90

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

EL USO DE LOS MODELOS DE REGRESION LOGISTICA

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I A

P R E S E N T A :

REGINA FABIOLA PIEDRA VELASCO

DIRECTOR DE TESIS:

M. EN C. INOCENCIO RAFAEL MADRID RIOS

2002

FACULTAD DE CIENCIAS
UNAM

DIVISION DE ESTUDIOS PROFESIONALES

FACULTAD DE CIENCIAS
SECCION ESCOLAR

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

M. EN C. ELENA DE OTEYZA DE OTEYZA
Jefa de la División de Estudios Profesionales de la Facultad de Ciencias
Presente

Comunico a usted que hemos revisado el trabajo escrito:
"El uso de los modelos de regresión logística".

realizado por Regina Fabiola Piedra Velasco

con número de cuenta 9423961-6 **, quién cubrió los créditos de la carrera de:** Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	M. en C. Inocencio Rafael Madrid Ríos
Propietario	Dr. Ignacio Méndez Ramírez
Propietario	Mat. Margarita Elvira Chávez Cano
Suplente	Act. Francisco Sánchez Villarreal
Suplente	Act. María del Rosario Espinosa Tufiño

Consejo Departamental de Matemáticas

FACULTAD DE CIENCIAS
CONSEJO DEPARTAMENTAL DE MATEMATICAS

M. en C. José Antonio Flores Díaz

Agradezco:

A mis padres, por poder contar con ustedes incondicionalmente.

A mi director de tesis, el maestro Rafael Madrid, por la motivación, paciencia y apoyo que me brindó durante el desarrollo de este trabajo.

A mis sinodales, Dr. Ignacio Méndez, Mat. Margarita Chávez, Act. Francisco Sánchez y Act. Ma. del Rosario Espinosa, por sus valiosos comentarios.

A mi alma máter, la Universidad Nacional Autónoma de México, y a quienes han puesto su confianza y apoyo para la realización de este trabajo.

Prefacio

La formación del actuario está sustentada por el lenguaje universal de las matemáticas, herramienta que le permite desenvolverse en varias áreas del conocimiento, planteando y aplicando modelos matemáticos, pero su labor no debe limitarse a ese planteamiento y aplicación, también debe promover la difusión de las técnicas que sean de interés para otros profesionistas y facilitar el trabajo interdisciplinario impidiendo que la información sea una cuestión de elite, buscando además que la recepción de información implique su uso si éste es adecuado.

El uso de los modelos de regresión logística, es el tema de este trabajo de tesis y fue desarrollado para divulgar la regresión logística como una técnica estadística que está siendo aplicada en diversas áreas, su utilidad ha sido aprovechada sobre todo en la medicina, la biología y las ciencias sociales, sin embargo, implícitamente se busca que situaciones de otras áreas puedan ser abordadas con esta técnica.

Durante el desarrollo del trabajo han surgido comentarios y preguntas, referentes a la aplicación del modelo y la interpretación de los resultados, detalles que en la mayoría de los textos pasan por alto, esto permitió estar conscientes de los problemas que muy a menudo se presentan con la aplicación de esta técnica y de los cuales se plantea una solución.

Índice

Apéndices

Introducción

Los modelos de regresión logística permiten estudiar la dependencia de una variable respuesta binomial respecto a otra u otras variables, su objetivo es analizar como cambia la probabilidad de ocurrencia o no ocurrencia de un evento de interés de la respuesta al cambiar los valores de los factores o variables que se propone lo expliquen, es decir, modela la probabilidad de ocurrencia o no ocurrencia de una variable categórica en términos de las variables explicativas. Las variables explicativas pueden ser categóricas (nominales u ordinales) y/o continuas y en este caso sólo nos enfocaremos a modelos con respuesta dicotómica, es decir, los modelos logísticos.

Este trabajo se compone de cuatro capítulos y cuatro apéndices. En el primer capítulo se introduce el modelo de regresión logística de manera intuitiva y el porque el modelo de regresión lineal no es el adecuado cuando la variable respuesta es dicotómica. Una ventaja del uso del modelo de regresión logística es la interpretación de los parámetros, los que una vez estimados informan de la aportación de las variables explicativas a la probabilidad del evento de interés, esto se abordará en el capítulo 2, también se menciona brevemente la inferencia sobre los parámetros y su interpretación cuando teniendo definido el éxito en la variable respuesta, éste se redefine como fracaso. En el capítulo 3 se aborda un ejemplo sobre la supervivencia de recién nacidos, que corresponde a un modelo de regresión logística con dos variables explicativas dicotómicas; también se trabajaron dos ejemplos más para comentar la aplicación del modelo en los bioensayos, desarrollando el concepto de dosis efectiva media y el de covariable. Las conclusiones generales se comentan en el capítulo 4. La versatilidad del modelo se aprecia en el apéndice A referente a la taxonomía de los modelos de regresión logística, es ahí donde se establece el alcance de este trabajo al seleccionar algunos modelos para ser ejemplificados; para completar la estructura del trabajo se introdujeron tres apéndices más que están vinculados con los capítulos 1 y 2.

Lista general de ejemplos

Ejemplo 1.1 Se estudia la recuperación de pacientes con cáncer apartir del nivel de LI que presenten, para introducir el modelo de regresión logística en la sección 1.1 y hacer su comparación con el modelo no lineal correspondiente a la ley de crecimiento logístico en el apéndice 1.1.

Variable respuesta: recuperación o no de pacientes con cáncer

Y=0 el paciente se recupera

Y=1 el paciente no se recupera

Variable explicativa continua: valor de LI que presenta el paciente

Ejemplo 1.2 Se explica la presencia de hipertensión en un individuo con su nivel de triglicéridos, para ilustrar las deficiencias del modelo de regresión lineal con variable respuesta dicotómica en la sección 1.2.

Variable respuesta: presencia o ausencia de hipertensión

Y=0 ausencia de hipertensión

Y=1 presencia de hipertensión

Variable explicativa continua: cantidad de triglicéridos

Ejemplo 2.1 Se estudia la probabilidad de ganar en dos juegos, el lanzamiento de un dado y el de una moneda, para introducir las medidas de asociación en la sección 2.1.

Variable respuesta: resultado del juego

Y=0 perder

Y=1 ganar

Variable explicativa dicotómica: tipo de juego

X=0 apostar a la salida de cierta cara en el lanzamiento de un dado

X=1 apostar a la salida de cierta cara en el lanzamiento de una moneda

Ejemplo 2.2 Supervivencia infantil (Y=1 sobrevive): Modelo de regresión logística con una variable explicativa dicotómica (M2). Se asocia la cantidad de cuidados prenatales con la supervivencia de recién nacidos, para ilustrar las medidas de asociación cuando la presentación de los datos corresponde a una tabla de contingencia de dos dimensiones.

Variable respuesta: supervivencia de recién nacidos

Y=0 el recién nacido muere

Y=1 el recién nacido sobrevive

Variable explicativa dicotómica: cantidad de cuidados prenatales

X=0 pocos cuidados prenatales

X=1 muchos cuidados prenatales

Ejemplo 2.3 Se compara la probabilidad de ganar en dos juegos de lotería, la nacional y la primitiva, para mostrar cuándo el riesgo relativo es aproximadamente igual a la razón de momios.

Variable respuesta: resultado del juego

Y=0 no ser premiado (perder)

Y=1 ser premiado (ganar)

Variable explicativa dicotómica: tipo de lotería que se juega

X=0 lotería primitiva, se premia una combinación de seis números de entre las que se pueden formar con 49 números.

X=1 lotería nacional, el premio es para un número extraído de entre 100,000

Ejemplo 2.4 Anemia aplásica: Modelo de regresión logística con una variable explicativa dicotómica (M2). Se estudia la relación entre la dosis de un injerto de médula y su aceptación o rechazo, para ilustrar la importancia del parámetro $\hat{\beta}_1$ en el tipo de relación entre X y $\hat{Pr}(X)$ en la sección 2.5.

Variable respuesta: rechazo o aceptación de un injerto de médula

Y=0 no rechaza el injerto

Y=1 sí rechaza el injerto

Variable explicativa dicotómica: dosis de médula

X=0 dosis alta

X=1 dosis baja

Ejemplo 3.1 Supervivencia infantil (Y=1 muere): Modelo de regresión logística con dos variables explicativas ambas dicotómicas (M9). Este ejemplo es similar al 2.2, sólo que ahora el éxito se refiere a la muerte del recién nacido y también se considera como variable explicativa la clínica en que se atendió la madre, con esto se ilustra el uso del modelo logístico con fines predictivos.

Variable respuesta: supervivencia de recién nacidos

Y=0 el recién nacido sobrevive

Y=1 el recién nacido muere

Variable explicativa X_1 dicotómica: cantidad de cuidados prenatales

X_1=0 pocos cuidados prenatales

X_1=1 muchos cuidados prenatales

Variable explicativa X_2 dicotómica: clínica en que se atendió la madre

X_2=0 clínica A

X_2=1 clínica B

Ejemplo 3.2 Bioensayos: La dosis efectiva media. Modelo de regresión logística con una variable explicativa continua (M1). Con el efecto que causa en ratones cierta dosis de droga se calcula la dosis efectiva media.

Variable respuesta: mortalidad de ratones

Y=0 el ratón sobrevive

Y=1 el ratón muere

Variable explicativa continua: dosis de droga

Ejemplo 3.3 Bioensayos: El concepto de covariable. Modelo de regresión logística con una variable explicativa dicotómica y una covariable continua (E1). Se trata de un bioensayo en el que se quiere comparar la potencia de dos drogas, pero debido a que las dosis con que se aplican son diferentes se introduce el concepto de covariable.

Variable respuesta: mortalidad de ratones

Y=0 el ratón sobrevive

Y=1 el ratón muere

Variable explicativa dicotómica: tipo de droga

X=0 droga estándar

X=1 droga nueva

Covariable: dosis de droga

Capítulo 1

Modelo de regresión logística

Cuando se requiere estudiar el cambio o la variabilidad de una respuesta bajo estudio que ocurre aleatoriamente y que se supone está relacionada con un conjunto de variables, de inmediato se piensa en estudiar esta relación a través del concepto de asociación con un modelo de regresión lineal, por ser el de mayor uso. Sin embargo en la literatura estadística se reporta que el modelo de regresión logística debe utilizarse cuando la variable respuesta es dicotómica[1], es decir, cuando es categórica y cada unidad bajo estudio puede clasificarse en dos categorías, así el tipo de la variable respuesta según sea continua o categórica determina el uso de alguno de los dos modelos. En este caso nos limitaremos a variables de respuesta dicotómicas, sin embargo, se aclara que cuando la respuesta sea de otra naturaleza, como lo son las variables continuas o las politómicas, y se quiera modelar su asociación con las variables explicativas induce el uso de otros modelos que no son los logísticos para respuesta binaria.

1.1 Introducción intuitiva al modelo de regresión logística

Para introducir el modelo de regresión logística pensemos en la siguiente situación: **Ejemplo 1.1** Uno de los factores que intervienen en la recuperación de pacientes con cáncer es el valor de LI que presenten, notemos que la recuperación o no de los pacientes puede ser indicada por una variable aleatoria binaria Y, y el valor de LI está representado por la variable explicativa X

[1] O bien si la respuesta es politómica, ya sea en una escala nominal u ordinal, el modelo de regresión que se utiliza en este caso es el modelo logístico generalizado.

que en este caso es continua, consideremos que varios pacientes pueden presentar el mismo valor de LI formando una población, entonces para un valor x_i de X, donde x_i corresponde a uno de los valores diferentes de X y existen tantas poblaciones como valores distintos de X existan, se tendrían n_i individuos definiendo una población de valores de Y (Méndez, 1976), donde $Y = 1$ indica recuperación y $Y = 0$ lo contrario, bajo esta idea y_{il} denota la observación $l - ésima$ de la población i, con $l = 1, ..., n_i$ e i que indica la población definida por x_i. Es natural asociar la recuperación de un paciente con su correspondiente valor de LI. Tomando toda la información de los individuos definamos una nueva variable $Z_i = \sum_{l=1}^{n_i} y_{il}$ que indique el número de pacientes que se recuperan con el valor x_i de LI, entonces surgiría la pregunta ¿al aumentar el valor de LI el número de pacientes recuperados aumenta? o de manera más general ¿cómo debe cambiar el valor de LI para que el número de pacientes recuperados aumente?. Por otro lado si tomamos la variable $\frac{Z_i}{n_i}$ estaríamos trabajando con la proporción de pacientes recuperados, la pregunta en este caso sería ¿con qué valores de LI la proporción de recuperaciones aumenta?, es evidente que en el contexto del ejemplo, nuestro interés es la recuperación de los pacientes, por lo que esos valores de LI deben hacer tender a uno a $\frac{Z_i}{n_i}$, y a cero los de $\left(1 - \frac{Z_i}{n_i}\right)$ que es la proporción de no recuperación, esto permite tener dos enfoques para el análsis de la información: la recuperación y la no recuperación según sea el contexto.

Hasta el momento se ha propuesto abordar el análisis de las dos siguientes maneras para la variable respuesta Y:

1. $Z_i = \sum_{l=1}^{n_i} y_{il}$ = número de pacientes que se recuperan dado que su valor de LI es x_i
2. $P_i = \frac{Z_i}{n_i}$ = proporción de pacientes recuperados

Para nuestro ejemplo de los pacientes con cáncer se reportaron los siguientes datos:

i	$LI = X$	P_i	Z_i	n_i
1	0.9	1	1	1
2	1.1	0.5	1	2

Observamos que $Z_1 = Z_2 = 1$, pero no podemos decir que la recuperación de pacientes se da igual para LI=0.9 y LI=1.1 porque los datos no son balanceados, es decir, el número de pacientes para cada valor de LI es distinto, por lo que las Z_i's no son comparables pues es una medida absoluta, entonces sólo cuando las n_i's son iguales las Z_i's como medidas absolutas pueden servir para compararse, pero independientemente de los valores que tengan las n_i's, P_i que es

una medida relativa nos permitirá trabajar para hacer las comparaciones de las proporciones de los diversos valores de X.

Como se ha mostrado, partiendo de una variable Y que tiene distribución Bernoulli (para cada x_i) se ha construido una variable de conteo de éxitos, que se distribuye Binomial, posteriormente se propuso la proporción de éxitos, que es una función de los conteos, ahora de manera general podemos pensar en la construcción de una función de las proporciones. Si trabajamos con la proporción de pacientes recuperados a la que nos referiremos como P_i, se obtendrán irregularidades en la uniformidad de la varianza de los errores (Draper, 1981), por lo que se han propuesto transformaciones sobre P_i que uniformizan la varianza. La transformación angular corresponde a la expresión $arcsen\sqrt{P_i}$ y genera el modelo lineal $arcsen\sqrt{P_i} = \beta_0 + \beta_1 X_i + \varepsilon_i$, las proporciones que caen en el intervalo (0.30, 0.70) tienen una varianza constante, motivo por el que la transformación angular no es necesaria cuando la mayoría de las proporciones observadas caen en ese intervalo, no siendo así para las proporciones del complemento, a las que la transformación angular esparce para aumentar su varianza (Snedecor, 1971); el inconveniente de esta transformación es la interpretación de los resultados, pues al aplicarla sobre la variable respuesta, ya no estamos trabajando el evento de interés y los resultados ya no son interpretables. Como pudo apreciarse la transformación angular es una función de las proporciones, mas no es la única que se ha sugerido, otras transformaciones son la logit y la probit:

3. $\text{logit}(P_i) = \ln(\frac{P_i}{1-P_i})$
4. $\text{probit}(P_i) = \Phi^{-1}(P_i)$

La transformación logit es la que se abordará en este trabajo y como se verá más adelante, en el capítulo dos, permitirá extender el análisis de los datos en facilidad y utilidad en la interpretación de los resultados. Antes de centrarnos en dicha transformación recordemos que no siempre un modelo lineal es el más adecuado para modelar la asociación de variables, motivo por el que se ha desarrollado la regresión no lineal en los parámetros (regresión curvilínea). En este enfoque hay modelos que ya están consolidados en muchas áreas de trabajo, por ejemplo una de las relaciones más comunes que se han observado es la ley de crecimiento logístico ajustada al censo poblacional de los Estados Unidos de Norteamérica del año 1790 al 1940, con la curva expresada en (1.1), obteniendo un buen ajuste del comportamiento de los datos con y_i el tamaño de la población al tiempo t_i (Norusis, 1990).

$$y_i = \frac{C}{1+e^{A+Bx_i}} + \varepsilon_i \qquad (1.1)$$

En particular cuando la respuesta y_i se refiere a P_i, C toma el valor 1, $A = -\beta_0$ y $B = -\beta_1$, de esta expresión se obtiene el siguiente modelo no lineal:

$$P_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}} + \varepsilon_i \qquad (1.2)$$

que produce el modelo ajustado $\hat{P}_i = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 X_i)}}$

Ahora trabajando la transformación logit con el modelo lineal de la expresión (1.3) veremos que es posible obtener P_i como un modelo no lineal en los parámetros (1.4).

Sea $\text{logit}(P_i) = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x + \varepsilon_i \qquad (1.3)$

tomando sólo la última igualdad

$$\begin{aligned}
\frac{P_i}{1-P_i} &= e^{\beta_0+\beta_1 x+\varepsilon_i} \\
P_i &= [1-P_i]e^{\beta_0+\beta_1 x+\varepsilon_i} \\
P_i &= e^{\beta_0+\beta_1 x+\varepsilon_i} - P_i e^{\beta_0+\beta_1 x+\varepsilon_i} \\
P_i + P_i e^{\beta_0+\beta_1 x+\varepsilon_i} &= e^{\beta_0+\beta_1 x+\varepsilon_i} \\
P_i[1+e^{\beta_0+\beta_1 x+\varepsilon_i}] &= e^{\beta_0+\beta_1 x+\varepsilon_i} \\
P_i &= \frac{e^{\beta_0+\beta_1 x+\varepsilon_i}}{[1+e^{\beta_0+\beta_1 x+\varepsilon_i}]} \\
P_i &= \frac{e^{-(\beta_0+\beta_1 x+\varepsilon_i)}}{e^{-(\beta_0+\beta_1 x+\varepsilon_i)}} \frac{e^{\beta_0+\beta_1 x+\varepsilon_i}}{[1+e^{\beta_0+\beta_1 x+\varepsilon_i}]} \\
P_i &= \frac{1}{1+e^{-(\beta_0+\beta_1 x+\varepsilon_i)}} \qquad (1.4)
\end{aligned}$$

Es decir, el modelo (1.4) implícitamente sí es lineal o linealizable ya que el desarrollo anterior se desprende de la expresión lineal (1.3). Las expresiones (1.2) y (1.4) permiten modelar P_i, pero aunque ambos modelos son muy parecidos esto no es así, la diferencia está en la posición del término ε_i. Obsérvese que el modelo (1.2) no es implícitamente lineal a menos que se omita el término ε_i. Es de esperarse que las estimaciones de P_i sean muy parecidas para ambos modelos, esto se muestra en el apéndice 1.1 donde se encuentran los resultados de un ejemplo hipotético al ser tratado con las dos expresiones.

En lo sucesivo sólo se abordará el modelo derivado de la transformación logit.

Dado que la variable respuesta es dicotómica, es natural pensar en una distribución Bernoulli de Y para cada valor x_i de LI, donde la probabilidad de éxito, que en este caso es la recuperación del paciente dado que su valor de LI es x_i, corresponde a P_i, es decir,

$$P_i = \Pr(Y = 1|x_i)$$

Sin pérdida de generalidad se expresará como $Pr(x) = \Pr(Y = 1|x)$ a la probabilidad de éxito dada $X = x$, donde x se refiere a un valor particular de la variable X. Es necesario hacer una aclaración referente a la notación, pues es usual encontrar en la literatura estadística que x_i denota el valor de la variable X que corresponde al individuo i-ésimo de la población bajo estudio, esta notación también se empleará en este trabajo previa aclaración y no debe confundirse con la notación empleada hasta el momento, donde x_i define una población de valores de Y.

Si $\hat{\beta}_0$ y $\hat{\beta}_1$ son los estimadores de los parámetros, el modelo ajustado que corresponde a la expresión (1.3) queda de la siguiente manera:

$$\ln\left(\frac{\hat{Pr}(x)}{1-\hat{Pr}(x)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x \qquad (1.5)$$

El cociente $\frac{\hat{Pr}(x)}{1-\hat{Pr}(x)}$ recibe el nombre de momio (ventaja) de éxito cuando $X = x$, y cuando dicho cociente es mayor que uno o menor que uno indica que el éxito y el fracaso no ocurren con la misma probabilidad, de acuerdo a la notación que seguiremos nos referiremos a este cociente por $\hat{\Psi}(x)$.

La parte izquierda de la igualdad (1.5) se conoce como la forma logit del modelo ajustado y se expresa como $\text{logit}\hat{Pr}(x) = \ln \hat{\Psi}(x)$.

De (1.5) la probabilidad estimada de éxito es:

$$\hat{Pr}(x) = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x)}} \qquad (1.6)$$

Esta expresión es de la forma $f(X) = \frac{1}{1+e^{-X}}$ que corresponde a la función logística.

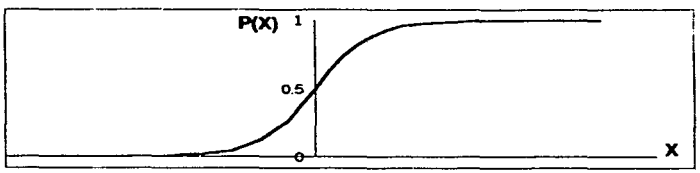

Figura 1.1 Gráfica de la función logística

En el apéndice 1.2 se explica porqué el modelo de regresión logística es un caso particular de los modelos lineales generalizados, estos modelos desarrollan una teoría más amplia en la que

varios modelos bajo ciertas características en común, entre ellos el modelo de regresión lineal y el modelo de regresión logística, son estudiados bajo un concepto general.

La expresión (1.6) se refiere al modelo de regresión logística simple, pero para cuando las variables respuesta son más de dos el modelo se llama de regresión logística múltiple y su expresión es la siguiente:

$$\hat{Pr}(\underline{x}) = \frac{1}{1+\exp[-(\hat{\beta}_0+\sum_{j=1}^{k}\hat{\beta}_j x_j)]} \qquad (1.7)$$

Para k variables explicativas y donde $\underline{x}= (x_1, x_2, ..., x_k)$ es el vector de valores que toman las k variables explicativas, en este caso $\hat{Pr}(\underline{x})$ es la probabilidad de éxito dado $\underline{x}$, su complemento corresponde a la probabilidad de fracaso dado el mismo valor de $\underline{x}$, y el logit correspondiente a $\hat{Pr}(\underline{x})$ es:

$$\text{logit}\hat{Pr}(\underline{x}) = \hat{\beta}_0 + \sum_{j=1}^{k}\hat{\beta}_j x_j$$

1.2 Deficiencias al usar el modelo de regresión lineal cuando la variable respuesta es dicotómica

Si estudiamos una respuesta binaria usando el modelo de regresión lineal usual, al realizar el diagnóstico con los residuales para saber si se validan los supuestos distribucionales de la fluctuación o variación aleatoria, podremos percatarnos de que algunos de estos supuestos no se cumplen. Veamos esto con detenimiento a través de un ejemplo.

Ejemplo 1.2 Modelo de regresión logística con una variable explicativa continua (M1). Se realiza un estudio para conocer la importancia que tienen los triglicéridos en la predicción de la hipertensión, para lo que se consideran 20 hombres adultos en actividad sedentaria, la variable explicativa indica la cantidad de triglicéridos y la respuesta la presencia de hipertensión (Y=1) o ausencia de hipertensión (Y=0), si usamos un modelo de regresión lineal obtendremos el siguiente ajuste:

$$\hat{Y} = -0.333437 + 0.006164 \cdot \textit{triglicéridos}$$

Es de nuestro interés saber si el modelo cumple con los supuestos distribucionales para el modelo de regresión lineal, que se refieren a la forma en que se distribuyen los errores aleatorios

ε_i, que suponemos tienen distribución normal, media cero y varianza constante σ^2, además de suponer que son independientes[2], esto se denota como $\varepsilon_i \sim NID(0, \sigma^2)$. Veamos ahora los supuestos distribucionales de los errores analizando los residuales.

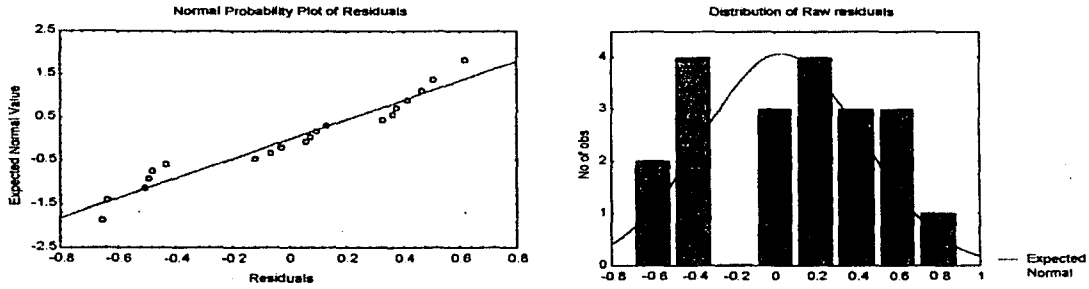

Figura 1.2 Gráfica de probabilidad normal de los residuales

Observamos que la gráfica de probabilidad normal de los residuales indica una ligera violación al supuesto de normalidad, pues esperariamos que no hubiese desviaciones en esta recta, es decir, que los puntos se distribuyeran sobre ella y sin tendencias, para verificar el cumplimiento de este supuesto aplicaremos la prueba de Lilliefors[3] para contrastar la hipótesis:

$$H_0 : F(X) \in \{N(\mu, \sigma^2)\} \quad \text{vs} \quad H_1 : F(X) \notin \{N(\mu, \sigma^2)\}$$

Donde F(X) indica la función de distribución de la variable aleatoria X que representa a los residuales y $\{N(\mu, \sigma^2)\}$ la familia de distribuciones normales sin especificar la media y la varianza. Nuestro tamaño de muestra es 20 y el vector de datos son los residuales al ajustar el modelo de regresión lineal, como no conocemos la media y la varianza entonces las estimamos $\hat{\sigma}^2_{residuales} = 0.16820$ y $\hat{\mu}_{residuales} = 0$.

[2] Por tanto las observaciones de la variable respuesta Y lo son también.

[3] La prueba de Lilliefors para bondad de ajuste se puede consultar en textos referentes a estadística no paramétrica.

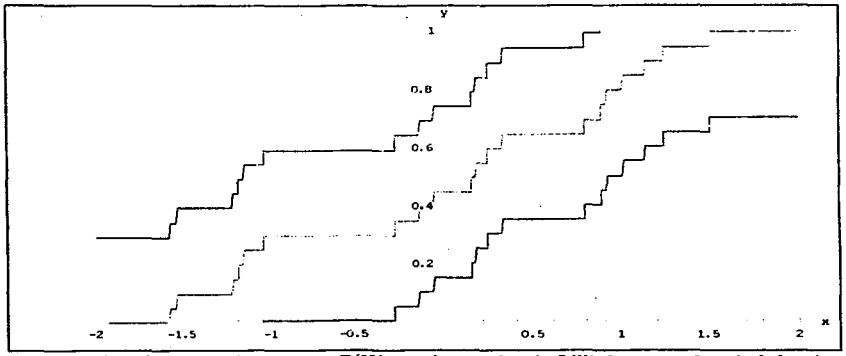

Figura 1.3 Banda de confianza para F(X) con la prueba de Lilliefors para bondad de ajuste

En la figura 1.3 observamos que la distribución empírica parece ajustarse a una distribución de la familia normal y al obtener las bandas de confianza al .95x100% ésta cae dentro, el valor del estadístico de prueba es D_{20}=0.153311 y el valor crítico es $D_{20}^{0.05}$=0.192 con nivel de significancia $\alpha = 0.05$, obtenido de tablas de la estadítica de Lilliefors, tenenos que D_{20}=0.153311<0.192=$D_{20}^{0.05}$ por lo tanto no se rechaza la hipótesis nula, es decir, acepto que la muestra de residuales provienen de una distribución normal, aunque la figura 1.2 sugiera una ligera falta de normalidad se concluye que estadísticamente este supuesto si se cumple para este caso particular. Veamos ahora el supuesto de varianza constante.

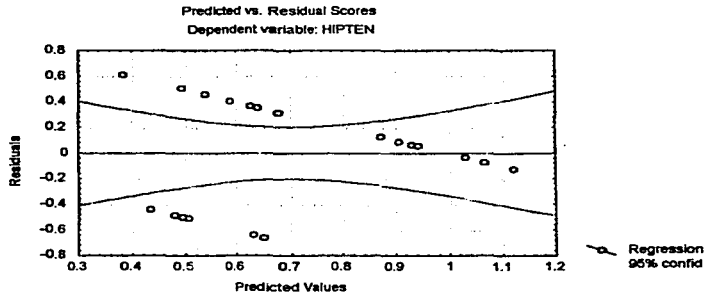

Figura 1.4 Gráfica de varianza de los residuales

Esta gráfica nos indica que la varianza no es constante, pues los residuales no se dispersan cerca de cero, dentro de una banda y de manera uniforme, podemos decir que la varianza para un conjunto de valores, los de la izquierda, es mayor a la de los últimos de la derecha, por lo tanto no es homogénea.

Hemos visto que hay problemas con los supuestos del modelo lineal al considerar una variable respuesta dicotómica, en particular con la homogeneidad de la varianza, ahora veamos esto en general. Tomemos el modelo de regresión lineal simple:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (1.8)$$

$$E(Y_i) = \beta_0 + \beta_1 x_i \qquad (1.9)$$

Consideremos para el modelo de regresión usual la presencia de hipertensión como una variable respuesta binaria[4] y como variable explicativa los triglicéridos, en este caso es de interés predecir que tan probable es que un individuo presente hipertensión a partir de la cantidad de triglicéridos que tenga, así entonces, sea:

$$Y_i = \begin{cases} 1 & \text{ocurrencia del evento de interés denominado éxito (el individuo tiene hipertensión)} \\ 0 & \text{no ocurrencia del evento de interés denominado fracaso (el individuo no tiene hipertensión)} \end{cases}$$

Donde la presencia de hipertensión tiene probablidad P_i y la no presencia $1-P_i$, la esperanza de Y_i es

$$E(Y_i) = 0 \cdot \Pr(Y_i = 0) + 1 \cdot \Pr(Y_i = 1) = P_i$$

entonces la ecuación (1.9) queda como

$$P_i = \Pr(Y_i = 1) = \beta_0 + \beta_1 x_i \qquad (1.10)$$

donde P_i depende del valor de x_i.

Veamos los supuestos distribucionales que hemos mencionado y el supuesto estructural para el modelo de regresión lineal; el supuesto estructural indica la relación funcional lineal de la variable respuesta Y con la variable explicativa X.

[4]En la base de datos empleada originalmente la variable respuesta era continua y representaba la presión sanguínea sistólica, dicha variable continua se transformó a un variable dicotómica, estableciendo dos categorías y clasificando a los individuos de la muestra según su presión sanguínea sistólica en hipertensos y no hipertensos. En este caso la clasificación permite analizar específicamente la hipertensión y esto tiene sentido, pues con la variable continua presión sanguínea sistólica y el modelo de regresión lineal, sólo se obtendría una estimación de ésta dependiendo del nivel de triglicéridos del individuo, en cambio, al tener la variable respuesta dicotómica hipertensión y analizarla con el modelo de regresión logística se podrá estimar la probabilidad de que un individuo presente hipertensión dado su nivel de triglicéridos, además de otras conclusiones como se verá más adelante. Esto deja abierta la opción de transformar una variable respuesta de continua a dicotómica dependiendo del contexto de investigación para analizar la información con el modelo de regresión logística.

$$
\begin{aligned}
&\text{De (1.8)} && \varepsilon_i = Y_i - \beta_0 - \beta_1 x_i \\
&\text{si } Y_i = 1 && \varepsilon_i = 1 - \beta_0 - \beta_1 x_i = 1 - P_i \qquad \text{por (1.10)} \\
&\text{de igual manera si } Y_i = 0 && \varepsilon_i = 0 - \beta_0 - \beta_1 x_i = -P_i \\
& && \Pr(\varepsilon_i = 1 - P_i) = \Pr(Y_i = 1) = P_i \\
& && \Pr(\varepsilon_i = -P_i) = \Pr(Y_i = 0) = 1 - P_i
\end{aligned}
$$

como vemos los errores tienen distribución discreta,

$$
\begin{aligned}
&E(\varepsilon_i) = (1 - P_i)(P_i) + (-P_i)(1 - P_i) = 0 \\
&Var(\varepsilon_i) \quad = E(\varepsilon_i^2) - E^2(\varepsilon_i) = [(1 - P_i)^2 P_i + (-P_i)^2(1 - P_i)] - 0 = (1 - P_i)[(1 - P_i)P_i + P_i^2] \\
&\qquad\qquad = (1 - P_i)[P_i(1 - P_i + P_i)] = (1 - P_i)P_i
\end{aligned}
$$

tienen media cero y su varianza es $(1 - P_i)P_i$ que no es constante, pues P_i depende del valor de x_i, donde $P_i \in [0; 1]$ por ser una probabilidad.

Vale la pena hacer hincapié en que existe diferencia entre lo que predeciremos al ajustar el modelo de regresión lineal con una variable respuesta continua que con una categórica, cuando nuestra variable respuesta bajo estudio es continua el modelo lineal predice a esta variable, y cuando la variable respuesta es dicotómica la predicción que obtenemos bajo la regresión lineal es la probabilidad de que dicha variable tome el valor que caracteriza a la categoría denominada éxito, en ambos casos dado un valor de la variable explicativa. Al ajustar el modelo de regresión lineal a datos donde la respuesta es categórica binaria, las estimaciones pueden no tener sentido debido a que $\beta_0 + \beta_1 x_i$ toma valores en los reales que exceden los límites de predicción de una probabilidad.

Retomando el ejemplo 1.2, en la siguiente tabla se muestran los datos observados y estimados por el modelo de regresión lineal, como se ha visto la esperanza de la variable respuesta (dicotómica) es la probabilidad de éxito dado un valor de X, es decir, la predicción a partir del modelo de regresión lineal es la probabilidad de que el paciente presente hipertensión dado que tiene cierto nivel de triglicéridos.

Tabla 1.1 Datos observados y estimados por el modelo de regresión lineal para el ejemplo 1.2

	Y	X (mg/dl)	$\hat{Y}$		Y	X (mg/dl)	$\hat{Y}$
	hiper	trigli			hiper	trigli	
1	1	205	0.930235	11	1	201	0.905578
2	1	195	0.868592	12	0	134	0.492573
3	1	236	1.121327	13	1	207	0.942563
4	1	156	0.628186	14	0	132	0.480244
5	1	134	0.492573	15	1	141	0.535723
6	0	157	0.634351	16	1	158	0.640515
7	1	164	0.677500	17	0	124	0.430930
8	1	116	0.381616	18	1	149	0.585037
9	0	136	0.504901	19	1	221	1.028863
10	0	160	0.652843	20	1	227	1.065849

Podemos observar que algunos valores estimados de la respuesta exceden el intervalo [0,1] y no tienen sentido al predecir una probabilidad. Este es un motivo junto con el incumplimiento de los supuestos del modelo de regresión lineal para plantear otro modelo.

Aplicando el modelo de regresión logística simple al ejemplo 1.2, el modelo ajustado queda como sigue en su forma logit:

$$logit\hat{P}r(X) = -6.5118 + 0.0472 \cdot triglicéridos$$

De donde -6.5118 es el estimador $\hat{\beta}_0$ para el parámetro β_0 y $\hat{\beta}_1$ =0.0472 para β_1, ahora que se ha aplicado la transformación logística los valores estimados de la variable respuesta tienen sentido sobre los reales, pues representan el logaritmo de un cociente de probabilidades.

El momio de presentar hipertensión dado que el individuo tiene cierta cantidad x de triglicéridos es $\hat{\Psi}(x) = e^{-6.5118+0.0472x}$, si el individuo no tiene triglicéridos. entonces para X=0 el modelo queda como $logit\hat{P}r(X = 0) = \hat{\beta}_0 = -6.5118$ y el momio correspondiente es $\hat{\Psi}(0) = \frac{\hat{P}r(0)}{1-\hat{P}r(0)} = e^{\hat{\beta}_0} = e^{-6.5118} = 0.0014 \doteq 1/673$ lo que se conoce como momio basal, pues este no depende de la variable explicativa, su logaritmo es representado por el coeficiente $\hat{\beta}_0$ del modelo y en este caso $\hat{P}r(0) < 1 - \hat{P}r(0)$ y $673\hat{P}r(0) \doteq 1 - \hat{P}r(0)$, es decir, la probabilidad de no presentar hipertensión es 673 veces la probabilidad de presentar hipertensión, cuando no se tienen triglicéridos.

Si nuestro interés es que la presencia de hipertensión sea menos probable, entonces de acuerdo a (1.6) $\hat{Pr}(X) = \frac{1}{1+e^{-(-6.5118+0.0472X)}}$ tiende a cero cuando a su vez $1+e^{-(-6.5118+0.0472X)}$ tienda a infinito y por tanto $e^{-(-6.5118+0.0472X)}$ también , esto ocurre si $X \longrightarrow -\infty$ ya que $\hat{\beta}_1 = 0.0472 > 0$, como $\hat{Pr}(X)$ está en función de X el coeficiente $\hat{\beta}_1 > 0$ indica que la relación entre la probabilidad de presentar hipertensión y la cantidad de triglicéridos es positiva, es decir, cuando la cantidad de triglicéridos es baja también la probabilidad de presentar hipertensión lo es; en la sección 2.5 se retomará este análisis interpretativo de $\hat{\beta}_1$.

Concluimos que la regresión lineal no es un modelo adecuado para estudiar la relación entre variables cuando la respuesta es dicotómica, por lo que se recurre al modelo de regresión logística, el siguiente capítulo se refiere a la interpretación de este modelo.

Capítulo 2

Interpretación de los parámetros de asociación en el modelo de regresión logística

2.1 Medidas de asociación en el modelo de regresión logística

Los modelos de regresión logística permiten estudiar la dependencia de una variable bernoulli con otra u otras variables explicativas, entonces nos cuestionamos ¿cuánta dependencia hay entre la respuesta y las variables explicativas por ellas mismas? podemos responder haciendo uso de las llamadas medidas de asociación. La medida de asociación que se emplea en el modelo de regresión logística es la razón de momios, sin embargo, existen otras medidas que vale la pena mencionar.

Medidas de asociación:

- riesgo atribuible $(\delta_{A,B})$
- riesgo relativo $(R_{A,B})$
- razón de momios $(\Psi_{A,B})$

Ejemplo 2.1 Sean dos juegos (V.Abraira, 1996), en uno se apuesta a la salida de cierta cara en el lanzamiento de un dado (X=0), y en el otro a la salida de cierta cara en el lanzamiento de una moneda (X=1), la probabilidad de ganar (Y=1) para el dado es $Pr(0) = 1/6$ y para la

moneda $Pr(1) = 1/2$. Es de nuestro interés saber si la probabilidad de ganar está asociada al tipo de juego. Obtenemos las medidas de asociación como sigue:

$$\begin{array}{ll} \text{riesgo atribuible} & \delta_{A,B} = Pr(A) - Pr(B) \\ \text{riesgo relativo} & R_{A,B} = \frac{Pr(A)}{Pr(B)} \\ \text{razón de momios} & \Psi_{A,B} = \frac{\frac{Pr(A)}{1-Pr(A)}}{\frac{Pr(B)}{1-Pr(B)}} \end{array}$$

$\delta_{1,0} = Pr(1) - Pr(0) = 1/2 - 1/6 = 1/3$

$R_{1,0} = \frac{Pr(1)}{Pr(0)} = \frac{1/2}{1/6} = 3$

$\Psi_{1,0} = \frac{\frac{Pr(1)}{1-Pr(1)}}{\frac{Pr(0)}{1-Pr(0)}} = \frac{\frac{1/2}{1/2}}{\frac{1/6}{5/6}} = 5$

En el caso en que $\delta_{1,0}$ es cero y cuando $R_{1,0}$ y $\Psi_{1,0}$ son ambos uno, se concluye que no hay diferencia entre la probabilidad de éxito dado que se tiene una situación u otra, en este ejemplo el riesgo atribuible $\delta_{1,0} = 1/3$ es distinto de cero por lo que decimos que la diferencia de probabilidades existe y sí importa que juego se tome para ganar, esta medida es la más intuitiva porque involucra únicamente la diferencia, pero si ahora vemos el riesgo relativo ($R_{1,0}$) podemos además saber que tanto más probable es ganar con un juego que con otro, como $R_{1,0} = 3$, quiere decir que es tres veces más probable ganar con la moneda que con el dado, es decir, la probabilidad de ganar depende del tipo de juego, y en este caso es más conveniente jugar un volado para ganar. La razón de momios como su nombre lo dice es un cociente de momios, recordemos que el momio $\frac{Pr(x)}{1-Pr(x)}$ de la probabilidad de éxito cuando $X = x$, indica cuánto más probable es el éxito que el fracaso dada $X = x$, siendo que cuando es mayor que uno la probabilidad de éxito es mayor que la de fracaso. Si tenemos dos distintos valores de X, digamos x_1 y x_2 y obtenemos los respectivos momios $\frac{Pr(x_1)}{1-Pr(x_1)}$ y $\frac{Pr(x_2)}{1-Pr(x_2)}$ tenemos estos para dos situaciones de X, tomando el cociente $\frac{\frac{Pr(x_1)}{1-Pr(x_1)}}{\frac{Pr(x_2)}{1-Pr(x_2)}} = \frac{Pr(x_1)(1-Pr(x_2))}{Pr(x_2)(1-Pr(x_1))}$ que recibe el nombre de razón de momios, sabemos como es el éxito respecto al fracaso comparando ambas situaciones, si estas afectan la probabilidad de éxito entonces la razón de momios es distinta de uno. Aunque esta medida de asociación es poco intuitiva, veremos cuan importante es en nuestro modelo. Para este ejemplo tenemos que $\Psi_{1,0} = 5$, razón de momios distinta de uno, por tanto el momio de éxito para la moneda es cinco veces el momio del dado y decimos que la probabilidad de ganar sobre la de perder es cinco veces mayor para la moneda que para el dado.

Ejemplo 2.2 Modelo de regresión logística con una variable explicativa dicotómica (M2).

Ahora supongamos que se quiere estudiar el efecto de cuidados prenatales en la supervivencia de recién nacidos, donde estos mueren (Y=0) o sobreviven (Y=1), la variable respuesta representa la supervivencia y definimos como éxito cuando el recién nacido sobrevive (como es de esperarse, sin embargo, la definición del éxito depende del contexto de investigación, véase la sección 2.4). Es de interés saber si existe asociación entre la supervivencia y la cantidad de cuidados prenatales recibidos por la madre (variable explicativa), que pueden ser pocos (X=0) o muchos (X=1), observemos que esta variable explicativa es categórica. Se ha tomado una muestra aleatoria de tamaño N y la información se presenta en la siguiente tabla:

Tabla 2.1 Tabla de contingencia de dos dimensiones

		variable explicativa cuidados prenatales		
		pocos (X=0)	muchos (X=1)	
variable respuesta supervivencia	muere (Y=0)	A	B	
	sobrevive (Y=1)	C	D	
	total	N_1	N_2	N

Esta presentación de los datos se conoce como tabla de contingencia y se emplea para determinar la asociación entre variables categóricas, como se ha mencionado la variable explicativa lo es y evidentemente la variable respuesta también, cada celda indica una frecuencia y dependiendo del número de variables la nombramos tabla de contingencia de dos dimensiones, cuando se trata de dos variables como en este caso, o tabla de contingencia multidimensional cuando se estudian tres o más, cabe mencionar que cada miembro de la población o de la muestra según el caso debe pertenecer sólo a una categoría (exclusividad). A partir de esta tabla estimamos la probabilidad de éxito cuando X=0 y X=1 como $\hat{Pr}(0) = \frac{C}{N_1}$ y $\hat{Pr}(1) = \frac{D}{N_2}$ respectivamente y calculamos:

$$\hat{\delta}_{1,0} = \hat{Pr}(1) - \hat{Pr}(0) = D/N_2 - C/N_1$$

$$\hat{R}_{1,0} = \frac{\hat{Pr}(1)}{\hat{Pr}(0)} = \frac{D/N_2}{C/N_1} = \frac{N_1 D}{N_2 C}$$

$$\hat{\Psi}_{1,0} = \frac{\frac{\hat{Pr}(1)}{1-\hat{Pr}(1)}}{\frac{\hat{Pr}(0)}{1-\hat{Pr}(0)}} = \frac{\frac{D/N_2}{B/N_2}}{\frac{C/N_1}{A/N_1}} = \frac{AD}{BC}$$

Supongamos los siguientes valores para interpretar resultados:

Tabla 2.2 Tabla de contingencia de dos dimensiones correspondiente a la supervivencia de recién nacidos y a los cuidados prenatales recibidos por la madre

variable respuesta supervivencia	variable explicativa cuidados prenatales: pocos (X=0)	muchos (X=1)	
muere (Y=0)	20	6	
sobrevive (Y=1)	373	316	
total	393	322	715

tenemos $\hat{Pr}(0) = \frac{373}{393} = 0.9491$ y $\hat{Pr}(1) = \frac{316}{322} = 0.9813$

$\hat{\delta}_{1,0} = 0.0322$

$\hat{R}_{1,0} = 1.0339$

$\hat{\Psi}_{1,0} = 2.8239 \simeq 3$

El riesgo atribuible es distinto de cero pero muy cercano a éste, mientras que el riesgo relativo es "casi uno" lo que indicaría no diferencia entre la probabilidad de éxito cuando X=0 y X=1, es decir, la cantidad de cuidados prenatales no influye en la probabilidad de sobrevivir, vemos que la razón de momios es distinta de uno entonces interpretamos que la razón de recién nacidos sobrevive/muere es aproximadamente tres veces mayor cuando se dan muchos cuidados prenatales que cuando los cuidados prenatales son pocos, esto significa que es más probable sobrevivir que morir con muchos cuidados prenatales que con pocos, por lo que concluimos que hay efecto del factor cuidados prenatales en la supervivencia. Se debe hacer hincapié en que estas conclusiones son sólo intuitivas, por lo que es necesario obtener para estas estimaciones puntuales sus correspondientes intervalos de confianza y probar hipótesis como $H_0 : \Psi_{1,0} = 1$ vs. $H_1 : \Psi_{1,0} \neq 1$ para verificar la precisión de la estimación y decidir correctamente sobre el efecto o no de las variables explicativas; véase la sección 2.3.

Hemos definido al éxito cuando el recién nacido sobrevive, pero supongamos ahora que el resultado en estudio es la mortalidad, entonces redefinimos éxito=muere, el análisis es análogo y su desarrollo se deja en manos del lector quién podrá verificar que el resultado obtenido es equivalente en interpretación[1], pues $\hat{\Psi}_{1,0} = 0.3541 = 1/2.8239 \simeq 1/3$, es decir, es más probable morir que sobrevivir cuando los cuidados prenatales son pocos, pues el momio de la probabilidad

[1] Revisar la sección 2.4

de éxito (morir) dado X=0 es tres veces el momio cuando X=1.

La razón de momios $\Psi_{A,B}$ puede aproximarse por el riesgo relativo cuando P_1 y P_2 son ambas muy pequeñas, teniendo que $1-P_1$ y $1-P_2$ son muy cercanas a uno, por lo que

$$\Psi = \frac{P_1(1-P_2)}{P_2(1-P_1)} \simeq \frac{P_1}{P_2} = R$$

Ejemplo 2.3 Comparemos ahora el juego de la lotería nacional (X=1) en el que el premio es para un número extraído de entre 100,000 con el de la lotería primitiva (X=0) en que se premia una combinación de seis números de entre las que se pueden formar con 49 números[2]. La probabilidad de ser premiado en la lotería nacional es $Pr(1) = \frac{1}{100000}$ y en la lotería primitiva $Pr(0) = \frac{1}{13983816}$ donde se consideran las 13,983,816 combinaciones de seis números formadas por 49, tenemos entonces

$$R_{1,0} = \frac{Pr(1)}{Pr(0)} = \frac{1/100000}{1/13983816} = 139.83816$$

$$\Psi_{1,0} = \frac{\frac{Pr(1)}{1-Pr(1)}}{\frac{Pr(0)}{1-Pr(0)}} = \frac{\frac{1/100000}{1-1/100000}}{\frac{1/13983816}{1-1/13983816}} = \frac{13983815}{99999} = 139.83954$$

Dado que las probabilidades $Pr(0)$ y $Pr(1)$ son muy pequeñas, el riesgo relativo es aproximadamente igual a la razón de momios e indican que es más probable ganar con la lotería nacional que con la primitiva.

Como se ha mencionado una medida de asociación es la razón de momios que refiere la comparación de dos subconjuntos que tienen valores de X diferentes, es decir, indica que tan probable es el éxito ante el fracaso dado que X toma un valor, respecto a cuanto toma otro en las mismas condiciones, permitiendo así concluir la aportación del factor para explicar el fenómeno en estudio o variable respuesta.

2.2 Interpretación de los parámetros

Supongamos que la variable explicativa X en un modelo simple también es binaria como la respuesta.

$$X = \begin{cases} 0 & \text{nivel de referencia} \\ 1 & \text{nivel de interés} \end{cases}$$

[2] Este ejemplo fue tomado de la página http://www.hrc.es/bioest/Reglog_1.html

Siendo así llamaremos a esta variable factor de exposición, en medicina se le conoce también como factor de riesgo.

En el caso de $logit\hat{P}r(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ los momios son:

$$\hat{\Psi}(0) = \frac{\hat{P}r(0)}{1-\hat{P}r(0)} = e^{\hat{\beta}_0} \qquad \text{momio basal}$$
$$\hat{\Psi}(1) = \frac{\hat{P}r(1)}{1-\hat{P}r(1)} = e^{\hat{\beta}_0+\hat{\beta}_1}$$

de donde

$$\ln \hat{\Psi}(0) = \hat{\beta}_0 \qquad y \qquad \ln \hat{\Psi}(1) = \hat{\beta}_0 + \hat{\beta}_1$$

El parámetro estimado $\hat{\beta}_0$ es el logaritmo natural del momio correspondiente al nivel de referencia, y la razón de momios queda expresada como:

$$\hat{\Psi}_{1,0} = \frac{\hat{\Psi}(1)}{\hat{\Psi}(0)} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{e^{\hat{\beta}_0}} = e^{\hat{\beta}_1}\ , \qquad \ln \hat{\Psi}_{1,0} = \hat{\beta}_1$$

E interpretamos el efecto de un factor de exposición, es decir, si $\hat{\Psi}_{1,0} = 1$ no hay efecto del factor pues los momios $\hat{\Psi}(1)$ y $\hat{\Psi}(0)$ son iguales, entonces la probabilidad de éxito respecto a la de fracaso es la misma para cualquier valor de X y para estudiar la asociación de las variables basta verificar si el coeficiente $\hat{\beta}_1$ es cero.

Si la variable X no es binaria y puede tomar más valores, entonces la interpretación de $\hat{\beta}_0$ es la misma siempre que el nivel de referencia sea $X = 0$ y $\hat{\beta}_1$ es el logaritmo de la razón de momios por aumento de una unidad en X. Supongamos que la variable X del ejemplo 2.2 ahora representa la edad de la madre, entonces un cambio de edad de un año en la supervivencia no suena muy interesante, pero una diferencia de quinquenios puede dar sentido a la asociación de las variables, sea $x_0 = 20$ y $x_1 = 25$

$$\ln \hat{\Psi}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0,$$
$$\ln \hat{\Psi}(x_1) = \hat{\beta}_0 + \hat{\beta}_1 x_1$$
$$\ln \hat{\Psi}_{x_1,x_0} = \ln[\frac{\hat{\Psi}(x_1)}{\hat{\Psi}(x_0)}] = \ln \hat{\Psi}(x_1) - \ln \hat{\Psi}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_1 - \hat{\beta}_0 - \hat{\beta}_1 x_0 = \hat{\beta}_1(x_1 - x_0)$$

Entonces $\hat{\Psi}_{x_1,x_0} = e^{5\hat{\beta}_1}$ e indica que el logaritmo de la razón de momios de edad 25 con respecto a edad 20 es cinco veces el logaritmo de la razón de momios por aumento de un año de edad, en general para un incremento Δ de la variable X, de x_0 a x_1 tenemos

$$\hat{\Psi}_{x_1,x_0} = \frac{\hat{\Psi}(x_1)}{\hat{\Psi}(x_0)} = e^{\Delta\hat{\beta}_1}$$

Ahora veamos la relación que tienen los momios con los parámetros en el modelo de regresión logística multiple, expresado en (1.7). Un cociente de momios compara el momio de un evento con el correspondiente a otro, en este caso los eventos estarán definidos en función de los valores que tomen las k variables explicativas, entonces se define $\underline{x}_A = (x_{A1}, x_{A2}, ..., x_{Ak})$ y $\underline{x}_B = (x_{B1}, x_{B2}, ..., x_{Bk})$, el momio de $\underline{x}_A$ es $\hat{\Psi}(\underline{x}_A) = \exp[\hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{Aj}]$ y el de $\underline{x}_B$ es $\hat{\Psi}(\underline{x}_B) = \exp[\hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{Bj}]$, con lo que se tiene una forma general para el cociente de momios:

$$\hat{\Psi}_{\underline{x}_A,\underline{x}_B} = \frac{\hat{\Psi}(\underline{x}_A)}{\hat{\Psi}(\underline{x}_B)} = \exp[\sum_{j=1}^{k} \hat{\beta}_j (x_{Aj} - x_{Bj})]$$

Si ahora los eventos $\underline{x}_A$ y $\underline{x}_B$ sólo difieren en el valor que tome una variable explicativa específica para los dos y las demás quedan fijas, entonces la comparación que hace el cociente de momios se refiere al cambio realizado en una variable de estudio y las otras variables explicaticas reciben el nombre de variables control, si $\underline{x}_A = (x_1, ..., x_{Aj}, ..., x_k)$ y $\underline{x}_B = (x_1, ..., x_{Bj}, ..., x_k)$ el cociente de momios es $\hat{\Psi}_{\underline{x}_A,\underline{x}_B} = e^{\hat{\beta}_j(x_{Aj} - x_{Bj})}$, si la variable X_j es binaria entonces el cociente de momios queda como $\hat{\Psi}_{\underline{x}_A,\underline{x}_B} = e^{\hat{\beta}_j}$, de donde el parámetro $\hat{\beta}_j$ representa el logaritmo natural de $\hat{\Psi}_{\underline{x}_A,\underline{x}_B}$ para este caso particular; nótese que aquí la interpretación de $\hat{\beta}_j$ es análoga a la de $\hat{\beta}_1$ en un modelo simple con varianle explicativa X_j, pero debe considerarse que las conclusiones están condicionadas a los valores que tomen las otras variables explicativas; también es importante tener en cuenta que $\hat{\beta}_1$ en el modelo simple no necesariamente es igual a $\hat{\beta}_j$ del modelo múltiple.

2.3 Inferencia sobre los parámetros

Como se ha mencionado en la sección 2.1 para probar la significancia de los parámetros estimados es necesario hacer inferencia estadística sobre ellos, en esta sección se mencionarán los estadísticos usuales sin hacer hincapié en el desarrollo teórico para su obtención. La estimación de los parámetros se hace con el método de máxima verosimilitud, por lo tanto la inferencia sobre los ellos y el buen ajuste del modelo estarán basados en ese principio.

En el modelo de regresión logística con k variables explicativas, sea $\underline{x}_i = (x_{i1}, x_{i2}, ..., x_{ik})$ uno de los m valores diferentes que toma $\underline{X}$ y definamos Z_i el número de éxitos observados dado

que presentaron el valor $\underline{x}_i$, es decir. $Z_i = \sum_{l=1}^{n_i} y_{il}$ donde n_i indica el número de observaciones con el valor $\underline{x}_i$, como se vio en la sección 1.1. siendo $Pr(\underline{x}_i) = \frac{1}{1+\exp[-(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_k x_{ik})]}$ la probabilidad de éxito dada $\underline{x}_i$ para $i = 1, ..., m$. Z_i se distribuye Binomial con parámetros $(n_i,\ Pr(\underline{x}_i))$ y la función de verosimilitud para $z_1, z_2, ... z_m$ es:

$$L(\beta_0, \beta_1, ..., \beta_k; z_1, z_2, ... z_m) = \prod_{i=1}^{m} \binom{n_i}{z_i} Pr(\underline{x}_i)^{z_i}\ [1 - Pr(\underline{x}_i)]^{n_i - z_i} \qquad (2.1)$$

El método de máxima verosimilitud tiene como objetivo obtener el vector $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ de estimadores del vector de coeficientes $\beta = (\beta_0, \beta_1, ..., \beta_k)$, que maximizan la función (2.1). Sea $L = L(\beta_0, \beta_1, ..., \beta_k; z_1, z_2, ..., z_m)$, el logaritmo de la función de verosimilitud es:

$$\ln L = C + \sum_{i=1}^{m} \{z_i \ln Pr(\underline{x}_i) + (n_i - z_i) \ln [1 - Pr(\underline{x}_i)]\}$$

Donde C es una constante que no depende de los parámetros, entonces para encontrar los valores de $\beta_0, \beta_1, ..., \beta_k$ que maximizan la función de verosimilitud se deriva $\ln L$ con respecto a los parámetros y se igualan a cero las expresiones obtenidas, éstas se muestran a continuación.

$$\begin{gathered} \sum_{i=1}^{m} [z_i - n_i Pr(\underline{x}_i)] = 0 \\ \sum_{i=1}^{m} x_{ij} [z_i - n_i Pr(\underline{x}_i)] = 0 \quad \text{para } j = 1, ..., k \end{gathered} \qquad (2.2)$$

Observemos que este sistema de ecuaciones (2.2) no es lineal en los parámetros y por lo tanto no se puede obtener una expresión despejada de este sistema para los parámetros, siendo necesario para su solución recurrir a métodos numéricos iterativos, en particular al de Newton-Rapson. Para obtener los resultados de los ejemplos que ilustran este trabajo se recurrió al paquete de cómputo estadístico SPSS versión 10.0, en el mercado de software estadístico existen otros paquetes que permiten el análisis de regresión logística y esto se comenta brevemente en el apéndice 2.1.

Supongamos que se ha ajustado un modelo de regresión logística múltiple con k variables explicativas (modelo completo) a cierta base de datos y que la pretensión es determinar que variables regresoras son significativas, esto permite tener un modelo más sencillo con un múmero menor de variables el cual pudiéramos utilizar (modelo reducido), entonces asociemos a cada modelo su correspondiente función de verosimilitud, sea $\hat{L}_C = L(\hat{\beta}_C; z_1, z_2, ..., z_m)$ con $\hat{\beta}_C$ el

vector de coeficientes estimados para el modelo con k variables explicativas, la función de verosimilitud maximizada del modelo completo y $\hat{L}_R = L(\hat{\beta}_R; z_1, z_2, ..., z_m)$ la correspondiente al modelo reducido donde $\hat{\beta}_R$ es el vector de los coeficientes estimados, es un hecho que $\hat{L}_R \leq \hat{L}_C$ (Flury, 1997) de manera que $0 \leq -2\ln\frac{\hat{L}_R}{\hat{L}_C} < \infty$, de donde $-2\ln\frac{\hat{L}_R}{\hat{L}_C}$ es el estadístico de prueba del cociente de verosimilitudes donde la hipótesis a probar es $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ para toda $j \in J$ donde J es el conjunto de subíndices correspondientes a los coeficientes de las variables que aparecen en el modelo completo pero no en el modelo reducido (considere que el modelo reducido puede obtenerse del modelo completo igualando a cero los coeficientes de las variables sobrantes), este estadístico de prueba se distribuye aproximadamente Ji-cuadrada con grados de libertad el resultado de la diferencia del número de parámetros del modelo completo menos el número de parámetros del modelo reducido, es decir, los que deben anularse en el modelo completo para obtener el reducido, todo bajo la hipótesis nula. Si el valor de dicho estadístico es mayor al cuantil de orden $(1-\alpha)$ de la distribución Ji-cuadrada con los correspondientes grados de libertad, entonces la hipótesis nula debe rechazarse con un nivel de prueba α.

La prueba anterior en particular nos permite probar la hipótesis $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ para $j = 0, .., k$, nótese que $j = 0$ indica la prueba sobre β_0, donde el cociente de verosimilitudes se distribuye Ji-cuadrada con un grado de libertad. Otra forma de probar esta hipótesis es recurrir al estadístico de Wald cuya forma general es $\frac{\hat{\beta}_j - \beta_j}{S(\hat{\beta}_j)}$ y se distribuye como una variable aleatoria normal estándar siempre que la muestra sea lo suficientemente grande, con $S(\hat{\beta}_j)$ el error estándar de $\hat{\beta}_j$, para este caso el estadístico queda como $\frac{\hat{\beta}_j}{S(\hat{\beta}_j)}$ bajo la hipótesis nula y la regla de decisión es comparar el valor del estadístico de prueba con el valor que tome cierto cuantil de la distribución normal estándar, si el estadístico de Wald es mayor al cuantil de orden $(1-\frac{\alpha}{2})$ entonces se concluye que la hipótesis nula debe rechazarse al nivel de prueba α[3]. Dado que el estádistico de Wald tiene una distribución normal estándar, entonces se tiene que $\left(\frac{\hat{\beta}_j}{S(\hat{\beta}_j)}\right)^2$ se distribuye aproximadamente como una variable aleatoria Ji-cuadrada con un grado de libertad bajo $H_0 : \beta_j = 0$. Notemos que con la prueba de Wald queda probado $H_0 : e^{\beta_j} = 1$ vs. $H_1 : e^{\beta_j} \neq 1$ para $j = 0, .., k$. Los estadísticos de Wald y del cociente de

[3]En este caso el estadístico de Wald hace una prueba de dos colas, por lo que también se debe considerar que si el estadístico es menor al cuantil $\frac{\alpha}{2}$ de la distribución normal estándar la hipótesis nula debe rechazarse al mismo nivel de prueba α.

verosimilitudes para probar la nulidad de un parámetro son asintóticamente equivalentes sólo cuando se trabaja con muestras grandes.

El intervalo correspondiente a β_j con $100(1-\alpha)\%$ de confianza es $\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}} \cdot S(\hat{\beta}_j)$ de donde se obtiene el intervalo de confianza para e^{β_j} como $\exp(\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}} \cdot S(\hat{\beta}_j))$ donde $Z_{1-\frac{\alpha}{2}}$ se refiere al cuantil de orden $(1-\frac{\alpha}{2})$ de la distribución normal estándar.

Una forma de probar el buen ajuste del modelo es calculando la devianza, que consiste en comparar para cada $\underline{x}_i$ los éxitos observados con los predichos usando la función de verosimilitud. Con los estimadores de máxima verosimilitud de los parámetros podemos obtener la probabilidad estimada de éxito dada $\underline{x}_i$, esto es $\hat{Pr}(\underline{x}_i)$, entonces el valor máximo del logaritmo de la función de verosimilitud en términos de $\hat{Pr}(\underline{x}_i)$ es:

$$\ln L(\hat{Pr}(\underline{x})) = \ln \hat{L} = C + \sum_{i=1}^{m} \left\{ z_i \ln \hat{Pr}(\underline{x}_i) + (n_i - z_i) \ln \left[1 - \hat{Pr}(\underline{x}_i)\right] \right\} \tag{2.3}$$

En esta expresión implícitamente está el número predicho de éxitos, que se calcula como $\hat{z}_i = n_i \hat{Pr}(\underline{x}_i)$, y como se busca compararlo con el número observado de éxitos mediante la función de verosimilitud, entonces ésta se plantea para el modelo "perfecto" (Flury, 1997), es decir, en el que la probabilidad estimada de éxito sea $\tilde{Pr}(\underline{x}_i) = \frac{z_i}{n_i}$, la siguiente expresión corresponde al valor máximo del logaritmo de la función de verosimilitud en términos de $\tilde{Pr}(\underline{x}_i)$.

$$\ln L(\tilde{Pr}(\underline{x})) = \ln \tilde{L} = C + \sum_{i=1}^{m} \left[z_i \ln \tfrac{z_i}{n_i} + (n_i - z_i) \ln(1 - \tfrac{z_i}{n_i}) \right] \tag{2.4}$$

La devianza del modelo con parámetros estimados $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ y probabilidad estimada de éxito $\hat{Pr}(\underline{x}_i)$ es: menos dos veces el logaritmo del cociente de verosimilitudes obtenido de las expresiones (2.3) y (2.4).

Devianza:
$$\begin{aligned} D &= -2 \ln \left[\tfrac{L(\hat{Pr}(\underline{x}))}{L(\tilde{Pr}(\underline{x}))} \right] \\ &= -2 \left[\ln L(\hat{Pr}(\underline{x})) - \ln L(\tilde{Pr}(\underline{x})) \right] \\ &= -2 \sum_{i=1}^{m} \left[z_i \ln \tfrac{\hat{z}_i}{z_i} + (n_i - z_i) \ln(\tfrac{n_i - \hat{z}_i}{n_i - z_i}) \right] \end{aligned}$$

La devianza en la regresión logística tiene el mismo objetivo que la suma de cuadrados de los residuales en la regresión lineal (Hosmer, 1989), es una medida de la falla de ajuste del modelo, a mayor devianza más falla en el ajuste. La distribución de la devianza es aproximadamente Ji-cuadrada con (m-k) grados de libertad (la diferencia entre el número de valores distintos de

X y el número de parámetros), siempre que todas las n_i's sean grandes y bajo la hipótesis de que el modelo se ajusta a los datos correctamente.

Retomando la expresión del estadístico para la prueba del cociente de verosimilitudes ésta se puede reescribir como la diferencia entre la devianza del modelo reducido y la del completo.

$$\begin{aligned} -2\ln\frac{\hat{L}_R}{\hat{L}_C} &= -2\left[\ln\hat{L}_R - \ln\hat{L}_C\right] \\ &= -2\left\{\left[\ln\hat{L}_R - \ln\tilde{L}\right] - \left[\ln\hat{L}_C - \ln\tilde{L}\right]\right\} \\ &= D_R - D_C \end{aligned}$$

La devianza es calculada por la mayoría de los paquetes de cómputo referentes a regresión logística, así como también el estadístico de Wald y su nivel de significancia descriptivo, en particular el paquete SPSS calcula los intervalos de confianza para e^{β_i}.

Una vez que se ha ajustado el modelo es posible obtener una tabla de clasificación de los datos observados con los datos estimados respecto a las categorías de la variable respuesta, esto se hace con la probabilidad estimada de éxito que explícitamente se obtiene del modelo, es de esperarse que si el modelo ajusta adecuadamente los datos en alguna dirección (por ejemplo predice bien los éxitos), entonces el porcentaje de buena clasificación en algún sentido sea alto, es decir, si se observa un éxito en la variable respuesta se espera que la probabilidad de que esa observacion sea éxito bajo el modelo sea alta y que dicha probabilidad sea baja si se ha observado fracaso; esta tabla de clasificación nos permitirá saber que tanto porcentaje de las observaciones están bien clasificadas y de que manera, fijando una probabilidad para clasificar (punto de corte). En la sección 3.1 se muestra el uso de las tablas de clasificación.

2.4 Interpretación de los resultados cuando en la variable respuesta el éxito es redefinido

Trabajando la variable respuesta hemos definido el éxito como Y=1, definirlo no siempre resulta fácil, pero nos apoyaremos en el contexto del problema, supongamos que nuestra variable respuesta representa el evento supervivencia de recién nacidos, donde hay dos alternativas, morir o vivir, entonces parece lógico pensar que el éxito es vivir, pero si nuestro interés es la incidencia de la mortalidad el éxito sería morir, dado que esto puede causar confusión mostraremos que los resultados obtenidos al definir el éxito al que también llamaremos problema son los mismos en

cuanto a interpretación para ambos casos. Se tienen dós posibles modelos según sea definido el éxito en la variable respuesta dicotómica:

$\hat{Pr}(Y=1|x) = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x)}}$ donde Y=1 indica que el recién nacido vive

$\hat{Pr}(Y'=1|x) = \frac{1}{1+e^{-(\hat{\alpha}_0+\hat{\alpha}_1 x)}}$ donde Y'=1 indica que el recién nacido muere

Estos son los dos enfoques que puede tomar el análisis según sea el contexto, en principio la variable respuesta se define como Y=1 éxito, lo que se mostrará es como al redefinir el éxito Y=0 como Y'=1 la interpretación de los resultados es la misma.

$$\begin{aligned}\hat{\alpha}_0+\hat{\alpha}_1 x &= \ln[\tfrac{\hat{Pr}(Y'=1|x)}{1-\hat{Pr}(Y'=1|x)}] = \ln[\tfrac{\hat{Pr}(Y=0|x)}{1-\hat{Pr}(Y=0|x)}] \\ &= \ln[\tfrac{1-\hat{Pr}(Y=1|x)}{\hat{Pr}(Y=1|x)}] = \ln[\tfrac{\hat{Pr}(Y=1|x)}{1-\hat{Pr}(Y=1|x)}]^{-1} \\ &= -\ln[\tfrac{\hat{Pr}(Y=1|x)}{1-\hat{Pr}(Y=1|x)}] = -(\hat{\beta}_0+\hat{\beta}_1 x)\end{aligned}$$

$\hat{\alpha}_0+\hat{\alpha}_1 x = -(\hat{\beta}_0+\hat{\beta}_1 x)$ sea $\hat{\alpha}_0 = -\hat{\beta}_0$ y $\hat{\alpha}_1 = -\hat{\beta}_1$

El estimador del parámetro β_1 indica la relación que guarda la probabilidad de que el recién nacido viva con el valor que toma la variable X, y $\hat{\alpha}_1$ indica la relación entre la probabilidad de que el recién nacido muera y el valor de la variable X, debido a que el signo de $\hat{\alpha}_1$ es contrario al de $\hat{\beta}_1$ las relaciones son inversas, sin embargo, el valor absoluto de ambos estimadores es el mismo $|\hat{\alpha}_1| = |\hat{\beta}_1|$, lo mismo ocurre con $|\hat{\alpha}_0| = |\hat{\beta}_0|$. Tenemos que $\hat{\Psi} = e^{\hat{\beta}_1}$ es la razón de momios y $\hat{\Psi}(0) = e^{\hat{\beta}_0}$ el momio basal, pero con el cambio de éxito $\hat{\Psi}' = e^{\hat{\alpha}_1} = e^{-\hat{\beta}_1} = \frac{1}{e^{\hat{\beta}_1}} = \frac{1}{\hat{\Psi}}$ y $\hat{\Psi}'(0) = e^{\hat{\alpha}_0} = e^{-\hat{\beta}_0} = \frac{1}{e^{\hat{\beta}_0}} = \frac{1}{\hat{\Psi}(0)}$ son los inversos tanto de la razón de momios como del momio basal inicial. Si los momios de que el recién nacido viva con X=1 respecto a que viva con X=0 son cinco a uno, entonces los momios de que el recién nacido muera con los mismos valores de X son uno a cinco, visto de otra forma,

$$\begin{aligned}\hat{\Psi} &= \frac{\frac{\hat{Pr}(Y=1|X=1)}{1-\hat{Pr}(Y=1|X=1)}}{\frac{\hat{Pr}(Y=1|X=0)}{1-\hat{Pr}(Y=1|X=0)}} = \frac{\frac{\hat{Pr}(Y'=0|X=1)}{1-\hat{Pr}(Y'=0|X=1)}}{\frac{\hat{Pr}(Y'=0|X=0)}{1-\hat{Pr}(Y'=0|X=0)}} = \frac{\frac{1-\hat{Pr}(Y'=1|X=1)}{\hat{Pr}(Y'=1|X=1)}}{\frac{1-\hat{Pr}(Y'=1|X=0)}{\hat{Pr}(Y'=1|X=0)}} \\ &= \frac{\hat{Pr}(Y'=1|X=0)[1-\hat{Pr}(Y'=1|X=1)]}{\hat{Pr}(Y'=1|X=1)[1-\hat{Pr}(Y'=1|X=0)]} = \frac{\frac{\hat{Pr}(Y'=1|X=0)}{1-\hat{Pr}(Y'=1|X=0)}}{\frac{\hat{Pr}(Y'=1|X=1)}{1-\hat{Pr}(Y'=1|X=1)}} = \left[\frac{\frac{\hat{Pr}(Y'=1|X=1)}{1-\hat{Pr}(Y'=1|X=1)}}{\frac{\hat{Pr}(Y'=1|X=0)}{1-\hat{Pr}(Y'=1|X=0)}}\right]^{-1} = \frac{1}{\hat{\Psi}'}\end{aligned}$$

Si la probabilidad de vivir con X=0 es mayor que con X=1, entonces la probabilidad de morir es menor con X=0 que con X=1, esto nos dará la misma interpretación de resultados sin importar como se haya definido el éxito.

2.5 Importancia del parámetro $\hat{\beta}_1$ en la interpretación de las relaciones de X con $logit\hat{P}r(X)$ y $\hat{P}r(X)$

Recordemos que la probabilidad de éxito cuando $X = x$ es $\hat{P}r(x) = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x)}}$, para lo cual teníamos que el logaritmo del momio cuando $X = x$ es $logit[\hat{P}r(x)] = \ln[\hat{\Psi}(x)] = \hat{\beta}_0 + \hat{\beta}_1 x$, lo que llamamos logit de la probabilidad de éxito cuando $X = x$, la relación que hay entre los valores que toma X y $logit[\hat{P}r(x)]$ puede ser indicada por $\hat{\beta}_1$, consideremos la siguiente notación: $A \uparrow$ significa que A tiende a su máximo valor (crece) y $A \downarrow$ significa que A tiende a su mínimo valor (decrece), conociendo el valor de $\hat{\beta}_1$, si $\hat{\beta}_1 > 0$ y $X \uparrow$ entonces $logit[\hat{P}r(x)] \uparrow$, pero si $X \downarrow$ entonces $logit[\hat{P}r(x)] \downarrow$, lo que sugiere que la relación es directa. Ahora veamos el caso en que $\hat{\beta}_1 < 0$, si $X \uparrow$ entonces $logit[\hat{P}r(x)] \downarrow$ y si $X \downarrow$ entonces $logit[\hat{P}r(x)] \uparrow$, es decir, la relación es inversa. Con lo anterior tenemos un indicador $\hat{\beta}_1$ de como es la relación entre X y $logit[\hat{P}r(x)]$, pero en la práctica es de interés estudiar la relación de X con $\hat{P}r(x)$ y no con $logit[\hat{P}r(x)]$, esto genera dos alternativas, que la relación se mantenga o que cambie, es decir, si $X \uparrow$ y $logit[\hat{P}r(x)] \uparrow$ implican que $\hat{P}r(x) \to 1$ la relación se mantiene directa, pero si implican que $\hat{P}r(x) \to 0$ entonces la relación cambia a inversa; por otro lado si $X \uparrow$ y $logit[\hat{P}r(x)] \downarrow$ implican que $\hat{P}r(x) \to 0$ la relación se mantiene inversa, pero si implican que $\hat{P}r(x) \to 1$ entonces la relación cambia a directa.

Para saber si la relación se mantiene o cambia, basta ver que ocurre con $\hat{P}r(x)$ cuando $logit[\hat{P}r(x)] \uparrow$ y cuando $logit[\hat{P}r(x)] \downarrow$, partimos de que $logit[\hat{P}r(x)] = \ln\left[\frac{\hat{P}r(x)}{1-\hat{P}r(x)}\right]$, para que $logit[\hat{P}r(x)] \uparrow$ es necesario que $\frac{\hat{P}r(x)}{1-\hat{P}r(x)} \uparrow$, pues el logaritmo es una función creciente, para lo que requerimos que $1 - \hat{P}r(x) \to 0$, cosa que ocurre cuando $\hat{P}r(x) \to 1$, esto significa que $logit[\hat{P}r(x)] \uparrow$ cuando $\hat{P}r(x) \to 1$, por lo que la relación se mantiene directa. Cuando $logit[\hat{P}r(x)] \downarrow$ tenemos que $\frac{\hat{P}r(x)}{1-\hat{P}r(x)} \to 0$, y esto ocurre si y sólo si $\hat{P}r(x) \to 0$, entonces cuando $logit[\hat{P}r(x)] \downarrow$, $\hat{P}r(x) \downarrow$ también, en resumen tenemos:

si $\hat{\beta}_1 > 0$ la relación es directa:

$$\begin{array}{llcll} \text{cuando X}\uparrow & logit[\hat{Pr}(x)] \uparrow & \Longleftrightarrow & \text{cuando X}\uparrow & \hat{Pr}(x) \longrightarrow 1 \\ \text{cuando X}\downarrow & logit[\hat{Pr}(x)] \downarrow & \Longleftrightarrow & \text{cuando X}\downarrow & \hat{Pr}(x) \longrightarrow 0 \end{array}$$

si $\hat{\beta}_1 < 0$ la relación es inversa:

$$\begin{array}{llcll} \text{cuando X}\uparrow & logit[\hat{Pr}(x)] \downarrow & \Longleftrightarrow & \text{cuando X}\uparrow & \hat{Pr}(x) \longrightarrow 0 \\ \text{cuando X}\downarrow & logit[\hat{Pr}(x)] \uparrow & \Longleftrightarrow & \text{cuando X}\downarrow & \hat{Pr}(x) \longrightarrow 1 \end{array}$$

A continuación se ilustra una de estas situaciones.

Ejemplo 2.4 Modelo de regresión logística con una variable explicativa dicotómica (M2). Tomemos el caso de 68 pacientes con anemia aplásica, a quienes se les aplica un injerto de médula, la dosis de células de médula puede ser de menos de (dosis baja) o al menos (dosis alta) $3.0x10^8$ células/kg, nuestro interés es el rechazo del injerto (Y=1) que es representado por la variable respuesta, para lo que consideraremos los datos presentados en la siguiente tabla (Matthews, 1988).

Tabla 2.3 Pacientes clasificados de acuerdo a si hubo aceptación o rechazo del injerto de médula al considerar dos dosis

variable respuesta: rechazo de injerto	variable explicativa: dósis de médula: alta (X=0)	baja (X=1)	total
no (Y=0)	28	19	47
sí (Y=1)	4	17	21
total	32	36	68

El rechazo o no del injerto se clasifica en dos categorías, no Y=0 o sí Y=1, al ajustar el modelo de regresión logística la forma logit que obtenemos es:

$logit\hat{P}r(X) = -1.9459 + 1.8347 \cdot$ *dosis de médula*, el estimador del parámetro β_1 es mayor que cero, entonces concluimos que la relación entre la probabilidad de sí rechazar el injerto y los valores que toma la variable explicativa X es directa, es decir, a menor dosis (X=1) la probabilidad de rechazar el injerto aumenta, podemos obtener el valor del momio de éxito (rechazo del injerto) para ambos valores de X, $\hat{\Psi}(0) = e^{\hat{\beta}_0} = 0.1428$ y $\hat{\Psi}(1) = e^{\hat{\beta}_0+\hat{\beta}_1} = 0.8947$, entonces el cociente de momios es $\hat{\Psi}_{1,0} = e^{\hat{\beta}_1} = 6.2654$ e indica que la probabilidad de éxito cuando la dosis de médula es baja (X=1) es seis veces la probabilidad de éxito cuando la dosis de médula es alta (X=0), siendo más probable la aceptación del injerto cuando la dosis es

baja, tenemos que la probabilidad de sí rechazar el injerto de una dosis alta de célula (X=0) es $\hat{Pr}(0) = \frac{1}{1+e^{1.9459}} = 0.1250$ y la probabilidad de rechazar el injerto de una dosis baja (X=1) es $\hat{Pr}(1) = \frac{1}{1+e^{-(-1.9459+1.8347)}} = 0.4722$, confirmando la relación directa entre X y $\hat{Pr}(X)$ descrita por el signo positivo de $\hat{\beta}_1$.

Capítulo 3

Aplicaciones

En este capítulo se ilustrará el uso de algunos modelos que se reportan en el apéndice Taxonomía de los modelos de regresión logística.

3.1 Supervivencia infantil

Ejemplo 3.1 Modelo de regresión logística con dos variables explicativas dicotómicas (M9). Supongamos que deseamos evaluar la supervivencia de recién nacidos según la cantidad de cuidados prenatales recibidos por la madre y la clínica donde ésta fue atendida, para lo que se considera la información de archivo que corresponde a 715 mujeres y se presenta a continuación:

Tabla 3.1 Frecuencia de supervivencia de recién nacidos según la cantidad de cuidados prenatales recibidos por la madre y la clínica donde ésta fue atendida

clínica	cuidados prenatales	supervivencia	
		sobrevive Y=0	muere Y=1
A=0	pocos=0	176	3
	muchos=1	293	4
B=1	pocos=0	197	17
	muchos=1	23	2

La variable respuesta es la supervivencia de los recién nacidos, donde sobrevive $= 0$ y muere $= 1$, las variables explicativas son clínica A $= 0$, B $= 1$ y cuidados prenatales recibidos pocos $= 0$ y muchos $= 1$, como estas dos variables son de tipo categórico las llamaremos factores. A continuación se muestra de forma gráfica la información.

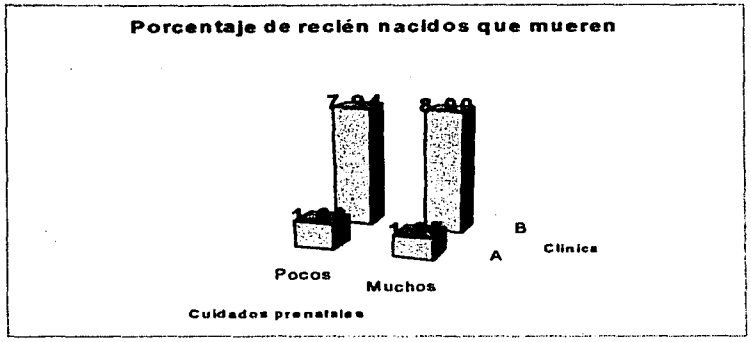

Figura 3.1 Distribución porcentual de recién nacidos que mueren según la cantidad de cuidados prenatales recibidos por la madre y la clínica donde ésta fue atendida

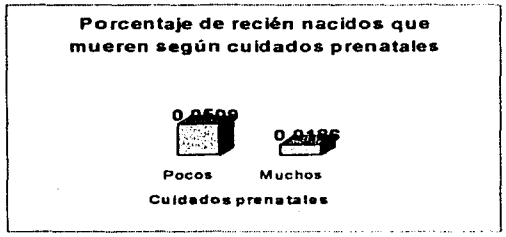

Figura 3.2 a) Distribución porcentual de recién nacidos que mueren según cuidados prenatales

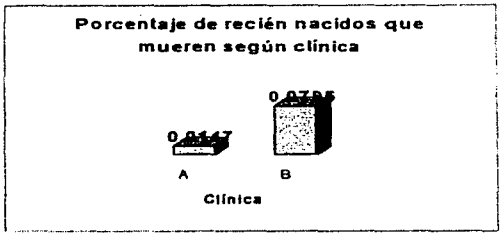

Figura 3.2 b) Distribución porcentual de recién nacidos que mueren según clínica

La figura 3.1 nos muestra que el porcentaje de recién nacidos que mueren es más alto en la clínica B sin importar la cantidad de cuidados prenatales recibidos por la madre, si los cuidados prenatales son pocos entonces el aumento que produce el factor clínica en el porcentaje de recién nacidos que mueren es de 6.27% y si los cuidados prenatales son muchos entonces es de 6.65%, al parecer el aumento es el mismo y en la misma dirección, lo que al parecer indica que la interacción entre ambos factores es nula. También se observa que el porcentaje de recién nacidos que mueren con pocos cuidados prenatales es muy parecido al porcentaje de los que recibieron muchos cuidados prenatales en cada clínica, lo que indicaría que el factor cuidados prenatales no tiene efecto en la supervivencia de recién nacidos. Al analizar las gráficas simples, es decir, con sólo un factor (figuras 3.2), se observa que el porcentaje de mortalidad es más alto cuando los cuidados prenatales son pocos y cuando la atención se dio en la clínica B. A continuación se hace uso del modelo de regresión logística para modelar los datos y verificar estadísticamente las conjeturas anteriores.

El modelo por ajustar es: $logit[Pr(\mathbf{x})] = \beta_0 + \beta_1(cuidados\ prenatales) + \beta_2(clínica) + \varepsilon$, que corresponde al modelo M9 de la taxonomía mostrada en el apéndice, es decir, es un modelo con dos variables explicativas dicotómicas, el modelo ajustado es:

$\widehat{logit}[\hat{Pr}(\mathbf{x})] = -4.137 - 0.110(cuidados\ prenatales) + 1.699(clínica)$ y las estimaciones de los coeficientes se obtienen de la tabla 3.2.

Tabla 3.2 Estimación de los coeficientes del modelo $logit[Pr(\mathbf{x})]=\beta_0+\beta_1(cuidados\ prenatales)+\beta_2(clínica)+\varepsilon$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	X1	-.110	.561	.039	1	.844	.895	.298	2.689
	X2	1.699	.531	10.252	1	.001	5.469	1.933	15.474
	Constant	-4.137	.508	66.411	1	.000	.016		

a. Variable(s) entered on step 1: X1, X2.

La tabla anterior reporta el nivel de significancia descriptivo para la prueba de Wald referente a la nulidad de los coeficientes, considerando que el nivel de prueba es 0.05, el coeficiente β_1 en este caso es el único que no es significativo porque el mínimo riesgo para rechazar la hipótesis nula es 0.844, el cual es mayor que el riesgo prederminado, y esto puede verificarse con el intervalo de confianza para el cociente de momios para cuidados prenatales controlado por el factor clínica $e^{\hat{\beta}_1} = \Psi_{1,0}$ que incluye el valor 1, lo que indica asociación nula de X_1 (cuidados prenatales) con la probabilidad de éxito, entonces la probabilidad de morir dados muchos cuidados prenatales es igual que cuando se reciben pocos cuidados (recibidos en la misma clínica), por lo que no hay efecto de cuidados prenatales en la supervivencia, ahora expresando el intervalo de confianza para $\Psi_{1,0}$ en fracciones tenemos aproximadamente que $\Psi_{1,0} \in (1/3, 3)$, esto dice que el intervalo de confianza abarca las situaciones en que el momio de pocos cuidados prenatales es al menos tres veces el de muchos y el momio de muchos cuidados prenatales es a lo mas tres veces el momio de pocos cuidados prenatales[1], esto no permite concluir que el cociente de momios sea distinto de uno. El coeficiente del factor clínica sí es significativo con el nivel de prueba del 0.05, y el cambiar de la clínica A a la B incrementa la probabilidad de muerte, es decir, la mortalidad es más alta en la clínica B, esto lo vemos en el

[1] Los límites de confianza se pueden expresar como:
$\Psi_{1,0} \in (3\Psi(X_1 = 1) = \Psi(X_1 = 0);\ \Psi(X_1 = 1) = 3\Psi(X_1 = 0))$

valor positivo del estimador[2] y en el cociente de momios para el factor clínica controlado por el factor cuidados prenatales cuyo valor es $\Psi_{1,0} = e^{\hat{\beta}_2} = 5.469$, pues es mayor que uno e indica que la probabilidad de morir en la clínica B es cinco veces mayor que la probabilidad de morir en la clínica A. La inferencia realizada sobre los coeficientes permite concluir que el modelo propuesto no es el indicado, sugiriendo ajustar un modelo sólo con el factor clínica, es decir, $logit[Pr(X_2 = x)] = \beta_0 + \beta_2(clínica) + \varepsilon$, para probar si este modelo reducido realmente ajusta mejor los datos se puede recurrir a la prueba del cociente de verosimilitudes, como en este caso sólo se excluye una variable del modelo completo la hipótesis a probar es la misma que se prueba con el estadístico de Wald, es decir, $H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$ y se procede calculando el estadístico de prueba como la diferencia entre la devianza del modelo reducido y la devianza del modelo total, que se distribuye χ^2 con un grado de libertad (dado que sólo se considera un parámetro bajo la hipótesis nula), se tiene entonces $\chi^2_{(1)_{Calculada}} = 205.634 - 205.595 = 0.039$ y el valor en tablas para una χ^2 con un grado de libertad y probabilidad acumulada $1 - \alpha$, donde $\alpha = 0.05$ es el nivel de significancia, es $\chi^2_{(1)_{Tablas}} = 3.84$, por lo que el valor calculado del estadítico es menor al de tablas y aceptamos la hipótesis nula, es decir, el parámetro no es significativo, confirmando el resultado de la prueba de Wald, es decir, los cuidados prenatales no afectan significativamente la probabilidad de morir teniendo en cuenta la información de las clínicas, lo que confirma que el modelo adecuado es $logit[Pr(X_2 = x)] = \beta_0 + \beta_2(clínica) + \varepsilon$.

La siguiente tabla muestra los resultados del ajuste de este modelo.

Tabla 3.3 Estimación de los coeficientes del modelo $logit[Pr(X_2=x)]=\beta_0+\beta_2(clínica)+\varepsilon$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	X2	1.755	.450	15.244	1	.000	5.786	2.397	13.968
	Constant	-4.205	.381	121.937	1	.000	.015		

a. Variable(s) entered on step 1: X2.

El modelo ajustado es $logit[\hat{Pr}(X_2 = x)] = -4.205 + 1.755(clínica)$ y ambos coeficientes son significativos con un nivel de prueba 0.05, el valor estimado del coeficiente del factor clínica es positivo, lo que indica que la probabilidad de muerte se incrementa en la clínica B ($X_2 = 1$),

[2] Revisar la sección 2.5

y dicha probabilidad es al menos 2.397 y a lo más 13.968 veces mayor en esta clínica que en la clínica A.

Hemos obtenido un modelo para la predicción de la supervivencia de recién nacidos, basado en la clínica en que fue atendida la madre, las predicciones se hacen a partir de la probabilidad de éxito (el recién nacido muere) y para que éstas sean creibles es necesario analizar como son con respecto a lo observado.

Con el modelo ajustado se obtiene la probabilidad estimada de que un recién nacido muera (éxito) dado que la madre fue atendida en la clínica A $\hat{Pr}(X_2 = 0) = 0.01471$ y también la probabilidad de éxito dado que la madre fue atendida en la clínica B $\hat{Pr}(X_2 = 1) = 0.07950$.

Con los valores estimados de la probabilidad de éxito es posible agrupar a los individuos de la muestra y así obtener el número de defunciones esperadas en cada grupo, en este caso se forman dos, uno con probabilidad de éxito 0.01471 que se integra por los 476 recién nacidos cuyas madres fueron atendidas en la clínica A y con un número esperado de defunciones (0.01471)(476)=7, el otro grupo tiene probabilidad de éxito 0.07950 y está integrado por los 239 recién nacidos de madres que fueron atendidas en la clínica B, tiene (0.07950)(239)=19 defunciones esperadas. Si el modelo predice correctamente, entonces para cada grupo el número de defunciones esperadas coincide con el de observadas. A la comparación del número de éxitos predichos, en cada grupo definido por la probabilidad de éxito estimada por el modelo, con el número observado se le conoce como calibración, siendo que ha medida que el número de éxitos observados coincide con el de predichos el modelo está bien calibrado. En este ejemplo podemos obtener de la tabla 3.1 el número de defunciones para cada grupo y con lo cual concluir que el modelo está perfectamente calibrado. Para evaluar la calibración de un modelo se recurre a la prueba estadística de Hosmer y Lemeshow, esta prueba es reportada por el paquete y consiste en comparar el número de éxitos y de fracasos esperados con el de observados, para cada grupo definido por las probabilidades de éxito, a través de la prueba Ji-cuadrada. La tabla 3.4 reporta las frecuencias observadas y esperadas de éxito y fracaso para los grupos.

Tabla 3.4 Tabla de frecuencias para la prueba de Hosmer y Lemeshow

Contingency Table for Hosmer and Lemeshow Test

		supervivencia = sobrevive		supervivencia = muere		
		Observed	Expected	Observed	Expected	Total
Step 1	1	469	469.000	7	7.000	476
	2	220	220.000	19	19.000	239

En este caso no se tienen discrepancias por lo que la calibración del modelo es perfecta como ya se mencionó. La tabla 3.5 contiene el valor del estadístico de prueba, los grados de libertad y el nivel de significancia. Los grados de libertad son el número de grupos menos dos, en este caso es cero y el paquete no reportó información.

Tabla 3.5 Resultados de la prueba de Hosmer y Lemeshow

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1			

Dado que la variable respuesta bajo estudio es dicotómica, se tienen dos posibles resultados para la predicción de la supervivencia de recién nacidos: positivo o negativo, que corresponden a muere o sobrevive respectivamente; estableciendo un punto de corte ρ para clasificar y considerando la tabla 3.6, la probabilidad de obtener positivo dado que el recién nacido realmente muere, es decir, la probabilidad de un verdadero positivo es $\frac{d}{c+d}$ y la probabilidad de obtener positivo dado que el recién nacido realmente sobrevive es $\frac{b}{a+b}$ la probabilidad de un falso positivo; su complemento es $\frac{a}{a+b}$ y corresponde a la probabilidad de obtener un negativo dado que el recién nacido realmente sobrevive, o dicho de otra forma, la probabilidad de un verdadero negativo, esto se conoce como especificidad y la probabilidad de un verdadero positivo como sensibilidad.

Tabla 3.6 De clasificación con punto de corte ρ

Classification Table[a]

Observed			Predicted		
			supervivencia		
			sobrevive	muere	Percentage Correct
Step 1	supervivencia	sobrevive	a	b	a/(a+b) x 100%
		muere	c	d	d/(c+d) x 100%
	Overall Percentage				(a+d)/(a+b+c+d) x 100%

a. The cut value is p

Notemos en la tabla 3.6 que la columna de porcentaje de buena clasificación (percentage correct) reporta el porcentaje de verdaderos negativos, el de verdaderos positivos y el del total de casos bien clasificados, es decir, nos da el valor de la especificidad y el de la sensibilidad en porcentaje.

Para clasificar a los recién nacidos con el modelo es inmediato pensar que el criterio de clasificación consiste en asignar el grupo de pertenencia dependiendo de cual tiene probabilidad de éxito más grande, lo que corresponde a establecer el punto de corte en 0.5.

Tabla 3.7 De clasificación con punto de corte 0.5 correspondiente a la supervivencia de recién nacidos y al modelo con variable explicativa clínica

Classification Table[a]

Observed			Predicted		
			supervivencia		
			sobrevive	muere	Percentage Correct
Step 1	supervivencia	sobrevive	689	0	100.0
		muere	26	0	.0
	Overall Percentage				96.4

a. The cut value is .500

En este ejemplo la sensibilidad del modelo en las predicciones, considerando para el criterio de clasificación el punto de corte 0.5 y recurriendo a la tabla 3.7, es cero y la especificidad es uno; estos valores indican que el modelo es perfectamente específico ya que para todos los recién nacidos que sobreviven su predicción es correcta, sin embargo, no es sensible, es decir, clasifica erróneamente todas la defunciones.

Si la sensibilidad y el complemento de la especificidad coinciden, entonces la probabilidad de obtener un positivo dado que el recién nacido sobrevive es la misma que la de obtener positivo dado que el recién nacido muere, por lo que las predicciones hechas con el modelo no permiten distinguir entre los recién nacidos que mueren y los que sobreviven.

Como para el modelo ajustado en este ejemplo, considerando el punto de corte 0.5, el valor de la sensibilidad es igual al del complemento de la especificidad, entonces el modelo no distingue los éxitos de los fracasos. Discriminación es el concepto que se asigna a la evaluación del grado en que el modelo hace tal distinción, entonces en este caso se concluye que el modelo tienen discriminación nula.

Valores altos de la sensibilidad y la especificidad en conjunto, se refieren a una buena clasificación de los recién nacidos, es decir, que los que sobrevivieron sean clasificados como sobrevivientes y los que no entonces se clasifiquen como no sobrevivientes, teniendo así un alto porcentaje de buena clasificación.

En la tabla 3.7 se reporta un 96.4% de los casos bien clasificados, esto parece indicar que el modelo es altamente sensible y específico, sin embargo, no es así, al analizar la tabla nos percatamos de que todos los recién nacidos son predichos como sobrevivientes debido a que las probabilidades de éxito estimadas son menores a 0.5 y entonces la probabilidad de fracaso, es decir, que el recién nacido sobreviva, es mayor a la de éxito; además el número de observaciones para cada grupo no es igual siendo más los recién nacidos que sobreviven (689) que los que mueren (26) produciendo un porcentaje de buena clasificación alto, esto último no permite concluir que el porcentaje de buena clasificación reportado con un punto de corte de 0.5 sea representativo. La proporción observada de sobrevivientes es $\frac{689}{715} = 0.9636$ y la de recién nacidos que mueren es $\frac{26}{715} = 0.0363$, debido a que estas proporciones no son iguales establecer en este caso el punto de corte para la clasificación en 0.5 no es lo más indicado, pues si se propone como probabilidad a priori de éxito el valor 0.0363 entonces las observaciones no están clasificadas con el punto de corte 0.5 pues de ser así todas caerían en el grupo de recién nacidos que sobreviven contradiciendo la clasificación real. Una vez que se han hecho las reflexiones anteriores, se propone como posible punto de corte adecuado el valor 0.0363, que corresponde a la proporción de exitos observados y se propuso como probabilidad a priori de éxito, la tabla de clasificación obtenida se muestra a continuación.

Tabla 3.8 De clasificación con punto de corte 0.0363 correspondiente a la supervivencia de recién nacidos, y al modelo con variable explicativa clínica

Classification Table[a]

Observed			Predicted		
			supervivencia		Percentage Correct
			sobrevive	muere	
Step 1	supervivencia	sobrevive	469	220	68.1
		muere	7	19	73.1
	Overall Percentage				68.3

a. The cut value is .030

Nótese que el porcentaje de buena clasificación ha disminuido de 96.4% a 68.3% con este nuevo punto de corte, pero más relevante es que el modelo ya no clasifica el total de las observaciones en sobrevive el recién nacido, ahora la especificidad es 0.681y a diferencia de cuando el punto de corte es 0.5 ésta se ha decrementado, sin embargo, la sensibilidad se incrementó de cero a 0.731; con estos valores la sensibilidad y el complemento de la especificidad (0.319) difieren, indicando que el modelo ha mejorado su discriminación. Tener un 68.3% de buena clasificación no es muy satisfactorio, pero en comparación con el 96.4% obtenido anteriormente donde la discriminación de los grupos de la variable respuesta no es adecuada por el punto de corte elegido, este porcentaje es más representativo. Entonces un alto porcentaje global de buena clasificación no siempre significa alta sensibilidad y especificidad simultaneamente, es necesario establecer un punto de corte adecuado para la clasificación, y considerar si el número de observaciones para cada categoría de la variable respuesta es en ambos casos el mismo (la variable respuesta es binaria).

Se ha introducido el concepto de discriminación cuando el criterio de clasificación se establece coniderando un determinado punto de corte, la variación de dicho punto de corte como se ha visto genera cambios en la discriminación del modelo, pero sin importar el punto de corte establecido la buena discriminación busca valores altos en la senibilidad y por supuesto bajos para el complemento de la especificidad, si se calculan estos valores para varios puntos de corte ordenados ascendentemente y se ubican en el plano los puntos (sensibilidad , 1-especificidad) entonces se puede trazar una curva, que es conocida como la curva ROC (receiver operating characteristic), la identidad indica que el modelo no discrimina, entonces se busca que la curva

ROC esté lo más alejada de ésta para que el modelo discrimine acertadamente. Es el área bajo la curva ROC la medida de discriminación, cuanto mayor es el área (está acotada por 1) se concluye que el modelo discrimina mejor, un área mayor a 0.7 indica que el modelo es aceptable.

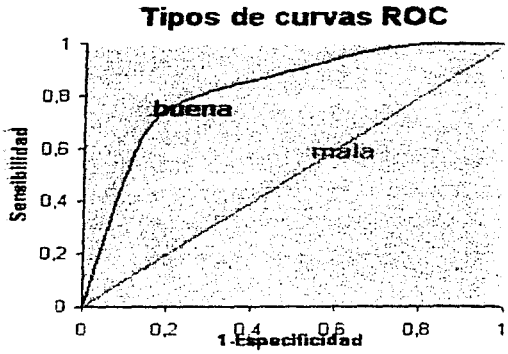

Figura 3.3 Tipos de curvas ROC. Figura tomada de (http://www.hrc.es/bioest/M_docente.html)

Es posible obtener el área bajo la curva ROC con el paquete de cómputo SPSS, en este caso es 0.706 (tabla 3.9) entonces concluimos que el modelo tiene discriminación aceptable.

Tabla 3.9 Reporte correspondiente al área bajo la curva ROC

Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.706	.051	.000	.605	.806

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

La validación del modelo corresponde a analizar si las predicciones de la respuesta bajo estudio son correctas, y ésta se compone de la validez y la generalizabilidad, lo primero corresponde al grado de coincidencia entre lo observado y lo predicho, que se refleja en dos aspectos: la calibración y la discriminación, como se han trabajado hasta el momento, notemos que tales conceptos se han desarrollado únicamente con la información contenida en la muestra,

por lo que podríamos referirnos a la validez como validación interna, se hace hincapié en este adjetivo porque el otro componente de la validación, la generalizabilidad, estudia la capacidad predictiva del modelo en individuos ajenos a la muestra empleada para obtener el ajuste, entonces se habla de validación externa, esta validación es en dos sentidos, con individuos de la población de la que fue tomada la muestra y con individuos de otra población, esto se refiere a la reproducibilidad y a la transportabilidad del modelo respectivamente. Entonces la generalizabilidad extiende la validez del modelo y por lo tanto su utilidad.

Para analizar la generalizabilidad de un modelo es necesario recurrir a técnicas estadísticas que no se tratarán en este trabajo, es importante señalar que el análisis de la validez se ha hecho superficialmente y puede extenderse aún más, la validación es un tema amplio que requiere un estudio aparte, sin embargo, debe tenerse en cuenta al aplicar un modelo de regresión logística con fines predictivos.

3.2 Bioensayos: La dosis efectiva media

La regresión logística tiene una de sus aplicaciones en el campo de la biología, y en los bioensayos podemos encontrar un basto campo de aplicación.

Los bioensayos se refieren al estudio de los efectos causados por un estímulo (biológico, químico, físico, fisiológico o psicológico) en un ambiente experimental controlado, sobre un organismo vivo que actua como agente de prueba (unidad experimental). En este caso la respuesta bajo estudio es el efecto causado por el estímulo, por ejemplo: el agente de prueba muere o sobrevive, o incrementa o no su peso, también podría darse como respuesta el cambio en el peso o en la presión arterial; las primeras dos respuestas son de tipo dicotómico por lo que la regresión logística es una herramienta adecuada, en cambio las últimas dos, los cambios en el peso y la presión arterial, caen dentro de una escala continua y no son tratadas por nuestro modelo bajo estudio. Un típico caso en que se recurre a los bioensayos es cuando se pretende analizar la magnitud del efecto que tienen determinadas dosis del estímulo bajo estudio, es decir, la relación entre la dosis y la respuesta correspondiente.

La dosis efectiva media indica la dosis requerida para que el 50% de las unidades experimentales presente el efecto bajo estudio como respuesta, también es conocida como dosis letal

media cuando se trabaja con dosis de veneno por ejemplo y se registra la muerte como éxito.

La dosis efectiva media bajo el Modelo de regresión logística con una variable explicativa continua (M1) es estimada por $-\frac{\hat{\beta}_0}{\hat{\beta}_1}$; recordemos la expresión (1.6) y busquemos la solución cuando $\hat{Pr}(x) = \frac{1}{2}$,

$$\frac{1}{2} = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x)}} \iff e^{-(\hat{\beta}_0+\hat{\beta}_1 x)} = 1 \iff -(\hat{\beta}_0+\hat{\beta}_1 x) = \ln(1) = 0 \iff x = -\frac{\hat{\beta}_0}{\hat{\beta}_1}$$

Ejemplo 3.2 Modelo de regresión logística con una variable explicativa continua (M1). Supogamos que se quiere estudiar el efecto que causa en ratones, agentes de prueba, cierta droga; el estímulo es la dosis de droga D que se aplica y se fijan siguiendo una progresión geométrica (1, 2, 4, 8, 16), el efecto se representa con la variable respuesta Y y se considera éxito la muerte del ratón (Y=1 si el ratón muere, Y=0 si sobrevive). En la siguiente tabla se muestran los datos correspondientes a este bioensayo.

Tabla 3.10 Datos observados de la respuesta de 75 ratones a los que se aplicó una dosis de droga

Dosis	Agentes de prueba totales	Número observado de éxitos
$D_1 = 1$	15	1
$D_2 = 2$	15	3
$D_3 = 4$	15	9
$D_4 = 8$	15	13
$D_5 = 16$	15	15

Como en este caso se ha definido Y=1 si el ratón muere, se entiende que el objetivo del análisis es estudiar el efecto letal que tiene la droga en los ratones, entonces el modelo de regresión logística va a modelar la probabilidad de que un ratón muera dado que se le aplicó cierta cantidad de droga. El modelo ajustado es el siguiente:

$$\hat{Pr}(X) = \frac{1}{1+\exp(2.566-0.622X)}$$

Donde los estimadores son $\hat{\beta}_0 = -2.566$ y $\hat{\beta}_1 = 0.662$, ambos significativos al nivel de prueba 0.5, el signo del último indica que la relación entre la dosis y la probabilidad de muerte es directa, por lo tanto a mayor dosis mayor mortalidad, calculemos ahora la siguiente razón de momios:

$$\hat{\Psi} = \frac{\hat{\Psi}(x_1)}{\hat{\Psi}(x_0)} = e^{\hat{\beta}_1} = 1.863 \text{ siempre que } x_1 - x_0 = 1$$

Esto indica que hay efecto de la dosis en los ratones, pues al comparar dos dosis que difieren en una unidad los correspondientes momios son diferentes, y como la razón de momios es mayor a uno, el momio de la dosis mayor excede al de la menor, siendo así para cualquier par de dosis que se comparen, entonces la probabilidad de muerte respecto a la de no muerte es mayor a medida que aumenta la dosis.

Tabla 3.11 Probabilidades estimadas de éxito y número esperado de éxitos para cada dosis

Dosis	Probabilidad estimada de éxito	Número esperado de éxitos
$D_1 = 1$	0.12520	1.88
$D_2 = 2$	0.21049	3.16
$D_3 = 4$	0.48060	7.21
$D_4 = 8$	0.91766	13.76
$D_5 = 16$	0.99938	14.99

La dosis letal media en este caso es 3.8761, es decir, con una dosis de droga aproximadamente de 4 unidades la mitad de los agentes de prueba mueren, en la tabla 3.10 se reporta que con la dosis D_4 de 15 individuos observados 9 mueren, esto corresponde al 60% de las observaciones para esa dosis. La tabla 3.11 reporta que la probabilidad estimada de éxito para la dosis D_4 es $\hat{Pr}(X = 4) = 0.4806$, aproximadamente 0.5 y el número esperado de éxitos es 7, lo que concuerda con el concepto de dosis letal media, la mitad de los individuos a los que se les aplicó la dosis D_4 se espera que presenten como respuesta el éxito.

3.3 Bioensayos: El concepto de covariable

Ejemplo 3.3 Modelo de regresión logística con una variable explicativa dicotómica y una covariable continua (E1). A continuación se reportan las observaciones obtenidas para realizar un bioensayo que consiste en estudiar la potencia de una nueva droga respecto a la droga estándar, para la droga estándar se tienen cinco dosis y para la nueva se establecen cuatro; cada dosis de cada droga es inyectada a 20 ratones, que actuan como agentes de prueba, la respuesta de interés es la letalidad que tienen las drogas.

Tabla 3.12 Frecuencias observadas de la respuesta de ratones a dos tipos de droga

Droga	Dosis	Frecuencias	
		Vive	Muere
Estándar	1	19	1
	2	15	5
	4	11	9
	8	4	16
	16	1	19
Nueva	1	16	4
	2	12	8
	4	5	15
	8	2	18

En este caso se tiene un factor de estudio X_1 que es el tipo de droga, $X_1 = 1$ para la droga nueva y $X_1 = 0$ para la droga estándar; y una variable continua X_2 que representa la dosis, como se puede observar en la tabla 3.12 las dosis siguen una progresión geométrica, así que al transformar con el logarítmo base 2 se tendrá una progresión lineal, para este ejemplo se trabajará con dicha transformación.

Una vez que se han definido las variables explicativas y la variable respuesta, el modelo de regresión logística que de inmediato se piensa ajustar es:

$$Pr(\text{muerte} \mid \text{tipo de droga, dosis}) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\epsilon)}} \qquad (3.1)$$

el ajuste del modelo es:

$$\hat{Pr}(\text{muerte} \mid \text{tipo de droga, dosis}) = \frac{1}{1+e^{-(-2.626+1.034X_1+1.313X_2)}}$$

Sin embargo, antes de continuar con la interpretación de este modelo ajustado, es importante hacer notar algunos conceptos.

Nuestro objetivo es saber si la droga nueva es más eficaz en la mortalidad de ratones, que la droga estándar, para lo que bastaría con ajustar un modelo de regresión logística simple, pero al conocer la forma como es administrada cada droga, nos percatamos de que la droga estándar considera cinco tipos de dosis y la droga nueva sólo los primeros cuatro, por lo que no

podemos hacer una comparación inmediata, y es necesario introducir al modelo la variable X_2 dosis de la droga, como covariable, es decir, una variable explicativa que influye en la variable respuesta sin que haya sido controlada. Para comprender mejor este concepto, pensemos en el siguiente caso hipotético, supongamos que se lleva a cabo un maratón, y se tienen dos atletas, entonces nos interesa saber cual de los dos recorre en menor tiempo la distancia correspondiente, supongamos que el atleta A es el vencedor, todo marcha bien, hasta que alguien comenta que no obstante su edad ha ganado, entonces nos damos cuenta de que el atleta A tiene 40 años, y el atleta B tiene 25 años, es evidente que el redimiento no es el mismo para ambos atletas, y es precisamente esto lo que propicia que al atleta A se le de una ventaja de ciertos kilómetros, entonces resulta que la edad del competidor también es una variable que explica el resultado de la competencia, sin embargo no había sido contralada pero es conocida, es precisamente a está variable a la que llamaremos covariable. Una vez que hemos introducido la covariable al modelo, se tienen dos variables explicativas y entonces surge el concepto de interacción, es decir, el efecto conjunto del tipo de droga y la dosis, en este caso la interacción debe suponerse nula, pues de no ser así los resultados obtenidos pueden no ser coherentes; a continuación se ejemplifican ambas situaciones. Sea d=dosis y D=droga, la tabla 3.13 a) contiene los datos en el caso de interacción nula, y la tabla 3.13 b) cuando sí hay interacción y es antagónica, ambas muestran la frecuencia de mortalidad de ratones.

Tabla 3.13 a) Ejemplo de interacción nula

	D_1	D_2	promedio
d_1	8	6	7
d_2	6	4	5

Tabla 3.13 b) Ejemplo de interacción antagónica

	D_1	D_2	promedio
d_1	8	6	7
d_2	6	8	7

En el ejemplo de interacción nula, tenemos que en promedio la dosis d_1 produce un mayor efecto que la dosis d_2 de ambas drogas, veamos que ambas dosis decrementan su efectividad idénticamente (dos unidades) al pasar de la droga D_1 a la D_2, esto indica que la interacción es nula, si ese decremento no fuera idéntico entonces sí habría interacción y ésta sería sinérgica; por otro lado cuando el efecto de una dosis decrece de una droga a otra, pero el efecto de la otra dosis se incrementa, se trata de una interacción antagónica, como lo es el ejemplo correspondiente a la tabla 3.13 b), en donde concluiríamos que el efecto de las dosis es el mismo en ambas drogas, pues en promedio así es, el detalle aquí es que tal conclusión es errónea, debido a que

la interacción no es nula.

Lo anterior justifica la introducción de las dos variables explicativas (una de ellas como covariable), sin embargo podemos extender aún más el modelo, de manera que nos permita dar conclusiones más acertadas; recordemos que es necesario que el efecto de la dosis sea el mismo en ambas drogas, lo que correspondería a una interacción nula, para así poder comparar el efecto del tipo de droga, esto no se puede probar en el modelo de la expresión (3.1), por lo que Freeman (1987) propone el siguiente:

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right) = D_1X_1 + D_2X_2 + \beta_1X_3 + \beta_2X_4 \qquad (3.2)$$

donde:

P_{ij} = Pr(Y=1 | i'ésima droga, j'ésima dosis)

d_{ij} =j'ésima dosis de la i'ésima droga

$X_1 = 1$ si $i = 1$ (droga estándar), 0 en otro caso

$X_2 = 1$ si $i = 2$ (droga nueva), 0 en otro caso

$X_3 = \log_2(d_{ij})$ si $i = 1$, 0 en otro caso

$X_4 = \log_2(d_{ij})$ si $i = 2$, 0 en otro caso

Bajo el modelo (3.2) se plantea la siguiente hipótesis nula:

$$H_0 : \beta_1 - \beta_2 = 0 \quad \text{(interacción nula)} \qquad (3.3)$$

Es decir, que el efecto de la dosis es el mismo en las dos drogas, buscando aceptar la hipótesis nula para así poder continuar el análisis. Debido a que en el módulo de regresión logística de los paquetes de cómputo no se tiene la opción de realizar esta prueba de hipótesis y el logit de la proporción de éxito observada para todos los casos está definido, entonces se plantea como alternativa para la prueba de hipótesis estimar un modelo de regresión lineal con variable respuesta el logit de la proporción de éxito observada y variables explicativas las mismas que considera el modelo logístico (3.2), el paquete SYSTAT realiza la prueba de hipótesis reportando que con el nivel de prueba 0.05 se acepta que el efecto de las dosis de la droga estándar es el mismo que el efecto de las dosis de la droga nueva, por lo que se concluye que la interacción es nula. Como $\beta_1 = \beta_2 = \beta$, el modelo (3.2) se reduce al siguiente:

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right) = D_1X_1 + D_2X_2 + \beta(X_3 + X_4) \qquad (3.4)$$

y entonces se compara el efecto de las drogas probando la siguiente hipótesis:

$$H_0 : D_1 - D_2 = 0 \qquad (3.5)$$

Para saber si existe diferencia significativa, recurriendo al paquete SYSTAT con la misma estrategia pero con las variables explicativas del modelo (3.4) se rechaza la hipótesis nula (3.5), es decir, se concluye que el efecto de las drogas no es el mismo. Freeman (1987) estudia la diferencia entre las drogas con la "potencia relativa" de la droga nueva, que se defiene como $\mathbf{P} = exp(\frac{D_2 - D_1}{\beta})$. El modelo ajustado de regresión logística para la expresión (3.4) es:

$logit[\hat{P}r(\underset{\sim}{X})] = -2.626X_1 - 1.592X_2 + 1.313(X_3 + X_4)$, con todos sus coeficientes significativos al nivel de prueba 0.05. De donde se calcula $\mathbf{P} = exp(\frac{-1.592+2.626}{1.313}) = 2.1979$ y se concluye que la potencia de la droga nueva es dos veces la de la droga estándar. En conclusión el efecto de dosis es el mismo para las dos drogas, sin embargo, el efecto de las dos drogas no es el mismo, siendo más potente la droga nueva que la estándar.

Capítulo 4

Conclusiones

El desarrollo de este trabajo nos permite concluir que: Cuando se quiere estudiar la relación de una variable respuesta dicotómica con otras variables explicativas continuas y/o categóricas el modelo de regresión lineal no es el adecuado, por lo que se recurre al modelo de regresión logística.

En este trabajo el enfoque con que se abordó el modelo corresponde a la transormación logit, mas no es el único pues el modelo de regresión logística pertenece a la familia de los modelos lineales generalizados que es una teoría más general.

El modelo logístico implícitamente sí es lineal o linealizable mediante la transformación logit.

Existen varias medidas para estudiar la asociación de variables, una de ellas es la razón de momios y es la que el modelo de regresión logística implícitamente maneja.

La información que aportan los parámetros del modelo, como el tipo de relación entre una variable X y la probabilidad de éxito $\hat{Pr}(X)$, o la asociación de variables mediante las razones de momios por ejemplo, hacen que el modelo de regresión logística sea práctico en cuanto a interpretación.

Para establecer en la variable respuesta el evento correspondiente al éxito, es necesario considerar el contexto y los objetivos de la situación bajo estudio, sin embargo, los resultados obtenidos de este modelo coinciden con los reportados por el modelo en el que se redefine el éxito como fracaso.

El campo de aplicación del modelo de regresión logística es extenso, una de sus aplicaciones

es el desarrollo de modelos con fines predictivos, debido a que el modelo logístico proporciona la probabilidad estimada de éxito dados ciertos valores de las variables explicativas, es posible predecir para un individuo su grupo de pertenencia (en la variable respuesta), según la probabilidad que se determine como mínima para aceptar que el individuo presenta el éxito (punto de corte). Para poner en práctica un modelo predictivo es necesario probar su validación, en este trabajo se menciona brevemente lo correspondiente a la validación interna, la calibración y la discriminación, pero también debe estudiarse su validación externa, este tema requiere un análisis aparte. Otra de las aplicaciones del modelo se encuentra en los bioensayos, específicamente en el cálculo de la dosis efectiva media y la potencia relativa de un tratamiento cuando es comparado con otro, el concepto de covariable también se trata en el caso de la regresión logística como se vio en el capítulo 3. El uso de la regresión logística es muy amplio y cada área en la que se aplica ha desarrollado un lenguaje propio y mayor interés en algunos conceptos.

Como se pudo observar a lo largo de este trabajo, el análisis de regresión logística va más allá de la estimación de los parámetros, la estructura del modelo genera una amplia gama de posibilidades y como se aprecia en la taxonomía (apéndice A) el modelo puede variar según el número y tipo de las variables explicativas.

Un comentario referente a la información obtenida de la revisión bibliográfica que no se presenta en este trabajo es el siguiente: Como se ha mencionado el modelo logístico puede aplicarse con un carácter predictivo y entonces establecer una regla de clasificación o discriminación, otro recurso estadístico para estos fines es el análisis discriminante, que proporciona una regla para la clasificación de individuos en dos o más grupos a diferencia de la regresión logística revisada que sólo acepta dos debido a que la variable respuesta es dicotómica, pero esta herramienta alternativa sólo acepta variables explicativas cuantitativas para la discriminación, en cambio la regresión logística acepta variables (explicativas) tanto cuantitativas como cualitativas. Con esto se pretende estar conscientes de que existen otras técnicas y pueden emplearse según sean los objetivos y el tipo de información a analizar.

Apéndices

Apéndice A Taxonomía de los modelos de regresión logística

Tabla A.1 Tipos de modelos de regresión logística

Variable respuesta	Variable explicativa 1	Variable explicativa 2	Modelo	Ejemplo
dicotómica	continua		M1	Ejemplo 1.2 (p. 10) Ejemplo 3.2 Bioensayos: La dosis efectiva media (p. 43)
	dicotómica		M2	Ejemplo 2.2 (p. 18) Ejemplo 2.4 (p. 30)
	nominal		M3	
	ordinal		M4	
	continua	continua	M5	
		dicotómica	M6	
		nominal	M7	
		ordinal	M8	
	dicotómica	dicotómica	M9	Ejemplo 3.1 Supervivencia infantil (p. 32)
		nominal	M10	
		ordinal	M11	
	nominal	nominal	M12	
		ordinal	M13	
	ordinal	ordinal	M14	

Tabla A.2 Caso especial del modelo de regresión logística

Variable respuesta	Variable explicativa	Covariable	Modelo	Ejemplo
dicotómica	dicotómica	continua	E1	Ejemplo 3.3 Bioensayos: El concepto de covariable (p. 44)

Los modelos de regresión logística como se ha mencionado ya, se caracterizan por el tipo de variable respuesta binaria, sin embargo es común encontrar situaciones en que es de interés

estudiar más de dos categorías en la respuesta, entonces el modelo adecuado es el de regresión logística generalizado, este trabajo sólo abordará el modelo para una variable respuesta binaria. En lo que respecta al tipo de variables explicativas, éstas pueden ser de tipo categórico y/o continuo, teniendo así cuatro tipos de variables explicativas: continuas, categóricas dicotómicas[1], categóricas politómicas no ordenadas (nominales) y categóricas politómicas ordenadas (ordinales); por lo tanto si modelamos con sólo una variable explicativa, tenemos cuatro posibles modelos de regresión logística, y para el caso de dos variables explicativas el número de posibles modelos se verá incrementado como se muestra en la tabla A.1.

Cuando se pretende ajustar un modelo de regresión logística se plantea cuantas variables explicativas se estudiarán y cual es su tipo, hay casos particulares en los que además de las variables explicativas consideradas existen otras que podrían tener influencia en la respuesta bajo estudio, éste es el caso de las covariables; en este trabajo se introdujo e ilustró este concepto en la sección 3.3, a través de un modelo de regresión logísta con una variable explicativa dicotómica y una covariable continua, en la tabla A.2 se introduce este caso especial.

Podemos observar que son varios los tipos de modelos y no todos se abordarán en el presente trabajo, el manejo de los resultados no es el mismo y por tanto es conveniente tratarlos por separado. En el capítulo tres se abordaron tres ejemplos para ilustrar algunas aplicaciones del modelo logístico, la supervivencia infantil como un modelo con fines predictivos y dos bioensayos para el concepto de dosis efectiva media y el de covariable respectivamente, además se introdujeron otros en el desarrollo del trabajo para mostrar algunos conceptos, como puede verse en la lista general de ejemplos.

[1] Una variable dicotómica puede ser nominal u ordinal, sin embargo, el análisis en ambos casos es el mismo, por lo que se trabaja indistintamente.

Apéndice 1.1 Comparación de las probabilidades de éxito obtenidas con un modelo no lineal con las obtenidas con un modelo de regresión logística para un ejemplo particular

En este apéndice se hace la comparación de las estimaciones obtenidas al ajustar los modelos (1.2) y (1.4) a la siguiente situación hipotética, continuando con el contexto del ejemplo de la sección 1.1, la siguiente tabla contiene los datos de 320 pacientes con cáncer, la columna LI indica el valor que toma la variable explicativa X valor de LI y ha sido redefinida en la siguiente columna por comodidad (puede verificarse que haciendo esto las estimaciones de la probabilidad de recuperación no se alteran, porque los valores originales de LI están equiespaciados), la columna nombrada éxitos corresponde a la frecuencia de pacientes para cada valor de LI que sí se recuperaron y la columna total indica el número de pacientes que presentaron el respectivo valor de LI, haciendo el cociente pacientes recuperados entre total de pacientes se obtuvo la proporción de recuperaciones para cada valor de LI. La variable respuesta en este caso es la proporción de pacientes recuperados.

LI	LI redefinido	éxitos	total	proporción	estimación RNL	estimación RL
0.4	0	12	20	0.60000	0.59448	0.59592
0.5	1	12	20	0.60000	0.61843	0.61958
0.6	2	13	20	0.65000	0.64182	0.64269
0.7	3	13	20	0.65000	0.66455	0.66515
0.8	4	14	20	0.70000	0.68655	0.68688
0.9	5	14	20	0.70000	0.70773	0.70782
1.0	6	15	20	0.75000	0.72806	0.72792
1.1	7	15	20	0.75000	0.74747	0.74713
1.2	8	15	20	0.75000	0.76594	0.76542
1.3	9	16	20	0.80000	0.78346	0.78277
1.4	10	16	20	0.80000	0.80000	0.79918
1.5	11	16	20	0.80000	0.81558	0.81464
1.6	12	17	20	0.85000	0.83020	0.82916
1.7	13	17	20	0.85000	0.84389	0.84277
1.8	14	17	20	0.85000	0.85666	0.85548
1.9	15	17	20	0.85000	0.86855	0.86732

La expresión (1.2) corresponde a un modelo no lineal, para obtener las estimaciones de los parámetros se recurrió al paquete de computo estadístico SPSS y al módulo correspondiente a regresión no lineal, para realizar la rutina es necesario introducir los valores iniciales de los parámetos, para X=0 la proporción correspondiente es 0.60 y entonces el valor inicial para β_0 es

0.405465, si ahora X=1 la proporción también es 0.60 y entonces el valor inicial de β_1 es cero; el modelo de regresión no lineal ajustado es $\hat{P}_x = \frac{1}{1+e^{-(0.38250+0.10038x)}}$ y las proporciones estimadas se encuentran en la penúltima columna de la tabla anterior. Para el modelo de regresión logística la variable explicativa es la misma, pero a diferencia del modelo no lineal que consideró una variable continua en la respuesta, aquí se toma la variable dicotómica Y=1 si el paciente se recuperó y Y=0 si el paciente no se recuperó, el modelo de regresión logística ajustado es $\hat{Pr}(x) = \frac{1}{1+e^{-(0.38851+0.09926x)}}$ y la última columna de la tabla contiene las probabilidades estimadas para cada valor de X.

Observamos que las estimaciones obtenidas con ambos modelos no difieren por mucho, como era de esperarse, teniendo dos opciones para el análisis de nuestra información, sin embargo como se puede leer en este trabajo los alcances de la regresión logística van más allá de la estimación de una probabilidad.

Apéndice 1.2 El modelo de regresión logística como caso particular de los modelos lineales generalizados

El modelo de la expresión (1.3) es un caso particular de los llamados modelos lineales generalizados que se caracterizan porque:

1) el evento en estudio tiene como función de distribución, a un elemento perteneciente a la familia exponencial, como es el caso de la distribución Binomial y el de la Poisson, esto se conoce como componente aleatorio,

2) se considera una función lineal de las variables explicativas, llamada componente sistemático, y

3) "una liga que describe la relación funcional entre el componente sistemático y el valor esperado del componente aleatorio" (Agresti, 1990).

Estos tres elementos permiten modelar la asociación entre una variable respuesta categórica dicotómica (ocurrencia del evento o no ocurrencia del mismo) y las variables explicativas. En este caso en que la variable respuesta Y es dicotómica y tiene distribución Bernoulli con parámetro $P_i = \Pr(Y_i = 1)$, su función de densidad es:

$f(Y_i; P_i) = P_i^{Y_i}(1 - P_i)^{1-Y_i} = (1 - P_i)[P_i/(1 - P_i)]^{Y_i} = (1 - P_i)\exp[Y_i \ln(\frac{P_i}{1-P_i})]$ y como se puede apreciar tiene la forma $f(Y_i; P_i) = a(P_i)b(Y_i)\exp[Y_iQ(P_i)]$ de la familia exponencial, donde $Q(P_i) = \ln(\frac{P_i}{1-P_i})$ es llamado el parámetro natural de la distribución; el componente sistemático se refiere a un conjunto de variables explicativas bajo una función lineal, esto es $\eta_i = \beta_0 + \beta_1 X_i$; el último componente se refiere a la liga que existe entre el componente aleatorio y el sistemático, sabemos que $E(Y_i) = P_i$, entonces tenemos que $\eta_i = \beta_0 + \beta_1 X_i$ está vinculado a $E(Y_i) = P_i$ mediante una función $\eta_i = G(P_i)$, entonces $G(P_i) = \beta_0 + \beta_1 X_i$, la función liga que transforma la media de la variable Y_i al parámetro natural es llamada liga canónica y es definida como $G(P_i) = Q(P_i) = \ln(\frac{P_i}{1-P_i})$, de donde $\ln(\frac{P_i}{1-P_i}) = \beta_0 + \beta_1 X_i$ es el modelo de regresión logística que usa la liga $\eta_i = \ln(\frac{P_i}{1-P_i})$ llamada logit.

Apéndice 2.1 Sobre el software para realizar el análisis de regresión logística

Como se ha mencionado el software que se uso para el ajuste del modelo de regresión logística en los ejemplos de este trabajo, es el paquete de cómputo SPSS versión 10.0, éste no sólo trata la regresión logística, es más completo y este tema corresponde a uno de sus tantos módulos, es un paquete comercial y muy socorrido por la academia; sin embargo, no es el único, podemos mencionar el S-PLUS 2000, el SYSTAT y el XLSTAT por ser con los que se tuvo relación, Silva (1995) menciona también los paquetes: BMDP, TRUE EPISTAT, SAS, EGRET y GLIM. Existen además paquetes específicos para la regresión logística, Silva (1995) menciona algunos de los que han sido desarrollados.

La parte central de los reportes desplegados por la paquetería, consisten en la estimación puntual de los parámetros, sus correspondientes errores estándar y el estadístico para probar la nulidad de los parámetros; dependiendo de a cual se recurra el reporte puede incluir la estimación del intervalo de confianza para los parámetros, la evaluación de la bondad del ajuste del modelo, tablas de clasificación, etc., algunos, entre esos SPSS, también calculan las probabilidades estimadas de éxito y la predicción del grupo de pertenencia (considerando un punto de corte para la clasificación). En el caso de SPSS para evaluar la discriminación del modelo, la curva ROC no pertenece al módulo de regresión logística, pero sí a la parte gráfica del paquete, lo que permite ampliar el análisis de la validez del modelo. Considerando la variedad de software y el acceso que se tenga a ellos, los reportes se pueden complementar para un mejor análisis.

Bibliografía

Agresti, Alan. **Categorical data analysis.** New York: John Wiley, 1990.

Draper, Norman y Harry Smith. **Applied regression analysis.** 2a. ed. New York: Jonh Wiley, 1981.

Flury, Bernard. **A first course in multivariate statistics.** New York: Springer, 1997.

Freeman, Daniel H. **Applied categorical data analysis.** New York: M. Dekker, 1987.

Hosmer, David y Stanley Lemeshow. **Applied logistic regression.** New York: John Wiley, 1989.

Justice, Amy et al. "Assessing the generalizability of prognostic information." Annals of Internal Medicine 130.6 (1999): 515-524.

Kleinbaum, David et al. **Applied regression analysis and other multivariable methods.** 3rd. ed. Pacific Grove, California; Mexico City: Duxbury Press, 1998.

Matthews, David, Santiago Pueyo y Vernon Farewell. **Estadística médica: Aplicación e interpretación.** Barcelona; México: Salvat, 1988.

Méndez, Ignacio. **Modelos estadísticos lineales. Interpretación y aplicaciones.** México: Foccavi/ Conacyt, 1976.

Norusis, Marija y SPSS. **SPSS advanced statistics user's guide.** Chicago, Illinois: Spss, 1990.

Rué Monné, Montserrat et al. "Utilización de los modelos probabilísticos de mortalidad (MPM II) para evaluar la efectividad de la atención a pacientes en estado crítico." Medicina Clínica (Barcelona) 106.15 (1996): 565-570.

Silva, Luis. **Excursión a la regresión logística en ciencias de la salud.** Madrid: Díaz de Santos, 1995.

Snedecor, George y William Cochran. **Métodos estadísticos.** México: Compañía continental, 1971.

http://www.hrc.es/bioest/M_docente.html