UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS MATEMÁTICAS

FACULTAD DE CIENCIAS

Identificación de los Factores de Riesgo de la Deserción Escolar en la Universidad del Caribe. Una aplicación del Modelo de Regresión Logística y del Método de Imputación Múltiple

T E S I S

QUE PARA OBTENER EL GRADO ACADÉMICO DE

MAESTRO EN CIENCIAS

P R E S E N T A

FELIPE JESÚS CUEVA DEL CASTILLO MENDOZA

DIRECTORA DE LA TESIS: (DRA, REBECA AGUIRRE HERNÁNDEZ)

MÉXICO, D.F. **2012**

Identificación de los Factores de Riesgo de la Deserción Escolar en la Universidad del Caribe. Una Aplicación del Modelo de Regresión Logística y del Método de Imputación Múltiple.

Felipe Jesús Cueva del Castillo Mendoza

2012

Índice general

Capítulo 1

Introducción

El tema de las causas de la deserción escolar se ha convertido en objeto de estudio para distintos especialistas en el área de la investigación educativa, lo anterior como respuesta a la necesidad de identificar las características inherentes al estudiante que lo hacen proclive a abandonar sus estudios. En este sentido el presente trabajo considera de interés poder identificar en términos de significancia estadística las características que pudieran considerarse como influyentes en la explicación del abandono escolar para el caso de una institución de educación superior en específico. La identificación de estas características, que durante el presente trabajo serán denominadas como factores de riesgo, se hará mediante la aplicación de un análisis estadístico que por la naturaleza de la información disponible (con la presencia de una variable dependiente dicotómica y la pérdida de información para ciertos niveles de respuesta de algunos factores considerados), involucrará la aplicación de un modelo de regresión logística y del método de imputación múltiple, por lo que dicho análisis, además de constituir la directriz para el desarrollo del presente trabajo, al combinar las técnicas del análisis de regresión logística con el método de imputación múltiple, da un aporte metodológico que hace posible realizar un análisis de regresión logística ante la presencia de información faltante en sus variables explicativas.

La información recabada para realizar este análisis fue obtenida de una base de datos generada a partir de un cuestionario aplicado a una población estudiantil particular. Dicha base (que consta de los datos recabados para 417 individuos) corresponde a un cuestionario que a su ingreso fue aplicado a los estudiantes de la primera y segunda cohortes de la Universidad del Caribe en Cancún Quintana Roo. Cabe hacer mención que la selección de las características consideradas como influyentes para explicar la deserción, fue en apego al sentido común así como a la experiencia profesional que el autor de este trabajo ha acumulado como docente en su paso por instituciones de educación media superior y superior. Dichos factores a considerar se enlistan a continuación:

- Sexo
- Nivel de escolaridad del padre

- Nivel de escolaridad de la madre
- Estado Civil
- Tiene dependientes económicos
- Cuenta con servicios culturales en su barrio o colonia
- Si la casa donde vive es propia o rentada
- Nivel del barrio o colonia donde vive
- Edad
- Trabaja
- Tipo de sistema de procedencia
- Proviene de escuela pública o privada
- Concluyó en más de tres años su bachillerato
- Promedio de bachillerato

El desarrollo del presente trabajo comienza en el Capítulo 2 con una revisión detallada del tipo de información disponible, se da una descripción de cada una de las variables a considerar dentro del análisis así como de sus niveles de no respuesta. En el Capítulo 3 se fundamenta el uso del modelo de regresión logística como una alternativa viable para el análisis de la información disponible, así mismo, se describen los fundamentos del modelo de regresión logística tanto desde el punto de vista del análisis de regresión logística asintótico como desde la perspectiva del análisis regresión logística exacto. Posteriormente en el Capítulo 4 se establecen los supuestos relacionados con el mecanismo de información faltante que dan sustento al método de imputación múltiple utilizado, dándose una descripción de las generalidades de este método de imputación múltiple así como las particularidades que este método adopta cuando se adecua a las características de la información disponible. El Capítulo 5 considera la aplicación de los métodos expuestos en los Capítulos 3 y 4, se exponen los resultados obtenidos derivados de su aplicación al conjunto de datos recabados, y se concluye con el comparativo de los resultados obtenidos del análisis con datos sin imputar con aquellos obtenidos del análisis con datos imputados. Para finalizar, en el Capítulo 6 se desarrollan las conclusiones obtenidas a partir de los resultados expuestos en el Capítulo 5.

Capítulo 2

Descripción de la información disponible

2.1. Introducción

En el presente capítulo se expone una descripción detallada de cada uno de los factores involucrados en el análisis así como de la variable respuesta objeto de estudio, se asigna una codificación para sus distintos niveles de respuesta que permita incorporarlos al análisis como variables de tipo categórico. La razón de categorizar todas las variables obedece a la naturaleza del método de imputación múltiple que se va a utilizar en el análisis, ya que éste está diseñado para la imputación de variables de tipo categórico. El capítulo finaliza con la descripción de los porcentajes de no respuesta observados para cada uno de los factores de riesgo descritos.

2.2. Descripción de los Factores de Riesgo considerados en el análisis

Sexo (SEXO)

Representa el sexo del estudiante.

Código	**(Niveles de Respuesta Considerados)**
0	Masculino
1	Femenino

Nivel de escolaridad del padre (ESCOPAD)

Mide el máximo nivel educativo alcanzado por el padre del estudiante.

Código	(Niveles de Respuesta Considerados)
1	No lee ni escribe Sin Estudios
2	Primaria Incompleta Primaria Completa
3	Secundaria Incompleta Secundaria Completa
4	Bachillerato o Preparatoria Completa Bachillerato o Preparatoria Incompleta
5	Licenciatura, Normal o Carrera Completa Licenciatura, Normal o Carrera Incompleta Posgrado Completo Posgrado Incompleto
6	Otra

Nivel de escolaridad de la madre (ESCOMAD)

Mide el máximo nivel educativo alcanzado por la madre del estudiante.

Código	(Niveles de Respuesta Considerados)
1	Sin Estudios
2	Primaria Incompleta Primaria Completa
3	Secundaria Incompleta Secundaria Completa
4	Bachillerato o Preparatoria Completa Bachillerato o Preparatoria Incompleta
5	Licenciatura, Normal o Carrera Completa Licenciatura, Normal o Carrera Incompleta Posgrado Completo Posgrado Incompleto
6	Otra

Antes de pasar a la descripción del siguiente factor es importante mencionar que para los factores anteriores, el nivel de respuesta " Otra " toma en cuenta las categorías que involucran una formación enfocada en la capacitación para el trabajo, siendo éstas las que se enlistan a continuación:

- Capacitación Técnica o Comercial Después de la Primaria
- Capacitación Técnica o Comercial Después de la Secundaria
- Técnico profesional incompleto
- Técnico profesional completo

Estado Civil (EDOCIVI)

Describe si el estudiante tiene o no algún tipo de compromiso de carácter marital o de concubinato.

Código	(Niveles de Respuesta Considerados)
1	Soltero, Viudo o Divorciado
2	Casado, Unión Libre

Tiene dependientes económicos (DEPECO)

Describe si el estudiante tiene algún tipo de compromiso de carácter económico, derivado de la manutención de algún familiar.

Código	(Niveles de Respuesta Considerados)
0	No tiene dependientes económicos
1	Tiene dependientes económicos

Cuenta con servicios culturales en su barrio o colonia (SERVCUL)

Describe la oportunidad que pueda llegar a tener el estudiante en el área dónde habita de acceder a este tipo de servicios.

Código	(Niveles de Respuesta Considerados)
0	Prácticamente sin servicios culturales
1	Prácticamente con todos

Si la casa donde vive es propia o rentada (TIPCASA)

Describe la situación de propiedad que el estudiante tiene con respecto al espacio donde habita.

Código	(Niveles de Respuesta Considerados)
0	Propia
1	No propia

Nivel del barrio o colonia donde vive (DESCCOL)

Describe bajo la perspectiva del estudiante, el nivel socioeconómico del lugar de residencia donde éste habita.

Código	(Niveles de Respuesta Considerados)
1	Bajo
2	Medio Bajo Medio Medio Alto
3	Alto

Edad (EDAD)

Representa la edad del estudiante.

Código	(Niveles de Respuesta Considerados)
1	17 a 24 años
2	25 o más años

Trabaja (TRABAJA)

Describe si el estudiante desempeña alguna actividad laboral por la cual percibe un ingreso.

Código	(Niveles de Respuesta Considerados)
0	No
1	Sí

Tipo de sistema de procedencia (SISTEMA)

Describe la condición de escolarización del sistema de bachillerato de origen del estudiante.

Código	(Niveles de Respuesta Considerados)
1	Escolarizado o Semiescolarizado
2	Abierto

Proviene de escuela pública o privada (TIPOREG)

Describe si el estudiante tuvo erogaciones por concepto de pago de colegiaturas en el sistema de bachillerato de procedencia.

Código	(Niveles de Respuesta Considerados)
1	Pública
2	Privada o Por cooperación

Concluyó en más de tres años su bachillerato (REZAGO)

Describe si el estudiante concluyó su bachillerato durante el tiempo estipulado en su programa de estudios.

Código	(Niveles de Respuesta Considerados)
0	No
1	Sí

Promedio de bachillerato (PROBACH)

Muestra el promedio con que el estudiante concluyó su bachillerato.

Código	**(Niveles de Respuesta Considerados)**
0	6.0 a 8.0
1	8.1 a 10

2.3. Descripción de la Varible Respuesta y de los porcentajes de no respuesta para los Factores de Riesgo

En la descripción de la variable respuesta, que para este caso es si el estudiante desertó o no, es importante hacer mención que para considerar a un estudiante como desertor se tomó en cuenta a aquellos estudiantes que abandonaron en forma definitiva la carrera en la que estaban inscritos durante el primer periodo lectivo de su ingreso a la institución, y como no desertores a aquellos que no cumplían con esta condición. La codificación para los dos niveles de respuesta para esta variable así como la etiqueta que permitirá su identificación durante el análisis se presenta a continuación:

Tabla 2.1: Cantidad y porcentaje de observaciones perdidas por cada factor de riesgo considerando como base para su cálculo los 417 individuos registrados en la base de datos.

Factor	Numero de observaciones perdidas	Porcentaje
TIPOREG	3	0.72
SISTEMA	5	1.20
REZAGO	13	3.12
SEXO	0	0.00
EDOCIVI	17	4.08
DEPECO	0	0.00
TRABAJA	34	8.15
PROBACH	18	4.32
EDAD	7	1.68
TIPCASA	28	6.71
SERVCUL	19	4.56
DESCCOL	9	2.16
ESCOPAD	80	19.18
ESCOMAD	83	19.90

Desertó (DESERTO)

Código	(Niveles de Respuesta Considerados)
0	No
1	Sí

Por último, en la Tabla 2.1 se da una descripción del porcentaje de no respuesta para cada una de los factores de riesgo que se acaban de definir.

Capítulo 3

El Modelo de Regresión Logística

3.1. Introducción

El modelo de regresión logística puede ser utilizado para modelar la posible relación de causa y efecto entre una variable respuesta con dos categorías y un conjunto de una o más variables explicativas. Por lo tanto, tomando en cuenta la naturaleza de la variable respuesta involucrada en este análisis, la aplicación de este modelo puede ser conveniente para alcanzar el objetivo deseado.

Existen antecedentes de la aplicación del modelo de regresión logística dentro del ámbito de la investigación educativa. Por ejemplo, en el artículo de Considine y Zappala (2002), se utiliza la regresión logística para estimar la magnitud del efecto de los factores socioeconómicos, familiares, individuales y contextuales en el desempeño escolar en estudiantes con antecedentes de desventajas económicas.

3.2. Descripción del Modelo

El siguiente paso se basa en presentar una descripción del modelo de regresión logística, para lo cual se comienza por definir las literales que representarán las variables que serán utilizadas en su conformación. Sea "Z" la variable dependiente que para este caso es si el estudiante desertó o no, la cual tomará los valores $Z = 0$ ó $Z = 1$ en el orden que fue considerado a través de los identificadores que condujeron a la codificación de sus niveles de respuesta. En cuanto a la descripción de las variables que representan los factores, en general se designa a cada una de éstas como "w" con su respectivo subíndice.

Con la finalidad de definir en términos matemáticos la expresión del modelo a utilizar, es importante retornar al objetivo en mente, el cual consiste en identificar la ocurrencia o no del hecho objeto de estudio (si el estudiante desertó o no) a través de las características consideradas como influyentes (las variables

consideradas). Por lo que este interés se traducirá en medir la probabilidad de que la variable dependiente "Z" tome el valor de "1" ó "0" como función de las variables que se consideran potencialmente influyentes para que esto suceda.

La probabilidad de $Z = 1$ se denota por $P[Z = 1|\mathrm{w}] = \pi(\mathrm{w})$, mientras que la probabilidad de Z=0 es simbolizada a través de $P[Z = 0|\mathrm{w}] = 1 - \pi(\mathrm{w})$, donde el vector $\mathrm{w}' = (w_1, w_2, ..., w_p)$ representa la colección de las "p" variables explicativas que van a ser consideradas en el análisis. De esta manera queda definida "Z" como una variable aleatoria que tiene una distribución Bernoulli con parámetro $\pi(\mathrm{w})$ $(0 \leq \pi(\mathrm{w}) \leq 1)$.

La forma analítica en que la probabilidad objeto de interés se vincula con las variables explicativas y que se define como el modelo de regresión logística es la siguiente:

$$\pi(\mathrm{w}) = \frac{\exp\big(g(\mathrm{w})\big)}{1 + \exp\big(g(\mathrm{w})\big)}, \tag{3.1}$$

en cuyo caso la ecuación

$$g(\mathrm{w}) = \beta_0 + \beta_1 w_1 + \beta_2 w_2 + ... + \beta_p w_p, \tag{3.2}$$

se conoce como el predictor lineal de la regresión logística y $\beta_0, \beta_1, \beta_2, ..., \beta_p$ son los parámetros del modelo.

3.2.1. Ajuste del Modelo

De la ecuación (3.2) se puede notar que la parte sustancial en el ajuste del modelo de regresión logística consiste en ser capaz de encontrar valores numéricos (o valores estimados) para el conjunto de parámetros involucrados en éste. La forma de lograrlo será a través de la aplicación del método de estimación de Máxima Verosimilitud.

Como primer paso en la estimación de los parámetros del modelo, considérese que se tiene una muestra de n sujetos para cada uno de los cuales se obtuvieron valores observados de la variable Z y del vector w. Si se define a N como el número de valores distintos del vector w observado (por lo que si algunos sujetos de la muestra tienen el mismo valor en el vector observado w entoces $N < n$), y con n_i al número de sujetos en la muestra que tienen asociado el mismo valor observado de w (a este conjunto de valores observados de w se le denomina como patrón de covariables[1]); esto es, el número de sujetos para los cuales $\mathrm{w} = \mathrm{w}_i$, con $\mathrm{w}_i' = (w_{i1}, w_{i2}, ..., w_{ip})$, $i = 1, 2, ..., N$ y $n = n_1 + n_2 + ... + n_N$. Entonces con base en la ecuación (3.1), se puede distinguir que $\pi(\mathrm{w}_i) = \frac{\exp\left(g(\mathrm{w}_i)\right)}{1+\exp\left(g(\mathrm{w}_i)\right)}$ representa la probabilidad de que algún sujeto que pertenezca al patrón de covariables i tenga como respuesta asociada $Z = 1$. Así entonces, la probabilidad de observar

[1] El término covariable tiene la misma connotación que el de variable explicativa, siendo esta última denominación la que se ha venido usando y se utilizará a lo largo del presente trabajo.

z_i sujetos con respuesta asociada $Z = 1$, de un total de n_i sujetos, es evaluada como

$$\left(\begin{array}{c} n_i \\ z_i \end{array} \right) \pi(\mathrm{w}_i)^{z_i} \left[1 - \pi(\mathrm{w}_i)\right]^{n_i - z_i}, \tag{3.3}$$

por lo que el predictor lineal en este caso queda expresado de la siguiente manera

$$g(\mathrm{w}_i) = \beta_0 + \beta_1 w_{i1} + \beta_2 w_{i2} + ... + \beta_p w_{ip}. \tag{3.4}$$

De (3.4) salta a la vista que el valor de la probabilidad $\pi(\mathrm{w}_i)$ tan sólo depende del vector de parámetros desconocidos $\boldsymbol{\beta'} = (\beta_0, \beta_1, ..., \beta_p)$.

Ahora únicamente resta construir la función de verosimilitud que permita encontrar los valores estimados de los parárametros desconocidos, para lo cual se considera a $\boldsymbol{\mathcal{Z}} = \{z_1, ..., z_N\}$, el vector de los valores observados z_i para cada uno de los n_i grupos. La probabilidad de $\boldsymbol{\mathcal{Z}}$ bajo el modelo binomial es entonces determinada por

$$P[\mathcal{Z}] = \prod_{i=1}^{N} \left(\begin{array}{c} n_i \\ z_i \end{array} \right) \pi(\mathrm{w}_i)^{z_i} \left[1 - \pi(\mathrm{w}_i)\right]^{n_i - z_i}, \tag{3.5}$$

pues $z_1, z_2, ..., z_N$ constituyen un conjunto de observaciones de N variables aleatorias independientes binomiales. Por simplicidad, el cálculo de la función de verosimilitud se llevará a cabo ignorando el coeficiente binomial que no depende de $\boldsymbol{\beta}$. En consecuencia la función de verosimilitud es proporcional a:

$$\begin{aligned}
&\prod_{i=1}^{N} \pi(\mathrm{w}_i)^{z_i} \left[1 - \pi(\mathrm{w}_i)\right]^{n_i - z_i} \\
&= \left\{ \prod_{i=1}^{N} \left(\frac{\pi(\mathrm{w}_i)}{1 - \pi(\mathrm{w}_i)} \right)^{z_i} \right\} \left\{ \prod_{i=1}^{N} \left[1 - \pi(\mathrm{w}_i)\right]^{n_i} \right\} \\
&= \left\{ \prod_{i=1}^{N} \exp\left[\log\left(\frac{\pi(\mathrm{w}_i)}{1 - \pi(\mathrm{w}_i)} \right)^{z_i} \right] \right\} \left\{ \prod_{i=1}^{N} \left[1 - \pi(\mathrm{w}_i)\right]^{n_i} \right\} \\
&= \left\{ \exp\left[\sum_{i=1}^{N} z_i \log\left(\frac{\pi(\mathrm{w}_i)}{1 - \pi(\mathrm{w}_i)} \right) \right] \right\} \left\{ \prod_{i=1}^{N} \left[1 - \pi(\mathrm{w}_i)\right]^{n_i} \right\} \\
&= \left\{ \exp\left[\sum_{i=1}^{N} z_i g(\mathrm{w}_i) \right] \right\} \left\{ \prod_{i=1}^{N} \left[1 - \pi(\mathrm{w}_i)\right]^{n_i} \right\}.
\end{aligned}$$

Por consiguiente al tomar el logaritmo en la última expresión se obtiene la función conocida como la log-verosimilitud, cuya fórmula resultante es

$$\ell(\beta) = \sum_{i=1}^{N} z_i g(\mathrm{w}_i) - \sum_{i=1}^{N} n_i \log\left[1 + \exp(g(\mathrm{w}_i))\right]. \tag{3.6}$$

Así, las ecuaciones de verosimilitud se obtienen al tomar $\frac{\partial \ell(\beta)}{\partial \beta_k} = 0$, para $k = 0, 1, 2, ..., p$. Estas derivadas parciales tienen la siguiente forma general

$$\frac{\partial \ell(\beta)}{\partial \beta_0} = \sum_{i=1}^{N} z_i - \sum_{i=1}^{N} n_i \left(\frac{\exp\left(g(\mathrm{w}_i)\right)}{1 + \exp\left(g(\mathrm{w}_i)\right)} \right)$$

y

$$\frac{\partial \ell(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} z_i w_{ik} - \sum_{i=1}^{N} n_i w_{ik} \left(\frac{\exp\left(g(\mathrm{w}_i)\right)}{1 + \exp\left(g(\mathrm{w}_i)\right)} \right)$$

para $k = 1, 2, ..., p$, por lo que las ecuaciones de verosimilitud quedan determinadas de la siguente forma:

$$\sum_{i=1}^{N} \left[z_i - n_i \pi(\mathrm{w}_i) \right] = 0 \tag{3.7}$$

y

$$\sum_{i=1}^{N} w_{ik} \left[z_i - n_i \pi(\mathrm{w}_i) \right] = 0 \tag{3.8}$$

para $k = 1, 2, ..., p$.

Nótese que estas ecuaciones no tienen una solución cerrada, por lo tanto su solución requiere de la aplicación de un método iterativo, como por ejemplo el método de Newton Rapson (ver Agresti (2000)).

Si se denota a $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$ como el vector de soluciones a estas ecuaciones (vector de parámetros estimados). Entonces los valores ajustados para las probabilidades del modelo en cuestión se obtienen inicialmente sustituyendo éstos en la ecuación (3.4), para posteriormente sustituir este último valor en la ecuación (3.1). La representación para estas probabilidades estimadas estará determinada a través de la expresión $\hat{\pi}(\mathrm{w}_i)$.

Para concluir con el ajuste del modelo se definirá la matriz de varianzas y covarianzas de los parámetros estimados, la cual más adelante será de utilidad en la determinación de la significancia estadística de los coeficientes estimados.

El método de estimación de la matriz de varianzas y covarianzas está basado en la matriz de segundas derivadas parciales de la función de log-verosimilitud (3.6). Ésta resulta ser la inversa de la matriz de información $\mathrm{I}(\beta)$ de tamaño $(p+1) \times (p+1)$(ver Dobson (1990)), cuyos elementos son de la forma

$$I(\beta)_{kl} = E[U_k U_l] = E\left[\frac{\partial L(\beta)}{\partial \beta_k}\frac{\partial L(\beta)}{\partial \beta_l}\right], \tag{3.9}$$

para $k, l = 0, 1, 2, ..., p$; con U_k y U_l definidas en la literatura estadística como el puntaje total con respecto a β_k y β_l respectivamente. Para encontrar el valor de la parte derecha de (3.9) se recurre al útil resultado

$$E\left[\frac{\partial L(\beta)}{\partial \beta_k}\frac{\partial L(\beta)}{\partial \beta_l}\right] = -E\left[\frac{\partial^2 L(\beta)}{\partial \beta_k \partial \beta_l}\right],$$

donde

$$\frac{\partial^2 L(\beta)}{\partial^2 \beta_k} = -\sum_{i=1}^{N} \frac{n_i w_{ik}^2 e^{g(\mathrm{w}_i)}}{(1+e^{g(\mathrm{w}_i)})^2} = -\sum_{i=1}^{N} n_i w_{ik}^2 \pi_i (1-\pi_i)$$

y

$$\frac{\partial^2 L(\beta)}{\partial \beta_k \partial \beta_l} = -\sum_{i=1}^{N} \frac{n_i w_{ik} w_{il} e^{g(\mathrm{w}_i)}}{(1+e^{g(\mathrm{w}_i)})^2} = -\sum_{i=1}^{N} n_i w_{ik} w_{il} \pi_i (1-\pi_i)$$

para $k, l = 0, 1, 2, ..., p$, representan los elementos de la matriz de información, y donde π_i representa a $\pi(\mathrm{w}_i)$.

Como se puede notar, estas ecuaciones no están dadas como funciones de z_i, así que la información esperada y observada son idénticas, y por lo tanto

$$I(\beta)_{kk} = \sum_{i=1}^{N} n_i w_{ik}^2 \pi_i (1-\pi_i) \tag{3.10}$$

$$I(\beta)_{kl} = \sum_{i=1}^{N} n_i w_{ik} w_{il} \pi_i (1-\pi_i) \tag{3.11}$$

para $k, l = 0, 1, 2, ..., p$, representan los elementos de la matriz de información observada. A partir de (3.10) y (3.11) se nota que la matriz de información toma la forma $I(\beta) = \mathcal{W}' V \mathcal{W}$, donde $\mathcal{W}$ es una matriz de tamaño $N \times (p+1)$ que contiene la información de los valores observados del vector w_i que fueron obtenidos para cada patrón de covariables i; $i = 1, 2, ...N$, V es una matriz diagonal de tamaño $N \times N$, cuyo elemento general en la diagonal principal es $\pi_i(1-\pi_i)$. Por lo tanto, la matriz $\mathcal{W}$ es

$$\mathcal{W} = \begin{bmatrix} 1 & w_{11} & w_{12} & \cdots & w_{1p} \\ 1 & w_{21} & w_{22} & \cdots & w_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & w_{N1} & w_{N2} & \cdots & w_{Np} \end{bmatrix}$$

y la matriz V es

$$V = \begin{bmatrix} n_1\pi_1(1-\pi_1) & 0 & \cdots & 0 \\ 0 & n_2\pi_2(1-\pi_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & n_N\pi_N(1-\pi_N) \end{bmatrix},$$

expresiones que junto con (3.10) y (3.11) permiten notar que la matriz de información depende del vector de parámetros β a través de los elementos π_i de la matriz V, por lo que en adelante se recurrirá al uso de $I(\hat{\beta}) = \mathcal{W}'\hat{V}\mathcal{W}$ como un estimador de $I(\beta)$, donde $\hat{V}$ resulta ser la matriz V evaluada en el vector $\hat{\beta}'$ mediante el uso de los valores estimados $\hat{\pi}_i$ que constituyen parte de los elementos en la diagonal principal de $\hat{V}$.

La matriz de varianzas y covarianzas del vector de parámetros estimados $\hat{\beta}$ es calculada a apartir de $\text{Cov}(\hat{\beta}) = I^{-1}(\hat{\beta})$, donde $I^{-1}(\hat{\beta})$ representa la inversa de la matriz de información estimada cuyo k-ésimo elemento de la diagonal se denotará por $\text{Var}(\hat{\beta}_k)$, el cual corresponde a la varianza de $\hat{\beta}_k$, y $\text{Cov}(\hat{\beta}_k, \hat{\beta}_l)$ indica a un término arbitrario fuera de la diagonal, el cual representa la covarianza entre $\hat{\beta}_k$ y $\hat{\beta}_l$, $k, l = 0, 1, 2, ..., p$.

Por último, considerando la distribución asintótica del vector de parámetros estimados $\hat{\beta}$ (ver Dobson (1990)), se tiene que mediante los términos de la diagonal de la matriz $\text{Cov}(\hat{\beta})$, se obtienen elementos para realizar inferencias acerca de los parámetros individuales, al permitir por ejemplo el poder evaluar la precisión de cada uno de los parámetros estimados β_k, esto a través del cálculo de sus errores estándar, cuyos valores quedan determinados por medio de la expresión

$$SCE(\hat{\beta}_k) = \left[\text{Var}(\hat{\beta}_k)\right]^{\frac{1}{2}}, \tag{3.12}$$

resultado que además permite obtener intervalos asintóticos de confianza del $100(1-\alpha)\,\%$ para cada uno de los parámetros, por medio de la expresión

$$\hat{\beta}_k \pm \zeta_{1-\alpha/2} SCE(\hat{\beta}_k), \tag{3.13}$$

donde $\zeta_{1-\alpha/2}$ representa el punto en el extremo superior correspondiente al $100(1-\alpha/2)\,\%$ del área en una distribución normal estándar.

3.2.2. Evaluación de la significancia estadística de los coeficientes del modelo

Una vez que se ha ajustado el modelo con las variables consideradas relevantes, el paso a seguir es la determinación del grado de influencia que cada una de éstas (así como de sus interacciones en caso de haber sido consideradas) tienen sobre la variable respuesta. Para lograr tal propósito será necesario analizar la significancia estadística de los parámetros estimados; esto es, probar si algunos de éstos o bien todos son estadísticamente cero. Este contraste lleva

a la necesidad de formular y probar hipótesis estadísticas sobre los parámetros de interés, situación que conduce a dar una descripción de la distribución muestral de algunos estadísticos que serán de utilidad en la realización de este diagnóstico.

Partiendo del supuesto de que se ha ajustado un modelo con $p+1$ parámetros estimados, a continuación se procede a describir los estadísticos de prueba así como sus contrastes asociados, los cuales serán de utilidad en la medición del nivel de significancia estadística de dichos parámetros.

Prueba de Wald

Con la finalidad de definir este contraste, se da la descripción del estadístico de Wald, el cual queda determinado como

$$\begin{aligned} W &= \hat{\beta}' \Big[\text{Cov}(\hat{\beta}) \Big]^{-1} \hat{\beta} \\ &= \hat{\beta}' I(\hat{\beta}) \hat{\beta} \\ &= \hat{\beta}' (\mathcal{W}' \hat{V} \mathcal{W}) \hat{\beta}, \end{aligned}$$

donde su distribución es asintótica ji-cuadrada con p grados de libertad bajo la hipótesis nula

$$H_0 : \beta = 0.$$

Este estadístico tiene su equivalente univariado que resulta ser de utilidad en la evaluación de la sigificancia estadística en forma individual para cada uno de los parámetros involucrados en el modelo. Éste es definido a través de la relación

$$W_k = \frac{\hat{\beta}_k}{SCE(\hat{\beta}_k),}$$

para $k = 0, 1, 2, ..., p$. Bajo la hipótesis $\beta_k = 0$, este estadístico seguirá una distribución asintótica normal estándar.

Prueba del Cociente de Verosimilitudes

Para concluir, se presenta otra prueba que permitirá contrastar hipótesis acerca de los parámetros estimados. Para ello se comienza definiendo el estadístico

$$\mathcal{D} = -2 \ln \left[\frac{L(\hat{\beta})}{L(\hat{\beta}_{sat})} \right], \tag{3.14}$$

expresión que en la literatura estadística se conoce como la devianza (o estadístico del cociente de verosimilitudes), y cuyos términos $L(\hat{\beta})$ y $L(\hat{\beta}_{sat})$ representan las funciones de verosimilitud para el modelo ajustado y saturado respectivamente, siendo evaluadas en sus correspondientes estimadores de máxima verosimilitud $\hat{\beta}$ y $\hat{\beta}_{sat}$. Este estadístico realiza una comparación entre el modelo ajustado y el modelo saturado, donde bajo el supuesto de que el modelo ajustado es el correcto, su distribución se considera como asintótica ji-cuadrada con $N - (p + 1)$ grados de libertad. Cabe aclarar que se considera como modelo saturado a aquel cuyo número de parámetros estimados es igual al número total de observaciones.

Partiendo del principio de parsimonia, el modelo ajustado puede servir como base para evaluar si existe la posibilidad de encontrar un modelo más simple; esto es, un modelo que contenga un número menor de parámetros pero que aún de una adecuada descripción de los datos. Esta evaluación conduce a particionar el vector de parámetros β en (β_1, β_2), donde β_2 representa el vector que contiene los parámetros asociados al modelo más simple y cuya dimensión se considera como q, con $q < p + 1$. La valoración de este resultado se realiza a través del planteamiento del contraste

$$H_0 : \beta_1 = 0,$$

el cual representa la hipótesis nula que considera a los parámetros no incluidos en modelo más simple como estadísticamente iguales a cero.

Para poder realizar esta prueba, se deben comparar los valores de la devianza para estos dos modelos en competencia. La devianza del modelo base es calculada a partir de la ecuación (3.14), mientras que el valor para la devianza del modelo más simple es obtenida a partir del vector de parámetros estimados $\hat{\beta}_2$, estimado bajo la restricción $\beta_1 = 0$, cuyo valor se representa por

$$\mathcal{D}_2 = -2 \ln \left[\frac{L(\hat{\beta}_2)}{L(\hat{\beta}_{sat})} \right].$$

Para realizar el contraste se debe comparar la diferencia de estos dos estadísticos, por lo que si se llama G a esta diferncia entonces

$$G = \mathcal{D} - \mathcal{D}_2 = -2 \left[\ln[L(\hat{\beta})] - \ln[L(\hat{\beta}_{sat})] \right] + 2 \left[\ln[L(\hat{\beta}_2)] - \ln[L(\hat{\beta}_{sat})] \right].$$

Como la verosimilitud del modelo saturado es común a ambos valores de la devianza, entonces ésta puede ser expresada como

$$G = -2 \ln \left[\frac{L(\hat{\beta})}{L(\hat{\beta}_2)} \right], \tag{3.15}$$

cuya distribución queda determinada en forma asintótica como ji-cuadrada con $(p+1)-q$ grados de libertad. En general, si el valor de este estadístico es consistente con el valor de dicha distribución, se favorece el modelo correspondiente a

H_0 por tratarse de una representación más simple, en caso contrario, se favorece al representado por los $p+1$ parámetros por proveer una mejor descripción de los datos (ver Hosmer y Lemeshow (2000)).

3.3. Regresión Logística Exacta

Como se ha venido puntualizando durante la presentación del modelo de regresión logística, el supuesto de normalidad en los parámetros estimados así como los supuestos distribucionales del estadístico del cociente de verosimilitudes y de la prueba de Wald son propiedades asintóticas. Por tal motivo, a lo largo del análisis será pertinente tener en consideración que dichos supuestos sean cabalmente aplicables al conjunto de datos que se va a analizar.

Si bien es cierto que el número de observaciones contempladas a ser utilizadas en el ajuste del modelo es lo suficientemente grande como para poder pensar que los supuestos distribucionales asintóticos son los correctos, se tiene que la proporción de respuestas ($z=1$) es cercana a 0 (aproximadamente 7 %), lo que indica que se está ante la presencia de una base de datos desbalanceada. Como la base de datos considerada en el presente trabajo posee esta característica y además esta conformada por variables explicativas de tipo categórico, para cada tabla de contingencia obtenida a partir del cruce de la respuesta ($z=0,1$) versus los k niveles de cada una de las variables explicativas, se pudieran generar celdas con ninguna o bien con pocas observaciones, derivando con ello en posibles problemas de estimación en el modelo, lo cual está caracterizado por presentar coeficientes estimados y/o errores estándar estimados inusualmente grandes. Por tal motivo, cuando en el Capítulo 5 se haga uso de la base datos, se aplicará un análisis de regresión logística exacto como una alternativa que permitirá validar las inferencias obtenidas por el análisis asintótico de los parámetros estimados, o en el supuesto de que haya estimadores inusualmente grandes, proporcionará un modelo alternativo de ajuste para el análisis de los datos disponibles (ver Collett (2002).

3.3.1. La Función de Verosimilitud Condicional

Para dar comienzo con la discusión del análisis de regresión logístico exacto, se comenzará por definir la función de verosimilitud condicional, la cual en analogía con su equivalente incondicional, es de gran utilidad en la estimación de parámetros así como en la realización de los contrastes de hipótesis requeridos para este tipo de análisis.

Si en la función de densidad de probabilidad conjunta definida en (3.5) se consideran los valores $t_o = \sum_{i=1}^{N} z_i$ y $t_j = \sum_{i=1}^{N} z_i w_{ij}$ para $j = 1, 2, ..., p$, entonces se tiene que ésta toma la forma

$$P[\mathcal{Z}] = \frac{\exp\left(\sum_{j=0}^{p} \beta_j t_j\right)}{\prod_{i=1}^{N} \left[1 + \exp\left(g(\mathrm{w}_i)\right)\right]^{n_i}} \times \prod_{i=1}^{N} \left(\begin{array}{c} n_i \\ z_i \end{array} \right), \tag{3.16}$$

donde las cantidades $t_0, t_1, ..., t_p$ representan los valores observados de la variable aleatoria $T_j = \sum_{i=1}^{N} w_{ij} Zi$. Al considerar a la ecuación (3.16) como una función de verosimilitud se puede observar que t_j es de hecho un estadístico suficiente para β_j. Este resultado se sigue directamente del Criterio de Factorización de Neyman, puesto que la verosimilitud representada por esta ecuación es simplemente una función de los parámetros y de los estadísticos suficientes.

A fin de estructurar una función de verosimilitud condicional, el interés se centra ahora en la estimación de los parámetros del modelo de regresión logística a partir del análisis exacto. Supóngase sin pérdida de generalidad que se elige a β_p como el primer parámetro a estimar. Una vez determinado éste, los parámetros restantes $\beta_0, ...\beta_{p-1}$ serán referidos como parámetros de ruido.

El procedimiento de estimación de este parámetro conduce en un principio a determinar su correspondiente función de verosimilitud condicional, siendo ésta obtenida al eliminar los parámetros de ruido de (3.16) condicionando en los valores de sus respectivos estadísticos suficientes $T_0, T_1, ..., T_{p-1}$. Para lograr esto, se comienza definiendo la función de verosimilitud condicional para β_p, cuya expresión queda expresada a partir de la ecuación

$$L_c(\beta_p) = P[T_p = t_p | T_0 = t_0, ..., T_{p-1} = t_{p-1}], \tag{3.17}$$

o bien como

$$L_c(\beta_p) = \frac{P[T_o = t_o, T_1 = t_1, ..., T_p = t_p]}{P[T_o = t_o, T_1 = t_1, ..., T_{p-1} = t_{p-1}]}. \tag{3.18}$$

Como ya se advirtió más arriba, los estadísticos suficientes considerados son funciones de las variables aleatorias $Z_1, Z_2, ..., Z_N$. Por consiguiente las cantidades observadas de éstos claramente dependen de las observaciones $z_1, z_2, ..., z_N$. Por lo tanto, es evidente que el valor de la probabilidad en el numerador de la función de verosimilitud condicional estará asociado con (3.16). Por otra parte, teniendo en mente que estas observaciones representan exclusivamente una realización de las variables aleatorias $Z_1, Z_2, ..., Z_N$, entonces el valor de dicho numerador también deberá tomar en cuenta el número total de posibles realizaciones de éstas variables aleatorias, para las cuales los estadísticos suficientes tomen los valores $t_0, t_1, ..., t_p$. Si se denota a este número por $c(t_0, t_1, ..., t_p)$, la probabilidad conjunta de $T_j = t_j$, para $j = 0, 1, ..., p$, quedará definida al sumar la verosimilitud incondicional sobre el número total de estas posibles realizaciones. De este modo, la expresión resultante queda definida como

$$P[T_0 = t_0, ..., T_p = t_p] = \frac{c(t_0, t_1, ..., t_p) \exp\left(\sum_{j=0}^{p} \beta_j t_j\right)}{\prod_{i=1}^{N} \left[1 + \exp\left(g(\mathrm{w}_i)\right)\right]^{n_i}} \times \prod_{i=1}^{N} \begin{pmatrix} n_i \\ z_i \end{pmatrix}. \quad (3.19)$$

El siguiente paso consiste en determinar la expresión para el denominador en (3.18), cuyo valor representa la distribución marginal de $T_0, T_1, ..., T_{p-1}$. Por lo tanto ésta queda definida al sumar las probabilidades sobre todos los posibles valores de T_p para aquellas observaciones de $Z_1, Z_2, ..., Z_N$ que cumplan con la condición $t_o = \sum_{i=1}^{N} z_i$ y $t_j = \sum_{i=1}^{N} z_i w_{ij}$ para $j = 1, 2, ..., p-1$. Por consiguiente, la distribución de probabilidad conjunta requerida queda determinada por la expresión

$$\frac{\sum_u c(t_0, t_1, ..., t_{p-1}, u) \exp\left(\beta_p u + \sum_{j=0}^{p-1} \beta_j t_j\right)}{\prod_{i=1}^{N} \left[1 + \exp\left(g(\mathrm{w}_i)\right)\right]^{n_i}} \times \prod_{i=1}^{N} \begin{pmatrix} n_i \\ z_i \end{pmatrix}. \quad (3.20)$$

Así, si se utilizan los resultados de las ecuaciones (3.19) y (3.20), la función de verosimilitud condicional definida en (3.18) queda representada por la ecuación

$$L_c(\beta_p) = \frac{c(t_0, t_1, ..., t_p) \exp\left(\sum_{j=0}^{p} \beta_j t_j\right)}{\sum_u c(t_0, t_1, ..., t_{p-1}, u) \exp\left(\beta_p u + \sum_{j=0}^{p-1} \beta_j t_j\right)}, \quad (3.21)$$

o bien por

$$L_c(\beta_p) = \frac{c(t_0, t_1, ..., t_p) \exp\left(\beta_p t_p\right)}{\sum_u c(t_0, t_1, ..., t_{p-1}, u) \exp\left(\beta_p u\right)}. \quad (3.22)$$

De la última ecuación se puede observar que la función de verosimilitud condicional sólo depende de un parámetro. En consecuencia, si se desea estimar el valor de éste, bastará con maximizar dicha función con respecto a β_p, obteniendo así el llamado estimador condicional de máxima verosimilitud. A partir de un procedimieto análogo se pueden determinar los estimadores de los restantes p parámetros y de esta forma construir el modelo de regresión logística.

3.3.2. Pruebas de Hipótesis Exactas

Una vez definido el procedimiento de ajuste de un modelo de regresión logística a partir de la realización de un análisis exacto, a continuación se proporciona un contraste estadístico de hipótesis que es acorde con su estructura. Dicho contraste, conocido como la prueba exacta de *scores* condicional basada en la

varianza exacta, será el utilizado para realizar el procedimiento de contraste de hipótesis sobre los parámetros del modelo de regresión logística, cuando éste haya sido ajustado desde la perspectiva del análisis exacto.

Prueba Exacta de *Scores* Condicional basada en la Varianza Exacta

Con la finalidad de realizar una posterior generalización, se comienza dando una formulación para el estadístico de prueba para el caso univariado, el cual se define como $U(\beta_p)^2/i(\beta_p)$, donde $U(\beta_p)$ es el valor del *score* para β_p, $U(\beta_p) = \partial \log L_c(\beta_p)/\partial\beta_p$, e $i(\beta_p)$ es el valor de la función de información dada por

$$i(\beta_p) = E[\{U(\beta_p)\}^2 | T_0 = t_0, ..., T_{p-1} = t_{p-1}].$$

$\log L_c(\beta_p)$ resulta ser la función de log-verosimilitud condicional para β_p, la cual queda definida a partir de (3.22) como

$$\log L_c(\beta_p) = \log c(t_0, t_1, ..., t_p) + \beta_p t_p - \log \left\{ \sum_u c(t_0, t_1, ..., t_{p-1}, u) \exp\left(\beta_p u\right) \right\}.$$

Si se deriva esta cantidad con respecto a β_p se obtiene como resultado

$$U(\beta_p) = t_p - \frac{\sum_u uc(t_0, t_1, ..., t_{p-1}, u) \exp\left(\beta_p u\right)}{\sum_u c(t_0, t_1, ..., t_{p-1}, u) exp\left(\beta_p u\right)}. \tag{3.23}$$

De las ecuaciones (3.17) y (3.22), el segundo término de esta expresión puede ser escrito como

$$m_p = \sum_u uP(T_p = u | T_0 = t_0, ..., T_{p-1} = t_{p-1}),$$

cantidad que define la esperanza condicional de la variable aleatoria T_p, cuya varianza condicional queda determinada por la ecuación

$$v_p = \sum_u (u - m_p)^2 P(T_p = u | T_0 = t_0, ..., T_{p-1} = t_{p-1}).$$

Si se considera a $U(\beta_p)$ como una función de la variable aleatoria T_p, esto es $U(\beta_p) = T_p - m_p$, entonces

$$E[\{U(\beta_p)\}^2 | T_0 = t_0, ..., T_{p-1} = t_{p-1}] = E[(T_p - m_p)^2 | T_0 = t_0, ..., T_{p-1} = t_{p-1}]$$

y así i_p define la varianza condicional v_p de T_p. Por lo tanto el estadístico *score* para el caso univariado queda expresado como

$$Q_{sc} = \frac{(T_p - m_p)^2}{v_p}, \tag{3.24}$$

de donde

$$q_{sc} = \frac{(t_p - m_p)^2}{v_p}, \tag{3.25}$$

es el valor observado del estadístico para $T_p = t_p$. Así entonces, bajo la hipótesis nula $H_0 : \beta_p = 0$, mediante la sustitución de este valor en (3.23), es posible obtener los valores de m_p y v_p y con ello el valor del estadístico (3.25). De igual forma bajo esta hipótesis nula también se puede calcular el valor del estadístico de prueba para todos los posibles valores de la variable T_p.

Debido a que cuanto mayor sea el valor de este estadístico más fuerte es la evidencia en contra H_0, el valor-p basado en la prueba del score condicional exacto es calculado como

$$p = P(Q_{sc} \geq q_{sc}).$$

Este valor es obtenido a partir de sumar todas las probabilidades condicionales $P(T_p = u|T_0 = t_0, ..., T_{p-1} = t_{p-1})$ asociadas con todos aquellos valores de u para los cuales el estadístico resultó ser mayor o igual a q_{sc}, obteniendose así un valor p exacto para una prueba bilateral de la hipótesis $H_0 : \beta_p = 0$.

En general, supóngase ahora que se tiene un vector de parámetros β particionado en $(\beta_{\mathbf{1}}, \beta_{\mathbf{2}})$, y que $(\mathbf{t_1}, \mathbf{t_2})$ son los valores observados de los correspondientes estadísticos suficientes $\mathbf{T_1}$ y $\mathbf{T_2}$. Si se desea probar la hipótesis

$$H_0 : \beta_{\mathbf{1}} = \mathbf{0},$$

el estadístico score para el caso multivariado queda expresado como

$$Q_{sc} = (\mathbf{T_1} - \mathbf{m_1})' V_1^{-1} (\mathbf{T_1} - \mathbf{m_1}), \tag{3.26}$$

o a su correspondiente valor observado

$$q_{sc} = (\mathbf{t_1} - \mathbf{m_1})' \mathbf{V_1}^{-1} (\mathbf{t_1} - \mathbf{m_1}), \tag{3.27}$$

donde $\mathbf{m_1}$ es la media y $\mathbf{V_1}$ es la matriz de varianzas-covarianzas de la distribución exacta condicional en $\mathbf{T_2} = \mathbf{t_2}$ del vector de estadísticos suficientes, $\mathbf{T_1}$, correspondiente al vector de parámetros,$\beta_{\mathbf{1}}$, no incluidos en el modelo más simple. De la misma forma en que lo hicimos para el caso univariado, el valor-p exacto para una prueba bilateral de la hipótesis

$$H_0 : \beta_{\mathbf{1}} = \mathbf{0},$$

se calcula como

$$p = P(Q_{sc} \geq q_{sc}).$$

cuyo valor es obtenido mediante la suma de las probabilidades condicionales $P(\mathbf{u_1}|\mathbf{t_2})$ para todos aquellos valores del vector $\mathbf{T_1} = \mathbf{u_1}$ que bajo la hipótesis nula dieron lugar a valores del estadístico de prueba mayores o iguales a q_{sc}.

Si los vectores $\mathbf{T_1}$, $\mathbf{m_1}$ y la la matriz de varianzas-covarianzas $\mathbf{V_1}$ son considerados como escalares en las expresiones para el estadístico *score* dadas en

las ecuaciones (3.26) y (3.27), éstas se particularizan a las expresiones dadas en (3.24) y (3.25) respectivamente (ver Collett (2002)).

Capítulo 4

El problema de los datos perdidos

4.1. Introducción

En algunas ocasiones, por circunstancias fuera de su control, el investigador o el analista se enfrenta a la problemática de analizar información contenida en bases de datos con registros faltantes. Tal situación lo conduce a enfrentar la difícil tarea de buscar cómo procesar, analizar y obtener conclusiones válidas con los datos disponibles, pues el llevar a cabo las técnicas tradicionales de modelación estadística a un conjunto de datos cuando estos están incompletos puede derivar en sesgos y errores en las conclusiones y resultados.

Dado que las técnicas estadísticas para el análisis de datos consideran conjuntos de datos completos, una alternativa (que en primera instancia posee un atractivo práctico ante esta eventualidad) es descartar del análisis a aquellas unidades muestrales que tienen uno o más datos faltantes, pues realizar esto conlleva a tener una base de datos completa y con ello la garantía de poder realizar los análisis estadísticos contemplados en forma directa. Sin embargo, cuando se presenta un escenario multivariado donde los valores perdidos ocurren en más de una variable, es muy probable que estos constituyan una porción sustancial de la base de datos; bajo este escenario, la ventaja práctica que supondría hacer esto se contrapone a los inconvenientes inferenciales derivados de su aplicación, pues eliminar una cantidad importante de información tenderá a introducir sesgos. Además, la base de datos resultante puede llegar a ser representativa de la población de casos sin datos faltantes, más que de la población de todos los casos (Schafer, 1997).

Tomando en cuenta las desventajas que conlleva eliminar casos, existe otra alternativa metodológica denominada con el nombre genérico de imputación, la cual contempla el llenado (imputación) de los datos faltantes con valores plausibles. Dichos valores pueden ser generados a partir de dos técnicas. La primera, conocida como imputación sencilla, considera el llenado a partir de la media

observada de la variable cuyos datos se desean imputar, o bien, usando valores predichos a partir de un modelo de regresión, generando de esta manera un conjunto de datos completos. La segunda, conocida con el nombre de imputación múltiple, consiste en realizar la imputación a partir de un modelo de predicción de los valores faltantes, obteniendo con ello m valores imputados para cada una de las observaciones faltantes, creando así m conjuntos de datos completos. Si bien es cierto que ambas técnicas comparten la habilidad de generar un conjunto completo de datos para llevar a cabo cualquier tipo de análisis a partir de la aplicación de métodos estándar para datos completos, el empleo de la primera posee la desventaja de originar valores engañosos para las medidas de cálculo de la incertidumbre (ej. errores estándar, valor-p) al no considerar éstas la incertidumbre debida a la variabilidad de los datos perdidos (Schafer, 1997). Esta desventaja se ve superada al hacer uso de la técnica de imputación múltiple, pues los resultados obtenidos para los m conjuntos de datos imputados pueden ser fácilmente combinados para crear una inferencia válida que refleje la variabilidad muestral debida a valores faltantes (Little y Rubin, 1987).

Considerando la desventaja arriba mencionada, la primera técnica será completamente descartada, por lo que en lo que resta de este trabajo se aplicará solamente la técnica de imputación múltiple; la cual, si bien, como se verá en su momento, presenta limitaciones que no deben ser pasadas por alto y por las cuales no puede ser considerada como la panacea, por sus características ofrece la opción más viable para tratar con el problema de informacion perdida en el conjunto de datos utilizados para el desarrollo del análisis.

El enfoque a lo largo del presente capítulo consiste en la revisión de los supuestos que dan sustento a la aplicación del método de imputación múltiple. Se revisará el método de estimación de parámetros conocido como EM así como el algoritmo de Aumento de Datos; los cuales, como se verá en su momento, se incertan en el proceso de imputación múltiple. Una vez detallado en forma general el método de imputación múltiple, se finalizará con la descripción de las particularidades que el proceso de imputación múltiple adopta derivado de su aplicación al esquema de información perdida considerado.

4.2. Supuestos

Antes de poder aplicar adecuadamente un proceso de imputación múltiple es fundamental conocer los supuestos sobre los cuales versa la aplicación de esta metodología; pues de la medida en que estos se cumplen depende la validez y confiabilidad de los resultados obtenidos. Por tal motivo se describen a continuación dichos supuestos.

4.2.1. El modelo para los datos completos

En el contexto de los datos faltantes se entenderá por datos completos a aquel conjunto de datos muestrales conformado tanto por los datos observados como por los no observados; esto es, se tiene un conjunto de variables no obser-

vadas total o parcialmente (variables latentes). Se supone que las observaciones (tanto presentes como ausentes) de dicho conjunto para cada uno de los individuos incluidos en la muestra son el resultado de realizaciones independientes e idénticamente distribuidas (iid) de una distribución de probabilidad conocida.

	1	2	3	$\cdots$	p
1		?			
2			?		?
3					
.	?	?			
.	?				
.		?	?		
.	?		?		?
.					
.	?	?			
.					
.					?
n	?		?		

Tabla 4.1: Base de datos multivariada con valores perdidos

Una representación esquemática de un conjunto de datos completos es mostrada en la Tabla 4.1. El cual representa un arreglo rectangular, al que se denotará por Y, conformado por una matriz de tamaño $n \times p$, cuyos n renglones representan unidades observadas y las p columnas representan las variables registradas para esas unidades. Dado que esta matriz no está completamente observada sus valores perdidos son representados por los signos de interrogación.

Si ahora se considera a cada uno de los renglones de la matriz Y como una realización y_i de una variable aleatoria con función de densidad de probabilidad f, $i = 1, ..., n$, por el supuesto de independencia se tiene que la función de densidad de probabilidad para los datos completos puede ser escrita como

$$P(Y|\theta) = \prod_{i=1}^{n} f(y_i|\theta),$$

donde θ es un vector de parámetros desconocidos.

4.2.2. Mecanismos de pérdida de infomación

Autores como Little y Rubin (1987) han definido tres tipos de mecanismos de generación de información incompleta: datos faltantes estrictamente aleatorios (DFEA), datos faltantes aleatorios (DFA) y datos faltantes no aleatorios (DFNA); este último también se conoce como datos faltantes con mecanismo no ignorable.

Antes de dar paso a la descripción de estos supuestos, se define a R como una matriz de tamaño $n \times p$ de variables indicadoras cuyos elementos toman valores

0 ó 1 dependiendo de si los correspondientes elementos de Y son observados o no observados, asociando de esta manera a cada muestra en Y una muestra en R; además, si se representa la parte observada y perdida de Y como Y_{obs} y Y_{per} respectivamente, entonces la matriz Y puede ser caracterizada de la forma $Y = (Y_{obs}, Y_{per})$. En base a esta notación, en la Tabla 4.2 se da una breve descripción de estos tres supuestos de pérdida de información.

Tabla 4.2: Mecanismos de generación de información incompleta

Mecanismo	Características
Datos faltantes estrictamente aleatorios (missing completely at random).	Mecanismo de pérdida de información caracterizado por considerar que la pérdida de información es independiente tanto de los valores observados como de los valores perdidos de Y. Formalmente esto significa que R es independiente de Y. Donde, $P(R\|Y) = P(R)$.
Datos faltantes aleatorios (missing at random).	Mecanismo de pérdida de información caracterizado por considerar que la pérdida de información es independiente de los valores perdidos de Y pero no de sus valores observados. Formalmente esto significa que R es independiente de Y_{per}. Donde, $P(R\|Y) = P(R\|Y_{obs})$.
Datos faltantes no aleatorios (not missing at random)	Mecanismo de pérdida de información caracterizado por considerar que la pérdida de información depende de los valores perdidos y observados de Y. Formalmente esto significa que R es dependiente de Y_{per} y Y_{obs}. Donde, $P(R\|Y) = P(R\|Y_{obs}, Y_{per})$.

El desconocimiento del mecanismo que generó la pérdida de información como consecuencia de la falta de supervisión al momento de aplicar el cuestionario, permiten considerar estos tres mecanismos de información faltante como plausibles. Sin embargo, si por su simplicidad se decidiera inclinarse por el mecanismo de pérdida de información DFEA, esto implicaría descartar del análisis a todas aquellas unidades muestrales que tienen uno o más datos faltantes, consideración que como ya se mencionó es desechada por la potencial pérdida de información que se deriva de descartar casos incompletos en una base de datos multivariada, lo que conlleva a prescindir del supuesto DFEA. Por otra parte, si bien es cierto que bajo el esquema de imputación múltiple los mecanismos de pérdida de información DFA y DFNA otorgan la posibilidad de obtener bases de datos completos sin necesidad de eliminar casos, la dependencia que este último tiene sobre la distribución de probabilidad $P(R|Y)$ lo hace muy sensible a la elección de este modelo probabilístico. Por lo tanto, si se considera la desventaja que representa agregar el supuesto no ignorable al tratar de asignar una distribución de probabilidad $P(R|Y)$ sin un conocimiento inicial del proceso que

dió origen a la pérdida de información, y dado que el supuesto DFA está caracterizado por ignorar este mecanismo, el presente trabajo se inclinará en favor de este supuesto para dar solución al problema de los datos faltantes. Es importante hacer notar que debido a que no se cuenta un procedimiento general que permita contrastar esta hipótesis, se debe tener en cuenta que los resultados obtenidos pudieran verse afectados por desviaciones de la hipótesis DFA, siendo necesario considerarlos con cautela.

4.2.3. Determinación de la función de verosimilitud bajo el supuesto DFA

En algún momento durante el proceso de imputación de datos resultará necesario establecer una función de verosimilitud que permita estimar el vector de parámetros θ asociado a la función de densidad de probabilidad $P(Y|\theta)$. En la formulación de dicha verosimilitud para la estimación del vector de parámetros θ bajo el supuesto DFA que se supone el hecho $P(R|Y_{obs}, Y_{per}, \xi) = P(R|Y_{obs}, \xi)$, con ξ un vector de parámetros desconocidos, se comienza por establecer la función de densidad de probabilidad conjunta para los datos observados Y_{obs} y R, la cual queda definida al integrar sobre Y_{per} la densidad conjunta de Y y R. Esto es,

$$\begin{aligned} P(R, Y_{obs}|\theta, \xi) &= \int P(R, Y|\theta, \xi)\, dY_{per} \\ &= \int P(R|Y, \xi)P(Y|\theta)\, dY_{per}, \end{aligned} \tag{4.1}$$

por lo que la función de verosimilitud queda definida por

$$L(\theta, \xi|Y_{obs}, R) \propto P(R, Y_{obs}|\theta, \xi). \tag{4.2}$$

Por otra parte, si se aplica el supuesto DFA en (4.1) la expresión correspondiente para esta densidad conjunta toma la forma

$$P(R, Y_{obs}|\theta, \xi) = P(R|Y_{obs}, \xi)P(Y_{obs}|\theta),$$

donde su correspondiente función de verosimilitud queda definida por

$$L(\theta, \xi|Y_{obs}, R) \propto P(R|Y_{obs}, \xi)P(Y_{obs}|\theta). \tag{4.3}$$

De las expresión anterior se puede observar que si se supone que los parámetros θ y ξ son distintos, es decir, si el conocer los valores de θ no provee de información adicional acerca de ξ y viceversa, entonces las inferencias sobre θ basadas en esta verosimilitud no se verán afectadas por el parámetro ξ. Por tal motivo, éste puede ser considerado un parámetro de ruido y el primer factor a la derecha de (4.3) se puede considerar irrelevante en la estructuración de la verosimilitud. Así entonces, para realizar la construcción de la verosimilitud

bastará con considerar el segundo factor en (4.3) correspondiente a θ o cualquier función proporcional a éste como la función de verosimiltud. Esto es,

$$L(\theta|Y_{obs}) \propto P(Y_{obs}|\theta). \tag{4.4}$$

Por lo tanto, la asignación de este supuesto paramétrico a la hipótesis DFA permite considerar el mecanismo de pérdida de información como ignorable para términos de la estimación del parámetro θ, y con ello el poder concluir que toda la información estadística relevante acerca de θ estará contenida en la verosimilitud de los datos observados (Schafer, 1997).

4.3. El algoritmo EM

En algunas aplicaciones se encuentra que la forma funcional asumida por la función de verosimilitud (4.4) de los datos observados es demasiado complicada como para aplicar los métodos tradicionales en el cálculo de los estimadores de máxima verosimilitud para θ. En tales circunstancias, la estimación de θ requiere de la aplicación del algoritmo EM (Expectation Maximization), el cual deriva su nombre de los dos pasos requeridos para su implementación: el paso E, correspondiente al cálculo de la esperanza; el paso M, correspondiente al paso de maximización.

Para proceder con la realización del paso E, es indispensable contar con un valor preliminar $\theta^{(t)}$ para el parámetro θ. Una vez determinado éste, el paso E consiste en encontrar el valor esperado de la log-verosimilitud para datos completos, $\ell(\theta|Y) = \log L(\theta|Y)$, con respecto a los datos perdidos Y_{per}, condicionando en los datos observados Y_{obs} y el valor $\theta^{(t)}$. Obteniéndose de esta forma una esperanza condicional para la función de log-verosimilitud $\ell(\theta|Y)$ para datos completos, cuya representación está dada por la la expresión

$$\mathcal{Q}(\theta|\theta^{(t)}) = \int \ell(\theta|Y) P(Y_{per}|Y_{obs}, \theta^{(t)}) dY_{per}. \tag{4.5}$$

Para realizar el paso M bastará con maximizar esta log-verosimilitud esperada para encontrar el valor $\theta^{(t+1)}$ que cumpla con la condición

$$\mathcal{Q}(\theta^{(t+1)}|\theta^{(t)}) \geq \mathcal{Q}(\theta|\theta^{(t)}), \qquad \text{para todo } \theta.$$

De cumplirse esta condición, se puede mostrar que el cambio de $\theta^{(t)}$ a $\theta^{(t+1)}$ incrementa la log-verosimilitud de los datos observados $\ell(\theta|Y_{obs}) = \log L(\theta|Y_{obs})$ (ver Schafer, 1997); esto es,

$$\ell(\theta^{(t+1)}|Y_{obs}) \geq \ell(\theta^{(t)}|Y_{obs}).$$

Así entonces, se tiene que en términos de la log-verosimilitud de los datos observados, $\theta^{(t+1)}$ es un mejor estimador que $\theta^{(t)}$.

Si estos dos pasos son repetidos múltiples veces comenzando con un valor de inicio $\theta^{(0)}$, se define de esta manera una sucesión $\{\theta^{(t)} : t = 0, 1, 2...\}$. Esta sucesión en problemas bien comportados converge a un punto estacionario que

resulta ser un máximo global para $\ell(\theta|Y_{obs})$; derivándose así un proceso iterativo que da como resultado un estimador de máxima verosimilitud único para θ (Schafer, 1997).

Familia Exponencial y algoritmo EM

Si la función de densidad de probabilidad que ha generado los datos completos pertenece a la familia exponencial regular, la aplicación del algoritmo EM será particularmente simple. Para poder mostrar la forma en la cual se da esta simplificación, se parte de que Y tiene una función de densidad de probabilidad perteneciente a la familia exponencial regular. Esto es, se está suponiendo que la función de densidad de probabilidad f para la observación y_i del renglón i de la matriz Y tiene la forma

$$f(y_i|\theta) = a(\theta)b(y_i)\exp\left(\sum_{j=1}^{k} c_j(\theta)s_j(y_i)\right); \tag{4.6}$$

para $i = 1, 2, ..., n$, y $\theta = (\theta_1, ..., \theta_k)$; a partir de este supuesto se concluye que la función de densidad de probabilidad para los datos completos queda determinada por la expresión

$$P(Y|\theta) = [a(\theta)]^n B(Y)\exp\left(c(\theta)^T S(Y)\right), \tag{4.7}$$

donde $c(\theta) = (c_1(\theta), c_2(\theta), ..., c_k(\theta))^T$, $S(Y) = (S_1(Y), S_2(Y), ..., S_k(Y))^T$ es el vector de estadísticos suficientes para los datos completos, con

$$S_j = \sum_{i=1}^{n} s_j(y_i)$$

para $j = 1, 2, ..., k$. Si en base a la expresión (4.7) se considera la log-verosimilitud para los datos completos, se tendrá que la expresión resultante para ésta puede ser escrita como

$$\ell(\theta|Y) = c(\theta)^T S(Y) + na(\theta). \tag{4.8}$$

Si ahora se calcula (4.5) en términos de esta expresión, el paso E puede ser escrito como

$$\mathcal{Q}(\theta|\theta^{(t)}) = c(\theta)^T S^{(t)} + na(\theta), \tag{4.9}$$

donde $S^{(t)} = E(S(Y)|Y_{obs}, \theta^{(t)})$ es un estimador de los estadísticos suficientes. De las ecuaciones (4.8) y (4.9), salta a la vista que ambas tienen la misma forma funcional. Por tal motivo, el procedimiento para encontrar el valor de $\theta^{(t)}$ a partir de maximizar (4.9) no es computacionalmente diferente de aquel que maximiza a (4.8). Si se considera el hecho de que para los miembros de la familia exponencial regular sus estimadores de máxima verosimilitud en (4.8) pueden ser encontrados como soluciones de las ecuaciones de momentos

$$E(S(Y)|\theta) = S,$$

donde S es el valor observado del vector $S(Y)$ y cuya esperanza es tomada con respecto a $P(Y|\theta)$ (Schafer, 1997), se tiene entonces que para maximizar (4.9) bastará con determinar a $\theta^{(t+1)}$ como solución de las ecuaciones de momentos

$$E(S(Y)|\theta) = S^{(t)}.$$

Por lo tanto, bajo estas condiciones, el algoritmo EM tiene la siguiente forma simplificada:

Paso E: Estimar el estadístico suficiente S(Y) para el conjunto de datos completos, resolviendo la ecuación

$$S^{(t)} = E(S(Y)|Y_{obs}, \theta^{(t)}). \tag{4.10}$$

Paso M: Determinar $\theta^{(t+1)}$ como la solución de las ecuaciones

$$E(S(Y)|\theta) = S^{(t)}. \tag{4.11}$$

4.3.1. EM para la moda de una distribución final

Como se verá más adelante, en algunas circunstancias el uso del estimador de máxima verosimilitud para θ no será el más adecuado para lograr la máxima eficiencia del proceso de imputación múltiple. De darse el caso, se debe recurrir a una versión modificada del algoritmo EM que permita encontrar un estimador θ alternativo. La versión modificada a considerar es la correspondiente al algoritmo EM para la moda de una distribución final, cuya definición está fundamentada en elementos de la inferencia estadística Bayesiana, por lo que antes de dar comienzo con la descripción del algoritmo modificado EM para la moda de una distribución final, se parte primero de dar una breve descripción de los conceptos de la inferencia estadística Bayesiana que dan sustento a dicha modificación.

Algunos conceptos de Estadística Bayesiana

Bajo el paradigma Bayesiano de la estadística inferencial se considera al parámetro poblacional θ (parámetro que puede ser considerado como un vector n-dimensional) respecto al cual se desea realizar inferencias como una variable aleatoria, situación que deriva en la asignación de una distribución de probabilidad para dicho parámetro; a esta distribución de probabilidad se le denomina distribución de probabilidad inicial de θ. Si se denota como $\pi(\theta)$ a la densidad asociada con la distribución de probabilidad inicial del parámetro θ, la cual representa el conocimiento que se tiene acerca del parámetro θ antes de obtener cualquier información respecto a los datos, y si ahora se considera que se ha obtenido información del parámetro desconocido a partir de una muestra aleatoria obtenida de una realización de la variable aleatoria Y asociada a un

modelo probabilístico que depende del parámetro θ; al cual se denota como $P(Y|\theta)$ (denominado también como función de verosimilitud), entonces por el Teorema de Bayes se tiene que

$$P(\theta|Y) \propto P(Y|\theta)\pi(\theta), \tag{4.12}$$

donde la distribución de probabilidad de θ representada por la función de densidad en la parte izquierda de (4.12) se le denomina la distribución final de θ, por ser la distribución de θ después de que se ha observado una realización de la variable aleatoria Y.

Descripción del algoritmo EM para la moda de una distribución final

En esencia el algoritmo EM para la moda de una distribución final consiste en encontar el valor de θ que hace que la densidad final para datos observados $P(\theta|Y_{obs})$ tome el valor más alto. La primera modificación en el algoritmo EM para lograr este objetivo parte de redefinir el paso E; considerando para ello la densidad final para datos completos

$$P(\theta|Y) \propto P(Y|\theta)\pi(\theta).$$

Si se calcula el logaritmo de esta densidad final se obtiene

$$\log P(\theta|Y) = c(\theta)^T S(Y) + na(\theta) + \log \pi(\theta) + c;$$

y si ahora se promedia esta ecuación con respecto a los datos perdidos Y_{per} condicionando en los datos observados Y_{obs} para un valor valor preliminar $\theta^{(t)}$. Se deriva el paso E modificado, cuya expresión se representa por

$$\mathcal{Q}^*(\theta|\theta^{(t)}) = \mathcal{Q}(\theta|\theta^{(t)}) + \log \pi(\theta). \tag{4.13}$$

Como consecuencia el paso M modificado consistirá en elegir la siguiente iteración $\theta^{(t+1)}$ como el valor que maximiza $\mathcal{Q}^*(\theta|\theta^{(t)})$. Esto es, se debe encontar el valor $\theta^{(t+1)}$ que satisfaga

$$\mathcal{Q}^*(\theta^{(t+1)}|\theta^{(t)}) \geq \mathcal{Q}^*(\theta|\theta^{(t)}), \qquad \text{para todo } \theta,$$

condición bajo la que se puede mostrar que $\log P(\theta|Y_{obs})$ incrementará su valor en cada iteración, y en problemas bien comportados la sucesión de parámetros estimados convergerá a la moda de $P(\theta|Y_{obs})$ (Schafer, 1997).

4.4. Imputación múltiple

Partiendo del hecho de que en el proceso de imputación múltiple considerado se realiza la imputación de datos faltantes a partir de un modelo de predicción de los valores perdidos, es necesario describir de manera general el procedimiento que hace posible obtener los valores a partir de dicho modelo de predicción. La descripción general del proceso de imputación múltiple parte de

la descripción general del algoritmo de Aumento de Datos, pues como se verá a continuación este proceso se encuentra incorporado como una fase dentro de la implementación del esquema iterativo de dicho algoritmo.

4.4.1. Algoritmo de Aumento de Datos

El esquema iterativo para la implementación del algoritmo de Aumento de Datos considera los siguientes dos pasos:

Paso I

Dado un valor de inicio $\theta^{(t)}$ se selecciona un valor Y_{per} de su distribución predictiva condicional $P(Y_{per}|Y_{obs}, \theta^{(t)})$. Esto es,

$$Y_{per}^{(t+1)} \sim P(Y_{per}|Y_{obs}, \theta^{(t)}), \tag{4.14}$$

donde a (4.14) lo referimos como la imputación o paso I.

Paso P

Una vez realizado el paso I, el paso P consiste en seleccionar un nuevo valor de θ desde su densidad final de datos completos. Esto es,

$$\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{per}^{(t+1)}), \tag{4.15}$$

donde a (4.15) lo referimos como la distribución final o paso P.

Al repetir (4.14) y (4.15) desde un valor de inicio $\theta^{(0)}$ se obtiene una sucesión estocástica $\{(\theta^{(t+1)}, Y_{per}^{(t)}) : t = 1, 2, ...\}$ cuya distribución estacionaria es $P(\theta, Y_{per}|Y_{obs})$, y las subsucesiones $\{\theta^{(t+1)} : t = 1, 2, ...\}$ y $\{Y_{per}^{(t)} : t = 1, 2, ...\}$ tienen a $P(\theta|Y_{obs})$ y $P(Y_{per}|Y_{obs})$ como sus respectivas distribuciones estacionarias. Para un valor suficientemente grante de t, se puede considerar a $\theta^{(t)}$ como una selección aproximada de $P(\theta|Y_{obs})$. Así mismo, $Y_{per}^{(t)}$ puede ser considerada como selección aproximada de $P(Y_{per}|Y_{obs})$ (Schafer, 1997), siendo esta última, el valor objeto de interés para la concreción del proceso iterativo de imputación múltiple.

4.5. Imputación múltiple sobre una base de datos categóricos con datos faltantes

Las secciones precedentes, se abocaron a la tarea de describir de manera general las características de los algolritmos que son necesarios para llevar a un buen término el proceso de imputación múltiple. A partir de ahora y en las subsecuentes secciones el interés se centrará en la descripción de cómo estos procedimientos se particularizan y adecuan a las características que ofrece la información disponible.

4.5.1. El modelo multinomial

Dado que la totalidad de las variables que han sido introducidas en el análisis son de tipo categórico, para poder realizar la imputación múltiple es necesario recurrir a un supuesto distribucional que se adecue a la naturaleza de las variables en consideración. Por tal motivo se supondrá que los datos completos considerados en el análisis han sido generados a partir de una distribución multinomial.

Con la finalidad de describir desde un enfoque general el manejo que se hará de este esquema distribucional, supóngase que se tiene una muestra de n individuos (unidades observadas) a cada uno de los cuales se les ha realizado un conjunto de p mediciones de tipo categórico, $Y_1, Y_2, ..., Y_p$, las cuales se tienen dispuestas en un arreglo rectangular denotado por Y, que representa una matriz de tamaño $n \times p$ la cual no es completamente observada y cuyos valores perdidos fueron generados por un mecanismo de pérdida de tipo DFA.

Si se supone que los niveles de respuesta para cada variable categórica considerada (los cuales para propósitos de la imputación serán asumidos como nominales o categóricos no ordenados) han sido codificados como enteros positivos. Esto es, $Y_j \in \{1, 2, ..., d_j\}$ para $j = 1, 2, ..., p$. Entonces, partiendo del supuesto de que las n unidades muestrales son independientes e idénticamente distribuidas (iid), sin pérdida de información se puede reducir Y a un vector con D componentes, donde $D = \prod_{j=1}^{p} d_j$ es el número de las distintas combinaciones posibles para los niveles de $Y_1, Y_2, ..., Y_p$. Si se indexan las componentes del vector resultante por el subíndice $d = 1, 2, ..., D$, y se define a x_d como el número de unidades muestrales que caen en la componente d, entonces, el vector D-dimensional $x = (x_1, x_2, ..., x_D)$ representa al conjunto total de frecuencias observadas para cada una de las distintas D combinaciones posibles. De nueva cuenta, si se supone que las n unidades muestrales son iid y además se considera al tamaño muestral $n = \sum_{d=1}^{D} x_d$ como fijo, entoces x tiene una distribución multinomial (Schafer, 1997). Por lo tanto, la función de densidad de probabilidad para x está dada por

$$p(x_1, x_2, ..., x_D|\theta) = \frac{n!}{x_1!x_2!...x_D!}\theta_1^{x_1}\theta_2^{x_2}...\theta_D^{x_D}, \tag{4.16}$$

con $x_d = 0, 1, 2, ..., n$ para cada d, $\sum_{d=1}^{D} x_d = n$ y parámetro $\theta = (\theta_1, \theta_2, ..., \theta_D)$. Donde θ_d es la probabilidad de que una unidad muestral caiga en la componente d.

4.5.2. Caracterización de una base de datos categóricos incompletos

Una vez definido el supuesto distribucional sobre el cual versará el análisis de imputación múltiple, es necesario definir la notación que caracteriza a una base de datos categóricos con valores perdidos, y que a su vez permita describir en forma general y de la manera más clara posible, la forma en la cual se adecuan

a ésta los algoritmos EM y de Aumento de Datos necesarios en el proceso de imputación múltiple.

Redefinición de los vectores x y θ como tablas de contingencia

Tomando en cuenta que siempre resulta posible pasar de un arreglo en forma de tabla de contingencia a un vector al asignar un orden lineal a cada una sus celdas, y considerando la conveniencia notacional que representará describir en forma de tabla de contingencia a los vectores x y θ para describir una base de datos con valores perdidos; se redefinirá la forma en la cual se han de considerar a los vectores x y θ. Así entonces, basándose por completo en la notación que para dicho propósito es descrita por Schafer (1997), se define en un principio a x_y, con $y = (y_1, y_2, ..., y_p)$, como el elemento que representa al número total de unidades (conteos) en la muestra para las cuales el evento $Y_1 = y_1, Y_2 = y_2, ...Y_p = y_p$ ocurre. En otras palabras, se considera a este elemento como la frecuencia observada en la celda y de una tabla de contingencia p-dimensional, obtenida a partir de la clasificación cruzada de p variables categóricas. Por otra parte, si además se define a θ_y como la probabilidad de que una unidad muestral pertenezca a la celda y; entonces, los vectores x y θ pueden ser cada uno redefinidos como tablas de contingencia p-dimensionales, las cuales son representadas a partir de las expresiones

$$x = \{x_y : y \in \mathcal{Y}\}, \qquad \theta = \{\theta_y : y \in \mathcal{Y}\},$$

donde $\mathcal{Y}$ representa al conjunto de todos los valores posibles de y. Esto es, $\mathcal{Y}$ representa el producto cartesiano de los conjuntos $\{1, 2, ..., d_j\}$ para $j = 1, 2, ..., p$.

Patrón de pérdida de datos

Un concepto que es necesario definir y entender antes de poder realizar la caracterización de una base de datos categóricos con datos perdidos, es el referente al significado que se le dará al término patrón de pérdida. Se entenderá por un patrón de pérdida al comportamiento de pérdida de información que se observa en cada uno de los renglones de la matriz Y, el cual puede ir desde el asociado con aquellos renglones que no presentan información faltante en ninguno de sus p registros, pasando por renglones cuyos valores faltantes sólo están presentes en alguno o algunos de sus p registros, hasta el caso extremo donde el patrón de pérdida observado es el asociado a renglones que no cuentan con información en alguno de sus p registros. Si ahora se identifican y se clasifican los distintos patrones de pérdida observados en Y, distinguiendo para ello el orden en que se da la pérdida sobre las variables consideradas, entonces será posible encontrar un número máximo de 2^p posibles patrones de pérdida. Sin embargo, es usual encontrar que cuando p es grande no todos los posibles patrones de pérdida se encuentran presentes en una muestra.

Arreglo de una base incompleta de datos categóricos por patrón de pérdida

Con base en la notación antes descrita, se define ahora la caracterización de una base de datos categóricos con información faltante; comenzando para ello con un escenario que muestre a las observaciones en Y reagrupadas de acuerdo a su patrón de pérdida. Si se supone que de los 2^p posibles patrones de pérdida han sido identificados S, entonces como se muestra en la Tabla 4.3 se tiene un arreglo rectangular de S bloques agrupando en cada uno de ellos los renglones de la base de datos que comparten el mismo patrón de pérdida.

	Y_1	Y_2	Y_3	$\cdots$	Y_p
	?	?	?	$\cdots$	
1	?	?	?	$\cdots$	
	.	.	.	$\cdots$	
	?	?	?	$\cdots$	
2	?	?		$\cdots$	?
		?	?	$\cdots$	
3		?	?	$\cdots$	
.	.	.	.	$\cdots$	
.	.	.	.	$\cdots$	
.	.	.	.	$\cdots$	
.	.	.	.	$\cdots$	
	?	?	?	$\cdots$	?
S	?	?	?	$\cdots$	?
	.	.	.	$\cdots$	
	?	?	?	$\cdots$	?

Tabla 4.3: Base de datos multivariable agrupada por patrón de pérdida

Es importante resaltar que de no existir información faltante en cada uno de los S bloques considerados en Y, sería posible realizar una clasificación cruzada de sus elementos para cada uno de ellos, pudiendo con ello construir un conjunto de S arreglos p-dimensionales (tablas de contingencia). Si se designa a cada uno de estos arreglos hipotéticos por $x^{(s)}$, con $x^{(s)} = \{x_y^{(s)} : y \in \mathcal{Y}\}$ para $s = 1, 2, ..., S$; entonces, $x^{(s)}$ resultará ser no observada salvo en el caso que existiera un bloque que no presente información faltante. Fuera de este caso, sólo podrán ser observadas tablas de menor dimensión obtenidas de la clasificación cruzada de las variables observadas en cada uno de los bloques.

4.5.3. Función de log-verosimilitud y algoritmo EM

Prosiguiendo con la descripción general del método de imputación múltiple para datos categóricos, a continuación se expone la forma que adoptan la función de log-verosimilitud así como el algoritmo EM bajo este esquema de información.

Para tal efecto, se comienza definiendo un conjunto de indicadores de respuesta binaria

$$r_{sj} = \begin{cases} 1 & \text{si} \quad Y_j \quad \text{es observada en el patrón } s \\ 0 & \text{si} \quad Y_j \quad \text{no es observada en el patrón } s; \end{cases}$$

así como las funciones de extracción $\mathcal{O}_s(y)$ y $\mathcal{M}_s(y)$, cuyo cometido para cada patrón de perdida s es extraer de $y = (y_1, y_2, ..., y_p)$ los elementos correspondientes a las variables que son observadas y perdidas. Éstas quedan definidas como:

$$\mathcal{O}_s(y) = \{y_j : r_{sj} = 1\}$$

y

$$\mathcal{M}_s(y) = \{y_j : r_{sj} = 0\}.$$

Asimismo, se considera a $\mathcal{M}_s$ y $\mathcal{O}_s$ como los conjuntos de todos los posibles valores de $\mathcal{O}_s(y)$ y $\mathcal{M}_s(y)$.

Si para cada patrón de pérdida s se define a $z^{(s)}$ como la versión colapsada de la tabla no observada, $x^{(s)}$, obtenida de la clasificación cruzada de sus variables observadas, entonces los conteos de la tabla $z^{(s)}$ quedan representadas por

$$z^{(s)}_{\mathcal{O}_s(y)} = \sum_{\mathcal{M}_s(y) \in \mathcal{M}_s} x^{(s)}_y \quad \text{para todo } \mathcal{O}_s(y) \in \mathcal{O}_s,$$

donde la probabilidad marginal de que una observación caiga dentro de la celda $\mathcal{O}_s(y)$ de esta tabla queda definida como

$$\beta_{\mathcal{O}_s(y)} = \sum_{\mathcal{M}_s(y) \in \mathcal{M}_s} \theta_y.$$

Partiendo del resultado para tablas colapsadas de una distribución multinomial (Schafer, 1997, página 243), ahora es posible construir la función de log-verosimilitud para los datos observados, la cual queda formulada por la expresión

$$\ell(\theta | Y_{obs}) = \sum_{s=1}^{S} \sum_{\mathcal{O}_s(y) \in \mathcal{O}_s} z^{(s)}_{\mathcal{O}_s(y)} \log \beta_{\mathcal{O}_s(y)}. \tag{4.17}$$

La dificultad que representa tratar de maximizar (4.17) por métodos de gradiente exige la aplicación del algoritmo EM para la estimación del parámetro θ. La forma particular que éste toma bajo este diseño de información queda caracterizado por los siguientes sencillos pasos E y M:

Paso E

Como el modelo multinomial supuesto es un miembro de la familia exponencial regular (Schafer, 1997, página 243), entonces cada uno de los conteos x_y representa un estadístico suficiente. Si se asume este resultado y se recapitula lo establecido en la ecuación (4.10), queda claro que para definir el paso E es

necesario determinar la esperanza condicional de cada conteo x_y dados los datos observados y un valor asumido para θ. Si en la definición de este valor esperado se considera que $x = \sum_{s=1}^{S} x^{(s)}$, resulta natural observar que cada uno de los conteos puede se representado por $x_y = \sum_{s=1}^{S} x_y^{(s)}$. Entonces la expresión para este valor esperado puede ser determinada como

$$E(x_y|Y_{obs}, \theta) = \sum_{s=1}^{S} E(x_y^{(s)}|Y_{obs}, \theta). \tag{4.18}$$

Si por otra parte se define

$$x^{(s)}_{\mathcal{O}_s(y)} = \{x_y^{(s)} : \mathcal{M}_s(y) \in \mathcal{M}_s\} \tag{4.19}$$

como el conjunto que representa la porción de $x^{(s)}$ que es obtenida fijando $\mathcal{O}_s(y)$ en un valor específico pero variando $\mathcal{M}_s(y)$ sobre $\mathcal{M}_s$; esto es, $x^{(s)}_{\mathcal{O}_s(y)}$ es simplemente el conjunto de todos los conteos en $x^{(s)}$ que mediante su suma contribuyen al conteo observado $z^{(s)}_{\mathcal{O}_s(y)}$, entonces por las reglas de particionamiento (Schafer, 1997, página 243), $x^{(s)}_{\mathcal{O}_s(y)}$ condicionada sobre $z^{(s)}_{\mathcal{O}_s(y)}$ tiene una distribución multinomial con índice $z^{(s)}_{\mathcal{O}_s(y)}$ y parámetros

$$\gamma_{\mathcal{O}_s(y)} = \{\theta_y/\beta_{\mathcal{O}_s(y)} : \mathcal{M}_s(y) \in \mathcal{M}_s\}; \tag{4.20}$$

esto es,

$$x^{(s)}_{\mathcal{O}_s(y)}|z^{(s)}_{\mathcal{O}_s(y)}, \theta \sim M(z^{(s)}_{\mathcal{O}_s(y)}, \gamma_{\mathcal{O}_s(y)}). \tag{4.21}$$

Por lo tanto, si $x_y^{(s)} \in x^{(s)}_{\mathcal{O}_s(y)}$, entonces

$$E(x_y^{(s)}|Y_{obs}, \theta) = E(x_y^{(s)}|z^{(s)}_{\mathcal{O}_s(y)}, \theta). \tag{4.22}$$

Por otra parte, como $E(x_y^{(s)}|z^{(s)}_{\mathcal{O}_s(y)}, \theta) = z^{(s)}_{\mathcal{O}_s(y)}\theta_y/\beta_{\mathcal{O}_s(y)}$, entonces

$$E(x_y^{(s)}|Y_{obs}, \theta) = z^{(s)}_{\mathcal{O}_s(y)}\theta_y/\beta_{\mathcal{O}_s(y)}. \tag{4.23}$$

Por lo tanto el paso E consiste en calcular (4.23) para cada $s = 1, ...S$, y posteriormente sumar los resultados para obtener

$$E(x_y|Y_{obs}, \theta) = \sum_{s=1}^{S} z^{(s)}_{\mathcal{O}_s(y)}\theta_y/\beta_{\mathcal{O}_s(y)}. \tag{4.24}$$

Paso M

A partir de la expresión (4.11), se deduce que

$$n\theta_y = E(x_y|Y_{obs}, \theta),$$

de donde, resolviendo para θ_y, se obtiene

$$\theta_y = E(x_y|Y_{obs}, \theta)/n. \tag{4.25}$$

Así entonces el paso M consiste simplemente en estimar θ_y por el valor de

$$E(x_y|Y_{obs}, \theta)/n \quad \text{para toda } y \in \mathcal{Y}.$$

4.5.4. Algoritmo modificado EM para la moda de una distribución final

Por razones del azar puede suceder que algunas de las tablas observadas $z^{(1)}, ..., z^{(S)}$ contengan celdas con ceros muestrales, propiciando con ello que, derivado de la aplicación del algoritmo EM, se encuentren estimadores θ_y que asumen el valor cero. Esto se traduce en un estimador de máxima verosimilitud de θ que cae en la frontera del espacio paramétrico. Dichos valores no son los más idóneos en la aplicación del algoritmo de Aumento de Datos, pues el estar alejados del centro de la distribución final de los datos observados puede propiciar una convergencia lenta de los valores imputados a su distribución límite. Es por ello que en circunstancias como éstas es necesario recurrir a la aplicación del algoritmo EM para la moda de una distribución final, pues su uso permite encontrar un estimador más cercano al centro de la distribución de los datos observados y con ello garantizar una convergencia más eficiente (Schafer, 1997).

De (4.13) se observa que derivado de la aplicación del paso E del algoritmo EM para la moda de una distribución final se obtiene como resultado una expresión de la log-final para datos completos evaluada en los estadísticos suficientes. Por tal motivo es evidente que el paso E del algoritmo para la moda de una distribución final será el mismo como el de su contraparte de máxima verosimilitud. Por lo tanto, para realizar el paso M bastará con maximizar la densidad final de los datos completos evaluada en los estadísticos suficientes estimados en el paso E.

Con la finalidad de determinar la densidad final para los datos completos que se requiere maximizar, se parte de una elección adecuada de la distribución inicial para el vector de parámetros θ. Esta elección considera al vector $\theta = (\theta_1, \theta_2, ..., \theta_D)$ como un vector aleatorio con distribución Dirichlet con parámetros $\alpha = (\alpha_1, \alpha_2, ..., \alpha_D)$, cuya densidad inicial se expresa sin considerar la constante de normalización como

$$\pi(\theta) \propto \theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1}...\theta_D^{\alpha_D - 1}. \tag{4.26}$$

Por otra parte, si se considera a

$$P(Y|\theta) \propto \theta_1^{x_1}\theta_2^{x_2}...\theta_D^{x_D} \tag{4.27}$$

como el valor de la distribución multinomial obtenido al sustituir en (4.16) los estadísticos suficientes estimados obtenidos del paso E, se tiene que al multiplicar (4.26) por (4.27) se obtiene la densidad final de los datos completos

$$P(\theta|Y) \propto \theta_1^{\alpha_1+x_1-1}\theta_2^{\alpha_2+x_2-1}...\theta_D^{\alpha_D+x_D-1}. \tag{4.28}$$

Ésta es una densidad Dirichlet con parámetros

$$\alpha^{'} = (\alpha_1 + x_1, \alpha_2 + x_2, ..., \alpha_D + x_D).$$

A partir de las propiedades de una distribución Dirichlet se sabe que esta función es maximizada en

$$\theta_d = \frac{\alpha_d + x_d - 1}{n + \alpha_0 - D};$$

donde $\alpha_0 = \sum_{d=1}^{D} \alpha_d$. Así entonces el paso M consiste en estimar el valor de θ_y por la expresión

$$\frac{\alpha_y + x_y - 1}{n + \alpha_0 - D} \quad \text{para toda } y \in \mathcal{Y}. \tag{4.29}$$

4.5.5. Algoritmo de Aumento de Datos

Para concluir, ahora sólo resta describir la forma en la cual bajo este esquema de información se realiza el proceso de imputación múltiple a partir de la aplicación del algoritmo de Aumento de Datos. Partiendo de la definición del paso I, de (4.14) se sabe que es necesario seleccionar una realización de la distribución de los datos perdidos dados los datos observados y el parámetro θ. Para ello se comienza seleccionando como valor de inicio de $\theta^{(0)}$ al estimador obtenido del algoritmo EM; o bien, del algoritmo EM para la moda una distribución final, según sea lo más conveniente para lograr el propósito de eficientar la convergencia. Una vez realizado esto, para cada patrón de pérdida s se selecciona una realización

$$x^{(s)}_{\mathcal{O}_s(y)}|z^{(s)}_{\mathcal{O}_s(y)}, \theta^{(0)} \sim M(z^{(s)}_{\mathcal{O}_s(y)}, \gamma_{\mathcal{O}_s(y)}) \quad \text{para toda } \mathcal{O}_s(y) \in \mathcal{O}_s.$$

Esto es, se elige una realización de la porción de la tabla no observada $x^{(s)}$. Si se hace esto para cada porción de la tabla $x^{(s)}$, se obtiene una versión observada de ésta; concluyendo con ello que la aplicación del paso I consiste en seleccionar, para cada s, $x^{(s)}$ desde su distribución producto multinomial y sumarlas para obtener una tabla de datos simulada completa $x = x^{(1)} + x^{(2)} + ... + x^{(S)}$. Bajo la inicial Dirichlet $\theta \sim D(\alpha)$, el paso P es tan sólo una simulación de θ desde su distribución final de datos completos $D(\alpha + x)$, cuya expresión es representada como en (4.28).

Después del número de iteraciones necesarias para alcanzar la distribución límite, el algoritmo de Aumento de Datos generará un vector x al cual se puede considerar como una realización simulada de la distribución final predictiva de la tabla de contingencia de datos completos $P(x|Y_{obs})$. Sin embargo, como la finalidad del algoritmo de Aumento de Datos es utilizarlo para la imputación múltiple, es necesario al final de la simulación llenar los elementos perdidos Y_{per}

de la matriz de datos Y. Para ello se deberá realizar una iteración final del paso I que rellene los elementos perdidos de Y_{per} a partir de elecciones simuladas desde la distribución $P(Y_{per}|Y_{obs})$.

Ejemplo

Considerando que la notación utilizada en la generalización del método de imputación múltiple para datos categóricos es difícil de seguir, para finalizar, se realizará la exposición de un ejemplo que permita dar mayor lucidez a la forma particular que adoptan los algoritmos EM y de Aumento de Datos para una base de datos categóricos con datos faltantes.

Supóngase por simplicidad que se tiene una muestra de n individuos a cada uno de los cuales se les ha realizado un conjunto de 2 mediciones de tipo categórico, Y_1 y Y_2. Supóngase además que la variable Y_1 toma valores 1 y 2; mientras que, la variable Y_2 toma los valores 1,2 y 3. Si las n unidades en la muestra son independientes e idénticamente distribuidas, el conjunto de datos completos puede ser reducido a un arreglo de la forma $x = (x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23})$, o bien a su equivalente en forma de tabla de contingencia como la mostrada en la Tabla 4.4, donde x tiene una distribución multinomial.

Tabla 4.4: Tabla de contingencia para los datos completos

	$Y_2 = 1$	$Y_2 = 2$	$Y_2 = 3$
$Y_1 = 1$	x_{11}	x_{12}	x_{13}
$Y_1 = 2$	x_{21}	x_{22}	x_{23}

Si se supone que en ambas variables han ocurrido valores perdidos para algunos individuos en la muestra y además que de los $2^2 = 4$ posibles patrones de pérdida de información se han tomado en cuenta $S = 3$ (pues cualquier unidad muestral que tenga a Y_1 y Y_2 no observada puede ser excluida del análisis bajo ignorabilidad), a los cuales se les denotará como A, B y C, donde A incluye a aquellas unidades muestrales que tienen ambas variables observadas, B incluye a aquellas que tienen solamente a la variable Y_1 observada y C incluye a aquellas que tienen solamente a Y_2 observada. Entonces, el conjunto de datos completos queda caracterizado como $x = x^{(A)} + x^{(B)} + x^{(C)}$, donde las tablas de contingencia $x^{(k)}$ para $k = A, B, C$, descritas en la Tabla 4.5, son las asociadas a sus respectivos patrones de pérdida. Al no existir información completa para los patrones de pérdida B y C sus correspondientes tablas de contingencia son no observadas.

Por lo tanto, bajo estas consideraciones, la forma particular que toma el algoritmo EM para este caso se particulariza de la siguiente manera: Dado que $x = x^{(A)} + x^{(B)} + x^{(C)}$ entonces $x_{ij} = x^A_{ij} + x^B_{ij} + x^C_{ij}$, para $i = 1, 2$ y $j = 1, 2, 3$. Como $x^{(B)}$ y $x^{(C)}$ son no observadas, sólo es posible observar sus totales marginales $x^B_{i.} = x^B_{i1} + x^B_{i2} + x^B_{i3}$ para $i = 1, 2$, y $x^C_{.j} = x^C_{1j} + x^C_{2j}$ para $j = 1, 2, 3$.

Tabla 4.5: Tabla de contingencia $x^{(k)}$ para el k-ésimo patrón de perdida de información, $k = A, B, C$.

	$Y_2 = 1$	$Y_2 = 2$	$Y_2 = 3$
$Y_1 = 1$	x^k_{11}	x^k_{12}	x^k_{13}
$Y_1 = 2$	x^k_{21}	x^k_{22}	x^k_{23}

Los datos observados $Y_{obs} = \{x^A_{ij}, x^B_{i.}, x^C_{.j} : i = 1, 2, j = 1, 2, 3\}$ son desplegados en la Tabla 4.6, con una tabla de clasificación cruzada de tamaño 2×3 para las unidades de A de Y_1 y Y_2, y una tabla de tamaño 1×2 que clasifica las unidades en B para la variable Y_1 solamente, y una tabla de tamaño 1×3 que clasifica las unidades en C para Y_2 únicamente.

Tabla 4.6: Clasificación de las unidades muestrales para dos variables incompletamente observadas

(a) Ambas variables observadas

	$Y_2 = 1$	$Y_2 = 2$	$Y_2 = 3$	
$Y_1 = 1$	x^A_{11}	x^A_{12}	x^A_{13}	$x^A_{1.}$
$Y_1 = 2$	x^A_{21}	x^A_{22}	x^A_{23}	$x^A_{2.}$
	$x^A_{.1}$	$x^A_{.2}$	$x^A_{.3}$	

(b) Y_2 perdida

$Y_1 = 1$	x^B_1
$Y_1 = 2$	x^B_2

(c) Y_1 perdida

$Y_2 = 1$	$Y_2 = 2$	$Y_2 = 3$
$x^C_{.1}$	$x^C_{.2}$	$x^C_{.3}$

Si se considera a $\theta = (\theta_{11}, \theta_{12}, \theta_{13}, \theta_{21}, \theta_{22}, \theta_{23})$, con θ_{ij} para $i = 1, 2$ y $j = 1, 2, 3$, como la probabilidad de que una unidad muestral pertenezca al i-ésimo renglón y la j-ésima columna de la Tabla 4.4; entonces, por las reglas de particionamiento (Schafer, 1997, página 243) aplicadas dentro de las partes B y C de la muestra, la distribución predictiva de los datos perdidos dado θ y los datos observados se convierte en un conjunto de distribuciones multinomiales independientes,

$$(x^B_{i1}, x^B_{i2}, x^B_{i3})|Y_{obs}, \theta \sim M(x^B_{i.}, (\theta_{i1}/\theta_{i.}, \theta_{i2}/\theta_{i.}, \theta_{i3}/\theta_{i.})) \quad i = 1, 2$$

y

$$(x^C_{1j}, x^C_{2j})|Y_{obs}, \theta \sim M(x^C_{.j}, (\theta_{1j}/\theta_{.j}, \theta_{2j}/\theta_{.j})), \quad j = 1, 2, 3,$$

con $\theta_{i.} = \theta_{i1} + \theta_{i2} + \theta_{i3}$ para $i = 1, 2$, y $\theta_{.j} = \theta_{1j} + \theta_{2j}$ para $j = 1, 2, 3$.

El paso E de EM reemplaza los conteos desconocidos x^B_{ij} y x^C_{ij} en x_{ij} por sus esperanzas condicionales bajo un valor asumido para θ,

$$\begin{aligned} E(x_{ij}|Y_{obs}, \theta) &= E(x^A_{ij} + x^B_{ij} + x^C_{ij}|Y_{obs}, \theta) \\ &= x^A_{ij} + x^B_{i.}\theta_{ij}/\theta_{i.} + x^C_{.j}\theta_{ij}/\theta_{.j}. \end{aligned}$$

El paso M estima θ_{ij} por $E(x_{ij}|Y_{obs}, \theta)/n$

Si fuera necesario recurrir a la aplicación del algoritmo EM para la moda de una distribución final, como ya fue mencionado, bastará únicamente con elegir un valor adecuado para $\alpha = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{22}, \alpha_{23})$ de la distribución inicial Dirichlet asociada con θ y sustituir el paso M anterior por la estimación de θ_{ij} a partir de la expresión

$$\frac{\alpha_{ij} + E(x_{ij}|Y_{obs}, \theta) - 1}{n + \alpha_0 - 6},$$

con $\alpha_0 = \sum_{i=1}^{2} \sum_{j=1}^{3} \alpha_{ij}$.

Una vez seleccionado el valor de inicio $\theta^{(0)}$ desde el algoritmo EM; o bien, a partir del algoritmo EM para la moda una distribución final, la aplicación del algoritmo de Aumento de Datos para $t = 0, 1, 2, ...$ queda descrito de la siguiente forma:

Paso I

Para cada uno de los patrones de pérdida B y C se elige una realizacion de las tablas de contingencia $x^{(B)}$ y $x^{(C)}$ como,

$$(x_{i1}^B, x_{i2}^B, x_{i3}^B)|Y_{obs}, \theta^{(t)} \sim M(x_{i.}^B, (\theta_{i1}^{(t)}/\theta_{i.}^{(t)}, \theta_{i2}^{(t)}/\theta_{i.}^{(t)}, \theta_{i3}^{(t)}/\theta_{i.}^{(t)})) \quad i = 1, 2$$

y

$$(x_{1j}^C, x_{2j}^C)|Y_{obs}, \theta^{(t)} \sim M(x_{.j}^C, (\theta_{1j}^{(t)}/\theta_{.j}^{(t)}, \theta_{2j}^{(t)}/\theta_{.j}^{(t)})), \quad j = 1, 2, 3.$$

Obteniéndose de esta manera la tabla de datos simulados completa a partir de $x = x^{(A)} + x^{(B)} + x^{(C)}$

Paso P

A partir de una elección adecuada de α, el paso P considera la elección de $\theta^{(t)} = (\theta_{11}^{(t)}, \theta_{12}^{(t)}, \theta_{13}^{(t)}, \theta_{21}^{(t)}, \theta_{22}^{(t)}, \theta_{23}^{(t)})$ $t > 0$, como una realización de la distribución final Dirichlet de datos completos

$$\theta^{(t)} \sim D(\alpha + x).$$

Por último, al considerar que el objetivo de esta imputación es llenar los elementos perdidos de las variables Y_1 y Y_2, entonces a partir de una iteración t lo suficientemente grande del algoritmo de Aumento de Datos, se seleccionan los valores perdidos de Y_1 condicionados a su correspondiente valor observado de Y_2, así como los valores perdidos de Y_2 condicionados a su correspondiente valor observado de Y_1, desde las distribuciones condicionales

$y_1 \sim P(Y_1|Y_2 = j)$ para $j = 1, 2, 3$

y

$y_2 \sim P(Y_2|Y_1 = i)$ para $i = 1, 2$.

Donde, $P(Y_1|Y_2 = j) = \theta_{ij}^{(t)}/\theta_{.j}^{(t)}$ para $i = 1, 2$ y $P(Y_2|Y_1 = i) = \theta_{ij}^{(t)}/\theta_{i.}^{(t)}$ para $j = 1, 2, 3$.

Capítulo 5

Análisis Estadístico

Antes de realizar el análisis estadístico, se parte de la selección de una muestra aleatoria obtenida a partir de un muestreo aleatorio simple que selecciona aproximadamente el 80 % de individuos sobre cada una de las cohortes escolares consideradas. La decisión de trabajar con una muestra aleatoria y no con la totalidad de los datos, obedece a que en su momento se tendrá el interés de evaluar el ajuste del modelo resultante a partir de valorar su capacidad predictiva con respecto a los individuos no contemplados en la muestra. Por lo que, de un total de 417 individuos, se contará para los análisis subsecuentes con un total de 331.

Tabla 5.1: Cantidad y porcentaje de observaciones perdidas por variable considerando como base para su cálculo una muestra de 331 individuos

Variable	Numero de observaciones perdidas	Porcentaje
TIPOREG	2	0.60
SISTEMA	4	1.21
REZAGO	11	3.32
SEXO	0	0.00
EDOCIVI	16	4.83
DEPECO	0	0.00
TRABAJA	26	7.85
PROBACH	14	4.23
EDAD	6	1.81
TIPCASA	23	6.95
SERVCUL	15	4.53
DESCCOL	9	2.72
ESCOPAD	65	19.64
ESCOMAD	70	21.15
DESERTO	0	0.00

La cantidad y porcentaje de datos perdidos por variable para esta muestra son presentados en la tabla 5.1.

Para el desarrollo del análisis estadístico se consideran dos etapas. La primera (correspondiente al análisis de datos sin imputar), consiste en la aplicación del análisis de regresión logística al conjunto de variables explicativas cuyos porcentajes de no respuesta no fueron tan elevados (menores a 10 %), excluyendo por ello de este análisis preliminar a las variables predictoras ESCOPAD y ESCOMAD. La segunda etapa (correspondiente al análisis de datos imputados), contempla la inclusión de dichas variables junto con aquellas variables que resultaron ser estadísticamente significativas en la primera etapa del análisis, con la finalidad de conformar una base de datos con un número menor de variables cuyos valores faltantes serán imputados y posteriormente analizados mediante la aplicación del análisis de regresión logística. La razón de no incluir en la segunda etapa del análisis a todas las variables explicativas obedeció a la relación de dependencia que el algoritmo de Aumento de Datos guarda con la cantidad de información perdida, ya que en un escenario multivariado como el considerado donde la mayoría de las variables presentan algún nivel de información perdida, la convergencia del algoritmo de Aumento de Datos a su distribución límite pudiera verse fuertemente comprometida.

Antes de dar comienzo con el análisis es importante mencionar que para el desarrollo del análisis estadístico serán utilizados los paquetes LogXact-5 y Stata-8. Por otra parte, para los procedimientos de imputación múltilple se hará uso de funciones en el lenguaje estadístico Splus desarrolladas por J.L. Schafer del Departamento de Estadística de la Universidad Estatal de Pensilvania EUA [1].

5.1. Análisis inferencial con datos sin imputar

Para dar paso al análisis de los datos con las variables predictoras consideradas, en una primera instancia se realiza el ajuste univariado de cada una de éstas con respecto a la variable respuesta. Se utilizará para ello el análisis de regresión logística asintótico y exacto. Los resultados del ajuste univariado para cada uno de estos análisis son mostrados en las Tablas 5.2 y 5.3 respectivamente. En estas tablas, para cada una de las variables listadas en la primera columna, se presenta la información correspondiente a cada una de las medidas relacionadas con el ajuste de éstas y la variable respuesta en el orden siguiente: El coeficiente β_0 para el modelo de regresión logística univariado que contiene sólo esta variable (coef); el error estándar estimado del coeficiente estimado de la pendiente (Err.Est); el cociente estimado de momios (CM) y su intervalo del 95 % de confianza (95 %IC); los estadísticos de prueba así como su valor p asociado. Recuérdese que G es el estadístico de prueba del cociente de verosimilitudes para el modelo logístico ordinario, y Q es el estadístico del Score Condicional para el modelo logístico exacto.

[1] Dichas funciones pueden ser obtenidas mediante descarga gratuita del paquete CAT el cual puede ser descargado desde el sitio web http://www.stat.psu.edu/ jls/misof twa.html#splus.

Tabla 5.2: Análsis de regresión logística asintótico univariado

Variable	Coef	Err.Est.	CM	95 % IC	G	p
TIPOREG	−0.476	0.6345	0.62	(0.18, 2.16)	0.62	0.430
SISTEMA	0.557	0.7846	1.75	(0.38, 8.12)	0.45	0.504
REZAGO	−0.196	0.5229	0.82	(0.30, 2.29)	0.14	0.704
SEXO	−0.072	0.4086	0.93	(0.42, 2.07)	0.03	0.860
EDOCIVI	*	*	*	*	*	*
DEPECO	−0.795	0.4518	0.45	(0.19, 1.10)	2.82	0.093
TRABAJA	0.168	0.5300	1.18	(0.42, 3.34)	0.10	0.749
PROBACH	1.899	0.5201	6.68	(2.41, 18.50)	16.63	< 0.001
EDAD	−0.215	0.7623	0.81	(0.18, 3.60)	0.08	0.772
TIPCASA	0.624	0.4653	1.87	(0.75, 4.64)	1.80	0.179
SERVCUL	0.511	0.4585	1.67	(0.68, 4.10)	1.18	0.277
DESCCOL2	−0.456	1.0792	0.63	(0.08, 5.26)		
DESCCOL3	1.099	1.3333	3.00	(0.22, 40.93)	2.71	0.258

*:No es posible calcularlo utilizando este análsis

Se sigue la estrategia sugerida por Hosmer y Lemeshow (2000) para la selección de variables potencialmente significativas para su inclusión en un análisis multivariado. Se distingue para tal fin a todas aquellas variables cuyas pruebas univariadas en alguno de los dos modelos considerados mostraron tener un valor $p <$0.25.

Al realizar una inspección visual de los resultados obtenidos en dichas tablas con la finalidad de buscar variables potencialmente significativas que puedan ser incluidas en un análisis multivariado, se nota que la magnitud de los coeficientes de regresión estimados y su desviación estándar es similar en ambos análisis. Sin embargo, los niveles de significancia exacto y asintótico presentan algunas diferencias. Por ejemplo, en la variable SERVCUL, el nivel de significancia asintótico es 0.277 y el exacto es 0.316. El primer valor es cercano a 0.25 mientras que el segundo está más alejado. La otra variable que produce resultados distintos en ambos métodos es EDOCIVI. En el primer caso el modelo no puede ser ajustado mientras que en el segundo la variable no es significativa al 25 %.

De los resultados de las Tablas 5.2 y 5.3 se concluye entonces que hay evidencia de que, bajo el criterio de selección antes mencionado, las variables: DEPECO, PROBACH, TIPCASA, DESCCOL y SERVCUL, podrían convertirse en predictores potenciales de la variable respuesta bajo un análisis multivariado. Por tal motivo serán candidatas para ser incluidas en este análisis. Si bien es cierto que en el caso de la variable SERVCUL, su nivel de significancia en el correspondiente análisis logístico asintótico es ligeramente superior a 0.25, éste será incluido en el análisis al considerar que su valor-p (0.277) no está tan alejado del valor crítico previamente fijado.

Considerando que para ambos análisis los coeficientes y sus errores estándar correspondientes no son tan distintos, no existe razón para pensar que el desba-

Tabla 5.3: Análisis de regresión logística exacto univariado

Variable	Coef	Err.Est.	CM	95 % IC	Q	p
TIPOREG	−0.474	0.6337	0.62	(0.12, 2.19)	0.57	0.589
SISTEMA	0.555	0.7828	1.74	(0.18, 8.29)	0.51	0.624
REZAGO	−0.195	0.5221	0.82	(0.23, 2.40)	0.14	0.808
SEXO	−0.072	0.4080	0.93	(0.38, 2.26)	0.03	1.000
EDOCIVI	−0.860	*	0.42	$(-\infty, 2.58)$	1.68	0.380
DEPECO	−0.792	0.4508	0.45	(0.18, 1.27)	3.22	0.104
TRABAJA	0.168	0.5292	1.18	(0.38, 4.07)	0.10	0.801
PROBACH	1.892	0.5194	6.64	(2.29, 23.54)	16.84	< 0.001
EDAD	−0.214	0.7613	0.81	(0.09, 3.55)	0.08	1.000
TIPCASA	0.621	0.4645	1.86	(0.68, 5.26)	1.84	0.239
SERVCUL	0.510	0.4577	1.66	(0.59, 4.39)	1.26	0.316
DESCCOL2	−0.454	1.0767	0.63	(0.08, 24.15)		
DESCCOL3	1.037	1.2955	2.82	(0.12, 196.28)	4.10	0.130

*:No es posible calcularlo utilizando este análisis

lanceo de los datos haya tenido algún efecto en la estimación de los coeficientes y errores estándar. Por tal motivo, en el ajuste del modelo final (salvo que durante el proceso de ajuste fuera necesario recurrir al análisis exacto), sólo se realizará el ajuste a partir del análisis asintótico. Los resultados de dicho ajuste están dados en la tabla 5.4, donde ζ representa el correspondiente valor observado del estadístico univariado de Wald para cada una de las variables involucradas.

Tabla 5.4: Resultados del ajuste de un modelo multivariado de regresión logística al conjunto de variables que fueron estadisticamente significativas a un nivel de 0.25 en las Tablas 5.2 y 5.3

Variable	Coef	Err.Est.	ζ	p
DEPECO	−0.742	0.5801	−1.28	0.201
PROBACH	1.687	0.5457	3.09	0.002
TIPCASA	0.564	0.4888	1.15	0.249
DESCCOL2	−0.548	1.1388	−0.48	0.631
DESCCOL3	0.673	1.4493	0.46	0.642
SERVCUL	0.800	0.5277	1.52	0.130
Constante	−3.011	1.3105	−2.30	0.022

De los resultados de la Tabla 5.4, se puede notar que en base a la prueba univariada de Wald, solamente las variables PROBACH y SERVCUL podrían ser consideradas en conjunto como posibles predictores de la variable respuesta; esto pensando en un nivel de significancia menos restrictivo de 0.13. Prosi-

guiendo con la selección del modelo se comparan ahora el modelo conteniendo todas las variables involucradas en la Tabla 5.4 contra el modelo que solamente contiene las variables PROBACH y SERVCUL. El resultado de esta prueba de

Tabla 5.5: Prueba de hipótesis

G	g.l	$P(\chi^2(g.l) > G)$
4.206	4	0.379

H_0: Modelo con SERVCUL+PROBACH
H_1: Modelo conteniendo todas las variables de la Tabla 5.4

hipótesis se muestra en la Tabla 5.5, donde a partir del valor p resultante (0.379), no se rechaza la hipótesis nula a un nivel de significancia de 0.05. Por lo tanto se asume al modelo conformado por las variables PROBACH y SERVCUL, como el que hasta el momento resulta el más simple que da una adecuada descripción de los datos.

Tabla 5.6: Prueba de hipótesis

G	g.l	$P(\chi^2(g.l) > G)$
1.868	1	0.172

H_0: Modelo con PROBACH
H_1: Modelo con SERVCUL+PROBACH

A partir del análisis univariado de las Tablas 5.2 y 5.3, se nota que la variable SERVCUL individualmente no resultó ser estadísticamente significativa, caso contrario a PROBACH que mostró serlo. Por tal motivo, para continuar con la selección de variables bastará con comparar el modelo que contiene las variables PROBACH y SERVCUL, contra el modelo que contiene únicamente la variable PROBACH. La prueba de hipótesis derivada de dicha comparación se muestra en la Tabla 5.6. Se puede observar que la hipótesis nula no es rechazada a un nivel de 0.05.

Con la finalidad de cerciorarse si cada una de las variables que fueron descartadas en base al análisis de la Tablas 5.2 y 5.3 pudieran convertirse en un importante predictor de la respuesta cuando cada una de éstas, por separado, es ajustada junto a la variable PROBACH, se realizan las comparaciones de cada uno de estos modelos contra el modelo que contiene a PROBACH como única variable explicativa. Cabe mencionar que dada la imposibilidad de realizar el contraste para el modelo conteniendo la variable EDOCIVI, fue necesario recurrir a un contraste de hipótesis exacto. Los resultados derivados de dichas comparaciones son mostrados en la Tabla 5.7. Al observar los resultados de dicha tabla se puede concluir que los distintos contrates presentados no rechazan la hipótesis nula a un nivel de 0.05. Así entonces, se tiene indicios para suponer que la variable PROBACH es la única que muestra evidencia de tener alguna

Tabla 5.7: Prueba de hipótesis

Variable	Q	p
TIPOREG	0.65	0.576
SISTEMA	2.35	0.168
REZAGO	0.12	0.777
SEXO	0.17	0.823
EDOCIVI	2.07	0.222
TRABAJA	0.39	0.605
EDAD	< 0.001	1.000

H_0: Modelo con PROBACH
H_1: Cada uno de los siete modelos que se pueden ajustar con PROBACH+cada variable incluida en la tabla

asociación con la variable respuesta. Por lo tanto, si se realiza el ajuste de ésta se encuentra que el logit estimado, $\hat{g}(w)$, está dado por la ecuación

$$\hat{g}(w) = -3.643 + 1.899 \times PROBACH, \tag{5.1}$$

donde los valores ajustados del modelo de regresión logística, $\hat{\pi}(w)$, son dados por la ecuación

$$\hat{\pi}(w) = \frac{e^{-3.643+1.899\times PROBACH}}{1 + e^{-3.643+1.899\times PROBACH}}. \tag{5.2}$$

A partir de la ecuación anterior resulta posible estimar la probabilidad de que un estudiante deserte con base a su promedio de bachillerato. Esto al evaluar los valores de las variables de diseño correspondientes a la variable PROBACH, utilizadas en el ajuste correspondiente a (5.2). Los valores obtenidos para dichas probabilidades son

$$\hat{\pi}(0) = 0.026$$

y

$$\hat{\pi}(1) = 0.149,$$

resultados que corresponden a la probabilidad de que un estudiante deserte dado que tuvo un promedio de bachillerato de 6 a 8 y de 8.1 a 10 respectivamente. Este resultado es explicado por el hecho de que de los 23 estudiantes que desertaron 18 de ellos tuvieron un promedio mayor o igual a 8.1.

5.1.1. Evaluación del ajuste del modelo

Dado que en el ajuste del modelo sólo una de las variables mostró ser estadísticamente significativa, el modelo ajustado (5.2) resulta ser saturado. Por

tal motivo sólo se puede evaluar su ajuste a partir de la medición de su capacidad para clasificar correctamente a cada estudiante como desertor o no desertor.

Sensibilidad y Especificidad

Desde la perspectiva de la capacidad clasificatoria del modelo, es necesario considerar al modelo ajustado (5.2) como una prueba diagnóstica dicotómica, pues éste permitirá clasificar a cada estudiante como desertor y no desertor; clasificación que se considerará como positiva (1) y negativa (0) respectivamente. La evaluación de la exactitud con la que se realiza dicha clasificación se obtiene a partir de las medidas de sensibilidad y especificidad. La medida de sensibilidad será en este caso la probabilidad de clasificar correctamente a un estudiante como desertor cuando efectivamente lo es; la medida de especificidad es la probabilidad de clasificar correctamente a un estudiante como no desertor cuando efectivamente no lo es.

Tabla 5.8: Tabla de clasificación basada en (5.2) para los casos no considerados en el ajuste del modelo, usando un punto de corte de 0.5

Clasificado	Observado		Total
	DESERTO=1	DESERTO=0	
DESERTO=1	0	0	0
DESERTO=0	2	80	82
Total	2	80	82

Sensibilidad=0/2=0 %, Especificidad =80/80=100 %

Tabla 5.9: Tabla de clasificación basada en (5.2) para los casos considerados en el ajuste del modelo, usando un punto de corte de 0.5

Clasificado	Observado		Total
	DESERTO=1	DESERTO=0	
DESERTO=1	0	0	0
DESERTO=0	23	294	317
Total	23	294	317

Sensibilidad=0/23=0 %, Especificidad =294/294=100 %

Para poder realizar dicha clasificación es necesario contar con un criterio que permita discriminar a partir de la prueba diagnóstica la pertenencia de un estudiante a alguno de los dos grupos. Para ello, se establece un nivel de decisión o valor de corte c, a partir del cual, un individuo cuya probabilidad estimada

por el modelo sea menor a c será clasificado como perteneciente al grupo de respuesta 0, en caso contrario como perteneciente al grupo de respuesta 1. Cuando los resultados obtenidos a partir de esta clasificación y las observaciones de la variable respuesta, Z ,son cruzadas en una tabla de contingencia, será posible estimar la sensibilidad y la especificidad de la prueba correspondientes al valor de corte c.

Partiendo del valor de corte más comúnmente utilizado en el análisis discriminante de 0.5, para los casos que no fueron considerados durante el ajuste del modelo como para los casos considerados en su ajuste, se realizan de manera separada dos tablas de contingencia (Tabla 5.8 y Tabla 5.9) que consideran los cruces entre las clasificaciones obtenidas por (5.2) en ambos conjuntos de datos y sus correspondientes valores observados de la variable respuesta. Cabe mencionar que en la conformación de estas tablas se eliminaron los datos faltantes (que para el caso de los datos no utilizados en el ajuste del modelo fueron 4, mientras que para el caso de los datos utilizados en el ajuste del modelo fueron 14). Al revisar la capacidad de discriminación del modelo ajustado a partir del valor de corte considerado para ambas tablas, se puede notar que la sensibilidad resulta ser de 0 %, mientras que la especificidad es de 100 %. Por lo tanto, bajo el valor de corte considerado se nota que la capacidad discriminatoria del modelo hacia la respuesta de interés es nula.

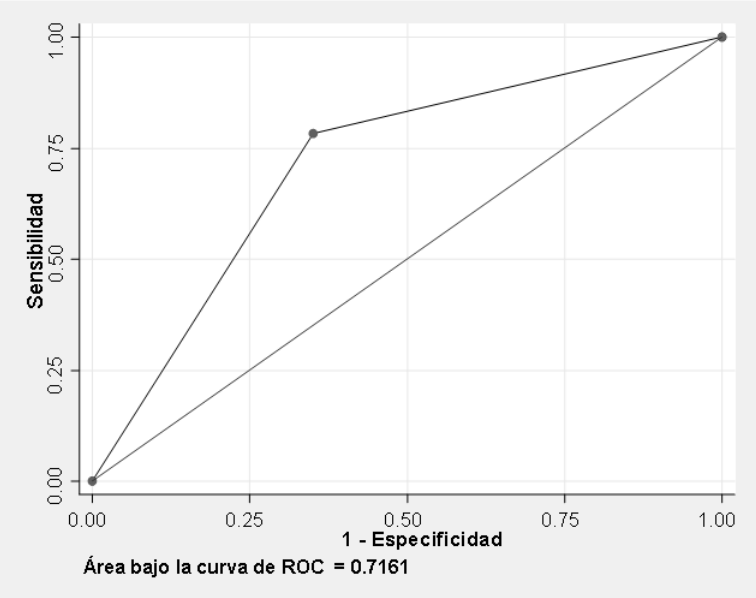
Análisis de sensibilidad a partir de la Curva de ROC

Debido a que la capacidad discriminatoria del modelo depende en gran medida del valor de corte seleccionado, el resultado obtenido con el valor de corte de 0.5 no es concluyente. Es por ello que, partiendo del mismo principio de clasificación, se realizará a continuación la aplicación del análisis conocido como Curva ROC, el cual premitirá medir la precisión discriminatoria del modelo no sólo desde la perspectiva de un único valor de corte y de un par de valores de sensibilidad y especificidad, sino más bien de manera global a partir de un conjunto de parejas de valores de sensibilidad y especificidad correspondientes a cada uno de los distintos niveles de decisión (valores de corte).

A partir de los datos usados en el ajuste de (5.2), se utilizará la curva ROC para medir la precisión clasificatoria de éste hacia la respuesta de interés. Desde esta perspectiva, para todos los posibles valores de corte, se tomarán en cuenta sus correspondientes estimaciones de sensibilidad y 1-especificidad, a fin de conformar los pares ordenados (1-especificidad, sensibilidad) que permitirán obtener la representación gráfica de la curva ROC. La gráfica de la Figura 5.1 representa la curva ROC para los datos considerados.

El área bajo la curva ROC proporciona una representación global de la exactitud diagnóstica, ya que su valor es susceptible a la relación que guarden las estimaciones de la sensibilidad contra las estimaciones de 1-especificidad para los distintos valores de corte. Si en todos los valores de corte dichas tasas fueran iguales, la curva ROC sería la diagonal que une los vértices inferior izquierdo y superior derecho, obteniéndose con ello una área bajo la curva de 0.5. Tal circunstancia indicaría que la capacidad del modelo para detectar acertadamente

Figura 5.1



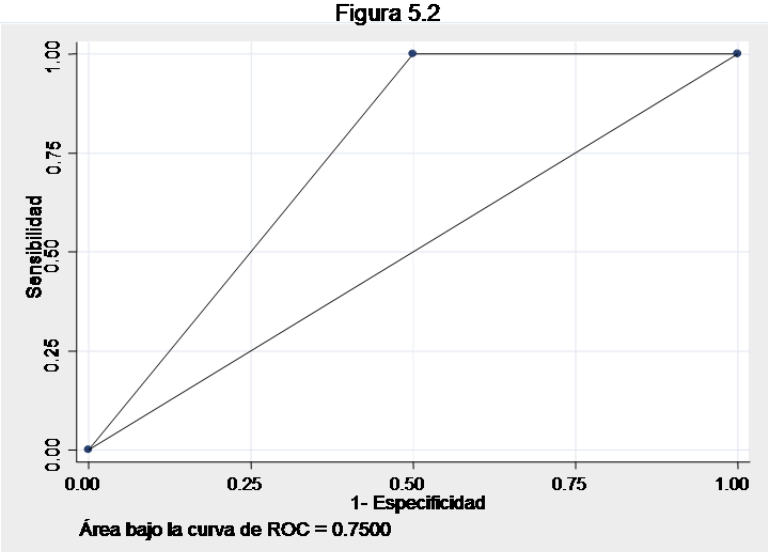
en favor de la respuesta de interés, resulta igual a la de no hacerlo. Es decir, no existe una tendencia favorable de la prueba para discriminar acertadamente en favor de la respuesta de interés. La exactitud de la prueba y por ende el área bajo la curva aumenta a medida de que la prueba tiene mayor capacidad de detectar acertadamente en favor de la respuesta de interés, desplazándose desde la diagonal hacia el vértice superior izquierdo. A medida que la discriminación tiende a ser perfecta (100 % de sensibilidad y 100 % de especificidad) el área bajo la curva tenderá a ser cercana a 1(ver Hosmer y Lemeshow (2000)).

De la Figura 5.1 se tiene que para los datos considerados, el área bajo la curva ROC es de 0.7161. Como una regla general (ver, Hosmer y Lemeshow 2000, página 162) si se obtienen valores $0.7 \leq ROC \leq 0.8$ se considera como aceptable discriminación, por lo que se concluye que bajo el modelo 5.1 existe una tendencia favorable a clasificar acertadamente a un estudiante como desertor cuando efectivamente lo es.

Con la finalidad de complementar lo hecho en la Tabla 5.8, se aplicará ahora el análisis de curva de ROC para medir la precisión discriminatoria de (5.2) para los casos que no fueron considerados en su ajuste. El resultado de dicho análisis es exhibido en la Figura 5.2 la cual muestra el área bajo la curva de ROC de 0.7500; cantidad que basada en la regla general antes mencionada permite considerar como una discriminación aceptable. Este resultado lleva a realizar la misma conclusión para (5.1) que en el caso anterior.

Como ya se mencionó, la base de datos que se ha utilizado está desbalanceada. Este inconveniente podría implicar que la elección de la liga logit por ser simétrica alrededor de 1/2, pudiera no otorgar con (5.2) el mejor ajuste a los

Figura 5.2



datos, por lo que para terminos comparativos, es conveniente utilizar la transformación log-log complementario. Transformación que por sus características involucra asimetría en la curva de respuesta para $\pi(w)$ (Agresti, 2000, página 248). El ajuste estimado para la liga log-log complementario queda determinado por la ecuación

$$\hat{g}(w) = -3.656 + 1.830 \times PROBACH. \tag{5.3}$$

Con la finalidad de comparar si en términos de ajuste la función liga log-log complementario estimada en (5.3) es más adecuada que la liga logit estimada en (5.2), se recurre al Criterio de Información de Akaike (AIC). Propuesto por Akaike (1973), este criterio permitirá realizar la comparación entre ambos modelos al seleccionar a aquel que maximice el AIC. Al calcular el AIC para ambos modelos se obtuvo el mismo valor (152.3319), por lo tanto, se concluye que no existe evidencia que sugiera que el modelo log-log complementario sea mejor que el modelo logístico.

5.2. Análisis inferencial con datos imputados

Para dar inicio con esta etapa del análisis, se parte con una descripción detallada de los procedimientos que permitirán llevar a cabo el proceso de imputación múltiple sobre los datos faltantes de las variables consideradas, para posteriormente concluir con el ajuste del modelo predictivo así como con la comparación de los resultados obtenidos de éste con aquellos derivados del modelo (5.2). Cabe mencionar que a partir de las consideraciones hechas al principio de

este capítulo, para el desarrollo de esta segunda etapa del análisis, se considerarán solamente a las variables PROBACH, ESCOPAD y ESCOMAD.

5.2.1. Determinación del parámetro de inicio

Antes de dar paso a la aplicación del algoritmo de Aumento de Datos y con ello realizar la imputación múltiple sobre los datos faltantes de las variables PROBACH, ESCOPAD y ESCOMAD, es necesario determinar un valor de inicio para su ejecución. Siguiendo la sugerencia dada por Schafer (1997) para la elección de un valor de inicio, se buscará aquella que garantice un valor cercano al centro de la distribución final de los datos observados, $P(\theta|Y_{obs})$, esto con el fin de garantizar que la convergencia sea más rápida hacia la distribución estacionaria. Por lo tanto, en una primera instancia, el candidato idóneo es el estimador de máxima verosimilitud $\hat{\theta}$. Una vez determinado éste a partir del algoritmo EM, en una inspección visual de sus $D = 144$ componentes se detectó que más de la mitad de éstos tomaron el valor de cero. Tales valores no son adecuados en este caso, pues seleccionar este estimador como valor de inicio representa elegir un valor en la frontera del espacio paramétrico, alejándolo con ello del centro de la distribución final de los datos observados. En relación a este último resultado es importante notar que la tabla de contingencia x que se genera de la aplicación del paso E es dispersa; esto es, un número considerable de celdas x_d no contienen observaciones.

La presencia de este estimador en la frontera del espacio paramétrico obliga a determinar otro valor de inicio mediante la aplicación del algoritmo EM para la moda de una distribución final. Como se puede observar de la relacion (4.28), los parámetros de la distribución final están conformados por los valores x_d de x obtenidos de la aplicación del paso previo E. Dado que, como ya se mencionó, la aplicación de este paso genera una tabla de contingencia x dispersa, la elección de una distribución inicial Dirichlet para θ deberá considerar valores α_d que garanticen la condición $\alpha_d + x_d > 1$, pues de esta condición depende la certeza de que la moda para la densidad posterior será única y caerá en el interior del espacio paramétrico (Schafer, 1997). Así entonces, será necesario elegir $\alpha_d > 1$ para obtener un estimador con las propiedades deseadas; esto es, que todos sus elementos sean distintos de cero.

Como la densidad expresada en (4.28) es equivalente a la función de verosimilitud de una tabla de contingencia multinomial

$$x' = (\alpha_1 + x_1 - 1, \alpha_2 + x_2 - 1, ..., \alpha_D + x_D - 1),$$

se puede notar que la elección de una distribución inicial con $\alpha_d > 1$, para $d = 1, 2, ..., D$, durante la aplicación del algoritmo EM para la moda de una distribución final, agregará el equivalente de $\alpha_d - 1$ observaciones iniciales a cada celda. Este valor, $\alpha_d - 1$, conocido como *flattening constant*, tiene el efecto de alisar el estimador de θ hacia una tabla uniforme en la cual todas las probabilidades de las celdas son iguales. Una distribución inicial que alisa parámetros

estimados hacia una tabla uniforme es llamada una distribución inicial *flattening*.

Atendiendo a la sugerencia dada por Schafer (1997), en el sentido de que en situaciones como la analizada donde se tiene una fuerte ausencia de conocimiento inicial acerca de θ no es recomendable agregar información inicial en montos mayores al 10 o 20 % del actual tamaño muestral, en la aplicación del algoritmo EM para la moda de una distribución final se asignará como distribución inicial una Dirichlet con $\alpha = (c, c, ..., c)$, para $c = 1 + \sqrt{n}/D$ ($n = 418$ y $D = 144$). Agregando con ello el equivalente aproximado de 20.4 observaciones iniciales que representan un total de aproximadadamente el 4.9 % del actual tamaño muestral. Cabe destacar que el valor $\sqrt{n}/D$, corresponde a una de las posibles elecciones empíricas de una *flattening constant*, las cuales son revisadas por Fienberg and Holland (1972).

5.2.2. Generando las imputaciones

Como ya se mencionó antes, el objetivo de la imputación múltiple es crear m ensayos aproximadamente independientes de Y_{per}, los cuales permitan obtener m bases de datos completos simulados. Para lograr este objetivo, en el presente caso, las m imputaciones se obtendrán a partir de un proceso de simulación de m cadenas paralelas de longitud k generadas a partir de la aplicación del algoritmo de Aumento de Datos, partiendo cada una de éstas desde un mismo valor de inicio; que para esta aplicación será la moda de la distribución final obtenida a partir del algoritmo EM bajo las condiciones antes descritas, y de un valor α de la distribución inicial Dirichlet que será el mismo que el utilizado en la determinación de dicha moda de la distribución final. Los valores imputados para Y_{per} serán los valores de la k-ésima iteración final de cada una de las m cadenas.

Determinación del número k de iteraciones

Una vez establecido el mecanismo bajo el cual se realizarán las imputaciones, en un principio es necesario determinar el número k de iteraciones del algoritmo de Aumento de Datos que garantice que los valores de Y_{per} obtenidos durante el proceso de imputación serán efectivamente realizaciones aproximadamente independientes de la distribución predictiva final $P(Y_{per}|Y_{obs})$. Dado que una condición suficiente para que la sucesión estocástica $\{(\theta^{(t)}, Y_{per}^{(t)}) : t = 0, 1, 2, ..., \}$ generada por una aplicación sucesiva del algoritmo de Aumento de Datos haya convergido a la distribución estacionaria $P(\theta, Y_{per}|Y_{obs})$ es que la distribución de $\theta^{(t)}$ haya convergido a $P(\theta|Y_{obs})$, distintas medidas basadas en los componentes de θ han sido propuestas para evaluar la convergencia de la distribución de las iteraciones de $\theta^{(t)}$ a su distribución estacionaria. Muchas de éstas se enfocan al diagnóstico de la estacionariedad a partir del monitoreo de la convergencia de cada una de las componentes individuales del vector de parámetros θ, así como de todas las funciones de éste que pudieran parecer relevantes. Sin embargo, tomando en consideración que se enfrenta un problema de carácter multivariado

con un vector paramétrico θ de dimensión 144, el intento de monitorear la convergencia de los componenetes individuales así como tratar de especificar todas las funciones de θ que sean relevantes, para el caso de este análisis resulta inoperante. Es entonces necesario la aplicación de una medida que permita diagnosticar la estacionariedad con respecto la distribución conjunta de θ. Dicha medida para este análisis será una función escalar de la forma $\xi(\theta) = v^T\theta$ para algún vector constante v. Sugerida por Schafer (1997), esta medida conocida como la peor función lineal de los parámetros, definida así en el sentido de que su distribución marginal converge más lentamente, fortaleciendo así con su uso la evidencia de convergencia global, es definida como

$$\xi(\theta) = \hat{v}_1^T(\theta - \hat{\theta}), \tag{5.4}$$

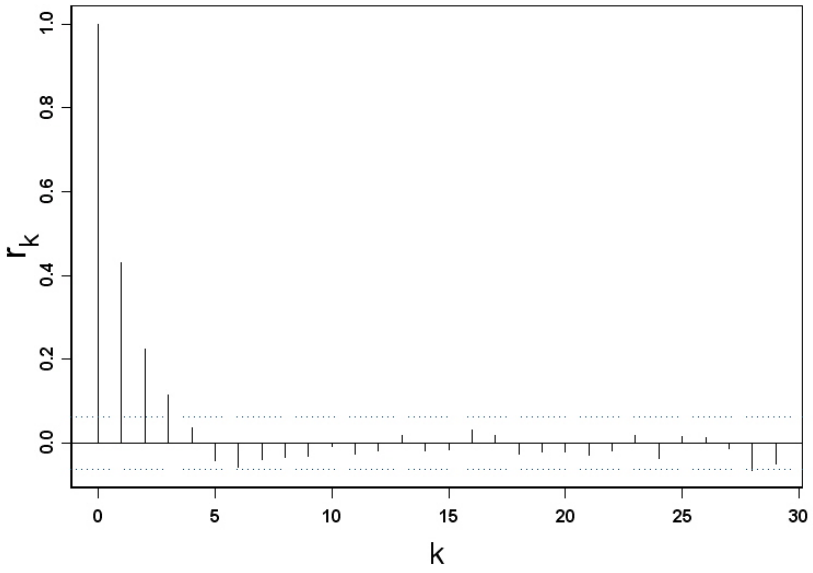
donde $\hat{\theta}$ es la moda de la distribución final obtenida a partir del algoritmo EM. La determinación de un estimador numérico $\hat{v}_1$ de v_1 para esta función, parte de la idea mostrada por Schafer (1997) en el sentido de que la tasa de convergencia de EM está gobernada por la fracción de información de Fisher perdida debido a la no respuesta. Para el caso multivariado puede ser representada por medio de una matriz $M = {I_c}^{-1}(\hat{\theta})I_m(\hat{\theta})$ de tamaño $D \times D$ (en nuestro caso D=144), donde $I_m(\hat{\theta})$ representa la información perdida debida a la no respuesta; mientras que $I_c(\hat{\theta})$ representa la información completa. Por lo tanto, al considerar a $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D$ como los eigenvalores ordenados de M, y a $v_1, v_2, ..., v_D$ como los eingenvectores de M correspondientes a estos eigenvalores ordenados, Schafer (1997) muestra que la tasa de convergencia de EM está gobernada por la fracción más grande de información perdida, que para este caso queda representada por λ_1, cantidad que representa la información perdida correspondiente a la dirección particular del vector v_1. Así entonces determina que para un número suficientemente grande de iteraciones de EM; supóngase t, se puede esperar que en una vecindad cercana a la moda se cumpla la condición

$$\varepsilon^{(t)} \approx cv_1, \tag{5.5}$$

donde $\varepsilon^{(t)} = \theta^{(t)} - \hat{\theta}$ representa el vector error y c una constante de proporcionalidad. Se concluye de esta forma que un estimador numérico $\hat{v}_1$ de v_1 puede ser simplemente obtenido tomando la diferencia entre el valor convergente $\hat{\theta}$ y cualquiera de las iteraciones finales de EM. Por lo tanto, partiendo de esta idea, para este caso se construye la peor función lineal mediante la determinación del estimador numérico para $\hat{v}_1$ a partir de la diferencia entre el estimador de θ un paso antes de la convergencia y el valor convergente $\hat{\theta}$.

Considerando que para asegurar la convergencia aproximada a la distribución estacionaria basta con encontrar un valor de k que garantice que $\theta^{(t+k)}$ sea aproximadamente independiente de $\theta^{(t)}$ para cualquier t, se corre a partir de la moda $\hat{\theta}$ junto con su respectiva α, una sola cadena de 1000 iteraciones para (5.4), analizando su estructura de correlación a partir de la función de autocorrelación (FAC) muestral definida como

Figura 5.2



$$r_k = \frac{\sum_{t=1}^{n-k}(\varepsilon^{(t)} - \overline{\varepsilon})(\varepsilon^{(t+k)} - \overline{\varepsilon})}{\sum_{t=1}^{n}(\varepsilon^{(t)} - \overline{\varepsilon})^2}, \quad k = 0, 1, 2, ... \tag{5.6}$$

Esta función mide la correlación entre los valores de una serie temporal distanciados en un lapso de tiempo (retraso) k, y $\overline{\varepsilon}$ es la media muestral de la serie observada (ver Guerrero, 2003). La aplicación de dicha medida permitirá determinar un valor k que grantice que la FAC muestral de (5.4) sea efectivamente cero, y con ello poder juzgar qué tantas iteraciones son necesarias para alcanzar la estacionalidad global aproximada.

Tabla 5.10: Función de autocorrelación muestral con retraso k para 1000 realizaciones de la peor función lineal de los parámetros

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
r_k	1.000	0.431	0.224	0.113	0.037	−0.043	−0.058	−0.039	−0.033	−0.031	−0.008	−0.028	−0.018	0.018	−0.019
$\mathcal{M}_k$		0.062	0.073	0.075	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076
k	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
r_k	−0.017	0.032	0.017	−0.026	−0.020	−0.020	−0.029	−0.019	0.017	−0.037	0.014	0.013	−0.014	−0.065	−0.050
$\mathcal{M}_k$	0.076	0.076	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077

A partir del uso de un contraste de hipótesis (Schafer, 1997), es posible

mostrar que la dependencia lineal serial para un valor máximo de $k = 30$ de las 1000 realizaciones de (5.4) decae más alla de un retraso $k = 4$, y con ello aproximar la convergencia del algoritmo de Aumento de Datos a su distribución estacionaria. Dicho contraste para un nivel $\alpha = 0.05$ queda planteado con la hipótesis nula de ausencia de correlación para un retraso k o mayor, $\rho_k = \rho_{k+1} = \rho_{k+2} = ... = 0$, contra la hipótesis alternativa $\rho_k \neq 0$, por lo que al apoyarse en el hecho de que la distribución de r_k es aproximadamente normal cuando $\rho_k = 0$, se rechaza la hipótesis nula si

$$\mid r_k \mid \geq \mathcal{M}_k, \tag{5.7}$$

con

$$\mathcal{M}_k = 1.96\left[\frac{1}{1000}\left(1 + 2\sum_{t=1}^{k-1} {r_t}^2\right)\right]^{\frac{1}{2}};$$

donde la expresión dentro de corchete representa la varianza estimada para r_k.

De los valores críticos obtenidos a partir de (5.7), mostrados como líneas punteadas en la Figura 5.2, así como de los valores de r_k y $\mathcal{M}_k$ para los distintos retrasos k en la Tabla 5.10, se concluye con un nivel $\alpha = 0.05$ de confianza que las correlaciones no difieren significativamente de cero más allá de $k = 4$. Por lo tanto, basándose en esta evidencia se puede concluir que para un número mayor a $k = 4$ de iteraciones del algoritmo de Aumento de Datos alcanzará efectivamente la estacionariedad.

Determinación del número m de imputaciones

Es posible medir la eficiencia de la estimación por imputación múltiple para cada uno de los parámetros según el número de imputaciones m y su correspondiente fracción de información perdida λ. Esto a partir de la medida propuesta por Rubin (1987) a través de la expresión

$$\left(1 + \frac{\lambda}{m}\right)^{-1}; \tag{5.8}$$

la cual mide la eficiencia relativa aproximada (en escala de varianza) entre un estimador puntual basado en un número finito de m imputaciones y uno basado en un número infinito de éstas. Se utiliza este resultado para la elección del número de imputaciones a realizar, siendo la elección de m aquella que garantiza que con un pequeño número se obtiene una ganancia en eficiencia no mucho menor a la que se obtendría de considerar un número infinito de imputaciones.

Con el objetivo de determinar la peor pérdida de eficiencia que se podría esperar para una determinada elección de m, se tomará como referencia para el cálculo de distintos valores de m el valor de λ_1. La determinación de este valor no es en forma directa, siendo necesario, al igual que como se hizo para el caso del vector v_1, recurrir al algoritmo EM para su estimación. Por lo tanto es necesario considerar las tasas de convergencia correspondientes a los elementos

individuales de $\theta = (\theta_1, \theta_2, ..., \theta_D)$. Estas tasas pueden ser estimadas de las iteraciones de EM por

$$\hat{\lambda}_j^{(t)} = \frac{\theta_j^{(t+1)} - \theta_j^{(t)}}{\theta_j^{(t)} - \theta_j^{(t-1)}}, \tag{5.9}$$

para $j = 1, 2, ..., D$ en valores adecuadamente grandes de t. Schafer (1997) hace notar que (5.9) hace posible estimar el eigenvalor más grande λ_1 de M.

Con la finalidad de estimar la peor fracción de información perdida, ante la presencia (como es el presente caso) de un número grande de parámetros, se atiende la sugerencia que para tal efecto da Schafer (1997). En primera instancia se calcula el valor de (5.9) para cada uno de los $D = 144$ parámetros sobre las primeras $t = 16$ iteraciones de EM, para posteriormente eliminar cualesquier valor fuera del intervalo $(0, 1)$, tomando entonces como valor estimado de la peor fracción de información perdida $\hat{\lambda}_1$, a la cantidad que resulte ser la mediana del conjunto de valores obtenidos a partir de las medidas correspondientes a cada uno de los $\hat{\lambda}_j^{(t)}$, $t = 1, 2, ..., 16$. Dicho valor para este caso resulta ser $\hat{\lambda}_1 = 0.22$.

Así entonces, si sobre este valor se avalua (5.8) para distintos valores de m, se obtendrán distintas cantidades estimadas para la pérdida de eficiencia. Como se puede observar de la Tabla 5.11, cuanto mayor sea el valor de m, la pérdida de eficiencia resulta ser menor. Sin embargo, debido a que en un análisis de imputación múltiple se involucran consideraciones de tipo práctico, el considerar un número excesivo de bases resulta engorroso y poco funcional. Por tal motivo, en el presente caso, se usará un valor de $m = 10$ bases de datos, por lo que al hacer esta elección se soportará una pérdida de eficiencia correspondientes al 97.85 % con respecto a una imputación que considere a m tendiendo a infinito.

Tabla 5.11:

$\hat{\lambda}$ \ m	3	5	10	20
0.22	0.9317	0.9579	0.9785	0.9891

Prueba de hipótesis con datos imputados

Una vez imputadas las 10 bases de datos obtenidas a partir de la generación de $m = 10$ cadenas paralelas de longitud $k = 20$; éste último valor fue considerado para proveer un margen extra de seguridad en la convergencia. Se retorna al objetivo inicial de este trabajo, que es el indentificar los factores de riesgo asociados con la variable respuesta. Para ello, se procede a realizar un análisis de regresión logística que junto a la variable PROBACH incorpore a las variables ESCOPAD y ESCOMAD que por sus niveles de no respuesta fueron eliminadas en el análisis preliminar, así como incluir la información contenida en las 10

bases de datos imputadas, de tal forma que sea posible ajustar un modelo de regresión logística que en términos de parsimonia resulte ser el mejor. Por lo antes dicho, es importante entonces partir de la idea de integrar los estadísticos de prueba utilizados en análisis de regresión logística usados en la determinación y selección del mejor modelo ajustado, a los estadísticos de prueba que permitan incorporar la información contenida en las 10 bases de datos imputadas.

Partiendo de esta idea, se comienza por definir un estadístico equivalente al de Wald, el cual combina los m estadísticos de Wald ,$W^{(t)}$, $t = 1, 2, ..., m$, correspondientes al contraste de hipótesis para un modelo de regresión logística con k parámetros, realizado sobre datos completos para cada una de m bases de datos imputadas. Tal estadístico, al cual se denota por Δ, así como sus medidas asociadas son expuestos por Schafer (1997). En este caso se definie junto con sus medidas asociadas, de la siguiente manera:

$$\Delta = \frac{\overline{\delta}_w k^{-1} - (m+1)(m-1)^{-1} r}{1+r},$$

donde

$$\overline{\delta}_w = \frac{1}{m} \sum_{t=1}^{m} W^{(t)},$$

es el promedio de los estadísticos de Wald, y

$$r = (1 + m^{-1}) \left[\frac{1}{m-1} \sum_{t=1}^{m} \left(\sqrt{W^{(t)}} - \overline{\sqrt{W}} \right)^2 \right],$$

es $(1 + m^{-1})$ veces la varianza muestral de sus raíces cuadradas. El valor p combinado bajo la hipótesis nula es

$$p = P(F_{k,v} \geq \Delta),$$

cuyos grados de libertad para v quedan determinados a partir de

$$v = k^{-3/m}(m-1)(1 + r^{-1})^2.$$

La cantidad

$$\hat{\lambda} = \frac{r + 2/(v+3)}{r+1},$$

representa la fracción de información perdida estimada correspondiente al vector de parámetros asociados con las variables omitidas bajo la hipótesis nula.

En la Tabla 5.12 se muestra el contraste entre la hipótesis nula representada por modelo que contiene sólo la variable PROBACH contra las tres hipótesis alternativas correspondientes a los tres modelos obtenidos de las posibles combinaciones de la variable PROBACH con las variables ESCOPAD y ESCOMAD. Como se puede notar, a un nivel de significancia del 5 % no se rechaza la hipótesis nula en ninguno de los tres casos.

Tabla 5.12: Prueba de hipótesis para datos imputados H_0 : Modelo con PROBACH vs H_1 : Modelo con PROBACH+ Variables Incluidas

Variables Incluidas	Δ	k	v	p	$100r$	$100\hat{\lambda}$
ESCOPAD+ESCOMAD	1.2920	4	108	0.28	30.57	24.79
ESCOMAD	1.2497	2	603	0.29	12.37	11.30
ESCOPAD	1.2813	2	169	0.28	26.24	21.71

Cabe hacer mención que en el ajuste del modelo que incorpora las variables ESCOPAD y ESCOMAD se redujo el número de categorías de repuesta para cada una de estas dos variables. La finalidad de esto fue generar niveles de respuesta para ambas variables que agruparan individuos que compartieran características similares relacionadas con el nivel educativo de sus progenitores. Los nuevos tres niveles de respuesta obtenidos a partir de colapsar las categorías para cada una de las dos variables en consideración fueron generados de la siguiente manera: (1 y 2), (3 , 4 y 6) y 5. La forma de esta agrupación fue pensada para considerar tres niveles educativos: nulo o básico, medio y superior. La no obtención de un nivel de significancia estadística para las variables ESCOPAD y ESCOMAD podría deberse a la forma en la que se agruparon los tres nuevos niveles de respuesta, pues es posible pensar que el entorno familiar en el que se desarrolla un estudiante cuyos progenitores no tienen algún nivel educativo puede ser totalmente distinto de aquellos estudiantes donde el padre o madre si tuvo aunque sea un nivel básico de escolaridad. Por lo tanto otra posible agrupación de respuestas pudiera haber considerado en un sólo grupo a los individuos cuyos progenitores no tienen ningún tipo de instrucción educativa, separándolos de aquellos agrupados en la categoría que considera a los progenitores con instrucción educativa de básica a superior. Sin embargo, por la característica de la información disponible la posibilidad de incorporar las variables ESCOPAD y ESCOMAD como binarias bajo esta nueva forma de agrupación no es posible pues en ambos casos el número de estudiantes cuyos progenitores no tienen ningún nivel educativo es mucho menor en comparación de aquellos estudiantes cuyos progenitores tienen al menos un nivel básico de estudios lo que puede ocasionar problemas en el ajuste del modelo.

Ajuste del modelo final

En el ajuste del modelo final se hará uso de la medida propuesta por Schafer (1997), la cual permitirá estimar los parámetros del modelo final a partir de la combinación de los parámetros estimados correspondientes a cada una de las diez bases de datos que se han imputado. Dicha medida basada en la estimación de un modelo de regresión logística con $p + 1$ parámetros calculados sobre m bases de datos imputados es definida a partir de la expresión

$$\overline{\beta_k} = \frac{1}{m}\sum_{t=1}^{m} \hat{\beta_k}^{(t)},$$

para $k = 0, 1, 2, ..., p$, cuya varianza está representada por

$$Var(\overline{\beta_k}) = \overline{\mathcal{U}} + (1 + m^{-1})B,$$

donde

$$\overline{\mathcal{U}} = \frac{1}{m}\sum_{t=1}^{m} Var(\hat{\beta_k}^{(t)}),$$

representa la varianza dentro de la imputación y ,

$$B = \frac{1}{m-1}\sum_{t=1}^{m} (\hat{\beta_k}^{(t)} - \overline{\beta_k})^2$$

representa la varianza entre imputaciones.

El cálculo del valor p asociado a la prueba de la hipótesis nula $\beta_k = 0$ contra una alternativa bilateral es determinado por

$$p = P(F_{1,v} \geq \mathcal{T}),$$

donde $\mathcal{T} = (Var(\overline{\beta_k}))^{-1}\overline{\beta_k}^2$, y F representa las distribución de Fisher con grados de libertad

$$v = (m-1)\left[1 + \frac{\overline{\mathcal{U}}}{(1 + m^{-1})B}\right]^2.$$

Por último, se define una medida que pemitirá evaluar cómo los datos perdidos contribuyen a la incertidumbre inferencial de los β_k. Dicha medida, que al igual que en el caso multivariado cuantifica el incremento relativo en la varianza debido a la no respuesta, queda definida por la expresión

$$r = \frac{(1 + m^{-1})B}{\overline{\mathcal{U}}}.$$

En la Tabla 5.13 se muestran los resultados para el ajuste del modelo final conteniendo únicamente el intercepto y la variable explicativa PROBACH; pues ésta fue la única que resultó ser estadísticamente significativa al 5 % en el previo análisis multivariado. Para cada coeficiente, la tabla despliega sus respectivos estimadores puntuales y sus errores estándar, así como el valor p para el contraste de hipótesis que mide la significancia estadística de los coeficientes estimados. Integrando además el incremento relativo en la varianza debido a la no respuesta r y el estimador de la fracción de información perdida $\hat{\lambda}$.

Una vez finalizado el análisis por imputación múltiple, se puede concluir a partir del valor p de la tabla 5.13 correspondiente a la variable PROBACH, que ésta resulta ser un predictor estadísticamente distinto de cero a un nivel del 5 %.

Tabla 5.13: Inferencias por imputación múltiple para los coeficientes del modelo de regresión logística

Variable	β	$\sqrt{T}$	v	p	$100r$	$100\hat{\lambda}$
Intercepto	-3.4832	0.4340	99	0.000	9.96	4.65
PROBACH	1.7940	0.5106	79	0.000	12.94	4.97

Por lo tanto, considerando los coeficientes estimados en la tabla 5.13 tanto para el intercepto como para la variable PROBACH, es posible calcular el predictor lineal estimado $\hat{g}(w)$, cuya expresión para este caso queda determinada a partir de la ecuación

$$\hat{g}(w) = -3.483 + 1.794 \times PROBACH, \tag{5.10}$$

donde el correspondiente modelo de regresión logística ajustado, $\hat{\pi}(x)$, está dado por la ecuación

$$\hat{\pi}(w) = \frac{e^{-3.483+1.794\times PROBACH}}{1 + e^{-3.483+1.794\times PROBACH}}. \tag{5.11}$$

Como previamente fue realizado en el análisis de datos sin imputar, es posible ahora a partir de esta expresión calcular la probabilidad de que un estudiante deserte en base a su promedio de bachillerato.

$$\hat{\pi}(0) = 0.030$$

y

$$\hat{\pi}(1) = 0.156.$$

Este resultado corresponde a la probabilidad de que un estudiante deserte dado que tuvo un promedio de bachillerato de 6 a 8 y de 8.1 a 10 respectivamente. Por lo tanto las probabilidades de deserción fueron de nueva cuenta estimadas más altas para el grupo de estudiantes cuyo promedio de bachillerato fue de 8.1 a 10 mientras que la probabilidad menor fue asignada a los estudiantes con promedio de bachillerato de 6 a 8. Siendo este resultado acorde con el resultado obtenido para el modelo (5.2).

Dada la similitud entre los coeficientes estimados del modelo ajustado en (5.2) con aquellos estimados en el modelo (5.10), se concluye que existe razón para pensar que el modelo ajustado para datos sin imputar pueda ser considerado adecuado.

Capítulo 6

Conclusiones

La utilización del método de imputación múltiple como complemento a la aplicación del análisis de regresión logística mostró ser una metodología adecuada como apoyo en la búsqueda de factores de riesgo asociados con la deserción escolar, pues permitió la incorporación en el análisis (y con ello poder determinar su nivel de significancia estadística) de las variables relacionadas con la escolaridad del padre y de la madre del estudiante; variables consideradas de importancia en la explicación de la deserción escolar y que por sus altos niveles de no respuesta hubieran tenido que ser descartadas del análisis; o bien, ser incorporadas a éste, no sin recurrir a la alternativa indeseable de eliminar todos los registros asociados con información perdida en alguna de estas variables.

Más allá del atractivo práctico que puede representar la aplicación del algoritmo de Aumento de Datos por sí mismo o en combinación con el análisis de regresión logística, es de resaltar las limitantes que surgen o pueden surgir al momento de aplicar el algoritmo de Aumento de Datos o una combinación de éste con el análisis de regresión logística; limitantes que siempre deben tenerse presentes pudieran llegar a dificultar o incluso hacer inviable la implementación de ambas metodologías en bases de datos como la considerada en el presente trabajo.

Es un hecho que la aplicación del método de imputación múltiple junto con el análisis de regresión logística exacto está completamente descartado; esto debido a que los estimadores de los parámetros del modelo de regresión logística para dicho análisis no son de máxima verosimilitud, supuesto que el método de imputación múltiple considera necesario para hacer válidas las estimaciones e inferencias sobre dichos parámetros, y que lo hacen incompatible con las inferencias sobre los parámetros estimados derivados del análisis de regresión logística exacto. Si bien es cierto que por fortuna durante el desarrollo del análisis de datos imputados no fue necesario recurrir al ajuste de parámetros mediante un análisis de regresión logística exacto, pues el desbalance de los datos no produjo inestabilidad en los coeficientes y errores estándar estimados por máxima verosimilitud para cada uno de los conjuntos de datos imputados, de haber sido necesario usar estimadores exactos hubiera propiciado que se dejara inconcluso

el análisis. Así entonces, si se desea aplicar un análisis de regresión logística con variables explicativas de tipo categórico en combinación con el método de imputación múltiple, se debe tomar en cuenta que si en la variable respuesta objeto de estudio la proporción de respuestas (z = 1) es cercana a 0, se puede correr el riesgo de no poder aplicar el análisis de regresión asintótico y con ello dejar inconcluso el análisis. En este sentido, se abre paso a la investigación en la búsqueda de una metodología de estimación que haga compatible las inferencias del análisis de regresión logística exacto con el método de imputación múltiple.

Por otra parte, la dependencia en la cantidad de información perdida que el algoritmo de Aumento de Datos tiene para alcanzar su distribución estacionaria, y contar con una base de datos en donde la mayor parte de las variables consideradas presentaron en mayor o menor medida pérdida de información, condujo a aplicar el algoritmo de Aumento de Datos junto con el análisis de regresión logística a sólo un número reducido de estas variables, dejando con ello de lado la posibilidad de replicar el análisis que sobre detección de factores de riesgo se hiciera sobre la base de datos sin imputar, a un conjunto de bases de datos que incluyeran todos los registros perdidos imputados para todas las variables explicativas consideradas en el presente análisis. Con base en esto, se puede concluir que el no poder imputar de forma confiable bases de datos completos con una pérdida de información importante en cada una de sus variables, representa una limitante del algoritmo de Aumento de Datos, ya que con ello se dejan fuera variables que con información imputada podrían convertirse en predictores potenciales de la respuesta, comprometiendo de esta manera la confiabilidad de los resultados obtenidos de la aplicación del análisis de regresión logística.

También es de destacar que, si se pretende aplicar un análisis de imputación múltiple como el realizado donde el algoritmo de Aumento de Datos basa sus imputaciones en variables de tipo categórico, esto puede llegar a representar una desventaja ante la presencia de variables explicativas con escala de medición continua, ya que para considerarlas en el proceso de imputación múltiple, será necesario categorizarlas lo cual implica pérdida de información.

Al momento de aplicar un análisis de imputación múltiple como el desarrollado en el presente trabajo, siempre se debe tener presente la importancia que tienen los supuestos que le dan sustento a esta metodología, pues de su cabal cumplimiento depende la fiabilidad de los resultados obtenidos. En este sentido y en base a la experiencia derivada del presente análisis, se puede concluir que, dada la imposibilidad que se tiene de contar con una prueba de hipótesis que permita validar al mecanismo DFA como el que generó la pérdida de información, resulta de importancia la participación del analista de la información en el diseño y aplicación del cuestionario. Una adecuada supervición por parte del analista de la información del diseño de las preguntas del cuestionario, así como del llenado del mismo por parte del personal encargado de aplicarlo, puede prevenir la presencia de información faltante (pues contar con toda la información completa siempre es preferible a aplicar cualquier método de imputación) y en caso de que ésta se presente, el analista puede tener más indicios de cuál es el mecanismo que dio origen a la pérdida de información y con ello aplicar

el método más adecuado para modelarla. Con esto se evita la introducción de supuestos muchas veces difíciles de comprobar y de cuyo cabal cumplimiento depende la validez de los resultados obtenidos.

Bibliografía

[1] Agresti, A. (2000) Categorical Data Analysis . J.Wiley & Sons, New York.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory. eds. Petrov B. and Cski F., Akadmiai Kiad, Budapest, 267-281.

[3] Allison, P.D. (2002) Missing Data. Sage, Thousand Oaks, CA.

[4] Collett, D. (2002) Modelling Binary Data. Chapman & Hall. London.

[5] Considine, G. and Zappla, G. (2002). The influence of social and economic disadvantage in the academic performance of school students in Australia. Joournal of Sociology, 38, 129-148.

[6] Cox, D.R. and Hinkley, D.V. (1970) The Analysis of Binary Data. Chapman & Hall. London.

[7] Dobson, A.J. (1990) An Introduction to Generalized Linear Models. Chapman & Hall. London.

[8] Fienberg, S. E. and Holland, P. W. (1972). On the Choice of Flattening Constants for Estimating Multinomial Probabilities. Journal of Multivariate Analysis, 2, 127- 134.

[9] Guerrero, V. (2003) Análisis Estadístico de Series de Tiempo Económicas. Segunda Edición. Thomson

[10] Hosmer, D.W and Lemeshow, S. (2000) Applied Logistic Regression. J.Wiley & Sons, New York.

[11] Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. J.Wiley & Sons, New York.

[12] Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data. Chapman & Hall. London.