Notas de clase de Probabilidad y Estadística

Volumen 8: Regresión logística binaria

Versión 2 (Julio, 2019)

Dr. rer. nat. Humberto LLinás Solano
Doctor en Estadística (Mainz-Alemania)
Profesor Titular
Investigador Asociado (Colciencias)
hllinas@uninorte.edu.co

Departamento de Matemáticas y Estadística
Universidad del Norte
(www.uninorte.edu.co).

ÍNDICE GENERAL

Prefacio

Introducción

Estas notas de clase hacen parte de un compendio de varios volúmenes y están dirigido a todo tipo de público que requiere de algún conocimiento básico en Estadística.

El autor

Humberto Jesús LLinás Solano es Licenciado en Ciencias de la Educación, con énfasis en Matemáticas, Física y Estadística de la Universidad del Atlántico (Colombia). Magister en Matemáticas, convenio Universidad del Valle-Universidad del Norte (Colombia). Doctor en Estadística (Dr. rer. nat.) de la Universidad Johannes Gutenberg de Mainz (Alemania). Desde 1998 se desempeña como profesor de tiempo completo de la Universidad del Norte y forma parte de los grupos de investigación Matemáticas y Enfermedades tropicales de dicha institución. Autor de los productos[1]:

- *Estadística descriptiva y distribuciones de probabilidad* (2005, [6])
- *Estadística inferencial* (2006, [9])
- *Una visión histórica del concepto moderno de integral* (como editor, 2006, [4])
- *Medida e integración* (2007, [10])
- *Applets de estadística* (2007, [12])
- *Introducción a la estadística con aplicaciones en Ciencias Sociales* (2012, [13])
- *Procesos estocásticos con aplicaciones* (como coautor, 2013, [2])
- *Introducción a la estadística matemática* (2014, [14])
- *Introducción a la teoría de la probabilidad* (2014, [15])

[1] Se cita el título del texto o applet, el año de publicación y la referencia bibliográfica respectiva. Cuando sea necesario, un comentario adicional.

1 El modelo logístico

1.1 Preliminares

1.1.1. Introducción

Los modelos logísticos son adecuados para situaciones donde se quiere explicar la probabilidad p de ocurrencia de un evento de interés por medio de los valores de ciertas variables explicativas. Si se asocia al evento de interés una variable dicotómica, entonces, ésta es una variable de Bernoulli con esperanza condicional p. Es un tipo particular de los modelos lineales generalizados, abreviados por MLG.

Además, con base en una teoría asintótica para las ML-estimaciones, se han encontrado aproximaciones para diferentes deviaciones, es decir, para (-2) veces los logaritmos de las ML-funciones. Estas se usan para obtener diferentes pruebas de hipótesis estadísticas que tienen distribuciones asintóticas chi-cuadrada.

Por una parte, en algunos libros se mencionan estos resultados dando sólo pocos detalles. En [1] ya se encuentran más detalles con mayor enfoque para el caso de variables explicativas categóricas. De todas formas, falta el desarrollo detallado de la teoría asintótica de ML-estimaciones para el caso de variables independientes y no idénticamente distribuidas. En otros libros clásicos de Estadística Matemática, se detalla sólo el caso de variables independientes e idénticamente distribuidas. Esto último no se presenta en MLG.

1.1.2. Conceptos básicos de la estadística

El lector debe estar familiarizado conlos siguientes términos,entre otros (ver LLINÁS [6] y [9]):

- Variable aleatoria
- Esperanza y varianza de una variable aleatoria
- Distribuciones de Bernoulli, Binomial, normal y Chi-cuadrada
- Distribuciones exactas y asintóticas
- Independencia de variables aleatorias

- Varianza de una suma de variables aleatorias
- Método de máxima verosimilitud
- Estimador vs Estimación
- Intervalos de confianza y pruebas de hipótesis

1.1.3. La función logaritmo y la logística

La función logística viene dada por

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

y su inversa es la función logit:

$$g(x) = Logit(x) = \ln\left(\frac{x}{1-x}\right)$$

Ejercicio. Gráfique de las siguientes funciones. Además, encuentre su dominio y rango.

- La función logartimo natural: $f(x) = \ln x$.
- La función logística.
- La función logit.

1.1.4. Logístico vs Lineal vs Anova

No.	TEMA	LOGISTICO	LINEAL	ANOVA
1.	Y	Categórica discreta (de Bernoulli)	Numérica continua	Numérica continua
2.	X:	Categórica o numérica	Numérica	categórica con niveles
3.	Explican:	$P(Y=1\|\star) = E(Y=1\|\star)$	$E(Y\|\star)$	$E(Y)$
4.	Método de Estimación	Máxima verosimilitud (ML)	Mínimos cuadrados (LS)	Mínimos cuadrados (LS)
5.	Estimación de parámetros y pruebas de hipótesis	Sí requiere teoría asintótica	No requiere teoría asintótica	No requiere teoría asintótica
6.	Estadísticos	Distribución χ^2	Distribución F	Distribución F

Cuadro 1.1: Comparación de los modelos lineal, logístico y anova

1.2 Modelos logísticos y modelos relacionados

1.2.1. Ejemplo introductorio

Ejemplo 1.1

Ingrese a la página [18]. Descargue y abra el archivo de datos **chdage** que corresponde a un estudio realizado para investigar las causas de enfermedades coronarias. La lista de las variables que contiene se muestra abajo:

Columnas	Variable	Abreviación
1-3	Código de identificación	ID
9-10	Edad	Age
17	Enfermedad coronaria	CHD
		(0=ausente, 1=presente)

Use Statgraphics para explorar la relación entre las variables **Age** y **CHD** siguiendo las instrucciones que se proponen abajo.

1. Escoja la opción *Plot - Scatterplots - XY Plots - ....*, realize un diagrama de dispersión de **Age** versus **CHD** y responda las siguientes preguntas:
   (a) ¿Tendencia para los individuos con evidencia de **CHD**?
   (b) ¿Describe claramente este diagrama la naturaleza dicotómica de **CHD**?
   (c) ¿Provee este diagrama una clara relación entre las dos variables?
   (d) ¿Es pequeña la variabilidad en **CHD** en todas las edades?
   (e) ¿Qué es lo que dificulta describir la relación entre las dos variables?
   (f) ¿Qué posible método podemos utilizar para remover alguna variación pero sin modificar la posible relación entre las dos variables?
2. Escoja la opción *Describe - Numeric Data - One Variable Analysis - ....* y construya una tabla de frecuencias agrupadas para la variable **Age**, digamos con 9 clases, comenzando con (dato menor - 1) y terminando con (dato mayor + 1) y analízela.
3. Categorize los intervalos de clases (las edades) y defina una nueva variable **Age-Cat**. Mediante la opción *Describe - Categorical Data - Crosstabulation - .....* realize una tabulación cruzada entre **Age-Cat** y **CHD** (por lo menos debe aparecer frecuencia y proporción de individuos con presencia de **CHD**). ¿Es esta proporción creciente o decreciente a medida que la edad aumenta?
4. Ahora, con la opción *Plot - Scatterplots - XY Plots - ....* realize un diagrama de dispersión de las marcas de clases de los intervalos de clases (en el eje X) versus esta proporción.
   (a) ¿Provee este diagrama una clara relación entre las dos variables?
   (b) ¿Cuál podría ser esta relación? ◄

1.2.2. El modelo de Bernoulli

La variable de interés Y es de Bernoulli. En símbolo, $Y \sim \mathscr{B}(1,p)$, siendo $p := E(Y) = P(Y=1)$ la probabilidad de que ocurra Y.

Haciendo n observaciones independientes de Y, se obtiene la muestra $Y = (Y_1,\ldots,Y_n)$ con los datos $y_i \in \{0,1\}$, $i = 1,\cdots,n$, donde y_i es un posible valor de Y_i, las cuales son independientes entre sí.

Se llega a un modelo estadístico de Bernoulli:

$$Y_i = p_i + e_i \quad \sim \quad \mathscr{B}(1,p_i), \quad i = 1,\cdots,n.$$

Fijando $y = (y_1,\cdots,y_n)^T$ obtenemos la función de verosimilitud en el parámetro $p = (p_1,\cdots,p_n)^T$:

$$L(p) = \prod_{i=1}^{n} [p_i^{y_i}(1-p_i)^{1-y_i}]$$

El logaritmo de la función de máxima verosimilitud será:

$$\mathscr{L}(p) := \ln L(p) = \sum_{i=1}^{n} [y_i \ln p_i + (1-y_i)\ln(1-p_i)] \tag{1.1}$$

Como $0 \le f(y,p) \le 1$, se tiene que $-\infty \le \mathscr{L}(p) \le 0$.

Hay varias situaciones que se pueden presentar en un modelo de Bernoulli. Se dice que éste se puede identificar como alguno de los siguientes modelos: completo, nulo o saturado.

1.2.3. El modelo completo

El *modelo completo* es caracterizado por el supuesto de que todos p_i, $i = 1,\cdots,n$ son considerados como parámetros.

Teorema 1.1

En el modelo completo, las ML-estimaciones de p_i son $\hat{p}_i = Y_i$ con valores $\hat{p}_i = y_i$, $\quad i = 1,\cdots,n$. Además, $\mathscr{L}_c := \mathscr{L}(y) = 0$.

Ejemplo 1.2

Para los datos del archivo **chdage** (consúltese [18]), en el modelo completo, se tiene que $\mathscr{L}(y) = 0$. ◄

1.2.4. El modelo nulo

El *modelo nulo* es caracterizado por el supuesto de que todos los p_i, $i = 1,\cdots,n$ son considerados iguales; es decir, se tiene un solo parámetro $p = p_i, i = 1,\cdots,n$. En este caso, (1.1) será:

$$\mathscr{L}(p) \quad = \quad n[\overline{y}\ln p + (1-\overline{y})\ln(1-p)] \tag{1.2}$$

Teorema 1.2

En el modelo nulo, la ML-estimación de p es $\hat{p} = \overline{Y}$ con valor $\hat{p} = \overline{y}$. Además, $\mathscr{L}_o := \mathscr{L}(\overline{y}) < 0$ si y sólo si $0 < \overline{y} < 1$.

Ejemplo 1.3

Para los datos del archivo **chdage** (consúltese [18]), en el modelo nulo:

(a) $\hat{p} = \overline{y} = 43/100 = 0,43$

(b) $\mathscr{L}(\hat{p}) = \mathscr{L}(0,43) = -68,3315$ ◄

1.2.5. El modelo saturado y supuesto

El modelo saturado está caracterizado por los siguientes supuestos:

1. Se supone que:

   a) Se tienen K variables explicativas $X_1, \cdots, X_K$ (algunas pueden ser numéricas y otras categóricas) con valores $x_{1i}, \cdots, x_{Ki}$ para $i = 1, \cdots, n$ (fijadas u observadas por el estadístico, según sean variables determinísticas o aleatorias).

   b) Entre las n kuplas $(x_{1i}, \cdots, x_{Ki})$ de los valores de la variable explicativa X haya J kuplas diferentes, definiendo las J poblaciones. Por tanto, $J \leq n$.

   Notación Para cada población $j = 1, \cdots, J$ se denota:

   - el número de observaciones Y_{ij} en cada población j por n_j, siendo $n_1 + \cdots + n_J = n$;
   - la suma de las n_j observaciones Y_{ij} en j por $Z_j := \sum_{i=1}^{n_j} Y_{ij}$ con valor $z_j = \sum_{i=1}^{n_j} y_{ij}$, siendo $\sum_{j=1}^{J} z_j = \sum_{i=1}^{n} y_i$.

   Para mayor simplicidad en la escritura, se abreviará la j-ésima población $(x_{1j}, \cdots, x_{Kj})$ por el símbolo $\star$.

2. Para cada población $j = 1, \cdots, J$ y cada observación $i = 1, \cdots, n$ en j, se supone que:

   - $(Y_{ij}|\star) \sim \mathscr{B}(1, p_j)$
   - Las variables $(Y_{ij}|\star)$ son independientes entre sí
   - $p_j = P(Y_{ij} = 1|\star) = E(Y_{ij}|\star)$ y $V(Y_{ij}|x_j) = p_j(1 - p_j)$

   A continuación, se oprimirá el símbolo $\star$.

   El supuesto 2 implica:

   a) Todos los p_{ij}, $i = 1, \cdots, n$ dentro de cada población j son iguales. Es decir, se tiene como parámetro el vector $p = (p_1, \cdots, p_J)^T$.

b) Para cada población $j = 1, \cdots, J$:

- $Z_j \sim \mathscr{B}(n_j, p_j)$
- Las variables Z_j son independientes entre las poblaciones

En el modelo saturado, el logaritmo de la función de máxima verosimilitud será

$$\begin{aligned} \mathscr{L}(p) &= \sum_{j=1}^{J} \left(\sum_{i=1}^{n_j} [y_{ij} \ln p_j + (1 - y_{ij}) \ln(1 - p_j)] \right) \\ &= \sum_{j=1}^{J} [z_j \ln p_j + (n_j - z_j) \ln(1 - p_j)] \end{aligned} \tag{1.3}$$

Teorema 1.3

En el modelo saturado, las ML-estimaciones de p_j son $\tilde{p}_j = \frac{Z_j}{n_j}$, con valores $\tilde{p}_j = \frac{z_j}{n_j}$, $\quad j = 1, \cdots, J$. Además,

$$\mathscr{L}_s := \mathscr{L}(\tilde{p}) = \sum_{j=1}^{J} n_j [\tilde{p}_j \ln \tilde{p}_j + (1 - \tilde{p}_j) \ln(1 - \tilde{p}_j)] \tag{1.4}$$

También se cumple: $\mathscr{L}_s < 0$ para $0 < \tilde{p}_j < 1$.

Ejemplo 1.4

Para los datos del archivo **chdage** (consúltese [18]), en el modelo saturado,hay $J = 43$ poblaciones y se cumple que $\mathscr{L}(\tilde{p}) = -41,7994$. ◄

1.3 El modelo logístico

1.3.1. Supuestos

Se hacen los supuestos 1 y 2 de la sección 1.2.5, donde adicionalmente se supone que la matriz de diseño

$$C = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1K} \\ 1 & x_{21} & \cdots & x_{2K} \\ \vdots & \vdots & & \vdots \\ 1 & x_{J1} & \cdots & x_{JK} \end{pmatrix}$$

tiene rango completo $Rg(C) = 1 + K \leq J$. Para llegar a un modelo logístico se hace el supuesto adicional

$$3. \quad \text{Logit}(p_j) := \ln\left(\frac{p_j}{1 - p_j}\right) = \delta + \beta_1 x_{j1} + \cdots + \beta_K x_{jK} \tag{1.5}$$

Sea $\alpha = (\delta, \beta_1, \ldots, \beta_K)^T$ el vector de parámetros en el modelo.

Nótese que el supuesto sobre $Rg(C) = 1 + K$, hace identificable al parámetro α.

Sea $g_j := \delta + \beta_1 x_{1j} + \cdots + \beta_K x_{Kj}$. Entonces:

$$p_j = \text{Logit}^{-1}(g_j) = \frac{e^{g_j}}{1 + e^{g_j}} \tag{1.6}$$

El logaritmo de la función de verosimilitud se puede escribir en función de α como:

$$\mathscr{L}(\alpha) = \sum_{j=1}^{J} z_j g_j - \sum_{j=1}^{J} n_j \ln\left[1 + e^{g_j}\right]. \tag{1.7}$$

1.3.2. Relaciones entre el logístico y el saturado

Las ecuaciones del supuesto 3 de la sección 1.3 se pueden escribir así:

$$\begin{pmatrix} \text{Logit}(p_1) \\ \text{Logit}(p_2) \\ \vdots \\ \text{Logit}(p_J) \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1K} \\ 1 & x_{21} & \cdots & x_{2K} \\ \vdots & \vdots & & \vdots \\ 1 & x_{J1} & \cdots & x_{JK} \end{pmatrix} \cdot \begin{pmatrix} \delta \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} = C\alpha,$$

Con base en lo anterior, se pueden distinguir los dos siguientes casos:

1. $J = 1 + K$
   En este caso, C es una matriz invertible. Por lo tanto,

   $$\alpha = C^{-1} \cdot \begin{pmatrix} \text{Logit}(p_1) \\ \vdots \\ \text{Logit}(p_J) \end{pmatrix}$$

   Es decir, hay una relación uno a uno entre los parámetros del modelo saturado y los del logístico. O sea, los dos modelos expresan lo mismo.

   Particularmente, las ML-estimaciones de las probabilidades p_j son iguales en ambos modelos: $\hat{p}_j = \tilde{p}_j$ para cada $j = 1, 2, \ldots, K$.

2. $J > 1 + K$
   En este caso, primero hay que calcular $\hat{\alpha}$ y a partir de éstas, se pueden calcular las $\hat{p}_j$ mediante:

   $$\hat{p}_j = \text{Logit}^{-1}(\hat{g}_j), \quad j = 1, \cdots, J,$$

   donde $\hat{g}_j := \hat{\delta} + \hat{\beta}_1 x_{j1} + \cdots + \hat{\beta}_K x_{jK}$. En general, resultan que $\hat{p}_j \neq \tilde{p}_j$.

Ejemplo 1.5

Considere la siguiente tabla de datos:

CHD	Edad	Talla
1	15	130
1	15	130
0	15	130
1	20	140
1	20	140
1	20	140
0	20	140
1	20	140
0	30	165
1	30	165
0	30	165
1	30	165

Halle $\hat{p}_j$, $\tilde{p}_j$ y $\hat{\alpha}$ utilizando el método explicado en 1.3.2. Observe que

$$C = \begin{pmatrix} 1 & 15 & 130 \\ 1 & 20 & 140 \\ 1 & 30 & 165 \end{pmatrix}, \qquad C^{-1} = \begin{pmatrix} -36 & 57 & -20 \\ -1 & 1,4 & -0,4 \\ 0,4 & -0,6 & 0,2 \end{pmatrix}$$

son las correspondientes matriz de diseño y su inversa. ◄

1.3.3. Casos agrupado y no agrupado

1. Cuando se trabaja con el modelo saturado, se tiene el caso de utilizar *datos agrupados*.
2. Cuando se tiene el caso especial $n_j = 1$, para todo j (lo que implica que $J = n$) se habla de *datos no agrupados*.
3. La distinción entre datos agrupados y no agrupados es importante por dos razones:
   a) Algunos métodos de análisis apropiados a datos agrupados no son aplicables a datos no agrupados.
   b) Las aproximaciones asintóticas pueden estar basados en uno de estos dos casos distintos:
   - $n \to \infty$
   - o $J \to \infty$, caso que es únicamente es apropiado para datos no agrupados.
4. En la práctica:
   - Cuando se tienen datos agrupados es importante tener en cuenta que J *debe ser fijo*. Por esta razón, debe tomarse como base el modelo saturado. Es decir, se empieza el análisis usando los vectores Z_j, $j = 1, \cdots, J$.
   - Si $J \to \infty$ (por ejemplo, si $J = n$), entonces, en el modelo saturado no se puede considerar a J como fijo. Obsérvese que esta situación se presenta cuando se tienen datos no agrupados. En este caso, no se puede tomar como base el modelo saturado. Ahora se empezaría el análisis utilizando, de una vez, las observaciones Y_i, $i = 1, \cdots, n$.

1.3.4. Estimación de los parámetros logísticos

El método que se propone para calcular las ML-estimaciones en un modelo logístico es *el método iterativo de Newton-Raphson*. Generalmente, el método requiere:

- Una estimación inicial para el valor que maximiza la función.
- La función es aproximada en una vecindad de aquella estimación por un polinomio de segundo grado.
- Entonces, la siguiente estimación se calcula como el máximo de dicho polinomio.
- Luego, se repite el proceso, usando esta estimación como la estimación inicial.
- De esta manera, el método genera una sucesión de estimaciones. Estas estimaciones convergen a la localización del máximo cuando la función es adecuada y/o la estimación inicial es buena.

Para más detalles, ver teorema 8 en LLINÁS [7].

Ejemplo 1.6

Para los datos del archivo **chdage** (ver [18]):

(a) $\widehat{\beta} = 0,111$

(b) $\widehat{\delta} = -5,309$

(c) El logit estimado para un sujeto de edad 50 es $0,240$

(d) Proporción estimada de personas con presencia de CHD a la edad de 50:
$\widehat{P}(Y = 1/x = 50) = 0,560$

(e) El error estándar de $\widehat{\beta}$ es $\widehat{S}_{\widehat{\beta}} = 0,0240593$

(f) El error estándar de $\widehat{\delta}$ es $\widehat{S}_{\widehat{\delta}} = 1,13363$

(g) $\mathscr{L}(\hat{\alpha}) = -53,677$ ◄

1.3.5. ODDS

El cociente

$$O_j = \frac{p_j}{1-p_j},$$

es llamado ODDS. Siempre: $O_j > 0$.

Es la proporción entre las probabilidades de ocurrencia y no ocurrencia del evento que se relaciona con Y en la población j.

Ejemplo 1.7

Para los datos del archivo **chdage** (ver [18]), el odds estimado para individuos con edad de 50 es $0,560/(1-0,560) = 1,28$. ◄

1.3.6. Razones ODDS

El cociente

$$OR(x_i \text{ vs } x_j) = \frac{O_i}{O_j} = \frac{\frac{p_i}{1-p_i}}{\frac{p_j}{1-p_j}}$$

es llamado RAZÓN ODDS. Siempre: $RR(i,j) > 0$.

Se puede demostrar que

$$OR(i,j) = e^{\beta_1(x_{1i}-x_{1j}) + \beta_2(x_{2i}-x_{2j}) + \cdots + \beta_K(x_{Ki}-x_{Kj})} \tag{1.8}$$

Cuando $x_{ki} - x_{kj} = 1$ para todo k, entonces

$$OR := OR(i,j) = e^{\beta_1 + \cdots + \beta_K}$$

no depende de $X_1, \ldots, X_K$ y muestra el cambio proporcional en la variable de respuesta cuando las variables independientes se incrementen en 1 unidad.

Ejemplo 1.8

Para los datos del archivo **chdage** (ver [18]), la razón odds estimada cuando el incremento la edad se incrementa en 1 año es 1,1173. ◄

1.4 Intervalos de confianza

1.4.1. Intervalo de confianza para la pendiente β_k

Un intervalo de confianza del $(1-\alpha)100\,\%$ para β_k es:

$$\widehat{\beta}_k - Z_{\alpha/2}\,\widehat{S}_{\widehat{\beta}_k} < \beta_k < \widehat{\beta}_k + Z_{\alpha/2}\,\widehat{S}_{\widehat{\beta}_k}$$

Aquí, $\widehat{S}_{\widehat{\beta}_k}$ es el error estándar del estimador $\widehat{\beta}_k$.

Ejemplo 1.9

Para los datos del archivo **chdage** (consúltese [18]), Un intervalo de confianza del 95 % para β es: $(0,064; 0,158)$. ◄

1.4.2. Intervalo de confianza para el intercepto δ

Un intervalo de confianza del $(1-\alpha)100\,\%$ para δ es:

$$\widehat{\delta} - Z_{\alpha/2}\,\widehat{S}_{\widehat{\delta}} < \delta < \widehat{\delta} + Z_{\alpha/2}\,\widehat{S}_{\widehat{\delta}}$$

Aquí, $\widehat{S}_{\widehat{\delta}}$ es el error estándar del estimador $\widehat{\delta}$.

Ejemplo 1.10

Para los datos del archivo **chdage** (consúltese [18]), Un intervalo de confianza del 95 % para δ es: $(-7,531; -3,087)$. ◄

1.4.3. Intervalo de confianza para $Logit(p_j)$

Para una población $x_{j1}, \ldots, x_{jK}$ dada, un intervalo de confianza del $(1-\alpha)100\,\%$ para $Logit(p_j)$ es:

$$Logit(\widehat{p_j}) - Z_{\alpha/2}\,\widehat{S}_{Logit(\widehat{p_j})} \quad < \quad Logit(p_j) \quad < \quad Logit(\widehat{p_j}) + Z_{\alpha/2}\,\widehat{S}_{Logit(\widehat{p_j})}$$

Aquí, $\widehat{S}_{Logit(\widehat{p_j})}$ es el error estándar del estimador $Logit(\widehat{p_j})$.

Recordar: Para una población $x_{j1}, \ldots, x_{jK}$ dada, el estimador de la varianza de $Logit(\widehat{p_j})$ es

$$\widehat{Var}\big(Logit(\widehat{p_j})\big) = \widehat{Var}\big(\widehat{\delta}\big) + \sum_{i=1}^{K} x_{ji}^2\,\widehat{Var}\big(\widehat{\beta_i}\big) + 2\sum_{i=o}^{K}\sum_{k=i+1}^{K} x_{ji}\,x_{jk}\,\widehat{Cov}\big(\widehat{\beta_i}, \widehat{\beta_k}\big)$$

Aquí $\beta_0 := \delta$ y $x_{j0} := 1$.

Ejemplo 1.11

Para los datos del archivo **chdage** (consúltese [18]):

(a) Varianza estimada del estimador de la pendiente: $0,000579$

(b) Varianza estimada del estimador del intercepto: $1,28517$

(c) Covarianza estimada entre los estimadores del intercepto y de la pendiente: $-0,026677$

(d) Varianza estimada del estimador del logit a la edad de 50: $0,0650$

(e) Error estándar estimado del logit a la edad de 50: $0,2549$

(f) Un intervalo de confianza del 95 % para el logit a la edad de 50 viene dado por: $(-0,260; 0,7345)$ ◄

1.4.4. Intervalo de confianza para p_j

Para una población $x_{j1}, \ldots, x_{jK}$ dada, un intervalo de confianza del $(1-\alpha)100\,\%$ para p_j es:

$$\frac{e^{Logit(\widehat{p_j}) - Z_{\alpha/2}\,\widehat{S}_{Logit(\widehat{p_j})}}}{1 + e^{Logit(\widehat{p_j}) - Z_{\alpha/2}\,\widehat{S}_{Logit(\widehat{p_j})}}} \quad < \quad p_j \quad < \frac{e^{Logit(\widehat{p_j}) + Z_{\alpha/2}\,\widehat{S}_{Logit(\widehat{p_j})}}}{1 + e^{Logit(\widehat{p_j}) + Z_{\alpha/2}\,\widehat{S}_{Logit(\widehat{p_j})}}}$$

Aquí, $\widehat{S}_{Logit(\widehat{p_j})}$ es el error estándar del estimador $Logit(\widehat{p_j})$.

Ejemplo 1.12

Para los datos del archivo **chdage** (consúltese [18]), un intervalo de confianza del 95 % para la proporción de individuos con presencia de CHD a la edad de 50 es: $(0,435; 0,677)$. ◄

1.4.5. Intervalo de confianza para O_j

Para una población $x_{j1},\ldots,x_{jK}$ dada, un intervalo de confianza del $(1-\alpha)100\,\%$ para O_j es:

$$e^{Logit(\widehat{p_j}) - Z_{\alpha/2}\,\widehat{S}_{Logit(\widehat{p_j})}} \quad < \quad O_j \quad < e^{Logit(\widehat{p_j}) + Z_{\alpha/2}\,\widehat{S}_{Logit(\widehat{p_j})}}$$

Aquí, $\widehat{S}_{Logit(\widehat{p_j})}$ es el error estándar del estimador $Logit(\widehat{p_j})$.

Ejemplo 1.13

Para los datos del archivo **chdage** (consúltese [18]), un intervalo de confianza del 95 % para el odds a la edad de 50 es: $(0,771;2,096)$. ◄

1.4.6. Intervalo de confianza para OR

Un intervalo de confianza del $(1-\alpha)100\,\%$ para OR es:

$$e^{\widehat{\beta} - Z_{\alpha/2}\,\widehat{S}_{\widehat{\beta}}} \quad < \quad OR \quad < e^{\widehat{\beta} + Z_{\alpha/2}\,\widehat{S}_{\widehat{\beta}}}$$

Aquí, $\widehat{S}_{\widehat{\beta}}$ es el error estándar del estimador $\widehat{\beta}$.

Ejemplo 1.14

Para los datos del archivo **chdage** (consúltese [18]):

(a) Una estimación de la razón odds (para incremento de 1 año) es 1,117.

(b) Un intervalo de confianza del 95 % para la razón odds poblacional es: $(1,065;1,172)$. ◄

1.5 Pruebas de comparación de modelos y selección de modelos

En esta sección se presentan estadísticas para distintas pruebas de comparación de modelos:

- H_0: Logístico vs H_1: Saturado,
- H_0: Nulo vs H_1: Logístico,
- H_0: Logístico vs H_1: Completo,
- H_0: Submodelo vs H_1: Logístico,
- H_0: Submodelo con una variable explicativa menos vs H_1: Logístico.

Estas estadísticas tienen distribución asintótica chi-cuadrada.

1.5.1. Comparación de un modelo logístico con el modelo saturado correspondiente

Teorema 1.4

La LR-estadística de prueba (según el método de cocientes de funciones de verosimilitud) para la hipótesis

H_0: el modelo logístico (con $X_1, \ldots, X_K$),
vs la alternativa
H_1: el modelo saturado correspondiente (con sus J poblaciones)

es equivalente a la llamada deviación que tiene el modelo logístico del modelo saturado

$$D^*(M) := 2\ln\left(\frac{L(\tilde{p})}{L(\hat{\alpha})}\right) = 2[\mathscr{L}(\tilde{p}) - \mathscr{L}(\hat{\alpha})]$$

la cual tiene distribución asintótica chi-cuadrada con $\nu = J - (1+K)$ grados de libertad cuando $n \to \infty$ y J es fijo.

Observaciones:

1. Aquí se requiere que $J > 1 + K$.

2. Para el caso en que $J = 1 + K$, el análisis en un modelo logístico es el mismo que en el modelo saturado.

3. Esta prueba únicamente se cumple para datos agrupados porque J es fijo (lo que no sucede para el caso de datos no agrupados).

4. Se espera que esta prueba no rechace H_0 (p-valor alto), o sea, que los datos obtenidos no estén en contra del modelo logístico. Es decir, que al pasar del modelo saturado al modelo logístico no se pierde información estadísticamente significativa.

Ejemplo 1.15

Para los datos del archivo **chdage** (consúltese [18]):

(a) $\mathscr{L}(\tilde{p}) = -41,7994$

(b) $\mathscr{L}(\hat{\alpha}) = -53,677$

(c) Un valor del estadístico de prueba es: $D^*(M) = 2[-41,7994 - (-53,677)] = 23,7552$

(d) El estadístico $D^*(M)$ tiene distribución asintótica chi-cuadrada con $\nu = 43-2 = 41$ grados de libertad (hay $J = 43$ poblaciones).

(e) P-valor $= P(\chi^2_{41} > 23,7552) = 0,9857$

(f) No se rechaza H_0 (es decir, el modelo logístico vale). ◄

1.5.2. Comparación de un modelo logístico con el modelo nulo

Teorema 1.5

Para la hipótesis

H_0: el modelo nulo (sólo con el intercepto),
vs la alternativa
H_1: el modelo logístico (con $X_1, \ldots, X_K$)

la estadística de prueba es

$$D^*(0) = 2\ln\left(\frac{L(\hat{\alpha})}{L(\hat{\delta}_o)}\right) = 2[\mathscr{L}(\hat{\alpha}) - \mathscr{L}(\hat{\delta}_o)]$$

y tiene distribución asintótica chi-cuadrada con K grados de libertad cuando $J \to \infty$.

Aquí $\hat{\delta}_o = logit(\overline{Y})$ es la estimación de δ en el modelo nulo.

Observaciones:

1. La hipótesis es equivalente a la hipótesis $H_o : \beta = 0$.
2. Esta prueba sólo es válida para el caso de datos no agrupados ($J = n$).
3. En el trabajo práctico, se espera que esta prueba sí rechace H_o (p-valor bajo). Es decir, que las variables explicativas $X_1, \ldots, X_K$ del modelo logístico, tiene una explicación más informativa que sólo el intercepto.
4. En caso contrario (si la prueba no rechaza H_0), que no es muy común en problemas prácticos, se tendría que chequear otro modelo logístico con más o con otras variables.

Ejemplo 1.16

Para los datos del archivo **chdage** (consúltese [18]):

(a) $\delta_o = \overline{y} = 43/100 = 0,43$

(b) $\mathscr{L}(\hat{\delta}_o) = \mathscr{L}(0,43) = -68,3315$

(c) $\mathscr{L}(\hat{\alpha}) = -53,677$

(d) Un valor del estadístico de prueba es: $D^*(0) = 2[-53,677 - (-68,3315)] = 29,309$

(e) El estadístico $D^*(M)$ tiene distribución asintótica chi-cuadrada con $\nu = 1$ grado de libertad.

(f) P-valor $= P(\chi^2_1 > 29,309) \approx 0$

(g) Se rechaza H_0 (es decir, el modelo logístico vale). ◄

1.5.3. Comparación de un modelo logístico con el modelo completo

Teorema 1.6

Para la hipótesis

H_0: el modelo logístico (con $X_1,\ldots,X_K$),
vs la alternativa
H_1: el modelo completo (que no se basa en poblaciones)

la estadística de prueba es

$$D(M) := 2\ln\left(\frac{L(\hat{p})}{L(\hat{\alpha})}\right) = 2[\mathscr{L}(\hat{p}) - \mathscr{L}(\hat{\alpha})] = -2\,\mathscr{L}(\hat{\alpha})$$

y tiene distribución asintótica chi-cuadrada con $\nu = n-(1+K)$ grados de libertad cuando $n \to \infty$.

Observaciones:

1. Se espera que esta prueba no rechace H_0 (p-valor alto), o sea, que los datos obtenidos no estén en contra del modelo logístico. Es decir, que al pasar del modelo completo al modelo logístico no se pierde información estadísticamente significativa.

Ejemplo 1.17

Para los datos del archivo **chdage** (consúltese [18]):

(a) $\mathscr{L}(\hat{p}) = 0$

(b) $\mathscr{L}(\hat{\alpha}) = -53,677$

(c) Un valor del estadístico de prueba es: $D^*(M) = -2(-53,677) = 107,354$

(d) El estadístico $D^*(M)$ tiene distribución asintótica chi-cuadrada con $\nu = n-2 = 98$ grados de libertad (hay $n = 100$ observaciones).

(e) P-valor $= P(\chi^2_{98} > 107,354) = 0,2434$

(f) No se rechaza H_0 (es decir, el modelo logístico vale). ◄

1.5.4. Comparación de un modelo logístico con algún submodelo

Teorema 1.7

Para la hipótesis

H_0: un submodelo logístico con $X_1,\cdots,X_{\tilde{K}}$,
vs la alternativa
H_1: el modelo logístico con $X_1,\cdots,X_K$ con $\tilde{K} < K$,

(a) La estadística de prueba es equivalente a la estadística

$$D^*(L) := 2\log\left(\frac{L(\hat{\alpha})}{L(\hat{\alpha}_o)}\right) = 2[\mathscr{L}(\hat{\alpha}) - \mathscr{L}(\hat{\alpha}_o)]$$

Aquí: $\hat{\alpha} = (\hat{\delta}, \hat{\beta}_1, \cdots, \hat{\beta}_K)^T$ es la ML-estimación en el modelo logístico de la alternativa H_1 y $\hat{\alpha}_o = (\hat{\delta}_o, \hat{\beta}_{o1}, \cdots, \hat{\beta}_{o\tilde{K}})^T$ es la ML-estimación en el submodelo logístico de la hipótesis H_0.

(b) $D^*(L)$ tiene distribución asintótica chi-cuadrada con $K - \tilde{K}$ grados de libertad cuando $J \to \infty$.

(c) Para la situación anterior, una estadística asintóticamente equivalente es la de Wald:

$$W(L) := \hat{\gamma}^T \cdot \hat{Cov}^{-1}(\hat{\gamma}) \cdot \hat{\gamma}$$

la cual también tiene distribución asintótica chi-cuadrada con $K - \tilde{K}$ grados de libertad cuando $J \to \infty$.

Aquí $\hat{\gamma}$ es la estimación de γ, que es la parte $(K - \tilde{K})$-dimensional del vector α que se anula bajo H_0 y $\hat{Cov}(\hat{\gamma})$ es la matriz de covarianzas estimada de $\hat{\gamma}$.

Observación:

1. Nótese que la hipótesis de la primera parte del teorema es equivalente a la hipótesis $H_o : \gamma = 0$.
2. Esta prueba sólo es válida para datos no agrupados. Aunque, también, es posible realizarla teniendo en cuenta el modelo saturado. Pero, como en la prueba únicamente se considera el modelo logístico, no tiene mucho sentido comparar éste con un submodelo teniendo que pasar por el modelo saturado.
3. Se espera que no se rechace la prueba (p-valor alto).

1.6 Comparación de un modelo logístico con un submodelo que tiene una variable explicativa menos

Teorema 1.8

Para la hipótesis

H_0: el submodelo (con $X_1, \cdots, X_K$ sin un X_k),
vs la alternativa
H_1: el modelo logístico (con $X_1, \cdots, X_K$)

se puede tomar, alternativamente, una de las dos estadísticas de pruebas siguientes (ambas tienen distribución asintótica chi-cuadrada con 1 grado de libertad cuando $J \to \infty$):

(a) $2\log\left(\frac{L(\hat{\alpha})}{L(\hat{\alpha}_o)}\right) = 2[\mathscr{L}(\hat{\alpha}) - \mathscr{L}(\hat{\alpha}_o)]$,

donde $\hat{\alpha}_o$ es la estimación bajo H_o

(b) $\frac{\hat{\beta}_k^2}{\hat{V}(\hat{\beta}_k)} = \left(\frac{\hat{\beta}_k}{SE(\hat{\beta}_k)}\right)^2$,

siendo $\hat{\beta}_k$ la estimación de β_k, para cada $k = 0,1,\cdots,K$ en el modelo (bajo H_1) con su varianza estimada $\hat{V}(\hat{\beta}_k)$ y su error estándar $SE(\hat{\beta}_k) = \sqrt{\hat{V}(\hat{\beta}_k)}$

Observación:

1. En este teorema se está considerando el caso de datos no agrupados.
2. Con base en todas las pruebas parciales, para cada $k = 0,1,\cdots,K$ se eliminará la variable explicativa que menor aporte individual tenga en la explicación. Es decir, la variable que tenga p-valor parcial más alto. Así, se sigue eliminando variable tras variable hasta que se rechacen todas las pruebas parciales (todas las tengan p-valores bajos).

1.7 Análisis de deviaciones(ANODEV)

ANODEV para un modelo logístico con respecto al modelo saturado correspondiente

Con el fin de analizar la bondad de un modelo logístico fijo (con variable explicativa X que define J poblaciones), se le compara hacia los dos lados, es decir, con el modelo saturado correspondiente (con estas J poblaciones) y con el modelo nulo (con sólo el intercepto).

Para esta situación, puede orientarse en la llamada tabla de ANODEV (en inglés: **AN**alysis **O**f **DEV**iance), en analogía a la ANOVA para modelos lineales.

Esta tabla es la 1.2 (la notación DF viene del inglés **D**egree of **F**reedom).

Teorema		DF	Estadístico
1.5.2	Diferencia de deviaciones	K	$D^*(0) - D^*(M) = 2[\mathscr{L}(\hat{\alpha}) - \mathscr{L}(\hat{\delta}_0)]$
1.5.1	Deviación del modelo logístico	$J-(1+K)$	$D^*(M) := 2[\mathscr{L}(\tilde{p}) - \mathscr{L}(\hat{\alpha})]$
	Deviación total (del modelo nulo)	$J-1$	$D^*(0) := 2[\mathscr{L}(\tilde{p}) - \mathscr{L}(\hat{\delta}_0)]$

Cuadro 1.2: ANODEV del modelo logístico vs saturado

ANODEV para un modelo logístico con respecto al modelo completo

Si se quiere comparar dos o más modelos logísticos, se propone analizar la bondad de un modelo logístico fijo (con variables explicativas $X_1,\ldots,X_K$ que define J poblaciones), al compararlo con el modelo completo (que no se basa en poblaciones) y con el modelo nulo (sólo con el intercepto).

Recuerde que el logaritmo de la función de máxima verosimilitud del modelo completo es igual a cero: el máximo valor posible (es decir, $\mathscr{L}(\hat{p}) = 0$).

En este caso, la deviación del modelo logístico será:

$$D(M) := 2[\mathscr{L}(\hat{p}) - \mathscr{L}(\hat{\alpha})] = -2[\mathscr{L}(\hat{\alpha})]$$

Esta situación se orienta en la tabla 1.3 de ANODEV.

Teorema		DF	Estadístico
1.5.2	Diferencia de deviaciones	K	$D(0) - D(M) = 2[\mathscr{L}(\hat{\alpha}) - \mathscr{L}(\hat{\delta}_0)]$
1.5.3	Deviación del modelo logístico	$n-(1+K)$	$D(M) := -2[\mathscr{L}(\hat{\alpha})]$
	Deviación total (del modelo nulo)	$n-1$	$D(0) := -2[\mathscr{L}(\hat{\delta}_0)]$

Cuadro 1.3: ANODEV del modelo logístico vs completo

1.8 Criterio para la escogencia de un buen submodelo logístico

1.8.1. El análisis de un modelo logístico M fijo

Se compara este modelo logístico (con sus 2 parámetros) con el modelo saturado correspondiente (con J poblaciones/ parámetros) y con el modelo nulo (con 1 parámetro, el intercepto), respectivamente.

Para que el modelo logístico pueda considerarse como aceptable, debe estar cerca del modelo saturado y lejos del modelo nulo. Es decir,

1. No debe ser rechazado el modelo logístico vs el saturado, y
2. Sí debe ser rechazado el modelo nulo vs el logístico.

1.8.2. De un modelo logístico M hacia un buen submodelo logístico M_o

Al eliminar sucesivamente variable tras variable, empezando con un modelo inicial, se llega a submodelos. Para escoger el mejor submodelo logístico, se debe comparar cada submodelo con:

1. Su modelo saturado (el número de las poblaciones baja) y con el modelo nulo, según lo mencionado anteriormente. Como criterio sirve el teorema 1.6. Es decir, cada eliminación debe llevar a un submodelo que no esté rechazado vs el modelo anterior.

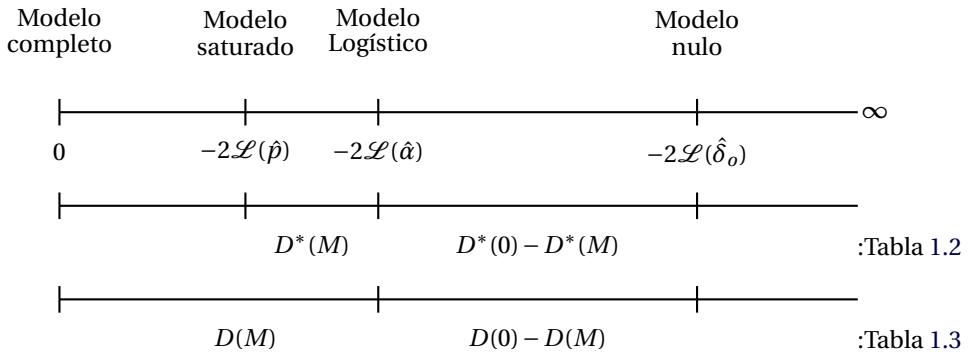

Figura 1.1: Gráfica para el análisis de un modelo logístico M fijo

2. El modelo inicial, según la hipótesis y con la estadística dada en el teorema 1.5.4.

3. La sucesión de todos los submodelos anteriores. Para esto no sirve la estadística RDS, como se observó en la sección 1.7. En este caso, se debe calcular la deviación relativa RDC para cada submodelo. De esta manera, se obtiene una sucesión (finita) decreciente de valores que sirve como un criterio (junto al anterior) para decidir cuándo y porqué se detiene el proceso de eliminación. Es decir, para decidir cuál submodelo puede ser mejor que los anteriores, incluso, que el modelo inicial.

La situación final puede visualizarse en la figura 1.2.

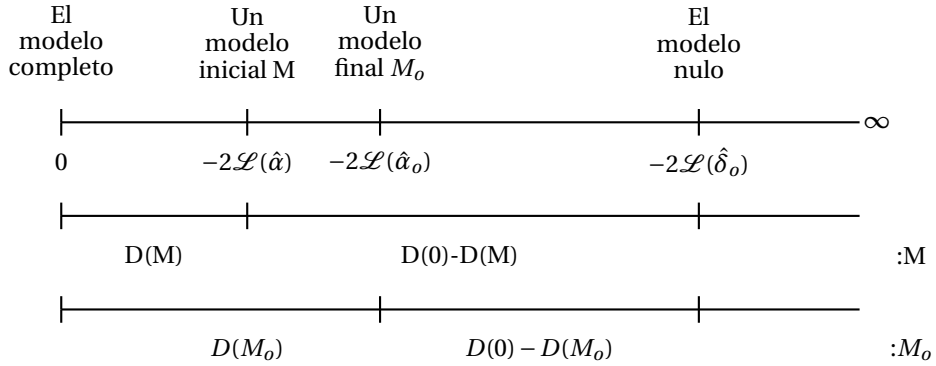

Figura 1.2: Gráfica para escoger el mejor submodelo logístico M_o

1.9 ✍ Ejercicios

Para la solución de los siguientes ejercicios, téngase en cuenta los siguientes comentarios:

- Todos los datos mencionados aparecen en [18].
- Siempre debe detallar el análisis del conjunto de datos (con las variables especificadas) basado en lo realizado en el capítulo 1.
- Utilize Statgraphics y Excel para sus cálculos.
- Los resultados que no presente Statgraphics deben ser calculados con ayuda de Excel.

- Verifique cómo Statgraphics obtiene las estimaciones correspondientes, los logaritmos de las funciones de máxima verosimilitud, los estadísticos de pruebas, los p-valores, razones odds, intervalos de confianza (para pj, ODDS, razones ODDS, intercepto, pendiente, ...), etc.

1.1 Demuestre el teorema 1.2.3.

1.2 Demuestre el teorema 1.2.4.

1.3 Demuestre el teorema 1.2.5.

1.4 Los datos **ICU** corresponden a una muestra de 200 sujetos que hicieron parte de un estudio de supervivencia de pacientes que fueron remitidos a una unidad de cuidados intensivos (intensive care unit - ICU). La meta principal de este estudio fue desarrollar un modelo de regresión logístico para predecir la probabilidad de supervivencia de estos pacientes en el hospital y estudiar los factores de riesgos asociados con el índice de mortalidad ICU. En estos datos tome a la variable **AGE** como independiente y **STA** como dependiente.

(a) Escriba la ecuación general para el modelo de regresión logístico de **STA** contra **AGE** y para el logit transformado de este modelo. ¿Qué características de **STA** nos pone a pensar que debamos considerar el modelo de regresión logístico en vez del usual modelo de regresión lineal para describir la relación entre **STA** y **AGE**?

(b) Forme un diagrama de dispersión de **STA** contra **AGE**.

(c) Usando los intervalos [15,24], [25,34], [35,44], [45,54], [55,64], [65,74], [75,84], [85,94] para **AGE**, calcule la media de **STA** de los sujetos dentro de cada intervalo. Grafique estos valores de la media de **STA** contra el punto medio del intervalo de **AGE** usando el mismo conjunto de ejes que se utilizaron en la parte (b).

(d) Escriba una expresión para la función de verosimilitud y del logaritmo de esta función para el modelo de regresión logístico de (a) usando los 200 datos no agrupados. Obtenga una expresión para las dos ecuaciones de verosimilitud.

(e) Usando Statgraphics obtenga las estimaciones de los parámetros del modelo de regresión logístico de (a). Usando estas estimaciones, escriba las correspondientes ecuaciones para los valores ajustados. Grafique la ecuación para los valores ajustados utilizando los mismos ejes como en (b) y (c).

(f) Resuma (describa en palabras) los resultados presentados en la gráfica obtenida en (b), (c) y (e).

(g) Usando los resultados de la salida de Statgraphics usada para la parte (e), verifique la significancia del coeficiente de AGE. ¿Qué supuestos se necesitan para realizar dicha prueba?

(h) Usando los resultados de (e), halle un intervalo del 95 % de confianza para la pendiente y la constante. Escriba una interpretación con respecto al intervalo encontrado para la pendiente.

(i) Obtenga la matriz de covarianzas estimada para el modelo en (e). Calcule el logit y la probabilidad logística estimada para una persona de 60 años. Calcule un intervalo del 95 % de confianza para el logit y la probabilidad logística estimada. Interprete la probabilidad estimada y su intervalo de confianza.

(j) Use un paquete estadístico para obtener el logit estimado y su error estándar para cada persona en el estudio ICU. Grafique el logit estimado y los límites del intervalo del 95 % de confianza versus AGE para cada persona. Explique (en palabras) similaridades y diferencias entre las apariencias de esta gráfica y una gráfica de una gráfica de un modelo de regresión ajustado y sus límites del intervalo del 95 % de confianza.

1.5 Considere los datos **ICU**. Repita el ejercicio 1.4 utilizando la variable **TYP** (como variable dependiente) en vez de **STA**.

1.6 Considere los datos **ICU**. Repita todos los análisis realizados en el capítulo 1, pero considerando ahora las variables **AGE** (como variable independiente) y **STA** (como variable dependiente).

1.7 Considere los datos **ICU**. Haga el análisis correspondiente tomando a **STA** como variable dependiente y a **AGE**, **SYS** y **HRA** como independientes.

1.8 Detalle el análisis para los datos **UIS** tomado a **DFREE** como variable dependiente y **AGE**, **BECK** y **NDRUGTX** como variables independientes.

1.9 Los datos **PROS** corresponden a un estudio realizado pacientes con cáncer de próstata para determinar si las variables medidas en un examen básico pueden ser usadas para predecir si el tumor ha penetrado la cápsula prostática. Los datos fueron recogidos teniendo en cuenta 380 individuos, 153 de los cuales tuvieron un cáncer que penetró la cápsula prostática. En estos datos, una variable que se pensó que era particularmente predicitva para para la penetración de cápsula es el nível de antígeno prostático, **PSA**. Repita los pasos (a)-(g) y (j) del ejercicio 1.4 usando **CAPSULE** como variable dependiente y utilize para **PSA** los intervalos [0,0; 2,4], [2,5; 4,4], [4,5; 6,4], [6,5; 8,4], [8,5; 10,4], [10,5; 12,4], [12,5; 20,4], [20,5; 140].

1.10 De todas las variables que aparecen en los datos **PROS** sólo considere a **CAPSULE** (como variable dependiente) y **PSA** (como variable independiente).

(a) Responda:

I. ¿Cuál es la ecuación para el modelo de regresión logística?

II. ¿Cuál es la ecuación para la transformación logit de este modelo?

III. ¿Qué características de la variable dependiente nos conduce a considerar la regresión logística como más apropiada que el modelo de regresión lineal para describir la relación entre las dos variables mencionadas anteriormente?

(b) Calcule:

I. $\mathscr{L}(\widehat{p})$ en el modelo completo.

II. $\mathscr{L}(\widehat{p})$ en el modelo nulo.

III. $\mathscr{L}(\widetilde{p})$ en el modelo saturado.

IV. $\mathscr{L}(\widehat{\alpha})$ en el modelo logístico.

(c) Construya intervalos del 95 % de confianza para los siguientes parámetros e interprételos (justifique en forma clara y precisa todas sus afirmaciones):

I. La pendiente β. ¿Es apropiado el modelo?

II. El intercepto δ. ¿Pasa la curva de regresión logística por el origen?

III. $P(CAPSULE = 1 \,/\, PSA = 11{,}2 \text{ mg/ml})$.

IV. $P(CAPSULE = 0 \,/\, PSA = 11{,}2 \text{ mg/ml})$.

V. La razón odds OR. ¿Es PSA estadísticamente significativa en el modelo?

(d) Realize las siguientes pruebas de comparación de modelos resumiendo en una tabla las pruebas realizadas, el valor y la distribución muestral del estadístico de prueba, los grados de libertad, el P-valor y su decisión.

I. Nulo vs Logístico.

II. Logístico vs Completo.

III. Logístico vs Saturado.

1.11 Considere los datos **PROS**. Realize un análisis de regresión logística tomando a **CAPSULE** como variable dependiente y **VOL** como variable independiente.

1.12 Considere los datos **PROS**. Realize un análisis de regresión logística tomando a **CAPSULE** como variable dependiente y **AGE** como variable independiente.

1.13 Considere los datos **PROS**, tomando a **CAPSULE** como variable dependiente y **AGE**, **PSA** y **VOL** como variables independientes.

1.14 Los datos **LOWBWT** corresponden a un estudio realizado para identificar factores de riesgos asociados a nacimientos de bebés con bajo peso (peso menor que 2.500 gramos). Los datos fueron recogidos teniendo en cuenta 189 mujeres, 59 de las cuales tuvieron bebés con bajo peso y 130 de las cuales tuvieron bebés con peso normal. De todas las variables que aparecen sólo considere a **LOW** (como variable dependiente) y **LWT** (como variable independiente).

(a) Responda:

I. ¿Cuál es la ecuación para el modelo de regresión logística?

II. ¿Cuál es la ecuación para la transformación logit de este modelo?

III. ¿Qué características de la variable dependiente nos conduce a considerar la regresión logística como más apropiada que el modelo de regresión lineal para describir la relación entre las dos variables mencionadas anteriormente?

(b) Calcule:

I. $\mathscr{L}(\widehat{p})$ en el modelo completo.

II. $\mathscr{L}(\widehat{p})$ en el modelo nulo.

III. $\mathscr{L}(\widetilde{p})$ en el modelo saturado.

IV. $\mathscr{L}(\widehat{\alpha})$ en el modelo logístico.

(c) Construya intervalos del 95 % de confianza para los siguientes parámetros e interprételos (justifique en forma clara y precisa todas sus afirmaciones):

I. La pendiente β. ¿Es apropiado el modelo?

II. El intercepto δ. ¿Pasa la curva de regresión logística por el origen?

III. $P(LOW = 1 \,/\, LWT = 100,3$ libras).

IV. $P(LOW = 0 \,/\, LWT = 100,3$ libras).

V. La razón odds OR. ¿Es LWT estadísticamente significativa en el modelo?

(d) Realize las siguientes pruebas de comparación de modelos resumiendo en una tabla las pruebas realizadas, el valor y la distribución muestral del estadístico de prueba, los grados de libertad, el P-valor y su decisión.

I. Nulo vs Logístico.

II. Logístico vs Saturado.

III. Logístico vs Completo.

1.15 Considere los datos **LOWBWT**, tomando a **LOW** como variable dependiente y **AGE** como variable independiente.

1.16 Considere los datos **LOWBWT**, tomando a **LOW** como variable dependiente y **LWT** como variable independiente.

1.17 Considere los datos **LOWBWT**, tomando a **LOW** como variable dependiente y **BWT** como variable independiente.

1.18 Considere los datos **LOWBWT**, tomando a **LOW** como variable dependiente y **AGE**, **LWT** y **BWT** como variables independientes.

1.19 Considere los datos **LOWBWTM11**, tomando a **LOW** como variable dependiente y **AGE** como variable independiente.

1.20 Considere los datos **LOWBWTM11**, tomando a **LOW** como variable dependiente y **LWT** como variable independiente.

1.21 Considere los datos **LOWBWTM11**, tomando a **LOW** como variable dependiente y **AGE** y **LWT** como variables independientes.

1.22 En los datos **CLSLOWBWT** una variable que los físicos consideraron importante para el control el peso del bebé (variable dependiente **LOW**) fue el peso de la madre durante el último periodo menstrual, **LWT**. Repita los pasos (a)-(g) del ejercicio 1.4, pero para la parte (c) utilize los intervalos [80,99], [100,109], [110,114], [115,119], [120,124], [125,129], [130,250].

(h) La gráfica en la parte (c) no parece en forma de S. La razón principal es que el rango de los valores graficados está aproximadamente entre 0,2 y 0,56. Explique por qué un modelo para la probabilidad de LOW como una función de LWT pudiese ser el modelo de regresión logístico.

1.23 Considere los datos **CLSLOWBWT**, tomando a **LOW** como variable dependiente y **AGE** como variable independiente.

1.24 Considere los datos **CLSLOWBWT**, tomando a **LOW** como variable dependiente y **LWT** como variable independiente.

1.25 Considere los datos **CLSLOWBWT**, tomando a **LOW** como variable dependiente y **BWT** como variable independiente.

1.26 Considere los datos **CLSLOWBWT**, tomando a **LOW** como variable dependiente y **AGE**, **LWT** y **BWT** como variables independientes.

2 El caso de variables independientes discretas

2.1 Introducción

Si alguna de las variables independientes son discretas o variables en escala nominal tales como sexo, raza, sexo, grupo de tratamiento, etc., es inapropiado incluirlas en el modelo como si ellas fuesen variables en escala de intervalo. Los números usados para representar los diferentes niveles son identificaciones y no tienen significado numérico. En esta situación, el método de escogencia es usar una colección de variables de diseño (o variables dummy).

Raza	D_1	D_2
Blanca	0	0
Negra	1	0
Otro	0	1

Cuadro 2.1: Codificación de las variables de diseño para **Raza** (con tres niveles)

Por ejemplo, suponga que una de las variables independientes es raza que ha sido codificado como blanco, negro y otro. En este caso, se necesitan dos variables de diseño, digamos, D_1 y D_2 (véase la tabla 2.1).

En general, si una variable escala nominal tiene k posibles valores, entonces se necesitan $k-1$ variables de diseño.

Para ilustrar la notación usada para las variables de diseño en estas notas, suponga que la j-ésima variable independiente x_j tiene k_j niveles. Las $k_j - 1$ variables de diseño serán denotadas por D_{jl} y los coeficientes para estas variables de diseño como β_{jl}. Por consiguiente, el logit para un modelo probabilístico con k variables y la j-ésima variable discreta será

$$g(x) := \text{Logit}(p_j) = \delta + \beta_1 x_1 + \cdots + \sum_{m=1}^{k_j-1} \beta_{jm} D_{jm} + \beta_K x_K \tag{2.1}$$

2.2 Variable independiente dicotómica

Consideraremos el caso donde el modelo contiene sólo una variable independiente y que ésta es nominal y dicotómica (es decir, medida en dos niveles).

Asumamos que la variable independiente x está codificada como uno o cero. Entonces el logit para una proporción con $x = 1$ y $x = 0$ es

$$g(1) - g(0) = (\delta + \beta_1) - \delta = \beta_1$$

El álgebra mostrado en esta ecuación es muy importante. Se presenta en este nivel de detalles para enfatizar que el primer paso al interpretar el efecto de la variable independiente en un modelo es expresar la diferencia de logit deseada en términos del modelo. En este caso, la diferencia de logit es igual a β_1. Para interpretar este resultado necesitamos introducir y discutir la *razón odds*.

Los posibles valores de las probabilidades logísticas pueden ser convenientemente organizados en una tabla de 2×2 como se muestra en la tabla 2.2. Para simplificar la notación, hemos usado la notación $\pi(x) := P(Y = 1/X = x)$ para representar la probabilidad condicional de $Y = 1$ dado $X = x$ cuando se utiliza la regresión logística y, como ya se sabe, viene dada por

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

En la tabla 2.2 vemos sus valores.

	$x = 1$	$x = 0$
$y = 1$	$\pi(1) = \frac{e^{\delta+\beta_1}}{1+e^{\delta+\beta_1}}$	$\pi(0) = \frac{e^{\delta}}{1+e^{\delta}}$
$y = 0$	$1 - \pi(1) = \frac{1}{1+e^{\delta+\beta_1}}$	$1 - \pi(0) = \frac{1}{1+e^{\delta}}$
Total	1	1

Cuadro 2.2: Valores de $\pi(x)$ cuando la variable independiente X toma valores $x = 0, 1$

2.2.1. Odds y razón odds

Los ODDS de los resultados cuando están presente individuos con $x = 1$ está definido como

$$O(1) = \frac{\pi(1)}{1 - \pi(1)}$$

Similarmente, los odds de los resultados cuando están presente individuos con $x = 0$ está definido como

$$O(0) = \frac{\pi(0)}{1 - \pi(0)}$$

La RAZÓN ODDS se define como el cociente entre el odds para $x = 1$ y el odds para $x = 0$ y está dada por la ecuación:

$$OR = \frac{O(1)}{O(0)} = \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]} \tag{2.2}$$

Sustituyendo las expresiones para el modelo de regresión logístico mostradas en la tabla 2.2 en (2.2), obtenemos

$$OR = e^{\beta_1} \tag{2.3}$$

Por consiguiente, para la regresión logística con una variable independiente codificada con 0 y 1, la relación entre la razón odds y el coeficiente de regresión es como se muestra en (2.3).

Ejemplo 2.1

(a) Suponga que Y representa la presencia o ausencia de cáncer de hígado y X denota si una persona es fumadora. Entonces un $\widehat{O} = 2$ estima que el cáncer de hígado es dos veces más probable que ocurra entre los fumadores que entre los no fumadores en la población de estudio.

(b) Suponga que Y representa la presencia o ausencia de infarto y X denota si una persona está comprometida a realizar ejercicios físicos de manera extenuante. Si el odds estimado es $\widehat{O} = 0,5$, entonces la ocurrencia de infarto es 0,5 de probable que ocurra entre los que hacen ejercicio que entre aquéllos que no lo hacen en la población de estudio. ◄

Un ejemplo puede ser útil para aclarar qué son las razones odds y cómo se puede calcular de los resultados de un programa de regresión logística o de una tabla de 2×2.

Ejemplo 2.2

Considere los datos del archivo **chdage** (ver [18]). Crear una nueva variable, **AGED**, que toma valores 1 si la edad del sujeto es mayor o igual que 55 y 0 de otro modo. El resultado de cruzar la variable edad dicotomizada **AGED** con la variable de respuesta **CHD** se presenta en la tabla 2.3 (en Statgraphics se escoge la opción *Describe - Categorical Data - Crosstabulation - .....*).

	Edad ≥ 55 ($x = 1$)	Edad < 55 ($x = 0$)	Total
Presente ($y = 1$)	21	22	43
Ausente ($y = 0$)	6	51	57
Total	27	73	100

Cuadro 2.3: Tabla de contingencia de la edad dicotomizada **AGED** (x) a 55 años y **CHD** (y) para 100 sujetos

(a) Para estos datos, la función de verosimilitud en el modelo saturado es

$$L(p) = \pi(1)^{21} \cdot [1 - \pi(1)]^6 \cdot \pi(0)^{22} \cdot [1 - \pi(0)]^{51}$$

(b) A través de Statgraphics (*Relacionar - Datos de atributo - Regresión logística-AGED en factor cuantitativo*) podemos obtener las estimaciones $\widehat{\delta} = -0,841$ y $\widehat{\beta}_1 = \widehat{\beta}_{AGED=1} = 2,09355$. Además, $\widehat{S}_{\widehat{\beta}_1} = 0,5285$ y $\widehat{S}_{\widehat{\delta}} = 0,2551$.

(c) Según la ecuación (2.3), la estimación de la razón de odds es $\widehat{OR} = e^{2,094} = 8,1$. Este valor puede obtenerse directamente de la tabla 2.3 de la siguiente manera:

$$\widehat{OR} = \frac{21/6}{22/51} = 8,11$$ ◄

2.2.2. Intervalo de confianza para razones odds

Un intervalo de confianza del $(1-\alpha)100\,\%$ para OR es:

$$e^{\widehat{\beta}_1 - Z_{\alpha/2}\widehat{S}_{\widehat{\beta}_1}} < OR < e^{\widehat{\beta}_1 + Z_{\alpha/2}\widehat{S}_{\widehat{\beta}_1}}$$

Aquí, $\widehat{S}_{\widehat{\beta}_1}$ es el error estándar del estimador $\widehat{\beta}_1$.

Ejemplo 2.3

Considere los datos del archivo **chdage** (ver [18]). Un intervalo del 95 % para OR es (2,9; 22,9). Este intervalo es sesgado a la derecha porque excede el 1. Entonces con un grado de confianza del 95 %, el CHD ocurre entre aquellas personas con edad mayor o igual que 55 años en la población de estudio por lo menos 2,9 veces o a lo más 22,9 veces más probable que aquéllos por debajo de 55. ◄

2.3 Variable independiente policotómica

En esta sección supondremos que, en vez de dos categorías, la variable independiente tiene $k > 2$ valores diferentes.

Ejemplo 2.4

Suponga que en un estudio de CHD la variable RACE es codificada en cuatro niveles y que la tabulación cruzada de RAZA por CHD produce los datos que aparecen en la tabla 2.4.

	Blanco	Negro	Hispánico	Otro	Total
Presente ($y = 1$)	5	20	15	10	50
Ausente ($y = 0$)	20	10	10	10	50
Total	25	30	25	20	100
Razón odds	1	8	6	4	
IC 95 %		(2,3; 27,6)	(1,7; 21,3)	(1,1; 14,9)	
$\ln(\widehat{OR})$	0	2,08	1,79	1,39	

Cuadro 2.4: Tabla de contingencia de datos hipotéticos sobre RAZA y CHD para 100 sujetos

En esa misma tabla se da la razón odds para cada raza usando Blanco como grupo de referencia. Por ejemplo, para los hispánicos la razón odds estimada es

$$\frac{(15)(20)}{(5)(10)} = 6$$

La referencia de grupo es indicada por un valor de 1 para la razón odds. Estas mismas estimaciones para las razones odds se pueden obtener de un programa de regresión logística (por ejemplo, Statgraphics) con una apropiada escogencia de variables de diseño. El método para especificar las variables de diseño sugiere establecerlas igual a ceros para el grupo de referencia y, luego, establecer 1 a una sola variable de diseño para cada una de los tres grupos. Esto se ilustra en la tabla 2.5.

RAZA (Código)	RAZA2	RAZA3	RAZA4
Blanca (1)	0	0	0
Negra (2)	1	0	0
Hispánica (3)	0	1	0
Otra (4)	0	0	1

Cuadro 2.5: Especificación de las variables de diseño para RAZA usando códigos de referencia y tomando a BLANCA como grupo de referencia

(a) Usando Statgraphics con las variables codificadas que se muestran en la tabla 2.5, obtenemos los resultados que se indican abajo

```
Estimated Regression Model (Maximum Likelihood)
-----------------------------------------------------------------------
                                        tandard    Estimated
Parameter              Estimate          Error      Odds Ratio
-----------------------------------------------------------------------
CONSTANT               -1,38629        1,23153
RAZA_2=1                1,79176        0,632455     0,125
RAZA_3=1                2,0944         0,645497     0,166667
RAZA_4=1                1,38629        0,67082      0,25
-----------------------------------------------------------------------


              Analysis of Deviance
-----------------------------------------------------
Source          Deviance            Df      P-Value
-----------------------------------------------------
Model           14,042                3       0,0028
Residual        124,587            96        0,0265
-----------------------------------------------------
Total (corr.)   138,629            99
```

(b) Observe que el error estándar del coeficiente estimado para la variable de diseño RAZA2 es

$$S_{\hat{\beta}_1} = \left(\frac{1}{5}+\frac{1}{20}+\frac{1}{20}+\frac{1}{10}\right)^{0,5} = 0,6325$$

(c) En general, los límites para un intervalo del $(1-\alpha)100\,\%$ de confianza para los coeficientes son de la forma

$$\hat{\beta}_j \pm Z_{\alpha/2} S_{\hat{\beta}_j}$$

y para las razones odds, son

$$\exp\{\hat{\beta}_j \pm Z_{\alpha/2} S_{\hat{\beta}_j}\}$$

Se deja al lector el cálculo de los intervalos de confianza presentados en la tabla 2.4.

(d) Se deja al lector las pruebas de comparación del modelo logístico con los modelos nulo, saturado y completo. ◄

2.4 ✍ Ejercicios

Para la solución de los siguientes ejercicios, téngase en cuenta los siguientes comentarios:

- Todos los datos mencionados aparecen en [18].

- Siempre debe detallar el análisis del conjunto de datos (con las variables especificadas) basado en lo realizado en el capítulo 2.
- Utilize Statgraphics y Excel para sus cálculos.
- Los resultados que no presente Statgraphics deben ser calculados con ayuda de Excel.
- Para el mejor submodelo, verifique cómo Statgraphics obtiene las estimaciones correspondientes, los logaritmos de las funciones de máxima verosimilitud, los estadísticos de pruebas, los p-valores, razones odds, intervalos de confianza (para pj, ODDS, razones ODDS, intercepto, pendiente, ...), etc.

2.1 Considere los datos **ICU**. Crear una nueva variable, **AGED**, que toma el valor 1 si la edad es mayor o igual que 60 y el valor 0, de otro modo. Repita todos los análisis realizados en el capítulo 2, pero considerando **AGED** como variable independiente y **STA** como variable dependiente.

2.2 Considere los datos **ICU**. Repita los inciso (a), (b), (d), (e) y (g) del ejercicio 1.4 utilizando la variable **TYP** como variable independiente (en vez de **AGE**) y **STA** como variable dependiente.

2.3 Considere los datos **ICU**. Use **STA** como variable dependiente y **CPR** como variable independiente.

(a) Demuestre que el valor del logaritmo de las razones odds obtenido de la tabla de contingencia de **STA** contra **CPR** es idéntico a la pendiente estimada del modelo de regresión logístico de **STA** sobre **CPR**.

(b) Verifique que el error estándar estimado de la pendiente estimada para **CPR** obtenido en Statgraphics es idéntico a la raiz cuadrada de la suma de los inversos de las frecuencias de cada celda de la tabla de contingencia de **STA** contra **CPR**.

(c) ¿Qué aspecto relacionado con la codificación de la variable **CPR** hace que los cálculos para los dos métodos en (a) y (b) sean equivalente.

(d) Obtenga un intervalo del 95 % de confianza para la razón odds.

(e) Para propósitos de ilustración, use una transformación de datos para recodificar (sólo en este inciso) la variable **CPR** como sigue: 4= no y 2=si. Desarrolle la regresión logística de **STA** sobre **CPR** (recodificada). Demuestre cómo los cálculos de la diferencia logit de **CPR**=si versus **CPT**=no es equivalente al valor del logaritmo de la razón odds obtenido en (a). Use los resultados de la regresión logística con el fin de obtener un intervalo del 95 % de confianza para la razón odds y verificar que ellos son los mismos límites obtenidos en (d).

2.4 Considere los datos **ICU**. Use **STA** como variable dependiente y **RACE** como variable independiente.

(a) La variable **RACE** está codificada en tres niveles. Prepare una tabla mostrando la codificación de dos variables de diseño usando como grupo de referencia el valor RACE=1 (White).

(b) Muestre que los logaritmos de las razones odds obtenido de la tabla de contingencia de **STA** contra **RACE**, usando **RACE**=1 como grupo de referencia, son idénticos a las pendientes estimadas para las dos variables de diseño del modelo de regresión logístico de **STA** sobre **RACE**.

(c) Verifique que los errores estándar estimados de la pendiente estimada para las dos variables de diseño del modelo son idénticos a la raiz cuadrada de la suma de los inversos de las frecuencias de cada celda de la tabla de contingencia de **STA** contra **RACE** que se utilizó para calcular las razones odds.

(d) Obtenga un intervalo del 95 % de confianza para las razones odds.

2.5 Considere los datos **ICU**. Use **STA** como variable dependiente y nuevamente **RACE** como covariable.

(a) La variable **RACE** está codificada en tres niveles. Crear variables de diseño para **RACE** usando como grupo de referencia el valor RACE=1 (White).

(b) Desarrolle la regresión logística de **STA** contra **RACE**.

(c) Muestre que las diferencias logit estimadas de RACE=2 versus RACE=1 y RACE=3 versus RACE=1 son equivalentes a los valores del logaritmo de la razón odds obtenidos en el problema 2.4b.

(d) Use los resultados de la regresión logística para encontrar el 95 % de confianza para las razones odds y verifique que son los mismos límites a los encontrados en el problema 2.4d.

(e) Halle la matriz de covarianza estimada y obtenga las varianzas de las diferencias logit estimadas de RACE=2 versus RACE=1 y RACE=3 versus RACE=1. Encuentre los errores estándar correspondientes.

(f) Use los resultados de la regresión logística para obtener un intervalo del 95 % de confianza para las razones odds y verificar que ellos son los mismos límites obtenidos en 2.4d.

2.6 Considere la variable **AGE** en los datos **ICU**. Use **STA** como variable dependiente.

(a) Prepare una tabla mostrando la codificación de tres variables de diseño basadas en los cuartiles empíricos de **AGE** usando el primer cuartil como referencia de grupo.

(b) Ajuste la regresión logística de **STA** sobre **AGE** (recodificada).

(c) Grafique las tres pendientes estimadas versus el punto medio de la edad cuartil respectiva y grafique como un cuarto punto un valor de cero en el punto medio del primer cuartil de la edad. ¿Sugiere esta gráfica que el logit es lineal en la edad?

2.7 Use los datos **ICU** y considere el modelo de regresión logístico multivariado de **STA** (variable dependiente) sobre **AGE**, **CAN**, **CPR**, **INF** y **RACE**.

(a) La variable **RACE** está codificada en tres niveles. Prepare una tabla mostrando la codificación de dos variables de diseño necesaria para incluir esta variable en un modelo de regresión logístico. Tome como referencia la categoría White.

(b) Escriba la ecuación general para el modelo de regresión logístico de STA contra **AGE**, **CAN**, **CPR**, **INF** y **RACE** y para el logit transformado de este modelo. ¿Cuántos parámetros tiene este modelo?

(c) Escriba una expresión general para la función de verosimilitud y del logaritmo de esta función para el modelo de regresión logístico de (b). ¿Cuántas ecuaciones de verosimilitud hay? Obtenga una expresión para una ecuación de verosimilitud típicaÂ´para este problema.

(d) Usando un paquete estadístico obtenga las estimaciones de máxima verosimilitud de los parámetros del modelo de regresión logístico de (b). Usando estas estimaciones, escriba las correspondientes ecuaciones para los valores ajustados. Es decir, las probabilidades logísticas estimadas.

(e) Usando los resultados de la salida de Statgraphics usada para la parte (d), verifique la significancia del coeficiente de AGE. ¿Qué supuestos se necesitan para realizar dicha prueba?

(f) Utilize la estadística de Wald para obtener una aproximación a la significancia individual de los coeficientes de las variables independientes en el modelo. Ajuste un modelo reducido que elimine aquellas variables que no tienen significancia. Ajuste significancia conjunta (condicional) de las variables que quedaron en el modelo.

(g) Usando los resultados de (f), halle un intervalo del 95% de confianza para todos los coeficientes del modelo. Escriba una interpretación con respecto a los intervalos encontrados para las pendientes.

(h) Obtenga la matriz de covarianzas estimada para el modelo final en (f). Escoja un conjunto de valores para las covariables en ese modelo y calcule el logit y la probabilidad logística estimada para una persona con estas características. Calcule un intervalo del 95 % de confianza para el logit y la probabilidad logística estimada. Interprete la probabilidad estimada y su intervalo de confianza.

2.8 Use los datos **PROS** y considere el modelo de regresión logístico multivariado de **CAPSULE** (variable dependiente) sobre **AGE**, **RACE**, **DPROS**, **DCAPS**, **PSA**, **GLEASON** y **VOL**.

(a) La variable **DPROS** está codificada en cuatro niveles. Prepare una tabla mostrando la codificación de tres variables de diseño necesaria para incluir esta variable en un modelo de regresión logístico. Tome como referencia la categoría No nodule.

(b) La variable **DCAPS** está codificada en dos niveles (1 y 2). ¿Puede ser usada esta variable en su codificación original o debe ser creada una variable de diseño? Explore esta pregunta comparando las estimaciones de los coeficientes obtenidas en un modelo que contenga **DCAPS** como se codificó originalmente con aquellas estimaciones encontradas al utilizar una nueva variable de diseño **DCAPSnew**=**DCAPS**-1, la cual está codificada con 0 y 1.

(c) Repita las partes (b)-(h) del ejercicio 2.7.

2.9 Los datos **UIS** se recogieron con el propósito de comparar dos programas de tratamiento A y B para reducir el abuso de la droga y prevenir sus riesgos. Realize un análisis de regresión logística para encontrar el mejor submodelo logístico, tomando a **DFREE** como variable dependiente y al resto como variables independientes.

2.10 Considere los datos **CLSLOWBWT**, tomando a **LOW** como variable dependiente y al resto como variables independientes.

2.11 Considere los datos **LOWBWT**, tomando a **LOW** como variable dependiente y al resto como variables independientes.

A Apéndice

A.1 Pseudo-R-cuadrado

A.1.1. Introducción

Como punto de partida, recordemos que un (no-pseudo) R^2 es un estadístico generado en la regresión ordinaria. A menudo es usado como una medida de bondad de ajuste y es definida como

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}$$

donde n es el número de observaciones en el modelo, $\overline{y}$ es la media de las observaciones y_i y $\widehat{y}_i$ es el valor predicho por el modelo.

Cuando analizamos los datos con el modelo de regresión logística, no existe un R^2 equivalente. Las estimaciones de los parámetros logísticos son estimaciones de máxima verosimilitud obtenidos por procesos iterativos. Ellos no son calculados con el fin de minimizar la varianza, así que no se puede aplicar la aproximación de la regresión ordinaria para la bondad de ajuste. Sin embargo para evaluar la bondad de ajuste de los modelos logísticos se han desarrollado los llamados PSEUDO-R^2, los cuales son medidas basadas en estimaciones de máxima verosimilitud, como ya se explicó anteriormente. Desafortunadamente, algunas de estas medidas no alcanzan 0 ó 1, muchos de ellos pueden tener valores diferentes y es importante reclacar que no se pueden interpretar como se haría con el R^2 de la regresión ordinaria. En la práctica, se habla de un ajuste aceptable para valores desde 0,2 y como bueno de 0,4.

Algunos de los pseudos-R^2 más importantes y utilizados son los de Mc-Fadden, Mc-Fadden ajustado, Cox-Snell, Nagelkerke, enttre otros. Explicaremos a continuación cada uno de ellos.

A.1.2. Pseudo-R^2 de Mc-Fadden

Se calcula con la fórmula:

$$R^2_{MF} = 1 - \frac{\mathscr{L}(\widehat{\alpha})}{\mathscr{L}(\widehat{\alpha}_0)}$$

Aquí $\mathscr{L}(\widehat{\alpha}_0)$ es el Log-verosimilitud en el modelo nulo y $\mathscr{L}(\widehat{\alpha})$ es el Log-verosimilitud en el modelo logístico. Toma sólo valores menores que 1.

A.1.3. Pseudo-R^2 de Mc-Fadden ajustado

En analogía al R^2 ajustado de la regresión lineal (que también tiene en cuenta el número de parámetros de la regresión y el número de casos), se propone la correspondiente versión para el Pseudo-R^2 de Mc-Fadden. Desafortunadamente, al respecto, la discusión no termina allí porque en la literatura podemos encontrar muchas otras propuestas. Una de estas propuestas es la siguiente:

$$R^2_{MFADJ} = 1 - \frac{\mathscr{L}(\widehat{\alpha}) - m}{\mathscr{L}(\widehat{\alpha}_0)}$$

Aquí m es el número de parámetros en el modelo (sin la constante). Esta medida puede tomar valores negativos.

A.1.4. Pseudo-R^2 de Cox-Snell

Se calcula con la fórmula:

$$R^2_{CS} = 1 - \left(\frac{L(\widehat{\alpha}_0)}{L(\widehat{\alpha})}\right)^{2/n}$$

Aquí $L(\widehat{\alpha}_0)$ es la verosimilitud (no el Log-verosimilitud) en el modelo nulo y $L(\widehat{\alpha})$, la del modelo logístico.

A.1.5. Pseudo-R^2 de Mc-Fadden

Se calcula con la fórmula:

$$R^2_N = \frac{R^2_{CS}}{1 - [L(\widehat{\alpha}_0)]^{2/n}}$$

Bibliografía

[1] AGRESTI, A., *Categorical data analysis*. John Wiley and Sons, Inc., New York, 1990.

[2] BARBOSA, R.; LLINÁS, H., *Procesos estocásticos con aplicaciones*, Barranquilla: Editorial Universidad del Norte, 2013.

[3] HOSMER, D. and LEMESHOW S., *Applied Logistic Regression*, Segunda edición, John Wiley and Sons, 2000.

[4] KALB, K. y KONDER, P., *Una visión histórica del concepto moderno de integral*, Barranquilla: Editorial Universidad del Norte, 2006 (editor: Dr. rer. nat. Humberto LLinás).

[5] KLEINBAUM, D. and KLEIN, M., *Logistic Regression: A self Learning Text*, Segunda edición, Ed. Springer, 2002.

[6] LLINÁS, H.; ROJAS, C., *Estadística descriptiva y distribuciones de probabilidad*. Barranquilla: Ediciones Uninorte, 2005.

[7] LLINÁS, H., *Precisiones en la teoría de los modelos logísticos*, Revista Colombiana de Estadística, Volumen 29, Número 2, pág. 239-265, 2006.

[8] LLINÁS, H., *Estadística inferencial*. Barranquilla: Ediciones Uninorte, 2006.

[9] LLINÁS, H., *Estadística inferencial*, Barranquilla: Editorial Universidad del Norte, 2006.

[10] LLINÁS, H., *Medida e integración*. Barranquilla: Editorial Universidad del Norte, 2007.

[11] LLINÁS, H., *Applet: La ley de los grandes números*. Se puede encontrar en el siguiente link: http://ylang-ylang.uninorte.edu.co/Objetos/Estadistica/LeyDeGrandesNumeros/index.html

[12] LLINÁS, H., *Applets de estadística*, 2007. Se puede encontrar en el siguiente link: *http://ylang-ylang.uninorte.edu.co:8080/drupal/?q=node/238*

[13] LLINÁS, H.; ALONSO, J. FLÓREZ, K., *Introducción a la estadística con aplicaciones en Ciencias Sociales*, Barranquilla: Editorial Universidad del Norte, 2012.

[14] LLINÁS, H., *Introducción a la estadística matemática*, Barranquilla: Editorial Universidad del Norte, 2014.

[15] LLINÁS, H., *Introducción a la teoría de probabilidad*, Barranquilla: Editorial Universidad del Norte, 2014.

[16] NELDER, J.A. and WEDDERBURN, R.W.M., *Generalized linear models*. The Journal of the Royal Statistical Society, serie A 135, pág.370-384, 1972.

[17] PÉREZ, C., *Estadística práctica con Statgraphics*. España: Prentice Hall, 2002.

[18] Página web de datos estadísticos del Institute for Digital Research and Education (IDRE) de la Universidad de California en Los Angeles (UCLA): *https://stats.idre.ucla.edu/*. En especial, consultar: *https://stats.idre.ucla.edu/other/examples/alr2/*