UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

"REGRESION LOGISTICA APLICADA A UN ESTUDIO DE HIPERTENSION"

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I A

P R E S E N T A :

DORA MARIA CANTU ORTEGA

FACULTAD DE CIENCIAS
UNAM

DIRECTOR DE TESIS: DRA. GUILLERMINA ESLAVA GOMEZ

2004

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

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

Comunicamos a usted que hemos revisado el trabajo escrito:

"Regresión Logística aplicada a un estudio de hipertensión"

realizado por Cantú Ortega Dora María

con número de cuenta 9851463-2 , quien cubrió los créditos de la carrera de:

Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Dra. Guillermina Eslava Gómez	
Propietario	Dra. Rebeca Aguirre Hernández	
Propietario	Mat. Margarita Elvira Chávez Cano	
Suplente	M.en C. Inocencio Rafael Madrid Ríos	
Suplente	Act. Francisco Sánchez Villareal	

Consejo Departamental de Matemáticas

M.en C. José Antonio Flores Díaz

Dedicatorias y Agradecimientos

A **Dios** por ser mi guía y mi esperanza siempre.

Especialmente a ustedes **Pa** y **Ma** porque sin su apoyo incondicional no sería posible recoger este pequeño fruto de tantos años de esfuerzo mutuo. En verdad, muchas gracias por ser los padres que son, por su ejemplo y entrega total. Por estar conmigo en las buenas y en las malas, por ser tan pacientes y amorosos.

A ti **Amor** porque eres parte de este sueño. Gracias por tu gran apoyo, por alentarme y caminar conmigo siempre pero sobretodo cuando perdía el rumbo. De corazón mil gracias y para ti un logro más juntos.

A **Guillermina Eslava**, mi directora de tesis, por todo su apoyo, paciencia y confianza para realizar esta tesis. Por compartirme su conocimiento y experiencia, pero sobretodo por su espléndida dirección para realizar este trabajo.

A Rebeca Aguirre, Margarita Chávez, Rafael Madrid y Francisco Sánchez, mis **sinodales**, por su valioso tiempo para la revisión de la tesis, pero sobretodo por sus importantes observaciones.

A todos mis maestros por su invaluable conocimiento.

A **Bita** y el **Abuelo**, mis queridos hermanos, por los gratísimos momentos juntos en todo este tiempo de escuela. Por todo lo que me han enseñado con su alegría, entusiasmo y buen sentido del humor.

A mi familia porque cada uno me enseña a luchar dignamente por lo que uno quiere en las circunstancias que le toca vivir.

A mis grandes amigos de la facultad, la **Nana**, **Viole**, **Orie**, **Gaby**, **Pia**, **Mich**, **Argé**, **Oscar**, **Compa**, **Pedro** y **Sama** por la buena y sincera amistad que creamos.

A la **UNAM** por la enorme oportunidad de formar parte de ella, por hoy poder tener una carrera profesional.

INDICE

Resumen

Se expone brevemente la teoría relacionada con la Regresión Logística para poder aplicarla en un estudio de hipertensión arterial. Este estudio fue realizado por la Subdirección en Salud Pública del Instituto Nacional de Perinatología entre el 2001 y 2002 dirigido a un grupo de hombres y mujeres entre 20 y 69 años. Se presenta un análisis exploratorio de las variables antropométricas y las relacionadas con antecedentes de diabetes, obesidad e hipertensión arterial en relación con la presencia de hipertensión arterial. Esto a través de algunas gráficas de dispersión para investigar algún comportamiento respecto a la presencia de hipertensión arterial así como la revisión de algunas tablas cruzadas para las variables categóricas.

Se ajustaron algunas regresiones logísticas para hombres y para mujeres, primero a partir de las variables antropométricas intactas y después estas variables de manera compuesta. Debido a la dificultad práctica de interpretación de las variables compuestas se trabaja con las variables originales. Las bases de datos correspondientes al subgrupo de hombres y mujeres se reducen un 34% y 20% respectivamente al considerar las variables de herencia por lo que sólo se presentan los resultados para el subgrupo de hombres mostrando la regresión que incluye las variables de herencia.

Se obtuvieron modelos a partir de los métodos forward, backward y la metodología propuesta por Hosmer y Lemeshow de estos se seleccionaron los modelos para hombres y mujeres. Se realizan diferentes análisis para la evaluación del ajuste a través de diferentes estadísticas de prueba y algunas gráficas. Se explica de manera detallada la interpretación de los modelos mediante las razones de momios, logitos y probabilidades estimadas. Los modelos propuestos sólo permiten descubrir algunas medidas antropométricas asociadas a la presencia de hipertensión arterial.

Introducción

Con el fin de representar de alguna manera la realidad, el uso de modelos se ha vuelto una herramienta estadística muy usada. Algunos modelos permiten estudiar cómo un conjunto de variables llamadas explicativas influyen en otra variable llamada respuesta. Cabe señalar que los modelos son una simplificación de la realidad por lo que en estricto sentido ningún modelo es del todo correcto. Pero se busca el modelo que mejor ayude a entender al fenómeno.

Un primer acercamiento se tiene al estudiar el modelo de regresión lineal, aplicado a variables respuesta continuas, las cuales se supone tienen distribución normal y varianza constante. Sin embargo, es conveniente tener en cuenta los modelos aplicados a variables respuesta categóricas por su amplio uso en diferentes áreas de estudio.

Por ejemplo, la recolección de información a través de cuestionarios para saber la opinión sobre preferencias (de partidos, de marcas, de comida), características (uso de métodos anticonceptivos, nivel de escolaridad, número de hijos, carrera estudiada, condición de actividad económica, tipos de cáncer), presencia o ausencia de una característica (enfermedad, funcionamiento, beca, crédito financiero) son diferentes tipos de variables categóricas que comúnmente se manejan en investigaciones sociales, epidemiológicas, ecológicas, de salud pública, ingeniería o negocios por mencionar algunas.

En el presente trabajo se expone en el primer capítulo la teoría referente al modelo de regresión logística: en la primera sección se explica brevemente el modelo logístico comparándolo con el modelo de regresión lineal, en la siguiente sección se presenta formalmente el modelo, a continuación se explica la estimación de los parámetros, se mencionan los diferentes métodos de selección de variables, implementados en la mayoría de los paquetes estadísticos: forward y backward, así como la metodología propuesta por Hosmer y Lemeshow. En la última sección se presentan las diferentes interpretaciones de los coeficientes

estimados del modelo logístico. En el segundo capítulo se presenta la aplicación del modelo logístico a un conjunto de datos resultado de una encuesta realizada entre el 2001 y 2002 a un grupo de 2388 individuos en una clínica del ISSSTE al sur de la Ciudad de México. El estudio tuvo por objetivo determinar la asociación entre la estatura, el Índice de Masa Corporal y la presencia de hipertensión arterial en hombres y mujeres entre 20 y 69 años de edad. Para esto se obtuvo mediante un cuestionario una historia clínica, una evaluación antropométrica y la toma de presión arterial. En la historia clínica se recolectaron datos sociodemográficos, antecedentes familiares, de enfermedades crónicas degenerativas, presencia de un diagnóstico previo de hipertensión arterial y, en su caso, el manejo recibido, información obstétrica y ginecológica (paridad, uso de métodos anticonceptivos hormonales, menopausia y terapia de reemplazo hormonal), hábitos como el tabaquismo y el consumo de alcohol, entre otros. Así, la base cuenta con 84 variables de las cuales se seleccionaron 20, las relacionadas con las medidas antropométricas y con los antecedentes de obesidad, diabetes e hipertensión arterial por ser las variables con mayor número de casos válidos, además de que fueron las propuestas por los investigadores para realizar los análisis de regresión logística. En el presente trabajo se pretende ver el efecto de algunas variables antropométricas en la presencia o ausencia de hipertensión arterial. El análisis estadístico se realiza por separado para cada sexo, ya que en las investigaciones epidemiológicas es necesario llevarlas a cabo de manera independiente debido a las diferencias por sexo en el desarrollo de cualquier enfermedad crónica. Se explican ampliamente los resultados obtenidos para el caso de los hombres y se presentan de forma resumida los de mujeres.

Capítulo 1. El modelo de regresión logística

1.1 El modelo de manera intuitiva

El modelo de regresión logística tiene como objetivo encontrar una expresión para explicar la probabilidad de que un evento ocurra dado un conjunto de valores en las variables explicativas. Como la variable dependiente es dicotómica, es decir, toma dos valores que pueden ser etiquetados como "0" o "1", los valores estimados tomarán valores de proporciones o probabilidades condicionales en los valores de las variables independientes.

Si se tuviera el modelo más sencillo, el modelo lineal, se pueden tener estimaciones de probabilidad no lógicas ya que si los valores de las variables independientes son muy grandes o muy pequeños pueden producir probabilidades superiores a uno o inferiores a cero. Es decir, si $p_i = \mathrm{P}[\mathrm{Y_i}=1]$ es la probabilidad de que el individuo i presente la característica de interés y x_{ij} es el valor que toma el individuo i en la variable j, se tiene el modelo lineal:

$$\mathrm{p_i} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \qquad (1.1)$$

Este modelo supone una relación lineal entre las variables explicativas y la dependiente. Pero generalmente un cambio en la variable independiente tiene un menor efecto en la probabilidad cuando la probabilidad de la variable independiente es cercano a cero o a uno que cuando la probabilidad está alrededor a 0.5.

Por ejemplo, si se desea estimar la probabilidad de comprar una casa según el ingreso de una persona, se tendría para personas con ingresos altos que: aumentar su salario en una unidad (digamos $100,000) no modificará mucho la probabilidad de comprar una casa será muy parecida y cercana a uno. De igual

manera sucederá para las personas con muy bajos ingresos pues a pesar de un aumento en los ingresos es muy poco probable que alcance para la compra de una casa por lo que la probabilidad de comprarla será muy parecida y cercana a cero. En cambio para una persona con ingresos medios, un incremento en el ingreso puede producir un efecto mayor en las probabilidad de comprar una casa. (Pampel, 2000. p. 6).

El problema de los límites de la variable respuesta estimados entre cero y uno y la no relación lineal con las variables independientes dan lugar a la transformación logística de la variable estimada que consiste en el logaritmo natural de los momios de presentar la característica,

$$\ln\left[\frac{p_i}{1-p_i}\right] \qquad (1.2)$$

Los momios son un cociente de probabilidades, expresa la probabilidad de éxito, $p_i = P[Y_i = 1]$, relativa a la probabilidad de fracaso, $1-p_i = \mathrm{P}[\mathrm{Y_i}=0]$,

$$\frac{p_i}{1-p_i} = \frac{\mathrm{P}[\mathrm{Y_i}=1]}{1\text{-}\mathrm{P}[\mathrm{Y_i}=1]} = \frac{\mathrm{P}[\mathrm{Y_i}=1]}{\mathrm{P}[\mathrm{Y_i}=0]}$$

Por definición, los momios son no negativos pero lo interesante de sus valores es saber si son mayores o menores a uno. Si los momios son mayores a uno indican que la probabilidad de éxito es mayor a la de fracaso. En cambio si los momios son menores a uno implica que la probabilidad de fracaso es mayor a la de éxito.

Con este primer paso en la transformación, se logra eliminar el límite superior en la probabilidad estimada, teniendo valores mayores o iguales a cero. Al aplicar el logaritmo natural a números menores a uno se tienen valores menores a cero y por lo tanto el lado izquierdo de la ecuación en (1.1) (aplicando esta transformación a la probabilidad estimada) toma valores entre $-\infty$ e ∞, al igual que el lado derecho de la ecuación. A esta transformación se le conoce

también como logito.

La relación entre las variables independientes y las probabilidades estimadas no es lineal por lo que se busca una función que para un cambio en la variable independiente se tenga un menor cambio en las probabilidades cercanas a cero o a uno y mayores en las probabilidades cercanas a 0.5. Al aplicar la transformación logística se logra la relación no lineal de las variables independientes con las probabilidades. Un cambio en una unidad en el logito resulta en menores diferencias en las probabilidades cercanas a cero y a uno que en las cercanas a 0.5. En la Tabla 1.1 se puede observar este comportamiento.

Tabla 1.1
Ralación entre logitos y probabilidades

logito=ln[Pi/(1-Pi)]	-4	-3	-2	-1	0	1	2	3	4
Pi	0.0180	0.0474	0.1192	0.2689	0.5000	0.7311	0.8808	0.9526	0.9820
cambio		*0.0294*	0.0718	0.1497	*0.2311*	0.2311	0.1497	0.0718	*0.0294*

Así la linealidad en los logitos define teóricamente la no linealidad con las probabilidades.

Por lo que la relación lineal de los logitos con las variables independientes está dada por:

$$\ln\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}. \qquad (1.3)$$

Si se despeja la probabilidad se tiene:

$$p_i = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}. \qquad (1.4)$$

1.2 Modelo de Regresión Logística Binaria

El modelo logístico pertenece a la familia de los Modelos Lineales Generalizados por lo que en primer lugar se mencionan las componentes de estos modelos para después presentar el caso Logístico.

1.2.1 Modelos Lineales Generalizados

Los Modelos Lineales Generalizados (MLG) son una familia que contiene una gran variedad de modelos para variables de respuesta tanto categóricas como continuas. A este conjunto de modelos pertenecen el más conocido para variables continuas, el Modelo Lineal y el Modelo Logístico para respuestas binarias.

Todos se caracterizan por ser lineales en alguna transformación de la media de la variable respuesta cuya distribución pertenece a la familia exponencial. Están determinados por tres componentes:

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

$$f(y_i;\theta_i) = \mathrm{a}(\theta_i)\mathrm{b}(y_i)\exp\left[\mathrm{y_i Q}(\theta_i)\right] \tag{1.5}$$

donde $\mathrm{Q}(\theta_i)$ es el parámetro natural. En la siguiente sección se ejemplifica el caso de la distribución Binomial.

2.- Sistemática. También se le conoce como predictor lineal de las variables explicativas, es una función lineal del vector de parámetros $\boldsymbol{\beta}_{p\times 1}$. Las variables explicativas $X_1, X_2, \cdots, X_{p-1}$ pueden ser cualitativas o cuantitativas. El vector de parámetros $\boldsymbol{\beta}_{p\times 1}$ es desconocido.

$$\eta_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} \quad i = 1, 2, \cdots, n \tag{1.6}$$

3.- Función Liga. Relaciona la media de la componente aleatoria, $E(y_i) = \mu_i$, con la sistemática, η_i.

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} \tag{1.7}$$

$g(\mu_i)$ es una transformación de la media expresada como una función lineal de los parámetros $\beta_0, \beta_1, \cdots, \beta_{p-1}$ y además es una función monótona diferenciable.

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

En la Tabla 1.2 se muestran los diferentes casos particulares de los MLG. Se puede ver que se tienen modelos para variables respuesta continuas, binarias, de conteo o politómicas cuyas distribuciones asociadas: Normal, Binomial, Poisson o Multinomial pertenecen a la familia exponencial. También se presentan las ligas aplicadas a la media de la variable respuesta, los tipos de variables explicativas que manejan y el nombre del modelo.

Tabla 1.2 Algunos casos particulares de los Modelos Lineales Generalizados

Componente aleatoria	Función liga	Componente sistemática	Modelo
Normal	Identidad	Cont. y Cat.	Regresión
Normal	Identidad	Categóricas	Análisis de Varianza
Normal	Identidad	Cat. y Cont.	Análisis de Covarianza
Binomial	Logit	Cont. y Cat.	Regresión Logística
Poisson	Log	Cont. y Cat.	Loglineal
Multinomial	Logit	Cont. y Cat.	Respuesta multinomial

(Ver Agresti, 2000. p. 118)

1.2.2 El modelo logístico

Sea una muestra aleatoria de n- observaciones independientes $Y_1, Y_2, \cdots, Y_n$ con:

$$Y_i \sim \text{Bin}(n_i,\ p_i)$$

con i=1,2,$\cdots$,n y $p_i = P[Y_i = 1]$, es decir,

p_i es la probabilidad de observar un éxito para cada uno de los n_i individuos con vector de variables explicativas $\mathbf{X}_i = (X_{i1}, X_{i2}, \cdots, X_{ip})$, $0 \le p_i \le 1$. Y

$$E(Y_i) = n_i p_i = \mu_i$$
$$Var(Y_i) = n_i p_i (1 - p_i)$$

Se dice que las observaciones no están agrupadas si $n_i = 1$ para toda i, es decir, la información está por individuo.

Por lo tanto, el Modelo Logístico o Modelo de Regresión Binaria Logística puede expresarse como:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} \qquad i = 1, 2, \cdots, n \tag{1.8}$$

$$\Rightarrow p_i = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij}\right)} \qquad (1.9)$$

De la expresión en (1.9) se tiene que no es un modelo lineal. Cabe señalar que si entre las variables explicativas alguna es categórica no es conveniente incluirla como si fuera variable continua por lo que se deben construir variables indicadoras para las diferentes categorías. Así, si una variable nominal tiene k posibles categorías, entonces k-1 variables indicadoras serán necesarias.

Dentro de los MLG existen varias funciones liga que permiten hacer la relación entre la media de la respuesta Y_i y el predictor lineal. En el modelo logístico, la función liga es la logística, y se denota como logit(p_i).

$$g(\mu_i) = \ln\left(\frac{\mu_i}{n_i - \mu_i}\right) = \ln\left(\frac{p_i}{1-p_i}\right) = \text{logit}(p_i)$$

A continuación se muestra la distribución para una variable respuesta binomial, con el fin de identificar su parámetro natural que define a la función liga del modelo logístico.

Si $n_i = 1$, la distribución para la variable aleatoria Y es:

$$\begin{aligned} f(y_i; p_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ &= (1-p_i)\left[\frac{p_i}{1-p_i}\right]^{y_i} \\ &= (1-p_i)\exp\left[y_i \ln\left(\frac{p_i}{1-p_i}\right)\right] \end{aligned}$$

Por analogía con los elementos de la distribución de un MLG en (1.5) se tiene:

$$a(\theta_i) = a(p_i) = (1-p_i)$$
$$b(y_i) = 1$$
$$Q(\theta_i) = Q(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$$

Entonces el parámetro natural es $\ln\left(\frac{p_i}{1-p_i}\right)$ y es liga canónica. A los modelos lineales generalizados que usan esta función se les llama Modelos Logito, (Agresti, 2002. p. 117).

Otras funciones liga que describen el comportamiento de la probabilidad de éxito en términos de las variables explicativas son:

i) Función Identidad.

$$g(\mu_i) = p_i$$

$$\Rightarrow p_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij}$$

Se le llama Modelo lineal de probabilidad, es un MLG con componente aleatoria Binomial y función liga identidad. Tiene el inconveniente de los límites en la probabilidad de éxito entre cero y uno.

ii) La Inversa de la Distribución Normal.

$$g(\mu_i) = \Phi^{-1}(p_i)$$

$$\Rightarrow p_i = \int_{-\infty}^{\beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij}} \frac{1}{\sqrt{2\Pi}} \exp\left(-\frac{1}{2} z^2\right) dz$$

Se conoce como modelo Probit, no es un modelo lineal.

iii) El Doble Logaritmo Complementario

$$g(\mu_i) = \ln\left[-\ln(1-p_i)\right]$$

$$\Rightarrow p_i = 1 - \exp\left[-\exp\left(\beta_0 + \sum_{j=1}^{p-1}\beta_j X_{ij}\right)\right]$$

Esta función se utiliza para tratar asimétricamente a los éxitos y fracasos.

1.3 Estimación de los parámetros

Para ajustar el modelo logístico (1.9) según un conjunto de datos se requiere estimar los valores de los parámetros desconocidos β_i y uno de los métodos más utilizados es el de Máxima Verosimilitud (MV).

De manera general, el método de MV consiste en maximizar una función llamada "Función de Verosimilitud" que es la función de probabilidad asociada a un conjunto de datos en función de los parámetros desconocidos. De modo que los estimadores que se obtienen maximizan la probabilidad de obtener los datos observados.

Para el modelo logístico se supone una muestra aleatoria de n-observaciones $Y_1, Y_2, \cdots, Y_n$ independientes con distribución Bernoulli(p_i)

$$f_{Y_i}(y_i, p_i) = p_i^{y_i}(1-p_i)^{1-y_i} \quad y_i : 0,1. \qquad (1.10)$$

La función de densidad conjunta de $Y_1, Y_2, \cdots, Y_n$ es:

$$f_{Y_1, Y_2, \cdots, Y_n}(y_1, y_2, \cdots, y_n) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i},$$

por lo que la Función de Verosimilitud para p_i es:

$$l(p_1,\cdots,p_n)=\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

$$\text{con} \quad p_i=\frac{\exp(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})}{1+\exp(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})},$$

la función de verosimilitud para el modelo logístico está en función de los parámetros desconocidos $\beta_0,\beta_1,\cdots,\beta_p$ y se busca maximizar esta función respecto a estos parámetros, $\boldsymbol{\beta}=(\beta_0,\beta_1,\cdots,\beta_p)$.

$$l(\boldsymbol{\beta})=\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{1.11}$$

Para encontrar los estimadores MV se necesita derivar esta función respecto a cada parámetro desconocido para ello primero se obtiene el logaritmo de la función porque es una transformación monótona y por tanto la solución es la misma.

$$\begin{aligned} L(\boldsymbol{\beta})=\ln\left[l(\boldsymbol{\beta})\right] &= \sum_{i=1}^{n}\left\{y_i\ln(p_i)+(1-y_i)\ln(1-p_i)\right\} \\ &= \sum_{i=1}^{n}\left\{y_i\ln\frac{p_i}{(1-p_i)}+\ln(1-p_i)\right\} \\ &= \sum_{i=1}^{n} y_i(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})+\sum_{i=1}^{n}\ln\left(\frac{1}{1+\exp(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})}\right) \\ &= \sum_{i=1}^{n} y_i(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})-\sum_{i=1}^{n}\ln\left(1+\exp(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})\right) \end{aligned} \tag{1.12}$$

Derivando esta ecuación respecto a β_0 y β_j se obtiene las p+1 ecuaciones de

verosimilitud:

$$\frac{\partial \mathrm{L}(\boldsymbol{\beta})}{\partial(\beta_0)} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}$$

$$\frac{\partial \mathrm{L}(\boldsymbol{\beta})}{\partial(\beta_j)} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \frac{x_{ij} \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})} \quad \mathrm{j}=1,2,\cdots,\mathrm{p}$$

En regresión lineal, las ecuaciones de verosimilitud, son lineales en los parámetros y por lo mismo fáciles de resolver pero en el caso del modelo logístico las ecuaciones de verosimilitud no son lineales en los parámetros por lo que se resuelven por métodos numéricos iterativos que usualmente están incluidos en los paquetes estadísticos.

Generalmente el programa inicia con un modelo en el que todos los coeficientes son iguales a los estimados por Mínimos Cuadrados. Después sucesivamente el algoritmo escoge un nuevo conjunto de coeficientes que producen un mayor valor en $L(\boldsymbol{\beta})$. Este proceso continúa hasta que el incremento en el valor de $L(\boldsymbol{\beta})$ es muy pequeño y cuando los coeficientes cambian poco, (Pampel, 2000. p. 44).

El método para estimar las varianzas y covarianzas de los coeficientes estimados es a partir del definido en la teoría de estimación por MV. Se obtiene la segunda derivada parcial de $L(\boldsymbol{\beta})$. Las segundas derivadas parciales tienen la forma :

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j^{\,2}} = -\sum x_{ij}^{\,2} p_i (1-p_i)$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_l} = -\sum x_{ij} x_{il} p_i (1-p_i) \qquad j,l=0,1,2,\cdots,p$$

Estas ecuaciones dan lugar a la matriz de Información Observada $\mathrm{I}(\boldsymbol{\beta})$ y las varianzas y covarianzas de los coeficientes estimados se obtienen de la inversa de esta matriz $\mathrm{Var}(\boldsymbol{\beta})=\hat{\mathrm{I}}^{-1}(\boldsymbol{\beta})$.

La fórmula de la matriz de información es:

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

donde X es una matriz n x (p+1) que contiene la información de cada individuo y V es una matriz n x n diagonal con los elementos $\hat{p}_i(1-\hat{p}_i)$.

$$\mathrm{X}=\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \quad \text{y} \quad \mathrm{V}=\begin{bmatrix} \hat{p}_1(1-\hat{p}_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{p}_n(1-\hat{p}_n) \end{bmatrix}$$

1.4 Evaluación del ajuste del modelo

Para saber si el modelo se ajusta a los datos, se pueden hacer por lo menos tres tipos de evaluaciones: 1) Ver si existe de manera conjunta alguna relación entre las variables explicativas y la dependiente. 2) Una vez que se sabe que el modelo tiene sentido interesa explorar qué tanto influye cada una de las variables explicativas en la variable respuesta. 3) Sobre cada una de las observaciones, para ver si el modelo se comporta bien para todas las observaciones.

1.4.1 Medidas de bondad de ajuste

Las medidas de bondad de ajuste indican si en términos generales el modelo ajustado describe apropiadamente la relación entre la variable respuesta y las variables independientes. Los MLG usan como medidas a las estadísticas Ji-Cuadrada de Pearson Generalizada (χ^2) y la Devianza (D), ambas miden la distancia entre los valores observados y los ajustados.

A través de estas estadísticas se contrastan las hipótesis:

H_0 = El modelo describe apropiadamente los datos

o $\text{logit}(p_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$

vs

H_1 = El modelo es inadecuado

o $\text{logit}(p_i) \neq \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$

La Ji-Cuadrada de Pearson Generalizada es:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\text{V}}\text{ar}(y_i)}$$

para el Modelo Logístico

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

$$\text{con} \quad r_i = \frac{y_i - n_i \hat{p}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}} \qquad (1.13)$$

Para la Devianza de cualquier MLG se tiene:

$L(\hat{\mu}_i; y_i)$, el máximo del logaritmo natural de la función de verosimilitud del modelo ajustado.

$L(y_i; y_i)$, el máximo del logaritmo natural de la función de verosimilitud del modelo saturado. A pesar de que este modelo no es parsimonioso (tiene un parámetro para cada observación), sirve como comparación con otros modelos porque tiene el ajuste perfecto (los valores observados y los ajustados bajo el modelo saturado son iguales), (Agresti, 2002. p. 139).

se define como sigue:

$$D = -2\ln\left[\frac{\text{Func. Veros mod. Ajustado}}{\text{Func. Veros mod. Saturado}}\right] = -2\left[L(\hat{\mu}_i; y_i) - L(y_i; y_i)\right] = -2\left[L_m - L_s\right]$$

Indica qué tan bien describe el modelo ajustado al modelo saturado. Se aplica –2ln para tener una cantidad que se distribuye asintóticamente como χ^2 y el número de grados de libertad es igual a la diferencia entre el número de patrones de covariables (número de combinaciones diferentes observadas de las variables explicativas) y el número de parámetros estimados.

Para el modelo logístico la función de verosimilitud se define como en (1.12) y para diferenciar el modelo saturado del ajustado se denotan las probabilidades de éxito como sigue:

En el modelo saturado:

$$\tilde{p}_i = \frac{y_i}{n_i}$$

y para el modelo ajustado:

$$\hat{p}_i = \frac{\exp\left(\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j X_{ij}\right)}{1 + \exp\left(\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j X_{ij}\right)}$$

Sean

$L(\tilde{p}_i; y_i) = L_s = \ln$ [Función de Verosimilitud del modelo saturado]

$L(\hat{p}_i; y_i) = L_m = \ln$ [Función de Verosimilitud del modelo ajustado]

Entonces la Devianza para el modelo Logístico es:

$$D = -2[L_m - L_s] = 2[L_s - L_m]$$

$$\Rightarrow D = 2\sum_{i=1}^{n} \left\{ \begin{array}{l} \left[y_i \ln(\tilde{p}_i) + (n_i - y_i)\ln(1-\tilde{p}_i) \right] - \\ \left[y_i \ln(\hat{p}_i) + (n_i - y_i)\ln(1-\hat{p}_i) \right] \end{array} \right\}$$

$$= 2\sum_{i=1}^{n} \left[y_i \ln\left(\frac{\tilde{p}_i}{\hat{p}_i}\right) + (n_i - y_i)\ln\left(\frac{(1-\tilde{p}_i)}{(1-\hat{p}_i)}\right) \right]$$

$$= 2\sum_{i=1}^{n} \left[y_i \ln\left(\frac{n_i\tilde{p}_i}{n_i\hat{p}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i(1-\tilde{p}_i)}{n_i(1-\hat{p}_i)}\right) \right]$$

$$\Rightarrow D = 2\sum_{i=1}^{n} \left[y_i \ln\left(\frac{y_i}{n_i\hat{p}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i(1-\hat{p}_i)}\right) \right]$$

Se trata de una medida de bondad de ajuste que compara los valores observados con los ajustados. Cabe señalar que cuando se desarrolla para el modelo lineal se obtiene la expresión de la suma de cuadrados de la desviación de los valores observados y ajustados, $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. Entonces la estadística de bondad de ajuste basada en los residuos de la Devianza es:

$$D = \sum_{i=1}^{n} d_i^2$$

$$\text{con} \quad d_i = \pm \sqrt{2 \sum_{i=1}^{n} \left[y_i \ln\left(\frac{y_i}{n_i \hat{p}_i} \right) + (n_i - y_i) \ln\left(\frac{n_i - y_i}{n_i (1 - \hat{p}_i)} \right) \right]} \qquad (1.14)$$

Tanto la estadística de Pearson como la Devianza, bajo el supuesto de H_0 y de n_i "grandes" en un número fijo de niveles o categorías en las variables explicativas se distribuyen asintóticamente χ^2 con grados de libertad igual a la diferencia del número de patrones de covariables y el número de parámetros estimados, (Agresti, 2002. p. 175).

Cuando se tiene alguna variable explicativa continua la estadística D no es asintóticamente χ^2 y en lugar de D se usa la estadística de Hosmer-Lemeshow. La mayoría de las variables a analizar en la aplicación de la regresión logística son continuas por lo que conviene ilustrar en este apartado el uso de la estadística D a través de un ejemplo que contiene variables categóricas.

Ejemplo de modelo saturado

Un modelo saturado es aquel que tiene igual número de parámetros estimados que de patrones de covariables (en el caso de solo tener variables categóricas) o bien tiene un parámetro por cada observación (en el caso de tener variables continuas).

Supóngase que se realizarán votaciones electorales en tres diferentes regiones sean Zona1, Zona2 y Zona3 y se desea conocer la probabilidad de votar. También se conoce el número de votantes por zona y sexo de encuestas previas. De esta manera se tienen 6 combinaciones posibles de observar: el número de votantes en la zona1 y son hombres, en la zona1 y son mujeres y así sucesivamente. El modelo saturado para este ejemplo sería el que tiene igual número de parámetros estimados como de patrones de covariables, es decir, 6. Si

se consideran como categorías de referencia a los hombres para la variable sexo y a la zona3 para las regiones (en la sección 1.6.2 se explica la construcción de variables indicadoras para variables politómicas), se tiene el modelo saturado:

$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{muj} + \beta_2 \text{zona1} + \beta_3 \text{zona2} + \beta_4 \text{muj*zona1} + \beta_5 \text{muj*zona2}$$

y un modelo reducido para este ejemplo puede ser cualquier caso particular del saturado: el que no considera ninguna variable explicativa; aquel que solo tiene efectos principales (sexo o regiones); aquel que considera ambas variables explicativas. A continuación se enuncian.

$$\text{logit}(p_i) = \beta_0,$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{muj},$$

$$\text{logit}(p_i) = \beta_0 + \beta_2 \text{zona1} + \beta_3 \text{zona2},$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{muj} + \beta_2 \text{zona1} + \beta_3 \text{zona2}.$$

Cabe señalar que las estadísticas de bondad de ajuste D y χ^2 pueden usarse cuando el número de patrones de covariables es pequeño comparado con el número total de individuos, lo cual generalmente sucede cuando todas las variables explicativas son cualitativas. En el caso de que el número de patrones de covariables es "aproximadamente" igual al número total de individuos estudiados se propone agrupar los datos con base en las probabilidades estimadas y construir la estadística $\hat{C}$, (Hosmer y Lemeshow, 2000. p.147).

Se ordenan los individuos según sus probabilidades estimadas en orden creciente y se forman g-grupos (se recomienda $g \geq 6$) y la estadística se obtiene:

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)} \qquad (1.15)$$

donde $\bar{\pi}_k = \sum_{i=1}^{C_k} \frac{n_i \hat{p}_i}{\dot{n}_k}$ es la probabilidad promedio de observar un éxito en el grupo k

con C_k = Número de patrones de covariables en el grupo k

$\dot{n}_k$ = Número de individuos en el grupo k

O_k = Número de éxitos observados en el grupo k

Hosmer y Lemeshow observaron a través de simulaciones que cuando el número de patrones de covariables se mantiene fijo conforme crece el tamaño de la muestra y el modelo ajustado es apropiado, la distribución de la estadística se aproxima a una distribución χ^2 con (g-2) grados de libertad. En la aplicación se ilustra el uso de esta estadística, la cual generalmente la calculan los paquetes que incluyen el módulo de regresión logística.

1.4.2 Pruebas sobre los coeficientes

Las pruebas de bondad de ajuste solo dicen si el modelo ajustado describe o no a los datos por lo que la potencia de la prueba para rechazar Ho es débil. Ahora interesa saber si cada una de las variables explicativas influyen de manera importante en la variable respuesta.

Una manera más efectiva de probar la significancia de las variables explicativas en el modelo es a través de pruebas de hipótesis condicionales, que consisten en comparar modelos donde uno es un caso particular del otro, es decir, modelos anidados. Para llevar a cabo estas pruebas se utiliza el Cociente de Verosimilitudes y se plantean pruebas de hipótesis como la siguiente, donde se compara el modelo de la media ($\beta_j = 0 \quad \forall j=1,2,\cdots p$) contra el modelo que incluye variables explicativas ($\beta_j \neq 0$, para alguna j):

$$H_0: \text{ logit } (p_i) = \beta_o$$

vs

$$H_1: \text{ logit } (p_i) = \beta_o + \sum_{i=1}^{p} \beta_{ij} x_j$$

que equivale a decir, si tiene sentido incluir en el modelo al menos una de las variables explicativas.

La estadística de prueba basada en el Cociente de Verosimilitudes se calcula como la diferencia de las devianzas de los dos modelos:

$$D_{H_0} = -2\left(L_{H_0} - L_s\right)$$
$$D_{H_1} = -2\left(L_{H_1} - L_s\right)$$

$$\Rightarrow G^2 = D_{H_0} - D_{H_1} = -2\left(L_{H_0} - L_{H_1}\right) \qquad (1.16)$$

El modelo definido en H_0 es el caso particular del modelo definido en H_1 por lo que la devianza en el modelo definido en H_0 es mayor a la del modelo con más parámetros estimados, el definido en H_1. Por lo que si la diferencia es pequeña los modelos son equivalentes y si es grande indica que las variables en el modelo general ayudan a describir la variable respuesta. Para conocer los grados de libertad se tiene:

Los grados de libertad de la D_{H_0} = Número de patrones de covariables-

Número parámetros estimados en H_0

Los grados de libertad de la D_{H_1} = Número de patrones de covariables-

Número parámetros estimados en H_1

$\Rightarrow gl\left(D_{H_0}\right) - gl\left(D_{H_1}\right)$ = Núm de parámetros estimados en H_1 - Núm de parámetros estimados en H_0

$$\therefore \; G^2 \sim \chi^2_{(\#\text{ parám } H_1 \; - \; \#\text{ parám } H_0)}$$

Otra forma de probar la significancia de las variables independientes es a través de la estadística de Wald:

$$W = \frac{\beta_j}{SE\left(\beta_j\right)} \tag{1.17}$$

la cual bajo la hipótesis nula de $\beta_j = 0$, se distribuye como una Normal Estándar. También se pueden obtener intervalos de confianza para cada coeficiente estimado para ver si está contenido el cero $\hat{\beta}_j \pm z_{\alpha/2} SE\left(\hat{\beta}_j\right)$.

1.4.3 Medidas de diagnóstico

Una vez que se verifica que el modelo describe adecuadamente los datos y que las variables independientes son significativas se procede a examinar los residuos para determinar qué datos están bien representados por el modelo y cuáles no.

Las medidas de diagnóstico son índices que se calculan a partir de:

1.-Los residuos (De Pearson o de la Devianza)

2.-Las palancas

h_i = i-ésimo elemento en la diagonal de la matriz sombrero H

$$H = V^{1/2} X \left(X' V X\right)^{-1} X' V^{1/2}$$

Si se desea ver la influencia del patrón de covariables 'i' en el vector de

coeficientes estimados $\left(\hat{\boldsymbol{\beta}}\right)$ se tiene la medida:

$$\Delta\boldsymbol{\beta}_{(-i)} = \frac{r_i^2 h_i}{(1-h_i)^2} \tag{1.18}$$

Para detectar posibles influencias en las estadísticas de Bondad de Ajuste:

a) En la estadística χ^2:

$$\Delta\,\chi^2_{(-i)} = \frac{r_i^2}{1-h_i}$$

b) En la Devianza:

$$\Delta D_{(-i)} = d_i^2 + \frac{r_i^2 h_i}{(1-h_i)}$$

Una vez que se calculan las medidas se grafican los índices contra el patrón de covariables para detectar posibles patrones de covariables influyentes con el fin de analizar que hacer con ellos o bien contra las probabilidades estimadas, en el caso de tener alguna variable explicativa continua, (Agresti, 2002. p. 225).

1.5 Selección del modelo

Existen varios procedimientos para seleccionar el modelo que permitirá hacer algunas inferencias sobre el fenómeno estudiado. Sin embargo, también es necesario tener en cuenta la opinión de los expertos en el fenómeno de estudio para no descartar variables que son importantes en la explicación de la variable de interés y que tal vez los criterios estadísticos no logran evidenciarlo.

Como un primer acercamiento al modelo pueden servir, los procedimientos de selección automática de variables que generalmente vienen incluidos en los

paquetes estadísticos, Forward y Backward. Aunque es conveniente tomar con reserva los resultados y a la vez consultar a los expertos.

Los procedimientos automáticos de selección o eliminación de variables consisten en un algoritmo que revisa la importancia de las variables para incluirlas o eliminarlas según un valor de referencia que depende de los supuestos del modelo. Para la regresión Logística se supone una distribución binomial y su significancia se mide a través del valor de significancia (p-value) de la prueba basada en el Cociente de Verosimilitudes. Así, en cualquier paso del algoritmo la variable más importante, en términos estadísticos, es aquella que produce el mayor cambio en el logaritmo natural de la función de verosimilitud respecto al logaritmo natural de la función de verosimilitud del modelo que no contiene a la variable, esto es, la variable que produce el mayor valor en la estadística del cociente de verosimilitudes G^2, en el método Forward. De manera equivalente, para el método Backward elimina la variable que produce el menor cambio en el logaritmo natural de la función de verosimilitud respecto al logaritmo natural de la función de verosimilitud del modelo que contiene todas las variables, es decir, la variable que produce el menor valor en la estadística del cociente de verosimilitudes.

A continuación se explica paso a paso el algoritmo del método forward para la selección de variables.

Paso (0). Inicia una vez que se tienen todas las p-variables consideradas potencialmente influyentes en la variable respuesta. Se hace un primer ajuste en el que no se incluye ninguna variable explicativa, es decir, el modelo que contiene solo a la constante y se obtiene el logaritmo natural de su función de verosimilitud, llámese L_0. Se ajustan p-regresiones logísticas univariadas que incluyen a cada variable, se compara el logaritmo natural de la función de verosimilitud de cada

modelo $L_j^{(0)}$[1] con L_0, es decir, se obtiene la estadística de cociente de verosimilitudes para cada modelo $G_j^{(0)} = -2\left(L_0 - L_j^{(0)}\right)$, que compara el modelo que contiene cada variable con el que solo tiene la constante. El nivel de significancia de la estadística de prueba se calcula como, $P\left[\chi^2_{(\nu)} > G_j^{(0)}\right] = p_j^{(0)}$, donde $\nu = 1$ si X_j es continua y $\nu = k-1$ si X_j es politómica con k categorías. Se elige la variable con el menor nivel de significancia X_{e_1}, es decir, $p_{e_1}^{(0)} = \min\left(p_j^{(0)}\right)$. Esto indica que la variable X_{e_1} es la que más diferencia provoca en la diferencia de verosimilitudes con respecto al modelo de la constante y por ende la variable debe ser considerada. Cabe señalar que la variable seleccionada como importante no necesariamente es estadísticamente significativa, es decir $p_{e_1}^{(0)}$ no siempre es menor a 0.05.

La importancia de las variables se decide a partir de un nivel de significancia preestablecido p_E. Se ha encontrado que un $p_E = 0.05$ es un nivel muy estricto porque excluye variables importantes del modelo, se recomienda un nivel entre 0.15 y 0.20. Una vez elegido este nivel la variable es considerada como importante si el nivel de significancia para G es menor a p_E.

Si $p_{e_1}^{(0)} < p_E$ el algoritmo pasa al siguiente paso, si no ahí termina.

Paso (1). Ahora se tiene el modelo ajustado que contiene a X_{e_1} y el valor del logaritmo natural de su función de verosimilitud $L_{e_1}^{(1)}$. Para determinar si alguna de las otras p-1 variables son importantes, una vez que entró la variable X_{e_1}, se ajustan p-1 regresiones logísticas que contienen X_{e_1} y X_j $j = 1,2,3,\cdots,p$ y $j \neq e_1$.

---

[1] El número en el paréntesis indica el paso en el procedimiento.

Sea $L^{(1)}_{e_1}$ el logaritmo natural de la función de verosimilitud del modelo que contiene a X_{e_1} y X_j; $G^{(1)}_j = -2\left(L^{(1)}_{e_1} - L^{(1)}_{e_1 j}\right)$ es el valor de la estadística del cociente de verosimilitudes del modelo. Su nivel de significancia se denotará como $p^{(1)}_j$. Supóngase X_{e_2} es la variable con el menor nivel de significancia en este paso, es decir, $p^{(1)}_{e_2} = \min\left(p^{(1)}_j\right)$. Si este nivel es menor a p_E se pasa al siguiente paso de otra manera aquí termina el algoritmo.

Paso (2). Se tiene el modelo que incluye X_{e_1} y X_{e_2}. En este punto puede suceder que la variable X_{e_1} deje de ser importante al entrar la variable X_{e_2}. Para revisar este aspecto se examina la importancia de cada variable mediante pruebas parciales de cociente de verosimilitudes donde se prueba el modelo que incluye a una de las variables contra el que contiene a ambas. Así, se tiene el caso de aplicar el método backward ya que a partir de un modelo más complejo se trata de ver si es posible tener uno más sencillo, es decir, se revisa si alguna de las dos variables no es importante. Para esto es necesario establecer un nuevo nivel de significancia p_R que tiene que ser mayor a p_E para evitar que el programa meta y saque la misma variable en diferentes pasos. De manera explícita se tiene:

Sea $L^{(2)}_{-e_j}$ el logaritmo natural de la función de verosimilitud del modelo con X_{e_j} removida; $G^{(2)}_{-e_j} = -2\left(L^{(2)}_{-e_j} - L^{(2)}_{e_1 e_2}\right)$ la estadística de cociente de verosimilitudes del modelo sin la variable j contra el que contiene a ambas y $p^{(2)}_{-e_j}$ su nivel de significancia. El programa selecciona la variable, X_{r_2}, que al ser removida deja el mayor nivel de significancia, es decir, $p^{(2)}_{r_2} = \max\left(p^{(2)}_{-e_j}, p^{(2)}_{-e_2}\right)$. Para decidir si X_{r_2} debe ser removida, el programa compara $p^{(2)}_{r_2}$ con p_R. Entonces si $p^{(2)}_{r_2} > p_R$ la variable es removida del modelo ya que indica una contribución mínima al modelo.

Si se tienen las dos variables como importantes, se examinan las p-2

regresiones logísticas que contienen X_{e_1}, X_{e_2} y $X_j \quad j=1,2,3,\cdots,p \quad j \neq e_1, e_2$. El programa evalúa el logaritmo natural de la función de verosimilitud de cada modelo que contiene sólo X_{e_1} y X_{e_2} y el del modelo con las tres variables explicativas y calcula el correspondiente nivel de significancia. Sea X_{e_3} la variable con el mínimo nivel de significancia en la prueba, esto es, $p_{e_3}^{(2)} = \min\left(p_j^{(2)}\right)$. Si este nivel es menor que p_E, el programa pasa al paso (3), de otro modo termina.

Paso (3). Es idéntico al paso (2). El programa ajusta el modelo incluyendo la variable seleccionada en los pasos previos, hace una revisión de las variables anteriormente introducidas y continúa la selección. Así prosigue hasta el paso (S).

Paso (S). Se llega a este paso cuando: 1) Todas las variables entraron al modelo; 2) Todas las variables en el modelo tienen niveles de significancia para remover menores a p_R y las variables no incluidas en el modelo tienen niveles de significancia mayores a p_E, (Hosmer y Lemeshow, 2000. p.116-119).

El método de eliminación o Backward se diferencia del Forward en que inicia con un modelo con todas las variables que se considere influyen potencialmente en la variable respuesta y secuencialmente va eliminando términos en lugar de incrementar y lo va haciendo como en la revisión parcial del paso(2). En cada etapa, elimina el término que menos aporta al ajuste del modelo. que ahora es identificado por el mayor nivel de significancia, en las pruebas de cociente de verosimilitudes.

En cualquier proceso, para variables cualitativas con más de dos categorías, considera a la variable entera en lugar de variables dicotómicas, es decir, incluye o elimina a la variable completa. La misma condición aplica para las interacciones que contienen esa variable, (Agresti, 2002. p.214).

Es importante usar estos métodos de selección con cuidado . Por ejemplo, el tamaño de una muestra puede subestimar el efecto real de algunas variables y

reflejar en ellas efectos débiles. Por lo que una selección cuidadosa en la formulación de modelos también es necesaria.

Hosmer y Lemeshow proponen una serie de pasos para la selección del modelo a explicar:

1. Estimar modelos logísticos univariados para todas las variables independientes que se consideren influyen en la variable respuesta y seleccionar aquellas que tengan un nivel de significancia menor a 0.25 y también aquellas que se consideren importantes en el área de estudio aunque no cumplan con la regla del nivel de significancia. La recomendación de 0.25 se basa en el trabajo de Bendel y Afifi (1977) en Regresión Lineal y en el trabajo de Mickey y Greenland (1989) en Regresión Logística. Ellos mostraron que el uso de 0.05 provocaba muy a menudo no considerar variables que se sabía eran importantes y el uso de un nivel mayor tenía la desventaja de incluir variables que no se consideraban muy importantes.

2. Se construye un nuevo modelo múltiple con las variables seleccionadas y se estiman sus coeficientes. Es necesario corroborar la significancia de las variables ya sea con la estadística de Wald o mediante la comparación de la estimación de los coeficientes en los modelos univariados con los de este último. Las variables que no cumplan con las condiciones deben eliminarse y ajustar un nuevo modelo. El nuevo modelo debe ser comparado con el anterior (más grande) mediante una prueba de Cociente de Verosimilitudes. Nuevamente se deben comparar los valores de los estimadores del modelo reducido con el general. Es importante identificar las variables que difieren mucho en su estimación. Esto indica que una o más de las variables excluídas eran importantes en el sentido de que proveían un mejor ajuste a la variable que queda en el modelo. Este proceso de eliminar, reevaluar y verificar continúa hasta que parece que todas las variables importantes están incluídas en el modelo y eliminadas aquellas no tan importantes

estadísticamente o así se consideran dentro del estudio.

3. Se sugiere que cualquiera de las variables no incluídas en el modelo múltiple original sean integradas al modelo para identificar variables que por sí solas no son significativas respecto a la variable respuesta y que pueden serlo en presencia de otras variables.

4. Una vez que se tiene un modelo con las variables más importantes se pueden revisar las interacciones entre las variables del modelo. Para decidir la inclusión de una interacción se necesita considerar las cuestiones prácticas y estadísticas, (Hosmer y Lemeshow, 2000. p.92-99).

1.6 Interpretación de los coeficientes

La interpretación de un modelo ajustado requiere la capacidad de extraer inferencias prácticas de los coeficientes estimados en el mismo, es decir, se desea saber qué indican los coeficientes estimados acerca de la variable respuesta o pregunta de investigación.

Generalmente cuando la variable respuesta tiene una transformación no lineal como la logística se pueden tener tres formas para interpretar los coeficientes estimados para el modelo. Esto es, el efecto sobre los logitos que es el más directo pero con menos significado intuitivo, el efecto en las razones de momios que es el más utilizado y el efecto en las probabilidades.

1.6.1 Efecto en los logitos

Los coeficientes estimados asociados a las variables independientes representan la pendiente de una función de la variable dependiente por unidad de cambio en la variable independiente. Así, la interpretación involucra dos aspectos: Determinar la relación funcional entre la variable dependiente y las variables independientes, y una definición apropiada de unidad de cambio para la variable independiente.

En el caso de la regresión logística la función que enlaza la variable dependiente con el predictor lineal es la transformación logística.

$$g(x)=\ln\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right)=\beta_0+\beta_1 x \qquad (1.19)$$

Para el modelo de regresión lineal el coeficiente que representa la pendiente, β_1, es igual a la diferencia entre el valor de la variable respuesta en $x+1$ y el valor de la variable respuesta en x para algún valor de X. En este caso, la interpretación de los coeficientes es relativamente directa, ya que representa un cambio en la escala de medida de la variable dependiente por unidad de cambio en la variable independiente.

El coeficiente β_1 de la regresión logística muestra el cambio en el logaritmo natural de los momios de presentar una característica, al cambiar una unidad en la variable independiente.

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

Cabe señalar que la interpretación del coeficiente dependerá de la habilidad para dar significado a la diferencia de dos logitos.

A continuación se considera la interpretación de las razones de momios estimadas para los tres tipos de variables independientes posibles: dicotómica, politómica o continua.

1.6.2 Efecto en la razón de momios

Para explicar la interpretación de los coeficientes estimados resulta muy útil iniciar con el caso en que la variable independiente es dicotómica ya que este caso permite entender el concepto fundamental para los otros casos.

Variable independiente dicotómica

Se supone que la variable independiente está codificada como cero o uno, la diferencia en los logitos para un individuo con X=1 y X=0 es:

$$g(1)-g(0)=[\beta_0+\beta_1]-[\beta_0]=\beta_1$$

Para poder interpretar este resultado se necesita introducir la medida de asociación llamada Razón de Momios (RM).

Sean:

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

$$\pi(x)=P(Y=1|X=x)$$

Los momios de presentar la característica para una persona con X=1:

$$\frac{\pi(1)}{1-\pi(1)} \tag{1.20}$$

Los momios de presentar la característica para una persona con X=0.

$$\frac{\pi(0)}{1-\pi(0)} \tag{1.21}$$

Entonces la Razón de Momios se define como el cociente de los dos momios (1.20) y (1.21):

$$RM=\frac{Momio(Y=1|X=1)}{Momio(Y=1|X=0)}=\frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(0)}{1-\pi(0)}} \tag{1.22}$$

Los valores que toma el modelo logístico en cada momio se resumen en la siguiente tabla:

Tabla 1.3 Valores en el modelo logístico cuando la variable independiente es dicotómica

	Var Independiente	
Var Respuesta	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	$1-\pi(1)=\frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\pi(0)=\frac{1}{1+e^{\beta_0}}$

Si se sustituyen estos valores en la ecuación (1.22) se obtiene:

$$RM=\frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\Big/\frac{1}{1+e^{\beta_0+\beta_1}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}}\Big/\frac{1}{1+e^{\beta_0}}}=\frac{\frac{\left(e^{\beta_0+\beta_1}\right)\cdot\left(1+e^{\beta_0+\beta_1}\right)}{\left(1+e^{\beta_0+\beta_1}\right)}}{\frac{\left(e^{\beta_0}\right)\cdot\left(1+e^{\beta_0}\right)}{\left(1+e^{\beta_0}\right)}}=\frac{\left(e^{\beta_0+\beta_1}\right)}{\left(e^{\beta_0}\right)}=e^{\beta_1}$$

Por lo tanto, para una regresión logística con variable independiente binaria codificada por ejemplo como cero o uno, la relación que existe entre la razón de momios y el coeficiente de regresión es la exponencial del coeficiente.

$$RM=e^{\beta_1} \tag{1.23}$$

La razón de momios es una medida de asociación que se ha usado ampliamente en epidemiología porque aproxima cuánto es más probable (o

improbable) que la variable respuesta esté presente en individuos con X=1 que en aquellos con X=0.

La interpretación dada a las razones de momios se basa en el hecho de que en muchas ocasiones se aproxima a una cantidad llamada Riesgo Relativo (RR) que es igual a :

$$RR = \frac{\pi(1)}{\pi(0)} \tag{1.24}$$

Entonces se tiene que la Razón de Momios se aproxima al Riesgo Relativo cuando

$$\frac{1-\pi(0)}{1-\pi(1)} \approx 1$$

lo cual sucede cuando $\pi(0)$ y $\pi(1)$ son pequeñas.

Para obtener más información del valor del parámetro se recomienda usar intervalos de confianza para la razón de momios, además de su estimador puntual.

En teoría, para muestras grandes, la distribución de la RM es normal. Pero desafortunadamente este requisito no se tiene en la mayoría de las investigaciones. Por lo que, las inferencias se hacen con base en la distribución muestral del $\ln\left(\hat{RM}\right) = \hat{\beta}_j$, el cual tiende a seguir una distribución Normal, en muestras relativamente pequeñas, (Hosmer y Lemeshow, 2000. p.52).

Un intervalo de confianza del $100\times(1-\alpha)\%$ de confianza estimado para la Razón de Momios se obtiene primero calculando los puntos extremos del intervalo para el coeficiente estimado, $\hat{\beta}_j$, y después tomando su exponencial.

Entonces los puntos extremos del intervalo de confianza para estimar la

razón de momios están dados por la expresión:

$$\exp\left\{\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} S\hat{E}\left(\hat{\beta}_j\right)\right\} \tag{1.25}$$

Debido a la importancia que tienen las razones de momios como medida de asociación, los paquetes estadísticos como SPSS y STATA, por mencionar algunos, proveen tanto sus estimadores puntuales como sus intervalos.

Variable independiente politómica

Ahora se supone que la variable independiente tiene más de dos categorías. En este caso es indispensable tratarla de un modo especial, es decir, no se le puede considerar como variable categórica. Se debe construir un conjunto de variables indicadoras para las variables politómicas.

Si se tiene una variable con k-categorías, se necesitará construir $k-1$ variables indicadoras, (aquellas a las que se les asigna uno en caso de tener la característica y cero cuando no) ya que la k-ésima variable indicadora resulta redundante al quedar determinadas las anteriores. Se determina una de las k-categorías como variable de referencia (aquella que queda determinada por la condición de las otras variables indicadoras), la cual estará representada dentro de cada variable indicadora por un cero de modo que las $k-1$ variables indicadoras estarán formadas por ceros y un uno en la categoría de la variable indicadora en cuestión.

Por ejemplo, si se tiene a la variable Partido Político y se tienen 4 posibles respuestas, el partido A, B, C o D, se puede elegir a cualquiera de los 4 partidos como categoría de referencia , por ejemplificar se elige al partido A. Entonces se tienen 3 nuevas variables indicadoras para la variable Partido Político con la siguiente forma:

Tabla 1.4 Ejemplo de variables indicadoras para una variable politómica

	Variables indicadoras		
	Partido_1	Partido_2	Partido_3
Partido A	0	0	0
Partido B	1	0	0
Partido C	0	1	0
Partido D	0	0	1

De modo que una persona que elige PartidoA tendrá valor cero para las tres variables indicadoras y si elige cualquier otro partido tendrá ceros en dos de las variables indicadoras y uno en la que representa al Partido elegido.

De esta manera se obtendrán $k-1$ coeficientes estimados, es decir, uno para cada variable indicadora con el fin de hacer comparaciones con el grupo de referencia. Esto es, obtener la Razón de Momios para el grupo en cuestión respecto al grupo de referencia.

Así, se vuelve a tener el caso de la variable independiente dicotómica, en donde el cambio en el logito es, de 1 a 0, de presentar a no presentar la característica de la variable independiente. Aunque en este caso se refiere a comparar el grupo en cuestión con el de referencia.

En el ejemplo de la variable Partido Político, si se quiere conocer la razón de momios de votar por el Partido C en lugar del Partido A (porque es la categoría de referencia) se tiene:

Primero se obtiene:

$$\ln\left[\hat{RM}(\text{ParC, ParA})\right] = \hat{g}(\text{ParC}) - \hat{g}(\text{ParA})$$
$$= \left[\hat{\beta}_0 + \hat{\beta}_1(\text{Par\_1=0}) + \hat{\beta}_2(\text{Par\_2=1}) + \hat{\beta}_3(\text{Par\_3=0})\right] -$$
$$\left[\hat{\beta}_0 + \hat{\beta}_1(\text{Par\_1=0}) + \hat{\beta}_2(\text{Par\_2=0}) + \hat{\beta}_3(\text{Par\_3=0})\right]$$
$$= \hat{\beta}_2$$

$\hat{\beta}_2$ corresponde al coeficiente estimado para la variable indicadora del Partido C.

Ahora se obtiene la Razón de Momios:

$$\hat{RM}(\text{ParC, ParA}) = \exp\left(\hat{\beta}_2\right),$$

entonces se dice que es más (menos), probable que una persona vote a favor del Partido C en comparación con el Partido A.

De igual manera que en el caso dicotómico, se obtienen los Intervalos de Confianza para las diferentes Razones de Momios estimadas para cada variable indicadora (Ver 1.24).

Cuando las categorías de la variable cualitativa guardan una relación de orden, es decir, la variable es ordinal se tiene el mismo manejo que las variables politómicas pero además si se observa que existe alguna tendencia lineal en sus parámetros estimados o bien si los parámetros estimados son muy parecidos, pueden tratarse como cuantitativas con lo que se obtiene un modelo más sencillo siempre y cuando también ajuste bien a los datos (ver ejemplo en Agresti, 2002. p.190).

Variable independiente continua

Cuando un modelo de regresión logística tiene una variable independiente continua la interpretación del coeficiente estimado depende de cómo ingresó al modelo (orden o función de la variable) y de la unidad de la variable.

Para propósitos de interpretación del caso de una variable continua se

supone que el logito es lineal en la variable independiente, es decir, $g(x)=\beta_0+\beta_1 x$. De aquí se sigue que el coeficiente de la pendiente, β_1, da el cambio en el logito para un cambio en una unidad en X, esto es, $\beta_1=g(x+1)-g(x)$ para cualquier valor de X.

Muy a menudo el valor de uno, clínicamente, no es muy interesante. Por ejemplo un año de incremento en la edad o un mmHg de incremento en la presión sanguínea. Por otro lado si el rango de X es de cero a uno entonces un cambio de uno es demasiado grande. Por lo tanto para dar una interpretación más útil a variables continuas se recomienda estimar de manera puntual y por intervalo el cambio de "c" unidades en la variable.

El logito para un cambio de "c" unidades en X se obtiene de la diferencia de los logitos: $g(x+c)-g(x)=\hat{\beta}_0+\hat{\beta}_1 x+c\hat{\beta}_1-\left[\hat{\beta}_0+\hat{\beta}_1 x\right]=c\hat{\beta}_1$ y la razón de momios asociada se obtiene al aplicarle la función exponencial al logito,

$$\hat{RM}(c)=\hat{RM}(x+c,x)=\exp\left(c\hat{\beta}_1\right)$$

Y los puntos extremos del Intervalo de Confianza del $100\times(1-\alpha)\%$ para estimar la Razón de Momios se da por:

$$\exp\left\{c\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} c\hat{SE}\left(\hat{\beta}_j\right)\right\},$$

y se dice que por un incremento en "c" unidades en la variable independiente, es $\exp(c\beta_1)$ veces más (menos) probable de ocurrir la variable respuesta. Cabe señalar que la validez de este razonamiento se cuestiona porque tal vez las probabilidades de ocurrencia no se manejan de igual manera a lo largo de los diferentes valores de la variable continua pero esto sucede cuando se modela linealmente la variable independiente con el logito. Si se cree que la variable independiente no es lineal con el logito se puede agrupar y usar la variable como

dicotómica, o bien, usar términos de mayor orden (x^2, x^3, ...) u otra relación no lineal (log(x)), (Hosmer y Lemeshow, 2000. p.63)

Cabe señalar que el efecto de las variables independientes sobre la Razón de Momios es multiplicativo. Así el valor estimado de la variable respuesta no cambia cuando se multiplica por uno. Es más, la exponencial de un número positivo, $\hat{\beta}_j > 0$, es mayor a uno y de un número negativo, $\hat{\beta}_j < 0$, es menor a uno pero mayor a cero. Por lo tanto un coeficiente estimado igual a cero no altera la razón de momios pero los coeficientes mayores a uno aumentan la razón de momios y los menores a uno la disminuyen. Además mientras más alejado esté el coeficiente de uno en cualquier dirección mayor es el efecto en la Razón de Momios.

1.6.3 Efecto en las probabilidades

Como la relación entre la variable independiente y las probabilidades no es lineal, no puede ser representada por un solo coeficiente. Así, el efecto sobre las probabilidades depende del punto en el cual se mide el cambio.

Una manera directa de ver la influencia de una variable continua en las probabilidades es calculando la recta tangente a la curva no lineal en un punto determinado. Para variables dicotómicas, el cambio relevante ocurre de 0 a 1, y la recta tangente debido a pequeños cambios en X tiene menos sentido. Por lo que se calculan probabilidades estimadas para cada grupo para medir la diferencia en probabilidades .

Sin embargo, los cambios en las probabilidades por unidad de cambio en X difieren de los de la recta tangente a la curva logística. Así, en lugar de utilizar la derivada parcial se pueden utilizar probabilidades estimadas tanto para variables continuas como dicotómicas.

Las diferentes formas de interpretar los efectos en las probabilidades indican la dificultad de resumir la relación no lineal. Algunos autores recomiendan evitar estos tipos de interpretación y enfocarse en los cambios en los momios.

Capítulo 2. Aplicación del modelo logístico a un estudio de hipertensión arterial

Introducción

Una vez presentados los conceptos más importantes sobre el ajuste del modelo de regresión logística se ilustra su uso e interpretación en un estudio de hipertensión arterial.

Dentro de la investigación epidemiológica el uso de los modelos logísticos se ha vuelto una herramienta muy útil en la identificación y estimación de posibles factores de riesgo en el desarrollo de algunas enfermedades. Esto con el fin de tener mejores programas de prevención de las mismas.

Se cuenta con una base de datos que fue proporcionada por el Dr. Héctor Ávila Rosas y la M. en C. Martha Kaufer Horwitz quienes pertenecen al departamento de Salud Pública de la Facultad de Medicina en la UNAM. Ellos realizaron una investigación prospectiva para describir la relación que existe entre el Índice de Masa Corporal y la estatura baja con la presencia de hipertensión arterial. Por lo que se consideró oportuno llevar a cabo un análisis de Regresión Logística a este tipo de datos, con el fin de explotar la información contenida en la base, además de ejemplificar su uso en un problema médico.

Los modelos que a continuación se presentan tratan de explicar la hipertensión arterial en función de un conjunto de mediciones en los individuos, y algunas variables relacionadas con la herencia. Además se interpretan con detalle los principales estimadores de los modelos.

2.1 Planteamiento del problema

A pesar de que la obesidad es un factor etiológico de enfermedades como la hipertensión arterial, se han encontrado evidencias de que el índice de masa corporal, principal indicador de obesidad, no refleja el porcentaje de grasa corporal total de la misma manera a lo largo de las diferentes estaturas. Esto es, el IMC no es independiente de la estatura y por lo tanto individuos con estatura baja tienden a acumular mayor cantidad de grasa en la región central que los de mayor estatura. También se sabe que a mayor contenido de grasa corporal – sobretodo en la región central – mayor es el riesgo de desarrollar algunas enfermedades crónicas como la hipertensión arterial. Por lo que resulta más informativo la distribución de la grasa que su indicador como tal.

En el presente estudio se pretende encontrar algunos factores de riesgo en la presencia de hipertensión arterial, pero principalmente factores relacionados con la antropometría de los individuos. En este caso, de hombres y mujeres entre 20 y 69 años.

2.2 Metodología

Se trata de un estudio realizado por la Subdirección de Investigación en Salud Pública del Instituto Nacional de Perinatología. La recolección de la información se llevó a cabo durante el 2001 y 2002 en la Clínica de Medicina Familiar "Dr Ignacio Chávez" del ISSSTE, ubicada al sur de la Ciudad de México.

Los individuos sujetos de medición fueron hombres y mujeres entre 20 y 69 años que acudieron a la clínica y voluntariamente aceptaron formar parte del estudio y quedaron excluídos aquellos individuos que de alguna manera su condición de salud influía en el objetivo del estudio que era encontrar la asociación entre el IMC, la composición corporal y la presencia de hipertensión arterial, es decir personas que:

Se habían sometido a una dieta de reducción de peso en los últimos seis meses y habían perdido peso

Habían sufrido una amputación o con ausencia congénita de algún miembro

Realizaban actividad física intensa o ejercicios como el físicoculturismo o levantamiento de pesas

Mujeres embarazadas o con lactancia exclusiva o que habían dado a luz en los seis meses previos al estudio.

El estudio consistió en la valoración de un grupo de 2388 individuos de ambos sexos. La Tabla2.1 muestra la distribución de la muestra por grupos de edad y por sexo.

Tabla 2.1 Distribución de la muestra por grupos de edad y sexo

			SEXO DE LOS INDIVIDUOS		
			MASCULINO	FEMENINO	Total
GRUPOS DE EDAD	20 A 29 AÑOS	Num de casos	220	243	463
		%	19.8%	19.0%	19.4%
	30 A 39 AÑOS	Num de casos	221	252	473
		%	19.9%	19.7%	19.8%
	40 A 49 AÑOS	Num de casos	227	276	503
		%	20.5%	21.6%	21.1%
	50 A 59 AÑOS	Num de casos	220	270	490
		%	19.8%	21.1%	20.5%
	60 A 69 AÑOS	Num de casos	222	237	459
		%	20.0%	18.5%	19.2%
Total		Num de casos	1110	1278	2388
		%	100.0%	100.0%	100.0%

La evaluación se llevó a cabo en una sola sesión con excepción de aquellos en los que se detectó una Presión Arterial (PA) normal alta (**PA sistólica ≥ 130 mmHg y/o PA diástolica ≥85mmHg**), en cuyo caso se les citó a una segunda evaluación de la presión arterial (dentro de los siguientes 8 días) con el fin de confirmar o descartar el diagnóstico de hipertensión arterial (HTA). En la Tabla 2.2 se muestra el número de casos diagnosticados con HTA por sexo.

Tabla 2.2 Proporciones observadas de HTA por sexo

			hta si y no		
			normal	hta	Total
SEXO DE LOS INDIVIDUOS	MASCULINO	Núm de casos	850	218	1068
		%	79.6%	20.4%	100.0%
	FEMENINO	Núm de casos	987	259	1246
		%	79.2%	20.8%	100.0%
Total		Núm de casos	1837	477	2314
		%	79.4%	20.6%	100.0%

La evaluación antropométrica consistió de la medición del peso, la estatura, la talla sentado, los panículos adiposos (tricipital, bicipital, subescapular, suprailíaco y del muslo), las circunferencias de la cintura y de la cadera, anchura de codo y de rodilla, el diámetro biacromial y bicrestal. A partir de estas mediciones se obtuvieron algunos indicadores de composición corporal como: el Índice de Masa Corporal, el Índice Cintura Talla y el Índice Cintura Cadera. La descripción de estas variables se puede ver en la Tabla 2.4

Antes de iniciar con los análisis de regresión logística fue necesario preparar las bases de datos para el subgrupo de hombres y mujeres para lo cual primero se revisaron los casos válidos en la variable de interés, la presencia de HTA, posteriormente para las variables antropométricas y por último para las variables de herencia: antecedentes de diabetes, obesidad o hipertensión arterial en alguno o en ambos padres. Se encontraron 74 casos faltantes en la variable de HTA, 150 en las variables antropométricas y 400 al considerar las variables de herencia. La Tabla 2.3 muestra el porcentaje de casos válidos para cada base respecto a la base inicial.

Tabla 2.3 Casos válidos para cada base filtrada

Variables consideradas		Sexo		Total
		Masculino	Femenino	
Todas	Núm de casos	1110	1278	2388
	%Base inicial	100.0%	100.0%	100.0%
HTA	Núm de casos	1068	1246	2314
	%Base inicial	96.2%	97.5%	96.9%
HTA y antropométricas	Núm de casos	1004	1160	2164
	%Base inicial	90.5%	90.8%	90.6%
HTA, antrop y de herencia	Núm de casos	735	1029	1764
	%Base inicial	66.2%	80.5%	73.9%

2.3 Selección de los modelos logísticos

En este estudio, la variable respuesta binaria corresponde al diagnóstico de hipertensión arterial: 1 si fue diagnosticado y 0 si no (HTA) y las variables explicativas son las medidas antropométricas de cada individuo (BICIPII, TRICIPII, SUBESI, SUPRAI, PESOI, IMC, CINTURAI, CADERAI, ICC, ICT, TALLAI, SENTADOI, BIACROI, BICRESI, RODILLAI, CODOI), la EDAD y las relacionadas con antecedentes de hipertensión arterial (ANTHTA), de obesidad (ANTOBES) y de diabetes (ANTDIAB) donde éstas son variables dicotómicas y representan 1 si alguno o ambos padres presentaron la enfermedad y 0 si no, para estas variables la categoría de referencia es la codificada como "0", esto quiere decir que las comparaciones en las Razones de Momios son respecto a los que no tienen algún antecedente de hipertensión en sus padres. Cabe señalar que las 3 variables dicotómicas fueron recodificadas a partir de una variable politómica, con el fin de ver el efecto de la presencia o no de alguna enfermedad en alguno de sus padres o en ambos. En la Tabla 2.4 se describen brevemente estas variables.

Tabla 2.4 Descripción de las variables explicativas para llevar a cabo los diferentes análisis de regresión logística

Grupos de variables antropométricas	Descripción	Variable	Valores y códigos
	Hipertensión arterial	HTA	1-sí, 0-No
	Edad	EDAD	20-69 (años)
Panículos adiposos	Pliegue Bicipital.- Cara frontal del brazo	BICIPII	mm
	Pliegue tricipital.- Cara posterior del brazo	TRICIPII	mm
	Pliegue subescapular.- Pliegue en la espalda	SUBESI	mm
	Pliegue suprailíaco.- Pliegue en la cadera	SUPRAI	mm
Masa Corporal	Peso corporal	PESOI	Kg
	Indice de Masa Corporal.- peso/estatura^2	IMC	Kg/cm^2
Masa Central	Circunferencia de cintura	CINTURAI	cm
	Circunferencia de cadera	CADERAI	cm
	Indice Cintura Cadera.- cintura/cadera	ICC	-
	Indice Cintura Talla.- cintura/estatura	ICT	-
Complexión	Estatura	TALLAI	cm
	Longitud sentado.- Segmento superior	SENTADOI	cm
	Diámetro biacromial.- En los hombros	BIACROI	cm
	Diámetro bicrestal.- En la cadera	BICRESI	cm
	Anchura de rodilla	RODILLAI	mm
	Anchura de codo	CODOI	mm
	1a. Comp de las vars antropométricas	pcom_gpo	-
	2a. Comp de las vars antropométricas	scom_gpo	-
	1a. Comp del grupo de panículos adiposos	pcom_pli	-
	2a. Comp del grupo de panículos adiposos	scom_pli	-
	1a. Comp del grupo de masa corporal	pcom_imc	-
	2a. Comp del grupo de masa corporal	scom_imc	-
	1a. Comp del grupo de masa central	pcom_cad	-
	2a. Comp del grupo de masa central	scom_cad	-
	1a. Comp del grupo de complexión	pcom_ose	-
	2a. Comp del grupo de complexión	scom_ose	-
	Antecedentes de obesidad en los padres	ANTOBESI	1-sí, 0-No
	Antecedentes de diabetes en los padres	ANTDIAB	1-sí, 0-No
	Antecedentes de hta en los padres	ANTHTA	1-sí, 0-No

2.3.1 Análisis exploratorio

Primeramente se decidió trabajar con las bases que tienen casos válidos para las variables antropométricas debido a que sólo se pierde alrededor del 10% de la base original para hombres y mujeres. Además se conservan las proporciones de la variable de interés, HTA, de la base completa (ver Tabla 2.5). Cabe señalar que para la base que además incluye las variables relacionadas con la herencia también se conservan estas proporciones.

Se obtuvieron las gráficas de dispersión para las variables continuas respecto a la variable HTA de donde se pudieron identificar 6 casos atípicos, los cuales se revisaron con detalle respecto a otras variables antropométricas para confirmar su eliminación por error de captura o de medición. Con lo cual la base se redujo a 2158 casos válidos. En la Tabla 2.5 se puede ver la proporción de HTA por sexo para estos casos.

Tabla 2.5 Proporciones observadas de HTA por sexo, en la base con casos válidos para las variables antropométricas y sin casos atípicos

			hta si y no		
			normal	hta	Total
SEXO DE LOS INDIVIDUOS	MASCULINO	Núm de casos	798	203	1001
		%	79.7%	20.3%	100.0%
	FEMENINO	Núm de casos	916	241	1157
		%	79.2%	20.8%	100.0%
Total		Núm de casos	1714	444	2158
		%	79.4%	20.6%	100.0%

Para las variables categóricas se obtuvieron tanto tablas cruzadas respecto a HTA como gráficas de barra para ver si existía alguna diferencia en proporciones de presencia de HTA por tener algún antecedente de obesidad, diabetes o hipertensión arterial. Gráficamente se pudo observar que sólo el antecedente de HTA en alguno o ambos padres presenta una mayor proporción de HTA que los

NORMALES. En el caso de las mujeres se observa diferencia de proporción de HTA en la variable referente al estado de menopausia, de modo que a medida que la etapa de la menopausia avanza, mayor es la proporción de mujeres con presencia de HTA. Ver gráficas en el Anexo 3.

De las gráficas para las variables continuas podría pensarse que se tiene una diferencia en el comportamiento de las variables: edad, pliegue subescapular, pliegue suprailíaco, los diámetros biacromial y bicrestal , la cintura, el peso, la rodilla y los índices: cintura-cadera, cintura-talla y el índice de masa corporal ya que para todas ellas las observaciones correspondientes a los HTA están ligeramente por arriba de los NORMALES. Para revisar de manera numérica estas variables se obtuvieron sus medias por sexo y por su diagnóstico de HTA o NORMALES. Estas medias se presentan en la Tabla 2.6. Las medias de cada variable para hombres y mujeres está por debajo de la media para el grupo de HTA de cada sexo a excepción de la variable TALLAI y SENTADOI lo que confirma que las personas con menor estatura tienden a acumular más grasa en la región central y por ende son más propensos a presentar HTA. Comparando las medias de los Normales con la de los HTA se tienen las mayores diferencias en las variables: EDAD, SUBESI, CINTURAI, CADERAI, PESOI, ICT. Por lo que estas variables pueden ser factores importantes en la presencia de HTA.

Tabla 2.6 Medias de las variables antropométricas por sexo y por HTA

Variables	Hombres			Mujeres		
	Total	Normales	HTA	Total	Normales	HTA
EDAD	44.8	**41.9**	**56.1**	44.5	**41.2**	**56.7**
T.A. Sistólica	119.5	116.6	133.0	113.7	110.3	129.1
T.A. Diastólica	77.2	75.5	85.2	73.1	71.5	80.4
TALLAI	167.8	**168.1**	**166.8**	154.2	**154.7**	**152.5**
BICIPII	9.7	9.4	11.0	14.9	14.2	17.5
TRICIPII	12.4	12.0	13.7	22.5	22.0	24.3
SUBESI	21.2	**20.4**	**24.5**	25.0	24.5	26.7
SUPRAI	19.4	18.8	21.8	24.6	24.0	27.2
BIACROI	40.8	40.7	40.9	36.1	36.1	36.3
BICRESI	31.1	30.9	32.0	29.8	29.5	31.1
CINTURAI	94.1	**92.6**	**100.2**	85.7	**84.2**	**91.5**
CADERAI	100.4	**99.7**	**103.2**	101.7	**100.6**	**105.9**
SENTADOI	87.2	**87.4**	**86.5**	82.8	**83.1**	**81.9**
RODILLAI	9.7	9.6	9.8	8.9	8.9	9.2
CODOI	6.7	6.6	6.8	5.8	5.8	6.0
PESOI	77.2	**76.0**	**81.7**	65.7	**64.6**	**69.9**
IMC	27.4	26.9	29.3	27.6	27.0	30.0
ICC	0.9	0.9	1.0	0.8	0.8	0.9
ICT	56.1	55.1	60.1	55.7	**54.5**	**60.1**

A partir de estas variables analizadas se pretende encontrar un modelo para el subgrupo de hombres y uno para el de mujeres que pueda identificar, en lo posible, las relaciones de éstas con la presencia de HTA.

2.3.2 Selección de variables

En principio, se hizo una regresión exploratoria con las variables originales, esto es, tomando a las variables antropométricas con sus mediciones intactas y a la edad. En la tabla 2.7 se puede ver que según la estadística de Wald sólo la edad, el pliegue subescapular y el diámetro bicrestal son estadísticamente significativas al 5%. Sin embargo al entrar todas las variables antropométricas y a la vez sus índices pueden tenerse problemas de multicolinelidad, esto es, algunos factores pueden estar altamente correlacionados y resulta redundante tomar a ambos como es el caso del peso con el IMC (ver en el Anexo 3 la Tabla de

correlaciones). Por esta razón se ajusta nuevamente la regresión considerando únicamente las variables antropométricas y se observa que además de las variables significativas en la regresión anterior aparece el peso como variable significativa.

Tabla 2.7 Parámetros estimados para la regresión logística considerando todas las variables antropométricas, para hombres

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	**EDAD**	**.092**	**.010**	**93.998**	**1**	**.000**	**1.097**	**1.077**	**1.117**
	BICIPII	-.049	.030	2.607	1	.106	.952	.898	1.010
	TRICIPII	.028	.023	1.486	1	.223	1.029	.983	1.076
	SUBESI	**.053**	**.017**	**9.576**	**1**	**.002**	**1.055**	**1.020**	**1.091**
	SUPRAI	.000	.016	.000	1	.983	1.000	.970	1.032
	PESOI	.054	.164	.109	1	.741	1.056	.766	1.456
	CINTURAI	.122	.377	.106	1	.745	1.130	.540	2.364
	CADERAI	.155	.205	.573	1	.449	1.168	.782	1.744
	TALLAI	-.220	.139	2.507	1	.113	.802	.611	1.054
	SENTADOI	.039	.027	2.185	1	.139	1.040	.987	1.095
	BIACROI	.012	.043	.079	1	.779	1.012	.930	1.101
	BICRESI	.077	.038	4.068	1	.044	1.080	1.002	1.164
	RODILLAI	-.134	.164	.664	1	.415	.875	.634	1.207
	CODOI	.078	.224	.120	1	.729	1.081	.696	1.677
	IMC	-.036	.455	.006	1	.937	.964	.395	2.354
	ICT	-.496	.614	.651	1	.420	.609	.183	2.030
	ICC	17.644	21.784	.656	1	.418	4.6E+07	.000	2.E+26
	Constant	5.327	27.791	.037	1	.848	205.766		

a. Variable(s) entered on step 1: EDAD, BICIPII, TRICIPII, SUBESI, SUPRAI, PESOI, CINTURAI, CADERAI TALLAI, SENTADOI, BIACROI, BICRESI, RODILLAI, CODOI, IMC, ICT, ICC.

Esta primera regresión se comentó con los investigadores del estudio y de ahí se realizaron dos nuevas regresiones: con las variables antropométricas como variables compuestas ya que ellos piensan que alguna de las variables originales pesa demasiado como el IMC o el peso corporal y no permite descubrir variables también importantes en la explicación de la hipertensión. Para esto, se realizaron cinco Análisis de Componentes Principales (ACP's), uno que incluía a todas las variables antropométricas y otros cuatro resultado de la subdivisión de este grupo donde uno estaba formado por variables relacionadas con la masa corporal

(PESOI, IMC), otro con los panículos adiposos (TRICIPII, SUBESI, BICIPII, SUPRAI), otro con la masa central (CINTURAI, CADERAI, ICC, ICT) y un último con la complexión (TALLAI, SENTADOI, BIACROI, BICRESI, CODO, RODILLA). De ellos se observó que la primera componente presentaba pesos muy parecidos y positivos por lo que representaban un promedio de las variables consideradas, mientras que las otras componentes tenían pesos positivos y negativos. Este comportamiento es característico en variables relacionadas con variables antropométricas y de manera general se dice que la primera componente habla del tamaño de la persona y las siguientes de la forma del cuerpo (Blackitch y Reyment, p.170).

Para llevar acabo las regresiones logísticas con estas nuevas variables se tomaron de cada análisis la primera y segunda componente. Así se observó que los resultados en las variables significativas en cada regresión eran equivalentes. Por ejemplo para el subgrupo de los hombres cuando se ajustó la regresión con las variables intactas resultaron significativas el peso, el pliegue subescapular y la edad. Por otro lado al ajustar la regresión con la primera y segunda componente del ACP's de todas las variables, se obtuvieron estas dos variables junto con la edad, lo cual era de esperarse porque implícitamente incluyen al peso y al pliegue. Por su parte al ajustar la regresión con la edad y las variables representantes de cada grupo se obtuvieron como significativas a la primera componente del grupo de masa corporal y la primera componente del grupo de panículos adiposos los cuales representan implícitamente al peso y al pliegue subescapular, respectivamente. (Ver Tabla 2.8)

Tabla 2.8 Variables en los diferentes análisis y sus correspondientes variables estadísticamente significativas en el subgrupo de hombres

	Análisis 1	Análisis 2	Análisis 3
Variables	***edad***	***edad***	***edad***
	bicipii	***pcom_gpo***	***pcom_pli***
	tricipii	scom_gpo	***scom_pli***
	subesi		
	suprai		
	pesoi		***pcom_imc***
	imc		scom_imc
	biacroi		pcom_ose
	bicresi		scom_ose
	rodillai		
	codoi		
	tallai		
	sentadoi		
	cinturai		pcom_cad
	caderai		scom_cad
	icc		
	ict		

Variables seleccionadas por el método backward y forward en SPSS

Se presentaron estos nuevos análisis a los investigadores y debido a la dificultad en la interpretación práctica de las componentes principales se decidió manejar las variables antropométricas originales. Así, se ajustaron regresiones logísticas en el paquete SPSS bajo los métodos de selección de variables de forward y backward para cada sexo. Sin embargo solo se presenta ampliamente el caso de los hombres para ejemplificar la teoría de regresión logística. Los resultados correspondientes al subgrupo de mujeres se presentan resumidos al final de los de hombres y se comentan algunos aspectos en las conclusiones.

Bajo ambos métodos resultaron significativas las mismas variables: EDAD, PESOI y SUBESI. Cabe señalar que para los hombres se siguieron las estrategias

propuestas por Hosmer y Lemeshow (Ver Anexo 1) y resultaron significativas las mismas variables que bajo los métodos de selección automática.

Como una segunda etapa se incluyeron en las regresiones las variables relacionadas con la herencia. Cabe señalar que los investigadores consideran que la información de la variable de ANTHTA carece de precisión por la forma en que se recabó la misma. Por esto, los resultados serán interpretados, a reserva de este hecho.

2.4 Prueba del Cociente de Verosimilitudes

Para decidir qué modelo se interpretará resulta oportuno hacer una prueba de Cociente de Verosimilitudes (G^2) donde el modelo en la hipótesis nula será un caso particular del definido en la hipótesis alternativa. En este caso se definirá en la hipótesis nula los obtenidos bajo los métodos forward, bakward y de Hosmer-Lemeshow y en la alternativa el modelo que además incluye a la variable BICRESI ya que resultó significativa al 5% al entrar todas las variables antropométricas (ver Tabla 2.7).

El planteamiento de las hipótesis es como sigue:

H_0: logit $(p_i) = \beta_o + \beta_1 edad_i + \beta_2 subesi_i + \beta_3 pesoi_i$

vs

H_1: logit $(p_i) = \beta_o + \beta_1 edad_i + \beta_2 subesi_i + \beta_3 pesoi_i + \beta_4 bicresi_i$

O bien los modelos con las variables de anthta.

A partir de estos modelos se obtuvieron algunas estadísticas que se resumen en la Tabla 2.9. De ahí se observa que la contribución explicativa de la variable BICRESI no es significativa ya que la estadística del Cociente de

Verosimilitudes (G^2) tiene un p-value mayor a 0.05 por lo que no existe evidencia para decir que el modelo que además incluye a la variable BICRESI explique mejor la presencia de HTA. Así, se concluye que el modelo propuesto a interpretar es el definido en la hipótesis nula:

$$\ln\left(\frac{p_i}{1-p_i}\right) = -9.647 + 0.09edad + 0.053subesi + 0.033pesoi$$

ó

$$\ln\left(\frac{p_i}{1-p_i}\right) = -10.578 + 0.091edad + 0.056subesi + 0.038pesoi + 0.835anthta$$

Tabla 2.9 Prueba de cociente de verosimilitudes de los modelos propuestos para Hombres

Hip	Vars sig	-2 Log FV	G^2=-2(L_{Ho}-L_{H1})	$p > \chi^2_{(1)}$
H_0	EDAD, PESOI, SUBESI	776.31586	1.40446	0.2360
H_1	EDAD, PESOI, SUBESI, BICRESI	774.9114		

Tabla 2.9 Cotinuación

Modelo	Vars sig	$\hat{C}$	$p > \chi^2_{(4)}$
H_0	EDAD, PESOI, SUBESI	1.81	0.7704
H_1	EDAD, PESOI, SUBESI, BICRESI	2.16	0.9760

Cabe señalar que en la siguiente sección sólo se presenta la evaluación del modelo que no incluye la variable de ANTHTA por su imprecisión ya que la obtención de ésta, según los investigadores, fue a partir de lo que el paciente sabe y no de algún diagnóstico de HTA en los padres.

2.5 Análisis confirmatorios de los modelos propuestos

2.5.1 Estadística de Hosmer y Lemeshow

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

Para evaluar la estadística de Hosmer–Lemeshow (H-L) primero se obtuvo la estadística con 10 grupos a partir de las probabilidades estimadas pero en tres de los 10 grupos se tuvieron frecuencias menores a cinco por lo que se obtuvo la misma estadística pero a partir de 6 grupos (mínimo sugerido), de los cuales solo el primero tuvo menos de 5 observaciones, como lo muestra la Tabla 2.10.

Tabla 2.10 Frecuencias observadas y esperadas en cada grupo para el subgrupo de hombres bajo el modelo2

_Group	_Prob	_Obs_1	_Exp_1	_Obs_0	_Exp_0	_Total
1	0.0297	3	2.8	164	164.2	167
2	0.0637	8	7.6	159	159.4	167
3	0.1308	17	15.9	150	151.1	167
4	0.2393	26	30.3	141	136.7	167
5	0.4045	59	53.7	108	113.3	167
6	0.885	90	92.7	76	73.3	166

number of observations	1001
numb er of groups	6
Hosmer-Leme show chi2(4)	1.81
Prob > chi2	0.7704

Resultados obtenidos por Stata

Así, la frecuencia observada en el cuarto grupo diagnosticados con HTA es de 26 individuos. Este valor se obtiene de la suma de las personas con Y=1 en este grupo de 167 individuos de igual manera la frecuencia estimada en este

grupo es de 30.3 que es la suma de las 167 probabilidades estimadas de este grupo. La frecuencia observada para el grupo de personas sin HTA es 167-26=141 y la frecuencia estimada para este grupo es 167-30.3=136.7

El valor de la estadística de H-L (ver ecuación 1.15) calculada a partir las frecuencias en la Tabla 2.10 es $\hat{C} = 1.81$ y el nivel de significancia correspondiente a la distribución chi-cuadrada con 4 grados de libertad es 0.7704 esto indica que el modelo describe bien a los datos. Si se comparan las frecuencias observadas y estimadas en las 12 celdas en la tabla son muy parecidas lo cual corrobora el resultado de la estadística.

2.5.2 Tablas de clasificación

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

Para obtener la variable dicotómica se necesita definir un punto de corte, ***c***, y comparar cada probabilidad estimada con dicho valor ***c***. Si la probabilidad estimada es mayor a ***c*** entonces la variable dicotómica toma el valor de 1; en otro caso es igual a 0. El valor más común es 0.5 .

La tabla cruzada para el Modelo 2 para el subgrupo de hombres se presenta en la Tabla 2.11, el porcentaje global de casos clasificados correctamente es de 82.6% =[(762+65)/1001]*100 pero solo el **32%** =[(65)/(138+65)]*100 de los casos con HTA fueron clasificados correctamente y el 95.5%= [(762)/(762+36)]*100 de los que no fueron diagnosticados con HTA fueron clasificados correctamente bajo el modelo. Cabe señalar que el grupo de HTA es el grupo más pequeño, es sólo una cuarta parte del de los Normales, por lo que la proporción clasificada correctamente de HTA es baja. Así en epidemiología usualmente se dice que el índice de sensibilidad del modelo es de 32% y el índice de sensitividad de 95%, si el punto de corte es 0.05. De esta tabla se puede

obtener otro indicador para evaluar el ajuste del modelo que es el Valor Pronóstico Positivo (VPP) que se calcula en base a los valores estimados, consiste en dividir el número de estimados correctamente como HTA entre el total de casos estimados como HTA, (Feinstein, 1985. p. 419 y 434). Así en este caso se dice que casi 3 veces es mejor el pronóstico de HTA mediante el modelo propuesto que sin él porque se tiene una prevalencia en la muestra de 203/1001=0.20 y un VPP de 0.64. Tanto el VPP como el Valor Pronóstico Negativo se refieren a la precisión en predecir resultados positivos y negativos, (Ver Tabla 2.12).

Tabla2.11 Tabla de clasificación según el modelo[a]

			Predicted		
			hta si y no		Percentage
	Observed		normal	hta	Correct
Step 1	hta si y no	normal	762	36	95.5
		hta	138	65	32.0
	Overall Percentage				82.6

a. El valor de corte es 0.5

Tabla 2.12 Índices usados en epidemiología a partir de la tabla cruzada

Pronóstico	HTA		Total	Valores Pronósticos
	si	no		
Positivo	65	36	101	= 65/101=**0.644**
Negativo	138	762	900	=762/900=**0.847**
Total	203	798	**1001**	

Sensibilidad =65/203=**0.32**

Especificidad =762/798=**0.955**

2.5.3 Pruebas sobre los coeficientes

Además de probar la significancia conjunta de las variables es importante revisar la significancia de cada variable. De la tabla 2.13 se puede decir que para todos los efectos los intervalos de confianza para sus coeficientes asociados no contienen al cero por lo que se puede afirmar con un 95% de confianza que la

edad, el pliegue subescapular y el peso corporal tienen un efecto significativo en la probabilidad de presentar hipertensión arterial. Asimismo, la estadística de Wald para probar la hipótesis de que la variable considerada, no tiene un efecto significativo en la probabilidad de presentar HTA fueron grandes, mayores de 13 con un nivel de significancia prácticamente de cero por lo que se concluye que la edad, el pliegue subescapular y el peso se asocian con la presencia de HTA.

Tabla 2.13 Coeficientes estimados para el modelo 2.

	Coef estim	S.E.	Wald=(estim/se)^ 2	df	Sig.	IC 95% para coef estim inferior	superior
EDAD	0.09	0.008	124.664	1	0	0.074	0.106
SUBESI	0.053	0.014	13.713	1	0	0.026	0.080
PESOI	0.033	0.008	16.131	1	0	0.017	0.049
Constant	-9.647	0.818	138.926	1	0	-11.250	-8.044

Salida en SPSS para el modelo que incluye EDAD, SUBESI, PESOI

Otra manera de revisar el ajuste del modelo es graficando el logito $\ln\left[\frac{\hat{P}(Y_i=1)}{1-\hat{P}(Y_i=1)}\right]$ o predictor lineal estimado $\hat{\beta}_o + \hat{\beta}_1 edad_i + \hat{\beta}_2 subesi_i + \hat{\beta}_4 pesoi_i$ contra las variables significativas. Las figuras 1, 2 y 3 muestran una tendencia lineal por lo que no es necesario transformar estas variables.

Figura 1.
Relación entre la edad y el predictor lineal estimado para el modelo 2 del subgrupo de hombres

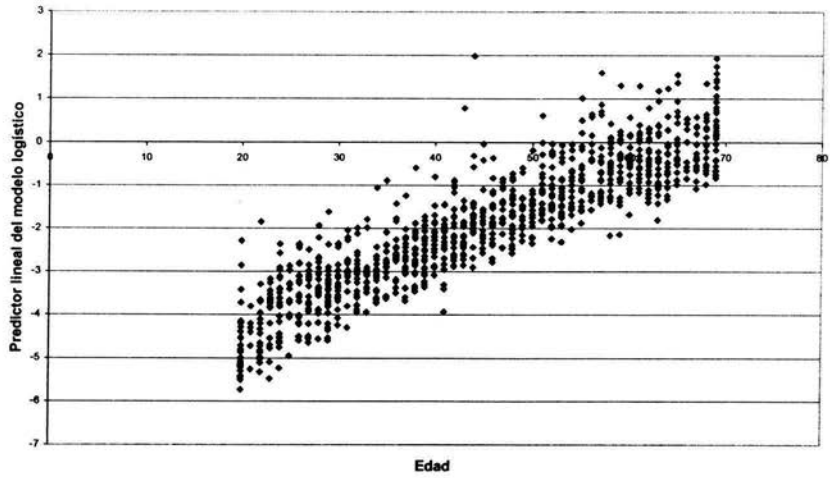


Figura 2.
Relación entre el pliegue subescapular y el predictor lineal estimado para el modelo 2 del subgrupo de hombres

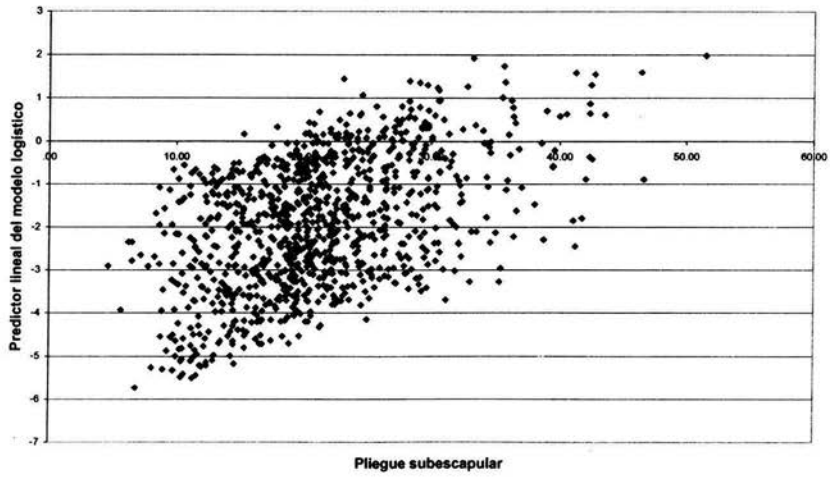
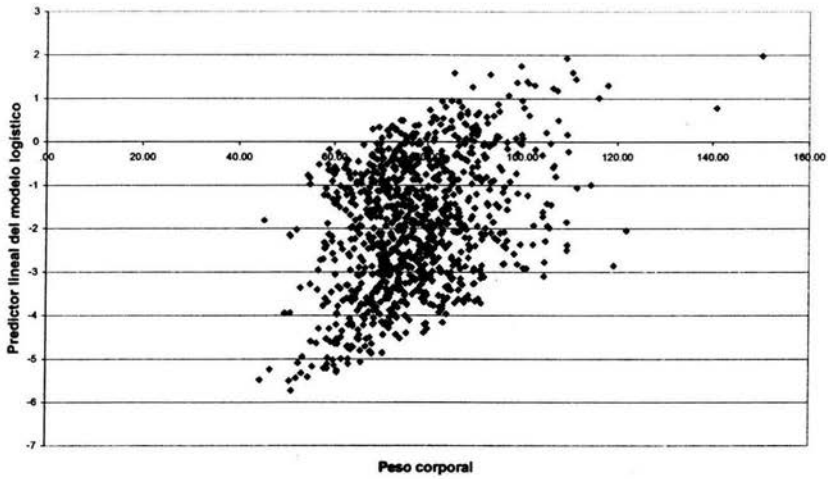


Figura 3.
Relación entre el peso corporal y el predictor lineal estimado para el modelo2 del subgrupo de hombres



2.5.4 Influencia de las observaciones

Para realizar el análisis de los residuos se grafican los residuos de Pearson o de la Devianza contra las variables explicativas pero al tener variables continuas resulta poco útil graficarlas para examinar el ajuste del modelo debido a que existen pocas observaciones con el mismo valor en cada variable con lo que los residuos son positivos cuando existe presencia de HTA y negativos cuando no existe HTA, (Agresti, 2002. p. 221). Este comportamiento puede verse en las siguientes gráficas donde se forman dos grupos de nubes en las cuales los puntos mayores a cero son los diagnosticados como HTA y los negativos son los Normales. En el caso de poder agrupar los datos, las gráficas de los residuos pueden ser más informativas sobre el ajuste ya que permite identificar patrones de covariables o niveles que tienen una fuerte influencia en el ajuste del modelo. De las gráficas también se puede decir que los residuos estandarizados (ver ecuación

1.13) se encuentran entre –3 y 3 por lo que se puede considerar un buen ajuste.

Figura4. Residuos estandarizados para el modelo ajustado

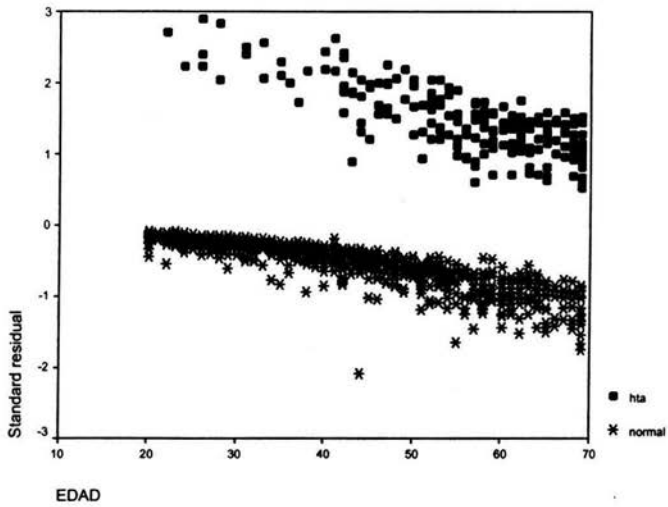


Figura5. Residuos estandarizados para el modelo ajustado

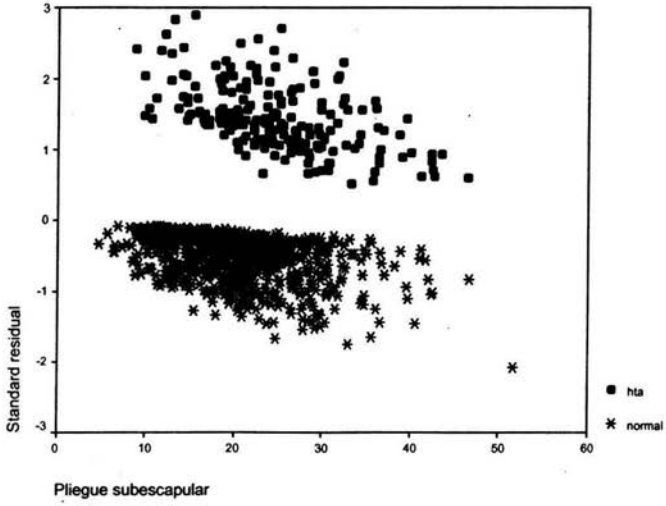
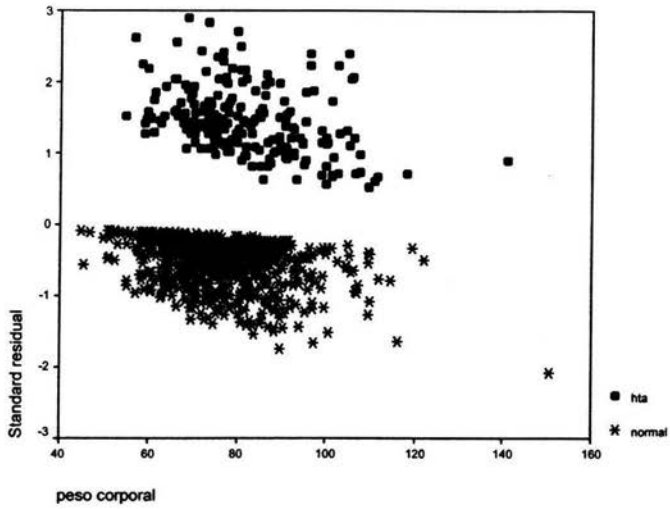
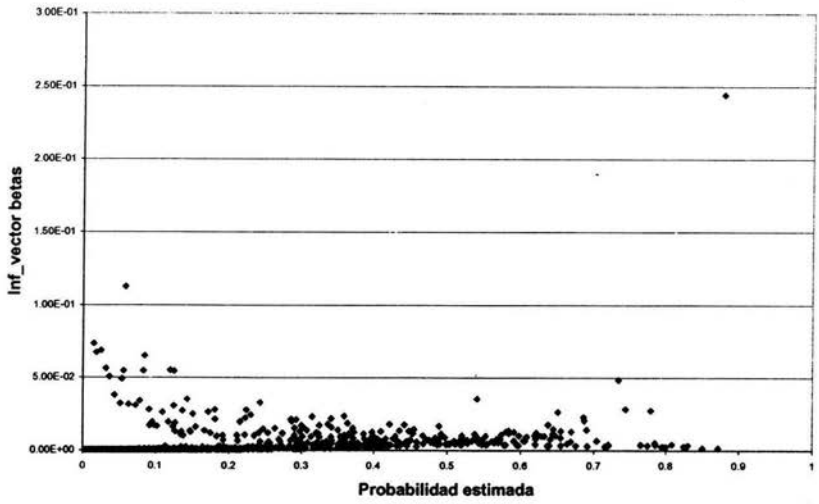


Figura 6. Residuos estandarizados para el modelo ajustado



En la Figura 7 se muestran las probabilidades estimadas contra el índice de influencia sobre el vector de coeficientes estimados (ver ecuación 1.18) y se puede ver que el punto más influyente es el que corresponde al individuo 3423. Aunque al eliminarlo y ajustar nuevamente el modelo no cambiaron mucho los estimadores. El individuo identificado como 3423 corresponde a un hombre de 44 años con medidas muy altas en comparación con las medias de las variables. Su pliegue subescapular mide 51.5 mm, la cintura 143.9 cm, la cadera 140 cm y pesa 150 kg por lo que es un individuo que se aleja de la mayoría. Se observa que la probabilidad estimada para él es alta y se debe principalmente a las elevadas medidas en el pliegue subescapular y peso.

Figura 7. Influencia de las observaciones en los coeficientes de regresión del modelo 2 para el subgrupo de hombres



2.6 Interpretación del modelo

Una vez que se concluye que el modelo propuesto describe bien a los datos, se procede a su interpretación. En este caso se interpretará el modelo que incluye a la variable ANTHTA para ver la interpretación de variables continuas y de una dicotómica. Cabe señalar que los coeficientes estimados se conservan del modelo evaluado anteriormente por lo que se procede con su interpretación.

En la Tabla 2.14 se muestra con más detalle los coeficientes estimados para el modelo 4 de la tabla 2.9 y en la Figura 8A se pueden ver las razones de momios para las variables significativas. Sin embargo, el efecto del antecedente de HTA que es el más fuerte no permite ver los efectos de las otras variables por lo que se presenta en la Figura 8B las variables significativas estadísticamente pero excluyendo a la variable de ANTHTA para apreciar los efectos de las variables restantes.

Figura 8A. Razones de momios estimadas de vars significativas

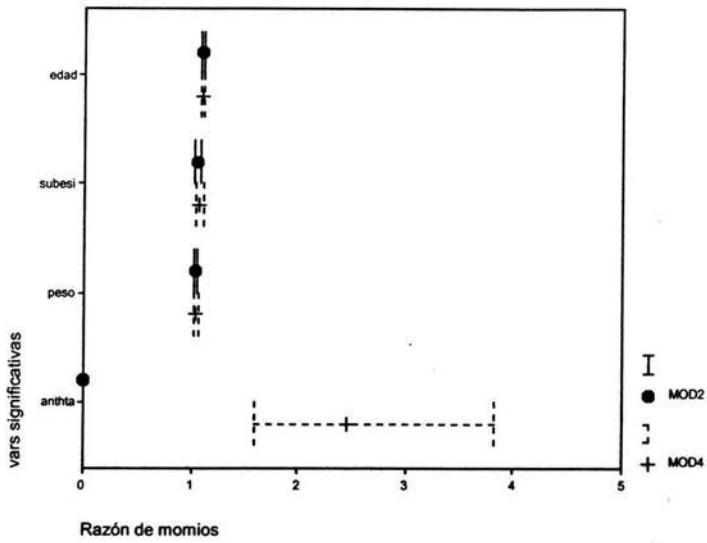
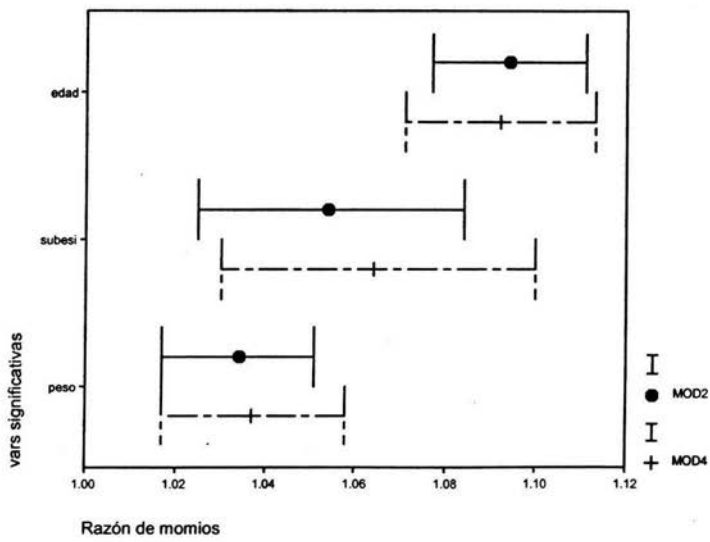


Figura 8B. Razones de momios estimadas sin la variable ANTHTA



Gráficamente se puede ver que, después de ANTHTA, la edad es la variable que influye más en la presencia de HTA, mientras que el pliegue subescapular y el peso tienen un efecto muy similar y menor, aunque el efecto del peso es más preciso que el del pliegue subescapular.

Tabla 2.14 Coeficientes estimados para el modelo 4.

	Coef estim	S.E.	Wald=(estim/se)^2	df	Sig.	Exp(est)	IC 95% para Exp(est) inferior	superior
EDAD	0.091	0.01	88.818	1	0	1.096	1.075	1.117
SUBESI	0.056	0.016	12.229	1	0	1.057	1.025	1.091
PESOI	0.038	0.01	15.402	1	0	1.039	1.019	1.059
ANTHTA(1)	0.835	0.218	14.633	1	0	2.304	1.502	3.534
Constant	-10.578	0.977	117.258	1	0	0		

Salida en SPSS para el modelo que incluye EDAD, SUBESI, PESOI Y ANTHTA.

2.6.1 Razones de momios y probabilidades estimadas

Como se explicó en la sección seis del capítulo uno, la interpretación de las razones de momios son las más usadas por lo que se inicia la interpretación de ellas.

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

Para la variable ANTHTA la categoría de referencia es no tener antecedentes por parte de los padres de HTA por lo que es respecto a esta categoría con la que se hace la comparación en la interpretación de la razón de momios. Explícitamente la razón de momios asociada a esta variable es:

$$RM = \frac{Momio(\text{HTA}|\text{ANTHTA}=1,\text{PESOI}=x,\ \text{SUBESI}=x,\ \text{EDAD}=x)}{Momio(\text{HTA}|\text{ANTHTA}=0,\text{PESOI}=x,\ \text{SUBESI}=x,\ \text{EDAD}=x)} = e^{\beta_4} = 2.304$$

La interpretación correcta es que los momios de presentar hipertensión

arterial (HTA) cuando alguno o ambos padres la padecieron, manteniendo constantes las otras variables, es 2.304 veces más que los momios de presentar HTA cuando los padres no la padecieron. De manera equivalente se dice que la probabilidad de presentar HTA en lugar de no presentarla cuando alguno de los padres padeció hipertensón arterial es 2.304 veces la probabilidad de que un hombre sin antecedentes de HTA en padres presente HTA en lugar de no presentarla. En muchas ocasiones esta interpretación puede darse de una manera más directa, pero no del todo correcta, diciendo que los hombres donde al menos uno de los padres padecieron HTA son 2.304 veces más probables de presentar HTA que aquellos hombres cuyos padres no padecieron HTA. Esta interpretación es válida cuando la razón de momios se aproxima a la medida de asociación llamada riesgo relativo, pero esto sucede sólo cuando la variable respuesta es "rara", es decir, que la probabilidad de presentar HTA sea pequeña. Esto no sucede en la prevalencia de HTA en la muestra porque alrededor del 20% de los individuos presentan HTA (Ver Tabla 2.2).

El intervalo de confianza estimado sugiere que los momios de presentar HTA para individuos con padres que padecieron HTA puede ser tan pequeño como 1.5 o tan grande como 3.5 veces más que los momios de presentar HTA cuando no se tiene algún antecedente de HTA en sus padres. En otras palabras se dice que en comparación con las personas cuyos padres no padecieron HTA, las personas que alguno o ambos padres padecieron HTA tienen una mayor probabilidad de presentar HTA.

Figura 9. Probabilidad estimada vs pliegue subescapular

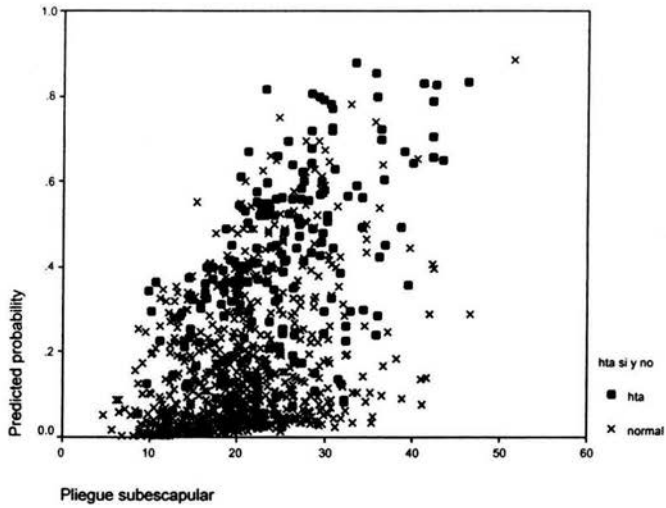


Figura 10A. Probabilidad estimada vs peso

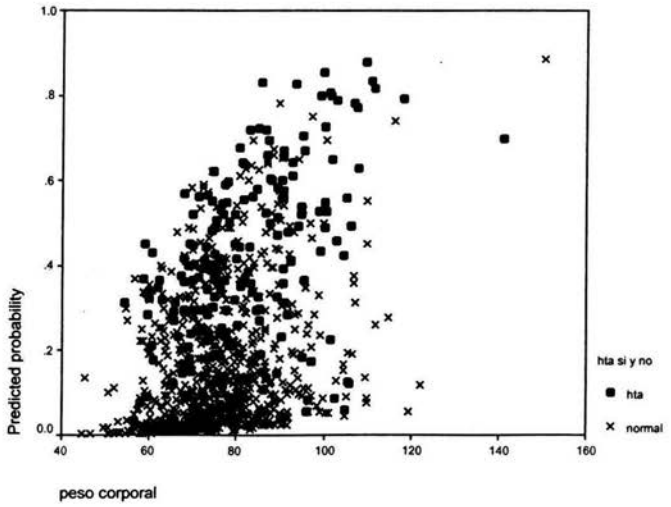


Figura 10B. Probabilidad estimada vs peso

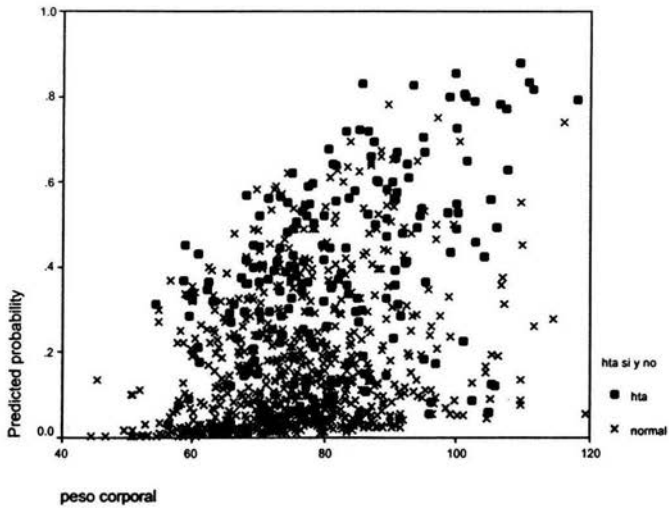
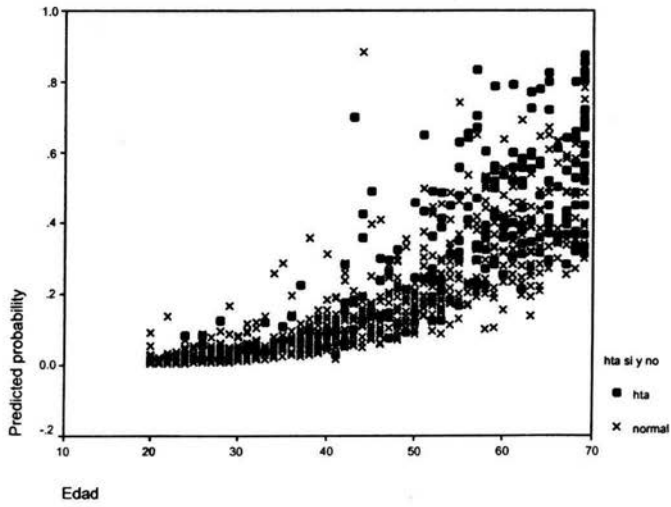


Figura 11. Probabilidad estimada vs edad



Las figuras 9, 10A o 10B y 11 muestran la probabilidad estimada de presentar HTA para cada individuo según su pliegue subescapular, peso y edad. A grandes rasgos se observa que la probabilidad de presentar HTA aumenta conforme aumenta cada variable, también se puede ver que las personas con diagnóstico observado de HTA son las que presentan mayor probabilidad estimada de presentar hipertensión arterial. Estos resultados se confirman en los resultados de la Tabla 2.14. Estas tres variables tienen razones de momios mayores a uno. Así, las variables PESOI, SUBESI y EDAD están asociadas de manera positiva con la presencia de HTA, es decir, conforme aumentan las unidades de estas variables mayor es la probabilidad de presentar HTA. En el caso de la variable PESOI y considerando a las demás constantes, una persona con una unidad más (1kg) es más probable de presentar HTA que una persona con una unidad menos (1kg).

$$RM = \frac{Momio(\text{HTA}|\text{PESOI}=x+1,\ \text{SUBESI}=x,\ \text{EDAD}=x, \text{ANTHTA}=x)}{Momio(\text{HTA}|\text{PESOI}=x,\ \text{SUBESI}=x,\ \text{EDAD}=x, \text{ANTHTA}=x)} = e^{\beta_3} = 1.039$$

De manera precisa se dice que los momios de presentar HTA cuando se tiene x+1 unidades (kg) es 1.04 veces los momios de presentar HTA cuando se tienen x unidades, considerando constantes el resto de las variables involucradas. Los momios estimados de presentar HTA se multiplica por 1.04 por cada incremento de una unidad en el PESOI, es decir, por cada unidad que aumenta el PESOI los momios aumentan 4%, manteniendo constantes las demás variables. De manera análoga se interpreta para la variable SUBESI y EDAD.

2.6.2 Logitos

De los parámetros estimados, coeficientes de regresión, sus signos positivos nos dicen que las probabilidades de presentar HTA crecen conforme los valores en cada variable crecen, para las variables continuas y para la variable dicotómica al ser mayor a cero que es el coeficiente implícitamente asociado a la

categoría de no antecedentes de HTA en al menos uno de los padres, indica que es más probable presentar HTA en caso de tener algún antecedente de HTA en sus padres que si no lo hay. De este modo un incremento de una unidad (1 año) en la variable EDAD, el logito de presentar HTA aumenta por 0.091, de igual manera se interpreta para las variables PESOI y SUBESI.

Respecto a la variable ANTHTA se dice que el logaritmo natural de los momios de presentar HTA es 0.835 mayor para aquellos con antecedentes de HTA en alguno de sus padres que para los que no tienen antecedentes.

2.7 Análisis para el subgrupo de mujeres

Se realizó una primera regresión exploratoria con todas las variables antropométricas e incluso los índices y la variable referente a la variable menopausia por lo observado en las gráficas del análisis exploratorio. De ahí se observó que sólo la edad y el diámetro bicrestal resultaron significativas al 5%. Ver Tabla 2.15. Se ajustó una nueva regresión sin considerar a los índices porque quizás estos índices resultan equivalentes a sus variables independientes y no permiten ver los efectos de éstas. Como en el caso de los hombres, considerar únicamente las variables antropométricas permite descubrir dos factores más: el peso y la talla. Por último se realizaron las regresiones logísticas bajo el método forward y backward resultando estadísticamente significativas en ambos métodos: EDAD, PESOI y BICRESI.

Se evaluó la estadística de Hosmer y Lemeshow para este modelo pero a pesar de que se obtuvo a partir de 6 grupos, dos de ellos contienen sólo una observación por lo que la distribución asintótica a una χ^2 no se tiene y no es posible usar esta estadística para probar la bondad de ajuste del modelo. Ver Tabla 2.16.

Tabla 2.15 Regresión con todas las variables antropométricas, para mujeres.

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	**EDAD**	**.096**	**.014**	**48.400**	**1**	**.000**	**1.100**	**1.071**	**1.130**
	BICIPII	.019	.019	1.065	1	.302	1.020	.983	1.058
	TRICIPII	.012	.021	.344	1	.557	1.012	.972	1.054
	SUBESI	-.013	.014	.771	1	.380	.987	.960	1.016
	SUPRAI	.007	.015	.232	1	.630	1.007	.978	1.038
	PESOI	.074	.148	.248	1	.619	1.076	.806	1.438
	CINTURAI	.012	.301	.002	1	.967	1.012	.561	1.827
	CADERAI	.041	.095	.190	1	.663	1.042	.866	1.255
	TALLAI	-.122	.121	1.017	1	.313	.885	.698	1.122
	SENTADOI	.023	.032	.508	1	.476	1.023	.961	1.089
	BIACROI	-.003	.053	.003	1	.960	.997	.899	1.107
	BICRESI	**.103**	**.039**	**7.118**	**1**	**.008**	**1.109**	**1.028**	**1.196**
	RODILLAI	-.134	.127	1.114	1	.291	.874	.681	1.122
	CODOI	-.031	.222	.019	1	.891	.970	.627	1.500
	IMC	-.037	.349	.012	1	.915	.963	.486	1.909
	ICT	-.156	.453	.118	1	.731	.856	.352	2.079
	ICC	8.239	11.253	.536	1	.464	3784.425	.000	1.4E+13
	MENODX2			1.670	2	.434			
	MENODX2(1)	.399	.335	1.421	1	.233	1.490	.773	2.873
	MENODX2(2)	.379	.316	1.440	1	.230	1.461	.786	2.716
	Constant	1.091	20.801	.003	1	.958	2.978		

a. Variable(s) entered on step 1: EDAD, BICIPII, TRICIPII, SUBESI, SUPRAI, PESOI, CINTURAI, CADERAI TALLAI, SENTADOI, BIACROI, BICRESI, RODILLAI, CODOI, IMC, ICT, ICC, MENODX2.

Tabla 2.16 Frecuencias observadas y esperadas en cada grupo para el subgrupo de mujeres

_Group	_Prob	_Obs_1	_Exp_1	_Obs_0	_Exp_0	_Total
1	0.0197	1	2.2	192	190.8	193
2	0.0523	1	6.5	192	186.5	193
3	0.1228	15	16.7	178	176.3	193
4	0.2627	50	36.1	143	156.9	193
5	0.4421	65	67.9	128	125.1	193
6	0.8623	109	111.5	83	80.5	192

number of observations	1157
numb er of groups	6
Hosmer-Leme show chi2(4)	12.68
Prob > chi2	0.0129

Resultados obtenidos por STATA

En este caso Agresti (2002, p.199) sugiere comparar el modelo ajustado con uno más complicado. Resulta oportuno probar el modelo que incluye a la variable TALLAI además de las obtenidas bajo los métodos forward y backward: EDAD, PESOI y BICRESI.

La prueba de hipótesis es la siguiente:

$$H_0: \text{ logit } (p_i) = \beta_o + \beta_1 edad_i + \beta_2 pesoi_i + \beta_3 bicresi$$

vs

$$H_1: \text{ logit } (p_i) = \beta_o + \beta_1 edad_i + \beta_2 pesoi_i + \beta_3 bicresi_i + \beta_4 Tallai$$

Tabla2.17 Prueba de cociente de verosimilitudes de los modelos propuestos para mujeres

Hip	Vars sig	-2 Log FV	G^2=-2(L_{Ho}-L_{H1})	$p > \chi^2_{(1)}$
H_o	EDAD, PESOI, BICRESI	877.147	3.218	0.0728
H_1	EDAD, PESOI, BICRESI, TALLAI	873.929		

La estadística del cociente de verosimilitudes igual a 3.218 tiene una significancia de 0.07 mayor a 0.05 lo que lleva a no rechazar la hipótesis nula por lo que el modelo que incluye la EDAD, PESOI y BICRESI se ajusta a los datos.

Por lo tanto el modelo propuesto para las mujeres es:

$$\text{logit } (p_i) = \beta_o + \beta_1 edad_i + \beta_2 pesoi_i + \beta_3 bicresi$$

y sus coeficientes estimados son:

Tabla 2.18 Coeficientes estimados para el modelo propuesto para mujeres

								95.0% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	EDAD	.107	.008	163.144	1	.000	1.113	1.095	1.131
	PESOI	.029	.008	11.650	1	.001	1.029	1.012	1.046
	BICRESI	.085	.034	6.457	1	.011	1.089	1.020	1.164
	Constant	-11.211	.989	128.374	1	.000	.000		

a. Variable(s) entered on step 1: EDAD, PESOI, BICRESI.

Las gráficas de las variables independientes contra el predictor lineal estimado $\hat{\beta}_o + \hat{\beta}_1 edad_i + \hat{\beta}_2 pesoi_i + \hat{\beta}_3 bicresi$ muestran una tendencia lineal por lo que no es necesario transformar las variables .

Figura12. Relación entre el predictor lineal y la edad

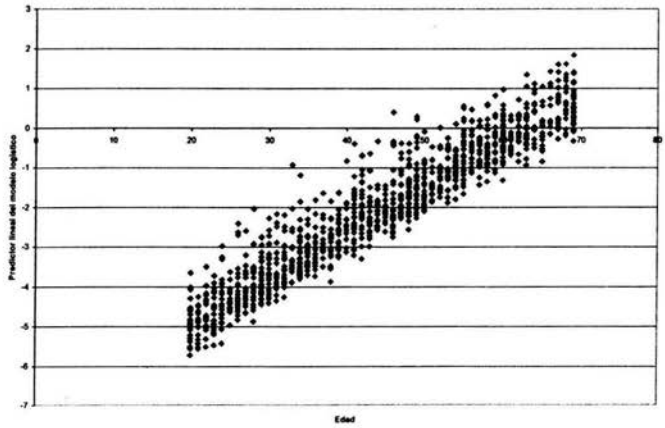


Figura13. Relación entre el predictor lineal y el diámetro bicrestal

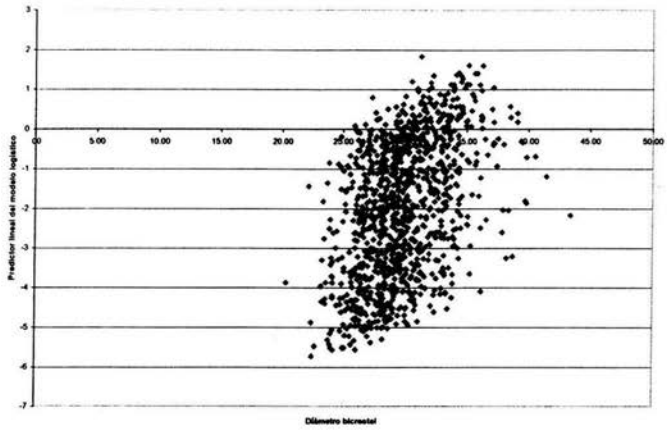
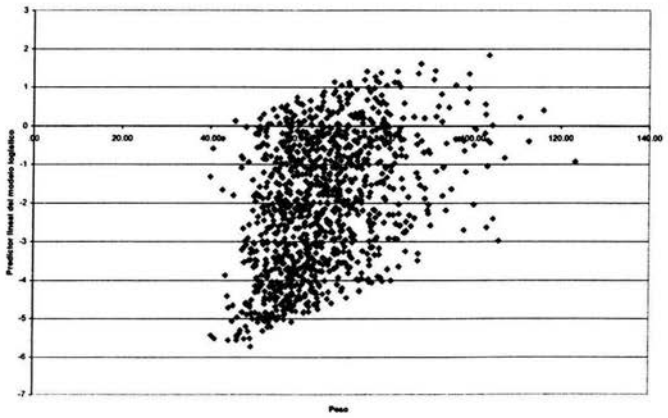


Figura14. Relación entre el predictor lineal y el peso



De la Tabla 2.18 se tiene que las variables EDAD, BICRESI y PESOI influyen en la presencia de HTA: De modo que a mayor edad, diámetro bicrestal o peso se tiene una mayor probabilidad de presentar HTA. En el Anexo 4 se pueden ver las fotografías de la toma de medición del pliegue subescapular y el diámetro bicrestal.

Conclusiones

El conocimiento de una técnica de regresión aplicada a variables respuesta de tipo cualitativo como la Logística es de mucha utilidad ya que en la práctica el uso de variables dicotómicas es muy amplio y no se limita al área médica, por ejemplo se aplica en la política, en el área social, industrial o financiera.

Se trató de encontrar algunas medidas antropométricas que estén relacionadas con la presencia de HTA en hombres y mujeres. Se encontró que para los hombres la presencia de HTA está asociada con el peso y el pliegue subescapular que son medidas de grasa central lo cual corrobora la hipótesis de que la distribución de la grasa en la región central es un factor de riesgo. Para las mujeres sobresalen el diámetro bicrestal y el peso. Es interesante la significancia del diámetro bicrestal por ser una medida más precisa que los pliegues, al tratarse de un hueso. En ambos casos la edad fue el factor que más influye.

Los modelos propuestos no pueden ser predictivos de la presencia de HTA ya que no se analizaron aspectos que se sabe son influyentes como los hábitos de ejercicio, de tabaquismo, de consumo de alcohol, de alimentación, por mencionar algunos.

Aunque se consideraron los antecedentes de enfermedades como la diabetes, obesidad e hipertensión arterial, la precisión de esta información no permite hacer alguna inferencia confiable sobre ellos. Cabe señalar que en el caso de los hombres se presentó su interpretación para ejemplificar el uso de variables explicativas dicotómicas.

El uso de variables continuas no permite el uso de estadísticas como la Chi Cuadrada o la Devianza para la evaluación del ajuste del modelo. En un futuro análisis podrían formarse categorías para las variables continuas y ver su efecto en la presencia de HTA.

Cabe señalar que la obtención de las variables antropométricas es relativamente sencilla y barata pero su variabilidad es grande porque son mediciones muy sensibles a factores como la hora y la forma en que se obtienen por lo que se deben tomar con reserva estos resultados.

La prevalencia de HTA en la ciudad de México para 1993 según la Encuesta Nacional de Enfermedades Crónicas es de 28.5% y 25.1% en hombres y mujeres respectivamente, mientras que la muestra estima un 20% para ambos sexos. Sin embargo no son comparables estas prevalencias porque corresponden a dos poblaciones diferentes, las dos primeras a una abierta y la otra a una clínica del ISSSTE al sur de la ciudad de México.

El poder discriminatorio de la regresión logística resulta muy parecido al obtenido bajo un análisis de discriminante cuando se tienen variables explicativas cuantitativas. Bajo este análisis para el subgrupo de los hombres, el 31.5% de los diagnosticados con hipertensión fueron clasificados correctamente y para las mujeres el 33.6%. Mientras que la regresión logística logra clasificar correctamente 32% y 35%, respectivamente.

De acuerdo con el análisis estadístico se puede concluir que a partir de la información de la encuesta, los modelos propuestos no logran predecir la presencia de HTA. Aunque auxilian a identificar algunos factores de riesgo asociados a la presencia de HTA.

Anexo 1. Construcción del modelo logístico bajo Hosmer y Lemeshow

A continuación se presenta la aplicación de la estrategia propuesta por Hosmer y Lemeshow para el subgrupo de hombres.

Primero se analizó cada variable individualmente mediante modelos de regresión logística univariada. En la Tabla A1 se presenta para cada variable la información resumida.

Tabla A1. Modelos univariados para explicar HTA.

Variable	Coef Estim	Error Est	LH1	G^2=-2(LHo-LH1)	GL	p-val	Z=(CoefEst/EE)^2	p-val
Edad	0.084	0.007	-417.534	174.453	1	0.000	129.368	0.000
bicipi	0.070	0.016	-495.460	18.601	1	0.000	19.198	0.000
tricipi	0.053	0.014	-497.532	14.456	1	0.000	14.919	0.000
subesi	0.079	0.011	-478.696	52.129	1	0.000	49.691	0.000
suprai	0.045	0.010	-493.813	21.894	1	0.000	21.941	0.000
pesoi	0.036	0.006	-488.042	33.435	1	0.000	32.434	0.000
cinturai	0.068	0.008	-464.338	80.843	1	0.000	71.077	0.000
caderai	0.057	0.010	-488.480	32.560	1	0.000	31.366	0.000
tallai	-0.030	0.012	-501.719	6.081	1	0.014	5.991	0.014
sentadoi	-0.040	0.017	-501.907	5.706	1	0.017	5.737	0.017
biacroi	***0.021***	***0.029***	***-504.509***	***0.501***	***1***	***0.479***	***0.500***	***0.479***
bicresi	0.096	0.024	-496.446	16.628	1	0.000	16.601	0.000
rodillai	0.260	0.102	-501.502	6.516	1	0.011	6.500	0.011
codoi	0.535	0.151	-498.487	12.545	1	0.000	12.483	0.000

L_{Ho}=-504.76006

De acuerdo con la Prueba de cociente de Verosimilitudes o la estadística de Wald, casi todas las variables explicativas tienen un nivel de significancia menor a 0.25, excepto BIACROI por lo que existe evidencia de que individualmente todas las demás tiene un efecto significativo en la probabilidad de presentar HTA.

La prueba del cociente de verosimilitudes está probando que el coeficiente de la variable en cuestión sea cero, es decir, se realiza la siguiente prueba de hipótesis:

H_0: logit $(p_i) = \beta_o$

vs

H_1: logit $(p_i) = \beta_o + \beta_j X_{ji}$ j=1, 2, ⋯,14 ,

y la estadística para probar la hipótesis se calcula: $G^2 = -2\left(L_{H_0} - L_{H_1}\right)$ y se compara contra una Chi cuadrada con un grado de libertad que es la diferencia de parámetros de los modelos de la prueba.

Después se ajustó el modelo múltiple que incluye a todas las variables explicativas significativas simultáneamente, ver Tabla A2.

Tabla A2. Modelo múltiple con las variables significativas a un nivel de 0.25 en la Tabla A1.

Variable	Coef.	Std. Err.	z=coef/se	P>|z|	[95% Conf.	Interval]
edad	***0.093***	***0.009***	***9.900***	***0.000***	***0.075***	***0.111***
bicipii	-0.045	0.030	-1.490	0.135	-0.104	0.014
tricipii	0.024	0.023	1.060	0.290	-0.021	0.070
subesi	***0.054***	***0.017***	***3.130***	***0.002***	***0.020***	***0.087***
suprai	0.000	0.015	0.020	0.983	-0.030	0.031
pesoi	**0.049**	**0.021**	**2.310**	**0.021**	**0.007**	**0.090**
cinturai	-0.006	0.022	-0.260	0.794	-0.048	0.037
caderai	-0.015	0.025	-0.590	0.556	-0.063	0.034
tallai	-0.036	0.021	-1.720	0.085	-0.078	0.005
sentadoi	0.036	0.026	1.380	0.166	-0.015	0.088
bicresi	**0.077**	**0.037**	**2.050**	**0.040**	**0.004**	**0.150**
rodillai	-0.137	0.163	-0.840	0.402	-0.457	0.183
codoi	0.071	0.219	0.320	0.746	-0.359	0.501
_cons	***-7.478***	***3.734***	***-2.000***	***0.045***	***-14.797***	***-0.159***

Resultados obtenidos en STATA

Se examinan los niveles de significancia de la estadística de Wald para cada variable explicativa y se compara el valor del coeficiente asociado a cada variable explicativa con el coeficiente estimado en el modelo univariado de modo

que no deben variar mucho en su valor y tampoco el signo de lo contrario se descarta la variable explicativa.

Los resultados en la Tabla A2 al compararlos con los de la Tabla A1 indican que hay una débil asociación para la mayoría de las variables al controlar las otras variables. Esto se puede ver en sus niveles de significancia mayores a 0.05. De este modelo sólo las variables EDAD, SUBESI, PESOI y BICRESI deben ser incluidas en el modelo.

Una vez que se tiene el primer modelo múltiple se revisa si la variable que no fue significativa en los modelos univariados (BIACROI) lo es ahora en presencia de otras variables. (Ver Tabla A3). Las variables BICRESI y BIACROI resultan no significativas por lo que se descarta la posibilidad de incluirlas en el modelo.

Tabla A3.
Modelo múltiple incluyendo las variables con p<0.25 en la Tabla A1.

Variable	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
edad	0.089	0.008	10.780	0.000	0.073	0.105
subesi	0.057	0.015	3.840	0.000	0.028	0.085
pesoi	0.028	0.010	2.700	0.007	0.008	0.049
bicresi	**0.039**	**0.033**	**1.180**	**0.237**	**-0.026**	**0.105**
biacroi	**-0.004**	**0.040**	**-0.100**	**0.919**	**-0.083**	**0.075**
_cons	-10.331	1.601	-6.450	0.000	-13.469	-7.194

Resultados obtenidos en STATA

Ahora corresponde explorar la escala de las variables significativas con el logito, para ello se obtiene una gráfica de dispersión entre cada variable y el logito estimado. (Ver figuras 1, 2 y 3). Las gráficas muestran un incremento lineal por lo que sugieren tratar a cada variable lineal al logito. Por lo tanto, según las estrategias propuestas por Hosmer y Lemeshow para construir el modelo de regresión logística, las variables EDAD, SUBESI y PESOI pueden explicar a la hipertensión, a continuación se debe revisar qué tan bien la describen.

Anexo 2. Comandos en SPSS y STATA para obtener los resultados

En SPSS

```
**Abre base original.

GET
  FILE='C:\dora\tesis\hta.sav'.

***Para conjuntar vars sobre herencia.
**Para herencia de obesidad.
*fre antobesi.
fre antob9  .
recode antob9  (4 5 6  7 8=1) (0 =0) (1 2 3=9) (else=copy) into antobesi.
mis val antobesi (9).
formats antobesi(f1).
var lab antobesi Sabe si alguno de sus padres son obesos.
val lab antobesi
0 NO
1 SI.
*fre antobesi.
***De manera similar se crean las variables para antecedentes de diabetes e HTA.

**Análisis de CPs.
FACTOR
  /VARIABLES bicipii tricipii subesi suprai pesoi imc cinturai caderai icc ict tallai sentadoi
biacroi bicresi rodillai
  codoi /MISSING LISTWISE
 /ANALYSIS bicipii tricipii subesi suprai pesoi imc cinturai caderai icc ict tallai sentadoi
biacroi bicresi rodillai
  codoi
  /PRINT INITIAL EXTRACTION
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PC
  /ROTATION NOROTATE
  /SAVE REG(ALL)
  /METHOD=CORRELATION .
***De manera similar se obtienen los CPs para los diferntes grupos de variables
antropométricas.


****Regresiones exploratorias.

**solo vars antropométricas.
```

```
GET
  FILE='C:\dora\tesis\FINAL\HTADOS_ANTROP_CASOS VAL1004h.sav'.

***Con vars de antecedentes.
GET
  FILE='C:\dora\tesis\FINAL\HTADOS_ANTROP_ANTEC_CASOS VAL(2)hoes735.sav'.

**Regresión con las 16 variables antropométricas.
LOGISTIC REGRESSION VAR=htados
  /METHOD=ENTER edad bicipii tricipii subesi suprai pesoi cinturai caderai tallai sentadoi
biacroi bicresi rodillai
  codoi  imc ict icc
  /PRINT=GOODFIT CI(95)
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .

**Regresión con variables antropométricas sin índices.
LOGISTIC REGRESSION VAR=htados
  /METHOD=ENTER edad bicipii tricipii subesi suprai pesoi cinturai caderai tallai sentadoi
biacroi bicresi rodillai
  codoi
  /PRINT=GOODFIT CI(95)
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .

**Regresiones por métodos forward y bakward.
LOGISTIC REGRESSION VAR=htados
  /METHOD=FSTEP edad bicipii tricipii subesi suprai pesoi cinturai caderai tallai sentadoi
biacroi bicresi rodillai
  codoi
  /PRINT=GOODFIT CI(95)
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .

LOGISTIC REGRESSION VAR=htados
  /METHOD=BSTEP edad bicipii tricipii subesi suprai pesoi cinturai caderai tallai sentadoi
biacroi bicresi rodillai
  codoi
  /PRINT=GOODFIT CI(95)
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .

***FACTORES SELECCIONADOS.

LOGISTIC REGRESSION VAR=htados
  /METHOD=ENTER edad subesi pesoi
  /PRINT=GOODFIT CI(95)
  /SAVE PRED COOK LEVER DFBETA RESID LRESID SRESID ZRESID DEV
  /PRINT=GOODFIT CI(95)
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .

***REGRESIONES PARA MUJERES.
GET
  FILE='C:\dora\tesis\FINAL\HTADOS_ANTROP_CASOS VAL1156M.sav'.
**De manera análoga al subgrupo de los hombres.
```

En STATA.

***Para obtener las estadísticas de Hosmer

*. use "C:\dora\Datos cat\base_lista_hoes.dta", clear.

*. logistic htados edad subesi pesoi,coef.

*. lfit, group (10) table.

*. lfit, group (6) table.

***Para obtener probabilidades, predictores lineales, residuos, coeficientes de influencia.

predict probestim, p

predict predlin, xb

predict respears, residuals

predict resdev, deviance

predict respearest, rstandar

predict his, hat

predict vecbetas, dbeta

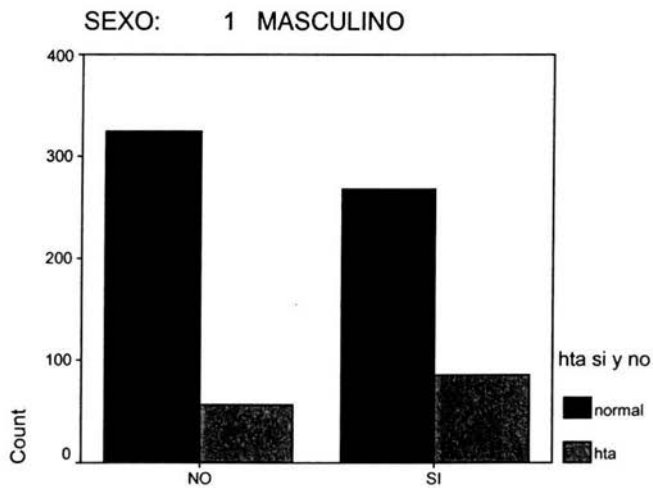
predict estchi2, dx2

Anexo 3. Algunas estadísticas descriptivas de las observaciones

Sabe si alguno de sus padres padece hta * hta si y no Crosstabulation

			hta si y no		
			normal	hta	Total
Sabe si alguno de sus padres padece hta	NO	Núm de casos	325	56	381
		%	54.8%	39.4%	51.8%
	SI	Núm de casos	268	86	354
		%	45.2%	60.6%	48.2%
Total		Núm de casos	593	142	735
		%	100.0%	100.0%	100.0%

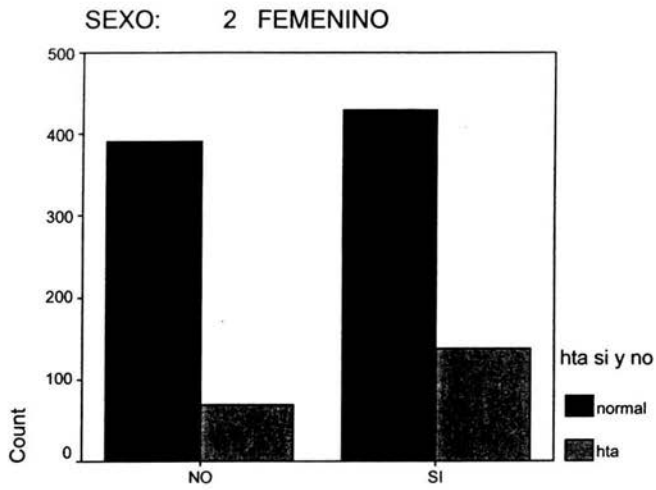
a. SEXO DE LOS INDIVIDUOS = MASCULINO



Sabe si alguno de sus padres padece hta * hta si y no Crosstabulation[a]

			hta si y no		
			normal	hta	Total
Sabe si alguno de sus padres padece hta	NO	Núm casos	391	69	460
		%	47.6%	33.5%	44.8%
	SI	Núm casos	430	137	567
		%	52.4%	66.5%	55.2%
Total		Núm casos	821	206	1027
		%	100.0%	100.0%	100.0%

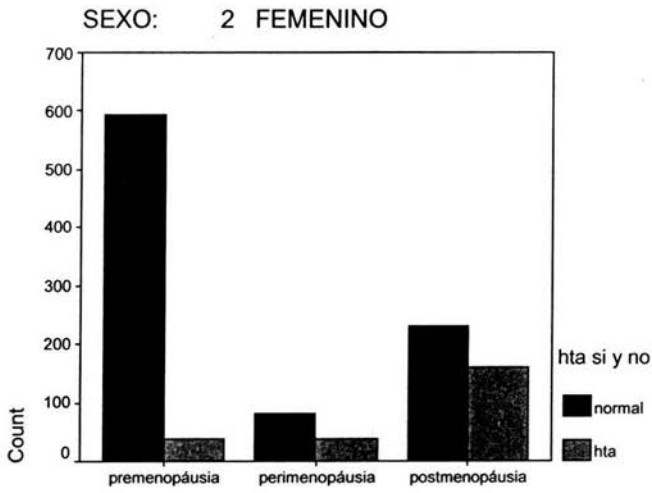
a. SEXO DE LOS INDIVIDUOS = FEMENINO



Condición de menopausia por HTA

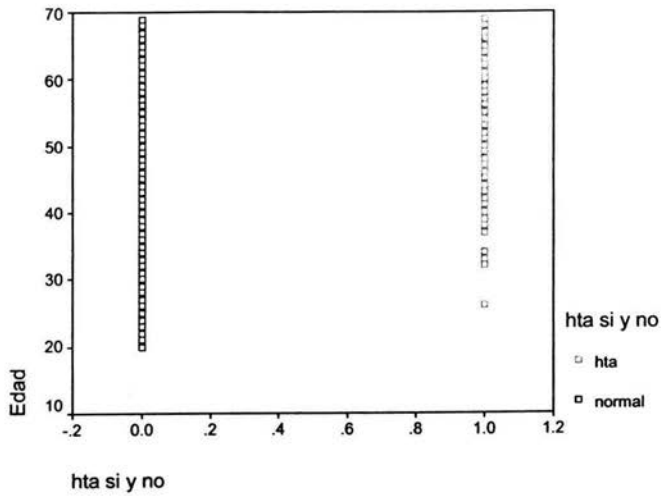
			hta si y no		
			normal	hta	Total
Condición de menopausia	premenopáusia	Núm casos	593	38	631
		%	65.5%	16.0%	55.2%
	perimenopáusia	Núm casos	82	39	121
		%	9.1%	16.4%	10.6%
	postmenopáusia	Núm casos	231	161	392
		%	25.5%	67.6%	34.3%
Total		Núm casos	906	238	1144
		%	100.0%	100.0%	100.0%

a. SEXO DE LOS INDIVIDUOS = FEMENINO

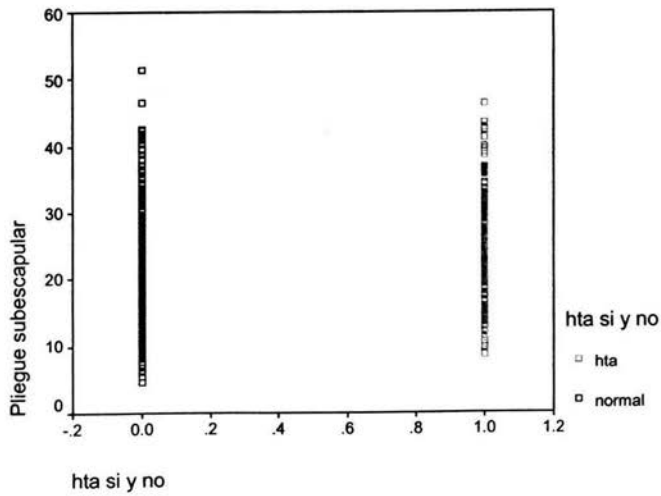


dx menopausia corregida por edad

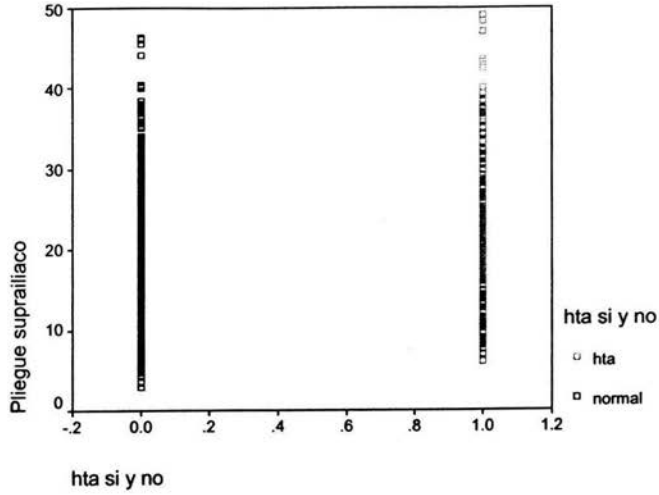
Edad según HTA



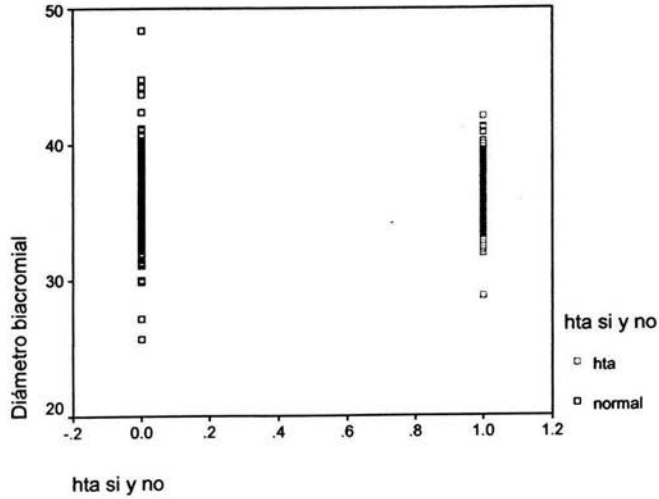
Pliegue subescapular según HTA



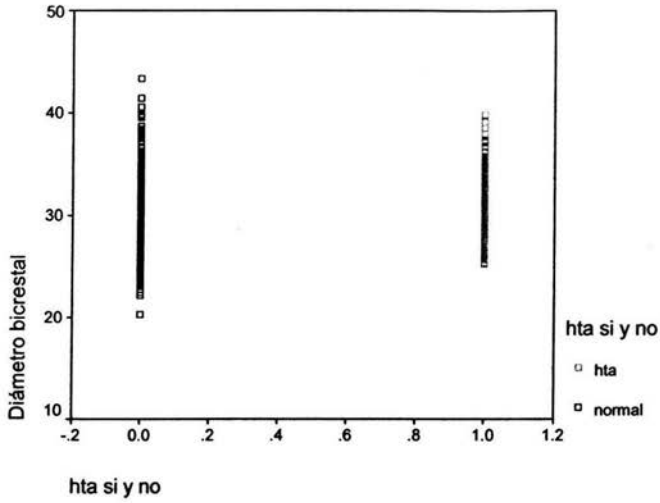
Pliegue suprailiaco según HTA



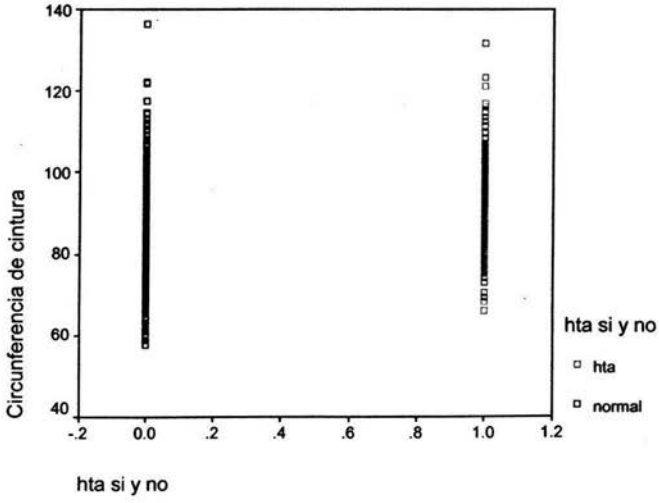
Diámetro biacromial según HTA



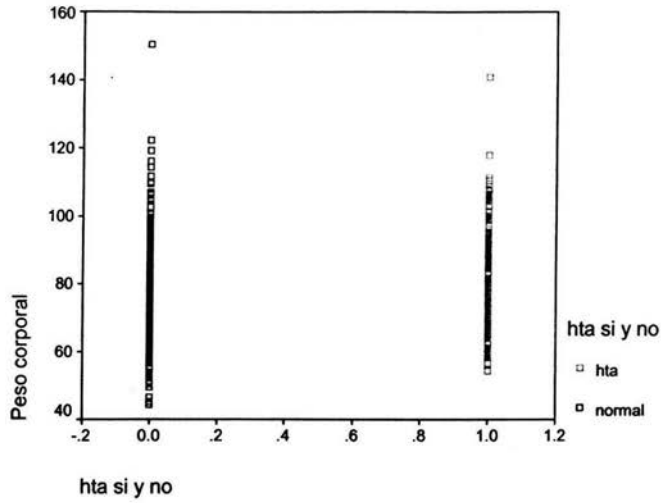
Diámetro bicrestal según HTA



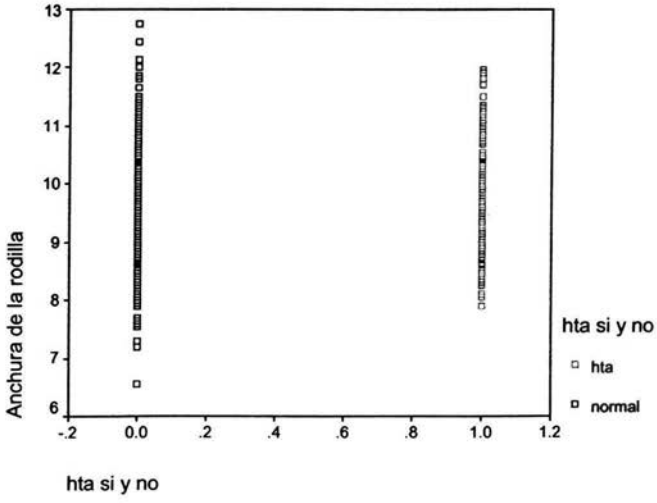
Circunferencia de cintura según HTA

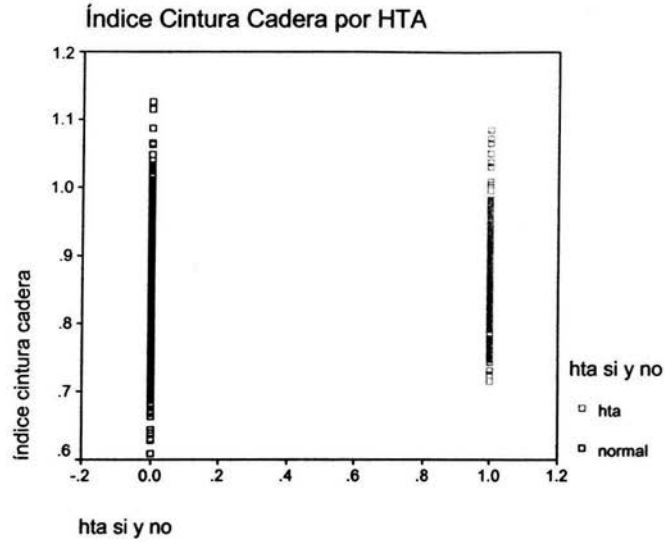


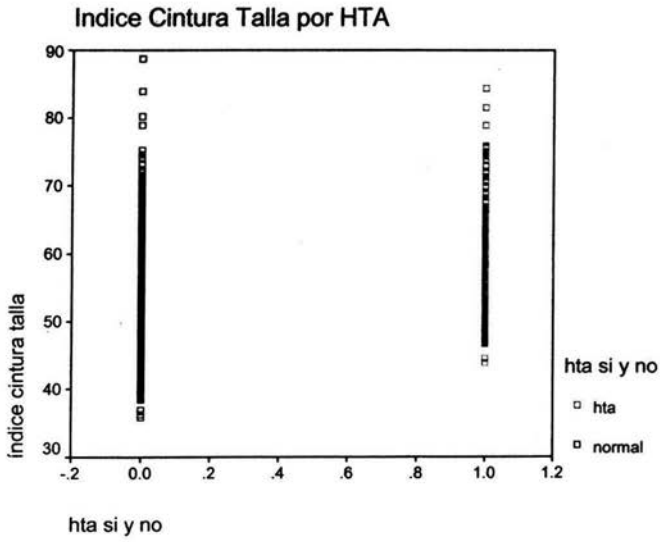
Peso según HTA

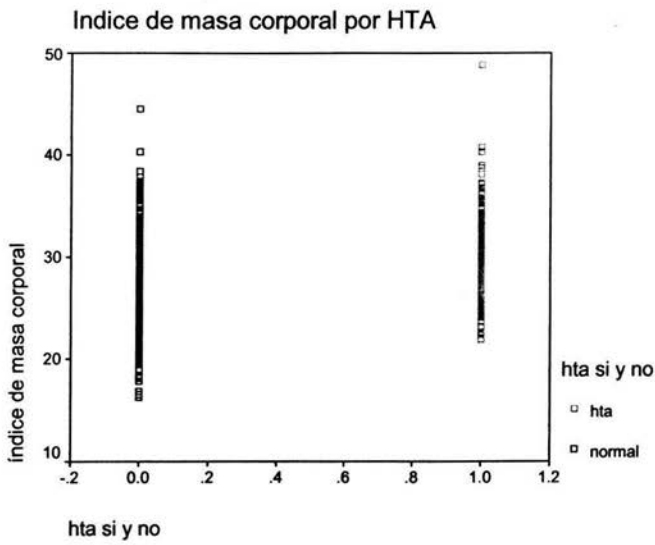


Rodilla por HTA




Índice Cintura Cadera por HTA
índice cintura cadera
1.2
1.1
1.0
.9
.8
.7
.6
-.2
0.0
.2
.4
.6
.8
1.0
1.2
hta si y no
hta si y no
hta
normal


Indice Cintura Talla por HTA
índice cintura talla
90
80
70
60
50
40
30
-.2
0.0
.2
.4
.6
.8
1.0
1.2
hta si y no
hta si y no
hta
normal


Indice de masa corporal por HTA
índice de masa corporal
50
40
30
20
10
-.2
0.0
.2
.4
.6
.8
1.0
1.2
hta si y no
hta si y no
hta
normal

Matriz de correlaciones de las variables antropométricas

	EDAD	TALLAI	BICIPII	TRICIPII	SUBESI
EDAD	1	-0.206	0.145	0.039	0.118
TALLAI	-0.206	1	-0.289	-0.446	-0.174
BICIPII	0.145	-0.289	1	0.724	0.507
TRICIPII	0.039	-0.446	0.724	1	0.513
SUBESI	0.118	-0.174	0.507	0.513	1
SUPRAI	0.127	-0.219	0.605	0.576	0.617
BIACROI	-0.074	0.685	-0.103	-0.341	-0.034
BICRESI	0.137	0.259	0.265	0.043	0.139
CINTURAI	0.315	0.264	0.3	0.081	0.423
CADERAI	0.149	0.058	0.557	0.437	0.485
SENTADOI	-0.28	0.72	-0.147	-0.195	-0.059
RODILLAI	0.073	0.404	0.185	-0.015	0.117
CODOI	0.143	0.573	-0.006	-0.25	0.012
PESOI	0.073	0.531	0.28	0.072	0.352
IMC	0.235	-0.099	0.546	0.415	0.544
ICC	0.324	0.315	-0.084	-0.282	0.161
ICT	0.411	-0.169	0.435	0.28	0.509

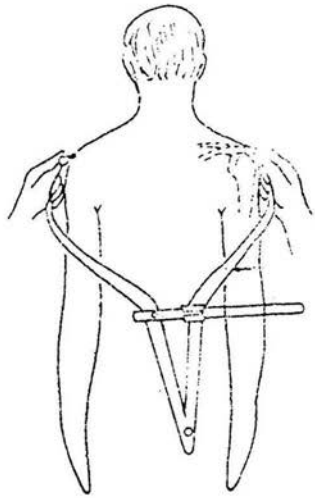
	SUPRAI	BIACROI	BICRESI	CINTURAI	CADERAI	SENTADOI
EDAD	0.127	-0.074	0.137	0.315	0.149	-0.28
TALLAI	-0.219	0.685	0.259	0.264	0.058	0.72
BICIPII	0.605	-0.103	0.265	0.3	0.557	-0.147
TRICIPII	0.576	-0.341	0.043	0.081	0.437	-0.195
SUBESI	0.617	-0.034	0.139	0.423	0.485	-0.059
SUPRAI	1	-0.005	0.315	0.363	0.507	-0.1
BIACROI	-0.005	1	0.438	0.472	0.243	0.464
BICRESI	0.315	0.438	1	0.537	0.543	0.105
CINTURAI	0.363	0.472	0.537	1	0.71	0.155
CADERAI	0.507	0.243	0.543	0.71	1	0.071
SENTADOI	-0.1	0.464	0.105	0.155	0.071	1
RODILLAI	0.198	0.545	0.481	0.499	0.492	0.27
CODOI	0.022	0.68	0.46	0.534	0.326	0.354
PESOI	0.295	0.626	0.551	0.837	0.749	0.42
IMC	0.507	0.236	0.465	0.79	0.848	-0.021
ICC	0.05	0.436	0.26	0.763	0.091	0.152
ICT	0.468	0.18	0.437	0.904	0.702	-0.157

	RODILLAI	CODOI	PESOI	IMC	ICC	ICT
EDAD	0.073	0.143	0.073	0.235	0.324	0.411
TALLAI	0.404	0.573	0.531	-0.099	0.315	-0.169
BICIPII	0.185	-0.006	0.28	0.546	-0.084	0.435
TRICIPII	-0.015	-0.25	0.072	0.415	-0.282	0.28
SUBESI	0.117	0.012	0.352	0.544	0.161	0.509
SUPRAI	0.198	0.022	0.295	0.507	0.05	0.468
BIACROI	0.545	0.68	0.626	0.236	0.436	0.18
BICRESI	0.481	0.46	0.551	0.465	0.26	0.437
CINTURAI	0.499	0.534	0.837	0.79	0.763	0.904
CADERAI	0.492	0.326	0.749	0.848	0.091	0.702
SENTADOI	0.27	0.354	0.42	-0.021	0.152	-0.157
RODILLAI	1	0.606	0.641	0.465	0.254	0.332
CODOI	0.606	1	0.643	0.341	0.456	0.293
PESOI	0.641	0.643	1	0.785	0.494	0.616
IMC	0.465	0.341	0.785	1	0.344	0.85
ICC	0.254	0.456	0.494	0.344	1	0.639
ICT	0.332	0.293	0.616	0.85	0.639	1

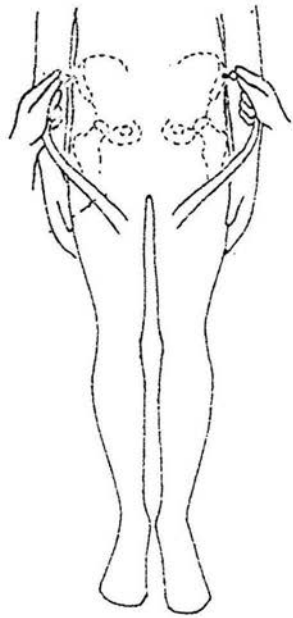
Anexo 4. Imágenes de algunas mediciones antropométricas

Técnica para la medición del pliegue subescapular

Técnica para la medición del diámetro biacromial



Técnica para la medición del diámetro bicrestal



Bibliografía

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

R.E. Blackitch; R.A Reyment. 1971. Multivariate Morphometrics. Academic Press Inc. New York.

Bendel, R. B. , y Afifi, A. A.(1977). Comparison of stopping rules in forward regresión. Journal of the American Statistical Association, 72.

Feinstein, Alvan 1985. Clinical Epidemiology. W.B. Saunders Company.

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

Mickey, J. y Greeland, S. (1989). A study of the impact of confounder-selection criteria on effect estimation. American Journal of Epidemiology, 129.

Pampel, F.C. 2000. Logistic Regresión: A Primer. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-132. Thousand Oaks, CA:Sage.

Kaufer Horwitz Martha, Protocolo de la investigación: Estatura, índice de masa corporal, grasa corporal total y riesgo de hipertensión arterial en adultos de ambos sexos.