Universidad Nacional Autónoma de México

Facultad de Ciencias

Deserción escolar en la educación media superior: una aproximación logística

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
ACTUARIO

PRESENTA:
JUAN PABLO DÍAZ MARTÍNEZ

DIRECTOR DE TESIS:
DOCTORA RUTH SELENE FUENTES GARCÍA

México D.F. Ciudad Universitaria, 2015

Agradecimientos

A Él, que con su sangre derramada nos dio la vida eterna

Índice general

Introducción

Pronosticar o dar aproximaciones a futuros eventos ha sido siempre una práctica frecuente para los seres humanos. En tiempos remotos estos pronósticos se realizaban mediante métodos un poco ortodoxos. Con el paso del tiempo y gracias a los avances teóricos y tecnológicos de la ciencia, estas aproximaciones han ido cambiando hasta llegar a metodologías rigurosamente científicas y bien fundamentadas teóricamente.

La estadística puede dar respuesta a muchas de las necesidades que la sociedad actual nos plantea. Su tarea fundamental es la reducción de datos, con el objetivo de representar la realidad y describirla, predecir su futuro o simplemente conocerla. En nuestros días, la estadística se ha convertido en un método efectivo para describir con exactitud los valores de datos económicos, políticos, sociales, psicológicos, biológicos y físicos, y sirve como herramienta para relacionar y analizar dichos datos. El trabajo del experto estadístico no consiste ya sólo en reunir y tabular los datos, sino sobre todo en el proceso de interpretación de esa información. El desarrollo de la teoría de la probabilidad ha aumentado el alcance de las aplicaciones de la estadística. Muchos conjuntos de datos se pueden aproximar, con gran exactitud, utilizando modelos probabilísticos; los resultados de éstas se pueden utilizar para analizar datos estadísticos. La probabilidad es útil para comprobar la fiabilidad de las inferencias estadísticas y para predecir el tipo y la cantidad de datos necesarios en un determinado estudio estadístico.

El análisis de regresión es una técnica estadística para investigar y modelar la relación entre variables. Son numerosas las aplicaciones de la regresión, y las hay en casi cualquier campo, incluyendo en ingeniería, ciencias físicas y químicas, economía, administración, ciencias biológicas y de la vida y en las ciencias sociales. De hecho, puede ser que el análisis de regresión sea la técnica estadística más usada.

La deserción escolar en la educación media superior en nuestro país es un grave problema para el desarrollo de México. La deserción escolar en el nivel medio superior pasó del 19.3 % en el ciclo escolar 1994-1995, al 14.4 %, en el ciclo 2011-2012.

Para conocer la realidad de la educación media superior, el gobierno federal por medio de la Secretaria de Educación Pública (SEP) realizó la Encuesta Nacional de Deserción en la Educación Media Superior, realizada en julio de 2011 (ENDEMS), que es la primera en ser representativa a nivel nacional y para cuya elaboración se contó con el respaldo del Consejo para la Evaluación de la Educación del Tipo Medio Superior (COPEEMS A.C.). Algo muy valioso de esta encuesta es que nos proporciona información sobre el momento en que los jóvenes desertores estudiaban, a diferencia de otras, que muestran la situación al momento de la entrevista. Además, presenta entrevistas a jóvenes que no han desertado y a aquellos que nunca ingresaron al nivel medio superior.

Para esta encuesta, 6,472 mujeres y 6,542 hombres de todo el país fueron entrevistados. De ellos, 4,779 jóvenes nunca se matricularon en el nivel medio superior, 2,549 desertaron de este nivel y 5,686 lo concluyeron o que continúan estudiando.

En la presente tesis se utilizará la regresión logística, ésta es adecuada cuando la variable de respuesta sólo tiene dos resultados posibles, llamados, en forma genérica, "éxito" y "fracaso" y se representan por 0 y 1. El presente trabajo buscará encontrar las variables más importantes que inciden en la deserción escolar, así como proveer de resultados para que, de ser necesario, se puedan realizar políticas públicas para reducir la deserción.

El trabajo se divide de la siguiente forma:
En el capítulo 1 se presentan nociones básicas y definiciones del análisis de regresión logística, así como las nociones matemáticas para ajustar el modelo y análisis de residuales.
En el capítulo 2 se describirá las estadísticas descriptivas de la ENDEMS. tratando de entender la deserción en nuestro país y resultados que serán de utilidad durante el desarrollo del presente trabajo.
En el capítulo 3 se presenta el modelo que se aborda en esta tesis, se muestran las estimaciones de la regresión así como las pruebas para el ajuste del modelo. Para finalizar se presentan los apéndices correspondientes y las conclusiones alcanzadas.

Capítulo 1

Entendiendo la regresión logística

Los métodos de regresión se han vuelto una componente principal de cualquier tipo de análisis relacionado con datos, tratando siempre de relacionar una variable de respuesta con una o mas variables explicativas. La regresión logística es un tipo de regresión para predecir el resultado de una variable *binaria o dicotómica* en función de variables explicativas.

Lo que distingue a la regresión logística de la regresión lineal es el modelo paramétrico y las hipótesis de los modelos. Una vez dichas estas diferencias, los métodos empleados en el análisis de la regresión logística son muy parecidos a los de la regresión lineal.

Cómo se explico antes, la variable de respuesta en la regresión logística es de tipo *binaria* y sólo tiene dos tipos de respuestas cualitativas y por lo tanto, puede ser representada por una variable indicadora tomando los valores 0 y 1. Algunos ejemplos en los que se puede utilizar la regresión logística son los siguientes:

- En el estudio de mujeres casadas que trabajan, cómo una función de la edad, número de hijos e ingreso del marido. La variable de respuesta definida tiene dos posibles salidas: mujeres casadas que trabajan y mujeres casadas que no trabajan.
- En un estudio longitudinal de enfermedades del corazón, cómo una función de la edad del paciente, género, si fuma o no, nivel de colesterol y la presión arterial. La variable de respuesta definida tiene dos posibles salidas: personas con enfermedad del corazón al momento del estudio y personas sin enfermedad del corazón al momento del estudio.

Esto ejemplos muestran el gran rango de aplicaciones en donde al variable de respuesta es binaria y por lo tanto, puede ser representada por los valores 0 y 1. Consideremos primero el significado de la "función respuesta" cuando la variable de respuesta es binaria, después veremos los problemas que se pueden tener con este tipo de variables.

1.1. Preliminares

Recordemos el modelo de regresión simple:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad Y_i = 0, 1 \tag{1.1}$$

donde la variable Y_i es binaria, la esperanza de esta variable $\mathbb{E}[Y_i]$ tiene un significado especial en este caso. Dado que $\mathbb{E}[\varepsilon_i] = 0$ tenemos que:

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i \tag{1.2}$$

Ahora consideremos que Y_i es una variable aleatoria que se distribuye Bernoulli, por lo cual podremos estipular la distribución de probabilidad como sigue:

Y_i	Probabilidad
1	$\mathbb{P}[Y_i = 1] = \pi_i$
0	$\mathbb{P}[Y_i = 0] = 1 - \pi_i$

Esto significa que π_i es la probabilidad de que la variable $Y_i = 1$ y que 1-π_i es la probabilidad de que la variable $Y_i = 0$. Recordando la definición de esperanza matemática tenemos que:

$$\mathbb{E}[Y_i] = 1(\pi_i) + 0(1 - \pi_i) = \pi_i = \mathbb{P}[Y_i = 1] \tag{1.3}$$

Igualando (1.2) y (1.3) tenemos que:

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i = \pi_i \tag{1.4}$$

Habiendo visto lo anterior, los problemas empiezan a aparecer cuando la variable de respuesta es una variable binaria. A continuación veremos tres de estos problemas, usando el modelo de regresión lineal simple como ilustración.

1. *No-normalidad de los errores:* Para una variable de respuesta binaria 0,1 cada error $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$ puede tomar únicamente dos valores:

$$\text{Cuando } Y_i = 1: \qquad \varepsilon_i = 1 - \beta_0 - \beta_1 X_i$$

$$\text{Cuando } Y_i = 0: \qquad \varepsilon_i = -\beta_0 - \beta_1 X_i$$

Claramente, en este modelo los errores no se distribuyen normal.

2. *Varianza de los errores no constante:* Otro problema es que los errores ε_i no tienen varianza constante cuando la variable de respuesta es de tipo binaria. Veamos cual es la varianza de la ecuación (1.1):

$$Var\left[Y_i\right] = \mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i\right]\right)^2\right] = (1-\pi_i)^2\pi_i + (0-\pi_i)^2(1-\pi_i) \qquad \Longrightarrow$$

$$Var\left[Y_i\right] = \pi_i(1-\pi_i) = \mathbb{E}\left[Y_i\right]\left(1 - \mathbb{E}\left[Y_i\right]\right) \tag{1.5}$$

Veamos ahora que la varianza de los errores ε_i es igual a la variable de respuesta Y_i, primero despejando ϵ_i de (1.1) e igualando con (1.4) tenemos que $\epsilon_i = Y_i - \pi_i \qquad \Longrightarrow$

$$Var\left[\varepsilon_i\right] = Var\left[Y_i - \pi_i\right] = 1^2 Var\left[Y_i\right] = \pi_i(1-\pi_i) = \mathbb{E}\left[Y_i\right]\left(1 - \mathbb{E}\left[Y_i\right]\right) \Longrightarrow$$

$$Var\left[\varepsilon_i\right] = (\beta_0 + \beta_1 X_i)(1 - \beta_0 + \beta_1 X_i) \tag{1.6}$$

podemos notar en esta última ecuación que la $Var\left[\varepsilon_i\right]$ depende de la variable X_i. Esto implica que la varianza de los errores difiere para cada valor de X y por lo tanto la estimación de los parámetros por mínimos cuadrados ordinarios no será óptimo.

3. *Restricciones en la función de respuesta:* Dado que la función de respuesta representa probabilidades cuando la variable es binaria, el valor esperado de la variable de respuesta estará restringida de la siguiente manera:

$$0 \leq \mathbb{E}\left[Y\right] = \pi \leq 1 \tag{1.7}$$

Varias funciones no necesariamente cumplen esta restricción. Las dificultades creadas por la restricción (1.7) en la función de respuesta son las más importantes entre los otros dos problemas antes anunciados. Se podrían utilizar otro tipo de métodos para estimar los parámetros (mínimos cuadrados ponderados) para manejar el problema de la desigualdad de varianzas en los errores. Además, con muestras demasiado grandes el método de mínimos cuadrados provee estimadores que son asintóticamente normales, aunque los errores no se dristribuyan normales. Por estas razones, se introduce la función logística que ayuda a modelar cualquier tipo de respuesta binaria.

1.2. La función logística

Antes de empezar a explicar el modelo de regresión logística, hemos visto en la sección anterior los problemas de modelar variables de respuesta binarias con el modelo de regresión lineal. Por esta razón, introduciremos la función logística. Las características de esta función es que está acotada entre 0 y 1, su gráfica tiene una forma peculiar en forma de S (sigmoidal) y se aproximan a 0 y a 1 asintóticamente. En la figura 1.1 podemos ver la gráfica de una normal estándar y de una función de densidad logística, cada una con media 0 y varianza 1.

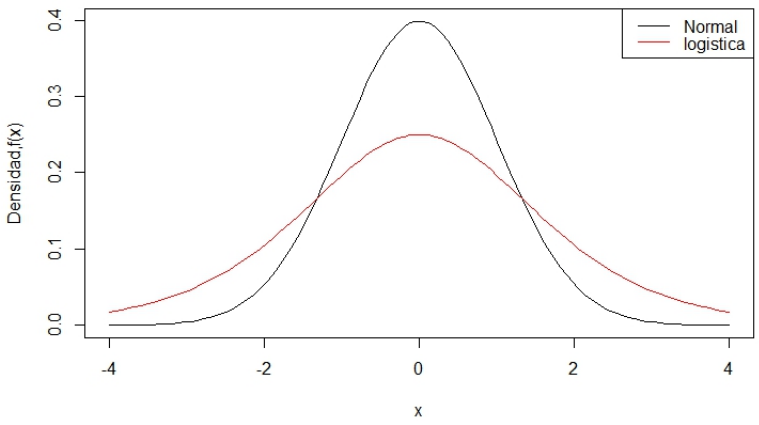

Figura 1.1: Función de densidad de la dristribución normal y logística

Ambas gráficas son casi indistinguibles, la única diferencia es que al parecer la función logística tiene las colas más pesadas.

Una aproximación de la regresión logística se puede entender con la función logística:

$$F(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

Si t es visto como una función lineal de una variable dependiente x, la función logística puede ser vista de la siguiente forma:

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

La ecuación anterior es interpretada como la probabilidad de que la variable independiente sea un "éxito". Aplicando la función inversa de la ecuación anterior, obtenemos lo siguiente:

$$F^{-1}(x) = \beta_0 + \beta_1 x = \pi' \tag{1.8}$$

La transformación anterior es conocida como la "transformación logística de la probabilidad π" y está dado por lo siguiente:

$$F^{-1}(x) = log\left(\frac{\pi}{1 - \pi}\right)$$

donde la razón $\pi/(1 - \pi)$ es conocida como la "razón de momios"

La ecuación 1.8 tiene varias características que hacen importante a la función logística:

1. Esta función está acotada entre el 0 y 1, y se acerca asintóticamente a estos límites.
2. Mientras $\beta_1$crece, la ecuación 1.8 toma la forma de una S-sigmoidal, cambiando muy rápido en el centro.
3. Cuando β_1 cambia de positivo a negativo,la ecuación 1.8 cambia de una función monótona creciente a una función monótona decreciente.
4. Por último, notemos la simetría de la ecuación. Si la variable de respuesta es codificada usando $Y' = 1-Y$, esto es, cambiando los unos por los ceros (y viceversa), los signos para todos los coeficientes se invertirán.

En las siguientes secciones del capítulo se explica el desarrollo e interpretación de los modelos de regresión logística.

1.3. Regresión logística simple

Para empezar con este modelo, primero establezcamos un modelo formal de la regresión logística simple. Recalquemos que cuando la variable de respuesta es binaria, tomando los valores 1 y 0 con probabilidad π y 1-π respectivamente, la variable aleatoria Y se distribuye Bernoulli con parámetro $\mathbb{E}[Y] = \pi$. Establezcamos el modelo de regresión logística de la siguiente forma:

$$Y_i = \mathbb{E}[Y_i] + \varepsilon_i$$

Dado que la distribución de los errores dependen de la distribución Bernoulli de la variable Y_i, es preferible establecer el modelo de regresión logística de la siguiente manera: las variables Y_i son variables aleatorias independientes que se distribuyen Bernoulli con valores esperados $\mathbb{E}[Y_i] = \pi_i$ donde:

$$\mathbb{E}[Y_i] = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 \mathrm{X}_i)} \tag{1.9}$$

Las observaciones de la variable X se asumen como constantes conocidas. Alternativamente, si las observaciones de la variable X son aleatorias, $\mathbb{E}[Y_i]$ será vista como una esperanza condicional de cada valor X_i.

1.3.1. Función de máxima verosimilitud

Dado que cada Y_i es una variable aleatoria Bernoulli, donde:

$$\mathbb{P}[Y_i = 1] = \pi_i$$
$$\mathbb{P}[Y_i = 0] = 1 - \pi_i$$

podemos representar su función de densidad como sigue:

$$f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \qquad Y_i = 0, 1; \qquad i = 1, \ldots, n \tag{1.10}$$

Notemos que $f_i(1) = \pi_i$ y que $f_i(0) = 1 - \pi_i$. Por lo tanto, $f_i(Y_i)$ simplemente representa la probabilidad de que $Y_i = 1$ ó 0.

Dado que las observaciones Y_i son independientes, la función de probabilidad conjunta queda de la siguiente manera:

$$g(Y_i, \ldots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \tag{1.11}$$

Será más fácil para después encontrar la estimación de los parámetros por máxima verosimilitud si trabajamos con el logaritmo en la función de probabilidad conjunta:

$$\begin{aligned} \log g(Y_1, \ldots, Y_n) &= \log \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \\ &= \sum_{i=1}^{n} \left[Y_i \log \pi_i + (1 - Y_i) \log(1-\pi_i) \right] \\ &= \sum_{i=1}^{n} \left[Y_i \log \left(\frac{\pi_i}{1-\pi_i} \right) \right] + \sum_{i=1}^{n} \log(1-\pi_i) \end{aligned}$$

Como $\mathbb{E}[Y_i] = \pi_i$ para una variable binaria, se sigue de la ecuación (1.11) que:

$$1 - \pi_i = [1 + \exp(\beta_0 + \beta_1 X_i)]^{-1} \tag{1.12}$$

Además, de la ecuación (1.12) obtenemos:

$$\log \left(\frac{\pi_i}{1-\pi_i} \right) = \beta_0 + \beta_1 X_i \tag{1.13}$$

Por lo tanto, de la ecuación (1.11) puede ser expresada de la siguiente manera:

$$\begin{aligned} \log L(\beta_0, \beta_1) &= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) + \sum_{i=1}^{n} \log([1 + \exp(\beta_0 + \beta_1 X_i)]^{-1}) \\ &= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log(1 + \exp(\beta_0 + \beta_1 X_i)) \end{aligned} \tag{1.14}$$

donde $\log L(\beta_0, \beta_1)$ sustituye a la función $\log g(Y_1, \ldots, Y_n)$ para mostrar específicamente que ahora vemos esta función como la función de verosimilitud con los parámetros que deseamos estimar.

1.3.2. Estimación por máxima verosimilitud

El método de máxima verosimilitud estima los parámetros β_0 y β_1 de la ecuación (1.14), el problema que ahora se presenta es que no existe una solución cerrada para los parámetros a estimar en nuestra función de verosimilitud. Esto será resuelto por métodos numéricos que fueron anteriormente propuestos y que explicaremos brevemente en el apéndice B. Dejaremos entonces a los paquetes estadísticos estimar los parámetros y así poder encontrar b_0 y b_1 que serán los estimadores encontrados por el método de máxima verosimilitud.

Denotemos ahora a $\hat{\pi}_i$ como el valor ajustado (estimado) para el i-ésimo caso:

$$\hat{\pi}_i = \frac{\exp(b_0 + b_1 X_i)}{1 + \exp(b_0 + b_1 X_i)} \tag{1.15}$$

La función logística ajustada (estimada) quedaría de la siguientes manera:

$$\hat{\pi} = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)} \tag{1.16}$$

Si utilizamos la transformación logística vista en (1.8), podemos expresar (1.16) de la siguiente manera:

$$\hat{\pi'} = b_0 + b_1 X \tag{1.17}$$

donde $\hat{\pi'} = \log(\frac{\hat{\pi}}{1-\hat{\pi}})$

Una vez que la función logística ajustada ha sido encontrada, veremos en las siguientes secciones la bondad de ajuste de esta función para poder hacer inferencias sobre los parámetros y predicciones. Antes de eso veremos que interpretación tiene b_1 y como quedaría el modelo de regresión logística múltiple una vez visto el simple.

1.3.3. Interpretación del estimador b_1

La interpretación del coeficiente estimado b_1 en la función logística ajustada (1.16) no es parecida a la que teníamos con el modelo de regresión lineal. La razón es que el efecto de incrementar una unidad en X varía para el modelo de regresión logística ya que depende del lugar donde esté el punto de inicio en X. Una interpretación de b_1 es con la propiedad de la función logística ajustada que dice que las probabilidades estimadas $\frac{\hat{\pi}}{1-\hat{\pi}}$ son multiplicadas por $\exp(b_1)$ por cada unidad incrementada en X.

Para ver esto, fijemos un valor en la función (1.17) de la siguiente manera $X = X_j$

$$\hat{\pi'}(X_j) = b_0 + b_1 X_j$$

La notación $\hat{\pi}'(X_j)$ indica específicamente el nivel en la variable X asociado con el valor estimado. Ahora fijemos un valor en la misma función ahora en $X = X_j + 1$:

$$\hat{\pi}'(X_j + 1) = b_0 + b_1(X_j + 1)$$

La diferencia entre los valores $X_j + 1$ yX_j es:

$$\hat{\pi}'(X_j + 1) - \hat{\pi}'(X_j) = b_1(X_j + 1 - X_j) = b_1$$

Recordando que $\hat{\pi}' = \log(\frac{\hat{\pi}}{1-\hat{\pi}})$, $\hat{\pi}'(X_j)$ es el logaritmo de las probabilidades estimadas cuando $X = X_j$; denotémoslo como $\log(odds_1)$. De la misma manera, $\hat{\pi}'(X_j + 1)$ es logaritmos de las probabilidades estimadas cuando $X = X_j + 1$; denotemoslo como $\log(odds_2)$. Por lo tanto, la diferencia entre los dos valores fijos puede ser expresada de la siguiente manera:

$$\log(odds_2) - log(odds_1) = \log\left(\frac{odds_2}{odds_1}\right) = b_1$$

Multiplicando la ecuación anterior por la función exponencial, vemos que el cociente estimado de probabilidades, llamados razón de momios (odd ratios) y denotado $\widehat{OR}$, es igual a $\exp(b_1)$:

$$\widehat{OR} = \frac{odds_2}{odds_1} = \exp(b_1)$$

1.4. Regresión logística múltiple

El modelo de regresión logística simple (1.9) puede ser fácilmente extendido a más de una variable predictora. De hecho, es mejor tener mas variables predictoras en el modelo para mejorar el modelo a estudiar. Extendamos el modelo de regresión logística simple, simplemente cambiemos $\beta_0 + \beta_1 X$ en () por $\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$. Para simplificar notación, será mejor utilizar notación de matrices con los siguientes tres vectores:

$$\boldsymbol{\beta}_{px1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \boldsymbol{X}_{px1} = \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_{p-1} \end{bmatrix} \qquad \boldsymbol{X}_{i_{px1}} = \begin{bmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{i,p-1} \end{bmatrix} \tag{1.18}$$

Con esto tenemos que:

$$\boldsymbol{X}'\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1}$$
$$\boldsymbol{X_i}'\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{i,p-1}$$

Con esta notación, la función logística (1.9) es extendida a una nueva función logística múltiple de la siguiente manera:

$$\mathbb{E}[Y] = \frac{\exp(\boldsymbol{X}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}'\boldsymbol{\beta})} \tag{1.19}$$

y la ecuación (1.19) puede ser extendida de la siguiente forma:

$$\mathbb{E}[Y] = [1+\exp(-\boldsymbol{X}'\boldsymbol{\beta})]^{-1} \tag{1.20}$$

De la misma manera, la transformación logística (1.8) queda así:

$$\pi' = \boldsymbol{X}'\boldsymbol{\beta} \tag{1.21}$$

donde $\pi' = \log(\frac{\pi}{1-\pi})$

El modelo de regresión múltiple puede ser formulado de la siguiente manera; las Y_i son variables aleatorias independientes que se distribuyen Bernoulli con valores esperados $\mathbb{E}[Y_i] = \pi_i$ donde:

$$\mathbb{E}[Y_i] = \pi_i = \frac{\exp(\boldsymbol{X_i'}\boldsymbol{\beta})}{1+\exp(\boldsymbol{X_i'}\boldsymbol{\beta})} \tag{1.22}$$

Como en el modelo de regresión logística simple, las observaciones X son consideradas como constantes. Si en dado caso X es aleatoria, $\mathbb{E}[Y_i]$ podrá ser visto como una esperanza condicional para cada valor de $X_{i1}, X_{i2}, \ldots, X_{i,p-1}$. La ventaja de este modelo es que las variables predictoras pueden ser cuantitativas o cualitativas. Esta flexibilidad implica que el modelo de regresión múltiple sea muy atractivo y usado para modelar fenómenos científicos como sociales.

1.4.1. Ajuste del modelo

Habiendo obtenido la función logística múltiple (1.22), procederemos a estimar los parámetros por el método de máxima verosimilitud. La función encontrada en la regresión simple (1.14) puede ser extendida directamente al modelo de regresión logística múltiple:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\boldsymbol{X_i'}\boldsymbol{\beta}) - \sum_{i=1}^{n} \log\left[1+\exp(\boldsymbol{X_i'}\boldsymbol{\beta})\right] \tag{1.23}$$

Al igual que en modelo de regresión logística simple, la manera de encontrar los estimadores que maximicen la función $\log L(\boldsymbol{\beta})$ es por medio de métodos numéricos. Estos estimadores serán denotados por $b_0, b_1, \ldots, b_{p-1}$. En forma de vector quedaría de la siguiente manera:

$$\boldsymbol{b}_{px1} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

La función logística ajustada (estimada) y los valores ajustados (estimados) serán expresados como siguen:

$$\hat{\pi} = \frac{\exp(\boldsymbol{X}'\boldsymbol{b})}{1+\exp(\boldsymbol{X}'\boldsymbol{b})} = \left[1+\exp(-\boldsymbol{X}'\boldsymbol{b})\right]^{-1} \tag{1.24}$$

$$\hat{\pi_i} = \frac{\exp(\boldsymbol{X_i}'\boldsymbol{b})}{1+\exp(\boldsymbol{X_i}'\boldsymbol{b})} = \left[1+\exp(-\boldsymbol{X_i}'\boldsymbol{b})\right]^{-1} \tag{1.25}$$

donde:

$$\boldsymbol{X}'\boldsymbol{b} = b_0 + b_1 X_1 + \ldots, + b_{p-1} X_{p-1}$$
$$\boldsymbol{X_i}'\boldsymbol{b} = b_0 + b_1 X_{i1} + \ldots, + b_{p-1} X_{i,p-1}$$

1.5. Inferencias sobre los parámetros de regresión

Al igual que en el modelo de regresión lineal, nos interesa también poder hacer inferencias sobre los parámetros en el modelo de regresión logística, el procedimiento que presentaremos será válido para muestras grandes. Para muestras grandes, los estimadores encontrados por el método de máxima verosimilitud para la regresión logística se distribuirán normal, con un sesgo pequeño o nulo, y con varianzas y covarianzas que son a su vez funciones de las derivadas parciales de segundo orden del logaritmo de la función de verosimilitud.

Denotemos a $\boldsymbol{G}$ como la matriz de las derivadas parciales de segundo orden del logaritmo de la función de verosimilitud (1.22), las derivadas serán con respecto a los parámetros $\beta_0, \beta_1, \ldots, \beta_{p-1}$:

$$\boldsymbol{G}_{pxp} = [g_{ij}] \qquad i = 0, 1, \ldots, p-1; \qquad j = 0, 1, \ldots, p-1 \tag{1.26}$$

donde:

$$g_{00} = \frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \beta_0^2}$$
$$g_{01} = \frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \beta_0 \partial \beta_1}$$
$$etc.$$

Esta matriz es conocida como la matriz hessiana. Cuando las derivadas parciales de segundo orden en la matriz hessiana son evaluadas con los estimadores, i.e $\boldsymbol{\beta} = \boldsymbol{b}$, se puede obtener la matriz de varianzas-covarianzas estimada de la siguiente manera:

$$s^2[\boldsymbol{b}] = (-g_{ij})^{-1} \tag{1.27}$$

Las inferencias sobre los coeficientes del modelo de regresión logística (simple o múltiple) están basadas en el siguiente resultado para muestras grandes:

$$\frac{b_k - \beta_k}{s[b_k]} \sim z \qquad k = 0, 1, \ldots, p-1 \tag{1.28}$$

donde z es la variable aleatoria normal estándar y $s\left[b_k\right]$ es desviación estándar obtenida en (1.27).

1.5.1. Prueba para un sólo coeficiente β_k: La prueba de Wald

Una prueba para muestras grandes para un sólo parámetro de regresión puede ser construida basada en (1.28). Las hipótesis quedarían de la siguiente manera:

$$H_o : \beta_k = 0$$
$$H_a : \beta_k \neq 0$$

la estadística apropiada sería la siguiente:

$$z^* = \frac{b_k}{s\left[b_k\right]}$$

y la regla de decisión quedaría de la siguiente manera:

$$Si \mid z^* \mid \leq z(1 - \alpha/2), concluir\ H_o$$
$$Si \mid z^* \mid > z(1 - \alpha/2), concluir\ H_a$$

La prueba anterior es conocida como la prueba de Wald. En ocasiones, el cuadrado de z^* es usado en vez del valor absoluto, si eso ocurriera, la prueba será basada en una distribución ji-cuadrada con un grado de libertad.

1.5.2. Estimación de intervalos para un sólo coeficiente β_k

De la distribución en (1.28), podemos obtener directamente un intervalo de 1-α para β_k:

$$b_k \pm z(1 - \alpha/2)s\left[b_k\right] \tag{1.29}$$

donde $z(1-\alpha/2)$ es el percentil $(1-\alpha/2)100$ de una distribución normal estándar. El intervalo correspondiente para el radio de probabilidades $\exp(\beta_k)$ es:

$$\exp\left[b_k \pm z(1 - \alpha/2)s\left[b_k\right]\right] \tag{1.30}$$

1.5.3. Prueba para varios coeficientes $\beta_k = 0$: Prueba de razón de verosimilitud

Frecuentemente uno está más interesado en determinar si un subconjunto de las variables predictoras X en el modelo de regresión logística múltiple puede ser desechado, es decir, si esos coeficientes son igual cero. La prueba que utilizaremos es análoga al que se utiliza en el modelo de regresión lineal, esta prueba es llamada "prueba de cociente de verosimilitud" y al igual que en el

modelo de regresión lineal, se basa en comparar modelos completos contra modelos reducidos. Al igual que antes, esta prueba es válida para muestras grandes.

Empecemos primero con el modelo logístico completo:

$$\pi = \left[1 + \exp(-\boldsymbol{X}'\boldsymbol{\beta_C})\right]^{-1} \tag{1.31}$$

donde:

$$\boldsymbol{X}'\boldsymbol{\beta_C} = \beta_0 + \beta_1 + \cdots + \beta_{p-1}X_{p-1}$$

Ahora obtendremos los estimadores máximos verosímiles para el modelo completo, denotados con $\boldsymbol{b}_C$, y después los evaluaremos en la función de verosimilitud $L(\boldsymbol{\beta})$ donde $\boldsymbol{\beta}_{\mathrm{C}} = \boldsymbol{b}_C$. Denotemos este valor de la función de verosimilitud para el modelo completo por $L(C)$. La hipótesis que deseamos probar es la siguiente:

$$\begin{aligned} H_o &: \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0 \\ H_a &: \textit{no todas las } \beta_k \textit{ en } H_o \textit{ son cero} \end{aligned} \tag{1.32}$$

donde, por conveniencia, acomodamos el modelo de tal manera que los últimos $p - q$ coeficientes son aquellos probados. El modelo logístico reducido es el siguiente:

$$\pi = \left[1 + \exp(-\boldsymbol{X}'\boldsymbol{\beta_R})\right]^{-1} \tag{1.33}$$

donde:

$$\boldsymbol{X}'\boldsymbol{\beta_R} = \beta_0 + \beta_1 + \cdots + \beta_{q-1}X_{q-1}$$

De igual forma que con el modelo completo, obtendremos los estimadores máximos verosímiles para el modelo reducido, denotados con $\boldsymbol{b}_R$ y los evaluaremos en la función de verosimilitud con q parámetros donde $\boldsymbol{\beta}_R = \boldsymbol{b}_R$. Denotemos este valor de la función de verosimilitud para el modelo reducido por $L(R)$. La estadística que utilizaremos para la prueba de cociente de verosimilitud, denotada por G^2, es la siguiente:

$$G^2 = -2\log\left[\frac{L(R)}{L(C)}\right] = -2\left[\log L(R) - \log(F)\right] \tag{1.34}$$

notemos que si el cociente de $L(R)/L(C)$ es pequeño, concluyendo H_a, entonces G^2 será grande, esto implica que valores grandes en G^2 hará que aceptemos H_a.

Cuando el tamaño de muestra es muy grande, G^2 se distribuye como una ji-cuadrada con $p - q$ grados de libertad. La regla de decisión para esta prueba quedaría de la siguiente manera:

$$\begin{aligned} &Si\ G^2 \leq \chi^2(1-\alpha; p-q), concluir\ H_o \\ &Si\ G^2 > \chi^2(1-\alpha; p-q), concluir\ H_a \end{aligned} \tag{1.35}$$

1.6. Métodos automáticos para seleccionar el modelo

Durante varios años, diferentes métodos de selección de modelos han sido desarrollados para la regresión logística. Para la presente tesis, se utilizará el método stepwise que será desarrollado a continuación.

1.6.1. Método stepwise

Cuando el número de predictores es demasiado grande, la mejor manera de seleccionar un modelo es utilizar el método stepwise. Éste método es muy parecido al utilizado en los modelos lineales, el único cambio radica en la regla de decisión para añadir o quitar un predictor.

Para la regresión lineal múltiple, la decisión se basa en el estadístico t asociado al coeficiente b_k y a su vez a su p-value. Para la regresión logística, se obtiene un procedimiento análogo basándose en la regla de decisión del estadístico de Wald para el k-ésimo coeficiente y a su vez su p-value.

1.7. Pruebas de bondad de ajuste

Después de ajustar el modelo de regresión logística, éste necesita ser examinado antes de hacer inferencias en los resultados; en particular tendríamos que examinar si la variable de respuesta estimada es monótona y tiene forma sigmoidal. Las pruebas de bondad de ajuste proveen una medida total en el ajuste del modelo.

1.7.1. Prueba Ji-cuadrada para bondad de ajuste

La prueba ji-cuadrada para bondad de ajuste asume que solamente las observaciones Y_{ij} son independientes y que los datos replicados están disponibles. Las hipótesis de interés son las siguientes:

$$\begin{aligned} H_o :&\mathbb{E}\left[Y\right] = \left[1+\exp(-\boldsymbol{X_i'}\boldsymbol{\beta})\right]^{-1} \\ H_a :&\mathbb{E}\left[Y\right] \neq \left[1+\exp(-\boldsymbol{X_i'}\boldsymbol{\beta})\right]^{-1} \end{aligned} \tag{1.36}$$

En muchos casos, un número de observaciones repetidas son obtenidas en varios niveles de la variable predictora X. Denotemos los X niveles en donde las observaciones repetidas son obtenidas por $X_1, \ldots, X_C$ y asumamos que hay n_j respuestas binarias en cada nivel X_j, esto implica que el valor observado de la i-ésima respuesta binaria en X_j es denotada por Y_{ij}, donde $i = 1, \ldots, n_j$ y $j = 1, \ldots, c$. El número de unos en el nivel X_j es denotado por Y_{1j}:

$$Y_{1j} = \sum_{i=1}^{n_j} Y_{ij}$$

y la proporción de unos en el nivel X_j es denotado por p_j:

$$p_j = \frac{Y_{1j}}{n_j}$$

Con lo anterior, el número de casos en j-ésimo nivel con respuesta 1 será denotado por O_{j1} y el número de casos en el j-ésimo nivel con repuesta 0 será denotado por O_{j0}. Debido a que la variable de respuesta Y_{ij} es una variable aleatoria Bernoulli cuyas respuestas son 0 y 1, el número de O_{j1} y O_{j0} están dados por lo siguiente:

$$\begin{array}{ll} O_{j1} = & \sum_{i=1}^{n_j} Y_{ij} = Y_{1j} \\ O_{j0} = & \sum_{i=1}^{n_j}(1 - Y_{ij}) = n_j - Y_{1j} = n_j - O_{j1} \end{array}$$

para $j = 1, \ldots, c$.

Si la función logística de respuesta es apropiada, el valor esperado de Y_{ij}está dada por:

$$\mathbb{E}\left[Y_{ij}\right] = \pi_j = \left[1 + \exp(-\boldsymbol{X_i'\beta})\right]^{-1}$$

y el valor estimado $\hat{\pi_j} = \left[1 + \exp(-\boldsymbol{X_i'b})\right]^{-1}$.

Si la función logística de respuesta es apropiada, el número de casos esperados con $Y_{ij} = 1$ y $Y_{ij} = 0$ para el nivel j-ésimo se estiman de la siguiente manera:

$$\begin{aligned} E_{j1} =& n_j\hat{\pi_j} \\ E_{j0} =& n_j(1 - \hat{\pi_j}) = n_j - E_{j1} \end{aligned}$$

donde E_{j1} denota el número esperado estimado de unos en el nivel j-ésimo y E_{j0} denota el número esperado estimado de ceros en el nivel j-ésimo.

La estadística de prueba es la ji-cuadrada para bondad de ajuste:

$$X^2 = \sum_{j=1}^{c}\sum_{k=0}\frac{(O_{jk} - E_{jk})^2}{E_{jk}}$$

Si la función logística de respuesta es apropiada, X^2se aproxima a una distribución ji-cuadrada χ^2 con c-p grados de libertad cuando $n_j$es demasiado grande y p<c. Valores grandes de la estadística X^2indican que función logística de respuesta no es apropiada. La regla de decisión para probar las alternativas en (1.36), con un nivel de significancia α, quedaría de la siguiente manera:

$$\begin{aligned} &Si\ X^2 \leq \chi^2(1-\alpha; c-p),\ concluir\ H_o \\ &Si\ X^2 > \chi^2(1-\alpha; c-p),\ concluir\ H_a \end{aligned} \tag{1.37}$$

1.7.2. Prueba de la devianza para bondad de ajuste

La prueba de la devianza para bondad de ajuste en los modelos de regresión logística, es completamente análoga a la prueba F es los modelos de regresión lineales. De la misma manera que la prueba ji-cuadrada para bondad de ajuste, se asume que solamente hay c únicas combinaciones de las variables predictoras $X_1, \ldots, X_C$, el número de observaciones binarias repetidas en $X_j$se denota como n_j y la i-esima respuesta en el predictor X_j será denotada por Y_{ij}.

Ésta prueba se basa en la prueba de razón de verosimilitud para el modelo reducido:

$$\mathbb{E}\left[Y_{ij}\right] = \left[1 + \exp(-\boldsymbol{X_j'}\boldsymbol{\beta})\right]^{-1}$$

en contra del modelo completo:

$$\mathbb{E}\left[Y_{ij}\right] = \pi_j \qquad j = 1, \ldots, c$$

El modelo completo en la prueba de la devianza para bondad de ajuste permite una única probabilidad π_j para cada predictor. Este modelo completo en la regresión logística es usualmente referido como modelo saturado.

Recordando la prueba de razón de verosimilitud (1.34), para la prueba de la devianza necesitamos obtener los valores de máxima verosimilitud para el modelo reducido y completo, $L(F)$ y $L(R)$, $L(R)$ es obtenido al ajustar el modelo de regresión logística y los estimadores máximos verosímiles de los c parámetros en el modelo completo están dados por las siguientes proporciones:

$$p_j = \frac{Y_{1j}}{n_j}$$

Denotando como $\hat{\pi_j}$ al estimador de π_j en el nivel X_j, de la misma manera que la estadística para la prueba de razón de verosimilitud se puede obtener lo siguiente:

$$\begin{aligned} G^2 &= -2\left[logL(R) - logL(F)\right] \\ &= -2\sum_{j=1}^{c}\left[Y_j log\left(\frac{\hat{\pi_j}}{p_j}\right) + (n_j - Y_j) log\left(\frac{1-\hat{\pi_j}}{1-p_j}\right)\right] \\ &= DEV(X_0, X_1, \ldots, X_{p-1}) \end{aligned} \tag{1.38}$$

La prueba de razón de verosimilitud anterior es llamada como la devianza, ésta mide las desviación, en términos de $-2LogL$, entre el modelo saturado y el modelo estimado.

Si la función logística es correcta y n_j es suficientemente grande, esto implicaría que la devianza se aproximaría a una ji-cuadrada con c-p grados de libertad. Valores grandes de la devianza indicarán que el modelo estimado no es correcto. Por lo tanto, las hipótesis son las siguientes:

$$H_o : \mathbb{E}[Y] = \left[1 + \exp(-\boldsymbol{X_j'}\boldsymbol{\beta})\right]^{-1}$$
$$H_a : \mathbb{E}[Y] \neq \left[1 + \exp(-\boldsymbol{X_j'}\boldsymbol{\beta})\right]^{-1}$$

así, la regla de decisión será la siguiente:

$$Si\ DEV(X_0, X_1, \ldots, X_{p-1}) \leq \chi^2(1-\alpha; c-p),\ concluir\ H_o$$
$$Si\ DEV(X_0, X_1, \ldots, X_{p-1}) > \chi^2(1-\alpha; c-p),\ concluir\ H_a$$

1.7.3. Bondad de ajuste Hosmer-Lemeshow

Siguiendo con la idea anterior de reagrupar los datos, consideremos ajustar un modelo de regresión logística así como cálculo de todos los predictores estimados $\hat{y}_i$, y agrupemos los individuos de acuerdo al orden de los predictores $\hat{y}_i$, de menor a mayor, por decir.

Por ejemplo, si hay 100 individuos diferentes, cada uno con su predictor estimado $\hat{y_j} \quad j = 1, 2 \ldots, 100$, creamos 10 grupos, el primero con las diez $\hat{y_j}'s$ más bajas, después las diez siguientes y así sucesivamente. De esta manera, se puede crear una (2 x10, en este caso) tabla de números observados de éxitos en cada grupo (o un número promedio, si uno se divide por 10 en este caso) en comparación con la predicción promedio.

El estadístico de Hosmer-Lemeshow quedaría de la siguiente manera:

$$G^2_{HL} = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)} \sim \chi^2$$

donde:

$$n_j = número\ de\ observaciones\ en\ el\ j-ésimo\ grupo$$
$$O_j = \sum_i y_{ij}$$
$$E_j = \sum_i \hat{\pi_{ij}}$$

de

1.7.4. Curva ROC

Antes de desarrollar a fondo la curva ROC, empezaremos con algunos conceptos fundamentales que ayudarán a entender de una mejor manera la construcción en interpretación de las curvas ROC.

Consideremos una prueba de diagnóstico que ayuda a detectar la presencia o ausencia de algún tipo de enfermedad, esta prueba tendrá dos resultados: positivo $T+$o negativo$T-$. De la misma manera, habrá dos resultados si la enfermedad está presente o no: $D+$si esta presente o $D-$si no lo está.

La sensibilidad se define como la proporción de pacientes con la enfermedad presente que hayan tenido un resultado positivo de la prueba diagnóstica; se denota de la siguiente manera $\mathbb{P}[T+|D+]$. La especificidad se define como la proporción de pacientes sin la enfermedad que hayan tenido un resultado negativo de la prueba diagnóstica; se denota de la siguiente manera $\mathbb{P}[T-|D-]$.

Con lo anterior, la sensibilidad y la especificidad solo se basa en un punto de corte para clasificar un resultado de la prueba como positivo. Una manera más completa para describir la exactitud de la clasificación está dada por el área bajo la curva ROC (Receiver Operating Characteristic). Esta curva, que se originó de la teoría de detección de señales, muestra como el receptor funciona con la existencia de alguna señal habiendo ruido. Ésta grafica la probabilidad de detectar una señal verdadera (sensibilidad) y una señal falsa (1-especificidad) para un rango amplio de posibles puntos de corte.

El área bajo la curva ROC, cuyo rango va de cero a uno, provee una medida para discriminar entre aquellos sujetos que experimentan un resultado de interés versus aquellos que no.

Si el objetivo fuera escoger un punto de corte óptimo para clasificar, se debería elegir un punto que maximice tanto la sensibilidad como la especificidad. Este punto puede ser encontrado graficando la probabilidad del punto de corte versus la sensibilidad y la especificidad; donde se crucen, ese será el punto de corte óptimo para la clasificación.

De la misma manera, la gráfica de la sensibilidad versus 1-especificidad sobre todos los posibles puntos de corte es denotada como la curva ROC y el área bajo la curva provee una medida de discriminación. Como regla general tenemos lo siguiente:

- Si ROC=0.5, esto sugiere ningún tipo de discriminación
- Si 0.7≤ROC<0.8, esto sugiere una discriminación aceptable
- Si 0.8≤ROC<0.9, esto sugiere una excelente discriminación
- Si ROC≥0.9, esto sugiere una excepcional discriminación

En la práctica es muy raro (por no decir imposible) observar áreas bajo la curva ROC mayores a 0.9.

1.8. Análisis residual del modelo de regresión logística

El análisis residual para el modelo de regresión logística es un poco más complicado comparado con el modelo de regresión lineal dado que la variable de respuesta Y_i solamente puede tomar los valores 0 y 1. Esto implica que el residual i-ésimo e_i, puede tomar únicamente dos valores:

$$e_i = \begin{cases} 1 - \hat{\pi}_i & si\, Y_i = 1 \\ -\hat{\pi}_i & si\, Y_i = 0 \end{cases} \tag{1.39}$$

Estos residuos claramente no se distribuirán normal y la gráfica de estos residuales vs valores estimados o variables predictoras no brindarán ningún tipo de información.

1.8.1. Residuales de Pearson

Una alternativa para este problema son los residuales de Pearson, éstos se obtienen dividiendo los residuales entre los errores estándar de Y_i, $\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}$. Los residuales de Pearson quedarían de la siguiente manera:

$$r_{P_i} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} \tag{1.40}$$

Los residuales de Pearson están directamente relacionados con estadística de la prueba ji-cuadrada de bondad de ajuste. Para ver esto recordemos ésta estadística:

$$X^2 = \sum_{j=1}^{c}\sum_{k=0}^{1} \frac{(O_{jk} - E_{jk})^2}{E_{jk}} = \sum_{j=1}^{c} \frac{(O_{j0} - E_{j0})^2}{E_{j0}} + \sum_{j=1}^{c} \frac{(O_{j1} - E_{j1})^2}{E_{j1}} \tag{1.41}$$

Para datos binarios, la estadística anterior cambiará de la siguiente forma, $j = 1, c = n, O_{j1} = Y_i, O_{j0} = 1 - Y_i, E_{j1} = \hat{\pi}_i, E_{j0} = 1 - \hat{\pi}_i$, esto implica que (1.41) queda de la siguiente forma:

$$\begin{aligned} X^2 &= \sum_{i=1}^{n} \frac{[(1-Y_i) - (1-\hat{\pi}_i)]^2}{1-\hat{\pi}_i} + \sum_{i=1}^{n} \frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i} \\ &= \sum_{i=1}^{n} \frac{(Y_i - \hat{\pi}_i)^2}{1-\hat{\pi}_i} + \sum_{i=1}^{n} \frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i} \\ &= \sum_{i=1}^{n} \frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)} \end{aligned} \tag{1.42}$$

Por lo tanto, podemos ver que la suma de los cuadrados de los residuales de Pearson (1.40) es numéricamente igual a la estadística de la ji-cuadrada (1.41).

Esto implica que el cuadrado de cada residual de Pearson mide la contribución de cada respuesta binaria en la estadística de la prueba ji-cuadrada.

Los residuales de Pearson no tienen varianza unitaria ya que no hubo variación en los valores ajustados $\hat{\pi}_i$. Un mejor procedimiento es dividir los residuales entre su desviación estándar. Este valor es aproximadamente $\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})}$ donde h_{ii} es el i´-ésimo elemento de la diagonal de la siguiente matriz:

$$\boldsymbol{H} = \hat{\boldsymbol{W}}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{W}}^{\frac{1}{2}} \tag{1.43}$$

donde $\hat{W}$ es una matriz nxn con elementos en su diagonal $\hat{\pi}_i(1-\hat{\pi}_i)$, X pendiente y $\hat{W}^{\frac{1}{2}}$ es una matriz nxn con elementos en su diagonal que son las raíces cuadradas de la diagonal en $\hat{W}$. Los residuales de Pearson studentizados quedan de la siguiente manera:

$$r_{sP_i} = \frac{r_{P_i}}{\sqrt{1-h_{ii}}}, \tag{1.44}$$

Recordemos que en los modelos de regresión lineal múltiple, la matriz gorro satisface la siguiente expresión $\hat{\mathbf{Y}} = \mathbf{HY}$. La matriz gorro para la regresión logística se desarrolló de la misma manera; ésta satisface la expresión $\widehat{\pi}' = \mathbf{HY}$, donde $\widehat{\pi}'$ es un vector (nx1) de predictores.

1.8.2. Residuales de la devianza

El modelo de la devianza (1.38) fue obtenido utilizando la prueba de razón de verosimilitud donde el modelo reducido es el modelo estimado y el modelo completo es el modelo saturado para los datos agrupados. Para los resultados binarios, se toma el número de categorías en X de la siguiente manera: $c = n,\ n_j = 1,\ j = i,\ Y_{1j} = Y_i,\ p_j = Y_{1j}/n_j = Y_i$, y por lo tanto, la expresión (1.38) toma la siguiente forma:

$$\begin{aligned} G^2 &= -2\sum_{i=1}^{n}\left[Y_i log\left(\frac{\hat{\pi}_i}{Y_i}\right) + (1-Y_i)log\left(\frac{1-\hat{\pi}_i}{1-Y_j}\right)\right] \\ &= -2\sum_{i=1}^{n}\left[Y_i log(\hat{\pi}_i) + (1-Y_i)log(1-\hat{\pi}_i) - Y_i log(Y_i) - (1-Y_i)log(1-Y_i)\right] \\ &= -2\sum_{i=1}^{n}\left[Y_i log(\hat{\pi}_i) + (1-Y_i)log(1-\hat{\pi}_i)\right] \end{aligned}$$

dado que $Y_i log(Y_i) = (1-Y_i)log(1-Y_i) = 0$ para $Y_i = 0$ *o* $Y_i = 1$. Esto implica que para datos dicotómicos, el modelo (1.38) queda de la siguiente forma:

$$DEV(X_0, X_1, \ldots, X_{p-1}) = -2\sum_{i=1}^{n}\left[Y_i log(\hat{\pi}_i) + (1-Y_i)log(1-\hat{\pi}_i)\right] \tag{1.45}$$

El residual de la devianza para el i-ésimo caso, denotado por dev_i, se define como la raíz cuadrada de la contribución del i-ésimo caso del modelo expuesto en (1.45):

$$dev_i = sign(Y_i - \hat{\pi}_i)\sqrt{-2\left[Y_i log(\hat{\pi}_i) + (1 - Y_i) log(1 - \hat{\pi}_i)\right]} \tag{1.46}$$

donde el operador *sign* es positivo cuando $Y_i \geq \hat{\pi}_i$ y negativo en el otro caso. Por lo tanto, la suma de los residuales de la devianza al cuadrado es igual al modelo de la devianza (1.45):

$$\sum_{i=1}^{n}(dev_i)^2 = DEV(X_0, X_1, \ldots, X_{p-1} \tag{1.47}$$

Esto implica que el cuadrado de cada residual mide la contribución de cada respuesta dicotómica al estadístico para la prueba de bondad de ajuste de la devianza en (1.45).

1.8.3. Detección de observaciones influyentes

En esta sección se hablará brevemente de 3 medidas que pueden ser usadas para identificar observaciones que influyan en la estimación. Se considerará la influencia de datos binarios individuales en 3 aspectos del análisis:

1. La estadística ji-cuadrada de Pearson
2. El estadístico de la devianza
3. El predictor lineal estimado, $\hat{\pi}_i$

1.8.3.1. Influencia en la estadística ji-cuadrada y en la devianza

Sean X^2 y DEV los estadísticos antes mencionados basados en los datos estimados, y sean X_i^2 y DEV_ilos valores de estos estadísticos cuando el i-ésimo caso es eliminado. El estadístico delta ji-cuadrado se define como el cambio en el estadístico ji-cuadrada cuando el i-ésimo caso es eliminado.

$$\triangle X_i^2 = X^2 - X_i^2$$

De la misma manera, el estadístico delta de la devianza se define análogamente que el estadístico delta ji-cuadrado.

$$\triangle dev_i = DEV - DEV_i$$

La determinación de la enésima estadística delta (ji-cuadrada o devianza) requiere maximizar n veces la verosimilitud, lo cual requiere de mucho trabajo. Para mayor rapidez en el cómputo, se han desarrollado las siguientes aproximaciones:

$$\triangle X_i^2 = r_{SP}^2 \tag{1.48}$$

$$\triangle dev_i = h_{ii} r_{SP}^2 + dev_i^2 \tag{1.49}$$

En resumen, los estadísticos $\triangle X_i^2$ y $\triangle dev_i$ brindan el cambio en el estadístico ji-cuadrada y el de la devianza cuando el i-ésimo caso es eliminado. Por lo tanto, proveen medidas de influencia para cada caso en estos estadísticos.

Capítulo 2

La Encuesta Nacional de Deserción en la Educación Media Superior

La ENDEMS es la primera encuesta que se levanta en México específicamente sobre deserción en la Educación Media Superior y que es representativa a nivel nacional. Por ello, se estima que los resultados aquí presentados proveerán información útil para comprender el fenómeno y diseñar estrategias de prevención y atención a la población estudiantil que cursa este nivel educativo.

Es importante señalar que al tener un enfoque cuantitativo, la encuesta aporta información valiosa acerca de las pautas generales que se manifiestan de este fenómeno a nivel nacional, es decir, los resultados que aquí se presentan servirán para conocer el problema de manera global, para mostrar las características que son comunes a nivel agregado y que habrá que considerar al aproximarse al tema. Por lo tanto, el presente análisis no agota las variadas expresiones que del problema se presentan en los distintos niveles, sean locales, estatales o de instituciones educativas específicas[1].

De acuerdo a la Secretaria de Educación Pública, *se define como desertor a aquella persona que inició el grado o el nivel educativo correspondiente, no lo concluyó y no se encuentra realizando estudios para alcanzar dicha conclusión.* Hay que tener en cuenta que no se debe tomar esta definición como un juicio de valor.

[1] Reporte de la EDNEMS, 2011

2.1. Metodología de la Encuesta

Brevemente en la introducción se explicó la metodología que se utilizó para realizar la Encuesta Nacional de Deserción en la Educación Media Superior (ENDEMS), ya que el motivo de tesis es tratar de explicar los factores de deserción, se explicará de manera general cuál fue la metodología a seguir para poder diseñar la ENDEMS.

La población específica de interés para diseñar la encuesta fueron los jóvenes que desertaron de la Educación Media Superior. Para conocer sus problemas y estar en posibilidad de contrastar de manera efectiva las situaciones a las que esta población se enfrenta, es indispensable contar también con información de los jóvenes que no desertaron de este nivel educativo.

El esquema de muestreo fue probabilista, estratificado, por conglomerados y polietápico. La población objetivo de la ENDEMS estuvo compuesta por jóvenes, hombres y mujeres, de entre 14 y 25 años de edad que residían permanentemente en viviendas particulares ubicadas en localidades con más de 500 habitantes dentro del territorio nacional. Las entrevistas se realizaron entonces a las siguientes subpoblaciones:

1. Los desertores: aquellos jóvenes que iniciaron la Educación Media Superior y al momento de la entrevista no la habían concluido ni se encontraban realizando estudios para concluir este nivel educativo.

2. Los no desertores: aquellos estudiantes que iniciaron la Educación Media Superior y al momento de la entrevista: a) ya la habían terminado, o b) no la habían terminado pero seguían estudiando para completarla.

3. Los no matriculados: aquellos jóvenes que al momento de la entrevista no estaban inscritos en la Educación Media Superior, ya sea porque seguían estudiando y todavía no terminaban la secundaria o porque no estaban estudiando y abandonaron sus estudios en algún momento anterior a la media superior.

Se visitaron 44,000 viviendas donde se entrevistó a 6,472 mujeres y a 6,542 hombres. Las entrevistas se realizaron en el mes de julio de 2011. De este universo se lograron 4,779 entrevistas a jóvenes que nunca se matricularon en el nivel medio superior; 2,549 entrevistas a jóvenes que desertaron de este nivel y 5,686 entrevistas a jóvenes que lo concluyeron o que continúan estudiando.

Considerando el diseño muestral, la distribución porcentual de las subpoblaciones entrevistadas se integró como se indica en la Figura 2.1:

Las edades de los entrevistados fluctuaron entre los 14 y los 25 años. Los entrevistados de 14 años fueron en su mayoría jóvenes no matriculados, como se muestra en la Cuadro 2.1:

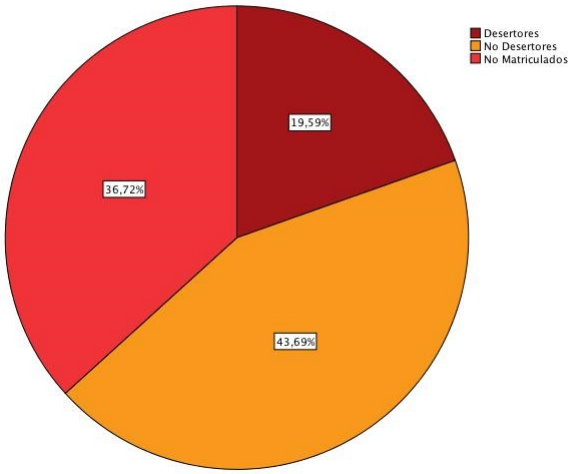

Figura 2.1: Población entrevistada

Edad entrevistados	Desertores (%)	No desertores (%)	No matriculados (%)	Total (%)
14	0.0	0.6	27.4	10.6
15	2.1	5.2	23.9	11.8
16	7.4	14.8	9.2	11.6
17	13.1	16.6	6.1	12.1
18	12.7	15.1	5.3	11.0
19	12.8	11.4	4.3	9.0
20	13.1	9.5	5.3	8.5
21	8.8	7.0	3.9	6.1
22	9.1	6.1	3.6	5.6
23	7.9	5.6	4.4	5.5
24	7.0	4.6	3.3	4.5
25	6.0	3.3	3.4	3.7
NA	0.1	0.1	0.0	0.1

Cuadro 2.1: Edad de los entrevistados

Tiempo	Desertores (%)
3 meses o menos	6.6
Entre 3 y 6 meses	10.6
Entre 6 y 12 meses	16.8
Entre 1 y 2 años	19.7
Entre 2 y 4 años	23.2
Entre 4 y 6 años	13.4
Más de 6 años	9.8

Cuadro 2.2: Tiempo transcurrido entre que dejaron la escuela y la entrevista

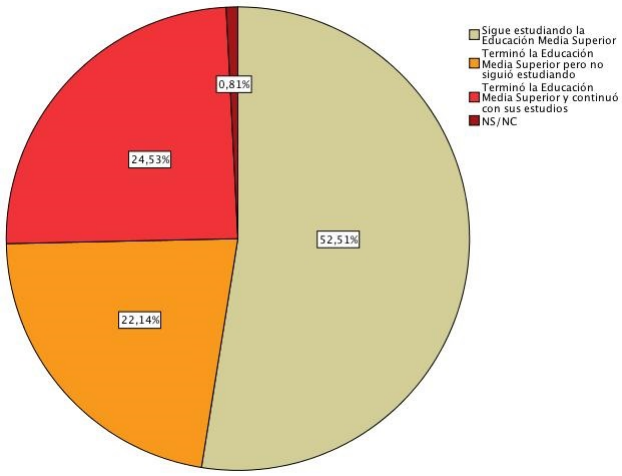

Figura 2.2: Situación académica de los jóvenes no desertores

El tiempo transcurrido desde que los desertores dejaron la escuela hasta el momento de la entrevista (ver Cuadro 2.2) muestra que el 34 % de los entrevistados había abandonado la escuela en los últimos doce meses:

La situación académica de los no desertores al momento de la entrevista se presenta de la siguiente forma:

El 50.8 % de los jóvenes del grupo de no desertores continuaban estudiando la Educación Media Superior al momento del levantamiento. De este grupo, algunos pudieran ser "desertores en potencia", por ejemplo, porque un no desertor de 15 años al momento de la entrevista podría eventualmente desertar a los 17. A pesar de esta consideración su carácter de no desertor al momento de la entrevista no cambia y por este motivo la estadística descriptiva tanto de desertores como de no desertores se presenta con respecto a todos los entrevistados, es decir, considera el grupo total de edades de 14 a 25 años (Figura 2.2).

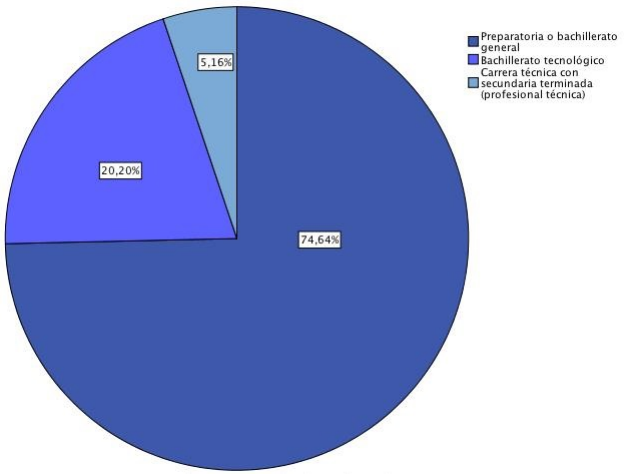

Figura 2.3: Tipo de educación media superior

2.2. Describiendo la ENDEMS

Antes de ajustar el modelo de regresión logística, es importante y conveniente realizar estadística descriptiva para conocer más a fondo algunas variables de la ENDEMS 2011; esta estadística nos ayudará para poder seleccionar las variables que podrían ser incluidos en el modelo que explique la deserción escolar. Las estadísticas presentadas a continuación toman en cuenta los factores de expansión calculados previamente e incluidos en la base de datos, la inferencia tiene representatividad a nivel nacional.

Primero veamos en que tipo de educación media superior se encontraban los desertores (Figura 2.3)

7 de cada 10 jóvenes mexicanos que desertaron pertenecían a una preparatoria o bachillerato general; cabe resaltar que solo el 20 % que alguna vez estuvo inscrito en algún tipo de bachillerato tecnológico.

La ENDEMS 2011 preguntó a los desertores el principal motivo por el que abandonaron los estudios y se les ofrecieron 23 opciones de respuestas posibles, con el fin de que pudieran señalar de manera más precisa la causa de su abandono. En la encuesta se solicitó a los desertores que indicaran, la principal razón que los llevó a abandonar la escuela, aparte de ésto, se les solicitó cuáles otros motivos consideraban importantes (en total fueron tres menciones por encuestado).

Observando la Figura 2.4, "la falta de dinero en el hogar para útiles, pasajes o inscripción", fue mencionada como la principal razón por 36 % de los desertores, e indicada entre las tres principales razones en casi un 50 %. La segunda

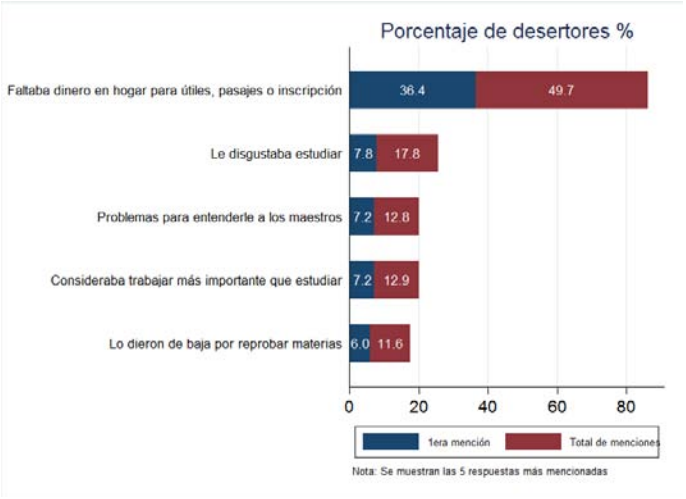

Figura 2.4: Principales motivos para abandonar la EMS (Desertores)

razón mencionada como la principal, fue "le disgustaba estudiar" y la tercera "consideraba trabajar más importante que estudiar". Un 10 % de desertores no mencionaron motivo alguno, porcentaje que no se muestra en el gráfico. Veamos ahora la misma razón principal de deserción por hombres y mujeres.

La Figura 2.5 nos muestra que en el caso de los hombres, "la falta de dinero en el hogar para útiles, pasajes o inscripción" sigue siendo la primera razón para el 39 % de los desertores, también la segunda razón fue que "les disgustaba estudiar" y lo curioso es que para los varones la tercera mención fue que "Lo dieron de baja por reprobar materias" (para el general fue la quinta razón).

En el caso de las mujeres, se sigue la tendencia de "la falta de dinero en el hogar..." como la primera razón para casi 34 % de las mujeres desertoras. No obstante, la segunda razón fue "se embarazó o tuvo un hijo" y la tercera "se casó". Cabe destacar que ninguna de las razones son curriculares, es decir, no tienen que ver con la escuela. Observemos ahora la principal razón de deserción por grupos de edad (Figura 2.7).

Para la Figura 2.7, los jóvenes entre 15 y 19 años señalan como primera razón "la falta de dinero en el hogar para útiles, pasajes o inscripción" (35 %) seguida ¨le disgustaba estudiar" y como tercer razón "lo dieron de baja por reprobar materias"

En el Figura 2.8, entre los jóvenes entre 20 y 25 años es notable la importancia que dan a la falta de dinero, pues 37.3 % manifestó ésta como la principal razón para dejar la escuela. En segundo lugar mencionaron la baja por reprobar

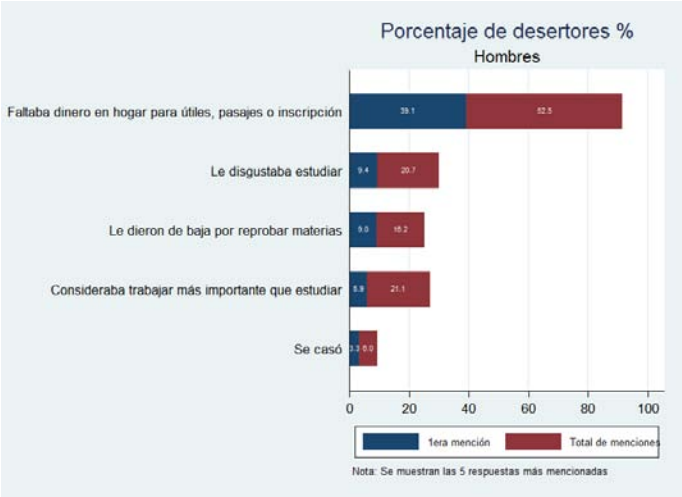

Figura 2.5: Principales motivos para abandonar la EMS (Hombres)

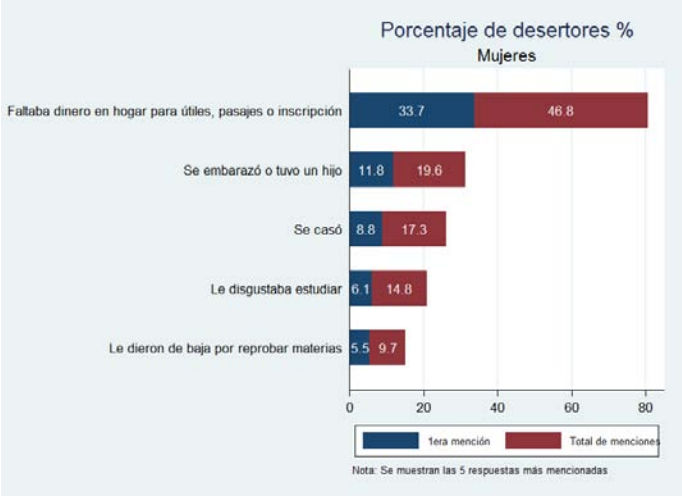

Figura 2.6: Principales motivos para abandonar la EMS (Mujeres)

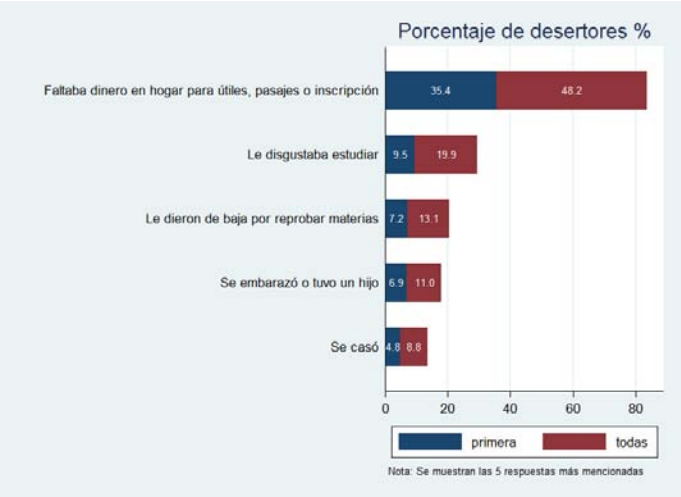

Figura 2.7: Principales motivos para abandonar la EMS (15 a 19 años)

materias y el embarazo o tener un hijo.

Al momento de abandonar la escuela, el 57.9 % de los jóvenes quería seguir estudiando mientras que el 39 % de los jóvenes querían salirse de la escuela (Figura 2.9). Sólo un pequeño porcentaje no sabía la respuesta.

La mayoría de los desertores (62 %) no mencionó a persona alguna como influencia en la decisión de desertar (Figura 2.10). Los jóvenes que mencionaron a alguien reportaron en primer lugar a su pareja y amigos como influencia en su decisión, seguidos de su papá y de su mamá. Observando el Figura 2.11, alrededor del 20 % de los desertores fueron buscados por miembros de su comunidad escolar para convencerlos de seguir estudiando o para conocer sus motivos.

El 41.3 % y el 28.4 % de los jóvenes que desertaron consideran que dejar de estudiar fue una muy mala o una mala decisión respectivamente (Figura 2.12); en conjunto 69.7 % considera que fue una mala o muy mala decisión. También, el 44.8 % de los desertores considera que sus relaciones familiares fueron afectadas negativamente mucho o algo por haber dejado de estudiar (Figura 2.13).

Después de haber visto algunas variables que afectan a los jóvenes desertores, comparemos y veamos algunas diferencias entre los jóvenes desertores y no desertores que nos ayudarán a construir el modelo de regresión logística.

Los padres de los jóvenes no desertores tienen un mayor nivel de estudios que los padres de los desertores. En particular la diferencia se acentúa a partir de la Educación Superior, mientras 25.5 % de los no desertores tienen uno de sus padres con algún grado de licenciatura o mayor, solo 9.6 % de los desertores tienen un padre con esta característica. Por el contrario 65.2 % de los jóvenes desertores tienen padres que no iniciaron la Educación Media Superior, comparado

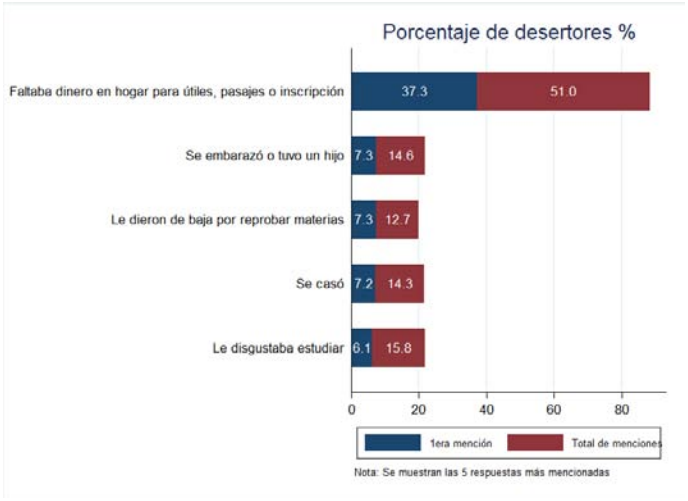

Figura 2.8: Principales motivos para abandonar la EMS (20 a 25 años)

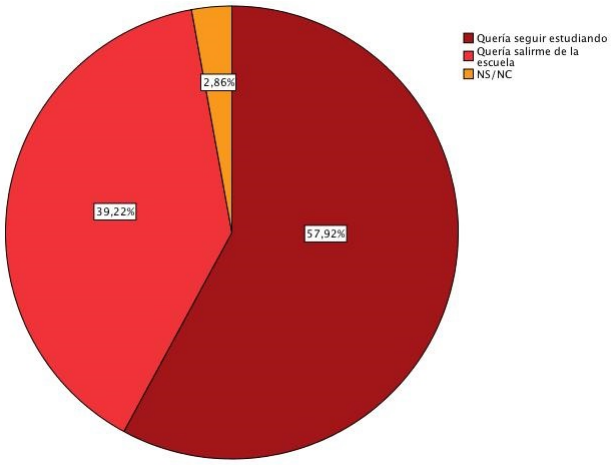

Figura 2.9: ¿Tú querías seguir estudiando o te querías salir de la escuela

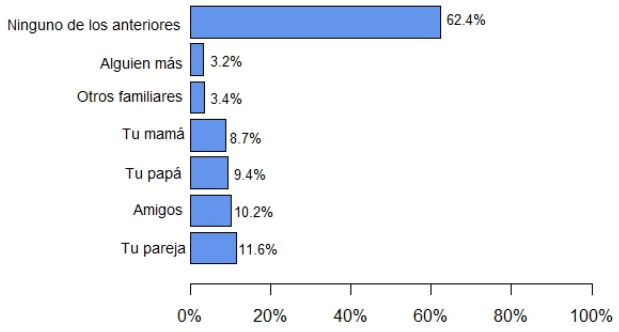

Figura 2.10: ¿Quién influyó en la decisión de que dejaras de estudiar?

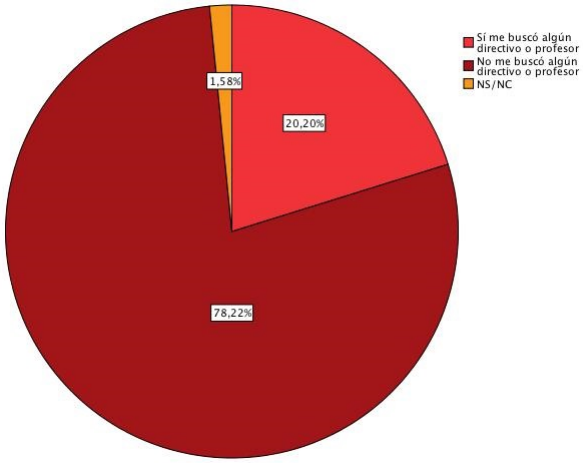

Figura 2.11: ¿Te buscó algún profesor de la escuela o directivo para convencerte de seguir estudiando?

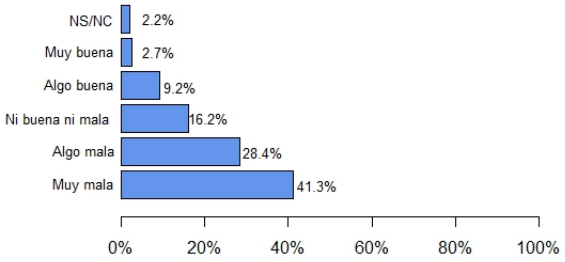

Figura 2.12: ¿Crees que dejar de estudiar fue una buena o una mala decisión?

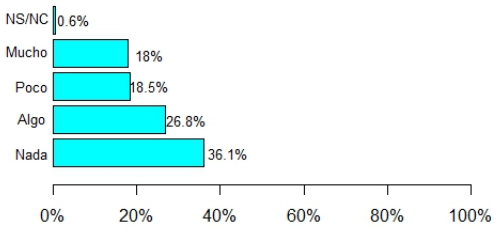

Figura 2.13: ¿Qué tanto crees que haber dejado de estudiar ha afectado negativamente...?

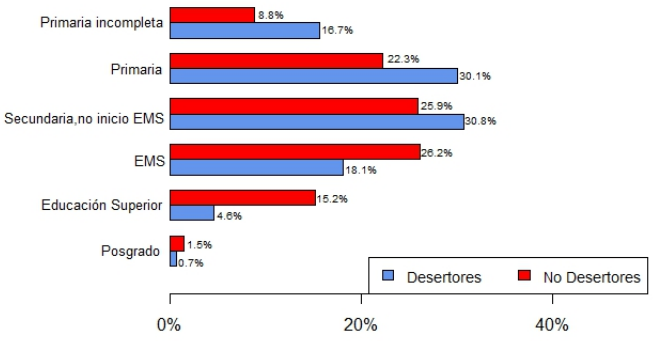

Figura 2.14: Nivel de estudios alcanzado por la madre

con 44.8 % de los jóvenes no desertores con padres con la misma característica. (Figura 2.14 y 2.15)

Observando la Figura 2.16, el 67.4 % de los no desertores dijeron asistir siempre a clases, en contraste con el 42.1 % de los desertores. Faltar con cierta frecuencia o faltar mucho a clases fue dicho por 16.5 % de los desertores en comparación con solo 2.7 % de los no desertores. La reprobación es más frecuente también, entre los jóvenes desertores ya que 26.5 % de ellos dijeron haber reprobado varias materias o más de las permitidas; en contraste con solamente 6.4 % de los no desertores (Figura 2.17).

Para la Figura 2.18, el grupo de jóvenes que desertó observaba un menor promedio en sus calificaciones que el grupo que no desertó. El promedio para los desertores fue de 7.7 en comparación del 8.3 de los no desertores.

En cuanto a becas, el 23.5 % de los jóvenes que no desertaron recibían algún tipo de ayuda durante la educación media superior, en comparación con un 12.5 % de los jóvenes que desertaron que también recibían la ayuda (Figura 2.19). La beca más frecuenta era la del programa "Oportunidades".

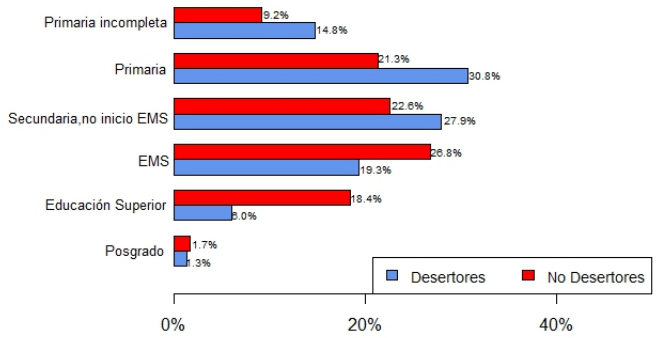

Figura 2.15: Nivel de estudios alcanzado por el padre

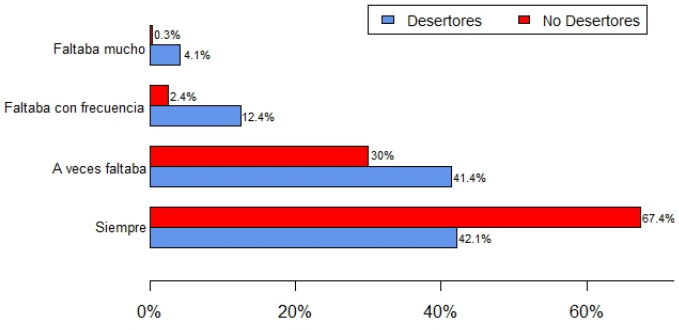

Figura 2.16: Asistencia a clases

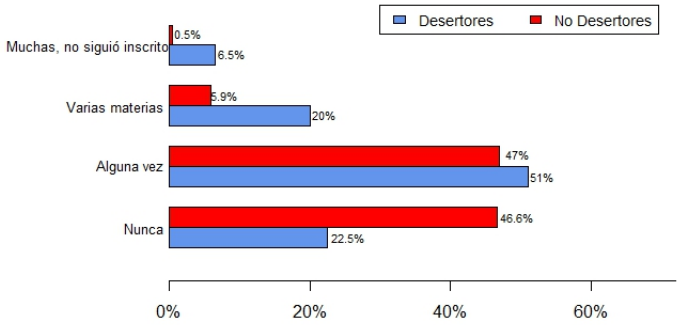

Figura 2.17: Reprobación en la educación media superior

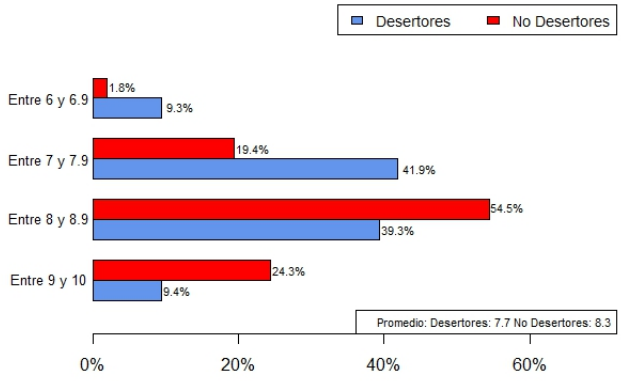

Figura 2.18: Promedio escolar cuando estudiaba en la educación media superior

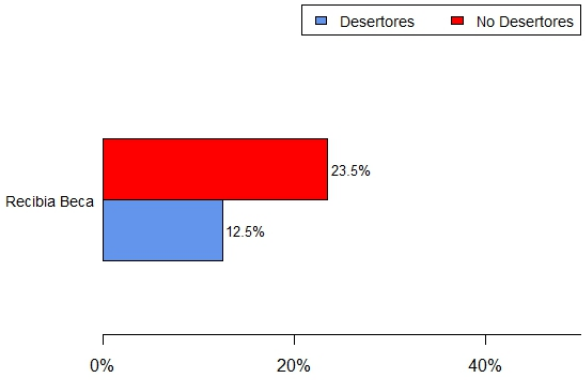

Figura 2.19: Jóvenes que tenían beca mientras estudiaban

Capítulo 3

Aplicando el modelo de regresión logística para la ENDEMS

En el capítulo 2 de la presente tesis, se mostró con estadística descriptiva algunas diferencias importantes entre la población desertora y no desertora, al momento de estar estudiando la Educación Media Superior. Para poder identificar a los jóvenes en riesgo de deserción, y en particular, saber cuales son las características que mejor predicen la deserción de los estudiantes de la Educación Media Superior en México, se utilizó el modelo de regresión logística en el presente capítulo para modelar la deserción.

El Cuadro 3.1 muestra las variables utilizadas en el modelo de regresión logística. La mayoría de las variables toman sólo dos valores posibles (variables dicotómicas) que indican el contar o no con una característica; también hay variables categóricas que como su nombre lo dice, presentan varias categorías dentro de la misma variable.

Durante el desarrollo de este capítulo, se presentará el modelo que mejor predice la deserción escolar; primero se presenta la estimación del modelo de regresión logística para la deserción en la EMS, después se presentan los ajustes al modelo y por último algunas postestimaciones del propio modelo.

3.1. Desarrollando el modelo

Para el desarrollo de esta tesis, el modelo que predice la deserción escolar es el siguiente:

$$\mathbb{P}\left[Y_i = 1\right] = \frac{\exp(\boldsymbol{X_i'}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{X_i'}\boldsymbol{\beta})} \tag{3.1}$$

VARIABLE	DEFINICIÓN
Asistencia	La frecuencia en asistencia a clases, categorizada de la más alta a la más baja
Reprobación	El nivel de reprobación en la EMS, categorizada del más alto al más bajo
Promedio	El promedio en la EMS, categorizada del más alto al más bajo
Beca	Si contaba o no con beca el alumno durante la EMS
Trabajo	Si el alumno consideraba más importante trabajar que estudiar o viceversa
Casarse	Si se casó o no durante la EMS
Embarazo	Se embarazó, embarazó a alguien o tuvo un hijo
Escolaridad de los padres	La escolaridad de los padres más alta, categorizada de la más alta a la más baja

Cuadro 3.1: Variables utilizadas en el modelo

donde $\mathbb{P}[Y_i = 1]$ es la probabilidad de desertar y $X_i,\ i = 1, \ldots, 7$ son las variables predictoras descritas en la Tabla 3.1.

3.1.1. Construyendo el modelo de regresión logística para la deserción

Como primer aproximación para poder explicar la deserción escolar, el Cuadro 3.2 muestra las razones de momios (odd ratios), los p-valores de cada variable y el intervalo de confianza al 95 %; además, también muestra algunos datos del ajuste del modelo que por el momento no se discutirán sino más adelante en este capítulo (el Anexo muestra la codificación de cada variable en el modelo):

Las variables categóricas siempre deben de ser evaluadas en términos de una referencia. La referencia siempre puede ser cambiada como uno quiera, pero en todos los casos, esta variable referencia debe quedar fuera del modelo. El Cuadro 3.2 nos deja interpretar lo siguiente: los alumnos que faltaban con cierta frecuencia a clases tienen 2.45 más veces de posibilidad de desertar que aquellos que asistían con frecuencia; los alumnos que faltaban mucho a clase tienen 8 más veces de posibilidad de desertar que aquellos que asistían con frecuencia.

En cuestión de reprobación de materias, los alumnos que reprobaron varias materias tienen 2.43 más veces de posibilidad de desertar que aquellos que no reprobaron o reprobaron alguna materia; los alumnos que reprobaron más de las materias reprobadas permitidas y no les permitieron continuar en la escuela tienen casi 9 más veces de posibilidad de desertar que aquellos que fueron constantes en sus materias en la escuela.

Para el promedio en la educación media superior, los alumnos que tuvieron un promedio ni alto ni bajo, presentan 2.42 más veces de posibilidad de desertar que aquellos que tuvieron un promedio alto; se puede observar que la razón de momios va aumentando dependiendo del promedio hasta tener casi 8 más veces de posibilidad de desertar aquellos alumnos que tuvieron su promedio dentro de los más bajos en comparación de los más altos.

Logistic	regression	Number of obs =	7851
		LR chi2(15) =	2027.71
		Prob > chi2 =	0.0000
Log likelihood =	3830.7346	Pseudo R2 =	0.2093

Deserción	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
asis1			*Referencia*			
asis2	2.4548	0.3144	7.01	0.000	1.9098	3.1552
asis3	8.0222	2.5439	6.57	0.000	4.3090	14.9353
reprob1			*Referencia*			
reprob2	2.4332	0.2248	9.62	0.000	2.0302	2.9161
reprob3	8.9094	2.2599	8.62	0.000	5.4192	14.6474
prom1			*Referencia*			
prom2	2.4243	0.1555	13.81	0.000	2.1379	2.7491
prom3	3.1735	0.2767	13.24	0.000	2.6750	3.7649
prom4	7.9588	2.7023	6.11	0.000	4.0910	15.4834
Beca	0.4440	0.0363	-9.92	0.000	0.3782	0.5212
Trabajo	2.2692	0.2000	9.3	0.000	1.9093	2.6970
Casarse	3.3246	0.4594	8.69	0.000	2.5359	4.3587
Embarazo	4.2701	0.5525	11.22	0.000	3.3136	5.5027
esc_padres1			*Referencia*			
esc_padres2	0.7619	0.0884	-2.34	0.019	0.6070	0.9564
esc_padres3	0.7055	0.0796	-3.09	0.002	0.5656	0.8800
esc_padres4	0.3879	0.0441	-8.33	0.000	0.3104	0.4848
esc_padres5	0.1861	0.0254	-12.32	0.000	0.1424	0.2431
_cons	0.3335	0.0370	-9.9	0.000	0.2683	0.4145

Cuadro 3.2: Modelo de regresión logística para la deserción escolar

Siempre hay cuestiones sociales y económicas que afectan el desempeño del alumno durante su trayectoria escolar, ejemplo de esto lo vemos a continuación: los alumnos que contaron con beca durante sus estudios tienen 56 % menos de posibilidad de desertar que aquellos que no contaron con beca; los alumnos que consideraron más importante trabajar que estudiar tiene 2.26 más veces de posibilidad de desertar en comparación con aquellos alumnos que pusieron como primordial el estudio antes que el trabajo; adicional a esto, los alumnos que se casaron durante la educación media superior, presentan 3.32 más posibilidades de desertar en comparación con los solteros; por último, los alumnos que embarazaron a alguien o se embarazaron tienen 4.27 más posibilidad de desertar que aquellos que no tuvieron un embarazo durante sus estudios.

Por último, la escolaridad más alta de los padres también influye en la deserción de los alumnos; a más escolaridad, menor posibilidad de desertar. Los alumnos con padres que estudiaron la educación superior (completa o incompleta) tienen 88 % menos de posibilidad de desertar que aquellos alumnos con padres sin primaria.

En primer instancia el modelo parece estar bien ajustado, es decir, todos los p-valores son estadísticamente significativos al 95 %. Sin embargo, puede ser el caso de que el modelo no es apropiado para los datos y los p-valores aparentan contribuir al modelo. O también puede suceder que las variables predictoras contribuyen al modelo pero éste no es apto para los datos utilizados. Para asegurarnos de que el modelo está bien ajustado, en la sección de este capítulo se desarrolló más a fondo esta cuestión.

3.1.2. Coeficientes estandarizados

Estandarizar los coeficientes es un método muy común usado en la regresión lineal, ésto sirve para que las variables predictoras tengan la misma "medida" y por lo tanto puedan ser comparadas. El propósito de esta estandarización es ver que predictores tienen mayor influencia en la variable de respuesta; cada predictor es medido en términos de su desviación estándar y en consecuencia pueden ser comparados entre ellos.

Esta misma lógica se mantiene en la regresión logística. Sin embargo, el hecho de que la variable de respuesta sea binaria y no continua complica un poco las cosas. La versión de los coeficientes estandarizados para la regresión logística se define como sigue:

$$Std\left(\beta_i\right) = \beta_i \frac{SD\left(X_i\right)}{\sqrt{\frac{\pi^2}{3}}}$$

donde β es el parámetro estimado de la variable predictora X y $SD(X)$ es la desviación estándar de las variables predictoras. Hay que notar que $\pi^2/3$ es la varianza fija de la distribución logística. Aplicando ésto a nuestro modelo para explicar la deserción, el Cuadro 3.3 nos muestra los coeficientes estandarizados por cada variable predictora en el modelo:

Variable	Coef	OR	St.Coef	Prob(z)
Constant	-1.0981			
asis2	0.8980	2.4548	0.1122	0.000
asis3	2.0822	8.0222	0.1317	0.000
reprob2	0.8892	2.4332	0.1469	0.000
reprob3	2.1871	8.9094	0.1777	0.000
prom2	0.8856	2.4243	0.2391	0.000
prom3	1.1548	3.1735	0.2253	0.000
prom4	2.0743	7.9588	0.1327	0.000
Beca	-0.8119	0.4440	-0.1758	0.000
Trabajo	0.8194	2.2692	0.1333	0.000
Casarse	1.2014	3.3246	0.1427	0.000
Embarazo	1.4516	4.2701	0.1846	0.000
esc_padres2	-0.2719	0.7619	-0.0595	0.019
esc_padres3	-0.3488	0.7055	-0.0846	0.002
esc_padres4	-0.9469	0.3879	-0.2415	0.000
esc_padres5	-1.6817	0.1861	-0.3340	0.000

Cuadro 3.3: Coeficientes estandarizados

Basándonos en los coeficientes estandarizados, la escolaridad de los padres es lo que más influye en la deserción, teniendo mayor peso los padres con educación media superior y superior; de la misma manera el promedio de los alumnos influye en la deserción, teniendo mayor peso los alumnos con un promedio ni alto ni bajo. Como se dijo antes, estandarizar los coeficientes de las variables predictoras del modelo, ayuda a poder compararlos dado que provienen de diferentes mediciones.

3.1.3. Los métodos stepwise

Los métodos stepwise ha jugado un papel importante en las ciencias sociales. Esto principalmente se debe a que los investigadores en el área de ciencias sociales generalmente ponen más de 50 variables en el modelo y utilizan estos métodos para desechar variables que no contribuyen de manera significativa a la variable de respuesta. Esto generalmente pasa cuando la variable de respuesta, o el objeto del estudio relacionada a la variable de respuesta, es relativamente nueva y no existen estudios que determinen la relación entre las variables.

Los métodos stepwise proveen de una vía rápida para desechar variables cuando no hay evidencia de que estén relacionadas a la variable de respuesta o a las variables predictoras. Este método puede ahorrar tiempo y dinero. Hay varios métodos stepwise, los más comunes son los siguientes:

- Método forward

- Método backward
- Método backward y forward

Existen también mecanismos para forzar un número específico de variables predictoras en el modelo. Generalmente esto pasa cuando se cuenta con un grupo de variables que se desean mantener en el modelo. Uno quisiera mantener aquellas variables predictoras que pueden influir en la respuesta, o creer que dado estudios anteriores estas variables pueden influir en la variable de respuesta, y por lo tanto, después de algún metodo stepwise que haya desechado variables ayudará a contar con un modelo final.

La clave para usar los métodos stepwise es el criterio usado para la inclusión o exclusión, En la regresión lineal generalmente se usan las estadísticas R^2 o la F como criterio de selección. Probablemente el criterio más popular para los modelos de regresión logística es el p-valor. En cualquier etapa del desarrollo del modelo, las variables predictoras que tengan un p-valor mayor al nivel de confianza especificado, serán excluidas del modelo. De la misma manera, las variables predictoras que tengan un p-valor menor al nivel de confianza especificado, serán retenidos en el modelo.

El método forward (utilizado en esta tesis) comienza solamente con una constante en el modelo, después una variable predictora es añadida en el modelo y el modelo resultante es re-evaluado decidiendo si se mantiene o se rechaza el predictor. A continuación otro predictor es añadido al modelo y se evalúa en conjunto con el predictor previamente retenido, y es excluido o retenido. Este procedimiento continua hasta totalizar las variables predictoras que serán evaluadas.

Para nuestro propósito, el método forward fue utilizado para retener o excluir las variables predictoras en nuestro modelo de deserción antes descrito. El Cuadro 3.4 muestra el algoritmo utilizado para nuestro modelo de deserción, utilizando un nivel de confianza al 95 %:

El algoritmo anterior del método forward, muestra las mismas variables predictoras dentro del modelo propuesto para predecir la deserción escolar (Tabla 3.2). Si consideramos alguna variable que pueda contribuir significativamente al modelo, siempre es posible evaluarla utilizando los resultados de los métodos stepwise. Para el ajuste del modelo, es necesario evaluarlo con algún tipo de bondad de ajuste, ésto se verá en las secciones siguientes (requisito necesario para obtener el modelo final).

Begin with empty model				
p = 0.0000 < 0.0500 adding	reprob2			
p = 0.0000 < 0.0500 adding	Embarazo			
p = 0.0000 < 0.0500 adding	esc_padres5			
p = 0.0000 < 0.0500 adding	reprob3			
p = 0.0000 < 0.0500 adding	Trabajo			
p = 0.0000 < 0.0500 adding	Beca			
p = 0.0000 < 0.0500 adding	esc_padres4			
p = 0.0000 < 0.0500 adding	prom3			
p = 0.0000 < 0.0500 adding	prom2			
p = 0.0000 < 0.0500 adding	Casarse			
p = 0.0000 < 0.0500 adding	prom4			
p = 0.0000 < 0.0500 adding	asis2			
p = 0.0000 < 0.0500 adding	asis3			
p = 0.0451 < 0.0500 adding	esc_padres3			
p = 0.0191 < 0.0500 adding	esc_padres2			
Logistic regression		Number of obs	=	7851
		LR chi2(15)	=	2027.71
		Prob > chi2	=	0
Log likelihood = -3830.7346		Pseudo R2	=	0.2093

Cuadro 3.4: Método stepwise para seleccionar el modelo

3.2. Análisis para el ajuste del modelo

Ajustar un modelo de regresión tradicionalmente ha sido realizado por varios estadísticos o analizando los residuales. Dado que el ajuste de modelo es muy importante para obtener un modelo final, en esta sección se discutirán las siguientes categorías (previamente en el capítulo 2 se habló de la parte teórica) para ajustar el modelo en nuestra predicción de la deserción:

- Pruebas de ajuste tradicionales
- Prueba de bondad de ajuste Hosmer-Lemeshow
- Análisis de residuales

3.2.1. Pruebas de ajuste tradicionales para la regresión logística

Los estadísticos de ajuste fueron primeramente desarrollados para los modelos lineares generalizados (GLM) para la regresión logística. Estos fueron por primera vez integrados en los paquetes estadísticos para los modelos lineares generalizados, aproximadamente en 1970. Recordemos que los modelos lineares generalizados utilizan el algoritmo de los mínimos cuadrados iterativamente reponderados (IRLS) para la estimación de los parámetros, errores estándar y

Logistic regression	Number of obs	=	7851
	LR chi2(15)	=	2027.71
	Prob > chi2	=	0
Log likelihood = -3830.7346	Pseudo R2	=	0.2093

Cuadro 3.5: Pseudo R^2y coeficiente de verosimilitud obtenidos en el modelo

algunos estadísticos. Las pruebas de ajuste tradicionales para la regresión logística son las siguientes:

- Estadístico pseudo R^2
- Prueba del coeficiente de verosimilitud

El estadístico estándar para la bondad de ajuste la regresión lineal por mínimos cuadrados por excelencia es el estadístico R^2, también conocido como coeficiente de determinación. Recordando lo visto en el capítulo 2, la pseudo R^2 utilizada para la regresión logística es la siguiente:

$$Pseudo\ R^2 = 1 - LL_F/LL_C$$

donde: $LL_C = modelo\ de\ regresión\ con\ la\ constante$ y
$LL_F = modelo\ de\ regresión\ completo$

Recordemos ahora que la prueba del coeficiente de verosimilitud nos ayuda a incluir o no variables predictoras en el modelo. La prueba de hipótesis también puede ser formulada con la función log-verosimilitud reducida estando unicamente la constante en el modelo:

$$-2\left[LL_C - LL_F\right]$$

Este estadístico tiene una distribución ji-cuadrada con sus grados de libertad iguales a la diferencia entre el número de predictores entre los dos modelos (reducido y completo). Usualmente comparamos un modelo con determinados predictores versus el modelo reducido. Cuando los predictores son significativos (p-valores$<$.05), el p-valor del estadístico ji-cuadrada es cercano a 0.

En nuestro modelo de regresión para tratar de explicar la deserción, la estadística pseudo R^2 y prueba del coeficiente de verosimilitud del modelo se muestran en el Cuadro 3.5:

Con los datos anteriores (Cuadro 3.5), se puede concluir que el modelo está bien ajustado.

3.2.2. Prueba de bondad de ajuste Hosmer-Lemeshow

La estadística para bondad de ajuste más importante para el modelo de regresión logística es sin duda la Hosmer-Lemeshow, junto con su respectiva tabla de deciles. La tabla divide el rango de probabilidades, μ, en grupos. Hosmer y Lemeshow recomiendan 10 grupos para datos relativamente grandes, pero el número puede variar dependiendo de los datos. El objetivo es proveer un contador, para cada grupo respectivo, de cada unos observados, de unos esperados al igual que ceros esperados y ceros observados. Teóricamente, los valores esperados y observados contados deben ser muy parecidos.

Basada en la distribución ji-cuadrada, el estadístico Hosmer-Lemeshow (H-L) es construido para medir la total correspondencia de los valores contados. A menor estadístico H-L, menor varianza en el modelo ajustado y mejor ajustado será el modelo.

Un problema que muchas veces se presenta son los empates (múltiples valores iguales); cuando esto sucede, seleccionar una partición de los deciles induce aleatoriedad a la estadística. Es necesario asegurarse que esta aleatoriedad no afecte afecte una aplicación particular a la prueba de hipótesis.

El método más común para combatir la dificultad de los empates es utilizar diferente número de grupos. Si no hay cambios sustanciales en la estadística después de agrupar, la estadística H-L será confiable y se podrá concluir que el modelo estará bien ajustado.

Para el modelo de deserción, el estadístico H-L se observa en el Cuadro 3.6 habiendo dividido en 10 grupos:

El estadístico Hosmer-Lemeshow es relativamente bajo (13.89), resultando en un p-valor de .0846. Tanto el estadístico como el p-valor del modelo infieren que este modelo están bien ajustados. Observando el Cuadro 3.6, pongamos atención en los valores observados y los valores esperados en cada grupo. Los ceros observados son muy parecidos a los ceros esperados para casi todos los grupos, siendo la diferencia mayor en los grupos 1,9 y 10. De nuevo, el menor ajuste total implica un estadístico H-L pequeño y un p-valor$>$0.05.

3.2.2.1. Matriz de clasificación

La matriz de clasificación es una de las primeras pruebas de bondad de ajuste utilizada en la regresión logística. Originalmente fue utilizada en el análisis de discriminantes, de hecho, dado que la distribución logística restringe los valores predichos a el rango de probabilidad 0-1, el análisis de discriminantes ha usado la distribución logística en su matriz de clasificación.

Logistic modelo for Deserción, goodness-of-fit-test

(Table collapsed on quantiles of estimated probabilities)

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0548	33	45.3	830	817.7	863
2	0.1146	112	121.5	1021	1011.5	1133
3	0.1308	61	56.6	381	385.4	442
4	0.2021	137	139.5	611	608.5	748
5	0.2388	248	244.7	839	842.3	1087
6	0.3073	113	118.4	327	321.6	440
7	0.3818	416	407.2	697	705.8	1113
8	0.4471	237	230.2	299	305.8	536
9	0.6644	431	403.9	277	304.1	708
10	0.9943	626	646.7	155	134.4	781

number of observations = 7851

number of groups = 10

Hosmer-Lemeshow chi2(8) = 13.89

Prob > chi2 = 0.0846

Cuadro 3.6: Bondad de ajuste Hosmer-Lemeshow

La matriz de clasificación se basa en un punto de corte, que responde a la siguiente pregunta, ¿Cuál es la probabilidad optima de separar los valores predichos versus valores observados y fallos?. Dado que las definiciones de sensibilidad y especificidad están basadas en estas relaciones, éstas se usan para determinar el punto de corte.

Una manera de determinar el punto de corte es correr el modelo de regresión, predecir un valor μ que es la probabilidad de éxito (en nuestro caso de desertar), y utilizar la media de μ como punto de corte. A continuación se presenta la media de μ con el modelo ajustado:

Mean estimation			Number of obs	7851
	Mean	Std. Err. [95% Conf.	[95% Conf.	Interval]
mu	0.3074768	0.0026041	0.302372	0.3125815

El punto de corte óptimo para nuestro modelo de deserción es aproximadamente 0.30. Después de haber encontrado el punto óptimo, se graficaron los valores de sensibilidad y especificidad contra probabilidad del punto de corte. La Figura 3.1 muestra el rango de los valores de sensibilidad y especificidad que serán observados si calculamos una tabla de clasificación para diferentes puntos de corte entre 0 y 1.

El gráfico 3.1 muestra que el punto de corte es alrededor de .30 (muy parecido al que habíamos calculado antes); se muestra a continuación los valores de

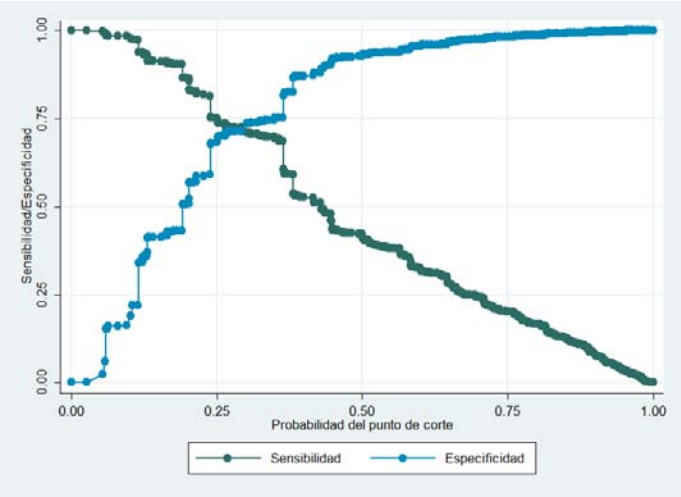

Figura 3.1: Sensibilidad/Especificidad

la sensibilidad, especificidad y la probabilidad de los puntos de corte donde la sensibilidad y especificidad se encuentran muy cercanos:

	se	sp	cutp
440	0.724938	0.714181	0.28116
441	0.724938	0.714548	0.291072
442	0.725352	0.713629	0.270409
443	0.725352	0.713813	0.272771
444	0.726595	0.712157	0.269691
445	0.730323	0.707927	0.264863

La probabilidad del punto de corte donde la sensibilidad y la especificidad se encuentran más cercanos es aproximadamente .29, éste valor es el punto de corte óptimo para la matriz de clasificación. Al igual que antes, el valor es muy parecido al calculado previamente.

La matriz de clasificación (Cuadro 3.7) muestra que nuestro modelo para explicar la deserción clasifica el 71.77 % correctamente. Este es un valor aceptable, reforzando lo visto anteriormente de que nuestro modelo está bien clasificado. Adicionalmente a nuestra matriz de clasificación, la sensibilidad y especificidad ayudan a construir curvas ROC.

3.2.2.2. Curva ROC

La curva ROC (Receiver Operator Characteristic) es generalmente utilizada para que el modelo de regresión logística clasifique casos. Recordando que la

Logistic modelo for Deserción

	True		
Classified	D	~D	Total
+	1750	1552	3302
-	664	3885	4549
Total	2414	5437	7851
Classified + if predicted Pr(D)	>= .29		
True D defined as Deserción	!= 0		
Sensitivity		Pr(+|D)	72.49 %
Specificity		Pr(-|~D)	71.45 %
Positive predictive value		Pr(D|+)	53.00 %
Negative predictive value		Pr(~D|-)	85.40 %
False + rate for true ~D		Pr(+|~D)	28.55 %
False - rate for true D		Pr(-|D)	27.51 %
False + rate for classified +		Pr(~D|+)	47.00 %
False - rate for classified -		Pr(D|-)	14.60 %
Correctly classified			71.77 %

Cuadro 3.7: Matriz de clasificación obtenida en el modelo

curva ROC está definida como la sensibilidad (eje vertical) entre la especificidad inversa (eje horizontal). Un modelo sin un valor predicho tendrá una pendiente igual a 1, resultando en un área bajo la curva de 0.5.

En nuestro modelo de deserción, la curva ROC (Figura 3.2) tiene un p-valor de 0.7959, reforzando la idea de que nuestro modelo está bien ajustado.

3.2.3. Análisis de residuales

El análisis de residuales es una parte importante para el ajuste del modelo y es una de las maneras más antiguas que se han utilizado para validar modelos de regresión. El análisis de residuales para el modelo de regresión logística se basan en los residuales de los modelos lineales generalizados; con estos modelos (GLM) se calculan los residuales para analizar el ajuste del modelo de deserción escolar.

3.2.3.1. Residual de Pearson

El residual de Pearson se calcula restando el valor observado menos el valor predicho, ajustado simplemente por la raíz cuadrada de la varianza. Recordando que la varianza de una variable aleatoria binomial es $np(1-p)$ donde p es la probabilidad de éxito, el residual de Pearson queda de la siguiente forma:

$$r^P = (y - \mu)/\sqrt{n\mu(1-\mu)} \tag{3.2}$$

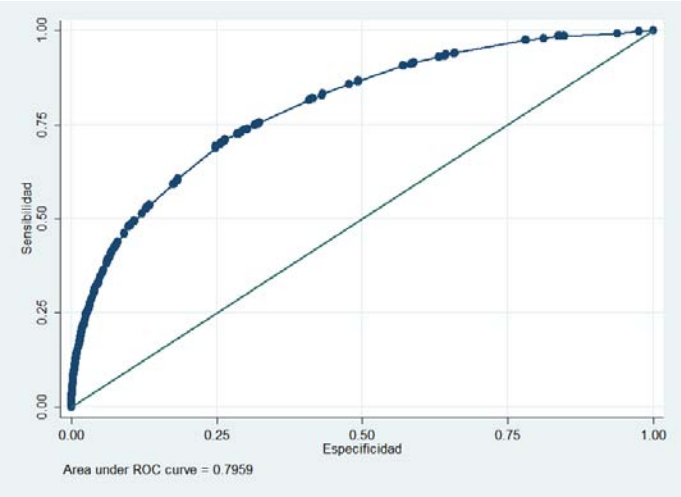

Figura 3.2: Área bajo la curva ROC

	Deserción	Mu	Varianza	Pearson
1	Desertor	0.291	0.2063	-1.5955
2	Desertor	0.8165	0.1497	-0.596
3	Desertor	0.7524	0.1862	1.1472
4	Desertor	0.565	0.2457	1.2407
5	Desertor	0.2269	0.1754	0.7254

Cuadro 3.8: Primeros cinco residuales de Pearson estimados en el modelo

El estadístico ji-cuadrado de Pearson, se define en términos de los residuales de Pearson sumando el cuadrado de éstos. Este estadístico era popular para la bondad de ajuste del modelo, pero ha perdido uso en los años recientes debido a su sesgo. El Cuadro 3.8 muestra los primeros cinco residuales correspondientes a los primeros cinco individuos del modelo de deserción escolar.

3.2.3.2. Residual de Pearson estandarizado

Los residuales de Pearson estandarizados son normalizados a una desviación estándar de 1. Esta normalización es aproximada dividiendo el residual de Pearson entre $\sqrt{1-h}$ donde h es una medida de influencia de un predictor al modelo. El término estandarizado puede ser interpretado como un ajuste entre la correlación que existe entre el valor observado y el valor predicho. El Cuadro 3.9 compara los residuales no estandarizados y estandarizados de Pearson de los primeros cinco individuos.

	R. Pearson	R. Estand. Pearson
1	-1.5955	-1.8199
2	-0.596	-0.6058
3	1.1472	1.1584
4	1.2407	1.2462
5	0.7254	0.7545

Cuadro 3.9: Residuales de Pearson estandarizados (primeros cinco individuos)

Al observar la tabla anterior, la diferencia entre los residuales estandarizados y los no estandarizados es mínima.

3.2.3.3. Residual de la devianza estandarizado

Los residuales de la devianza es uno de los residuales más importantes usados en los modelos de regresión basados en los modelos lineares generalizados. Estos residuales casi siempre se distribuyen normal en comparación con los otros residuales utilizados en el análisis. Una gran ventaja de los residuales estandarizados es que pueden ser graficados e interpretados, de tal manera que en un diagrama de dispersión, se pueden graficar los residuales estandarizados de la devianza versus el valor predicho.

	Deviance	Sdeviance
1	-1.632128	-1.861672
2	-0.572103	-0.5814926
3	1.508498	1.523279
4	1.511065	1.51778
5	0.7090576	0.7375798

De nuevo, la diferencia entre la devianza estandarizada y la devianza es mínima. Para graficar los residuales de la devianza estandarizada, es conveniente elevar al cuadrado estos residuales y compararlos con los valores predichos; los valores por arriba de 4 serán considerados como "outliers", en nuestro modelo, son pocos los "outliers" para aquellos alumnos desertores y no desertores (Figura 3.3).

3.2.3.4. Matriz gorro m-asintótica

Los residuales m-asintóticos son aquellos que se basan en los patrones de las covariables en vez de las observaciones individuales. Previamente se discutió en el capítulo 2 la matriz gorro para estandarizar estadísticos, ahora se expande la discusión para darle un contexto m-asintótico.

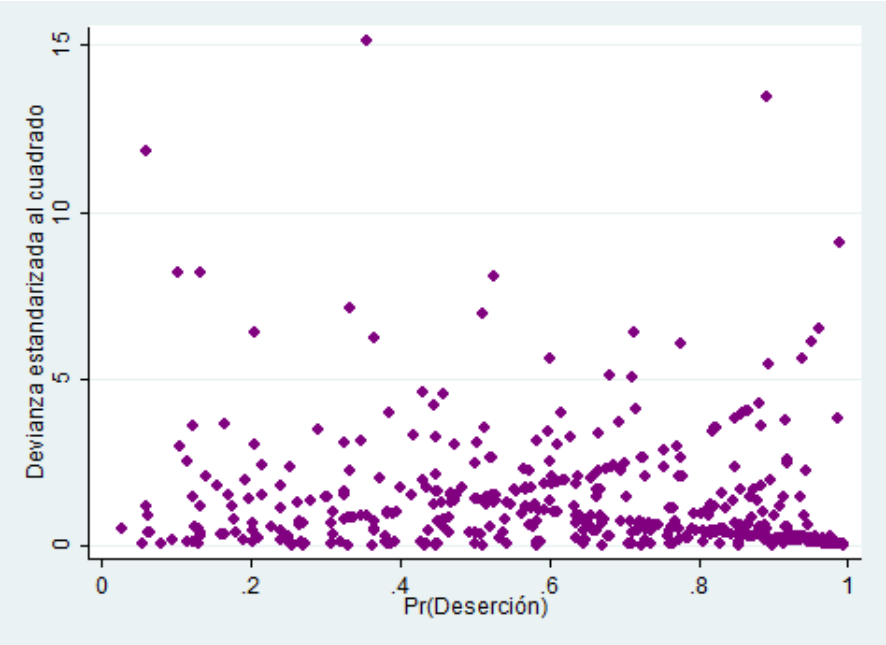

Figura 3.3: Residuales de la devianza estandarizada

Valores grandes de la matriz gorro indican que los patrones en las covariables se encuentran lejos del promedio de estos patrones, independientemente de los residuales. Cuando los valores de la matriz gorro son graficados versus los residuales de Pearson estandarizados, valores en los extremos horizontales se entienden como residuales altos. De la misma manera, valores grandes de la matriz gorro también se entienden como residuales altos.

En el modelo de deserción escolar, la Figura 3.4 muestra pocos residuales mayores a +-2, es decir, el modelo de deserción está bien ajustado.

3.2.3.5. Otros residuales m-asintóticos

Los residuales DeltaX indican la disminución en el estadístico ji-cuadrado de Pearson debido a la eliminación de los patrones en las covariables, asociados con su respectivo residual DeltaX (en el Capítulo 2 se dearrolló más a fondo este residual). La Figura 3.5 muestra el residual DeltaX al cuadrado versus los valores predichos, de la misma manera, valores de los residuales por arriba de 4 son considerados como "outliers" (esto se debe a que el percentil 95 de una distribución ji-cuadrada es 3.84).

La medida dbeta (o db), sirve para evaluar el efecto de eliminar un individuo en el valor de los coeficientes estimados. La medida de Pregibon sirve

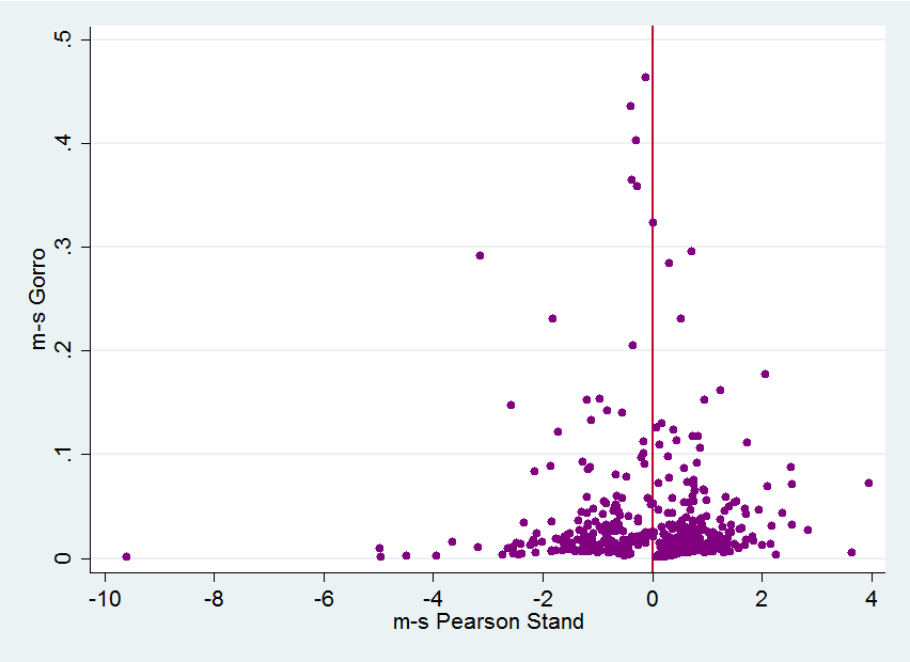

Figura 3.4: Residuales m-asintóticos estimados en el modelo

para observar el cambio en el vector de coeficientes estimados que se produciría al eliminar individuos de la muestra; esto se puede ver al graficar los residuales DeltaX al cuadrado versus la probabilidad de desertar, pero ahora siendo ponderados por la medida dbeta. En la Figura 3.6 podemos observar que solo algunos de los outliers tienen un peso que podría influenciar en las estimaciones. Con esto podemos concluir que el modelo sigue estando bien ajustado

3.3. Resultados

Después de haber analizado las causas de deserción escolar en nuestro país con los datos del 2012, esta sección analiza el mismo modelo de regresión logística pero tomando en cuenta el diseño de muestreo complejo utilizado en la presente encuesta. la Cuadro 3.10 muestra las razones de momios (odd ratios), los p-valores de cada variable y el intervalo de confianza al 95 %, tomando en cuenta el diseño de muestra (estratos, unidades primarias de muestreo y factores de expansión).

De la misma manera que el Cuadro 3.1, los odd ratios antes mostrados no cambian prácticamente en comparación con el modelo de regresión sin el diseño de muestreo compleja; hay casi 10 más veces de posibilidad de desertar si no se asistía con regularidad a clases en comparación con los que asistían frecuen-

Survey:	Logistic regression	Number of obs =	7851
Number of strata =	29	Population size =	13366.82
Number of PSUs =	541	Design df =	512
		F(15, 498) =	57.83
		Prob > F =	0

Deserción	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
asis1	*Referencia*					
asis2	2.7889	0.4205	6.8	0.000	2.0739	3.7505
asis3	9.7821	3.6828	6.06	0.000	4.6689	20.4952
reprob1	*Referencia*					
reprob2	2.3241	0.2405	8.15	0.000	1.8966	2.8480
reprob3	8.6526	2.5246	7.4	0.000	4.8775	15.3494
prom1	*Referencia*					
prom2	2.4456	0.1757	12.45	0.000	2.1237	2.8164
prom3	3.1226	0.3178	11.19	0.000	2.5567	3.8136
prom4	7.0114	2.9139	4.69	0.000	3.0989	15.8636
Beca	0.4712	0.0445	-7.96	0.000	0.3914	0.5673
Trabajo	2.4418	0.2821	7.73	0.000	1.9460	3.0639
Casarse	3.0214	0.5524	6.05	0.000	2.1096	4.3271
Embarazo	3.8116	0.6773	7.53	0.000	2.6884	5.4041
esc_padres1	*Referencia*					
esc_padres2	0.6902	0.0926	-2.76	0.006	0.5303	0.8982
esc_padres3	0.6650	0.0893	-3.04	0.003	0.5107	0.8659
esc_padres4	0.3692	0.0482	-7.63	0.000	0.2857	0.4771
esc_padres5	0.1716	0.0278	-10.89	0.000	0.1249	0.2358
_cons	0.2619	0.0340	-10.32	0.000	0.2030	0.3380

Cuadro 3.10: Odd Ratios del modelo de regresión logística para la deserción (diseño de muestreo complejo)

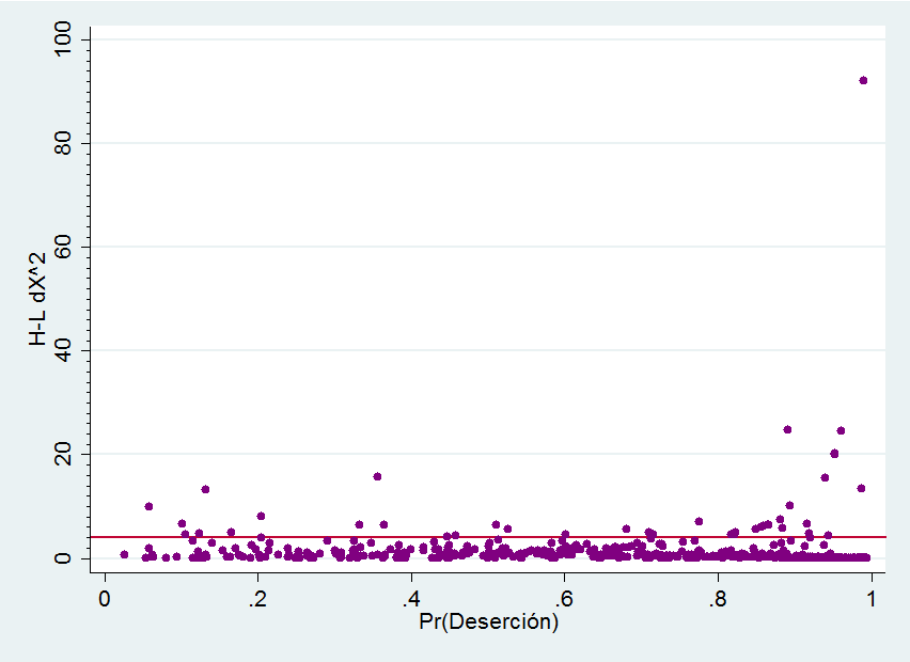

Figura 3.5: Residuales Delta-X al cuadrado

temente, de la misma manera quien tenía de los peores promedios tiene 7 más veces de posibilidad de desertar en comparación con los que tenían los mejores promedios. Por último, hay 93 % de posibilidad de no desertar si tus padres tienen una escolaridad alta.

Todos las variables predictoras son estadísticamente significativas en el modelo, ya que una a una su p-valor es menor a .05; con esto podemos concluir que no hay diferencia en las estimaciones hechas utilizando o no el diseño complejo.

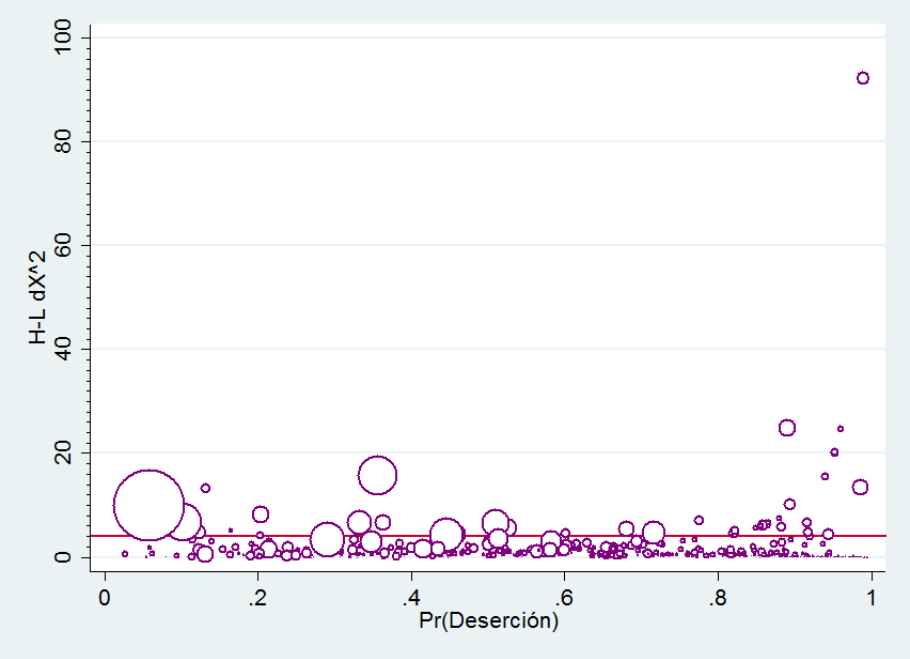

Figura 3.6: Residuales DeltaX ponderados por Deltab

Capítulo 4

Discusión y conclusiones

Como se ha visto, dado el impacto que tiene la deserción en la educación media superior, estudiar este fenómeno debe de ser considerado un objetivo angular,el diagnóstico presentado en esta tesis ayudará en el diseño y ejecución de políticas públicas.

La ENDEMS 2012 ofreció, por primera vez en nuestro país, una base de datos, de representatividad nacional, que ayuda a entender los factores que envuelven e influyen la deserción escolar en nuestro país. El objetivo de este trabajo no fue solo presentar las características principales de los jóvenes estudiantes que han desertado en la educación media superior en nuestro país, sino tomar medidas con el análisis antes expuesto, para abatir la deserción escolar. Como se vio en el capítulo anterior, la deserción escolar, al tener varios factores que influyen en ésta, no puede ser abordada desde un único flanco. El enfoque deberá ser desde una perspectiva cultural, social, económica y educativa.

Los modelos de regresión logística son modelos estadísticos en los que se desea conocer la relación entre una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) y una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas, siendo la ecuación inicial del modelo de tipo exponencial, si bien su transformación logarítmica (logit) permite su uso como una función lineal.

Con el modelo de regresión logística, presentado anteriormente, los tres factores que mejor predicen la deserción de los jóvenes son: la asistencia a clases, el promedio y la reprobación. De acuerdo a los resultados de la presente tesis, los jóvenes que faltan "mucho a clase" tienen 8 veces más posibilidad de desertar; de modo similar, el aumento para los que faltan "con cierta frecuencia" es de 2.5 veces más. En cuanto al promedio escolar, quienes tienen "de las más bajas" calificaciones tienen 8 veces más posibilidades de desertar, mientras que quienes tienen las calificaciones más altas tienen 2 veces más posibilidades de desertar. Finalmente, los que reprueban varias veces tienen casi 93veces más posibilidades

de desertar y para quienes reprueban en más ocasiones de las permitidas esta posibilidad de desertar aumenta a casi 9 veces.

Después de haber estimado la probabilidad de desertar o no con el modelo de regresión logística, se utilizaron métodos para la bondad de ajuste del modelo; con éstos se corroboró que el modelo de deserción escolar propuesto en esta tesis, estuvo bien ajustado y el modelo predice de manera correcta las inferencias sobre la deserción escolar en la educación media superior en nuestro país.

Por último, no hubo gran diferencia en las estimaciones utilizando el diseño de muestreo complejo, concluyendo que las estimaciones hechas asumiendo una muestra aleatoria simple aproximan de manera correcta las inferencias hechas sobre la deserción escolar.

La presente tesis contribuye a conocer el panorama de la deserción a nivel nacional y, en este sentido, representa un insumo pertinente para reforzar las políticas educativas existentes y para el diseño de aquellas que sean necesarias. Si bien es una aproximación de la realidad estudiantil mexicana, permite conocer más a fondo las características principales por las cuales los jóvenes mexicanos desertan.

Bibliografía

[1] Agresti, A. (2002). *Categorical Data Analysis,* 2nd edition. New York: John Wiley & Sons.

[2] Archer, K.J.; Lemeshow, S. (2006). Goodness-of-fit for logistic regression model fitted using survey sample data. *Stata Journal* 6 (1): 97-105.

[3] Collet, D. (2003). *Modelling Binary Data,* 2nd edition. London: Chapman & Hall.

[4] Hilbe, J.M. (1994). Generalized linear models. *The American Statistician* 48: 255-265.

[5] Hilbe, J.M. (2009). *Logistic Regression Models.* London: Chapman & Hall.

[6] Hoffmann, J.P. (2004). *Generalized Linear Models: An Applied Approach.* Boston: Allyn & Bacon.

[7] Hosmer, D.W.; Lemeshow S. (2000). *Applied Logistic Regression,* 2nd edition. New York: John Wiley & Sons.

[8] Kleinbaum, D.G. (1994). *Logistic Regression: A Self-Teaching Guide.* New York: Springer-Verlag.

[9] Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics* 9: 705-724.

[10] Stone, C. (1985). Additive regression and other nonparametric models. *Annals of Statistics* 13: 689-705

[11] .Woodward, M. (2005). *Epidemiology: Study Design and Data Analysis.* Boca Raton, Fl: Chapman & Hall

Apéndice A

El modelo lineal generalizado

El modelo de regresión logística expuesto en este trabajo, pertenece a una familia de modelos de regresión, llamada modelo lineal generalizado (GLM, de *generalized linear modelo*). Este modelo es en realidad una técnica unificadora de modelos de regresión y de diseño de experimentos, que une los modelos acostumbrados de regresión con la teoría normal y los modelos no lineales. Una hipótesis clave en el GLM es que la distribución de la variable de respuesta es un miembro de una familia exponencial de distribuciones, que incluye, entre otras, la normal, binomial, de Poisson, normal inversa, exponencial y gamma.

Las distribuciones que son de la familia exponencial tienen la forma general $f(y_i, \theta_i, \phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + h(y_i, \phi)\}$ donde ϕ es un parámetro de escala y, θ_i se llama parámetro natural de localización. Para los miembros de la familia exponencial, se tiene lo siguiente:

$$\mu = \mathbb{E}[y] = \frac{db(\theta_i)}{d(\theta_i)}$$

$$\begin{aligned} Var(y) &= \frac{d^2b(\theta_i)}{d(\theta_i^2)}a(\phi) \\ &= \frac{d\mu}{d\theta_i}a(\phi) \end{aligned}$$

Sea

$$Var(\mu) = \frac{Var(y)}{a(\phi)}\frac{d\mu}{d\theta_i} \tag{4.1}$$

donde $Var(\mu)$ representa la dependencia de la varianza de la respuesta a su media. Ésta es una característica de todas las distribuciones que son una familia exponencial, a excepción de la distribución normal. Como resultado de la ecuación (1.41), se tiene que:

$$\frac{d\theta_i}{d\mu} = \frac{1}{Var(\mu)}$$

El concepto básico de un GLM es desarrollar un modelo lineal para una función adecuada del valor esperado de la variable de respuesta. Sea η_i el predictor lineal, definido por:

$$\eta_i = g[\mathbb{E}(y_i)] = g(\mu_i) = \boldsymbol{X_i^{'}\beta}$$

Nótese que la respuesta esperada no es más que:

$$\mathbb{E}(y_i) = g^{-1}(\eta_i) = g^{-1}(X_i^{'}\beta)$$

La función g se llama función liga. Hay muchas alternativas posibles para la función cadena, pero si se escoge:

$$\eta_i = \theta_i$$

se dice que η_i es una liga canónica. La liga canónica que nos va a interesar será la relacionada con la distribución logística:

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

Una idea muy básica es que hay dos componentes en un sistema lineal general: la distribución de la respuesta y la función liga. Se puede considerar la selección de la función liga en una forma muy parecida a la de una transformación en la respuesta, sin embargo, a diferencia de una transformación, la función liga aprovecha la distribución natural de la respuesta. Así como el no usar una transformación adecuada puede acarrear problemas con un modelo lineal ajustado, la elección inadecuada de la función liga también puede ocasionar grandes problemas con un modelo lineal general.

El método de máxima verosimilitud es la base teórica de la estimación de parámetros en el modelo lineal general, sin embargo, la implementación real de la máxima verosimilitud da como resultado un algoritmo basado en los mínimos cuadrados iterativamente reponderados (IRLS). Es exactamente lo que se vio antes para el caso especial de la regresión logística.

Observaciones importantes acerca del modelo lineal general:

1. En el caso normal, cuando los experimentadores y los analistas de datos usan una transformación, usan mínimos cuadrados ordinarios para ajustar realmente el modelo en la escala transformada.

2. En un modelo lineal general se reconoce que la varianza de la respuesta no es constante y se usan mínimos cuadrados ponderados como base de la estimación de parámetros.

3. Lo anterior sugiere que un modelo lineal general debe ser mejor que el análisis estándar, usando transformaciones cuando un problema queda con varianza constante después de hacer la transformación.

4. Toda la inferencia que se describió antes acerca de la regresión logística se aplica en forma directa al modelo lineal general. Esto es, se puede usar la desviación del modelo para probar el ajuste general, y la diferencia de las desviaciones entre un modelo completo y uno reducido se puede usar para probar hipótesis acerca de subconjuntos de parámetros en el modelo. Se puede aplicar la inferencia de Wald para probar hipótesis y establecer intervalos de confianza para parámetros individuales del modelo.

Apéndice B

Estimación de los parámetros en la regresión logística

El logaritmo de la verosimilitud para un modelo de regresión logística se presentó en la ecuación (1.22) como sigue:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\boldsymbol{X_i'}\boldsymbol{\beta}) - \sum_{i=1}^{n} \log\left[1 + \exp(\boldsymbol{X_i'}\boldsymbol{\beta})\right]$$

En muchas aplicaciones de modelos de regresión logística se tienen intentos u observaciones repetidas en cada valor de las variables x. Sea Y_i la cantidad de unos observada para la i-ésima observación, y sea n_i la cantidad de intentos en cada observación. Entonces, el logaritmo de la verosimilitud queda de las siguiente manera:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i\pi_i + \sum_{i=1}^{n} n_i log\,(1-\pi_i) - \sum_{i=1}^{n} Y_i log\,(1-\pi_i)$$

Los estimados de máxima verosimilitud (MLE, de *maximum likelihood estimate*) se pueden calcular con un algoritmo de mínimos cuadrados iterativamente reponderados (IRLS, de *iteratively reweighted least squares*). Para visualizarlo, recuérdese que los estimados MLE son las soluciones:

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = 0$$

que se puede expresar de la siguiente forma:

$$\frac{\partial L}{\partial \pi_i}\frac{\partial \pi_i}{\partial \boldsymbol{\beta}} = 0$$

Hay que notar lo siguiente:

$$\frac{\partial L}{\partial \pi_i} = \sum_{i=1}^{n} \frac{n_i}{\pi_i} - \sum_{i=1}^{n} \frac{n_i}{1-\pi_i} + \sum_{i=1}^{n} \frac{Y_i}{1-\pi_i}$$

y también:

$$\frac{\partial \pi_i}{\partial \boldsymbol{\beta}} = \left\{ \frac{\exp(\boldsymbol{X_i'}\boldsymbol{\beta})}{1+\exp(\boldsymbol{X_i'}\boldsymbol{\beta})} - \left[\frac{\exp(\boldsymbol{X_i'}\boldsymbol{\beta})}{1+\exp(\boldsymbol{X_i'}\boldsymbol{\beta})} \right]^2 \right\} X_i$$

Con lo anterior se deduce lo siguiente:

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\beta}} &= \left[\sum_{i=1}^{n} \frac{n_i}{\pi_i} - \sum_{i=1}^{n} \frac{n_i}{1-\pi_i} + \sum_{i=1}^{n} \frac{Y_i}{1-\pi_i} \right] \pi_i \left(1-\pi_i\right) X_i \\ &= \sum_{i=1}^{n} \left[\frac{n_i}{\pi_i} - \frac{n_i}{1-\pi_i} + \frac{Y_i}{1-\pi_i} \right] \pi_i \left(1-\pi_i\right) X_i \\ &= \sum_{i=1}^{n} \left(Y_i - n_i \pi_i\right) X_i \end{aligned}$$

Por lo tanto, estimador de máxima verosimilitud resuelve la ecuación:

$$\boldsymbol{X}^{'} \left(\boldsymbol{Y} - \boldsymbol{\mu}\right) = 0$$

en donde $\boldsymbol{Y}^{'} = [Y_1, Y_2, ..., Y_n]$ y $\boldsymbol{\mu}^{'} = [n_1\pi_1, n_2\pi_2, ..., n_n\pi_n]$. Este conjunto de ecuaciones se llama con frecuencia *ecuaciones de puntuación de máxima verosimilitud*. Son en realidad de la misma forma de las ecuaciones normales para los mínimos cuadrados lineales.

El método de *Newton-Raphson* es el que se usa para resolver las ecuaciones de puntuación para el modelo de regresión logística. En este procedimiento se observa que en la proximidad de la solución se puede usar un desarrollo en serie de Taylor de primer orden para formar la aproximación:

$$P_i - \pi_i \approx \left(\frac{\partial \pi_i}{\partial \boldsymbol{\beta}} \right)^{'} \left(\boldsymbol{\beta^*} - \boldsymbol{\beta}\right) \tag{4.2}$$

donde:

$$P_i = \frac{Y_i}{n_i}$$

y $\boldsymbol{\beta}^*$ es el valor de $\boldsymbol{\beta}$ que resuelve las ecuaciones de puntuación. Sea ahora $\phi_i = X_i^{'}\boldsymbol{\beta}$, esto implica que $\frac{\partial \phi_i}{\partial \boldsymbol{\beta}} = X_i$. Recordando que

$$\pi_i = \frac{\exp(\phi_i)}{1+\exp(\phi_i)} \tag{4.3}$$

utilicemos la regla de la cadena:

$$\frac{\partial \pi_i}{\partial \boldsymbol{\beta}} = \frac{\partial \pi_i}{\partial \phi_i} \frac{\partial \phi_i}{\partial \boldsymbol{\beta}} = \frac{\partial \pi_i}{\partial \phi_i} X_i$$

Por lo anterior, la ecuación () se puede escribir de la siguiente forma:

$$\begin{aligned} P_i - \pi_i &\approx \left(\frac{\partial \pi_i}{\partial \phi_i}\right) X_i^{'}(\boldsymbol{\beta^*} - \boldsymbol{\beta}) \\ &\approx \left(\frac{\partial \pi_i}{\partial \phi_i}\right) X_i^{'}\boldsymbol{\beta^*} - \boldsymbol{X_i^{'}\beta} \\ &\approx \left(\frac{\partial \pi_i}{\partial \phi_i}\right) (\phi_i^* - \phi_i) \end{aligned}$$

donde ϕ_i^* es el valor de ϕ_i evaluado en β^*. Notemos ahora lo siguiente:

$$Y_i - n_i\pi_i = (n_iP_i - n_i\pi_i) = n_i(P_i - \phi_i)$$

y utilizando la ecuación (), se obtiene lo siguiente:

$$\frac{\partial \pi_i}{\partial \phi_i} = \frac{\exp(\phi_i)}{1 + \exp(\phi_i)} - \left[\frac{\exp(\phi_i)}{1 + \exp(\phi_i)}\right]^2$$

Esto implica:

$$Y_i - n_i\pi_i \approx [n_i\pi_i(1 - \pi_i)]\,(\phi_i^* - \phi_i)$$

Ahora bien, la varianza del predictor lineal $\phi_i^* = X_i^{'}\boldsymbol{\beta^*}$ es, con una primera aproximación lo siguiente:

$$Var\,[\phi_i^*] = \frac{1}{n_i\pi_i(1 - \pi_i)}$$

Con lo anterior desarrollado, se obtiene lo siguiente:

$$Y_i - n_i\pi_i \approx \left[\frac{1}{Var\,[\phi_i^*]}\right](\phi_i^* - \phi_i)$$

y se pueden expresar las ecuaciones de puntuación en la siguiente forma:

$$\sum_{i=1}^{n}\left[\frac{1}{Var\,[\phi_i^*]}\right](\phi_i^* - \phi_i) = 0$$

o bien en su forma matricial:

$$\boldsymbol{X^{'}V^{-1}}(\phi^* - \phi) = 0$$

siendo $\boldsymbol{V}$ una matriz diagonal de los factores de ponderación obtenidos con las varianzas de las ϕ_i. Como $\phi = \boldsymbol{X\beta}$, las ecuaciones de puntuación se pueden escribir como sigue:

$$\boldsymbol{X^{'}V^{-1}}(\phi^* - \boldsymbol{X\beta}) = 0$$

y el estimador de máxima verosimilitud de $\boldsymbol{\beta}$ es:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{(X^{'}V^{-1}X)^{-1}X^{'}V^{-1}}\phi^*$$

Sin embargo, se presenta el problema que no se conoce ϕ^* . La solución de este problema emplea la ecuación ():

$$P_i - \pi_i \approx \left(\frac{\partial \pi_i}{\partial \phi_i}\right)(\phi_i^* - \phi_i)$$

de donde se pueden despejar las ϕ^*:

$$\phi_i^* \approx \phi_i + (P_i - \pi_i)\left(\frac{\partial \pi_i}{\partial \phi_i}\right)$$

Sean $z_i = \phi_i + (P_i - \pi_i)\left(\frac{\partial \pi_i}{\partial \phi_i}\right)$ y $\boldsymbol{z}' = [z_1, z_2, \ldots, z_n]$. Entonces, el estimado de NewtonRaphson para β es:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{z}$$

Notemos que la parte aleatoria de z_i es:

$$(P_i - \pi_i)\left(\frac{\partial \pi_i}{\partial \phi_i}\right)$$

Con esto la varianza queda de la siguiente forma:

$$\begin{aligned} Var\left[(P_i - \pi_i)\left(\frac{\partial \pi_i}{\partial \phi_i}\right)\right] &= \left[\frac{\pi_i(1-\pi_i)}{n_i}\right]\left(\frac{\partial \pi_i}{\partial \phi_i}\right)^2 \\ &= \left[\frac{\pi_i(1-\pi_i)}{n_i}\right]\left(\frac{1}{\pi_i(1-\pi_i)}\right)^2 \\ &= \frac{1}{n_i\pi_i(1-\pi_i)} \end{aligned}$$

Entonces, $\boldsymbol{V}$ es la matriz diagonal de los factores de ponderación obtenidos con las varianzas de la parte aleatoria de $\boldsymbol{z}$. Por consiguiente, el algoritmo IRLS basado en el método de Newton-Raphson se puede describir como sigue:

1. Usar mínimos cuadrados ordinarios para obtener un estimado inicial de β, por ejemplo $\hat{\beta}_0$
2. Usar $\hat{\beta}_0$ para estimar $\boldsymbol{V}$ y π
3. Definir $\phi_0 = \boldsymbol{X}\hat{\beta}_0$
4. Basar $\boldsymbol{z_1}$ en ϕ_0
5. Obtener un nuevo estimado de $\hat{\beta}_1$ e iterar, hasta que se satisfaga un criterio adecuado de convergencia.

Apéndice C

Ajuste de la regresión logística para datos provenientes de una muestra compleja

Lo desarrollado en el capítulo 1 de la presente tesis asume que los datos a analizar provienen de una muestra aleatoria simple. De unos años para acá se ha hecho un progreso considerable para extender el modelo de regresión logística para otro tipo de muestras. Algunos de los más recientes avances en los paquetes estadísticos para la regresión logística son que pueden realizar análisis con datos obtenidos de una muestra compleja. Estas rutinas pueden ser encontradas en paquetes como SPSS, STATA, SUDAAN, etc.; el utilizado para la presente tesis fue el paquete estadístico STATA, que al igual que los anteriores presenta un paquete para analizar este tipo de muestras.

La idea principal es configurar una función que aproxime la función de verosimilitud de una población finita con una función de verosimilitud formada de una muestra observada y con factores de expansión conocidos. Supongamos que asumimos que una población puede partida en $k = 1, 2, \ldots, K$ estratos, en $j = 1, 2, \ldots, M_k$ unidades primarias de muestreo en cada estrato y en $i = 1, 2, \ldots, N_{kj}$ elementos en la kj-ésima unidad primaria de muestreo. Supongamos también que nuestros datos observados consisten en n_{kj} elementos de m_k unidades primarias de muestro de un estrato k. Denotemos el número total de observaciones como $n = \sum_{k=1}^{K} \sum_{j=1}^{m_k} n_{kj}$, denotemos los factores de expansión conocidos para la kji-ésima observación como w_{kji}, el vector de variables predictoras como $\boldsymbol{X}_{kji}$ y la variable de respuesta como Y_{kji}. La función de verosimilitud logarítmica aproximada quedaría de la siguiente forma:

$$\sum_{k=1}^{K} \sum_{j=1}^{m_k} \sum_{i=1}^{n_{kj}} [w_{kji} Y_{kji}] \log [\pi (X_{kji})] + [w_{kji}(1 - Y_{kji})] \log [1 - \pi (X_{kji})] \quad (4.4)$$

Diferenciando esta ecuación con respecto a los coeficientes de regresión desconocidos deja el vector de p+1 ecuaciones de puntuación:

$$\boldsymbol{X}' \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{\pi}) = 0 \quad (4.5)$$

donde $\boldsymbol{X}$ es una matriz de tamaño nx$(p+1)$ con los valores de las variables predictoras, $\boldsymbol{W}$ es una matriz diagonal de tamaño nxn que contiene los factores de expansión, $\boldsymbol{Y}$es un vector de tamaño nx1 con las variables de respuesta y $\boldsymbol{\pi} = \left(\pi(X_{111}), \ldots, \pi(X_{Km_kn_{kj}})\right)^{'}$ es un vector de tamaño nx1 que contiene las probabilidades logísticas.

En teoría, cualquier paquete estadístico para analizar la regresión logística puede obtener las soluciones de la ecuación (56). El problema se presenta al tratar de obtener el estimador de la matriz de covarianzas de los coeficientes. Ingenuamente el uso de cualquier paquete para analizar la regresión logística con matriz de factores de expansión $\boldsymbol{W}$ would yield estimates en la matriz $\left(\boldsymbol{X^{'}DX}\right)^{\boldsymbol{-1}}$ donde $\boldsymbol{D = WV}$ es una matriz diagonal de tamaño nxn con el elemento general $w_{kji}\hat{\pi}\left(X_{kji}\right)\left[1 - \hat{\pi}\left(X_{kji}\right)\right]$. El estimador correcto de la varianza es el siguiente:

$$\hat{Var}(\hat{\boldsymbol{\beta}}) = \boldsymbol{(X^{'}DX)^{-1}S(X^{'}DX)^{-1}} \tag{4.6}$$

donde $\boldsymbol{S}$ es un estimador agrupado dentro del estrato de la matriz de covarianzas el lado izquierdo de la ecuación (45). Denotemos un elemento general en el vector de la ecuación (45) como $z_{kji}^{'} = X_{kji}^{'} w_{kji}\left(Y_{kji} - \pi\left(X_{kji}\right)\right)$, también denotemos la suma sobre las unidades muestrales n_{kj} en la j-ésima unidad primaria de muestreo y en el k-ésimo estrato como $z_{kj} = \sum_{i=1}^{n_{kj}} z_{kji}$ y su específica media por estrato como $\bar{z}_k = \frac{1}{m_k}\sum_{j=1}^{m_k} z_{kji}$. El estimador dentro del estrato para el k-ésimo estrato es el siguiente:

$$S_k = \frac{m_k}{m_k - 1}\sum_{j=1}^{m_k}(z_{kj} - \bar{z}_k)(z_{kj} - \bar{z}_k)'$$

El estimador agrupado es $\boldsymbol{S} = \sum_{k=1}^{K}(1 - f_k)S_k$, a $(1 - f_k)$ se le conoce como el corrector por finitud donde $f_k = m_k/M_k$ es la tasa del número unidades primarias muestrales observadas y el total de unidades primarias muestrales para el k-ésimo estrato.

La función de verosimilitud (44) es solamente una aproximación a la verdadera función de verosimilitud. Esto implica que podríamos esperar que las inferencias sobre los parámetros deben de estar basadas en la prueba de Wald. Sin embargo, algunas simulaciones (Korn y Graubard 1990, Thomas y Rao 1987) mostraron que cuando los datos provienen de muestras complejas, el uso de de la estadística Wald modificada y la distribución F son mejores para el alfa que buscamos contrastar. R provee resultados de la estadística de Wald modificada. El problema es que ninguna de estas simulaciones examina el modelo de regresión logística usando variables categóricas y continuas con los estimados obtenidos en las ecuaciones (45) y (46).

En la simulación de de Korn y Graubard utilizan una regresión lineal con errores que se distribuyen normal, Thomas y Rao examinan modelos con va-

riables de respuesta dicotómicas y pocas variables predictoras categóricas. Otro problema es que los paquetes estadísticos, por ejemplo R, utiliza la distribución t para evaluar signicancia en la prueba de Wald para los coeficientes. Dada la escasez de simulaciones apropiadas y de teoría, no hay suficiente evidencia para soportar el uso de la estadística de Wald modificada con la distribución F para los modelos de regresión logística. Una posible justificación es que el uso de la estadística de Wald modificada con la distribución F es conservadora en los niveles de significancia es, en general, mayor en comparación con los obtenidos utilizando la estadística de Wald al ser una normal multivariada para muestras grandes. A continuación se presentan resultados basados en las dos pruebas (modificada y no modificada).

La relación entre la prueba de Wald y la prueba modificada es la siguiente; sea W la estadística de Wald para probar que todos los p-coeficientes en el modelo ajustado son iguales a 0, es decir:

$$W = \hat{\boldsymbol{\beta}}' \left[\hat{Var}\left(\hat{\boldsymbol{\beta}}\right)_{pxp} \right]^{-1} \hat{\boldsymbol{\beta}}$$

donde $\hat{\boldsymbol{\beta}}$ denota el vector de los coeficientes y $\hat{Var}\left(\hat{\boldsymbol{\beta}}\right)$ es la submatriz de tamaño pxp obtenida de la matriz (46). Esto implica que se deja afuera la fila y columna para el término constante. El p-valor se obtiene utilizando la distribución ji-cuadrada con p grados de libertad, es decir, $p - value = \mathbb{P}\left[\chi^2(p) \geq W\right]$.

La estadística de Wald modificada es la siguiente:

$$F = \frac{(s - p + 1)}{sp} W$$

donde $s = \left(\sum_{k=1}^{K} m_k\right) - K$ es el número total de unidades primarias de muestreo observadas menos el número de estratos. El p-valor se obtiene usando la distribución F con p y $(s - p + 1)$ grados de libertad,es decir, $p - value = \mathbb{P}\left[\mathrm{F}(p, s - p + 1) \geq F\right]$.