UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

"EVALUACIÓN DE LA DEPRESIÓN A TRAVÉS DE REGRESIÓN LOGÍSTICA. ENCUESTA DE EVALUACIÓN DEL DESEMPEÑO. MÉXICO 2002-2003"

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 :

KARINA RINCÓN RENTERÍA

DIRECTOR DE TESIS: M. EN A. P. MARÍA DEL PILAR ALONSO REYES

2006

1. Datos del alumno
   Rincón
   Rentería
   Karina
   57 99 30 01
   Universidad Nacional Autónoma de México
   Facultad de Ciencias
   Actuaría
   094115471

2. Datos de tutor
   M. en A. P.
   María del Pilar
   Alonso
   Reyes

3. Datos de sinodal 1
   M. en C.
   José Antonio
   Flores
   Díaz

4. Datos de sinodal 2
   M. en C.
   Rina Betzabeth
   Ojeda
   Castañeda

5. Datos de sinodal 3
   Act.
   Jaime
   Vázquez
   Alamilla

6. Datos de sinodal 4
   Act.
   Lucio Gerardo
   Chávez
   Heredia

7. Datos del trabajo escrito
   Evaluación de la Depresión a través de regresión logística
   Encuesta de Evaluación del Desempeño. México 2002-2003.
   84p
   2006

Agradecimientos

Agradezco

A Pilar por la atención, la paciencia, el apoyo, la confianza y el tiempo dedicado a contribuir enormemente al desarrollo de este trabajo.
¡Muchas Gracias!

Al Dr. Esteban Puentes y a la Mtra. Mariana Bello por facilitarme la información, las bases de datos y sobre todo por su disposición para resolver cualquier pregunta.

A los miembros del jurado que revisaron este trabajo, gracias por su tiempo y contribuciones.

A mi madre por todo su cariño, en el breve tiempo que compartimos.

A mi padre y mis primos por todo el apoyo incondicional en todos estos años.

A Julián por su apoyo incondicional, su amor y compresión.

A todos mis amigos, sobrinos y compañeros que siempre me apoyaron.

Índice

Introducción

Dentro del área estadística, el análisis de regresión logística es un tema que ha adquirido fuerza en los últimos años por ser una herramienta que puede ser aplicada en diversos campos de la investigación, siempre y se cuando cumplan con los supuestos y requerimientos necesarios para su uso.

El análisis de regresión logística permite trabajar con diversos tipos de variables, cuya escala de medida no necesariamente es numérica. Con ellas es posible cuantificar la relación que existe entre la probabilidad de ocurrencia de un determinado suceso que sea de interés, para conocer el comportamiento que podría tener la intervención de otros factores, con un resultado positivo o negativo en relación al suceso en cuestión.

Esta tesis se ha escrito de la forma más sencilla posible, con la finalidad de que sea fácil para la comprensión y el entendimiento de aquellos estudiantes que comienzan a tener inquietudes sobre una de tantas aplicaciones que puede llegar a tener la estadística, pero cabe señalar que se requiere de un mínimo y suficiente material matemático para comprender todos los procesos llevados a cabo.

En el presente documento se aborda el análisis de regresión logística en el contexto de su aplicación en un estudio epidemiológico, en especial en lo que se refiere a trastornos mentales, dentro de los que se encuentran los trastornos depresivos, en donde se buscarán factores que incidan en la prevalencia o no de eventos depresivos en adultos.

En el capítulo I, se describe la metodología necesaria para la aplicación del análisis de regresión logística; se determina el modelo de respuesta dicotómica con 2 posibles valores, así como la interpretación de los coeficientes estimados del modelo propuesto y las correspondientes pruebas de significancia. También se establecen las estadísticas necesarias

para determinar si el modelo propuesto se encuentra ajustado correctamente y por último se establecerá el modelo para una variable politómica.

En el capítulo II se lleva a cabo la descripción de la base metodológica de la Encuesta Nacional de Evaluación del Desempeño 2002-2003, de la cual se extrajo la información que se utilizó en el análisis, tomando únicamente las variables de interés necesarias para definir la depresión en adultos, el cual esta basado en una cédula con criterios de diagnostico definidos en el manual Diagnostico y Estadístico de la Enfermedades Mentales (DSM IV) de la Asociación Americana de Psiquiatría.

En el capítulo III se lleva a cabo la aplicación del análisis de regresión logística, para determinar los factores que pudieran incidir en los episodios depresivos. En este capítulo se describe el procedimiento llevado a cabo para la obtención del mejor modelo propuesto, la interpretación de los coeficientes, las pruebas de hipótesis sobre los parámetros, la preparación de la base de datos y las pruebas de significancia del modelo.

Por último, se presentan las conclusiones del estudio completo, los anexos que encierran todas las corridas que se hicieron en el paquete STATA Statistical Software for Profesional versión 8 para que el lector pueda consultarlas y la bibliografía que se usó para elaborar el material de los capítulos I y II.

Capítulo I

Modelo de Regresión Logística

En este capítulo se cubren los aspectos conceptuales y teóricos esenciales para la comprensión del modelo de regresión logística, que es una herramienta estadística que puede usarse, en los aquellos casos en donde la variable de estudio tiene la característica de ser dicotómica, es decir únicamente tiene dos resultados o valores como "éxito-fracaso", "vivo-muerto", "positivo-negativo" o bien si la variable explicativa es discreta, tomando más de dos posibles valores, para la cual se presenta el modelo logístico multinomial.

1.1 Breve historia de la regresión logística

El concepto de regresión es uno de los pilares más importantes de la estadística y data al menos de principios de 1800 con los trabajos de Legendre, Gauss y Laplace. Es posible que el término de regresión se deba a Francis Galton, quién acuñó el término "***r**egresión* hacia la media" para describir la observación de que los hijos de padres muy altos tienden a ser algo más bajos que sus progenitores, y por el contrario los hijos de padres muy bajos suelen ser algo más altos y por lo tanto acercarse en ambos casos más a la media de la población.

Este fenómeno que se produce en muchos aspectos en la naturaleza, es explicado por Stephen M. Stiegler con el siguiente ejemplo: Supóngase que dos momentos diferentes se efectúa un examen sobre una materia específica a un alumno del que no se tiene referencia, observando que obtiene una nota mucho más alta que la media de sus compañeros de clase. ¿Qué tan buena se esperará que sea la puntuación en el segundo examen para este alumno? Probablemente alta, pero también probablemente no tan alta como en la primera ocasión, ya que probablemente el gran éxito en la primera ocasión se deba a dos componentes: por un lado la capacidad del alumno, componente estable o permanente y por otro un cierto grado de suerte, componente transitorio y en cierta medida aleatorio. El coeficiente que medía esa ***r**egresión* hacia la media pasó desde entonces a indicarse con la letra ***r***.

Aunque este primer concepto de regresión no tenga nada que ver con el sentido que actualmente se utiliza para esa palabra, que designa las técnicas empleadas para construir funciones matemáticas que permiten calcular o predecir el resultado de una o más variables denominadas variables de respuesta o dependiente a partir de otras variables nombradas variables independientes o explicativas.

Una de las técnicas de regresión más utilizadas actualmente en varios campos de estudio como la medicina, la sociología, mercadotecnia entre otros, es la regresión logística. Ya en 1937 Bartlett utilizó la transformación $\log\left[\frac{y}{1-y}\right]$ para analizar proporciones. También Fisher y Yates sugieren en 1938 el uso de esta transformación para analizar datos binarios. El término *logit* fue introducido por Joseph Berkson en 1944 para designar a esta transformación y sus trabajos popularizaron su uso. Jerome Cornfield utilizó la regresión logística para el cálculo de los cocientes de momios como valores aproximados del riesgo relativo en estudios de casos y controles, en 1967 Walker y Duncan contribuyen a abordar el tema de estimar la probabilidad de ocurrencia de cierto acontecimiento en función de varias variables. Uno de los principales difusores de la regresión logística fue David R. Cox en 1970 con su libro *"The Analysis of Binary Data"*. El uso de la regresión logística se expande desde principios de los ochenta debido, especialmente, a las facilidades informáticas con las que se contaba desde ese entonces y ha llegado ha ser, en muchos campos de estudio, el método estándar de análisis, cuando se trata de describir la relación entre una variable de respuesta dicotomica y una o más variables explicativas.

1.2 El problema básico que resuelve la regresión logística

Los modelos de regresión han llegado a ser una técnica integral para cualquier análisis estadístico que trata de describir la relación entre una variable de respuesta y una o más variables explicativas. La variable de respuesta puede tener tres tipos de escala y el modelo de regresión más usado es el que incluye una variable de respuesta con escala cardinal. Sin embargo, hay varios casos donde es frecuente que la variable de respuesta es discreta,

tomando dos o más posibles valores. La forma de este tipo de datos es muy común y estos pueden tener clasificaciones tan comunes como "vivo-muerto", "empleado-desempleado", "éxito-fracaso". Para este último caso, el modelo de *regresión logística* permite predecir la probabilidad de ocurrencia como una función de variables independientes para un fenómeno en el cual la variable de estudio es discreta con dos o más respuestas.

1.3 Modelos lineales generalizados

Los modelos estadísticos clásicos para regresión, series de tiempo y análisis de datos longitudinales son útiles generalmente, en situaciones donde los datos siguen una distribución normal y pueden ser explicados mediante alguna estructura lineal. Estos modelos son simples de interpretar y los métodos son bien interpretados e investigados teóricamente. Sin embargo, los supuestos subyacentes pueden ser demasiado estrictos y la aplicación de los métodos puede ser errónea en situaciones en donde los datos son claramente no-normales.

Los modelos lineales generalizados son una extensión de los modelos lineales clásicos. Una extensa clase de *Modelos Lineales Generalizados* (*GLM)*[1], fueron introducidos por los británicos Nelder y Wedderburn (1972). Los GLM quedan especificados por tres componentes: *aleatorio, sistemático y función liga.*

El componente aleatorio está formado por observaciones $y_1, y_2 .....y_N$ independientemente distribuidas con

$$E(y_i) = \mu_i$$

[1] Por sus siglas en ingles

A partir de una distribución dentro de la familia exponencial, cada observación y_i tiene una función de densidad de la siguiente forma:

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp(y_i Q(\theta_i)) \qquad \ldots\ldots (1.1)$$

La familia exponencial incluye varias distribuciones importantes como la Poisson, la Binomial y la Normal entre otras. Los valores del parámetro θ_i pueden variar entre 1 a N $i = (1,2,\ldots,N)$, dependiendo de los valores de las variables explicativas. El término $Q(\theta_i)$ es llamado parámetro natural de la distribución.

El componente sistemático de un GLM relaciona un vector $\eta = \eta_1,\ldots\ldots,\eta_N$ a una serie de variables explicativas a través de un modelo lineal. Este vector η es llamado predictor lineal.

$$\eta = X\beta$$

Donde *X* es la matriz que contiene los valores de las variables explicativas para las *N* observaciones y β es un vector de parámetros desconocidos.

El tercer componente de un GLM es la liga entre el componente aleatorio y el componente sistemático. Entonces μ_i es ligado al vector η_i por una función monótona diferenciable $g(.)$ es decir:

$$\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij} \qquad \ldots\ldots (1.2)$$

Así, el modelo junta valores esperados de observaciones a variables explicativas. Algunas distribuciones tienen una función liga especial, para la cual existe un estadístico suficiente

para los parámetros en el predictor lineal $\eta = \sum_i \beta_i x_i$. Éstas son llamadas ligas canónicas y ocurren cuando:

$$\theta = \eta$$

Donde θ es el parámetro canónico, los estadísticos suficientes están dados por $\sum_j Yx_j$ $j = 1,.....p$ haciendo la suma sobre unidades.

Dentro de la familia de GLM se incluyen los modelos para datos categóricos. A continuación se detallan dos de ellos, y se ilustran para cada uno sus tres componentes.

a) Función Poisson

Las celdas contenidas en una tabla de contingencia son frecuentemente tratadas como una variable aleatoria con distribución Poisson. Sea y_i el contenido en la i-ésima celda y $E(y_i) = \theta_i$ para $i = 1,2,\ldots, N$, la función de probabilidad es de la forma:

$$f(y_i;\theta_i) = \frac{\exp(-\theta_i)(\theta_i)^{y_i}}{y_i!} = \exp(-\theta_i)\left(\frac{1}{y_i!}\right)\exp[y_i \log(\theta_i)] \qquad \ldots\ldots (1.3)$$

La ecuación (1.3) tiene la forma de la ecuación (1.1), entonces el componente aleatorio es:

$$Q(\theta_i) = \log(\theta_i) \qquad \ldots\ldots (1.4)$$

La liga canónica es la función:

$$\log(\theta_i) = \sum_j \beta_j x_{ji} \qquad \ldots\ldots (1.5)$$

La liga canónica es llamada modelo log lineal para una tabla de contingencia y pertenece a los modelos log lineales.

b) Función Bernoulli

Existen muchas variables de respuestas categóricas que tienen solamente dos clases. La observación para cada sujeto puede ser descrita como un éxito o como un fracaso. La representación de estos posibles resultados se pueden definir numéricamente por 1 y 0. La distribución Bernoulli para variables aleatorias binarias define las probabilidades:

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

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

Donde:

π	es la probabilidad de éxito
$1-\pi$	es la probabilidad de fracaso

Cuando Y_i tiene una distribución Bernoulli con parámetro π_i la función de probabilidad es:

$$f(y_i;\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = (1-\pi_i)\left[\frac{\pi_i}{1-\pi_i}\right]^{y_i} \quad \ldots\ldots(1.6)$$

$$= (1-\pi\ )\exp\left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right)\right] \quad \ldots\ldots(1.7)$$

Para $y_i = 0$ y 1. Esta distribución pertenece a la familia exponencial, el parámetro natural o componente aleatorio es:

$$Q(\theta) = \log\left[\frac{\pi}{(1-\pi)}\right] \qquad \ldots\ldots (1.8)$$

Que es llamada transformación del logit ,y está definida en términos de una transformación de $\pi(x)$ y es el estudio central de la regresión logística.

1.4 Modelo Logístico

Supóngase que se tiene una muestra de n observaciones independientes de la pareja de variables (x_i, y_i) donde $i = 1,2,\ldots,n$, y_i indica el valor de la variable de respuesta y x_i el valor de la variable explicativa para el i-ésimo sujeto. Dado que Y es una variable que sólo toma dos valores 0 y 1, la relación que existe entre la variable de respuesta y la variable explicativa se determina por el valor esperado de la variable de respuesta. A este valor se le conoce como la esperanza condicional.

$$E(Y \mid x) \qquad \ldots\ldots (1.9)$$

Dado el hecho de que se quiere expresar la probabilidad de que ocurra el fenómeno en cuestión en función de cierta variable x, el modelo de regresión sería el siguiente:

$$E(Y \mid x) = \pi(x) = \beta_0 + \beta_1 x \qquad \ldots\ldots (1.10)$$

Dicho modelo podría interpretarse como un modelo de regresión lineal y tratar de estimar a partir de los datos, el procedimiento de mínimos cuadrados, los coeficientes β_0 y β_1 de la ecuación. Sin embargo aunque esto es matemáticamente posible, conduciría a resultados absurdos, ya que cuando se calcule la función obtenida para diferentes valores de x_i, se podrían obtener valores de $\pi < 0$ o bien $\pi > 1$, lo cual carece de sentido, tratándose de una

probabilidad y esto sucede porque dicha restricción no se impone en una regresión lineal, en la que la respuesta puede en principio tomar cualquier valor.

Muchas funciones de distribución han sido propuestas para usarse en un análisis de variables de respuesta dicotómicas. En el libro de Hosmer y Lemeshow, 1989, Cox (1970) discute algo sobre este tema. Algunas de las razones para escoger la distribución logística son:

- Satisface que $0 < \pi < 1$, debido a que la función exponencial produce valores mayores que cero para cualquier real.
- Las variables explicativas pueden ser de naturaleza: dicotómicas, ordinales o nominales,
- Generalmente se espera una relación no lineal entre x y $\pi(x)$, es decir, un cambio en x puede tener menos impacto cuando $\pi(x)$ se encuentra cerca de su valor intermedio.

La forma específica del modelo de regresión logística para el caso cuando una variable es dicotómica es el siguiente:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad \ldots\ldots (1.11)$$

Donde:

e es la base del logaritmo natural

β_0 y β_1 son parámetros a estimar

x es el valor de la variable explicativa

Dado que la variable de respuesta es dicotómica para el caso contrario la forma específica está dada por:

$$1-\pi(\mathrm{x})=\frac{1}{1+\mathrm{e}^{\beta_0+\beta_1}} \qquad \ldots\ldots (1.12)$$

Los *momios* asociados a cierto suceso se definen como la razón entre la probabilidad $\pi(\mathrm{x})$ de que el suceso ocurra y la probabilidad $1-\pi(\mathrm{x})$de que no ocurra, es decir; el momio se calcula como:

$$\frac{\pi(\mathrm{x})}{1-\pi(\mathrm{x})} \qquad \ldots\ldots (1.13)$$

Este cociente expresa cuánto más probable es que se produzca el suceso comparado con que no se produzca. Si se sustituye las ecuaciones (1.11) y (1.12) en (1.13) se obtiene:

$$\frac{\pi(x)}{1-\pi(x)}=e^{\beta_0+\beta_1 x} \qquad \ldots\ldots (1.14)$$

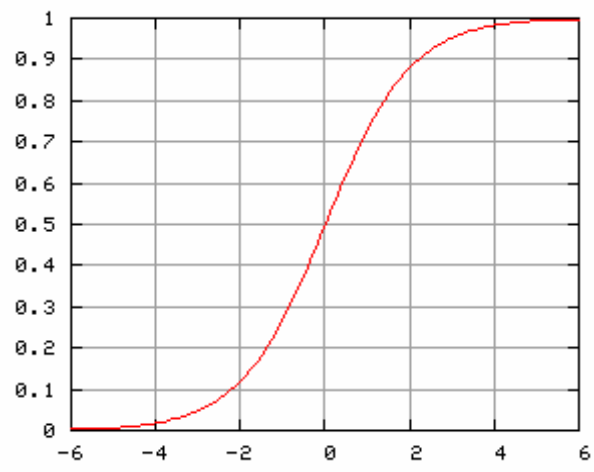
Si a la ecuación anterior se le aplica el logaritmo se obtiene la transformación *logit* o transformación logística siguiente:

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right]=\beta_0+\beta_1 x \qquad \ldots\ldots (1.15)$$

La importancia de la transformación del *logit* es que tiene muchas de las propiedades deseables de un modelo de regresión lineal. El *logit* es lineal en los parámetros y toma valores de $(-\infty,\infty)$. Dependiendo del rango de *x* gráficamente se tiene una curva en forma de ***S*** alargada (figura 1) que representa la relación que existe entre la variable de respuesta y la variable explicativa, además la curva logística es asintótica a los valores 0 y 1, es decir,

se acerca progresivamente a estos valores pero nunca los alcanza y la relación entre la exposición y la probabilidad de ocurrencia del evento es una relación monótona creciente.

Figura 1. Curva logística.



Fuente: http://en.wikipedia.org/wiki/Image:Logistic-curve.png#file

Donde si $x \to \infty$

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

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

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

Entonces se tiene que el modelo de regresión logística consiste en una ecuación matemática que relaciona a una variable explicativa x_i con la probabilidad de ocurrencia de una variable de respuesta Y, relación que se supone es lineal como se observa en (1.15).

Donde las betas son los parámetros desconocidos del componente lineal del modelo, x es el valor de la variable explicativa, Y la variable de respuesta representada como:

$$Y = log\left[\frac{\pi(x)}{1-\pi(x)}\right] \qquad \ldots\ldots (1.16)$$

1.5 Estimación de los parámetros

El método más usado para la estimación de los parámetros de un modelo de regresión logística es generalmente el método de máxima verosimilitud. Este escogerá como valor estimado de los parámetros, aquéllos que tienen mayor probabilidad de ocurrir según los datos observados. Con el fin de aplicar este método, primero se construye una función llamada *función de verosimilitud*, que expresa la probabilidad con base en el valor de los datos observados como una función de parámetros desconocidos. Los estimadores de máxima verosimilitud de cada uno de estos parámetros son aquellos valores que maximizan dicha función de probabilidad. Así los estimadores resultantes son aquellos que concuerdan más fielmente con los datos observados.

Si Y es una variable codificada con valores 0 y 1, entonces la expresión para $\pi(x)$ dada en la ecuación (1.10) proporciona la probabilidad condicional para un valor arbitrario de $\beta' = (\beta_0, \beta_1)$, el vector de parámetros.

$$P(Y = 1|x) = \pi(x) \quad y \quad P(Y = 0|x) = 1 - \pi(x)$$

Una manera conveniente para expresar la contribución a la función de verosimilitud para la pareja (x_i, y_i) es a través del término.

$$f(x_i) = \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \qquad \ldots\ldots (1.17)$$

Puesto que las observaciones son asumidas como independientes, la función de verosimilitud $l(\beta')$ es obtenida como el producto de los términos dados en la expresión (1.18) como:

$$l(\beta') = \prod_{i=1}^{n} f(x_i)$$

$$= \prod_{i=1}^{n} \left[\pi(x_i)^{y_i} \left(1 - \pi(x_i)\right)^{1-y_i} \right] \qquad \ldots\ldots (1.18)$$

El estimador de máxima verosimilitud es obtenido al maximizar el logaritmo de la función de densidad de probabilidad conjunta a partir de la ecuación (1.19), Se realiza de este modo ya que resulta más sencillo que si se tomara solamente la función de probabilidad, además de que la función logaritmo es una función monótona creciente.

$$L(\beta') = \log[l(\beta')] = \sum_{i=1}^{n} y_i \log(\pi(x_i)) + \sum_{i=1}^{n} (1 - y_i) \log(1 - \pi(x_i)) \qquad \ldots\ldots (1.19)$$

$$= \sum_{i=1}^{n} \{ y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i)) \}$$

$$= \sum_{i=1}^{n} y_i \log\left(\frac{\pi(x_i)}{1 - \pi(x_i)} \right) + \sum_{i=1}^{n} \log(1 - \pi(x_i)) \qquad \ldots\ldots (1.20)$$

Sustituyendo el valor de $\pi(x)$ de (1.11) en (1.21) queda como:

$$L(\beta') = log(l(\beta')) = \sum_{i=1}^{n} y_i (\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} log(1 + exp(\beta_0 + \beta_1 x_i)) \qquad \ldots\ldots (1.21)$$

Para encontrar los valores del vector β' que maximicen $L(\beta')$, se deriva la ecuación (1.21) con respecto a β_0 y β_1, igualando a cero la ecuación, queda de la siguiente manera:

$$\frac{\partial \ln(L(\beta_0, \beta_1))}{\partial \beta_0} = \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} = 0 \qquad \ldots\ldots (1.22)$$

$$\frac{\partial \ln\left(L\left(\beta_0,\beta_1\right)\right)}{\partial \beta_1} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{x_i \exp\left(\beta_0 + \beta_1 X_i\right)}{1+\exp\left(\beta_0 + \beta_1 X_i\right)} = 0 \qquad \ldots\ldots (1.23)$$

Las dos ecuaciones se deben resolver simultáneamente, y como son no lineales en β_0 y β_1, el cálculo se realiza utilizando un método numérico iterativo.

Por consecuencia, se escriben las ecuaciones en forma matricial, lo cual ayuda en la transición a la regresión logística múltiple. Sea X una matriz de $n \times 2$ con cada reglón dado por $(1, x_i)$, Y como el vector respuesta y $E(y) = \pi$, la ecuación de probabilidad puede ser escrita como:

$$\frac{\partial L(\beta)}{\partial \beta} = X'(Y - \pi) \qquad \ldots\ldots (1.24)$$

Donde $L(\beta') = \log[l(\beta_0, \beta_1)]$.

Si $\frac{\partial L(\beta)}{\partial \beta} = 0$, de (1.25) se sigue que:

$$X'\pi = X'Y \qquad \ldots\ldots (1.25)$$

Si $\pi = Y$, es decir, si el valor esperado de y_i es igual a la probabilidad estimada para $Y = 1$. La solución de la ecuación (1.25) satisface la condición

$$X'(Y - \hat{Y}) = 0 \qquad \ldots\ldots (1.26)$$

Lo anterior es tomado en el caso de la regresión lineal simple y múltiple. La ecuación (1.26) se resuelve generalmente usando el método de Newton Raphson, el cual resuelve ecuaciones no lineales, para mayor detalle consultar en Agresti (2002).

1.6 Modelo múltiple de regresión logística

Anteriormente se introdujo el modelo de regresión logística en el contexto univariado. Como en el caso de la regresión lineal, la fuerza de la técnica del modelo se halla en la capacidad para poder modelar muchas variables, algunas de éstas pueden estar en diferentes escalas de medida., es decir se pueden tener variables explicativas continuas, o bien discretas, pudiendo ser variables de tipo dicotómico o de escala nominal u ordinal.

La extensión al modelo múltiple que incluye más variables explicativas dentro del modelo de regresión logística es inmediata.

Considérese una colección de *p* variables independientes las cuales serán denotadas por el vector $x' = (x_1, x_2, \ldots, x_p)$. La probabilidad condicional está dada:

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

$$\pi(x') = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p}} = \frac{e^{\beta' x'}}{1 + e^{\beta' x'}} \qquad \ldots\ldots (1.27)$$

Donde:

e es la base del logaritmo natural

β' es el vector de los parámetros a estimar

x' es el vector de las variables explicativas

La ecuación (1.28) es llamada la función de regresión logística y es no lineal. Sin embargo al aplicar la transformación logit produce una función lineal en los parámetros del vector β'.

$$logit(\pi(x')) = log\left(\frac{\pi(x')}{1 - \pi(x')}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad \ldots\ldots (1.28)$$

Se debe recordar que el rango de π se encuentran entre 0 y 1, mientras que el rango de los valores de $\log\left[\frac{\pi(x')}{(1-\pi(x'))}\right]$ se encuentra entre $(-\infty,\infty)$. Los coeficientes β' calculados en el componente lineal referido como $\text{logit}(\pi(x'))$, el cual, es difícil de interpretar. Una forma sencilla de interpretar los coeficientes del modelo es usando las proporciones, las cuales pueden ser derivadas de la ecuación (1.28).

$$\frac{\pi(x')}{1-\pi(x')} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) \qquad \ldots\ldots (1.29)$$

Donde $\pi(x')$ es la probabilidad de que un evento ocurra y $[1-\pi(x')]$ es la probabilidad complemento, es decir, la probabilidad de que un evento no ocurra. Los valores de los parámetros β que son los coeficientes para las variables explicativas muestran cambios en el $\text{logit}(\pi(x'))$ que es asociado con una unidad de cambio en la variable explicativa cuando las demás variables en el modelo son considerados como constantes.

1.7 Estimación de los parámetros para el modelo generalizado

Se considerarán las N respuestas como variables aleatorias Bernoulli. Sea $x_i = (x_{i0}, x_{i1}, \ldots, x_{ik})$ que indica el i-ésimo conjunto de valores de k variables explicativas, $i = 1, \ldots, N;$ donde $x_{i0} = 1$. Cuando las variables explicativas son continuas, puede existir un conjunto diferente para cada sujeto. Se puede expresar el modelo de regresión como:

$$\pi(x_i) = \frac{\exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)}{\left[1 + \exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)\right]} \qquad \ldots\ldots (1.30)$$

Cuando más de una observación sobre Y ocurre en un valor fijo x_i, es suficiente registrar el número de observaciones n_i y el número de resultados "1". De esta manera se considera que Y_i se refiere a una suma de "éxitos" en lugar de respuestas binarias individuales. Las $\{Y_i, i = 1,\ldots, I\}$ son variables aleatorias binomiales independientes con $E(Y_i) = n_i\pi(x_i)$, donde $n_1 + \cdots + n_I = N$. La función L de probabilidad conjunta de $(Y_1,\ldots,Y_I)$, es proporcional al producto de las funciones binomiales.

$$L = \prod_{i=1}^{I} \pi(x_i)^{y_i} (1-\pi(x_i))^{n_i - y_i}$$

$$= \left\{ \prod_{i=1}^{I} (\pi(x_i))^{y_i} \right\} \left\{ \prod_{i=1}^{I} [1-\pi(x_i)]^{n_i - y_i} \right\} \qquad \ldots\ldots (1.31)$$

Al aplicarle el logaritmo a la ecuación (1.31) se obtiene:

$$= \sum_{i=1}^{I} y_i \log(\pi(x_i)) + \sum_{i=1}^{I} (n_i - y_i)\log[1-\pi(x_i)]$$

$$= \sum_{i=1}^{I} y_i \log(\pi(x_i)) + (n_i - y_i)\log[1-\pi(x_i)]$$

$$= \sum_{i=1}^{I} y_i \log\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] + n_i \log[1-\pi(x_i)] \qquad \ldots\ldots (1.32)$$

Sustituyendo $\pi(x_i)$ en la ecuación (1.32)

$$log(L(\beta)) = \sum_{i=1}^{I} y_i \left(\sum_{j=0}^{k} \beta_j x_{ij} \right) - n_i \, log\left[1 + exp\left(\sum_{j=0}^{k} \beta_j x_{ij} \right) \right]$$

$$log(L(\beta)) = \sum_{j=0}^{k} \sum_{i=1}^{I} (y_i x_{ij})\beta_j - \sum_{i=1}^{I} n_i \, log\left[1 + exp\left(\sum_{j=0}^{k} \beta_j x_{ij} \right) \right] \qquad \ldots\ldots (1.33)$$

Lo anterior depende únicamente de las cantidades binomiales a través del estadístico suficiente $\sum_{j=0}^{k} y_i x_{ij}$.

Para maximizar el logaritmo de *L* se deriva la ecuación de probabilidad con respecto a los elementos de β_a se tiene:

$$\frac{\partial L}{\partial \beta_a} = \sum_i y_i x_{ia} - \sum_i n_i x_{ia} \left[\frac{\exp\left(\sum_j \beta_j x_{ij} \right)}{1 + \exp\left(\sum_j \beta_j x_{ij} \right)} \right] \qquad \ldots\ldots (1.32)$$

Por lo que las ecuaciones de verosimilitud son:

$$\sum_i y_i x_{ia} - \sum_i n_i \hat{\pi}_i x_{ia} = 0, \;\; a = 0, \ldots, k \qquad \ldots\ldots (1.33)$$

Sea X la matriz $I \times (k+1)$ de valores $\{x_{ij}\}$. La ecuación (1.33) de verosimilitud tiene la forma:

$$X' y = X' n_i \hat{\pi}_i \qquad \ldots\ldots (1.34)$$

La ecuación anterior es similar a la utilizada en el método de mínimos cuadrados para el modelo de regresión lineal, es decir $(X'X)\beta = X'Y,$ donde $\hat{\beta} = (X'X)^{-1} X'Y.$ La ecuación (1.33) ilustra un resultado fundamental para los modelos lineales generalizados, que usan la liga canónica. Las ecuaciones de verosimilitud corresponden al estadístico suficiente para la estimación de sus valores esperados.

La matriz de información es el valor esperado negativo de la matriz de segundas derivadas. Bajo condiciones regulares, los estimadores de máxima verosimilitud de parámetros tienen una distribución normal para muestras grandes con matriz de covarianza igual a la inversa de la matriz de información. Para el modelo de regresión logística.

$$\frac{\partial^2 L(\beta)}{\partial\beta_a \partial\beta_b} = -\sum_i \frac{x_{ia}x_{ib}n_i \exp\left(\sum_j \beta_j x_{ij}\right)}{\left[1+\exp\left(\sum_j \beta_j x_{ij}\right)\right]}$$

$$= -\sum_i x_{ia}x_{ib}n_i\pi_i(1-\pi_i) \qquad \ldots\ldots (1.35)$$

La expresión (1.35) no es una función de $\{y_i\}$, la matriz teniendo elementos iguales no negativos de (1.33), e invirtiéndose se tiene:

$$\hat{\mathrm{C}}\mathrm{ov}(\hat{\beta}) = \{X'\mathrm{Diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)]X\}^{-1} \qquad \ldots\ldots (1.36)$$

Donde $\mathrm{Diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)]$ denota la matriz que tiene elementos $[n_i\hat{\pi}_i(1-\hat{\pi}_i)]$ en la diagonal principal de (1.36) son los errores estándar estimados del modelo.

En un conjunto X fijo, la varianza estimada del *logit* predicho $\hat{L} = X\hat{\beta}$ es $\hat{\sigma}^2(\hat{L}) = X\ Cov(\hat{\beta})\ X'$. Para muestras grandes $\hat{L} \pm z_{\alpha/2}\hat{\sigma}(\hat{L})$ es un intervalo de confianza para el *logit*.

Y como la ecuación que se tiene es no lineal se resuelve de forma iterativa, a través del método de Newton Raphson, para mayor detalle del cálculo iterativo consultar, Agresti 2002.

1.8 Interacciones y no linealidad

Silva (1995) muestra que en la regresión lineal múltiple, es posible incluir interacciones y términos no lineales en un modelo de regresión logística. Las interacciones pueden ser explicadas por términos adicionales que son incluidos en el componente sistemático del modelo, el cual muestra el producto de variables explicativas que interactúan. Por ejemplo, si hay una interacción entre x_1 y x_2, el componente lineal del modelo puede ser representado como:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \qquad \ldots\ldots (1.37)$$

Dicha maniobra procede en caso de que se piense que la influencia de x_1 sobre la variable de respuesta se modifique en función de cuál sea el valor de x_2 o viceversa.

Naturalmente, esta idea puede extenderse a más variables. Podrían incorporarse términos que involucren a tres o más de ellas. Una regla general que se ha dado es que si en un ajuste se incluye un término de cierto orden, se incluyan entonces todos los de orden inferior. Por ejemplo, si se incluye el término $x_1 x_2 x_3$, de orden 3; entonces se deben incluir todos los de orden 2; $x_1 x_2, x_1 x_3, x_2 x_3$, además de x_1, x_2, x_3.De no hacerse de este modo, la interpretación de los parámetros se torna cuando menos confusa.

En la ecuación (1.37) se aprecia que el coeficiente de x_1 no es ahora constante, si no que depende de x_2 o viceversa. Concretamente puesto que:

$$\beta_0 + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2$$

El grado en que influye el aumento de x_1 en una unidad (es decir, el cociente de momios asociado a x_1) es igual a: $\exp(\beta_1 + \beta_3 x_2)$. Por las propiedades de la función exponencial, se tiene que:

$$\exp(\beta_1 + \beta_3 x_2) = \exp(\beta_1)\exp(\beta_3 x_2) \qquad \ldots\ldots (1.38)$$

Agregando a las interacciones las relaciones no lineales también pueden ser explicadas por términos polinomiales incluidos en el componente sistemático del modelo. Por ejemplo, si la variable x_2 mostró una relación no lineal con la variable de respuesta, un término cuadrático como x_2^2 puede ser incluido en la ecuación. Un modelo que contiene dos variables explicativas, una de las cuales muestra una relación no lineal con el predictor lineal puede ser representado como:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 \qquad \ldots\ldots (1.39)$$

1.9 Interpretación de los coeficientes

Después del ajuste del modelo es necesario conocer lo que representan los coeficientes estimados, sobre todo enfocarse en 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 un suceso frente a que no ocurra dicho suceso. Así los coeficientes estimados asociados a las variables independientes representan la pendiente de la 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 vincula la variable independiente con el predictor lineal.

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x \qquad \ldots\ldots (1.51)$$

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

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

a) Variable independiente dicotómica

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 un éxito o un fracaso.

$$P(Y=1) = \pi(x) \qquad \text{y} \qquad P(Y=0) = 1-\pi(x)$$

Estos valores se pueden ver en la tabla 1.1 de 2×2

Tabla 1.1 Valores del modelo de RL cuando la variable independiente es dicotómica

	$x=1$	$x=0$
$y=1$	$\pi(1)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0)=\frac{e^{\beta_0}}{1+e^{\beta_0}}$
$y=0$	$1-\pi(1)=\frac{1}{1+e^{\beta_o+\beta_1}}$	$1-\pi(0)=\frac{1}{1+e^{\beta_o}}$

Fuente: Hosmer y Lemeshow, Applied Logistic Regression.

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

$$\frac{\pi(1)}{1-\pi(1)} \qquad \ldots\ldots (1.52)$$

Similarmente lo mismo para $x=0$

$$\frac{\pi(0)}{1-\pi(0)} \qquad \ldots\ldots (1.53)$$

El logaritmo de momios, es llamado el logit

$$\ln\left(\frac{\pi(1)}{1-\pi(1)}\right) \qquad \ldots\ldots (1.54)$$

$$\ln\left(\frac{\pi(0)}{1-\pi(0)}\right) \qquad \ldots\ldots (1.55)$$

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

$$RM = \frac{\dfrac{\pi(1)}{[1-\pi(1)]}}{\dfrac{\pi(0)}{[1-\pi(0)]}} \qquad \ldots\ldots (1.56)$$

Aplicando el logaritmo natural

$$\ln(RM) = \ln\left(\frac{\dfrac{\pi(1)}{[1-\pi(1)]}}{\dfrac{\pi(0)}{[1-\pi(0)]}}\right) \qquad \ldots\ldots (1.57)$$

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

$$RM = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} \qquad \ldots\ldots .(1.58)$$

Si a la ecuación (1.58) se le aplica el logaritmo se tiene que:

$$\ln(RM) = \ln\left(e^{\beta_1}\right) = \beta_1 \qquad \ldots\ldots (1.59)$$

De la ecuación anterior se obtiene 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 (RR), la cual indica cuánto 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 $\pi(x)$ expresa el riesgo de que se produzca el suceso y $1-\pi(x)$ de que no ocurra.

$$RR = \frac{\pi(1)}{1-\pi(1)} \qquad \ldots\ldots (1.60)$$

Teniendo que:

$$riesgo\ relativo = 1$$

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 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.

b) Variable independiente de escala nominal

Ahora supóngase que la variable independiente tiene $k > 2$ valores distintos. Por ejemplo podrían existir variables que denotan la raza, la religión que se profesa, el lugar de residencia entre otras. Cada una de estas variables tiene un número fijo de respuesta discreta y la escala de medida es nominal y se vuelve inapropiado incluirlas en el modelo como si fueran variables con escala de intervalo, ya que comúnmente se suelen representar con números que suelen ser sólo identificadores y carecen de significancia.

En general, si una variable de escala nominal tiene *k* posibles valores, entonces será necesario crear $k-1$ variables de diseño, a menos que la variable en cuestión tenga el mismo valor en todos los casos observados.

Si se supone que la j-ésima variable independiente, x_j, tiene k_j niveles. Las $k_j - 1$ variables de diseño serán denotadas como D_{ju} $u = 1,\ldots,k_j - 1$ y los coeficientes para estás variables de diseño serán denotados como B_{ju}, $u = 1,2,\ldots,k_j - 1$. Para este caso, la forma logit para un modelo con *p* variables y la *j*-ésima variable en escala nominal sería:

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{u=1}^{k_j - 1} \beta_{ju} D_{ju} + \beta_p x_p \qquad \ldots\ldots(1.63)$$

Sin embargo existen diferentes métodos para crear variables de diseño para variables independientes politómicas. La decisión de un método específico dependerá del alcance del análisis y la fase de desarrollo del modelo, a continuación se presentan algunos métodos.

Método de codificación de celda de referencia

Este método para especificar variables de diseño implica hacer a todas ellas iguales a cero para el grupo de referencia y habilitar una variable de diseño simple con valor igual a 1 para cada uno de los otros grupos. Como se ilustra en el tabla 1.2:

Tabla 1.2. Ilustración de la codificación de variables de diseño usando el método de codificación de celda de referencia

Lugar de origen	D_1	D_2
Europa	0	0
América	1	0
Otro	0	1

Fuente: Hosmer y Lemeshow, Applied Logistic Regression.

Este método de codificación de variables de diseño es el comúnmente más usado. La razón primordial de la utilización del método es el interés por estimar el riesgo de un grupo "expuesto" respecto a un grupo "control" o "no expuesto".

Desviación de codificación de medias

Este segundo método de codificación de variables de diseño, es utilizado con mayor frecuencia en análisis de varianza y regresión lineal que en la regresión logística. Esto expresa tanto el efecto como la desviación de la "media del grupo" de la "media global". En el caso de la regresión logística la "media del grupo" es la logit del grupo y la "media global" es el promedio logit. El promedio logit se obtiene, ajustando el valor de todas las variables de diseño igual a -1 para una de las categorías y utilizando la codificación 0, 1 para las categorías restantes. Como se ilustra en la tabla 1.3.

Tabla 1.3. Ilustración de la codificación de variables de diseño usando el método de codificación de medias.

Lugar de origen	D_1	D_2	D_3
Europa	-1	-1	-1
América	1	0	0
África	0	1	0
Otro	0	0	1

Fuente: Hosmer y Lemeshow, Applied Logistic Regression.

Sin embargo la interpretación de los coeficientes no es tan sencilla como en la situación cuando un grupo referente es usado. La exponenciación de los coeficientes estimados produce el cociente de momios para un grupo en particular a la media geométrica de los momios. Sin embargo el número resultante no es precisamente el cociente de momios porque la cantidad en el numerador y el denominador no representa los momios para dos distintas categorías. La exponenciación de la estimación de los coeficientes expresa el momio relativo a un promedio.

c) Variable independiente continua

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

Bajo la suposición de que el logit es lineal en la variable independiente, la ecuación para el logit es:

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

El coeficiente de la pendiente β_1 da el cambio en el logit para un incremento en "1" unidad en *x*, esto es:

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

Para cualquier valor de *x.*

Muy a menudo el valor de "1" no será muy interesante. Por ejemplo, un incremento de 1 año en edad o de 1mm Hg. en presión sanguínea puede ser muy pequeño para ser considerado importante, en comparación con un cambio de 10 años o de 10mm Hg. Por lo tanto, para la proporción que se incrementarse *c* unidades en *x*, queda:

$$\beta_1 = g(x+c) - g(x) = c\beta_1$$

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

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

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

1.10 Intervalos de Confianza

En el modelo de regresión logística, también se pueden estimar intervalos de confianza que ayuden a proporcionar un rango alrededor de los coeficientes β, para el cambio de momios y para π_j, en donde se espera que esté contenido a un determinado nivel de certeza el valor "verdadero" (poblacional).

El cálculo de los intervalos de confianza está basado en una aproximación al método de máxima verosimilitud.

***Intervalos de confianza para* los parámetros β**

Un intervalo de confianza para los parámetros β es un intervalo de confianza para el cambio en el logaritmo de las proporciones. El intervalo para β es obtenido como:

$$\hat{\beta} \pm z_{\alpha/2} s_{\beta} \qquad \ldots\ldots (1.40)$$

Donde $z_{(1-\alpha/2)}$ denota la desviación estándar de una distribución de probabilidad normal con un área de $\alpha/2$. En este caso z se usa en lugar de t, la cual es usada para intervalos de confianza en regresión lineal. Esto es porque no existe normalidad en la regresión logística como es asumida en la regresión lineal. S_{β} se refiere al error estándar $\left(\sqrt{Var(\hat{\beta}_i)}\right)$ del estimador correspondiente. Un intervalo para $\hat{\beta}$ debería ser usado cuidadosamente, especialmente si el tamaño de la muestra no es grande.

Intervalo de confianza para el cambio de momios

Un intervalo de confianza para el cambio en las proporciones es similar a un intervalo de confianza para el cambio en los momios. Por ejemplo para un sólo regresor,

$\log\left[\frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1\right]$, se tiene $\frac{\pi}{1-\pi} = exp(\beta_0 + \beta_1 x_1)$. Esto es $exp(\beta_1)$ representa el cambio en los momios por unidad de cambio en x.

Un intervalo de confianza para $exp(\beta_1)$ es un intervalo de confianza para el cambio en los momios. Una forma para obtener dicho intervalo es:

$$exp\left[c\hat{\beta}_1 \pm z_{\alpha/2}(cs_{\beta_1})\right]$$

En donde *c* representa el incremento en *x* para el intervalo deseado. Se debe notar que el intervalo no es una función de *x*, lo cual parece absurdo. Esto es, porque el resultado está basado en el hecho de que el incremento en el error de Y es independiente al valor de X.

Intervalo de confianza para π

Además de conocer el intervalo de confianza para el cambio en los momios es importante conocer si el riesgo está siendo incrementado en *x*, se necesita conocer cual es el riesgo de un valor de *x* dado.

Dado que se conoce $\hat{\pi}_i$ que es una estimación obtenida de un subconjunto de la población objetivo, es necesario un intervalo de confianza para π_i. La aproximación más obvia a usarse debería ser $\hat{\pi}_i \pm zs_\pi$ pero $0 \le \pi \le 1$, por lo que $\hat{\pi}$ no debería ser aproximada a una distribución normal para muchos valores de π. Este método para obtener el intervalo de confianza no es el más apropiado. Aunque si se tiene una muestra muy grande este método es de gran utilidad. Supóngase que existe un sólo regresor. Como $\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 x_1$, un intervalo de confianza para $\log\left[\frac{\pi}{1-\pi}\right]$ debería estar dado por $\beta_0 + \beta_1 x \pm zs_{\beta_0+\beta_1 x}$. Al tomar el exponente se tiene:

$$\text{Límite inferior } \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x - zs\sqrt{h_{ij}}\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x - zs\sqrt{h_{ij}}\right)}$$

$$\text{Límite superior } \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x + zs\sqrt{h_{ij}}\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x + zs\sqrt{h_{ij}}\right)}$$

Donde $Var\left(\hat{Y} \setminus X_i\right) = \sigma^2 h_{ij}$, donde h_{ij} es definida previamente

Intervalo de confianza para ŷ fija

Estos intervalos son calculados para el predictor lineal, $\text{logit}(\pi_i)$, y son transformadas en un intervalo de confianza para π.

$$\hat{y} = \text{logit}(\pi_i) = \text{logit}(\hat{\pi}_i) \pm 1.96\text{ASE} \qquad \ldots\ldots (1.41)$$

Donde ASE es el error estándar del $\text{logit}(\hat{\pi}_i)$ y el valor 1.96 es la aproximación normal de la distribución de la muestra de un estadístico t con un nivel del 95% de confianza La dificultad para el cálculo de los intervalos de confianza para el modelo logístico radica en determinar el ASE asociado con la predicción. Cuando los intervalos de confianza no se pueden obtener directamente a través de software, una de las dos siguientes métodos pueden ser usadas para calcular el ASE para $\text{logit}(\hat{\pi}_i)$:

Método 1

Si se puede obtener la matriz de varianza-covarianza, el ASE puede ser calculado usando:

$$\text{ASE para logit}(\pi_i) = \sqrt{\sum_{j=1}^{k} x_{j0}^2 \cdot Var\left(\hat{\beta}_j\right) + 2\sum_{j=1}^{k}\sum_{n=1}^{j} x_{n0} \cdot x_{j0} \cdot Cov\left(\hat{\beta}_n, \hat{\beta}_j\right)} \qquad \ldots\ldots (1.42)$$

Donde $Var(\hat{\beta}_j)$ y $Cov(\hat{\beta}_n, \hat{\beta}_j)$ son valores obtenidos de la matriz varianza-covarianza y x es el valor de la variables explicativa.

Como se puede ver, el cálculo del ASE con este método es complejo y necesita ser calculado para cada combinación de variable explicativas que interesen.

Método 2

Si no se puede obtener la matriz de varianza-covarianza, el ASE para $\text{logit}(\pi_i)$ puede ser calculado usando un procedimiento con el cual se transforma cada variable explicativa y se recalcula el modelo. Usando este método se puede obtener una estimación del ASE para una combinación particular de valores de x. El procedimiento es el siguiente:

1. Crear nuevas variables explicativas, es decir, elegir un valor y restarlo a cada uno de los casos de las variables explicativas. Supóngase en el caso del modelo simple para un valor particular de x, a cada valor original de la variable explicativa se le resta el valor de 30, es decir, ahora cada valor es igual a $x = 30$.
2. Recalcular el modelo usando las nuevas variables explicativas, en este caso $x = 30$.
3. El ASE asociado con la constante da una estimación del ASE asociado para el $\text{logit}(\pi_i)$.
4. Finalmente el intervalo de confianza para $\text{logit}(\pi_i)$ puede ser calculado usando la ecuación (1.41).

1.11 Pruebas de significancia

La prueba de significancia del coeficiente de una variable en cualquier modelo obliga a preguntar. ¿El modelo que incluye la variable explicativa en cuestión dice más sobre la variable dependiente que un modelo que no incluye esta variable? Esta pregunta tiene respuesta al comparar los valores observados de la variable dependiente con los predichos

por cada modelo. Matemáticamente la función usada para comparar los valores observados y predichos depende de un problema particular. Si los valores predichos con la variable en el modelo son mejores o más precisos en algún sentido, que cuando la variable no está en el modelo, entonces se dice que la variable explicativa en cuestión es "significativa".

En el modelo de regresión logística después de la estimación de los coeficientes, la primera observación al modelo ajustado comúnmente concierne al cálculo de la significancia de las variables en el modelo. Esto usualmente involucra la formulación de pruebas de hipótesis para determinar si la variable independiente en el modelo tiene una asociación con la variable de respuesta, para estos casos se tiene la prueba de Wald y la prueba del cociente de verosimilitud, ambas requieren del cálculo de la estimación de máxima verosimilitud.

a) Prueba del cociente de verosimilitud

La estadística de prueba del cociente de verosimilitud da una medida de la devianza del modelo, que se basa en el logaritmo de la función de verosimilitud comparando los valores observados y estimados de la variable de respuesta, también puede ser utilizada como la estadística de bondad de ajuste.

El estadístico de bondad de ajuste es comúnmente citado como: -2 veces el logaritmo de verosimilitud $(-2LL)$, se aproxima a una distribución Ji-cuadrada, esto permite evaluar la significancia. La interpretación de $-2LL$ es sencilla, si su valor es pequeño, es decir es menor al valor en tablas, dependiendo de la presición deseada, entonces el modelo se ajusta muy bien y si es igual a cero indica, que no hay devianza.

La forma general la prueba del cociente de verosimilitud, generalmente se encuentra expresada de la forma $0 < -2\log\left(\frac{\hat{L}_0}{\hat{L}_1}\right) < \infty$, donde $\hat{L}_0$ es el valor máximo de la función de verosimilitud para el modelo nulo y $\hat{L}_1$ es el valor máximo de la función de verosimilitud para el modelo ajustado, cumpliéndose que $\hat{L}_0 < \hat{L}_1$, está estadística de prueba se distribuye

aproximadamente como una $\chi^2_{(m-k)}$, con $m-k$ grados de libertad, donde m es igual al número de parámetros del modelo ajustado y k es igual al número de parámetros del modelo nulo.

El indicador de ajuste se obtiene al sacar la diferencia del modelo nulo con respecto al modelo ajustado, esto es:

$$-2LL_{dif} = (-2LL_0) - (-2LL_1) \qquad \ldots\ldots (1.43)$$

Siendo el valor de $-2LL_0$ la medida de la devianza del modelo nulo $\text{logit}(\pi_i) = \beta_0$ y $-2LL_1$ el valor del modelo ajustado $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

El impacto que tienen las variables individuales o variables en grupo sobre el ajuste del modelo puede ser determinado comparando el ajuste de modelos anidados. La cantidad por la cual $-2LL$ decrece cuando variables adicionales son añadidas al modelo, indica el tamaño del efecto que estas variables tienen. La significancia del cambio en $-2LL$ está determinada por la prueba χ^2 con grados de libertad igual a la diferencia en el número de términos entre los dos modelos $-2LL$ puede ser derivado usando la siguiente ecuación

$$-2LL_{dif} = (-2LL_p) - (-2LL_{p+q}) \qquad \ldots\ldots (1.44)$$

Donde p es el modelo anidado más pequeño y $p+q$ es el modelo que contiene q variables explicativas, se evalúa el efecto individual y en grupo de variables sobre el modelo además de ser usado para seleccionar el modelo. Al igual que en la regresión lineal, la cantidad de devianza en el modelo puede ser minimizada incluyendo algunas variables explicativas. Maximizar el poder explicativo del modelo en esta forma no siempre es benéfico ya que se

puede caer en la inclusión de variables irrelevantes que pueden dar muy poco poder explicativo pero que pueden incrementar el error estándar asociado con la predicción. Por esta razón es útil reducir el número de variables en el modelo eliminando las que no tengan influencia significativa.

El estadístico $-2LL_{dif}$ también sirve como criterio para decidir que variables pueden ser eliminadas dentro del modelo de regresión logística. Usando la ecuación (1.44) los modelos de regresión anidados pueden ser comparados con el cálculo del efecto que variables explicativas individuales o en grupo tienen en la variable de respuesta.

b) pseudo $-$ R^2

La proporción de incremento del logaritmo de la función de verosimilitud del modelo propuesto en relación al modelo nulo se define como $pseudo - R^2$

$$pseudo - R^2 = \frac{1-\left(\frac{L_0}{L_1}\right)^{2/n}}{1-L_0^{2/n}} = \frac{L_1^{2/n} - L_0^{2/n}}{L_1^{2/n}\left(1-L_0^{2/n}\right)} \qquad \ldots\ldots (1.48)$$

Donde L_0 denota la devianza para el modelo nulo y L_1 es la devianza para el modelo ajustado y *n* es el tamaño de la muestra. El valor mínimo que puede tomar $pseudo - R^2$ es cero cuando el ajuste es malo, esto sucede cuando $L_1 = L_0$, y el valor máximo es uno cuando el ajuste es bueno y esto sucede cuando $L_1 = 1$. La definición de $pseudo - R^2$ no es prueba formal de significancia. Otra definición fue sugerida en 1974 por McFadden, y es la siguiente:

$$pseudo - R^2 = 1 - \left(\frac{LL_0}{LL_1}\right)^{2/n} \qquad \ldots\ldots (1.49)$$

Que al igual que la anterior no es una prueba formal de significancia. La definición que maneja la paquetería de STATA para el cálculo de la estadística $pseudo - R^2$ es la siguiente:

$$pseudo - R^2 = \frac{LL_1 - LL_0}{LL_0} \qquad \ldots\ldots (1.50)$$

Esta estadística tiene varios inconvenientes:

1. Su interpretación no es muy intuitiva, a diferencia de la R^2 del modelo de regresión lineal.
2. Se interpreta en términos de la función de verosimilitud que no es fácil de expresarse.
3. Tiende a reportar valores más bajos que la R^2 que se obtiene en los modelos lineales
4. Esto puede ser desconcertante en ocasiones al presentar datos por la familiaridad que existe con la regresión lineal.
5. Existen muchas formas de calcular $pseudo - R^2$, lo cual puede dar diferentes resultados para un mismo conjunto de datos

Por lo anterior no se recomienda el uso de dicha estadística excepto cuando se utilice como un criterio más en la selección de modelos alternativos, sin embargo se considera necesario mencionarla porque está incluida en el paquete de STATA.

c) Prueba de la estadística de Wald

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

$$H_o = \beta_j = 0$$

vs

$$H_a = \beta_j \neq 0 \qquad \text{Para } j = 0,12,\ldots, p$$

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:

$$W = \frac{\hat{\beta}_j - \beta_j}{S(\hat{\beta}_j)} \qquad \ldots\ldots.(1.45)$$

La estadística de Wald se distribuye como una normal estándar, cuando el 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.

$$\frac{\hat{\beta}_j}{S(\hat{\beta}_j)} \qquad \ldots\ldots (1.46)$$

La regla de decisión: se rechaza H_o al nivel de significancia α si $W > Z_{1-\alpha}$ Un estadístico de Wald significativo sugiere que la variable explicativa tiene un efecto en la variable de respuesta, sin embargo debe ser usado con precaución ya que tiende a exagerar la significancia de las variables, las cuales tienen coeficientes altos, y hay que considerar que puede también ser poco confiable para muestras pequeñas.

1.12 Regresión logística multinomial

El análisis logístico binario es ideal cuando la variable independiente tiene únicamente dos categorías, pero ¿qué sucede si hay más de dos categorías en la variable independiente? En algunos casos, podría ser razonable colapsar las categorías para tener únicamente dos y

aplicar el modelo logístico binario, pero quizás esta estrategia podría implicar perder información relevante en el estudio o bien la variable en estudio podría carecer de sentido con lo cual se llegarían a conclusiones totalmente erróneas.

La aplicación del modelo logit multinomial se basa en determinar los efectos de variables explicativas en una elección sometida a un conjunto de opciones por ejemplo la elección del candidato presidencial por partido político (PRI, PAN, PRD) o elegir una carrera en la facultad de ciencias (Actuaría, Biología, Física, Matemáticas).

Modelo logit multinomial general

La idea básica del modelo logit multinomial es comparar dos resultados o selecciones al mismo tiempo. La base principal para la construcción del modelo logit multinomial es el llamado "*baseline"* o riesgo.

Sea Y una variable de respuesta con *J* categorías $(j = 1,\ldots,J)$, el modelo logit para una variable de respuesta nominal simultáneamente describe el momio para todos las *J* categorías tomadas de dos en dos, teniendo como elección $J-1$.

Se tiene una variable de respuesta multinomial con probabilidades de respuesta $(\pi_1,\ldots,\pi_J)$ con *J* categorías, denotando como:

$$\pi_j(x) = P(Y = j|x) \qquad j = 1,\ldots J$$

Con *x* fijo de las variables explicativas, sea el logit de riesgo la comparación de la j-ésima categoría con la primera es decir:

$$BL_j = \log\left[\frac{P_j(y=j)}{P_1(y=1)}\right] = \log\left(\frac{\pi_j}{\pi_1}\right) \qquad j = 2,\ldots,J \qquad \ldots\ldots (1.64)$$

Donde P_j y P_1 indican las probabilidades para la j-ésima y primer categoría respectivamente. El uso de la primera categoría como *baseline* es arbitrario, cualquier otra categoría pudo haber sido *baseline*.

Para el caso de una variable independiente *x* con *I* categorías, es decir $(X = 1,\ldots, I)$ para cada valor de *x* el logit-*baseline* es:

$$BL_{ij} = \log\left[\frac{P_j\left(y = j|x = i\right)}{P_1\left(y = 1|x = i\right)}\right] = \log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) \qquad \ldots\ldots (1.65)$$

La ecuación anterior representa un modelo saturado, la estimación de la ecuación se obtiene a partir de:

$$\log\left(\frac{F_{ij}}{F_{1j}}\right) = \log\left(\frac{f_{ij}}{f_{1j}}\right) \qquad \ldots\ldots (1.66)$$

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

La ecuación (1.66) se puede expresar en términos del modelo lineal generalizado:

$$BL_{ij} = \sum_{i=1}^{I} \log\left(\frac{F_{ij}}{F_{i1}}\right) \cdot I\left(x = i\right) \qquad \ldots\ldots (1.67)$$

Donde $I(x=i)$ es la función indicadora, con $i=1$ como verdadero y 0 en otro caso, con una variable dummy tomando la primera categoría como referencia, la ecuación (1.67) queda escrita como:

$$BL_{ij} = \alpha_j + \sum_{j=1}^{J} \beta_{ij} \cdot I(x=i) \qquad x>1 \qquad \ldots\ldots (1.68)$$

Donde:

α_j es el logit-*baseline* para $x=1$

β_{ij} es la diferencia entre logit-*baseline* de $x=1$ y $x=i$

α_j y β_{ij} pueden ser estimados separadamente para todo *i, j.* Para modelos no saturados las estimaciones simultáneas producen resultados diferentes.

Modelo logit mulinomial estándar

El modelo logit multinomial pueden ser visto como una extensión del modelo logit binario. Sea Y una variable de respuesta con J categorías, es decir $j=1,\ldots,J$. Sea p_{ij} la probabilidad de que un individuo *i* caiga en la categoría j. El modelo entonces es:

$$\log\left(\frac{p_{ij}}{p_{iJ}}\right) = \beta_j x_i \qquad j=1,\ldots,J-1 \qquad \ldots\ldots (1.69)$$

Donde:

x_i es el vector columna de las variables que describen al individuo *i*

β_j es el vector reglón de los coeficientes para la categoría *j*

Se observa que cada categoría *j* es comprada con el último valor de la variable categórica, la ecuación anterior es una generalización del modelo logit binario y puede ser descrita como:

$$p_{ij} = \frac{e^{\beta_j x_i}}{\sum_{k=1}^{J} e^{\beta_k x_i}} \qquad \ldots\ldots (1.70)$$

Con $\sum_{j=1}^{J} p_{ij} = 1$ para cualquier *i* con la usual normalización $\beta_1 = 0$ lo que significa que en la ecuación (1.70) se tiene:

$$P[y_i = j|x_i] = \frac{e^{\beta_j x_i}}{1 + \sum_{k=1}^{J-1} e^{\beta_k x_i}} \qquad j = 1,\ldots,J-1 \qquad \ldots\ldots (1.71)$$

$$P[y_i = 1|x_i] = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\beta_k x_i}} \qquad \ldots\ldots (1.72)$$

Se puede ver que cuando $J = 2$, se estima un sólo conjunto de parámetros correspondientes al resultado $y = 2$ con la primera categoría $(y = 1)$, en este caso se habla del modelo logit binario, aunque en el modelo logit binario la variable dependiente es comunmente codificada como $(0, 1)$ en lugar de $(1, 2)$, por tanto el modelo logit binario puede ser visto como un caso especial del modelo logit multinomial.

Una alternativa para codificar la variable de respuesta categórica es tomar valores $0,\ldots,J-1$ en lugar de $1,\ldots,J$, esto hace que el modelo logit multinomial se parezca más al modelo logit binario. Con la codificación de $0,\ldots,J-1$, se puede seguir la convención de

tomar como primer categoría a $(y=0)$ como la categoría de referencia, así que se tiene a β_0 en lugar de β_1.

1.13 Estimación de los parámetros

La estimación de los parámetros para el modelo logístico multinomial se realiza a través del método de máxima verosimilitud.

La construcción de la función de verosimilitud se realiza de la siguiente manera, se construyen J variables codificadas como 0 y 1 para indicar al grupo al que pertenece cada observación. Se debe tomar en cuenta que estas variables son introducidas solamente para explicar la función de verosimilitud y no son construidas ni usadas en el modelo de regresión multinomial. Sea $Y=k$ entonces $Y_k=1$ y 0 para el resto de las variables. Este resultado es uno y solamente uno para cada observación, en donde la $\sum_{j=1}^{J} Y_j = 1$. La función de verosimilitud para *n* observaciones queda:

$$l(\beta)=\prod_{i=1}^{n}\left[\pi_0(x_i)^{Y_{0i}}+\cdots+\pi_k(x_i)^{Y_{ki}}+\cdots+\pi_J(x_i)^{Y_{Ji}}\right] \qquad \ldots\ldots (1.73)$$

Aplicando logaritmo y factorizando $\sum Y_{ji}=1$

$$L(\beta)=\sum\left[Y_1 g_1(x_i)+\cdots+Y_k g_k(x_i)+\cdots+Y_J g_J(x_i)-\log\left(1+e^{g_1(x_i)}+\cdots+e^{g_J(x_i)}\right)\right] \quad \ldots. (1.74)$$

Las ecuaciones de verosimilitud se obtienen de la primera derivada parcial de $L(\beta)$ con respecto a cada $2(p+1)$ parámetros desconocidos. La forma general de las ecuaciones es la siguiente:

$$\frac{\partial L(\beta)}{\partial \beta_{jk}} = \sum_{i=1}^{n} x_{ki}\left(y_{ji} - \pi_j(x_i)\right) \qquad \ldots\ldots (1.75)$$

Para $j = 1,\ldots,J$ y $k = 0,1,2,\ldots,p$, se renombra $x_{0i} = 1$ para cada sujeto.

Hosmer y Lemeshow se detalla la forma de obtener el máximo estimador de verosimilitud de $\hat{\beta}$, igualando las ecuaciones a cero y resolviendo para $\hat{\beta}$, de manera iterativa.

1.14 Interpretación de los coeficientes

En el modelo logístico multinomial, el cociente de momios entre las categorías *j* y 1 para una *i* dada es:

$$\frac{\pi_{ij}}{\pi_{i1}} = \exp\left(x_i \beta_j\right) \qquad j = 2,\ldots,J \qquad \ldots\ldots (1.76)$$

El logit, es la función lineal de x_i

$$log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = x_i \beta_J \qquad j = 2,\ldots,J \qquad \ldots\ldots .(1.77)$$

La interpretación de los coeficientes en el modelo logístico multinomial no es tan sencilla como en la regresión logística binaria.

Por ejemplo supóngase que β_j es positiva en la ecuación (1.77). Entonces una unidad de incremento en x_i es causada porque se incrementa $\log\left(\frac{\pi_{ij}}{\pi_{i1}}\right)$, es decir, se incrementa en β_j unidades.

$\left(\frac{\pi_{ij}}{\pi_{i1}}\right)$ puede incrementarse cuando π_{ij} y π_{i1} decrecen. Un valor positivo de β_i no significa precisamente que una unidad incrementada en x_i incrementa en π_{ij}.

En la regresión multinomial, los efectos de las variables predictivas en $\log\left(\frac{\pi_{ij}}{\pi_{i1}}\right)$ y $\left(\frac{\pi_{ij}}{\pi_{i1}}\right)$ pueden ser engañosos, ya que π_{ij} puede estar en la dirección opuesta.

La interpretación de β_{jk} como el logaritmo de momios cuando x_k es una variable continua requiere una comparación de $x_k = x_k^0$, el cual es un valor arbitrario de x_k:

$$log\left[\frac{\frac{\left(\pi_j \middle| x_k = x_k^0 + 1\right)}{\left(\pi_1 \middle| x_k = x_k^0 + 1\right)}}{\frac{\left(\pi_j \middle| x_k = x_k^0\right)}{\left(\pi_1 \middle| x_k = x_k^0\right)}}\right] = \beta_{jk} \qquad \ldots\ldots (1.78)$$

La relación anterior interesa por el contraste entre las categorías *j* y *k*.

Lo anterior se puede extender a contrastar cualesquiera dos categorías *j* y *k* incluyendo la aplicación de coeficientes para las categorías *j* y *k*. La ecuación (1.76), puede ser extendida a:

$$\frac{\pi_{ij}}{\pi_{ik}} = \exp\left[x_i\left(\beta_j - \beta_k\right)\right] \qquad \ldots\ldots (1.77)$$

Para una variable explicativa x_i la diferencia entre los coeficientes $\left(\beta_{ji} - \beta_{ki}\right)$ determina la dirección del cambio en las proporciones dentro de las categorías *j* y *k*. Una diferencia positiva significa que cuando x_i incrementa, hay una proporción más grande de observar la alternativa *j* que la *k*.

1.15 Pruebas de significancia

En el modelo logístico multinomial la prueba de cociente de verosimilitud es la más conveniente para trabajar.

Se tienen dos modelos de regresión logística L_1 y L_2, con la misma variable de respuesta pero diferentes variables predictivas, L_2 tiene todas las variables predictivas incluidas en el primer modelo más una más, es decir L_1 está contenido en L_2, entonces $L_1 < L_2$.

Un estadístico para el cual la distribución muestral es conocida es:

$$-\log\left(\frac{L_1}{L_2}\right)^2 = -2\log\left(\frac{L_1}{L_2}\right) = -2\left[\log(L_1) - \log(L_2)\right] \qquad \ldots\ldots (1.78)$$

Se usa (1.78) para probar que el segundo modelo se ajusta a los datos significativamente mejor que en el segundo modelo. La prueba es una χ^2, es decir $-\log\left(\frac{L_1}{L_2}\right)^2$ se distribuye como una χ^2 con grados de libertad igual a la diferencia entre el número de coeficientes a ser estimados en los modelos.

Capítulo II

Encuesta Nacional de Evaluación del Desempeño. Base Metodológica

El presente capítulo muestra las características principales de la Encuesta Nacional de Evaluación del Desempeño (ENED) 2002-2003, con el fin de introducir al lector en las particularidades de este instrumento. También se resalta que no se ocuparán todas las variables medidas en la encuesta, solamente se consideran las siguientes: lugar de residencia, tipo de localidad, características especificas de los encuestados como edad, género, grado de estudio, ocupación, estado civil, y variables que abordan preguntas relacionadas al estado de ánimo del encuestado como sentimiento de tristeza y duración del mismo, perdida de interés en las cosas que realiza, falta de apetito o lo contrario y fatiga.

Posteriormente se presenta el objetivo del presente estudio y se definen de manera específica los síntomas característicos para diagnosticar depresión así como la importancia que ha tomado en los últimos años el estudio de los trastornos mentales, incluyéndose dentro de éstos los episodios depresivos y finalmente se muestran los resultados de algunos estudios realizados.

2.1 Base metodológica de la encuesta

La encuesta a utilizar en el análisis es una encuesta de salud que emplea la Organización Mundial de la Salud (OMS) con el nombre de *World Health Survey* (WHS), esta encuesta aborda varios temas de salud y la capacidad de respuesta de los sistemas de salud de los países participantes. En México, la WHS fue realizado por primera vez en el año 2001, por el Instituto Nacional de Salud Pública (INSP) con el nombre de Encuesta de Salud y Capacidad de Respuesta del Sistema de Salud Mexicano, la cual fue una muestra probabilística con representatividad nacional de aproximadamente 5 000 hogares. Para el año 2002, la Secretaría de Salud de México (SSA) decidió

ampliar los alcances del estudio realizado en 2001 y elaboró una encuesta que abarcaba un mayor número de temas e incorporaba nuevos elementos; dicha encuesta se dio a conocer con el nombre de Encuesta Nacional de Evaluación del Desempeño (ENED) y esta es la encuesta base del análisis.

Objetivo de la encuesta

Su objetivo es proporcionar información adecuada para apoyar la toma de decisiones orientadas al mejoramiento de la calidad en los servicios de salud, conocer el estado de salud de la población, así como su percepción sobre ésta y su experiencia con los proveedores de servicios de salud, construir un marco basal que muestre la morbilidad, la mortalidad y las barreras de acceso a los servicios mencionados, así como la calidad de la atención.

Descripción de la encuesta

Se aplicaron dos cuestionarios: uno para el hogar y otro individual, estos se encuentran en el Anexo A y únicamente incluye las partes de los cuestionarios que se usaron; el primero contiene información sobre el número de habitantes en el hogar, revisión general y demográfica de los integrantes del mismo, seguridad médica, así como indicadores de ingresos y gastos del hogar.

El cuestionario individual se aplicó a un informante mayor de 18 años de edad, miembro del hogar seleccionado. El informante permitió captar los datos generales de cada informante seleccionado dentro de la vivienda, su percepción sobre la salud mediante la aplicación de viñetas de descripción y valoración de estados de salud. También capta los factores de riesgo a los que se encontró sometido y las defunciones que se presentaron en el hogar; obtuvo información referente a la capacidad de respuesta del sistema de salud y del alcance de los objetivos sanitarios, mediante la percepción de los usuarios, lo que conlleva a la evaluación de la calidad de la atención; ésto, nuevamente, con la aplicación de las viñetas correspondientes a esta sección. Este cuestionario estuvo integrado por nueve secciones: Las características sociodemográficas del entrevistado, la descripción del estado de salud, valoración del estado de salud, factores de riesgo en la salud del entrevistado como tabaco, alcohol,

actividad física y medioambientales o de agua y saneamiento, mortalidad de algún miembro del hogar, cobertura sobre cuestiones de salud que afectan a los integrantes del hogar, así como el tipo de asistencia y tratamiento que recibieron, la capacidad de respuesta del sistema de salud, la asistencia médica, evaluación general de los sistemas de salud, gastos de bolsillo, tiempos de traslado y espera, confidencialidad, autonomía, comunicación, elección del proveedor de salud y comodidades básicas, la importancia que tienen para la población los objetivos de los sistemas de salud.

Diseño muestral

La encuesta cuenta con un diseño muestral probabilístico, polietápico, estratificado y por conglomerados. Los estratos considerados por la ENED fueron ciudad o área metropolitana, complemento urbano y rural. La definición de estos estratos aparece en el tabla 2.1.

Tabla 2.1. Estratos de localidades

Estrato de localidad	Descripción
Ciudad o área metropolitana	Ciudades o áreas metropolitanas o ciudades con más de 100,000 mil habitantes o capitales de los estados.
Complemento urbano	Localidades de 2,500 a 99,999 habitantes
Rural	Localidades menores a 2,500 habitantes

Fuente: Elaboración propia

El tamaño de muestra para cada estrato se asignó de manera proporcional al número de habitantes dentro del estrato.

Tamaño de muestra

Para cumplir con los objetivos de la ENED se consideró que el indicador de importancia más pequeño era una proporción de aproximadamente 9%. Además, considerando que la encuesta debería permitir obtener estimaciones estatales con un error relativo máximo de 25%, una confianza de 95%, una tasa de respuesta de 15% y un efecto de diseño de

1.7, se determinó un tamaño de muestra de aproximadamente 1 243 hogares usando la fórmula:

$$n = \frac{Z^2_{\alpha/2}(1-P)deff}{r^2 P\, TR}$$

Donde:

n = Tamaño de muestra

P = Proporción a estimar

$Z^2_{\alpha/2}$ = Cuantil de distribución asociado a un nivel de confianza deseado α

r = Error relativo máximo que se está dispuesto a aceptar

deff = Efecto de diseño, que es la pérdida o ganancia en la eficiencia del diseño, por tratarse de un diseño compuesto.[2]

TR = Tasa de respuesta esperada

Este tamaño de muestra fue redondeado a 1 250 viviendas por estado, lo que dio por resultado un tamaño de muestra de 40 000 viviendas a escala nacional y se calcularon estimadores para cada una de las entidades federativas como dominios de estudio, es decir, los resultados obtenidos por la encuesta son representativos para poder hacer inferencia estadística, por lo que pueden ser expandidos a la población de cada estado. El levantamiento de la información se llevó a cabo entre noviembre de 2002 y marzo de 2003.

Esquema de selección de la muestra

El esquema de selección se manejó de dos maneras diferentes, según se tratara de una área urbana o rural. Por área urbana se entiende las ciudades o áreas metropolitanas y las del complemento urbano, como se definieron en la tabla 2.1

[2] Cociente de la varianza del diseño utilizado entre la varianza del muestreo aleatorio simple para un mismo tamaño de muestra.

Esquema de selección en áreas urbanas

• En la primera etapa se seleccionaron *k* Áreas Geoestadísticas Básicas (AGEB)[3] con probabilidad proporcional al tamaño (PPT).
• Para cada una de las AGEB seleccionadas se eligieron siete manzanas usando muestreo con PPT.
• Posteriormente, en cada una de las manzanas seleccionadas se escogieron siete viviendas por medio de muestreo sistemático con arranque aleatorio.
• Finalmente, en cada uno de los hogares de las viviendas seleccionadas se eligió al entrevistado entre todos los informantes considerados adecuados.

La selección del entrevistado se llevó a cabo de acuerdo con las tablas de Kish.

En cada una de las etapas de selección el tamaño de la unidad de muestreo se definió como el número de habitantes en ella.

Las constantes *m* y *k* quedaron en función del tamaño del estrato en la entidad.

Esquema de selección en áreas rurales

• En la primera etapa se seleccionaron *k* AGEB rurales con probabilidad proporcional al tamaño.
• Para cada AGEB seleccionada, mediante muestreo sistemático, se eligieron cinco segmentos de 10 viviendas cada uno.
• En cada uno de los hogares de las viviendas que integraban los segmentos seleccionados se usaron las tablas de Kish para escoger al entrevistado de entre los informantes adecuados en la vivienda.

En este caso la constante *k* estuvo en función del tamaño del estrato en la entidad.

[3] Área Geoestadística Básica. Constituye la unidad básica del Marco Geoestadístico Nacional y se clasifican en dos tipos. Área Geoestadística Básica Urbana o Área Geoestadística Básica Rural. La primera se define como el área geográfica ocupada por un conjunto de manzanas generalmente de 1 a 50, perfectamente delimitadas por avenidas, calles, andadores o cualquier otro rasgo fácil de identificación y cuyo uso del suelo sea principalmente habitacional, industrial de servicios, etc. Y la segunda se ubican en la parte rural, cuya extensión es de 8 500 hectáreas y se caracterizan por el uso de suelo agropecuario. INEGI

Resultados de la encuesta

En total se realizaron entrevistas a 38 746 de las 40 000 viviendas seleccionadas en la muestra. Esto implica que la tasa de respuesta del estudio fue de 96.9% y la de no respuesta 3.1%, la distribución de la tasa de no respuesta por entidad federativa fue menor al 15% que era lo que se tenía previsto. La mayor de estas tasas, se obtuvo en el estado de Guerrero que fue de 8.6%. Lo anterior se tradujo en una mejora de las capacidades de inferencia de la encuesta al permitir la estimación de proporciones estatales más pequeñas y se mantiene el resto de los parámetros fijos para el cálculo de los factores.

La información se almacenó en 7 bases de datos; 1 base contiene la información de los hogares y 6 bases para la información del cuestionario individual.

2.2 Definición de depresión mayor

La palabra depresión viene del latín *depressus* que significa abatido, derribado. Básicamente, la característica esencial de un episodio depresivo es un período de al menos dos semanas en el que existe un estado de ánimo deprimido o una pérdida de interés o placer por casi todas las actividades.

La definición de depresión a usar en este trabajo está basada en una cédula con criterios diagnósticos definidos por el Manual Diagnóstico y Estadístico de las Enfermedades Mentales (DSM IV)[4]

Se considera que existe depresión mayor cuando el individuo entrevistado refiere tener todos los síntomas siguientes: haberse sentido triste o vacío la mayor parte del día (Criterio A1) o bien; haber perdido interés por casi todas las cosas, incluyendo las que normalmente solía disfrutar (Criterio A2); además de sentirse con falta de energía o cansado constantemente (Criterio A6); que estos síntomas se presentaran durante la mayor parte del día, casi todos los días y que persistieran por un período mínimo de dos

[4] Por sus siglas en inglés

semanas, presentar alteraciones del apetito (Criterio A3), dificultad cognitiva (Criterio A8).

2.3 Importancia del diagnóstico de depresión mayor

La depresión se integra en el conglomerado de trastornos mentales que cada día cobran mayor importancia en el mundo y Murray (1997) estima que en 2020 será la segunda causa de años de vida saludable perdidos a escala mundial y la primera en países desarrollados.

Los trastornos mentales tienen un fuerte impacto sobre la vida de los individuos, la familia y la sociedad en su conjunto. Se calcula que más de 20% de la población mundial padecerá algún trastorno afectivo que requiera tratamiento médico en algún momento de su vida (Remick 2002). La Organización Mundial de la Salud (OMS) en el *Informe Mundial sobre la Salud* de 2001, refiere que la prevalencia puntual de depresión en el mundo en los hombres es de 1.9% y de 3.2% en las mujeres; la prevalencia de depresión para un periodo de 12 meses es de 5.8% y 9.5%, respectivamente.

En México se han llevado a cabo algunos estudios epidemiológicos para estimar la prevalencia de trastornos mentales, incluidos los trastornos y episodios depresivos, identificando, además, el proceso de búsqueda de ayuda. Entre los trabajos previos cabe destacar un estudio llevado a cabo como parte de la Encuesta Nacional de Adicciones en 1988, en el cual se incluyó una sección para investigar la prevalencia de trastornos mentales en personas de entre 18 y 65 años de edad. Medina-Mora (1992) refiere como uno de los principales hallazgos que 34% de la población estudiada presentó uno o más síntomas de depresión durante el mes anterior al estudio. El 13% de la población presentó sintomatología severa con importantes variaciones de acuerdo con el sexo del entrevistado: 8.5% entre los hombres y 17% en las mujeres.

En otro estudio con base en la Encuesta Nacional de Epidemiología Psiquiátrica, llevada a cabo en 2002 entre población urbana de 18 a 65 años de edad, concluyó que los trastornos afectivos (dentro de los que se incluyen los trastornos depresivos), se ubican,

respecto al resto de los trastornos investigados, en tercer lugar en frecuencia para prevalencia alguna vez en la vida (9.1%), después de los trastornos de ansiedad (14.3%) y los trastornos por uso de sustancias (9.2%). Al limitar el análisis de la encuesta a los 12 meses previos a su aplicación, los trastornos más comunes fueron los de ansiedad, seguidos por los afectivos.

Al analizar los trastornos individualmente, el episodio depresivo pasa a un quinto lugar (luego de las fobias específicas, los trastornos de conducta, la dependencia al alcohol y la fobia social), con una prevalencia de 3.3% alguna vez en la vida. Sin embargo entre las mujeres, la depresión mayor ocupa el segundo lugar (Medina y Col. 2003).

2.4 Objetivo del estudio

El objetivo del estudio en este trabajo es presentar las estimaciones de prevalencia de depresión mayor en la población adulta en México, así como el porcentaje de personas deprimidas que han sido diagnosticadas médicamente y a través de un modelo múltiple de regresión logística encontrar factores determinantes que pudieran incidir en que un individuo presente episodios de depresión mayor, a partir de los datos generados por la encuesta nacional de evaluación del desempeño cuyo trabajo de campo se llevó a cabo entre noviembre de 2002 y marzo de 2003 y que se detalló anteriormente.

Capítulo III

Aplicación del Modelo de Regresión Logística en el Diagnóstico de Depresión en Población Adulta en México

En este capítulo se hace la aplicación del modelo de regresión logística para determinar el efecto conjunto de factores que pudieran incidir en el diagnóstico de episodios de depresión mayor definido en el Capítulo II.

En primer lugar se lleva a cabo la descripción de los datos que se usarán en el análisis, posteriormente se muestra cuales fueron los criterios que se usaron para diagnosticar la prevalencia de eventos depresivos en los adultos en México. Posteriormente se describen las características tanto de la variable dependiente como de las variables independientes; para la ejecución del procedimiento se usó el paquete estadístico STATA 8, con base en lo que se vio en el Capítulo I, con la finalidad de identificar las variables explicativas que resultan ser significativas en el modelo establecido y finalmente se hace la interpretación de los coeficientes obtenidos con sus intervalos de confianza.

3.1 Información utilizada

La información se obtuvo de la Encuesta Nacional de Evaluación del Desempeño que se detalló en el Capítulo II. Para el análisis de datos de esta tesis, únicamente se usó la información del cuestionario individual del cual se extrajo información de características sociodemográficas del individuo, tales como la residencia, el tipo de localidad en el que reside, el sexo, la edad, el estado civil, el grado de escolaridad, desempleo, el estado de salud que refiere el entrevistado.

La información necesaria para el calculo de depresión mayor se definió en el Capítulo II a través de la parte del cuestionario que se refiere a cobertura: estados crónicos:

diagnóstico y tratamiento, en donde se aplicaron las siguientes preguntas, acotadas a un período de los últimos 12 meses previos a la fecha de aplicación del cuestionario, el levantamiento de la información se llevó a cabo entre noviembre de 2002 y marzo de 2003. En el cuestionario se preguntó específicamente:

- Algún periodo de varios días en el que la mayor parte del día se sintiera triste, vacío o deprimido (DEPRIMIDO).
- Un período de varios días en el que perdiera el interés por casi todas las cosas de las que suele disfrutar, como sus aficiones, sus relaciones personales o su trabajo (DESINTERES).
- Un período de varios días en el que se sintiera con falta de energía o cansado constantemente. (FATIGA).

Entre paréntesis se describe el nombre de la variable, además si el entrevistado contestaba afirmativamente alguna de estas preguntas, el encuestador debía preguntar además:

- Si ese período de pérdida de interés o falta de energía duró más de dos semanas (SEM2DEPRI).
- Si ese período de pérdida de interés o falta de energía se produjo durante la mayor parte del día, y casi todos los días (FATIDIA).
- Si había perdido el apetito durante ese período (APETITO).
- Si el entrevistado notó que su pensamiento se hacía más lento durante ese período (PENSAMIENTO).

Además se preguntó si el individuo había sido previamente diagnosticado por algún médico por padecer depresión, en caso de contestar afirmativamente se le preguntó si éste había recibido algún tratamiento o medicamento. En todos los casos se tomó en cuenta el procedimiento de muestreo para ponderar las observaciones y para estimar la varianza compuesta de las estimaciones.

3.2 Variable de respuesta o variable dependiente

La construcción de la variable de respuesta se basó en los criterios definidos en el Capítulo II, sección 2.2 que define los episodios depresivos, basados en una cédula con criterios definidos por el Manual Diagnóstico y Estadístico de las Enfermedades Mentales, considerando que existía depresión mayor cuando el individuo entrevistado refería tener los síntomas de la forma siguiente:

$$VAR_1 = DEPRIMIDO \cup DESINTERES$$

$$VAR_2 = FATIGA \cap SEM2DEPRI \cap FATDIA \cap APETITO \cap PENSAMIENTO$$

$$DEPRE\_CAL = VAR_1 \cap VAR_2$$

La variable DEPRE_CAL se define como una variable dicotómica de la siguiente manera:

$$\begin{cases} 1 = \text{Prevalencia de depresión} \\ 0 = \text{No existe depresión} \end{cases}$$

Con base en los datos que proporciona la encuesta, la prevalencia de depresión mayor en adultos en México es de 4.5% y es de 25% la prevalencia de adultos que padecen depresión y ya han sido diagnosticados.

Encontrando que por cada individuo diagnosticado existen 3 individuos que padecen depresión, pero que no han sido diagnosticados por algún especialista de la salud.

3.3 Variables explicativas o variables independientes

En el modelo logístico no existe ninguna restricción sobre la escala de medición de las variables independientes, éstas pueden ser dicotómicas, numéricas discretas o continuas. Para el caso en que se tenga una variable independiente categórica u ordinal, la propuesta sería generar las variables indicadoras para cada una de las categorías de la variable.

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 tener cualquier valor que sea posible dentro de un rango de valores de números reales. Las variables categóricas pueden ser de diversos tipos: dicotómicas por ejemplo: el sexo (hombre o mujer), politómicas no ordenadas, politómicas ordenadas.

Las variables categóricas que se tienen están constituidas entre 2 y 6 categorías y es necesario establecer la categoría de referencia o de comparación, técnicamente cualquier valor puede serlo, pero debe tomarse en cuenta aquella que tiene sentido desde el punto de vista del problema que se trabaje; ya que se puede cambiar la interpretación de cada coeficiente (β_i), en general, se cuantifica el efecto de cada categoría con respecto al valor de referencia que se haya establecido.

En la tabla 3.1 se establece la etiqueta con la que se nombra a cada variable, su descripción, se especifican las categorías, el tipo de la variable, como está codificada en la base de datos y por último el porcentaje muestral.

La definición de las categorías de los estratos de localidad se muestran en el tabla 2.1 del Capítulo II. La variable de escolaridad se refiere a escolaridad básica cuando el individuo indica haber cursado al menos el kinder o la primaria o secundaria completa, para el caso de educación media si el entrevistado contestó haber cursado la preparatoria o bachillerato o bien una carrera técnica o comercial y la escolaridad superior si el

entrevistado hizo referencia de contar con estudios profesionales o bien de maestría y/o doctorado.

Tabla 3.1. Descripción de las variables independientes

Variable independientes					
Etiqueta	Descripción	Categorias	Tipo	Codificación	% de muestra
Stloc	Tipo de localidad	Rural	Categórica	0	23.03
		Complemento Urbano		1	24.28
		Ciudad o área metropolitana		2	52.70
Sex	Sexo del encuestado	Hombre	Dicotómica	0	42.01
		Mujer		1	57.99
Age	Edad del ecncuestado	Mayores de 18 años	Numerica	[18, 100]	100.00
Civil2	Estado civil del encuestado	Soltero	Categórica	0	20.08
		Casado		1	57.10
		Separado		2	3.86
		Divorciado		3	1.30
		Viudo		4	6.13
		Unión libre		5	11.53
Escol	Escolaridad	Ninguna	Categórica	0	12.36
		Básica		1	63.93
		Media		2	15.56
		Superior		3	8.15
Desem	Desempleo	Trabaja	Dicotómica	0	49.12
		No trabaja		1	50.88
Edo_salud1	El entrevistado opina sobre su estado de salud	Bueno	Categórica	0	66.19
		Regular		1	28.41
		Malo		2	5.39

Fuente: Elaboración propia

3.4 Análisis del modelo de regresión logística para medir la asociación de las variables independientes.

El análisis se realiza a través del modelo de regresión logística para cada una de las variables independientes referidas en la tabla 3.1, con el fin de encontrar cuales de ellas resultan ser estadísticamente significativas para lo cual se usa la prueba de Wald y la del cociente de verosimilitud con un nivel de significancia del 95%.

Las variables estadísticamente significativas a un nivel de confianza del 95% fueron: sexo, edad, estado civil, escolaridad, desempleo, el estado de salud que refiere el

entrevistado. Las corridas del programa que arrojan estos resultados se encuentran en el Anexo B.1.

Una observación importante es que la variable de estrato de localidad (STLOC) no resultó significativa $|Z| = |1.71|$ que es menor al valor en tablas de 1.96 y por lo tanto la variable no es significativa.

La observación que puede hacerse es en relación a la definición de "rural", esta definición se basa exclusivamente en el tamaño poblacional que se define como localidades menores a 2 500 habitantes, que ciertamente no coincide con otras definiciones que pudieran existir de "rural" en donde asocian ciertas variables demográficas o económicas, de acuerdo con los resultados que se obtuvieron de la encuesta la prevalencia de depresión en áreas rurales es casi de la misma magnitud que la prevalencia de depresión en zonas urbanas, en el caso de las mujeres e incluso más alta en el caso de los hombres como puede apreciarse en la tabla 3.2.

Tabla 3.2. Descripción de las variables independientes

Tipo de localidad	Mujeres	Hombres	Ambos
Rural	5.9%	3.5%	4.9%
Cuidades medianas	6.0%	2.9%	4.7%
Cápitales o áreas metropolitanas	5.7%	1.9%	4.2%

Fuente: Elaboración propia

3.5 Análisis del modelo de regresión logística múltiple

Una vez identificadas las variables independientes significativas, se construye el modelo de regresión logística múltiple a partir de las siguientes variables explicativas señaladas en el tabla 3.1: sexo, edad, estado civil, escolaridad, desempleo y la opinión que el entrevistado refiere de su estado de salud, con la finalidad de estimar la fuerza de asociación entre la exposición y el evento de interés y determinar si esta asociación es estadísticamente diferente del valor nulo.

La salida del programa se encuentra en el Anexo B.2.1 La tabla siguiente muestra las variables explicativas, con su respectiva etiqueta y los valores de los coeficientes, el error estándar, la estadística de Wald y el valor P asociado a la prueba en cuestión.

Tabla 3.3 Coeficientes, estadística de Wald

Descripción	Variable	Coeficiente (b)	Error estándar	Estadística de Wald (z)	P>z
Sexo	sex	0.6872	0.1118	6.15	0.000
Edad	age	0.0067	0.0027	2.50	0.012
Desempleo	desem	0.0433	0.1051	0.41	0.681
Básica	_Iescol_1	-0.2321	0.1031	-2.25	0.024
Media	_Iescol_2	-0.8335	0.1719	-4.85	0.000
Superior	_Iescol_3	-0.7075	0.2063	-3.43	0.001
Casado	_Icivil2_1	0.0075	0.1412	0.05	0.958
Separado	_Icivil2_2	0.4913	0.2207	2.23	0.026
Divorciado	_Icivil2_3	0.3030	0.2477	1.22	0.221
Viudo	_Icivil2_4	0.1905	0.1815	1.05	0.294
Unión libre	_Icivil2_5	0.1307	0.1770	0.74	0.460
Regular	_Iedo_salu~1	1.2811	0.0963	13.30	0.000
Malo	_Iedo_salu~2	2.0157	0.1246	16.18	0.000

Fuente: Elaboración propia

Al considerar todas las variables, la aportación de la variable desempleo (véase tabla 3.3), no resulta ser estadísticamente significativa, en donde el valor de $|Z| = |0.41|$ que es menor al valor de Z en tablas de 1.96.

Por otro lado en la tabla 3.3 se observa que la variable estado civil resulta significativa para la categoría casado, al excluir la variable del modelo, se comprueba que su contribución al modelo es irrelevante como se aprecia en la tabla 3.4.

Tabla 3.4. Comparación de modelos

Log pseudo-likelihood	
Modelo nulo	-7050.8851
7 variables	-6308.8592
12 variables	-6297.2028

Fuente: Elaboración propia

Entonces las variables que resultaron asociadas significativamente fueron sexo, edad, escolaridad y la opinión que refiere el entrevistado de su estado de salud, como se muestra en la tabla 3.5.

Tabla 3.5 Coeficientes, estadística de Wald e intervalos de confianza

Descripción	Variable	Coeficiente (b)	Error estándar	Estadística de Wald (z)	P>z	Intervalo de confianza 95%	
Sexo	sex	0.7516	0.0915	8.21	0.000	0.5722	0.9309
Edad	age	0.0081	0.0024	3.38	0.001	0.0034	0.0128
Básica	_lescol_1	-0.2371	0.1034	-2.29	0.022	-0.4398	-0.0345
Media	_lescol_2	-0.8533	0.1686	-5.06	0.000	-1.1837	-0.5228
Superior	_lescol_3	-0.7367	0.2031	-3.63	0.000	-1.1349	-0.3385
Regular	_ledo_salu~1	1.2811	0.0964	13.29	0.000	1.0922	1.4701
Malo	_ledo_salu~2	2.0172	0.1249	16.15	0.000	1.7724	2.2619

Fuente: Elaboración propia

Para efectos de verificar que tan bueno es el ajuste del modelo a los datos se realizó una prueba de bondad de ajuste basada en el estadístico χ^2 de Pearson;

Lo que se desea es confrontar la hipótesis nula y la hipótesis alternativa

$$H_0 = \beta_j \neq 0 \qquad \text{vs} \qquad H_a = \beta_j = 0$$

El valor obtenido para el estadístico χ^2 es de 0.2532 apoya la hipótesis nula y sugiere que el modelo propuesto ajusta razonablemente bien a los datos. Esto aunado a que el número de patrones de covariables $(J = 1505)$ es considerablemente menor al tamaño de la muestra $(n = 38656)$ por lo que se puede concluir que el modelo tiene un buen ajuste.

3.6 Interpretación de los coeficientes

Los momios asociados a cierto suceso se definen como la razón entre la probabilidad de que el suceso ocurra y la probabilidad de que el suceso no ocurra, es decir, un número que expresa cuánto más probable es que se produzca frente a que no se produzca el suceso en cuestión y se sabe que una vez conocidos los momios se puede obtener la probabilidad de que el suceso ocurra.

De modo que ambas informaciones son de alguna manera equivalentes y expresan la misma noción: cuantifican cuan probable es que algo ocurra, en particular, el riesgo de

un acontecimiento. En este punto es menester recordar el concepto de riesgo relativo que sintetiza cuánto más probable es desarrollar algún síntoma o padecimiento si se estuvo expuesto a cierto factor y no haber estado expuesto a él.

Por este motivo, en el contexto de la regresión logística es de gran importancia la interpretación de los momios 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.

Si se parte de la variable dependiente:

$$Y_i = \begin{cases} 1 & \text{deprimido} \\ 0 & \text{no deprimido} \end{cases}$$

Y la variable explicativa:

$$X_i = \begin{cases} 1 & \text{Mujer} \\ 0 & \text{Hombre} \end{cases}$$

El cociente de momios queda de la siguiente manera:

$$\text{RM} = \frac{\text{Momio}\left(\text{depre} \middle| \text{sex} = \text{m}, \text{age} = \text{x}, \text{esl1} = \text{x}, \text{esl2} = \text{x}, \text{sal1} = \text{x}, \text{sal2} = \text{x}\right)}{\text{Momio}\left(\text{depre} \middle| \text{sex} = \text{h}, \text{age} = \text{x}, \text{esl1} = \text{x}, \text{esl2} = \text{x}, \text{sal1} = \text{x}, \text{sal2} = \text{x}\right)} = e^{.751559} = 2.12$$

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 que una mujer presente eventos de depresión es de 2.1 veces la probabilidad que de un hombre.

En cuanto a la variable explicativa edad, la probabilidad de que una persona presente eventos de depresión se incrementa conforme incrementa la edad.

$$\psi(c) = \exp(c\beta_{age})$$

Para la variable de escolaridad la categoría de referencia es cuando el entrevistado contestó no contar con ningún grado de escolaridad, teniéndose una exposición protectora para las demás categorías. Véase la tabla 3.6.

Tabla 3.6 Cociente de momios e intervalos de confianza

Depre_cal	Cociente de momios	[Intervalo de confianza 95%]	
sex	2.12	1.77	2.54
age	1.01	1.00	1.01
_lescol_1	0.79	0.64	0.97
_lescol_2	0.43	0.31	0.59
_lescol_3	0.48	0.32	0.71
_ledo_sal1	3.60	2.98	4.35
_ledo_sal2	7.52	5.89	9.60

Fuente: Elaboración propia

Para esta variable la probabilidad de riesgo de que una persona sin estudios presente episodios depresivos es 1.2 veces en comparación con una persona que cuenta con estudios básicos y de 2.3 y 2.1 en comparación con una persona que cuenta con estudios medio y superior respectivamente.

Y por último, la variable sobre la opinión que una persona refiere de su estado de salud, muestra que es 3.6 y 7.5 veces más probable presentar eventos depresivos en personas que refieren estados de salud regulares o malos que aquellas que refieren contar con un estado bueno de salud.

Al repetir los modelos de manera diferenciada para cada uno de los sexos se encontró, como puede observarse en el Anexo B.3.2, únicamente en el caso de los hombres que la condición de desempleo resultó significativamente asociada a la presencia de depresión siendo 1.6 veces más probable que un individuo sin empleo sufra de episodios depresivos que un individuo que tiene empleo, además de que la variable edad en el caso de los hombres no resultó significativa siendo significativa únicamente en las mujeres.

Conclusiones

El carácter aleatorio del procedimiento muestral permite que diversas estimaciones sean representativas del fenómeno en diferentes niveles de desagregación. Sin embargo, la medición mediante encuestas tiene sus problemas particulares sobre todo cuando una encuesta pretende abarcar varios temas, dejando de profundizar en ciertos tópicos.

La medición llevada a cabo con la ENED 2002, se basó en siete reactivos, puede tener ciertos problemas relacionados con la especificidad del instrumento para el cálculo de la depresión, debe tomarse en cuenta que la selección de los reactivos incluidos en la encuesta se basó en los criterios diagnósticos del DSM IV, el cual se detalló en el capítulo II, dicha definición en este trabajo cumple cabalmente con esta disposición. No obstante, la encuesta no permitió utilizar algunos criterios de exclusión como son los asociados a que la sintomatología depresiva esté asociada a un duelo reciente o a un deterioro fisiológico. Esta última consideración indudablemente puede estar afectando la especificidad del instrumento. Adicionalmente, el instrumento utilizado tampoco tiene capacidad para distinguir el nivel de severidad del trastorno depresivo, pero de momento es el que está disponible.

Los resultados presentados muestran que el diagnóstico de trastornos depresivos es más frecuente entre las mujeres siendo la prevalencia de depresión en las mujeres más o menos el doble de la observada en los hombres. La mayor parte de los antecedentes, nacionales e internacionales, coinciden en esta diferencia y confirman el éxito de los procesos estadísticos aplicados en este trabajo.

La relación de la edad con la depresión, que muestra una clara asociación positiva en los resultados, ha mostrado un comportamiento consistente en otros estudios; los resultados de está tesis coinciden con la Encuesta Nacional de Epidemiología Psiquiátrica en México al identificar mayor prevalencia conforme se incrementa la edad (Caraveo y Col 1997).

Con relación al tipo de comunidad, el primer comentario que debe hacerse es con relación a la definición de "rural". En México esta definición se basa exclusivamente en el tamaño poblacional de la comunidad por lo que no necesariamente tiene el mismo significado que en otros países y, por lo tanto, son difíciles las comparaciones en ese sentido. De acuerdo con los resultados, la prevalencia en las áreas rurales es de la misma magnitud que en las zonas urbanas en el caso de las mujeres e incluso más alta en el caso de los hombres. Evidentemente hay una importante asociación entre las condiciones de marginación que son frecuentes en las zonas rurales y la alta prevalencia de depresión. Sin embargo la variable de estrato de localidad no resultó significativa.

Algunas de las variables asociadas encontradas en el análisis como asociadas a la presencia de depresión son indicadores de vulnerabilidad o marginación social. La alta prevalencia asociada a los niveles más bajos de escolaridad, a las edades más avanzadas y al sexo femenino puede interpretarse como la manifestación de una mayor vulnerabilidad de estos grupos a los diversos factores que pueden condicionar el desarrollo de un episodio depresivo.

Una variable adicional que ha sido referida como detonador de este tipo de episodios, y que en el estudio mostró asociación sólo en el caso de los hombres, es el desempleo.

De cualquier modo, es notable que algunas variables que determinan una mayor probabilidad de desarrollar depresión también se asocien con una probabilidad menor de ser diagnosticados en ámbitos clínicos. Esta combinación cierra así un círculo perverso donde los más vulnerables son asimismo los que más dificultad tienen para acceder al tratamiento de su padecimiento. Este hecho multiplica las consecuencias negativas de la enfermedad, disminuyendo la calidad de vida de los individuos afectados en particular y de la población en lo general, ya que los individuos deprimidos disminuyen su productividad, afectan la dinámica familiar y, si no son atendidos, tienen mayor probabilidad de sufrir un desenlace fatal.

Anexo A
Cuestionarios

SECRETARÍA DE SALUD

DIRECCIÓN GENERAL DE EVALUACIÓN DEL DESEMPEÑO

ENCUESTA NACIONAL DE EVALUACIÓN DEL DESEMPEÑO 2002

CUESTIONARIO DE HOGAR

IDENTIFICACIÓN GEOGRÁFICA

ID. Núm. de cuestionario: |___|___|___|___|___|___| **Ent. Fed.**_____ |___|___|

Mun. o Del._____|___|___|___| **Estrato** |___|

1-Urbano
2-Rural

Datos generales

¿Cuántas personas viven permanentemente en esta casa? |___|___|

A. REVISIÓN GENERAL Y DEMOGRÁFICA

	A	B	C	D	E	F	J
				Mayores de 6 años	**Mayores de 15 años**		
NÚMERO DE REGISTRO	1¿Cuáles son los nombres de las personas que viven en esta casa? *(por favor dígame los nombres de los integrantes de la familia empezando con los hombres de mayor a menor y después las mujeres en el mismo orden)* NOMBRE (variable omitida en la base de datos)	2. Relación con el jefe de familia el/ella.....01 esposo(a).....02 hijo(a).....03 nuera/yerno.....04 nieto(a).....05 padre/madre.....06 suegro(a).....07 hermano(a).....08 compañero(a).....09 abuelo(a).....10 otro familiar.....11 sin parentesco.....12 desconocido.....13	3.¿Cuántos años cumplidos tiene _____? ***(nombre)***	4.¿Hasta qué año estudió? ***(Anotar el código y el total de años estudiados*)** No fue a la escuela..... 0 Primaria..... 1 Tec. C/Prim.……..2 Secundaria..... 3 Tec C/Sec.…..…4 Prepa o Vocacional.....5 Tec. C/Prepa.…..…6 Licenciatura..... 10 Postgrado..... 11 N/S…..…..…….....88 N/C…...……99 Nivel | Años	5.¿Cuál es su estado civil? Soltero.....1 Casado.....2 Separado.....3 Divorciado.....4 Viudo.....5 Unión Libre.....6 N/S.....….....8 N/C.....….....9	6. ¿En que trabaja actualmente? Empleado público.....1 Empleado del sector privado.2 Independiente.3 Empresario.....4 Otros.....5	¿En promedio cual es su ingreso en el último mes?
Hombres							
0400		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|
0401		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|
0402		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|
0403		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|
0404		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|
0405		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|
0406		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|
Mujeres							
0407		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|
0408		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|
0409		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|
0410		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|
0411		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|
0412		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|
0413		|__|__|	|__|__|	|__|__|	|__|__|	|__|	|__|	|__|__|__|__|

Preguntas para todos los hogares

Q0500. ¿Quién es el jefe del hogar?	Anote el número de registro que le corresponde_____.	|__|__|__|__|

0600 SEGURO MÉDICO

Quisiera hacerle algunas preguntas acerca del seguro médico. Cuando decimos que alguien está cubierto por un seguro médico nos referimos a que esta persona esta protegido(a) por una organización que cubre sus gastos médicos.

Entrevistador haga las preguntas para cada miembro del hogar en el mismo orden en que fueron anotados (respetando su número de registro)

NÚMERO DE REGISTRO	Q0600. ¿Está cubierta esta persona por algún plan de seguro médico? Sí.....1 No.....5	Q0601. ¿Está cubierta esta persona por algún plan de seguro médico obligatorio? Nota: por prestación laboral (ya sea derechohabiente o beneficiario) Sí, a cual? Ninguno.....0 IMSS.....1 ISSSTE.....2 PEMEX.....3 SEDENA.....4 MARINA.....5 OTROS.....6	Q0602. ¿Está cubierta esta persona por algún plan de seguro médico voluntario? Ninguno.....0 ↓ p.p.Q0604 Privado.....1 Seguro Popular.....2 IMSS voluntario.....3 Otros.....4	Q0603. ¿Cuánto paga al año por el seguro médico voluntario (NOMBRE)?	Q0604. El único motivo de que esta persona esté asegurada ¿es por su relación con alguien que tiene seguro médico? Sí.....1 No.....5 ↓ Pase a la siguiente persona	Q0605. ¿Quien está suscrito al plan de seguro médico que cubre a esta persona? INDIQUE el número de registro del hogar de la persona en cuestión. Sí la persona no está en el hogar, escriba 9999
Hombres						
0400	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I
0401	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I
0402	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I
0403	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I
0404	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I
0405	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I
0406	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I
Mujeres						
0407	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I
0408	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I
0409	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I
0410	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I
0411	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I
0412	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I
0413	I__I	I__I	I__I	I__I__I__I__I	I__I	I__I__I__I__I

0700. INDICADORES DE INGRESOS FIJOS (Bajos)

Me gustaría hacerle algunas preguntas sobre su vivienda.

Q0700. Por favor puede usted decirme ¿cuántos cuartos hay en esta casa?	Número de cuartos_____	|___|___|
Q0701. ¿Cuántas sillas hay en la casa?	Número de sillas_____ Ninguna.....00	|___|___|
Q0702. ¿Cuántas mesas hay en la casa?	Número de mesas_____ Ninguna.....00	|___|___|
Q0703. ¿Cuántos coches hay en el hogar?	Número de coches_____ Ninguno.....00	|___|___|
Q0704. ¿Cuenta su vivienda con electricidad?	Sí..... 1 No..... 5	|___|
¿Tiene alguien en su hogar.....		
Q0705. bicicleta?	Sí..... 1 No..... 5	|___|
Q0706. reloj?	Sí..... 1 No..... 5	|___|
Q0707. una cubeta?	Sí..... 1 No..... 5	|___|
Q0708. una lavadora?	Sí..... 1 No..... 5	|___|
Q0709. un lavavajillas?	Sí..... 1 No..... 5	|___|
Q0710. un refrigerador?	Sí..... 1 No..... 5	|___|
Q0711. un teléfono fijo?	Sí..... 1 No..... 5	|___|
Q0712. un teléfono celular?	Sí..... 1 No..... 5	|___|
Q0713. un televisor?	Sí..... 1 No..... 5	|___|
Q0714. una computadora?	Sí..... 1 No..... 5	|___|

0800. GASTOS DEL HOGAR

Ahora quisiera hacerle algunas preguntas acerca del gasto de su hogar.

¿En las últimas 4 semanas.....		
Q0800. cuánto ha gastado su hogar en total? (incluya todo lo que gastó en alimentos, servicios, etc.)	Cantidad_____	|___|___|___|___|
Q0801. ¿Cuánto ha gastado su hogar en: Alimentos naturales no procesados como (semillas, frutas y verduras frescas, carne fresca etc.? No incluya alimentos procesados como salchichonería, refrescos, botanas etc.	Cantidad_____	|___|___|___|___|
Q0801a. ¿Cuánto ha gastado en combustibles (gasolina, gas, carbón, petróleo, etc.)?	Cantidad_____	|___|___|___|___|
Q080ab. ¿Cuánto ha gastado en.....	Cigarrillos con filtro $:_____ Cigarrillos sin filtro $:_____ Puros $:_____	|___|___|___| |___|___|___| |___|___|___|
Q0801c. ¿Cuánto ha gastado en bebidas que contengan alcohol como...	Cerveza $:_____ Pulque $:_____ Whisky $:_____ Ron $:_____ Brandy $:_____ Tequila $:_____ Vino de mesa $:_____ Mezcal $:_____ Aguardiente $:_____ Otros $:_____	|___|___|___| |___|___|___| |___|___|___| |___|___|___| |___|___|___| |___|___|___| |___|___|___| |___|___|___| |___|___|___| |___|___|___|
Q0802. ¿Cuánto ha gastado en renta, electricidad, agua, teléfono?	Cantidad_____	|___|___|___|___|
Q0803. ¿En educación, clases particulares o cursos y material escolar?	Cantidad_____	|___|___|___|___|
Q0804. ¿En gastos de salud sin contar los reembolsos de los seguros (devolución de los seguros)?	Cantidad_____	|___|___|___|___|
Q0805. Primas de seguros voluntarios y cotizaciones a programas de salud	Cantidad_____	|___|___|___|___|
Q0806. Otros gastos por ejemplo vacaciones, tintorería, lavandería, pago a empleadas domésticas, etc.	Cantidad_____	|___|___|___|___|

Quisiera hacerle algunas preguntas más concretas acerca del gasto de salud en su hogar. Al contestar las preguntas, piense en todas las veces que algún miembro del hogar ha requerido atención médica en los últimos 3 meses. Por favor, no cuente los gastos de reembolso por algún seguro médico y los gastos de transporte.

¿Cuánto ha gastado su hogar durante los últimos 3 meses en:

Q0807. Hospitalización de algún miembro del hogar?	Cantidad_____	|___|___|___|___|
Q0808. Cuidados de médicos, enfermeras o comadronas que no requirieron hospitalización?	Cantidad_____	|___|___|___|___|

Q0809. Cuidados de curanderos o médicos alternativos	Cantidad_____	|___|___|___|___|
Q0810. Dentista	Cantidad_____	|___|___|___|___|
Q0811. Tratamientos o medicamentos	Cantidad_____	|___|___|___|___|
Q0812. Cuánto ha gastado en lentes, dispositivos para el oído, prótesis, etc.	Cantidad_____	|___|___|___|___|
Q0813. Diagnóstico y pruebas de laboratorio como rayos X y análisis de sangre	Cantidad_____	|___|___|___|___|
Q0814. Cualquier otro producto o servicio médico que no se haya mencionado	Cantidad_____	|___|___|___|___|
Ahora vamos a referirnos al último año; esto es de (mes) a (mes)		
Q0815. En los últimos 12 meses, ¿cuántas veces ha sido hospitalizado algún miembro de su hogar? **ANOTE EL NÚMERO DE VECES PARA TODOS LOS MIEMBROS DEL HOGAR EN SU TOTALIDAD SI NO HAY NINGUNO ESCRIBA ¨0¨**	Número de veces_____ **Si no hay ninguno p.p.Q0817**	|___|___|
Q0816. Descontando lo que ya me dijo de los gastos de hospitalización y gastos de seguro médico de las últimas 4 semanas ahora dígame ¿cuánto han pagado en este hogar por gastos de hospitalización en los últimos 12 meses?	Cantidad_____	|___|___|___|___|
<u>En los últimos 12 meses</u>, ¿Con que ingresos pagó su hogar los gastos de salud?		
Q0817. Sueldos o salarios actuales de los miembros del hogar	Sí..... 1 No..... 5	|___|
Q0818. Ahorros familiares (cuenta bancaria, tandas, etc.)	Sí..... 1 No..... 5	|___|
Q0819. Reembolso o devolución de un plan de seguro médico	Sí..... 1 No..... 5	|___|
Q0820. Venta de propiedades como muebles, animales, joyas, coches, etc.	Sí..... 1 No..... 5	|___|
Q0821. Apoyo o regalo en dinero por parte de otros familiares o amigos que no conviven en su hogar	Sí..... 1 No..... 5	|___|
Q0822. Dinero prestado por algún familiar o amigo	Sí..... 1 No..... 5	|___|
Q0822a. Crédito bancario	Sí..... 1 No..... 5	|___|
Q0823. Otro.	Sí..... 1 No..... 5	|___|

SECRETARÍA DE SALUD

DIRECCIÓN GENERAL DE EVALUACIÓN DEL DESEMPEÑO

ENCUESTA NACIONAL DE EVALUACIÓN DEL DESEMPEÑO 2002

CUESTIONARIO INDIVIDUAL A / B

IDENTIFICACIÓN GEOGRÁFICA

ID. Núm. de cuestionario: |___|___|___|___|___|___| **Ent. Fed.**_____ |___|___|

Mun. o Del._____|___|___|___| **Estrato** |___|

1-Urbano
2-Rural

Renglón de la persona entrevistada _____

Tipo de Cuestionario |___|

A
B

Sección 1000. Características sociodemográficas del entrevistado

Antes de preguntarle por su estado de salud quisiera hacerle algunas preguntas de carácter general.

Q1000. ¿Cuál es su idioma materno?	_____	
Encuestador Q1001. Circule el sexo según corresponda	Mujer..... 1 Hombre..... 2	|___|
Q1002. Me podría decir. ¿Cuántos años cumplidos tiene usted?	Años_____ → p.p.1008 N/S..... 888 N/R..... 777	|___|___|___|
Q1008. ¿Cuál es su estado civil actual...	Soltero (a) ?..... 1 Casado (a)? ..... 2 Separado (a)? ..... 3 Divorciado (a)? ..... 4 Viudo (a)? ..... 5 Unión libre? 6	|___|
Q1009. ¿Hasta que año aprobó en la escuela?	Ninguno..... 0 Preescolar o kínder ..... 1 Primaria completa..... 2 Secundaria completa..... 3 Preparatoria o bachillerato completo.... 4 Carrera técnica o comercial..... 5 Profesional..... 6 Maestría o Doctorado..... 7	|___| Nivel
Q1010. ¿En total cuantos años estudió?	Años:_____	|___|___|

Q1011. ¿De las siguientes opciones usted se identifica con alguna...	Sí	¿Cuál?	No	N/R	
a) Religión?	(1)	_____	(5)	(7)	|___| --|___|___|
b) Región del país?	(1)	_____	(5)	(7)	|___| --|___|___|
c) Raza?	(1)	_____	(5)	(7)	|___| --|___|___|
d) Con algún otro grupo?	(1)	_____	(5)	(7)	|___| --|___|___|

Ahora quisiera hacerle algunas preguntas acerca de su ocupación laboral.

Q1012. ¿ En qué trabaja actualmente?	Empleado público (federal, estatal)..... 1 Empleado particular (de empresa)..... 2 Independiente (por su cuenta)..... 3 Empresario..... 4 Trabajador voluntario (sin pago)..... 5 No trabaja…… 6 →	|___| p.p.Q1014

Q1013. ¿Cual de las siguientes opciones describe mejor su principal ocupación en los últimos 12 meses?	Legislador, alto funcionario, director de empresa..... 01 Profesional (Ingeniero ,médico, profesor, sacerdote, etc.).. 02 Profesional, técnico o asociado profesional (inspector, agente financiero, etc.)..... 03 Empleado (secretario, cajero, etc.)..... 04 Empleado de servicios o ventas (cocinero, agente de viajes, dependientes, etc.)..... 05 Trabajador en la agricultura o pesca..... 06 Artesano o comerciante (carpintero, pintor, joyero, carnicero, etc.)..... 07 Operador o montador en una planta o maquina (montador de equipos, operador de máquinas de coser, conductor, etc.)..... 08 Trabajador manual (vendedor ambulante, boleros, etc.)..... 09 Fuerzas armadas (militar gubernamental)..... 10 Trabajo doméstico..... 11	|___|___| → p.p.Q2000
Q1014. ¿Cuál es el motivo principal por el que usted no trabaja?	Cuidado de la casa / de la familia..... 1 Ha buscado pero no ha encontrado trabajo..... 2 Trabaja como voluntario..... 3 Estudia / formación profesional..... 4 Jubilado / pensionado / demasiado grande para trabajar..... 5 Por tener problemas de salud..... 6 Otra..... 7	|___|

Sección 2000. Descripción del estado de salud (Estado de salud general)

Las primeras preguntas se refieren a su estado de salud general, incluida la salud física y la mental.

Q2000. En general ¿cómo calificaría hoy su estado de salud?	Muy bueno..... 1 Bueno..... 2 Moderado..... 3 Malo..... 4 Muy malo..... 5	|____|
Q2001. En general durante los últimos 30 días, ¿qué grado de dificultad ha tenido para realizar las tareas del trabajo y del hogar?	Ninguno..... 1 Leve..... 2 Moderado..... 3 Alto..... 4 Extremo imposible de realizar..... 5	|____|

Sección Q6000. COBERTURA

Ahora quisiera Hacerle una serie de preguntas acerca de algunos problemas y cuestiones de salud que hayan podido afectarle a usted y a sus hijos menores de su hogar y el tipo de asistencia o tratamiento que hayan recibido.

ESTADOS CRÓNICOS: DIAGNÓSTICO Y TRATAMIENTO

Q6000. ¿Alguna vez le han diagnosticado artritis (una enfermedad de las articulaciones)?	Sí..... 1 No..... 5 No sabe..... 8	|____| → p.p.Q6003
Q6001. ¿Le han puesto alguna vez en tratamiento por ello?	Sí..... 1 No..... 5 No sabe..... 8	|____|
Q6002.Durante las últimas 2 semanas ¿ha tomado algún médicamento o ha segido otro tratamiento?	Sí..... 1 No..... 5 No sabe..... 8	|____|

Q6002a. ¿Cuánto le costó el tratamiento para la artritis?	Cantidad :_____	|____|____|____|____|
Q6009. ¿Alguna vez le ha dicho un médico u otro profesional de salud que tuviera angina de pecho (enfermedad del corazón)?	Sí..... 1 No..... 5 No sabe..... 8 → p.p.Q6012	|____|
Q6010. ¿Le han puesto alguna vez en tratamiento por ello?	Sí..... 1 No..... 5 No sabe..... 8	|____|
Q6011. Durante <u>las últimas 2 semanas</u>, ¿ha tomado algún medicamento <u>o ha seguido otro tratamiento</u>?	Sí..... 1 No..... 5 No sabe..... 8	|____|
Q6011a. ¿Cuánto le costó el tratamiento para la angina de pecho?	Cantidad:_____	|____|____|____|____|
Q6017. ¿Alguna vez le han diagnosticado asma (una enfermedad alérgica respiratoria)?	Sí..... 1 No..... 5 No sabe..... 8 → p.p.Q6020	|____|
Q6018. ¿Le han puesto alguna vez en tratamiento por ello?	Sí..... 1 No..... 5 No sabe..... 8	|____|
Q6019. Durante las <u>últimas 2 semanas</u>, ¿ha tomado algún medicamento o ha seguido otro tratamiento?	Sí..... 1 No..... 5 No sabe..... 8	|____|
Q6019a. ¿Cuánto le costó el tratamiento para el asma?	Cantidad:_____	|____|____|____|____|
Q6025. ¿Alguna vez le ha dicho un médico u otro profesional de salud que sufriera depresión?	Sí..... 1 No..... 5 No sabe..... 8 → p.p.Q6028	|____|
Q6026. ¿Ha estado alguna vez en tratamiento ?	Sí..... 1 No..... 5 No sabe..... 8	|____|
Q6027. Durante <u>las últimas 2 semanas</u>, ¿ha tomado algún medicamento o ha seguido otro tratamiento?	Sí..... 1 No..... 5 No sabe..... 8	|____|
Q6027a. ¿Cuánto le costó este tratamiento para la depresión?	Cantidad:_____	|____|____|____|____|

Durante los últimos 12 meses ¿ha tenido...

Q6028. Algún periodo de <u>varios días</u> en el que la mayor parte del día se sintiera <u>triste, vacío o deprimido</u>?	Sí..... 1 No..... 5 No sabe..... 8	|____|
Q6029. Un periodo de <u>varios días</u> en el que perdiera el interés <u>por casi todas las cosas</u> de las que suele disfrutar, como sus aficiones, sus relaciones personales o su trabajo?	Sí..... 1 No..... 5 No sabe..... 8	|____|

Q6030. Un periodo de varios días en el que se sintiera con falta de energía o cansado(a) constantemente?	Sí..... 1 No..... 5 No sabe..... 8	|____|
Encuestador si la respuesta es NO o NO SABE en las preguntas Q6028, Q6029 y Q6030 pase a Q6035		
Q6031. ¿Este periodo de pérdida de interés o falta de energía duró más de dos semanas?	Sí..... 1 No..... 5	|____|
Q6032. ¿Este periodo de pérdida de interés o falta de energía se produjo durante la mayor parte del día, y casi todos los días?	Sí..... 1 No..... 5	|____|
Q6033. ¿Perdió el apetito durante este periodo?	Sí..... 1 No..... 5	|____|
Q6034. ¿Notó Usted que su pensamiento se hacía más lento durante este periodo?	Sí..... 1 No..... 5	|____|
Q6035. ¿Alguna vez le han diagnosticado un problema de esquizofrenia o psicosis?	Sí..... 1 No..... 5 No sabe..... 8	|____| → p.p.Q6038
Q6036. ¿Ha estado alguna vez en tratamiento?	Sí..... 1 No..... 5 No sabe..... 8	|____|
Q6037. Durante las últimas 2 semanas, ¿ha tomado algún medicamento o ha seguido otro tratamiento?	Sí..... 1 No..... 5 No sabe..... 8	|____|
Q6037a. ¿Cuánto le costó el tratamiento para este padecimiento?	Cantidad:_____	|____|____|____|____|

Anexo B

Resultados de las corridas del paquete Stata 8

B.1. Corridas del modelo de regresión logística a través del paquete STATA para cada variable explicativa.

B.1.1 Modelo de regresión logística variable dependiente depresión y variable explicativa estrato de localidad.

```
. logit depre_cal stloc [pweight=p_int ]

(sum of wgt is   7.5575e+07)
Iteration 0:   log pseudo-likelihood = -7050.8851
Iteration 1:   log pseudo-likelihood = -7047.1517
Iteration 2:   log pseudo-likelihood = -7047.1469

Logit estimates                                   Number of obs   =      38656
                                                  Wald chi2(1)    =       2.92
                                                  Prob > chi2     =     0.0875
Log pseudo-likelihood = -7047.1469                Pseudo R2       =     0.0005
------------------------------------------------------------------------------
             |               Robust
  depre_cal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       stloc |  -.0814096   .0476412    -1.71   0.087    -.1747846    .0119654
       _cons |  -2.960081   .0671106   -44.11   0.000    -3.091616   -2.828547
------------------------------------------------------------------------------
```

B.1.2 Modelo de regresión logística variable dependiente depresión y variable explicativa sexo del entrevistado.

```
. logit depre_cal sex [pweight=p_int ]

(sum of wgt is   7.5575e+07)
Iteration 0:   log pseudo-likelihood = -7050.8851
Iteration 1:   log pseudo-likelihood = -6927.4323
Iteration 2:   log pseudo-likelihood = -6922.8468
Iteration 3:   log pseudo-likelihood = -6922.8343

Logit estimates                                   Number of obs   =      38656
                                                  Wald chi2(1)    =      92.62
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -6922.8343                Pseudo R2       =     0.0182


------------------------------------------------------------------------------
             |               Robust
  depre_cal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .8643051   .0898087     9.62   0.000     .6882833    1.040327
       _cons |  -3.642975   .0767878   -47.44   0.000    -3.793476   -3.492473
------------------------------------------------------------------------------
             |               Robust
  depre_cal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   2.373356    .213148     9.62   0.000     1.990296    2.830142
------------------------------------------------------------------------------
```

B.1.3 Modelo de regresión logística variable dependiente depresión y variable explicativa edad.

```
. logit depre_cal age[pweight=p_int ]

(sum of wgt is   7.5575e+07)
Iteration 0:   log pseudo-likelihood = -7050.8851
Iteration 1:   log pseudo-likelihood = -6918.7742
Iteration 2:   log pseudo-likelihood = -6909.5655
Iteration 3:   log pseudo-likelihood = -6909.5368

Logit estimates                                   Number of obs   =      38656
                                                  Wald chi2(1)    =     131.92
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -6909.5368                Pseudo R2       =     0.0200

------------------------------------------------------------------------------
             |               Robust
   depre_cal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0238689   .0020781    11.49   0.000     .0197958     .027942
       _cons |  -4.083175   .1085591   -37.61   0.000    -4.295947   -3.870403
------------------------------------------------------------------------------
             |               Robust
   depre_cal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.024156   .0021283    11.49   0.000     1.019993    1.028336
------------------------------------------------------------------------------
```

B.1.4 Modelo de regresión logística variable dependiente depresión y variable explicativa desempleo.

```
. logit depre_cal desem[pweight=p_int ]

(sum of wgt is   7.5575e+07)
Iteration 0:   log pseudo-likelihood = -7050.8851
Iteration 1:   log pseudo-likelihood = -6971.2575
Iteration 2:   log pseudo-likelihood = -6969.5969
Iteration 3:   log pseudo-likelihood = -6969.5961

Logit estimates                                   Number of obs   =      38656
                                                  Wald chi2(1)    =      56.83
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -6969.5961                Pseudo R2       =     0.0115

------------------------------------------------------------------------------
             |               Robust
   depre_cal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       desem |   .6445589    .085498     7.54   0.000      .476986    .8121319
       _cons |  -3.438124   .0706114   -48.69   0.000     -3.57652   -3.299728
------------------------------------------------------------------------------

             |               Robust
   depre_cal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       desem |   1.905147   .1628862     7.54   0.000     1.611211    2.252705
------------------------------------------------------------------------------
```

B.1.5 Modelo de regresión logística variable dependiente depresión y variable explicativa escolaridad.

```
. xi: logit depre_cal i.escol[pweight=p_int ]
i.escol           _Iescol_0-3         (naturally coded; _Iescol_0 omitted)

(sum of wgt is   7.5575e+07)
Iteration 0:   log pseudo-likelihood = -7050.8851
Iteration 1:   log pseudo-likelihood = -6925.5569
Iteration 2:   log pseudo-likelihood = -6911.5166
Iteration 3:   log pseudo-likelihood = -6911.3996
Iteration 4:   log pseudo-likelihood = -6911.3996
```

```
Logit estimates                                   Number of obs   =      38656
                                                  Wald chi2(3)    =     116.32
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -6911.3996                Pseudo R2       =     0.0198

------------------------------------------------------------------------------
             |               Robust
   depre_cal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iescol_1 |    -.57198   .0966156    -5.92   0.000     -.761343   -.3826169
   _Iescol_2 |  -1.457652   .1559573    -9.35   0.000    -1.763323   -1.151981
   _Iescol_3 |  -1.418452   .1927792    -7.36   0.000    -1.796293   -1.040612
       _cons |   -2.44069   .0826398   -29.53   0.000    -2.602661   -2.278719
------------------------------------------------------------------------------
             |               Robust
   depre_cal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iescol_1 |   .5644068   .0545305    -5.92   0.000     .4670388    .6820742
   _Iescol_2 |   .2327823   .0363041    -9.35   0.000     .1714742    .3160101
   _Iescol_3 |   .2420884   .0466696    -7.36   0.000     .1659128    .3532385
------------------------------------------------------------------------------
```

B.1.6 Modelo de regresión logística variable dependiente depresión y variable explicativa estado civil.

```
. xi: logit depre_cal i.civil2[pweight=p_int ]
i.civil2          _Icivil2_0-5        (naturally coded; _Icivil2_0 omitted)

(sum of wgt is   7.5575e+07)
Iteration 0:   log pseudo-likelihood = -7050.8851
Iteration 1:   log pseudo-likelihood =  -6960.832
Iteration 2:   log pseudo-likelihood = -6940.6014
Iteration 3:   log pseudo-likelihood = -6940.3177
Iteration 4:   log pseudo-likelihood = -6940.3176

Logit estimates                                   Number of obs   =      38656
                                                  Wald chi2(5)    =     101.02
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -6940.3176                Pseudo R2       =     0.0157

------------------------------------------------------------------------------
             |               Robust
   depre_cal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Icivil2_1 |   .4890937   .1338732     3.65   0.000      .226707    .7514803
  _Icivil2_2 |   1.154154   .2030846     5.68   0.000     .7561157    1.552193
  _Icivil2_3 |   .8793742   .2330002     3.77   0.000     .4227021    1.336046
  _Icivil2_4 |   1.359186    .156774     8.67   0.000     1.051915    1.666458
  _Icivil2_5 |     .55901   .1724416     3.24   0.001     .2210306    .8969894
       _cons |  -3.602266   .1224435   -29.42   0.000     -3.84225   -3.362281
------------------------------------------------------------------------------
             |               Robust
   depre_cal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Icivil2_1 |   1.630837   .2183254     3.65   0.000     1.254462    2.120136
  _Icivil2_2 |    3.17134   .6440503     5.68   0.000     2.129987    4.721812
  _Icivil2_3 |   2.409391   .5613888     3.77   0.000      1.52608    3.803974
  _Icivil2_4 |   3.893023    .610325     8.67   0.000     2.863128    5.293383
  _Icivil2_5 |    1.74894   .3015901     3.24   0.001     1.247362    2.452209
------------------------------------------------------------------------------
```

B.1.7 Modelo de regresión logística variable dependiente depresión y variable explicativa opinión de su estado de salud.

```
. xi: logit depre_cal i.edo_salud1[pweight=p_int ]
i.edo_salud1      _Iedo_salud_0-2     (naturally coded; _Iedo_salud_0 omitted)

(sum of wgt is   7.5575e+07)
Iteration 0:   log pseudo-likelihood = -7050.8851
Iteration 1:   log pseudo-likelihood = -6885.5366
Iteration 2:   log pseudo-likelihood = -6481.7908
```

```
Iteration 3:   log pseudo-likelihood = -6470.8628
Iteration 4:   log pseudo-likelihood = -6470.8032

Logit estimates                                   Number of obs   =      38656
                                                  Wald chi2(2)    =     413.29
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -6470.8032                Pseudo R2       =     0.0823
------------------------------------------------------------------------------
             |               Robust
   depre_cal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Iedo_salu~1 |   1.450314   .0931442    15.57   0.000     1.267755    1.632874
_Iedo_salu~2 |   2.295545   .1205443    19.04   0.000     2.059282    2.531807
       _cons |  -3.902464   .0743526   -52.49   0.000    -4.048193   -3.756736
------------------------------------------------------------------------------
             |               Robust
   depre_cal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Iedo_salu~1 |   4.264454   .3972093    15.57   0.000     3.552867    5.118562
_Iedo_salu~2 |   9.929845   1.196987    19.04   0.000     7.840341    12.57622
------------------------------------------------------------------------------
```

B.2. Corridas del modelo de regresión logística múltiple a través del paquete STATA para cada variable explicativa.

B.2.1 Modelo de regresión logística variable dependiente depresión y variables explicativas sexo, edad, estado civil, escolaridad, desempleo, condición de aseguramiento y la opinión de su estado de salud.

```
. xi: logit depre_cal  sex  age  desem i.escol i.civil2  i.edo_salud1[pweight=p_int ]
i.escol           _Iescol_0-3         (naturally coded; _Iescol_0 omitted)
i.civil2          _Icivil2_0-5        (naturally coded; _Icivil2_0 omitted)
i.edo_salud1      _Iedo_salud_0-2     (naturally coded; _Iedo_salud_0 omitted)

(sum of wgt is   7.5575e+07)
Iteration 0:   log pseudo-likelihood = -7050.8851
Iteration 1:   log pseudo-likelihood = -6784.8073
Iteration 2:   log pseudo-likelihood = -6309.5993
Iteration 3:   log pseudo-likelihood = -6297.0394
Iteration 4:   log pseudo-likelihood = -6296.9675
Iteration 5:   log pseudo-likelihood = -6296.9675

Logit estimates                                   Number of obs   =      38656
                                                  Wald chi2(13)   =     579.11
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -6296.9675                Pseudo R2       =     0.1069

------------------------------------------------------------------------------
             |               Robust
   depre_cal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .6872499   .1118152     6.15   0.000      .468096    .9064037
         age |   .0066765   .0026692     2.50   0.012      .001445     .011908
       desem |   .0432651   .1051224     0.41   0.681     -.162771    .2493012
   _Iescol_1 |  -.2320965    .103115    -2.25   0.024    -.4341982   -.0299948
   _Iescol_2 |  -.8334929   .1719294    -4.85   0.000    -1.170468   -.4965176
   _Iescol_3 |  -.7075241   .2063239    -3.43   0.001    -1.111912   -.3031366
  _Icivil2_1 |   .0075046   .1411695     0.05   0.958    -.2691824    .2841917
  _Icivil2_2 |   .4913335   .2206583     2.23   0.026     .0588512    .9238158
------------------------------------------------------------------------------
             |               Robust
   depre_cal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Icivil2_3 |   .3030023   .2477124     1.22   0.221    -.1825052    .7885097
  _Icivil2_4 |   .1904655   .1815139     1.05   0.294    -.1652951    .5462262
  _Icivil2_5 |   .1306753   .1770333     0.74   0.460    -.2163036    .4776542
_Iedo_salu~1 |   1.281123   .0963194    13.30   0.000      1.09234    1.469905
_Iedo_salu~2 |   2.015653   .1246085    16.18   0.000     1.771425    2.259881
       _cons |  -4.355173   .2124902   -20.50   0.000    -4.771646   -3.938699
------------------------------------------------------------------------------
```

B.2.2 Modelo de regresión logística variable dependiente depresión y variables explicativas sexo, edad, escolaridad, estado civil y opinión de su estado de salud.

```
. xi: logit depre_cal  sex  age i.escol i.civil2  i.edo_salud1[pweight=p_int ]
i.escol           _Iescol_0-3         (naturally coded; _Iescol_0 omitted)
i.civil2          _Icivil2_0-5        (naturally coded; _Icivil2_0 omitted)
i.edo_salud1      _Iedo_salud_0-2     (naturally coded; _Iedo_salud_0 omitted)

(sum of wgt is   7.5575e+07)
Iteration 0:   log pseudo-likelihood = -7050.8851
Iteration 1:   log pseudo-likelihood = -6785.0726
Iteration 2:   log pseudo-likelihood = -6309.8544
Iteration 3:   log pseudo-likelihood = -6297.2747
Iteration 4:   log pseudo-likelihood = -6297.2028
Iteration 5:   log pseudo-likelihood = -6297.2028

Logit estimates                                   Number of obs   =      38656
                                                  Wald chi2(12)   =     578.25
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -6297.2028                Pseudo R2       =     0.1069

------------------------------------------------------------------------------
             |               Robust
   depre_cal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .7110191   .0939974     7.56   0.000     .5267876    .8952505
         age |   .0068768   .0026581     2.59   0.010     .0016671    .0120866
   _Iescol_1 |  -.2335912   .1033687    -2.26   0.024    -.4361901   -.0309923
   _Iescol_2 |  -.8351985   .1716326    -4.87   0.000    -1.171592   -.4988047
   _Iescol_3 |  -.7137729   .2050587    -3.48   0.000    -1.115681   -.3118652
  _Icivil2_1 |   .0129608    .141903     0.09   0.927     -.265164    .2910857
  _Icivil2_2 |   .4848752   .2177702     2.23   0.026     .0580534    .9116969
  _Icivil2_3 |   .2980277   .2462818     1.21   0.226    -.1846758    .7807313
  _Icivil2_4 |   .1925874   .1816131     1.06   0.289    -.1633677    .5485425
  _Icivil2_5 |   .1356488    .176871     0.77   0.443     -.211012    .4823096
_Iedo_salu~1 |   1.281797   .0964122    13.29   0.000     1.092833    1.470762
_Iedo_salu~2 |   2.017484   .1243764    16.22   0.000      1.77371    2.261257
       _cons |  -4.356859   .2129922   -20.46   0.000    -4.774316   -3.939402
------------------------------------------------------------------------------
```

B.2.3 Modelo de regresión logística variable dependiente depresión y variables explicativas sexo, edad, escolaridad y opinión de su estado de salud.

```
. xi: logit depre_cal  sex  age i.escol i.edo_salud1[pweight=p_int ]
i.escol           _Iescol_0-3         (naturally coded; _Iescol_0 omitted)
i.edo_salud1      _Iedo_salud_0-2     (naturally coded; _Iedo_salud_0 omitted)

(sum of wgt is   7.5575e+07)
Iteration 0:   log pseudo-likelihood = -7050.8851
Iteration 1:   log pseudo-likelihood = -6780.5313
Iteration 2:   log pseudo-likelihood = -6322.0966
Iteration 3:   log pseudo-likelihood = -6308.9396
Iteration 4:   log pseudo-likelihood = -6308.8592
Logit estimates                                   Number of obs   =      38656
                                                  Wald chi2(7)    =     536.62
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -6308.8592                Pseudo R2       =     0.1052

------------------------------------------------------------------------------
             |               Robust
   depre_cal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |    .751559    .091511     8.21   0.000     .5722007    .9309172
         age |   .0080966   .0023988     3.38   0.001     .0033949    .0127982
   _Iescol_1 |  -.2371124   .1033903    -2.29   0.022    -.4397536   -.0344713
   _Iescol_2 |  -.8532536   .1685984    -5.06   0.000      -1.1837   -.5228068
   _Iescol_3 |  -.7367057   .2031459    -3.63   0.000    -1.134864    -.338547
_Iedo_salu~1 |   1.281139   .0964127    13.29   0.000     1.092173    1.470104
_Iedo_salu~2 |   2.017156   .1248637    16.15   0.000     1.772428    2.261884
       _cons |  -4.359799   .1814516   -24.03   0.000    -4.715438   -4.004161
------------------------------------------------------------------------------
```

B.2.4 Prueba de bondad de ajuste χ^2 de Pearson

```
. lfit

Logistic model for depre_cal, goodness-of-fit test

       number of observations =  38656
 number of covariate patterns =  1505
           Pearson chi2(1497) =  1532.98
                 Prob > chi2 =  0.2532
```

B.2.5 Modelo de regresión logística sexo masculino variable dependiente depresión y variables explicativas edad, desempleo escolaridad y opinión de su estado de salud.

```
xi: logit depre_cal age desem i.escol  i.edo_salud1 [pweight=p_int ] if sex==0
i.escol           _Iescol_0-3         (naturally coded; _Iescol_0 omitted)
i.edo_salud1      _Iedo_salud_0-2     (naturally coded; _Iedo_salud_0 omitted)

(sum of wgt is   3.1746e+07)
Iteration 0:   log pseudo-likelihood = -1939.6763
Iteration 1:   log pseudo-likelihood = -1933.9537
Iteration 2:   log pseudo-likelihood = -1784.4259
Iteration 3:   log pseudo-likelihood = -1776.5727
Iteration 4:   log pseudo-likelihood = -1776.4311
Iteration 5:   log pseudo-likelihood = -1776.4311

Logit estimates                                   Number of obs   =      16333
                                                  Wald chi2(7)    =     137.53
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -1776.4311                Pseudo R2       =     0.0842

------------------------------------------------------------------------------
             |               Robust
   depre_cal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0071945   .0041227     1.75   0.081    -.0008858    .0152749
       desem |   .3850446   .1755404     2.19   0.028     .0409918    .7290975
   _Iescol_1 |  -.2476773   .1830106    -1.35   0.176    -.6063715     .111017
   _Iescol_2 |  -.6546645    .310783    -2.11   0.035    -1.263788    -.045541
   _Iescol_3 |  -.6170778   .3325202    -1.86   0.063    -1.268805    .0346499
_Iedo_salu~1 |   1.334363   .1777201     7.51   0.000     .9860384    1.682688
_Iedo_salu~2 |   1.871909   .2400939     7.80   0.000     1.401333    2.342484
       _cons |  -4.439833    .315642   -14.07   0.000     -5.05848   -3.821186
------------------------------------------------------------------------------
```

B.2.6 Modelo de regresión logística sexo femenino variable dependiente depresión y variables explicativas edad, escolaridad y opinión de su estado de salud.

```
. xi: logit depre_cal age i.escol  i.edo_salud1 [pweight=p_int ] if sex==1
i.escol           _Iescol_0-3         (naturally coded; _Iescol_0 omitted)
i.edo_salud1      _Iedo_salud_0-2     (naturally coded; _Iedo_salud_0 omitted)

(sum of wgt is   4.3829e+07)
Iteration 0:   log pseudo-likelihood = -4973.2555
Iteration 1:   log pseudo-likelihood = -4748.4643
Iteration 2:   log pseudo-likelihood = -4520.2854
Iteration 3:   log pseudo-likelihood = -4516.1559
Iteration 4:   log pseudo-likelihood = -4516.1454

Logit estimates                                   Number of obs   =      22323
                                                  Wald chi2(6)    =     357.07
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -4516.1454                Pseudo R2       =     0.0919

------------------------------------------------------------------------------
             |               Robust
   depre_cal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
        age |   .0071502   .0029357     2.44   0.015     .0013963    .0129041
   _Iescol_1 |  -.2260176   .1233884    -1.83   0.067    -.4678543    .0158192
   _Iescol_2 |  -.9311535   .1999897    -4.66   0.000    -1.323126   -.5391808
   _Iescol_3 |  -.8116661   .2559967    -3.17   0.002     -1.31341   -.3099217
_Iedo_salu~1 |   1.259286   .1139403    11.05   0.000     1.035967    1.482605
_Iedo_salu~2 |   2.042219   .1472181    13.87   0.000     1.753677    2.330761
       _cons |  -3.558015   .1959389   -18.16   0.000    -3.942048   -3.173982
------------------------------------------------------------------------------
```

B.3 Cociente de momios

B.3.1 Cociente de momios variable dependiente depresión y variables explicativas género (categoría de referencia es hombre), edad, escolaridad (categoría de referencia es ningún grado de escolaridad) y opinión de su estado de salud (bueno).

```
. xi: logistic depre_cal sex age i.escol  i.edo_salud1 [pweight=p_int ]
i.escol           _Iescol_0-3         (naturally coded; _Iescol_0 omitted)
i.edo_salud1      _Iedo_salud_0-2     (naturally coded; _Iedo_salud_0 omitted)

Logistic regression                               Number of obs   =      38656
                                                  Wald chi2(7)    =     536.62
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -6308.8592                Pseudo R2       =     0.1052

------------------------------------------------------------------------------
             |               Robust
   depre_cal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   2.120303    .194031     8.21   0.000     1.772163    2.536835
         age |   1.008129   .0024183     3.38   0.001     1.003401     1.01288
   _Iescol_1 |   .7889026   .0815648    -2.29   0.022     .6441951    .9661161
   _Iescol_2 |   .4260265   .0718274    -5.06   0.000     .3061438    .5928542
   _Iescol_3 |   .4786882   .0972436    -3.63   0.000     .3214657    .7128053
_Iedo_salu~1 |   3.600738    .347157    13.29   0.000     2.980746    4.349689
_Iedo_salu~2 |   7.516917   .9385897    16.15   0.000     5.885124    9.601164
------------------------------------------------------------------------------
```

B.3.2 Cociente de momios sexo masculino variable dependiente depresión y variables explicativas desempleo, edad, escolaridad (categoría de referencia es ningún grado de escolaridad) y opinión de su estado de salud (bueno).

```
. xi: logistic depre_cal desem i.escol  i.edo_salud1 [pweight=p_int ] if sex==0
i.escol           _Iescol_0-3         (naturally coded; _Iescol_0 omitted)
i.edo_salud1      _Iedo_salud_0-2     (naturally coded; _Iedo_salud_0 omitted)

Logistic regression                               Number of obs   =      16333
                                                  Wald chi2(6)    =     135.03
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -1778.8153                Pseudo R2       =     0.0829

------------------------------------------------------------------------------
             |               Robust
   depre_cal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       desem |    1.60807   .2601102     2.94   0.003     1.171171     2.20795
   _Iescol_1 |   .7159716   .1276889    -1.87   0.061     .5047653    1.015552
   _Iescol_2 |   .4449193   .1305963    -2.76   0.006      .250282    .7909205
   _Iescol_3 |   .4719479   .1521338    -2.33   0.020     .2509035    .8877309
_Iedo_salu~1 |   3.973743   .7195687     7.62   0.000     2.786528    5.666776
_Iedo_salu~2 |   6.968193   1.743498     7.76   0.000     4.267196    11.37883
------------------------------------------------------------------------------
```

B.3.3 Cociente de momios sexo femenino variable dependiente depresión y variables explicativas edad, escolaridad (categoría de referencia es ningún grado de escolaridad) y opinión de su estado de salud (bueno).

```
. xi: logistic depre_cal age i.escol  i.edo_salud1 [pweight=p_int ] if sex==1
i.escol           _Iescol_0-3         (naturally coded; _Iescol_0 omitted)
i.edo_salud1      _Iedo_salud_0-2     (naturally coded; _Iedo_salud_0 omitted)

Logistic regression                               Number of obs   =      22323
                                                  Wald chi2(6)    =     357.07
                                                  Prob > chi2     =     0.0000
Log pseudo-likelihood = -4516.1454                Pseudo R2       =     0.0919

------------------------------------------------------------------------------
             |               Robust
   depre_cal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.007176   .0029568     2.44   0.015     1.001397    1.012988
   _Iescol_1 |   .7977041   .0984274    -1.83   0.067     .6263448    1.015945
   _Iescol_2 |   .3940989   .0788157    -4.66   0.000     .2663015    .5832258
   _Iescol_3 |   .4441175   .1136926    -3.17   0.002     .2689014    .7335044
_Iedo_salu~1 |   3.522905   .4014007    11.05   0.000      2.81783    4.404402
_Iedo_salu~2 |   7.707695   1.134712    13.87   0.000     5.775802    10.28577
------------------------------------------------------------------------------
```

Bibliografía

Agreti Alan, Categorical Data Analysis, 2da Edición, New York Wiley Interscience, 2002.

Allison, Paul D. Logistic Regression using the SAS system, Theory and Aplication.

American Psychiatric Association: Diagnostic and statistical manual of mental disorders 4a ed. Washington, DC: American Psychiatric Association; 1995.

Caraveo J, Martínez N, Rivera B, Polo A. Prevalencia en la vida de episodios depresivos y utilización de servicios especializados. Salud Mental 1997.

Hosmer and Lemeshow, Applied Logistic Regression, 1989 by Jonh Wiley & Sons, Inc.

Kish L. Survey sampling. Nueva York (NY): John Wiley and Sons; 1965.

Kleinbaum David G. Logistic Regression, A Self-Learning Text, 1994 Springer-Verlang, New York, Inc.

Medina-Mora ME, Borges G, Lara C, Benjet C, Blanco J, Fleiz C *et al*. Prevalencia de trastornos mentales y uso de servicios: resultados de la Encuesta Nacional de Epidemiología Psiquiátrica en México. Salud Mental 2003.

Medina-Mora ME, Rascón ML, Tapia R, Mariño M, Juárez F, Villatoro J *et al*. Trastornos emocionales en población urbana mexicana: resultados de un estudio nacional. 1992.

Murray CJ, López AD. Alternative projections of mortality and disability by cause 1990-2020: Global burden of disease study. Lancet 1997.

Organización Mundial de la Salud. Informe sobre la Salud en el Mundo 2001. Salud Mental: nuevos conocimientos, nuevas esperanzas. Ginebra, Suiza: Organización Mundial de la Salud; 2001.

Remick RA. Diagnosis and management of depression in primary care: A clinical update and review. CMAJ 2002.

Silva Aycaguer Luís Carlos, Excursión a la regresión logística en ciencias de la salud. Madrid, España: Ed. Díaz de Santos. 1995.

Stata Corporation 4905 Lakeway Drive College Station, Texas 77845 USA

http://en.wikipedia.org/wiki/Image:Logistic-curve.png#file , 30 de mayo de 2006