UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

REGRESIÓN LOGÍSTICA APLICADA A LA EVALUACIÓN DE RIESGOS

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 :

MEDRANO ORTÍZ MARÍA GUADALUPE

DIRECTORA DE TESIS:

DRA. GUILLERMINA ESLAVA GÓMEZ

2004

UNIVERSIDAD NACIONAL AVENOMA DE MEXICO

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:

"Regresión Logística aplicada a la evaluación de riesgos"

realizado por María Guadalupe Medrano Ortíz

con número de cuenta 9505380-8 , 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	Mat. Margarita Elvira Chávez Cano	
Propietario	M.en A.P. María del Pilar Alonso Reyes	
Suplente	Act. Francisco Sánchez Villarreal	
Suplente	Act. Eduardo Hernández Pérez	

Consejo Departamental de

FACULTAD DE CIENCIAS

Act. Jaime Vázquez Alamilla

Agradezco a mis padres y hermanos por el
apoyo, la constante motivación, el cariño y
sobre todo quiero agradecer que siempre han
estado a mi lado, por eso, a ellos les dedico esta obra.

También quiero agradecer a la Dra. Guillermina Eslava Gómez
por su atención y ayuda en la realización de esta tesis, asimismo
a la Maestra Victoria Caraveo Enríquez y al Doctor Luís David
Sánchez Velásquez, por facilitarme las bases de datos utilizadas
con fines estrictamente académicos.

ÍNDICE

RESUMEN

La aplicación del modelo de regresión logística es el tema de tesis porque resulta ser de gran utilidad para estudiar las relaciones de asociación entre una variable dependiente categórica y un conjunto de variables independientes, las cuales pueden ser continuas o categóricas. El modelo de regresión logística es considerado una técnica estadística que puede ser aplicada a cualquier campo de interés.

Se realizó la aplicación del análisis a dos campos de estudio distintos entre si pero con la misma finalidad, de analizar el cambio que se da en la probabilidad de ocurrencia (éxito) de un suceso de interés (variable respuesta), con base en un conjunto de variables.

El propósito de esta tesis es mostrar la aplicación de modelos estadísticos en dos campos de estudio, el sector médico y el sector social para identificar e indagar las diferencias entre las distintas características que son consideradas según el tipo de investigación que se esté realizando.

En las dos bases de datos trabajadas se obtuvieron variables cualitativas y cuantitativas por lo tanto se realizó una inspección inicial de las observaciones sacando las frecuencias de las variables explicativas de mayor interés para el estudio con la finalidad de obtener el valor que contiene el mayor número de observaciones para establecerlo como categoría de referencia, etiquetándola con el valor de cero y así continuar hasta llegar con la categoría de menor número de observaciones.

Los pasos a seguir para la realización de cada análisis de regresión logística son los siguientes:

1) Plantear el problema a resolver determinando cuál es la variable dependiente y cuáles son las variables independientes.
2) Estimar y seleccionar el modelo más compatible con los datos y con el problema.
3) Estudiar la evaluación del ajuste del modelo seleccionado analizando la existencia de observaciones influyentes.
4) Interpretar los resultados obtenidos.

El primer análisis realizado fue la elección de carrera de los estudiantes provenientes de 2 instituciones, la Universidad Autónoma Metropolitana y la Universidad Iberoamericana, se realizó con el propósito de identificar los factores que intervienen en la elección de 3 distintas carreras, licenciatura de administración, licenciatura de nutrición e ingeniería.

Se construyeron 5 modelos logísticos, los 3 primeros corresponden a cada una de las carreras teniendo como valor en la variable respuesta, 0 cuando no se elige la carrera y 1 cuando se eligió, la cuarta corresponde a una variable de respuesta tricotómica considerando a las tres distintas carreras y se toma como categoría de referencia a la carrera de administración porque fue la que contenía el mayor número de observaciones y por tener una población estudiantil con aproximadamente el mismo número de hombres y de mujeres, el quinto modelo corresponde a la variable respuesta tipo de universidad a la que ingresaron los estudiantes en el año de 1996.

El propósito de la aplicación del modelo de regresión logística fue para identificar algunos de los factores socioeconómicos, culturales, familiares y proyectos de vida que tuvieron los estudiante al elegir carrera, la modelación permitió corroborar que el sexo del estudiante, la educación de ambos padres, la ocupación que tiene la madre y activad que el estudiante piensa realizar al terminar la carrera fueron factores influyentes para la elección de carrera.

El segundo análisis realizado trató sobre la calidad de vida que tienen los pacientes al salir de la Unidad de Terapia Intensiva (UTI) provenientes de dos instituciones, el hospital Centro Médico Nacional Siglo XXI y Centro Médico Nacional la Raza, se construyó solamente un modelo logístico teniendo como objetivo identificar las características, demográficas, clínicas, paraclínicas terapéuticas y de morbilidad con respecto a los pacientes que tuvieron mala calidad de vida a los 3 meses del egreso hospitalario. El análisis corroboró que las características de la edad del paciente, hospital en el cual estaban internados, servicio que requirieron, la estancia posterior a la UTI, si tenían enfermedad de cáncer o enfermedad pulmonar y calidad de vida a los 2 meses previos de la UTI fueron factores que tuvieron una relación de asociación con la variable respuesta.

INTRODUCCIÓN

Esta tesis se encuentra dentro del área estadística, especialmente dentro del campo de Análisis de Regresión Logística, siendo un tema que ha adquirido fuerza en los últimos años por ser una herramienta estadística que puede ser aplicada en diversos campos de la investigación, siempre y cuando cumpla con los requerimientos necesarios para su uso.

El Análisis de Regresión Logística permite trabajar con un gran número de variables para cuantificar la relación que existe entre la probabilidad de ocurrencia de un determinado suceso que sea de interés y así conocer el comportamiento que podría tener ante la intervención de otros factores, con un resultado positivo o negativo en relación al suceso.

En la búsqueda de tema para la realización de tesis se encontró que no existe un gran uso del Análisis de Regresión Logística ya que solamente se encontraron 10 tesis referentes al mismo tema, están registradas en la Biblioteca Central de la UNAM, siendo ésta una motivación por el cual me interese en explorar y especificar cual es la metodología necesaria para determinar que factores influyen en la ocurrencia de un determinado suceso.

Se ha escrito en forma sencilla con la finalidad de que este trabajo sea útil para la comprensión y el entendimiento de aquellos estudiantes que comienzan a tener inquietudes sobre una de las tantas aplicaciones que puede llegar a tener la estadística, así como a toda persona que esté ávida de conocimientos.

La metodología es aplicable a dos temas, el primero se encuentra dentro del sector social, que trata de la elección de carrera de los jóvenes estudiantes

de preparatoria dicha elección involucra factores económicos, sociales, hasta las expectativas de vida que se hayan establecido.

El segundo tema a tratar, se encuentra dentro del sector médico, factor importante para la población en general, ya que al realizar este tipo de investigaciones los hospitales al igual que los médicos van a estar mejor capacitados para dar un buen servicio a las personas que lo necesitan. El factor de interés a investigar es la calidad de vida a los 3 meses de egreso hospitalario de aquellos pacientes que salen de la Unidad de Terapia Intensiva (UTI) de dos instituciones diferentes del IMSS.

Por lo tanto, se describirá brevemente la metodología a seguir para la aplicación del análisis de regresión logística, se determinará el modelo para la variable de respuesta dicotómica (con 2 posibles valores), así como la determinación de las consideraciones a tomar para la obtención de los coeficientes estimados del modelo y las pruebas de significancia, también se establecerán las estadísticas necesarias para determinar si el modelo propuesto se encuentra ajustado correctamente y por último se establecerá el modelo para una variable respuesta de tipo politómico (con más de 2 posibles valores).

Posteriormente en los siguientes capítulos se llevará a cabo la aplicación del análisis de regresión logística, en los dos temas a utilizar. En forma individual se describirá el procedimiento llevado a cabo para la obtención del mejor modelo propuesto, interpretación de los coeficientes, pruebas de hipótesis sobre los parámetros, preparación de la base de datos y las pruebas de significancia del modelo.

El segundo tema a tratar, se encuentra dentro del sector medico, factor importante para la población en general, ya que al realizar este tipo de investigaciones los hospitales al igual que los médicos van a estar mejor capacitados para dar un buen servicio a las personas que lo necesitan. El factor de interés a investigar es la calidad de vida (que puedan valerse por si solos a los 3 meses de egreso hospitalario), de aquellos pacientes que salen

de la Unidad de Terapia Intensiva (UTI) de dos instituciones diferentes del IMSS.

Por lo tanto se describe brevemente la metodología a seguir para llevar a cabo la aplicación del análisis de regresión logística, se determina el modelo para la variable de respuesta dicotómica (con 2 posibles valores), así como la determinación de las consideraciones a tomar para la obtención de los coeficientes estimados del modelo y las pruebas de significancia, también se establecen las estadísticas necesarias para determinar si el modelo propuesto se encuentra ajustado perfectamente para no tener una sobre estimación y por último se establece el modelo para una variable de respuesta de tipo politómico (con más de 2 posibles valores).

Posteriormente en los siguientes capítulos se lleva a cabo la aplicación del análisis de regresión logística, en los dos temas a utilizar. En forma individual se describe el procedimiento llevado a cabo para la obtención del mejor modelo propuesto, interpretación de los coeficientes, pruebas de hipótesis sobre los parámetros, preparación de la base de datos y las pruebas de significancia del modelo.

CAPÍTULO 1

MODELOS DE REGRESIÓN LOGÍSTICA PARA LA EVALUACIÓN DE RIESGO

En este capítulo se describen las herramientas estadísticas para llevar a cabo el ajuste de un modelo a las variables explicativas (pueden ser dicotómicas o continuas) con respecto a una variable de respuesta (binaria) y así determinar la influencia de las variables explicativas sobre la probabilidad de ocurrencia de un suceso.

Se iniciará por la especificación de la metodología para la aplicación del análisis de regresión logística, considerando el más usado dentro de los modelos lineales generalizados para una variable de respuesta binaria (caso dicotómico, con dos posibles valores), después se establecen las herramientas necesarias para la estimación de los coeficientes al igual que el intervalo de confianza para cada uno de ellos; esto se obtiene a partir del significado de las tablas de contingencia de las probabilidades ajustadas, finalmente se presenta el modelo logístico para la variable de respuesta multinomial (variable dependiente politómica con más de 2 posibles valores), al igual que se establecen las estadísticas necesarias para llevar a cabo el ajuste y la evaluación requerida según el tipo de investigación.

1.1. Modelo de Regresión Logística

Modelos Lineales Generalizados

Dentro de los modelos lineales generalizados (MLG) se encuentran los modelos logísticos, siendo éstos los más utilizados para el caso cuando se tiene una variable de respuesta binaria.

Los modelos lineales generalizados se componen de tres partes, éstas son: *aleatoria, sistemática y la función liga.*

En primera instancia se encuentra el componente *aleatorio* que parte de las observaciones: $y_1, y_2, \ldots y_n$ de una distribución de la familia exponencial con n variables independientes, cada observación y_i tiene función de densidad de la forma:

$$f(y_i;\theta) = a(\theta)b(y_i)\exp(y_i Q(\theta))$$

El valor del parámetro θ puede tomar los valores de $\theta_1,\ldots,\theta_n$ y el término $Q(\theta)$ es llamado el parámetro natural de la distribución o componente aleatorio.

El *sistemático* o predictor lineal de las variables explicativas, relaciona un vector $\beta = (\beta_0, \beta_2, \ldots \beta_{k-1})$ con las variables independientes $X_1, X_2, \ldots X_{k-1}$, quedando de la siguiente manera:

$$\eta_i = \beta_0 + \sum_{i=1}^{k-1} \beta_j X_{ij}, \quad \text{para toda } i = 1,2,\ldots n.$$

Por último se tiene la *función liga,* la cual relaciona la componente aleatoria con el componente sistemático. Sea $\mu_i = E(y_i)$ con $i = 1,2,\ldots n.$, estando μ ligada a β_i por $\beta_i = g(\mu)$, siendo g una función monótona diferenciable.

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{k-1} \beta_j X_{ij}$$

Se le llama función canónica a la función liga que transforma la media del parámetro natural (Ver Agresti, 2000, Pág. 116).

$$g(\mu_i) = Q(\theta_i) = \beta_0 + \sum_{j=1}^{k-1} \beta_j X_{ij}$$

1.2. MODELO LOGÍSTICO

Supóngase que se tiene una muestra de n observaciones independientes de la pareja (x_i, y_i) con $i = 1, 2, \ldots, n$, donde y_i denota el valor de la variable respuesta y x_i el valor de la variable explicativa para el $i-iésimo$ sujeto.

Sea: Y la variable de respuesta, con dos posibles valores 0 y 1.

Por lo tanto la relación que existe entre la variable de respuesta y la variable independiente se determina por el valor esperado de la variable respuesta.

A este valor se le conoce como la esperanza condicional, esto es:

$$E(Y|X)$$

Siendo Y la variable respuesta y X la variable independiente, en una regresión lineal se tiene que la esperanza puede ser expresada como una función lineal para una variable independiente X, tal que:

$$E(Y\ X) = \beta_0 + \beta_1 X, \qquad \text{tal que,} \qquad -\infty < E(Y\ X) < \infty$$

O bien, el modelo se puede escribir como:

$$Y = \beta_o + \beta_1 X + \varepsilon$$

El modelo anterior no puede ser utilizado para la evaluación de valores específicos de la variable independiente ya que se obtendría en el mayor número de casos un valor diferente de 1 y de 0 (los valores posibles de Y), lo cual carece de todo sentido. Por esta razón la regresión lineal no puede ser utilizada para la situación descrita por lo tanto se va a utilizar el método de regresión logística.

Se ha hablado de la utilización del modelo de regresión logística ya que tiene como prioridad expresar la probabilidad de que ocurra un riesgo descrito en función de ciertas variables que se presumen relevantes o influyentes para la variable respuesta o variable dependiente, A continuación se establecerá el modelo para el caso en que la variable dependiente únicamente cuenta con dos posibles valores $[0,1]$, caso dicotómico, esto es:

Sea:

$$Y = 1, \quad la\ ocurrencia\ del\ suceso\ (éxito)$$
$$Y = 0, \quad la\ no\ ocurrencia\ del\ suceso\ (fracaso)$$

La distribución Bernoulli para variables aleatorias binarias corresponde a:

$$P(Y = 1) = \pi$$
$$P(Y = 0) = 1 - \pi$$

Con:

$$E(Y) = \pi$$

Entonces la probabilidad de la relación que existe entre π y X bajo un modelo de regresión logística binaria, puede ser representada por:

$$\pi(X) = \frac{exp(\beta_0 + \beta_1 X_1 + \varepsilon)}{1 + exp(\beta_0 + \beta_1 X_1 + \varepsilon)} \qquad (A)$$

Satisfaciendo que: $0 \leq \pi \leq 1$.

El modelo de forma equivalente a (A) puede ser escrito usando la transformación logística o transformación ***logit*** como:

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X + \varepsilon$$

Por lo anterior se tiene que el modelo de Regresión Logística consiste de una ecuación matemática que relaciona a un conjunto de variables llamadas explicativas con la probabilidad de ocurrencia de una variable llamada respuesta, que se supone están relacionadas de manera lineal, de la forma siguiente:

$$\ln\left[\frac{\pi(X)}{(1-\pi(X))}\right] = \beta_0 + \beta_1 X + \varepsilon_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (B)$$

Donde los $\beta_i's$ son los parámetros desconocidos del componente lineal del modelo, X el valor de la variable explicativa, Y la variable respuesta representada como: $\ln\left[\frac{\pi(X)}{(1-\pi(X))}\right]$ y por último ε el error aleatorio.

La transformación ***logit*** produce un modelo que es lineal en los parámetros, y toma valores entre $-\infty$ e ∞, dependiendo del rango que tenga x.

Gráficamente se tiene una curva que representa la relación que existe entre la variable respuesta y las variables explicativas, esto es, $E\left(Y \mid X\right)$, se muestra la gráfica para una variable explicativa de $X = x$.

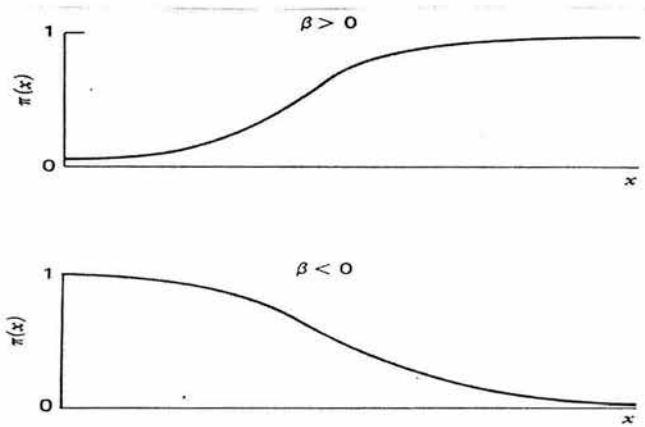


FIGURA 1. Curva de Regresión Logística, con relación creciente o decreciente según sea el signo de β.

Se observa que cuando $x \to \infty$

$$\pi(x) \to 0, \text{ si } \beta < 0$$

y

$$\pi(x) \to 1, \text{ si } \beta > 0$$

Si $\beta \to 0$, la curva tiende a una línea horizontal.

1.2.1. ESTIMACIÓN DE LOS PARÁMETROS

FUNCIÓN DE VEROSIMILITUD

Comúnmente el método más usado para la estimación de los parámetros del modelo de regresión logística es el método de máxima verosimilitud.

Donde la función de verosimilitud como función de los parámetros desconocidos está definida como la función de probabilidades conjunta de las variables aleatorias.

El estimador de máxima verosimilitud corresponde al valor máximo de la función de verosimilitud.

A partir de la variable respuesta con dos posibles valores 0 y 1, la probabilidad condicional es:

$$P(Y=1 \mid x) = 1-\pi(x)$$

y

$$P(Y=0 \mid x) = \pi(x)$$

La función de verosimilitud correspondiente para una muestra de tamaño n con observaciones de $y_1, y_2, \ldots, y_n$, con variables aleatorias de $Y_1, Y_2, \ldots, Y_n$, con Y_i independientes por lo tanto la función de densidad queda:

Se tiene que $f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$, $Y_i = 0,1$; $i = 1,2,\ldots,n$.

$$L(\beta_0,\beta_1) = \prod_{i=1}^{n} \left[\pi_i^{Y_i}(1-\pi_i)^{1-Y_i}\right]\ldots$$

$$= \prod_{i=1}^{n}\left(\pi_i^{Y_i}\right)\left[\prod_{i=n_1+1}^{n}(1-\pi_i)^{1-Y_i}\right] \ldots\ldots\ldots\ldots (C)$$

o equivalentemente

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \left(\frac{e^{\beta_0 + \beta_1 x}}{1 - e^{\beta_0 + \beta_1 x}} \right)^{Y_i} \left(\frac{e^{\beta_0}}{1 - e^{\beta_0}} \right)^{1 - Y_i}$$

El estimador de máxima verosimilitud es obtenido de forma equivalente al maximizar el logaritmo de la función de densidad de probabilidad conjunta a partir de (C), esto es:

$$ln < \left(\beta_0, \beta_1\right) = \sum_{i=1}^{n} Y_i \ln(\pi_i) + \sum_{i=1}^{n} (1 - Y_i) \ln(1 - \pi_i)$$

$$= \sum_{i=1}^{n} Y_i \ln\left(\frac{\pi_i}{1 - \pi_i} \right) + \sum_{i=1}^{n} \ln(1 - \pi_i)$$

Cuando el valor de $\pi(X) = \dfrac{exp\left(\beta_0 + \beta_1 X_1 + \varepsilon\right)}{1 + exp\left(\beta_0 + \beta_1 X_1 + \varepsilon\right)}$, nos queda el resultado como:

$$\ln(L(\beta_0, \beta_1)) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 x_i) + \sum_{i=1}^{n} \ln(1 + \exp(\beta_0 + \beta_1 x_i))^{1} \ldots\ldots\ldots\ldots\ldots\ldots (D)$$

Para encontrar los valores de β que maximicen $L(\beta_0, \beta_1)$, se deriva la ecuación (D) con respecto a β_0 y β_1, igualando la ecuación a cero, quedando de la siguiente manera:

$$\frac{\partial \ln(L(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} (Y_i) - \sum_{i=1}^{n} \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = 0 \qquad (E)$$

y

$$\frac{\partial \ln(L(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} Y_i x_i - \sum_{i=1}^{n} \frac{x_i \exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = 0 \qquad (F)$$

---

[1] *Para más información sobre la demostración presentada revisar Agrest, pag. 258-259.*

Las dos ecuaciones se deben resolver simultáneamente, no son lineales en β_0 y β_1 por lo tanto se debe llevar a cabo la aplicación de un método numérico para darles solución, por ejemplo el Método de Newton, ya que presenta una mayor eficiencia para este tipo de problemas.

1.3. GENERALIZACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA

El modelo logístico puede ser generalizado directamente a la situación cuando se tienen $X = (x_1, x_2, \ldots, x_p)$ variables independientes. La probabilidad de π es modelada como:

$$\pi(x) = P(Y = 1 X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$$

$$= \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}} \qquad \text{(G)}$$

La ecuación (B) es llamada la función de regresión logística no es lineal en los parámetros, pero al aplicar la transformación *logito* produce una función lineal en los parámetros.

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_o + \beta_1 X_1 + \ldots + \beta_p X_p$$

Si π es la probabilidad de que un evento ocurra, la razón queda $\frac{\pi}{1-\pi}$, llamado el coeficiente de momios del evento, se tiene:

$$\frac{\pi(Y=1)}{1-\pi(Y=0)} = \frac{\pi}{1-\pi} = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}$$

Aplicando el logaritmo natural:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Se debe recordar que los rangos de π se encuentran entre 0 y 1, mientras que el rango de los valores de $ln\left(\pi_{1-\pi}\right)$ se encuentran entre $(-\infty,\infty)$.

1.3.1. ESTIMACIÓN DE LOS PARÁMETROS.

FUNCIÓN DE VEROSIMILITUD

Para la obtención de los estimadores de máxima de verosimilitud procede de la siguiente manera:

Se tiene una muestra (x_i, y_i), de n observaciones independientes de la pareja (X,Y), con $i = 1, 2, \ldots, n$, se quiere obtener la estimación de los parámetros $\beta_0, \beta_1, \ldots, \beta_p$. Se obtiene la función la función verosimilitud.

$$l(\beta) = \prod_{i=1}^{n} (\pi_i)^{Y_i} [1-\pi_i]^{1-Y_i}$$

El logaritmo de verosimilitud:

$$\ln(l(\beta)) = \sum_{i=1}^{n} \{Y_i \ln[\pi_i] + (1-Y_i)\ln[1-\pi_i]\}$$

Al derivar se obtiene $p+1$ ecuaciones de verosimilitud, las ecuaciones resultantes son expresadas en términos de π, como a continuación se presentan:

$$\sum_{i=1}^{n}(Y_i - \pi_i) = 0$$

y

$$\sum_{i=1}^{n}(x_{ij}[Y_i - \pi_i]) = 0 \quad \text{Para } j = 1, 2, \ldots, p$$

El método de estimación de las varianzas y covarianzas de los coeficientes estimados se obtienen a partir de la matriz cuyos elementos son la segunda derivada parcial del logaritmo de la función de verosimilitud. Las derivadas parciales son:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i) = 0$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_u} = -\sum_{i=1}^{n} x_{ij} x_{iu} \pi_i (1 - \pi_i) = 0$$

A estas dos ecuaciones se resuelven en forma simultánea e interactivamente.

La matriz de información es denotada como $I(\beta)$. Las covarianzas y varianzas se obtienen a partir de la inversa de la matriz, $I^{-1}(\beta)$. La estimación de las varianzas y covarianzas a partir de $I^{-1}(\hat{\beta})$.

La estimación del error estándar de los coeficientes estimados es:

$$\hat{SE}(\hat{\beta}_j) = (\hat{\sigma}^2(\hat{\beta}_j))^{1/2}, \text{ para } j = 0, 1, \ldots, p. \quad [2]$$

Por lo tanto la matriz queda de la siguiente manera.

$$\hat{I}(\hat{\beta}) = X'VX$$

[2] *Para ver como se obtuvieron los resultados expuestos ver Hosmer y Lemeshow, pag.27-30.*

Donde X es una matriz de dimensión $n\times(p+1)$, que contiene a las observaciones de cada sujeto y V es una matriz diagonal de $n\times n$ con elementos generales de $\hat{\pi}_i(1-\hat{\pi}_i)$, quedando de la siguiente manera:

Con diagonal $(\hat{\pi}_1(1-\hat{\pi}_1),\hat{\pi}_2(1-\hat{\pi}_2),\ldots,\hat{\pi}_n(1-\hat{\pi}_n))$

$$X=\begin{pmatrix}1 & x_{11} & \cdots & x_{1p}\\ 1 & x_{21} & \cdots & x_{2p}\\ \vdots & & \vdots & \\ 1 & x_{n1} & \cdots & x_{np}\end{pmatrix},$$

$$V=\begin{pmatrix}\hat{\pi}(x_1)(1-\hat{\pi}(x_1) & 0 & \cdots & 0\\ 0 & \hat{\pi}(x_2)(1-\hat{\pi}(x_2)) & \cdots & 0\\ \vdots & & \vdots & \\ 0 & 0 & \ldots & \hat{\pi}(x_n)(1-\hat{\pi}(x_n))\end{pmatrix}$$

1.4. *PRUEBAS DE SIGNIFICANCIA*

Para determinar que una variable explicativa desde el punto de vista estadístico, tiene una asociación con la variable de respuesta, se requiere de la utilización de pruebas de hipótesis, se tiene la prueba de Wald y la prueba del coeficiente de verosimilitud siendo más potente que la prueba de Wald.

1.4.1. *PRUEBA DEL COEFICIENTE DE VEROSIMILITUD*

La estadística de prueba del coeficiente de verosimilitud se basa en el logaritmo de la función de verosimilitud comparando los valores observados y esperados (estimados) de la variable repuesta, también puede ser utilizado como una estadística de bondad de ajuste.

El método de máxima verosimilitud consiste en maximizar una función llamada función de verosimilitud, que es la función de probabilidad asociada a un conjunto de datos en función de los parámetros desconocidos, de tal modo que los estimados obtenidos maximizan la probabilidad de obtener los datos observados.

La estadística de comparación del modelo saturado y el ajustado el cual es un submodelo del modelo saturado, se conoce como el cociente de verosimilitud es la devianza, el cual se define como:

$$D = -2\ln\left[\frac{Función\ de\ verosimilitud\ del\ modelo\ ajustado}{Función\ de\ verosimilitud\ del\ modelo\ saturado}\right] \qquad (H)$$

$$D = -2\left[ln\ fun.\ de\ Ver.\ del\ modelo\ ajustado - ln\ fun.\ de\ Ver.\ del\ modelo\ saturado\right].(I)$$

La forma general la prueba del coeficiente de verosimilitud, generalmente se encuentra expresada de la forma $0 < -2\ln(\hat{L}_0 / \hat{L}_1) < \infty$, donde $\hat{L}_0$ es el valor máximo de la función de verosimilitud para el modelo ajustado y $\hat{L}_1$ es el valor máximo de la función de verosimilitud para el modelo saturado, cumpliéndose que $\hat{L}_0 < \hat{L}_1$, esta estadística de prueba se distribuye aproximadamente como una χ^2_{m-k}, con $m-k$ grados de libertad, m es igual al número de parámetros del modelo saturado y k es igual al número de parámetros del modelo ajustado.

El indicador de ajuste se obtiene al sacar la diferencia del modelo ajustado con respecto al modelo saturado, se obtiene a partir de (I), esto es:

$$-2LnL_{dif} = (-2LnL_0) - (-2LnL_1)$$

Siendo el valor de $-2LnL_0$ la medida de la devianza del modelo uno y $-2LnL_1$ la medida de la devianza del modelo dos.

Como ejemplo se tienen dos modelos a comparar por medio de la prueba del coeficiente de verosimilitud, teniendo el modelo 0 (ajustado) como un caso especial del modelo 1 (saturado).

Bajo la hipótesis:

$$H_o : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_p = 0$$

vs.

$$H_a : \beta_i \neq 0 \qquad \text{Para alguna } i = 0, 1, 2, \ldots, p.$$

Si el valor de la estadística es mayor al cuantil de orden $(1-\alpha)$ de la distribución χ^2_{m-k}, entonces la hipótesis nula debe rechazarse con un nivel de significancia α.

Si el valor de la función de Verosimilitud del modelo ajustado es menor al valor de la función de verosimilitud del modelo saturado implica que la devianza será grande por lo tanto el modelo ajustado resulta inadecuado para la determinación de los datos.

La significancia estadística de una variable independiente se obtiene a partir de la comparación de cada variable X_i de los éxitos observados con los estimados usando la función de verosimilitud. A partir de la ecuación

$$\ln(l(\beta)) = \sum_{i=1}^{n} \{Y_i \ln(\pi_i) + (1 - Y_i) \ln(1 - \pi_i)\}^3$$

Y de (H) se obtiene que la devianza del modelo de los parámetros estimados es menos dos veces el logaritmo del cociente de verosimilitud obtenido de las dos expresiones anteriores, esto es:

---

[3] *Hosmer y Lemeshow, pag. 13-15.*

$$D = -2\sum_{i=1}^{n}\left\{Y_i \ln \frac{\hat{\pi}_i}{y_i} + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right\}$$

Donde $\hat{\pi}_i = \hat{\pi}(x_i)$.

La devianza en la regresión logística tiene el mismo objetivo que la suma de cuadrados de los residuales en la regresión lineal, es una medida de la falla de ajuste del modelo a mayor devianza más falla en el ajuste, bajo condiciones de regularidad. La distribución de la devianza se aproxima a una χ^2_{m-k} con $(m-k)$ grados de libertad, siempre y cuando las n_i sean grandes y bajo la hipótesis de que el modelo se ajusta a los datos correctamente.

A partir de la expresión del cociente de Verosimilitud se tiene:

$$\begin{aligned} -2\ln\frac{\hat{L}_0}{\hat{L}_1} &= -2\left(\ln \hat{L}_0 - \ln \hat{L}_1\right) \\ &= -2\left(\left(\ln \hat{L}_0 - \ln \hat{L}\right) - \ln \hat{L}_1 - \hat{L}\right) \\ G &= D_0 - D_1 \end{aligned}$$

1.4.2. PRUEBA DE LA ESTADÍSTICA DE WALD

Esta estadística de prueba se obtiene al comparar el estimador máximo verosimil del parámetro β_0, con el error estándar estimado, en base a las siguientes hipótesis:

$$H_0 : \beta_j = 0$$

vs

$$H_a : \beta_j \neq 0 \qquad \text{Para } j = 0,1,2,\ldots,p$$

Nótese que. $j = 0$ indicando que la prueba es sobre β_0.

La prueba de hipótesis anterior se puede realizar a través de la estadística de Wald cuya forma general es:

$$\frac{\hat{\beta}_j - \beta_j}{S(\hat{\beta}_j)}$$

Se distribuye como una Normal estándar, cuando le tamaño de muestra es suficientemente grande con $S(\hat{\beta}_j)$ el error estándar de $\hat{\beta}_j$, para este caso se tiene la siguiente estadística.

$$\frac{\hat{\beta}_j}{S(\hat{\beta}_j)}$$

La regla de decisión: se rechaza H_0 al nivel de signifiancia α si $W > Z_{1-\alpha}$.

Ya que la estadística de Wald tiene una distribución normal estándar, se tiene que:

$$\left(\frac{\hat{\beta}_j}{\hat{S}(\hat{\beta}_j)}\right)^2 \qquad (4)$$

La hipótesis nula se distribuye aproximadamente como una variable aleatoria χ_1^2 con un grado de libertad, siempre y cuando el tamaño de la muestra sea suficientemente grande.

La prueba de hipótesis es equivalente a:

$$H_0 : e^{\beta_j} = 1$$

vs

$$H_a : e^{\beta_j} \neq 1 \quad \text{Para } j = 0,1,\ldots,p.$$

---

[4] *Agresti Alan, pág. 185.*

1.5. AJUSTE DEL MODELO

Cuando se utiliza el método de regresión logística para la descripción e inferencia acerca de los efectos que se puedan tener en una variable de respuesta binaria, lo que interesa conocer del modelo, es si efectivamente resulta ser el más apropiado sin llevar a cabo una sobreestimación en la modelación de las variables explicativas, esto es, las medidas de ajuste indican que el modelo ajustado describe adecuadamente la relación entre la variable de respuesta y las variables independientes.

1.5.1. CÁLCULO DE RESIDUALES

El residual es una estadística que corresponde a la diferencia entre el valor observado y el valor estimado.

Los residuales pueden ser expresados como:

$$e_i = y_i - \hat{y}_i \qquad \text{para} \quad i = 1, 2, \ldots, n$$

Siendo:

e_i el residual, para el caso i

y_i es el valor observado para el caso i,

$\hat{y}_i$ el valor estimado para el caso i

En la regresión logística el resultado es engañoso, porque la precisión de $\hat{y}_{i.}$, está dada en función del número de observaciones n_i y de la probabilidad $\hat{p}_i$; mientras más grande sea n_i, será más preciso el valor de $\hat{y}_i$, recordando que se

tiene que $\hat{p}_i = \frac{\hat{y}_i}{n_i}$, de donde $\hat{y}_i = \hat{p}_i n_i$ y el error estándar de $\hat{y}_i$ está en función del error de $\hat{p}_i$.

Existen diferentes opciones para el cálculo de residuales en regresión logística, dependiendo de como sean calculados; los que se presentan a continuación son los residuales de Pearson o devianza.

a) El residual de Pearson se basa en la prueba de hipótesis:

$H_o = El\ modelo\ describe\ los\ datos\ apropiadamente.$

vs.

$H_a = El\ modelo\ es\ inadecuado.$

La χ^2 de Pearson para el modelo logístico se basa en los residuos siguientes:

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

con

$$r_i = \frac{y_i - n_i \hat{p}_i}{n_i \hat{p}_i (1 - \hat{p}_i)} \qquad \text{para} \quad i = 1, \ldots, n$$

donde:

y_i es el valor observado para el caso i,

$\hat{p}_i$ es la $P(Y = 1)$ del éxito para el modelo.

$n_i \hat{p}_i$ valor estimado del suceso.

La estadística de Pearson para la prueba del modelo corresponde a:

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

Por lo tanto el cuadrado del residual de Pearson es una componente de χ^2, cuando el tamaño de la muestra es grande la distribución de χ^2 se aproxima a una distribución normal.

b) El residual de devianza: se aproxima a una distribución normal, por lo que es más conveniente su uso en modelos de Regresión Logística para la comparación con respecto a otros modelos.

Por lo tanto el residual de la devianza para la observación i es:

$$d_i \times signo(y_i - n_i\hat{\pi}_i)$$

Donde

$$d_i = 2\left(y_i \, ln \frac{y_i}{n_i\hat{\pi}_i} + \left((n_i - y_i) ln \frac{n_i - y_i}{n_i - n_i\hat{\pi}_i} \right) \right)$$

Bajo H_o.

Para variables explicativas continuas la estadística de devianza no se distribuye asintóticamente χ^2, lo que generalmente se usa es la estadística G^2 basada en la diferencia de devianzas entre dos modelos que difieren por un número pequeño de parámetros.

c) Hosmer y Lemeshow.

En el caso de la regresión logística una idea intuitiva es calcular la probabilidad de ocurrencia del suceso. Si el ajuste es bueno, es de esperar que un valor alto de probabilidad se asocie con el suceso del éxito y viceversa, si el valor

de esa probabilidad calculada es bajo entonces se asocia con el fracaso del suceso.

Existen diferentes pruebas de bondad de ajuste, presentamos la que propusieron Hosmer y Lemeshow (*ver Hosmer y Lemeshow, 2ª., 2000, pag. 136-142*).

La bondad de ajuste *HL* se forma a partir de g grupos de individuos con base en n probabilidades estimadas, generalmente se trabaja con 10 déciles, es decir, se tiene:

$J = n$ con n columnas de los valores de las probabilidades estimadas.

Todos los grupos deben tener aproximadamente el mismo número de observaciones las cuales deben tener el mismo patrón de covariables aproximadamente igual al número total de observaciones.

Una fórmula para calcular HL es:

$$HL = \sum_{k=1}^{g} \frac{(O_k - E_k)^2}{n_k \pi(x_k)(1 - \pi(x_k))},$$

donde:

n número de individuos en el grupo g

$O_k = \sum_{j=1}^{n_k'} y_j$ número de éxitos observados en el grupo g

$E_k = n_k \hat{\pi}(x_k)$ número de éxitos estimados en el grupo g.

$\pi_k = \sum_{i=1}^{C_k} \frac{n_i \hat{p}_i}{n_k}$ probabilidad promedio estimada de observar un éxito en el grupo g.

La distribución de la estadística de prueba se aproxima a una distribución χ^2 con $g-2$ grados de libertad.

$$HL \sim \chi^2_{(g-2)}$$

Regla de decisión: se rechaza la hipótesis nula al nivel de significancia de α.

$H_o = El\ modelo\ describe\ los\ datos\ apropiadamente.$, sí;

$$HL > \chi^2_{(g-2)}.$$

Donde $w_{1-\alpha}$ es el cuantil de $1-\alpha$ de una χ^2 con $g-2$ grados de libertad.

1.5.2. MEDIDAS DE INFLUENCIA

Cuantifican la influencia que cada observación ejerce sobre la estimación del vector de parámetros o sobre las predicciones hechas a partir del mismo que forma en cuanto más grande son, mayor es la influencia ejercida en una observación de estimación del modelo, dentro de este tipo de medidas nos encontramos con: residuales de apalancamiento y residuales delta-beta o DFbetas.

a) *Residual Delta –Beta o DFBetas*

El residual Delta-Beta, evalúa el cambio obtenido, si se estimará un modelo por cada $i-$ésima observación eliminada; teniendo $j=1,2,\ldots,p$, para p parámetros estimados, $i=1,2,\ldots,p$ para n observaciones, es decir, para cada

parámetro del modelo, evalúa el cambio en el parámetro estimado si la i-ésima observación es eliminada, este cambio es dividido entre el error estándar. Por lo tanto la estadística Delta-Beta es la diferencia estandarizada en cada parámetro estimado cuando se elimina la i-ésima observación.

Entonces el residual de DFBeta se obtiene a partir del residual de Pearson:

$$DFBeta = \frac{e_i}{1-h_k} = \frac{y_i - n_i\hat{\pi}_i}{n_i\hat{\pi}_i(1-\pi_i)(1-h_k)}$$

Donde el e_i es el residual de Pearson obtenido al ajustar el modelo, eliminando la observación i-ésima de la muestra con h_i el elemento en la diagonal de la matriz sombrero H para la regresión logística.

1.6. INTERPRETACIÓN DE LOS COEFICIENTES

Después del ajuste del modelo es necesario conocer lo que representa la significancia de los coeficientes estimados, sobre todo enfocarse a la interpretación de los valores sobre el tipo de estudio que se planteó.

El término Momios es asociado a un suceso que se define como la razón entre la probabilidad de que dicho suceso ocurra y la probabilidad de que no ocurra; es decir, un número que expresa cuanto más probable es que ocurra frente a que no ocurra el hecho en cuestión. Así los coeficientes estimados asociados a las variables independientes representan la pendiente de una función de la variable dependiente por unidad de cambio en la variable independiente.

Se sabe que el modelo de regresión logística es lineal en los parámetros por medio de la transformación *logit,* esta transformación enlaza la variable independiente con el predictor lineal.

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

El coeficiente en el modelo de regresión lineal representa la pendiente β_1, siendo igual a la diferencia entre el valor de la variable respuesta en $x+1$ y el valor de la variable respuesta en x para algún valor de X.

En el modelo de regresión logística el coeficiente β_1, muestra el cambio en el logaritmo natural del momio de presentar una característica, al cambiar una unidad en la variable independiente.

$$\beta_1 = \beta_0 + \beta_1 x + \beta_1 - (\beta_0 + \beta_1 x) = g(x+1) - g(x)$$

A continuación se determinará la interpretación de los coeficientes del modelo de regresión logística para cada una de las posibilidades que se tenga para la *variable independiente, dicotómica y continua.*

VARIABLE INDEPENDIENTE DICOTOMICA

Se va a iniciar con la interpretación de los coeficientes del modelo de regresión logística para las variables independientes dicotómicas.

Sea X una variable explicativa , codificada como 0 y 1.

Y una variable de respuesta Y de un suceso con sus respectivas probabilidades de que ocurra éxito o fracaso.

$\pi(x)$, $(1-\pi(x))$.

A partir del modelo:

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

Los valores se muestran en la siguiente tabla.

	Variable Independiente		
		x=1	x=0
Variable de Respuesta	y=1	$\pi(1)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0)=\frac{e^{\beta_0}}{1+e^{\beta_0}}$
	y=0	$1-\pi(1)=\frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\pi(0)=\frac{1}{1+e^{\beta_0}}$

Tabla 2. Valores del modelo de Regresión Logística, con la variable independiente dicotómica

El momio que representa la categoría entre los individuos con $x=1$ está definido por:

$$\frac{\pi(1)}{1-\pi(1)}$$

Ocurriendo lo mismo para $x=0$, por lo tanto el logaritmo corresponde a:

$$ln\left(\frac{\pi(1)}{1-\pi(1)}\right)$$

Si el coeficiente de momios se denota por RM, se tiene que en el cociente de momios para $x=1$ y $x=0$, es:

$$RM=\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}$$

Aplicando el logaritmo natural:

$$\ln(RM) = \ln\left(\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}\right)$$

Al sustituir los valores de $\pi(x)$, con respecto a lo obtenido en la tabla 2, se tiene que:

$$RM = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Aplicando logaritmo nos queda:

$$\ln(RM) = \ln(e^{\beta_1}) = \beta_1$$

Se obtuvo que la relación que existe entre el cociente de momios y el coeficiente de regresión es la exponencial del coeficiente. La interpretación del cociente de momios está basada en el hecho de que se tiene una aproximación a la variable riesgo relativo el cual nos indica cuanto más probable es que ocurra el riesgo a que no ocurra para las observaciones con X=0 en lugar de las observaciones de X=1.

Se supone que $\pi(1)$ denota el riesgo de que se produzca el suceso y $1-\pi(1)$ de que no ocurra.

$$RR = \frac{\pi(1)}{1-\pi(1)}$$

Teniendo que:

$$Riesgo\,\mathrm{Re}lativo = 1$$

Si el riesgo relativo es igual a 1 indica que la proporción de individuos se mantiene constante en los diferentes niveles de la variable antecedente (x_i) por lo que no existe relación entre las dos variables. Los cocientes mayores a 1 indican

el número de veces que resultó ser mayor la incidencia del éxito, con respecto al fracaso, mientras que los menores a uno indican que la incidencia es menor entre los individuos que se encuentran en el éxito del suceso, tratándose de un factor de protección.

Ya que el cociente de momios tiene una interpretación independiente resulta ser importante obtener el intervalo de confianza para cada parámetro.

El intervalo de confianza con $100(1-\alpha)\%$, para el coeficiente β_1, está dado por:

$$\hat{\beta} - Z_{1-\alpha_2} s.e.(\hat{\beta}) < \beta_1 < \hat{\beta} + Z_{1-\alpha_2} s.e.(\hat{\beta})$$

Siendo $se(\hat{\beta})$ el error estándar asociado a la estimación $\hat{\beta}$.

El intervalo de confianza correspondiente para el cociente de momios se obtiene a partir de:

$$P\left[\exp\left[(\hat{\beta}_1) - Z_{1-\alpha_2} s.e.(\hat{\beta})\right] < \exp((\beta_1) < \exp(\hat{\beta}_1) + Z_{1-\alpha_2} s.e.(\hat{\beta})\ ] = 1-\alpha\right.$$

1.6.2. *VARIABLE INDEPENDIENTE CONTINUA*

La interpretación de los coeficientes estimados para una variable independiente continua de un modelo de regresión logística es el siguiente:

Bajo la suposición de que el logito es lineal en la variable independiente:

$$logit(\pi_i) = \beta_0 + \beta_1 X$$

El coeficiente de la pendiente β_1 da el cambio en el logito para un cambio en una unidad en X, al incrementarse c unidades en X, queda:

$$\beta_1 = g(x+c) - g(x) = c \cdot \beta_1, \qquad \text{para el valor de } x$$

La asociación del cociente de momios se obtiene por la exponencial del logit, esto nos queda:

$$RM(c) = RM(x+c, X) = \exp(c \cdot \beta_1)$$

La interpretación para un coeficiente estimado de una variable explicativa continua cuando se da un incremento de c unidades en la variable independiente, es $\exp(c\hat{\beta})$ veces más probable de ocurrir la variable respuesta

El intervalo de confianza para el cociente de momios queda:

$$\left[\exp\left[(\hat{\beta}_1) - Z_{1-\alpha_2} s.e.(\hat{\beta})\right] < \exp((\beta_1) < \exp(\hat{\beta}_1) + Z_{1-\alpha_2} s.e.(\hat{\beta})\right]$$

1.6. Modelo de Regresión Logística Politómica

El modelo logit politómico es una generalización del modelo logit binario, el modelo de regresión logístico, multinomial con una variable de respuesta, con $c \geq 2$ categorías, existen diferentes métodos para la obtención de los logit multinomial.

El modelo logit multinomial es usado para comparar 2 modelos ajustados respecto a un tercero, considerando una variable respuesta tricotómica, la base principal para la construcción del modelo se le conoce como *riesgo* o *"baseline"*.

Sea Y una variable respuesta con J categorías, el modelo logit para una variable de respuesta nominal simultáneamente describe el momios para todos los $\binom{J}{2}$ pares de categorías, teniendo como elección $J-1$.

Logit categoría de riesgo.

Se tiene para una variable de respuesta multinomial con probabilidades de respuesta $(\pi_1,\ldots,\pi_j)$ con j categorías, denotando como $\pi_j(x)=P(Y=j|x)$, x fijo de las variables explicativas. Sea y_i la variable de respuesta con categorías codificadas: como $1,2,\ldots,j$, sea el logit de riesgo la comparación de la $j-ésima$ categoría con la primera, es decir:

$$BL_j=\ln\left[\frac{(P_j(y=j))}{(P_k(y=k))}\right]=\log\left(\frac{\pi_j}{\pi_k}\right)$$

o

$$\ln\left[\frac{\left(\pi_j(\pi_j+\pi_k)\right)}{\left(\pi_k(\pi_j+\pi_k)\right)}\right]=\log\left(\frac{\pi_j}{\pi_k}\right) \qquad (1)$$

Donde π_j y π_k indican las probabilidades para la $j-ésima$ y $k-ésima$ categoría, la "*baseline*" es arbitraria (π_k).

Para el caso de una variable independiente X con r categorías, se tiene para "*logit-baseline*"

$$BL_{rj}=\ln\left[\frac{(P_j(y=j|x=r))}{(P_k(y=k|x=r))}\right]=\log\left(\frac{\pi_{rj}}{\pi_{rk}}\right)$$

Se obtuvo un modelo saturado, la estimación de la ecuación se obtiene a partir de:

$$ln\left(\frac{F_{ij}}{F_{kj}}\right)=ln\left(\frac{f_{ij}}{f_{kj}}\right)$$

Siendo las observaciones y frecuencias esperadas en el $i-ésimo$ renglón y $j-ésima$ columna para la tabla de clasificación $X\ y\ Y$.

Expresado en términos del modelo lineal generalizado queda:

$$BL_{ij}=\sum_{r=1}^{j} ln\left(\frac{F_{ij}}{F_{kj}}\right)\cdot R(x=r)$$

Siendo R la función indicadora, con $i=1$ como verdadero y cero en otro caso, con una variable codificada como "*dummy*" tomando la primera categoría como referencia, la ecuación queda escrita como:

$$BL_{ij}=\alpha_j+\sum_{j=1}^{j}\beta_{rj}\cdot R(x=r) \qquad x\succ 1$$

Donde:

α_j es el "*logit-baseline*" para $x=k$

β_{ij} es la diferencia entre "*logit-baseline*" de $x=k$ y $x=r$

Pueden ser estimados separadamente para todo i, j. Para los modelos no saturados las estimaciones simultáneas producen resultados diferentes.

El modelo logit para la categoría j de la variable respuesta de la ecuación (1), en lugar de la categoría k, dado que la observación cae en una de estas dos categorías básicas. Todo logit de este tipo puede ser derivado de un ajuste básico del tamaño $c-1$. Por ejemplo si:

$$L_j = log(\pi_j|\pi_c), \qquad j = 1,\ldots,c-1 \qquad \text{(a)}$$

Entonces:

$$\log(\pi_j|\pi_c) = L_j - L_k \qquad para\ 1 \le j \prec k \le c-1$$

Los logit pueden ser formados por categorías de grupos que son continuos en la escala ordinal.

Ejemplos:

a) La que ordena las categorías es "C*umulative*" logits:

$$L_j = log\left[\frac{(\pi_{j+1} + \ldots + \pi_c)}{(\pi_1 + \ldots + \pi_j)}\right], \quad j = 1,\ldots,c-1$$

b) El "*Ratio "continuation-ratio*" logits:

$$L_j = log\left[\frac{\pi_{j+1}}{(\pi_1 + \ldots + \pi_j)}\right], \quad j = 1,\ldots,c-1$$

c) En las "*Adjacent-categories*" logits:

$$L_j = log\left[\frac{\pi_{j+1}}{\pi_j}\right], \quad j = 1,\ldots,c-1$$

Cuando c=2, los primeros tres tipos de logit se simplifican al estándar logit que es $\left(\pi_2 / \pi_1\right)$, siendo equivalente a (1).

1.7. SELECCIÓN DEL MODELO POR MEDIO DEL PAQUETE ESTADÍSTICO SPSS

Existen diferentes algoritmos de selección automática de las variables explicativas, en el paquete estadístico SPSS encontramos dos: método "*Forward*" (adicionando variables) y "*Backward*"[1] (eliminando variables), dicho procedimiento se basa en el algoritmo de determinar que variables van a ser incluidas o excluidas del modelo ya que resulta conveniente eliminar las que no sean tan significativas para el modelo y se puedan ajustar un modelo más simple (o más parsimonioso) obteniendo así una mayor precisión de los estimadores.

En la regresión logística se tiene una distribución binomial, el nivel de significancia se establecerá por medio de la prueba del cociente de verosimilitud el cual determinará que variable resulta ser relevante estadísticamente para ingresarla en la obtención de resultados, y eliminar a la variable que produce el menor valor de la estadística del cociente de verosimilitud.

El método de selección automático de "*Forward*", sigue los siguientes pasos:

1. Inicia considerando a todas la p-variables potencialmente influyentes de la variable respuesta, comienza con un primer ajuste tomando en cuenta a la variable constante obteniendo el logaritmo natural de la función de verosimilitud (L_o), después se ajustan p-regresiones logísticas univariadas que incluyen a cada variable obteniendo a su vez i funciones de verosimilitud (L_i).

   Se obtiene la estadística del cociente de Verosimilitud

   $$G_i = -2(L_o - L_i)$$

   el valor de significancia estadística

   $$P[\chi_v^2 > G_i\ ] = p_i \qquad \text{con} \qquad \begin{matrix} v = 1, \textit{si es continua} \\ v = k-1 \textit{ si es politómica} \end{matrix}$$

[1] *Hosmer y Lemeshow, pag.106-110.*

Se elige la variable con menor valor de significancia

$$X_r \quad con \quad p_r = min[p_i]$$

Para determinar que variable resulta ser significativa se compara el valor p_r con respecto al valor de 0.05 establecido para p_R. Si se obtiene que $p_r < p_R$ se acepta la variable y se continua con el siguiente paso.

2. A partir del modelo ajustado que contiene X_{r1} y el logaritmo natural de la función de verosimilitud L_r, se determina si existe otra de las $p-1$ variables importantes, por lo tanto se ajustan $p-1$ regresiones logísticas que contienen a X_{r1} y X_j $con\ j = 1,\ldots n$, $j \neq r$.

   Sea L_r el logaritmo natural de la función de verosimilitud del modelo que contiene a X_{r1} y $X_j : G = -2(L_{r1} - L_{rj})$, el valor de la estadística del cociente de verosimilitud de cada modelo con valor de significancia p_j. Suponiendo que X_{r1} es la variable con el menor valor de significancia en este paso, es decir $p_{r2} = min[p_i]$, si este valor es menor que p_R, se pasa al siguiente paso o ahí se termina el algoritmo al no tener otra variable.

3. Se realiza el mismo procedimiento del punto 2, con la diferencia de que el modelo ajustado ahora va ser X_{r1} y X_{r2}, se realiza la revisión de las variables anteriormente introducidas y se continúa con la selección, siguiendo el procedimiento hasta llegar al paso (t).

4. En el paso (t) se localizan todas las variables que entraron al modelo, con valores de significancia.

El método de eliminación "*Backward*" es diferente al *"Forward"* ya que el primero inicia con todas la variables que se consideran potencialmente influyentes

con respecto a la variable de respuesta y repetidamente va eliminado los términos en lugar de incrementarlos, con el procedimiento de identificar a la variable que tiene el mayor valor de significancia de las pruebas de cociente de verosimilitud.

Para el caso cuando se tiene una variable cualitativa con más de 2 categorías se considera a la variable completa para la eliminación o inclusión del modelo propuesto.

Los procedimientos anteriores aparecen implementadas solamente para una variable respuesta binaria, implementadas en el paquete estadístico SPSS.

CAPÍTULO 2

APLICACIÓN DEL MODELO LOGÍSTICO PARA OBTENER EL MEJOR AJUSTE QUE REPRESENTE LA VARIABLE DEPENDIENTE A PARTIR DE LAS VARIABLES CATEGÓRICAS

CAPÍTULO 2

APLICACIÓN DEL MODELO LOGÍSTICO PARA OBTENER EL MEJOR AJUSTE QUE REPRESENTE LA VARIABLE DEPENDIENTE A PARTIR DE LAS VARIABLES CATEGÓRICAS

En este capítulo se presenta un análisis sobre las características o factores que intervienen en la elección de carrera ya que es considerado como un fenómeno multicausal en el que se presume intervienen determinantes sociodemográficas, familiares y subjetivas (proyecto de vida, valores y habilidades del individuo), se va a realizar el análisis bajo una perspectiva de género y así detectar la posible relación existente entre el fenómeno de elección de carrera y algunas características psicosociales del individuo al igual que se expondrán los pasos requeridos para dicha investigación, el proceso de análisis de regresión logística es una herramienta que ayuda a modelar las variables explicativas con respecto a la variable respuesta (elección de una determinada carrera de las tres posibilidades que se van a manejar), teniendo una influencia tanto positiva como negativa.

En primer lugar se va a establecer la descripción de los datos con los que se trabajó, después se obtendrán los resultados con la ayuda del paquete estadístico SPSS, con base en la metodología que se vio en el *Capítulo 1* para identificar cuales son las variables que resultan ser significativas para el modelo establecido al igual que el cálculo de la razón de momios e intervalos de confianza, teniendo como finalidad evaluar el grado de asociación que existe entre las variables explicativas y la variable respuesta. Se va a presentar en forma individual para cada tipo de carrera y posteriormente se dará en forma conjunta y

así observar las similitudes y diferencias de los distintos modelos propuestos con la intención de establecer conclusiones coherentes y apegados a la realidad. Es importante notar que el cuestionario se aplicó a los alumnos de dos tipos de universidades, cada una cuenta con poblaciones estudiantiles diferentes tanto socialmente como económicamente por esa razón resulta importante hacer un anexo con la variable respuesta tipo de universidad al que acude el estudiante y establecer una relación con la decisión de elegir una de las tres carrera a considerar, obteniendo así un resultado global sobre los factores que influyen en dicho riesgo.

El siguiente punto dará a conocer la información utilizada para llevar a cabo dicho proceso.

2.1. Información Utilizada

Los datos utilizados fueron obtenidos a partir de una encuesta realizada en agosto de 1996 en algunos estudiantes que ingresaron a la educación nivel superior de dos universidades, estas son: Universidad Iberoamericana (UIA) y la Universidad Autónoma Metropolitana (UAM), dicha base de datos fue proporcionada por la Maestra Victoria Caraveo Enríquez[6], con fines estrictamente académicos.

El cuestionario se aplicó tanto a hombres como mujeres, las preguntas realizadas tenían como objetivo principal representar los posibles factores que pueden intervenir en la toma de elección de carrera, como son: los aspectos económicos, culturales, familiares, sociales y al igual que los proyectos de vida que desean realizar en un futuro no muy lejano, es decir cuando terminen de estudiar.

La muestra consistió de 344 estudiantes, hombres y mujeres que ingresaron, en agosto de 1996 a dos Instituciones, siendo una del sector público y

[6] *Profesora en la facultad de medicina de la UNAM por asignatura.*

la otra del sector privado de la Ciudad de México, cada alumno contestó un cuestionario diseñado especialmente para determinar los posibles factores de la elección de carrera, de la misma manera también se contesto el Inventario de IMAFE[7] diseñado por Dra. Lara Cantú (1993)[(2)], el cual va enfocado a determinar el perfil de masculinidad y feminidad del individuo.

Las instituciones tuvieron carreras equiparables, como se ve a continuación:

UNIVERSIDAD	CARRERA	TAMAÑO DE MUESTRA
UIA	Nutrición y ciencia de los Alimentos.	35
UAM	Nutrición	35
UIA	Electrónica y Telecomunicaciones.	41
UAM	Electrónica	50
UIA	Administración de empresas	106
UAM	Administración	77

El cuestionario se aplicó en un periodo de 15 días al ingresar a las actividades académicas de la nueva escuela para llevar a cabo sus estudios a nivel licenciatura.

2.2. VARIABLE RESPUESTA (VARIABLE DEPENDIENTE)

Se ha manejado la elección de carrera como la ocurrencia de un determinado suceso, esto es;

$$Y_i = \begin{cases} 1 & Se\ elige\ la\ carrera \quad i \\ 0 & e.o.c. \end{cases}$$

La aplicación del análisis de regresión logística para cada uno de los modelos se llevo a cabo en forma individual, 4 de los modelos son con respuesta binaria y el quinto es de una variable respuesta tricotómica.

---

[7] *Inventario de Masculinidad - Feminidad*

Primero se verá el análisis realizado para la carrera de la licenciatura de nutrición con la finalidad de identificar los posibles factores que pudieron haber intervenido en la elección de la misma, en segundo lugar se aplicará la misma metodología para la carrera de ingeniería en electrónica, continúa la licenciatura en administración y posteriormente se verá para el tipo de universidad y por último se considera la variable respuesta tricotómica (con 3 categorías).

En el siguiente cuadro se determina la variable respuesta para cada uno de los 5 modelos.

VARIABLE RESPUESTA TIPO DE CARRERA	CATEGORÍAS	CODIFICACIÓN
1.-Nutrición	Nut y ciencias de los alimentos, Nut.	1
	Electrónica y Tele, Electrónica.	0
	Adm. de empresas, Administración	0
2.-Ingeniería	Nut y ciencias de los alimentos, Nut.	0
	Electrónica y Tele.., electrónica.	1
	Adm. de empresas, Administración	0
3.-Administración	Nut. y ciencias de los alimentos, Nut.	0
	Electrónica y Tele., electrónica.	0
	Adm. de empresas, Administración	1
4.-Universidad	Universidad Privada (UIA)	0
	Universidad Pública (UAM)	1

Y por último vamos a ver la variable respuesta para el caso tricotómico.

VARIABLE RESPUESTA	CATEGORÍAS	CODIFICACIÓN
5.-Carrera	Nutrición	2
	Ingeniería	1
	Administración	0

***Nota**: 1 significa el éxito del suceso, es decir, la preferencia por esa categoría y 0 significaría la desaprobación de la categoría que esté representando o dicho de otra manera el fracaso del suceso.*

2.3. *Variables Explicativas*

De todas las preguntas que contenía el cuestionario, por ejemplo: estado civil del estudiante, si estaba trabajando cuando realizaba sus estudios, tipo de vivienda en la que habita (departamento, casa, etc.), lugar donde le gustaría trabajar, etc., solamente 14 factores reflejaron las características de mayor interés para el investigador, por esa razón se trabajó en la composición e integración de cada una de las variables seleccionadas para contar con una mayor facilidad de manejo e interpretación de los resultados.

Para empezar con cualquier tipo de análisis, se debe tener un cuidado especial sobre las variables que se vayan a utilizar, ya que se debe especificar el tipo de variable, ya sea categórica o continua. Una variable continua es aquella que puede contener cualquier valor que sea posible dentro de un rango de valores con números reales positivos o negativos. Las variables categóricas pueden ser de diversos modos: dicotómicas, por ejemplo: el sexo (hombre o mujer), politómicas no ordenas, politómicas ordenadas.

Las variables categóricas que se tienen están constituidas entre 2 y 4 valores, por lo tanto es necesario establecer la categoría de referencia o de comparación, técnicamente cualquier valor puede serlo, pero debe tomarse aquella categoría que tiene sentido desde el punto de vista del problema que se trabaje; ya que puede cambiar la interpretación de cada coeficiente (β_i), en general, se cuantifica el efecto de cada categoría que tiene la variable con respecto al valor de referencia que se haya establecido, en este caso se eligió como categoría de referencia el que contiene el mayor número de observaciones para obtener una buena estimación y un intervalo de confianza relativamente angosto y no cometer errores en la elaboración del modelo.

En las siguientes tablas se establece el nombre de cada variable (etiqueta), a que se refiere cada una de ellas, la especificación de las categorías y por último el porcentaje muestral.

Las variables continuas son:

Etiqueta	Variable Explicativa (Descripción)	Valores Observados
Promedio	Promedio que obtuvieron al salir de preparatoria los estudiantes	El promedio varia entre el rango de 6.4 - 10
Edad	Edad que tienen al entrar a la Universidad	La edad de los estudiantes se encuentra en un rango de 17 - 53 años

Tabla A. Variables explicativas continúas.

A continuación se dan a conocer las variables categóricas con sus respectivos valores, debe mencionarse que la categoría de referencia es la que está codificada como "0".

Variables Explicativas (Descripción)	Etiqueta	Categorías	Codificación	% DE Muestra
Sexo de los estudiantes	Sexo	Masculino	0	55.2
		Femenino	1	44.8
Universidades que imparten las carreras de Nut, Ing. y Adm.	*Universidad*	UIA	0	52.9
		UAM	1	47.1
ASPECTO ECONOMICO				
Número de autos que tiene la familia del estudiante.	*Auto*	Dos o más autos	0	56.1
		A lo más un auto	1	21.8
		Ningún auto	2	20.1
ASPECTO SOCIOCULTURAL				
El estudiante ya había considerado escogerla	*Quiso la carrera*	No	0	65.1
		Si	1	33.7
Se eligió la carrera por las condiciones del país.	*País*	En desacuerdo	0	91.3
		Deacuerdo	1	7.8
ASPECTO FAMILIAR				
Ocupación a la que se dedica la mamá.	*Ocupación de la madre*	Trabajar	0	50.6
		Hogar	1	31.7

Variables Explicativas (Descripción)	Etiqueta	Categorías	Codificación	% DE Muestra
Comunicación que tiene el estudiante con su padre	*Comunicación padre*	Siempre	0	73
		Nunca	1	20.9
Comunicación que tiene el estudiante con su madre	*Comunicación madre*	Siempre	0	89
		Nunca	1	8.4
Grado de estudio que tiene la madre del estudiante	*Esc.madre*	A lo más sec.	0	35.2
		Prep. o carrera Tec.	1	32
		Licenciatura	2	25.3
		Maestría o Doct.	3	4.9
Grado de estudio que tiene el padre del estudiante	*Esc.padre*	Licenciatura	0	40.7
		A lo más sec.	1	26.5
		Prep.o carrera Tec.	2	16.9
		Maestría o Doct.	3	13.1
PROYECTOS DE VIDA				
Actividad que van a desempeñar al terminar la licenciatura	*Al terminar la carrera*	Seguir estudiando	0	64.2
		Trabajar	1	32
		Otra Actividad	2	3.8
ÍNDICE DE MASCULINIDAD Y FEMINIDAD				
Mide las características femeninas y masculinas de la personalidad de cada individuo.	*IMAFE*	Androgina	0	26.7
		Indiferente	1	23.3
		Femenino	2	22.7
		Masculino	3	18

Tabla B. Distribución de las variables explicativas categorías según tamaño de muestra y clasificación por aspectos sociales, económicos, etc.

Las codificaciones presentadas en la *Tabla B*, se obtuvieron a partir de un proceso de exploración de variables así como de la distribución de la muestra, ya que de esta forma pudieron ser modeladas para la obtención de resultados satisfactorios.

2.4. Correlación de las Variables Explicativas

Antes de comenzar con el análisis es aconsejable revisar las correlaciones de Pearson entre las variables explicativas porque podría ser un problema para la obtención del modelo de regresión logística ya que estaría desprovisto de sentido en algunos valores de los coeficientes para su interpretación. Se tiene como

finalidad identificar variables altamente correlacionadas entre sí para aplicar el criterio del investigador sobre que variables resultan ser más relevantes para el estudio.

Hacemos notar que la correlación lineal entre pares de variables es solo un tipo de colinealidad. Puede existir colinealidad entre tres o más variables y que no se detecta a través de correlaciones lineales por pares.

Correlations

Correlaciones de Pearson	Edad	Promedio	Univer.	Termino carrera	Quiso carrera	País	Autos	Educ. padre	Educ. madre	Sexo	Ocupación madre	IMAFE	Com. padre	Com. madre
Edad	1.000	-.274	.503	.062	.280	.012	.274	.023	-.396	-.188	.320	.040	.054	.047
Promedio	-.274	1.000	-.284	-.038	-.159	-.065	-.146	.100	.225	.368	-.194	-.013	.045	.096
Universidad	.503	-.284	1.000	.167	.185	-.019	.650	.069	-.552	-.165	.289	.081	-.162	-.090
Terminar	.062	-.038	.167	1.000	-.004	-.116	.240	.045	-.185	.010	.085	.057	-.002	-.017
Quiso esta car	.280	-.159	.185	-.004	1.000	-.013	.077	.077	-.133	-.264	.083	-.051	.113	.006
Pais	.012	-.065	-.019	-.116	-.013	1.000	-.046	-.050	-.033	-.106	.011	.110	-.038	-.165
Autos	.274	-.146	.650	.240	.077	-.046	1.000	.093	-.513	-.111	.214	.098	-.212	-.145
Educ.padre	.023	.100	.069	.045	.077	-.050	.093	1.000	-.001	-.033	-.015	.109	-.023	-.052
Educ.madre	-.396	.225	-.552	-.185	-.133	-.033	-.513	-.001	1.000	.179	-.475	.002	.019	.047
Sexo	-.188	.368	-.165	.010	-.264	-.106	-.111	-.033	.179	1.000	-.191	-.091	.033	.145
Ocupación ma	.320	-.194	.289	.085	.083	.011	.214	-.015	-.475	-.191	1.000	.007	.133	.049
IMAFE	.040	-.013	.081	.057	-.051	.110	.098	.109	.002	-.091	.007	1.000	-.001	.063
Com.padre	.054	.045	-.162	-.002	.113	-.038	-.212	-.023	.019	.033	.133	-.001	1.000	.452
Com.madre	.047	.096	-.090	-.017	.006	-.165	-.145	-.052	.047	.145	.049	.063	.452	1.000

Tabla C. Correlación lineal entre las variables explicativas propuestas en la tabla B.

Se observa que existe alta correlación lineal entre algunas variables, como es el caso de la variable *universidad* que está altamente correlacionada con; *edad, número de autos y educación que tiene la madre* del estudiante; ocurriendo lo mismo con la variable *número de automóviles* y la variable *educación que tiene la madre* del estudiante que a su vez esta se encuentra relacionada con la variable *ocupación de la madre* y por último también se encuentra una correlación alta con respecto a la *comprensión que tiene la madre* y *comprensión que tiene el padre* hacia los estudiantes (hijos).

Después de conocer que variables se encuentran altamente correlacionadas es necesario establecer cual de ellas resulta relevante para el estudio, en primer lugar tenemos que era importante conocer la relevancia que tienen las variables educación de la madre para la elección de carrera, como esta se encuentra relacionada con otras 3 variables fue necesario construir una nueva

variable llamada escolaridad de los padres obtenida a través de la variable escolaridad de la madre y escolaridad del padre.

Correlations

Correlación de Pearson	Universidad	Autos	Ocupación madre	Educ. ambos padres
Universidad	1.000	.650	.289	.033
Autos	.650	1.000	.214	.077
Ocup. de la madre	.289	.214	1.000	-.056
Educ.ambos padres	.033	.077	-.056	1.000

Tabla D. Correlación lineal entre las variables explicativas Universidad, auto, ocupación madre y Escolaridad ambos padres.

Para la construcción de la nueva variable escolaridad de ambos padres no se tuvo problemas de correlación por lo tanto resulta más propicio la utilización de está para la obtención de modelos que se encuentren correctamente estimados.

Con respecto a las otras correlaciones solamente se va a considerar la variable comprensión que tiene la madre y universidad, en el siguiente punto se establecerá cuales son las variables a considerar para el análisis según el tipo de variable respuesta.

2.5. Análisis de la Licenciatura de Nutrición

Después de haber obtenido las correlaciones de las variables explicativas, se construirá el modelo de regresión logística múltiple a partir de las siguientes variables explicativas:

Etiqueta	Categoría		Elección de carrera		total
			Otro	Nutrición	
Promedio	*Continua*	Media	8.13	8.67	
		Tamaño de muestra	243	59	302

Etiqueta	Categorías		Elección de carrera		% DE
			Otro	Nutrición	Muestra
Sexo	*Masculino*	Observaciones	163	6	169
		%	96.40%	3.60%	100.00%
	Femenino	Observaciones	80	53	133
		%	60.20%	39.80%	100.00%
Universidad	*UIA*	Observaciones	133	32	165
		%	80.60%	19.40%	100.00%
	UAM	Observaciones	110	27	137
		%	80.30%	19.70%	100.00%
Quiso la carrera	*No*	Observaciones	149	54	203
		%	73.40%	26.60%	100.00%
	Si	Observaciones	94	5	99
		%	94.90%	5.10%	100.00%
País	*En desacuerdo*	Observaciones	26		26
		%	100.00%		100.00%
	Deacuerdo	Observaciones	217	59	276
		%	78.60%	21.40%	100.00%
Ocupación de la madre	*Trabajar*	Observaciones	109	40	149
		%	73.20%	26.80%	100.00%
	Hogar	Observaciones	90	10	100
		%	90.00%	10.00%	100.00%
Com.madre	*Siempre*	Observaciones	222	54	276
		%	80.40%	19.60%	100.00%
	Nunca	Observaciones	21	4	25
		%	84.00%	16.00%	100.00%
Al terminar la carrera	*Seguir estudiando*	Observaciones	152	39	191
		%	79.60%	20.40%	100.00%
	Trabajar	Observaciones	81	18	99
		%	81.80%	18.20%	100.00%
	Otra Actividad	Observaciones	10	2	12
		%	83.30%	16.70%	100.00%
IMAFE	*Androgina*	Observaciones	68	14	82
		%	82.90%	17.10%	100.00%
	Indiferente	Observaciones	60	13	73
		%	82.20%	17.80%	100.00%
	Femenino	Observaciones	44	23	67
		%	65.70%	34.30%	100.00%
	Masculino	Observaciones	48	6	54
		%	88.90%	11.10%	100.00%
Escolaridad ambos padres	*Al menos un padre con licenciatura*	Observaciones	100	27	127
		%	78.70%	21.30%	100.00%
	Ambos a lo más secundaria	Observaciones	65	8	73
		%	89.00%	11.00%	100.00%
	Al menos un padre con preparatoria	Observaciones	43	7	50
		%	86.00%	14.00%	100.00%
	Al menos un padre con Maestría o Doctorado.	Observaciones	32	15	47
		%	68.10%	31.90%	100.00%

Tabla E. Porcentajes y promedios de la distribución de los valores de las variables explicativas a utilizar en la construcción de los modelos.

Uno de los objetivos del análisis de regresión logística es cuantificar la relación entre la variable respuesta y las variables explicativas para establecer el grado de confianza que afirma la cuantificación realizada por el método y sobre todo que se ajuste a la realidad observada.

Se comenzará el análisis a partir del modelo propuesto inicialmente, el cual queda como:

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{10} x_{10}.$$

Como variable respuesta se tiene:

$$Y = \begin{cases} 1 & \textit{Elección de la carrera de nutrición} \\ 0 & \textit{En Otro caso.} \end{cases}$$

Para la comprobación de significancia de los parámetros estimados se usa la estadística de Wald (*tabla 1, Anexo I*), las variables estadísticamente significativas fueron *promedio del estudiante, siempre había considerado estudiar la carrera de nutrición, al terminar la carrera se va a dedicar a trabajar, sexo del estudiante, universidad a la que acuden y por último escolaridad de ambos padres con a lo más secundaria.*

El procedimiento a seguir es:

1. Se seleccionará el modelo de regresión logística para cada modelo propuesto usando el método de selección backward o forward implementados en el paquete estadístico SPSS:
2. Para cada variable respuesta se utilizarán las diez variables explicativas que se presentaron en la *tabla E*.
3. Los resultados que se obtendrán serán con base a lo que se vio en el capítulo 1 y posteriormente se escogerá el modelo que resulte ser el más prometedor para la investigación.

2.5.1. MODELO PROPUESTO POR UN MÉTODO DE SELECCIÓN

El siguiente cuadro muestra las variables explicativas que resultaron ser significativas para el modelo por el método de selección backward, con su respectiva etiqueta y los valores de los coeficientes, el error estándar, la estadística de Wald y el p-value a considerar para rechazar o aceptar la prueba de hipótesis, como a continuación se muestra:

Factor (i)	Etiqueta	Variable	Coeficiente de regresión	Error estándar	$\left(\frac{\beta}{Se(\beta)}\right)^2$	
Promedio	*Promedio*	**Continua**	*0.62*	*0.267*	*5.413*	$P < 0.03$
Sexo del estudiante	*Sexo(Fem)*	Hombre=0 * **Mujer=1**	*2.641*	*0.496*	*28.329*	$P < 0.001$
Quiso esta carrera	*Qestcarr(Si)*	*No=0* * ***Si=1***	*-1.503*	*0.541*	*7.73*	$P < 0.01$
Escolaridad ambos padres	*Edpadres(Sec)*	Lic.=0 * ***Sec=1*** Prep=2 ** MoD=3 **	*-1.364*	*0.626*	*4.756*	$P < 0.03$
Universidad	*Univer(UAM)*	UIA=0 * ***UAM=1***	1.878	0.523	12.893	$P < 0.001$

Cuadro 1. Resultado del ajuste logístico de la variable elección de la carrera de Nutrición en función de cinco variables explicativas, por el método de selección backward.

El *cuadro 1*, presenta las 5 variables explicativas que resultaron ser estadísticamente significativas para el modelo, teniendo como categoría de referencia el valor con un asterisco (*) y las categorías que no resultaron ser significativas para el modelo con 2 asteriscos (**).

El modelo queda determinado como

$$\ln\left(\frac{\pi(Y=1)}{1-\pi(Y=1)}\right) = -8.756 + 0.62\ Promedio + 1.878\ Universidad\ (UAM) + 2.641\ Sexo(Fem)$$
$$- 1.503\ Quiso\ esta\ carrera(Si) - 1.364\ Esc.ambos\ padres\ (Sec)$$

2.5.2. *PRUEBA DE SIGNIFICANCIA DEL MODELO*

PORCENTAJE DE CLASIFICACIÓN

La clasificación correcta del modelo propuesto determina el número de observaciones de la muestra que el modelo clasifica correctamente. Se da el porcentaje de las observaciones consideradas correctamente entre el total de los datos clasificados.

Una aproximación sería la construcción del modelo logístico de los datos calculándose el *logit* estimado, para la construcción del logit se calcularían las probabilidades estimadas para cada observación, se parte de la variable Y, con 2 posibles valores $(0 \acute{o} 1)$, se juzga un modelo con buen poder predictivo, si el número de observaciones de la muestra que el modelo clasifica correctamente es alto. Se determina un valor fijo de P_0 entre 0 y 1, si se obtiene la probabilidad estimada de una observación mayor que 0.5 se asignará al grupo 1 $(Y = 1)$, y si es menor que 0.5 se va a clasificar al grupo 0 $(Y = 0)$, por lo tanto se va a determinar que proporción de los datos es clasificado correctamente[8].

Como consecuencia de tal mecanismo se produce *falsos positivos* (predice el éxito, pero realmente no lo obtuvieron) y *falsos negativos* (predice como no éxito cuando realmente si lo obtuvieron), además de aciertos en uno u otro sentido, esto se aprecia mejor en el siguiente cuadro.

		Realidad		
Predicción		*Éxito (Éxito=1)*	*No Éxito (Éxito=0)*	*Total*
Éxito	$P \succ P_0$	a	b	a+b
No éxito	$P \leq P_0$	c	d	c+d
TOTAL		a+c	b+d	n

Cuadro 2. Tabla de configuración según la clasificación de los datos de una muestra con respecto a los resultados reales y vaticinados, para cada P_0.

[8] Regression Analysis by example, Chatterjee Hadi Price, 3ª, 2000, pg. 328-330.

Por lo tanto los indicadores clásicos: la *Sensibilidad* y la *Especificidad*, son:

$$A = \frac{a}{a+c} \qquad B = \frac{d}{b+d}$$

Si se obtiene un valor de sensibilidad igual al valor de especificidad nos indica que el modelo no distingue los éxitos de los fracasos, con valores altos en la sensibilidad y especificidad indica una buena clasificación y por ende en el modelo logístico se pueden extraer buenos resultados y una proporción baja de clasificación indicaría que fue un mal ajuste de las variables.

Para el modelo propuesto en el cuadro 1 se tiene la siguiente clasificación:

Classification Table[a]

			Predicted		
			los que escogieron la carrera de Nutrición		Percentage Correct
	Observed		OTRO	NUT.	
Step 1	los que escogieron la carrera de Nutrición	OTRO	230	10	95.8
		NUT.	31	26	45.6
	Overall Percentage		.		86.2

a. The cut value is .500

Cuadro 3. Porcentaje de clasificación para el modelo propuesto.

Del *cuadro 3* se obtiene que la *especificidad* es igual a $.95 = (230/240)$, esto indica que los estudiantes que eligieron las carreras de administración e ingeniería son clasificados correctamente en la elección de las carreras de administración e ingeniería. La *sensibilidad* es de $.46 = (26/57)$, estudiantes que se encuentran en la carrera de nutrición que son clasificados correctamente en la licenciatura de nutrición bajo el modelo.

Otro indicador para la evaluación del modelo es el *valor de pronóstico positivo* el cual se basa en el cálculo de los valores estimados, ya que divide el número estimado correctamente de una de las categorías de la variable respuesta

entre el total de casos estimados correctamente, como se verá en el siguiente cuadro.

Pronóstico	Carrera		Total	Valor pronóstico
	Nut.	Adm. E Ing.		
Nut.	26	10	36	(26/36)=.72
Adm. E Ing.	31	230	261	(230/261)=.88
Total	57	240	297	

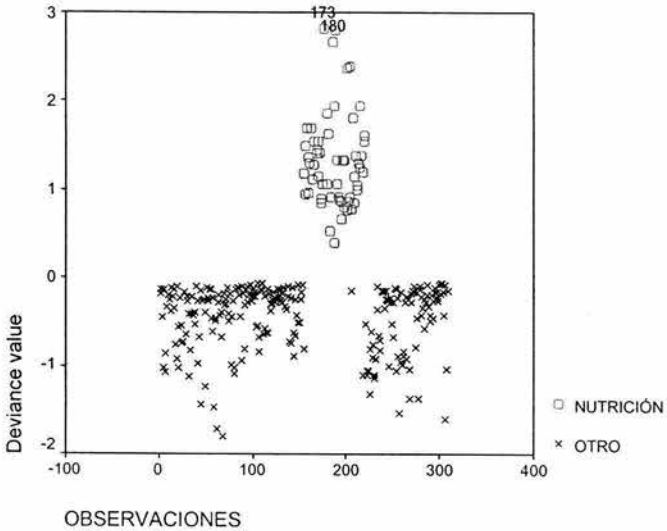
Cuadro 4. Porcentaje de clasificación correcta de los valores estimados.

Por lo tanto el valor de pronóstico positivo es = .72 y *la prevalencia* de la muestra es de $(57/297)=.19$, lo cual nos indica que fue mejor en más de 3 veces el pronóstico de elección de carrera mediante el modelo propuesto que sin él, este valor es obtenido a partir del cociente del VPP/Prevalencia=$.72/.19=3.79$

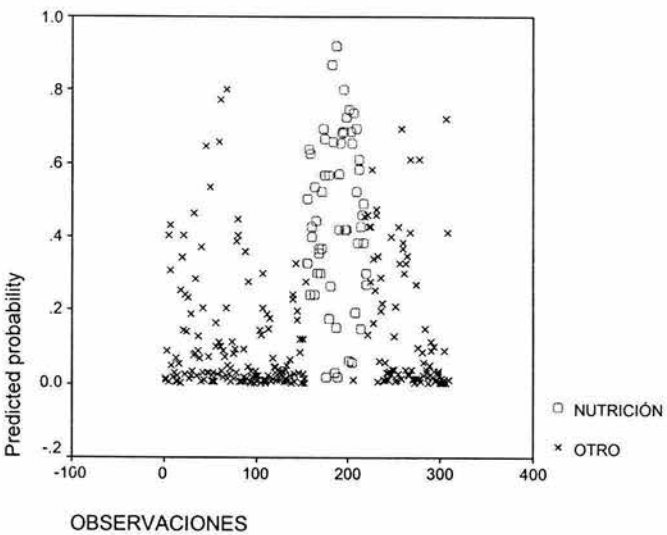
RESIDUALES

Después de que se obtuvo el modelo de regresión logística es bueno aplicar las medidas de diagnóstico, entre ellos se encuentran los residuales *estandarizados de devianza* y las *medidas de "influence"*, con la finalidad de identificar los comportamientos sistemáticos que pudieran dar indicios de un modelo inadecuado, como se verá en las siguientes gráficas de dispersión, obtenidas a partir del paquete estadístico SPSS.

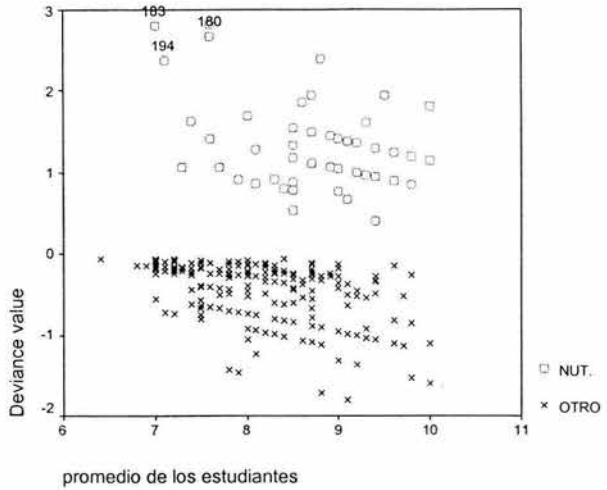
En la gráfica 1, se forman dos grupos, los valores ***positivos*** pertenecen a la ***carrera de nutrición*** y los valores ***negativos*** corresponden a la ***carrera de administración e ingeniería***.



Gráfica de dispersión 1. Residuos estándarizados vs observación.



Gráfica de dispersión 2.Probabilidad estimada vs observación.



Gráfica de dispersión 3. Residuos estándarizados vs promedio de los estudiantes.

Las gráficas faltantes para las variables categóricas se encuentran en el *anexo I, gráficas del 1-4*. En las gráficas obtenidas por los residuales se aprecia que el modelo esta ajustado correctamente ya que los 2 grupos se encuentran entre 2 y -2.

Los valores que son considerados puntos influyentes son los que se distinguen de los grupos de dispersión, las observaciones son: **180** (hombre, 20 años, nutrición, UAM), **173** (mujer, 18 años, nutrición, UIA), **183** (mujer, 19 años, nutrición, UAM) y **194** (mujer, 21 años, nutrición, UAM), como son pocos los casos al llevar a cabo la eliminación de estos no influyen significativamente en la obtención de resultados, por lo tanto no fue necesario llevar a cabo otro proceso de ajuste.

Otra prueba de ajuste es la de Hosmer y Lemeshow, es decir, que el modelo calculado se ajuste efectivamente a los datos observados. Como se muestra en el *anexo I, tabla 2,* el ajuste es bueno en general ya que los valores esperados son similares a los observados.

PRUEBA DE SIGNIFICANCIA

La prueba de significancia del modelo se realiza a través de la prueba del Cociente de Verosimilitud basándose en la estadística $G = -2(L_o - L_a)$, que es la diferencia entre la devianza del modelo ajustado y la devianza del modelo saturado teniendo una distribución χ_1^2.

Se quieren probar las siguientes hipótesis:

$$H_0 : \beta_0 + \beta_1 \text{Promedio} + \beta_2 \, Universidad + \beta_3 \, Sexo + \beta_4 Quiso\ esta\ carrera$$

o $H_0 : \beta_5 = 0$

vs.

$$H_a : \beta_0 + \beta_1 \text{Promedio} + \beta_2 \, Universidad + \beta_3 \, Sexo + \beta_4 Quiso\ esta\ carrera + \beta_5 \, Esc.ambospadres$$

o

$H_a : \beta_5 \neq 0$

Quedando de la siguiente manera:

Modelo	(-2log-likelihood)	$G = -2(L_o - L_a)$	χ_1^2
H_0	196.082	12.731	3.840
H_a	208.813		

Como el valor en tablas para la χ_1^2 con un grado de libertad es igual a 3.840 y se tiene que $\chi_1^2 < G$, se rechaza la hipótesis nula y nos quedamos con el modelo que contiene a las 5 variables explicativas (promedio, universidad, sexo, quiso la carrera y escolaridad de ambos padres), para extraer las conclusiones correspondientes al modelo propuesto.

2.5.3. INTERVALO DE CONFIANZA

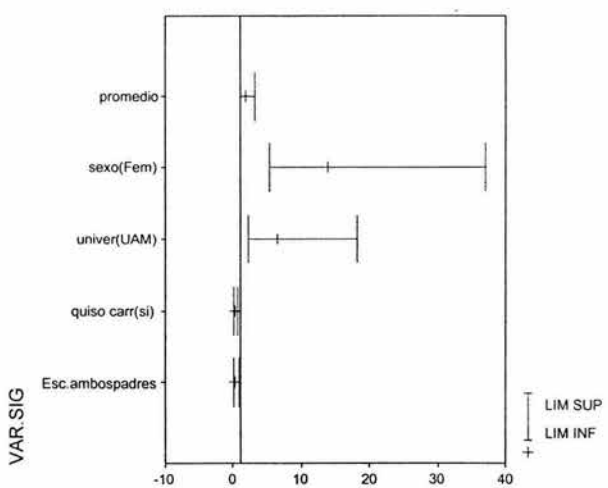
Después de proponer un modelo para la elección de carrera de nutrición, revisaremos el grado de significancia estadística de cada una de las variables con un intervalo de confianza del 95%, para el cociente de momios ($exp(\hat{\beta})$).

Variables in the Equation

		B	Exp(B)	95.0% C.I.for EXP(B)	
				Lower	Upper
Step 1[a]	PROMEDIO	.620	1.859	1.103	3.135
	QESTCARR(SI)	-1.503	.222	.077	.642
	SEXO(FEM)	2.641	14.025	5.303	37.087
	UNIVER(UAM)	1.878	6.541	2.346	18.233
	EDPADRES(SEC)	-.271	.762	.237	2.455
	Constant	-8.756	.000		

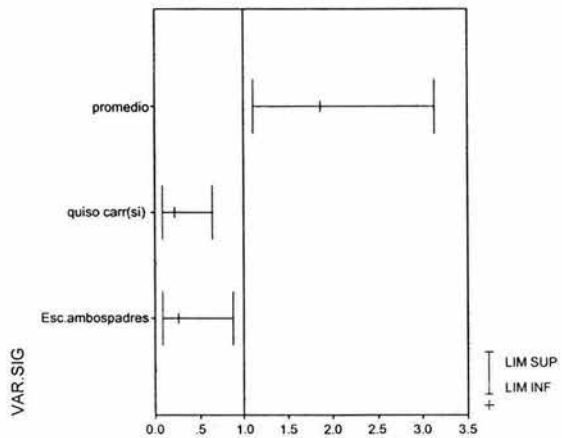
a. Variable(s) entered on step 1: PROMEDIO, QESTCARR, SEXO, UNIVER, EDPADRES.

Cuadro 5. Intervalo de confianza del 95%, de las variables estadísticamente significativas de los "Odds Ratio".



GRÁFICA 4.Intervalo de confianza del 95% de la $exp(\hat{\beta})$.

La siguiente gráfica se presenta para su mejor apreciación de las variables explicativas: promedio, quiso la carrera y escolaridad ambos padres.



GRÁFICA 4a. Intervalo de confianza del 95% de la $exp(\hat{\beta})$.

Aquellos intervalos de confianza que se encuentran del lado derecho del valor de referencia de 1, significa que los $\hat{\beta}_i$>0, teniendo un efecto positivo con respecto al factor riesgo de elección de carrera de nutrición, ocurriendo lo contrario para aquellos intervalos de confianza con $\hat{\beta}_i$<0, corroborándose que ningún intervalo de confianza contiene el valor de uno lo cual nos indica que son estadísticamente significativas.

2.5.4. INTERPRETACIÓN DEL MOMIO

Los momios asociados a un suceso se definen como la razón entre la probabilidad de que el suceso ocurra y la probabilidad de que no ocurra, por ese motivo son de gran importancia ya que por medio de este estimador se establece la relación directa que existe entre la variable respuesta y las categorías

explicativas seleccionadas. Cuando se tiene un valor resultante de $\hat{\beta}_i$>0 la función es creciente y si el valor resultante es $\hat{\beta}_i$<0 la función es decreciente, esto nos indica que p crece o decrece según como lo hace la variable independiente.

Partimos de la Variable respuesta:

$$Y_i = \begin{cases} 1 & Elección\ de\ la\ carrera\ nutrición \\ 0 & Otra\ (administración\ e\ ingeniería) \end{cases}$$

Variable explicativa:

$$X_1 = \begin{cases} 1 & Universidad\ autónoma\ metropolitana \\ 0 & Universidad\ iberoamericana \end{cases}$$

Por lo tanto el cociente de momios queda de la siguiente manera:

$$RM = \frac{Momio(Carrera\ Nutrición \quad Universidad = UAM, prom. = x, sexo = x, quisocarr = x, edcpadres = x)}{Momio(Carrera\ Nutrición \quad Universidad = UIA, prom. = x, sexo = x, quisocarr = x, edcpadres = x)}$$

$$= e^{1.878} = 6.541$$

La interpretación del momio se tiene que realizar para una de las variables estadísticamente significativas mientras que las demás permanecen constantes, entonces la probabilidad de riesgo de elección de carrera de nutrición en lugar de las otras dos (administración e ingeniería) con respecto a los estudiantes que ingresaron a la Universidad Autónoma Metropolitana (UAM) es 6.541 veces la probabilidad de que un estudiante hubiera preferido ingresar a la Universidad Iberoamericana (UIA) con una elección de la carrera de nutrición en lugar de las otras dos. De la misma manera el intervalo de confianza (*gráfica 4*) nos confirma lo citado anteriormente ya que se tiene que el momio de elegir la carrera de nutrición de los estudiantes que ingresaron a la UAM puede ser aproximadamente 2 a 18 veces más que el momio de elegir la carrera de nutrición e ingresar a la UIA.

2.5.5. Conclusiones

El procedimiento anterior se realizó con la finalidad de encontrar el mejor modelo ajustado ya que se tuvieron que hacer modificaciones durante la elaboración del modelo *"logit"* con respecto a las variables y así contar con una buena estimación de los factores que se consideraron para la construcción.

La encuesta se realizó a 59 estudiantes de la carrera de nutrición, de acuerdo a la distribución de la *tabla 1, anexo I* se tiene que el 54.2% son de la Universidad Iberoamericana mientras que la Universidad Autónoma Metropolitana (UAM) tiene 45.8%, es importante notar que no existen hombres en la encuesta realizada de la Universidad Iberoamericana (UIA), lo cual no infiere que eran los únicos hombres de la generación de 1996 de las 2 instituciones. Por lo tanto de acuerdo al modelo logit presentado, se tiene que la probabilidad de elección de la carrera de nutrición al de las otras dos carreras (Adm. e Ing.) con respecto a los estudiantes que ingresaron a la UAM fue mayor en comparación a la probabilidad de haber elegido la UIA la carrera de nutrición en lugar de las otras dos carreras.

La variable promedio de los estudiantes tiene que la probabilidad de elección de la carrera de nutrición aumenta conforme aumenta la calificación de los estudiantes al salir de preparatoria (*gráfica de dispersión 5, anexo I).*

El sexo de los estudiantes resulta significativo que aparezca ya que la probabilidad de elección de la carrera de nutrición en lugar de las otras dos carreras de los estudiantes del sexo femenino fue mayor a la probabilidad de los estudiantes del sexo masculino que eligieron la carrera de nutrición en lugar de las otras dos opciones de carreras.

De los 302 estudiantes que contestaron la pregunta de que si ya tenían en mente que carrera elegir antes de salir de preparatoria el 67.2% contesto no,

mientras que el 32.8% contestó que si, el modelo logit indica es que la probabilidad de elección de la carrera de nutrición en lugar de las otras dos carreras de aquellos estudiantes que ya tenían en mente que carrera elegir antes de salir de preparatoria fue menor en comparación de la probabilidad de los estudiantes que no sabían al elegir la carrera de nutrición en lugar de las otras dos carreras.

Por último se ha considerado la escolaridad de ambos padres siendo una característica importante que apareció en el modelo, la probabilidad de elección de la carrera de nutrición en lugar de las otras dos carreras con respecto a los estudiantes que tienen ambos padres con a lo más secundaria fue menor en comparación de los estudiantes que tienen al menos un padre con licenciatura de la carrera de nutrición en lugar de las otras dos carreras. De los 57 alumnos que contestaron la educación que tienen sus padres el 47.4% tiene al menos un padre con licenciatura, el 14% tiene a ambos padres con a lo más secundaria, los alumnos que tienen al menos un padre con preparatoria son el 12.3% y el 26.3% pertenece a los alumnos que tienen al menos un padre con maestría o doctorado.

2.6. *Análisis Correspondiente a la carrera de Ingeniería en Electrónica*

2.6.1. Modelo Propuesto

Se va a proceder de misma forma del punto *2.5.*, para la obtención del modelo logit de la carrera de Ingeniería, a partir de las variables explicativas que se encuentran en la *tabla E.*

El modelo a ajustar es el siguiente:

$$Y = \begin{cases} 1 & \textit{Elección de la carrera de ingeniería} \\ 0 & \textit{En otro caso} \end{cases}$$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	PROMEDIO	-.094	.256	.136	1	.712	.910	.551	1.503
	TERCARRE(EST)			2.524	2	.283			
	TERCARRE(TRAB)	.040	.429	.009	1	.926	1.040	.449	2.412
	TERCARRE(OTRA)	1.242	.788	2.486	1	.115	3.463	.739	16.222
	QESTCARR(SI)	.541	.387	1.956	1	.162	1.718	.805	3.666
	PAIS(NO)	.693	.645	1.154	1	.283	2.000	.565	7.080
	SEXO(FEM)	-2.741	.525	27.217	1	.000	.065	.023	.181
	OCUPMADRE(HOG)	-1.028	.438	5.520	1	.019	.358	.152	.843
	IMAFE(AND)			2.638	3	.451			
	IMAFE(IND)	-.077	.504	.023	1	.879	.926	.345	2.488
	IMAFE(FEM)	.528	.547	.932	1	.334	1.695	.581	4.948
	IMAFE(MASC)	-.392	.533	.539	1	.463	.676	.238	1.923
	COMMADRE(NO)	-.747	.679	1.210	1	.271	.474	.125	1.794
	ESCPADRE(LIC)			8.176	3	.043			
	ESCPADRES(SEC)	1.226	.663	3.420	1	.064	3.407	.929	12.493
	ESCPADRES(PREP)	-.312	.544	.329	1	.566	.732	.252	2.126
	ESCPADRES(MoD)	-.906	.616	2.165	1	.141	.404	.121	1.351
	UNIVER(UAM)	-.362	.498	.528	1	.467	.696	.262	1.848
	Constant	.761	2.109	.130	1	.718	2.140		

a. Variable(s) entered on step 1: PROMEDIO, TERCARRE, QESTCARR, PAIS, SEXO, MADTRAB, IMAFE, COMMADRE, EDPADRES, UNIVER.

Cuadro 6. Modelo ajustado de 10 variables explicativas para ajustar la probabilidad de elección de carrera de Ingeniería.

Se obtuvieron como variables estadísticamente significativas; sexo del estudiante, ocupación que tiene madre y por último la escolaridad de ambos padres.

Usando el método de selección backward se obtuvieron los siguientes resultados, los valores con el asterisco (*) son los valores de referencia y los de doble asterisco (**) son valores no significativos para el modelo, como se muestra a continuación.

Factor (i)	Etiqueta	Variable	Coeficiente de regresión	Error estándar	$\left(\frac{\beta}{SE(\beta)}\right)^2$	
Sexo del estudiante	*Sexo(Fem)*	Hombre=0 * ***Mujer=1***	*-2.606*	*0.455*	*32.758*	$P < 0.001$
Ocupación que tiene la madre	*Ocupmadre(Hog)*	*Trab=0* * ***Hogar=1***	*-1.005*	*0.41*	*6.012*	$P < 0.02$
Escolaridad ambos padres	*Edpadres(Sec)*	Lic.=0 * ***Sec=1*** Prep=2 ** MoD=3 **	*0.941*	*0.489*	*3.694*	$P < 0.06$

Cuadro 7. Resultados de ajustar la probabilidad de elección de la carrera de ingeniería en función del sexo, ocupación que tiene la madre y escolaridad ambos padres, según la regresión logística por el método de selección backward.

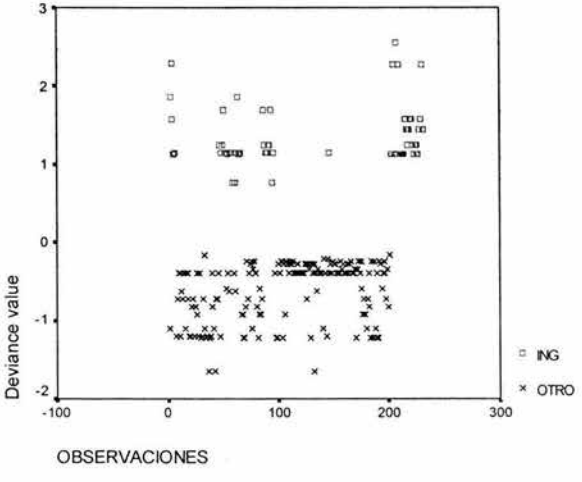
Se obtuvieron las mismas variables significativas que el cuadro anterior, dicha significancia también se puede comprobar por medio del intervalo de confianza del coeficiente de regresión, por ejemplo para la variable sexo se tiene: $-2.606 \pm 1.96(.455) = (-1.71, -3.49)$ como no incluye el valor de 0 el intervalo de confianza del 95% la variable resulta tener una importancia para la variable respuesta lo cual difiere con respecto a la categoría escolaridad de ambos padres de a lo más ambos con secundaria ya que su intervalo de confianza del 95% incluye el valor de cero $.941 \pm 1.96(.489) = (-0.017, 1.899)$, esto significa que la variable explicativa es independiente de la variable respuesta, estadísticamente resulta ser no significativa para el modelo. Para determinar la inclusión de la variable escolaridad de ambos padres al modelo se va ha realizar la prueba del coeficiente de verosimilitud.

2.6.2. *Prueba de significancia del modelo*

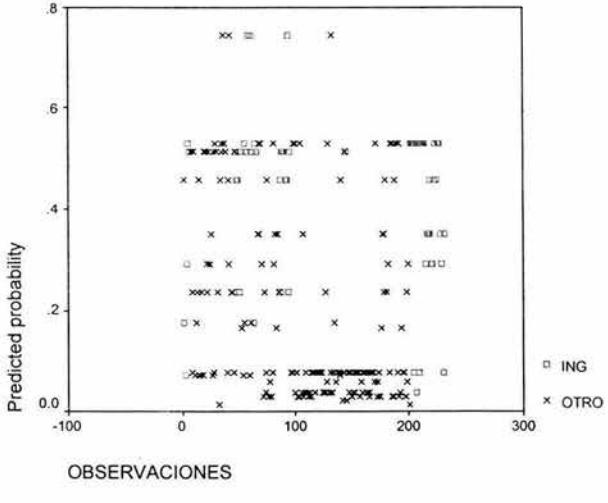
Análisis de residuales

A continuación se presenta el análisis de residuales para ver el comportamiento sistemático del modelo ajustado y notar si hay inicios de un modelo inadecuado o detectar casos de observaciones de puntos influyentes.

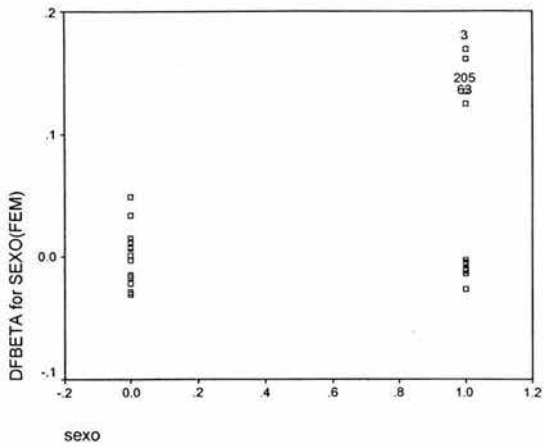
Los valores ***positivos*** de la gráfica siguiente representa la ***carrera de ingeniería*** y los valores ***negativos*** son de las ***carreras de administración*** *y* ***nutrición.***



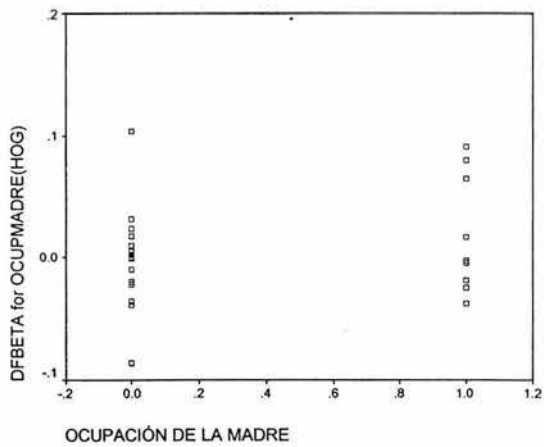
Gráfica dispersión 4. Residuos estándarizados vs el número de observación.



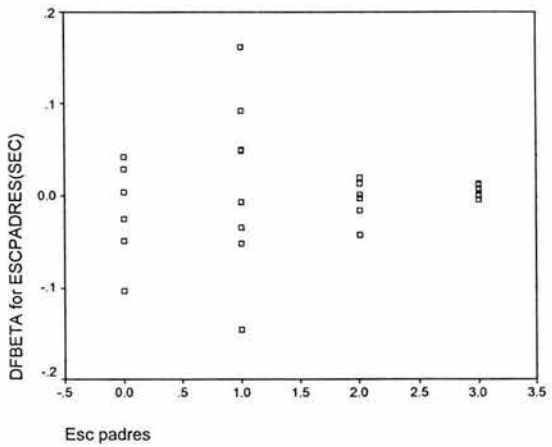
Gráfica de dispersión 5.Probabilidad estimada vs el número de observación.



Gráfica de dispersión 6. Df-Beta de sexo vs Sexo de cada estudiante (0=Masc., 1=Fem.)



Gráfica de dispersión 7. Df-Beta Ocupación de la madre vs ocupación de la madre (0=Trabaja., 1=Hogar)



Gráfica de dispersión 8. Df-Beta escolaridad de los padres vs escolaridad ambos padres (0=Lic., 1=Sec., 2=prep., 3=Maestría o Doct.

Por las gráficas anteriores se puede percatar que el ajuste del modelo es bueno por la razón de que las observaciones se encuentran entre 2 y -2. El ajuste del modelo también se puede corroborar por medio de la prueba de Hosmer y Lemeshow, ya que el modelo calculado se ajusta efectivamente a los datos estimados (similares en los valores), como se muestra en la *tabla 5, anexo I*.

PRUEBA DE SIGNIFICANCIA DEL MODELO

La finalidad de realizar esta prueba es con el objetivo de identificar si la variable escolaridad de ambos padres de a lo más ambos con secundaria es de relevancia ingresarla al modelo, a continuación se establecen los dos modelos (ajustado y saturado) para realizar la prueba de las devianzas por la estadística $G = -2(L_o - L_a)$.

Se van a probar las siguientes hipótesis:

$$H_0 : \beta_0 + \beta_1 Sexo + \beta_2\ Ocupación\ madre$$

Vs.

$$H_a : \beta_0 + \beta_1 Sexo + \beta_2\ Ocupación\ madre + \beta_3 Esc.ambos\ padres$$

Queda de la siguiente manera:

Modelo	(-2log-likelihood)	$G = -2(L_o - L_a)$	χ_1^2
H_0	207.49	8.803	3.840
H_a	216.293		

El valor en tablas para la χ_1^2 con un grado de libertad con nivel de significancia del 0.05 es igual a 3.840, se tiene que $\chi_1^2 < G$ por lo tanto se rechaza la hipótesis nula y se propone la hipótesis alternativa para la extracción de conclusiones, esto indica que la variable escolaridad de ambos padres resulta ser relevante para la explicación de la variable respuesta.

2.6.3. Modelo Propuesto

Modelo logit:

$$\ln\left(\frac{\hat{\pi}(Y=1)}{1-\hat{\pi}(Y=1)}\right) = 0.122 - 2.606\ Sexo\ (Femenino) - 1.505\ Ocupación\ \ madre\ (Hogar) + 0.941\ Escolaridad\ \ padre(Sec)$$

2.6.4. *PORCENTAJE DE CLASIFICACIÓN*

El siguiente cuadro determina el porcentaje de clasificación correcta.

Classification Table[a]

Observed			Predicted		
			carrera de Ingenieria		Percentage Correct
			OTRO	ING	
Step 1	carrera de Ingenieria	OTRO	142	29	83.0
		ING	29	31	51.7
	Overall Percentage				74.9

a. The cut value is .500

Cuadro 8. Tabla de Clasificación para el modelo elección de la carrera de ingeniería.

Se obtiene que el valor de *especificidad* es $0.83 = (142/171)$, indicando que el modelo de clasificación es correcta, esto es, que los estudiantes que eligieron las carreras de administración y nutrición son clasificados correctamente en la elección de las carreras de las mismas.

Para el valor de *sensibilidad* se tiene $0.52 = (31/60)$, estudiantes que se encuentran en la carrera de nutrición que son clasificados correctamente en la licenciatura de nutrición, por lo tanto el modelo propuesto es aceptable. Se observa que el tamaño de los valores de la variable respuesta son distintos entre si.

Para la obtención del valor de pronóstico positivo se tiene el siguiente cuadro.

Pronóstico	Carrera		Total	Valor pronóstico
	Adm. Y Nut.	Ing.		
Adm. Y Nut.	142	29	171	(142/171)=.83
Ing.	29	31	60	(31/60)=.52
Total	171	60	231	

Cuadro 9. Valor pronóstico sobre los datos que se están clasificando como correctos.

El *valor de pronóstico positivo* es .52 y la *prevalencia* de la muestra es de $(60/231) = 0.26$, esto indica que fue mejor en más de 2 veces el pronóstico de elección de carrera mediante el modelo propuesto que sin él.

2.6.4. *Intervalo de Confianza*

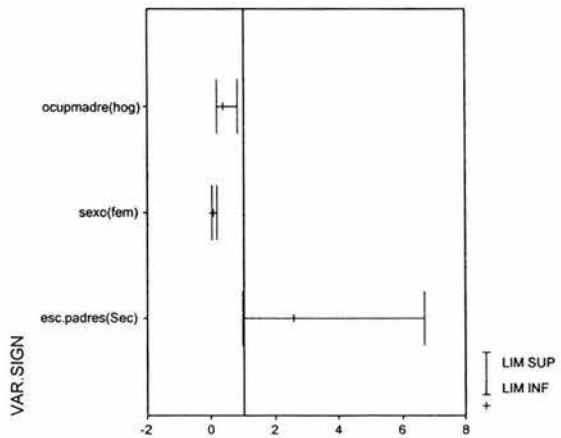
El grado de significancia que tienen las variables del cuadro 2, se pueden apreciar con una mayor claridad a través de las gráficas *"high low chart"*, como se vera a continuación.

Variables in the Equation

		B	Exp(B)	95.0% C.I.for EXP(B)	
				Lower	Upper
Step 1[a]	SEXO(FEM)	-2.606	.074	.030	.180
	OCUPMADRE(HOG)	-1.005	.366	.164	.817
	ESCPADRES(SEC)	.941	2.562	.982	6.687
	Constant	.122	1.130		

a. Variable(s) entered on step 1: SEXO, MADTRAB, EDPADRES.

Cuadro 10. Valores del intervalo de confianza para la $exp(\hat{\beta})$.



GRÁFICA 9 Intervalo de confianza del 95% de la $exp(\hat{\beta})$.

Como se había mencionado en el punto 2.6.1. los intervalos de confianza de las variables ocupación que tiene la madre del estudiante y el sexo del estudiante no incluye el valor de uno en cambio la categoría escolaridad de a lo más ambos con secundaria contiene el valor de 1 el intervalo de confianza, pero por la prueba del cociente de verosimilitud la variable fue considerada para la construcción del modelo logit.

2.6.5. INTERPRETACIÓN DEL MOMIO

La interpretación de los "odds ratio" o "razón de momios" corresponde a la ocurrencia de un suceso bajo cierta condición entre los que corresponden bajo otra condición. Se parte de la variable respuesta:

$$Y_i = \begin{cases} 1 & Elección\ de\ la\ carrera\ ingeniería \\ 0 & Otra\ (administración\ y\ nutrición) \end{cases}$$

Se obtuvieron como variables estadísticamente significativas: sexo, *ocupación que tiene madre y escolaridad ambos padres.*

La razón de momio para un caso en particular se tiene:

$$RM = \frac{Momio(Carrera\ Ingenieria \mid Ocupación\ Madre = hogar, sexo = x, esc.ambos\ padre = x)}{Momio(Carrera\ Ingenieria \mid Ocupación\ Madre = trabaja, sexo = x, esc.ambos\ padre = x)}$$

$$= e^{-1.005} = 0.366$$

La probabilidad de elegir la carrera de ingeniería al no elegirla cuando los estudiantes tienen a su madre que se dedica al hogar es 0.366 veces la probabilidad de que el estudiante tenga a su madre que se dedica a trabajar y eligió la carrera de ingeniería en lugar de no elegirla, siempre y cuando las demás variables explicativas permanezcan constantes, lo que se acaba de decir se puede apreciar por la *gráfica 9* ya que indica que el momio de elegir la carrera de ingeniería de los estudiantes que tuvieron a su madre que se dedica al hogar puede ser tan chico como 0.16 o grande como .817 veces más que el momio de elegir la carrera de ingeniería y que tuvieran a su madre que se dedica a trabajar.

2.6.6. Conclusiones

Esta carrera es preferida por la población estudiantil masculina ya que de las 60 observaciones obtenidas el 88.3% pertenece al sexo masculino mientras que el 11.7% son del sexo femenino (*tabla 6, anexo I*). En esta carrera se dio un igual número de observaciones obtenidas de la Universidad Iberoamericana y de la Universidad Autónoma Metropolitana con 30 observaciones para cada universidad. *(tabla 7, anexo I)*

El momio de la variable sexo del estudiante indica que la probabilidad de elegir la carrera de ingeniería al no elegirla si es del sexo femenino es menor en comparación a la probabilidad de los estudiantes del sexo masculino de elegir la carrera de ingeniería al no elegirla.

Como se había mencionado al inicio de este capítulo la familia tiene un papel importante en la decisión que tiene el estudiante al elegir carrera ya que la madre es considerada la persona de mayor influencia hacia los hijos, dicha característica la encontramos en la ocupación que desempeña la madre, ya que de los 60 alumnos que contestaron cual es la ocupación que tiene la madre cuando eligieron la carrera de ingeniería el 35% se dedican al hogar mientras que el 65% se encontraba trabajando (*tabla 8, anexo I)*. Por el modelo se obtiene que la probabilidad de elección de la carrera de ingeniería al no elegirla cuando los estudiantes tienen a su madre que se dedica al hogar es menor a la probabilidad de los estudiantes que tienen a su madre que se dedica a trabajar y elegir la carrera de ingeniería en lugar de las otras dos (nutrición y administración).

Otro factor importante que apareció en el modelo logístico es la escolaridad de ambos padres, teniendo en la muestra que de los 60 alumnos que escogieron la carrera de ingeniería el 30% corresponde a los padres que tuvieron a lo más secundaria mientras que el 10% a los padres con al menos maestría o doctorado, el 20% a los padres que al menos uno tiene la preparatoria y el 40% le corresponde a los padres que al menos uno tiene la licenciatura *(tabla 9, anexo I)*, el momio de la variable escolaridad de ambos padres indica que la probabilidad de elección de la carrera de ingeniería al no elegirla cuando los estudiantes tienen a sus padres con ambos a lo más secundaria fue mayor en comparación a la probabilidad de los estudiantes de al menos uno de los padres tiene la licenciatura y que hayan elegido la carrera de ingeniería en lugar de no elegirla.

2.7. *Análisis de la Licenciatura de Administración*

Las variables explicativas que se van ha utilizar son las que se presentan en la *tabla E* de este capítulo, el análisis se realizó de la misma forma que los anteriores.

$$Y_i = \begin{cases} 1 & \textit{Elección de la carrera de administración} \\ 0 & \textit{En otro caso} \end{cases}$$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	PROMEDIO	-.351	.209	2.822	1	.093	.704	.467	1.060
	UNIVER(UAM)	-.861	.407	4.475	1	.034	.423	.190	.939
	TERCARRE(EST)			7.145	2	.028			
	TERCARRE(TRAB)	.869	.359	5.867	1	.015	2.384	1.180	4.813
	TERCARRE(OTRA)	-.655	.799	.671	1	.413	.520	.108	2.490
	QESTCARR(SI)	.053	.333	.025	1	.873	1.054	.549	2.024
	PAIS(NO)	.808	.579	1.944	1	.163	2.243	.721	6.981
	SEXO(FEM)	.498	.364	1.870	1	.171	1.645	.806	3.356
	OCUPMADRE(HOG)	1.430	.376	14.496	1	.000	4.178	2.001	8.721
	IMAFE(AND)			4.501	3	.212			
	IMAFE(IND)	-.114	.425	.072	1	.788	.892	.388	2.052
	IMAFE(FEM)	-.726	.409	3.147	1	.076	.484	.217	1.079
	IMAFE(MASC)	.150	.448	.112	1	.738	1.162	.483	2.796
	COMMADRE(SI)	.243	.571	.182	1	.670	1.276	.417	3.906
	ESCPADRES(LIC)			2.967	3	.397			
	EDPADRES(SEC)	-.556	.557	.997	1	.318	.573	.193	1.708
	EDPADRES(PREP)	.342	.458	.557	1	.455	1.408	.574	3.454
	EDPADRES(MoD)	.277	.424	.425	1	.514	1.319	.574	3.029
	Constant	2.504	1.719	2.122	1	.145	12.229		

a. Variable(s) entered on step 1: PROMEDIO, UNIVER, TERCARRE, QESTCARR, PAIS, SEXO, MADTRAB, IMAFE, COMMADRE, EDPADRES.

Cuadro 11. Valores de los coeficientes de las 10 variables explicativas incluidas para ajustar la probabilidad de elección de la carrera de administración.

Se obtuvieron 4 variables estadísticamente significativas y ahora se va a realizar el análisis por el método de selección backward.

Factor (i)	Etiqueta	Variable	Coeficiente de regresión	Error estándar	$\left(\frac{\beta}{SE(\beta)}\right)^2$	
Universidad	*Univer(UAM)*	UIA=0 *				
		UAM=1	*-0.915*	*0.322*	*8.069*	$P < 0.01$
Ocupación que tiene la madre	*Ocupmadre(Hog)*	*Trab=0* *				
		Hogar=1	*1.241*	*0.328*	*14.286*	$P < 0.001$
Al terminar la carrera	*Tercarre(Trab)*	Estudiar=0 *				
		Trabajar=1	*0.581*	*0.327*	*3.159*	$P < 0.07$
		Otra=2 **				

Cuadro 12. Resultados de ajustar la probabilidad de elección de la carrera de ingeniería en función del sexo, ocupación que tiene la madre y al terminar la carrera, según la regresión logística por el método de selección backward.

Modelo logístico:

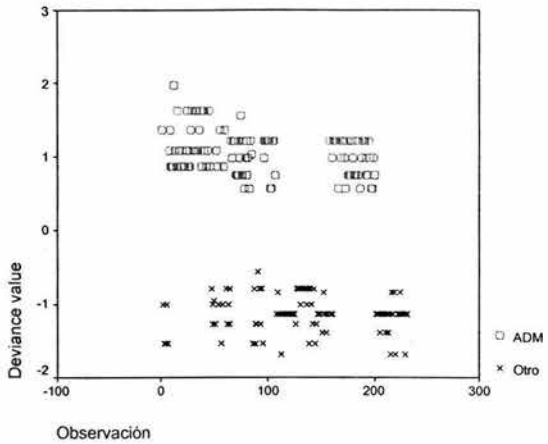
$$\ln\left(\frac{\hat{\pi}(Y=1)}{1-(\hat{\pi}(Y=1))}\right) = -.096 - .915\ Universidad\ (UAM) + 1.241\ Ocupación\ madre\ (Hogar) + .581 Ter.carr.\ (Trab)$$

2.7.1. PRUEBA DE SIGNIFICANCIA DEL MODELO

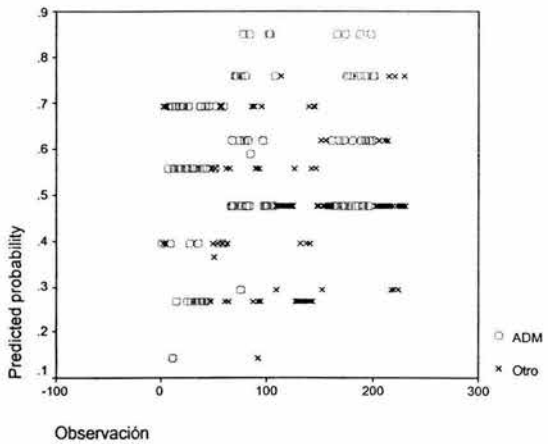
ANÁLISIS DE RESIDUALES

Por medio del análisis del residual de Pearson y el DF-beta, se revisará si el ajuste del modelo obtenido es bueno, de suceder lo contrario se realizará el procedimiento necesario para mejorar dicho ajuste.

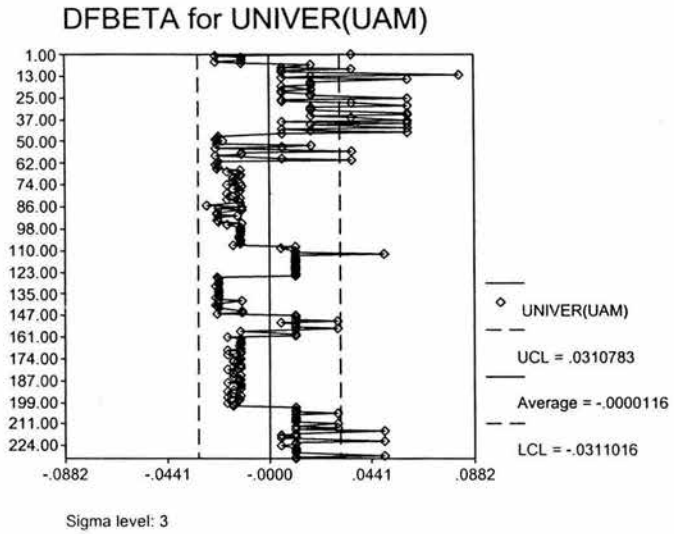
En la *gráfica 10* se tiene que los valores ***positivos*** representan a los estudiantes de ***administración,*** mientras que los ***negativos*** son los estudiantes de ***nutrición e ingeniería.***



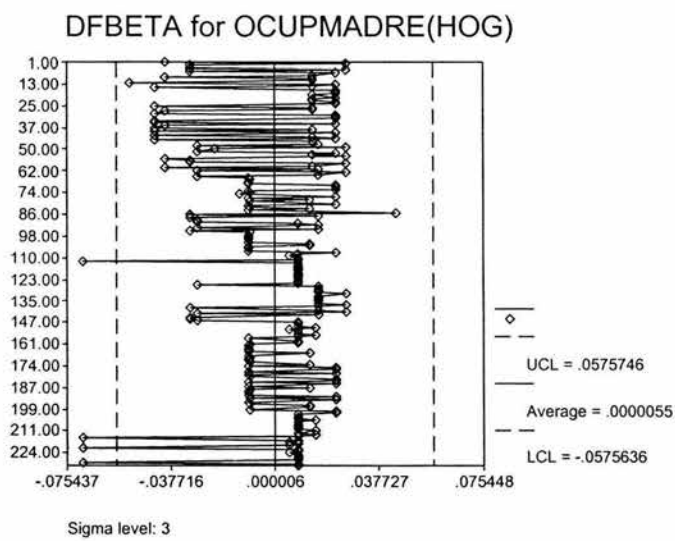
Gráfica de dispersión 10.Residuos estándarizados vs Número de observación.



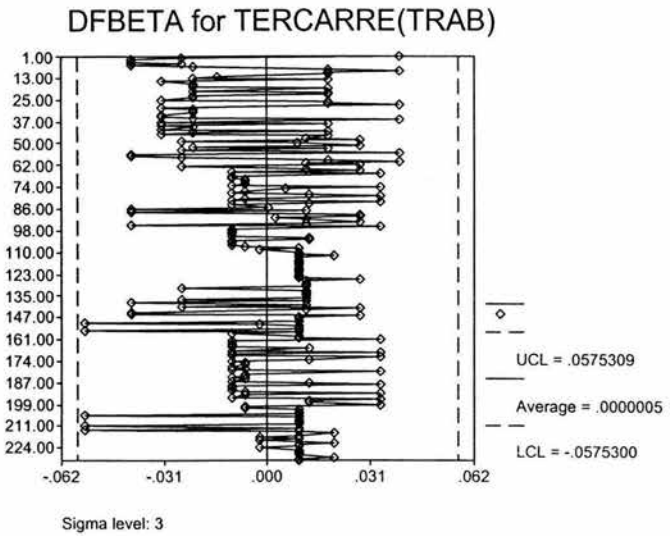
Gráfica de dispersión 11. Probabilidad estimada vs Número de observación.



Gráfica 12. DF-Beta de Universidad (UAM).



Gráfica 13. DFBeta para la ocupación que tiene la madre (hogar).



Gráfica 14. DF-Beta actividad que van a realizar al terminar la carrera (Trab).

El ajuste es bueno en general para el modelo propuesto, por la *gráfica 10* se obtuvo que las observaciones se encuentran entre 2 y -2.

La prueba de bondad de ajuste de Hosmer y Lemeshow *(Tabla 10, Anexo I)*, también indica que el ajuste es bueno ya que los valores observados son muy similares a los estimados.

PORCENTAJE DE CLASIFICACIÓN

El porcentaje de clasificación correcto se presenta en el siguiente cuadro:

Classification Table[a]

Observed			Predicted		
			CARRERA DE ADM		Percentage Correct
			OTRO	ADM	
Step 1	CARRERA DE ADM	OTRO	77	30	72.0
		ADM	49	68	58.1
	Overall Percentage				64.7

a. The cut value is .500

Cuadro 13. Clasificación de la tabla para el modelo de elección de la carrera de administración.

Se tiene que la proporción de los valores de *sensibilidad* y *especificidad* son: *especificidad* es igual a $(77/107)=.72$, esto indica que la clasificación correcta de los datos observados de las carreras de nutrición e ingeniería se estimaron correctamente 77 observaciones de un total de 107 casos, el valor de *sensibilidad* es $(68/117)=.58$, se obtuvo un valor relativamente bajo pero esto se debe a que las observaciones de los valores de la variable respuesta son distintas, en una se tiene un mayor número de observaciones que en la otra, por lo tanto se dice que es aceptable el modelo obtenido.

Para el indicador del valor de *pronóstico positivo* se tiene .69, mientras que la *prevalencia* es igual a $(117/224)=.52$, esto dice que es mejor en más de una vez el pronóstico de elección de la carrera de administración mediante el modelo propuesto que sin haberlo propuesto, obteniéndose este valor a partir de la diferencia del cociente entre VPP/prevalencia esto es igual a $.69/.52=1.32$

Pronóstico	Carrera		Total	Valor pronóstico
	OTRO	ADM		
OTRO	77	49	126	(77/126)=.61
ADM	30	68	98	(68/98)=.69
Total	107	117	224	

Cuadro 14. Valor pronóstico sobre los datos que se están clasificando como correctos.

PRUEBA DEL COCIENTE DE VEROSIMILITUD

Se establece el estadístico $G = -2(L_o - L_a)$ para la aplicación de la prueba del cociente de verosimilitud, a partir de la especificación del modelo ajustado y el modelo saturado.

Prueba de Hipótesis:

$$H_0 : \beta_0 + \beta_1 Universidad + \beta_2\ Ocupación\ madre$$

VS.

$$H_a : \beta_0 + \beta_1 Universidad + \beta_2\ Ocupación\ madre + \beta_3 Terr\ carrera$$

Se tiene:

Modelo	(-2log-likelihood)	$G = -2(L_o - L_a)$	χ_1^2
H_0	290.67	4.993	3.840
H_a	285.677		

El valor para la χ_1^2 con 1 grado de libertad con nivel de significancia del 0.05 en tablas es igual a 3.840, como el valor $\chi_1^2 < G$ se rechaza la hipótesis nula y se acepta la hipótesis alternativa que contiene las variables *universidad, ocupación que tiene la madre* y actividad que van a realizar *al terminar la carrera*.

2.7.2. INTERPRETACIÓN DE LOS COEFICIENTES

INTERVALO DE CONFIANZA

El intervalo de confianza del 95% para las variables estadísticamente significativas es el siguiente:

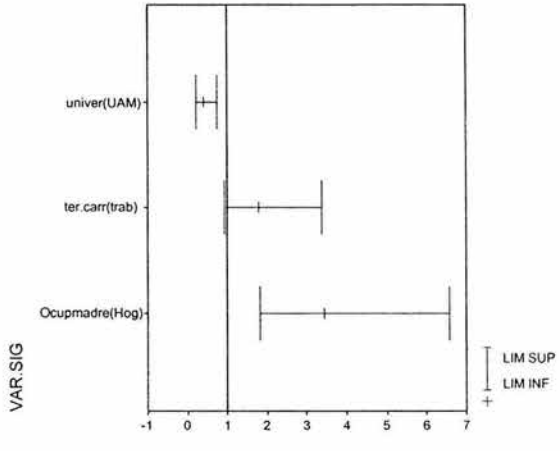
Variables in the Equation

		B	Exp(B)	95.0% C.I.for EXP(B)	
				Lower	Upper
Step 1[a]	UNIVER(UAM)	-.915	.401	.213	.753
	OCUPMADRE(HOG)	1.241	3.458	1.817	6.580
	TERCARRE(TRAB)	.581	1.788	.942	3.394
	Constant	-.096	.908		

a. Variable(s) entered on step 1: UNIVER, MADTRAB, TERCARRE.

Cuadro 15. Resultado de los coeficientes para el intervalo de confianza del 95%.

Gráficamente se tiene:



Gráfica 15. Intervalo de confianza del 95%, para los cociente de momios.

Recordemos que se tiene como variable respuesta:

$$Y_i = \begin{cases} 1 & \textit{Eleccion de la carrera administración} \\ 0 & \textit{Otra (nutrición e ingeniería)} \end{cases}$$

Las variables explicativas estadísticamente significativas son: *al terminar la carrera (trabajar), universidad y por último ocupación que tiene la madre (hogar).*

Tenemos como razón de momios:

$$RM = \frac{Momio(Carrera\ Administración \mid Ocupación\ de\ la\ madre = Hogar, Universidad = x, terminar\ carr = x)}{Momio(Carrera\ Administración \mid Ocupación\ de\ la\ madre = trabaja, Universidad = x, terminar\ carr = x)}$$

$$= e^{1.241} = 3.46$$

La interpretación correcta del momio indica que la probabilidad de elegir la carrera de administración en lugar de otro caso de los estudiantes que tienen a su madre que se dedica al hogar fue de 3.46 veces en comparación a la probabilidad de los estudiantes que tienen a su madre que se dedica a trabajar y que hayan escogido la carrera de administración en lugar de las otras dos, esto se puede corroborar por medio del intervalo de confianza ya que el momio de elegir la carrera de administración de los estudiantes que tienen a su madre que se dedica al hogar se encuentre entre un valor chico de aproximadamente 2 o puede ser un valor grande de 6.5 veces más que el momio de elegir la carrera de administración con una madre que se dedica a trabajar.

2.7.3. Conclusiones

En la carrera de administración no existe diferencia con respecto a las dos primeras (nutrición e ingeniería), ya que se da en forma equivalente la asistencia tanto de hombres como de mujeres en las dos universidades, de las 45 observaciones que se obtuvieron de la Universidad Autónoma Metropolitana (UAM) de la carrera de administración, se tiene del sexo masculino el 66.7% y del sexo femenino el 33.3% mientras que en la Universidad Iberoamericana (UIA) de la carrera de administración se registro que el 44.4% son del sexo masculino y el

55.6% son del sexo femenino de las 72 observaciones registradas de la carrera de administración en la UIA. (*tabla 11, anexo I).* Para la variable universidad se tiene en el modelo que la probabilidad de elegir la carrera de administración en lugar de las otras dos con respecto a los estudiantes que ingresaron a la UAM es menor en comparación de los que eligieron la UIA de los que eligieron la carrera de administración en lugar de no elegirla.

En la elección de la carrera de administración se tuvo que de las 147 observaciones de los estudiantes que tienen en mente seguir con sus estudios al terminar la carrera el 49% ingreso a la carrera de administración, de las 67 observaciones que se encontraron de los estudiantes que piensan salir a buscar trabajo el 62.7% corresponde a la carrera de administración, *(Tabla 12, Anexo I).* Por el modelo logístico se obtuvo que la probabilidad de elección de la carrera de administración de los estudiantes que piensan salir a buscar trabajo es mayor en comparación de la probabilidad de los estudiantes que tienen en mente seguir estudiando y que escogieron la carrera de administración en lugar de las otras opciones.

Para la carrera de administración se tiene que de 143 observaciones de los estudiantes que tienen a su madre que se dedica a trabajar el 43.4% corresponde a los que ingresaron a la carrera de administración, mientras que de las 81 observaciones de los estudiantes que tienen a su madre que se dedica al hogar el 67.9% corresponde a los que ingresaron a la carrera de administración. (*Tabla 13, Anexo I)*, entonces la probabilidad de elección de la carrera de administración en lugar de no elegirla de los estudiantes que tienen a su madre que se dedica al hogar es mayor a la probabilidad de los estudiantes que tienen a su madre que se dedica a trabajar y que escogieron la carrera de administración en lugar de no elegirla.

2.8. Análisis de la variable respuesta Carrera (Nutrición, Ingeniería y Administración)

A continuación se presentan los resultados obtenidos para el caso tricotómico, es decir, cuando se tiene una variable respuesta con tres posibles valores, teniendo como finalidad explorar un poco más sobre el comportamiento de las variables explicativas que se están manejando para la toma de elección de carrera en el año de ingreso de 1996, siempre y cuando se cumplan las condiciones establecidas inicialmente.

Se sabe que son tres categorías para la variable respuesta: administración, ingeniería y nutrición, por lo tanto, se ajustaron 3 modelos con categorías de referencia distintas como se mostrará a continuación. A cada modelo se aplicó el estadístico de prueba de Pearson, considerando las variables explicativas que se determinaron en la *tabla E*.

El número de observaciones totales con las que se cuenta son 344, siendo el 53.2% de la carrera de administración, el 26.5% de la carrera de ingeniería y por último 20.3% de la carrera de nutrición.

Para la obtención del primer modelo se considera como categoría de referencia la carrera de Ingeniería para la variable respuesta:

$$Y = \begin{cases} y_0 = 0 & \textit{Carrera de ingeniería} \\ y_1 = 1 & \textit{Carrera de nutrición} \\ y_2 = 2 & \textit{Carrera de administración} \end{cases}$$

Parameter Estimates		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower	Upper
NUT.	Intercept	-9.335	3.43	7.408	1	0.006			
	PROMEDIO	0.805	0.393	4.199	1	0.04	2.236	1.036	4.827
	[UNIVERSI=UIA]	2.716	0.813	11.17	1	0.001	15.113	3.074	74.310
	[ENDCARR=TRAB]	-1.294	0.676	3.663	1	0.056	0.274	0.073	1.032
	[QCARRERA=SI]	-1.825	0.7	6 792	1	0.009	0.161	0.041	0.636
	[SEXO=FEM]	4.596	0.779	34.83	1	0	99.072	21.531	455.872
	[EDPADRES=SEC]	-2.774	1.011	7.537	1	0.006	0.062	0.009	0.452
ADM.	Intercept	0.244	2.147	0.013	1	0.91			
	[SEXO=FEM]	2.271	0.537	17.92	1	0	9.691	3.386	27.738
	[OCUPMAD=HOG]	0.921	0.434	4.514	1	0.034	2.512	1.074	5.877
a	This parameter is set to zero because it is redundant.								

Cuadro 16. Modelo ajustado para 10 variables explicativas teniendo como categoría de referencia de la variable respuesta la carrera de ingeniería.

Parameter Estimates		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower	Upper
ADM	Intercept	9.55	3.295	8.397	1	0.004			
	PROMEDIO	-0.8	0.326	6.502	1	0.011	0.435	0.230	0.825
	[QESTCARR=SI]	-1.4	0.619	4.976	1	0.026	0.251	0.075	0.846
	[EDPADRES=SEC]	1.68	0.84	4.004	1	0.045	5.365	1.035	27.812
	[SEXO=FEM]	2.33	0.637	13.321	1	0	10.223	2.934	35.625
	[UNIVERSI=UAM]	-2.4	0.685	12.103	1	0.001	0.092	0.024	0.353
ING	Intercept	7.13	3.854	3.42	1	0.064			
	PROMEDIO	-0.8	0.393	4.199	1	0.04	0.447	0.207	0.966
	[QESTCARR=SI]	-1.8	0.7	6.792	1	0.009	0.161	0.041	0.636
	[EDPADRES=SEC]	2.77	1.011	7.537	1	0.006	16.025	2.211	116.132
	[SEXO=FEM]	-0.6	0.779	34.827	1	0	99.072	21.531	455.872
	[UNIVERSI=UAM]	-2.7	0.813	11.167	1	0.001	0.066	0.013	0.325
a	This parameter is set to zero because it is redundant.								

Cuadro 17. Modelo ajustado de regresión logística con 10 variables explicativas teniendo como categoría de referencia la carrera de nutrición para la variable respuesta.

		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								lower	Upper
ING	Intercept	-2.549	2.524	1.020	1	0.312			
	[SEXO=FEM]	-2.271	0.543	17.900	1	0.000	0.103	0.036	0.295
	[MADTRAB=HOG]	-0.921	0.434	4.514	1	0.034	0.398	0.170	0.931
	[IMAFE=FEM]	1.013	0.574	3.118	1	0.077	2.753	1.005	8.474
NUT	Intercept	-9.812	3.319	8.737	1	0.003			
	PROMEDIO	0.832	0.326	6.502	1	0.011	2.297	1.212	4.354
	[ENDCARR=trab]	-1.428	0.567	6.344	1	0.012	0.240	0.079	0.729
	[EDPADRES=SEC]	-1.68	0.84	4.004	1	0.045	0.186	3.60E-02	0.966
	[SEXO=FEM]	2.325	0.637	13.321	1	0	10.223	2.934	35.625
	[UNIVERSI=UAM]	2.383	0.685	12.103	1	0.001	10.838	2.83	41.496
	[QCARRERA=si]	-1.381	0.619	4.976	1	0.026	0.251	0.075	0.846

Cuadro 18. Modelo de regresión logística con 10 variables explicativas tomando como Categoría de referencia la carrera de administración de la variable respuesta.

De los tres modelos propuestos se obtuvieron para cada uno entre 2 y 7 variables explicativas estadísticamente significativas, pero se debe proponer un modelo, la elección se va a realizar de acuerdo a la metodología utilizada en los análisis anteriores, esto es, se va escoger el modelo que tiene como categoría de referencia el mayor número de observaciones por lo tanto el modelo propuesto se encuentra en el *Cuadro 18*, teniendo como categoría de referencia la carrera de administración. Esto también ayuda para tener una mejor estimación de los coeficientes, ya que, se cuenta con una población similar tanto del sexo masculino como del femenino y no enfocar la modelación de las variables explicativas a una tendencia del tipo de sexo de los estudiantes.

Se proponen los siguientes modelos logísticos:

Para la ***carrera de ilngeniería:***

$$\ln\left(\frac{\hat{\pi}(Y=1)}{1-\hat{\pi}(Y=0)-\hat{\pi}(Y=2)}\right) = -2.549 - 2.271\,Sexo(Fem) - 0.921\,Ocupación\ madre\ (Hog) + 1.013\,IMAFE\ (Fem)$$

Para la ***carrera de nutrición:***

$$\ln\left(\frac{\hat{\pi}(Y=2)}{1-\hat{\pi}(Y=0)-\hat{\pi}(Y=1)}\right) = -9.812 + 2.325\,Sexo(femenino) + 0.832\,promedio - 1.428\,ter.carre(Trab) - 1.68\,Esc.padres(Sec) + 2.383\,Universidad(UAM) - 1.381\,quiso\,esta\,carr\,(Si)$$

2.8.1. *Porcentaje Correcto de clasificación*

Se tiene que el porcentaje correcto de clasificación de la variable respuesta de la estimación del Modelo de Regresión Logística es el siguiente:

Classification

Observed	Predicted			
	ing	nut	adm	Percent Correct
ing	25	2	36	39.7%
nut	1	39	15	70.9%
adm	22	12	97	74.0%
Overall Percentage	19.3%	21.3%	59.4%	64.7%

Cuadro 19. Tabla de Clasificación para el modelo elección de carrera teniendo como categoría de referencia la carrera de administración.

El porcentaje de clasificación se obtiene a partir de las observaciones estimadas correctamente entre el total de observaciones, por ejemplo, de 131 observaciones para la carrera de administración soló 97 observaciones son estimados correctamente, entonces el porcentaje de clasificación es $(97/171 = .74)(100) = 74\%$, mientras que el valor bajo se encuentra en la carrera de ingeniería, esto sucede por la distribución de las observaciones.

Existe otro indicador que ayuda a evaluar el ajuste del modelo que es el valor de pronóstico positivo (VPP), se calcula con base a los valores estimados correctamente del total de observaciones estimadas, quedando de la siguiente manera.

	VALOR ESTIMADO.				
VALOR OBSERVADO	ING	NUT	ADM	TOTAL	PRONÓSTICO ESTIMADO
ING	25	1	22	48	52%
NUT	2	39	12	53	73.50%
ADM	36	15	97	148	65.50%
TOTAL	63	55	131	249	

Cuadro 20. Valor pronóstico sobre los datos que se están clasificando como correctos.

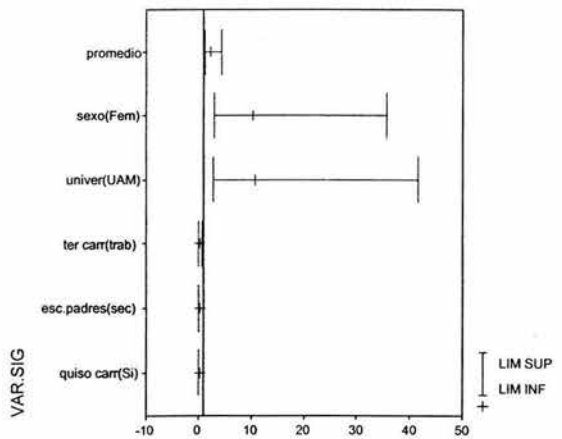
El *valor de pronóstico positivo* para el modelo logit con categoría de referencia la carrera de administración para la variable respuesta carrera, se tiene para el modelo de la carrera de nutrición el 73.50% con una *prevalencia* de $(55/249) = 0.22$. Para la carrera de ingeniería se tiene como *valor de pronóstico positivo* el 52% con una *prevalencia* de 0.25, lo anterior nos indica que en más de 2 veces es mejor el pronóstico de los modelo propuestos que sin ellos.

2.8.2. *INTERPRETACIÓN DE LOS COEFICIENTES*

El intervalo de confianza para cada modelo es:

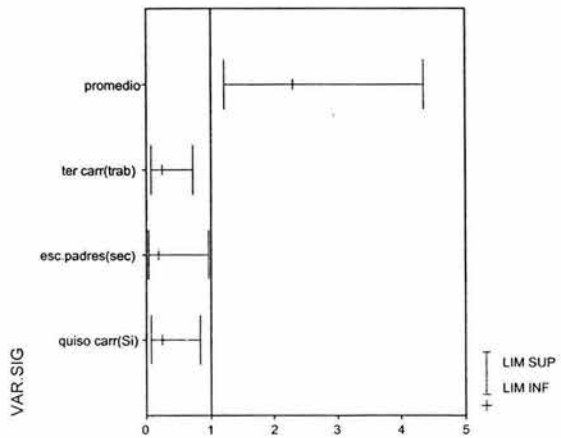
		B	Exp(B)	95% Confidence Interval for Exp(B)	
				lower	Upper
NUT	Intercept	-9.812			
	PROMEDIO	0.832	2.297	1.212	4.354
	[ENDCARR=trab]	-1.428	0.240	0.079	0.729
	[EDPADRES=SEC]	-1.68	0.186	0.036	0.966
	[SEXO=FEM]	2.325	10.223	2.934	35.625
	[UNIVERSI=UAM]	2.383	10.838	2.83	41.496
	[QCARRERA=si]	-1.381	0.251	0.075	0.846

Cuadro 21. Resultado de los valores del intervalo de confianza para los cocientes de momios.



Gráfica 16. Intervalo de confianza del 95%, de la carrera de nutrición.

La siguiente gráfica se presenta para observar las variables promedio, al terminar la carrera, escolaridad ambos padres y siempre quiso esta carrera de la gráfica 16.



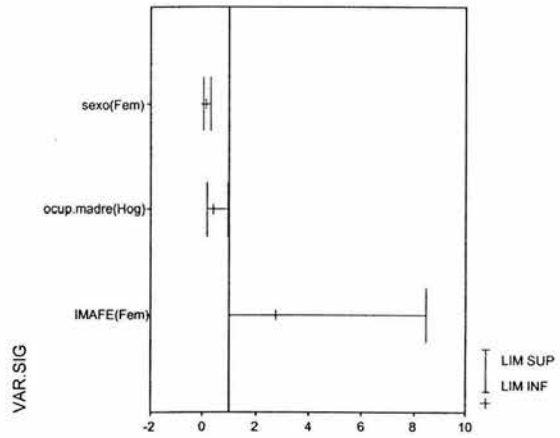
Gráfica 16a. Intervalo de confianza del 95%, de la carrera de nutrición de las variables promedio, al termino de la carrera, escolaridad de ambos padres y quiso esta carrera.

Para la carrera de nutrición se obtuvieron 6 variables estadísticamente significativas en donde ningún intervalo de confianza contiene el valor de uno por lo tanto se afirma que se tiene un efecto significativo en la probabilidad de riesgo. La interpretación del momio también se puede obtener a partir del intervalo de confianza de los cocientes de momios, por ejemplo se tiene el momio de la variable universidad, esto es el momio de elegir la carrera de nutrición en lugar de elegir la carrera administración de los estudiantes que ingresaron a la Universidad Autónoma Metropolitana puede ser tan pequeño como aprox. 3 o tan grande como aprox. 41 veces más que el momio de elegir la carrera de nutrición en lugar de la carrera de administración de los estudiantes que ingresaron a la Universidad Iberoamericana, la interpretación que se acaba de señalar se aplica de la misma forma para las demás variables.

El intervalo de confianza del modelo propuesto para la carrera de ingeniería queda:

		B	Exp(B)	95% Confidence Interval for Exp(B)	
				lower	Upper
ING	Intercept	-2.549			
	[SEXO=FEM]	-2.271	0.103	0.036	0.295
	[MADTRAB=HOG]	-0.921	0.398	0.170	0.931
	[IMAFE=FEM]	1.013	2.753	1.005	8.474

Cuadro 22. Resultado de los valores del intervalo de confianza para los cocientes de momios de la carrera de ingeniería.



Gráfica 17. Intervalo de confianza del 95%, de la carrera de ingeniería.

En las 3 variables estadísticamente significativas del modelo propuesto para la carrera de ingeniería, no contienen el valor de uno, por lo que a continuación se darán las conclusiones correspondientes para cada modelo propuesto.

2.8.3. *Conclusiones*

Se aplicó la regresión logística para una variable respuesta tricotómica, estableciendo como categoría de referencia la carrera de administración por el número de observaciones y la población estudiantil que la conforma. Los resultados obtenidos fueron de ayuda ya que se respaldó lo que se tenía en forma individual para cada una de las carreras. Solamente se identificó una nueva variable estadísticamente significativa distinta a lo que ya se tenía en los análisis anteriores.

El análisis realizado para la variable respuesta elección de carrera se identifica para el modelo de nutrición una nueva variable que anteriormente no había salido significativa, la variable actividad que van a realizar los estudiantes al terminar la carrera, por lo tanto la probabilidad de elección de la carrera de nutrición en lugar de la carrera de administración de los estudiantes que piensan salir a buscar trabajo es menor en comparación de la probabilidad de los estudiantes que piensan seguir con sus estudios y eligieron la carrera de nutrición en lugar de la carrera de administración.

Para el modelo propuesto de la carrera de ingeniería también se obtuvo una nueva variable como significativa siendo el índice de masculinidad y feminidad (IMAFE), lo que el momio indica es que la probabilidad de elección de la carrera de ingeniería en lugar de la carrera de administración de los estudiantes que tuvieron un valor alto en la personalidad con características femeninas fue mayor en comparación de los estudiantes que tuvieron valores altos en la personalidad con características femeninas como masculinas (andróginas) y que eligieron la carrera de ingeniería en lugar de la carrera de administración.

En este capitulo se ha presentado el ajuste de varios modelos. Primero, tomando cada categoría de la variable respuesta y comparándola con las otras

dos categorías colapsadas en una sola (Nut. contra Ing. y Adm; Ing. contra Nut. y Adm; y por último Adm. contra Nut. e Ing.). Segundo, al considerar un modelo con variable respuesta tricotómica, es decir, con tres categorías, se tomó una categoría y se comparó solamente con la tercera (Nut. contra Adm.) y también la segunda con la tercera (Ing. contra Adm).

Esto muestra que la interpretación de los coeficientes en la regresión logística es relativa tanto a la categoría de referencia de la variable explicativa correspondiente como a la categoría o categorías de referencia de la variable respuesta. Por esa razón se tuvo cuidado en la presentación del modelo final, considerando como categoría de referencia el de mayor número de observaciones, también se analizó que los resultados obtenidos aportaran nueva información o que respaldarán las conclusiones anteriores y fuera de ayuda para la investigación.

2.9. Análisis de la Variable Respuesta Universidad (Universidad Autónoma Metropolitana y Universidad Iberoamericana)

Para el análisis de la variable respuesta universidad se tienen 2 categorías:

1) Ingreso a la Universidad Autónoma Metropolitana (UAM).
2) Ingreso a la Universidad Iberoamericana (UIA).

Lo que se busca establecer al trabajar la variable respuesta universidad, es conocer si el factor elección de carrera al ingresar a nivel licenciatura en el año de 1996, podría haber tenido una influencia negativa o positiva para la variable respuesta; porque se pensaría que en muchos de los casos cuando se elige una carrera que se imparte en una Universidad Privada (UIA), es por la razón de tener una economía estable para abarcar los costos de la educación, por lo tanto, la tarea fue el de especificar que factores fueron los que intervinieron a la hora de escoger carrera y saber si están relacionados con la variable universidad.

El análisis se esbozará de la misma forma que los casos anteriores, se dará a conocer las variables explicativas que resultaron ser significativas para el modelo final. A continuación se presentan las modificaciones realizadas a algunas variables para la obtención de resultados correctos del modelo.

1. Variable respuesta ***universidad:***

VARIABLE RESPUESTA	CATEGORÍAS	CODIFICACIÓN
4.-Universidad	Universidad Privada (UIA)	0
	Universidad Pública (UAM)	1

2. Variables explicativas son: *promedio, sexo, termino de carrera, quiso esta carrera, país, ocupación de la madre, IMAFE, comprensión de la madre y escolaridad ambos padres*, que se encuentran determinados en la *tabla E* de este capítulo.

3. Se agregó la variable ***carrera***:

Etiqueta	Categoría	Codificación
Carrera	Administración	0
	Ingeniería	1
	Nutrición	2

Después de dar una breve explicación de las modificaciones realizadas para el análisis de la variable respuesta universidad se dará la presentación del modelo logístico.

Los resultados obtenidos ha partir de la modelación de las 10 variables explicativas son los siguientes:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	PROMEDIO	-.381	.269	2.001	1	.157	.683	.403	1.158
	CARRERA(ADM)			11.529	2	.003			
	CARRERA(ING)	-.276	.534	.267	1	.605	.759	.266	2.162
	CARRERA(NUT)	2.130	.642	11.005	1	.001	8.414	2.391	29.616
	SEXO(FEM)	-1.283	.630	4.145	1	.042	.277	.081	.953
	TERCARRE(EST)			3.120	2	.210			
	TERCARRE(TRAB)	.564	.467	1.455	1	.228	1.757	.703	4.391
	TERCARRE(OTRA)	-1.046	.954	1.202	1	.273	.351	.054	2.279
	QESTCARR(SI)	.912	.437	4.351	1	.037	2.489	1.057	5.865
	PAIS(NO)	.371	.894	.172	1	.678	1.449	.251	8.367
	OCUPMADRE(HOG)	.210	.459	.210	1	.647	1.234	.502	3.036
	IMAFE(AND)			1.958	3	.581			
	IMAFE(IND)	.340	.578	.346	1	.557	1.404	.453	4.357
	IMAFE(FEM)	.491	.559	.773	1	.379	1.635	.546	4.889
	IMAFE(MASC)	.755	.556	1.848	1	.174	2.128	.716	6.324
	COMMADRE(SI)	-.652	.877	.552	1	.458	.521	.093	2.909
	EDPADRES(LIC)			16.031	3	.001			
	EDPADRES(SEC)	11.814	23.292	.257	1	.612	135107	.000	9.E+24
	EDPADRES(PREP)	1.717	.483	12.616	1	.000	5.567	2.159	14.357
	EDPADRES(MoD)	-.468	.565	.688	1	.407	.626	.207	1.894
	Constant	1.045	2.228	.220	1	.639	2.842		

a. Variable(s) entered on step 1: PROMEDIO, CARRERA, SEXO, TERCARRE, QESTCARR, PAIS, MADTRAB, IMAFE, COMMADRE, EDPADRES.

Cuadro 23. Resultado del modelo ajustado de las 10 variables explicativas para la variable respuesta universidad.

Por el método de selección automática backward.

Factor (i)	Etiqueta	Variable	Coeficiente regresión	Error estándar	$\left(\frac{\beta_i}{SE(\beta_i)}\right)^2$	
Sexo del estudiante	*Sexo*	Maculino=0 * ***femenino=1***	*-1.353*	*0.565*	*5.735*	$P<0.02$
Siempre quiso la	*Quiso la carr(Si)*	No=0 * ***Sí=1***	*0.865*	*0.41*	*4.453*	$P<0.04$
Escolariada ambos padres	*Esc.padres*	Lic.=0 * ***Sec=1*** Prep=2 ** MoD=3 **	*1.609*	*0.439*	*13.442*	$P<0.001$
Tipo de Carrer:	*Carrera (Nut.)*	Administración=0 * Ingeniería=1 ** ***Nutrición=2***	*1.833*	0.59	9.649	$P<0.003$

Cuadro 24. Coeficientes estimados de la regresión logística para un modelo potencialmente candidato, categoría de referencia con un (*) y categorías no significativas para el modelo (**).

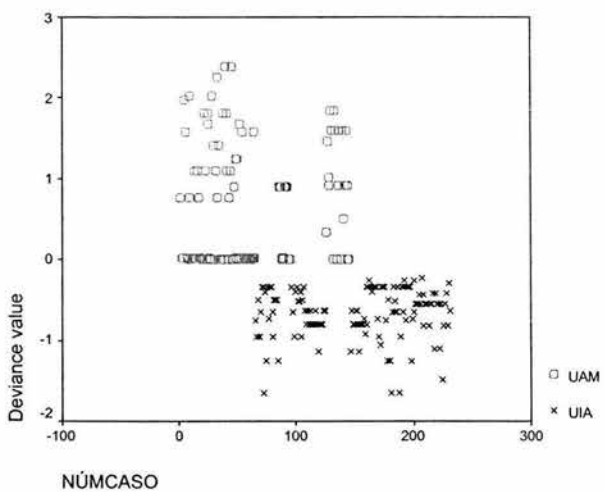
El modelo logístico queda determinado como:

$$\ln\left(\frac{\hat{\pi}(Y=1)}{1-\pi(Y=1)}\right) = -1.417 - 1.353 Sexo(fem) + 0.865 Quiso\,la\,carr\,(Si) + 1.609 Esc.padres\,(\mathrm{Pr}ep) + 1.833\,Carrerade\,Nutrición$$

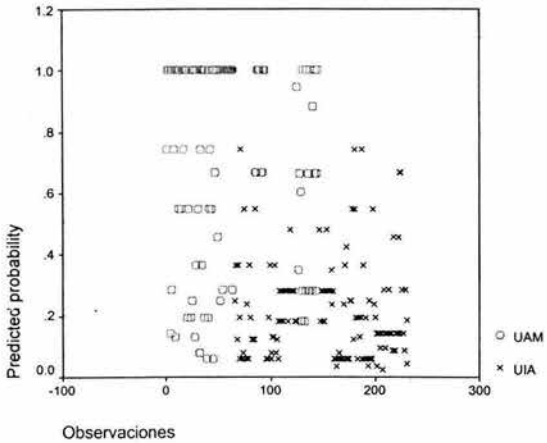
2.9.1. PRUEBA DE SIGNIFICANCIA DEL MODELO

ANÁLISIS DE RESIDUALES

Por medio de la aplicación de análisis de residuales, se corrobora que el ajuste de las variables que resultaron ser significativas son correctas, como se ve ha continuación.



Gráfica de dispersión 18.Residuales estándarizados vs número de observación.



Gráfica de dispersión 19. Probabilidad estimada vs número de observación.

Por las 2 gráficas anteriores y por las gráficas que se encuentran; en el *Anexo I, gráficas 6-9*, se obtiene que el ajuste del modelo es bueno para la obtención de resultados. La prueba de Hosmer y Lemeshow (*tabla 15, Anexo I)* comprobamos que el ajuste del modelo es bueno ya que los valores observados son similares a los esperados.

2.9.2. *Interpretación de los Coeficientes*

El intervalo de confianza del 95% de las categorías estadísticamente significativas quedan de la siguiente manera.

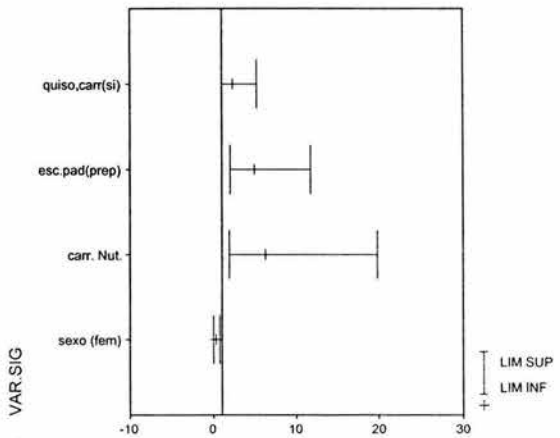
Variables in the Equation

		B	Exp(B)	95.0% C.I.for EXP(B)	
				Lower	Upper
Step 1[a]	CARRERA(NUT)	1.833	6.251	1.967	19.870
	SEXO(FEM)	-1.353	.259	.085	.782
	QESTCARR(SI)	.865	2.376	1.064	5.308
	EDPADRES(PREP)	1.609	4.999	2.115	11.817
	Constant	-1.417	.243		

a. Variable(s) entered on step 1: CARRERA, SEXO, QESTCARR, EDPADRES.

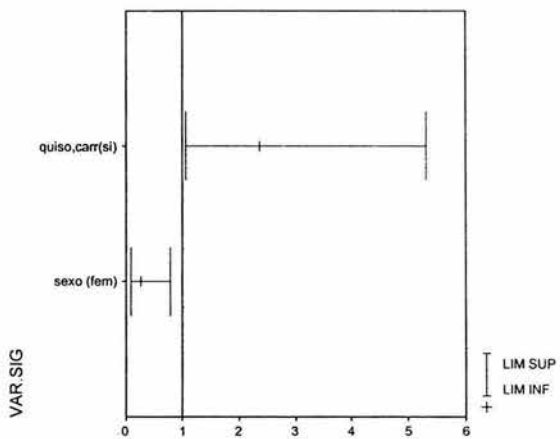
Cuadro 25. Valores del intervalo de confianza para los cocientes de momios.

Por medio de gráfica se tiene:



GRÁFICA 20. Intervalo de confianza del 95% de la $exp(\hat{\beta})$.

Se tuvo que recurrió a otra gráfica para la mejor apreciación del intervalo de confianza de las variables *sexo y quiso esta carrera.*



GRÁFICA 20a. Intervalo de confianza del 95% de la $exp(\hat{\beta})$ *de 2 categorías significativas (sexo y siempre quiso esta carrera).*

Los intervalos de confianza estimados de las categorías estadísticamente significativas no incluyen el valor de uno, por lo tanto, se puede extraer las conclusiones correspondientes de los coeficientes. Por ejemplo, se tiene la variable explicativa escolaridad de ambos padres, lo que sugiere el momio de la elección de la UAM para los alumnos que tienen a sus padres con escolaridad al menos un padre con preparatoria puede ser pequeño como aprox. 2 o grande como 11 veces más que el momio de ingresar la UAM cuando tienen a sus padres con al menos uno con licenciatura.

2.9.3. CONCLUSIONES

Después de ver en forma individual cada una de las carreras y al obtener sus categorías significativas se dio la tarea de encontrar la relación existente entre el tipo de universidad con respecto a la elección de carrera y efectivamente se encontró una relación entre la categoría carrera de nutrición y la variable universidad, el momio indica que la probabilidad de ingresar a la universidad Autónoma Metropolitana en lugar de la Universidad Iberoamericana de los estudiantes que eligieron la carrera de nutrición fue mayor en comparación a la probabilidad de los estudiantes que prefirieron la carrera de administración y que ingresaron a la UAM en lugar de la UIA. De las 96 observaciones registradas en la UAM se tiene que el 20.8% corresponde a la carrera de nutrición. (*tabla 16, anexo I*)

Para el momio de la variable siempre quiso esta carrera, se tiene que la probabilidad de elección de la UAM en lugar de elegir la UIA de los estudiantes que ya sabían que carrera elegir al salir de preparatoria fue mayor a la probabilidad de los estudiantes que no habían considerado que carrera elegir y que a su vez ingresaron a la UAM en lugar de la UIA. Se tuvieron 51

observaciones de la carrera de nutrición y el 88.2% corresponde a los estudiantes que no habían considerado que carrera elegir y el 11.8% son los estudiantes que ya tenían en mente la carrera de su preferencia. *(tabla 17, anexo I)*

Por último la probabilidad de elección de la UAM en lugar de elegir la UIA de los estudiantes del sexo femenino fue menor en comparación a la probabilidad de los estudiantes del sexo masculino y que ingresaron a la UAM en lugar de la UIA. De la UAM se tuvo 96 observaciones, el 65.6% fue de estudiantes del sexo masculino y el 34.4% del sexo femenino, en cambio en la UIA se tuvo 135 observaciones con el 43.7% del sexo masculino y 56.3% del sexo femenino. *(tabla 17, anexo I)*.

2.10. Conclusiones Generales

Los resultados presentados en los incisos anteriores no pueden ser generalizados para cualquier año de ingreso a nivel licenciatura, solamente abarca el periodo de ingreso de 1996, uno de los motivos es que la muestra es relativamente pequeña sin llevarse a cabo un muestreo estrictamente aleatorio para la aplicación del mismo.

En la licenciatura de nutrición se tiene que los estudiantes con promedio alto y que eligieron la Universidad Autónoma Metropolitana aumenta la probabilidad de elección de la carrera de nutrición, en cambio para la elección de la carrera de administración va ha disminuir la probabilidad de elección de la UAM, aquí entraría el análisis que se hizo sobre la variable respuesta universidad, la carrera de nutrición aumenta la probabilidad de ingreso a la UAM, algo similar ocurre con aquellas personas que siempre habían tenido en mente que era lo que quería estudiar; también se obtuvo la variable de que al menos un padre con preparatoria va aumentar la probabilidad de ingresar a la Universidad Autónoma Metropolitana en comparación de aquellos alumnos que tienen al menos un padre con licenciatura. Ya que al tener un grado de estudio de licenciatura se pensaría que se tiene una mayor oportunidad de encontrar un empleo con una renumeración económicamente alta, y por último se tiene la categoría sexo del estudiante ya que el sexo femenino va a disminuir la probabilidad de elección de la UAM en comparación del sexo masculino.

Existe otro hecho tal como se observa en la realidad, mientras sea mujer hay una mayor probabilidad de escoger la carrera de nutrición sucediendo algo distinto para la carrera de ingeniería ya que es preferida por el sexo masculino.

Otras variables que aparecieron en la probabilidad de no elección de la carrera de nutrición, es que si los estudiantes ya tenían en mente que carrera

escoger antes de salir de preparatoria, sucediendo lo contrario para el ingreso a la Universidad Autónoma Metropolitana y la escolaridad de ambos padres con secundaria ocurriendo algo distinto para la carrera de ingeniería ya que en esta aumenta la probabilidad de elección,

Pasando a los resultados de la licenciatura de ingeniería, que a diferencia de la primera este tipo de carrera es preferida por el sexo masculino al igual que los estudiantes que tienen a su madre que se dedica a trabajar, se tiene una discrepancia con la carrera de administración ya que los alumnos que la escogieron tienen a su madre que se dedica al hogar.

Por último se tiene la licenciatura de administración que aumenta la probabilidad de elección de la carrera para aquellos estudiantes que tienen a su madre que se dedica al hogar y los que tienen en mente salir a buscar trabajo sucediendo lo contrario con respecto a la carrera de nutrición ya que en ésta se tiene la idea de continuar con los estudios.

2.11. Comentarios Finales

Al tener la muestra de elección de carrera que es un subconjunto de la población estudiantil tanto de la Universidad Iberoamericana como de la Universidad Autónoma Metropolitana, lo que se busca obtener son resultados concretos que expliquen los hechos que acontecieron en la toma de elección de carrera de 1996, dicho proceso requiriere antes que nada que se haga una exploración de los datos que se utilizaron para el análisis por medio de las frecuencias que describen cualquier tipo de variables, esta exploración nos ayuda para no tener categorías dentro del análisis que contengan un número de datos relativamente bajos en comparación de otra categoría que tenga el mayor número de datos, esto es, si estamos hablando de variables con más de dos categorías. Dicha exploración es una herramienta necesaria para determinar que categoría se va a tomar como referencia, ya que se podría tomar a cualquiera de las que tenemos dentro de la variable, pero es importante notar que se busca establecer una interpretación adecuada de los resultados obtenidos y así evitar los errores que se podrían cometer, por ejemplo: el tomar una categoría que no sea del mayor interés para el estudio o tomar una categoría de referencia que tenga un número muy pequeño de observaciones en comparación de las demás categorías, ya que esto podría llevar a una mala estimación de los coeficientes o provocaría un intervalo de confianza demasiado grande.

Después de establecer como son las variables explicativas, tenemos que pasar al caso de la variable respuesta, ya que ésta se trata del caso tricotómico, como es más fácil manejar una variable respuesta con dos categorías para considerar la interpretación de los datos, primero tomamos como referencia una de las tres mientras que a la otras dos las colapsamos para tener finalmente una variable de respuesta del caso dicotómico y posteriormente pasar a lo que se refiere la variable respuesta del caso tricotómico con el previo conocimiento sobre

las categorías que resultaron tener una significancia estadísticamente importante para el modelo.

Para cada uno de los modelos se exploró el uso de interacciones como: escolaridad de la madre con ocupación de la madre, número de autos con tipo de carrera, escolaridad del padre con tipo de universidad, escolaridad de la madre con tipo de universidad, sexo con la ocupación de la madre, IMAFE con sexo y finalmente número de autos con educación de la madre, considerándose para distintos modelos, resultando ser no significativas para unos y en otras ocasiones resultaron ser las únicas variables significativas, por ejemplo, se consideró la variable IMAFE con sexo para la construcción del modelo logístico de la carrera de ingeniería y solamente se obtuvo como variable estadísticamente significativa IMAFE y la nueva variable significativa de IMAFE con sexo, sin aportar más información. Para la utilización de las interacciones se debe tener cuidado ya que para la construcción de éstas es conveniente estar completamente seguros de que dos variables que son consideradas explicativas se encuentran altamente correlaciones y no cometer errores en la modelación de las variables, por esa razón se tuvo una preferencia por los modelo propuestos en los puntos anteriores.

Lo único que hace falta mencionar es que para algunas variables explicativas con más de dos categorías se colapsaron para tener solamente dos opciones y así tener una mejor interpretación de los datos sobre los hechos que acontecieron en la variable respuesta como ocurrió en el modelo universidad, sin ocurrir algo similar en los demás modelos ya que en algunos de ellos se llego a perder información, por esa razón se utilizaron las variables explicativas determinadas en la *tabla E*, de este capítulo.

El proceso llevado a cabo fue con la finalidad de conocer en forma individual para cada carrera cuales son los posibles factores que auxilian a identificar algunos factores de riesgo asociados a la elección de una carrera, ya

que la población estudiantil con la que cuenta cada carrera varía según las condiciones en las que se desenvuelvan los miembros de cada familia y sobre todo en el sexo que tengan. De igual manera se debe considerar el nivel socioeconómico de cada individuo ya que este representa un factor diferencial que pesa sobre las decisiones de elección de carrera, el cual tiene una estrecha relación con la variable universidad a la que están acudiendo, este factor económico también influye en los proyectos que tienen en mente o sobre lo que piensan realizar los estudiantes al terminar la carrera. No se debe olvidar que el índice de masculinidad y feminidad se aplicó con la finalidad de poder identificar dentro de las tres carreras un patrón en común o el de identificar las característica de comportamiento que tienen los estudiantes de cada área.

CAPÍTULO 3

OBTENCIÓN DE UN AJUSTE REPRESENTATIVO DE LA CALIDAD DE VIDA DE LOS PACIENTES QUE EGRESAN DE LA UNIDAD DE TERAPIA INTENSIVA

CAPÍTULO 3

OBTENCIÓN DE UN AJUSTE REPRESENTATIVO DE LA CALIDAD DE VIDA DE LOS PACIENTES QUE EGRESAN DE LA UNIDAD DE TERAPIA INTENSIVA

Este capítulo abarca la aplicación del análisis de regresión logística a un diferente campo de estudio con respecto a lo que se vio en el capítulo anterior, la discrepancia que se tiene en la nueva base de datos es que ahora se trabajo sobre el sector Médico.

El cuestionario se aplicó en el año 2002 al 2004 a pacientes hospitalizados en las instituciones: Centro Médico Nacional Siglo XXI y Centro Médico Nacional la Raza, obteniendo un tamaño de muestra de 237 observaciones.

El objetivo primordial es identificar los posibles factores que permiten pronosticar el nivel de estado vital (recuperación aceptable o limitación, se está hablando de la calidad de vida a los 3 meses del egreso hospitalario), así como de la obtención de un modelo parsimonioso de los datos a utilizar.

Existen diferentes factores que influyen en que un paciente tenga una buena o mala calidad de vida cuando hayan salido de la Unidad de Terapia Intensiva (UTI), como son los factores demográficos, terapéuticos, etc. La literatura nos indica que el riesgo de tener una mala calidad de vida va relacionada con aquellos pacientes que tienen una edad avanzada (grande), que tenga un índice de morbilidad alto y que haya ingresado con una mala calidad de vida a los 2 meses previos de la hospitalización a la Unidad de Terapia Intensiva, dichos

factores al igual que otros a considerar van hacer utilizados para la elaboración y obtención de un modelo que determine el riesgo de tener una mala calidad de vida a los 3 meses del egreso hospitalario.

A continuación se especifican las variables explicativas y la variable respuesta para que posteriormente sean modeladas por medio de la aplicación del Análisis de Regresión Logística que resulta ser útil para la identificación de los factores de riesgo con la utilización del paquete estadístico SPSS.

3.1. DATOS

La información fue proporcionada por el Médico Cirujano Luis David Sánchez Velásquez estudiante del postgrado en medicina de la UNAM con fines estrictamente académicos. Además de contar con su asesoramiento.

El levantamiento de la encuesta se llevó a cabo entre los años 2002 al 2004, en las Unidades de Terapia Intensiva de dos hospitales distintos, realizándose un estudio observacional, aplicado a los enfermos hospitalizados en la Unidad de Terapia Intensiva, sin distinción de género (hombre o mujer), con una edad igual o mayor de 18 años y que viviera en el Distrito Federal o área Metropolitana.

Se tuvo cuidado de eliminar las observaciones de los pacientes que no contestaron por completo los cuestionarios de calidad de vida, enfermos o familiares que no aceptaron tener una participación en el estudio y los enfermos con diagnósticos de sepsis procedentes de otro hospital. También hacemos notar que se tomará en cuenta solo a los pacientes sobrevivientes, estos constituyen aproximadamente un 50% del total de los que se registraron y que fueron sometidos a la Unidad de Terapia Intensiva.

El número de observaciones obtenidas en los dos hospitales es como sigue.

	Observaciones	Porcentaje (%)
Siglo XXI	153	64.6
La Raza	84	35.4
Total	237	100.0

Cuadro A. Distribución de las 237 observaciones según Hospital.

3.2. Variable Respuesta

Variable respuesta: Calidad de Vida de los pacientes a los 3 meses del alta hospitalaria *(CV)*, dicotómica. Este factor nos indica en que condiciones se encuentra el paciente para realizar sus actividades diarias, es decir, si puede realizarlas el solo o necesita la ayuda de uno de sus familiares.

Etiqueta	Variable Respuesta (Descripción)	Categorías	Codificación
CV	Calidad de Vida	Buena	0
	a los 3 meses de alta	Mala	1

Cuadro B. Clasificación de las categorías de la variable respuesta2.

3.3. Variables Explicativas

Las características consideradas para evaluar el riesgo del nivel de Estado Vital (Calidad de Vida) son las siguientes:

1. *Demográficas*: género, edad, escolaridad (en años), estado civil, estado laboral, servicio, CV1 (estado vital a los 2 meses previos a la hospitalización), motivos de egreso de la UTI y por último el factor hospital.

2. *Clínicas:* Comorbilidad (calificación de Charlson), Falla orgánica (calificación de bruselas), escala de gravedad de la enfermedad (APACHE II), calificación fisiológica aguda (APS) y la calificación de coma (Glasgow).
3. *Paraclínicas:* química sanguínea y cultivos.
4. *Terapéuticas*: asistencia mecánica ventilatoria (VM), insulina, sedación, relajación, nutrición parenteral total (nutrición artificial) y traqueotomía.

Se tiene como base inicial más de 50 variables explicativas a considerar (*Cuadro C*), como el número es grande es necesario llevar a cabo una selección de las variables, considerando solamente a las que tienen una mayor relevancia e importancia para el estudio, esta selección se realizó por medio de la información documentada al igual que la ayuda de una persona debidamente capacitada e informada sobre el caso, ya que tiene el conocimiento teórico y práctico de las Unidades de Terapia Intensiva.

Antes de hospitalizarse		
Etiqueta	**Descripción**	**Valores**
Sexo	Sexo.	1=Mujer; 2=Hombre
Edad	Edad.	Continua
Escola	Escolaridad.	Continua
Edo.Civil.	Estado civil.	1=Solo; 2=Acompañado
Edo.Lab.	Estado laboral.	1=Trabaja; 2=No trabaja
Enf.pulmor	Enfermedad pulmonar.	0=No; 1=Sí
Diabetes	Diabetes mellitus.	0=No; 1=Sí
Enf.renal	Enfermedad renal.	0=No; 1=Sí
Charlson	Charlson (Número de enfermedades graves).	0=Ninguna; 1=1 enferme 2=2 ó más enfermedade
act.fisiol.	Actividad fisiológica básica 2 meses previos a la hospitalización.	Continua
act.diaria	Actividad diaria normal 2 meses previos a la hospitalización.	Continua
ee1	Estado emocional 2 meses previos a la hospitalización.	
cv1	Calidad de vida 2 meses previos a la hospitalización.	Continua
cv1codif	Calidad de vida 2 meses previos a la hospitalización codificada.	1=Buena; 2=Mala
ee1	Estado emocional 2 meses previos a la hospitalización.	Continua
cv1	Calidad de vida 2 meses previos a la hospitalización.	Continua
cv1codif	Calidad de vida 2 meses previos a la hospitalización codificada.	1=Buena; 2=Mala

	Durante la hospitalización, antes de ingresar a la terapia intensiva:	
Hospit	Hospital	1=Siglo XXI; 2=La Raza
Procede	Procedencia	1=Urgen.; 2=Piso; 3=Quiróf
Servicio	Servicio	1=Medicina; 2=Cirugía
Qxurgent	Si es cirugía, si fue de urgencia.	0=No urgente; 1=Urgente
estpre	Estancia hospitalaria previa al ingreso a la terapia.	Continua
	Durante la estancia en la terapia intensiva	
Apacheii	APACHE II (Calificación de la gravedad de la enfermedad).	Continua
Mortpre	Mortalidad predicha (Calculada por el modelo APACHE II).	Continua (expresada en %)
Aps1	Calificación fisiológica aguda en el 1er.día de estancia en la terapia.	Continua
Sepsis	Sepsis grave.	0=No; 1=Sí
Díasg	Día que se presentó sepsis grave al ingreso de terapia.	Continua
Sitioinf	Sitio de infección.	0=Ninguno; 1=Pulmón; 2=Otros
Bicho	Microorganismo causal de la infección.	0=Ninguno;1=BGN;2=CGP;3=Otro
Adquirid	Lugar de la adquisición de la infección.	0=Ninguno;1=Comunidad;2=Hospi
Aps2	Calificación fisiológica aguda en el día de la infección.	0=No; 1=Sí
Choqueco	Presencia de estado de choque.	0=No; 1=Sí
Díasch	Días en estado de choque.	Continua
Fneurol	Presencia de falla neurológica.	0=No; 1=Sí
Díasfneu	Días en falla neurológica.	Continua
Fhematol	Presencia de falla hematológica.	0=No; 1=Sí
Díasfhem	Días en falla hematológica.	Continua
Fendócr	Presencia de falla endócrina.	0=No; 1=Sí
Díasfend	Días en falla endócrina.	Continua
Frenal	Presencia de falla renal.	0=No; 1=Sí
Díasfren	Días en falla renal.	Continua
Fhepátic	Presencia de falla hepática.	0=No; 1=Sí
Díasfhep	Días en falla hepática.	Continua
Frespirc	Presencia de falla respiratoria.	0=No; 1=Sí
Díasfres	Días en falla respiratoria.	Continua
Br1	Bruselas 1 (Calificación total de falla orgánica el día 1 en estancia en TI).	Continua
Br2	Bruselas 2.	Continua
Br3	Bruselas 3.	Continua
Br4	Bruselas 4.	Continua
Br5	Bruselas 5.	Continua
Brus2	Bruselas del día de adquisición de la infección.	0=No; 1=Sí
	Uso de dobutamina.	0=No; 1=Sí
	Uso de norepinefrina.	0=No; 1=Sí
	Uso de sedante.	0=No; 1=Sí
	Días de uso de sedante.	Continua
	Uso de relajante.	0=No; 1=Sí
	Días de uso de relajante.	Continua
	Número de antibióticos usados.	
	Uso de cirugía.	0=No; 1=Sí
	Número de cirugías realizadas.	Continua
	Realización de traqueostomía.	0=No; 1=Sí
	Realización de gastrostomía.	0=No; 1=Sí

	Requerimiento de reanimación cardiopulmonar.	Continua
	Presión arterial sistólica más baja del primer día de estancia e	Continua
	Glasgow más bajo del primer día de estancia en la terapia.	Continua
	Balance hídrico del primer día de estancia en la terapia.	Continua
	Plaquetas más bajas del primer día de estancia en la terapia.	Continua
	Glucosa más alta del primer día de estancia en la terapia.	Continua
	Creatinina más alta del primer día de estancia en la terapia.	Continua
	Estancia en la terapia.	Continua
	Motivo de egreso de la terapia.	1=Mejoría; 5=Defunción
	Posterior al egreso de la terapia, aún hospitalizado	
Estposuti	Estancia hospitalaria posterior al egreso de la terapia.	Continua
Esthosp	Estancia hospitalaria total.	1=Mejoría; 5=Defunción
altah	Motivo de egreso del hospital.	Continua
	A los 3 meses del alta hospitalaria	
Edo.Civil.2	Estado civil a los 3 meses del egreso hospitalario.	
Edo.Lab.2	Estado laboral a los 3 meses del egreso hospitalario.	
Act.Fis.2	Actividad fisiológica básica a los 3 meses del alta del hospital.	
Act.Dia.2	Actividad diaria normal a los 3 meses del alta del hospital.	
Edo.Emocio.	Estado emocional a los 3 meses del alta del hospital.	

Cuadro C.-Descripción de las variables disponibles en la primera etapa del análisis.

A partir del *Cuadro C* se observan variables continuas y en otros casos se tienen categorías de 2 a 5 valores aproximadamente, por esa razón en la primera etapa fue necesario realizar un análisis exploratorio de la información; se revisaron los casos válidos, se obtuvieron correlaciones lineales, se colapsaron categorías, se aplicaron regresiones logísticas y por último se consideraron las variables de mayor interés para el estudio.

El cuadro que a continuación se muestra fue el resultado de muchas pruebas para la obtención de un modelo que represente la relación que existe entre un grupo de variables explicativas con respecto a una variable dicotómica, teniendo como finalidad conseguir un modelo que se ajuste a los objetivos establecidos inicialmente.

Etiqueta	Descripción	Valores
	Antes de Hospitalizarse	
Hospital	Hospital.	0=Siglo XXI; 1=La Raza
Sexo	Sexo.	0=Mujer; 1=Hombre
Edad	Edad.	Continua
Trabajo	Estado laboral.	0=No trabaja; 1=Sí trabaja
Enf.pulmona	Enfermedad pulmonar.	0=Sí; 1=No
Diabetes	Diabetes mellitus.	0=No; 1=Sí
Cáncer	Cáncer codificado.	0=Si; 1=No
CV 1	Calidad de vida 2 meses previos a la hospitalización	0=Mala; 1=Buena
	Durante la Hospitalización, antes de Ingresar a la UTI	
Hospital	Hospital.	0=Siglo XXI; 1=La Raza
Servicio	Servicio.	0=Medicina; 1=Cirugía
	Durante la Estancia en la Terapia Intensiva	
Apacheii (C)	APACHE II (Calificación de la gravedad de la enfermedad).	Continua
Apacheii (D)	APACHE II codificado	0=15 ó menos; 1=16 ó más
Mortpred(C)	Mortalidad predicha (Calculada por el modelo APACHE II). Expresada en porcentaje.	Continua (expresada en%)
Mortpred(D)	Mortalidad predicha codificada.	0=15 ó menos; 1=16 ó más
Aps(C)	APS1. Calificación fisiológica aguda en el primer día de estancia en la terapia.	Continua
Aps(D)	APS1 codificado.	0=10 ó menos; 1=11 ó más
Bruselas(C)	Bruselas día 2.	Continua
Bruselas(D)	Bruselas del día 2 codificado.	0=4 ó menos; 1=5 ó más
Sepsisgr	Sepsis grave.	0=Sí; 1=No
Bicho	Germen causante de la sepsis grave.	0=sin/aislamiento; 1=Con/A
F.neurol	Falla neurológica.	0=No; 1=Sí
F.renal	Falla renal.	0=No; 1=Sí
F.respir	Falla respiratoria.	0=No; 1=Sí
Días.fneurol	Días en falla neurológica.	Continua
Días.frenal	Días en falla renal.	Continua
Días.frespir	Días en falla respiratoria.	Continua
Vmdías	Días en ventilación mecánica asistida.	Continua
Cirugía	Cirugía durante la estancia en la UTI.	0=No; 1=Sí
Traqueo	Traqueotomía.	0=Si; 1=No
Usoantib	Uso de antibiótico en la UTI.	0=Sí; 1=No
Glasgow	Glasgow día 1.	Continua
Creatinina	Creatinina día 1.	Continua
pH arterial	pH arterial día 3.	Continua
ik.resp.(C)	Índice de Kirby respiratorio día 3.	Continua
ik resp.(D)	Índice de Kirby respiratorio día 3 codificado.	0=151 ó mayor; 1=150 ó menor
estuti	Estancia en UTI.	Continua
altauti	Motivo de egreso de la UTI.	0=Mejoría;1=Malas condiciones
	Posterior al egreso de la Terapia, aún Hospitalizado	
Hospital	Estancia hospitalaria posterior al egreso de la terapia.	Continua

Cuadro D. Conjunto de variables explicativas para ser consideradas después de hacer un análisis exploratorio.

Debe aclararse que la tabla anterior, tiene variables que son continuas y que a su vez estas fueron modificadas para que fueran dicotómicas, ya que en ocasiones resulta que es más relevante para la variable respuesta la primera en lugar de la segunda o viceversa.

3.4. VARIABLES CORRELACIONADAS

Ya que se determinaron las variables explicativas a considerar en el *Cuadro D,* se tiene que revisar que no haya problemas de colinealidad entre las variables independientes, esto es, identificar que variables se encuentran altamente correlacionadas, hecho que si ignoramos nos traería complicaciones para la obtención de los estimadores.

El procedimiento utilizado para la obtención de la colinealidad de las variables fue por medio de las correlaciones de Pearson.

Para el estudio es importante conocer la relación que existe entre las variables independientes: *BRUSELAS, APS, APACHE II* y *MORTALIDAD PRE*DICHA, con respecto al factor riesgo Calidad de Vida a los 3 meses del egreso hospitalario, pero la construcción de las primeras 4 variables se hizo de la forma siguiente:

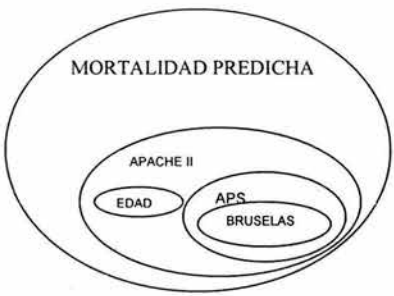


Diagrama 1. Muestra la construcción de las variables independientes; mortalidad predicha, bruselas, aps y apache ii.

Lo anterior se corrobora por medio del siguiente cuadro;

Correlations

Correlación de Pearson	APACHE II (C)	Mortalidad predicha (C)	APS (C)	Bruselas (C)	Sepsis grave (C)	Edad (C)
APACHE II (C)	1.000	.803**	.880**	.496**	.167*	.283**
Mort pred (C)	.803**	1.000	.724**	.450**	.245**	.174**
APS (C)	.880**	.724**	1.000	.532**	.185**	-.082
Bruselas (C)	.496**	.450**	.532**	1.000	.137*	.027
Sepsis grave	.167*	.245**	.185**	.137*	1.000	-.047
Edad (C)	.283**	.174**	-.082	.027	-.047	1.000

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Cuadro E. Correlación lineal entre variables continúas.

Como era de esperarse la correlación que existe entre las variables apache ii (C), mortalidad predicha, Aps. (C) y bruselas (C) es alta, por lo tanto si se incluyeran las 4 variables los resultados serían poco confiables, siendo ésta una razón por el cual se tiene que dar prioridad a la variable que resulte ser de mayor interés para el análisis o el que nos proporcione mayor información, por lo tanto solamente se incluirá para la modelación de las variables, *APACHE II.*

Otras variables que presentan correlación lineal alta son: falla respiratoria, días de ventilación mecánica, traqueotomía y estancia en la Unidad de Terapia Intensiva, considerando solamente para el modelo la variable *DÍAS DE VENTILACIÓN MECÁNICA.*

Correlations

Correlació de Pearson	Días Falla Respir.	Enf.pulmonar	Días en ventilación	Traqueotomía	Est. UTI	Est. Post.UTI
Días F. Respir.	1.000	.166*	.941**	.698**	.914**	.369**
Enf.pulmonar	.166*	1.000	.162*	.159*	.143*	.117
Días en ventilación	.941**	.162*	1.000	.732**	.954**	.374**
Traqueotomía	.698**	.159*	.732**	1.000	.692**	.358**
Est. UTI	.914**	.143*	.954**	.692**	1.000	.356**
Est. Post.UTI	.369**	.117	.374**	.358**	.356**	1.000

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Cuadro F. Correlaciones lineales de 6 variables independientes.

A parte de las dos tablas anteriores se encuentran otras correlaciones que se presentan en el *anexo II tabla 1.*

Las variables que son consideradas para la modelación con sus respectivos porcentajes son las siguientes:

ETIQUETA	VALORES	PORCENTAJE	CALIDAD DE VIDA		TOTAL
			Buena	Mala	
Sepsis grave	No	observaciones	58	63	121
		porcentaje	47.90%	52.10%	100.00%
	Sí	observaciones	65	51	116
		porcentaje	56.00%	44.00%	100.00%
		Total	123	114	237
Hospital	Siglo XXI	observaciones	65	88	153
		porcentaje	42.50%	57.50%	100.00%
	La Raza	observaciones	58	26	84
		porcentaje	69.00%	31.00%	100.00%
		Total	123	114	237
Sexo	Femenino	observaciones	60	61	121
		porcentaje	49.60%	50.40%	100.00%
	Masculino	observaciones	63	53	116
		porcentaje	54.30%	45.70%	100.00%
		Total	123	114	237
Diabetes	No	observaciones	87	90	177
		porcentaje	49.20%	50.80%	100.00%
	Sí	observaciones	36	24	60
		porcentaje	60.00%	40.00%	100.00%
		Total	123	114	237
Servicio	Cirugía	observaciones	55	74	129
		porcentaje	42.60%	57.40%	100.00%
	Medicina interna	observaciones	68	40	108
		porcentaje	63.00%	37.00%	100.00%
		Total	123	114	
Cirugía	No	observaciones	102	83	185
		porcentaje	55.10%	44.90%	100.00%
	Sí	observaciones	21	31	52
		porcentaje	40.40%	59.60%	100.00%
		Total	123	114	237
Enfermedad pulmonar	No	observaciones	116	98	214
		porcentaje	54.20%	45.80%	100.00%
	Sí	observaciones	7	16	23
		porcentaje	30.40%	69.60%	100.00%
		total	123	114	237

ETIQUETA	VALORES	PORCENTAJE	CALIDAD DE VIDA		TOTAL
			Buena	Mala	
Cáncer	No	observaciones	119	104	223
		porcentaje	53.40%	46.60%	100.00%
	Sí	observaciones	4	10	14
		porcentaje	28.60%	71.40%	100.00%
		total	123	114	237
CV 1	Buena	observaciones	110	73	183
		porcentaje	60.10%	39.90%	100.00%
	Mala	observaciones	13	41	54
		porcentaje	24.10%	75.90%	100.00%
		total	123	114	237

Cuadro G. Porcentaje de las variables explicativas por calidad de vida 2, utilizadas para el modelo.

ETIQUETA	CATEGORIA		CALIDAD DE VIDA		TOTAL
			BUENA	MALA	
Glasgow	Continua	Media	14.28	13.14	
		Tamaño de muestra	123	114	237
Edad	Continua	Media	47.46	56.01	
		Tamaño de muestra	123	114	237
Ventilación Mecánica	Continua	Media	4.85	6.86	
		Tamaño de muestra	123	114	237
Apacheii (C)	Continua	Media	14.93	16.47	
		Tamaño de muestra	123	114	237
Dias falla Renal	Continua	Media	0.57	0.28	
		Tamaño de muestra	123	114	237
Estancia posterior a la UTI	Continua	Media	15.32	22.5	
		Tamaño de muestra	123	114	237

Cuadro H. Promedio de las variables explicativas continuas para la construcción del modelo.

Se obtuvieron 15 variables explicativas para la aplicación del análisis de regresión logística, no se cuenta con datos faltantes, la categoría de referencia es el que contiene el mayor número de observaciones; por ejemplo, para la variable *cáncer* se tiene como categoría de referencia cuando *no tiene cáncer*.

3.5. Procedimiento para la obtención de un Modelo de Regresión Logística

Para iniciar el procedimiento de selección de las 15 variables consideradas en el *cuadro G y H*, se tiene que comenzar con la aplicación de la regresión logística al factor riesgo que en este caso es la variable dependiente como se muestra a continuación:

$$Y_i = \begin{cases} 1 & \textit{Mala calidad de vida} \\ 0 & \textit{Buena calidad de vida} \end{cases}$$

El modelo propuesto para la variable respuesta considerando las 15 variables explicativas queda:

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{15} x_{15}$$

Obteniéndose por la estadística de Wald como significativas las variables *edad, enfermedad pulmonar, cáncer, calidad de vida 1, estancia posterior a la UTI y glasgow.*

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	SEXO (MASC.)	.151	.333	.205	1	.651	1.163	.605	2.233
	EDAD	.027	.010	7.287	1	.007	1.028	1.008	1.048
	ENF.PULMONAR (SÍ)	1.342	.599	5.011	1	.025	3.826	1.182	12.388
	DIABETES (SÍ)	.074	.388	.036	1	.849	1.077	.503	2.305
	CÁNCER (SÍ)	1.429	.726	3.879	1	.049	4.176	1.007	17.319
	CV 1(MALA)	1.686	.424	15.8	1	.000	5.397	2.349	12.397
	SERVICIO (MED.INTERNA)	-.456	.344	1.761	1	.184	.634	.323	1.243
	APACHEII	-.035	.032	1.187	1	.276	.965	.906	1.028
	SEPSIS GRAVE (SÍ)	-.171	.371	.213	1	.645	.843	.408	1.743
	FALLA RESPIRATORIA	-.015	.036	.173	1	.678	.985	.917	1.058
	EST.POS.UTI.	.031	.011	8.173	1	.004	1.032	1.010	1.054
	FALLA RENAL	.002	.116	.000	1	.984	1.002	.798	1.259
	CIRUGÍA (SÍ)	.497	.433	1.319	1	.251	1.644	.704	3.839
	GLASGOW	-.294	.079	14.0	1	.000	.745	.639	.869
	HOSPITAL (LA RAZA)	-.708	.384	3.398	1	.065	.493	.232	1.046
	Constant	2.353	1.274	3.414	1	.065	10.521		

a. Variable(s) entered on step 1: SEXO, EDAD, EPOC, DIABETES, CÁNCERCO, CV1CODIF, SERVICIO, APACHEII, SEPSISGR, DIASFRES, ESTPOSUT, DIASFREN, CIRUGÍA, GLASG1, HOSPITAL.

Tabla 1. Valores de los Coeficientes de las variables incluidas para ajustar la probabilidad de tener una mala calidad de vida.

3.5.1. VARIABLES SIGNIFICATIVAS

Ya que se obtuvieron las variables estadísticamente significativas se va a proceder a utilizar el método de selección Backward implementado en el paquete estadístico SPSS

Los resultados se muestran a continuación:

FACTOR $\left(\frac{\hat{\beta}(i)}{SE(\hat{\beta})}\right)^2$	ETIQUETA	VARIABLE	COEFICIENTE REGRESIÓN	ERROR ESTANDAR		
MÉTODO	BACKWARD					
Edad del paciente	*EDAD*	***Continua***	*0.024*	*0.009*	*6.388*	*P < 0.02*
Enf.pulmonar	*ENFERMEDAD PULMONAR (Si)*	*No=0* * ***Sí=1***	*1.088*	*0.549*	*3.923*	*P < 0.05*
Enf. Cáncer	*CANCER(SÍ)*	*No=0* * ***Sí=1***	*1.476*	*0.699*	*4.454*	*P < 0.04*
CV a los 2 meses previos a la UTI.	*CV1 (Buena)*	*Buena=0* * ***Mala=1***	*1.548*	*0.397*	*15.18*	*P < 0.01*
Calificación de coma	*Glasgow*	***Continua***	*-2.65*	*0.071*	*13.77*	*P < 0.01*
Estancia posterior a la UTI	*EST.POS.UTI*	***Continua***	*0.031*	*0.01*	*9.082*	*P< 0.01*
Atención que necesita el paciente	*SERVICIO (Med.Interna)*	*Cirugia=0* * ***Med. Interna=1***	-0.63	0.357	3.722	*P < 0.06*
Tipo de Hospital	*HOSPITAL (La Raza)*	*Siglo XXI =0* * ***La Raza=1***	-0.611	0.357	2.926	P<.0.09

Tabla 2. Coeficientes estimados de la Regresión Logística para un modelo potencialmente candidato.

Son ocho variables estadísticamente significativas, teniendo como categoría de referencia el valor con el asterisco (*), dentro de las cuales dos resultan tener un p-valor de significancia mayor a 0.05, aunque estadísticamente no son significativas son necesarias para la construcción del modelo ya que el ajuste se mejora y se obtiene un mayor número de características asociadas a la Mala calidad de vida.

Modelo logístico:

$$\ln\left(\frac{\hat{\pi}(Y=1)}{1-\hat{\pi}(Y=1)}\right) = 1.127 - .024\ Edad + 1.088\ Enfermedad\ pulmonar(SÍ) + 1.476\ Enf.cáncer(SÍ) + 1.548\ CV1\ (MALA) - 2.65\ Glasgow + .031\ Est.postUTI - 0.630\ Servicio(Med.Interna) - 6.11\ Hospital(La\ Raza) \quad .....(A)$$

3.5.2. PRUEBA DE SIGNIFICANCIA DEL MODELO

En la prueba de significancia del cociente de verosimilitud se tienen que establecer dos modelos, uno referente al modelo ajustado y el otro al modelo saturado, la diferencia de la devianzas entre los modelos se establecerá el estadístico $G = -2(L_o - L_a)$ para la realización de la prueba.

Prueba de Hipótesis:

$$H_0 : \beta_0 + \beta_1\ Edad + \beta_2\ Enf.pulmonar + \beta_3\ Enf.cáncer + \beta_4\ CV\ 1 + \beta_5\ Glasgow + \beta_6\ Est.post.UTI$$

VS.

$$H_a : \beta_0 + \beta_1\ Edad + \beta_2\ Enf.pulmonar + \beta_3\ Enf.cáncer + \beta_4\ CV\ 1 + \beta_5\ Glasgow + \beta_6\ Est.post.UTI + \beta_7\ Servicio + \beta_8\ Hospital$$

Se tiene:

Modelo	(-2log-likelihood)	$G = -2(L_o - L_a)$	χ_1^2
H_0	258.552	8.175	5.990
H_a	250.377		

El valor para la χ_1^2 con 2 grados de libertad con nivel de significancia del 0.05 en tablas es igual a 5.990, como el valor $\chi_1^2 < G$, se rechaza la hipótesis nula y se acepta la hipótesis alternativa que contiene las variables servicio y hospital, esto nos indica que evidentemente las variables consideradas en la hipótesis nula contribuyen efectivamente a "explicar" las modificaciones que se producen en *P(Y=1)*.

3.5.3. *Porcentaje Correcto de Clasificación*

La utilidad de contar con tablas cruzadas para la obtención del porcentaje de clasificación correcto es para obtener el porcentaje de los casos considerados correctamente entre el total de casos clasificados, como se muestra en el siguiente cuadro.

Classification Table[a]

Observed			Predicted		
			Calidad de Vida		Porcentaje correcto
			BUENA	MALA	
Step 1	Calidad de vida	BUENA	94	29	76.4
		MALA	29	85	74.6
	Overall Percentage				75.5

a. The cut value is .500

Cuadro 8. Clasificación de la tabla para el modelo de Calidad de Vida a los 3 meses del egreso hospitalario del modelo (A) de la página 113.

Se tiene el valor de *especificidad* igual $94/123=.764$, indica que las personas que tuvieron una buena calidad de.vida a los 3 meses de haber salido de la UTI se clasificaron como correctas 94 observaciones, por lo tanto la predicción es aceptable.

Para el valor de *sensibilidad* se tiene $85/114=.745$, indica que las personas que tienen una mala calidad de vida a los 3 meses de egreso hospitalario se clasifican como correctos 81 observaciones de mala calidad de vida, por lo tanto la predicción es aceptable.

Se puede obtener otro indicador a partir de las observaciones estimadas conocido como el Valor pronóstico positivo, esto es, divide el número de observaciones clasificados correctamente entre el total de observaciones estimados, como se ve a continuación.

Pronóstico	Calidad de vida 2		Total	Valor pronóstico
	Buena	Mala		
Buena	98	33	131	(98/131)=.75
Mala	25	81	106	(81/106)=.76
Total	123	114	237	

Cuadro 9. Valor pronóstico sobre los datos que se están clasificando como correctos.

El valor de pronóstico positivo es igual a .76 y para el valor de prevalencia observada es igual a 114/237=0.48, esto nos indica que fue mejor en más de 1 vez el pronóstico de la calidad de vida a los 3 meses de egreso hospitalario mediante el modelo propuesto que sin él.

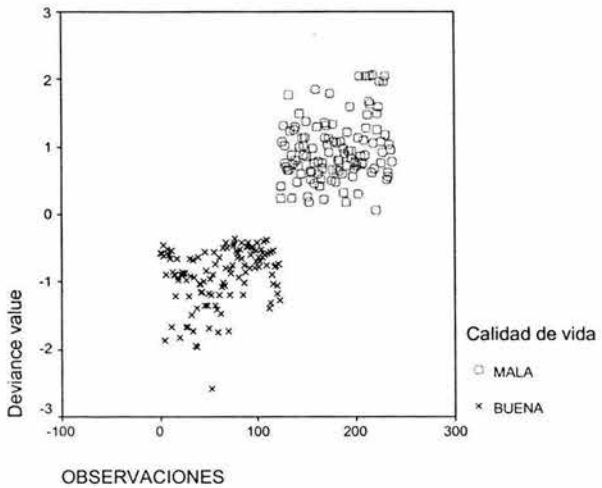
En la prevalencia de la muestra se tiene el 48% y el Valor de Pronóstico Positivo es del 76.4%, indicando que fue mejor el pronóstico de la Calidad de Vida 2 mediante el modelo propuesto que sin él.

3.5.4. ANÁLISIS DE RESIDUALES

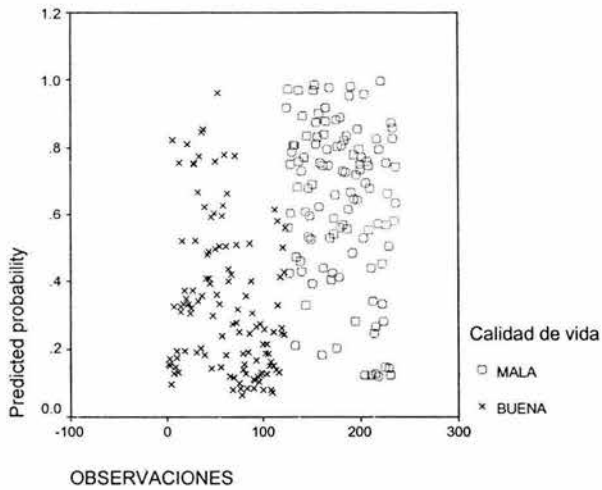
El análisis de residuales es recomendado para identificar el comportamiento sistemático, que pudieran dar inicios de un modelo inadecuado o también para detectar casos o conjuntos de observaciones aisladas.

Dicho análisis se realiza por medio de los residuos de Pearson, para el caso continuo y para los casos categóricos se utilizo al residual DF-Beta, como se determino en el capítulo 1.

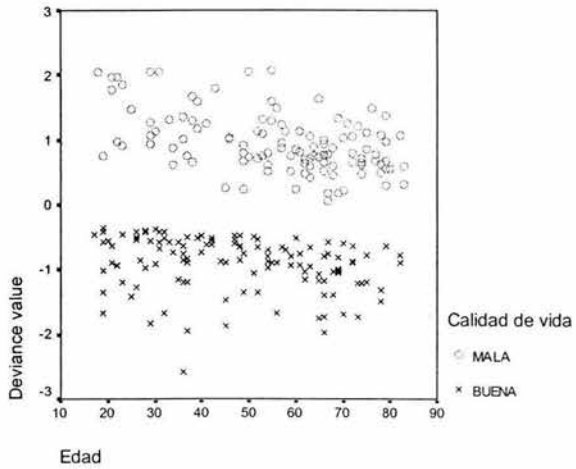
En las siguientes gráficas se muestran 2 grupos; del lado ***positivo*** están clasificados los valores de ***CALIDAD DE VIDA MALA***, a los 3 meses del egreso hospitalario mientras que en los valores ***negativos*** se tiene ***CALIDAD DE VIDA 2 BUENA.***



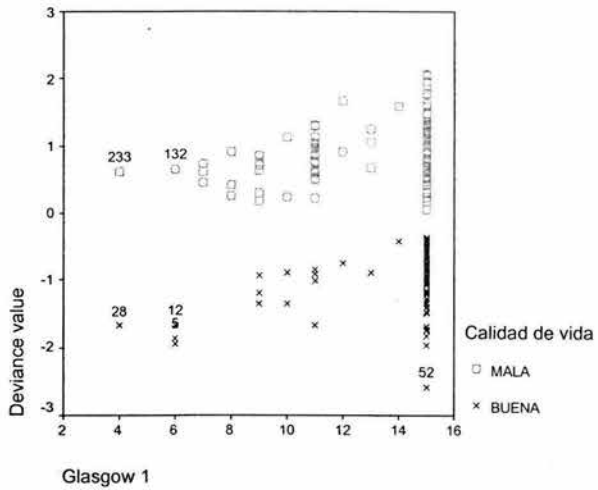
Gráfica de dispersión 1. Rresiduos de Pearson vs Número de observación.



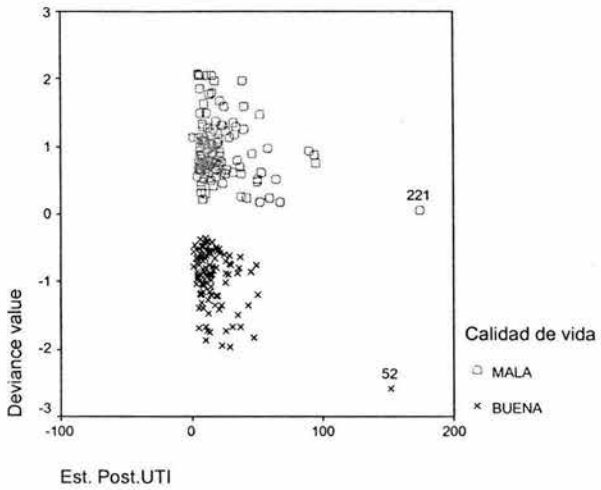
Gráfica de dispersión 2. Probabilidad estimada vs Número de observación.



Gráfica de dispersión 3. Residuos estándarizados vs edad.

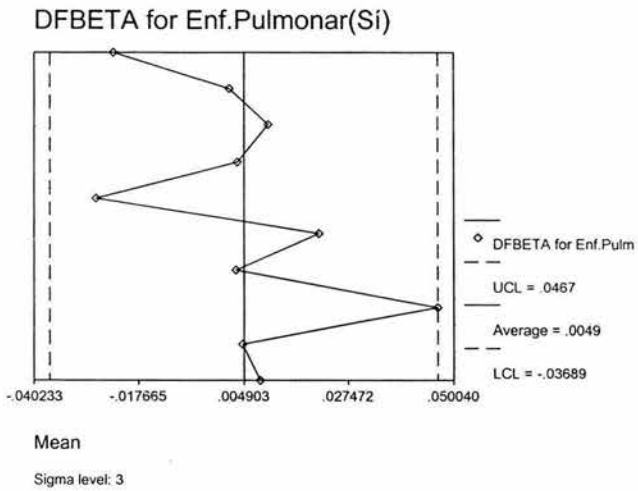


Gráfica de dispersión 4. Residuos estándarizados vs glasgow.

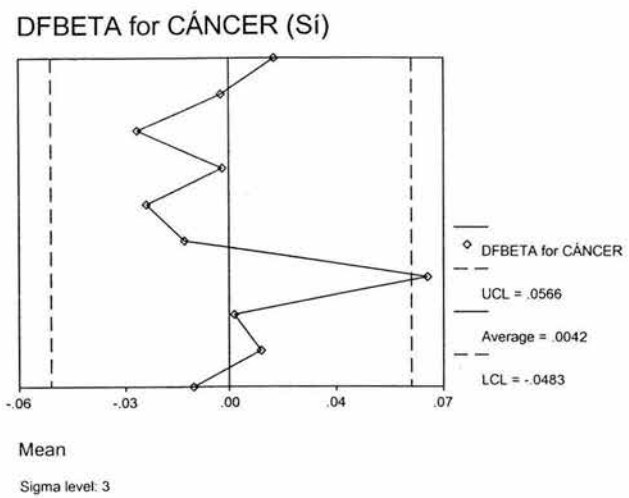


Gráfica de dispersión 5. Residuos estándarizados vs. estancia posterior a la UTI

A continuación se presenta los siguientes 2 gráficos para el caso categórico.



Gráfica 6. DFBeta de la variable enfermedad pulmonar



Gráfica 7. DFBeta de la variable enfermedad cáncer.

Las gráficas faltantes para las variables categóricas se encuentran en el *anexo II, gráficas del 1-3.*

No se tuvo que realizar algún tipo de ajuste adicional, ya que se obtuvo que el ajuste del modelo es bueno en general por la razón de que los 2 grupos de nubes se encuentran entre 2 y -2, con la excepción de 2 casos que son los puntos influyentes, correspondiendo a las observaciones **52** y **221** (*gráfica 5).* La observación 52 corresponde a una mujer de 36 años, proveniente del hospital Siglo XXI, con servicio de cirugía, contó con sepsis grave, falla respiratoria, traqueotomía, necesitó ventilación mecánica por 37 días, estuvo 44 días en la estancia de la Unidad de Terapia Intensiva (UTI) y por último permaneció en la estancia posterior a la UTI 152 días, considerándose como un caso grave al igual que la observación 221 que trata de un hombre de 67 años, proveniente del Hospital La Raza, tuvo enfermedad pulmonar, sepsis grave, necesitó cirugía y traqueotomía, permaneció en la estancia posterior a la UTI 174 días.

3.5.5. *Intervalo de Confianza*

Ya identificamos las variables que resultaron ser significativas, a continuación se presenta el intervalo de confianza del 95% de confianza de los cocientes de momios.

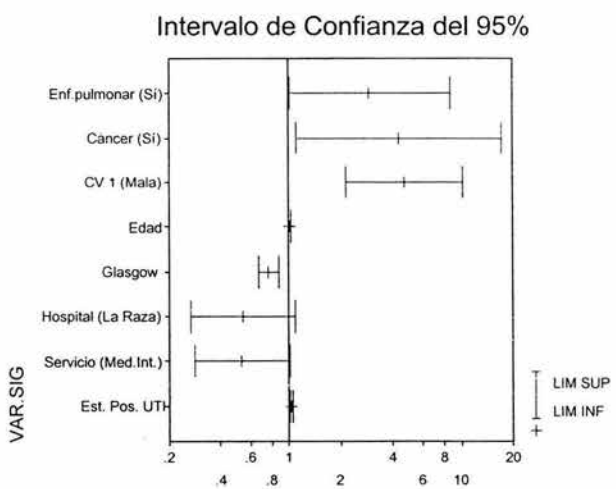
Variables in the Equation

		B	Exp(B)	95.0% C.I.for EXP(B)	
				Lower	Upper
Step 1[a]	EDAD	.024	1.024	1.005	1.043
	ENF.PULM (SÍ)	1.088	2.969	1.012	8.716
	CÁNCER(SÍ)	1.476	4.375	1.111	17.229
	CV 1(MALA)	1.548	4.700	2.158	10.238
	EST.POS.UTI	.031	1.031	1.011	1.052
	GLASGOW	-.265	.767	.667	.882
	SERVICIO(MED.INT.)	-.630	.533	.281	1.010
	HOSPITAL (LA RAZA)	-.611	.543	.269	1.093
	Constant	1.757	5.797		

a. Variable(s) entered on step 1: EDAD, EPOC, CÁNCERCO, CV1CODIF, ESTPOSUT, GLASG1, SERVICIO, HOSPITAL.

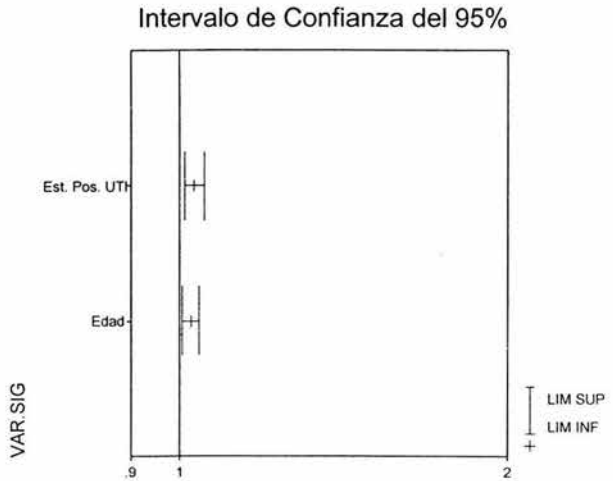
Cuadro 10. Valores del intervalo de confianza para la $exp(\hat{\beta})$.

Por medio de gráficas se tiene:

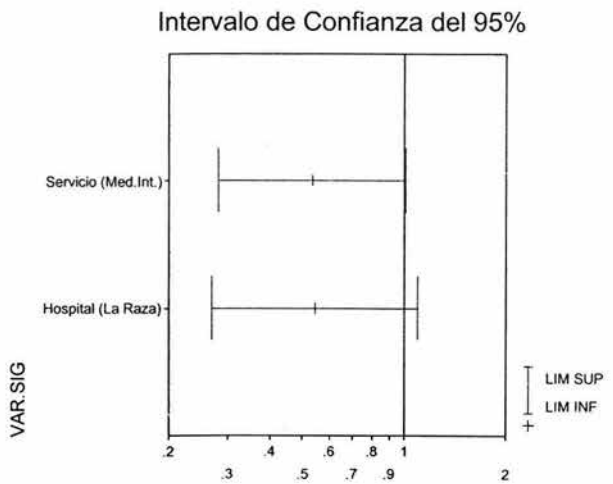


Gráfica 8. Intervalo de confianza del 95% para la $exp(\hat{\beta})$, *con escala logarítmica.*

A continuación se presentan otras dos gráficas para su mejor apreciación.



Gráfica 8a. Intervalo de confianza del 95% para la $exp(\hat{\beta})$, con escala logarítmica.



Gráfica 8b. Intervalo de confianza del 95% para la $exp(\hat{\beta})$, con escala logarítmica, de las variables servicio y hospital.

A partir de la *Gráfica 8* se aprecia que los intervalos de confianza de las $\exp(\hat{\beta})$ no contienen el valor de 1 por lo que se afirma que el 95% de confianza de las variables estadísticamente significativas tienen un efecto significativo en la probabilidad de riesgo. En la *Gráfica 8b*, se tiene que en el intervalo incluye el valor de 1, lo cual nos indicaría que no son significativas para el modelo estadísticamente, pero son consideradas por la relevancia que tienen para la obtención de resultados, como se vio en el *punto 3.5.2*.

3.5.6. Interpretación del Momio

Para la interpretación de los coeficientes de las variables es necesario tener en cuenta como se ha definido la variable de respuesta, ya que en la obtención de un parámetro estimado con signo positivo indica que la P(Y=1) crece cuando los valores de la variable crecen, pero depende tanto de la variable en cuestión como del suceso Y= Mala calidad de vida, es decir, se determina la relación funcional entre la variable respuesta con las variables explicativas.

Por ejemplo, se tiene un caso en particular que es la variable explicativa enfermedad pulmonar.

Variable respuesta:

$$Y = \begin{cases} 1 & Tiene\ Mala\ CV \\ 0 & Tiene\ Buena\ CV \end{cases}$$

Variable independiente:

$$X_1 = \begin{cases} 0 & No \\ 1 & Si \end{cases}$$

Queda el modelo de la siguiente manera:

$$RM = \frac{Momio\left(\begin{array}{l|l} Mala\,CV & Enf.Pulmonar = Si, enf.cáncer = x, CV1 = x, Edad = x, Glasgow = x, \\ & Servicio. = x, Hospital = x, Est.Pos.UTI = x \end{array}\right)}{Momio\left(\begin{array}{l|l} Mala\,CV & Enf.Pulmonar = No, enf.cáncer = x, CV1 = x, Edad = x, Glasgow = x, \\ & Servicio = x, Hospital = x, Est.Pos.UTI = x \end{array}\right)}$$

$$= e^{1.088} = 2.968$$

El momio nos indica que la probabilidad de tener una Mala calidad de vida (MCV) en pacientes que tuvieron enfermedad pulmonar es 2.968 veces la probabilidad de tener una mala calidad de vida en pacientes que no tuvieron enfermedad pulmonar, permaneciendo constantes las otras variables. Lo anterior también se puede verificar por medio del intervalo de confianza (*Gráfica 8*), indica que el momio de tener una MCV de los pacientes que tuvieron enfermedad pulmonar puede ser aproximadamente entre 1 o 8.7 veces más que el momio de tener una Mala CV sin tener una enfermedad pulmonar.

Para los demás casos se establecerá la interpretación del momio en las conclusiones.

3.5.7. Conclusiones

El total de observaciones consideradas fue de 237, de las cuales el 35.4% son del Hospital la Raza y el 64.6% del Hospital Siglo XXI (*tabla 1, anexo II*).

Al explorar la relación entre la calidad de vida (a los 3 meses de egreso de la UTI) y los factores: edad, glasgow, servicio, hospital, etc., (*cuadro G*), se encontró en el periodo 2002 al 2004 una correlación tanto positiva como negativa de las variables: *edad, enfermedad pulmonar, cáncer, CV a los 2 meses previos al*

ingreso de la UTI, glasgow (calificación de coma), estancia posterior a la UTI, servicio (atención que necesito el paciente) y por último la variable *Hospital.*

En la variable edad se tiene que la probabilidad de tener una mala calidad de vida aumenta cuando el paciente tiene mayor edad, como se ve en las *gráficas 4 y 5 Anexo II.*

De los 237 observaciones se tiene el 5.9% con cáncer y el 94.1% (*tabla 3, anexo II*), sin complicaciones de algún padecimiento de cáncer, esta variable resultó tener una influencia para el modelo, entonces la probabilidad de tener una mala calidad de vida en lugar de una buena con respecto a los pacientes que tienen cáncer fue mayor en comparación de la probabilidad de no tener cáncer y tener una mala CV en lugar de una buena.

Se registró el 77.2% de pacientes con una buena calidad de vida a los 2 meses previos de ingresar a la Unidad de Terapia Intensiva (UTI) y el 22.8% con una mala calidad de vida (*tabla 4, anexo II*), por lo tanto la probabilidad de tener una mala calidad de vida a los 3 meses del egreso hospitalario de los pacientes que tuvieron una mala calidad de vida a los 2 meses previos al ingreso hospitalario fue mayor en comparación de la probabilidad de haber tenido una buena calidad de vida a los 2 meses previos al ingreso hospitalario y tener una mala calidad de vida, observando que de los 54 pacientes que ingresaron con una mala calidad de vida al hospital solamente el 75.9% al salir del hospital siguió teniendo una mala calidad de vida *(tabla 5, anexo II*).

Para la calificación de coma, que es la variable glasgow se refiere al número de días que tardo en salir de coma y sobre como se encuentran sus condiciones físicas; esta variable nos indica que la probabilidad de tener una mala calidad de vida disminuye si el paciente salio del coma en un plazo no mayor de 10 días encontrándose vivo con un desempeño normal.

La probabilidad para la variable estancia posterior a la UTI de tener una mala calidad de vida es mayor cuando el paciente permanece en la UTI un mayor número de días (*gráfica 7, anexo II*).

El 54.4% de los pacientes requirieron de una intervención quirúrgica mientras que el 45.6% solamente necesito de la medicina interna *(tabla 7, anexo II)*. Se tiene que la probabilidad de tener una mala calidad de vida de los pacientes que necesitaron de la medicina interna fue menor en comparación de los pacientes que requirieron de cirugía.

Para la variable hospital se encontró que la probabilidad de tener una mala calidad de vida con respecto a los pacientes que se encontraron internados en el hospital la raza es menor en comparación a los que se encontraron en el hospital Siglo XXI.

Dicha asociación se puede explicar en parte porque de los 129 pacientes que requirieron de una cirugía el 62.7% se realizó en el Hospital Centro Médico Nacional (Siglo XXI), con un número mayor de cirugías neurológicas mientras que en el Hospital del Centro Médico la Raza se tuvo el 39.3%, teniendo como especialidad la cirugía cardiaca y la primera tiende a dar una mala calidad de vida a largo plazo (*tabla 8, anexo II),*

COMENTARIOS FINALES DEL CAPÍTULO

3.6. Comentarios Finales

La base de datos proporcionada por el Médico Cirujano Luis David Sánchez resultó ser buena ya que para la obtención de la información se tuvo cuidado de que se llenara correctamente y de que no hubiera ninguna duda al respecto de la persona que acepto ingresar al estudio, llevándose a cabo también un estudio detallado para la obtención de resultados fiables y de interés concerniente a un aspecto importante dentro del ámbito médico que es la calidad de vida que tienen los pacientes después del egreso Hospitalario de la Unidad de Terapia Intensiva.

La base de datos se sometió a un análisis exploratorio, para determinar modificaciones necesarias tanto en el número de variables explicativas como en su codificación. También se tuvo cuidado de elegir si se tenia una variable continua o en su caso la misma variable recodificada como categórica para establecer cual de la dos resultaba tener una relación estadística significativa.

Para la obtención del modelo final se debe concretar la idea sobre lo que el investigador requiere para consumar los objetivos expuestos desde un inicio, ya que se pueden obtener distintos modelos con respecto a las variables independientes que se estén considerando, pero a fin de cuentas se expondrá el más indicado para dicha investigación, la exploración del modelo final va tomado de la mano con respecto a la exploración de lo que significan y representan variables explicativas para la variable respuesta.

Por esta razón se presentó el mejor modelo obtenido con respecto a los intereses del investigador familiarizado con el tema, ya que es la persona indicada para explicar o determinar que tan relacionadas se encuentran dichas variables utilizadas y según sea el caso de determinar que variables entran al modelo o no según su conocimiento médico sobre las que tienen una mayor relevancia para la categoría de respuesta.

COMENTARIOS

COMENTARIOS.

Para la realización de esta tesis se realizo un gran trabajo ya que con anterioridad no se había manejado o más bien no se tenía un dominio sobre el tema de Análisis de Regresión Logística, por lo tanto se comenzó a partir de los estadísticos necesarios para la comprensión y el conocimiento del tema y así poder realizar una investigación con éxito.

Se cumplieron los objetivos y metas personales, que fue la terminación de la tesis demostrando los conocimientos adquiridos en el transcurso de la preparación académica a nivel Licenciatura en el ramo matemático. Se adquirió experiencia en el manejo y exploración de los datos para la obtención de resultados importantes, así como de la correcta aplicación del Análisis de Regresión Logística, también se obtuvo el conocimiento sobre que resultados se pueden obtener al aplicar dicho modelo, referente a que variables se pueden utilizar y de cómo se deben determinar para llevar a cabo el proceso de la aplicación de la metodogía (se determinan las estadísticas básicas de la Regresión Logística), al no tener contratiempos para la interpretación de cada uno de los procesos establecidos y extraer las conclusiones correspondientes a cada paso del análisis, también se adquirió el conocimiento sobre la utilización del manejo del paquete estadístico SPSS.

Se tuvo contacto con personas especialistas en los campos de estudios al igual que en el área estadística, llevando una buena comunicación e interrelación para el entendimiento mutuo sobre lo que se quería obtener en base a los datos presentados.

ANEXOS I Y II

ANEXO I

Correlations

Correlaciones de Pearson	Edad	Promedio	Univer.	Termino carrera	Quiso carrera	País	Autos	Educ. padre	Educ. madre	Sexo	Ocupación madre	IMAFE	Com. padre	Com. madre
Edad	1.000	-.274	.503	.062	.280	.012	.274	.023	-.396	-.188	.320	.040	.054	.047
Promedio	-.274	1.000	-.284	-.038	-.159	-.065	-.146	.100	.225	.368	-.194	-.013	.045	.096
Universidad	.503	-.284	1.000	.167	.185	-.019	.650	.069	-.552	-.165	.289	.081	-.162	-.090
Terminar	.062	-.038	.167	1.000	-.004	-.116	.240	.045	-.185	.010	.085	.057	-.002	-.017
Quiso esta car	.280	-.159	.185	-.004	1.000	-.013	.077	.077	-.133	-.264	.083	-.051	.113	.006
País	.012	-.065	-.019	-.116	-.013	1.000	-.046	-.050	-.033	-.106	.011	.110	-.038	-.165
Autos	.274	-.146	.650	.240	.077	-.046	1.000	.093	-.513	-.111	.214	.098	-.212	-.145
Educ.padre	.023	.100	.069	.045	.077	-.050	.093	1.000	-.001	-.033	-.015	.109	-.023	-.052
Educ.madre	-.396	.225	-.552	-.185	-.133	-.033	-.513	-.001	1.000	.179	-.475	.002	.019	.047
Sexo	-.188	.368	-.165	.010	-.264	-.106	-.111	-.033	.179	.000	-.191	-.091	.033	.145
Ocupación ma	.320	-.194	.289	.085	.083	.011	.214	-.015	-.475	-.191	1.000	.007	.133	.049
IMAFE	.040	-.013	.081	.057	-.051	.110	.098	.109	.002	-.091	.007	1.000	-.001	.063
Com.padre	.054	.045	-.162	-.002	.113	-.038	-.212	-.023	.019	.033	.133	-.001	1.000	.452
Com.madre	.047	.096	-.090	-.017	.006	-.165	-.145	-.052	.047	.145	.049	.063	.452	1.000

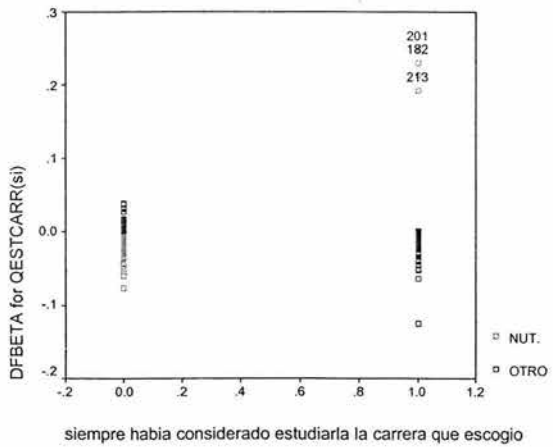
Tabla A. Correlación lineal entre las variables explicativas propuestas en la tabla B del capítulo 2.

Variables in the Equation

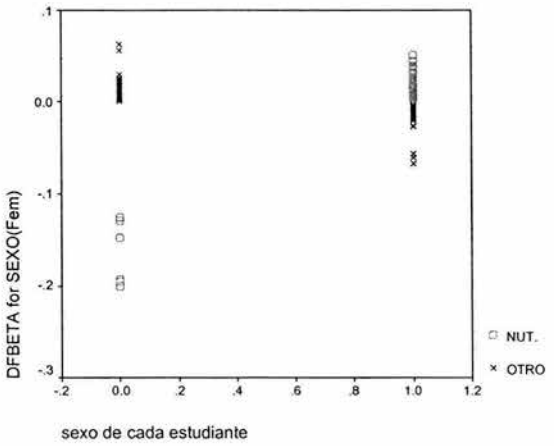
		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	PROMEDIO	.871	.337	6.662	1	.010	2.390	1.233	4.630
	TERCARRE (EST)			3.148	2	.207			
	TERCARRE(TRAB)	-.969	.553	3.064	1	.080	.380	.128	1.123
	TERCARRE(OTRA)	-.051	1.102	.002	1	.963	.951	.110	8.244
	QUISOESTCARR(SI)	-1.531	.656	5.454	1	.020	.216	.060	.782
	PAIS(NO)	8.742	20.233	.187	1	.666	6261	.000	1.046
	SEXO(FEM)	2.889	.663	18.959	1	.000	17.967	4.895	65.940
	OCUP.MADRE(HOGAR)	-.469	.528	.787	1	.375	.626	.222	1.763
	IMAFE(AND)			2.027	3	.567			
	IMAFE(IND)	.695	.681	1.043	1	.307	2.004	.528	7.612
	IMAFE(FEM)	.583	.597	.954	1	.329	1.791	.556	5.766
	IMAFE(MASC)	-.119	.760	.024	1	.876	.888	.200	3.941
	COMMADRE(SI)	.907	.991	.838	1	.360	2.478	.355	17.287
	EDPADRES (LIC)			6.922	3	.074			
	EDPADRES(SEC)	-2.166	.864	6.288	1	.012	.115	.021	.623
	EDPADRES(PREP)	-.758	.752	1.015	1	.314	.469	.107	2.048
	EDPADRES(MoD)	.315	.592	.283	1	.595	1.370	.430	4.372
	UNIVER(UAM)	2.579	.709	13.216	1	.000	13.180	3.282	52.927
	Constant	-19.55	20.459	.913	1	.339	.000		

a. Variable(s) entered on step 1: PROMEDIO, TERCARRE, QESTCARR, PAIS, SEXO, MADTRAB, IMAFE COMMADRE, EDPADRES, UNIVER.

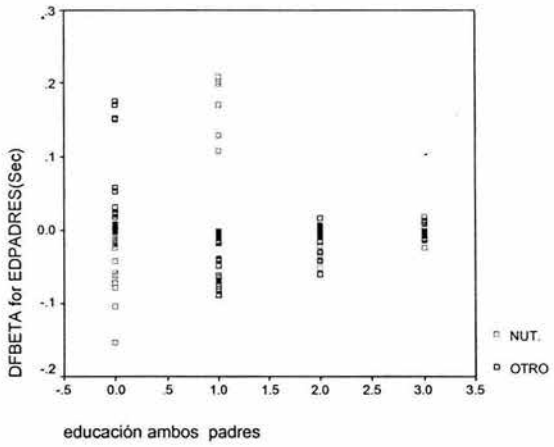
TABLA 1 Resultado de ajustar la probabilidad de mejoría en función de 10 variables explicativas para la carrera de nutrición.



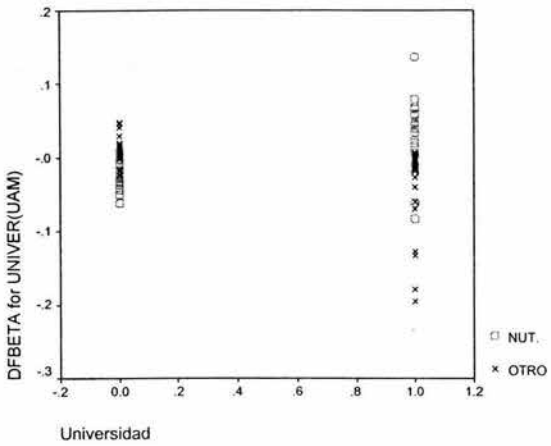
Gráfica de dispersión 1. Residuos estándarizados vs quiso esta carrera (0=no, 1=Sí)



Gráfica de dispersión 2. DFBeta del sexo del estudiante (Fem) vs sexo del estudiante (0=Masc. y 1=Fem)



Gráfica de dispersión 3. DFBeta de escolaridad ambos padres (Sec) vs Escolaridad ambos padres (0=Lic., 1=Sec, 2=Prep y 3=MoD)



Gráfica de dispersión 4. DFBeta Universidad (UAM) vs variable universidad (0=UIA, 1=UAM)

Contingency Table for Hosmer and Lemeshow Test

		OTRO		NUTRICIÓN		
		Observed	Expected	Observed	Expected	Total
Step 1	1	30	29.861	0	.139	30
	2	31	30.671	0	.329	31
	3	29	30.383	2	.617	31
	4	29	29.077	1	.923	30
	5	29	29.182	2	1.818	31
	6	28	26.310	2	3.690	30
	7	23	22.331	6	6.669	29
	8	21	19.923	10	11.077	31
	9	14	15.703	17	15.297	31
	10	6	6.557	17	16.443	23

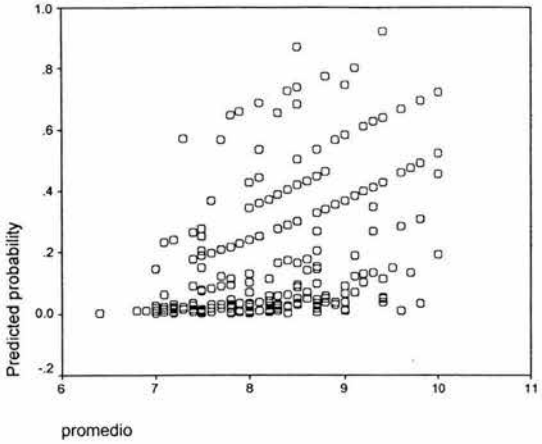
Tabla 2. Resultado de la prueba de Hosmer y Lemeshow para el modelo propuesto de la carrera de nutrición.

				UNIVERSIDAD		
				UIA	UAM	Total
OTRO	SEXO	MASC	Count	80	83	163
			%	49.1%	50.9%	100%
		FEM	Count	53	27	80
			%	66.3%	33.8%	100%
	Total		Count	133	110	243
			%	54.7%	45.3%	100%
NUT.	SEXO	MASC	Count		6	6
			%		100.0%	100%
		FEM	Count	32	21	53
			%	60.4%	39.6%	100%
	Total		Count	32	27	59
			%	54.2%	45.8%	100%

Tabla 3. Distribución de las observaciones según tipo de universidad por carrera de nutrición y sexo.

			EDUCACIÓN AMBOS PADRES				
			al menos un padre lic.	ambos a lo mas sec.	al menos un padre prep.	al menos un padre maestria o doctorado	Total
CARRERA	ADM	Count	70	35	29	23	157
		%	44.6%	22.3%	18.5%	14.6%	100.0%
	ING	Count	30	30	14	9	83
		%	36.1%	36.1%	16.9%	10.8%	100.0%
	NUT	Count	27	8	7	15	57
		%	47.4%	14.0%	12.3%	26.3%	100.0%
Total		Count	127	73	50	47	297
		%	42.8%	24.6%	16.8%	15.8%	100.0%

Tabla 4. Distribución de las observaciones según educación ambos padres por carrera.



Gráfica de dispersión 5. Probabilidad estimado vs la variable promedio del estudiante.

Contingency Table for Hosmer and Lemeshow Test

		OTRO		INGENIERIA		
		Observed	Expected	Observed	Expected	Total
Step 1	1	29	28.727	1	1.273	30
	2	73	73.271	6	5.729	79
	3	35	35.273	20	19.727	55
	4	34	33.727	33	33.273	67

Tabla 5. Resultado de la prueba de Hosmer y Lemeshow para la carrera de ingeniería.

			SEXO		
			masculino	femenino	Total
CARRERA DE INGENIERIA	OTRO	Count	69	102	171
		%	40.4%	59.6%	100.0%
	ING	Count	53	7	60
		%	88.3%	11.7%	100.0%
Total		Count	122	109	231
		%	52.8%	47.2%	100.0%

Tabla 6. Distribución de las observaciones según sexo por carrera de ingeniería.

			UNIVERSIDAD		
			UIA	UAM	Total
CARRERA DE INGENIERIA	OTRO	Count	105	66	171
		%	61.4%	38.6%	100.0%
	ING	Count	30	30	60
		%	50.0%	50.0%	100.0%
Total		Count	135	96	231
		%	58.4%	41.6%	100.0%

Tabla 7. Distribución de las observaciones según tipo de universidad por carrera de ingeniería.

			OCUPACIÓN DE LA MADRE		
			trabaja	hogar	Total
CARRERA DE INGENIERIA	OTRO	Count	107	64	171
		%	62.6%	37.4%	100.0%
	ING	Count	39	21	60
		%	65.0%	35.0%	100.0%
Total		Count	146	85	231
		%	63.2%	36.8%	100.0%

Tabla 8. Distribución de las observaciones según ocupación de la madre por carrera de ingeniería.

			EDUCACIÓN AMBOS PADRES				
			al menos un padre lic.	ambos a lo mas sec.	al menos un padre prep.	al menos un padre M o D	Total
CARRERA DE INGENIERIA	OTRO	Count	82	26	30	33	171
		%	48.0%	15.2%	17.5%	19.3%	100.0%
	ING	Count	24	18	12	6	60
		%	40.0%	30.0%	20.0%	10.0%	100.0%
Total		Count	106	44	42	39	231
		%	45.9%	19.0%	18.2%	16.9%	100.0%

Tabla 9. Distribución de las observaciones según escolaridad ambos padres por carrera de ingeniería.

Contingency Table for Hosmer and Lemeshow Test

		OTRO		ADMINISTRACIÓN		
		Observed	Expected	Observed	Expected	Total
Step 1	1	19	21.512	10	7.488	29
	2	15	13.962	7	8.038	22
	3	43	39.301	32	35.699	75
	4	10	11.072	15	13.928	25
	5	5	7.269	14	11.731	19
	6	11	7.693	14	17.307	25
	7	4	6.190	25	22.810	29

Tabla 10. Resultado de la prueba de Hosmer y Lemeshow para la carrera de administración.

				CARRERA		
UNIVER				OTRO	ADM	Total
UIA	SEXO	MASC	Count	25	32	57
			%	43.9%	56.1%	100.0%
			%	43.9%	44.4%	44.2%
		FEM	Count	32	40	72
			%	44.4%	55.6%	100.0%
			%	56.1%	55.6%	55.8%
	Total		Count	57	72	129
			%	44.2%	55.8%	100.0%
			%	100.0%	100.0%	100.0%
UAM	SEXO	MASC	Count	32	30	62
			%	51.6%	48.4%	100.0%
			%	64.0%	66.7%	65.3%
		FEM	Count	18	15	33
			%	54.5%	45.5%	100.0%
			%	36.0%	33.3%	34.7%
	Total		Count	50	45	95
			%	52.6%	47.4%	100.0%
			%	100.0%	100.0%	100.0%

Tabla 11. Distribución de las observaciones según tipo de carrera por universidad y sexo.

			CARRERA		
			NUT. Y ING	ADM	Total
TERCARRE	ESTUDIAR	Count	75	72	147
		%	51.0%	49.0%	100.0%
	TRABAJAR	Count	25	42	67
		%	37.3%	62.7%	100.0%
	OTRA	Count	7	3	10
		%	70.0%	30.0%	100.0%
Total		Count	107	117	224
		%	47.8%	52.2%	100.0%

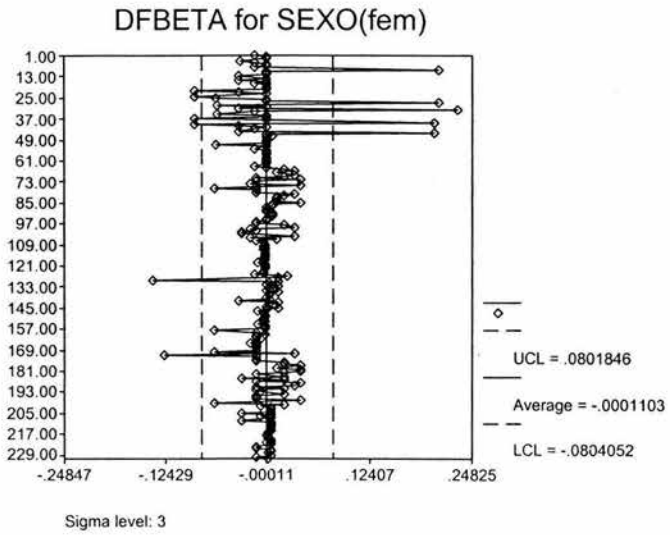
Tabla 12. Distribución de las observaciones según tipo de carrera por actividad al terminar la carrera

			CARRERA		
			OTRO	ADM	Total
OCUP DE LA MADRE	trabaja	Count	81	62	143
		%	56.6%	43.4%	100.0%
	hogar	Count	26	55	81
		%	32.1%	67.9%	100.0%
Total		Count	107	117	224
		%	47.8%	52.2%	100.0%

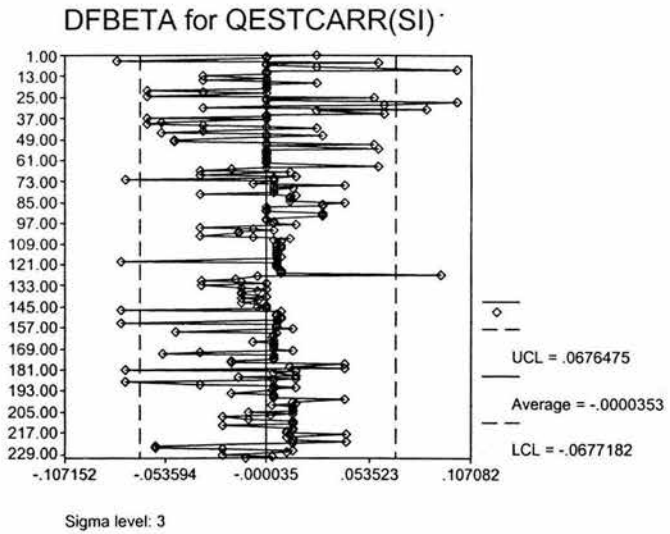
Tabla 13. Distribución de las observaciones según tipo de carrera por ocupación que tiene la madre.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ADM	183	53.2	53.2	53.2
	ING	91	26.5	26.5	79.7
	NUT	70	20.3	20.3	100.0
	Total	344	100.0	100.0	

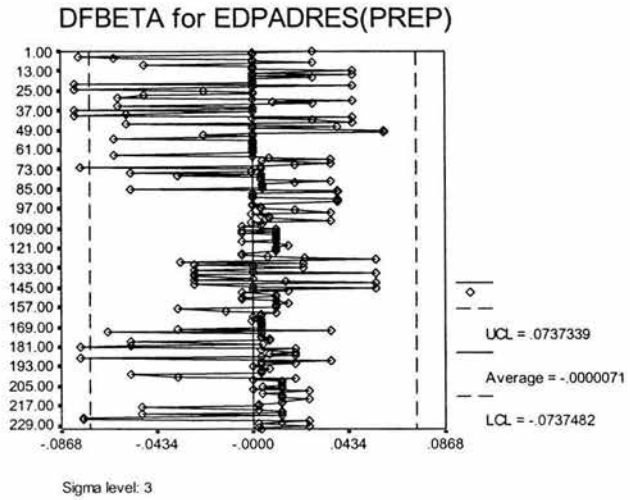
Tabla 1. Distribución de las observaciones por carrera.



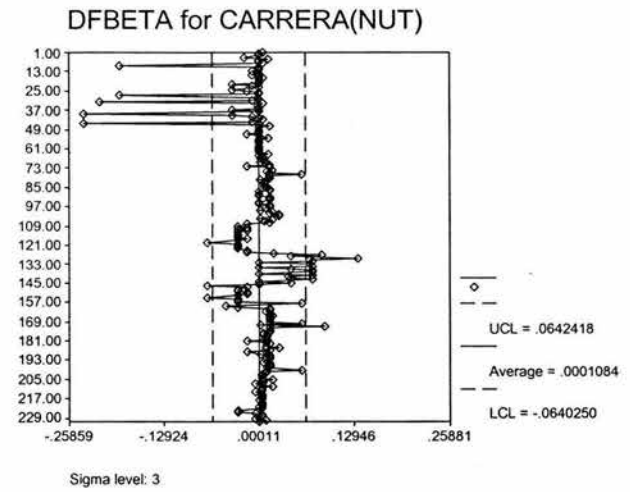
Gráfica 6. DFBeta para la categoría sexo femenino.



Gráfica7. DFBeta para la categoría ya había considerado la carrera



Gráfica 8. DFBeta para escolaridad ambos padres al menos uno con preparatoria.



Gráfica 9. DFBeta para la variable carrera.

Contingency Table for Hosmer and Lemeshow Test

		UIA		UAM		
		Observed	Expected	Observed	Expected	Total
Step 1	1	31	31.178	2	1.822	33
	2	19	19.684	3	2.316	22
	3	22	20.957	3	4.043	25
	4	17	18.026	6	4.974	23
	5	18	17.243	6	6.757	24
	6	15	14.748	8	8.252	23
	7	10	10.934	17	16.066	27
	8	3	2.234	20	20.766	23
	9	0	.002	31	30.998	31

Tabla 15. Resultado de la prueba de Hosmer y Lemeshow para la variable respuesta universidad.

			CARRERA			
			ADM	ING	NUT	Total
UNIVERSIDAD	UIA	Count	74	30	31	135
		%	54.8%	22.2%	23.0%	100.0%
	UAM	Count	46	30	20	96
		%	47.9%	31.3%	20.8%	100.0%
Total		Count	120	60	51	231
		%	51.9%	26.0%	22.1%	100.0%

Tabla 16. Distribución de las observaciones según carrera por universidad.

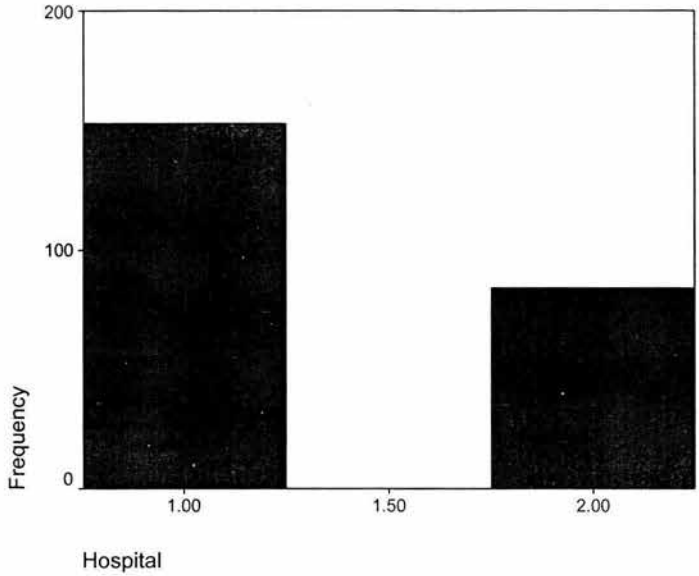
			CARRERA			
			ADM	ING	NUT	Total
QUISO ESTA CARRERA	NO	Count	80	32	45	157
		%	66.7%	53.3%	88.2%	68.0%
	SI	Count	40	28	6	74
		%	33.3%	46.7%	11.8%	32.0%
Total		Count	120	60	51	231
		%	100.0%	100%	100.0%	100.0%

Tabla 17. Distribución de las observaciones según carrera por siempre quiso esta carrera.

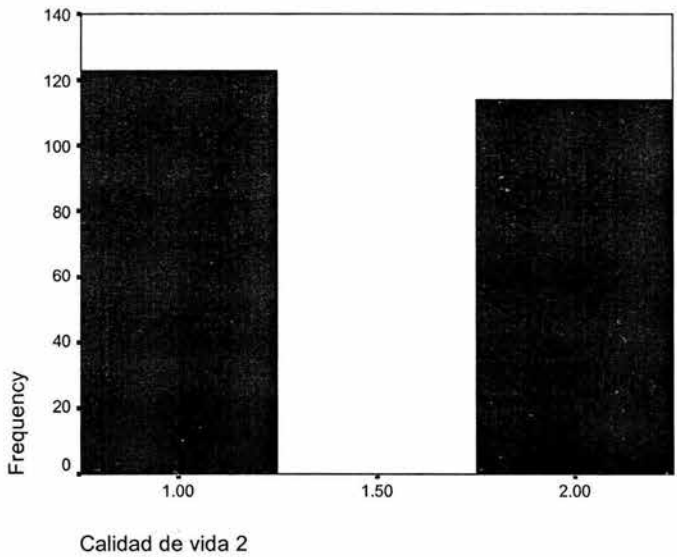
			SEXO		
			MASC	FEM	Total
UNIVERSIDAD	UIA	Count	59	76	135
		%	43.7%	56.3%	100.0%
	UAM	Count	63	33	96
		%	65.6%	34.4%	100.0%
Total		Count	122	109	231
		%	52.8%	47.2%	100.0%

Tabla 18. Distribución de las observaciones según sexo por universidad.

ANEXO II



Gráfica A. Frecuencia de la variable hospital(1=Siglo XXI, 2=la Raza)



Gráfica B. Frecuencia de la variable calidad de vida a los 3 meses del egreso hospitalario(1=buena, 2=mala)

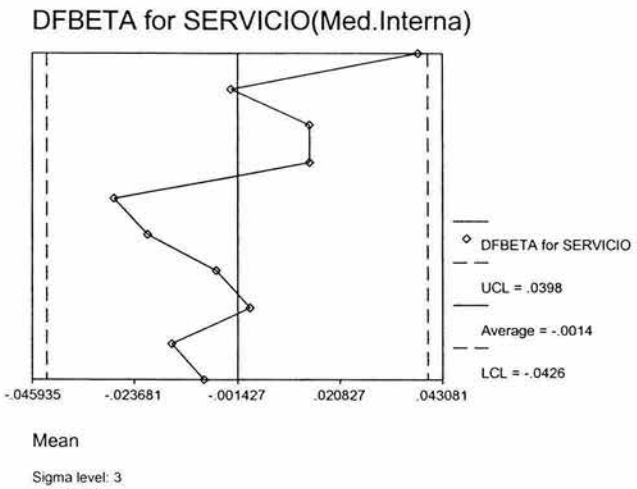
Correlations

	cv1codif.	Hospital	Sexo	Edad	Trabajo	Diabetes	Cáncer	Servicio	Días Falla Neurol.	Días Falla Renal.	Cirugía	Uso de antibiótico	Glasgow	pH 3
cv1codif.	1.000	-.108	-.129*	.260**	-.211**	-.062	-.008	.133*	-.074	-.038	-.094	-.146*	.080	.101
Hospital	-.108	1.000	.086	-.067	.080	.137*	-.186**	-.225**	-.172**	.068	-.158*	.082	.272**	-.420**
Sexo	-.129*	.086	1.000	-.070	-.424**	-.007	-.031	.015	-.055	-.025	-.132*	.107	.117	-.152*
Edad	.260**	-.067	-.070	1.000	-.328**	.189**	-.093	.157*	-.003	-.042	-.045	-.065	.040	.086
Trabajo	-.211**	.080	-.424**	-.328**	1.000	-.060	.040	-.187**	.039	-.065	.111	.059	-.165*	-.051
Diabetes	-.062	.137*	-.007	.189**	-.060	1.000	-.146*	-.110	-.094	.117	-.098	-.002	.095	-.015
Cáncer	-.008	-.186**	-.031	-.093	.040	-.146*	1.000	.157*	-.070	-.064	.127	.004	.042	.092
Servicio	.133*	-.225**	.015	.157*	-.187**	-.110	.157*	1.000	-.026	-.094	.137*	-.193**	.022	.231**
Días Falla Neurol.	-.074	-.172**	-.055	-.003	.039	-.094	-.070	-.026	1.000	-.003	.116	-.075	-.677**	.106
Días Falla Renal	-.038	.068	-.025	-.042	-.065	.117	-.064	-.094	-.003	1.000	.071	-.029	.050	.021
Cirugía	-.094	-.158*	-.132*	-.045	.111	-.098	.127	.137*	.116	.071	1.000	-.061	-.068	.133*
Uso antibiótico	-.146*	.082	.107	-.065	.059	-.002	.004	-.193**	-.075	-.029	-.061	1.000	.064	-.286**
Glasgow	.080	.272**	.117	.040	-.165*	.095	.042	.022	-.677**	.050	-.068	.064	1.000	-.162*
pH 3	.101	-.420**	-.152*	.086	-.051	-.015	.092	.231**	.106	.021	.133*	-.286**	-.162*	1.000

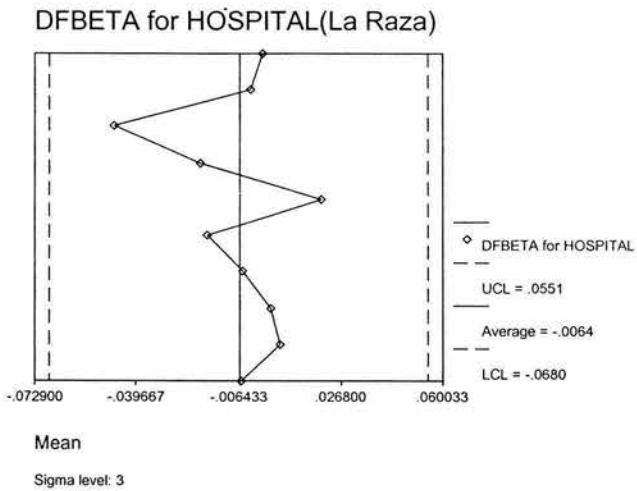
*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

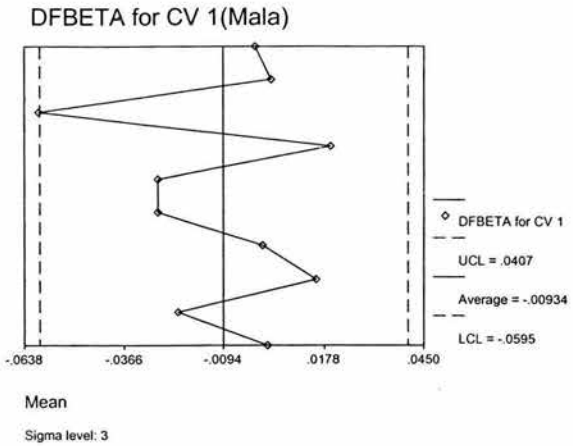
Tabla 1. Correlaciones de las variables explicativas a considerar para la modelación de las mismas.



Gráfica 1. DFBeta para la variable servicio (medicina Interna)



Gráfica 2. DFBeta para la variable hospital (La Raza).



Gráfica 3. DFBeta de la variable calidad de vida a los 2 meses previos de ingresar a la UTI (mala).

Hospital

		Frequency	Percent
Valid	SIGLO XXI	153	64.6
	LA RAZA	84	35.4
	Total	237	100.0

Tabla 2. Distribución de los pacientes según tipo de hospital.

Cáncer

		Frequency	Percent
Valid	No	223	94.1
	Si	14	5.9
	Total	237	100.0

Tabla 3.Distribución de los pacientes si presentan la enfermedad de cáncer o no.

Calidad de Vida 1

		Frequency	Percent
Valid	Buena	183	77.2
	Mala	54	22.8
	Total	237	100.0

Tabla 4. Distribución de los pacientes según calidad de vida que tuvieron a los 2 meses previos a la hospitalización.

			Calidad de vida (a los3 meses de egreso hospitalario)		
			BUENA	MALA	Total
Calidad de vida (a los 2 meses del ingreso hospitalario)	BUENA	Observaciones	110	73	183
		porcentaje	60.1%	39.9%	100%
	MALA	Observaciones	13	41	54
		porcentaje	24.1%	75.9%	100%
Total		Observaciones	123	114	237
		porcentaje	51.9%	48.1%	100%

Tabla5. Distribución de los pacientes según calidad de vida a los 3 meses posteriores a la hospitalización y la calidad de vida a los 2 meses previos de ingresar a la UTI.

Hospital

		Frequency	Percent
Valid	Siglo XXI	153	64.6
	La Raza	84	35.4
	Total	237	100.0

Tabla 6. Distribución de los pacientes que aceptaron ingresar al estudio según tipo de institución hospitalaria.

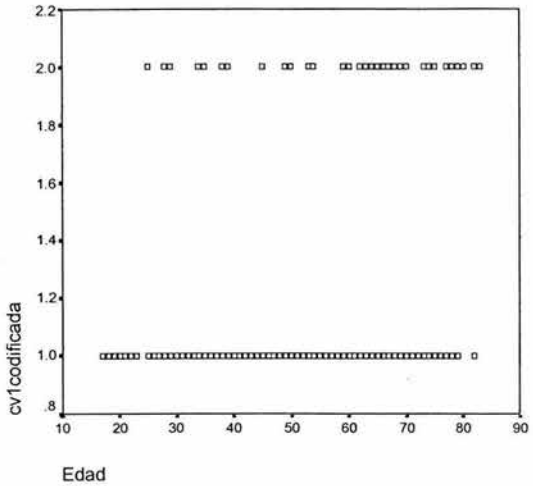
Servicio

		Frequency	Percent
Valid	Medicina Interna	108	45.6
	Cirugía	129	54.4
	Total	237	100.0

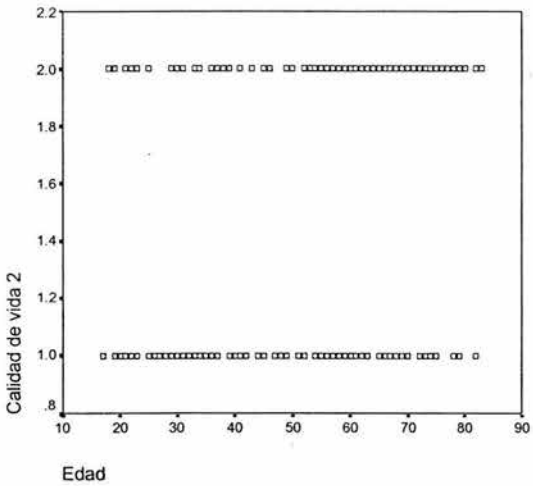
Tabla 7. Distribución de los pacientes según el tipo de servicio que necesitaron en la UTI,

			Servicio		
			Medicina Interna	Cirugía	Total
Hospital	Siglo XXI	Observaciones	57	96	153
		porcentaje	37.3%	62.7%	100%
	La Raza	Observaciones	51	33	84
		porcentaje	60.7%	39.3%	100%
Total		Observaciones	108	129	237
		porcentaje	45.6%	54.4%	100%

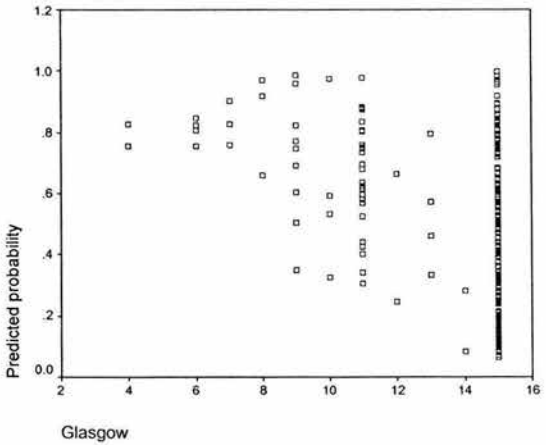
Tabla 8. Distribución de las observaciones según tipo de hospital y servicio que necesitaron en la UTI:



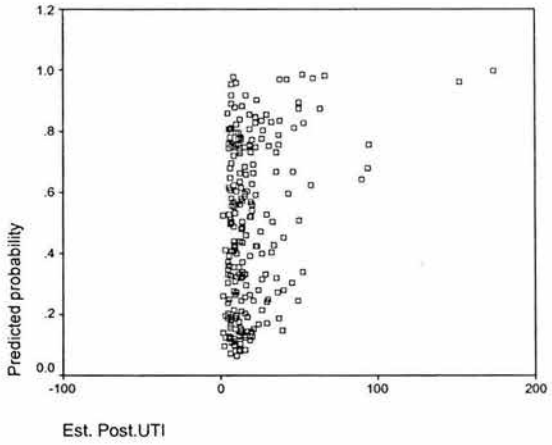
Gráfica de dispersión 4. Variable explicativa edad contra la calidad de vida a los 2 meses previos a la UTI. (1=buena,2=mala).



Gráfica de dispersión 5. Variable explicativa edad contra la calidad de vida a los 3 meses de egreso hospitalario. (1=buena,2=mala).



Gráfica de dispersión 6. De la probabilidad estimada contra Glasgow.



Gráfica de dispersión 7. De la probabilidad estimada contra estancia posterior a la UTI.

BIBLIOGRAFÍA

1. Agresti Alan, *Categorial Data Analysis*, 2ª Edición, New York, Wiley Interscience, 2002.
2. Hosmer and Lemeshow, *Applied Logistic Regression*, 2ª Edición, New York, J Wiley, 2000.
3. ANUIES (Asociación Nacional de Universidades e Instituciones de Educación Superior), *Elección de carrera*, Categoría de Ensayos elección de carrera, Colección: Biblioteca de la educación superior, 1ª Edición, 1998.
4. Catterjee Hadi P., *Analysis by example*, 3ª Edición, New York, J. Wiley, 2000.