: El modelo ajustado describe adecuadamente a los datos

Vs

H_1: El modelo ajustado no describe adecuadamente a los datos

Si H_0 es cierta, entonces la estadística de *H-L* se distribuye como una ji-cuadrada.

Tabla 2.10 Tabla de Contingencia para la prueba de Hosmer y Lemeshow (Modelo 2)

Contingency Table for Hosmer and Lemeshow Test

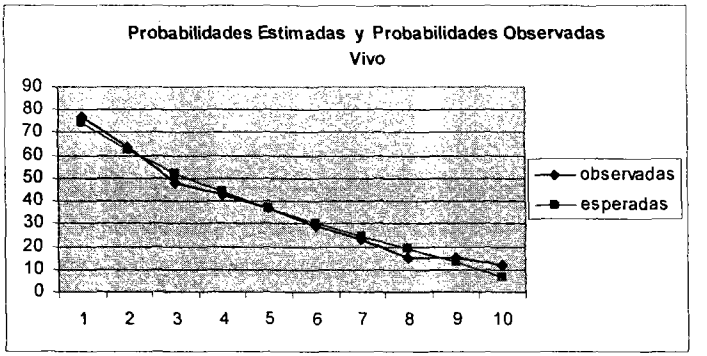
	VIVMUERT Variable blanco = 0 Vivo		VIVMUERT Variable blanco = 1 Muerto		
	Observed	Expected	Observed	Expected	Total
1	77	77	7	9.682	84
2	64	62.384	20	21.616	84
3	48	51.991	36	32.009	84
4	43	43.925	41	40.075	84
5	37	36.886	47	47.114	84
6	29	29.931	55	54.069	84
7	23	24.549	61	59.451	84
8	15	19.034	69	64.966	84
9	15	13.351	69	70.649	84
10	12	6.633	68	73.367	80

Para calcular esta estadística, se formaron 10 grupos con base en las probabilidades estimadas. Para cada grupo se calculó el número observado de éxitos así como el número esperado de éxitos.

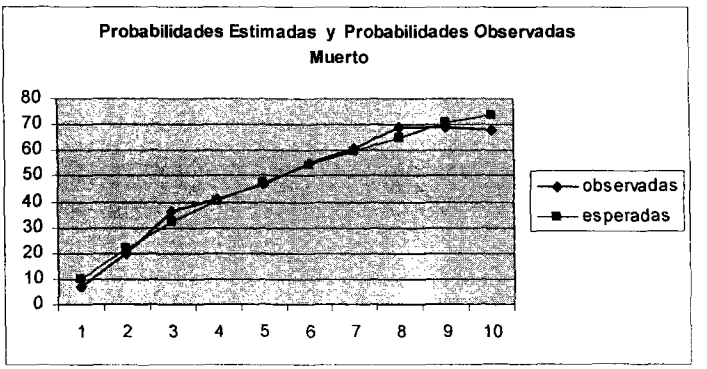
El valor de la estadística de Hosmer – Lemeshow calculada a partir de las frecuencias en la tabla 2.10 es $\hat{C} = 8.113$ y el valor en tablas de una ji - cuadrada con 8 grados de libertad es de 15.507. Como $8.113 < 15.507$ entonces no se rechaza H_0 lo cual indica que el modelo 2 efectivamente describe adecuadamente a los datos.

Gráficamente se puede corroborar lo anterior si comparamos las probabilidades observadas contra las estimadas.

Grafica 3.



Grafica 4.



Como puede verse gráficamente no existe gran diferencia entre las probabilidades observadas y las estimadas por lo tanto el modelo 2 ajusta adecuadamente a los datos.

Tabla 2.11 Tabla de Contingencia para la prueba de Hosmer y Lemeshow (Modelo 3)

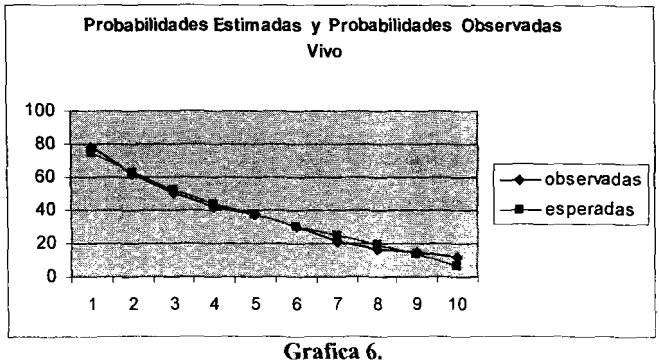
Contingency Table for Hosmer and Lemeshow Test

	VIVMUERT Variable blanco = 0 Vivo		VIVMUERT Variable blanco = 1 Muerto		
	Observed	Expected	Observed	Expected	Total
1	78	74.281	6	9.719	84
2	62	62.387	22	21.613	84
3	50	51.937	34	32.063	84
4	42	43.932	42	40.068	84
5	37	36.953	47	47.047	84
6	30	30.006	54	53.994	84
7	21	24.518	63	59.482	84
8	16	19.039	68	64.961	84
9	15	13.314	69	70.686	84
10	12	6.632	68	73.368	80

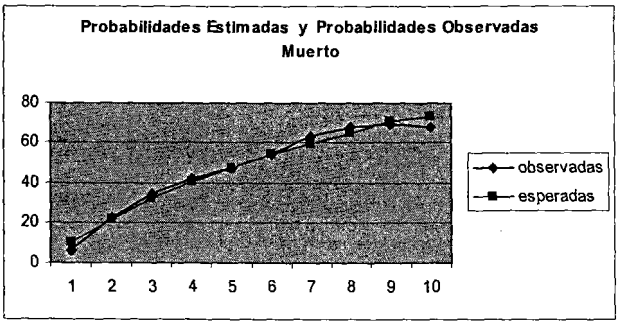
El valor de la estadística de Hosmer – Lemeshow calculada a partir de las frecuencias en la tabla 2.11 correspondiente al modelo 3 es $\hat{C}$ = 8.317 y el valor en tablas de una ji - cuadrada con 8 grados de libertad es de 15.507. Como 8.317 < 15.507 entonces no se rechaza H_0 lo cual indica que el modelo 3 efectivamente describe adecuadamente a los datos.

A continuación se presentan las gráficas de las probabilidades estimadas contra las observadas.

Grafica 5.



Grafica 6.



Hasta este punto, se ha visto que los modelos 2 y 3 ajustan adecuadamente a los datos y que de estos dos modelos el mejor es el modelo 3. Para comparar directamente a los modelos 2 y 3 podría usarse el cociente de verosimilitudes:

$$G = 914.267 - 914.199 = 0.068$$

H_0: El modelo 3 describe adecuadamente a los datos

vs

H_1: El modelo 2 describe adecuadamente a los datos

El valor en tablas de una χ^2 con un grado de libertad es de 3.84. Como $0.068 < 3.84$ no se rechaza el modelo 3.

Ahora hace falta probar cual de los tres modelos binarios que se ajustaron es el mejor, para ello se presenta una tabla con las devianzas de los modelos anteriores así como el cálculo del AIC para cada uno de ellos y el porcentaje global de acuerdo a sus tablas de clasificación.

Tabla 2.12 Criterio de Información de Akaike para los tres modelos

Modelo	Devianza	Número de Parámetros	AIC	% Global de Clasificación
1	936.282	8	952.282	71.5%
2	914.199	9	932.199	72.5%
3	914.267	8	930.267	72.5%

Como puede verse en la tabla anterior, el mejor de todos los modelos es el 3 ya que es el que tiene un AIC más pequeño.

2.4.5 Interpretación de la Razón de Momios

La razón de momios estimada, así como sus intervalos de confianza, se obtienen a partir de los coeficientes estimados y de los valores extremos en los intervalos de confianza al aplicarles la función exponencial.

En la tabla 2.13 aparecen los coeficientes estimados en el modelo 3 así como los intervalos de confianza del 95% para el exponente de dichos coeficientes.

Tabla 2.13 Coeficientes Estimados e Intervalo de Confianza para la Razón de Momios del modelo 3

Variables	Coeficientes Estimados	exp(B)	IC 95% para exp(B)	
			Inferior	Superior
EDAD	.020	1.020	1.011	1.030
INFNOS01(1)	.775	2.170	1.526	3.084
MORTPRED	.019	1.019	1.010	1.028
FCARD(1)	.509	1.663	1.188	2.330
CV1CODIF(1)	.431	1.538	1.066	2.219
FNEUROL(1)	1.134	3.108	1.925	5.017
FRESPIRC(1)	1.293	3.645	2.292	5.795
CONSTANTE	-3.321	.036		

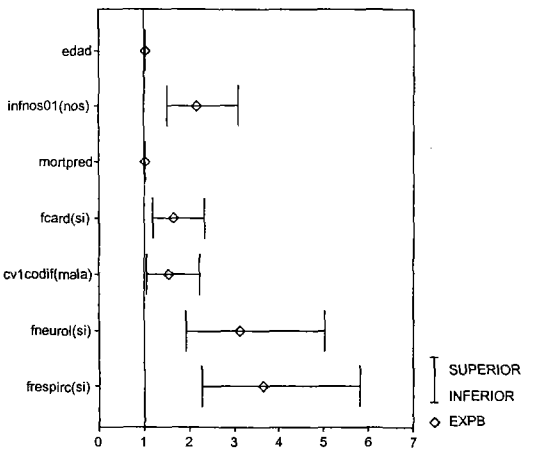
En el caso de la variable *infnos01* la categoría de referencia es si *infnos* = no tiene infección o la infección es comunitaria, respecto a esta categoría se hace la comparación al interpretar la razón de momios.

$$RM = \frac{Momios(vivomuer = muerto \mid \inf nos01 = nos, mortpred = x, edad = x, cv1codif = x, fcard = x, frespirc = x, fneurol = x)}{Momios(vivomuer = muerto \mid \inf nos01 = com, mortpred = x, edad = x, cv1codif = x, fcard = x, frespirc = x, fneurol = x)} = e^{\beta_1} = 2.170$$

Lo anterior se interpreta como el momio de que una persona muera cuando la infección es nosocomial, manteniendo constantes a las otras variables, se duplica a los momios de que una persona muera dado que no tiene infección o la infección es comunitaria. Análogamente se hace esta interpretación para las variables *fcard, cv1codif, fneurol y frespirc.*

El intervalo de confianza estimado sugiere que los momios de que una persona muera para todas aquellas personas que tienen infección nosocomial puede ser tan pequeño como 1.526 o tan grande como 3.084. A continuación se presenta la gráfica.

Gráfica 7. Intervalo de Confianza de las Razones de Momios



Para la variable *edad* se tiene lo siguiente:

$$RM = \frac{Momios(vivomuer = muerto \mid \inf nos01 = x, mortpred = x, edad = x+1, cv1codif = x, fcard = x, frespirc = x, fneurol = x)}{Momios(vivomuer = muerto \mid \inf nos01 = x, mortpred = x, edad = x, cv1codif = x, fcard = x, frespirc = x, fneurol = x)} = e^{\beta_1} = 1.020$$

Lo anterior indica que los momios de que una persona muera cuando la edad se incrementa en un año, manteniendo constantes a las otras variables, está asociado con un incremento del 2% en los momios de que una persona muera cuando tiene un año menos de edad, es decir, por cada año que aumenta la edad, los momios aumentan un 2%.

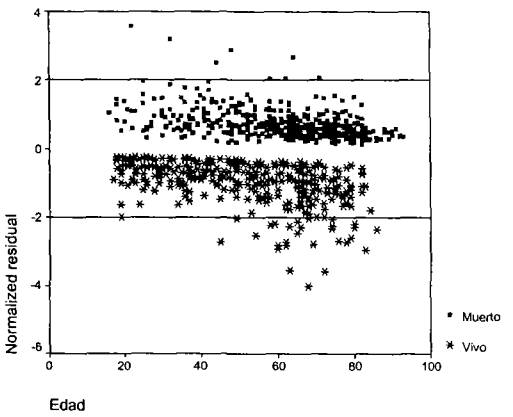
En el caso de la variable *mortpred,* por cada unidad que esta aumenta, los momios aumentan un 1.9%.

Respecto a los parámetros estimados tienen signos positivos lo cual indica que la probabilidad de que una persona muera aumenta conforme aumenta el valor de cada variable continua. Entonces, un incremento de una unidad en las variables *edad* y *mortpred* aumenta en 0.020 y 0.019 respectivamente. Mientras que para las variables *infnos01, fcard, cv1codif, fneurol y frespirc* también hay un incremento cuando existe la presencia de alguna de las fallas y cuando la infección es nosocomial.

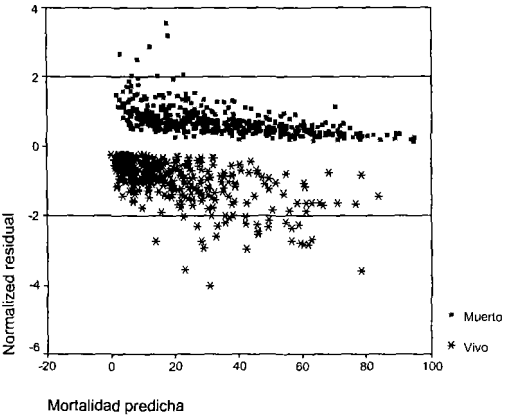
2.4.6. Análisis de los Residuos

Para llevar a cabo el análisis del modelo 3 por medio de los residuos se procedió a graficar los residuos estandarizados contra cada una de las variables continúas del modelo. A continuación se presentan las gráficas correspondientes a cada una de ellas.

Gráfica 8. Residuos Estandarizados vs Edad

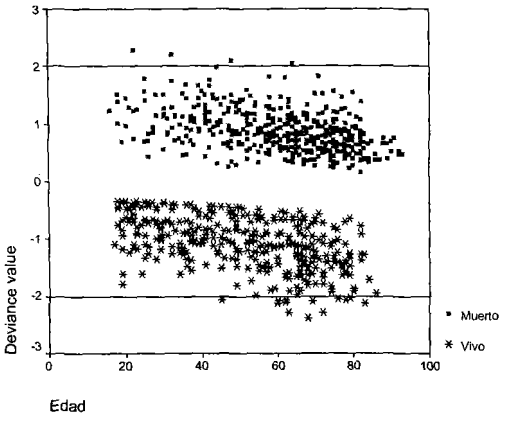


Gráfica 9. Residuos Estandarizados vs Mortalidad Predicha

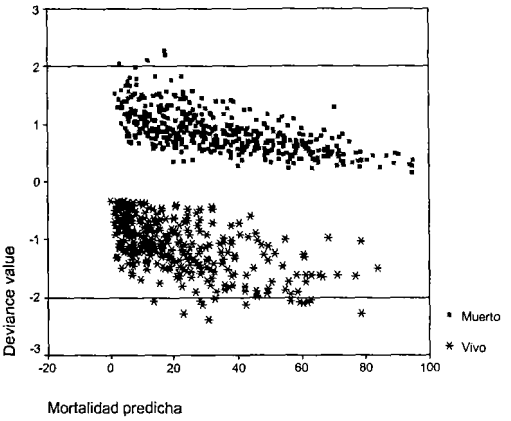


De acuerdo con las gráficas anteriores se puede decir que el ajuste es bueno ya que los residuos estandarizados en ambas gráficas se encuentran entre -2 y 2.

Gráfica 10. Residuos de la Devianza vs Edad



Gráfica 11. Residuos de la Devianza vs Mortalidad Predicha

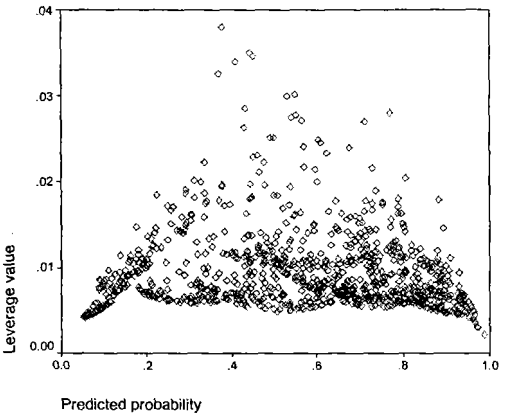


De igual manera los valores en las graficas 10 y 11 se encuentran en su mayoría entre -2 y 2 por lo que se puede considerar un buen ajuste.

2.4.7. Medidas de Influencia: Valor de Influencia y Distancia de Cook

Mediante este análisis, se detectan las observaciones que influyen en la estimación de los parámetros y en la calidad del ajuste del modelo.

Gráfica 12. Valor de Influencia (leverage) vs Probabilidades Estimadas

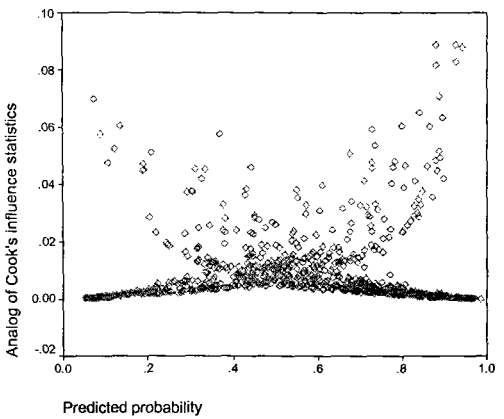


En la gráfica 12 se muestra la dispersión entre las probabilidades estimadas y los valores de influencia. Los valores que son considerados puntos influyentes son aquellos valores que se encuentran más alejados de la nube de puntos, estos puntos corresponden a las siguientes observaciones:

* 63: Mujer de 47 años, 6 días en *vm*, falla cardiaca, falla neurológica, murió.
* 393: Hombre de 25 años, falla neurológica, murió.
* 507: Mujer 29 años, con *irc*, falla cardiaca, falla renal, falla hematológica, sobrevivió
* 544: Mujer 25 años, con *irc*, sepsis, falla cardiaca, falla renal, falla hepática, sobrevivió.
* 628: Hombre 47 años, diabetes, sepsis, falla cardiaca, falla endocrina, falla hematológica, falla renal, sobrevivió.

Debido a que al eliminar estos casos no existe una influencia significativa en los parámetros al momento de ajustar el modelo, no fue necesario llevar a cabo otro proceso de ajuste.

Gráfica 13. Distancia de Cook vs Probabilidades Estimadas



La gráfica anterior muestra la dispersión entre las probabilidades estimadas y la distancia de Cook, al igual que en la gráfica 12, no hay puntos influyentes.

2.5 Ajuste del Modelo de Regresión Logística Tricotómico.

Una vez que se ajustó un modelo de regresión logística para conocer los factores de riesgo asociados a la probabilidad de que una persona muera, se decidió ajustar otro modelo para modelar las probabilidades de los pacientes con respecto a su calidad de vida después de haber sido dados de alta de la UTI ya que se considera que la medición de la calidad de vida es una variable de interés cada vez mayor debido a que los enfermos que sobreviven a la sepsis pueden tener una función orgánica gravemente comprometida, la cual resulta en síntomas persistentes (disnea, fatiga, depresión), estado funcional comprometido (función física, social y emocional) y reducción en la calidad de vida relacionada con la salud. Actualmente no existe ningún estudio que aborde el tema de la predicción del desempeño de los enfermos con sepsis grave al egreso hospitalario, con este nuevo modelo se pretende encontrar cuáles son los factores de riesgo que, en los enfermos con sepsis grave, pronostican diferentes resultados funcionales.

Ahora se va a ajustar un modelo de regresión logística multinomial nominal ya que en este caso, la variable de respuesta está codificada con tres categorías: 0 = Vivo Mala Calidad de Vida; 1 = Muerto y 2 = Vivo Buena Calidad de Vida. De aquí en adelante llamaremos a esta variable "shapiro".

Shapiro
- $Y_0 = 0$ Vivo Mala Calidad de Vida
- $Y_1 = 1$ Muerto
- $Y_2 = 2$ Vivo Buena Calidad de Vida

El total de observaciones para la variable shapiro es de 846, y se encuentran distribuidas de la siguiente manera: 178 observaciones corresponden a Vivo Mala Calidad de Vida, 483 a Muerto y por último se tienen 185 observaciones para Vivo Buena Calidad de Vida. En términos de porcentaje equivalen al 21.0%, 57.1% y 21.9% del total de observaciones respectivamente.

Se ajustó un modelo usando como variables explicativas: fcard, fneurol, frespirc, infnos01, cv1cod, edad, estuti, mortpred y cáncer y shapiro como variable de respuesta, tomando como referencia a "Vivo Buena Calidad de Vida" es decir, al ajustar el modelo se va a comparar la categoría Vivo Mala Calidad de vida contra Vivo Buena Calidad de Vida y de forma simultanea se va a comparar Muerto contra Vivo Buena Calidad de Vida.

Como resultado tenemos los siguientes modelos logísticos:

Modelo 4a.

Corresponde a comparar los grupos:

Vivo Mala Calidad de Vida vs Vivo Buena Calidad de Vida

$$\ln\left(\frac{\hat{P}(Y = mala\ calidad)}{\hat{P}(Y = buena\ calidad)}\right) = -2.767 + 0.02793edad - 0.0006525mortpred + 0.02691estuti + 1.308cáncer(si) + 1.041fneurol(si) - 0.612fcard(si) + 1.086frespirc(si) + 1.801cv1cod(mala) + 0.542\text{inf}nos01(nos)$$

Modelo 4b.

Corresponde a comparar los grupos:

Muerto vs Vivo Buena Calidad de Vida

$$\ln\left(\frac{\hat{P}(Y = muerto)}{\hat{P}(Y = buena\ calidad)}\right) = -4.233 + 0.03664edad + 0.01878mortpred + 0.03226estuti + 1.196cáncer(si) + 1.787fneurol(si) + 0.08838fcard(si) + 1.817frespirc(si) + 1.687cv1cod(mala) + 1.068\text{inf}nos01(nos)$$

Como puede verse en la tabla de clasificación, el modelo ajustado clasifica correctamente a los vivos que tienen buena calidad de vida con un porcentaje de 56.8% mientras que por otra parte, a las personas que viven con una mala calidad de vida las clasifica con el 11.8% y a los muertos los clasifica con el 89.9%.

Tabla 2.14 Tabla de Clasificación para el Modelo 4

Classification

Observed	Predicted			
	0 Vivo Mala calidad de vida	1 Muerto	2 Vivo Buena calidad de vida	Percent Correct
0 Vivo Mala calidad de vida	21	121	36	11.8%
1 Muerto	12	425	36	89.9%
2 Vivo Buena calidad de vida	6	74	105	56.8%
Overall Percentage	4.7%	74.2%	21.2%	65.9%

La tabla muestra la forma en que el modelo clasifica a las observaciones es decir, del total de observaciones de la categoría Vivo Mala Calidad de Vida, 21 de esas observaciones las clasifica como vivos con mala calidad de vida, 121 como muertos y 36 son clasificados como vivos con buena calidad de vida. Mientras que del total de la categoría Muerto, 12 son clasificadas como vivos con mala calidad de vida, 425 como muertos y 36 como vivo buena calidad de vida.

En el modelo 4a el coeficiente estimado para la variable *fcard* es negativo, esto quiere decir que si la persona tiene una falla cardiaca entonces la probabilidad de tener una mala calidad de vida disminuye lo cual es absurdo. El resultado anterior sugiere que *fcard* está correlacionada con otra de las variables incluidas en el modelo.

Tabla 2.15 Coeficientes Estimados en el Intervalo de Confianza Modelo 4a

	B	S.E	Wald	df	Sig.	Exp (B)	IC 95% para B Inferior	Superior
CONSTANTE	-2.767	.427	41.949	1	.000			
EDAD	.02793	.007	**16.383**	1	.000	1.028	1.015	1.042
MORTPRED	0.0006525	.007	.008	1	.929	.999	.985	1.014
ESTUTI	0.02691	.021	1.579	1	.209	1.027	.985	1.071
CANCER(si)	1.308	.553	5.597	1	.018	3.699	1.252	10.930
FNEUROL(si)	1.041	.426	5.982	1	.014	2.832	1.230	6.520
FCARD(si)	-0.612	.250	5.981	1	.014	.542	.332	.886
FRESPIRC(si)	1.086	.284	**14.634**	1	.000	2.963	1.698	5.169
CV1CODIF(si)	1.801	.333	**29.307**	1	.000	6.056	3.155	11.625
INFNOS01(nos)	.542	.288	3.535	1	.060	1.719	.977	3.024

Tabla 2.16 Coeficientes Estimados en el Intervalo de Confianza Modelo 4b

	B	S.E	Wald	df	Sig.	Exp (B)	IC 95% para B Inferior	Superior
CONSTANTE	-4.233	.442	91.717	1	.000			
EDAD	0.03664	.006	**33.124**	1	.000	1.037	1.024	1.050
MORTPRED	0.01878	.006	8.526	1	.004	1.019	1.006	1.032
ESTUTI	0.03226	.019	2.925	1	.087	1.033	.995	1.072
CANCER(si)	1.196	.553	4.682	1	.030	3.307	1.119	9.770
FNEUROL(si)	1.787	.382	**21.919**	1	.000	5.970	2.826	12.612
FCARD(si)	.08838	.232	.145	1	.704	1.092	.693	1.723
FRESPIRC(si)	1.817	.290	**39.357**	1	.000	6.152	3.488	10.853
CV1CODIF(si)	1.687	.322	**27.382**	1	.000	5.401	2.872	10.160
INFNOS01(nos)	1.068	.258	**17.157**	1	.000	2.911	1.756	4.826

En general, las variables más significativas de los modelos 4a y 4b son: *edad, fneurol, frespirc, cv1codif* e *infnos01* aunque las demás variables también son significativas.

Al comentar los resultados anteriores con el MC., se decidió ajustar tres modelos más usando a la edad codificada y tomando en cuenta a las variables significativas de los modelos 4a y 4b, se quitaron *fcard* y *mortpred* y en lugar de estas variables se incluyó la variable *qxurgent.*

Al ajustar los tres modelos se tomaron en cuenta las siguientes variables: *edad, fneurol, frespirc, cv1codif, infnos01, qxurgent, edad1 y edad2.*

Cabe mencionar que los tres modelos tienen a todas las variables en común excepto por la variable edad ya que dicha variable se codifico de dos formas distintas y esto dio como resultado a las variables *edad1* y *edad2*. De esta forma se obtuvo un modelo con edad continua, otro con *edad1* y por último un modelo con *edad2*. A continuación se presenta la codificación de las nuevas variables.

$$qxurgent = \begin{cases} 0 & \text{sin cirugía o cirugía electiva} \\ 1 & \text{cirugía urgente} \end{cases}$$

$$edad1 = \begin{cases} 1 & \text{0-40 años} \\ 0 & \text{41 o más} \end{cases}$$

$$edad2 = \begin{cases} 1 & \text{0-60 años} \\ 0 & \text{61 o más} \end{cases}$$

De los tres modelos ajustados, el que contiene a la variable *edad2* es el que escogió el MC. ya que a su criterio, es el modelo que aporta más información.

La siguiente tabla muestra el número de pacientes que tenían 61 años o más en cada categoría de la variable de respuesta.

Tabla 2.17

EDAD2 Edad codificada * SHAPIRO Variable blanco Crosstabulation

Count

		SHAPIRO Variable blanco			
		0 Vivo Mala calidad de vida	1 Muerto	2 Vivo Buena calidad de vida	Total
EDAD2 Edad codificada	0 61 o más	87	272	49	408
	1 0 a 60	91	211	136	438
Total		178	483	185	846

A continuación se presentan los resultados obtenidos con dicho modelo.

Se tomaron en cuenta las siguientes variables: *edad2*, *fneurol*, *frespirc*, *cv1codif*, *infnos01* y *qxurgent*.

Modelo 5a.

Corresponde a comparar los grupos:

Vivo Mala Calidad de Vida vs Vivo Buena Calidad de Vida

$$\ln\left(\frac{\hat{P}(Y = mala\ calidad)}{\hat{P}(Y = buena\ calidad)}\right) = -1.048 + 0.698 qxurgent(si) + 1.050 fneurol(si) + 0.956 frespirc(si) + 0.392 \inf nos01(nos)$$
$$+ 1.822 cv1codif(mala) - 0.634 edad2(0 - 60 años)$$

Modelo 5b.

Corresponde a comparar los grupos:

Muerto vs Vivo Buena Calidad de Vida

$$\ln\left(\frac{\hat{P}(Y = muerto)}{\hat{P}(Y = buena\ calidad)}\right) = -1.212 + 0.627 qxurgent(si) + 2.004 fneurol(si) + 1.998 frespirc(si) + 1.131 \inf nos01(nos) + 1.865 cvlcodif(mala) - 1.022 edad2(0-60 años)$$

Tabla 2.18 Tabla de Clasificación Modelo 5

Classification

Observed	Predicted			
	0 Vivo Mala calidad de vida	1 Muerto	2 Vivo Buena calidad de vida	Percent Correct
0 Vivo Mala calidad de vida	13	127	38	7.3%
1 Muerto	9	416	48	87.9%
2 Vivo Buena calidad de vida	5	69	111	60.0%
Overall Percentage	3.2%	73.2%	23.6%	64.6%

Como puede verse en la tabla de clasificación, el modelo ajustado clasifica correctamente a los vivos que tienen buena calidad de vida con un porcentaje de 60.0% mientras que por otra parte, a las personas que viven con una mala calidad de vida las clasifica con el 7.3% y a los muertos los clasifica con el 87.9%. Mientras que el porcentaje total de los casos que fueron clasificados correctamente por el modelo es de 64.6%.

Tabla 2.19 Coeficientes Estimados en el Intervalo de Confianza Modelo 5a

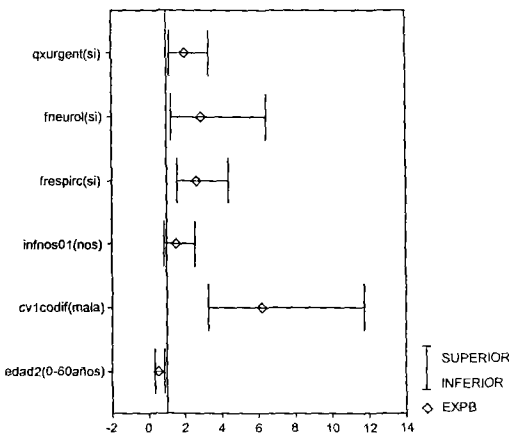
	B	S.E	Wald	df	Sig.	Exp (B)	IC 95% para expB	
							Inferior	Superior
CONSTANTE	-1.048	.294	12.729	1	.000			
QXURGENT	.698	.256	7.444	1	.006	2.009	1.217	3.316
FNEUROL(si)	1.050	.414	6.445	1	.011	2.859	1.271	6.431
FRESPIRC(si)	.956	.263	13.257	1	.000	2.602	1.555	4.354
INFNOS01(nos)	.392	.275	2.040	1	.153	1.481	.864	2.537
CV1CODIF(si)	1.822	.327	**31.121**	1	.000	6.187	3.261	11.736
EDAD2 (0-60AÑOS)	-.634	.243	6.782	1	.009	.531	.329	.855

Tabla 2.20 Coeficientes Estimados en el Intervalo de Confianza Modelo 5b

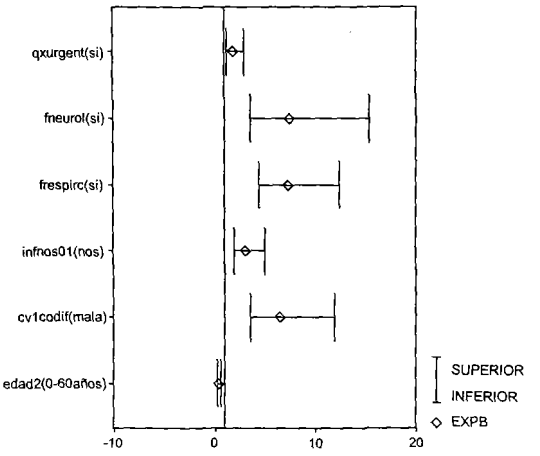
	B	S.E	Wald	df	Sig.	Exp (B)	IC 95% para B Inferior	Superior
CONSTANTE	-1.212	.292	17.176	1	.000			
QXURGENT	.627	.235	7.091	1	.008	1.872	1.180	2.969
FNEUROL(si)	2.004	.373	**28.927**	1	.000	7.420	3.574	15.402
FRESPIRC(si)	1.998	.267	**56.050**	1	.000	7.378	4.372	12.449
INFNOS01(nos)	1.131	.242	**21.750**	1	.000	3.098	1.926	4.983
CV1CODIF(si)	1.865	.314	**35.257**	1	.000	6.458	3.489	11.953
EDAD2 (0-60AÑOS)	-1.022	.222	**21.226**	1	.000	.360	.233	.556

Las variables que se encuentran en negritas son las que resultaron ser más significativas aunque también podemos decir que el resto de las variables también lo son. A continuación se presentan las gráficas correspondientes al intervalo de confianza de las razones de momios de los modelos 5a y 5b.

Gráfica 14. Intervalo de Confianza de las Razones de Momios Modelo 5a.



Gráfica 15. Intervalo de Confianza de las Razones de Momios Modelo 5b.



A continuación se presentan dos modelos binarios ajustados de forma individual con el fin de comparar los resultados con el modelo 5.

Para ello se creó a las siguientes variables:

malcal

1 Mala Calidad de Vida

0 Buena Calidad de Vida

$$muer \begin{cases} 1 \text{ Muerto} \\ 0 \text{ Buena Calidad de Vida} \end{cases}$$

Modelo 6.

Corresponde a comparar los grupos:

Mala Calidad de Vida vs Buena Calidad de Vida

$$\ln\left(\frac{P(Y = mala\ calidad)}{P(Y = buena\ calidad)}\right) = -0.975 - 0.630 edad2(0-60 años) + 0.871 fneurol + 0.841 frespirc(si) + 0.414 \inf nos01(si) + 1.877 cvlcodif(si) + 0.716 qxurgent(si)$$

Modelo 7.

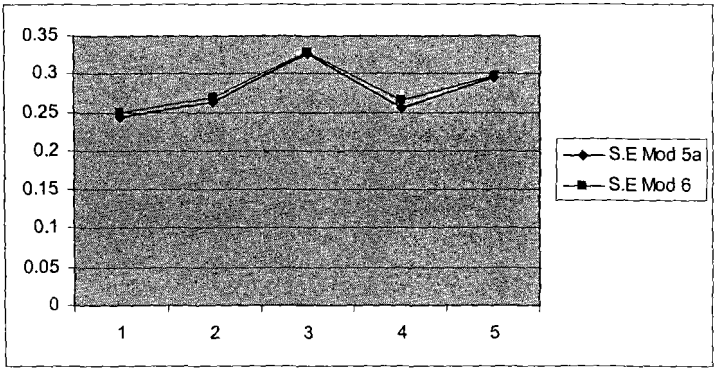
Corresponde a comparar los grupos:

Muerto vs Buena Calidad de Vida

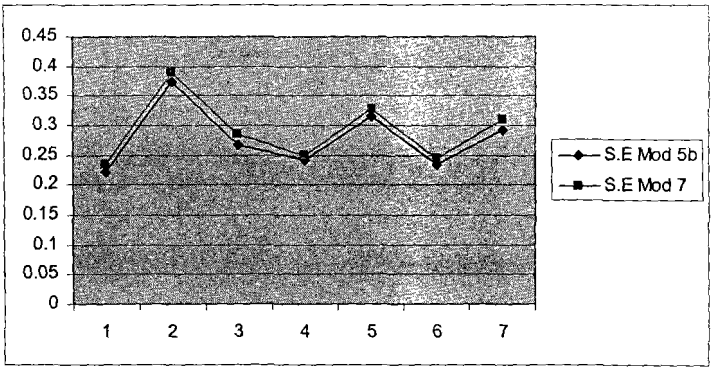
$$\ln\left(\frac{P(Y = muerto)}{P(Y = buena\ calidad)}\right) = -1.279 - 1.043 edad2(0-60 años) + 2.151 fneurol(si) + 2.090 frespirc(si) + 1.152 \inf nos01(si) + 1.715 cvlcodif(mala) + 0.632 qxurgent(si)$$

Los coeficientes estimados para ambos modelos son muy parecidos a los coeficientes estimados de los modelos 5a y 5b; además de que sus errores estándar también son parecidos, para corroborar lo anterior se procedió a graficar los errores estándar del modelo 5a y los errores estándar del modelo 6.

Gráfica 16. Errores Estándar (Modelo 5a y Modelo 6)



Gráfica 17. Errores Estándar (Modelo 5b y Modelo 7)



Efectivamente los errores estándar son muy parecidos aunque las desviaciones estándar estimadas para el modelo de estimación simultánea son todas ligeramente mayores a las desviaciones estándar para los modelos estimados por separado.

Prueba de Bondad de Ajuste de Hosmer y Lemeshow para el modelo 6

El valor de la estadísiica de Hosmer – Lemeshow calculada a partir de las frecuencias es $\hat{C}$ = 4.094 y el valor en tablas de una ji - cuadrada con 6 grados de libertad es de 12.592. Como 4.094 < 12.592 entonces no se rechaza H_0 lo cual indica que el modelo 6 describe adecuadamente a los datos.

Prueba de Bondad de Ajuste de Hosmer y Lemeshow para el modelo 7

El valor de la estadística de Hosmer – Lemeshow calculada a partir de las frecuencias es $\hat{C}$ = 8.039 y el valor en tablas de una ji - cuadrada con 8 grados de libertad es de 15.507. Como 4.094 < 15.507 entonces no se rechaza H_0 lo cual indica que el modelo 7 describe adecuadamente a los datos.

Los modelos 6 y 7 ajustan adecuadamente a los datos y a su vez son comparables a los modelos 5a y 5b, esto apoya el hecho de que estos últimos ajustan adecuadamente a los datos.

2.6 Técnicas de Regresión Lineal Múltiple y Poisson.

Adicionalmente, para completar el análisis de los factores de riesgo que conllevan a tener buena calidad de vida, mala calidad de vida o muerte en los pacientes hospitalizados en terapia intensiva se decidió ajustar un modelo de regresión lineal ya que una variable importante para determinar alguna de las tres características anteriores es *estuti* que se refiere a los días en los que se encuentra un paciente en terapia intensiva. En este caso la variable es discreta aunque por su tipo y valores que toma puede también considerarse como continua. Debido a esto se decidió ajustar en primer lugar una regresión lineal múltiple y en segundo lugar una regresión poisson, así que usamos a la variable *estuti* como la variable dependiente para determinar con que otras variables se encuentra asociada y en qué sentido se da dicha asociación es decir, si el valor de *estuti* tiende a aumentar o disminuir al aumentar los valores de alguna de las variables continuas o en que proporción se incrementa al haber presencia o ausencia en el caso de las variables categóricas.

A continuación se presenta una breve explicación en lo que se refiere a regresión lineal y regresión poisson.

2.6.1 El Modelo de Regresión Lineal Múltiple.

El objetivo de la regresión lineal es el de encontrar un plano que ajuste a la nube de puntos de las gráficas para predecir los valores de Y a partir de los valores de las variables explicativas.

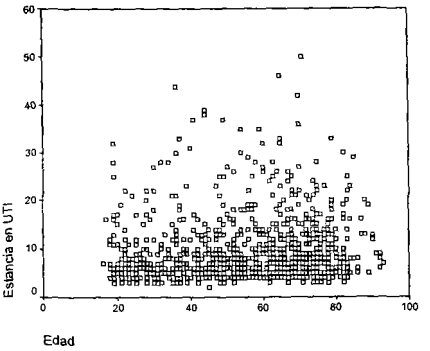
La ecuación general del modelo de regresión múltiple es de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_5 X_5 + \epsilon$$

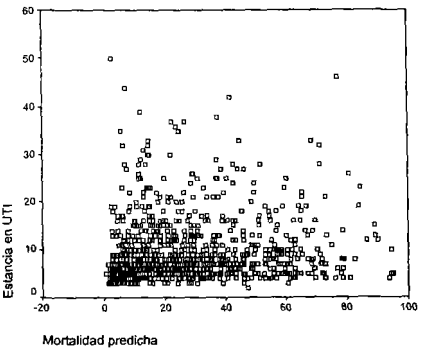
Por medio de un gráfico de dispersión se puede visualizar la relación existente entre las variables continuas.

A continuación se presentan las gráficas de dispersión de las variables continuas contra la variable dependiente *estuti*.

Gráfica 16.



Gráfica 17.



Las gráficas correspondientes a las variables cualitativas se presentan en el Anexo I.

En primer lugar se ajustó un modelo con las siguientes variables explicativas: *edad, qxurgent,epoc, irc, cáncer, apacheii, diasfneu, diasfhem, diasfend, diasfren, diasfhep, diasfres, vmdias, antibiót, nocirugí, npt, traqueos, gastrost, sepsisgr y mortpred.* Por medio del método forward en SPSS para identificar cuales eran las variables más significativas, de esta forma fue como se seleccionó a las variables.

2.6.2 Interpretación de los coeficientes de regresión y la tabla del Análisis de la Varianza

Al ajustar el modelo de regresión lineal tenemos los siguientes resultados:

Tabla 3.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.991	.133		22.497	.000
	DIASFEND Días en falla endócrina	.190	.058	.031	3.269	.001
	DIASFREN Días en falla renal	8.420E-02	.026	.032	3.284	.001
	VMDIAS Días en ventilación	.884	.014	.890	64.731	.000
	NPT Nutrición parenteral total	.509	.184	.029	2.770	.006
	ANTIBIÓT Antibióticos	.178	.048	.049	3.721	.000
	NOCIRUGÍ Número de cirugías	.319	.087	.040	3.666	.000
	MORTPRED Mortalidad predicha	-1.631E-02	.003	-.050	-5.246	.000

a. Dependent Variable: ESTUTI Estancia en UTI

De acuerdo a los resultados obtenidos en la tabla anterior, el modelo de regresión lineal queda de la siguiente forma:

Modelo 8.

$$Y = 2.991 + 0.190 diasfend + 0.08420 diasfren + 0.884 vmdias + 0.509 npt + 0.178 antibiot + 0.319 nocirugi - .01631 mortpred$$

A continuación se presenta la tabla del análisis de la varianza del modelo ajustado.

Tabla 4. Análisis de la Varianza.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	38285.811	7	5469.402	1546.567	.000[a]
	Residual	3211.124	908	3.536		
	Total	41496.934	915			

a. Predictors: (Constant), MORTPRED Mortalidad predicha, NOCIRUGÍ Número de cirugías, DIASFEND Días en falla endócrina, DIASFREN Días en falla renal, NPT Nutrición parenteral total, ANTIBIÓT Antibióticos, VMDIAS Días en ventilación

b. Dependent Variable: ESTUTI Estancia en UTI

Las hipótesis son las siguientes:

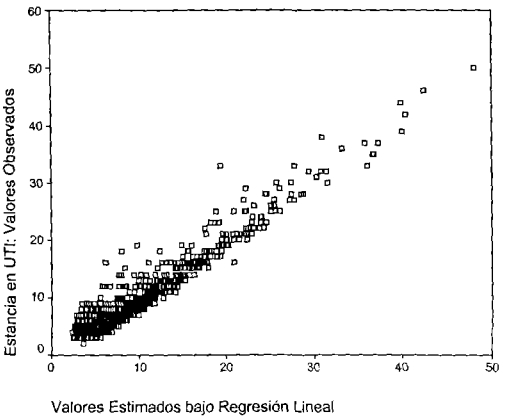
$$H_0: \beta_i = 0 \text{ con } i = 1,2,\ldots\ldots, p \text{ vs } H_1: \beta_i \neq 0 \text{ para al menos una } i,\ i = 1,2,\ldots\ldots, p$$

El valor de F en la tabla correspondiente al análisis de la varianza es de 1546.567 y el valor en tablas de la distribución F es de 2.020, como 1546.567 > 2.224 entonces rechazamos la hipótesis nula H_0: $\beta_0, \beta_1,\ldots\ldots \beta_5 = 0$, lo cual indica que si existe una relación lineal entre las variables ya que por lo menos uno de los coeficientes estimados es igual a cero, por lo tanto el modelo es bueno ya que se puede decir con una confianza del 95% que las variables tienen poder explicativo.

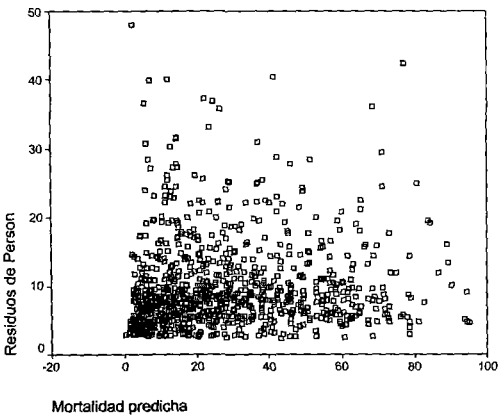
El coeficiente de determinación nos sirve para medir el poder explicativo del modelo, en este caso dicho coeficiente está expresado por $r^2 = 0.923$ este valor revela que el 92.3% de la variación de los días en terapia intensiva es explicado por las variables contenidas en el modelo. Por otro lado, el coeficiente de correlación r = 0.961 el cuál indica que existe una alta correlación positiva entre las variables explicativas y la variable dependiente *estuti*.

A continuación se presenta la gráfica correspondiente a los valores pronosticados contra la variable *estuti*. Como puede verse, la relación es lineal.

Gráfica 18.



Gráfica 19.



2.6.3 Regresión Poisson

El modelo de Regresión Poisson es también un modelo lineal generalizado y la variable de respuesta representa un conteo. Modela la relación entre los conteos observados y un conjunto de variables explicativas.

La distribución poisson está dada de la siguiente manera:

$$f_{yi}(y_i, \mu_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \quad \mu_i > 0, y_i = 0,1,2....$$

$$E(y_i) = \mu_i \quad y \quad Var(y_i) = \mu_i$$

El modelo de regresión poisson es el siguiente:

$$\ln(\mu_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Equivalentemente:

$$\mu_i = e^{\beta_0} e^{\beta_1 X_1} e^{\beta_2 X_2} .....e^{\beta_k X_k}$$

Para estimar los coeficientes de un modelo de Poisson generalmente se utiliza el método de máxima verosimilitud y se obtiene la solución usando métodos numéricos. Para la estimación usaremos la rutina *glm* del paquete STATA y un ejemplo de la sintaxis se presenta en el Anexo II. (Agresti.1996. Sección 4.3.1)

Por otro lado las estrategias para la selección del modelo son análogas a las usadas en el caso de regresión logística.

Una vez que se ajustó el modelo de regresión lineal múltiple, se procedió a ajustar un modelo de regresión poisson usando a las variables contenidas en el modelo 8, es decir:

diasfend, *diasfren*, *vmdias*, *npt*, *antibiot*, *norcirugi* y *mortpred*, con el fin de comparar las probabilidades estimadas por ambos modelos.

A continuación se presentan los resultados.

Tabla 5. Coeficientes Estimados para el Modelo de Regresión Poisson

estuti	Coef.	Std. Err.	z	P>z	[95% Conf. Intervalo]	
diasfren	0.0133945	.0034855	3.84	0.000	0.006563	0.020226
diasfend	0.0020868	.0081052	0.26	0.797	-0.0137991	0.0179727
norcirugias	-0.043866	0.0107625	-4.00	0.000	-0.0641806	-0.219925
vmdias	0.0602569	0.0018021	33.44	0.000	0.0567248	0.063789
mortpred	0.0000793	0.000517	0.15	0.878	-0.0009341	0.0010927
npt	0.1243637	0.273064	4.55	0.000	0.0708442	0.1778833
antibiot	0.0111795	0.0073205	1.53	0.127	-0.0031684	0.0255275
constante	1.666563	0.0235106	70.89	0.000	1.620483	1.712642

De acuerdo a los resultados obtenidos en la tabla anterior, el modelo de regresión poisson está dado de la manera siguiente:

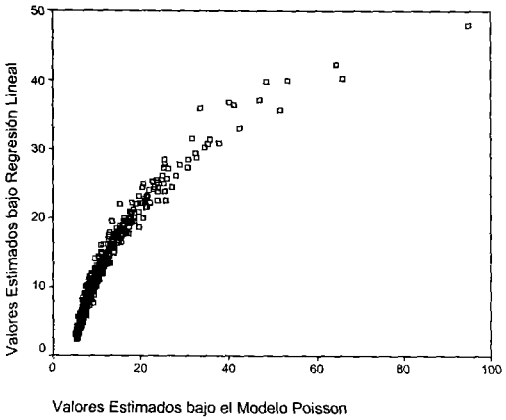
Modelo 9.

$$\ln(\mu_i) = 1.667 + .002\,diasfend + 0.013\,diasfren + .060\,vmdias + 0.024\,npt(si) + 0.011\,antibiot(si) - 0.044\,norcirugia\,s + 0.0000793\,mortpred$$

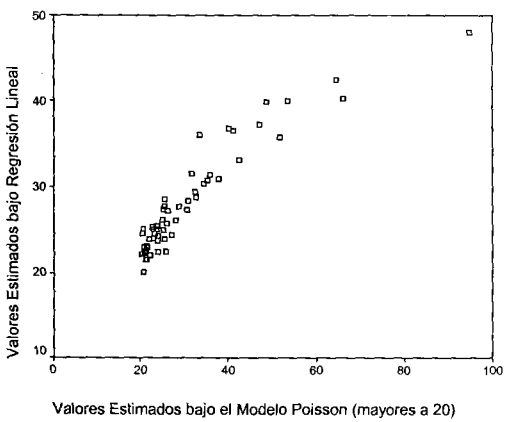
Del ajuste de los modelos de regresión lineal múltiple y regresión Poisson obtenemos resultados que nos llevan a las mismas conclusiones respecto a la dependencia de las variables explicativas y la variable de respuesta, ya que por un lado el mismo subconjunto de variables explicativas resultan estadísticamente significativas, y por otro lado al graficar las probabilidades estimadas de ambos modelos (gráfica 20a) se puede ver que la relación casi uno a uno para todos aquellos valores menores a veinte, mientras que para valores mayores a veinte (gráfica 20b), el modelo de regresión Poisson predice valores más altos que el modelo de regresión lineal.

Cabe mencionar que los coeficientes ajustados de ambos modelos no pueden ser comparados directamente ya que las escalas de medición son diferentes, es por eso que solamente se graficó las probabilidades estimadas de cada uno de ellos.

Gráfica 20a.

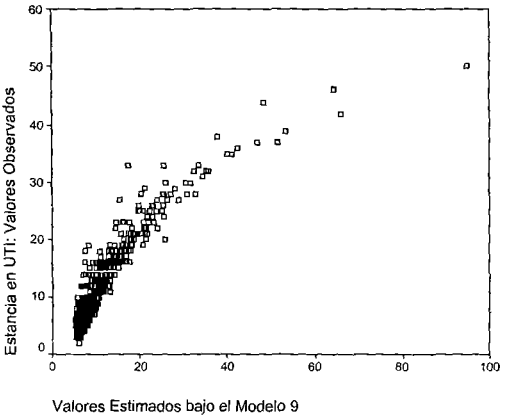


Gráfica 20b.



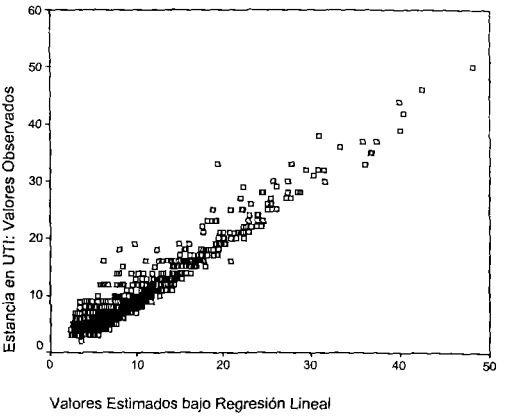
La siguiente gráfica muestra la relación entre los días de estancia en terapia intensiva y los valores estimados del modelo 9.

Gráfica 21.



La siguiente gráfica muestra la relación entre los días de estancia en terapia intensiva y los valores estimados bajo el modelo de regresión lineal.

Gráfica 18.



Independientemente de los resultados anteriores, se ajustó otro modelo de regresión Poisson usando las variables siguientes: *estuti*, *diasfren*, *diasfend*, *norcirugias*, *vmdias*, *mortpred*, *npt*, *antibiot*, *edad*, *traqueos*, *qxurgent*, *sepsisgr*, *epoc*, *cancer* y *diasfhem*.

Tabla 6. Coeficientes Estimados para el Modelo de Regresión Poisson Alternativo

estuti	Coef.	Std. Err.	z	P>z	[95% Conf. Intervalo]	
diasfren	0.0135939	0.0033127	4.10	0.000	0.0071011	0.0200867
traqueos	0.1029256	0.0287844	3.58	0.000	0.0465091	0.159342
norcirugias	-0.0293074	0.0107047	-2.74	0.006	-0.0502882	-0.0083265
vmdias	0.0559744	0.0017413	32.14	0.000	0.0525614	0.059873
sepsisgr	0.1347006	0.0266164	5.06	0.000	0.0825335	0.1868678
npt	0.1011698	0.0273655	3.70	0.000	0.0475345	0.1548052
constante	1.624149	0.0206623	78.60	0.000	1.583652	1.664646

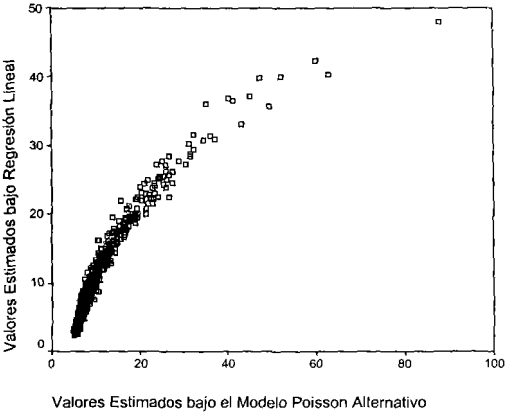
De acuerdo a la tabla anterior, el modelo resultante es:

Modelo 10.

$$\ln(\mu_i) = 1.624 + 0.014 diasfren + 0.103 traqueos\,(si) - 0.029 norcirugias + 0.056 vmdias + 0.135 sepsisgr\,(si) + 0.101 npt\,(si)$$

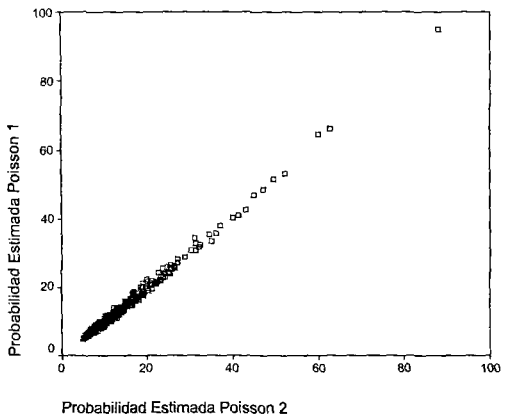
La gráfica que se presenta a continuación muestra la relación entre los días de estancia en terapia intensiva y las probabilidades estimadas del modelo 10.

Gráfica 22.



La siguiente gráfica muestra la relación entre los valores estimados del modelo 9 y los valores estimados del modelo 10. Como puede verse, la relación es aproximadamente de una recta a 45° que pasa por el origen, lo cual indica que no existe gran diferencia en la predicción de ambos modelos.

Gráfica 23.



Conclusiones

El total de las observaciones analizadas fue de 917, lo suficiente para ajustar modelos con el fin de detectar aquellas variables que influyen en: días de terapia intensiva, mortalidad y estado vital de los pacientes hospitalizados en el Centro Médico Nacional Siglo XXI y Centro Médico Nacional La Raza.

Con el modelo de regresión logística binaria se encontró que la edad, mortalidad predicha, falla neurológica, falla cardiaca, falla respiratoria, calidad de vida prehospitalaria e infección nosocomial son algunos de los factores más importantes que influyen en la muerte de los pacientes.

A. La edad resultó ser estadísticamente significativa ya que por cada año que aumenta dicha variable, se incrementa la probabilidad de morir

B. Una de las variables que se creía era muy importante para determinar la mortalidad de los pacientes era sepsis y efectivamente es importante ya que cuando la infección o sepsis es adquirida dentro del nosocomio la probabilidad de morir es mucho mayor que si hubiese sido adquirida de forma comunitaria o no tuviera infección.

C. Si los pacientes llevaban una mala calidad de vida antes de ser hospitalizados aumenta la probabilidad de que mueran.

D. Si se presenta una falla neurológica o una falla respiratoria, y considerando dentro del modelo a las variables: *edad*, *infnos01*, *mortpred*, *fcard*, *cv1codif*, *fneurol*, *frespirc*; los momios de que la persona muera son tres veces más grandes que los momios de que una persona muera dado que no tiene ninguna de las fallas anteriores.

El modelo de regresión logística multinomial no puede ser utilizado para predecir el estado vital de una persona después de salir de terapia intensiva ya que las mediciones que se están usando no lo permiten. Dicho modelo sirvió para ayudar a entender las asociaciones entre las variables. Esta observación se hace con base en los resultados obtenidos en la tabla de clasificación.

Las variables que influyen para determinar si alguno de los pacientes va a morir, vivir o tener una mala calidad de vida son: *qxurgent*, *fneurol*, *frespirc*, *infnos01*, calidad de vida antes de ser hospitalizados y *edad*.

Respecto al modelo de regresión lineal múltiple, se encontró que los factores que influyen en el número de días en que una persona se encuentre en terapia intensiva (estuti) son: los días en falla renal, días en ventilación mecánica, número de cirugías, antibióticos, mortalidad predicha, días en falla endocrina y nutrición parenteral total. Todas las variables anteriores tienen una relación lineal con los días en terapia intensiva lo cual indica que *estuti* depende de dichas variables.

Cuando hay presencia de nutrición parental y uso de antibióticos los días de terapia intensiva se incrementan en 0.509 y 0.178 fracciones de unidad respectivamente mientras que por cada día en falla endocrina, renal, ventilación mecánica y número de cirugías, los días en terapia aumentan 0.190, 0.084, 0.88 y 0.319 fracciones de unidad respectivamente; por otra parte, la asociación que existe entre la mortalidad predicha y los días en terapia intensiva es negativa, es decir, entre mayor sea la mortalidad predicha menor será el tiempo que la persona permanezca en terapia intensiva, esto se debe a valores altos de mortalidad predicha están asociados a una mayor probabilidad de muerte.

Con el modelo de regresión Poisson (Modelo 10) se llegó a la conclusión de que las variables que influyen en el tiempo en que una persona permanece en terapia intensiva son los días en falla renal, número de cirugías, días en ventilación mecánica, si tiene sepsis grave, si le hicieron una traqueotomía y si hay presencia de nutrición parental. Además se observó lo siguiente:

1) Por cada día en falla renal, se produce un incremento de 1.37% en el número promedio de días en terapia intensiva.

2) Si tiene sepsis grave, entonces el número promedio de días en terapia intensiva aumenta 14%. ($e^{0.1347006} = 1.1441$)

3) Si se le práctica una traqueotomía, el número promedio de días en terapia intensiva se incremente en un 11%.

4) Para el número de cirugías, días en ventilación mecánica y nutrición parental el número promedio aumenta en 96%, 6% y 13% respectivamente.

Finalmente comentamos que el uso de los Modelos Lineales Generalizados en particular de la regresión logística binaria, tricotómica, regresión lineal múltiple y regresión Poisson, son modelos que auxilian a entender relaciones de dependencia entre variables explicativas y una variable de respuesta. En este trabajo hemos ilustrado su uso y aplicación usando un conjunto de observaciones proveniente de un estudio médico actual.

ANEXO I

Anexo 1. Estadísticas Descriptivas de las Observaciones

Matriz de Correlación de algunas variables continuas

Correlation Matrix

		VMDIAS Días en ventilación	DIASSEDA Días en sedación	RELAJACI Relajación	DIASFRES Días en falla respiratoria
Correlation	VMDIAS Días en ventilación	1.000	.761	.447	.885
	DIASSEDA Días en sedación	.761	1.000	.480	.751
	RELAJACI Relajación	.447	.480	1.000	.440
	DIASFRES Días en falla respiratoria	.885	.751	.440	1.000

Tablas Cruzadas

Vivomuer * Edad codificada

VIVMUERT Variable blanco * EDADCOD Edad codificada Crosstabulation

Count

		EDADCOD Edad codificada			
		0 40 ó menos	1 41 a 60 años	2 61 ó más	Total
VIVMUERT Variable blanco	0 Vivo	119	108	136	363
	1 Muerto	69	142	272	483
Total		188	250	408	846

Vivomuer * Sexo

VIVMUERT Variable blanco * SEXO Sexo Crosstabulation

Count

		SEXO Sexo		
		0 Femenino	1 Masculino	Total
VIVMUERT Variable blanco	0 Vivo	178	185	363
	1 Muerto	270	213	483
Total		448	398	846

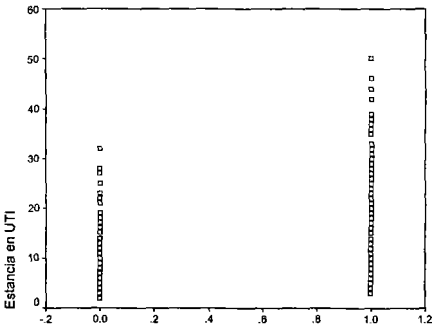
Sepsis grave * Vivomuer

SEPSISGR Sepsis grave * VIVMUERT Variable blanco Crosstabulation

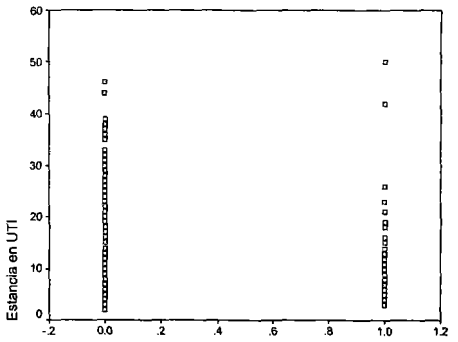
Count

		VIVMUERT Variable blanco		
		0 Vivo	1 Muerto	Total
SEPSISGR Sepsis grave	0 No	185	161	346
	1 Sí	178	322	500
Total		363	483	846

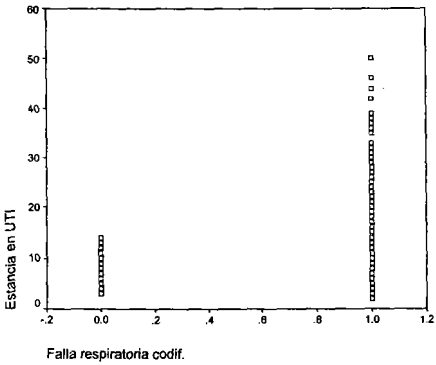
Gráficas de las Variables Categóricas



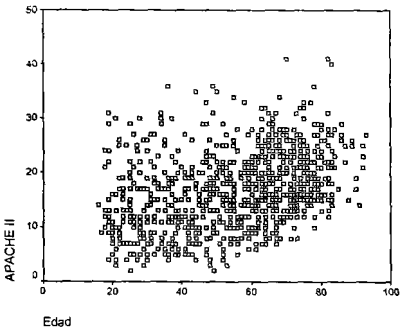
Falla cardiaca codificada

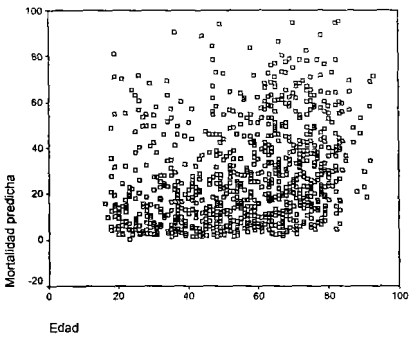


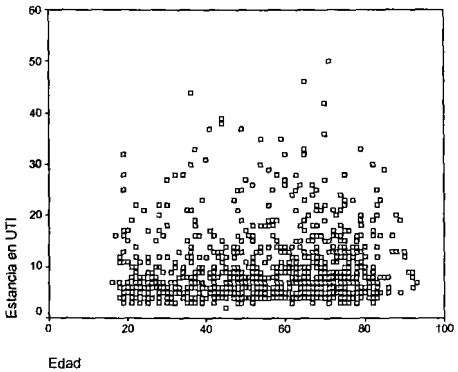
Falla neurológica codificada



Gráficas de las Variables Continuas

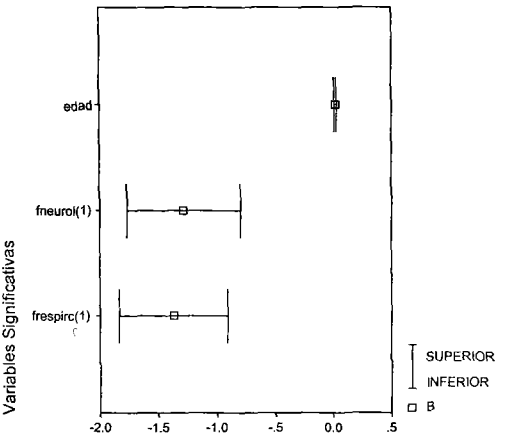




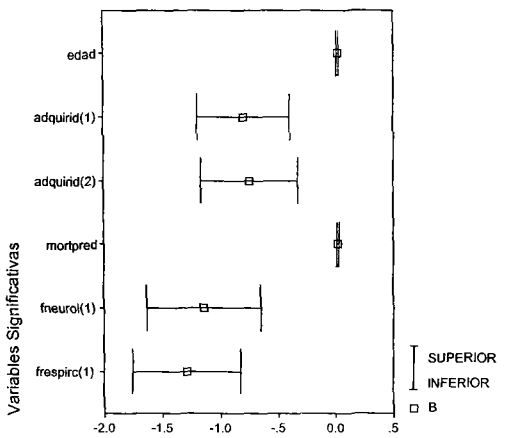


Graficas Correspondientes a los Intervalos de Confianza para los Coeficientes Estimados de los Modelos 1 y 2

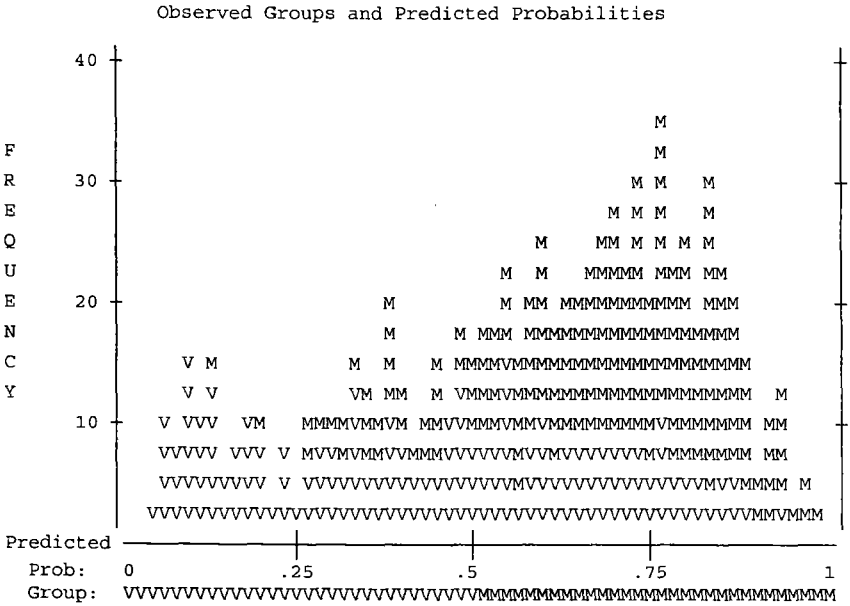
Intervalo de Confianza de los Coeficientes Estimados de las Variables Significativas del Modelo 1



Intervalo de Confianza de los Coeficientes Estimados de las Variables Significativas del Modelo 2

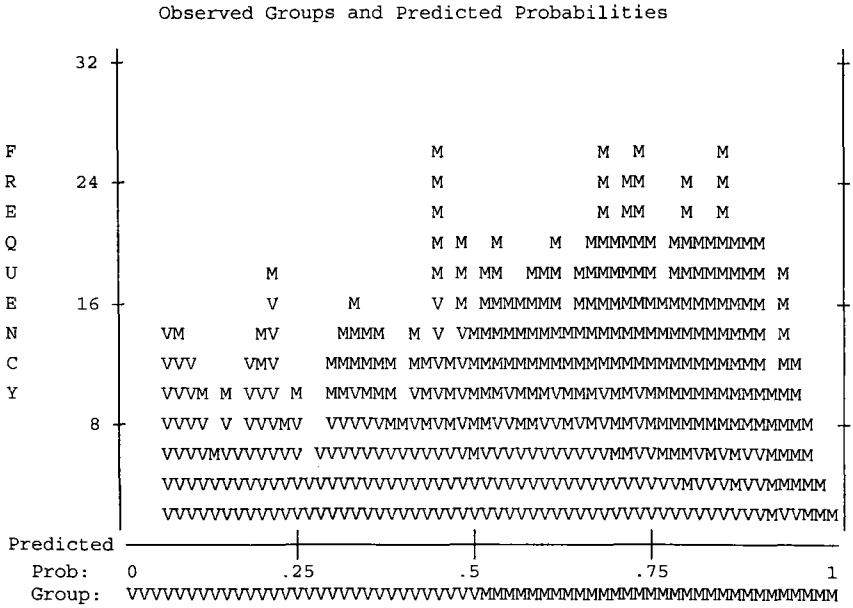


Histograma de Probabilidades Modelo 1



Predicted Probability is of Membership for Muerto
The Cut Value is .50
Symbols: V - Vivo
M - Muerto
Each Symbol Represents 2.5 Cases.

Histograma de Probabilidades Modelo 5



Predicted Probability is of Membership for Muerto
The Cut Value is .50
Symbols: V - Vivo
M - Muerto
Each Symbol Represents 2 Cases.

ANEXO II

Anexo2. Comandos que se utilizaron para estimar los coeficientes de los modelos usando el paquete estadístico SPSS.

★ Instrucciones para ajustar el modelo incluyendo a las enfermedades por medio del método Forward

```
LOGISTIC REGRESSION VAR=vivomuer
/METHOD=FSTEP(COND) epoc dm dmcompl ulcera irc cáncer icc esteroid edad
/CONTRAST (epoc)=Indicator /CONTRAST (dm)=Indicator /CONTRAST
(dmcompl)=Indicator /CONTRAST (ulcera)=Indicator /CONTRAST (irc)=Indicator
/CONTRAST (cáncer)=Indicator /CONTRAST (icc)=Indicator /CONTRAST
(esteroid)=Indicator
/CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

★ Instrucciones para ajustar el modelo 5 incluyendo la nueva variable infnos01

```
LOGISTIC REGRESSION VAR=vivomuer
 /METHOD=FSTEP(COND) edad cáncer estuti infnos01 mortpred fcard
 cv1codif fneurol frespirc
 /CONTRAST (fcard)=Indicator /CONTRAST (cáncer)=Indicator /CONTRAST
 (infnos01)=Indicator /CONTRAST (cv1codif)=Indicator /CONTRAST
 (fneurol)=Indicator /CONTRAST (frespirc)=Indicator /CONTRAST
 (cv2codif)=Indicator
 /CLASSPLOT
 /PRINT=GOODFIT CI(95)
 /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

★ Instrucciones para ajustar el modelo usando sepsis y mortpred por medio del método Backward

```
LOGISTIC REGRESSION VAR=vivomuer
 /METHOD= BSTEP(COND) edad cáncer estuti sepsisgr mortpred fcard
 cv1codif fneurol frespirc
 /CONTRAST (fcard)=Indicator /CONTRAST (cáncer)=Indicator /CONTRAST
 (sepsisgr)=Indicator /CONTRAST (cv1codif)=Indicator /CONTRAST
```

(fneurol)=Indicator /CONTRAST (frespirc)=Indicator /CONTRAST
(cv2codif)=Indicator
/CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .

Comandos que se utilizaron para estimar los coeficientes de los modelos usando el paquete estadístico STATA.

★ Instrucciones para ajustar el modelo 5

logistic vivmuert cv1codif mortpred infnos01 edad fneurol fcard frespirc cancer estuti, coef

★ Instrucciones para obtener los residuos de la devianza, pearson y probabilidades

predict rd1, deviance

predict rp1, pearson

predict probestim, p

★ Instrucciones para ajustar el modelo 9

glm estuti diasfend diasfren vmdias npt antibiot norcirugi mortpred, family(poisson) link(log)

★ Instrucciones para ajustar el modelo 10

sw glm estuti diasfend diasfren vmdias npt antibiot norcirugi sepsisgr diasfhem edad qxurgent cancer epoc traqueos mortpred, family(poisson) link(log) pr(.05)

Bibliografía

Hosmer, D. Lemeshow, S. 2000. Applied Logistic Regression. 2a ed. John Wiley & Sons. Nueva York.

Agresti, A. 2002. Categorical Data Analysis. John Wiley & Sons.

Agresti, A. 1996. An Introduction to Categorical Data Analysis. John Wiley & Sons.

Luque Martínez, Teodoro 2000. Técnicas de Análisis de Datos en Investigación de Mercados. ed. Pirámide.

Silva Alcayer, Luis Carlos.1994. Excursión a la Regresión Logística en Ciencias de la Salud. Madrid: Díaz de Santos.

D. Collet.1991. Modelling Binary Data. Chapman & Hall.

Thomas P. Ryan.1997. Modern Regression Methods. John Wiley & Sons.

Medrano Ortiz, María Guadalupe.2004. Modelos de Regresión Logística para la Evaluación de Riesgo