UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

REGRESIÓN LOGÍSTICA CON EFECTOS ALEATORIOS PARA DATOS AGRUPADOS

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

MATEMÁTICO

P R E S E N T A :

FRANCISCO REYES SANCHEZ

DIRECTOR DE TESIS:
DRA. GUILLERMINA ESLAVA GÓMEZ

CIUDAD DE MÉXICO, 2019

1. Datos del alumno
Reyes
Sánchez
Francisco
55 46 60 42 65
Universidad Nacional Autónoma de México
Facultad de Ciencias
Matemáticas
414007237

2. Datos del tutor
Dra.
Guillermina
Eslava
Gómez

3. Datos del sinodal 1
Mat.
Margarita Elvira
Chávez
Cano

4. Datos del sinodal 2
Dra.
Natalia Bárbara
Mantilla
Beniers

5. Datos del sinodal 3
Dra. Lizbeth
Naranjo
Albarrán

6. Datos del sinodal 4
M. en C.
María Fernanda Gil
Leyva
Villa

7. Datos del trabajo escrito
Regresión logística con efectos aleatorios para datos agrupados
94 p.
2019

Resumen

El modelo de Regresión logística con efectos aleatorios se utiliza para modelar conjuntos de observaciones con respuesta multinomial y donde las observaciones están agrupadas.

Para este trabajo de tesis, se cuenta con un conjunto de datos provenientes de la *Encuesta Nacional de Alimentación y Nutrición en el Medio Rural* llevada a cabo en áreas rurales de México en 1996. En este caso, para cada observación se cuenta con una variable respuesta binaria y un conjunto de variables explicativas, y se desea estimar los efectos de éstas en la variable respuesta. Sin embargo, el modelo logístico en su forma simple generalmente no es adecuado para estos datos debido a que las observaciones se encuentran agrupadas, y las respuestas de observaciones pertenecientes a un mismo grupo podrían no ser estocásticamente independientes. Por esta razón, en el presente trabajo, utilizamos al modelo logístico simple y lo comparamos con dos enfoques para el análisis de datos agrupados y correlacionados: el modelo de regresión logística con efectos aleatorios y el modelo de regresión logística ponderada (*Survey-weighted generalised linear model*).

En el análisis estadístico realizado, se considera como respuesta a una variable indicadora de desnutrición para niños de edad preescolar (peso para la edad: 1 = desnutrición, 0 = normal), y un conjunto de ocho variables explicativas correspondientes a aspectos socioeconómicos de la familia y algunas características de la madre: el gasto semanal per cápita en alimentos, el número de personas por habitación, edad de la madre, presencia de piso de algún material, disponibilidad de baño, disponibilidad de estufa de gas, escolaridad de la madre y un indicador de dialecto para la madre del infante. Los datos están agrupados en un total de 853 pueblos de zonas rurales y comprenden a mediciones de 17,865 niños menores de cinco años (un niño por madre).

Resultados. La capacidad predictiva de los ajustes de regresión logística en su forma simple, con intersección aleatoria y ponderado por encuesta resulta casi nula: todas las observaciones son clasificadas como normales. Sin embargo, la descripción y significancia de efectos de las variables predictoras en la respuesta coincide en dichos ajustes: las variables predictoras consideradas tienen un efecto significativo en la probabilidad de padecer desnutrición. Por otro lado, a pesar de que las estimaciones puntuales de los coeficientes de los tres ajustes resultan similares, la amplitud de los intervalos de confianza resulta superior (y más confiable) cuando se considera la estructura de correlación debida a la estructura de la distribución de la población.

Palabras clave: Regresión logística; Regresión logística con efectos aleatorios; Datos de encuesta complejas; Modelos lineales generalizados mixtos.

Agradecimientos

A mis padres y mi hermana, a quienes amo tanto, por su apoyo incondicional.

A la Dra. Guillermina por su accesibilidad, tiempo y sinceridad.

A cada uno de mis sinodales por el tiempo empleado en la revisión de este trabajo y por sus valiosos comentarios.

A mis amigos Gerardo, Gustavo, Mariana, Rossana y Saúl, a quienes aprecio mucho y cuya compañía hacía posible el pasar tardes enteras realizando tareas en la Facultad de Ciencias. A Irán por su presencia, cariño y apoyo.

A mis profesores de la Facultad de Ciencias, ya que el presente trabajo representa la culminación del proceso de aprendizaje resultante de su dedicación.

Índice general

Introducción

El presente trabajo se centra en la presentación y ajuste de modelos de regresión logística para datos agrupados.

La característica principal de esta clase de datos es el incumplimiento de la hipótesis de independencia entre observaciones de la variable respuesta. Esto ocurre, por ejemplo, cuando múltiples observaciones se encuentran expuestas bajo efectos similares que las diferencian de otras. Generando así, una distinción de observaciones por grupos ó *clusters*, en los cuales las observaciones de la variable respuesta pertenecientes a un mismo grupo suelen estar correlacionadas.

Para analizar las consecuencias que conlleva ignorar la correlación existente entre observaciones, se considera como referencia al modelo logístico en su forma simple (i.e. al modelo logístico bajo el supuesto de independencia), y se presentan dos modelos logísticos para datos agrupados: el modelo con intersección aleatoria y ponderado por diseño (*Survey-weighted logistic regression*). La aplicación y comparación de estos modelos se realiza usando datos provinientes de la *Encuesta Nacional de Alimentación y Nutrición en el Medio Rural* 1996 (ENAL 96), la cual fue llevada a cabo por el Instituto Nacional de la Nutrición Salvador Zubirán.

El propósito de los ajustes fue la estimación del riesgo de padecer desnutrición en niños menores de 5 años por medio de ocho variables predictoras (variables explicativas) correspondientes a aspectos socioeconómicos de la familia y algunas características de la madre, utilizando información proveniente de 17 865 observaciones de niños menores de seis años de edad. Cada uno de los niños fue seleccionado de una familia diferente (considerando un niño por madre) y pertenece a una de 853 comunidades rurales analizadas en la encuesta.

Para los ajustes se consideraron interacciones de segundo orden entre las variables predictoras. La selección de variables se realizó en el modelo logístico en su forma simple, utilizando métodos *stepwise*. Con fines comparativos, las variables seleccionadas fueron utilizadas en los ajustes de regresión logística con intersección aleatoria y ponderado por encuesta.

OBJETIVOS

Presentar dos enfoques de regresión logística para datos agrupados (utilizando como referencia al modelo logístico en su forma simple): el modelo logístico con intersección aleatoria y el modelo logístico ponderado por diseño.
Ilustrar los modelos logísticos en datos correspondientes a la ENAL 96. Lo anterior, usando el *software R*.
Comparar los resultados obtenidos de los ajustes realizados.

En el capítulo 1 se presenta la descripción del diseño muestral de la ENAL 96, así como la descripción de la base de datos utilizada. En los capítulos 2,3 y 4 se muestran el modelo

de regresión logística, el modelo de regresión logística con intersección aleatoria y el modelo logístico ponderado por diseño; respectivamente. En cada uno de estos se incluyen apartados ilustrativos del modelo abordado, en los cuales se introducen recuadros de código de *R*. En el capítulo 4 se presenta la aplicación e interpretación de los modelos logísticos, utilizando datos provinientes de la ENAL 96. La conclusión de los resultados obtenidos y la comparación de ajustes se discute en el capítulo 5. Por otro lado, en el Anexo I se incluye un análisis exploratorio de los datos. En el Anexo II se verifica una de las afirmaciones obtenidas por los modelos ajustados: la probabilidad de padecer desnutrición en los infantes disminuye por cada edad cumplido de la madre. Este resultado sólo es válido para algunos rangos de edad de las madres observadas, a saber, madres con edades entre (13,28), (35,38) y (40,44). Por último, el código utilizado en *R* se presenta en el Anexo III.

Capítulo 1

Conceptos previos e introducción de la base de datos utilizada

El objetivo de este capítulo es presentar el conjunto de datos considerados para la aplicación de los modelos de regresión logística abordados. Dicho conjunto proviene de la Encuesta Nacional de Nutrición 1996 (ENAL 96). Ésta fue realizada por el Instituto Nacional Salvador Zubirán para "conocer la situación nutricional de la población infantil en el medio rural" (Avila et al. 1997). Debido a que la ENAL 96 corresponde a una encuesta compleja, de manera previa su descripción, se repasan dos conceptos básicos propios de este tipo de encuestas, los cuales facilitarán el entendimiento del proceso realizado para su obtención: el diseño muestral y los factores de expansión.

1.1. Conceptos previos

1.1.1. Diseño muestral

Las inferencias de datos provenientes de una encuesta compleja son obtenidas considerando la forma en la que la selección de individuos fue realizada. Las muestras obtenidas corresponden a muestras probabilísticas, las cuales se caracterizan por las siguientes propiedades (Lumley 2010, p. 3):

1. Cada individuo en la población debe tener una probabilidad diferente de cero ($p_i \neq 0$) de pertenecer a la muestra.

2. La probabilidad p_i debe ser conocida para cada individuo que pertenezca a la muestra.

3. Cada pareja de individuos en la población debe tener una probabilidad $p_{ij} \neq 0$ de pertenecer a la muestra.

4. La probabilidad p_{ij} debe ser conocida para cada pareja que pertenezca a la muestra.

El diseño de la muestra es realizado de manera previa al proceso de muestreo y permite al investigador mantener bajo control el proceso aleatorio de selección, incluyendo componentes como estratos en los que puede aplicarse, de manera independiente, técnicas para la selección de unidades de muestreo. Con la finalidad de realizar inferencias, los datos a obtener se consideran desconocidos, pero fijos.

> El objetivo del análisis de encuestas complejas es estimar características de la población fija, y la inferencia basada en el diseño no admite la generalización de los resultados a otras poblaciones (Lumley 2010, p. 2).

1.1.2. Factores de expansión

Con el propósito de generalizar los resultados inferidos de una muestra probabilística, cada individuo muestreado se considera representativo de otros individuos de condiciones similares (Lumley 2010, p. 4): un individuo seleccionado con probabilidad p_i, representa $1/p_i$ individuos en la población. Por ejemplo, supongamos que en una población de 1000 individuos, se toma una muestra de 250 individuos utilizando muestreo aleatorio simple. En este caso, para cualquier individuo, la probabilidad de ser seleccionado es $p_i = 250/1000{=}1/4$. Así, cada persona seleccionada representaría a 4 individuos. El valor $1/p_i$ es conocido como factor de expansión o peso muestral (*sampling weight*).

En el ejemplo anterior, se supone que el peso muestral es el mismo para cada individuo. Sin embargo, en las encuestas complejas suelen asignarse pesos distintos a los sujetos muestreados. Por ejemplo, supongamos que se toma una muestra de un niño por familia, utilizando muestreo aletorio simple, en una población particular. Para una familia con 3 niños, el niño seleccionado tendría un peso muestral de 3. Mientras que para una familia con 5 niños, el peso muestral del niño seleccionado sería de 5.

1.2. Diseño muestral de la Encuesta Nacional de Nutrición 1996

La muestra fue obtenida de comunidades rurales con población de 500 a 2500 habitantes, y cuya población económicamente activa estuviera dedicada sobre todo a la agricultura, de acuerdo con la información del X Censo Nacional de Población y Vivienda, 1990 (Ávila et al. 1997). Tres etapas de muestreo fueron consideradas, en las cuales se seleccionaron estratos, comunidades (unidades primarias de muestreo) y viviendas (unidades secundarias de muestreo). La selección de unidades de muestreo en cada etapa fue realizada utilizando muestreo aleatorio simple.

372 estratos fueron definidos. En cada uno de éstos, entre 2 y 4 comunidades fueron seleccionadas, generando un registro de 854 comunidades rurales. En cada comunidad, 50 viviendas se eligieron para la aplicación de la encuesta. Sin embargo, debido a que algunas familias no pudieron ser localizadas, entre 2 y 49 viviendas fueron encuestadas por comunidad, reuniendo un total de 38 232 familias. En cada familia, una muestra de uno a tres niños menores de seis años de edad fue seleccionada (unidades terciarias de muestreo) para la toma de medidas antropométricas con la finalidad de construir indicadores de desnutrición: peso para la edad (*wfa*), talla para la edad (*hfa*), entre otros. Así, el total de niños registrados por la encuesta es de 31 601. También se recolectó información de las madres de familia y de las viviendas.

1.3. Descripción de la base de datos

La base de datos empleada en este trabajo corresponde a un subconjunto de los datos de la ENAL 96. Su construcción, incluyendo la elección de variables explicativas, se debe al trabajo realizado por Eslava y Tjur 2008.

Las observaciones de los niños fueron seleccionadas considerando dos criterios: (1) edad menor o igual a cinco años y (2) edad de la madre entre 12 y 50. Después de esta selección, las observaciones con datos faltantes fueron removidos, obteniendo así un total de 26 819 niños. Posteriormente, utilizando muestreo aleatorio simple, se eligió un niño por familia, generando una muestra con 18 774 infantes o familias. Finalmente, eliminando los datos con observaciones faltantes, se generaron un total de 17 865 observaciones agrupadas en 853 comunidades (*clusters*). Además, debido a que uno de los estratos considerados sólo contenía una comunidad, éste fue adjuntado a uno de sus estratos vecinos. En consecuencia, se trabajó con un total de 371 estratos. Las tablas 1.1 y 1.2 resumen el diseño de muestra y las descripciones dadas de la base de datos.

Tabla 1.1. Diseño de muestra y distribución de las unidades de muestreo. Las comunidades rurales son las unidades primarias de muestreo (*psu*) de la muestra y son consideradas como grupos (*clusters*) en el análisis. Fuente: Eslava 2002.

	Estratos	Comunidades por estratos	Familias por comunidad	Niños por Familia
		2–4	2–49	1–3
Total	372	854	18 774	26 819
En análisis	371	853	17 865	17 865

Tabla 1.2. Distribución del número de niños por familia. Fuente: Eslava 2002.

Niños por familia	Número de casos	Número total de niños	Número de niños en la sub-muestra
1	11 842	11 842	11 842
2	5 819	11 638	5 819
3	1 113	3 339	1 113
Total	18 774	26 819	18 774

Tabla 1.3. Descripción y rango de las variables utilizadas en el ajuste de los modelos de regresión. Nota: las medias presentadas en la tabla corresponden a las medias no ponderadas de los datos.

Variable	**Descripción**	**Rango**
wfa2	Indicador de desnutrición (variable respuesta)	1= con desnutrición; 0 = Normal; moda= 0
foodexp	Gasto semanal per cápita en alimentos	mín= 0.20, media= 27.35, máx= 266.67
persroom	Número de personas por habitación	mín= 0.1429, media= 4.24, máx= 14
agemoth	Edad de la madre	mín= 12, media= 26.90, máx= 49
floor01	Disponibilidad de piso	1= piso de algún material; 0=ninguno; moda= 1
wc01	Disponibilidad de letrina	1=disponible; 0=no disponible; moda=1
cooker01	Disponibilidad de estufa de gas	1=diponible; 0=no disponible; moda= 0
schoolm01	Escolaridad de la madre	1= al menos con primaria concluida; 0=en otro caso; moda= 1
languagem01	Indicador de dialecto hablado por la madre	1=sólo español; 0=español y un dialecto; moda= 1
f3bm18774b	Factores de expansión	
psu	Indicador de comunidad	

Los factores de expansión resultantes de la selección fueron agregados a la base de datos por medio de la variable *f3bm18774b*. Así mismo, se incluyó a la variable *psu*, cuyos valores corresponden a indicadores de comunidad para cada una de las observaciones. Como indicador de desnutrición se consideró a la variable, derivada del peso por edad (*wfa*): *wfa2*= 1 ($wfa \leq -2$) desnutrición y *wfa2*= 0 ($wfa > -2$) normal. Es importante mencionar que la variable peso por edad (*wfa*) fue originalmente medida en una escala que va de -5 a 5 (Eslava y Tjur 2013, p. 5).

La descripción de las variables anteriores, además de las variables explicativas consideradas, se resumen en la Tabla 1.3. Por otro lado, la Tabla 1.4, muestra las medias y proporciones estimadas, considerando los factores de expansión, de las variables utilizadas en el ajuste de los modelos de regresión. Los resultados de las estimaciones realizadas, bajo diseño, muestran que en las comunidades rurales de México:

1. 17 % de las familias tienen un niño que padece desnutrición.

2. El promedio del gasto semanal per cápita es 26.59.

3. El promedio de personas por habitación es aproximadamente 4.
4. El 58 % de las viviendas cuentan con piso de algún material.
5. El 66 % de la población dispone de baño en sus viviendas.
6. El 44 % de la población cuenta con estufa de gas.
7. 72 % de las madres de familia hablan sólo español.
8. 79 % de la madres de familia terminó al menos la primaria.

Tabla 1.4. Medias y proporciones estimados, bajo diseño, de las variables utilizadas en el ajuste de los modelos de regresión.

Variable	**Categoría**	Media	Proporción	IC 95 %
foodexp	Continuo	26.59	–	(25.66, 27.52)
persroom	Conteo	4.40	–	(4.34, 4.46)
agemoth	Conteo	26.74	–	(26.56, 26.93)
floor01	1	–	0.58	(0.56, 0.60)
wc01	1	–	0.66	(0.64, 0.68)
cooker01	1	–	0.43	(0.41, 0.45)
schoolm01	1	–	0.72	(0.70, 0.73)
languagem01	1	–	0.79	(0.76, 0.81)

En los siguientes capítulos, las variables $foodexp$ y $persroom$ son utilizadas como variables predictoras de la respuesta (*wfa2*) para ejemplificar los modelos presentados. Estos ajustes sólo son ilustrativos y no se pretende en ningún momento encontrar un modelo predictivo para el riesgo de desnutrición.

Capítulo 2

Regresión Logística

2.1. Descripción del modelo

Para una variable respuesta $Y \sim binomial(n, \pi)$ y un vector con p variables explicativas $\boldsymbol{x} = (x_0, x_1, \ldots, x_p)^T$, donde $x_0 = 1$; sea $\pi(\boldsymbol{x})$ la probabilidad de éxito π dado $\boldsymbol{x}$. Además, sea $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ un vector de parámetros. De esta forma, el modelo de regresión logística se representa por la ecuación (Agresti 2013, p.192)

$$\pi(\boldsymbol{x}) = \frac{exp(\boldsymbol{\beta}^T\boldsymbol{x})}{1 + exp(\boldsymbol{\beta}^T\boldsymbol{x})} = \frac{exp(\sum_{j=0}^{p}\beta_j x_j)}{1 + exp(\sum_{j=0}^{p}\beta_j x_j)} \tag{2.1}$$

De manera equivalente, la ecuación anterior se escribe como

$$logit[\pi(\boldsymbol{x})] = \boldsymbol{\beta}^T\boldsymbol{x} = \sum_{j=0}^{p}\beta_j x_j \tag{2.2}$$

donde

$$logit[\pi(\boldsymbol{x})] := log\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right)$$

Así, el modelo logístico tiene tres supuestos esenciales:

1. Para cada observación i en la muestra $Y_i \sim binomial(n_i, \pi_i)$.
2. Independencia estocástica de las variables respuesta Y_i.
3. Para cada observación i en la muestra, la función $logit[\pi_i(\boldsymbol{x})]$ resulta de una combinación lineal de las variables explicativas.

Bajo estos supuestos, puede obtenerse un estimador $\hat{\pi}_i$ de π_i utilizando la ecuación (2.1):

$$\hat{\pi}_i = \pi_i(\boldsymbol{x})$$

De esta forma, dado que $E[Y_i] = n_i\pi_i$ para toda i, el valor observado de la variable respuesta Y_i, se estima mediante la siguiente igualdad

$$\hat{y}_i = \widehat{E[Y_i]} = n_i\hat{\pi}_i \tag{2.3}$$

2.1.1. Regresión logística como modelo de clasificación

Cuando en cada observación de la muestra el número de ensayos de la variable Y_i es igual a uno ($n_i = 1$), se tiene que $Y_i \sim binomial(1, \pi_i)$ ó, equivalentemente, $Y_i \sim bernoulli(\pi_i)$. En este caso, utilizando la ecuación (2.3), un estimador para el valor observado de la respuesta y_i es

$$\hat{y}_i = n_i\hat{\pi}_i = \hat{\pi}_i$$

donde $\hat{\pi}_i \in (0, 1)$ para toda i. Sin embargo, la variable respuesta sólo toma dos posibles valores, los cuales no pertenecen al intervalo anterior: $Y_i = 1$ con probabilidad $\pi_i = P[Y_i = 1]$ ó $Y_i = 0$ con probabilidad $1 - \pi_i$. Por tal motivo, no es adecuado estimar el valor de la respuesta utilizando sólo el valor de $\hat{\pi}_i = \pi_i(\boldsymbol{x})$. De manera alternativa, para la estimación se considera un valor de referencia $0 \leq \tau \leq 1$ tal que se pronostica $Y_i = 1$ si y sólo si $\pi(\boldsymbol{x}) > \tau$. El valor τ es conocido como *umbral* (*threshold*). Usualmente se elige $\tau = 0.5$, indicando que se estima el valor de Y más probable con base en la probabilidad de éxito estimada.

2.2. Interpretación del modelo de regresión logística

Los efectos de las variables explicativas sobre la variable respuesta son interpretados por medio de los parámetros β_j. Sin embargo, debido a la relación existente entre β_j y Y_i (ecuación (2.1)), tales interpretaciones no son directas y suelen darse en términos de $\hat{\pi}_i$. Por este motivo, en esta sección, previamente a la interpretación de los coeficientes, se repasan las definiciones de *momios* y *razón de momios*, las cuales facilitarán la interpretación del modelo.

2.2.1. Momios y Razón de momios

Una forma de comparar la probabilidad de ocurrencia π de un evento con su respectiva probabilidad de no ocurrencia $(1 - \pi)$ resulta de considerar el cociente

$$\Omega = \frac{\pi}{1 - \pi} \qquad con \quad 0 \leq \pi < 1.$$

El cociente Ω es conocido por el nombre de $Momio$ (*odd*) y, por definición, $\Omega \geq 0$. Cuando $\Omega > 1$, se tiene que $\pi > 1 - \pi$, indicando que es más probable presenciar la ocurrencia del evento que la ausencia del mismo. Por otro lado, cuando $\Omega < 1$, la ausencia del evento es más probable que su presencia, esto es $1 - \pi < \pi$. Además, por definición se tiene que $\Omega = 1$ si y sólo si $\pi = 1/2$, es decir, $\Omega = 1$ si y sólo si la probabilidad de ocurrencia es la misma que la de ausencia ($\pi = 1/2 = 1 - \pi$). Para ejemplificar, supongamos que la probabilidad de que un alumno presente la enfermedad de las paperas en la facultad de Ciencias UNAM es $\pi = 0.8$. En este caso, $\Omega = 0.8/0.2 = 4$; mostrando que la probabilidad de presentar dicha enfermedad es cuatro veces mayor que la probabilidad de no presentarla. En consecuencia, se esperaría observar una persona no infectada por cada cuatro personas enfermas.

Utilizando momios es posible comparar probabilidades de ocurrencia de dos eventos distintos: consideremos el cociente dado por

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/1 - \pi_1}{\pi_2/1 - \pi_2}; \qquad \text{con} \quad \Omega_2 > 0.$$

El valor de θ es conocido como *razón de momios* (*odds ratio*). Por definición $\theta \geq 0$. Cuando $\theta = 1$, $\pi_i(1-\pi_j) = \pi_j(1-\pi_i)$. En consecuencia, $\pi_i = \pi_j$. Por otra parte, cuando $\theta > 1$, se tiene que $\pi_i > \pi_j$. Finalmente, $\theta < 1$ indica que $\pi_i < \pi_j$. Así, si

$$\theta = \frac{\Omega_1}{\Omega_2} = 3;$$

donde Ω_1 y Ω_2 son los momios de padecer paperas en las facultades de Ciencias y Química de la UNAM, respectivamente, el momio de padecer paperas en la facultad de Ciencias es tres veces mayor que el momio en la facultad de Química y, en consecuencia, la probabilidad de observar un alumno con paperas en la facultad de Ciencias es mayor que en Química.

2.2.2. Interpretación de los parámetros en el modelo

Recordando la definición de momio, podemos notar que la transformación *logit*, definida como

$$logit[\pi(\boldsymbol{x})] = log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right),$$

representa el logaritmo del momio $\pi(\boldsymbol{x})/1-\pi(\boldsymbol{x})$. Este momio compara la probabilidad de éxito contra la probabilidad de fracaso de cada ensayo Bernoulli de la variable respuesta binomial Y. Aplicando la función exponencial en la ecuación (2.2) se obtiene que

$$\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = exp\left(\sum_{j=0}^{p}\beta_j x_j\right) \tag{2.4}$$

De esta forma, utilizando la ecuación anterior, podemos interpretar la magnitud de cada parámetro β_k de la siguiente manera (Agresti 2013, p.164): el momio se multiplica por $exp(\beta_k)$ por cada unidad incrementada de la variable x_k ($k = 1, \ldots, p$) siempre que el resto de variables permanezcan constantes. En otras palabras, $exp(\beta_k)$ es la razón de momios resultante de dividir el momio (ecuación (2.4)) evaluado en x_k+1 entre el momio evaluado para x_k siempre que el resto de variables se mantengan fijas en ambos momios. Usando la ecuación (2.4), lo anterior dicho se muestra en términos matemáticos mediante la siguiente serie de igualdades

$$\begin{aligned}\frac{\textit{Momio evaluado en } x_k+1}{\textit{Momio evaluado en } x_k} &= \frac{exp\left[\left(\sum_{j=0;\, j\neq k}^{p}\beta_j x_j\right)+\beta_k(x_k+1)\right]}{exp\left[\sum_{j=0}^{p}\beta_j x_j\right]} \\ &= \frac{exp\left[\left(\sum_{j=0}^{p}\beta_j x_j\right)+\beta_k\right]}{exp\left[\sum_{j=0}^{p}\beta_j x_j\right]} = exp(\beta_k)\end{aligned}$$

En consecuencia, despejando el valor del momio evaluado en x_k+1,

$$\textit{Momio evaluado en } x_k+1 = exp(\beta_k) \times \textit{Momio evaluado en } x_k;$$

indicando que por cada unidad aumentada de la variable x_k, el momio se multiplica por $exp(\beta_k)$. Además de lo anterior, el signo del parámetro β_k es de gran importancia en el

modelo para determinar si $\pi(\boldsymbol{x})$ aumenta o decrece cuando la variable x_k aumenta. Esto se explica debido a que

$$\frac{\partial \pi(\boldsymbol{x})}{\partial x_k} = \beta_k \pi(\boldsymbol{x})[1 - \pi(\boldsymbol{x})],$$

donde $\pi(\boldsymbol{x})[1 - \pi(\boldsymbol{x})] > 0$ para todo vector $\boldsymbol{x}$. Mostrando de esta manera que si $\beta_k > 0$, la probabilidad $\pi(\boldsymbol{x})$ aumenta cuando x_k aumenta. En contraste, si $\beta_k < 0$, la probabilidad $\pi(\boldsymbol{x})$ disminuye cuando x_k aumenta.

2.3. Estimación y ajuste

Con la finalidad de describir el efecto que tienen las variables explicativas en la variable respuesta Y, es necesario estimar, primero, el vector de parámetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ a partir de la información muestral. El método de Máxima verosimilitud se utiliza para este propósito debido a las propiedades de sus estimadores (Cox y Hinkley 1974, pp. 283–311). Para los apartados siguientes, se introduce la siguiente notación: para una muestra de N elementos, denotemos por y_i al valor observado de la variable respuesta $Y_i \sim binomial(n_i, \pi_i)$ para $i = 1, \ldots, N$. Además, sea $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^T$ el vector de valores observados de las p variables explicativas para la observación i.

Así, el modelo logístico dado por la ecuación (2.1), es (Agresti 2013, p. 192)

$$\pi(\boldsymbol{x}_i) = \frac{exp(\sum_{j=0}^{p} \beta_j x_{ij})}{1 + exp(\sum_{j=0}^{p} \beta_j x_{ij})} = \frac{exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{1 + exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}. \tag{2.5}$$

2.3.1. Ecuaciones de verosimilitud del modelo

Bajo los supuestos del modelo (sección 2.1), la función de verosimilitud de la muestra es

$$l(\boldsymbol{\beta}; \boldsymbol{y}) = f(\boldsymbol{\beta}; \boldsymbol{y}) \;\; = \;\; \prod_{i=1}^{N} \binom{n_i}{y_i} \pi_i(\boldsymbol{x})^{y_i} [1 - \pi_i(\boldsymbol{x}_i)]^{n_i - y_i}, \tag{2.6}$$

donde $\boldsymbol{y} = (y_1, \ldots, y_N)$ es el vector de valores observados de la variable respuesta Y. Se desea encontrar los valores $\hat{\beta}_0, \ldots, \hat{\beta}_p$ que maximizan la ecuación anterior, i.e. se desea encontrar $\hat{\boldsymbol{\beta}} = arg\ máx\ l(\boldsymbol{\beta})$. El procedimiento anterior equivale a encontrar los valores de máximos de la ecuación de *log-verosimilitud* de la muestra dada por

$$L(\boldsymbol{\beta}) = log[l(\boldsymbol{\beta}; \boldsymbol{y})] = log\left[\prod_{i=1}^{N} \binom{n_i}{y_i} \pi_i(\boldsymbol{x})^{y_i} [1 - \pi_i(\boldsymbol{x}_i)]^{n_i - y_i} \right]. \tag{2.7}$$

Lo anterior es debido a que para toda $j = 1, \ldots, p$ (y suponiendo $f(\boldsymbol{\beta}; \boldsymbol{y}) \neq 0$),

$$0 = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \frac{1}{f(\boldsymbol{\beta}; \boldsymbol{y})} \frac{\partial f(\boldsymbol{\beta}; \boldsymbol{y})}{\partial \beta_j} \quad \text{si y sólo si} \quad 0 = \frac{\partial f(\boldsymbol{\beta}; \boldsymbol{y})}{\partial \beta_j}$$

Para encontrar los valores estimados de los coeficientes del modelo, se resuelven las ecuaciones

$$
\begin{aligned}
0 = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j}\left\{ log\left[\prod_{i=1}^{N} \pi_i(\boldsymbol{x})^{y_i}[1-\pi_i(\boldsymbol{x}_i)]^{n_i-y_i}\right]\right\} \\
&= \frac{\partial}{\partial \beta_j}\left\{\sum_{i=1}^{N}\left[y_i\left(\frac{exp\left\{\sum_{k=0}^{p}\beta_k x_{ik}\right\}}{1+exp\left\{\sum_{k=0}^{p}\beta_k x_{ik}\right\}}\right)+(n_i-y_i)\left(1-\left(\frac{exp\left\{\sum_{k=0}^{p}\beta_k x_{ik}\right\}}{1+exp\left\{\sum_{k=0}^{p}\beta_k x_{ik}\right\}}\right)\right)\right]\right\} \\
&= \sum_{i=1}^{N} y_i x_{ij}\left[n_i x_{ij}\left(\frac{exp\left\{\sum_{k=0}^{p}\beta_k x_{ik}\right\}}{1+exp\left\{\sum_{k=0}^{p}\beta_k x_{ik}\right\}}\right)\right] \\
&= \sum_{i=1}^{N} y_i x_{ij} - n_i \pi_i(\boldsymbol{x}) x_{ij};
\end{aligned}
$$

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

Estas ecuaciones son conocidas por el nombre de ecuaciones de verosimilitud, y las soluciones $\hat{\beta}_j$ se conocen como *estimadores máximo verosímiles de la muestra*. Debido a que las ecuaciones de verosimilitud no son lineales para los parámetros β_j, sus estimadores son calculados por métodos iterativos (véase Agresti 2015, pp.176–177).

Una de las propiedades más importantes de los estimadores $\hat{\beta}_j$ ocurre cuando el número de observaciones tiende a infinito ($N \to \infty$), y resulta como consecuencia de las suposiciones del modelo: el estimador máximo verosímil $\hat{\beta}_j$ de β_j se distribuye normal con media β_j (para $j = 1, 2, \ldots, n$) y varianza $\hat{V}(\hat{\beta}_j)$ igual a la inversa de la información de Fisher (Cox y Hinkley 1974, p. 294). Esta propiedad permite realizar inferencias sobre los efectos de las variables explicativas.

2.4. Pruebas de hipótesis y significancia estadística de efectos en el modelo

Como se mencionó con anterioridad, el valor β_j representa, de manera indirecta, el efecto que tiene la variable explicativa x_j sobre la variable respuesta Y. Así, bajo el modelo en consideración, cuando $\beta_j = 0$, la variable x_j no tiene ningún efecto sobre Y debido a que

$$
logit[\pi_i(\boldsymbol{x})] = \sum_{k=0}^{p} \beta_k x_{ik} = \sum_{k=0,\ k\neq j}^{p} \beta_k x_{ik}
$$

Es decir, el modelo que contiene a dicha variable resulta el mismo que aquél que no la considera. Cuando esto ocurre, se dice que la variable x_j no tiene un efecto estadísticamente significativo en la respuesta y a menudo, por simplicidad, sólo se dice que la variable x_j no es estadísticamente significativo. Resulta entonces imprescindible obtener evidencias estadísticas que prueben la significancia de efectos de las variables en el modelo. Por esta razón, dos pruebas de hipótesis clásicas se describen en seguida, en las cuales, de manera general, se contrasta la hipótesis nula $H_0 : \beta_j = \beta$ contra la hipótesis alternativa $H_1 : \beta_j \neq \beta$; i.e,

$$
H_0 : \beta_j = \beta \quad \text{vs} \quad H_1 : \beta_j \neq \beta;
$$

donde $\beta \in \mathbb{R}$.

2.4.1. Prueba de Wald

El estadístico de Wald se define como

$$z = \frac{\hat{\beta}_j - \beta}{\hat{se}(\hat{\beta}_j)},$$

donde $\hat{se}(\hat{\beta}_j)$ denota a la desviación estándar estimada $\hat{se}(\hat{\beta}_j) = \sqrt{\hat{v}(\hat{\beta}_j)}$ del parámetro estimado $\hat{\beta}_j$.
Recordemos que $\hat{\beta}_j$ tiene una distribución asintóticamente normal con media β_j y varianza igual a la inversa de la información de Fisher correspondiente (sección 2.3). Consecuentemente, bajo la hipótesis nula $H_0 : \beta_j = \beta$, z tiene una distribución normal estándar.

Criterio de rechazo: La hipótesis nula es rechazada al nivel de significancia α si y sólo si $z \leq z_{\frac{\alpha}{2}}$ ó $z \geq z_{1-\frac{\alpha}{2}}$.

Donde $z_{\frac{\alpha}{2}}$, $z_{1-\frac{\alpha}{2}}$ son los cuantiles $\frac{\alpha}{2}$, $(1 - \frac{\alpha}{2})$ de una distribución normal estándar, respectivamente. De manera equivalente, la hipótesis nula es rechazada si y sólo si el valor p (*p-value*) correspondiente a la hipótesis nula es menor que el nivel de significancia α. Debido a la simetría de la distribución normal, se tiene que $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$. Por lo tanto, el criterio de rechazo usualmente se escribe de la siguiente manera:

Criterio de rechazo: La hipótesis nula es rechazada al nivel de significancia α si y sólo si $z \leq -z_{1-\frac{\alpha}{2}}$ ó $z \geq z_{1-\frac{\alpha}{2}}$.

Una alternativa que deriva de la estadística z es considerar z^2, el cual tiene una distribución asintótica χ^2 con un grado de libertad. La diferencia de aplicación entre estos dos estadísticos resulta en el hecho de que si se utiliza a z, el contraste es bilateral (la región de rechazo se encuentra en los dos extremos de su distribución, esto conlleva a considerar dos cuantiles); mientras que si se utiliza el estadístico z^2, el contraste es unilateral, esto es debido a que la distribución χ^2 no resulta simétrica. El criterio de rechazo para z^2 se resume en lo siguiente: la hipótesis nula es rechazada al nivel de significancia α si y sólo si $z^2 \geq z_{1-\alpha}$, con $z_{1-\alpha}$ el cuantil $1 - \alpha$ de una distribución χ^2 con un grado de libertad.

2.4.2. Razón de verosimilitudes

En la prueba se requiere calcular el cociente o razón

$$\Lambda = \frac{l_0}{l_1},$$

donde l_0 es la función de verosimilitud bajo la hipótesis nula, i.e. evaluada en $\beta_j = \beta$ y $\beta_k = \hat{\beta}_k$ para $k \neq j$; mientras que l_1 es la función de verosimilitud maximizada en todo el espacio paramétrico, es decir, evaluada en $\beta_k = \hat{\beta}_k$ para $k = 1, ..., p$. Wilks (1938) probó, de forma más general, que $-2log\Lambda$ tiene una distribución asintótica χ^2 con un grado de libertad cuando el número de observaciones N tiende a infinito.

Criterio de rechazo: La hipótesis nula es rechazada al nivel de significancia α si y sólo si $-2log\Lambda \geq z_{1-\alpha}$, donde $z_{1-\alpha}$ representa el cuantil $(1 - \alpha)$ de una distribución χ^2 con un grado de libertad.

Análogamente a la prueba de Wald, y de manera equivalente al criterio anterior, la hipótesis nula es rechazada si y sólo si el valor de p, correspondiente a la hipótesis nula, es menor que el nivel de significancia α.

2.5. Intervalos de confianza

Por medio de las ecuaciones de verosimilitud, para cada parámetro β_k se ha encontrado un único estimador $\hat{\beta}_k$ de manera que éste maximice la probabilidad de obtener los registros observados en la muestra. Sin embargo, dicho valor podría diferir del verdadero valor del parámetro a estimar, introduciendo así un margen de incertidumbre. En general, la motivación de los intervalos de confianza es encontrar un rango de valores en donde, con cierto nivel de certeza, es posible ubicar al verdadero valor del parámeto (Kendall y Stuart 1961, p. 98). Utilizando los estadísticos de las dos pruebas de hipótesis presentadas en la sección anterior es posible construir intervalos de confianza considerando lo siguiente: un intervalo del $(1-\alpha)\times 100\,\%$ de confianza es el conjunto de valores β para los cuales la prueba $H_0 : \beta_j = \beta$ tiene un valor de p mayor que α; en otras palabras, es el conjunto de valores β tales que la hipótesis nula $H_0 : \beta_j = \beta$ no es rechazada al nivel de significancia α. De esta manera, se genera la siguiente equivalencia: $H_0 : \beta_j = \beta$ no es rechazada al nivel de significancia α si y sólo si β pertenece al intervalo con nivel de confianza $(1-\alpha)\times 100\,\%$. La construcción de intervalos de confianza utilizando los estadísticos z y $-2log\Lambda$ empleados en las pruebas respectivas de Wald y de Razón de verosimilitudes se presentan a continuación.

1. *Intervalo de confianza utilizando el estadístico de Wald z*:

   Considerando el criterio de rechazo correspondiente, el intervalo al nivel de confianza $(1-\alpha)\times 100\,\%$ para un parámetro, digamos β_j, se conforma por los valores β tales que

   $$-z_{1-\frac{\alpha}{2}} \leq z \leq z_{1-\frac{\alpha}{2}}$$

   Donde $z_{1-\frac{\alpha}{2}}$ representa el cuantil $1-\frac{\alpha}{2}$ de un distribución normal estándar.

   Recordando que

   $$z = \frac{\hat{\beta}_j - \beta}{\hat{se}(\hat{\beta}_j)},$$

   se obtiene el intervalo para β_j:

   $$\hat{\beta}_j - z_{1-\frac{\alpha}{2}} \times \hat{se}(\hat{\beta}_j) \leq \beta \leq \hat{\beta}_j + z_{1-\frac{\alpha}{2}} \times \hat{se}(\hat{\beta}_j) \tag{2.8}$$

2. *Intervalo de confianza utilizando el estadístico $-2log\Lambda$*:

   De manera análoga al intervalo anterior, el intervalo del parámetro β_j, al nivel de confianza $(1-\alpha)\times 100\,\%$ está conformado por los valores β tales que

   $$-2log\Lambda \leq z_{1-\alpha}$$

   donde $z_{1-\alpha}$ es el cuantil $1-\alpha$ de una distribución χ^2 con un grado de libertad.

2.5.1. Intervalo de confianza para las razones de momios $exp(\beta)$

Aplicando la función exponencial en el intervalo de Wald (ecuación (2.8)) es posible obtener un intervalo del $(1-\alpha)\times 100\,\%$ de confianza para $exp(\beta_j)$:

$$exp\left\{\hat{\beta}_j - z_{1-\frac{\alpha}{2}} \times \hat{se}(\hat{\beta}_j)\right\} \leq exp(\beta_j) \leq exp\left\{\hat{\beta}_j + z_{1-\frac{\alpha}{2}} \times \hat{se}(\hat{\beta}_j)\right\} \tag{2.9}$$

2.6. Ejemplo: Interpretación de los parámetros del modelo

Con fines ilustrativos, consideremos los datos provinientes de la *Encuesta Nacional de Alimentación y Nutrición en el Medio Rural 1996* (ENAL 96), ignorando, por ahora, el diseño muestral de la encuesta. Supongamos que se pretende analizar los efectos que tienen el gasto semanal per cápita en alimentos ($foodexp$) y el número de personas por habitación en casa ($persroom$) sobre la probabilidad de desnutrición en niños menores de 5 años de edad en zonas rurales de México (*wfa2* como indicador de desnutrición). El siguiente recuadro muestra las primeras 12 observaciones de estas variables.

```
-------------------------------------------------------------------
 #Primeras 12 observaciones de las variables en la base de datos:

    wfa2  foodexp   persroom          wfa2  foodexp   persroom
1    0  58.33333      2.0         7    0   37.50000      4.0
2    0  20.00000      7.0         8    0   40.00000      1.5
3    0  16.66667      3.0         9    0   50.00000      3.0
4    0  18.18182      5.5        10    0   28.57143      3.5
5    0  50.00000      2.0        11    0   11.11111      9.0
6    0  25.00000      4.0        12    0   16.66667      4.5
-------------------------------------------------------------------
```

La salida de R del ajuste logístico planteado se presenta en el siguiente recuadro. Como se mencionó con anterioridad, la solución numérica de las ecuaciones de verosimilitud del modelo logístico se logra utilizando métodos iterativos. El número de iteraciones realizadas para la obtención de dicha solución es indicada por *"Number of Fisher Scoring iterations: ... "* en la salida de R. En este caso, fueron necesarias cinco iteraciones para realizar el ajuste. El método utilizado para la estimación de la solución es llamado *Newton-Rapson* y, en el caso de la regresión logística (al igual que otros Modelos Lineales Generalizados), dicho método resulta equivalente al método iterativo *Fisher scoring* (Agresti 2015, pp. 140–143).

```
-------------------------------------------------------------------

#Importando base de datos:
```

```
> datos=read_excel("finfil1f_m17865.xls")

> attach(datos)
> enal=data.frame(wfa2, foodexp, persroom, agemoth, floor01, wc01,
                  cooker01,  schoolm01, languagem01,
                  psu,estratob,f3bm18774b)
> detach(datos)

#Ajuste logístico
> M1=glm(wfa2~foodexp+persroom, data=enal, family=binomial)
> summary(M1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.664179   0.068192 -24.404   <2e-16 ***
foodexp     -0.015817   0.001411 -11.213   <2e-16 ***
persroom     0.098769   0.010110   9.769   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of Fisher Scoring iterations: 5
-----------------------------------------------------------------------
```

Los resultados obtenidos indican que, para $i = 1, \ldots, N$;

$$logit[\hat{\pi}_i(\boldsymbol{x})] = -1.664 - 0.016 \times foodexp + 0.099 \times persroom. \tag{2.10}$$

Este modelo indica que la probabilidad estimada de observar a un niño con desnutrición disminuye cuando aumenta el gasto semanal per cápita en alimentos, multiplicando el momio por $exp(-0.016) = 0.984$ por cada unidad aumentada en dicho gasto cuando el número de personas en casa es constante. En otras palabras, el momio disminuye en promedio $(1 - 0.984)\times100$=1.6 por ciento por cada unidad invertida en el gasto semanal per cápita en alimentos. Lo anterior siempre que el número de personas en casa permanece fijo. En contraste, el número de personas por habitación en la casa tiene un efecto positivo en dicha probabalidad, incrementándola cuando el número de personas por habitación aumenta. En este caso, el momio se multiplica, en promedio, por $exp(0.099) = 1.104$ por cada persona agregada por habitación siempre que el gasto semanal per cápita permanece constante. De manera equivalente, el momio es $(1.104 - 1.000)\times100$=10.4 por ciento mayor por cada persona adicional por habitación. Los intervalos de confianza para los valores estimados de las razones de momios se muestran en la Tabla 2.1.

Tabla 2.1. Razones de momios ajustados para las variables predictoras del modelo logístico estimado.

Var. predictora	$exp(\hat{\beta})$	**IC 95 % para** $exp(\beta)$
foodexp	0.984	(0.982, 0.987)
persroom	1.104	(1.0821, 1.126)

Los valores del estadístico z (*z value*), así como los *p-values* ($Pr(>|z|)$) reportados por la salida de R corresponden a las pruebas de significancia de las variables explicativas en el modelo.

Dado que $p<0.05$ para cada coeficiente, los efectos de las variables explicativas sobre la respuesta son significativos si se elige $\alpha = 0.05$.

En el siguiente recuadro se muestran intervalos para los coeficientes del modelo al nivel de confianza del 95 %, utilizando las pruebas de Wald y de Razón de verosimilitudes correspondientes a la prueba $H_0 : \beta_i = \beta \quad vs \quad H_1 : \beta_i \neq \beta$; donde $\beta \in \mathbb{R}$. En cada uno de los intervalos encontrados, podemos notar que el valor $\beta = 0$ no pertenece a estos. Lo que implica que β_i es estadísticamente diferente de cero al nivel de confianza del 95 %, cumpliendo la equivalencia planteada en la sección 2.5.

```
-----------------------------------------------------------------------

##Intervalo de confianza Wald
#La función "print()" es aplicada para reducir el número de dígitos.

> print(confint(M1,type = "Wald",level=0.95), digits=3)
Waiting for profiling to be done...
              2.5 %  97.5 %
(Intercept) -1.7980 -1.5307
foodexp     -0.0186 -0.0131
persroom     0.0789  0.1185

##Intervalo de confianza razón de verosimilitudes

> print(confint(M1,type = c("profile"), level=0.95), digits=3)
Waiting for profiling to be done...
              2.5 %  97.5 %
(Intercept) -1.7980 -1.5307
foodexp     -0.0186 -0.0131
persroom     0.0789  0.1185
-----------------------------------------------------------------------
```

Los intervalos de confianza estimados para los coeficientes del modelo, usando los estadísticos de Wald y Razón de verosimilitudes, resultan numéricamente iguales (Recuadro anterior). Agresti (2015) menciona que "cuando $N \to \infty$, ... los intervalos [de Wald y Razón de verosimilitudes] tienen ciertas equivalencias asintóticas" (p. 131). En la aplicación presentada en el capítulo 5, el intervalo de confianza utilizado es el intervalo de Wald.

Los valores π_i pueden ser estimados empleando la igualdad (2.1). Por ejemplo, para la primera observación en la base de datos se tiene que $foodexp = 58.334$ y $persroom = 2$; en consecuencia

$$\hat{\pi} = \frac{exp\{-1.664 - 0.016(58.334) + 0.099(2)\}}{1 + exp\{-1.664 - 0.016(58.334) + 0.099(2)\}} = 0.084 \tag{2.11}$$

Así, si consideramos como umbral el valor $\tau = 0.5$, se estima $\widehat{wfa2}$=0 debido a que $\hat{\pi} < 0.5$. Esto es, para cualquier familia perteneciente a una zona rural de México con gasto semanal per cápita en alimentos igual a 58.334 y con 2 personas por habitación en casa, se espera que, en promedio, uno de los hijos padezca desnutrición si estos tienen menos de cinco años de edad. Dado que el valor estimado $\widehat{wfa2}$ =0 coincide con el observado *wfa2* =0, la primera observación en la base de datos es clasificada de manera correcta.
El modelo anterior supone que el efecto del gasto semanal no depende del número de personas por habitación. Por ejemplo, podría ocurrir que el efecto del gasto semanal per cápita en la respuesta sea mayor cuando el número de personas por habitación sea menor. Dicha dependencia se agrega en el modelo por medio de la interacción entre estas variables.

```
------------------------------------------------------------------------
> M2=glm(wfa2~ foodexp+persroom+foodexp:persroom,
data=enal, family=binomial)
> summary(M2)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.6927676  0.0876083 -19.322  < 2e-16 ***
foodexp          -0.0143849  0.0030946  -4.648 3.35e-06 ***
persroom          0.1058005  0.0169066   6.258 3.90e-10 ***
foodexp:persroom -0.0003834  0.0007402  -0.518    0.605
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of Fisher Scoring iterations: 5
------------------------------------------------------------------------
```

Los resultados indican que, para una familia con dos personas por habitación en casa, el efecto del gasto semanal es de $-0.015 = -0.0144 - 0.0003(2)$, implicando el siguiente modelo

$$logit[\hat{\pi}_i(\boldsymbol{x})] = -1.4811 - 0.015 \times foodexp$$

El valor -1.4811 resulta de sumar a la intersección (-1.6927) el efecto principal de *persroom* (0.1058×2). Por otro lado, para una familia con 7 personas por habitación en casa, el modelo es

$$logit[\hat{\pi}_i(\boldsymbol{x})] = -0.9521 - 0.0165 \times foodexp$$

En este caso, el nuevo efecto de *foodexp* resulta de la suma $-0.0144 - 0.0003(7) = -0.0165$. Así, la razón de momios estimada para *foodexp* es $exp(-0.0165)$=0.9836. Sin embargo, los

resultados obtenidos indican que la interacción entre las variables $foodexp$ y $persroom$ no es significativa para el ajuste ($p < 0.05$). Por lo tanto, tal interacción puede ser descartada del modelo.

2.7. Bondad de ajuste

Obtener los valores estimados de los parámetros del modelo y la significancia de cada uno de estos nos permite saber qué tan importantes resultan las variables, individualmente, en la descripción de la variable respuesta. Sin embargo, el análisis de significancia de efectos no equivale a verificar qué tan eficaz resulta el pronóstico del ajuste. En general, los valores pronosticados $\hat{y}_i$ no serán exactamente iguales a los valores observados de la respuesta. Por consiguiente, cuantificar, de alguna manera, qué tan diferentes son las estimaciones de los valores observados se convierte en una prioridad en el análisis del modelo. En las siguientes secciones se presentan alternativas para verificar el desempeño de los modelos logísticos.

2.7.1. Tablas de confusión y curva *ROC*

Una forma de analizar el desempeño del modelo es comparando directamente los valores ajustados con los valores observados de la variable respuesta Y. La tabla de confusión es un arreglo tabular de 2×2 que permite visualizar dichos valores. En la Tabla 2.2 se muestra su estructura general.

Tabla 2.2. Estructura general de una Tabla de confusión.

	Valores observados de Y	
Valores ajustados de Y	0	1
0	x_{11}	x_{12}
1	x_{21}	x_{22}

Los elementos de la diagonal principal x_{11} y x_{22}, representan a las observaciones clasificadas de manera correcta por el modelo. Por otro lado, los valores de la diagonal secundaria x_{12} y x_{21} representan a las observaciones clasificadas erróneamente por el ajuste. La proporción de estos errores, a saber,

$$P_e = \frac{x_{12} + x_{21}}{x_{11} + x_{12} + x_{21} + x_{22}},$$

es utilizada como una medida de error del modelo. Cuando el modelo es ajustado utilizando todas las observaciones de la muestra, P_e es conocido por el nombre de *error aparente*. Este error resulta ser una medida subestimada, debido a que el modelo es evaluado con las mismas observaciones con las que fue ajustado, minimizando así la magnitud del error. En contraste, P es llamado *error no aparente* cuando es calculado utilizando un conjunto de observaciones diferente al utilizado para el ajuste. De manera general se cumple que

$$\text{Error no aparente} > \text{Error aparente}.$$

Por otro lado,

$$Sp = P[\hat{y} = 0|y = 0],$$

es llamada *especificidad* (*Specificity*), y representa la proporción de observaciones con respuesta $y = 0$ que fueron correctamente clasificados ($\hat{y} = 0$). La especificidad puede ser estimada del ajuste utilizando la igualdad

$$\hat{Sp} = \hat{P}[\hat{y} = 0|y = 0] = \frac{x_{11}}{x_{11} + x_{21}}.$$

De manera análoga a la definición de especificidad, se define a la *sensibilidad* (*Sensitivity*) como la proporción de observaciones con respuesta $y = 1$ que fueron correctamente clasificados ($\hat{y} = 1$), i.e

$$Se = P[\hat{y} = 1|y = 1].$$

La sensibilidad es estimada mediante la igualdad

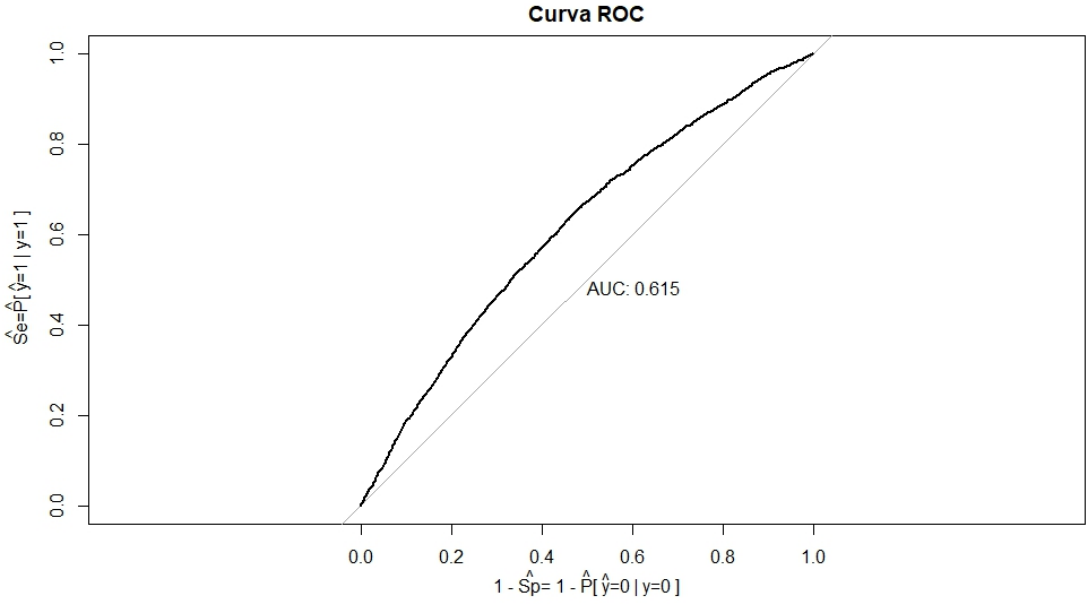
$$\hat{Se} = \hat{P}[\hat{y} = 1|y = 1] = \frac{x_{22}}{x_{12} + x_{22}}.$$

Las proporciones anteriores (P_e, $\hat{Sp}$, $\hat{Se}$) varían de acuerdo con el umbral (*threshold*) utilizado en el proceso de clasificación de observaciones. La curva *ROC* (*Receiver Operating Characteristics*) es una herramienta gráfica que permite visualizar dichas proporciones para todos los posibles umbrales. Esta curva corresponde a la gráfica de $\hat{Sp} = \hat{P}[\hat{y} = 0|y = 0]$ contra $1-\hat{Se}= 1-\hat{P}[\hat{y} = 1|y = 1]$ para cada posible umbral, y es útil para la evaluación del desempeño predictivo del modelo. El valor $1-\hat{Se}$ corresponde a la proporción estimada de observaciones con respuesta $y = 0$ que fueron clasificados erróneamente ($\hat{y} = 1$). La Figura 2.1 muestra la curva *ROC* del modelo ajustado en el ejemplo anterior.

```
-------------------------------------------------------------------------
library(pROC)
#Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique
#Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an
#open-source package for R and S+ to analyze and compare ROC curves.

#Modelo
M1=glm(wfa2~foodexp+persroom, data=enal,
family=binomial)

#Probabilidades estimadas
prob=predict(M1,type=c("response"))

#Curva ROC
prob=predict(M1,type=c("response"))
roc_glm<- roc(nutre$wfa2 ~ prob, plot=TRUE, print.auc = TRUE,
              legacy.axes = TRUE, main="Curva ROC",
              xlab=expression(paste("1 - ", hat(Sp), "= 1 - ",hat(P),
              "[ ",hat(y),"=0", " | y=0 ]")), ylab=expression(paste
              (hat(Se), "=", hat(P) ,"[ ",hat(y),"=1", " | y=1 ]")))
-------------------------------------------------------------------------
```

La recta identidad (Se = 1- Sp) representa el caso en el que no existe una relación entre las variables explicativas y la respuesta. Por este motivo, la recta identidad sirve como una curva de referencia de modelos "no informativos" (Hastie et al. 2013, pp. 147—148). Para resumir el desempeño general del modelo se utiliza el área bajo la curva (AUC). Una curva ROC "ideal" se espera con un AUC cercano a uno (≈ 1). En el ejemplo planteado se tiene que $AUC = 0.615$, por lo que se puede concluir que el desempeño predictivo del modelo es pobre.

Figura 2.1. *Curva ROC del modelo ajustado en el ejemplo anterior. El desempeño del ajuste no es muy bueno (AUC=0.615).*



2.7.2. Devianza

Cuando el número de parámetros es igual al número de observaciones, es posible lograr un ajuste perfecto por medio del modelo, donde los valores estimados son exactamente los valores observados de la variable respuesta. Sin embargo, este modelo no es útil en la práctica ya que se ajustan perfectamente a esos datos específicos y, en consecuencia, el modelo podría predecir erróneamente a la respuesta frente a nuevos datos. Dicho modelo es conocido como *modelo saturado* y el efecto que conlleva predecir de manera equivocada frente a nuevos datos es llamado *sobre-estimación* (*overfitting* en inglés). A pesar de sus inconvenientes, el modelo saturado es de gran utilidad como base de discrepancia para comparar desempeños de modelos con un menor número de variables: sea $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_0, \ldots, \tilde{\boldsymbol{\beta}}_N)$ el vector de los estimadores máximo verosímiles del modelo saturado, i.e. $\tilde{\boldsymbol{\beta}} = argmáx f(\boldsymbol{\beta}; \boldsymbol{y})$. Análogamente, sea $\hat{\boldsymbol{\beta}} = argmáx f(\boldsymbol{\beta}; \boldsymbol{y})$. Consideremos la siguiente expresión

$$D = -2log\left[\frac{f(\hat{\boldsymbol{\beta}}; \boldsymbol{y})}{f(\tilde{\boldsymbol{\beta}}; \boldsymbol{y})}\right] = 2[L(\tilde{\boldsymbol{\beta}}) - L(\hat{\boldsymbol{\beta}}] \tag{2.12}$$

El valor D en la ecuación anterior corresponde al estadístico de Wilks en la prueba de razón de verosimilitudes y, en el caso de la regresión logística, es conocido como la *devianza* del modelo. Dado que el espacio parametral del modelo ajustado se encuentra contenido en

el espacio parametral del modelo saturado (i.e. el conjunto de posibles valores que pueden tomar los coeficientes β_j en el modelo ajustado se encuentra contenido en el conjunto de posibles valores que pueden tomar los coeficientes β_j en el modelo saturado), se tiene que (Agresti 2015, p. 132)

$$f(\hat{\boldsymbol{\beta}}; \boldsymbol{y}) \leq f(\tilde{\boldsymbol{\beta}}; \boldsymbol{y})$$

Como consecuencia de lo anterior, $D \geq 0$, resultando que una mayor devianza en el modelo ajustado implica un peor desempeño. Para un modelo ideal con p variables explicativas se tendría $D = 0$, mostrando que el modelo ajustado tiene el mismo desempeño predictivo que el modelo saturado. En el caso de regresión logística, cuando se considera $Y_i \sim bernoulli(\pi_i)$, la función de verosimilitud del modelo saturado es

$$f(\text{modelo saturado}) = \prod_{i=1}^{N} f(\hat{y}_i; y_i) = \prod_{i=1}^{N} y_i^{y_i}(1-y_i)^{1-y_i} = 1$$

donde $f(\hat{y}_i; y_i)$ es la función de verosimilitud (ecuación 2.6) evaluada en los valores estimados $\hat{y}_i$. Lo anterior es debido a que $\hat{y}_i = y_i$, siendo $y_i = 0$ ó $y_i = 1$ para toda i. Así, el valor de la devianza es

$$D = -2log(\text{verosimilitud del modelo ajustado}) = -2[L(\hat{\boldsymbol{\beta}})].$$

La devianza, además, permite la comparación de modelos anidados. Dos modelos son llamados anidados si el conjunto de parámetros de uno de ellos es un subconjunto del conjunto de parámetros del otro, i.e. $\{\beta_0, \ldots, \beta_q\} \subseteq \{\beta_0, \ldots, \beta_p\}$.

Supongamos que se tienen dos modelos anidados M_1 y M_2 con un total de q y p variables explicativas respectivamente, siendo $q \leq p$. Así, el espacio parametral de M_1 se encuentra contenido en el de M_2 y, por consiguiente, la función de verosimilitud es mayor para M_2. De esta forma, si D_1 y D_2 son las devianzas de los modelos M_1 y M_2 respectivamente, puede comprobarse que $D_2 \leq D_1$ (Agresti 2015, p. 134). La desigualdad anterior muestra que ajustes más simples implican una mayor devianza. Por lo tanto, mientras mayor sea la diferencia $D_1 - D_2$, peor será el ajuste M_1 al ser comparado con M_2. En el caso de la regresión logística, $D_1 - D_2$ se distribuye asintóticamente ji cuadrada con $p - q$ grados de libertad ($\chi^2_{(p-q)}$). En consecuencia, es posible realizar el contraste de hipótesis $H_0 : D_1 - D_2 = 0$ *vs* $H_1 : D_1 - D_2 \neq 0$. La hipótesis nula H_0 es rechazada si y sólo si el modelo D_2 presenta un mejor desempeño que el modelo D_1, i.e. H_0 es rechazada si y sólo si hay evidencia estadística de que el modelo con más variables presenta un mejor desempeño que modelo con menos variables. La prueba anterior es equivalente al contraste de hipótesis $H_0 : \beta_{q+1}, \ldots, \beta_p = 0$ *vs* $H_1 : \beta_j \neq 0$ para alguna $j \in \{q+1, \ldots, p\}$.

2.7.3. *AIC* Y *BIC*

El *criterio de información de Akaike* (o *AIC* por sus siglas en inglés), es un valor que escencialmente penaliza a un modelo por tener muchos parámetros (Agresti 2015, p. 146). Si $\hat{\boldsymbol{\beta}}$ es el estimador máximo verosímil de $\boldsymbol{\beta}$ dado por el modelo ajustado $M = logit[\pi_i]$, el *AIC* se define como:

$$AIC = -2\left[L(\hat{\boldsymbol{\beta}}) - k\right],$$

donde k es el número de parámetros en M. Un menor valor de AIC indica un mejor desempeño. Una alternativa al AIC es el *Criterio de información de Bayes* (BIC). Éste se define como

$$BIC = -log(n)\left[L(\hat{\boldsymbol{\beta}}) - k\right].$$

El nombre "BIC" se debe al enfoque estadístico con el que éste se fundamenta: la Estadística Bayesiana. A diferencia del AIC, el BIC se caracteriza por penalizar más severamente por el número de parámetros del modelo.

2.8. Prueba de significancia del modelo

La prueba de significancia del modelo o significancia global del modelo corresponde a la prueba de hipótesis

$$H_0 : (\beta_1, \ldots, \beta_p) = \mathbf{0} \quad vs \quad H_1 : \beta_j \neq 0 \text{ para alguna } j \in \{1, \ldots, p\}.$$

Esta prueba permite determinar si existe un efecto entre la respuesta y cualquiera de las variables predictoras $(x_1, \ldots, x_p)$. Rechazar la hipótesis nula implica que al menos una de las variables explicativas tiene un efecto significativo en la respuesta. En otras palabras, rechazar H_0 indica que el ajuste planteado, con $p \geq 1$ variables explicativas, proporciona un mejor ajuste que el modelo sin variables predictoras dado por

$$logit[\pi(x)] = \beta_0. \tag{2.13}$$

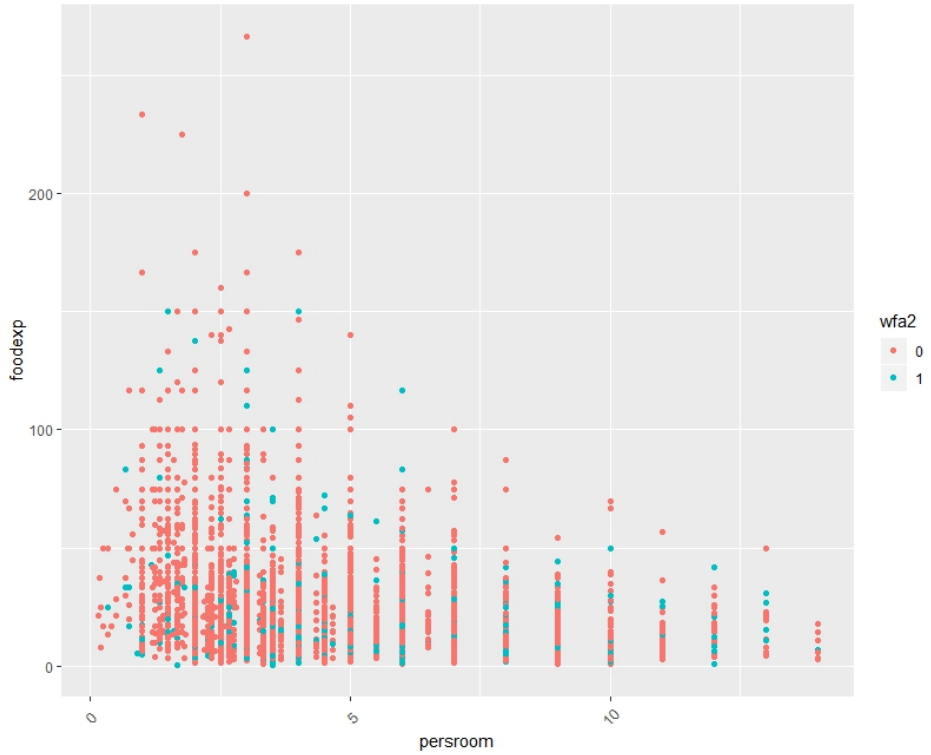
El modelo anterior es conocido por el nombre de *Modelo nulo* y su correspondiente devianza es llamada *devianza nula* (*Null deviance*). La prueba de significancia del modelo equivale a probar la diferencia de devianzas del modelo nulo y el modelo planteado: $H_0 : D_M - D_{Null} = 0$ *vs* $H_1 : D_M - D_{Null} \neq 0$.

Criterio de rechazo: La hipótesis nula es rechazada al nivel de significancia α si y sólo el valor de p (*p-value*) es menor que el nivel de significancia α.

2.9. Ejemplo: Evaluación de desempeño del modelo

En el ejemplo anterior, utilizando los datos de la ENAL 96, se ajustó un modelo de regresión logística para la presencia de desnutrición en niños menores de 5 años utilizando como variable respuesta al indicador *wfa2*. Las variables explicativas utilizadas fueron el gasto semanal per cápita en comida (*foodexp*) y el número de personas por habitación (*persroom*). La Figura 2.2 muestra la gráfica de la variable *persroom* contra la variable *foodexp*, distinguiendo observaciones por sus respectivos valores observados de *wfa2*. Podemos notar que no existe una distinción de las observaciones de acuerdo a los valores de la respuesta, ya que las observaciones con desnutrición (*wfa2*=1) se confunden con las observaciones sin desnutrición (*wfa2*=0). En consecuencia, podríamos esperar que el desempeño predictivo del modelo no sea destacable.

Figura 2.2. *Gráfica de dispersión del número de personas por habitación contra el gasto semanal per cápita en comida, distinguiendo observaciones por sus respectivos valores observados de wfa2 (presencia o ausencia de desnutrición).*



La función *anova*() en el programa R permite comparar la devianza de dos modelos anidados, con la posibilidad de incluir la prueba de hipótesis $H_0 : D_2 - D1 = 0$ *vs* $H_1 : D_2 - D1 \neq 0$. En este ejemplo compararemos los valores de la devianza y el AIC reportados por los modelos con y sin interacción realizados en el ejemplo previo.

```
-------------------------------------------------------------------------
#Modelo con efectos principales
> M1=glm(wfa2 ~ foodexp + persroom, data=enal,  family=binomial)
> (D1=deviance(M1))
[1] 15622.77

#Modelo con interacción
> M2=glm(wfa2 ~ foodexp + persroom + foodexp:persroom, data=enal,
            family=binomial)
> (D2=deviance(M2))
[1] 15622.5

#La diferencia de devianzas es de D2-D1=0.27
#Análisis de devianza

> anova(M1,M2,test="Chisq")
Analysis of Deviance Table
```

```
Model 1: wfa2 ~ foodexp + persroom
Model 2: wfa2 ~ foodexp + persroom + foodexp:persroom
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     17862      15623
2     17861      15622  1   0.2681   0.6046

#La devianza es reportada por el nombre de Residual Deviance por R
-----------------------------------------------------------------
```

El valor de la diferencia de devianzas entre los modelos anidados es de $D_{M2} - D_{M1} = 0.2681$. El *p-value* reportado por la función $anova()$ corresponde a la hipótesis nula $D_{M2} - D_{M1} = 0$. En este caso, dado que $p > 0.05$, no se rechaza H_0. Por lo tanto, se concluye que el desempeño del modelo M1 no es significativamente diferente al del modelo M2.

```
-----------------------------------------------------------------
#Criterio de información de Akaike
> AIC(M1)
[1] 15628.77

> AIC(M2)
[1] 15630.5

#Criterio de información de Bayes
> BIC(M1)
[1] 15652.14

> BIC(M2)
[1] 15661.67
-----------------------------------------------------------------
```

Los valores AIC y BIC del modelo con efectos principales $M1$ son menores que las del modelo con interacción $M2$, con diferencia de 1.73 unidades y 9.52 unidades, respectivamente. Por lo tanto, se tendría que el modelo $M1$ tiene un desempeño *ligeramente* superior al modelo $M2$. Sin embargo, debido a que las diferencias numérica de los valores del AIC y BIC de los modelos ajustados son muy pequeñas, cada una de éstas pueden no corresponder a una diferencia estadística, i.e. las diferencias pueden ser debidas a una variación muestral. En cambio, el resultado de la prueba de significancia $H_0 : D_{M1} - D_{M2} = 0$ *vs* $H_1 : D_{M1} - D_{M2} \neq 0$ provee una comparación más confiable entre los modelos $M1$ y $M2$. El resultado favorece al modelo con menos variables predictoras $M1$, mostrando que la interacción entre $foodexp$ y $persroom$ puede ser descartada del modelo $M2$.
La devianza y el AIC también son reportados por el resumen del modelo en R:

```
-----------------------------------------------------------------

> summary(M1)

Call:
glm(formula = wfa2 ~ foodexp + persroom, family = binomial, data = enal)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0422  -0.6385  -0.5643  -0.4636   2.7960

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.664179   0.068192 -24.404   <2e-16 ***
foodexp     -0.015817   0.001411 -11.213   <2e-16 ***
persroom     0.098769   0.010110   9.769   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


    Null deviance: 15974  on 17864  degrees of freedom  #Devianza nula = 15 974
Residual deviance: 15623  on 17862  degrees of freedom  #Devianza= 15 623
AIC: 15629  #AIC

Number of Fisher Scoring iterations: 57
```

Finalmente, se procede a realizar la prueba de significancia del modelo

$$H_0 : \beta_{foodexp} = 0 \;\; y \;\; \beta_{persroom} = 0 \qquad vs \qquad H_1 : \beta_{foodexp} \neq 0 \quad ó \quad \beta_{persroom} \neq 0$$

```
> Nulo=glm(wfa2~1, data=enal, family=binomial) #Modelo nulo
> anova(Nulo, M1, test="Chisq")
Analysis of Deviance Table

Model 1: wfa2 ~ 1
Model 2: wfa2 ~ foodexp + persroom
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     17864      15974
2     17862      15623  2   351.25 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Los resultados muestran que al menos una de las variables predictoras contribuye significativamente al modelo (p-*value*<0.05). Es decir, la hipótesis nula es rechazada.

2.10. Selección de modelo

La selección de modelo corresponde a un proceso de elección de variables predictoras con base en criterios estadísticos, con el objetivo de encontrar un subconjunto de variables explicativas con el mejor desempeño del ajuste. En esta sección se presentan los métodos *stepwise*,

los cuales consisten en la agregación ó remoción secuencial de variables explicativas en el modelo. De manera general, los métodos *stepwise*, coinciden en los siguientes dos aspectos: (1) la verificación del desempeño se desarrolla utilizando un criterio predeterminado, como el AIC, BIC o la devianza; (2) los criterios de selección *stepwise* consideran el principio de jerarquía: una interacción no pertenecerá al modelo si sus efectos principales o interacciones de orden inferior no pertencen al modelo.

2.10.1. *Backward stepwise selection*

El proceso es el siguiente: se inicia con todos las variables predictoras del modelo (modelo saturado). Después, de manera secuencial, varias de éstas son eliminadas del modelo utilizando el criterio de selección predeterminado. El método se detiene cuando la eliminación de cualquier variable no mejora significativamente el desempeño del modelo.

```
-----------------------------------------------------------------------
##Backward aplicado a moodelo con interacción de segundo orden
M2=glm(wfa2~ foodexp+persroom+foodexp:persroom,
        data=enal, family=binomial)

> Backward=step(M2,direction="backward")
Start:  AIC=15630.5
wfa2 ~ foodexp + persroom + foodexp:persroom

                   Df Deviance   AIC
- foodexp:persroom  1    15623 15629
<none>                   15622 15630

#Eliminar la interacción reduce en una unidad a la devianza y al AIC.

Step:  AIC=15628.77
wfa2 ~ foodexp + persroom

           Df Deviance   AIC
<none>           15623 15629
- persroom  1    15716 15720
- foodexp   1    15765 15769

#Si se elimina alguno de los efectos principales, el
#valor de la devianza y el AIC aumentarán.  Así, el modelo
#final es

> summary(backward)

Call:
glm(formula = wfa2 ~ foodexp + persroom, family = binomial, data = enal)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -1.664179   0.068192 -24.404   <2e-16 ***
foodexp     -0.015817   0.001411 -11.213   <2e-16 ***
persroom     0.098769   0.010110   9.769   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


    Null deviance: 15974  on 17864  degrees of freedom
Residual deviance: 15623  on 17862  degrees of freedom
AIC: 15629

Number of Fisher Scoring iterations: 5
```

2.10.2. *Forward stepwise selection*

Se inicia sin variables predictoras, sólo con una constante en el modelo (modelo Nulo). Posteriormente y de manera iterativa, se agregan las variables más significativas, esto considerando a las variables agregadas con anterioridad. El proceso se detiene cuando el desempeño del modelo no mejora significativamente de acuerdo al criterio establecido (por ejemplo: el *AIC* o el *BIC*).

```
#Forward selection:

M2=glm(wfa2~ foodexp+persroom+foodexp:persroom,
        data=enal, family=binomial)

> forward=step(M2, direction="forward")
Start:  AIC=15630.5
wfa2 ~ foodexp + persroom + foodexp:persroom

#No se eliminaron variables:

> summary(forward) #Resumen del modelo final

Call:
glm(formula = wfa2 ~ foodexp + persroom + foodexp:persroom,
family = binomial,  data = enal)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.6927     0.0876    -19.322  < 2e-16 ***
foodexp           -0.0143     0.0030     -4.648 3.35e-06 ***
persroom           0.1058     0.0169      6.258 3.90e-10 ***
foodexp:persroom  -0.0003     0.0007     -0.518    0.605
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
    Null deviance: 15974  on 17864  degrees of freedom
Residual deviance: 15623  on 17861  degrees of freedom
AIC: 15631

Number of Fisher Scoring iterations: 5
-------------------------------------------------------------
```

2.10.3. *Backward-forward selection*

Este proceso es una combinación de los procesos anteriores. Se inicia sin variables predictoras (modelo Nulo); consecuentemente, se agregan variables explicativas al modelo. Después de agregar una nueva variable, se elimina cualquier variable que no produzca una mejora en el desempeño del modelo.

```
 ------------------------------------------------------------------
 M2=glm(wfa2~ foodexp+persroom+foodexp:persroom,
        data=enal, family=binomial)

 > both=step(M2,direction="both")
Start:  AIC=15630.5
wfa2 ~ foodexp + persroom + foodexp:persroom

                   Df Deviance   AIC
- foodexp:persroom  1    15623 15629
<none>                   15622 15630

Step:  AIC=15628.77
wfa2 ~ foodexp + persroom     #Interacción eliminada

                   Df Deviance   AIC
<none>                   15623 15629
+ foodexp:persroom  1    15622 15630
- persroom          1    15716 15720
- foodexp           1    15765 15769

> summary(both)  #Resumen del modelo final

Call:
glm(formula = wfa2 ~ foodexp + persroom, family = binomial, data = enal)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.664179   0.068192 -24.404   <2e-16 ***
foodexp     -0.015817   0.001411 -11.213   <2e-16 ***
persroom     0.098769   0.010110   9.769   <2e-16 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



    Null deviance: 15974  on 17864  degrees of freedom
Residual deviance: 15623  on 17862  degrees of freedom
AIC: 15629

Number of Fisher Scoring iterations: 5
--------------------------------------------------------------
```

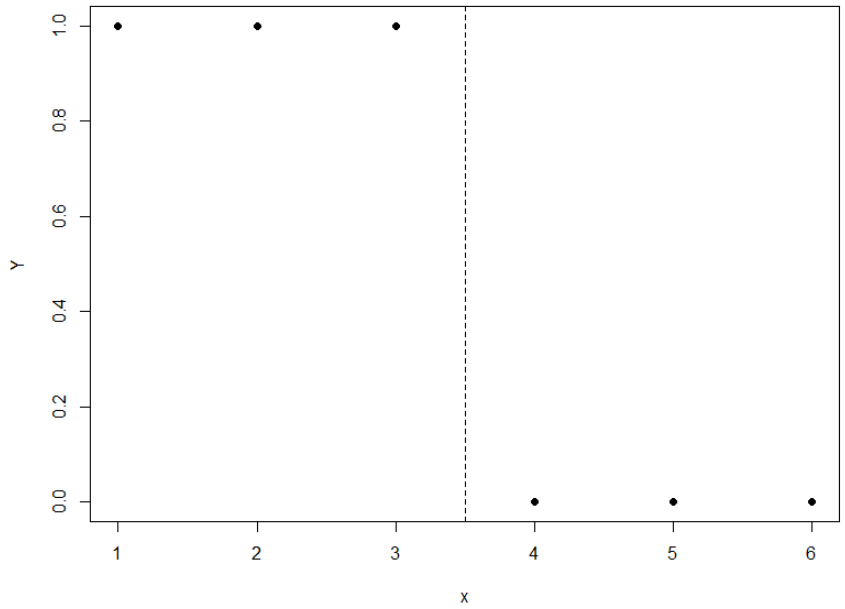
En los siguientes capítulos se referirá al modelo logístico tratado en este capitulo como el modelo logístico sin efectos aleatorios. Lo anterior con el objetivo de distinguirlo de los modelos que serán presentado posteriormente.

2.11. El problema de separación completa

El modelo de regresión logística es ampliamente utilizado por su simplicidad y el nivel de interpretabilidad con el que cuenta a partir de los coeficientes β_j. Sin embargo, cuando el número de observaciones N es muy pequeño o existen variables explicativas que permiten asignar correctamente a las observaciones (separándolas por regiones a través de un hiperplano), los estimadores máximo verosímiles no convergen a un valor finito. Este suceso es conocido como el *problema de separación completa*. Agresti (2015), muestra un ejemplo sencillo donde es posible notar esta situación (pp. 177–179): supongamos que $Y = 0$ para $x = 1, 2, 3$ y $Y = 1$ para $x = 4, 5, 6$. Podemos notar que $Y = 0$ si $x < 3.5$. Por otra parte si $x > 3.5$, $Y = 1$. Lo anterior muestra que las observaciones pueden separarse a través de la recta $x = 3.5$ (Figura 2.3). El modelo de regresión logística para estas observaciones resulta ser

$$logit(\pi_i) = \beta_0 + \beta_1 x_i$$

Figura 2.3. *Gráfica de ejemplo con separación completa, la recta $x = 3.5$ separa los datos en dos grupos distinguidos: $Y = 1$ si $x > 3.5$; $Y = 0$ cuando $x < 3.5$. Fuente: Agresti 2015*



Si tomamos $\pi = 0.5$ (*threshold*) para $x = 3.5$, entonces

$$\pi(3.5) = 0.5 = \frac{exp\{\beta_0 + \beta_1(3.5)\}}{1 + exp\{\beta_0 + \beta_1(3.5)\}} \Rightarrow 1 = exp\{\beta_0 + \beta_1(3.5)\} \Rightarrow 0 = \beta_0 + \beta_1(3.5)$$

De esta forma

$$\beta_0 = -\beta_1(3.5) \quad \text{para todo} \quad \beta_1$$

Así, si $\beta_1 > 0$, se tiene que

$$\begin{cases} \beta_0 + \beta_1 x_i > 0 \quad si \ x_i > 3.5 \\ \\ \beta_0 + \beta_1 x_i < 0 \quad si \ x_i < 3.5 \end{cases}$$

Recordando que

$$\pi(x_i) = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)} \ ;$$

se tiene el siguiente resultado

$$\lim_{\beta_1 \to \infty} \pi(x_i) = \begin{cases} 1 & si\ x_i > 3.5 \\ \\ 0 & si\ x_i < 3.5 \end{cases}$$

Debido a lo anterior, no existe un valor finito que maximice a la función de verosimilitud. Análogamente al caso anterior, cuando $\beta_1 < 0$, se tiene que $\lim_{\beta_1 \to -\infty} \pi(x_i) = 0$ si y sólo si $x_i >$3.5.

Cuando se presentan este tipo de casos, el programa estadístico R manda un mensaje de advertencia (a saber, *Warning message: glm.fit: fitted probabilities numerically 0 or 1 occurred*) al realizar el ajuste de regresión logística. Además, es posible identificar este problema debido a los valores de los errores estándar y *p-values*, los cuales resultan en valores estimados muy grandes.

```
-----------------------------------------------------------------------
#Separación completa en R
> x <- c(1,2,3,4,5,6); Y <- c(1,1,1,0,0,0)
> M <- glm(Y ~ x, family=binomial(link="logit"))
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred

> summary(M)

Call:
glm(formula = Y ~ x, family = binomial(link = "logit"))

Deviance Residuals:
         1           2           3           4           5           6
 2.110e-08   2.110e-08   1.052e-05  -1.052e-05  -2.110e-08  -2.110e-08

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    165.32  407521.43       0        1
x              -47.23  115264.41       0        1


   Null deviance: 8.3178e+00  on 5  degrees of freedom
Residual deviance: 2.2152e-10  on 4  degrees of freedom
AIC: 4

Number of Fisher Scoring iterations: 25
-----------------------------------------------------------------------
Fuente: Agresti 2015, p. 177.
```

Albert y Anderson (1984) prueban, de manera general (usando como variable respuesta a una variable $Y \sim multinomial(n, p_1, \ldots, p_q)$), que si un conjunto de datos presenta separación

completa, los estimadores máximo verosímiles no son finitos, y el valor máximo de la función de verosimilitud, como función de los parámetros β, es igual a uno, i.e.,

$$\max_{\boldsymbol{\beta}} f(\boldsymbol{\beta}; \boldsymbol{y}) = 1$$

En esta sección se muestra la demostración de este teorema para el caso particular de clasificación binaria, i.e. cuando $Y_i \sim bernoulli(\pi_i)$, $i = 1, \ldots, N$. De manera previa a la prueba y la definición del problema de separación completa, se introducen detalles técnicos y notación que serán de utilidad para las siguientes secciones.

1. Denotemos por E_j al conjunto de observaciones en la muestra tales que $y = j$ $(j = 0, 1)$. Así, la observación i en la muestra pertenece a E_1 si y sólo si el valor observado de la respuesta es igual a uno; i.e para $i = 1, \ldots, N$,

$$i \in E_1 \Leftrightarrow y = 1$$

2. Utilizando como umbral $\tau = 0.5$ en el proceso de clasificación, se tiene que la observación i se asigna al grupo cero $(\hat{y}_i = 0)$ si y sólo si

$$1 - \pi(\boldsymbol{x}_i) = \frac{1}{1 + exp(\boldsymbol{\beta}^T x_i)} \geq \frac{1}{2} \;\; \Leftrightarrow \;\; 1 + exp(\boldsymbol{\beta}^T x_i) \leq 2 \;\; \Leftrightarrow \;\; exp(\boldsymbol{\beta}^T x_i) \leq 1 \;\; \Leftrightarrow$$

$$\boldsymbol{\beta}^T x_i \leq 0$$

   Por consiguiente,

   a) $\hat{y}_i = 0 \Leftrightarrow \boldsymbol{\beta}^T x_i \leq 0$

   b) $\hat{y}_i = 1 \Leftrightarrow \boldsymbol{\beta}^T x_i > 0$

3. Como consecuencia de los puntos anteriores, una observación $i \in E_j$ es clasificada de manera correcta $(\hat{y}_i = j)$ si y sólo si $(-1)^j \boldsymbol{\beta}^T x_i < 0$ (con $j = 0, 1$), i.e, si se cumple la siguiente doble implicación

$$i \in E_j \Leftrightarrow (-1)^j \boldsymbol{\beta}^T x_i < 0$$

   Notemos que el caso en el que $(-1)^j \boldsymbol{\beta}^T x_i = 0$ no es incluido en la desigualdad anterior (para cada observación i en la muestra y para $j = 0, 1$). Esto es debido a que representa al caso en el que $\pi(\boldsymbol{x}_i) = \frac{1}{2} = 1 - \pi(\boldsymbol{x}_i)$. En este caso, no es claro el grupo al que debería ser asignada la observación i; usualmente, se decide que esta observación sea asignada al grupo 0 $(i \in E_0)$. Sin embargo, también es posible asignar a dicha observación en el grupo 1 si así se desea. En consecuencia, las estimaciones realizadas dependen de la toma de decisión considerada cuando $(-1)^j \boldsymbol{\beta}^T x_i = 0$ para alguna i cuando existen observaciones tales que $(-1)^j \boldsymbol{\beta}^T x_i = 0$.

4. Bajo los supuestos del modelo, la función de verosimilitud de la muestra (ecuación (2.6)) para el caso en el que $Y_i \sim bernoulli(\pi_i)$, con $i = 1, \ldots, N$, es

$$f(\boldsymbol{\beta};\boldsymbol{y}) = \prod_{i=1}^{N} \pi_i(\boldsymbol{x})^{y_i}[1-\pi_i(\boldsymbol{x}_i)]^{1-y_i}$$

Utilizando la notación introducida, podemos reescribir a la función de verosimilitud de la siguiente forma

$$\begin{aligned} f(\boldsymbol{\beta};\boldsymbol{y}) &= \left[\prod_{i\in E_0}(1-\pi_i(\boldsymbol{x}_i))\right]\left[\prod_{i\in E_1}\pi_i(\boldsymbol{x})\right] \\ &= \left[\prod_{i\in E_0}\left(\frac{1}{1+exp\left(\boldsymbol{\beta}^T\boldsymbol{x_i}\right)}\right)\right]\left[\prod_{i\in E_1}\left(\frac{exp\left(\boldsymbol{\beta}^T\boldsymbol{x_i}\right)}{1+exp\left(\boldsymbol{\beta}^T\boldsymbol{x_i}\right)}\right)\right]. \end{aligned}$$

Dado que, para $i = 1, \ldots, N$;

$$\frac{exp\left(\boldsymbol{\beta}^T\boldsymbol{x_i}\right)}{1+exp\left(\boldsymbol{\beta}^T\boldsymbol{x_i}\right)} = \frac{1}{1+exp(-\boldsymbol{\beta}^T\boldsymbol{x_i})},$$

se tiene que

$$\begin{aligned} f(\boldsymbol{\beta};\boldsymbol{y}) &= \left[\prod_{i\in E_0}\left(\frac{1}{1+exp\left(\boldsymbol{\beta}^T\boldsymbol{x_i}\right)}\right)\right]\left[\prod_{i\in E_1}\left(\frac{1}{1+exp\left(-\boldsymbol{\beta}^T\boldsymbol{x_i}\right)}\right)\right] \\ &= \prod_{j=0}^{1}\prod_{i\in E_j}\left[\frac{1}{1+exp\left[(-1)^j\boldsymbol{\beta}^T\boldsymbol{x_i}\right]}\right]. \end{aligned}$$

En consecuencia, la función de log-verosimilitud es

$$L(\boldsymbol{\beta}) = \sum_{j=0}^{1}\sum_{i\in E_j} log\left[\frac{1}{1+exp\left[(-1)^j\boldsymbol{\beta}^T\boldsymbol{x_i}\right]}\right].$$

Considerando los puntos anteriores se introduce la definición del problema de separación utilizando notación matemática.

Definición 2.1 Se dirá que un conjunto de datos presenta el problema de separación completa si existe un vector de parámetros $\boldsymbol{\beta}$ tal que para toda $j \in \{0, 1\}$ y para toda $i \in E_j$

$$(-1)^j \boldsymbol{\beta}^T \boldsymbol{x_i} > 0 \tag{2.14}$$

Es decir, cuando existe un vector $\boldsymbol{\beta}$ tal que cada observación es asignada correctamente a su grupo correspondiente. El siguiente teorema muestra que los estimadores máximo verosímiles $\hat{\beta}_1, \ldots, \hat{\beta}_p$ no convergen cuando se presenta dicho problema.

Teorema 1 (Albert y Anderson, 1984). *Si un conjunto de datos presenta separación completa, entonces los estimadores máximo verosímiles no son finitos y*

$$\max_{\boldsymbol{\beta}} f(\boldsymbol{\beta}; \boldsymbol{y}) = 1$$

Prueba. Dado que el problema de separación completa está presente en el conjunto de datos, existe $\boldsymbol{\beta}$ tal que satisface la ecuación (2.14). Así, para toda $k > 0$, $k\boldsymbol{\beta}$ también satisface (2.14), y

$$L(k\boldsymbol{\beta}) = \sum_{j=0}^{1} \sum_{i \in E_j} log\left(\frac{1}{1 + exp[(-1)^j k \boldsymbol{\beta}^T \boldsymbol{x_i}]}\right)$$

De donde, debido a que cada observación es clasificada de manera correcta (véase punto 3), se tiene que

$$(-1)^j k\boldsymbol{\beta}^T x_i = \begin{cases} k\boldsymbol{\beta}^T x_i < 0 & si\ j = 0 \\ \\ -k\boldsymbol{\beta}^T \boldsymbol{x_i} < 0 & si\ j = 1 \end{cases}$$

Así, para $j \in \{0, 1\}$,

$$\lim_{k \to \infty} exp[(-1)^j k \boldsymbol{\beta}^T x_i] = 0$$

Por lo tanto, como la función logaritmo es continua,

$$\begin{aligned} \lim_{k \to \infty} L(k\boldsymbol{\beta}) &= \sum_{j=0}^{1} \sum_{i \in E_j} log\left(\frac{1}{1 + \lim_{k \to \infty} exp[(-1)^j k \boldsymbol{\beta}^T x_i]}\right) \\ &= \sum_{j=0}^{1} \sum_{i \in E_j} log(1) = 0 \end{aligned} \tag{2.15}$$

Por otra parte, debido a que la función logaritmo es monótona creciente y

$$\left(\frac{1}{1 + exp[(-1)^j k \boldsymbol{\beta}^T x_i]}\right) \leq 1;$$

la función $L(k\boldsymbol{\beta})$ alcanza el valor máximo cuando

$$\left(\frac{1}{1 + exp[(-1)^j k \boldsymbol{\beta}^T x_i]}\right) = 1.$$

Lo cual, por la ecuación (2.15), se cumple cuando $k \to \infty$.

■

Debido a que los estimadores máximo verosímiles $\hat{\beta}_j$ no convergen, se ha discutido sobre qué se puede inferir de ellos. Como se mencionó antes, la separación completa puede deberse a que el número de observaciones N es muy pequeña. Por consiguiente, una solución podría ser aumentar el número de observaciones. Sin embargo, esto no siempre es posible. Una solución alternativa, con enfoque *Bayesiano*, es *la correción de Firth*. Ésta consiste en una penalización de los coeficientes para evitar que crezcan indefinidamente en valor absoluto (véase Firth, 1993).

Capítulo 3

Regresión logística con intersección aleatoria

En el modelo de regresión logística sin efectos aleatorios y, en general, en los modelos lineales generalizados (*GLM*) se supone que las observaciones de la variable respuesta pertenecientes a la muestra son estocásticamente independientes. Sin embargo, este supuesto no resulta válido para cierta clase de datos. Por ejemplo, en estudios en los que se analizan sujetos a través del tiempo, la evolución de cada individuo es caracterizada por medio de las observaciones anteriores de la respuesta y por condiciones o efectos específicos (e.g. clima, condiciones de salud, exposición a resinas, etc) a los que éste se encuentra expuesto. Así, debido a la similitud de condiciones de exposición, el conjunto de observaciones pertenecientes a un mismo individuo corresponden a un conjunto o grupo de variables correlacionadas.

De manera general, el supuesto de independencia resulta inválido cuando grupos de observaciones pueden ser distinguidas por factores comunes que los caracterizan y las distinguen de otros grupos: localidades, familias, tiempo, etc. En este capítulo revisamos al modelo de regresión logística con intersección aleatoria perteneciente a la familia de modelos conocida como *modelos lineales generalizados mixtos* (*GLMM*). Dicho modelo corresponde a un caso simple de los GLMM, en el cual se considera correlación de observaciones por grupos (*clusters*). En la sección 2.1 se presenta el modelo matemático del enfoque logístico. En las secciones 2.2 y 2.3 se abordan la interpretación y la estimación del modelo. Finalmente, en la sección 2.4 el modelo es ilustrado utilizando datos de la ENAL 96.

3.1. Regresión logística con intersección aleatoria

Denotemos por m al número de grupos (*clusters*) en la muestra. Además, sea d_i el número de observaciones en el grupo $i = 1, \ldots, m$. Para el individuo j perteneciente al grupo i, sean Y_{ij} la variable respuesta binaria y $\boldsymbol{x}_{ij} = (x_{ij0}, x_{ij1}, \ldots, x_{ijp})^T$ el vector de observaciones de las variables explicativas, donde $x_{ij0} = 1$. El modelo de *regresión logística con intersección aleatoria* se define como (Agresti 2015, p. 307)

$$logit\left[P(Y_{ij} = 1|u_i)\right] = u_i + \sum_{k=0}^{p} \beta_k x_{ijk} \tag{3.1}$$

donde $\{u_i\}$ es un conjunto de variables aleatorias independientes con distribución de probabilidad $N(0, \sigma^2)$, i.e. $u \sim N(0, \sigma^2)$. Estas variables son conocidas como *efectos aleatorios*

(*random effects*) y representan efectos no observados a los que cada grupo de observaciones estuvo en exposición. Por otro lado, los parámetros $\beta_0, \ldots, \beta_p$ son llamados efectos fijos y, al igual que en el modelo logístico sin efectos aleatorios (el cual fue visto en el capítulo anterior) representan la influencia de las variables explicativas en la respuesta. En el modelo logístico con intersección aleatoria se supone que las variables condicionadas $Y_{is}|u_i$ y $Y_{it}|u_i$ son independientes y tienen distribuciones binomiales, para $i = 1, \ldots, m$ y $s, t = 1, \ldots, d_i$; respondiendo a un modelo con respuesta binaria cuando $n_{ij} = 1$.

La correlación entre las variables respuesta pertenecientes a un mismo grupo se incluye en el modelo por medio de los efectos aleatorios (Agresti 2013, p. 494):

$$\begin{aligned} Cov(Y_{ij}, Y_{it}) &= E[Cov(Y_{ij}, Y_{it}|u_i)] + Cov(E[Y_{ij}|u_i], E[Y_{it}|u_i]) \\ &= 0 + Cov(E[Y_{ij}|u_i], E[Y_{it}|u_i]). \end{aligned}$$

Dado que $Y_{ij}|u_i \sim binomial(n_{ij}, \pi_{ij})$,

$$E[Y_{ij}|u_i] = n_{ij}\pi_{ij} = n_{ij}\left[\frac{exp(u_i + \sum_{k=0}^{p}\beta_k x_{ijk})}{1 + exp(u_i + \sum_{k=0}^{p}\beta_k x_{ijk})}\right] = \frac{n_{ij}}{1 + exp\{-(u_i + \sum_{k=0}^{p}\beta_k x_{ijk})\}}.$$

Por lo tanto, se tiene que

$$Cov(Y_{ij}, Y_{it}) =$$

$$Cov\left(\frac{n_{ij}}{1 + exp\{-(u_i + \sum_{k=0}^{p}\beta_k x_{ijk})\}}, \frac{n_{ik}}{1 + exp\left\{-(u_i + \sum_{k=0}^{p}\beta_k x_{itk})\right\}}\right).$$

Así, como las funciones de u_i involucradas en la igualdad anterior son monótonas crecientes, las variables Y_{ij} y Y_{it} se encuentran correlacionadas positivamente.

3.2. Interpretación del modelo

Considere la variable respuesta Y_{it} con vector de variables explicativas $\boldsymbol{x}_{it}$ y a la variable respuesta Y_{hs} con vector de variables explicativas $\boldsymbol{x}_{hs}$. El logaritmo de la razón de momios para estas dos observaciones es (Agresti 2013, p. 494)

$$\begin{aligned} logit[P(Y_{it} = 1|u_i)] - logit[P(Y_{hs} = 1|u_h)] &= \beta^T(x_{it} - x_{hs}) + (u_i - u_h) \\ &= \sum_{k=0}^{p}\beta_k(x_{itk} - x_{hsk}) + (u_i - u_h) \end{aligned} \tag{3.2}$$

De la ecuación anterior, cuando se comparan individuos pertenecientes a un mismo grupo ($i = h$) ó con el mismo efecto aleatorio, el valor de los efectos es anulado por el logaritmo de momios. Sin embargo, esto no ocurre cuando se comparan individuos con efectos aleatorios

distintos. En ese caso, considerando los supuestos $u_i, u_h \sim N(0, \sigma^2)$; se tiene que $(u_i - u_h) \sim N(0, 2\sigma^2)$. Por consiguiente, si $z_{\alpha/2}$ es el cuantil $(1-\alpha/2)$ de una distribución normal estándar, el $(1-\alpha) \times 100\,\%$ de los valores dados por la ecuación (3.2) se encontrarán entre

$$\sum_{k=0}^{p} \beta_k (x_{itk} - x_{hsk}) \pm z_{\alpha/2}\sqrt{2}\sigma \tag{3.3}$$

Por otro lado, el valor de $\sigma \geq 0$ representa el grado heterogeneidad de la muestra. Cuando $\sigma = 0$, el modelo (3.1) se simplifica al modelo de regresión logística sin efectos aleatorios, tratando a todas las observaciones como independientes. De esta forma, mientras mayor sea el valor de σ, mayor será la heterogeneidad entre grupos.

3.3. Estimación

La estimación de los parámetros, al igual que en el modelo de regresión logística sin efectos aleatorios, se realiza utilizando el método de máxima verosimilitud: sea $f(\boldsymbol{Y}|\boldsymbol{\mu};\boldsymbol{\beta})$ la función de densidad de probabilidad conjunta del vector de variables respuesta $\boldsymbol{Y}$ dado el vector de efectos aleatorios $\boldsymbol{u}$. Si $f(\boldsymbol{u};\sigma)$ denota a la función de densidad de probabilidad de $\boldsymbol{u}$, entonces la función de verosimilitud de la muestra $l(\boldsymbol{y};\boldsymbol{\beta},\sigma)$ es (Agresti 2013, p.519)

$$\begin{aligned}
l(\boldsymbol{y};\boldsymbol{\beta},\sigma) &= f(\boldsymbol{y};\boldsymbol{\beta},\sigma) \\
&= \int f(\boldsymbol{Y}|\boldsymbol{u};\boldsymbol{\beta}) f(\boldsymbol{u};\sigma) d\boldsymbol{u} \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\prod_{i=1}^{m} \prod_{j=1}^{n_i} f(Y_{ij}|u_i;\boldsymbol{\beta}) \right] \left[\prod_{i=1}^{m} f(u_i;\sigma^2) \right] du_1 \ldots du_m \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(Y_{ij}|u_i;\boldsymbol{\beta}) f(u_i;\sigma^2) du_i \\
&= \prod_{i=1}^{m} \left\{ \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f(Y_{ij}|u_i;\boldsymbol{\beta}) f(u_i;\sigma^2) du_i \right\}
\end{aligned}$$

Donde, para toda i, j;

$$f(Y_{ij}|u_i;\boldsymbol{\beta}) = \binom{n_{ij}}{y_{ij}} \left[\frac{exp(\beta^T x_{ij} + u_i)}{1 + exp(\beta^T x_{ij} + u_i)} \right]^{y_{ij}} \left[\frac{1}{1 + exp(\beta^T x_{ij} + u_i)} \right]^{n_{ij} - y_{ij}}$$

Los valores máximo-verosímiles de $\boldsymbol{\beta}$ y σ son estimados numéricamente debido a la complejidad de las soluciones de la función de verosimilitud.

3.4. Ejemplo: Ajuste de modelos con efectos aleatorios

En esta sección, se aplica el modelo con intersección aleatoria a las observaciones de la base de datos de la *Encuesta Nacional de Alimentación y Nutrición en el Medio Rural 1996*, considerando como variable respuesta al indicador de desnutrición *wfa2* y, como variables explicativas, a las variables *foodexp* y *persroom* utilizadas en el ejemplo del capítulo anterior.

En este caso se considera la agrupación de observaciones por comunidades rurales de México. Esto debido a que más de una observación fue tomada de una misma comunidad. Así, observaciones pertenecientes a una misma comunidad rural podrían no ser independientes debido a que podrían encontrarse en condiciones similares (e.g. condiciones sanitarias, alimentación, nivel socioeconómico) que las diferencian de otras comunidades. Los resultados del ajuste del modelo utilizando el programa R se presentan en los recuadros siguientes.

```
-------------------------------------------------------------------------
library(glmmML) #Göran Broström (2018), Generalized Linear Models
with Clustering.

enal$psu=as.factor(enal$psu) #para el ajuste, la variable respuesta
#debe transformarse a factor, en caso de que no lo esté.

#Ajuste
  M3=glmmML(wfa2 ~ foodexp + persroom, data=enal,  cluster = enal$psu,
            family=binomial)

> summary(M3)

Call:  glmmML(formula = wfa2 ~ foodexp + persroom,
              family = binomial, data = enal, cluster = enal$psu)


                coef se(coef)       z Pr(>|z|)
(Intercept) -1.89891 0.079894 -23.768 0.00e+00
foodexp     -0.01138 0.001557  -7.304 2.79e-13
persroom     0.08052 0.011026   7.302 2.83e-13

Scale parameter in mixing distribution:  0.7096 gaussian
Std. Error:                              0.03388

        LR p-value for H_0: sigma = 0:  1.807e-97

Residual deviance: 15190 on 17861 degrees of freedom  AIC: 15190

> M3$deviance
[1] 15185.18
-------------------------------------------------------------------------
```

El modelo con intersección aleatoria corresponde a

$$
\begin{aligned}
&logit[P(Y_{ij}=1|u_i)] = -1.8989 - 0.0114 \times foodexp \ + 0.0805 \times persroom + u_i \\
&u_i \sim N(0, 0.0339^2)
\end{aligned}
\tag{3.4}
$$

Los resultados indican que en cada comunidad rural, el momio se multiplica, en promedio, por $exp(-0.0114) = 0.989$ por cada unidad aumentada en el gasto semanal per cápita en alimentos cuando el número de personas por habitación permanece constante; mientras que éste se multiplica por $exp(0.0805) = 1.084$ por cada persona adicional por habitación cuando el gasto semanal per cápita permanece constante. Esto es, el momio disminuye, en promedio, 1.10 % por cada unidad invertida en el gasto semanal per cápita en alimentos y aumenta, en promedio, 8.4 % por cada unidad invertida en el gasto semanal per cápita en alimentos. Por otro lado, si se comparan dos familias de diferentes localidades o dos familias expuestas a condiciones diferentes, con respuestas Y_{it} y Y_{hs} $(i, h = 1, \ldots, m;\ t = 1, \ldots, d_i;\ s = 1, \ldots, d_h)$, tales que

$$foodexp_{it} = foodexp_{hs} + 1 \qquad \text{y} \qquad persroom_{it} = persroom_{hs},$$

se tendrá que en un 95 % de las veces (ecuación (3.3))

$$-0.0134 \leq logit[P(Y_{it} = 1|u_i)] - logit[P(Y_{hs} = 1|u_h)] \leq 0.1744$$

Lo anterior si en ambos hogares viven el mismo número de personas por habitación.
De igual manera que en el modelo de regresión sin efectos aleatorios (véase sección 2.6), un aumento en el gasto semanal en alimentos disminuye la probabilidad de desnutrición. En cambio, un aumento de personas por habitación implica un aumento en dicha probabilidad. La Tabla 3.1 muestra los coeficientes y sus respectivas desviaciones estándar estimados por los modelos de regresión logística sin efectos aleatorios y con efectos aleatorios. Las desviaciones estándar resultan ligeramente mayores en el caso del intersección aleatoria, implicando una amplitud mayor de sus intervalos de confianza.

Tabla 3.1. Coeficientes estimados y errores estándar obtenidos por el modelo de Regresión logística sin efectos aleatorios y el modelo de Regresión logística con intersección aleatoria (con efectos aleatorios).

	$\hat{\beta}$	$\hat{se}(\hat{\beta})$		$\hat{\beta}$	$\hat{se}(\hat{\beta})$
Sin efectos			**Con efectos**		
Intercept	-1.6642	0.0682	*Intercept*	-1.8989	0.0799
foodexp	-0.0158	0.0014	*foodexp*	-0.0114	0.0015
persroom	0.0988	0.0101	*persroom*	0.0805	0.0110

La salida de *R* incluye una prueba de significancia para $\sigma = 0.0339$ utilizando el método de Razón de verosimilitudes. Esta prueba corresponde al contraste de hipótesis $H_0 : \sigma = 0$ *vs* $H_1 : \sigma \neq 0$. El resultado indica que σ es significativamente diferente de cero ($p < 0.001$). Por consiguiente, existe evidencia estadística de que las observaciones de la respuesta se encuentran correlacionadas en cada comunidad rural (*cluster*). Sin embargo, dado que el valor estimado de σ es cercano a cero ($\sigma < 1$), la variación de los efectos u_i no es muy grande. En consecuencia, la probabilidad de desnutrición no es muy diferente entre comunidades. Por otro lado, dado que μ_i se supone con distribución $N(0, 0.0339^2)$, se espera que el 95 % de sus valores se encuentren entre $-1.96 \times 0.0339 = -0.0664$ y $1.96 \times 0.0339 = 0.0664$.

El resumen del modelo con interacción se presenta en el siguiente recuadro:

```
-----------------------------------------------------------------
> M4=glmmML(wfa2 ~ foodexp + persroom + foodexp:persroom,
         data=enal, cluster=enal$psu, family=binomial)
> summary(M4)

Call:  glmmML(formula = wfa2 ~ foodexp + persroom + foodexp:persroom,
        family = binomial, data = enal, cluster = enal$psu)


                      coef  se(coef)        z Pr(>|z|)
(Intercept)      -1.9382078 0.0967612 -20.0308 0.00e+00
foodexp          -0.0094023 0.0031532  -2.9818 2.87e-03
persroom          0.0905420 0.0177979   5.0872 3.63e-07
foodexp:persroom -0.0005457 0.0007625  -0.7157 4.74e-01

Scale parameter in mixing distribution:  0.7099 gaussian
Std. Error:                              0.03388

        LR p-value for H_0: sigma = 0:  1.599e-97

Residual deviance: 15180 on 17860 degrees of freedom  AIC: 15190

> M4$deviance
[1] 15184.67

> M4$aic
[1] 15194.67
-----------------------------------------------------------------
```

El modelo con interacción presenta una reducción de la devianza de 0.5116 unidades. No ocurre lo mismo con el AIC, siendo este menor para el modelo con efectos principales con diferencia de 1.4883 unidades. A pesar de lo anterior, la interacción entre las variables $foodexp$ y $persroom$ no resulta significativa (p= 0.474). Por lo tanto, ésta puede ser eliminada del modelo. Procedemos ahora con la prueba de significancia del modelo. Sean D_{Nulo} y D_{M3} las devianzas de los modelos *Nulo* y $M3$. Se desea probar

$$H_0 : D_{Nulo} - D_{M3} = 0 \qquad vs \qquad H_1 : D_{Nulo} - D_{M3} \neq 0$$

A diferencia de los modelos $glm()$ (ejemplo anterior), la función *anova()* no es aplicable a modelos *glmmML()* en R. Sin embargo, es posible calcular el *p-value* de la prueba recordando que, bajo la hipótesis nula, $(D_{Nulo} - D_{M3}) \sim \chi^2_{(2-0)}$.

```
-----------------------------------------------------------------------
> Nulo=glmmML(wfa2~1, data=enal, cluster=enal$psu,
family=binomial) #Modelo nulo

#Diferencia de devianzas
> deviance(Nulo)-deviance(M3)
```

```
[1] 150.2628

#P value
> pchisq(150.2628, df=2, lower.tail=FALSE)
[1] 2.348808e-33 #<0.05
```

Dado que $p<0.05$, se rechaza la hipótesis nula. Por lo tanto, el modelo es validado: al menos una de las variables es significativamente diferente de cero.

Para extraer los valores del BIC de los modelos ajustados (con la función *glmmML*), es necesario instalar el paquete *MuMIn* en R. El proceso para la extracción del BIC se muestra en el siguiente recuadro.

```
> library(MuMIn) #Kamil Barton (2018). MuMIn: Multi-Model Inference.

> BIC(M3)
[1] 15224.34

> BIC(M4)
[1] 15233.62
```

Al igual que el AIC, el BIC resulta mayor para el modelo con interacción. En el siguiente capítulo se presenta un modelo para analizar encuestas complejas. Éste permite la inclusión de una estructura de correlación más profunda ya que permite la inclusión de estratos y *clusters* como herramientas de división de la población.

Capítulo 4

Regresión logística para datos de encuestas complejas

Los modelos revisados en los capítulos anteriores y, en general, en gran parte de los métodos y modelos estadísticos, se supone que las observaciones pertenecientes a la muestra son seleccionadas aleatoriamente con la misma probabilidad, utilizando muestreo aleatorio simple. No obstante, existen casos en los que este tipo de selección no resulta la más adecuada para realizar inferencias estadísticas.

Las encuestas complejas se caracterizan por el proceso de selección de muestra (diseño muestral), el cual permite mantener una estructura de correlación entre los individuos de interés (población objetivo) por medio de la división de la población por grupos y estratos. En consecuencia, los métodos de inferencia utilizados para estas encuestas son diferentes a los métodos usuales, basándose en el diseño de encuesta desarrollado, a priori, por el investigador. En este capítulo se presenta el *modelo de regresión ponderada*, el cual permite realizar ajustes logísticos para encuestas complejas considerando la división de la población objetivo en estratos y grupos (*clusters*). Para las siguientes seccciones, denotemos por N al número de observaciones en la muestra, H al número de estratos considerados y M al número de grupos (*clusters*) en la muestra.

4.1. Modelo y estimación

El modelo logístico para datos de encuestas complejas o modelo logístico ponderado por diseño, corresponde al modelo de regresión logística sin efectos aleatorios visto en el Capítulo 2:

$$logit[\pi(\boldsymbol{x})] = log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \sum_{j=0}^{p} \beta_j x_j$$

En consecuencia, su interpretación es la misma que dicho modelo. Sin embargo, debido al diseño de encuesta, el enfoque de máxima verosimilitud para la estimación de los parámetros del modelo no es adecuado por ciertas razones: (1) las probabilidades de selección de las observaciones de la muestra no suelen ser iguales para cada individuo, y (2) la estratificación y agrupamiento (*clustering*) de observaciones de encuestas complejas infringen la hipótesis de independencia de las observaciones (Heeringa et al. 2017). Un enfoque desarrollado para la estimación de coeficientes es conocido como *estimación de pseudo máxima verosimilitud*

(*PMLE* por sus siglas en inglés): sea w_i el factor de expansión (peso muestral) de la observación i (con $i = 1, \ldots, N$). Los coeficientes son estimados maximizando la *función de pseudo verosimilitud ponderada*, definida como (Heeringa et al. 2017, p. 265)

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{N} \left\{ \pi(\boldsymbol{x}_i)^{y_i} [1 - \pi(\boldsymbol{x}_i)^{y_i}]^{1-y_i} \right\}^{w_i}$$

$$\text{con } \pi(\boldsymbol{x}) = \frac{exp(\sum_{j=0}^{p} \beta_j x_j)}{1 + exp(\sum_{j=0}^{p} \beta_j x_j)}$$

La solución de los coeficientes del modelo requiere la solución del siguiente vector de ecuaciones (Heeringa et al 2017, p. 266)

$$S(\boldsymbol{\beta}) = \sum_h \sum_m \sum_i w_{hmi} D_{hmi}^T [\pi_{hmi}(\boldsymbol{\beta})(1 - \pi_{hmi}(\boldsymbol{\beta})]^{-1} (y_{hmi} - \pi_{hmi}(\boldsymbol{\beta})) = 0, \qquad (4.1)$$

donde h refiere al estrato ($h = 1, \ldots, H$), m indica el grupo ($a = 1, \ldots, M$) en el estrato h, e i es un índice para las observaciones individuales ($i = 1, \ldots, N$). D_{hmi} es el vector de derivadas parciales $\partial(\pi_{hmi}(\boldsymbol{\beta}))/\partial\beta_j$ y el término w_{hmi} denota al peso muestral de la observación i ($i = 1, \ldots, N$). La solucion de la ecuación (4.1) es encontrada utilizando métodos numéricos.

Por otra parte, la matríz de covarianzas de los coeficientes β_i es calculada utilizando el estimador

$$\hat{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{J}^{-1})\, Var[S(\hat{\boldsymbol{\beta}})](\boldsymbol{J}^{-1}), \qquad (4.2)$$

donde $\boldsymbol{J}$ es la matríz de segundas derivadas con respecto a $\hat{\beta}_j$ de la función de log-pseudo verosimilitud de los datos y $S(\hat{\boldsymbol{\beta}})$ es el vector $S(\boldsymbol{\beta})$ evaluado en $\hat{\beta}$= *arg máx* $PL(\boldsymbol{\beta})$.

4.2. Ejemplo: Ajuste e interpretación del modelo

En esta sección, se ilustra el modelo logístico para datos de encuestas complejas, considerando, esta vez, el diseño de la encuesta *ENAL* 96. De igual forma que en los ejemplos anteriores, la variable dependiente es un indicador de desnutrición (*wfa2*), y las variables predictivas son el gasto semanal per cápita en alimentos ($foodexp$) y el número de personas por habitación en casa (*persroom*). La selección de individuos fue realizada considerando estratos y agrupamientos de familias por comunidades rurales (véase Capítulo 1). El factor de expansión y el estrato correspondiente a cada individuo se encuentran registrados en los vectores *estratob* y *f3bm18774b*, respectivamente. Se procede, primero por la especificación del modelo.

```
-------------------------------------------------------------------------
library(survey)  T. Lumley (2017) "survey: analysis of complex
survey samples"

#Especificación del diseño muestral
> (diseño=svydesign(id=~psu, strata = ~estratob,
```

```
                  weights = ~f3bm18774b  ,data=enal))

Stratified 1 - level Cluster Sampling design (with replacement)
With (853) clusters.
svydesign(id = ~psu, strata = ~estratob, weights = ~f3bm18774b,
    data = enal)
------------------------------------------------------------------------
```

Para el análisis de encuestas complejas se supone que los parámetros a estimar (e.g. media, proporciones, tamaño de la población, etc) son cantidades desconocidas, pero fijas, las cuales caracterizan a la población de interés. Al especificar el diseño de encuesta (*svydesign*) es necesario indicar el tamaño de la población, o, en su defecto, el vector de pesos muestrales (factores de expansión). En el recuadro anterior, la falta del tamaño de la población es indicada por "(with replacement)" en la salida de R (Lumley 2010, p. 21). En el siguiente recuadro, se estiman las medias de las variables *foodexp* y *persroom* considerando el diseño muestral. Para el cálculo de estas, es importante mencionar que en cada estrato considerado debe haber al menos 2 grupos o *clusters*.

```
------------------------------------------------------------------------
#Medicas estimadas
> svymean(~enal$foodexp,diseño) # Gasto semanal per cáp. en alimentos.
               mean     SE
enal$foodexp 26.594 0.4734
> svymean(~enal$persroom,diseño) # Personas por habitación
               mean     SE
enal$persroom 4.402 0.0316

#Intervalos de confianza para las medias
> confint(svymean(~enal$foodexp,diseño),df=degf(diseño))
               2.5 %   97.5 %
enal$foodexp 25.6635 27.52399
> confint(svymean(~enal$persroom,diseño),df=degf(diseño))
                 2.5 %   97.5 %
enal$persroom 4.339864 4.464042
------------------------------------------------------------------------
```

El procedimiento del ajuste logístico en R se muestra en el siguiente recuadro. Es importante determinar primero el diseño muestral.

```
------------------------------------------------------------------------
#Ajuste logístico

> svy1=svyglm(wfa2~foodexp+persroom,
                design=diseño, data= enal, family = quasibinomial)
> summary(svy1)

Call:
svyglm(formula = wfa2 ~ foodexp + persroom, design = diseño, data = enal,
```

```
    family = quasibinomial)

Survey design:
svydesign(id = ~psu, strata = ~estratob, weights = ~f3bm18774b,
    data = enal)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.586039   0.104230 -15.217  < 2e-16 ***
foodexp     -0.016221   0.002749  -5.901 6.82e-09 ***
persroom     0.089269   0.012890   6.925 1.40e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of Fisher Scoring iterations: 4

> svy1$deviance
[1] 16036.68

> AIC(svy1)
       eff.p          AIC     deltabar
    5.633236 16047.944989     2.816618
```

Podemos notar que los efectos de las variables explicativas sobre la variable respuesta son significativamente diferentes de cero ($p< 0.001$). El modelo indica que el momio es multiplicado, en promedio, por $exp(-0.0162) = 0.9839$ por cada unidad aumentada en el gasto semanal per cápita en alimentos, esto es, el momio disminuye 1.61 % por cada unidad invertida en dicho gasto si el número de personas por habitación es fijo. Por otro lado, el momio es multiplicado, en promedio, por $exp(0.0893) = 1.0934$ por cada persona agregada por habitación en casa, es decir, el momio se incrementa 9.34 % por cada persona extra por habitación siempre que el gasto semanal per capita es constante. Los intervalos del 95 % de confianza para los coeficientes estimados utilizando el estadístico de Wald están dados por

```
> print(confint(svy1,type = "Wald",level=0.95),digits=3)
              2.5 %  97.5 %
(Intercept) -1.7903 -1.3818
foodexp     -0.0216 -0.0108
persroom     0.0640  0.1145
```

Los valores estimados de las razones de momios $exp(\hat{\beta})$ se encuentran en la Tabla 4.1. Los resultados muestran que el momio de presentar desnutrición disminuye por cada unidad aumentada en el gasto semanal per cápita si se mantiene fijo el número de personas por habitación. Por otro lado, el momio aumenta por cada persona adicional por habitación. De manera general, los efectos de las variables *foodexp* y *persroom* coinciden con los encontrados en el modelo sin efectos aleatorios y el modelo con intersección aleatoria: un aumento en el

gasto semanal per cápita implica una disminución de la probabilidad de desnutrición $\pi(\boldsymbol{x})$, mientras que un aumento en el número de personas por habitación implica un aumento de dicha probabilidad.

Tabla 4.1. Razones de momios estimados para cada variable predictiva en el modelo logístico ponderado por diseño.

Var. predictiva	$exp(\hat{\beta})$	**IC** 95 % **para** $exp(\beta)$
foodexp	0.984	(0.982, 0.987)
persroom	1.104	(1.0821, 1.126)

La devianza y el AIC reportados son mayores que los reportados en el modelo con intersección aleatoria con diferencia de 852.01 y 853.27, respectivamente. Sin embargo, el cálculo del AIC es diferente al del modelo sin efectos aleatorios y al del modelo logístico con intersección aletoria. A diferencia del AIC, el BIC pierde su escencia al transladarse al modelo ponderado por diseño.

> Es posible construir un análogo natural del BIC con un fin más limitado, el cual consiste en seleccionar entre submodelos de un modelo de regresión dado (Lumley y Scott 2017, p. 7).

Los detalles técnicos de la construcción del AIC y el BIC para el modelo ponderado pueden consultarse en Lumley y Scott (2015).

De manera análoga a los ejemplos anteriores, la interacción entre las variables explicativas fue agregada al modelo ponderado ajustado. La interacción no resultó significativa (p= 0.6617), por lo que puede ser eliminada del ajuste. La prueba de significancia del modelo sin interacción (*svy1*) indica que al menos uno de las variables predictoras es significativamente diferente de cero (p<0.05). El procedimiento realizado se muestra en el siguiente recuadro.

```
-------------------------------------------------------------------------
#Modelo nulo
> Nulo=svyglm(wfa2~1,
+             design=diseño, data=enal, family = quasibinomial)

> anova(Nulo, svy1, test="Chisq")
Working 2logLR =  123.6177 p= < 2.22e-16

-------------------------------------------------------------------------
```

En el capítulo siguiente, se realiza la aplicación de los modelos presentados, utilizando los datos descritos en el capítulo 1 (ENAL 96). En las primeras secciones, se muestran y describen los modelos ajustados, y, posteriormente, se presenta la comparación de modelos con base en los resultados obtenidos.

Capítulo 5

Regresión logística para la estimación de la probabilidad de presentar desnutrición en niños de edad preescolar en zonas rurales de México 1996

5.1. Introducción

Los objetivos de este capítulo son ajustar y comparar los tres enfoques de regresión logística revisados en los capítulos anteriores, utilizando datos provenientes de la ENAL 96 (presentados en el capítulo 1). En el análisis realizado, se pretende describir los efectos que tienen 8 variables explicativas, correspondientes a aspectos socioeconómicas de la familia y algunas características de la madre, sobre la probabilidad de desnutrición en niños menores de cinco años de edad. Para hacer posible el contraste de ajustes, los tres enfoques de regresión han sido aplicados al mismo conjunto de variables explicativas. La selección de variables predictoras fue realizada en el modelo de regresión logística sin efectos aleatorios, considerando a todas las interacciones de segundo orden. El proceso iterativo de selección se llevó a cabo utilizando los pasos siguientes: (1) aplicación de los tres métodos *stepwise* al modelo logístico sin efectos aleatorios elegido, (2) selección del modelo resultante más parsimonioso, (3) remoción de variables no significativas, $\alpha = 0.05$, considerando el principio de jerarquía y (4) repetición de los pasos anteriores hasta que todas las variables seleccionadas o interacciones de orden superior sean significativas.

5.2. Regresión logística bajo el supuesto de independencia

Las estimaciones de los parámetros del modelo logístico sin efectos aleatorios elegido se muestran en la Tabla 5.1. Podemos notar que todos los efectos principales son significativos, con excepción de *foodexp* ($p = 0.868$) y *schoolm01* ($p = 0.337$). Sin embargo, la interacción *foodexp:schoolm01* resulta significativamente diferente de cero ($p<0.001$).

Tabla 5.1. Modelo de regresión logística sin efectos aleatorios estimado para el riesgo de desnutrición en niños de edad preescolar en zonas rurales de México. A pesar de que las variables *foodexp* y *schoolm01* no son significativas, éstas se incluyen en el modelo debido a que su interacción es significativa.

Var. predictora	**Categoría**	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	$\hat{se}(\hat{\beta})/\hat{\beta}$	z	p
Intercept	Constante	-0.5804	0.1218	-0.2099	-4.8	<0.001
foodexp	Continuo	0.0004	0.0023	6.0222	0.2	0.868
persroom	Conteo	0.0561	0.0108	0.1928	5.2	<0.001
agemoth	Continuo	-0.0101	0.0031	-0.3065	-3.3	0.001
floor	1	-0.2061	0.0465	-0.2255	-4.4	<0.001
wc01	1	-0.2320	0.0433	-0.1865	-5.4	<0.001
cooker01	1	-0.3717	0.0546	-0.1470	-6.8	<0.001
schoolm01	1	-0.0727	0.0756	-1.0406	-1.0	0.337
languagem01	1	-0.5258	0.0482	-0.0917	-10.9	<0.001
foodexp:schoolm01	Continuo × 1	-0.0110	0.0027	-0.2499	-4.0	<0.001

La interacción *foodexp:schoolm01* indica que el efecto que tiene el gasto semanal per cápita en alimentos (*foodexp*) sobre la probabilidad de desnutrición, varía en función de la escolaridad de la madre *schoolm01*. Pero, ¿cómo interpretar el modelo logístico si los efectos principales de las variables involucradas en la interacción no resultan significativas? Para responder esta pregunta, es importante recordar que cuando existe una interacción entre dos variables, éstas se encuentran relacionadas de manera que no es posible interpretar una de ellas sin considerar a la otra. En este caso, el coeficiente de *foodexp* (=0.0004) es el efecto de *foodexp* cuando *schoolm01* =0. Análogamente, el coeficiente de *schoolm01* (=-0.0727) es el efecto de *schoolm01* cuando *foodexp* =0. Así, el hecho de que *foodexp* no resulte significativo, en realidad quiere decir que *foodexp* no tiene un efecto significativo en la respuesta cuando *schoolm01* =0. Por lo tanto, para una madre sin estudios o con primaria inconclusa (*schoolm01=0*), la probabilidad de que uno de sus hijos padezca desnutrición está dada por

$$\begin{aligned} log\left(\frac{\pi}{1-\pi}\right) &= -0.5804 + 0.0004 \times foodexp + 0.0561 \times persroom - 0.0101 \times agemoth \\ &\quad - 0.2061 \times floor01 - 0.2320 \times wc01 - 0.3717 \times cooker01 - 0.0727 \times schoolm01 \\ &\quad - 0.5258 \times languagem01 - 0.0110 \times (foodexp \times schoolm01) \end{aligned}$$

$$\begin{aligned} &= -0.5804 + 0.0561 \times persroom - 0.0101 \times agemoth - 0.2061 \times floor01 - 0.2320 \times wc01 \\ &\quad - 0.3717 \times cooker01 - 0.5258 \times languagem01 \end{aligned}$$

El modelo anterior indica que el gasto semanal per cápita en alimentos no tiene un efecto significativo en la probabilidad de desnutrición en los infantes cuando la madre de familia no tiene estudios o no terminó la primaria.

Por otro lado, los resultados muestran que la interacción *foodexp:schoolm01* es significativa en el modelo. Por consiguiente, puede concluirse que *foodexp* tiene un efecto para otros valores de *schoolm01* ($\neq 0$). Dado que *schoolm01* es una variable binaria (con posibles valores 0 y 1), se concluye que *foodexp* tiene un efecto significativo cuando *schoolm01=1*. En consecuencia, para una madre con al menos primaria concluida (*schoolm01=1*), la probabilidad de que uno de sus hijos padezca desnutrición está dada por

$$\begin{aligned} log\left(\frac{\pi}{1-\pi}\right) &= -0.5804 + 0.0004 \times foodexp + 0.0561 \times persroom - 0.0101 \times agemoth \\ &\quad - 0.2061 \times floor01 - 0.2320 \times wc01 - 0.3717 \times cooker01 \\ &\quad - (0.0727 \times 1) - 0.5258 \times languagem01 - 0.0110 \times (foodexp \times 1) \\ &= -0.6531 - 0.0106 \times foodexp + 0.0561 \times persroom - 0.0101 \times agemoth \\ &\quad - 0.2061 \times floor01 - 0.2320 \times wc01 - 0.3717 \times cooker01 - 0.5258 \times languagem01 \end{aligned}$$

En contraste con el caso anterior, el modelo muestra que para las madres de familia con al menos primaria concluida, el gasto semanal per cápita en alimentos tiene un efecto significativo en la probabilidad de desnutrición en los infantes. El momio de padecer desnutrición se multiplica por $exp(\hat{\beta}) = 0.9894$ por cada unidad invertida en *foodexp* siempre que el resto de variables permanezcan fijas, i.e, el momio disminuye 1.06 % por cada unidad aumentada en el gasto semanal per cápita en alimentos.

Dado que los coeficientes estimados de las demás variables explicativas no varían en función de la escolaridad de la madre (*schoolm01*), sus respectivas interpretaciones son las mismas para cada caso. Los valores ajustados de las razones de momios $exp(\hat{\beta})$ de estas variables se presentan en la Tabla 5.2.

Tabla 5.2. Razones de momios estimados para las variables predictoras del modelo logístico sin efectos aleatorios.

Var. predictora	**Categoría**	$exp(\hat{\beta})$	**IC** 95 % **para** $exp(\beta)$
foodexp	Continuo	1.0004	(0.9960, 1.0048)
persroom	Conteo	1.0577	(1.0355, 1.0804)
agemoth	Continuo	0.9900	(0.9840, 0.9960)
floor01	1	0.8137	(0.7429, 0.8914)
wc01	1	0.7929	(0.7285, 0.8631)
cooker01	1	0.6895	(0.6195, 0.7675)
schoolm01	1	0.9299	(0.8018, 1.0785)
languagem01	1	0.5911	(0.5378, 0.6496)
foodexp:schoolm01	Continuo × 1	0.9890	(0.9838, 0.9944)

Con base en la información resumida en la tabla anterior, se deduce que las madres mayores de edad tienen una menor probabilidad de que sus hijos padezcan desnutrición

frente a las más jóvenes, reduciendo el momio 1.0 % por cada año de edad. Además, para los niños cuyos hogares disponen de piso de algún material, el momio de padecer desnutrición es 18.63 % menor que el momio para aquellos cuyas casas no tienen piso de algún tipo. Respecto a la disponibilidad de baño, los resultados indican que el momio de padecer desnutrición es 20.71 % menor para los niños que cuentan con este en sus hogares. Por otro lado, el momio es 31.05 % menor para los niños que disponen de estufa de gas. Finalmente, el modelo muestra que el momio de padecer desnutrición es 40.89 % mayor para las madres de familia que hablan alguna lengua indígena frente a aquellas que sólo hablan español.

Tabla 5.3. Tabla de confusión del modelo ajustado de Regresión logística sin efectos aleatorios. Error aparente = $\frac{2939}{17865} = 0.1645$

	Valores observados	
Valores ajustados	*Normal*	*Desnutrido*
Normal	14926	2939
Desnutrido	0	0

La Tabla 5.8 muestra la tabla de confusión del modelo ajustado. Es importante notar que, a pesar de que el modelo clasifica de manera errónea a todas las observaciones con desnutrición, el error aparente es muy bajo (proporción de error = 16.45 %). Esto ocurre porque el modelo asigna a todas las observaciones como "normales", siendo el número de estas (aproximademente cuatro veces) mayor que el número de observaciones con desnutrición. Este ejemplo muestra que se deben tomar precauciones en la interpretación del error aparente, ya que no resulta preciso y puede llevar a conclusiones erróneas. Respecto al desempeño del modelo, puede concluirse que el modelo ajustado carece de poder predictivo.

5.3. Regresión logística cuando el supuesto de independencia resulta inválido

5.3.1. Regresión logística con intersección aleatoria

El ajuste del modelo logístico con intersección aleatoria realizado sólo considera la división de la población rural por comunidades (*clusters*). Los resultados de las estimaciones se presentan en la Tabla 5.4

Tabla 5.4. Modelo de regresión logística con intersección aleatoria ajustado para el riesgo de desnutrición en niños de edad preescolar en zonas rurales de México 1996. Nota: $\hat{se}(\hat{\boldsymbol{\beta}}) = \sqrt{\hat{V}(\hat{\boldsymbol{\beta}})}$.

Var. predictora	**Categoría**	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	$\hat{se}(\hat{\beta})/\hat{\beta}$	z	p
Intercept	Constante	-0.7396	0.1389	-0.1878	-5.32	<0.001
foodexp	Continuo	0.0003	0.0025	8.8624	0.11	0.910
persroom	Conteo	0.0536	0.0114	0.2128	4.70	<0.001
agemoth	Continuo	-0.0095	0.0032	-0.3388	-2.95	0.003
floor01	1	-0.2074	0.0508	-0.2450	-4.08	<0.001
wc01	1	-0.1825	0.0499	-0.2737	-3.65	<0.001
cooker01	1	-0.3607	0.0597	-0.1654	-6.04	<0.001
schoolm01	1	-0.0723	0.0811	-1.1213	-0.89	0.372
languagem01	1	-0.5382	0.0655	-0.1219	-8.20	<0.001
foodexp:schoolm01	Continuo×1	-0.0096	0.0029	-0.3004	-3.33	<0.001

Al igual que con el modelo sin efectos aleatorios, los efectos principales son significativos (p<0.05) con excepción de *foodexp* y *schoolm01*. La variable *agemoth* es menos significativa que en el modelo sin efectos aleatorio ($p = 0.001$ para el modelo sin efectos y $p = 0.003$ para el modelo con intersección aleatoria). De nuevo, la interacción de segundo orden *foodexp:schoolm01* es también significativa, e indica que el gasto semanal per cápita en alimentos cambia de acuerdo a la escolaridad de la madre: (1) si la madre de familia no tiene estudios o no terminó la primaria (*schoolm01=0*), el gasto semanal per cápita en alimentos (*foodexp*) no resulta significativo para el modelo. En este caso, la probabilidad desnutrición está dada por

$$log\left(\frac{\pi}{1-\pi}\right) = -0.0740 + 0.0540 \times persroom - 0.0095 \times agemoth - 0.2074 \times floor01 - 0.1825 \times wc01 - 0.3606 \times cooker01 - 0.5381 \times languagem01 + a_i$$

$$a_i \sim N(0, 0.0326^2)$$

(2) Si la madre terminó al menos la primaria ($schoolm01 = 1$), la probabilidad de que uno de sus hijos padezca desnutrición está dada por

$$log\left(\frac{\pi}{1-\pi}\right) = -0.8124 - 0.0093 \times foodexp + 0.0540 \times persroom - 0.0095 \times agemoth - 0.2074 \times floor01 - 0.1825 \times wc01 - 0.3606 \times cooker01 - 0.5381 \times languagem01 + a_i$$

$$a_i \sim N(0, 0.0326^2)$$

En ambos modelos y en cada comunidad, la edad de la madre es un factor que disminuye la probabilidad de desnutrición en los hijos mientras mayor sea dicha edad. Además, la disponibilidad de piso, baño y estufa de gas reducen la probabilidad de desnutrición en niños pertenecientes a estas familias. Igualmente, el riesgo de desnutrición en los niños es mayor si la madre de familia habla una lengua indígena comparado con una madre de familia que sólo habla español. Los valores de los momios estimados para cada una de estas variables se muestran en la Tabla 5.5.

Tabla 5.5. Razones de momios estimados del modelo logístico con intersección aleatoria para las observaciones pertenecientes a una misma localidad.

Var. predictora	**Categoría**	$exp(\hat{\beta})$	**IC** 95 % **para** $exp(\beta)$
foodexp	Continuo	1.0003	(0.9954, 1.0051)
persroom	Conteo	1.0550	(1.0317, 1.0789)
agemoth	Continuo	0.9905	(0.9843, 0.9968)
floor01	1	0.8127	(0.7356, 0.8978)
wc01	1	0.8332	(0.7555, 0.9189)
cooker01	1	0.6972	(0.6202, 0.7837)
schoolm01	1	0.9302	(0.7935, 1.0905)
languagem01	1	0.5837	(0.5134, 0.6639)
foodexp:schoolm01	Continuo × 1	0.9904	(0.9848, 0.9960)

La prueba de significancia de $\sigma = 0.0326$ muestra evidencia de que las observaciones de la respuesta pertenecientes a una misma comunidad rural se encuentran correlacionadas ($p < 0.001$). En consecuencia, se espera que el 95 % de los valores de los efectos aleatorios $a_i \sim N(0, 0.0326^2)$ se encuentren entre $-1.96 \times 0.0326 = -0.064$ y $1.96 \times 0.0326 = 0.064$. El modelo con intersección aleatoria presenta una reducción en el *AIC*, *BIC* y en la devianza frente al modelo logístico sin efectos aleatorios (AIC=115127.49, Devianza=15107.49 para el modelo sin efectos y AIC=14927.09, Devianza= 14905.09 para el modelo con intersección aleatoria). Los valores numéricos de estos criterios se resumen en la Tabla 5.9.

5.3.2. Regresión logística ponderada

Para el ajuste ponderado por diseño se tomó en cuenta el diseño muestral de encuesta, conformado por tres etapas de muestreo, considerando estratos (*strata*), localidades rurales (*clusters*) y familias. Los resultados del ajuste se resumen en las Tablas 4.7 y 4.8. Podemos notar que, a pesar de que la variable predictora *floor01* es significativa para el modelo ponderado por diseño (p= 0.037), éste es menos significativo en comparación con los dos ajustes previos (p<0.001 en ambos). Nuevamente, los efectos de las variables *foodexp* y *schoolm01* no resultan significativos. No obstante, la interacción *foodexp:schoolm01* sí lo es (p<0.001), indicando que el gasto semanal per cápita varía de acuerdo a la escolaridad de la madre: (1) Para las madres de familia sin estudios o sin primara concluida, *schoolm01*= 0, el modelo de riesgo de desnutrición en los hijos es

$$log\left(\frac{\pi}{1-\pi}\right) = -0.5179 + 0.0492 \times persroom - 0.0119 \times agemoth - 0.1368 \times floor01 \\ - 0.2735 \times wc01 - 0.3731 \times cooker01 - 0.4074 \times languagem01$$

(2) Para las madres de familia con al menos primaria concluida, *schoolm01*= 0, la probabilidad de desnutrición en los hijos está dada por

$$log\left(\frac{\pi}{1-\pi}\right) = -0.5809 - 0.0619 \times foodexp + 0.0492 \times persroom - 0.0119 \times agemoth \\ -0.1368 \times floor01 - 0.2735 \times wc01 - 0.3731 \times cooker01 - 0.4074 \times languagem01$$

En el caso (2), el signo del efecto *foodexp* indica que un aumento en el gasto semanal per cápita en alimentos implica una reducción del riesgo de desnutrición en los hijos. El momio disminuye, aproximadamente, 6.00 % por cada unidad invertida en el gasto semanal per cápita en alimentos. En consecuencia, si el modelo es correcto, el riesgo de desnutrición disminuye 60 % por cada diez unidades invertidas en el gasto semanal per cápita. Este resultado muestra que *foodexp* tiene un gran peso en la reducción de la probabilidad de desnutrición para las madres con al menos primaria concluida.

Tabla 5.6. Modelo de regresión logística ponderada para el riesgo de desnutrición en niños de edad preescolar en zonas rurales de México.

Var. predictora	**Categoría**	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	$\hat{se}(\hat{\beta})/\hat{\beta}$	z	p
Intercept	Constante	-0.5179	0.1686	-0.3255	-3.07	0.0022
foodexp	Continuo	0.0011	0.0027	2.4729	0.40	0.6861
persroom	Conteo	0.0492	0.0131	0.2664	3.75	<0.001
agemoth	Conteo	-0.0119	0.0039	-0.3250	-3.08	0.0022
floor01	1	-0.1368	0.0655	-0.4787	-2.09	0.0373
wc01	1	-0.2735	0.0654	-0.2392	-4.18	<0.001
cooker01	1	-0.3731	0.0760	-0.2036	-4.91	<0.001
schoolm01	1	-0.0630	0.1054	-1.6718	-0.60	0.5500
languagem01	1	-0.4074	0.0946	-0.2321	-4.31	<0.001
foodexp:schoolm01	Continuo × 1	-0.0155	0.0037	-0.2360	-4.24	<0.001

Tabla 5.7. Razones de momios estimados para las variables explicativas del modelo logístico ponderado por diseño.

Var. predictora	**Categoría**	$exp(\hat{\beta})$	**IC** 95 % **para** $exp(\beta)$
foodexp	Continuo	1.0011	(0.9957, 1.0065)
persroom	Conteo	1.0504	(1.0238, 1.0777)
agemoth	Continuo	0.9881	(0.9806, 0.9957)
floor01	1	0.8722	(0.7671, 0.9916)
wc01	1	0.7607	(0.6692, 0.8648)
cooker01	1	0.6885	(0.5933, 0.7991)
schoolm01	1	0.9389	(0.7636, 1.1543)
languagem01	1	0.6654	(0.5528, 0.8009)
foodexp:schoolm01	Continuo × 1	0.9846	(0.9775, 0.9917)

La interpretación del modelo es análoga a la del modelo sin efectos aleatorios, los signos se mantienen y los valores estimados son similares. En comparación con los modelos previamente ajustados, el modelo ponderado por diseño presenta las deviaciones estándar más grandes de los tres modelos ajustados.

Tabla 5.8. Tabla de confusión del modelo ajustado de Regresión logística ponderado por diseño. Error aparente $= \frac{2939}{17865} = 0.1645$

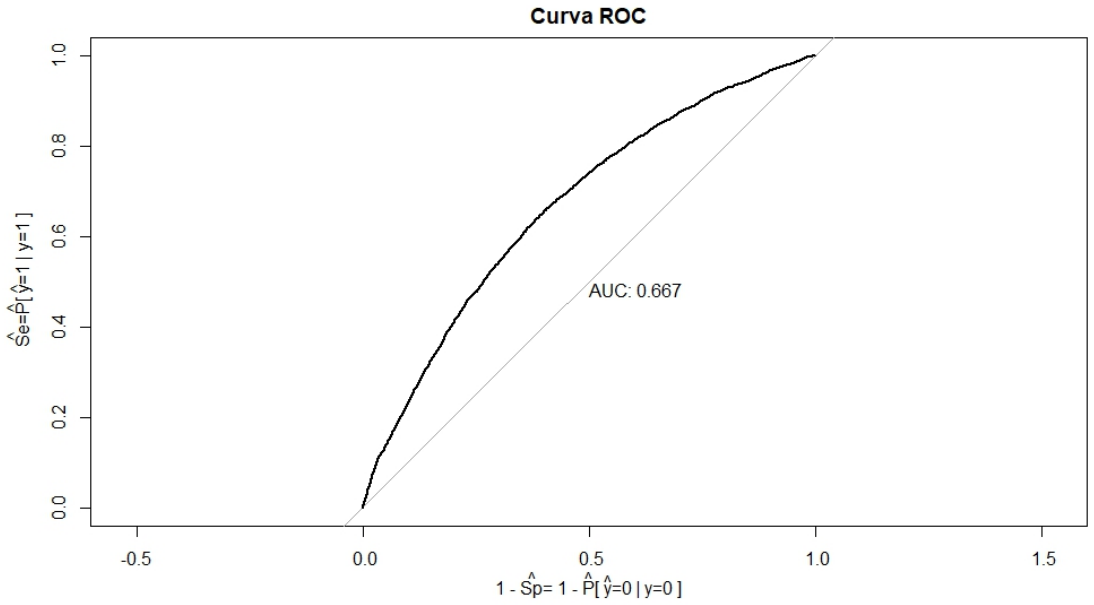
	Valores observados	
Valores ajustados	*Normal*	*Desnutrido*
Normal	14926	2939
Desnutrido	0	0

La tabla de confusión del modelo ponderado por diseño y, en consecuencia, el error aparente son los mismos que los reportados por el modelo sin efectos aleatorios. Es decir, al igual que este último, el desempeño predictivo del modelo ponderado por diseño es pobre, a pesar de que el error aparente indique lo contrario. Lo anterior es debido a que todas las observaciones de la base de datos son clasificadas como "normales" (Y=0). Los valores del AIC y la devianza (Devianza=15516.63, AIC=15556.53) resultan mayores que las del modelo con intersección aleatoria, pero menores que las de modelo sin efectos aleatorios (véase Tabla 5.9). Por último, se muestra la curva ROC del modelo ponderado por diseño. Los resultados obtenidos coinciden con los anteriores: el desempeño predictivo del modelo es pobre (AUC=0.667).

Tabla 5.9. Comparación de devianzas, AIC y BIC de los tres modelos ajustados.

Modelo	Devianza	AIC	BIC
Sin efectos	15107.49	15127.29	15205.39
Con intersección aleatoria	14905.09	14927.09	15012.79
Ponderado	15516.63	15556.53	–

Figura 5.1. *Curva ROC del modelo ponderado por diseño ajustado (AUC=0.667).*



5.4. Comparación de modelos

De manera general, las estimaciones puntuales de los coeficientes no presentan una tendencia constante de aumento o disminución para algún modelo particular, y, de hecho, éstas resultan similares en tres ajustes realizados (Tabla 5.10(b)). La diferencia más remarcable es la del coeficiente *foodexp* en el modelo ponderado por diseño, el cual presenta un aumento del 196.80 % respecto al modelo sin efectos aleatorios y 298.93 % respecto al modelo con intersección aleatoria (véase Tabla 5.10(c)). Sin embargo, este coeficiente corresponde al caso en el que (*schoolm01* =0), y no resulta significativo para el modelo. Por otro lado, los coeficientes de las variables *floor01*, *wc01*, *languagem01* y *foodexp:schoom01* presentan diferencias considerables en las estimaciones puntuales obtenidas por los tres modelos ajustados:

1. El coeficiente de la variable *floor01* estimado por el modelo logístico sin efectos aleatorios (β_a) presenta una reducción del 0.66 % respecto al modelo con intersección aleatoria (β_b), y un aumento del 33.63 % respecto al modelo ponderado por encuesta (β_c).

2. El coeficiente de *wc01* estimado por el modelo sin efectos aleatorios resulta 21.33 % mayor respecto al modelo con intersección aleatoria, pero 49 % menor respecto al modelo ponderado por diseño.

3. La diferencia entre los coeficientes estimados de *languagem01* por el modelo sin efectos aleatorios y por el modelo ponderado representa un 22.51 % del coeficiente estimado por el modelo sin efectos. Además, la diferencia entre los coeficientes estimados de esta misma variable por los modelos con intersección aleatoria y ponderado representa un 24 % del coeficiente estimado por el modelo con intersección.

4. El coeficiente estimado de la interacción *foodexp:schoolm01* por el modelo ponderado resulta 41.93 % mayor respecto al estimado por el modelo sin efectos aleatorios, y 61.32 % mayor respecto al estimado por el modelo con intersección aleatoria.

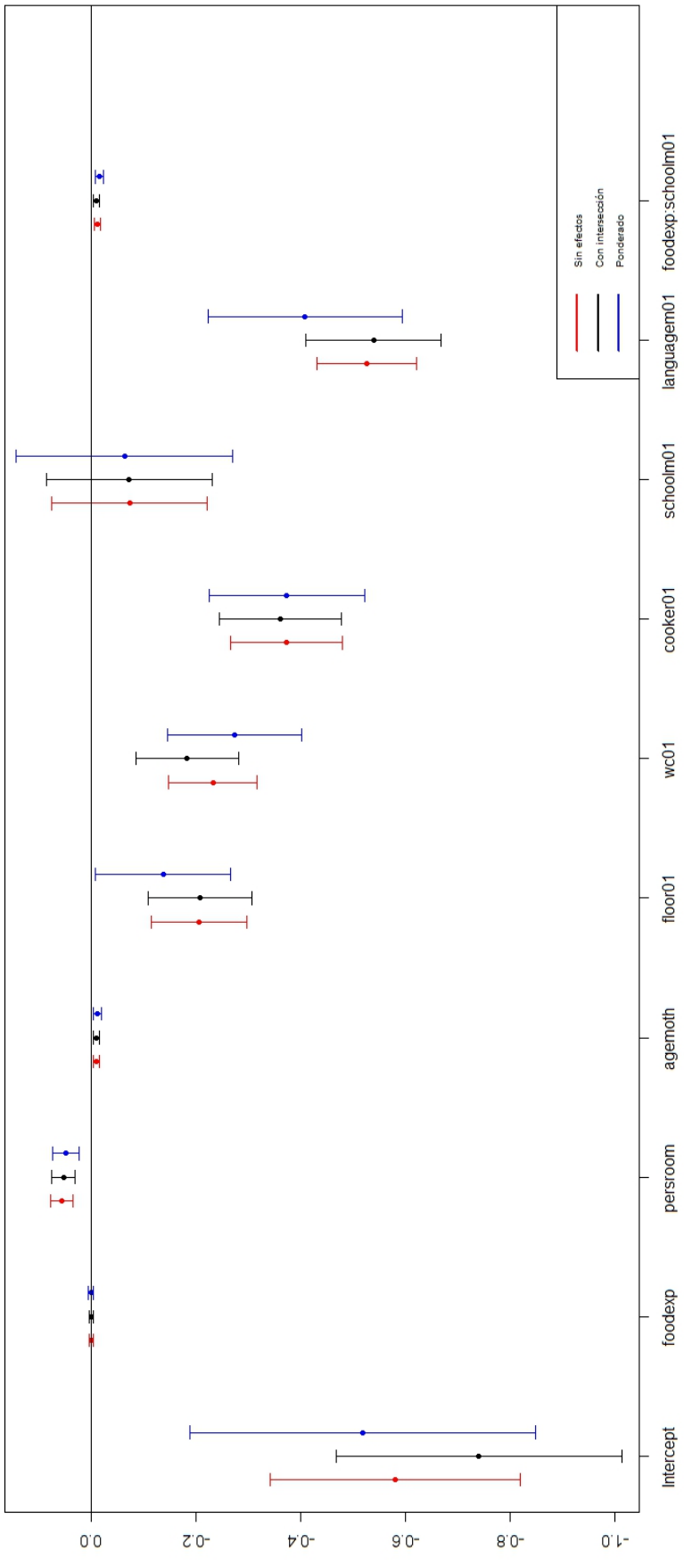


Figura 5.2. Gráfica comparativa de los intervalos de confianza estimados para los tres modelos ajustados (95 % de confianza). Podemos notar que los intervalos de los coeficientes de las variables foodexp y schoolm01 incluyen al valor cero. Esto muestra que ambas variables no son significativas para el modelo.

A diferencia de lo que sucede con las estimaciones puntuales de los coeficientes, las estimaciones de las desviaciones estándar son todas mayores en el modelo ponderado por diseño, respecto a los modelos sin efectos y con intersección aleatoria (Tabla 5.10(a)). Además, las estimaciones de las desviaciones estándar realizadas por el modelo con intersección aleatoria son mayores que las del modelo logístico sin efectos aleatorios. En consecuencia, los intervalos de confianza son más grandes para el modelo ponderado por diseño, seguidos por los intervalos de confianza del modelo con intersección aleatoria (véase Figura 5.2). Lo anterior, se debe a la correlación de observaciones existente en cada etapa de muestreo de la encuesta. Recordemos que el modelo ponderado por diseño fue ajustado tomando en cuenta la estructura completa de la encuesta (estratos y *clusters*); mientras que en el modelo con intersección aleatoria sólo se consideró el agrupamiento de observaciones por comunidades (*clusters*), y en el modelo logístico sin efectos aleatorios no fue considerado agrupamiento alguno de observaciones. Esto reafirma que, aunque las estimaciones puntuales de los parámetros β puedan ser similares, ignorar la estructura de correlación puede implicar estimaciones erróneas de las desviaciones estándar de los coeficientes. En conclusión, el modelo ponderado por diseño representa un ajuste más conservador, implicando intervalos de confianza más grandes en comparación con los ajustes de regresión logística sin efectos aleatorios y el modelo con intersección aleatoria. La prueba de significancia del modelo fue realizada en los tres ajustes:

$$H_0 : (\beta_{foodexp}, \beta_{persroom}, \ldots, \beta_{foodexp:schoolm01}) = \mathbf{0}$$
$$vs$$
$$H_1 : (\beta_{foodexp}, \beta_{persroom}, \ldots, \beta_{foodexp:schoolm01}) \neq \mathbf{0}$$

En los tres casos, la hipótesis nula H_0 fue rechazada (p<0.05). Es decir, al menos una de las variables predictoras tiene un efecto significativo en cada modelo.

Tabla 5.10(a). Comparación de estimaciones de los tres modelos ajustados de regresión logística para el cálculo de riesgo de desnutrición en zonas rurales de México: [a] Regresión logística sin efectos aleatorios, [b] Regresión logística con intersección aleatoria y [c] Regresión logística Ponderada.

Var. predictora	$\frac{\hat{\beta}_b}{\hat{\beta}_a}$	$\frac{\hat{\beta}_c}{\hat{\beta}_a}$	$\frac{\hat{\beta}_c}{\hat{\beta}_b}$	$\frac{\hat{se}(\hat{\beta}_b)}{\hat{se}(\hat{\beta}_a)}$	$\frac{\hat{se}(\hat{\beta}_c)}{\hat{se}(\hat{\beta}_a)}$	$\frac{\hat{se}(\hat{\beta}_c)}{\hat{se}(\hat{\beta}_b)}$
foodexp	0.744	2.968	3.989	1.095	1.219	1.113
persroom	0.955	0.876	0.917	1.054	1.211	1.148
agemoth	0.947	1.188	1.255	1.047	1.260	1.204
floor01	1.006	0.664	0.659	1.093	1.409	1.289
wc01	0.787	1.178	1.498	1.154	1.512	1.310
cooker01	0.970	1.004	1.035	1.092	1.390	1.273
schoolm01	0.995	0.868	0.871	1.073	1.394	1.299
languagem01	1.023	0.775	0.757	1.361	1.962	1.442
foodexp:schoolm01	0.880	1.419	1.613	1.058	1.340	1.267

Tabla 5.10(b). Comparación de estimaciones de los tres modelos de regresión logística ajustados. Para cada variable predictora, la primera fila muestra el estimador puntual $\hat{\beta}$ en el modelo. La segunda y tercera fila muestran el intervalo de confianza y la desviación estándar estimados de $\hat{\beta}$, respectivamente.

	Modelo de Regresión		
Var. predictora	Sin efectos	Efectos aleatorios	Ponderado por diseño
foodexp	0.0004 (-0.0041, 0.0047) 0.0023	0.0003 (-0.0045,0.0051) 0.0025	0.0011 (-0.0043, 0.0065) 0.0027
persroom	0.0561 (0.0349, 0.0773) 0.0108	0.0536 (0.0312, 0.07598) 0.0114	0.0492 (0.0235, 0.0749) 0.0131
agemoth	-0.0101 (-0.0161, -0.0040) 0.0031	-0.0095 (-0.0158, -0.0032) 0.0032	-0.0119 (-0.0196, -0.0043) 0.0039
floor01	-0.2061 (-0.2972, -0.1150) 0.0465	-0.2074 (-0.3070, -0.1078) 0.0508	-0.1368 (-0.2651, -0.0084) 0.0655
wc01	-0.2320 (-0.3167, -0.1471) 0.0433	-0.1825 (-0.2804, -0.0846) 0.0499	-0.2735 (-0.4017, -0.1453) 0.0654
cooker01	-0.3717 (-0.4790, -0.2648) 0.0546	-0.3607 (-0.4776,-0.2437) 0.0597	-0.3731 (-0.5220, -0.2242) 0.0760
schoolm01	-0.0727 (-0.2209, 0.0755) 0.0756	-0.0723 (-0.2313, 0.0866) 0.0811	-0.0630 (-0.2696, 0.1435) 0.1054
languagem01	-0.5258 (-0.6201, -0.4312) 0.0482	-0.5382 (-0.6668, -0.4096) 0.0655	-0.4074 (-0.5928 -0.2221) 0.0946
foodexp:schoolm01	-0.0110 (-0.0163, -0.0055) 0.0027	-0.0096 (-0.0153,-0.0039) 0.0029	-0.0155 (-0.0227, -0.0083) 0.0037

Tabla 5.10(c). Comparación de estimaciones de tres modelos de regresión logística para el cálculo de riesgo de desnutrición en zonas rurales de México: ^{a} Regresión logística sin efectos aleatorios, ^{b} Regresión logística con intersección aleatoria y ^{c} Regresión logística Ponderada.

Var. predictora	$100\left[\frac{\hat{\beta}_a - \hat{\beta}_b}{\hat{\beta}_a}\right]$	$100\left[\frac{\hat{\beta}_a - \hat{\beta}_c}{\hat{\beta}_a}\right]$	$100\left[\frac{\hat{\beta}_b - \hat{\beta}_c}{\hat{\beta}_b}\right]$
foodexp	25.60	-196.80	-298.93
persroom	4.50	12.38	8.267
agemoth	5.33	-18.80	-25.50
floor01	-0.66	33.63	34.06
wc01	21.33	-17.87	-49.83
cooker01	2.96	-0.40	-3.50
schoolm01	0.45	13.24	12.85
languagem01	-2.35	22.51	24.29
foodexp:schoolm01	12.02	-41.93	-61.32

Tabla 5.10(d). Comparación de estimaciones de tres modelos de regresión logística para el cálculo de riesgo de desnutrición en zonas rurales de México: ^{a} Regresión logística sin efectos aleatorios, ^{b} Regresión logística con intersección aleatoria y ^{c} Regresión logística Ponderada. Notación: $\hat{V}(\hat{\beta}) = \hat{se}(\hat{\beta})^2$.

Var. predictora	$100\left[\frac{\hat{V}(\hat{\beta}_a) - \hat{V}(\hat{\beta}_b)}{\hat{V}(\hat{\beta}_a)}\right]$	$100\left[\frac{\hat{V}(\hat{\beta}_a) - \hat{V}(\hat{\beta}_c)}{\hat{V}(\hat{\beta}_a)}\right]$	$100\left[\frac{\hat{V}(\hat{\beta}_b) - \hat{V}(\hat{\beta}_c)}{\hat{V}(\hat{\beta}_b)}\right]$
foodexp	-19.88	-48.54	-23.91
persroom	-11.13	-46.56	-31.88
agemoth	-9.55	-58.70	-44.86
floor01	-19.58	-98.56	-66.05
wc01	-33.28	-128.58	-71.50
cooker01	-19.31	-93.30	62.03
schoolm01	-15.07	94.26	-68.82
languagem01	-85.17	-284.88	-107.85
foodexp:schoolm01	-11.86	-79.68	-60.63

Capítulo 6

Conclusión

Los objetivos de este trabajo fueron la presentación, aplicación y comparación de modelos con variables respuesta binarias para datos agrupados.

Dos enfoques de regresión logística para tratar datos con observaciones correlacionadas de la respuesta fueron considerados: el modelo logístico con intersección aleatoria y el modelo logístico ponderado por diseño. Estos modelos se aplicaron, utilizando al mismo conjunto de variables predictoras, a datos provenientes de la *Encuesta Nacional de Alimentación y Nutrición en el Medio Rural* llevada a cabo en áreas rurales de México en 1996. Los modelos ajustados carecieron de poder predictivo: todos los individuos son pronosticados como "normales" ($Y = 0$). Sin embargo, dichos ajustes coinciden, de manera general, en la descripción de efectos de las variables predictoras en la respuesta: (1) la disponibilidad de piso, estufa de gas y wc son factores que reducen el riesgo de desnutrición; (2) el gasto semanal per cápita en alimentos sólo tiene un efecto significativo cuando la madre de familia concluyó al menos la primaria, dicho efecto reduce la probabilidad de padecer desnutrición en los infantes por cada unidad invertida en este gasto, (3) a mayor número de personas por habitación, mayor la probabilidad de desnutrición; (4) el riesgo de desnutrición en los niños es mayor si la madre habla alguna lengua indígena en comparación con las madres que sólo hablan español; (5) a mayor edad de la madre, menor el riesgo de presentar desnutrición.

Por otro lado, en el aspecto comparativo, se destaca lo siguiente: el modelo logístico sin efectos aleatorios, en el cual se supuso la hipótesis de independencia de observaciones de las variables respuesta, es el modelo que reporta las desviaciones estándar más pequeñas de los tres ajustes; en el modelo logístico con intersección aleatoria, donde se consideró el agrupamiento de observaciónes sólo por comunidades, se estimaron valores de desviaciones estándar mayores que las del modelo lineal, pero menores que las del modelo ponderado por diseño, en el que se consideró el diseño muestral completo (estrato, comunidad).

De manera general, los resultados estadísticos de los tres modelos ajustados resultan los mismos: las variables predictoras consideradas resultan significativas, y con una interpretación similar de efectos en la respuesta. Sin embargo, la precisión en la estimación de las desviaciones estándar (y en consecuencia, de los intervalos de confianza) pareciera verse afectada por la estructura de la distribución de la población. En consecuencia, el supuesto de independencia podría no cumplirse. Debido a lo anterior, se recomienda considerar al modelo logístico con intersección aleatoria y al modelo ponderado, los cuales toman en cuenta la posible correlación existente entre observaciones de la variable respuesta.

Anexos

Anexo A

Anexo I: Análisis exploratorio

En este apartado se muestra información descriptiva de las variables pertenecientes a la base de datos. El análisis realizado en las variables no categóricas (*foodexp, persroom* y *agemoth*) se enfoca principalmente en encontrar correlaciones entre las variables predictoras, además de la búsqueda de posibles valores atípicos. Por otro lado, en el apartado de las variables categóricas, se muestran las frecuencias observadas de cada una de sus categorías utilizando gráficas de barras. Se incluye también, un apartado exploratorio para las variables diseño.

A.1. Variables no categóricas

La Figura A.1 muestra la gráfica de dispersión por pares de las variables *foodexp*, *persroom* y *agemoth*. Los coeficientes de correlación (indicados en el panel superior) indican que no existe una correlación alta entre estas variables. La correlación más grande reportada es entre las variables *foodexp* y *persroom* (=0.32). Los histogramas, mostrados en el panel diagonal de la Figura A.1, muestran que:

1. la mayoría de las observaciones obtenidas (a saber, el 86.34 %) tienen un gasto semanal per cápita en alimentos entre cero y cincuenta.

2. la mayoría de las obseraciones (78.79 %) se caracterizan por tener un número de personas por habitación (*persroom*) menor que seis.

3. la mayoría de las madres de familia encuestadas (82.90 %) tienen edades entre quince y treinta y cinco años. Sólo el 10.80 % de las madres encuestadas tienen edades menores o iguales que quince años (y mayores que doce), y el 16.01 % tienen edades mayores o iguales que treinta y cinco (y menores que cincuenta).

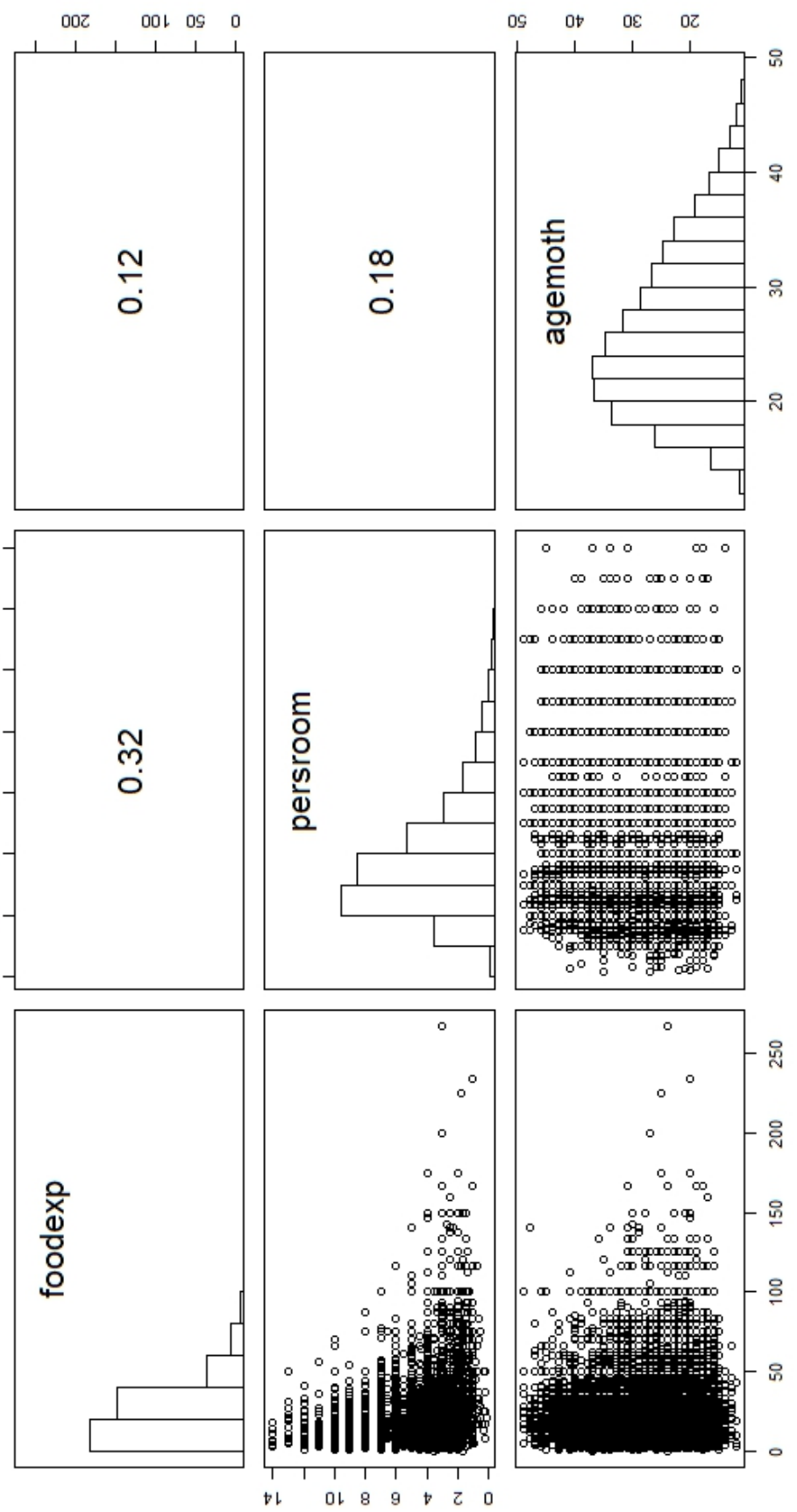


Figura A.1. *Gráfica de dispersión por pares de las variables no cateregóricas. El panel superior muestra la correlación por pares entre las variables. El panel diagonal muestra los histogramas de las variables, y el panel inferior las gráficas de dispersión.*

Para encontrar posibles valores atípicos (*outliers*) fueron utilizadas gráficas de caja y bigote para cada variable. Los resultados se muestran en Figura A.2. Podemos notar que, en el caso de *foodexp*, existen muchas observaciones atípicas, posiblemente una transformación ayude a eliminar el número de estas. Zuur et al. (2009) recomiendan utilizar una transformación logarítmica (pp. 532–533), esto debido a la asimetría positiva presentada en la distribución de frecuencias. La Figura A.3 muestra la gráfica de caja y bigote de la variable transformada *log(foodexp)*(logarítmo natural de *foodexp*). Podemos notar que a pesar de la transformación, el número de valores atípicos es considerable.

En el caso de *persroom*, se consideran como *outliers* a las observaciones que tienen un número de personas por habitación en casa mayor o igual a nueve. El caso extremo (*persroom* =14), se cumple para ocho observaciones de la base de datos. Por último, las observaciones tales que *agemoth* =49 son consideradas como *outliers*, lo cual se cumple para trece observaciones.

Figura A.2. *Gráficas de caja y bigote de las variables foodexp, persroom y agemoth*

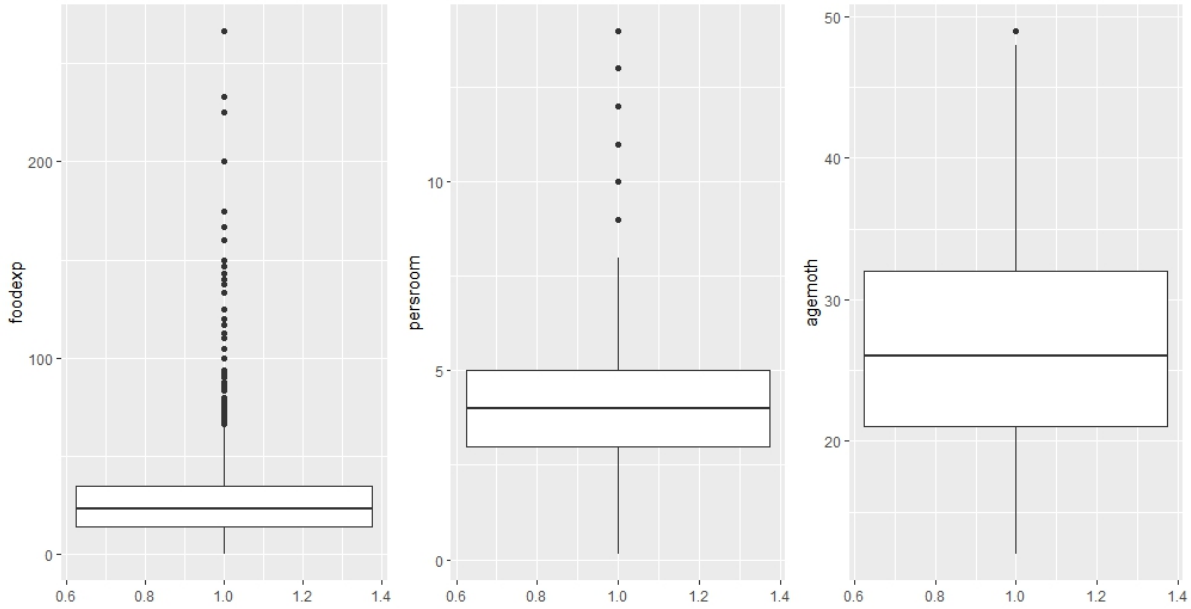
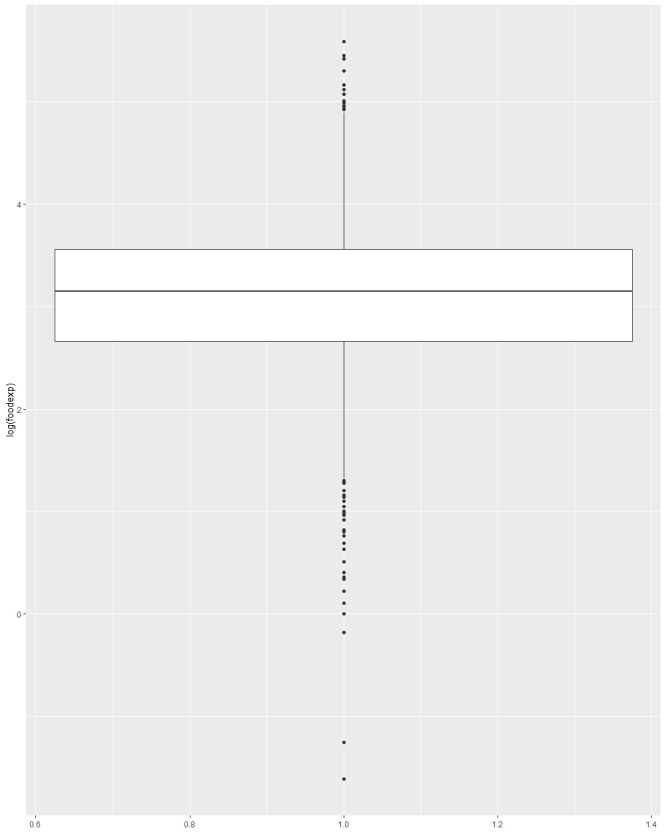


Figura A.3. *Gráfica de caja y bigote de la variable transformada ln(foodexp).*



A.2. Variables categóricas

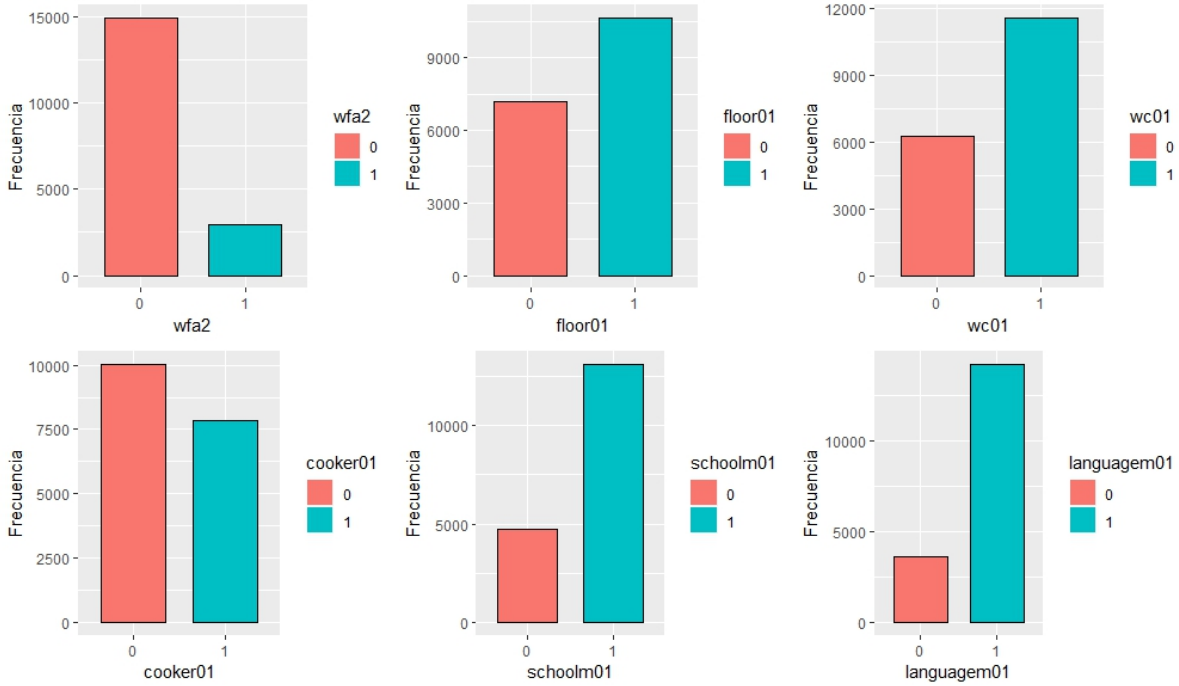
La Figura A.4 muestra las gráficas de frecuencias de las variables categóricas de la base de datos. Podemos notar que no existe un balance importante en las categorías de las variables *wfa2*, *schoolm01* y *languagem01*. En el caso de *wfa2*, tan sólo el 16.45 % de las observaciones pertenecen a la categoría 1, i.e el 16.45 % de los niños observados padecen desnutrición. Por otro lado, sólo el 26.69 % de las madres de familia no terminaron la primaria o no tiene estudios (*schoolm01* =0). Además, el 20.44 % de las madres de familia registradas habla un dialécto y el español (*languagem01* =0), mientras que el otro 79.55 % habla sólo español (*languagem01* = 1). Respecto a las demás variables se tiene:

1. El porcentaje de observaciones con disponibilidad de piso de algún material (*floor01*= 1) es 59.66 %.

2. El 64.84 % de las observaciones dispone de baño en casa (*wc01* =1).

3. El 43.90 % de las observaciones dispone de estufa de gas en casa (*cooker01* =1).

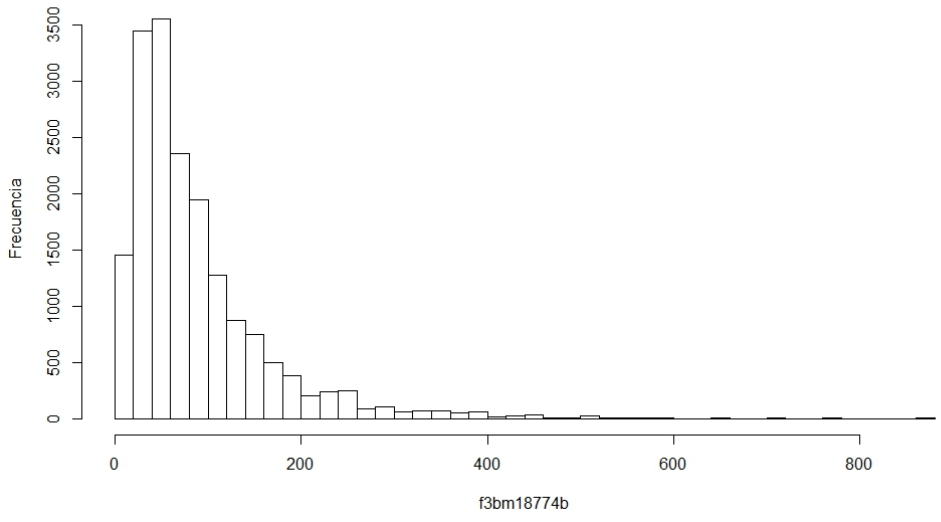
Figura A.4. *Gráfica de frecuencias para cada categoría de las variables binarias*



A.3. Variables de diseño

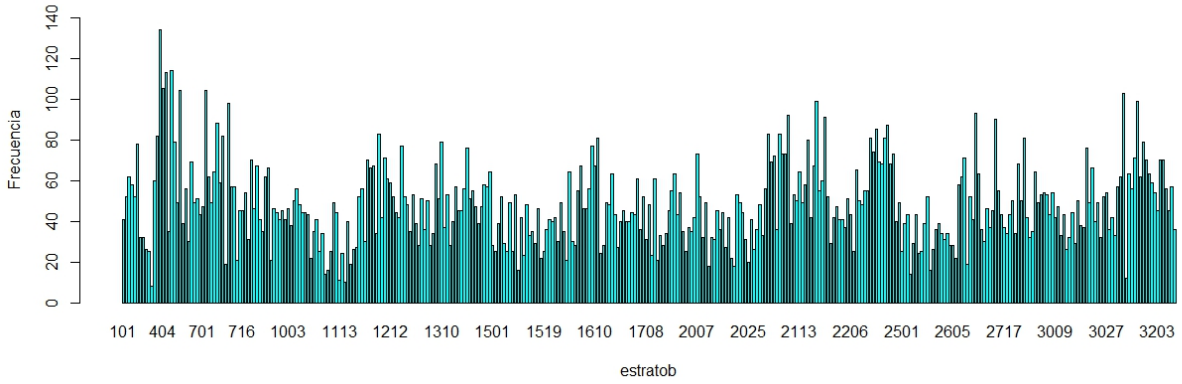
La Figura A.5 muestra el histograma de frecuencias de las pesos muestrales (factores de expansión) considerados en la ENAL 96. Ésta muestra que la distribución de frecuencias presenta asimetría positiva: el porcentaje de observaciones con pesos muestrales menores que cien (<100) es 71.46 %. En otras palabras, 71.46 % de las observaciones en la muestra fueron seleccionadas con una probabilidad mayor que 0.01.

Figura A.5. *Histograma de los pesos muestrales considerados en la ENAL96.*



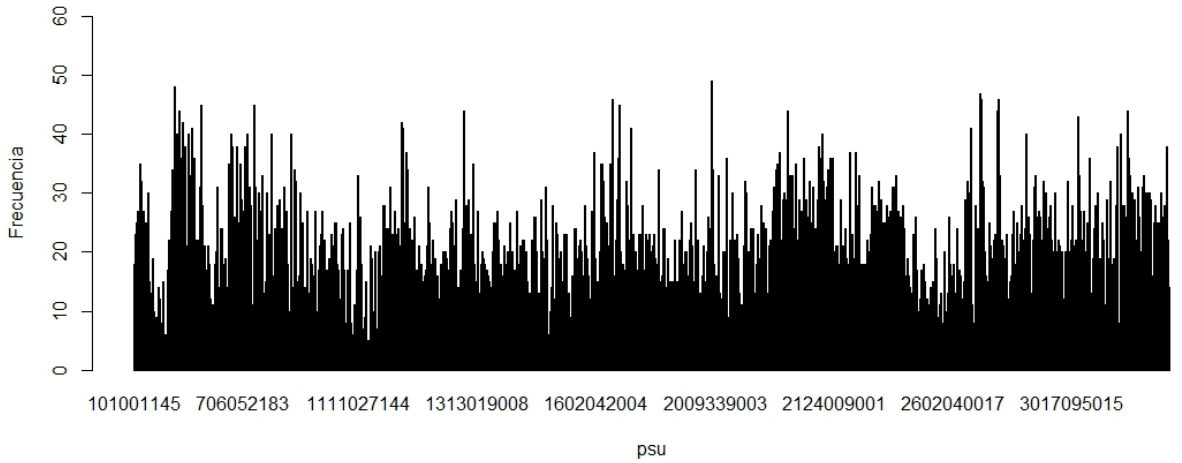
Respecto a los estratos considerados, en la Figura A.6 podemos observar que el número de observaciones por estrato es desproporcionada: hay estratos que cuentan con más de 80 observaciones, al igual que hay estratos con menos de 20 observaciones.

Figura A.6. *Gráfica de barras de la variable diseño estratob.*



De manera similar a la variable *estratob*, en la Figura A.7 podemos notar que el número de observaciones por *psu* no es balanceada: existen comunidades rurales (*psu*) en las que se encuestaron más de cuarenta personas, mientras que hay otras en las que se seleccionaron menos de diez.

Figura A.7. *Gráfica de barras de la variable diseño psu.*



Anexo B

Anexo II: Proporción de madres con niños con desnutrición

Uno de los resultados en los que coinciden los tres modelos ajustados indica que la probabilidad de que un infante padezca desnutrición disminuye por cada año de edad cumplido de la madre. Sin embargo, dicha afirmación sólo resulta válida para ciertos intervalos de valores de la edad de la madre (véase Figura B.1 y Tabla B.1): madres con edades entre (13, 28), (35, 38) y (40, 44). Las proporciones estimadas para las madres con 12, 13, 14, 47, 48, y 49 años parecen verse afectadas por sus frecuencias en la muestra.

Figura B.1. Gráfica de la proporción de madres con niños con desnutrición de acuerdo a la edad de la madre. Las proporciones son calculadas utilizando el promedio de los valores observados tales que *wfa2*=1 en la muestra.

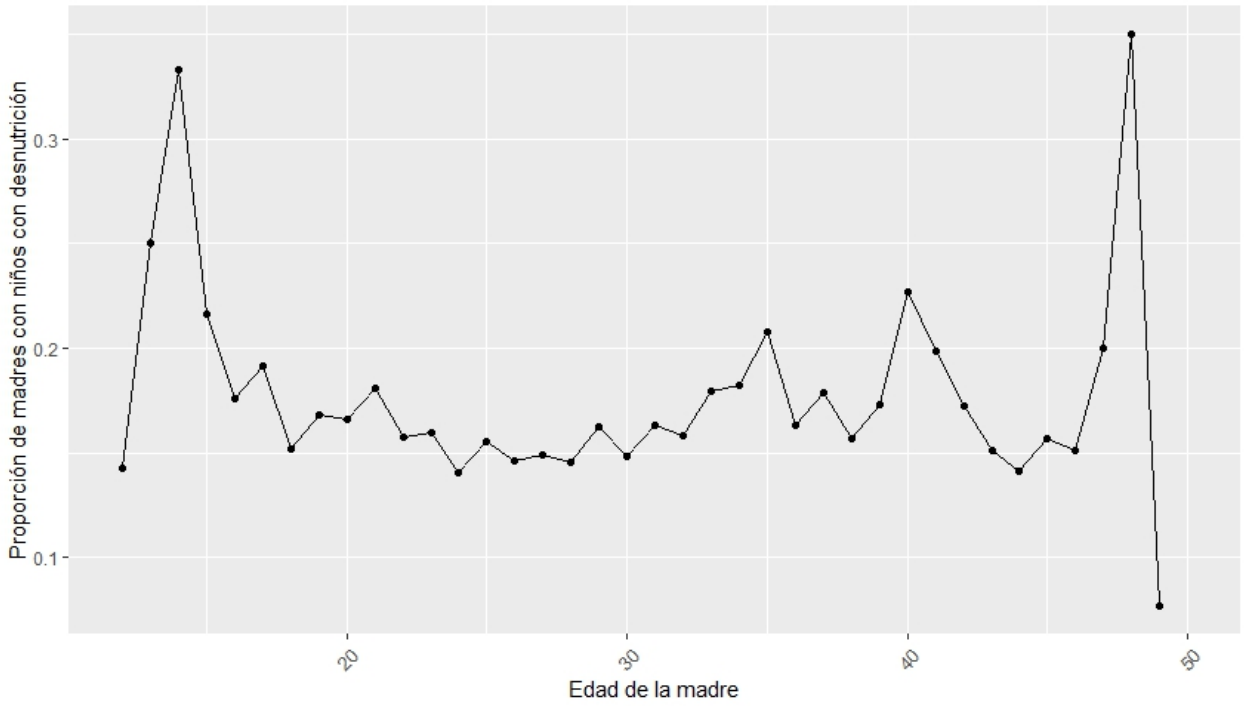


Tabla B.1. Anexo II: Proporción de madres con niños con desnutrición de acuerdo a la edad de la madre. Las proporciones son calculadas utilizando el promedio de los valores observados tales que *wfa2*=1 en la muestra.

Edad	**Frecuencia**	**Proporción**
12	7	0.143
13	16	0.250
14	45	0.333
15	125	0.216
16	335	0.176
17	517	0.191
18	703	0.152
19	862	0.168
20	935	0.166
21	974	0.181
22	1065	0.157
23	1034	0.160
24	1040	0.140
25	948	0.155
26	936	0.146
27	854	0.149
28	806	0.145
29	744	0.163
30	680	0.148
31	624	0.163
32	633	0.158
33	562	0.178
34	559	0.182
35	505	0.208
36	454	0.163
37	363	0.179
38	325	0.157
39	260	0.173
40	225	0.227
41	166	0.199
42	180	0.172
43	119	0.151
44	78	0.141
45	70	0.157
46	53	0.151
47	30	0.200
48	20	0.350
49	13	0.077
50	0	–

Tabla B.2. Frecuencia de estratos con una sóla comunidad de acuerdo a la edad de la madre.

Edad	**Número de madres**	**Núm. estratos con sólo un** *cluster*
12	7	7
13	16	16
14	45	44
15	125	112
16	335	262
17	517	363
18	703	441
19	862	492
20	935	517
21	974	536
22	1,065	573
23	1,034	553
24	1,040	576
25	948	543
26	936	527
27	854	535
28	806	497
29	744	476
30	680	465
31	624	419
32	633	432
33	562	386
34	559	378
35	505	373
36	454	346
37	363	286
38	325	258
39	260	218
40	225	197
41	166	147
42	180	155
43	119	102
44	78	72
45	70	67
46	53	53
47	30	30
48	20	19
49	13	13
Total	17,865	11,486

La Tabla B.2 presenta el número de estratos con sólo una comunidad rural (o *psu*) de acuerdo a la variable *agemoth*. En esta tabla podemos observar que, para cada edad de la

madre, existen estratos con sólo una comunidad (*cluster*). Por consiguiente, no es posible calcular las proporciones estimadas utilizando el diseño de la muestra.

Anexo C

Anexo III: Código de R utilizado

```
setwd("C:/Users/Paco/Desktop/Tesis/Base de datos")

#Paquetes
library(readxl)
library(ggplot2)
library(glmmML)
library(survey)
library(plotrix)
library(MuMIn)
library(pROC)
library(moments)
library(foreign)
library(xtable)

#Importando base de datos:
datos=as.data.frame(read_excel("finfil1f_m17865.xls"))


attach(datos)
enal=data.frame(wfa2, foodexp, persroom, agemoth, floor01, wc01,
                 cooker01,  schoolm01, languagem01,
                 psu,estratob,f3bm18774b)
detach(datos)


##Codificación de las variables binarias
names(enal)

enal$floor01=enal$floor01+1
enal$wc01=enal$wc01+1
enal$cooker01=enal$cooker01+1
enal$schoolm01=enal$schoolm01+1
enal$languagem01=enal$languagem01+1
```

```
enal$floor01[enal$floor01==2]=rep(0,length(enal$floor01[enal$floor01==2]))
enal$wc01[enal$wc01==2]=rep(0,length(enal$wc01[enal$wc01==2]))

enal$cooker01[enal$cooker01==2]=
rep(0,length(enal$cooker01[enal$cooker01==2]))

enal$schoolm01[enal$schoolm01==2]=
rep(0,length(enal$schoolm01[enal$schoolm01==2]))

enal$languagem01[enal$languagem01==2]=
rep(0,length(enal$languagem01[enal$languagem01==2]))

#Conversión de las variables binarias a factores
enal$floor01=as.factor(enal$floor01)
enal$wc01=as.factor(enal$wc01)
enal$cooker01=as.factor(enal$cooker01)
enal$schoolm01=as.factor(enal$schoolm01)
enal$languagem01=as.factor(enal$languagem01)
enal$psu=as.factor(enal$psu)
enal$estratob=as.factor(enal$estratob)

str(enal)

length(levels(enal$psu)) # Número de comunidades
length(levels(enal$estratob)) # Número de estratos del diseño

###Análisis descriptivo

names(enal)

##Histogramas
#foodexp
ggplot(data=enal, aes(enal$foodexp),xlim=c(10,100)) +
  geom_histogram(binwidth = 10, fill=I("blue"), col=I("black"))


#Persroom
ggplot(data=enal, aes(persroom),xlim=c(0,100)) +
  geom_histogram(binwidth = 1, fill=I("blue"), col=I("black"))

#agemoth
ggplot(data=enal, aes(agemoth)) +
  geom_histogram(binwidth = 1, fill=I("blue"), col=I("black"))
```

```
### Gráficas de barras

require(gridExtra)

#floor01
f=ggplot(enal, aes(x = floor01, fill = floor01,col=I("black"))) +
  geom_bar(width = 0.50)+ ylab("Frecuencia")

#wc01
w=ggplot(enal, aes(x = wc01, fill = wc01, col=I("black"))) +
  geom_bar(width = 0.50)+ ylab("Frecuencia")


#cooker01
c=ggplot(enal, aes(x = cooker01, fill = cooker01,col=I("black"))) +
  geom_bar(width = 0.50)+ ylab("Frecuencia")


#schoolm01
s=ggplot(enal, aes(x = schoolm01, fill = schoolm01,col=I("black"))) +
  geom_bar(width = 0.50)+ ylab("Frecuencia")


#languagem01
l=ggplot(enal, aes(x = languagem01, fill = languagem01,col=I("black"))) +
  geom_bar(width = 0.50)+ ylab("Frecuencia")

grid.arrange(wfa,f,w, c, s, l, ncol=2, nrow=3)

##Gráficas de caja y bigote

#foodexp
food=ggplot(enal, aes(x=1, y=foodexp)) + xlab("")+ geom_boxplot()

#ln(foodexp)
ggplot(enal, aes(x=1, y=log(foodexp))) + xlab("") + geom_boxplot()

#Persroom
Pers=ggplot(enal, aes(x=1, y=persroom)) + xlab("")+ geom_boxplot()

#agemoth
age=ggplot(enal, aes(x=1, y=agemoth)) + xlab("")+ geom_boxplot()

grid.arrange(food,Pers, age, ncol=3)


##Colinearidad
```

```
panel.hist <- function(x, ...)
{usr <- par("usr"); on.exit(par(usr))
par(usr = c(usr[1:2], 0, 1.5) )
h <- hist(x, plot = FALSE)
breaks <- h$breaks; nB <- length(breaks)
y <- h$counts; y <- y/max(y)
rect(breaks[-nB], 0, breaks[-1], y, ...)
}

panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste0(prefix, txt)
  if(missing(cex.cor)) cex.cor <- 1/strwidth(txt)
  text(0.5, 0.5, txt, cex = 2)
}

pairs(enal[,c(2,3,4)], diag.panel = panel.hist, upper.panel = panel.cor)

names(enal)
cor(enal[,c(2,3,4)])


###### Variables de diseño
barplot(table(enal$estratob), ylim=c(0,150),
        ylab="Frecuencia", col="cyan" , xlab="estratob")

barplot(table(enal$psu), ylab="Frecuencia", ylim=c(0,60), xlab="psu")

hist(enal$f3bm18774b, main="", xlab= "f3bm18774b",
    breaks=50, ylab="Frecuencia")

####Ajustes

#Regresión logística (sin interacciones de segundo orden)
M=glm(wfa2~.,family=binomial(link="logit"),data=enal[,c(-10,-11,-12)])

print(summary(M), digits= 2)

#Predicción
predicc=ifelse(M$fitted.values>0.5,1,0)
table(predicc, enal$wfa2) #Tabla de confusión
```

```
#Error aparente
2940/17865 # = 0.1645


##Regresión logística con interacciones
#Modelo con interacciones de segundo orden

M2=glm(wfa2~.^2,family=binomial(link="logit"),
       data=enal[,c(-10,-11,-12)])
print(summary(M2), digits= 2)


##Seleccion de variables

#Both (Backward and forward selection)
M_bo=step(M2,direction="both")
length(coef(M_bo)) #17 variables

#Backward
M_ba=step(M2,direction="backward")
length(coef(M_ba)) #17 variables

#Forward
M_fo=step(M2,direction="forward")
length(coef(M_fo)) # 37 variables (No reduce)

##Modelo elegido

print(summary(M_ba),digits=2) #La única interacción significativa es
#foodexp:schoolm


##Modelo resultante de eliminar interacciones no
#significativas

M_final=update(M, .~. + foodexp:schoolm01)

BIC(M_final)
#Extracción de desviaciones estándar
summ_glm=summary(M_final)
se_glm=as.data.frame(summ_glm$coefficients)$`Std. Error`

### se(beta)/beta
se_glm/M_final$coefficients
print(summary(M_final),digits=2) #Modelo final
```

```
#Intervalos del 95 % de confianza para los
#parámetros del modelo (Wald)

confint(M_final,type="Wald")

####### Significancia del modelo

#Modelo nulo
Nulo_glm=glm(wfa2~1,family=binomial(link="logit"),
             data=enal[,c(-10,-11,-12)])

anova(Nulo_glm,M_final, test="Chisq") # p<2.2e-16.


##Error aparente del modelo final

#Valores predichos
predicc=ifelse(M_final$fitted.values>0.5,1,0)
#Tabla de confusión
table(predicc, enal$wfa2)

#Error aparente
2939/17865 # = 0.1645


anova(M_final,M,test="Chisq")

#Momios (GLM)

(Momios_glm=exp(M_final$coefficients))

############   Intervalos 95% de
#confianza para los momios

## Cuantil 1-(0.5/2) = 1.96

#Lím ínf del intervalo
exp(M_final$coefficients-(1.96*se_glm))

#Lim sup del intervalo
exp(M_final$coefficients+(1.96*se_glm))


#### Gráfica de los intervalos del 95% de confianza de
#los coeficientes del modelo

#Límite inferior del intervalo
```

```
#El cuantil 0.975 es 1.96
liminf=double(length(M_final$coefficients))

for (i in 1:length(liminf)) {
  liminf[i]=M_final$coefficients[i]-(1.96*se_glm[i])
}

#Límite superior del intervalo

limsup=double(length(M_final$coefficients))

for (i in 1:length(limsup)) {
  limsup[i]=M_final$coefficients[i]+(1.96*se_glm[i])
}



#Gráfica

coef=c("Intercept", "foodexp", "persroom","agemoth",
       "floor01","wc01", "cooker01", "schoolm01",
       "languagem01","foodexp:schoolm01")

ggplot()+ geom_point(aes(x=factor(coef,levels=coef),
      M_final$coefficients)) +
  geom_errorbar(aes(x=factor(coef,levels=coef),
    ymin=liminf, ymax=limsup, width=5), width=0.2)+
  xlab("") + ylab("CI") +theme_bw()+
  theme(axis.text.x  = element_text(angle=45, vjust=0.5,
                                    size=10))



##Regresión logística con intersección aleatoria

#Modelo con interacciones de segundo orden

M_intercepto=glmmML(wfa2~ foodexp  + persroom + agemoth +
                floor01 + wc01 + cooker01+ schoolm01 +
                languagem01 + foodexp:schoolm01,
                family = binomial, cluster=enal$psu,
                data=enal)

print(summary(M_intercepto),digits=1)

BIC(M_intercepto)
```

```
M_intercepto$aic
M_intercepto$deviance


#### Significancia del modelo

Nulo_glmmML=glmmML(wfa2~1, data=enal, cluster=enal$psu,
                   family=binomial)

pchisq((deviance(Nulo_glmmML)-deviance(M_intercepto)),
       df=9, lower.tail=FALSE) # p=4.546e-87 < 0.05

##se(beta})/ beta

M_intercepto$coef.sd/M_intercepto$coefficients

###Intervalos de confianza de los coeficientes
##Límite inferior
#Vector de ceros que almacenará los límites inferiores
(liminf=double(length(M_intercepto$coefficients)))

for (i in 1:length(liminf)) {
  liminf[i]=M_intercepto$coefficients[i]-
    (1.96*M_intercepto$coef.sd[i])
  }

liminf

##Límite superior
#Vector de ceros
#que almacenará los límites superiores
(limsup=double(length(M_intercepto$coefficients)))

for (i in 1:length(limsup)) {
  limsup[i]=M_intercepto$coefficients[i]+
    (1.96*M_intercepto$coef.sd[i])
  }

limsup

#Gráfica

ggplot()+ geom_point(aes(x=factor(coef,levels=coef),
            y=M_intercepto$coefficients)) +
  geom_errorbar(aes(x=factor(coef,levels=coef),ymin=liminf,
      ymax=limsup), width=0.2)+ xlab("") + ylab("CI") +
  theme_bw() + theme(axis.text.x  = element_text(angle=45,
```

```
                                    vjust=0.5, size=10))

##Momios

(Momios_interc=exp(M_intercepto$coefficients))

############   Intervalos 95% de confianza para los momios
## Cuantil 1-(0.5/2) = 1.96


#Lím ínf
exp(M_intercepto$coefficients-
      (1.96*M_intercepto$coef.sd))

#Lim sup
exp(M_intercepto$coefficients+
      (1.96*M_intercepto$coef.sd))


####Modelo ponderado

str(enal)

(dis=svydesign(id=~psu, strata = ~estratob,
               weights = ~f3bm18774b  ,data=enal))

###Medias ponderadas

# Gasto semanal per cáp. en alimentos.
svymean(~enal$foodexp,dis)
# Personas por habitación
svymean(~enal$persroom,dis)
#Edad de la madre
svymean(~enal$agemoth,dis)

##Intervalo de confianza de las medias 95%

confint(svymean(~enal$foodexp,dis),df=degf(dis))
confint(svymean(~enal$persroom,dis),df=degf(dis))
confint(svymean(~enal$agemoth,dis),df=degf(dis))

###Proporciones estimadas de las variables categóricas
svymean(~enal$wfa2,dis)
svymean(~enal$floor01,dis)
svymean(~enal$wc01,dis)
svymean(~enal$cooker01,dis)
svymean(~enal$schoolm01,dis)
```

```
svymean(~enal$languagem01,dis)



##Intervalo de confianza de las proporciones 95%
confint(svymean(~enal$wfa2,dis))
confint(svymean(~enal$floor01,dis))
confint(svymean(~enal$wc01,dis))
confint(svymean(~enal$cooker01,dis))
confint(svymean(~enal$schoolm01,dis))
confint(svymean(~enal$languagem01,dis))

#Ajuste
svy=svyglm(wfa2~foodexp+persroom+agemoth+floor01+
          wc01+cooker01+schoolm01+languagem01+
            foodexp:schoolm01,
      design=dis, data=enal, family = quasibinomial)
summary(svy)

svy$deviance
AIC(svy)

#Probabilidades estimadas por el modelo

prob=predict(svy, enal, class="response",
             type="response")

#Predicción (umbral=0.5)
pred=ifelse(prob>0.5,1,0)
table(pred,enal$wfa2)

2939/17865


#####Curva ROC
prob_svy=predict(svy,type=c("response"))

roc_glm<- roc(enal$wfa2 ~ prob_svy, plot=TRUE,
  print.auc = TRUE,legacy.axes = TRUE,
  main="Curva ROC",
  xlab=expression(paste("1 - ", hat(Sp), "= 1 -
  ",hat(P),"[ ",hat(y),"=0", " |
  y=0 ]")), ylab=expression(paste(hat(Se),
  "=", hat(P) ,"[ ",hat(y),"=1", " | y=1 ]")))

####Significancia del modelo
```

```
#Modelo nulo
Nulo_svy=svyglm(wfa2~1, design=dis, data=enal,
              family = quasibinomial)

anova(Nulo_svy, svy, test="Chisq") # p < 2.22e-16



##Intervalos de confianza para los coeficientes del
## modelo
confint(svy,type="Wald")

##se(beta)/beta
#Extracción de desviaciones estándar
summ_svy=summary(svy)
se_svy=as.data.frame(summ_svy$coefficients)$`Std. Error`

se_svy/svy$coefficients

#####Momios
exp(svy$coefficients)

###########  Intervalos 95% de confianza para los momios.
## |Cuantil| = 1.96

#Lím ínf
exp(svy$coefficients-(1.96*se_svy))

#Lim sup
exp(svy$coefficients+(1.96*se_svy))


#### Gráfica de los intervalos del 95% de
#### confianza de los coeficientes del modelo

#Límite inferior del intervalo
#El cuantil 0.975 es 1.96
liminf=double(length(svy$coefficients))

for (i in 1:length(liminf)) {
  liminf[i]=svy$coefficients[i]-
    (1.96*se_svy[i])
}

#Límite superior del intervalo

limsup=double(length(svy$coefficients))
```

```
for (i in 1:length(limsup)) {
  limsup[i]=svy$coefficients[i]+
    (1.96*se_svy[i])
}


#Gráfica

ggplot()+ geom_point(aes(x=factor(coef,levels=coef),
 svy$coefficients)) + geom_errorbar(aes(x=factor(coef,
        levels=coef),ymin=liminf, ymax=limsup,
     width=5), width=0.2)+ xlab("") + ylab("CI") +
  theme_bw() +  theme(axis.text.x  =
          element_text(angle=45, vjust=0.5,
                       size=10))


####Comparación
#1
M_intercepto$coefficients/M_final$coefficients
svy$coefficients/M_final$coefficients
svy$coefficients/M_intercepto$coefficients

#2
(1-(M_intercepto$coefficients/M_final$coefficients))*100
(1-(svy$coefficients/M_final$coefficients))*100
(1-(svy$coefficients/M_intercepto$coefficients))*100

#Desviaciones estándar
###1
M_intercepto$coef.sd/se_glm
se_svy/se_glm
se_svy/M_intercepto$coef.sd

###2
#Con intersección aleatoria vs sin efectos
(1-((M_intercepto$coef.sd/se_glm)^2))*100
#Survey vs sin efectos aleatorios
(1-((se_svy/se_glm)^2))*100
#Survey vs Efectos
(1-((se_svy/M_intercepto$coef.sd)^2))*100



## Anexo III: Proporción de madres con niños con desnutrición vs
#edad de la madre
```

```
Propor=double(length(12:50))

for(i in 12:50){
  Propor[i-11]=mean(enal$wfa2[enal$agemoth==i])
}


ggplot()+ geom_line(aes(x=12:50,y=Propor))+
  geom_point(aes(x=12:50,y=Propor)) +
xlab("Edad de la madre") +
ylab("Proporción de madres con niños con desnutrición")+
theme(axis.text.x  = element_text(angle=45, vjust=0.5,
                                  size=10))
Propor

#Frecuencias de las edades (enteras)
frec_agem=double(length(12:50))

for(i in 12:50){
frec_agem[i-11]=sum(enal$agemoth==i)
}

frec_agem
### Intervalos de confianza para los coeficientes
#en una sóla gráfica


plotCI(x= (1:10), y=M_intercepto$coefficients,
       uiw=1.96*M_intercepto$coef.sd, pch=19,
       cex=0.63, xaxt="n", ylab="", xlab="",xlim=c(1,11), ylim=c(-1,0.12))

plotCI(x= (1:10)-0.17, y=M_final$coefficients, uiw= 1.96*se_glm, pch=19,
       cex=0.63, xaxt="n", ylab="", xlab="", add=TRUE, col="red")

plotCI(x= (1:10)+0.17, y=svy$coefficients,
       uiw=1.96*se_svy, pch=19,
       cex=0.63, xaxt="n", ylab="", xlab="",add=TRUE, col="blue")



#add axis
axis(1, at = c(1:10), labels = c("Intercept" , "foodexp" ,
                                 "persroom","agemoth" , "floor01", "wc01",
                                 "cooker01", "schoolm01", "languagem01",
                          "foodexp:schoolm01"),
     tick = TRUE, col = "black", cex.axis =1)
```

```
abline(h=0)

legend("bottomright",lty=c(1,1,1),lwd=2,col=c("red","black","blue"),
       legend=c("Sin efectos","Con intersección", "Ponderado"), cex=0.65)
```

Bibliografía

[1] Agresti, A. (2013) *Categorical data analysis*, tercera edición, New Jersey: John Wiley & Sons Inc.

[2] Agresti, A. (2015) *Foundations of linear and generalized linear models*, primera edición, Hoboken, New Jersey: John Wiley & Sons Inc.

[3] Albert, A. y Anderson, J. (1984) *On the existence of maximum likelihood estimates in Logistic regression models*. Biometrika, 71(1), 1–10.

[4] Avila Curiel, A. Shamah, T, y Chávez, A. (1997) *Encuesta Nacional de Alimentación y Nutrición en el Medio Rural 1996*. Instituto Nacional de la Nutrición Salvador Zubirán

[5] Cox, D. y Hinkey, D. (1974) *Theoretical statistics*, primera edición, Londres: Chapman & Hall.

[6] Firth, D. (1993) *Bias reduction of maximum likelihood estimates*, Biometrika 80, 27–38.

[7] Eslava, G. (2002). *The analysis of binary clustered sample survey data*, Cph: Copenhagen Business School, Dept. of Management Science and Statistics.

[8] Eslava, G. y Tjur, T. (2008) *A simple two-stage method as an alternative to random effects models for binary clustered data*, Reportes de investigación, Departamento de Matemáticas, Facultad de Ciencias, No 3–08.

[9] Hastie, T., Tibshirani, R., Witten, D. y James G. (2013) *An introduction to statistical learning: with applications in R*, primera edición, Springer Texts in Statistics.

[10] Heeringa, S., West, B. y Berglund, P. (2017) *Applied survey data analysis*, segunda edición, Chapman & Hall/CRC.

[11] Kendall, M. y Stuart, A. (1961) *The advanced theory of statistics* (Vol. 2), tercera edición, Hafner Publishing Company.

[12] Lumley, T. (2010) *Complex surveys: A guide to analysis using R*, primera edición, New Jersey: John Wiley & Sons, Inc.

[13] Lumley, T. y Scott, A. (2015) *AIC and BIC for modeling with complex survey data*, Journal of Survey Statistics and Methodology, 3, 1-18.

[14] Lumley, T. y Scott, A. (2017) *Fitting regression models to survey data*, Statistical Science 32: 265–278.

[15] Zuur, A., Ieno, E.N., Walker, N., Saveliev, A.A. y Smith, G.M. (2009) *Mixed effects models and extensions in ecology with R*, Statistical Science 32: 265–278