Universidad Nacional Autónoma de México

Facultad de Ciencias

Regresión logística y una aplicación en la salud mental

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Actuario

PRESENTA:

Alfredo Sepúlveda Sastré

TUTOR

M. en C. José Salvador Zamora Muñoz

Ciudad Universitaria, CD. MX., 2016

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Resumen

La regresión logística es una técnica estadística frecuentemente usada para modelar datos con variable de respuesta dicotómica. El objetivo de esta tesis es señalar los aspectos técnicos y prácticos más importantes que conforman a los modelos logísticos así como ejemplificar la aplicación de la regresión logística con datos reales.

En el primer capítulo se define la familia de modelos lineales generalizados, los cuales cuentan con tres componentes: una distribución dentro de la familia exponencial, un predictor lineal y una función de enlace. A partir de ello, cuando se elige la distribución binomial y la función de enlace logit se llega al modelo de regresión logística, es decir, los modelos logísticos son un caso particular de los modelos lineales generalizados. En el segundo capítulo se expone la definición y las técnicas de ajuste del modelo logístico, para los casos univariado y multivariado por separado. Asimismo, se explican las pruebas para verificar la significancia de las variables introducidas a un modelo logístico. En el tercer capítulo se presenta la razón de momios como principal medida de asociación del modelo y se define el modelo E, V, W. Además, se define el concepto de interacción multiplicativa entre dos variables. En el cuarto capítulo se realiza un análisis de las principales medidas de bondad de ajuste y se discuten las técnicas que existen para evaluar el desempeño discriminatorio del modelo, es decir, su capacidad de predecir respuestas. En el quinto capítulo se toman como ejemplo los datos de una muestra de adolescentes con trastorno por déficit de atención con hiperactividad (TDAH) y se analizan .

Agradecimientos

A mis padres por su apoyo incondicional...

Índice general

Capítulo 1

Modelo lineal generalizado

1.1. Modelo lineal

El modelo lineal es un caso particular de los modelos lineales generalizados (GLMs, por sus siglas en inglés), a su vez, éste es el modelo lineal más simple y se utilizará como referencia para comprender la generalización en GLM. El modelo lineal clásico busca explicar una variable de interés, $\boldsymbol{Y}$, por medio de una combinación lineal de variables independientes, también llamadas **covariables**, $x_1, \ldots, x_p$, que afectan directamente el comportamiento de $\boldsymbol{Y}$. Este modelo consta de dos partes: la parte sistemática y la parte aleatoria. Se asume que un vector de observaciones, $\mathbf{y}$, con n componentes es la realización de una variable aleatoria $\mathbf{Y}$ cuyos componentes son idénticamente distribuidos y con media $\boldsymbol{\mu}$. La parte sistemática del modelo es la especificación del vector $\boldsymbol{\mu}$ en términos de un número reducido de parámetros desconocidos, $\beta_0, \beta_1, \ldots, \beta_p$ y de variables explicativas o **covariables** . En el modelo lineal, $\boldsymbol{\mu}$ es de la forma

$$\boldsymbol{\mu} = \beta_0 + \sum_{i=1}^{p} x_j \beta_j$$

donde los parámetros, $\beta_j, j = 0, \ldots, p$, son desconocidos y tienen que ser estimados por los datos. Para cada observación, i, de la variable aleatoria $\boldsymbol{Y}$ se tiene que la esperanza

$$E(Y_i) = \mu_i = \beta_0 + \sum_{i=1}^{p} x_{ij}\beta_j; \qquad\qquad i = 1, \ldots, n$$

donde x_{ij} es el valor de la j-ésima covariable de la observación i. Resumido en su forma matricial esto es

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$$

donde $\boldsymbol{\mu}$, $\boldsymbol{X}$ y $\boldsymbol{\beta}$ son de dimensión $n \times 1$, $n \times p$ y $p \times 1$, respectivamente. El componente $\boldsymbol{X}$ es llamado *matriz de diseño* y $\boldsymbol{\beta}$ *vector de parámetros.*

Para la parte aleatoria se asume independencia y varianza constante de los errores.

Si además se asume que los errores siguen una *distribución normal* con varianza constante σ^2 se puede resumir el modelo lineal como:

$$E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}.$$

1.2. Modelos lineales generalizados

Una generalización del modelo lineal clásico es lo que se conoce como **modelos lineales generalizados**. El modelo lineal clásico admite dos extensiones en su estructura básica. La primera es que la distribución de la parte aleatoria puede ser cualquier distribución dentro de la familia exponencial, en contraste al componente aleatorio de los modelos clásicos que se toma únicamente con distribución normal. La segunda extensión es que se incluye una función que relaciona el componente sistemático del modelo con la esperanza de $\boldsymbol{Y}$, llamada **función de enlace**.

En resumen, el modelo lineal generalizado consta de tres partes:

1. El componente aleatorio del modelo $\boldsymbol{Y}$ tiene una distribución dentro de la familia exponencial.

2. El componente sistemático dado por las covariables $x_1, \ldots, x_p$ produce un predictor lineal $\boldsymbol{\eta}$ dado por:
$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}.$$

3. La función de enlace $g(\cdot)$ puede ser cualquier función monótona diferenciable y relaciona los componentes sistemático y aleatorio como sigue:
$$\boldsymbol{\eta} = g(\boldsymbol{\mu})$$

donde $\boldsymbol{\mu} = E(\boldsymbol{Y})$.

Las aplicaciones de los modelos lineales generalizados en distintas áreas de estudio son muy amplias. Como un ejemplo de su uso se muestra el siguiente modelo.

Considere un lenguaje que desciende de otro; por ejemplo, el italiano proviene del latín. Un modelo simple para el cambio de vocabulario es que si los lenguajes se encuentran separados por un tiempo t, entonces la probabilidad de que tengan palabras afines para un significado en particular es $e^{-\theta t}$ donde θ es un parámetro. Para una lista de N significados distintos comúnmente usados supongamos que un lingüista juzga, para cada significado, si las palabras correspondientes en dos idiomas son afines. Se puede crear un modelo para describir lo anterior.

Se definen las variables $Y_1, \ldots, Y_n$ como sigue.

$$Y_i = \begin{cases} 1 \text{ si los lenguajes tienen palabras afines para el significado i,} \\ 0 \text{ si las palabras no son afines.} \end{cases}$$

entonces

$$P(Y_i = 1) = e^{-\theta t}$$

y

$$P(Y_i = 0) = 1 - e^{-\theta t}.$$

Este es un caso especial de la distribución $binomial(n, \pi)$ con $n = 1$ y $E(Y_i) = \pi = e^{-\theta t}$. En este caso el enlace g se toma como la función logarítmica

$$g(\pi) = log(\pi) = -\theta t$$

para que $g(E(\boldsymbol{Y}))$ resulte lineal en el parámetro θ. En notación de modelos lineales, $x_i = -t$ para cada i y $\boldsymbol{\beta} = \theta$

1.3. Función de verosimilitud para miembros de la familia exponencial

Cualquier variable Y que define la parte aleatoria de un modelo lineal generalizado tiene una distribución dentro de la familia exponencial, es decir, su función de densidad es de la forma:

$$f_Y(y; \theta, \phi) = exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

para funciones especificas $a(\cdot)$, $b(\cdot)$, $c(\cdot)$. Si ϕ es conocido, entonces Y pertenece a la familia exponencial con parámetro canónico θ.

La función de log-verosimilitud para $f_Y(y; \theta, \phi)$ la escribimos como

$$l(\theta, \phi; y) = ln(f_Y(y; \theta, \phi)).$$

La media y la varianza se pueden obtener de las siguientes relaciones conocidas de la familia exponencial

$$E\left(\frac{\partial L}{\partial \theta}\right) = 0,$$

$$E\left(\frac{\partial^2 L}{\partial \theta^2}\right) + E\left(\frac{\partial L}{\partial \theta}\right)^2 = 0$$

respectivamente. Para obtener la media tenemos

$$L(\theta, \phi; y) = (y\theta - b(\theta))/a(\phi) + c(y, \phi)$$

de donde

$$\frac{\partial L}{\partial \theta} = (y - b'(\theta))/a(\phi)$$

y

$$\frac{\partial^2 L}{\partial \theta^2} = -b''(\theta)/a(\phi).$$

Así nos queda

$$0 = E\left(\frac{\partial L}{\partial \theta}\right) = (\mu - b'(\theta))/a(\phi)$$

y finalmente

$$E(Y) = \mu = b'(\theta).$$

Similar para la varianza

$$0 = E\left(\frac{b''(\theta)}{a(\phi)}\right) + E\left(\frac{y - b'(\theta)}{a(\phi)}\right)^2$$

que resulta

$$0 = -\frac{b''(\theta)}{a(\phi)} + \frac{Var(Y)}{a^2(\phi)}$$

con lo que obtenemos

$$Var(Y) = b''(\theta)a(\phi).$$

1.4. Funciones de enlace

La función de enlace relaciona el predictor lineal η_i al valor esperado μ_i de cada observación de la respuesta y_i. En el modelo lineal, la media y el predictor lineal son idénticos, y el uso de la función identidad como enlace es posible pues η y μ toman cualquier valor real. En el caso de la distribución binomial se tiene que $0 < \mu < 1$ y el enlace debe proyectar el intervalo $(0, 1)$ en toda la recta real.

Existen varias funciones que cumple con el requisito anterior, como son:

Logit

$$\eta = ln\left(\frac{\mu}{1 - \mu}\right)$$

Probit

$$\eta = \Phi^{-1}(\mu)$$

donde $\Phi^{-1}(\mu)$ es la inversa de la función de distribución de probabilidades asociada a una distribución normal estándar.

Log-log complementaria

$$\eta = ln(-ln(1-\mu)).$$

Sin embargo, durante esta presentación se presta completa atención al primer caso (logit), que corresponde a la función de enlace del modelo logístico.

Capítulo 2

Modelo de regresión logística

Una pregunta de investigación típica consiste en relacionar una variable (o más) de estudio o exposición (E), con una variable de respuesta (D).

Por ejemplo, considérese una variable de respuesta **dicotómica**, es decir, que sólo toma dos valores, con categorías, 0, que representa la ausencia de una enfermedad, y 1, que representa un sujeto enfermo. Esta variable puede ser el estatus de cardiopatía coronaria (CHD por sus siglas en inglés) con sujetos clasificados como 0 (no tiene CHD) y 1 (tiene CHD).

Supóngase también que se está interesado en una sola variable de exposición dicotómica; el estatus de fumador, cuyas categorías son 0 (sujeto no fumador) y 1 (sujeto fumador). La pregunta de investigación para este ejemplo radica en evaluar en qué magnitud el estatus de fumador se asocia con la cardiopatía coronaria.

Para realizar la valoración anterior entre el estatus de fumador y CHD, se deben considerar **variables de control** adicionales, tales como edad, sexo y raza (C_1,C_2 y C_3 respectivamente) que no son de interés primordial.

En este ejemplo, la variable de exposición (E) y los controles (C_1,C_2 y C_3) representan una colección de **variables independientes**, que deseamos usar para explicar la **variable dependiente** (D).

De forma más general, las variables independientes pueden ser denotadas como $X_1, X_2, \ldots, X_k$, donde k es el número de variables a considerar.

La elección de las $X^{'}s$ es flexible, puede representar cualquier colección de variables de exposición, de control e incluso de combinaciones de las anteriores.

Una elección válida de variables independientes es la siguiente:

- X_1 igual a la variable de exposición E.
- X_2, X_3 iguales a las variables de control C_1, C_2 respectivamente.
- X_4 igual al producto $E \times C_1$.

- X_5 igual al producto $C_1 \times C_2$.
- X_6 igual a E^2.

Siempre que se desea relacionar un conjunto de variables independientes con una variable dependiente se está considerando un problema multivariado o múltiple. En un estudio, el uso de modelos matemáticos es casi indispensable para analizar las complicadas relaciones entre varias variables.

La **regresión logística** es un acercamiento por medio de un modelo matemático que puede ser usado para describir la relación entre variables independientes y una variable dependiente dicotómica.

2.1. Definición del modelo

Debido a que la variable de interés en un modelo logístico es una variable binaria o dicotómica, que en su codificación más típica toma los valores 0 y 1 únicamente, se debe aproximar satisfactoriamente la esperanza condicional $E(Y|x)$ que debe ser mayor o igual que 0 y menor o igual que 1. Es decir, $0 \leq E(Y|x) \leq 1$.

El **error**, $\boldsymbol{\varepsilon}$, es decir, la parte aleatoria del modelo, denota la desviación de la observación con respecto de la esperanza condicional.

Por conveniencia, en adelante escribiremos $\pi(x) = E(Y|x)$ para referirnos a la esperanza condicionada. En el caso de las variables de respuesta dicotómicas, se puede expresar una observación de la variable de respuesta como $y = \pi(x) + \varepsilon$. En este caso el error asume dos posibles valores: si $y = 1$ entonces $\varepsilon = 1 - \pi(x)$ con probabilidad $\pi(x)$, y si $y = 0$ entonces $\varepsilon = -\pi(x)$ con probabilidad $1 - \pi(x)$. Así, la distribución del error tiene media cero y varianza $\pi(x)\,[1 - \pi(x)]$. Se sigue que la distribución condicional de Y, denotada por $f(Y|x)$, sigue una distribución binomial con probabilidad $\pi(x)$.

Gran parte de las bondades del modelo logístico son heredadas de la **función logística**, que describe matemáticamente la base del modelo. Esta función es

$$f(z) = \frac{1}{1 + e^{-z}}$$

con z que toma valores en $(-\infty,\infty)$. Los valores pequeños, z, en el domino de la función, se acercan asintóticamente al 0, y los valores grandes, se acercan al 1. Ya que $0 \leq f(z) \leq 1$, esta función es ideal para describir probabilidades.

Otra propiedad interesante de la función logística son los **umbrales**. Si se empieza a evaluar la función desde $z = -\infty$ tenemos que los valores de $f(z)$ son cercanos

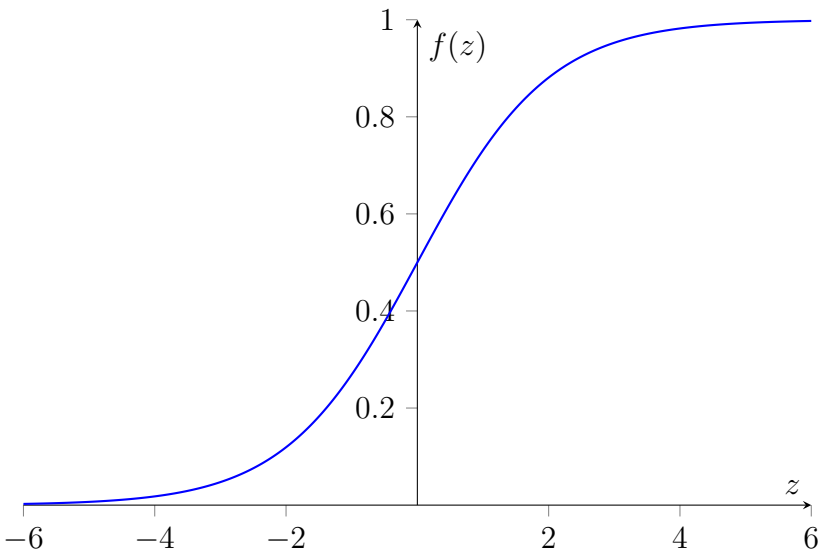

Figura 2.1: Gráfica de la función logística.

a 0, luego $f(z)$ incrementa drásticamente hasta llegar a valores cercanos a 1 donde su aumento se estabiliza hasta $z = \infty$. A estos puntos de inflexión se les llama umbrales.

El modelo logístico con una variable independiente o covariable, resulta de aplicar la función logística a la combianción lineal de x (covariable), $\eta = \beta_0 + \beta_1 x$, en donde β_0 y β_1 son términos constantes que representan parámetros desconocidos. Entonces

$$f(\eta) = f(\beta_0 + \beta_1 x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}.$$

Por lo anterior, el modelo logístico simple se define como:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}. \tag{2.1}$$

Alternativamente, el modelo logístico se puede definir a partir de la función **logit**, que en términos de modelos lineales generalizados es un enlace, $g(x)$, igual a la inversa de la función logística, es decir, $g(x) = f^{-1}(x)$. En general, el enlace es tal que

$$g(\pi(x)) = \eta$$

de donde se obtiene

$$g(\pi(x)) = ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x \tag{2.2}$$

que resulta equivalente a (2.1).

2.2. Ajuste del modelo de regresión logística simple

El ajuste de un modelo logístico simple requiere de la estimación de los parámetros desconocidos β_0 y β_1. Asúmase que se cuenta con n parejas de observaciones $(x_i, y_i), i = 1, \ldots, n$, de la variable independiente X y la variable de estudio Y, con codificación binaria $(0, 1)$, que expresa la presencia o ausencia de cierta característica de interés. El método usado para la estimación de tales parámetros es el de **máxima verosimilitud**. En términos generales, este método estima valores para los parámetros desconocidos que maximizan la probabilidad de obtener el conjunto de datos observado.

Para aplicar este método se debe encontrar la **función de verosimilitud**, que expresa la probabilidad de obtener un cierto conjunto de datos en función de los parámetros desconocidos.

Los estimadores **máximo-verosímiles** son las estimaciones de los parámetros desconocidos que maximizan tal función.

Dentro del contexto de regresión logística simple, el método es el siguiente: Si $\boldsymbol{\beta} = (\beta_0, \beta_1)$, la verosimilitud para un par de observaciones, (x_i, y_i), se puede expresar convenientemente como

$$l(\boldsymbol{\beta}) = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

pues integra ambos resultados de la variable de respuesta, de manera que

$$l(\boldsymbol{\beta}) = \begin{cases} P(Y = 1|x_i) = \pi(x_i) \text{ si } y_i = 1 \\ P(Y = 0|x_i) = 1 - \pi(x_i) \text{ si } y_i = 0 \end{cases}.$$

Si tomamos en cuenta n observaciones de la variable independiente y respuesta, la función de verosimilitud conjunta es

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

ya que se asume que las observaciones son independientes.

El principio de máxima verosimilitud sostiene que deben usarse como estimadores los valores del vector $\boldsymbol{\beta}$ que maximicen la función de verosimilitud, sin embargo, es equivalente maximizar el logaritmo de $l(\boldsymbol{\beta})$ para encontrar el máximo y muchas veces resulta más sencillo en la práctica. Esta nueva función es llamada **log-verosimilitud** y se define como

$$L(\boldsymbol{\beta}) = ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i ln[\pi(x_i)] + (1 - y_i) ln[1 - \pi(x_i)]\}. \tag{2.3}$$

Para encontrar el valor de $\boldsymbol{\beta}$ que maximice la función de log-verosimilitud se debe diferenciar $L(\boldsymbol{\beta})$ respecto de β_0 y β_1 e igualar a cero. Las **ecuaciones de verosimilitud**

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \tag{2.4}$$

y

$$\sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0 \tag{2.5}$$

resultan del proceso mencionado.

En las ecuaciones (2.4) y (2.5), los parámetros β_0 y β_1 resultan no lineales por lo que se deben usar métodos especiales para resolverlas. Nelder (1989) demuestra que se puede llegar a la solución de (2.4) y (2.5) usando un método iterativo de mínimos cuadrados ponderados.

El valor de $\boldsymbol{\beta}$ que soluciona las ecuaciones de máxima verosimilitud (2.4) y (2.5) es llamado **estimador máximo verosímil** y se simboliza $\widehat{\boldsymbol{\beta}}$.

Cualquier función del vector $\boldsymbol{\beta}$ se puede estimar una vez que se conocen los parámetros estimados $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)$. El procedimiento es evaluar la función deseada en las propias estimaciones paramétricas. Por ejemplo, el estimado de la esperanza condicional, $\pi(x)$, es

$$\widehat{\pi}(x) = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 x}}.$$

2.3. Significancia de las variables

Después de ajustar un modelo logístico simple, en particular, el primer paso es determinar la relevancia de las variables independientes, esto es, su **significancia**.

Un acercamiento para demostrar la significancia de una variable independiente dentro del modelo es contestar la pregunta: *¿El modelo que incluye dicha variable nos otorga más información acerca de la respuesta que el modelo que no la incluye?*. Esta pregunta se responde comparando los valores observados de la variable de respuesta y_i con los valores predichos por los modelos con y sin la variable evaluada, $\widehat{y_i}$ y $\widehat{y_i^*}$, respectivamente. Si por algún método se determina que los valores predichos por el modelo que incluye la variable son, en conjunto, *mejores* a los valores predichos por el modelo sin la variable, se dice entonces que la variable en cuestión es **estadísticamente significativa**.

Debe notarse que este acercamiento no evalúa qué tan buena representación hacen los valores predichos de los observados.

En regresión logística, la comparación entre los valores predichos y observados se realiza por medio de la función de log-verosimilitud ya definida en (2.3). Para comprender mejor esta comparación se debe pensar en los valores observados como valores predichos por el **modelo saturado**, es decir, un modelo el cual tiene la misma cantidad de parámetros que de observaciones.

La comparación entre valores observados y predichos se basa entonces en la siguiente expresión:

$$D = -2ln\left(\frac{verosimilitud\ del\ modelo\ ajustado}{verosimilitud\ del\ modelo\ saturado}\right). \tag{2.6}$$

La cantidad entre paréntesis dentro de la ecuación (2.6) es llamada **cociente de verosimilitudes**, y la transformación $h(x) = -2ln(x)$ se usa para obtener una distribución conocida con el fin de realizar pruebas de hipótesis. Más aún, cuando la variable de respuesta toma únicamente los valores cero y uno, la verosimilitud del modelo saturado, que se sigue de la ecuación (2.14), resulta

$$l(modelo\ saturado) = \prod_{i=1}^{n} y_i^{y_i}(1-y_i)^{y_i} = 1$$

pues por definición de un modelo saturado, los valores estimados y observados son iguales, es decir, $\widehat{\pi}(\boldsymbol{x}_i) = y_i$ (*Véase Bondad de Ajuste*).

Por lo tanto, siguiendo la expresión en (2.6)

$$D = -2ln(verosimilitud\ del\ modelo\ ajustado).$$

El estadístico D se conoce por el nombre de **devianza**.

Así, para valorar la significancia de una variable independiente o covariable dentro de un modelo de regresión logística comparamos el valor del estadístico D, incluyendo y excluyendo la variable en cuestión. Este es el método del **cociente de verosimilitudes**.

El cambio de D, cuando se incluye dicha variable es

$$G = D(modelo\ sin\ la\ variable) - D(modelo\ con\ la\ variable).$$

Al examinar el estadístico G se nota que se puede expresar de forma conveniente como

$$G = -2ln\left(\frac{verosimilitud\ sin\ la\ variable}{verosimilitud\ con\ la\ variable}\right). \tag{2.7}$$

Para el caso específico de un modelo simple, cuando la variable independiente no está en el modelo, el estimador máximo verosímil de β_0 es $ln(n_1/n_0)$ donde $n_1 = \sum y_i$ y $n_0 = \sum(1 - y_i)$ pues

$$\begin{aligned} L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{ y_i ln\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) + (1-y_i) ln\left(\frac{1}{1+e^{\beta_0}}\right) \right\} \\ &= \sum_{i=1}^{n} \left\{ y_i(\beta_0 - ln(1+e^{\beta_0})) - (1-y_i) ln(1+e^{\beta_0}) \right\} \\ &= \sum_{i=1}^{n} y_i\beta_0 - ln(1+e^{\beta_0}). \end{aligned}$$

Abusando de la notación respecto a la derivada parcial de β_0, (pues es el único parámetro)

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - \frac{e^{\beta_0}}{1+e^{\beta_0}})$$

igualando a cero según el método de máxima verosimilitud

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

y resolviendo para β_0 resulta

$$\widehat{\beta_0} = ln\left(\frac{n_1}{n-n_1}\right) = ln\left(\frac{n_1}{n_0}\right),\ n_0 = \sum_{i=1}^{n}(1-y_i),\ n_1 = \sum_{i=1}^{n} y_i.$$

Así, la función de log-verosimilitud para el modelo sin la variable es

$$\begin{aligned} L(\textit{modelo sin la variable}) &= \prod_{i=1}^{n} \widehat{\pi}(x_i)^{y_i} (1-\widehat{\pi}(x_i))^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{n_1}{n}\right)^{y_i} \left(\frac{n_0}{n}\right)^{1-y_i} \\ &= \left(\frac{n_1}{n}\right)^{\sum y_i} \left(\frac{n_0}{n}\right)^{\sum(1-y_i)} \\ &= \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_0}{n}\right)^{n_0}. \end{aligned}$$

En este contexto, el estadístico G para un modelo simple, resulta

$$G = -2ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\widehat{\pi}(x_i)^{y_i}(1-\widehat{\pi}(x_i))^{1-y_i}}\right]$$

o

$$G = 2\left\{\sum_{i=1}^{n}[y_i ln(\widehat{\pi}(x_i)) + (1-y_i)ln(1-\widehat{\pi}(x_i))]\right.$$

$$\left. - [n_1 ln(n_1) + n_0 ln(n_0)) - n ln(n)]\right\}.$$

Bajo la hipótesis de que $\beta_1 = 0$, el estadístico G se distribuye Ji-cuadrada con un grado de libertad y se simboliza, $G \sim \chi^2_{(1)}$.

Finalmente, se puede afirmar que para un modelo de regresión logística univariado, la variable independiente X es significativa si la probabilidad de que el valor de una distribución Ji-cuadrada con un grado de libertad sea mayor que el estadístico G, es menor a cierto valor de tolerancia que se elija, es decir, X es significativa a un grado de confianza α si

$$P[\chi^2_{(1)} > G] < (1 - \alpha),\ 0 < \alpha < 1.$$

La **prueba de Wald** es un equivalente de la prueba del cociente de máxima verosimilitud para el parámetro estimado $\widehat{\beta}_1$ usando una estimación de su error estándar. El estadístico para realizar esta prueba es comúnmente llamado W y se obtiene dividiendo el parámetro estimado $\widehat{\beta}_1$ entre la estimación de el error estándar de dicho parámetro $\widehat{SE}(\widehat{\beta}_1)$ (*véase sección 2.7.1*), esto es

$$W = \frac{\widehat{\beta}_1}{\widehat{SE}(\widehat{\beta}_1)}.$$

Bajo la hipótesis nula $H_0 : \beta_1 = 0$ e hipótesis de tamaño de muestra, el estadístico de Wald sigue aproximadamente una distribución Normal estándar, es decir, $W \sim N(0, 1)$. La prueba entonces consiste en verificar la condición

$$P(|z| > W) < (1 - \alpha),$$

donde z es una variable aleatoria con distribución Normal estándar, a un grado de significancia α; si la condición se cumple, la variable se considera significativa.

Nótese que esta prueba se puede realizar también usando la transformación W^2 que se distribuye ji-cuadrada con un grado de libertad.

2.4. Intervalos de confianza

La construcción de los intervalos de confianza se basa en la misma prueba estadística para corroborar la significancia de una variable. Para el caso univariado, la estimación de los intervalos de confianza para β_0 y β_1 usa el principio de la prueba de Wald y se refiere a éstos como **intervalos de confianza de Wald**.

Los intervalos tienen la forma

$$\widehat{\beta}_1 \pm z_{1-\alpha/2}\widehat{SE}(\widehat{\beta}_1) \tag{2.8}$$

y

$$\widehat{\beta}_0 \pm z_{1-\alpha/2}\widehat{SE}(\widehat{\beta}_0) \tag{2.9}$$

para $\widehat{\beta}_1$ y $\widehat{\beta}_0$ respectivamente, donde $z_{1-\alpha/2}$ denota el percentil $100(1-\alpha/2)\,\%$ de una distribución Normal estándar y $\widehat{SE}(\widehat{\beta}_i),\ i=0,1$, el estimador de error estándar de los parámetros involucrados.

El logit, $g(x) = \beta_0 + \beta_1 x$, la parte lineal del modelo de regresión logística, se estima calculándolo con los valores de las estimaciones paramétricas, es decir, $\widehat{g}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$. El estimador de la varianza para el logit requiere el siguiente cálculo directo que se deduce de la fórmula general

$$\begin{aligned}\widehat{Var}[\widehat{g}(x)] &= \widehat{Var}[\widehat{\beta}_0 + \widehat{\beta}_1 x] \\ &= \widehat{Var}(\widehat{\beta}_0) + x^2\widehat{Var}(\widehat{\beta}_1) + 2\widehat{Cov}(\widehat{\beta}_0, \widehat{\beta}_1)\end{aligned} \tag{2.10}$$

que se reduce a la estimación de varianzas y covarianzas de los parámetros.

Así entonces, el intervalo de confianza basado en la prueba de Wald para el logit es de la siguiente forma

$$\widehat{g}(x) \pm z_{1-\alpha/2}\widehat{SE}\,[\widehat{g}(x)]$$

y su notación es análoga a la de los intervalos de confianza para los parámetros.

El estimador del logit provee una forma de estimar el valor ajustado, en este caso, la probabilidad logística y su intervalo de confianza. El estimado de la probabilidad ajustada se obtiene al calcular la ecuación (2.1) con el estimado del logit (2.10), es decir

$$\widehat{\pi}(x) = \frac{e^{\widehat{g}(x)}}{1 + e^{\widehat{g}(x)}}. \tag{2.11}$$

y su intervalo de confianza basado en Wald es

$$\frac{e^{\widehat{g}(x) \pm z_{1-\alpha/2}\widehat{SE}[\widehat{g}(x)]}}{1 + e^{\widehat{g}(x) \pm z_{1-\alpha/2}\widehat{SE}[\widehat{g}(x)]}}$$

con notación análoga a los anteriores intervalos de confianza.

2.5. Modelo de regresión logística múltiple

Considérese una colección de p variables independientes denotadas por el vector $\boldsymbol{X}^T = (X_1, \ldots, X_p)$, donde cada X_i es una variable independiente incluida en el modelo. Asimismo, sea $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ el vector de parámetros del modelo. Tomando $P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \pi(\boldsymbol{x})$ como la probabilidad condicional de que la respuesta, $(Y = 1)$, se presente, con $\boldsymbol{x} = (1, x_1, \ldots, x_p)$ los valores observados del vector de

variables independientes $\boldsymbol{X}$, podemos denotar el modelo de regresión logística por medio de la ecuación

$$g(\boldsymbol{x}) = ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{2.12}$$

donde

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1+e^{g(\boldsymbol{x})}}. \tag{2.13}$$

2.6. Ajuste del modelo logístico múltiple

Asúmase una muestra de n observaciones independientes $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$. Como en el caso univariado, para ajustar el modelo multivariado se requiere estimar los parámetros desconocidos $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$. Como se describió anteriormente, el ajuste del modelo multivariado es análogo al caso simple. Para la estimación de $\boldsymbol{\beta}$ se utiliza el método de máxima verosimilitud. La versión de la función de máxima verosimilitud para el modelo completo es casi idéntica al modelo simple con excepción de que la probabilidad condicional $\pi(\boldsymbol{x})$ es ahora de la forma (2.13).

Conforme al método de máxima verosimilitud, se debe diferenciar la función de verosimilitud $l(\boldsymbol{\beta})$ con respecto a los parámetros y encontrar sus soluciones:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i}[1-\pi(\boldsymbol{x}_i)]^{1-y_i}. \tag{2.14}$$

Sustituyendo $\pi(\boldsymbol{x}_i)$ por su expresión como función de $\boldsymbol{\beta}$ resulta

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{\beta}\boldsymbol{x}_i^T}}{1+e^{\boldsymbol{\beta}\boldsymbol{x}_i^T}}\right)^{y_i} \left(\frac{1}{1+e^{\boldsymbol{\beta}\boldsymbol{x}_i^T}}\right)^{1-y_i}$$

$$= \prod_{i=1}^{n} \frac{\left(e^{\boldsymbol{\beta}\boldsymbol{x}_i^T}\right)^{y_i}}{1+e^{\boldsymbol{\beta}\boldsymbol{x}_i^T}}$$

y al obtener el logaritmo de (2.14) se obtiene la función de log-verosimilitud y $\boldsymbol{\beta}$ se hace lineal

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \boldsymbol{\beta}\boldsymbol{x}_i^T - \sum_{i=1}^{n} ln(1+e^{\boldsymbol{\beta}\boldsymbol{x}_i^T}).$$

Nótese que la combinación lineal de la variable $\boldsymbol{x}_i$ se puede escribir como un producto de vectores como $\boldsymbol{\beta}\boldsymbol{x}_i^T = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$.

Derivando con respecto a β_0 se obtiene

$$\begin{aligned}\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\boldsymbol{\beta}\boldsymbol{x}_i^T}}{1+e^{\boldsymbol{\beta}\boldsymbol{x}_i^T}} \\ &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi(\boldsymbol{x}_i)\end{aligned} \tag{2.15}$$

y con respecto a β_j, $j = 1, \ldots, p$, resulta

$$\begin{aligned}\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \frac{x_{ij} e^{\boldsymbol{\beta}\boldsymbol{x}_i^T}}{1+e^{\boldsymbol{\beta}\boldsymbol{x}_i^T}} \\ &= \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij}\pi(\boldsymbol{x}_i) \\ &= \sum_{i=1}^{n} x_{ij}(y_i - \pi(\boldsymbol{x}_i)).\end{aligned} \tag{2.16}$$

Luego, igualando a cero las ecuaciones (2.15), (2.16) se obtiene respectivamente

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi(\boldsymbol{x}_i) = 0 \tag{2.17}$$

y

$$\sum_{i=1}^{n} x_{ij}(y_i - \pi(\boldsymbol{x}_i)) = 0,\ j = 1, \ldots, p. \tag{2.18}$$

El sistema de $p+1$ ecuaciones resultantes de (2.17) y (2.18) es un sistema no lineal. Por este motivo es necesaria la implementación de métodos numéricos para su solución. En consecuencia, los valores ajustados para el modelo son $\widehat{\pi}(\boldsymbol{x}_i)$, $i = 1, \ldots, n$, es decir, la ecuación (2.13) calculada usando la solución $\widehat{\boldsymbol{\beta}}$ y el vector $\boldsymbol{x}_i$.

Nótese que de (2.17) se sigue

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \pi(\boldsymbol{x}_i)$$

Esto es, la suma de los valores observados de la variable de respuesta Y es igual a la suma de los valores esperados $\pi(\boldsymbol{x}_i)$. Dicho resultado es útil para aseverar el ajuste del modelo.

2.6.1. Estimación de la varianza de los parámetros

El método para estimar varianzas y covarianzas de los parámetros estimados se sigue de la teoría de estimación máximo-verosímil. Los estimadores se obtienen de la

matriz de segundas derivadas parciales de la función de verosimilitud. Las entradas de dicha matriz tienen la siguiente forma general

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j^2} &= -\sum_{i=1}^{n} x_{ij}\frac{x_{ij}e^{\boldsymbol{\beta}\boldsymbol{x}^T}(1+e^{\boldsymbol{\beta}\boldsymbol{x}^T}) - x_{ij}(e^{\boldsymbol{\beta}\boldsymbol{x}^T})^2}{(1+e^{\boldsymbol{\beta}\boldsymbol{x}^T})^2} \\ &= -\sum_{i=1}^{n} x_{ij}^2\pi(\boldsymbol{x}_i)(1-\pi(\boldsymbol{x}_i))\end{aligned} \tag{2.19}$$

y

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} &= -\sum_{i=1}^{n} x_{ij}\frac{x_{ik}e^{\boldsymbol{\beta}\boldsymbol{x}^T}(1+e^{\boldsymbol{\beta}\boldsymbol{x}^T}) - x_{ik}(e^{\boldsymbol{\beta}\boldsymbol{x}^T})^2}{(1+e^{\boldsymbol{\beta}\boldsymbol{x}^T})^2} \\ &= -\sum_{i=1}^{n} x_{ij}x_{ik}\pi(\boldsymbol{x}_i)(1-\pi(\boldsymbol{x}_i))\end{aligned} \tag{2.20}$$

para $j, k = 0, 1, \ldots, p$.

Sea la matriz de dimensión $(p+1)\times(p+1)$ que contiene los negativos de (2.19) y (2.20) denotada por $\mathbf{I}(\boldsymbol{\beta})$. Esta matriz es llamada **matriz de información observada de Fisher**. Las varianzas y covarianzas de los parámetros estimados se obtienen con el inverso de la matriz de información, denotada por $Var(\boldsymbol{\beta}) = \mathbf{I}^{-1}(\boldsymbol{\beta})$.

Los estimadores de las varianzas y covarianzas se obtienen al evaluar $Var(\boldsymbol{\beta})$ en $\widehat{\boldsymbol{\beta}}$ y se denotan $\widehat{Var}(\widehat{\boldsymbol{\beta}})$. El error estándar estimado de los parámetros estimados $\widehat{SE}(\widehat{\beta}_j)$ toma la siguiente forma

$$\widehat{SE}(\widehat{\beta}_j) = \left[\widehat{Var}(\widehat{\boldsymbol{\beta}})\right]^{1/2}$$

para $j = 0, 1, \ldots, p$.

Otra forma de obtener la matriz de información de Fisher es $\widehat{\mathbf{I}}(\boldsymbol{\beta}) = \mathbf{X}'\widehat{\mathbf{V}}\mathbf{X}$, donde $\mathbf{X}$ es una matriz de $n\times(p+1)$ que contiene las observaciones de cada individuo y $\widehat{\mathbf{V}}$ es una matriz diagonal de dimensión $n\times n$, con el elemento $\pi(\boldsymbol{x}_i)(1-\pi(\boldsymbol{x}_i))$. Esto es, la matriz $\mathbf{X}$ es

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{2.21}$$

y la matriz $\widehat{\mathbf{V}}$ es

$$\widehat{\mathbf{V}} = \begin{bmatrix} \widehat{\pi}_1(1-\widehat{\pi}_1) & 0 & 0 & \cdots & 0 \\ 0 & \widehat{\pi}_2(1-\widehat{\pi}_2) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \widehat{\pi}_n(1-\widehat{\pi}_n) \end{bmatrix} \tag{2.22}$$

donde $\widehat{\pi}_i = \widehat{\pi}(\boldsymbol{x}_i)$ es el valor de $\pi(\boldsymbol{x}_i)$ calculado usando $\widehat{\boldsymbol{\beta}}$ y la covariable del sujeto i, $\boldsymbol{x}_i$.

2.7. Significancia del modelo multivariado

La prueba del cociente de verosimilitudes para el modelo múltiple es análoga al caso univariado, sin embargo, la versión múltiple de la prueba tiene la finalidad de verificar la significancia de al menos una de las p variables independientes, $(X_1, \ldots, X_p)$, de manera simultánea. Ya que la función de verosimilitud de un modelo saturado siempre es igual a 1, el cálculo del estadístico G se basa en la misma expresión que (2.7), sin embargo, se compara el modelo con p variables independientes y un modelo con ninguna variable independiente y un parámetro β_0. La verosimilitud del modelo sin variables explicativas se conserva y G resulta

$$G = -2ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\widehat{\pi}(\boldsymbol{x}_i)^{y_i}(1-\widehat{\pi}(\boldsymbol{x}_i))^{1-y_i}}\right]$$

Bajo el supuesto de que todos los parámetros (excepto la constante del modelo β_0) son cero, es decir, $\beta_1 = \ldots = \beta_p = 0$, el estadístico G se distribuye Ji-cuadrada con p grados de libertad, $G \sim \chi^2_{(p)}$.

La prueba del cociente de verosimilitudes multivariada es entonces la comparación

$$P[\chi^2_{(p)} > G] < (1-\alpha) \tag{2.23}$$

a un grado de significancia α. Si la condición se cumple, se dice que al menos una variable es significativamente distinta de cero.

2.7.1. Prueba de Wald multivariada

La prueba de Wald en su forma univariada es también útil en el contexto de un modelo múltiple para comprobar la significancia de variables individualmente.

La expresión

$$W_j = \frac{\widehat{\beta}_j}{\widehat{SE}(\widehat{\beta}_j)}$$

denota el estadístico de Wald en su versión univariada. Esta prueba puede ser adoptada en los modelos logísticos múltiples cuando se desea evaluar la significancia individual de una cierta variable j.

El análogo del estadístico de Wald para el caso múltiple se puede calcular de la siguiente forma

$$\mathbf{W} = \widehat{\boldsymbol{\beta}}'[\widehat{Var}(\widehat{\boldsymbol{\beta}})]^{-1}\widehat{\boldsymbol{\beta}}$$

$$= \widehat{\boldsymbol{\beta}}'(\mathbf{X}'\widehat{\mathbf{V}}\mathbf{X})\widehat{\boldsymbol{\beta}}$$

donde $\mathbf{X}$ es la matriz de datos de las variables $\boldsymbol{x}_i,\ i = 1, \ldots, n$ previamente discutida en (2.21) y $\widehat{\mathbf{V}}$ es la matriz diagonal de varianzas (2.22).

Bajo la hipótesis nula ($H_0 : \boldsymbol{\beta} = 0$), el estadístico $\mathbf{W}$ en su versión múltiple (multivariada) se distribuye aproximadamente $\chi^2_{(p+1)}$.

La prueba del cociente de verosimilitudes encuentra un equivalente en la prueba de Wald múltiple si la matriz $\mathbf{X}$ y el vector de parámetros $\widehat{\boldsymbol{\beta}}$ se modifican para contener únicamente la información de los p parámetros asociados a la variables en el modelo, es decir, al eliminar $\widehat{\beta}_0$ de $\widehat{\boldsymbol{\beta}}$ y la primer fila y columna de la matriz $\mathbf{X}'\widehat{\mathbf{V}}\mathbf{X}$ relacionada con la información en $\mathbf{X}$ de β_0. En este caso, el estadístico $\mathbf{W}$ se distribuye Ji-cuadrada con p grados de libertad, $\mathbf{W} \sim \chi^2_p$.

De esta forma, para comprobar la significancia del modelo se utiliza la comparación

$$P[\chi^2_p > \mathbf{W}] < (1 - \alpha)$$

análoga a la comparación para la prueba del cociente de verosimilitudes en (2.23).

2.8. Intervalos de confianza

La generación de intervalos de confianza para los parámetros de un modelo logístico múltiple es idéntica que en el caso univariado y su expresión se conserva igual que en (2.8) y (2.9).

Estrictamente, los intervalos de confianza paramétricos del modelo múltiple resultan en la expresión

$$\widehat{\beta}_j \pm z_{1-\alpha/2}\widehat{SE}(\widehat{\beta}_j),\ j = 0, \ldots, p \tag{2.24}$$

para el parámetro j, donde $z_{1-\alpha/2}$ denota el percentil $100(1 - \alpha/2)\,\%$ de una distribución Normal estándar y $\widehat{SE}(\widehat{\beta}_j);\ j = 0, 1$ el estimador de error estándar del parámetro estimado, j.

El intervalo de confianza para el logit, $\widehat{g}(\boldsymbol{x})$, es más complejo en el caso múltiple que en el simple. Su añadida complejidad reside en la estimación de la varianza del logit, aunque se sigue de la misma idea que en (2.10).

Dada la expresión del logit para un modelo con p covariables

$$\widehat{g}(\boldsymbol{x}) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p$$

la estimación de la varianza del logit, $\widehat{Var}[\widehat{g}(\boldsymbol{x})]$, se calcula como sigue

$$\begin{aligned} \widehat{Var}[\widehat{g}(\boldsymbol{x})] &= \widehat{Var}[\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p] \\ &= \sum_{j=1}^{p} x_j^2 \widehat{Var}(\widehat{\beta}_j) + \sum_{j=1}^{p} \sum_{k=j+1}^{p} 2 x_j x_k \widehat{Cov}(\widehat{\beta}_j, \widehat{\beta}_k) \end{aligned} \tag{2.25}$$

con $x_0 = 1$.

Usando notación matricial, el logit se puede expresar como $\widehat{g}(\boldsymbol{x}) = \boldsymbol{x}'\widehat{\boldsymbol{\beta}}$ donde $\boldsymbol{x}' = (x_0, x_1, \ldots, x_p)$, con $x_0 = 1$, y el resultado (2.25) se resume en

$$\begin{aligned} \widehat{Var}[\widehat{g}(\boldsymbol{x})] &= \boldsymbol{x}' \widehat{Var}(\widehat{\boldsymbol{\beta}}) \boldsymbol{x} \\ &= \boldsymbol{x}' (\mathbf{X}' \widehat{\mathbf{V}} \mathbf{X})^{-1} \boldsymbol{x} \end{aligned}$$

que se sigue del resultado (2.6.1). Entonces, el intervalo de confianza para el logit multivariado es

$$\widehat{g}(\boldsymbol{x}) \pm z_{1-\alpha/2} \widehat{SE}\left[\widehat{g}(\boldsymbol{x})\right].$$

Del resultado anterior se deduce

$$\widehat{\pi}(\boldsymbol{x}) = \frac{e^{\widehat{g}(\boldsymbol{x})}}{1 + e^{\widehat{g}(\boldsymbol{x})}},$$

y su intervalo de confianza

$$\frac{e^{\widehat{g}(\boldsymbol{x}) \pm z_{1-\alpha/2} \widehat{SE}[\widehat{g}(\boldsymbol{x})]}}{1 + e^{\widehat{g}(\boldsymbol{x}) \pm z_{1-\alpha/2} \widehat{SE}[\widehat{g}(\boldsymbol{x})]}}.$$

Capítulo 3

Interpretación del modelo

3.1. Medidas de asociación

Después de realizar el ajuste de un modelo logístico, el siguiente tema de interés es otorgar un sentido cuantitativo útil de dicho modelo. La interpretación satisfactoria de un modelo requiere deducir un conjunto de conclusiones prácticas acerca de los datos de interés, esto, con base en los parámetros ajustados.

Así, la interpretación del modelo logístico involucra al menos dos problemas: determinar la dependencia funcional entre las variables dependiente e independiente, y definir apropiadamente la unidad de cambio de la variable independiente. En otras palabras, la interpretación de un modelo consiste en determinar cuantitativamente cómo afecta el cambio individual y grupal de las variables independientes, en la variable de estudio o respuesta.

La forma más natural de llegar a estas conclusiones es comparar directamente los riesgos predichos de un par de individuos diferentes, $\pi(\boldsymbol{x}_i)$ y $\pi(\boldsymbol{x}_j)$, y deducir cuántas veces mayor son los riesgos para conjuntos de variables independientes distintas. Esto da origen a la medida del **riesgo relativo (RR)**, definida como

$$RR = \frac{\pi(\boldsymbol{x}_i)}{\pi(\boldsymbol{x}_j)}, \quad i \neq j.$$

Sin embargo, su uso presenta dos condiciones. La primera es que sólo se puede calcular un riesgo individual, $\pi(\boldsymbol{x}_i)$, en estudios de seguimiento (follow-up). La segunda es que para calcular un riesgo individual se debe especificar cada valor del vector de variables independientes $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.

Aunque es recomendado usar la medida anterior en los estudios de seguimiento, sus condiciones prohíben su cálculo en otros casos, sin embargo, existen otras medidas de asociación más flexibles. Antes de introducir la **razón de momios**, una medida alternativa al RR, que puede ser usada en todos los casos de regresión

logística, se debe definir el concepto de **momios (odds)**.

3.1.1. Momios

Los **Momios** son la razón de la probabilidad que ocurra un evento entre la probabilidad de que no suceda el mismo evento. Su fórmula es

$$Momios(p) = \frac{p}{1-p}$$

donde p es la probabilidad de que suceda un evento. Por ejemplo, si $p = 0.25$ es la probabilidad de que ocurra un evento, entonces la probabilidad de que no ocurra el mismo evento es $1 - p = 0.75$, y los momios de ese evento resultan $\frac{1}{3}$. Alternativamente, se puede decir que los momios son 3 a 1 de que el evento no ocurra.

El concepto de momios se puede aplicar a la regresión logística reemplazando la probabilidad p por el riesgo ajustado $\pi(\boldsymbol{x})$, así obteniendo una medida para los momios de presentar una respuesta positiva ($Y = 1$), para un individuo con una especificación particular de $\boldsymbol{x}$. Los **momios del riesgo** (MR) se pueden expresar como

$$MR = \frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}$$

para cierta especificación de $\boldsymbol{x}$.

3.1.2. Razón de momios (RM)

El concepto de Momios, en específico los Momios del riesgo, resultan útiles al evaluar el riesgo de un sujeto específico, pero es necesario contar con una medida de asociación análoga al RR para verificar el impacto de las $\boldsymbol{X}'s$ al variar. La **Razón de Momios (RM)** compara los momios de un par de especificaciones de la variable independiente, $\boldsymbol{x}_i, \boldsymbol{x}_j$ de manera directa. Esta medida en general se calcula como

$$RM_{\boldsymbol{x}_i,\boldsymbol{x}_j} = \frac{MR(\boldsymbol{x}_i)}{MR(\boldsymbol{x}_j)}.$$

lo cual permite su cálculo en cualquier modelo logístico ajustado pues

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

se sigue de la definición del modelo (2.12).

Si $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ y $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$, son las variables medidas a los individuos i y j, respectivamente, entonces, en general, la razón de Momios se calcula como

$$RM_{\boldsymbol{x}_i,\boldsymbol{x}_j} = e^{\sum_{k=1}^{p} \beta_i(x_{ik}-x_{jk})}. \tag{3.1}$$

Alternativamente, la fórmula (3.1) se puede escribir como un producto de funciones de cada variable independiente por separado, forma útil para saber cómo contribuye cada variable de forma multiplicativa al RM, así

$$RM_{\boldsymbol{x}_i,\boldsymbol{x}_j} = \prod_{k=1}^{p} e^{\beta_i(x_{ik}-x_{jk})}.$$

Aunque se cuenta con una fórmula general para el cálculo de la Razón de Momios, es necesario definir algunos conceptos claves antes de mostrar el uso del RM en la interpretación de un modelo general de regresión logística, que incluye variables de exposición, de control y productos de variables; el modelo E,V,W.

3.1.3. Interacción multiplicativa

Al considerar un modelo logístico con más de dos variables independientes, se requiere considerar, además del efecto de cada variable en el modelo, un posible efecto que surge de su combinación o interacción.

Considérese, por ejemplo, un modelo logístico con dos variables dicotómicas independientes, A y B, y la variable de respuesta Y. Para este modelo existen cuatro posibles valores para el riesgo $\pi_{A,B} = P(Y = 1|A, B)$ que son $\pi_{0,0} = P(Y = 1|A = 0, B = 0)$, $\pi_{0,1} = P(Y = 1|A = 0, B = 1)$, $\pi_{1,0} = P(Y = 1|A = 1, B = 0)$ y $\pi_{1,1} = P(Y = 1|A = 1, B = 1)$. Dentro de este marco se pueden calcular las razones de momios para obtener el efecto de cada variable individualmente y en conjunto, usando como referencia los momios de $A = 0$ y $B = 0$, es decir

$$RM_{0,1} = \frac{\pi_{0,1}/(1-\pi_{0,1})}{\pi_{0,0}/(1-\pi_{0,0})},$$

$$RM_{1,0} = \frac{\pi_{1,0}/(1-\pi_{1,0})}{\pi_{0,0}/(1-\pi_{0,0})}$$

y

$$RM_{1,1} = \frac{\pi_{1,1}/(1-\pi_{1,1})}{\pi_{0,0}/(1-\pi_{0,0})}.$$

Entonces, la igualdad

$$RM_{1,1} = RM_{0,1} \times RM_{1,0} \tag{3.2}$$

expresa las aportaciones al modelo de cada variable por separado y conjuntamente. Si la ecuación anterior no se cumple, se afirma que existe interacción multiplicativa entre ambas variables.

En general, para demostrar la interacción entre dos variables A y B, basta con incluir una tercera **variable de interacción** $A\times B$ al modelo de estudio y comprobar

la significancia de su coeficiente, y al comprobar la igualdad (3.2) se puede escribir como la hipótesis nula

$$H_0 : \frac{RM_{A,B}}{RM_{0,B} \times RM_{A,0}} = 1$$

y es equivalente a comprobar

$$H_0 : ln\left(\frac{RM_{A,B}}{RM_{0,B} \times RM_{A,0}}\right) = \beta_k = 0$$

con β_k el coeficiente de $A \times B$ en el modelo logístico.

3.1.4. Modelo E, V, W

Para definir correctamente un modelo logístico general (por el tipo de variables que considera), se debe tomar en cuenta que no todas las variables independientes incluidas en un modelo, son de pleno interés al establecer una relación funcional entre una (o varias) variable explicativa y una respuesta. En muchos casos, se deben incluir variables que ajustan el modelo a circunstancias más específicas, a estas variables se les llama **variables de control** y se denotan como $C_1, C_2, \ldots, C_p$. Por ejemplo, para saber la relación entre el estatus de fumador y cardiopatías, se puede considerar un modelo logístico con esas dos variables, y además restringir el modelo por las edades y sexo de los sujetos.

Además de una variable de interés, E, el modelo E, V, W incluye p_1 variables denotadas como $V_1, V_2, \ldots, V_{p_1}$. Las $V's$ son funciones de los controles, $C's$, llamadas **confusoras**, y distorsionan la asociación entre E y la respuesta, Y (p.g. $V_1 = C_1, V_2 = C_3^2$). El modelo general también incluye p_2 variables que son productos de la forma $E \times W_1, E \times W_2, \ldots, E \times W_{p_2}$. Las $W's$ también son funciones de las $C's$ que interactúan con E llamadas **efectos modificadores**.

El modelo logístico que incorpora las variables E, $V's$ y $W's$ definidas anteriormente se escribe en su forma logit (2.12) como

$$logit(\pi(\boldsymbol{x})) = \alpha + \beta E + \sum_{i=1}^{p_1} \gamma_i V_i + E \sum_{j=1}^{p_2} \delta_j W_j.$$

La expresión de la **razón de momios ajustada** para este modelo, es decir, que compara $E = 0$ con $E = 1$, mientras mantiene constantes los controles, $C_1, C_2, \ldots, C_p$, se deduce de la formula general de la razón de momios (3.1) y es

$$RM_{E=0,E=1} = e^{(\beta + \sum_{j=1}^{p_2} \delta_j W_j)}.$$

Esta fórmula de RM expresa que si el modelo contiene términos de interacción $(E \times W_k)$, entonces el valor de la RM varía con los coeficientes $\delta's$ y depende de los

valores de W_j que se elijan. Esta propiedad debe hacer sentido pues el concepto de interacción implica que el efecto E es distinto para cada valor de las $W's$.

El caso anterior es la versión más simple del modelo E, V, W que integra únicamente una variable de interés, E, y además esta variable es dicotómica con codificación $E = (0, 1)$. Hallar la RM para el mismo modelo, pero con E codificada como (a, b) resulta trivial al remontarse a la fórmula general de la RM (3.1) y verificar que

$$RM_{E=a,E=b} = e^{((b-a)\beta+(b-a)\sum_{j=1}^{p_2}\delta_j W_j)}.$$

La fórmula anterior resulta válida también en el caso de una variable E **continua** u **ordinal**, con excepción de que para obtener una RM se deben de elegir dos valores (grupos de interés) de la variable de interés a comparar, a saber $E^* = a$ y $E^{**} = b$.

También es posible incluir una variable E de tipo **nominal** en el modelo anterior, es decir, una variable con k categorías no ordenables (no es posible determinar una escala). Un ejemplo de una variable nominal es la ocupación de un individuo. Para incorporar este tipo de variables en el modelo E, V, W se deben antes crear $k-1$ variables dicotómicas llamadas **variables dummys**, denotadas $E_1^f, E_2^f, \ldots, E_{k-1}^f$, que se definen como

$$E_i^f = \begin{cases} 1 \text{ si E es la categoría i} \\ 0 \text{ cualquier otro caso} \end{cases}$$

y se incluyen en el modelo como

$$logit(\pi(\boldsymbol{x})) = \alpha+\beta_1 E_1^f+\beta_2 E_2^f+\ldots+\beta_{k-1}E_{k-1}^f+\sum_{i=1}^{p_1}\gamma_i V_i+\sum_{j=1}^{p_2}\delta_j W_j; i = 1, \ldots, k-1$$

donde la k-ésima categoría se autodefine cuando $E_i^f = 0; i = 1, \ldots, k-1$.

Para encontrar la expresión de la RM en este caso es necesario especificar dos categorías $\boldsymbol{E}^*$ y $\boldsymbol{E}^{**}$ de la variable de interés nominal a ser comparadas, en términos de las $k-1$ variables dummys. En general $\boldsymbol{E}^* = E_1^{f*}, E_2^{f*}, \ldots, E_{k-1}^{f*}$ y $\boldsymbol{E}^{**} = E_1^{f**}, E_2^{f**}, \ldots, E_{k-1}^{f**}$.

Entonces la RM se sigue nuevamente de (3.1) y es

$$RM_{\boldsymbol{E}^*,\boldsymbol{E}^{**}} = e^{((b-a)\beta+(b-a)\sum_{j=1}^{p_2}\delta_j W_j)}.$$

Posterior a la definición de una razón de momios para el caso general de un modelo logístico con una variable de interés, es posible definir el modelo E, V, W **completo**, que incluye q variables de interés denotadas $E_1, E_2, \ldots, E_q$ que pueden

ser de tipo ordinal (incluyendo dicotómicas) y continuas. El caso de variables nominales se excluye pues una variable nominal debe ser transformada en variables dummys según se muestra en el caso anterior.

El modelo E, V, W completo se puede escribir en su forma logit como

$$\begin{aligned} logit(\pi(\boldsymbol{x})) = \alpha + \beta_1 E_1 + \beta_2 E_2 + \ldots + \beta_q E_q + \sum_{i=1}^{p_1} \gamma_i V_i \\ + E_1 \sum_{j=1}^{p_2} \delta_{1j} W_j \\ + E_2 \sum_{j=1}^{p_2} \delta_{2j} W_j \\ + \ldots \\ + E_q \sum_{j=1}^{p_2} \delta_{qj} W_j \end{aligned}$$

Esta expresión incluye las interacciones de cada variable de interés, E_k, en forma de una suma resumida como $E_k \sum_{j=1}^{p_2} \delta_{kj} W_j$ en donde δ_{kj} es un parámetro desconocido (ajustado por el modelo) y W_j es la j-ésima variable modificadora de efecto.

Nótese que el modelo anterior utiliza las mismas variables modificadoras de efecto para cada variable de interés. Aunque es posible definir un modelo más general que acepte diferentes modificadoras para las variables de interés, se prefiere el modelo anterior por su escritura conveniente.

Al igual que en el caso de un modelo con una variable nominal, para calcular la RM del modelo anterior se deben especificar dos categorías $\boldsymbol{E}^* = (E_1^*, \ldots, E_q^*)$ y $\boldsymbol{E}^{**} = (E_1^{**}, \ldots, E_q^{**})$ a comparase; la fórmula para la RM del modelo E, V, W completo es entonces

$$\begin{aligned} RM_{\boldsymbol{E}^*, \boldsymbol{E}^{**}} = exp[(E_1^* - E_1^{**})\beta_1 + (E_2^* - E_2^{**})\beta_2 + \ldots + (E_q^* - E_q^{**})\beta_q + . \\ + (E_1^* - E_1^{**}) \sum_{j=1}^{p_2} \delta_{1j} W_j \\ + (E_2^* - E_2^{**}) \sum_{j=1}^{p_2} \delta_{2j} W_j \\ + \ldots \\ + (E_q^* - E_q^{**}) \sum_{j=1}^{p_2} \delta_{qj} W_j]. \end{aligned} \tag{3.3}$$

Capítulo 4

Bondad de ajuste

Después de ajustar cualquier modelo estadístico a un conjunto de datos es necesario evaluar qué tan parecidas son las estimaciones efectuadas por el modelo a las observaciones reales. A este concepto se le conoce como bondad de ajuste. Generalmente, si las estimaciones son semejantes a los datos reales, entonces se dice que el ajuste del modelo es bueno. En caso contrario, si las estimaciones son disímiles a los datos reales, entonces el ajuste del modelo no se considera bueno.

Una medida de bondad de ajuste provee una comparación general entre las respuestas observadas (Y_i) y las estimadas ($\widehat{Y}_i$). Bajo este razonamiento, un ajuste perfecto ocurre cuando la diferencia entre cada respuesta observada y su análoga estimada resulta nula, es decir, cuando $Y_i - \widehat{Y}_i = 0$ para toda i. Los modelos que logran un ajuste perfecto se conocen como modelos saturados. En general, el modelo saturado para un conjunto de datos se define como cualquier modelo que contiene la misma cantidad de parámetros que de observaciones (tamaño de muestra). Los modelos saturados tiene la capacidad de predecir perfectamente la magnitud de la variable de respuesta para cada sujeto de la muestra.

Alternativamente, otro enfoque es considerar un ajuste perfecto por grupos formados por individuos que comparten patrones de covariables. Un patrón de covariables se define como una configuración particular de los valores de las covariables. Por ejemplo, si un modelo tiene las covariables binarias (X_1,X_2) que pueden tomar los valores 0 y 1, entonces el modelo puede tener a lo más 4 patrones de covariables distintos (1: $X_1 = 0$, $X_2 = 0$; 2: $X_1 = 0$, $X_2 = 1$; 3:$X_1 = 1$, $X_2 = 0$; 4:$X_1 = 1$, $X_2 = 1$).

Profundizando, supongamos que nuestro modelo contiene p covariables, $\boldsymbol{X} = (X_1, \ldots, X_p)$, y sea J el número de patrones de covariables distintos de n individuos. Si algún sujeto comparte un patrón de covariables con otro, entonces $J < n$. Se denota al número total de sujetos que cumplen $\boldsymbol{X} = \boldsymbol{X}_j$ como m_j, $j = 1, \ldots, J$. Se

sigue que $\sum_1^J m_j = n$. También, sea y_j el número de respuestas que cumplen $Y_i = 1$, $i = 1, \ldots, m_j$, dentro de los sujetos con $\boldsymbol{X} = \boldsymbol{X}_j$, entonces $\sum_1^J y_j = n_1$, el número total de sujetos con $Y = 1$.

Un modelo con ajuste perfecto por grupos no tiene la capacidad de predecir la respuesta individual sino la proporción de respuestas positivas por grupo, o equivalentemente, la cantidad de individuos con respuesta positiva en cada grupo. A estos modelos se les conoce como modelos saturados por grupos o modelos totalmente parametrizados. Los modelos saturados por grupos contienen todas las covariables que es posible definir con base en las covariables principales, también, su cantidad de parámetros es igual a la cantidad de patrones de covariables diferentes posibles para ese modelo. Los modelos totalmente parametrizados tienen la propiedad de ser los modelos más grandes (con mayor cantidad de parámetros) que se pueden ajustar usando las covariables de interés elegidas. Por ejemplo, si tenemos las covariables binarias X_1 y X_2 codificadas con los valores 0 y 1 y una respuesta Y, entonces el modelo totalmente parametrizado resulta $Y = a + bX_1 + cX_2 + dX_1X_2$. El acercamiento clásico de la bondad de ajuste considera como punto de referencia los modelos saturados. Como ya se explicó, estos modelos predicen, de manera exacta, la respuesta de cada individuo. Un modelo saturado ajustado para n sujetos es de la forma $g(\boldsymbol{X}) = w_1X_1 + w_2X_2 + \ldots + w_nX_n$ en donde

$$X_i = \begin{cases} 1 \text{ para el sujeto } i, i = 1, \ldots, n \\ 0 \text{ c.o.c} \end{cases} .$$

La función de verosimilitud para el modelo saturado L_{MS} es por definición

$$L_{MS} = \prod_{i=1}^{n} P(\boldsymbol{X}_i)^{Y_i}[1 - P(\boldsymbol{X}_i)]^{1-Y_i}. \tag{4.1}$$

Ya que para el sujeto i, $X_i = 1$ y $X_k = 0$ para $i \neq k$, la probabilidad de respuesta positiva para el sujeto i, $P(\boldsymbol{X}_i)$ puede expresarse únicamente en términos del coeficiente de regresión w_i haciendo el despeje

$$g(\boldsymbol{X}_i) = log\left(\frac{P(\boldsymbol{X}_i)}{1 - P(\boldsymbol{X}_i)}\right) = w_i$$

y

$$P(\boldsymbol{X}_i) = \frac{1}{1 + e^{-w_i}}$$

Como el modelo saturado se ajusta a los datos perfectamente entonces por definición $\widehat{Y}_i = \widehat{\pi}_i = Y_i$ para cada sujeto i, se sigue que la fórmula para el estimador máximo verosímil del modelo saturado (4.1) se resuelve fácilmente al reemplazar $P(\boldsymbol{X}_i)$ por Y_i. Si Y_i es una variable binaria que toma los valores 0 y 1 entonces la

expresión $Y_i^{Y_i}[1-Y_i]^{1-Y_i}$ siempre resulta 1; en consecuencia, la máxima verosimilitud para un modelo saturado resulta $L_{MS} = 1$

4.1. Estadístico de devianza D

El estadístico de devianza, por definición, es una medida de cociente de verosimilitudes que compara el modelo propuesto con un modelo referencia que provee un ajuste perfecto, en este caso, un modelo saturado.

La fórmula de devianza es entonces:

$$Dev(\widehat{\beta}) = -2ln(L_e/L_{max})$$

donde L_e y L_{max} son la verosimilitud del modelo propuesto y saturado, respectivamente.

Nótese que en un ajuste perfecto la devianza equivale a cero pues $-2Ln(1) = 0$. Por otro lado, si el modelo no se ajusta adecuadamente, entonces la relación entre L_e y L_{max} es pequeña y el valor del estadístico es relativamente grande.

La comparación de dos modelos no saturados por medio del método del cociente de verosimilitudes con el fin de averiguar cuál de dos modelos candidatos: L_R y L_F, presenta un mejor ajuste, equivale a restar sus estadísticos de devianza, como se muestra:

$$Dev_R(\widehat{\beta}) - Dev_F(\widehat{\beta}) = [-2ln(L_R/L_{max})] - [-2ln(L_F/L_{max})]$$

$$= -2ln(L_R) - [-2ln(L_F)] = -2ln(L_R/L_F).$$

Como se mencionó anteriormente, se sabe que $L_{max} = 1$, por lo que la forma explícita del estadístico de devianza resulta:

$$Dev(\widehat{\beta}) = -2\sum_{i=1}^{n}\left[Y_i ln\left(\frac{Y_i}{\widehat{\pi}(X_i)}\right) + (1-Y_i)ln\left(\frac{1-Y_i}{1-\widehat{\pi}(X_i)}\right)\right]$$

La fórmula anterior encuentra un equivalente luego de algo de álgebra y cálculo, esta segunda fórmula para la devianza es especialmente útil en el computo de la medida pues involucra la respuesta de la regresión logística, y las probabilidades estimadas. La siguiente fórmula hace uso únicamente de las probabilidades estimadas $\widehat{\pi}(X_i)$ de cada individuo, dejando fuera la respuesta observada Y_i. Esto implica que las fórmulas de devianza no aportan ninguna información acerca de la correspondencia entre valores observados y predichos para la variable Y.

$$Dev(\widehat{\beta}) = -2\sum_{i=1}^{n}\left[\widehat{\pi}(X_i)ln\left(\frac{\widehat{\pi}(X_i)}{1-\widehat{\pi}(X_i)}\right) + ln\left(1-\widehat{\pi}(X_i)\right)\right].$$

Existe una fórmula alternativa para el estadístico de devianza que se enfoca en los grupos de sujetos que comparten patrones de covariables y no en los sujetos como unidad de estudio. En esta versión de la devianza, al modelo saturado por grupos se le considera como el modelo de referencia o ajuste perfecto y los individuos se reúnen en patrones de covariables. Si se asume que el modelo tiene J patrones de covariables $X_j = (X_{j1}, X_{j2}, \ldots, X_{jp})$ con n_j sujetos para el patrón j; $\widehat{m}_j = n_j\widehat{\pi}(X_j)$ denota los casos esperados, donde $\widehat{\pi}(X_j)$ es la probabilidad de respuesta positiva $(Y = 1)$ para $X = X_j$ y m_j denota el número observado de respuestas positivas para el sub-grupo j, entonces:

$$Dev_{grupo}(\widehat{\beta}) = -2lnL_{e,grupo} - [-2lnL_{max,grupo}]$$

$$= -2\sum_{j=1}^{J}\left[m_j ln\left(\frac{m_j}{\widehat{m}_j}\right) + (n_j - m_j)ln\left(\frac{n_j - m_j}{n_j - \widehat{m}_j}\right)\right]$$

Donde $L_{e,grupo}$ denota función de verosimilitud para el modelo propuesto considerando cada patrón de covariables como un sujeto y $L_{max,grupo}$ la verosimilitud del modelo saturado por grupos.

Resulta trivial que las versiones de la devianza para grupos e individuos, no resultan equivalentes a menos que la cantidad de patrones de covariables sea igual al número de sujetos en los datos. Sin embargo, estas dos cantidades difieren únicamente por una constante, K, que cumple

$$-2lnL_{e,grupo} = -2lnL_e - 2k$$

donde

$$K = ln\left[\sum_{j=1}^{J}\frac{n_j!}{m_j!(n_j - m_j)!}\right]$$

El cálculo de la constante, K, no involucra al estimado paramétrico, $\widehat{\beta}$, pues su cálculo resulta igual, sin importar si se toma como unidad de estudio a los individuos o a los grupos.

Cuando el número de patrones de covariables J es considerablemente menor a n, el estadístico de devianza por grupos se asume con una distribución aproximada a una χ^2 con $J-p-1$ grados de libertad. Dicha aproximación es útil al hacer pruebas de hipótesis respecto a la devianza donde la hipótesis nula, H_0, es que el modelo

presenta un buen ajuste y la hipótesis alternativa, H_a, es que el modelo no ajusta satisfactoriamente. Si J es cercana a n, entonces la aproximación es cuestionable; ya que el modelo de Regresión Logística admite variables continuas, la igualdad de J y de n resultan un problema frecuente cuando únicamente se conoce esta técnica para realizar una estimación de la bondad de ajuste del modelo.

4.2. Estadística Ji-cuadrada de Pearson χ^2

La estadistica Ji-Cuadrada de Pearson compara frecuencias observadas y esperadas internamente para cada patrón de covariables. A cada comparación se le conoce como **residuo de Pearson** y se define como

$$r_j = \frac{m_j - n_j\widehat{\pi}_j}{\sqrt{n_j\widehat{\pi}_j(1-\widehat{\pi}_j)}}$$

para el patrón de covariables j.

La estadística Ji-cuadrada de Pearson se calcula como

$$\chi^2 = \sum_{j=1}^{J} r_j^2$$

y sigue una distribución igual a la del estadístico de Devianza ($\chi^2(J-p-1)$).

4.3. Estadística de Hosmer-Lemeshow $\widehat{C}$

Hosmer y Lemeshow (1980) propusieron un estadístico de bondad de ajuste para el modelo de regresión logística que se basa en la agrupación de los sujetos por probabilidades estimadas, $\widehat{\pi}(X_i)$. El **estadístico de Hosmer-Lemeshow**, $\widehat{\boldsymbol{C}}$, se creó para evitar el uso del estadístico de devianza en casos donde su aproximación a una distribución conocida (χ^2) resulta cuestionable. La implementación de esta medida de bondad de ajuste requiere que el modelo considere al menos tres patrones de covariables, $J \geq 3$; rara vez resulta significativo si la cantidad de patrones es menor a seis, $J < 6$, y funciona mejor si J se aproxima al número total de sujetos, n.

La estadística $\widehat{C}$ tiene la propiedad de que siempre es cero para un modelo completamente parametrizado, en otras palabras, el modelo de referencia de un ajuste perfecto es un modelo saturado por grupos.

La idea general es agrupar las probabilidades estimadas por el modelo logístico, $\widehat{\pi}_1, \ldots, \widehat{\pi}_n$, según su tamaño, definiendo puntos de corte para cada grupo correspondientes a Q percentiles. El número de respuestas positivas y negativas observadas,

en el grupo percentil q, son

$$O_{1q} = \sum_{i=1}^{n_q} Y_i$$

y

$$O_{0q} = \sum_{i=1}^{n_q} (1 - Y_i)$$

respectivamente, donde n_q es la cantidad de individuos en el grupo percentil q.

El número de respuestas positivas y negativas esperadas, por sub-grupo percentil q, son

$$E_{1q} = \sum_{i=1}^{n_q} \widehat{\pi}(X_{iq})$$

y

$$E_{0q} = \sum_{i=1}^{n_q} (1 - \widehat{\pi}(X_{iq}))$$

respectivamente, donde n_q es la cantidad de individuos en el grupo percentil q y X_{iq} son los valores para las covariables del i-ésimo sujeto en el mismo grupo.

El estadístico $\widehat{C}$ se calcula comparando los valores observados y esperados por grupo percentil de manera que

$$\widehat{C} = \sum_{q=1}^{Q} \frac{(O_{1q} - E_{1q})^2}{E_{1q}} + \sum_{q=1}^{Q} \frac{(O_{0q} - E_{0q})^2}{E_{0q}}.$$

Se demostró por simulaciones extensivas que el estadístico $\widehat{C}$ tiene una distribución aproximada $\chi^2(Q-2)$.

4.4. Diagnósticos

Las técnicas para comprobar la bondad de ajuste de un modelo, como la devianza y la estadística Ji-cuadrada, se basan en medidas que resumen la información otorgada por todos los sujetos de una muestra. Una desventaja de las medidas de bondad de ajuste es que carecen de un análisis individual de los patrones de covariables. Por lo tanto, antes de concluir que un modelo se ajusta a un conjunto de datos, se deben examinar otras medidas para saber si todo el rango de patrones de covariables, de forma individual, otorgan evidencia del ajuste. A dichas medidas se les llama *diagnósticos de regresión.*

Antes de introducir los diagnósticos utilizados en regresión logística se deben recordar dos cantidades necesarias para el cálculo de las medidas de diagnostico: los residuos de devianza y de Pearson. Ambas cantidades representan la comparación

entre la respuesta observada y estimada, por un patrón de covariables, bajo un esquema específico.

El residuos de devianza para el patrón de covariables X_j es:

$$d_j = \pm \left\{ 2 \left[m_j ln \left(\frac{m_j}{\widehat{m}_j} \right) + (n_j - m_j) ln \left(\frac{n_j - m_j}{n_j - \widehat{m}_j} \right) \right] \right\},$$

donde m_j es la cantidad de respuestas positivas ($Y = 1$) del patrón de covariables X_j, $\widehat{m}_j = n_j \widehat{\pi}(X_j)$, n_j la cantidad de individuos con patrón de covariables X_j y $\pm$ es el signo de $(m_j - \widehat{m}_j)$.

El residuo de Pearson para el patrón de covariables X_j es:

$$r_j = \frac{m_j - n_j \widehat{\pi}_j}{\sqrt{n_j \widehat{\pi}_j (1 - \widehat{\pi}_j)}}.$$

Tanto la estadística de bondad de ajuste de la devianza como la Ji-Cuadrada se obtienen por medio de los residuos de devianza y de Pearson pues

$$\chi^2 = \sum_{j=1}^{J} r_j^2$$

y

$$Dev = \sum_{j=1}^{J} d_j^2.$$

Además de los residuales, otras cantidades centrales para el cálculo de las medidas de diagnóstico son la *matriz sombrero* y los *valores de apalancamiento*. En regresión lineal, la matriz sombrero provee los valores ajustados de $\widehat{Y}$ por medio de una proyección de la variable independiente, Y, sobre el espacio de covariables. Sea X, que denota a la matriz de dimensión $J \times (p+1)$ que contiene los J patrones de covariables formados por los valores de p covariables y donde la primera columna son unos, denotando un intercepto en el modelo. En regresión lineal, la matriz sombrero es $\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$, y claramente, $\widehat{\mathbf{y}} = \mathbf{Hy}$. Los residuos de la regresión lineal, $(\mathbf{y} - \widehat{\mathbf{y}})$, expresados en términos de la matriz sombrero son $(\mathbf{I} - \mathbf{H})\mathbf{y}$, donde $\mathbf{I}$ es la matriz identidad de dimensión $J \times J$.

Pregibon (1981), usando una aproximación lineal de los valores ajustados de la regresión logística, deriva una matriz sombrero para la regresión logística. Esta matriz es

$$\mathbf{H} = V^{1/2}\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}V^{1/2}, \tag{4.2}$$

donde $\mathbf{V}$ es la matriz diagonal de dimensión $J \times J$ con elemento general

$$v_j = n_j \widehat{\pi}(X_j) \left[1 - \widehat{\pi}(X_j) \right].$$

En regresión lineal, los valores de la diagonal de la matriz **H** son llamados valores de apalancamiento y representan la distancia entre un valor de covariables, X_j, y la media de los datos, $\bar{X}$. La extensión del concepto de *distancia de la media*, a la regresión logística, requiere una discusión adicional.

Sea h_j el j-ésimo elemento diagonal de la matriz **H** definida en (4.2), es decir, los valores de apalancamiento de la regresión logística. Se puede demostrar que

$$\begin{aligned} h_j &= n_j \widehat{\pi}(X_j)\left[1-\widehat{\pi}(X_j)\right] X_j'(\mathbf{X'VX})^{-1}X_j \\ &= v_j \times b_j \end{aligned}$$

donde $v_j = \widehat{\pi}(X_j)\left[1-\widehat{\pi}(X_j)\right]$ es la varianza estimada de Y_j y

$$b_j = X_j'(\mathbf{X'VX})^{-1}X_j$$

es la distancia ponderada de X_j al vector de medias de los datos $\tilde{X}$. Al interpretar la magnitud de los valores de apalancamiento se debe de tener presente el efecto combinado de v_j y b_j en el apalancamiento h_j. Pregibon (1981) afirma que los puntos con valores de apalancamiento son puntos extremos en el espacio de covariables y, por lo tanto, son lejanos a la media. Lesaffre (1986) refuta lo anterior y señala que el valor de v_j no puede ser ignorado. Ambos puntos de vista son correctos. Aunque la distancia a la media de los patrones de covariables con una probabilidad estimada cercana a 0 y 1 es grande, su apalancamiento tiende a 0. En consecuencia, para interpretar correctamente el apalancamiento, se debe saber si la probabilidad estimada es pequeña (< 0.1) o grande (> 0.9). Si la probabilidad estimada está en el intervalo $(0.1, 0.9)$, el valor del apalancamiento puede ser visto como una distancia. En caso contrario, el valor del apalancamiento no mide una distancia.

En general, la varianza de los residuos de Pearson no es igual a 1 cuando no son estandarizados. Si se denota al residuo de Pearson del patrón de covariables X_j como r_j, el residuo de Pearson estandarizado es

$$r_{sj} = \frac{r_j}{\sqrt{1-h_j}}.$$

Otros diagnósticos importantes son los que examinan el efecto que tiene eliminar todos los sujetos con un patrón de covariables en particular, en los coeficientes estimados, y en las medidas de bondad de ajuste χ^2 y D.

El cambio en los coeficientes estimados, $\Delta\widehat{\boldsymbol{\beta}}_j$, se obtiene vía la diferencia estandarizada entre $\widehat{\boldsymbol{\beta}}$ y $\widehat{\boldsymbol{\beta}}_{(-j)}$, que representan el vector de coeficientes estimados y el vector de coeficientes estimados excluyendo a los sujetos con patrón de covariable X_j, respectivamente. La diferencia anterior es estandarizada por la matriz de covarianzas de $\widehat{\boldsymbol{\beta}}$. Pregibon (1981) muesta que una aproximación lineal de la diferencia estandarizada es

$$\begin{aligned}\Delta\widehat{\boldsymbol{\beta}}_j &= (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-j)})'(\mathbf{X}'\mathbf{V}\mathbf{X})(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-j)}) \\ &= \frac{r_j^2 h_j}{(1-h_j)^2} \\ &= \frac{r_{sj}^2 h_j}{(1-h_j)}.\end{aligned} \tag{4.3}$$

Para encontrar el cambio en la estadística Ji-cuadrada de Pearson, $\Delta\boldsymbol{\chi}_j^2$, se utiliza una aproximación lineal similar:

$$\begin{aligned}\Delta\boldsymbol{\chi}_j^2 &= \frac{r_j^2}{(1-h_j)} \\ &= r_{sj}^2\end{aligned} \tag{4.4}$$

Análogamente, el cambio en la devianza, $\Delta\boldsymbol{D}_j$, resulta

$$\Delta\boldsymbol{D}_j = d_j^2 + \frac{r_j^2 h_j}{(1-h_j)}$$

y reemplazando r_j^2 por d_j^2 se obtiene una forma similar a la ecuación (4.4)

$$\Delta\boldsymbol{D}_j = \frac{d_j^2}{(1-h_j)}. \tag{4.5}$$

Los diagnósticos de residuos son deseables pues ayudan a encontrar los patrones de covariables que no se ajustan adecuadamente ($\Delta\boldsymbol{\chi}_j^2$,$\Delta\boldsymbol{D}_j$) y aquellos que ejercen una influencia importante en los coeficientes estimados ($\Delta\widehat{\boldsymbol{\beta}}_j$). Los diagnósticos, en la mayoría de los casos, siguen un comportamiento que se puede esbozar. El cambio en la estadística Ji-cuadrada, $\Delta\boldsymbol{\chi}_j^2$, es menor cuando y_j se acerca a $m_j\widehat{\pi}(X_j)$. Esto ocurre si se cumple que $y_j = 0$ y $\widehat{\pi}(X_j) < 0.1$, ó, $y_j = m_j$ y $\widehat{\pi}(X_j) > 0.9$. Análogamente, $\Delta\boldsymbol{\chi}_j^2$ es grande cuando y_j se aleja de $m_j\widehat{\pi}(X_j)$, lo que ocurre cuando $y_j = 0$ y $\widehat{\pi}(X_j) > 0.9$, o, $y_j = m_j$ y $\widehat{\pi}(X_j) < 0.1$. En estos casos el cambio en los coeficientes estimados, $\Delta\widehat{\boldsymbol{\beta}}_j$ es pequeño pues $\Delta\widehat{\boldsymbol{\beta}}_j$ se aproxima a $\Delta\boldsymbol{\chi}_j^2 h_j$ y el apalancamiento, h_j, se acerca a 0. $\Delta\widehat{\boldsymbol{\beta}}_j$ es grande cuando $\Delta\boldsymbol{\chi}_j^2$ y h_j son al menos moderados. Estos valores son hallados cuando $0.1 < \widehat{\pi}(X_j) < 0.3$, o $0.7 < \widehat{\pi}(X_j) < 0.9$. También, son los intervalos donde h_j es mayor. Cabe mencionar que el comportamiento general de los diagnósticos puede no ser seguido en todos los casos y sólo debe ser tomado como una guía para la interpretación de los diagnósticos.

La interpretación de diagnósticos de la regresión logística es principalmente visual. A diferencia de la regresión lineal, en que se pueden obtener la distribución de los diagnósticos, en la regresión logística, la distribución de los diagnósticos, bajo la hipótesis de que el modelo se ajusta, sólo se conoce para algunos casos particulares.

Martin y Pardo (2009) sugieren utilizar $(2p/n)$ como el valor crítico para los valores de apalancamiento, el percentíl $\chi^2_{(0.05)}(p+1)$ para el diagnóstico $\Delta\chi^2$, y $\overline{hh} \times \chi^2_{(0.95)}(1)$ para $\Delta\widehat{\beta}$, donde $\overline{hh}$ es el promedio de los J valores de $\frac{h_j}{1-h_j}$. Hosmer y Lemeshow (2013) encuentran demasiado rigurosos los anteriores valores extremos.

Existen muchas gráficas recomendadas para la interpretación de diagnósticos pero resulta impráctico revisarlas todas. Los siguientes ejemplos se consideran la base del análisis de diagnósticos por su facilidad de obtener y su relevancia en la mayoría de las regresiones:

1. h_j contra $\widehat{\pi}_j$
2. $\Delta\chi^2_j$ contra $\widehat{\pi}_j$
3. ΔD_j contra $\widehat{\pi}_j$
4. $\Delta\widehat{\beta}_j$ contra $\widehat{\pi}_j$

Las gráficas $\Delta\chi^2_j$ contra $\widehat{\pi}_j$ y ΔD_j contra $\widehat{\pi}_j$ son similares y muestran curvas tipo cuadráticas. En cada gráfica es normal observar dos curvas, una creciente que corresponde a los patrones de covariables con $y_j = 0$, y otra decreciente que corresponde a los patrones con $y_j = 1$. Los puntos con mal ajuste típicamente se observan en las esquinas superiores de la gráfica. La interpretación de las gráficas es parcialmente visual y parcialmente numérica. Cuando m_j es grande las distribuciones de $\Delta\chi^2_j$ y ΔD_j se aproximan a $\chi^2(1)$ por lo que se utiliza 4 como una aproximación del 95 percentil $(\chi^2_{0.95}(1) = 3.84)$ de la distribución.

Otras gráficas frecuentemente utilizadas son:

1. $\Delta\chi^2_j$ contra h_j
2. ΔD_j contra h_j
3. $\Delta\widehat{\beta}_j$ contra h_j

Las últimas tres gráficas son muy utilizadas porque permiten evaluar directamente la contribución del apalancamiento en los diagnósticos. Por último, otra gráfica que resulta especialmente útil es $\Delta\chi^2_j$ contra $\widehat{\pi}_j$ con el tamaño del símbolo proporcional a $\Delta\widehat{\beta}_j$.

Tras encontrar patrones de covariables con ajustes pobres o con gran influencia en los coeficientes, se debe analizar su efecto en los coeficientes individuales. Usualmente, luego de borrar los patrones de covariables individualmente, se realiza el mismo proceso por grupos. Se borran todos los patrones de covariables con ajustes pobres, se borran los patrones de covariables con influencia en los coeficientes

estimados y, por último, se borran todos los patrones de covariables identificados. Finalmente, con ayuda de un experto en la materia de estudio, se debe determinar el rol que juegan los patrones de covariables identificados en el modelo final.

4.5. Desempeño discriminatorio

Una manera intuitiva de resumir los resultados del ajuste de un modelo logístico es vía una **tabla de clasificación cruzada**. Esta tabla resulta de concatenar las observaciones de la respuesta, Y, con una variable que clasifica las probabilidades estimadas por el modelo en dos grupos, 0 y 1. En este enfoque, las $\pi(\boldsymbol{x})$ sirven para predecir una respuesta binaria, más que para estimar un riesgo. La variable de clasificación, denotada $\tilde{Y}$, se obtiene a través de las $\pi(\boldsymbol{x})$. Primero, es necesario definir un punto de corte, c, y comparar cada probabilidad estimada contra c. Si la probabilidad estimada excede c entonces la variable de clasificación es igual a 1; en caso contrario, es igual a 0.

En regresión logística es posible utilizar la variable $\tilde{Y}$ como una predicción de la respuesta, Y, del modelo. Por ejemplo, a partir de estudios clínicos de la sangre y el ajuste un modelo logístico se podría predecir que si la probabilidad estimada de algún sujeto, $\pi(\boldsymbol{x})$, supera 0.7, entonces ese sujeto desarrollaría leucemia.

Cuadro 4.1: Tabla de clasificación cruzada 2 x 2

		Respuesta Observada	
		$Y = 1$	$Y = 0$
Respuesta	$\tilde{Y} = 1$	n_{VP}	n_{FP}
Predicha	$\tilde{Y} = 0$	n_{FN}	n_{VN}
		n_1	n_0

El cuadro 4.1 muestra gráficamente la distribución de una tabla de clasificación que combina las observaciones de Y y una cierta variable de clasificación o predicción $\tilde{Y}$ inducida por un punto de corte específico. En la tabla anterior resaltan dos cantidades: el número de **verdaderos positivos** (VP), o el conteo de respuestas positivas ($R+$) predichas correctamente; y el número **verdaderos negativos** (VN), las respuestas negativas ($R-$) observadas que fueron predichas correctamente.

A la proporción de VP dentro de todas las respuestas positivas observadas, denotado como n_1, se le conoce como **sensibilidad** (**Se**). Análogamente, la **especificidad** (**Sp**) es la proporción de VN dentro de las respuestas negativas observadas, n_0. Resumiendo

$$Se = \frac{n_{VP}}{n_1}$$

y

$$Sp = \frac{n_{VN}}{n_0}.$$

La Se y Sp pueden ser interpretadas como la proporción de predicciones correctas para cada tipo de respuesta, inducidas por un punto de corte específico. Entonces, entre más cercano a 1 sean ambas medidas, mayor será la capacidad discriminatoria o predictiva del modelo. Por ejemplo, para un punto de corte, $c = 0.2$; entre un modelo A, con $Se = 0.7$ y $Sp = 0.8$ y un modelo B, con $Se = 0.6$ y $Sp = 0.5$, se dice que el modelo A, presenta mejor desempeño discriminatorio que el modelo B.

Una desventaja del análisis del rendimiento discriminatorio de un modelo logístico a través de la sensibilidad y especificidad es que estas dos medidas están sujetas a un punto de corte fijo, y varían según su elección.

Cuando se selecciona un valor alto como punto de corte, la clasificación más rigurosa para las respuestas positivas resulta en una sensibilidad más baja; en cambio, la especificidad aumenta. Cuando el punto de corte es bajo, se predicen más respuestas positivas y la sensibilidad aumenta; por otro lado, la especificidad disminuye.

Por ejemplo, para un modelo logístico correspondiente a un estudio médico, se eligen dos puntos de corte, $c_1 = 0.2$ y $c_2 = 0.6$. Las tablas de clasificación 4.2 y 4.3 resultan de la discriminación de c_1 y c_2 respectivamente.

Cuadro 4.2: Tabla de clasificación cruzada: punto de corte 0.2

		Respuesta	**Observada**
		$Y = 1$	$Y = 0$
Respuesta	$\tilde{Y} = 1$	92	170
Predicha	$\tilde{Y} = 0$	28	210
		120	380

Cuadro 4.3: Tabla de clasificación cruzada: punto de corte 0.6

		Respuesta	**Observada**
		$Y = 1$	$Y = 0$
Respuesta	$\tilde{Y} = 1$	4	373
Predicha	$\tilde{Y} = 0$	116	7
		120	380

Entonces, la sensibilidad y especificidad de cada caso son $c_1 : Se = \frac{92}{120} = 76.6\,\%$ y $Sp = \frac{170}{380} = 44.73\,\%$; $c_2 : Se = \frac{4}{120} = 3.33\,\%$ y $Sp = \frac{373}{380} = 98.15\,\%$.

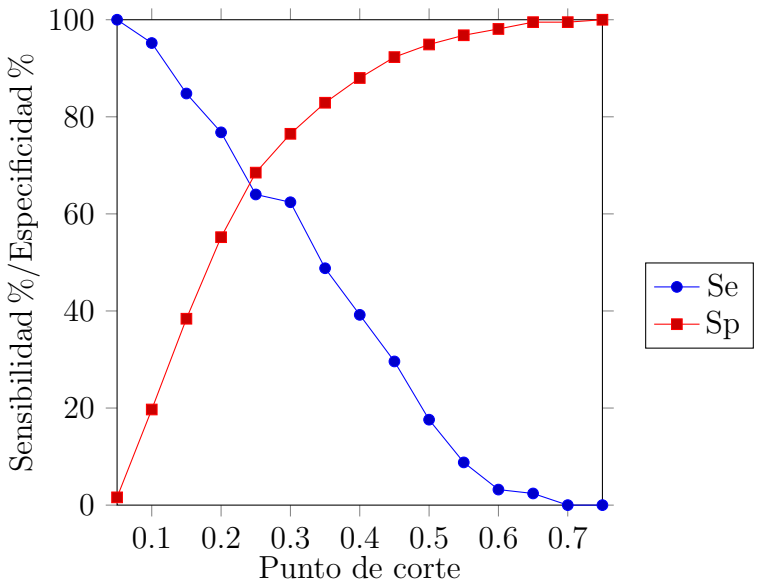

Figura 4.1: Gráfica de sensibilidad y especificidad contra todos los posibles puntos de corte en un estudio médico

Punto de corte óptimo

En el conjunto de puntos de corte posibles para el ajuste de una regresión logística, C, es posible encontrar un punto que maximiza la suma de la sensibilidad y especificidad. A este punto se le llama **punto de corte óptimo** y se denota como c^*. Si Se^c y Sp^c son la sensibilidad y especificidad inducidas por el corte c, entonces, c^* es

$$c^* = \max_C \{Se^c + Sp^c\}.$$

Un acercamiento alternativo al análisis de tablas de clasificación es el uso de una medida de resumen que incorpore todos los puntos de corte posibles para un modelo dado. Dicha medida se obtiene a través de una gráfica conocida como la curva ROC.

4.5.1. Curvas ROC

Una curva **Característica Operativa del Receptor** (o **ROC** por sus siglas en inglés) es la gráfica de la sensibilidad contra 1−especificidad, para los puntos de corte posibles de un modelo logístico. Aunque su origen se deriva de la teoría de detección de señales, en el contexto de regresión logística, las curvas ROC proveen un estimado de qué tan preciso es un modelo para distinguir entre respuestas observadas y predichas; o sea, cómo es el desempeño predictivo de un modelo.

Se debe notar que 1−**especificidad** es la proporción de respuestas, $(Y = 0)$, observadas predichas erróneamente. De forma alternativa, 1−Sp es la proporción de

falsos positivos (FP) y se calcula como

$$1 - Sp = \frac{n_{FP}}{n_0}.$$

Teniendo en cuenta que la aproximación de Sp a 1 indica una buena discriminación, se busca que 1−Sp sea cercano a 0; y más aún, en un modelo con capacidad predictiva se espera que la sensibilidad sea mayor a 1−especificidad para todos los puntos de corte. Lo anterior es una consecuencia de que en un modelo que provee una buena discriminación, se supone que en general, las respuestas positivas tengan una probabilidad estimada mayor a las respuestas negativas. En la curva ROC, esta propiedad se refleja en que la gráfica de la curva debe estar por encima de la función identidad, que es la representación de $Se = 1 - Sp$. La diagonal designada por la función identidad sirve de referencia para expresar qué se espera de un modelo el cual sus probabilidades estimadas no son informativas. Entre mayor sea la diferencia entre Se y 1−Sp, mayor será la capacidad predictiva del modelo, y mayor será el área bajo la curva ROC. De esta manera, el **área bajo la curva ROC (AUC)** es una medida que resume la sensibilidad y especificidad de un modelo logístico a través de todos los puntos de corte posibles.

En la construcción de cualquier curva ROC para un modelo logístico, se debe notar que las $\pi(\boldsymbol{x})$ inducen los puntos de corte posibles; asimismo, como las $\pi(\boldsymbol{x})$ son calculadas a través de cada patrón de covariables, la cantidad de puntos de corte posibles es igual a la cantidad de patrones de covariables de un modelo.

Por ejemplo, considérese un modelo logístico que incluye dos covariables dicotómicas, X_1, X_2, ambas con codificación $(0, 1)$. Este modelo sólo admite cuatro patrones de covariables y por lo tanto cuatro puntos de corte posibles, además de un quinto definido por $\pi(\boldsymbol{x}) = 1$ donde ninguna respuesta se predice como positiva pues $\pi(\boldsymbol{x}) \leq 1$ siempre. En la tabla 4.4 se muestran los puntos de corte en forma de probabilidades estimadas, la sensibilidad y la especificidad para el ejemplo anterior. La figura 4.2 resulta de graficar la sensibilidad contra la especificidad de los datos de la tabla 4.4; es decir, es la curva ROC del modelo anterior.

4.5.2. Área bajo la curva ROC (AUC)

La curva ROC es un método gráfico útil al establecer la capacidad predictiva de un modelo logístico; sin embargo, es necesario una medida que resuma los resultados proporcionados por dicha curva. La medida utilizada más comúnmente para este propósito es el **área bajo la curva ROC (AUC)**.

Para calcular el ABC se utiliza el método **trapezoidal**. Este método consiste en

Cuadro 4.4: Sensibilidad y 1-especificidad por patrón de covariable

$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	**Punto de Corte**	$\boldsymbol{\pi(x)}$	**Sensibilidad %**	**1-Especificidad %**
1	1	c_0	0.75	0	0
1	0	c_1	0.60	10	1
0	1	c_2	0.45	60	25
0	0	c_3	0.20	80	50
-	-	c_4	0.0	100	100

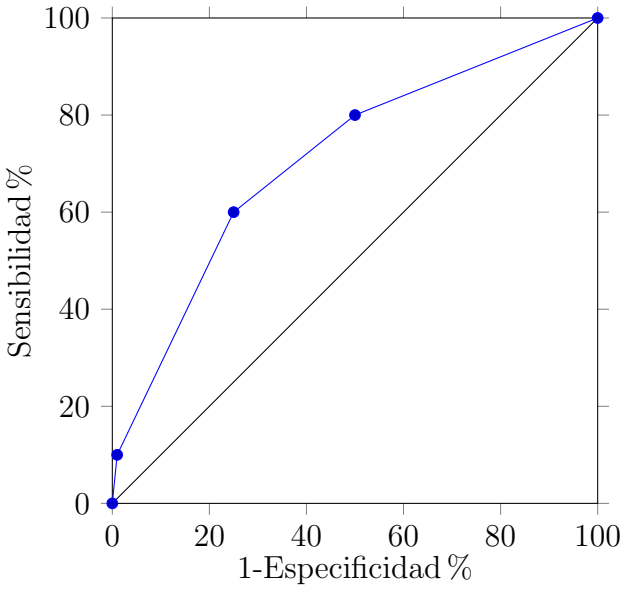

Figura 4.2: Curva ROC

formar trapezoides de esquinas definidas por puntos consecutivos de la curva ROC, denotados como $t_1, \ldots, t_i$. El área de cada trapezoide se calcula con las coordenadas de las esquinas. Finalmente, se suman las áreas de los trapezoides.

Por ejemplo, para calcular el ABC de la curva ROC definida por la tabla 4.4, se construyen cuatro trapezoides (fig. 4.3), t_1, t_2, t_3, t_4. Las coordenadas de las esquinas de los trapezoides están dadas por la sensibilidad y $1-$especificidad registrada para puntos de corte sucesivos. Si denotamos a la sensibilidad registrada para el punto de corte c_i como Se_i, y a $1-$ especificidad registrada para el mismo punto de corte como $1 - Sp_i$, entonces, el área de t_i se calcula

$$A(t_i) = \frac{((1 - Sp_i) - (1 - Sp_{i-1})) \times ((Se_i) + (Se_{i-1}))}{2}.$$

Entonces, para el ejemplo anterior

$$\begin{aligned} ABC &= t_1 + t_2 + t_3 + t_4 \\ &= \frac{(0.01-0)(0.1+0)}{2} + \frac{(0.25-0.01)(0.6+0.1)}{2} \\ &+ \frac{(0.5-0.25)(0.8+0.6)}{2} + \frac{(1-0.5)(1+0.8)}{2} \\ &= 0.7095. \end{aligned}$$

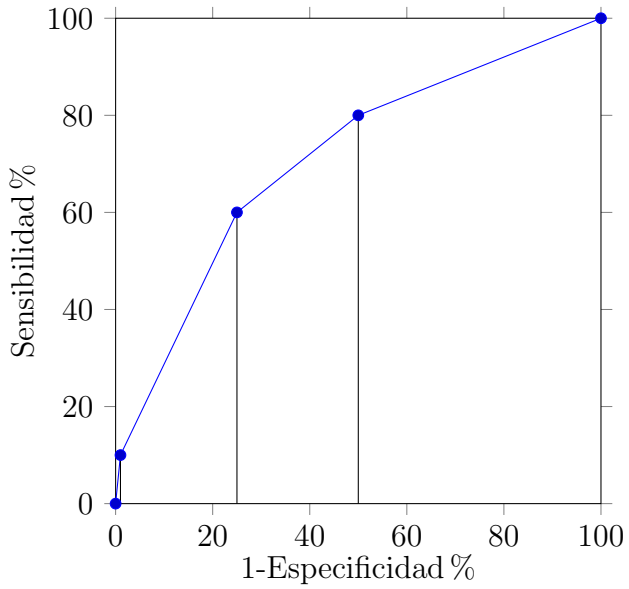

Figura 4.3: División de la curva ROC 2.1 en áreas trapezoidales

Una interpretación alternativa del significado del ABC relaciona la probabilidad estimada de los sujetos con respuesta positiva ($R+$) y negativa ($R-$). En un modelo con capacidad predictiva se espera que, en general, la probabilidad estimada de un sujeto con $R+$ sea mayor a la de un sujeto con $R-$. Si la condición anterior se cumple, esa pareja de sujetos es llamada **pareja concordante**. En caso de la probabilidad estimada de un sujeto con $R-$ sea mayor a la de un sujeto con $R+$ es llamada **pareja discordante**. También existen las **parejas empatadas**, que son parejas de sujetos con respuesta opuesta y probabilidad estimada igual.

El ABC se calcula de forma alternativa a través de la proporción de parejas concordantes y empatadas dentro de todas las posibles parejas de sujetos con respuesta opuesta; es decir, si se denota al número de parejas de sujetos con estas características como $n_p = n_1 \times n_2$, el número de parejas concordantes como n_{pc} y el número de parejas empatadas como n_{pe}, entonces

$$ABC = \frac{n_{pc} \times (0.5) n_{pe}}{n_p} \tag{4.6}$$

Debido a que la curva ROC se construye interpolando linealmente las coordenadas de sensibilidad y 1-especificidad consecutivas, la formula (4.6) sólo se incluye la mitad de las parejas empatadas. Por lo anterior, se agrega un factor de 0.5 a n_{pe} en la formula anterior.

Para ejemplificar, se toman los datos de la tabla 4.5. En esta tabla se muestra la cantidad de $R+$ y $R-$ registrados por cada punto de corte en una regresión logística. Si se suman los valores de la columna $\boldsymbol{Y} = 1$ se obtiene $n_1 = 10+50+20+20 = 100$. Análogamente, para la columna $\boldsymbol{Y} = 0$ se obtiene $n_0 = 200$.

Cuadro 4.5: Número de $R+$ y $R-$ por cada punto de corte

Punto de Corte	$\pi(\boldsymbol{x})$	**Sensibilidad**	**1-Especificidad**	$\boldsymbol{Y} = 1$	$\boldsymbol{Y} = 0$
c_0	0.75	0	0	0	0
c_1	0.60	10	1	10	2
c_2	0.45	60	25	50	48
c_3	0.20	80	50	20	50
c_4	0.0	100	100	20	100

Para c_1 se tienen $10 \times 2 = 20$ parejas empatadas, pero las mismas 10 $R+$ resultan mayores a $48+50+50 = 198$ $R-$ que resultan en $10 \times 198 = 1980$ parejas concordantes. Para c_2 se tienen $50 \times 48 = 2400$ parejas empatadas y $50 \times 150 = 7500$ parejas concordantes. Para c_3 hay $20 \times 50 = 1000$ parejas empatadas y $20 \times 100 = 2000$ parejas concordantes. Entonces $n_{pe} = 20 + 2400 + 100 + 200 = 5420$ y $n_{pc} = 1980 + 7500 + 2000 = 11480$. El número total de parejas con respuestas opuestas es $n_p = 100 \times 200 = 20000$. Por lo tanto, para este modelo

$$ABC = \frac{11480 + (0.5)5420}{20,000} = 0.7095.$$

La interpretación del ABC es empírica y depende de los requerimientos del estudio en que se emplee esta técnica. Idealmente, en un modelo que presenta una discriminación perfecta, la medida del ABC resulta 1; sin embargo, incluso un $ABC = 0.90$ implica una separación casi completa de los datos. De ser así, el ajuste de un modelo logístico es innecesario. Un $ABC = 0.50$ indica que un modelo no provee discriminación. También es posible que el ABC sea menor que 0.50. Esto indica una discriminación negativa, es decir, el modelo predice mejor las respuestas negativas. Para otorgar un criterio de calificación al ABC se considera lo siguiente. Un ABC de $0.9 - 1.0$ equivale a una discriminación **excelente**. De $0.8 - 0.9$ equivale a una discriminación **buena**, de $0.7 - 0.8$ una discriminación **aceptable**, de $0.6 - 0.7$ una discriminación **pobre**, y de $0.5 - 0.6$ se considera que el modelo **falla** al discriminar respuestas.

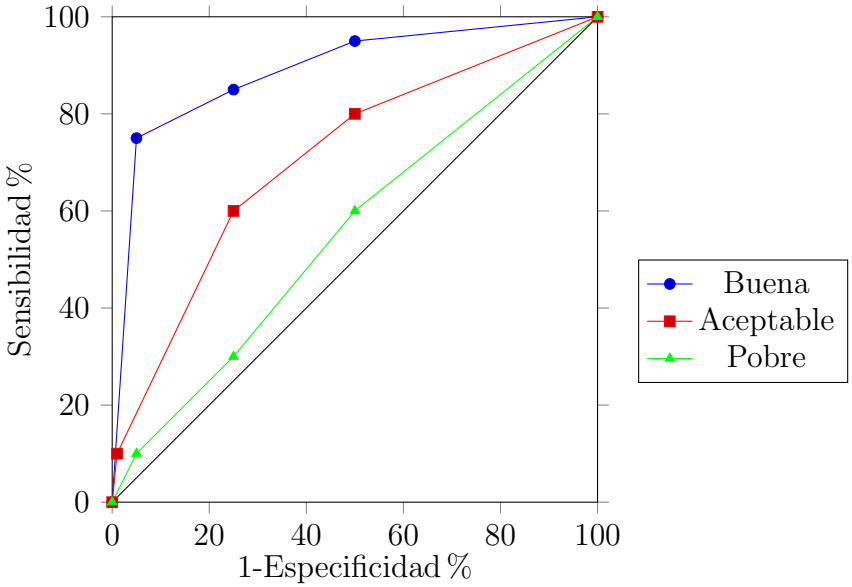

Figura 4.4: Calificación del desempeño discriminatorio de cada curva ROC.

Capítulo 5

Aplicación con datos reales

5.1. Introducción al trastorno por déficit de atención con hiperactividad

El trastorno por déficit de atención con hiperactividad (TDAH) es un trastorno del neurodesarrollo que comúnmente inicia antes de los 12 años de edad. El TDAH es un trastorno altamente prevalente ya que incluyendo niños, adolescentes y adultos, afecta alrededor del 6 de la población mundial [16]. La población afectada se caracteriza por presentar niveles de inatención, hiperactividad e impulsividad mayores a los esperados en su grupo de pares de la misma edad y sexo. Los individuos que presentan este trastorno, frecuentemente muestran afectación en aspectos asociados con su desempeño general, los cuales incluyen el área académica, familiar y social [5].

Se ha observado que el TDAH presenta una relación amplia con otros trastornos psiquiátricos y padecimientos médicos en general. En el área médica, a la presencia de más de una condición distinta en una misma persona se le conoce como **comorbilidad**. El TDAH se cataloga como un trastorno altamente comórbido ya que al menos el 75 % de sujetos con TDAH presentan otro trastorno psiquiátrico o médico. En infantes, los trastornos psiquiátricos comórbidos con el TDAH incluyen los trastornos de conducta, oposicionista desafiante, del estado de ánimo, de ansiedad y del aprendizaje [23]. En adultos, la prevalencia de trastornos de ansiedad comórbidos alcanza el 50 %. Asimismo, los trastornos del estado de ánimo, trastorno antisocial o el abuso de alcohol y sustancias llegan a tasas altas de prevalencia [8]. Los adultos diagnosticados en su infancia muestran mayor probabilidad de presentar invalidez psicosocial, comorbilidades psiquiátricas y falla escolar. También se han observado fuertes asociaciones del TDAH con problemas de aprendizaje y del neurodesarrollo, incluyendo discapacidad para leer, problemas en el discurso y en el lenguaje,

dificultades motoras y bajo coeficiente intelectual. Los adolescentes con este trastorno presentan mayor riesgo de mostrar un desempeño académico deficiente, baja autoestima, problemas en sus relaciones interpersonales, conflictos con sus padres, delincuencia, uso de alcohol, tabaco y otras sustancias ilícitas [5].

El sexo desempeña un papel importante al analizar este trastorno; se ha observado que la proporción de hombres con relación a mujeres que presentan TDAH en muestras comunitarias es aproximadamente de dos a uno, y en muestras clínicas supera los nueve a uno [6].

En cuanto a los factores relacionados con su etiología o sus causas, al ser un trastorno complejo en donde interactúan aspectos biológicos y medioambientales, se estima que el TDAH presenta un coeficiente de heredabilidad de alrededor del 76 %, o sea, el 76 % de la varianza relacionada con este trastorno se explica por el factor genético [27]. Por otro lado, las comorbilidades asociadas al TDAH pueden o no compartir un fuerte vínculo familiar o incluso genético. Por ejemplo, el TDAH y el trastorno depresivo mayor se asemejan en sus vulnerabilidades familiares mientras que el TDAH es independiente a la heredabilidad del trastorno por ansiedad [3].

Se ha demostrado que ciertos factores, además de los biológicos, influyen en la manifestación y persistencia del trastorno. Al conjunto de estos factores se le conoce como factores de**adversidad psicosocial**. Entre estos factores se incluyen: psicopatología del padre o la madre, problemas legales de los padres, nivel socioeconómico, disfunción familiar, nivel educativo de los padres, entre otros. Diversos estudios muestran que los sujetos con este trastorno están más expuestos a eventos estresantes y a factores relacionados con eventos de adversidad psicosocial que la población ordinaria [7]. En general, el TDAH parece ser un marcador de mal pronóstico para el nivel de funcionamiento encontrado en individuos afectados.

Como se menciona anteriormente, es imposible ignorar la contribución medioambiental en los modelos etiológicos del TDAH. En estudios familiares de este trastorno, se pueden observar fácilmente las similitudes de los factores externos que afectan a los integrantes de cada familia, por ejemplo, las similitudes entre hermanos. Estas similitudes o **medio ambiente compartido** resultan obvias pues son una consecuencia directa del estilo de vida de cada familia; sin embargo, también resulta interesante observar las diferencias ambientales entre hermanos. Los **ambientes no compartidos** resultan principalmente de la diferencia de cada hermano al asimilar eventos de su vida, y no tanto de advertir eventos distintos. Por ejemplo, el divorcio de los padres puede ser experimentado de forma muy distinta por hermanos de la misma familia. Los ambientes no compartidos pueden incluir: composición familiar, trato de los padres, interacciones, influencias externas, entre otros.

La importancia de estudiar este trastorno y sobre todo los factores relacionados con su predicción en poblaciones en riesgo, estriba en su gran carga social en términos de costo económico, estrés familiar y efectos adversos en el desempeño académico y vocacional [5]. El estudio de este trastorno y sus factores predictores es necesario para crear estrategias preventivas para la población en riesgo y tratar de contrarrestar parte de su impacto negativo[1].

5.2. Definición de variables y descripción de los datos

En este capítulo se ajusta un conjunto de modelos logísticos con la finalidad de encontrar las diferencias entre un grupo de probandos con TDAH y sus hermanos, que no padecen del mismo trastorno. Estos grupos son de pleno interés pues están formados por hermanos, y que comparten la misma adversidad genética y medioambiental, pero a su vez, sólo uno de ellos desarrolla TDAH.

Las variables elegidas para explicar y predecir el trastorno corresponden a dos grupos que se consideran etiológicamente importantes, la inteligencia emocional y el funcionamiento ejecutivo. Para cada modelo ajustado, se utiliza un subconjunto de las variables dentro de los grupos como variables de interés. Los grupos son:

- **Funcionamiento ejecutivo:** Medido por el inventario de evaluación del comportamiento de funciones ejecutivas (Behavior Rating Inventory of Executive Function, BRIEF) [17].
- **Inteligencia emocional:** Medida por el test de inteligencia emocional Mayer-Salovey-Caruso (Mayer-Salovey-Caruso Emotional Intelligence Test, MSCEIT) [20].

La variable dependiente en todos los modelos corresponde al estatus diagnóstico del trastorno para cada sujeto (TDAHDXSDEF). En algunos modelos, el sexo y la edad del sujeto se utilizan como variables de control. Las variables que corresponden a cada grupo se describen a continuación.

5.2.1. Variables de control

En algunos modelos se utilizan variables de control, que son:

[1]La mayoría de los datos presentados en esta sección fueron obtenidos de [4] sin embargo en el texto se incluyen las referencias originales.

- **SX**: La variable (SX) corresponde al sexo del sujeto. Su codificación es: 0: Femenino y 1: Masculino. La razón por la que el sexo masculino tiene valor de 1 es que usualmente se le otorga el riesgo a los sujetos masculinos por su mayor respresentación del TDAH.

- **Edad** La variable (Edad) corresponde a la edad del sujeto. Esta variable no esta codificada, únicamente se toman las edades enteras de los sujetos en años al tiempo de la entrevista.

5.2.2. Variables del funcionamiento ejecutivo

Las siguientes variables componen las áreas de disfunción del sujeto dentro de las subescalas de funcionamiento ejecutivo. Todas las variables de funciones ejecutivas son variables binarias y se codifican como 0: sin disfunción y 1: área disfuncional. A continuación se describen las variables. A continuación se describen las variables.

- **BRIEFADOL_IN**: Disfunción en el área de inhibición.
- **BRIEFADOL_SH**: Disfunción en el área de cambio de tarea.
- **BRIEFADOL_EC**: Disfunción en el área del control emocional.
- **BRIEFADOL_IA**: Disfunción en el área de inicio de actividad.
- **BRIEFADOL_WN**: Disfunción en el área de memoria de trabajo.
- **BRIEFADOL_PO**: Disfunción en el área de organización y planeación.
- **BRIEFADOL_MO**: Disfunción en el área de monitoreo.
- **BRIEFADOL_OM**: Disfunción en el área de organización de materiales.
- **BRIEFADOL_BRI**: Disfunción en el área de regulación de conducta.
- **BRIEFADOL_MCI**: Disfunción en el índice metacognitivo.
- **BRIEFADOL_GEC**: Disfunción en el índice ejecutivo global.

5.2.3. Variables de inteligencia emocional

Las siguientes variables componen las subescalas de inteligencia emocional. Al igual que las variables de funcionamiento ejecutivo son variables binarias y se codifican como 0: buen desempeño y 1: desempeño deficiente.

- **MSCEIT_PE**: Desempeño en la percepción emocional.
- **MSCEIT_FE**: Desempeño en la facilitación emocional.
- **MSCEIT_CE**: Desempeño en la comprensión emocional.
- **MSCEIT_ME**: Desempeño en el manejo emocional.
- **MSCEIT_EXP**: Desempeño en el área de experiencia.
- **MSCEIT_EST**: Desempeño en el área de estrategia.

Los datos están conformados por 94 observaciones, que pueden ser datos proporcionados por probandos con diagnóstico de TDAH o su hermano (o hermanos) que no tiene TDAH. La muestra se tomó en tres centros de atención a la salud mental: Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz (INPRFM), Hospital Psiquiátrico Infantil Juan N. Navarro (HPIJNN) y Fundación Federico Hoth AC (FFHAC); todas ellas ubicadas en la Ciudad de México, México. Los datos fueron facilitados por el Dr. Lino Palacios Cruz de la Clínica de Adolescencia del INPRMF [22]. Las edades de los sujetos se encuentran en el rango de 12 a 21 años, con una media de 15.79. El 56 % de los sujetos son hombres; el sexo masculino se codificó con 1 porque se espera que se le atribuya el riesgo. En la tabla 5.1 se muestran los estadísticos descriptivos de las variables incluidas en los modelos ajustados. No se encontraron datos perdidos para ninguna de las variables elegidas.

Cuadro 5.1: Tabla de descripción de los datos. Las columnas representan: número de sujetos, media, desviación estándar (D.S), mediana, mínimo y máximo por variable.

	n	Media	D.S	Mediana	Min	Max
TDAHDXSDEF	94	0.51	0.50	1.0	0	1
SX	94	0.56	0.50	1.0	0	1
Edad	94	15.79	2.26	16.0	12	21
BRIEFADOL_IN	94	0.37	0.49	0.0	0	1
BRIEFADOL_SH	94	0.40	0.49	0.0	0	1
BRIEFADOL_EC	94	0.33	0.47	0.0	0	1
BRIEFADOL_IN2	94	0.33	0.47	0.0	0	1
BRIEFADOL_WN	94	0.50	0.50	0.5	0	1
BRIEFADOL_PO	94	0.38	0.49	0.0	0	1
BRIEFADOL_MO	94	0.34	0.48	0.0	0	1
BRIEFADOL_OM	94	0.21	0.41	0.0	0	1
BRIEFADOL_BRI	94	0.46	0.50	0.0	0	1
BRIEFADOL_MCI	94	0.39	0.49	0.0	0	1
BRIEFADOL_GEC	94	0.44	0.50	0.0	0	1
MSCEIT_EXP	94	0.27	0.44	0.0	0	1
MSCEIT_EST	94	0.33	0.47	0.0	0	1
MSCEIT_PE	94	0.27	0.44	0.0	0	1
MSCEIT_FE	94	0.32	0.47	0.0	0	1
MSCEIT_CE	94	0.34	0.48	0.0	0	1
MSCEIT_ME	94	0.32	0.47	0.0	0	1

5.3. Modelos ajustados

Se ajustaron tres modelos considerando la cantidad de variables independientes introducidas y el tipo de variables divididas en principales, confusoras y de interacción. Uno de los tres modelos fue seleccionado y posteriormente se discuten sus características e interpretación.

5.3.1. Modelo 1

El Modelo 1 resulta del ajuste de un modelo logístico con la variable de Índice Ejecutivo Global (BRIEFADOL_GEC). Esta es la principal variable de interés. Aunque el Modelo 1 resulta un modelo sencillo, este ajuste es útil para ejemplificar el caso más elemental de regresión logística: el modelo de regresión logística simple (2.2).

$$g(P(TDAHXSDEF = 1|X)) = \beta_0 + \beta_1(X = BRIEFADOL\_GEC)$$

5.3.2. Modelo 2

El Modelo 2 incorpora las variables independientes de índice ejecutivo global ($BRIEFADOL\_GEC$) y el área de experiencia y estrategia de inteligencia emocional ($MSCEIT\_EXP$ y $MSCEIT\_EST$). También se introduce el sexo del sujeto (SX) como variable de ajuste (control).

$$\begin{aligned} g(P(TDAHXSDEF = 1|X)) = \beta_0 &+ \beta_1(X_1 = BRIEFADOL\_GEC) \\ &+ \beta_2(X_2 = MSCEIT\_EXP) \\ &+ \beta_3(X_3 = MSCEIT\_EST) \\ &+ \beta_4(X_4 = SX) \end{aligned}$$

5.3.3. Modelo 3

En numerosos estudios se observa que una variable, aunque no sea significativa por sí misma, puede modificar ampliamente el efecto que tiene otra variable en el desenlace. A esto se le llama **interacción multiplicativa**. Cuando se presenta interacción multiplicativa en un par de variables, es necesario introducir al modelo una nueva variable que resulta del producto de ambas. Se ha observado que el riesgo de las disfunciones en el funcionamiento ejecutivo es modificado por la edad

de los pacientes, por lo tanto la variable de interacción entre el índice ejecutivo global y la edad resulta interesante para un modelo explicativo. El Modelo 3 resulta de introducir la variable confusorada de *Edad*, así como la variable de interacción $BRIEFADOL\_GEC \times Edad$ al Modelo 2.

$$\begin{aligned} g(P(TDAHXSDEF = 1|X)) = \beta_0 &+ \beta_1(X_1 = BRIEFADOL\_GEC) \\ &+ \beta_2(X_2 = MSCEIT\_EXP) \\ &+ \beta_3(X_3 = MSCEIT\_EST) \\ &+ \beta_4(X_4 = SX) \\ &+ \beta_5(X_5 = Edad) \\ &+ \beta_6(X_6 = BRIEFADOL\_GEC \times Edad) \end{aligned}$$

5.4. Análisis del Modelo 3

El Modelo 3 fue elegido para la discución por su complejidad en número de variables y tipos de variables independientes, y en patrones de covariables. El diagnóstico de este modelo es un procedimiento similar al de los modelos 1 y 2, sin embargo, el cálculo de la razón de momios no se deriva directamente de un coeficiente. En general, los procedimientos utilizados para diagnosticar el Modelo 3 pueden ser utilizados en cualquier análisis de regresión logística.

El ajuste de los datos se realiza por medio de las ecuaciones de verosimilitud (2.17) y (2.18). La prueba de Wald univariada comprueba que las variables de índice ejecutivo global ($W = 2.32$; $p = .0202$) y sexo ($W = 2.53$; $p = 0.0115$) son significativas. La interacción entre el índice ejecutivo global y la edad ($W = -1.82$; $p = 0.0693$), y la variable de experiencia de inteligencia emocional ($W = 1.80$; $p = 0.0732$), presentan una tendencia a ser significativas.

Cuadro 5.2: Resumen del Modelo 3

	Estimado	IC 2.5 %	IC 97.5 %	Error Std	Valor z	Pr(>|z|)
Intercepto	-1.6801	-7.0560	3.5054	2.6524	-0.6334	0.5264
BRIEFADOL_GEC	11.0528	2.3640	21.4506	4.7577	2.3231	0.0202
MSCEIT_EXP	1.4518	-0.0990	3.1206	0.8104	1.7916	0.0732
MSCEIT_EST	-0.6801	-2.2603	0.7659	0.7623	-0.8922	0.3723
SX	1.4969	0.3783	2.7342	0.5923	2.5274	0.0115
Edad	-0.0237	-0.3408	0.2871	0.1575	-0.1507	0.8802
BRIEFADOL_GEC:Edad	-0.5206	-1.1387	0.0131	0.2866	-1.8164	0.0693

En las pruebas de significancia de las variables en conjunto, tanto en la prueba de Wald múltiple ($W = 23.00$; $p = 0.0000$) como en la del cociente de verosimilitudes ($G = 48.34$; $p = 0.0008$), las variables resultan significativas en comparación de un modelo nulo.

Cuadro 5.3: Prueba del cociente de verosimilitudes del Modelo 3

	#Gl	LogVer	Gl	Chi-Cuad	P(>Chi-Cuad)
1	7	-40.9650			
2	1	-65.1346	-6	48.3391	0.0000

Cuadro 5.4: Prueba de Wald del Modelo 3

	Gl Res	Gl	Chi-Cuad	P(>Chi-Cuad)
1	87			
2	93	-6	23.0063	0.0008

Los intervalos de confianza para los parámetros asociados a las variables independientes a un grado de confianza del 95 % se obtienen a través de la ecuacion (2.24). Los resultados se muestran el la tabla 5.2.

5.4.1. Interpretación y razón de momios

Las medidas de asociación, presentadas en el capítulo 3, son indispensables para comparar el riesgo entre patrones de covariables. En especifico, la razón de momios es una medida de asociación comunmente utilizada en los análisis de regresión logística por su versatilidad.

El cálculo de la razón de momios de este modelo considera, además de las variables independientes y de control, la variable de interacción $BRIEFADOL\_GEC \times Edad$. Para calcular la razón de momios de la variable $BRIEFADOL\_GEC$ es necesario otorgar valores fijos a las variables independientes adicionales. También es necesario otorgar valores a la variable con la que el índice ejecutivo global tiene una interacción, en este caso, la variable de control $Edad$. Siguiendo la fórmula (3.3), la RM del índice ejecutivo global es

$$\begin{aligned} RM_{BRIEFADOL\_GEC=0,1} &= exp[(1-0)\beta_1 + (1-0)\beta_5 \times Edad] \\ &= exp[11.05 - 0.52 \times Edad]. \end{aligned}$$

Se puede observar que el efecto de la interacción de la edad del sujeto con el índice ejecutivo global tiene un enorme impacto en los momios de la respuesta pues los momios aumentan 122 veces a una edad de 12 años. Sin embargo, la razón de momios disminuye según aumenta la edad del sujeto hasta 1, para una edad de 21 años.

También es posible calcular la razón de momios entre dos colecciones diferentes de valores de las variables independientes. En este caso se eligen dos patrones de covariables de interés. Sean $\boldsymbol{x}_1 = (BRIEFADOL\_GEC = 0, MSCEIT\_EXP = 0, MSCEIT\_EST = 0)$ y $\boldsymbol{x}_2 = (BRIEFADOL\_GEC = 1, MSCEIT\_EXP = 1, MSCEIT\_EST = 1)$, ambas especificaciones con el sexo (SX) fijo, siguendo nuevamente la formula (3.3) la razón de momios que compara $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$ es

$$\begin{aligned} RM_{\boldsymbol{x}_1,\boldsymbol{x}_2} &= exp[(1-0)\beta_1 + (1-0)\beta_2 + (1-0)\beta_3 + (1-0)\beta_5 \times Edad] \\ &= exp[11.05 + 1.45 - 0.68 - 0.52 \times Edad] \\ &= exp[11.82 - 0.52 \times Edad]. \end{aligned}$$

Comparando las especificaciones de variables independientes $\boldsymbol{x}_1$ contra $\boldsymbol{x}_2$, se observa que el efecto de las tres variables principales $BRIEFADOL\_GEC$, $MSCEIT\_EXP$ y $MSCEIT\_EST$ es mayor que el efecto de la variable $BRIEFADOL\_GEC$ por sí sola. A la edad de 12 años, los momios de $\boldsymbol{x}_2$ son 264 veces mayor que los de $\boldsymbol{x}_1$. Se observa una disminución considerable de la razón de momios para edades mayores. A la edad de 21 años la razón de momios entre $\boldsymbol{x}_2$ y $\boldsymbol{x}_1$ se reduce a 2.44.

En un modelo simple se puede observar el efecto directo de una variable construyendo una gráfica de las probabilidades estimadas contra los valores de la variable independiente. En modelos múltiples, esta técnica depende de los valores fijos que se otorguen a las variables independientes adicionales. En las gráficas 5.1 y 5.2 se observa que el efecto de el índice ejecutivo global sobre la probabilidad estimada es modificado por la edad del sujeto cuando las variables de inteligencia emocional son cero. Se concluye que a mayor edad del sujeto, la variable $BRIEFADOL\_GEC$ tiene menor impacto en las probabilidades estimadas, llegando a un efecto casi nulo a la edad de 21 años. En los hombres, a la edad de 13 años la probabilidad estimada con $BRIEFADOL\_GEC = 0$ es aproximadamente 0.4 y con $BRIEFADOL\_GEC = 1$ es 1.0. En cambio, a la edad de 21 años, la probabilidad estimada con $BRIEFADOL\_GEC = 0$ y $BRIEFADOL\_GEC = 1$ es muy similar. El mismo efecto se puede observar para el caso de las mujeres. Como se esperaba, el riesgo en mujeres es menor que en hombres a lo largo de todas las edades. El intervalo de confianza de $\widehat{\pi}(x)$ se calcula a través de la ecuación (2.11).

Cuadro 5.5: *:Razón de momios del índice ejecutivo global para valores de la variable de edad, controlado por el sexo. **:Razón de momios del índice ejecutivo global y los índices de inteligencia emocional para valores de la variable de edad, controlado por el sexo.

	Edad	R. Momios *	R. Momios **
1	12	122.23	264.45
2	13	72.63	157.13
3	14	43.15	93.36
4	15	25.64	55.47
5	16	15.24	32.96
6	17	9.05	19.59
7	18	5.38	11.64
8	19	3.20	6.91
9	20	1.90	4.11
10	21	1.13	2.44

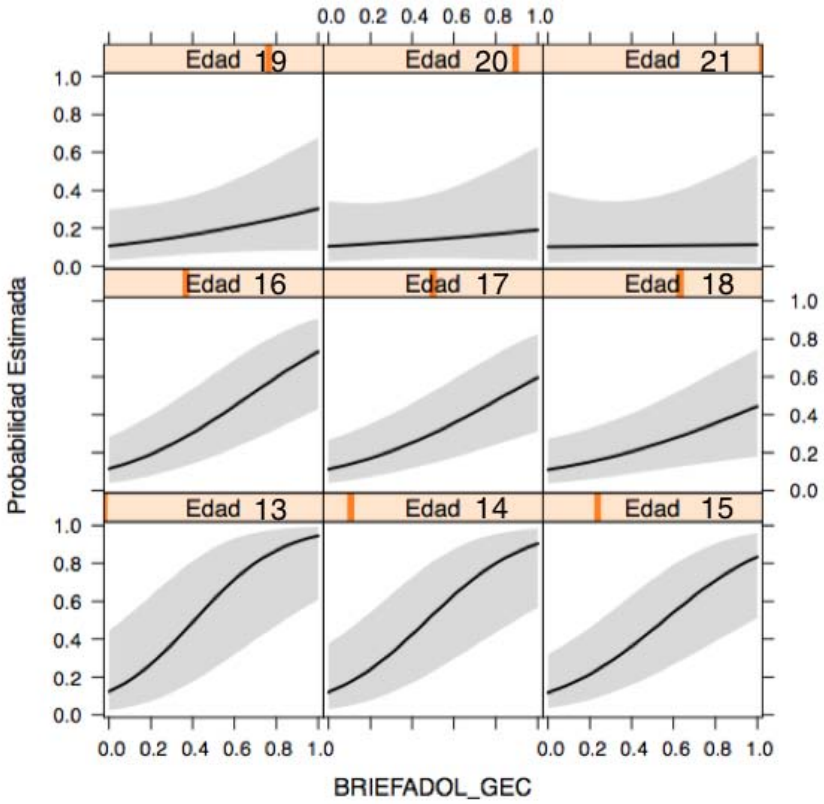

Figura 5.1: Efecto de *Edad* en la probabilidad estimada ($\pi(x)$) por nivel en $BRIEFADOL\_GEC$ en mujeres.

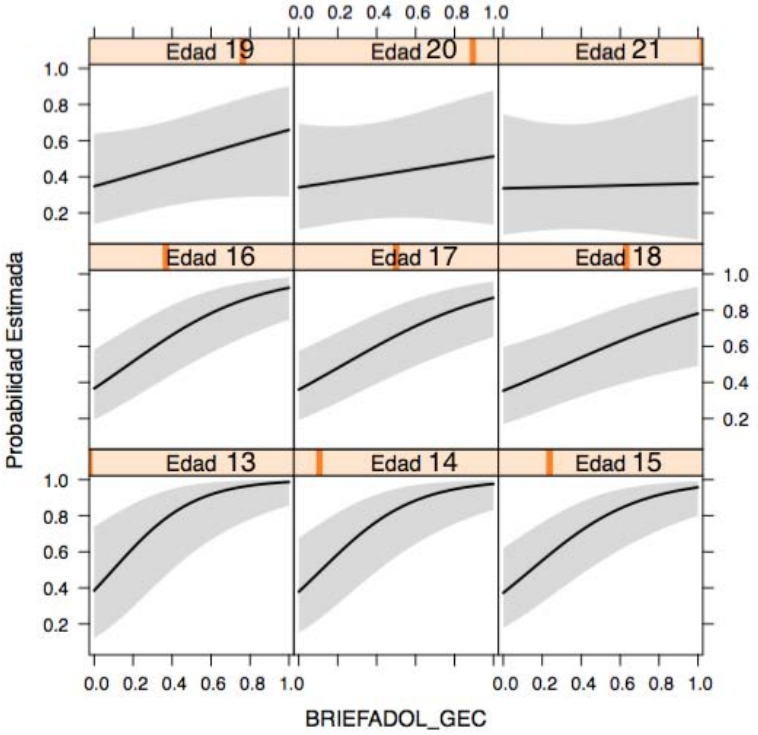

Figura 5.2: Efecto de *Edad* en la probabilidad estimada ($\pi(x)$) por nivel en $BRIEFADOL\_GEC$ en hombres.

5.4.2. Bondad de ajuste

Como se desarrolló en el capítulo 4, la finalidad de las medidas de bondad de ajuste es comparar las estimaciones generadas por un modelo ajustado con observaciones reales. Tal comparación se realiza de manera general y no observación por observación. En este análisis se utilizan tres medidas de bonad de ajuste: el estadístico de devianza, Ji-cuadrada de Pearson y Hosmer-Lemeshow.

La prueba de bondad de ajuste basada en la devianza ($D = 81.93$, $p = 0.6335$) resulta no significativa para el Modelo 3 por lo que se puede afirmar que no hay evidencia de falta de ajuste del modelo. Esta prueba se realiza obteniendo la probabilidad de que D sea mayor que una distribución Ji-cuadrada con grados de libertad igual a los del modelo.

La prueba Ji-Cuadrada ($\chi^2 = 94.12$,$p = 0.2824$) resulta no significativa, por lo tanto no hay evidencia de falta de ajuste, lo que concuerda con el resultado de la prueba de la Devianza.

Ya que el Modelo 3 produce 54 patrones de coviariables, la exactitud de la prueba de Hosmer-Lemeshow se incrementa considerablemente comparada con el Modelo 2, que sólo cuenta con 15 patrones de covariables. La prueba de Hosmer-Lemeshow indica que no hay evidencia de falta de ajuste del Modelo 3. La prueba de H-L depende de la cantidad de grupos (Q) de valores de $\widehat{\pi}(x)$, aunque comúnmente se realiza con $Q = 10$. En la tabla 5.6 se muestran los valores de la probabilidad de la prueba para valores de Q que van de 5 a 15. En algunos casos no es posible calcular la prueba para ciertos valores de Q porque algunos grupos quedan vacíos; estos casos se representan con el símbolo $-$ en la columna de probabilidades de la tabla 5.6.

Cuadro 5.6: Valores de probabilidad en la prueba Hosmer-Lemeshow para distintos valores de gurpos Q

	Grupos	gl	C	p
1	5	3.00	6.64	0.0842
2	6	4.00	4.07	0.3969
3	7	5.00	5.77	0.3289
4	8	6.00	7.51	0.2766
5	9	7.00	7.21	0.4077
6	10	8.00	9.92	0.2704
7	11	9.00	7.04	0.6334
8	12	10.00	6.99	0.7262
9	13	11.00	9.53	0.5728
10	14	12.00	12.57	0.4014
11	15	13.00	15.92	0.2533

5.4.3. Diagnósticos

Los diagnósticos, presentados en la sección 4.4, son las pruebas sobre los patrones de covariables individuales para conocer puntos que van en contra del ajuste general del modelo. Los dos aspectos que consideran los diagnósticos son el ajuste y la influencia que tienen los puntos en los coeficientes estimados del modelo. El análsis de diagnósticos se realiza haciendo observaciones sobre gráficas de la diferencia que hay en la medidas de bondad de ajuste o en los coefientes estimados cunado se excluye un patrón de covariables del modelo contra la probabilidad estimada, entre otras gráficas. Respecto a la falta de ajuste de los patrones de covariable, las gráficas $\Delta\chi^2$ (ecuación (4.4)) contra la probabilidad estimada y en la de ΔD (ecuación (4.5)) contra la probabilidad estimada se prefiren a las gráficas de residuos pues al tener el término del residuo al cuadrado eliminan el signo del residuo y acentúan una posible falta de ajuste en algún patrón de covariables. En las figuras 5.5 y 5.6 , la mayoría de los puntos presenta un ajuste bueno. Únicamente se observan tres puntos mayores a 4 en la primera y uno en la segunda.

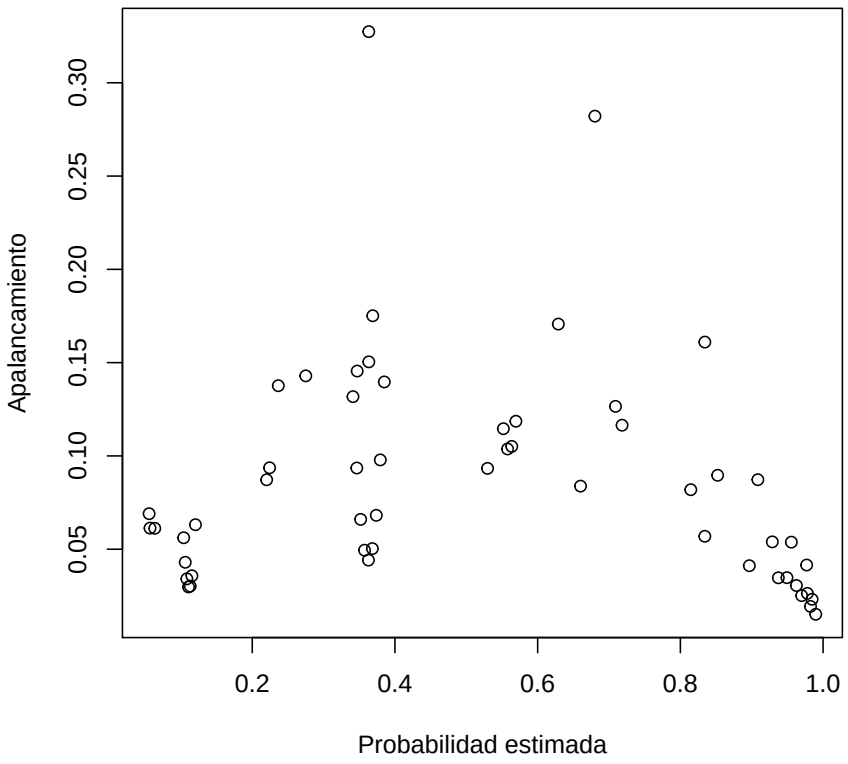

Figura 5.3: Gráfica del apalancamiento contra la probabilidad estimada, $J = 54$.

En la gráfica de $\Delta\widehat{\beta}$ (ecuación (4.3)) contra la probabilidad estimada se puede observar tres puntos que tienen la mayor influencia en los coeficientes estimados, aunque sólo dos de éstos son superiores a 1.0. Hosmer y Lemeshow (2013) observan que sólo los puntos con influencia en los coeficientes superior a 1.0 tienen un efecto importante en los coeficientes, sin embargo, existen excepciones y es importante notar un comportamiento fuera de lo común en otros datos. La gráfica de $\Delta\chi^2$

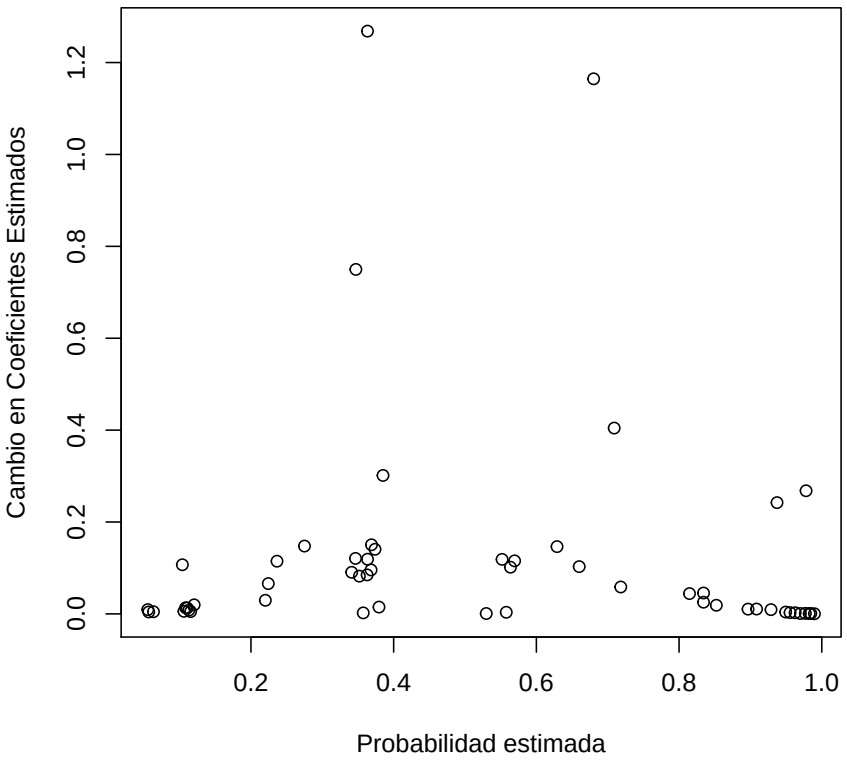

Figura 5.4: Gráfica de $\Delta\widehat{\beta}$ contra la probabilidad estimada, $J = 54$.

contra $\widehat{\pi}$ donde el símbolo es proporcional a $\Delta\widehat{\beta}$ se presenta en la Figura 5.7. Esta gráfica es útil al evaluar el papel que juegan los residuos y el apalancamiento en el cambio de los coeficientes estimados. Se puede observar que los dos circulos más grandes ($\widehat{\pi} \approx 0.36$,$\widehat{\pi} \approx 0.68$) corresponde a una diferencia en la estadística Ji-cuadrada moderada ($\Delta\chi^2 = 2.60$, $\Delta\chi^2 = 2.96$) pero a los apalancamientos más elevados ($h = 0.32$, $h = 0.28$). El tercer circulo más grande ($\widehat{\pi} \approx 0.34$) corresponde a un cambio en la Ji-Cuadrada grande ($\Delta\chi^2 = 4.4$) y a un apalancamiento de moderado a pequeño ($h = .14$). También se debe observar que los dos puntos con un cambio en la Ji-Cuadrada más grande, presentan una $\Delta\widehat{\beta}$ moderada pues su probabilidad estimada es cercana a 1, lo que implica que el apalancamiento es bajo.

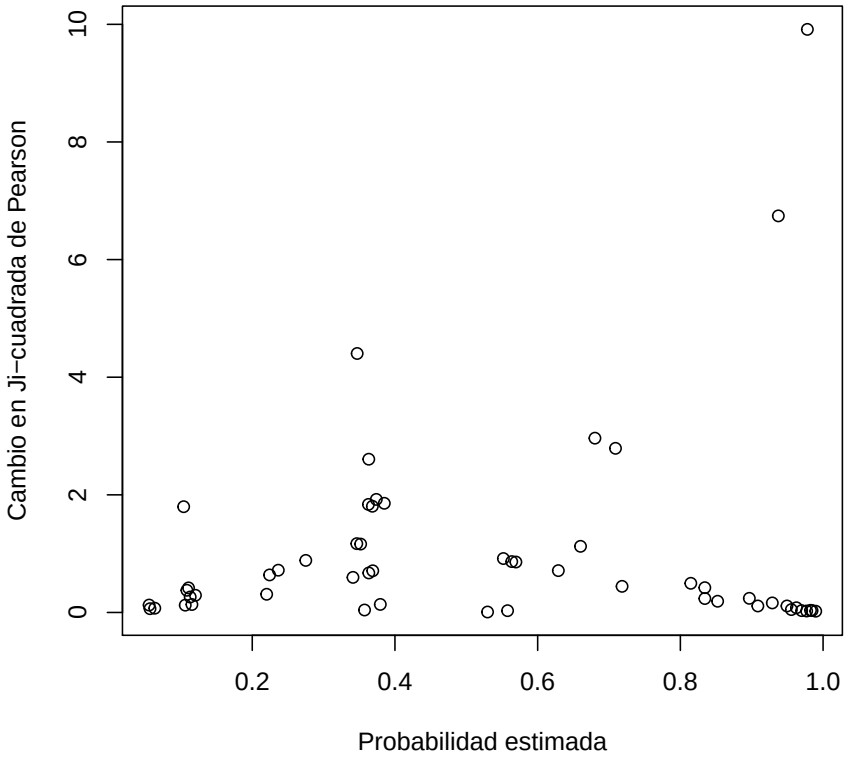

Figura 5.5: Gráfica de $\Delta\chi^2$ contra la probabilidad estimada, $J = 54$.

Hasta ahora hemos identificado cinco patrones de covariables que se destacan por su mal ajuste y su influencia. Los patrones número $48, 49, 52$ son los puntos con $\Delta\chi^2$ más alto. El patrón 52 es el único punto con ΔD mayor a 4. Los patrones $52, 53, 54$ son los que presentan la mayor influencia en los coeficientes estimados. Por último, los patrones 52 y 53 son los que tienen el apalancamiento más grande. El sigiuente paso en el análisis diagnóstico es evaluar el efecto de los patrones de covariables que identificamos, cuando son eliminados del modelo, en los coeficientes individuales. En la tabla 5.7 se pueden encontrar la influencia de los patrones de covariables nùmero 48,49,52,53,54 en los coeficientes individuales. Especialmente los patrones 52 y 53 que corresponden a patrones con $\Delta\widehat{\beta}$ más alta son los que presentan mayor influencia en los coeficientes individuales, sobre todo en la variable *Edad*.

En la tabla 5.8 se presentan los datos del porcentaje de cambio en los coeficientes individuales cuando se borran grupos de puntos identificados. En este análisis se consideran tres grupos. Los puntos con mal ajuste, los puntos con influencia grande en los coeficientes estimados y el grupo de todos los puntos identificados. Borrar el primer grupo tiene una gran influencia en la variable *Edad* (-300), en la varible del

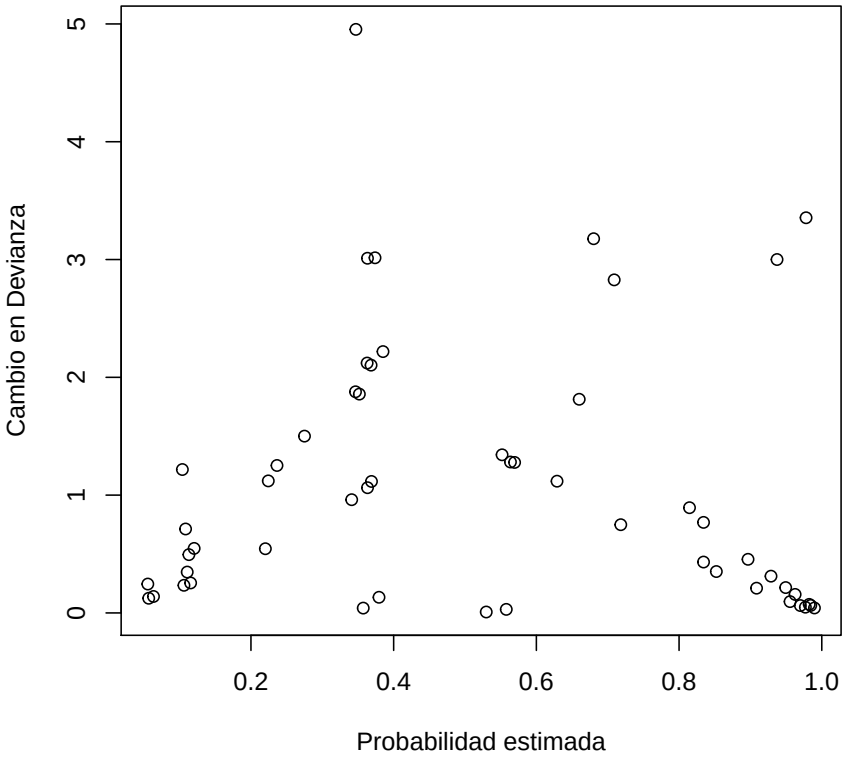

Figura 5.6: Gráfica de ΔD contra la probabilidad estimada, $J = 54$.

Cuadro 5.7: Cambio que ejerecen los patrones de covariables identificados en los coeficientes individuales en porcentaje.

	#48	#49	#52	#53	#54
(Intercept)	-7	-12	42	-43	-4
BRIEFADOL_GEC	-21	-48	11	17	-36
MSCEIT_EXP	-12	-21	64	-21	-4
MSCEIT_EST	-45	-65	87	18	11
SX	-17	-23	-14	-11	7
Edad	-7	0	-257	131	23
BRIEFADOL_GEC:Edad	-23	-56	16	25	-50

funcionamiento ejecutivo $BRIEFADOL\_GEC$ (-85) y en la interacción de estas dos variables. Borrar el conjunto de patrones de covariables con $\Delta\widehat{\beta}$ más elevado representa un cambio fuerte en las variables $MSCEIT\_EST$ (106) y $Edad$ (-157). Borrar el grupo de todas las variables identificadas tiene un efecto fuerte en las variables $BRIEFADOL\_GEC$ (-2664) $Edad$ (-200) y en la interacción de estas dos

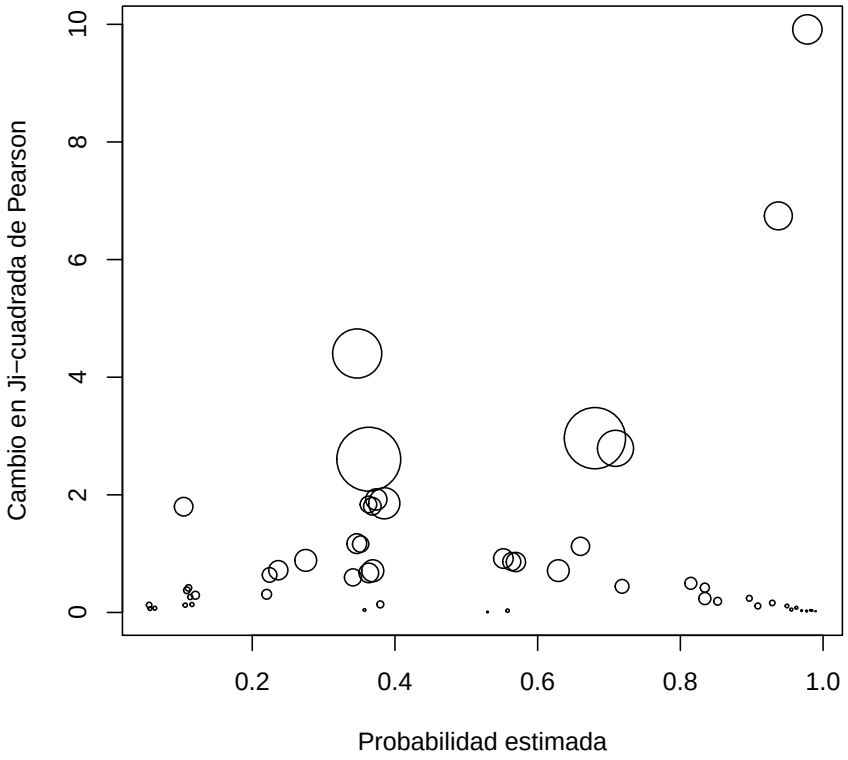

Figura 5.7: Gráfica de $\Delta\chi^2$ contra la probabilidad estimada donde el radio de cada símbolo es proporcional a $\Delta\beta$, $J = 54$.

variables (-3322). Se debe tener presente que el último grupo representa aproximadamente un 10 % del total de patrones de covariables en el modelo lo cual podría explicar los cambios drásticos que se observan en los coeficientes individuales.

Cuadro 5.8: Cambio que ejerecen los patrones de covariables identificados por grupos, en los coeficientes individuales en porcentaje. Grupo A: patrones con mal ajuste. Grupo B: patrones con influencia en coeficientes estimados. Grupo C: patrones con mal ajuste e influencia en los coeficientes.

	Grupo A	Grupo B	Grupo C
(Intercept)	17	13	-5
BRIEFADOL_GEC	-85	4	-2664
MSCEIT_EXP	31	46	7
MSCEIT_EST	-31	106	-21
SX	-76	-16	-66
Edad	-303	-157	-200
BRIEFADOL_GEC:Edad	-91	7	-3322

5.4.4. Desempeño discriminatorio

En la regresión logística es interesante clasificar las probabilidades estimadas de los patrones de covariables dentro de uno de los valores de la variable de respuesta de un modelo logístico y luego observar que tan acertada es la clasificación con respecto a las observaciones, es decir, evaluar el desempe??o discriminatorio del modelo, para lo cual se utliza la curva ROC y el área bajo la curva, presentadas en la sección 4.5.

En el Modelo 3 se observa que la variable de interacción introducida aumenta la cantidad de patrones de covariables y la capacidad de predicción del modelo comparado con los modelos 1 y 2.

Para ejemplificar la construcción de la cruva ROC y el cálculo del área bajo la curva (AUC) se toma la sensibilidad y especificidad obtenidas del Modelo 1. Ya que la variable $BRIEFADOL\_GEC$ es una variable binaria, el Modelo 1 produce dos patrones de covariables. La tabla 5.9 de sensibilidad, 1−especificidad y puntos de corte se obtiene de las probabilidades estimadas de los patrones de covariables y un punto de corte ficticio $c_0 : \pi(x) = 1$.

Cuadro 5.9: Tabla de sensiblidad, 1-especificidad por punto de corte del Modelo 1. El punto de corte está dado por la probabilidad estimada calculada para cada patrón de corvariables producido por los datos.

	Sensibilidad	1-Especificidad	Punto de corte
1	1.0000	1.0000	0.0000
2	0.7083	0.1522	0.2642
3	0.0000	0.0000	0.8293

El área bajo la curva ROC se calcula según el método trapezoidal (4.5.2). En este caso la división de la curva ROC resulta en dos trapezoides tales que las coordenadas de sus esquinas son (0;0), (0.1522;0) y (0.1522;0.7083) para el primer trapezoide; (0.1522;0), (0.1522;0.7083), (1;0) y (1;1) para el segundo trapezoide. Entonces

$$\begin{aligned} ABC &= t_1 + t_2 \\ &= \frac{(0.1522 - 0)(0.7083 + 0)}{2} + \frac{(1 - 0.1522)(1 + 0.7083)}{2} \\ &= 0.7781 \end{aligned}$$

lo que representa una capacidad de discriminación **aceptable** según el criterio de calificación definido en (4.3.2).

La construcción de la curva ROC y el cálculo del área bajo la curva para el Modelo 3 es equivalente a la del ejemplo anterior; a través de la tabla de la sensibilidad y

1−especificidad. En este caso la tabla 5.10 muestra los valores calculados para el Modelo 3.

La curva ROC de la figura 5.8 del Modelo 3 es más detallada que la del Modelo 1 porque se producen más patrones de covariables y por consiguiente, más puntos de corte. Un $AUC = 0.8793$ muestra una discriminación superior a los modelos 1 y 2. En conclusión, la discriminación del Modelo 3 roza a una discriminación **excelente** según el sistema de calificaciones en la sección 4.3 .

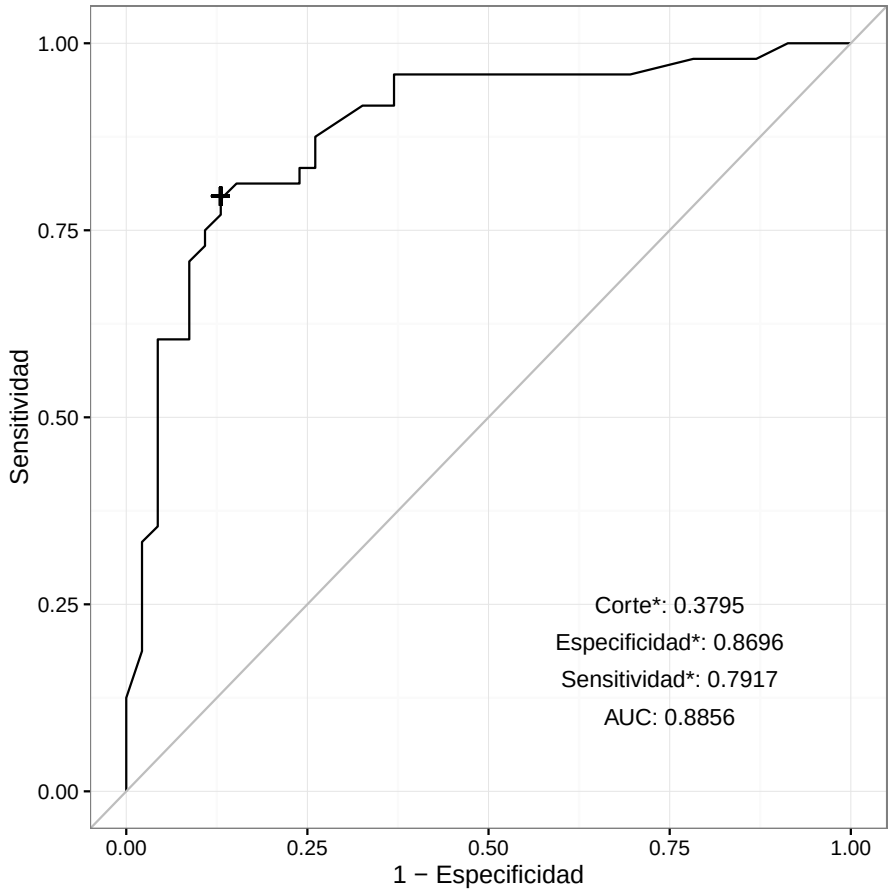

Figura 5.8: Curva ROC del Modelo 3. Corte*:Punto de corte óptimo. Sensibilidad* y Especificidad*: Sensibilidad y especificidad inducidas por el punto de corte óptimo. AUC: Área bajo la curva ROC. +:Punto de la curva ROC inducido por el corte óptimo.

El punto de corte que induce una clasificación óptima se obtiene encontrando el valor que maximiza la suma de la sensibiliad y la especificidad. El corte óptimo tiene la propiedad de que clasifica las respuestas en general con la mayor tasa de aciertos entre todos los cortes. Es posible mostrar gráficamente cómo se clasifican las respuestas dependiendo de un corte específico en una gáfica de las probabilidades estimadas contra las respuestas observadas. La gráfica 5.9 muestra la clasificación

de respuestas del Modelo 3 según el punto de corte óptimo.

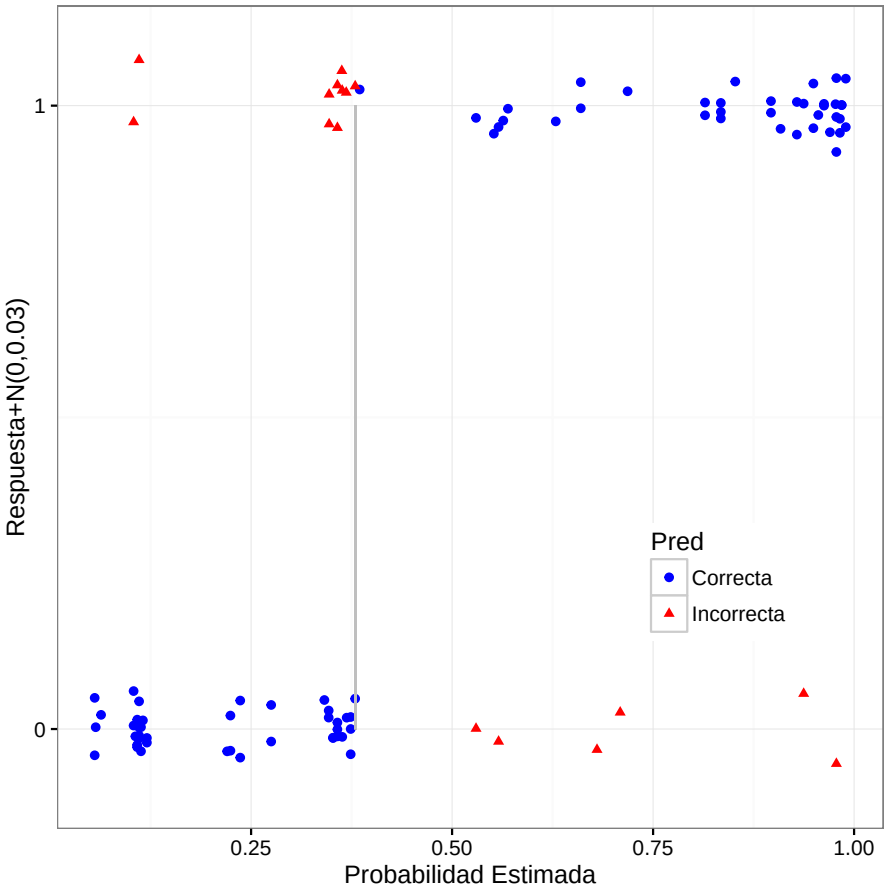

Figura 5.9: Valores de la variable de respuesta más una distribución N(0,0.03) clasificados por el punto de corte óptimo como predicciones correctas e incorrectas.

Cuadro 5.10: Tabla de sensiblidad, 1-especificidad y puntos de corte del Modelo 3

Sensibilidad	1-Especificidad	Punto de corte
1.0000	1.0000	0.0000
1.0000	0.9565	0.0555
1.0000	0.9348	0.0567
1.0000	0.9130	0.0634
0.9792	0.8696	0.1039
0.9792	0.8478	0.1061
0.9792	0.7826	0.1084
0.9583	0.6957	0.1107
0.9583	0.6522	0.1130
0.9583	0.6304	0.1154
0.9583	0.5870	0.1204
0.9583	0.5652	0.2202
0.9583	0.5217	0.2243
0.9583	0.4783	0.2365
0.9583	0.4348	0.2750
0.9583	0.4130	0.3412
0.9583	0.3696	0.3465
0.9167	0.3696	0.3471
0.9167	0.3261	0.3519
0.8750	0.2609	0.3574
0.8542	0.2609	0.3628
0.8333	0.2609	0.3633
0.8333	0.2391	0.3634
0.8125	0.2391	0.3683
0.8125	0.2174	0.3689
0.8125	0.1522	0.3739
0.7917	0.1304	0.3795
0.7708	0.1304	0.3851
0.7500	0.1087	0.5298
0.7292	0.1087	0.5520
0.7083	0.0870	0.5578
0.6875	0.0870	0.5637
0.6667	0.0870	0.5695
0.6458	0.0870	0.6290
0.6042	0.0870	0.6601

0.6042	0.0652	0.6803
0.6042	0.0435	0.7091
0.5833	0.0435	0.7183
0.5417	0.0435	0.8146
0.5000	0.0435	0.8343
0.4792	0.0435	0.8343
0.4583	0.0435	0.8522
0.4167	0.0435	0.8966
0.3958	0.0435	0.9086
0.3542	0.0435	0.9288
0.3333	0.0217	0.9373
0.2917	0.0217	0.9494
0.2708	0.0217	0.9556
0.2292	0.0217	0.9626
0.2083	0.0217	0.9700
0.1875	0.0217	0.9770
0.1250	0.0000	0.9780
0.0833	0.0000	0.9824
0.0417	0.0000	0.9846
0.0000	0.0000	0.9897

Conclusiones

El objetivo principal de esta investigación ha sido exhibir las bondades del modelo de regresión logística en estudios de salud mental. Con este propósito se exploraron las técnicas de ajuste, pruebas de significancia de variables, pruebas de bondad de ajuste y técnicas para evaluar el desempeño discriminatorio de un modelo logístico. Asimismo, se utilizó el ajuste de modelos con datos reales para ejemplificar los conceptos anteriores.

En los primeros cuatro capítulos de este trabajo se logró describir de manera breve los aspectos técnicos de la regresión logística. En el quinto capítulo se mostraron los aspectos prácticos de un análisis de regresión logística. Se utilizó modelo de regresión logística, que incluyó variables de control y de interacciones, con el fin de encontrar las principales diferencias entre un grupo de probandos con el trastorno por déficit de atención con hiperactividad y sus hermanos que no padecian dicho trastorno. Se encontró que la variable que indica la disfunción en el índice ejecutivo global con un peso de 11.06 es significativa ($W = 2.32$, $p = 0.02$). También se observó que la variable de edad modifica el efecto de la variable del índice ejecutivo global, disminuyendo su riesgo a medida que la edad aumenta. Por último, se evaluó el desempeño discriminatorio del modelo através de la curva ROC, dando como resultado una discriminación excelente ($AUC = .89$).

En conclusión, se mostró que la regresión logística es una alternativa para aquellos que estudian modelos explicativos y, en particular, en la investigación médica, en la que se encuentra repetidas veces una variable de respuesta dicotómica.

En estadística y en el análisis de datos en general, es imposible encontrar un modelo definitivo; los modelos sólo son una aproximación matemática de la realidad. No obstante, el modelo de regresión logística es, sin duda, una herramienta muy útil en la actualidad. Sus aplicaciones son vastas y suele utilizarse en áreas de la investigación tan diversas como el estudio de materiales, exploración geoquímica y prevención de accidentes. Por ésta y muchas otras razones, el estudio y comprensión de este modelo es imprescindible para enriquecer cualquier acervo de técnicas estadísticas.

Bibliografía

[1] A Albert and Emmanuel Lesaffre. Multiple group logistic discrimination. *Computers & mathematics with applications*, 12(2):209–224, 1986.

[2] Steven C Bagley, Halbert White, and Beatrice A Golomb. Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain. *Journal of clinical epidemiology*, 54(10):979–985, 2001.

[3] Joseph Biederman. Familial association between attention deficit disorder. *Am J Psychiatry*, 148:251–256, 1991.

[4] Joseph Biederman. Attention-deficit/hyperactivity disorder: a selective overview. *Biological psychiatry*, 57(11):1215–1220, 2005.

[5] Joseph Biederman, Stephen V Faraone, Michael C Monuteaux, Marie Bober, and Elizabeth Cadogen. Gender effects on attention-deficit/hyperactivity disorder in adults, revisited. *Biological psychiatry*, 55(7):692–700, 2004.

[6] Joseph Biederman, Eric Mick, Stephen V Faraone, Ellen Braaten, Alysa Doyle, Thomas Spencer, Timothy E Wilens, Elizabeth Frazier, and Mary Ann Johnson. Influence of gender on attention deficit hyperactivity disorder in children referred to a psychiatric clinic. *American Journal of Psychiatry*, 159(1):36–42, 2002.

[7] Joseph Biederman, Sharon Milberger, Stephen V Faraone, Kathleen Kiely, Jessica Guite, Eric Mick, Stuart Ablon, Rebecca Warburton, and Ellen Reed. Family-environment risk factors for attention-deficit hyperactivity disorder: A test of rutter's indicators of adversity. *Archives of general psychiatry*, 52(6):464–470, 1995.

[8] J Bierderman, S Faraone, T Spencer, et al. Patterns of psychiatric comorbidity, cognition and psychosocial functioning in adults with adhd. *Am J Psychiatry*, 150:1792–1797, 1993.

[9] Jeffrey D Burke, Rolf Loeber, and Benjamin B Lahey. Which aspects of adhd are associated with tobacco use in early adolescence? *Journal of Child Psychology and Psychiatry*, 42(04):493–502, 2001.

[10] Ronald Christensen. *Log-linear models and logistic regression*. Springer Science & Business Media, 2006.

[11] Delphine S Courvoisier, Christophe Combescure, Thomas Agoritsas, Angèle Gayet-Ageron, and Thomas V Perneger. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *Journal of clinical epidemiology*, 64(9):993–1000, 2011.

[12] James S Cramer. The origins and development of the logit model. *Logit models from economics and other fields*, pages 149–158, 2003.

[13] Jan Salomon Cramer. The origins of logistic regression. 2002.

[14] Klein David, Kleinbaum y Mitchel. Logistic regression: A self learning text, 1994.

[15] Adrian Dobson, Annette J y Barnett. *An introduction to generalized linear models*. CRC press, 2008.

[16] Stephen V Faraone, Joseph Sergeant, Christopher Gillberg, Joseph Biederman, et al. The worldwide prevalence of adhd: is it an american condition. *World psychiatry*, 2(2):104–113, 2003.

[17] Steven C Guy, Gerard A Gioia, and Peter K Isquith. *Behavior Rating Inventory of Executive Function-: Self-report Version*. Psychological Assessment Resources, 2004.

[18] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[19] Nirian Martin and Leandro Pardo. On the asymptotic distribution of cook's distance in logistic regression models. *Journal of Applied Statistics*, 36(10):1119–1146, 2009.

[20] JD Mayer. Mayer-salovey-caruso emotional intelligence test (msceit), version 2.0. *Toronto, Canada: Multi-Health Systems*, 2002.

[21] John A Nelder and RJ Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972.

[22] Lino Palacios-Cruz, Adriana Arias-Caballero, Rosa Elena Ulloa, Norma González-Reyna, Pablo Mayer-Villa, Miriam Feria, Liz Sosa, Francisco R de la Peña, Alfonso Cabrera-Lagunes, Alejandra Fragoso, et al. Adversidad psicosocial, psicopatología y funcionamiento en hermanos adolescentes en alto riesgo (har) con y sin trastorno por déficit de atención con hiperactividad (tdah). *Salud mental*, 37(6):467–476, 2014.

[23] Steven R Pliszka. Comorbidity of attention-deficit/hyperactivity disorder with psychiatric disorder: an overview. *The Journal of clinical psychiatry*, 59(suppl 7):50–58, 1998.

[24] Robert Plomin and Denise Daniels. Why are children in the same family so different from one another? *Behavioral and Brain Sciences*, 10(01):1–16, 1987.

[25] Daryl Pregibon. Logistic regression diagnostics. *The Annals of Statistics*, pages 705–724, 1981.

[26] H-C Steinhausen. The heterogeneity of causes and courses of attention-deficit/hyperactivity disorder. *Acta Psychiatrica Scandinavica*, 120(5):392–399, 2009.

[27] Erik G Willcutt, Alysa E Doyle, Joel T Nigg, Stephen V Faraone, and Bruce F Pennington. Validity of the executive function theory of attention-deficit/hyperactivity disorder: a meta-analytic review. *Biological psychiatry*, 57(11):1336–1346, 2005.

[28] Li-Kuang Yang and Chi-Yung Shang. Psychiatric comorbidities in adolescents with attention-deficit hyperactivity disorder and their siblings. *Canadian Journal of Psychiatry*, 56(5):281, 2011.