UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

INTRODUCCIÓN A LA REGRESIÓN LOGÍSTICA : APLICACIÓN EN CASOS BIOLÓGICOS

T E S I S

QUE PARA OBTENER EL TÍTULO DE

ACTUARIA

P R E S E N T A :

CLAUDIA LETICIA SÁNCHEZ SÁNCHEZ

DIRECTOR DE TESIS: M. en A.P. MARÍA DEL PILAR ALONSO REYES

2003 FACULTAD DE CIENCIAS
SECCION ESCOLAR

A

PAGINACION DISCONTINUA

DRA. MARÍA DE LOURDES ESTEVA PERALTA
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:

" Introducción a la regresión logística: Aplicación en casos Biológicos "

realizado por SANCHEZ SANCHEZ CLAUDIA LETICIA

con número de cuenta 0-9437817-5, quién cubrió los créditos de la carrera de ACTUARIA.

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis
Propietario M. en A.P. MARIA DEL PILAR ALONSO REYES
Propietario M. en C. JOSE ANTONIO FLORES DIAZ
Propietario ACT. JAIME VÁZQUEZ ALAMILLA
Suplente ACT. GERARDO CHÁVEZ HEREDIA
Suplente ACT. MARYPAOLA JANETT MAYA LÓPEZ

Consejo Departamental de Matemáticas

M. en C. JOSE ANTONIO FLORES DIAZ

B

Agradecimientos.

A mis padres porque me dieron la oportunidad de llegar a este punto tan importante en mi vida, este logro es tanto suyo como mío. A ti papá por enseñarme que en esta vida hay que luchar por lo que se quiere y que hay que obtenerlo de la mejor manera. A ti mamá porque me enseñaste a ser paciente con todo lo que hago.

A mis hermanos Memo y Miriam, con mucho cariño y amor, esperando ver que también le den esta satisfacción a papá y mamá.

A mi Manolo, por tu apoyo incondicional durante la carrera , por ser mi ejemplo a seguir y por hacer de mí una mejor persona.

Mi más sincero agradecimiento a la M. en A. P. María del Pilar Alonso Reyes por su inapreciable colaboración en la dirección de esta tesis.

C

INTRODUCCIÓN

El análisis de regresión nace y se desarrolla en dos matices culturales distintas: la francesa y la inglesa. En la primera vinculado a la astronomía y en la segunda a estudios eugenésicos.

El desarrollo del análisis de regresión, en Francia, estuvo asociado a tres problemas planteados en el siglo XIX:

1) Representar y determinar matemáticamente los movimientos de la Luna
2) Dar cuenta de una desigualdad no periódica en los movimientos de los planetas Júpiter y Saturno
3) Determinar la forma de la Tierra

Se trataba entonces de ajustar las ecuaciones derivadas de la teoría a los resultados de las observaciones astronómicas. El ajuste implicaba más ecuaciones que incógnitas en la medida que se tenía un número apreciable de observaciones y las segundas se reducían a unos pocos parámetros. Pareciera que en estos estudios no se dudaba acerca de la teoría, ésta era tomada por buena, las desviaciones entre los resultados que arrojaba el modelo y los datos observados, se suponía, tenían su origen en los errores de medición, los cuales, obviamente, eran considerados aleatorios.

La solución matemática a este problema se debe a Adrien Legendre, quién propone para resolver este problema "la técnica mínimo cuadrática". La estimación mínimo cuadrática ordinaria así como sus variantes se han

utilizado profundamente en el ajuste de modelos de regresión, ya sea lineales o susceptibles de ser linealizados.

El desarrollo del análisis de regresión a mediados del siglo XIX en Inglaterra, en un ambiente cultural claramente distinto al francés; se enmarcó en el debate sobre las diferencias de clase, de raza y de inteligencia. Francis Galton (1822-1921), quién trabajó en estadística, genética y sicología de las diferencias individuales estuvo signado por su interés en mejorar la raza. La técnica de regresión le permitió predecir las características de los hijos a partir de los rasgos de los padres. El trabajo de Galton, lo continuó Karl Pearson y R. A. Fisher. El programa social y político que orientó su quehacer de investigación descansaba en las siguientes premisas:

1) Suponía que la estructura de clases inglesa reflejaba las habilidades innatas de sus habitantes
2) Que la cúspide de la estructura social estaba formada por la elite profesional y la base por los pobres, los desempleados y los criminales
3) Debían alentarse los nacimientos en los primeros y limitarse en los segundos

Aumentar la inteligencia y las habilidades de los nacionales de un país le permitiría a éste, por ser el país más apto, enfrentar con mayor probabilidad de éxito la lucha darwiniana entre las naciones.

El planteamiento clásico del modelo de regresión supone que todas las variables son métricas, es decir, medidas en escala de intervalo o de razón. Esta condición se erige en una fuerte barrera para aplicar ésta técnica

estadística a diversos problemas. Sin embargo esta limitación fue superada al introducir primero, *variables explicativas ficticias(dicotómicas)* que dieran cuenta de la presencia o ausencia de un evento particular.

Es así como la regresión ganó en ductilidad, haciéndose más atractiva a los ojos de los científicos y estudiosos de este tipo de problemas. Sin embargo, el modelo, desarrollado hasta este punto, aún requería que la variable dependiente fuese métrica.

El ajuste de un modelo de regresión lineal cuando la variable dependiente es dicotómica conlleva una serie de anomalías en el modelo de regresión estándar. Estos problemas se superaron aplicando transformaciones logit a la variable dependiente. La teoría estadística se ha generalizado para el caso en que la variable dependiente contiene más de dos categorías, pero su aplicación se ve restringida por la escasa disponibilidad de paquetes de cómputo que incluyan las rutinas de cálculo que permitan su empleo.

La técnica de la regresión logística se originó en la década de los 60's con el trabajo de Cornfield, Gordon y Smith, en 1967 Walter y Duncan la utilizan ya en su forma actual.

Su uso se incrementa desde principios de los 80's como consecuencia de los adelantos ocurridos en el campo de la computación.

El objetivo de esta técnica estadística es expresar la probabilidad de que ocurra un hecho en función de ciertas variables, que se consideran potencialmente influyentes. La regresión logística, al igual que otras técnicas estadísticas multivariadas, da la posibilidad de evaluar la influencia de cada

una de las variables independientes sobre la variable respuesta y controlar el efecto del resto. Se tiene, por tanto, una variable dependiente, llamada Y que puede ser dicotómica o politómica y una o más variables independientes llamadas X.

Al ser la variable Y dicotómica, podrá tomar el valor de 0 si el hecho no ocurre y 1 si el hecho ocurre; el asignar los valores de esta manera o a la inversa es intrascendente, pero es muy importante tener en cuenta esta asignación al momento de interpretar los resultados. En el caso de una Y politómica, podrá tomar valores desde 0 hasta n en donde cada uno representa una categoría. Las variables independientes (también llamadas explicativas) pueden ser de cualquier naturaleza: cualitativas o cuantitativas.

Capitulo 1

MEDIDAS DE ASOCIACIÓN

INTRODUCCIÓN A LA ESCALA CATEGÓRICA

Una variable categórica es aquella en la cual la escala de medida consiste de una colección de categorías. Por ejemplo, la filosofía política puede estar medida como "liberal", "moderada" o "conservadora"; la recuperación de una operación podría estar medida en tal caso como "completamente recuperado", "casi recuperado", "algo recuperado" y "no del todo recuperado".

Las escalas categóricas son comunes en las ciencias sociales y en las ciencias biomédicas, sin embargo no significa que estén restringidas a estas áreas. Esto tiene lugar frecuentemente en ciencias de la conducta, salud pública, ecología, educación, investigación de mercados y muchos otros casos más. Incluso hay campos cuantitativos tales como las ciencias de las ingenierías y el control de calidad industrial que pueden usarlas, como por ejemplo contestar o medir que tan suave al tacto es una cierta tela, que tan fácil un trabajador encuentra una tarea segura a hacer.

1.1 *CLASES DE VARIABLES*

1.1.1 Variables dependientes e independientes

La mayor parte del análisis estadístico distingue entre variables respuesta (o "dependientes") y variables explicativas (o "independientes"). Por ejemplo, los modelos de regresión describen la distribución de una respuesta continua, tal como el tiempo de sobrevivencia después de una operación de transplante de corazón, en este caso las variables explicativas podrían ser edad, presión arterial, etc. y la variable respuesta podría ser el tiempo de sobrevivencia.

1.1.2 Variables cuantitativas y cualitativas

Una variable es *cualitativa* cuando niveles distintos difieren en calidad y no en cantidad, es decir, las etiquetas de los niveles de la variable pueden ser usadas para identificar las distintas categorías de dicha variable, pero éstos no representan magnitudes entre los niveles, no se puede hablar de que un nivel es más grande que, grande que o más pequeña que otro nivel.

Una variable es *cuantitativa* cuando sus posibles niveles pueden ser comparados en magnitud, es decir, cuando a cada uno de ellos se les asigna un valor representativo. Cuando se tiene una variable de este tipo se puede hablar de que un nivel es más grande o más pequeño que otro.

1.1.3 Variables categóricas

La gente puede ser clasificada de acuerdo a la religión que profesa, (Católica, Judía, Protestante) o a su modo de transportación (automóvil, autobús, metro, bicicleta, otro), tipo de residencia (casa, departamento, condominio, otro), por su raza, por su sexo, por su estado civil o tipo de empleado (fijo, variable). En cada uno de estos ejemplos cada una de las características son colocadas en categorías.

Religión que profesa, modo de transportación, tipo de residencia, raza, sexo y estado civil son ejemplos de variables categóricas. Las categorías pueden ser de forma natural, como en el caso de sexo (Hombre, Mujer), o arbitrarias como en el caso tipo de empleado (fijo, variable).

Una variable categórica puede ser referida como una variable cualitativa.

1.1.4 Variables numéricas

Cuando la característica de interés puede ser expresada por un número y se hace la distinción entre un número que es obtenido simplemente de contar de uno que requiere una medida, se tiene la siguiente clasificación.

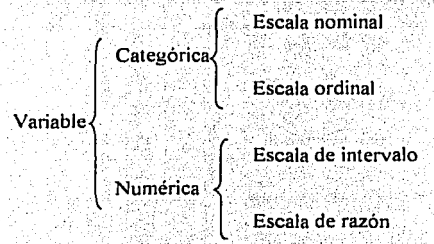
Variables discretas. El número de miembros en una familia, el número de autos que posee, el número de accidentes en el metro en diferentes horas, la población total en el Distrito Federal, son los resultados de contar; los ejemplos anteriores son ejemplos de variables discretas.

Una variable discreta es una variable en donde sus posibles valores son enteros o es un conjunto finito.

Variables continuas: Variables como estatura, peso o temperatura que pueden ser medidas con exactitud son denominadas variables continuas, dichas variables son intrínsicamente diferentes de una variable discreta.

1.2 ESCALAS DE MEDICIÓN

Las escalas de medición dependen del tipo de variable que se tenga, es decir, existen escalas para aquéllas que son categóricas y numéricas. El siguiente cuadro describe la clasificación.



Las variables categóricas pueden ser clasificadas como nominal u ordinal dependiendo de si existe o no un orden entre las categorías. En el caso de las variables numéricas estás pueden ser clasificadas como de intervalo o de razón dependiendo de si existe o no un cero absoluto.

1.2.1 Escalas para variables categóricas

1.2.1.1 Escala nominal

Variables cuyos niveles no tienen un orden natural, es decir, que son definidas en una *escala nominal* son aquéllas en donde las categorías o niveles son simplemente nombres, dichas variables también son llamadas variables categóricas. Ejemplos de variables nominales son las afiliaciones religiosas (Categorías: católica, judía, protestante, otra); modo de transportación (automóvil, autobús, metro, bicicleta, otro), tipo de residencia (casa, departamento, condominio, otro), raza, sexo, estado civil y tipo de empleado (fijo, variable). Para variables nominales, el orden de lista es irrelevante para el análisis estadístico.

1.2.1.2 Escala ordinal

Si las categorías de las variables pueden ser puestas en orden, la escala es llamada *escala ordinal.* Ejemplos de variables ordinales son el tamaño del automóvil (subcompacto, compacto, mediano, grande), clases sociales (alta, media, baja), actitud hacia la legalización del aborto (totalmente desaprobado, desaprobado, aprobado, totalmente aprobado), evaluación del nivel de inventario de una compañía (muy bajo, más o menos bien, muy alto) y diagnóstico de un paciente si tiene o no esclerosis múltiple (cierto, probable, poco probable, definitivamente no). Las variables ordinales claramente ordenan las categorías, pero las distancias absolutas entre categorías son desconocidas. Como se puede ver este tipo de clasificación implica un problema de calificación ya que por ejemplo una persona clasificada como

"moderada" es más liberal que una persona clasificada como "conservadora", pero no se puede dar un valor numérico para el que tanto más liberal sea esta persona con relación a la otra.

1.2.2 *Escalas para variables numéricas*

1.2.2.1 Escala de intervalo

Una variable de *intervalo* además de incorporar un orden, tiene la propiedad de que existe una distancia numérica entre cada par de niveles o categorías. Con este tipo de escala no solamente se puede comparar si es más grande o más pequeño una categoría que otra, sino que también se puede saber que tan grande o tan pequeño es. En la escala de intervalo, las distancia entre los niveles o categorías son iguales, además de que no posee un cero absoluto Por ejemplo, 0° Fahrenheit es un cero artificial ya que en ° Celsius representa 32 °.

1.2.2.2 Escala de razón

La *escala de razón* posee las características de las escalas anteriores, la diferencia que existe con la escala de intervalo, es que la de razón posee un cero absoluto, demás de que se puede hablar de proporciones entre categorías o niveles. Por ejemplo, $ 200 es 2 veces $ 100, en este caso $ 0 es ausencia de dinero.

En la jerarquía de medición, las variables de razón son las más altas, seguida de las variables de intervalo, variables ordinales, y las variables nominales son las más bajas. Métodos estadísticos diseñados para variables de un tipo también pueden ser usados con variables de niveles más altos, pero no en niveles más bajos. Por ejemplo, métodos estadísticos para variables ordinales pueden también ser usados con variables de intervalo (usando solamente el orden de los niveles y no sus distancias); ellos no pueden ser usados con variables nominales, ya que categorías de tales variables no tienen un orden significativo. Normalmente, lo mejor es aplicar métodos apropiados para la escala que se maneje.

1.3 CLASIFICACIÓN EN TABLAS

Una de las herramientas más usadas para el análisis bivariado o multivariado de datos ordinales y nominales es la clasificación en tabla. Dicha tabla puede ser construida cuando las observaciones en una muestra han sido clasificadas de acuerdo a sus valores en dos o más variables categóricas. La clasificación es usualmente presentada en una tabla de dos dimensiones, cada renglón y columna indica una categoría de cada variable. Cada combinación única de categorías de las variables es representada por una celda en la tabla.

Hay algunas convenciones en la presentación o descripción de esta herramienta. Primero, son a menudo identificadas por su tamaño, es decir, en términos del número de renglones (denotada por r), el cual es dado primero, seguida por el de columnas (denotada por c). El producto de éstos da el número de celdas en la tabla. Los totales de renglones y de columnas son llamados las frecuencias marginales o simplemente marginales. Éstas

representan la distribución de la frecuencia univariada en el renglón y en la columna respectivamente, también llamadas distribución marginal.

La clasificación en tablas es a menudo usada como un primer paso para detectar la relación entre variables. En el resto de este apartado se tratarán algunos procesos descriptivos para resumir las frecuencias en la clasificación en tablas como el siguiente paso en la investigación entre dos variables categóricas.

Si la tabla tiene I renglones y J columnas entonces se tiene una tabla de I × J, donde I y J toman algún valor entero.

La clasificación de más de dos variables es llamada tabla multidimensional. Por convención, la variable renglón (la cual es usualmente la dependiente) aparece primero, luego la columna (o independiente).

La figura 1 muestra una tabla general de I × J, en la cual la variable independiente es llamada X y la dependiente Y. Las etiquetas del renglón comienzan en 1 hasta I, las cuales indican las categorías de Y y, similarmente las etiquetas corren de 1 hasta J para las

TABLA 1

TABLA GENERAL CON I RENGLONES Y J COLUMNAS

X

(Variable Independiente)

Y (Variable Dependiente)	1	2	. . .	J	. . .	J	
1	n_{11}	n_{12}	. . .	N_{1j}	. . .	n_{1J}	n_{1+}
2	n_{21}	n_{22}	. . .	N_{2j}	. . .	n_{2J}	n_{2+}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
I	n_{i1}	n_{i2}	. . .	n_{ij}	. . .	n_{iJ}	n_{i+}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
I	n_{I1}	n_{I2}	. . .	n_{Ij}	. . .	n_{IJ}	n_{I+}
	n_{+1}	n_{+2}	. . .	N_{+j}	. . .	n_{+J}	N

Figura 1

Una combinación especifica de las variables renglón y columna son designadas por subíndices, la primera letra indica la categoria del renglón y la segunda la de la columna.

Las n_{ij} pequeñas representan las frecuencias de cada celda en una clasificación particular renglón-columna.

Los totales en el último renglón (n_{1+}, n_{2+} ...; n_{+1}, n_{+2}...; etc) constituye la distribución marginal de *X* y *Y*. En la tabla 1 (fig. 1), note que n_{1+} significa la suma de todas las observaciones en el primer renglón. El signo más en el subíndice indica que todas las entradas en el primer renglón han sido sumadas sobre las J columnas. Del mismo modo, n_{+1} es el total de todos los casos en la primer columna.

El número total de observaciones es N, esta cantidad es encontrada sumando todas las n_{ij} en la tabla o sumando los totales marginales de las variables en consideración.

Además de trabajar con tablas de frecuencias, es necesario o útil tener probabilidades. Sea $P(Y_i, X_j)$ que denota la probabilidad de que un individuo esté en la i-ésima clase de Y y la j-ésima clase de X en una tabla. $P(Y_1, X_1)$ en otras palabras, representa la probabilidad de tener la característica 1 en la variable Y y tener la característica 1 en la variable X. Para simplificar la notación, Pij frecuentemente reemplaza a $P(Y_i, X_j)$.

El diseño de tablas de probabilidad se sigue de la misma pauta de las de contingencia. P_{i+} por ejemplo, significa la probabilidad marginal de estar en la i-ésima categoría de Y Es obtenido de sumar las probabilidades en el i-ésimo renglón:

$$P_{i+} = P_{i1} + P_{i2} + \ldots + P_{iJ}$$

Asumiendo que la tabla tiene J columnas. Da la probabilidad de estar en la i-ésima categoría de Y, independientemente de X.

La probabilidad, de que un miembro de la población sea clasificado en la celda (i,j) es P_{ij} así que $E(n_{ij}) = E_{ij} = nP_{ij}$ representa el número de la muestra que puede caer dentro de cada celda. Ahora el conjunto $\{P_{ij}\}$ es hipotético para la población. Existe una forma estadística que permite decidir si las correspondientes frecuencias hipotéticas $\{nP_{ij}\}$ son consistentes con las frecuencias observadas $\{n_{ij}\}$. Dichaopción está basada en el estadístico de bondad de ajuste de la ji-cuadrada, el cual tiene la forma general

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - nP_{ij}\right)^2}{nP_{ij}} \qquad [1.1]$$

En el caso de que X y Y sean independientes, $P_{ij} = P_{i+} \cdot P_{+j}$ para cada i y j.

Sustituyendo el estimado $P_{ij} = P_{i+}P_{+j} = \frac{n_{i+}}{n} \cdot \frac{n_{+j}}{n}$ en la ecuación 1.1 se tiene

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - n \cdot \left(\frac{n_{i+} \cdot n_{+j}}{n^2}\right)\right)^2}{n \cdot \left(\frac{n_{i+} \cdot n_{+j}}{n^2}\right)} \qquad [1.2]$$

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(\frac{n \cdot n_{ij} - n_{i+} \cdot n_{+j}}{n}\right)^2}{\left(\frac{n_{i+} \cdot n_{+j}}{n}\right)}$$

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n \cdot n_{ij} - n_{i+} \cdot n_{+j}\right)^2}{n \cdot \left(n_{i+} \cdot n_{+j}\right)}$$

Si X y Y son independientes, la ecuación 1.2 tiene aproximadamente la distribución de una ji-cuadrada con (I-1)×(J-1) grados de libertad. El estadistico ji-cuadrada puede ser usado para establecer asociación, es decir, si χ^2 es más grande que un punto aceptable (nivel de significancia) de la distribución apropiada ji-cuadrada, entonces X y Y no son independientes.

1.4 COCIENTE DE MOMIOS

Una medida de asociación que puede ayudar a interpretar los modelos y describir la fuerza de asociación entre variables son los momios. Considérese una tabla de 2×2 teniendo frecuencias denotadas como sigue:

a	b
c	d

La probabilidad de estar clasificado en la columna 1 en lugar de la columna 2 es igual a a/b para el primer renglón, y c/d para el segundo renglón. La proporción de estas dos probabilidades es referida como el cociente de momios. Denotando por:

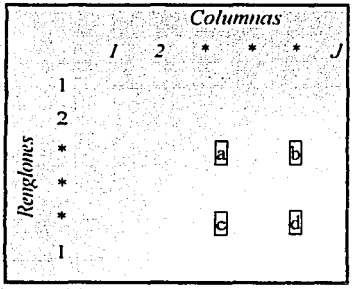
$$\theta = \frac{a/b}{c/d} = \frac{ad}{bc} \qquad [1.3]$$

Un nombre alternativo para esta medida es la proporción de producto cruzado. Cuando el momio es menor (o mayor) que 1.0, los miembros en el primer renglón son menos (o más) probables a ser clasificados en la columna 1 que los miembros en el segundo renglón. Independencia entre las variables columna y renglón corresponde a un valor poblacional de θ=1. Los momios se encuentran entre 0 e ∞, estos valores extremos ocurren cuando alguna de las entradas de las celdas es cero. El estimador muestral de $\theta = 0$ ó $\theta = \infty$ son no deseables cuando se cree que todas las probabilidades en las celdas de la población son cero. Un mejor estimador es obtenido mediante el cálculo de θ para una tabla ajustada en la cual .5 es agregado a cada una de las cantidades en cada celda.

Si los renglones (o columnas) son intercambiados, el cociente de momios correspondiente es el inverso (es decir, $1/\theta$) del cociente original Aquí para algún número positivo m, $\theta = m$ representa la misma fuerza de asociación que $\theta = 1/m$, pero en dirección opuesta.

Para este caso las frecuencias marginales son obtenidas cuando se suma el contenido de las celdas sobre los niveles de cada variable, es preferible ignorarla que controlarla en el análisis.

Para tablas I×J, los momios pueden ser calculados si se toman celdas que formen un rectángulo utilizando cada par de renglones en combinación con cada par de columnas. Por ejemplo viendo la siguiente tabla



En este caso se habla de independencia cuando todos los momios son iguales a 1 en la población.

1.5 MEDIDAS DE ASOCIACIÓN

1.5.1 Medidas de asociación para variables nominales

Mucho esfuerzo humano ha sido dedicado a descubrir relaciones importantes o asociaciones. Pocas veces es suficiente conocer únicamente algunas relaciones existentes, así que los investigadores tienden a cuantificar la asociación. La meta final puede ser establecer una relación de causa-efecto, pero la cuantificación es frecuentemente el primer paso hacia el logro de esta meta.

Las medidas de asociación para datos nominales no dependen en lo particular del orden en el cual las categorías son enlistadas, y todas las medidas discutidas en este apartado cumplen con esta propiedad. De las medidas más antiguas para datos nominales se encuentran las que están basadas en el estadístico de la ji-cuadrada.

1.5.1.1 Medidas basadas en el estadístico de la ji-cuadrada

El estadístico χ^2 para probar independencia se basa en un esquema de muestreo tal que, una muestra aleatoria independiente de tamaño N, es clasificada con respecto a dos características simultáneamente. Además, en la tabla de contingencia resultante, ambos conjuntos de frecuencias marginales totales son variables aleatorias, por lo que los supuestos de esta prueba son:

1. Los datos son independientes, de tal forma que el resultado que proporciona cada elemento de la muestra para ambas variables no influye en el dato de cualquier otro elemento.

2. Los totales de filas y de columnas, en la tabla de contingencia, son aleatorios. En otras palabras son contingentes por circunstancias que escapan del control del investigador.

1.5.1.2 *Tabla de contingencia con las frecuencias esperadas.*

Otra hipótesis que se suele probar es la de independencia entre dos variables de clasificación. En este procedimiento nuevamente se comparan las frecuencias observadas y las esperadas.

Para calcular la tabla con las frecuencias esperadas, considérese que, por ejemplo, X y Y son las variables categóricas para las filas y columnas, respectivamente, de la tabla de contingencia y sean:

P_{ij} la probabilidad, en la población, de que una observación pertenezca a la i-ésima categoría de la variable X y a la j-ésima categoría de la variable Y.

$P_{i\bullet}$ la probabilidad, en la población, de que una observación pertenezca a la i-ésima categoría de la variable X (sin hacer referencia a la variable Y).

$P_{\bullet j}$ la probabilidad, en la población, de que una observación pertenezca a la j-ésima categoría de la variable Y (sin hacer referencia a la variable X).

De la regla de la multiplicación de probabilidades de eventos independientes, se cumple que:

$$P_{ij} = P_{i\bullet} \cdot P_{\bullet j} \qquad [1.4]$$

Sea E_{ij} la frecuencia esperada, de elementos de la población, que pertenecen a la celda (i, j) de la tabla de contingencia donde se determinación se da considerando que si P_{ij} representa la probabilidad de que un miembro de la población en estudio pertenezca a la celda ij, entonces se esperaría que de los n individuos nP_{ij} estuvieran en la celda en cuestión por tal razón si existe independencia entre las variables categóricas, entonces:

$$E_{ij} = nP_{ij} = n \cdot \left(P_{i\bullet} \cdot P_{\bullet j}\right) = n\left(\frac{n_{i\bullet} \cdot n_{\bullet j}}{N^2}\right) \qquad [1.5]$$

donde N, $n_{i\bullet}$ y $n_{\bullet j}$, se refieren a los valores de tamaños de muestra poblacionales y marginales de toda la población.

Para calcular la tabla de contingencia con las frecuencias esperadas, de los datos de la muestra, los valores de las probabilidades poblacionales $P_{i\bullet}$ y $P_{\bullet j}$ son desconocidos, pero se pueden estimar de las frecuencias observadas.

Cuando dos variables son independientes, las frecuencias esperadas E_{ij} y las observadas O_{ij} deberán diferir únicamente por cantidades atribuibles a factores aleatorios. Si dos variables no son independientes, se esperarían grandes diferencias entre ambas frecuencias. En consecuencia, parece razonable basar la prueba de independencia en una estadística de prueba que considere las diferencias entre los conjuntos de frecuencias O_{ij} y E_{ij} para todas las celdas de las tablas.

1.5.1.3 Frecuencias esperadas pequeñas

La aproximación de χ_c^2 a una distribución χ^2 se cumple para valores grandes de n. Usualmente, se utiliza, como regla general empírica, que esta aproximación es satisfactoria cuando la frecuencia esperada de cada celda tiene un valor de al menos 5. Cochran, en 1954, enfatizó que esta regla es muy estricta y sugirió que si, relativamente pocos valores esperados son menores que 5 (por decir 1 celda de 5), entonces la aproximación de χ_c^2 a una distribución χ^2 es válida. Posteriormente en 1965, Lewontin y Felsenstein demostraron que para tablas de 2×J (I = 2), se puede utilizar la aproximación χ_c^2 si todos los valores esperados son de 1 o más.

Para poder usar la aproximación de χ_c^2 por una distribución χ^2, en ocasiones se ha resuelto el problema de las frecuencias esperadas pequeñas, mediante la combinación de categorías. Sin embargo, al hacer esto, generalmente se pierde información y cambia el objetivo inicial de la aplicación de esta prueba. Por otro lado, es factible que la aleatoriedad de la muestra se vea afectada, ya que las observaciones pueden caer en categorías que se están eligiendo de antemano. En consecuencia, es preferible evitar las combinaciones de categorías.

Cabe hacer énfasis en que esta prueba ji-cuadrada de independencia sólo se puede utilizar para establecer únicamente la significancia de la asociación entre las variables categóricas.

En los apartados anteriores se estudió la prueba de independencia χ^2 bajo un esquema de muestreo en el que las unidades observadas aleatoriamente son

clasificadas, simultáneamente, con respecto a dos características. En muchas ocasiones, además de cuestionarse la significancia de la prueba ji-cuadrada, se estará interesado en determinar el grado de asociación entre las dos variables cualitativas consideradas en la tabla de contingencia.

Algunas de las medidas de asociación que se han sugerido, se basan en el estadístico de prueba χ^2, debido a que un valor de χ^2 cercano a cero, indica independencia (no asociación) entre las variables. Pero dado que la magnitud de χ^2 depende del tamaño de muestra n y no puede ser comparable para diferentes tablas, estas medidas utilizan alguna transformación de χ^2.

Cabe mencionar que, en la literatura, se han sugerido muchas medidas de asociación para tablas de contingencia de las cuales no hay una considerada como la mejor. Realmente, la elección de una de éstas depende de las preferencias personales.

1.5.1.4 *Estadístico de prueba para variables nominales*

Para probar la independencia entre dos variables se utiliza un estadístico χ^2-cuadrada, el cual fue sugerido por Pearson en 1904. Este estadístico consiste en comparar las frecuencias observadas con las esperadas y está dada por:

$$\chi_c^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad [1.6]$$

La magnitud del estadistico de prueba χ_c^2 depende de los valores de las diferencias $(O_{ij} - E_{ij})$. Cuando las dos variables son independientes, estas diferencias son menores que cuando no son independientes y en consecuencia, χ_c^2 toma un valor más pequeño cuando Ho es cierta que cuando es falsa.

Bajo el supuesto de que la hipótesis nula de independencia es cierta, la estadistica de prueba χ_c^2 sigue aproximadamente una distribución χ^2 con (I-1)(J-1) grados de libertad. Por lo tanto, para rechazar o aceptar Ho, hay que basarse en la probabilidad del valor obtenido de χ_c^2. La aceptación de la hipótesis nula se da si el valor de χ_c^2 resulta ser menor o igual que le percentil χ_{1-p}^2 de una distribución χ^2 con (I-1) (J-1) grados de libertad, esto es

$$p = P\left(\chi_c^2 \leq \chi_{1-p}^2\right) = 1 - p.$$

En caso contrario se rechazará la hipótesis nula

1.5.1.4.1 Coeficiente de contingencia cuadrado medio

La forma más simple de transformar el valor de χ^2, propuesto por Pearson, consiste en dividir a χ^2 por n; la cual es denotada por ϕ^2:

$$\phi^2 = \frac{\chi^2}{n} \qquad [1.7]$$

El valor de la estadística de prueba χ^2 es igual a cero cuando hay independencia entre las dos variables, por lo que ϕ^2 tomará su menor valor cuando éstas no estén asociadas. Sin embargo, ϕ^2 no es una medida de asociación satisfactoria, puesto que puede tomar valores mayores a 1 y no hay límite superior para ϕ^2.

1.5.1.4.2 Coeficiente de contingencia

Pearson, también sugirió una variación de ϕ^2, denominado coeficiente de contingencia, el cual está dado por:

$$P = \sqrt{\frac{\phi^2}{1+\phi^2}} = \sqrt{\frac{\chi^2}{n+\chi^2}} \qquad [1.8]$$

En general, el máximo valor de χ^2 es n(q-1), donde q está dada por:

$$q = \min(r,c)$$

De aquí que, el máximo valor de P sea:

$$P = \sqrt{\frac{n(q-1)}{n+n(q-1)}} = \sqrt{\frac{q-1}{q}} \qquad [1.9]$$

el cual, en muchos casos es cercano a 1. Como el mínimo valor de χ^2 es 0, entonces P se encuentra entre los valores:

$$0 \leq P \leq \sqrt{\frac{q-1}{q}} < 1 \qquad [1.10]$$

De esta expresión se puede notar que no hay un límite superior para el coeficiente de contingencia P, por lo que Kendall y Stuart sugirieron el siguiente coeficiente de contingencia.

1.5.1.4.3 Coeficiente de contingencia de Kendall y Stuart

$$T = \frac{\chi^2}{n\sqrt{(r-1)(c-1)}}. \qquad [1.11]$$

Cuando hay independencia entre las dos variables, T toma el valor de 0. Cuando existe una asociación perfecta entre éstas, T es igual a 1, siempre y cuando el número de renglones sea igual al de las columnas. Si r ≠ c, esto último no se cumple.

1.5.1.4.4 Coeficiente de contingencia de Cramér

Cramér sugirió una modificación a T, dada por el siguiente coeficiente de contingencia: $C = \frac{\chi^2}{n[\min(r-1, c-1)]}$ el cual, en caso de completa asociación, toma el valor de uno para cualesquiera valores de r y c.

1.5.2 Medidas de asociación para variables ordinales

Varias de las medidas de asociación que han sido propuestas son: la gamma, la tau-b y tau-c de Kendall, rho-b y rho-c de Spearman, y la d de Sommers. Todas estas medidas son completamente similares en sus propósitos y características. Algunas de éstas son:

1. Medidas de asociación ordinal toman valores entre -1 y 1. Cuando las variables están en el nivel ordinal más bajo, es posible distinguir entre dos tipos de asociación positiva y negativa. Es una asociación positiva entre variables X y Y cuando al aumentar X aumenta Y, en cambio es una asociación negativa cuando al aumentar X disminuye Y.

2. Los valores de la población de medidas ordinales son iguales a 0 cuando las variables son estocásticamente independientes. Sin embargo si la medida ordinal es igual a cero no necesariamente las variables son estocásticamente independientes.

3. A excepción de la d de Sommers las medidas ordinales mencionadas anteriormente son simétricas.

1.5.2.1 Parejas concordantes y parejas discordantes

La mayoría de las medidas ordinales de asociación están basadas en la información contenida en la clasificación relativa para todas las parejas de observaciones.

Dos parejas (x_i, y_i) y (x_j, y_j) son concordantes si:

$$x_i < x_j, \text{ y } y_i < y_j$$
$$\text{o } x_i > x_j, \text{ y } y_i > y_j$$

Se dice discordantes si:

$$x_i < x_j, \text{ y } y_i > y_j$$
$$\text{o } x_i > x_j, \text{ y } y_i < y_j$$

Ocurre un empate si $x_i = x_j$, y/o $y_i = y_j$.

Sea C que denota el número de parejas concordantes de observaciones y sea D que denota el número total de parejas discordantes de observaciones.

1.5.2.2 Gamma

Todas las medidas de asociación que se tratarán en este apartado se basan de alguna manera en la diferencia C – D. Una diferencia positiva para C- D indica una asociación positiva, en el sentido de que hay más parejas concordantes que discordantes. Una diferencia negativa refleja una asociación

negativa. Tamaño de muestras grandes llevaría a números grandes de parejas con diferencias absolutas grandes en C – D, por lo tanto, es necesario estándarizar esta diferencia, así que es más fácil interpretar que tan fuerte es la asociación. Para poder estándarizar lo que hay que hacer es dividir C – D por su valor máximo posible, C + D, el cual es el número total de parejas concordantes y discordantes. Esto da la medida de asociación referida como ***gamma***, introducido por Goodman y Kruskal (1954). Esta fórmula muestral es:

$$\gamma = \frac{C-D}{C+D} \qquad [1.12]$$

El valor de gamma se encuentra entre –1 y 1, una asociación positiva fuerte (un valor de 1) corresponde cuando no hay parejas discordantes (D = 0). Gamma es igual a –1 cuando C = 0, y es igual a cero cuando C = D.

El signo indica si hay una asociación positiva o negativa entre las variables. La magnitud indica la fuerza de la asociación. Esta última puede ser interpretada haciendo uso de la expresión [1.13]:

$$\gamma = \frac{C-D}{C+D} = \frac{C}{C+D} - \frac{D}{C+D}. \qquad [1.13]$$

Donde $\frac{C}{C+D}$ es la proporción de los C + D pares que son concordantes, $\frac{D}{C+D}$ es la de los discordantes, y γ es la diferencia entre estas dos proporciones.

Para una tabla de 2x2, la medida se convierte en un caso especial la cual fue propuesta por el estadistico G. Udny Yule, en este caso se le llama a la gamma como la Q de Yule.

1.5.2.3 *Tau-b de Kendall*

En la clasificación cruzada de variables ordinales, no todos los pares de variables son concordantes o discordantes. Individuos colocados en la misma categoria se dice que están empatados en la misma variable. Dos observaciones que caen en el mismo renglón de una tabla están empatadas con respecto a la variable renglón, y dos observaciones que caen en la misma columna de una tabla están empatas con respecto a la variable columna.

Para una muestra de tamaño n, hay $\frac{n(n-1)}{2}$ pares de observaciones.

Para dos variables X y Y, sea T_x que denota el número de pares de observaciones empatadas , sea T_y que denota el numero de parejas empatadas en Y, sea T_{xy} que denota el número de pares empatadas en X y Y.

Ahora, $T_x + T_y - T_{xy}$ es el número total de pares de observaciones empatadas. La razón para restar T_{xy} en esta expresión es porque T_{xy} representa el número de pares que estan empatadas en ambas variables y aquí se están contando dos veces en la suma $T_x + T_y$. El número total de pares también se puede expresar como:

$\frac{n(n-1)}{2}$ = (Número de pares no empatados) + (Número de pares empatados)

$$= C + D + T_x + T_y - T_{xy} \qquad [1.14]$$

También se puede observar que $\frac{n \cdot (n-1)}{2} - T_x$ es el número total de pares menos el número total de pares empatados en X, el cual es el número de pares que no están empatados. Similarmente, el número de pares que no están empatadas en Y es $\frac{n(n-1)}{2} - T_y$.

La Tau-b de Kendall es una medida de asociación entre dos variables ordinales que toman en cuenta el número de pares empatados en cada variable. Este valor muestral está dado por la fórmula:

$$t_b = \frac{C - D}{\sqrt{\left[\frac{n(n-1)}{2} - T_X\right]\left[\frac{n(n-1)}{2} - T_Y\right]}} \qquad [1.15]$$

La magnitud t_b puede ser interpretada de muchas formas al igual que la gamma. La t_b toma valores entre –1 y 1.

Para completar el cálculo de t_b se necesita calcular T_X y T_Y. Se denota por t_i al número de observaciones empatadas en X en el i-ésimo nivel de las I columnas de una tabla de clasificación cruzada, por lo que se tienen $\frac{t_i \cdot (t_i - 1)}{2}$ pares empatados y al sumarlos se obtiene la siguiente expresión:

$$T_x = \sum_{i=1}^{c} \frac{t_i \cdot (t_i - 1)}{2} \qquad [1.16]$$

Para el caso de los renglones, sea r_j las observaciones en el j-ésimo renglón de una tabla de clasificación, y el total de empates por renglones será:

$$T_y = \sum_{j=1}^{c} \frac{r_j \cdot (r_j - 1)}{2} \qquad [1.17]$$

g

Capítulo 2

MODELO DE REGRESIÓN LOGÍSTICA

INTRODUCCIÓN

Variables que son registradas en una escala continua pueden ser modelados utilizando regresión lineal (múltiple). Algunas variables que se desea modelar son, sin embargo, discontinuas o dicotómicas y necesitan ser analizadas usando una técnica alternativa. La técnica que se considerará en este apartado permite a este tipo de datos que se encuentran restringidos entre dos valores, ser modelados. La forma de este tipo de datos es tan común y pueden consistir de clasificaciones dicotómicas tales como "vivo - muerto", "empleado - desempleado", "éxito - fracaso", o medidas de proporción restringidas entre 0 y 1; por ejemplo, proporción de asesinatos, proporción de preguntas correctamente contestadas.

La regresión logística es considerada como el modelo más importante de los modelos lineales generalizados (MLG) para variables de respuestas binarias.

2.1 MODELOS LINEALES GENERALIZADOS

Los modelos lineales generalizados son especificados por tres componentes, el aleatorio, el sistemático y una función liga. La ecuación de un modelo lineal generalizado está dado por:

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)]$$

El componente aleatorio de un MLG consiste de observaciones $Y = (y_1, \ldots, y_n)$ de una distribución en la familia exponencial. Esto es, cada observación y_i tiene la función de densidad de probabilidad o función generalizada de la forma:

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)]$$

El valor del parámetro θ_i puede variar para $i = 1, \ldots, n$ dependiendo de los valores de las variables explicativas. El término $Q(\theta_i)$ es llamado el parámetro natural de la distribución o componente aleatorio.

El componente sistemático de un MLG relaciona un vector $\gamma = (\gamma_1, \ldots, \gamma_n)'$ con el conjunto de variables explicativas a través de un modelo lineal.

$$\begin{aligned}
\gamma_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} \\
\gamma_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} \\
&\vdots \\
\gamma_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk}
\end{aligned}$$

El cual se puede expresar en forma matricial:

$$\gamma = X\beta \quad \text{donde}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots x_{1k} \\ 1 & x_{21} & x_{22} & \cdots x_{2k} \\ \vdots & \vdots & \vdots & \cdots \vdots \\ 1 & x_{n1} & x_{n1} & \cdots x_{nk} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}.$$

El vector γ es llamado el predictor lineal.

El tercer componente de un MLG es la función liga la cual representa la relación entre el componente aleatorio y el sistemático.

Sea $\mu = E[y_i]$ $i = 1,..., n$. Entonces μ está ligada a γ_i por $\gamma_i = g(\mu)$ donde g es una función monótona y diferenciable.

$$g(\mu_i) = \sum_{j=0}^{k} \beta_j x_{ij} \qquad i = 1, \ldots, n.$$

La función liga que transforma la media de y_i al parámetro natural se llama liga canónica.

$$g(\mu_i) = Q(\theta_i) = \sum_{j=0}^{k} \beta_j x_{ij}$$

En resumen, un MLG es un modelo lineal para la transformación de la media de una variable Y que pertenece a la familia exponencial.

2.2 MODELO LOGIT.

Algunas variables categóricas tienen solamente dos categorías. La observación para cada sujeto puede ser clasificada como un "éxito" o "fracaso". Representando estos posibles resultados por 1 ó 0. La distribución Bernoulli para variables aleatorias binarias especifica probabilidades $P[Y=1]=\pi$ y $P[Y=0]=1-\pi$ para los dos resultados, para el cual $E[Y]=\pi$.

Cuando y_i tiene una distribución Bernoulli con parámetro π_i, la función generalizada de probabilidad es:

$$f(y_i;\pi_i)=\pi_i^{y_i}(1-\pi_i)^{1-y_i}=(1-\pi_i)\left[\frac{\pi_i}{1-\pi_i}\right]^{y_i}$$

$$=(1-\pi_i)\exp\left[y_i\log\left(\frac{\pi_i}{1-\pi_i}\right)\right]$$

para $y_i = 0$ y 1. Esta distribución está en la familia exponencial. El parámetro lineal $Q(\pi)=\log\left(\frac{\pi_i}{1-\pi_i}\right)$ es el logaritmo de la proporción de respuesta 1, el cual es llamado el logit de π_i. Los MLG que utilizan la liga logit son llamados modelos logit.

2.3 AJUSTE DEL MODELO

2.3.1 Modelo de Probabilidad Lineal.

Para una variable de respuesta binaria el modelo de regresión

$$E[Y] = \pi(x) = \alpha + \beta \cdot x \qquad [2.1]$$

es llamado el modelo de probabilidad lineal. Cuando las observaciones en Y son independientes, este modelo es un MLG con la función liga identidad.

Pero este modelo presenta inconvenientes como:

- Las probabilidades deben estar entre 0 y 1, mientras que este modelo toma valores sobre la recta real.
- El modelo predice $\pi(x) < 0$ y $\pi(x) > 1$ para valores suficientemente pequeños o grandes para x.
- Generalmente se espera una relación no lineal entre x y $\pi(x)$, es decir, un cambio en x puede tener menos impacto cuando $\pi(x)$ es cercana a cero o a uno que cuando $\pi(x)$ se encuentra cerca de su valor intermedio.

El modelo [2.1] puede ser válido sobre un rango finito de valores de x y hay problemas para ajustar el modelo si se usa el método de mínimos cuadrados ya que las condiciones que hace que los estimadores sean óptimos no son satisfechas, así que es más provechoso estudiar modelos que implican una relación no lineal entre x y $\pi(x)$. Cuando se espera una relación

monótona, la curva apropiada es la que tiene la forma de *S*. Una función que tiene esta propiedad es:

$$\text{probabilidad de que un evento ocurra} = \pi(x) = \frac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}} \qquad [2.2]$$

llamada la *función de regresión logística*.

Donde:

e es la base del logaritmo natural,
α y β son parámetros del componente lineal del modelo
x es el valor de la variable explicativa.

Cuando $x \to \infty$, $\pi(x) \to 0$ cuando $\beta < 0$ y $\pi(x) \to 1$ cuando $\beta > 0$. Cuando $\beta \to 0$ la curva tiende a una línea horizontal. Cuando en el modelo $\beta < 0$, la respuesta binaria es independiente de x.

La curva de regresión logística tiene como primera derivada:

$$\frac{\partial \pi(x)}{\partial x} = \frac{(1+e^{\alpha+\beta x})\beta e^{\alpha+\beta x} - e^{\alpha+\beta x}\beta e^{\alpha+\beta x}}{\left(1+e^{\alpha+\beta x}\right)^2}$$

$$= \frac{\beta e^{\alpha+\beta x}(1+e^{\alpha+\beta x}-e^{\alpha+\beta x})}{\left(1+e^{\alpha+\beta x}\right)^2} = \frac{\beta e^{\alpha+\beta x}}{\left(1+e^{\alpha+\beta x}\right)^2}$$

$$= \frac{\beta}{\left(1+e^{\alpha+\beta x}\right)}\pi(x) = \beta\pi(x)\left(1-\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right) = \beta\pi(x)[1-\pi(x)]$$

Obteniendo la segunda derivada de la curva de regresión logística se tiene:

$$\frac{\partial^2 \pi(x)}{\partial x} = \beta \cdot \pi(x) \cdot \left(-\pi'(x)\right) + \beta \cdot \pi'(x) \cdot (1 - \pi(x))$$

$$\frac{\partial^2 \pi(x)}{\partial x} = \pi'(x)\{-\beta \cdot \pi(x) + \beta \cdot [1 - \pi(x)]\}$$

$$\frac{\partial^2 \pi(x)}{\partial x} = \pi'(x)\{\beta - 2 \cdot \beta \cdot \pi(x)\}$$

Sustituyendo a $\pi(x)$ y $\pi'(x)$ e igualando a cero la segunda derivada se tiene:

$$\frac{\beta \cdot \exp(\alpha + \beta \cdot x)}{[1 + \exp(\alpha + \beta \cdot x)]^2} \cdot \left\{\beta - 2 \cdot \beta \cdot \frac{\exp(\alpha + \beta \cdot x)}{1 + \exp(\alpha + \beta \cdot x}\right\} = 0$$

$$\frac{\beta^2 \cdot \exp(\alpha + \beta \cdot x)}{[1 + \exp(\alpha + \beta \cdot x)]^2} = 2 \cdot \beta^2 \cdot \frac{[\exp(\alpha + \beta \cdot x)]^2}{[1 + \exp(\alpha + \beta \cdot x]^2}$$

$$\frac{2 \cdot \exp(\alpha + \beta \cdot x)}{1 + \exp(\alpha + \beta \cdot x)} = 1 \Rightarrow \frac{\exp(\alpha + \beta \cdot x)}{1 + \exp(\alpha + \beta \cdot x)} = \frac{1}{2}$$

De la expresión anterior se puede ver que $\exp(\alpha + \beta \cdot x) = 1$ despejando a x se obtiene que $x = \frac{-\alpha}{\beta}$.

Por lo que la función se maximiza cuando $\pi(x) = \frac{1}{2}$ y $x = \frac{-\alpha}{\beta}$.

Se necesita primero transformar la medida de probabilidad en una medida de proporción, la cual está definida como la probabilidad de que un evento pase dividida por la probabilidad de que ese evento no pase. Dicha proporción puede ser representada como:

$$\text{proporción de que un evento pase} = \frac{\pi(x)}{1-\pi(x)} \qquad [2.3]$$

De la ecuación [2.2] la proporción puede ser derivada como:

$$\text{proporción de que un evento pase} = e^{(\alpha+\beta x)} = e^{\alpha}\,(e^{\beta})^{x}.$$

Esta fórmula da una interpretación básica para β. La proporción se incrementa cuando se multiplica por e^{β} para cada unidad incrementada en x. Si se toma el logaritmo a la proporción se tiene la siguiente relación lineal

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \qquad [2.4]$$

La función liga apropiada es el log(proporciones) la cual es comúnmente conocida como el *logit*.

Las ecuaciones [2.1] a [2.4] demuestran como una relación no lineal entre una variable de respuesta binaria y una variable explicativa continua puede ser representada linealmente cuando la variable respuesta es descrita en términos del logaritmo de las proporciones de probabilidad de un resultado particular.

Usando logit($\pi(x)$) en lugar de $\pi(x)$ permite a los parámetros ser estimados usando análisis de regresión.

El método para estimar los parámetros en la regresión logística es diferente al utilizado en la regresión lineal (múltiple), la cual utiliza el método de mínimos cuadrados. En la regresión logística, una forma para estimar los parámetros es a través de los estimadores de máxima verosimilitud.

2.3.2 Método de Máxima Verosimilitud

El método más comúnmente usado para estimar los parámetros de un modelo de regresión logística es el método de máxima verosimilitud. La función de probabilidad es, en general, definida como la función de probabilidad conjunta de las variables aleatorias.

Para una muestra de tamaño n cuyas observaciones son (y_1, y_2, ... , y_n) las variables aleatorias correspondientes son (Y_1, Y_2, ..., Y_n).

Sea Y_i la i-ésima variable independiente, su función de densidad conjunta está dada por:

$$L = g(Y_1, Y_2, \ldots Y_n) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i} \qquad [2.5]$$

En otras palabras la ecuación [2.1] da la probabilidad de una secuencia particular de ceros y unos.

Los estimadores de máxima verosimilitud son generalmente obtenidos de maximizar el logaritmo de la función de probabilidad conjunta, se realiza de este modo, ya que es más sencillo que si se tomará solamente la función de probabilidad, además de que la función logaritmo es una función monótona creciente. Aplicando el logaritmo a la ecuación [2.5] se tiene:

$$\log(L) = \log\left[\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}\right]$$

$$= \sum_{i=1}^{n} y_i \cdot \log(\pi_i) + \sum_{i=1}^{n} (1-y_i) \cdot \log(1-\pi_i)$$

$$= \sum_{i=1}^{n} y_i \cdot \log\left(\frac{\pi_i}{1-\pi}\right) + \sum_{i=1}^{n} \log(1-\pi_i)$$

$$= \sum_{i=1}^{n} y_i \cdot (\beta_0 + \beta_1 \cdot x_i) + \sum_{i=1}^{n} \log[1+\exp(\beta_0 + \beta_1 \cdot x_i)]$$

[2.6]

Derivando la ecuación [2.6] con respecto a β_0 y β_1, se obtiene lo siguiente:

$$\frac{\partial \log[L(\beta_0, \beta_1]}{\partial \beta_0} = \sum y_i - \sum \frac{\exp(\beta_0 + \beta_1 \cdot x_i)}{1+\exp(\beta_0 + \beta_1 \cdot x_i)}$$

$$\frac{\partial \log[L(\beta_0, \beta_1)]}{\partial \beta_1} = \sum y_i \cdot x_i - \sum \frac{x_i \cdot \exp(\beta_0 + \beta_1 \cdot x_i)}{1+\exp(\beta_0 + \beta_1 \cdot x_i)} \qquad [2.7]$$

Los estimadores de máxima verosimilitud para β_0 y β_1 se obtienen igualando a cero las ecuaciones [2.7] y resolviéndolas simultáneamente e iterativamente.

Por consecuencia se escriben las ecuaciones en forma matricial, lo cual ayuda en la transición a la regresión logística múltiple. Sea X una matriz de n×2 con cada renglón dado por $(1, x_i)$, Y como el vector respuesta y a π que denota la E[y], la ecuación de probabilidad puede ser escrita como:

$$\frac{\partial \ell(\beta)}{\partial \beta} = X'(Y - \pi) \qquad [2.8]$$

donde $\ell(\beta) = \log[L(\beta_0, \beta_1)]$. De [2.8] se sigue que:

$$X'\pi = X'Y \qquad [2.9]$$

$\frac{\partial \ell(\beta)}{\partial \beta} = 0$ Si $\pi = Y$, es decir, si el valor esperado de y_i es igual a la probabilidad estimada para Y=1.

La solución de la ecuación [2.9] satisface la siguiente condición

$$X'(Y - \hat{Y}) = 0, \qquad [2.10]$$

lo anterior también es tomado en el caso de la regresión lineal simple y múltiple.

La ecuación [2.10] es resuelta generalmente usando el método de Newton – Raphson. Esto es, primero se determina $\frac{\partial X' \cdot (Y - \pi)}{\partial \beta}$ la cual es equivalente a calcular $-\frac{\partial X' \cdot \pi}{\partial \beta} = -\left(\frac{\partial \pi}{\partial \beta}\right) \cdot X$.

De la ecuación [2.2] se puede obtener:

$$\frac{\partial \pi}{\partial \beta_0} = \pi \cdot (1 - \pi) \quad \text{y} \quad \frac{\partial \pi}{\partial \beta_1} = X \cdot \pi \cdot (1 - \pi)$$

Por lo que:

$$-\left(\frac{\partial \pi}{\partial \beta}\right) \cdot X = X' \cdot W \quad \text{asi} \quad -\frac{\partial X' \cdot \pi}{\partial \beta} = X' \cdot W \cdot X$$

donde W es una matriz diagonal con elementos $\pi_i(1 - \pi_i)$ que tendrian que ser estimados y X es previamente definida.

Los estimados de β son obtenidos como :

$$\beta = (X' \cdot W \cdot X)^{-1} \cdot X' \cdot W \cdot Z \qquad [2.11]$$

donde Z juega el papel de Y.

Específicamente se tiene $z_i = \hat{\eta}_i + \frac{y_i - \hat{\pi}_i}{\hat{\pi}_i \cdot (1-\hat{\pi}_i)}$ donde $\hat{\eta}_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$.

Note que $\hat{\eta}_i$ juega el papel de $\hat{y}_i$ y $\frac{(y_i - \hat{\pi}_i)}{\hat{\pi}_i \cdot (1-\hat{\pi}_i)}$ es el correspondiente residual de y_i dividido por la varianza estimada de y_i.

La ecuación [2.11] puede escribirse en términos de $\hat{\beta}$, como:

$$Z = X \cdot \beta^{(i)} + W^{-1} \cdot e \quad \text{con} \quad e = Y - \hat{\pi}$$

Se puede obtener:

$$\beta^{(i+1)} = (X'WX)^{-1} \cdot X'W \cdot \left(X\beta^{(i)} + W^{-1} \cdot e\right)$$

$$= \beta^{(i)} + (X'WX)^{-1} X'e \qquad [2.12]$$

La fórmula dada por la ecuación anterior es usada hasta que los estimadores convergen. El primer paso es obtener el estimador inicial, $\hat{\beta}^{(0)}$. Existen varios métodos o aproximaciones para obtenerlo, incluyendo el uso del análisis discriminante. El análisis es usado frecuentemente cuando la finalidad del estudio es la clasificación. El análisis discriminante no es muy usado debido a que es totalmente sensible a la normalidad de los regresores. Solamente si se conoce que los regresores tiene una distribución normal (un escenario poco probable) podría usarse el análisis discriminante. Los estimadores iniciales son obtenidos usando la siguiente expresión:

$$\hat{\beta}^{(0)} = \begin{bmatrix} \hat{\beta}_0^{(0)} \\ \hat{\beta}_1^{(0)} \end{bmatrix} = \begin{bmatrix} \ln\left(\frac{\theta_1}{\theta_0}\right) - 0.5 \cdot \frac{(\mu_1^2 - \mu_0^2)}{\hat{\sigma}^2} \\ \frac{(\mu_1^2 - \mu_0^2)}{\hat{\sigma}^2} \end{bmatrix} \qquad [2.13]$$

Los estimadores de la ecuación anterior que son usados en el calculo $\hat{\beta}_0^0$ y $\hat{\beta}_1^0$ pueden ser explicados como sigue, la distribución condicional de X dado Y=0 y X dado Y=1 son asumidos normales con medias μ_0 y μ_1 respectivamente. Lógicamente los estimadores de estos dos parámetros son $\mu_0 = X_0$ y $\mu_1 = X_1$, donde X_0 y X_1 son los promedios de los valores de x cuando Y = 0 y Y = 1, respectivamente. Los estimadores θ_0 y θ_1 estiman P(Y = 0) y P(Y = 1), respectivamente, y son definidas como $\theta_1 = \bar{Y}$ y $\theta_0 = 1 - \theta_1$. Estos, θ_0 y θ_1 son los porcentajes de unos y ceros, respectivamente, en el conjunto de datos.

El estimador de σ^2 es obtenido como:

$$\hat{\sigma}^2 = \frac{\left[(n_0 - 1) \cdot s_0^2 + (n_1 - 1) \cdot s_1^2\right]}{n_0 + n_1 - 2}$$

donde s_0^2 y s_1^2 son las varianzas muestrales calculadas usando Y=1 y Y=0, respectivamente, y n_0 y n_1 son los tamaños muestrales correspondientes.

2.3.3 *Intervalos de Confianza*

En esta sección se presentarán los intervalos de confianza para β_1, para el cambio en los momios y para π_j. El cálculo de los intervalos será basado en una aproximación al método de máxima verosimilitud.

2.2.3.1 *Intervalo de confianza para β_1*

Un intervalo de confianza para β_1 es un intervalo de confianza para el cambio en el logaritmo de las proporciones.

El intervalo para β_1 es obtenido como:

$$\beta \pm z_{\alpha/2} s_{\beta_1}$$

donde $z_{\alpha/2}$ denota la desviación estándar de una normal con un área de $\alpha/2$. En este caso z se usa en lugar de t, la cual es usada para intervalos de confianza en regresión lineal. Esto es porque no hay normalidad en regresión logística, como es asumida en la regresión lineal.

Un intervalo para β_1 deberia ser usado cuidadosamente, especialmente si el tamaño de la muestra no es grande.

2.2.3.2 Intervalo de confianza para el cambio en momios

Un intervalo de confianza para el cambio en las proporciones es similar a un intervalo de confianza para el cambio en los momios. Por ejemplo para un sólo regresor, $\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 x$, se tiene $\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 x)$. Esto es, $\exp(\beta_1)$ representan el cambio en los momios por unidad de cambio en X.

Un intervalo de confianza para $\exp(\beta_1)$ es un intervalo de confianza para el cambio en los momios. Una forma para obtener dicho intervalo es:

$$\exp\left(c\hat{\beta}_1 \pm z_{\alpha/2}\left(cs_{\beta_1}\right)\right)$$

donde c representa el incremento en X para el intervalo deseado. Se debe notar que el intervalo no es una función de X, lo cual parecería ilógico. Esto es, porque el resultado es basado en el hecho de que el incremento en el error de Y es independiente al nivel de X.

2.2.3.3 Intervalo de confianza para π

Aunque un intervalo de confianza para el cambio en los momios es obviamente importante, también lo es conocer si el riesgo está siendo incrementado. Especificamente, si se dice que el riesgo incrementa n veces por cada m unidades incrementadas en X, se necesita conocer cual es el riesgo es un valor de X dado.

Se necesita conocer $\hat{\pi}_i$, y se necesita como π_i es estimada. Por lo que, se necesita un intervalo de confianza para π_i . La aproximación más obvia a usarse debería ser $\hat{\pi} \pm zs_{\hat{\pi}}$, pero $0 \le \hat{\pi} \le 1$, por lo que $\hat{\pi}$ no debería ser aproximada a una distribución normal para muchos valores de π. Este método para obtener el intervalo de confianza no es el más apropiado. Aunque si se tiene una muestra grande este método es de gran utilidad. Asúmase que existe un solo regresor. Como $\log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 * x$, un intervalo de confianza para $\log\left[\frac{\pi}{1-\pi}\right]$ debería estar dado por $\beta_0 + \beta_1 * x \pm zs_{\beta_0+\beta_1*x}$. Al tomar el exponente se tiene:

Limite inferior $$\frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 * x - zs\sqrt{h_{ij}}\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 * x - zs\sqrt{h_{ij}}\right)}$$

Limite superior $$\frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 * x + zs\sqrt{h_{ij}}\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 * x + zs\sqrt{h_{ij}}\right)}$$

Donde $Var(\hat{Y}|X_i) = \sigma^2 h_{ij}$, donde h_{ij} es definida previamente.

2.2.3.4 Intervalo de confianza para una y fija

Estos intervalos son calculados para el predictor lineal, logit(π), y son transformados en un intervalo de confianza para π.

Intervalo de confianza para logit(π) = $\text{logit}(\hat{\pi}) \pm 1.96 * ASE$

donde ASE es el error estándar del logit $(\hat{\pi})$ y el valor 1.96 es la aproximación normal de la distribución de la muestra al estadístico t con un nivel del 95% de confianza.

2.3.3.5 Cálculo del ASE para logit(π)

Cuando la matriz de varianza y covarianza se puede calcular, el ASE para el logit(π) con una variable explicativa puede ser calculada usando:

$$ASE = \sqrt{Var(\hat{\beta}_0) + x^2 * Var(\hat{\beta}_1) + 2x * Cov(\hat{\beta}_0, \hat{\beta}_1)}$$

donde $Var(\hat{\beta}_0), Var(\hat{\beta}_1)$ y Covar$(\hat{\beta}_0, \hat{\beta}_1)$ son valores obtenidos de la matriz de varianza - covarianza, x es el valor de la variable explicativa.

Si la matriz de varianza - covarianza para el modelo no se puede calcular, el ASE para logit(π) puede ser calculado usando un procedimiento que transforma la variable explicativa y recalcula el modelo. Usando este método uno puede obtener un estimador del ASE para un valor particular de X. Por ejemplo, supóngase que x = 50.

1. A cada valor original de la variable explicativa se le resta el valor de 50, es decir, ahora cada valor es igual a x=50.
2. Se calcula el valor de regresión logística usando los nuevos valores (x = 50).

3. En los resultados estadísticos, el ASE asociado con la constante da una estimación del ASE para el logit.

2.4 PRUEBAS DE SIGNIFICANCIA

Hay un número de formas en los cuales la prueba de bondad de ajuste de un modelo de regresión logística puede ser calculado. Quizá el más usado y más poderoso es el estadístico log-verosimilitud. El estadístico de Wald y el análisis basado en la clasificación en tablas son comúnmente usados para evaluar el ajuste del modelo y por lo tanto también son discutidas como medidas de bondad de ajuste.

2.4.1 El estadístico log-verosimilitud

El estadístico log-verosimilitud da una medida de la devianza para un modelo de regresión logística (esto es, una medida de la diferencia entre los valores observados y los predecidos del modelo) y puede ser utilizada como un estadístico de bondad de ajuste. El estadístico de bondad de ajuste es usualmente citado como –2 veces el log-verosimilitud (-2LL) que se aproxima a la distribución de una Ji- cuadrada, esto permite evaluar la significancia. La interpretación de –2LL es sencilla, si su valor es pequeño entonces el modelo se ajusta muy bien, que –2LL sea igual a cero indica que se tiene el mejor modelo ajustado, es decir, no hay devianza.

Un indicador del ajuste del modelo puede ser obtenido comparando su valor –2LL con el valor –2LL para el modelo nulo. Este estadístico se puede renombrar como $-2LL_{dif}$.

$$-2LL_{dif} = (-2LL_0)-(2LL_1)$$

donde $-2LL_0$ es la medida de la devianza en el modelo nulo $\log it(\pi) = \beta_0$ y $-2LL_1$ es la medida de la devianza en el modelo nulo $\text{logit}(\pi) = \beta_0 + \beta_1 \cdot x$.

El cambio en el valor de –2LL representa el efecto que la variable explicativa tiene en la devianza del modelo, el cual puede ser evaluado para significancia usando la distribución Ji-cuadrada con grados de libertad igual a la diferencia en el número de términos entre los dos modelos. El efecto que tiene la variable x en el modelo puede ser calculada usando el estadístico $-2LL_{dif}$ con un grado de libertad.

El estadístico –2LL tiene una relación cercana con el usado en la regresión lineal (múltiple). De hecho, la suma de cuadrados del residual, el cual es una medida de devianza en la regresión lineal (múltiple), puede ser visto como un análogo a –2LL, la cual es la medida de la devianza para la liga logit. Similarmente, el estadístico F usado en la regresión lineal (múltiple) puede ser visto como el análogo al estadístico Ji-cuadrado en la regresión logística. Lo anterior se resume en el siguiente cuadro.

	Medida de la devianza	Distribución de referencia
Regresión Lineal	Suma de cuadrados del residuo	F
Regresión logística	Log-verosimilitud (-2LL)	χ^2

2.4.2 El estadístico Wald

El estadístico Wald es comparable con –2LL. Pruebas de hipótesis de que los coeficientes de regresión para la variable explicativa es cero, (es decir, la variable explicativa o tiene efecto alguno en la variable respuesta). El estadístico Wald puede ser calculado como:

$$\text{Estadístico Wald} = \frac{\beta^2}{ASE_\beta}$$

donde β es el coeficiente de regresión, y ASE_β es el error estándar de β.

El estadístico de Wald se aproxima a la distribución Normal estándar y es utilizada para pruebas de significancia usando una tabla de la distribución Normal estándar. Un estadístico Wald significativo sugiere que la variable explicativa tiene un efecto en la variable respuesta, esa es fácil de calcular e interpretar, aunque debería ser usado con precaución ya que tiende a exagerar la significancia de las variables, las cuales tienen coeficientes altos, y puede también ser poco fidedigno para muestras pequeñas. Dadas las restricciones para su uso, se recomienda que la prueba de bondad al ajuste para un modelo logístico sea calculado usando el estadístico basado en –2LL.

2.4.3 Análisis de datos a través de su clasificación en tablas

La eficiencia predictiva de un modelo particular puede ser calculado usando una tabla de clasificación, la cual compara las frecuencias observadas con las predecidas del modelo. Cuando se trabaja con una variable de respuesta binaria esta información es presentada en una tabla de contingencia de 2x2.

Hay un número de pruebas estadísticas que pueden ser usadas para calcular la significancia de la relación entre los valores predecidos y observados mostrados en la tabla de clasificación, por ejemplo λ de Goodman y Kruskal, τ de Kendall, ϕ, r de Pearson y los momios. Pero estas pruebas no son discutidas en este apartado ya que usualmente son consideradas como inferiores en comparación a la prueba basada en el estadístico –2LL.

Capítulo 3

MODELO DE REGRESIÓN LOGÍSTICA MÚLTIPLE

INTRODUCCIÓN

Los modelos de regresión logística múltiple pueden ser construidos para una variable de respuesta binaria, usando algún número de variables explicativas. Como en el análisis de regresión múltiple, estas variables explicativas pueden ser continuas o discretas. El componente sistemático (ver capitulo anterior) del modelo es similar al modelo de regresión lineal y hay aquí algunas consideraciones que se aplican a los modelos de regresión logística múltiple.

3.1 LA ECUACIÓN DEL MODELO

La ecuación del modelo de regresión logística para variables explicativas es similar al caso donde hay solamente una variable explicativa excepto que se permite que sean introducidas más de una variable en el modelo (el componente sistemático del modelo puede ser determinado por más de una variable). Cuando hay más de una variable explicativa, el modelo de regresión logística puede ser escrito como se muestra en la ecuación [3.1]

$$\text{Probabilidad de que un evento ocurra } \pi(z) = \frac{e^{(z)}}{1+e^{(z)}} \qquad [3.1]$$

llamada la función de regresión logística múltiple.

Donde: e es la base del logaritmo natural

z es el componente lineal (sistemático) del modelo y es igual a $\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$

La relación entre la probabilidad de que un evento ocurra (π) y el predictor lineal (z) en la ecuación [3.1] es no lineal. Sin embargo, siπ es transformada a logit(π (z)), esta relación puede ser transformada en una relación lineal.

$$\text{logit}(\pi(z)) = \text{logit}(\pi(x_i)) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k \qquad [3.2]$$

Los coeficientes calculados en el componente lineal referido como $\text{logit}(\pi(x_i))$, el cual, es difícil de interpretar. Una forma sencilla de interpretar los coeficientes del modelo es usando las proporciones, las cuales pueden ser derivadas de la ecuación [3.3]

$$\frac{\pi(x_i)}{1-\pi(x_i)} = \exp(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k) \qquad [3.3]$$

donde $(\pi(x_i))$ es la probabilidad de que un evento ocurra, y (1-π (x_i)) es la probabilidad complemento, es decir, la probabilidad de que un evento no ocurra.

Los valores de β (los coeficientes para las variables explicativas) muestran los cambios en $\text{logit}(\pi(x_i))$ que es asociada con una unidad de cambio en la variable explicativa cuando las demás variables en el modelo son considerados como constantes.

3.2 *AJUSTE DEL MODELO*

Se tratarán las N respuestas binarias como variables aleatorias Bernoulli. Sea $x_i = (x_{i0}, x_{i1}, \ldots, x_{ik})$ que denota el i-ésimo conjunto de valores de k variables explicativas, i = 1, . . . , I; donde $x_{i0} = 1$, ya que en el modelo existen k+1 parámetros ha estimar y por cada β_i existe un $x_{i\bullet}$ y $x_{i1} = 1$.

El modelo de regresión logística se expresa como sigue:

$$\pi(x_i) = \frac{\exp(\sum_{j=0}^{k} \beta_j \cdot x_{ij})}{\left[1 + \exp(\sum_{j=0}^{k} \beta_j \cdot x_{ij})\right]} \qquad [3.4].$$

Cuando más de una observación en Y ocurre en un valor fijo de x_i, es suficiente registrar el número de observaciones n_i y el número de resultados " 1". Sea y_i la i-ésima respuesta. El modelo asume que los y_i son variables aleatorias binomiales con $E[y_i] = n_i\pi(x_i)$, donde $n_1+n_2+\ldots+n_I = N$. La función generalizada de probabilidad conjunta es proporcional al producto de las I funciones binomiales, lo anterior es equivalente a la función de verosimilitud

$$\prod_{i=1}^{I} \pi(x_i)^{y_i} \cdot [1-\pi(x_i)]^{n_i - y_i} = \left\{\prod_{i=1}^{I} [1-\pi(x_i)]^{n_i}\right\}\left\{\prod_{i=1}^{I} \exp\left[\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)^{y_i}\right]\right\}$$

$$= \left\{\prod_{i=1}^{I} [1-\pi(x_i)]^{n_i}\right\}\left\{\exp \sum y_i \cdot \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)\right\} \qquad [3.5]$$

para el modelo [3.4], el i-ésimo logit es $\sum_j \beta_j \cdot x_{ij}$, así que el término exponencial en la última expresión es igual a

$$\exp\left[\sum_i y_i\left(\sum_j \beta_j \cdot x_{ij}\right)\right] = \exp\left[\sum_j\left(\sum_i y_i \cdot x_{ij}\right)\beta_j\right]$$

De forma similar, $[1-\pi(x_i)] = \left[1+\exp\left(\sum_j \beta_j \cdot x_{ij}\right)\right]^{-1}$, el logaritmo natural de la probabilidad es igual a:

$$L(\beta) = \sum_j\left(\sum_i y_i \cdot x_{ij}\right)\cdot \beta_j - \sum n_i \log\left[1+\exp\left(\sum_j \beta_j \cdot x_{ij}\right)\right] \qquad [3.6]$$

Lo anterior depende únicamente de las cantidades binomiales a través del estadístico suficiente $\sum_i y_i \cdot x_{ij}$; j = 0, . . . , k.

Derivando la ecuación de probabilidad con respecto a los elementos de β_a se tiene:

$$\frac{\partial L(\beta)}{\partial \beta_a} = \sum_i y_i \cdot x_{ia} - \sum_i n_i \cdot x_{ia}\left[\frac{\exp\left(\sum_j \beta_i \cdot x_{ij}\right)}{1+\exp\left(\sum_j \beta_i \cdot x_{ij}\right)}\right] \qquad [3.7]$$

e igualando a cero [3.6] para obtener el máximo de la función de verosimilitud [3.5], se tiene que las ecuaciones de probabilidad son:

$$\sum_i y_i \cdot x_{ia} - \sum_i n_i \cdot \hat{\pi}_i \cdot x_{ia} = 0; \quad a = 0, \ldots, k \qquad [3.8]$$

donde $\hat{\pi}_i = \dfrac{\exp\left(\sum_j \hat{\beta}_j \cdot x_{ij}\right)}{\left[1 + \exp\left(\sum_j \hat{\beta}_j \cdot x_{ij}\right)\right]}$ denota el estimador de máxima verosimilitud de π (x_i).

Sea X la matriz de Ix(k+1) valores de $\{x_{ij}\}$. La ecuación de probabilidad [3.6] tiene la forma

$$X'Y = X'\hat{m} \qquad [3.9]$$

donde $\hat{m}_i = n_i \cdot \hat{\pi}_i$. La ecuación anterior es similar a la utilizada en el método de mínimos cuadrados para el modelo de regresión lineal, es decir, $X'Y = X'\hat{Y}$, donde $\hat{Y} = X\hat{\beta}$ y $\hat{\beta} = (X'X)^{-1}X'Y$. La ecuación [3.8] ilustra un resultado fundamental para los modelos lineales generalizados, que usan la liga canónica. Las ecuaciones de probabilidad son iguales al estadístico suficiente para la estimación de sus valores esperados.

La matriz información es el valor esperado negativo de la matriz de segundas derivadas. Bajo condiciones regulares, los estimadores de máxima verosimilitud de parámetros tienen una distribución normal para muestras grandes con matriz de covarianza igual a la inversa de la matriz información. Para el modelo de regresión logística

$$\frac{\partial^2 L(\beta)}{\partial \beta_a \partial \beta_b} = -\sum_i \frac{x_{ia} \cdot x_{ib} \cdot n_i \cdot \exp\left(\sum_j \beta_j \cdot x_{ij}\right)}{\left[1+\exp\left(\sum_j \beta_j \cdot x_{ij}\right)\right]^2}$$

$$= -\sum_i x_{ia} \cdot x_{ib} \cdot n_i \cdot \pi_i \cdot (1-\pi_i) \qquad [3.9]$$

La expresión [3.9] no es una función de $\{y_i\}$, la matriz de valores esperados y observados son iguales. Esto pasa para todos los modelos lineales generalizados que usan la liga canónica.

Para estimar la matriz de covarianza se sustituye a β por $\hat{\beta}$ en la matriz teniendo elementos iguales a los negativos de [3.8], e invirtiendo se tiene:

$$\hat{Cov}(\hat{\beta}) = \{X' Diag[n_i \cdot \hat{\pi}_i \cdot (1-\hat{\pi}_i)] \cdot X\}^{-1} \qquad [3.10]$$

donde $Diag[n_i \cdot \hat{\pi}_i \cdot (1-\hat{\pi}_i)]$ denota la matriz diagonal que tiene elementos $\{n_i \cdot \hat{\pi}_i \cdot (1-\hat{\pi}_i)\}$ en la diagonal principal. La raiz cuadrada de los elementos de la diagonal principal de [3.10] son los errores estándar estimados del modelo.

En un conjunto **X** fijo, la varianza estimada del logit predecido $\hat{L} = X \cdot \hat{\beta}$ es $\hat{\sigma}^2(\hat{L}) = X \cdot Cov(\hat{\beta}) \cdot X'$. Para muestras grandes, $\hat{L} \pm z_{\alpha/2}\hat{\sigma}(\hat{L})$ es un intervalo de confianza para el logit.

3.2.1 *Interacciones y no linealidad*

Como en la regresión lineal múltiple, es posible incluir interacciones y términos no lineales en un modelo de regresión logistica. Interacciones que pueden ser explicados por términos adicionales que son incluidos en el componente sistemático del modelo, el cual muestra el producto de variables explicativas que interactúan. Por ejemplo, si hay una interacción entre x_1 y x_2, el componente lineal del modelo puede ser representado como $\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_1 \cdot x_2$. El efecto principal de cada variable explicativa, esto es, el efecto que las variables x_1 y x_2 tienen sobre la variable respuesta, son incluidas en el modelo junto con la interacción x_1x_2.

Agregando a las interacciones, las relaciones no lineales también pueden ser explicadas por términos polinomiales incluidos en el componente sistemático del modelo. Por ejemplo, si la variable x_2 mostró una relación no lineal con la variable respuesta, un término cuadrático como x_2^2 puede ser incluida en la ecuación. Un modelo que contiene dos variables explicativas, una de las cuales muestra una relación no lineal con el predictor lineal puede ser representado como:

$$\log it(\pi_i) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_2^2.$$

3.3 *Intervalos de confianza*

3.3.1 *Intervalos de confianza para β*

Intervalos de confianza para modelos de regresión logística que contienen múltiples variables explicativas pueden ser calculadas para el coeficiente de regresión β de la misma forma en que se calculó cuando sólo se tenía una sola variable explicativa.

$$\text{Intervalo de confianza para } \beta = \hat{\beta} \pm 1.96 \cdot ASE \text{ (ver pág. 45)}$$

Donde 1.96 es la aproximación normal de la distribución de la muestra al estadístico t a un nivel del 95% de confianza y ASE es el error estándar asintótico de $\hat{\beta}$.

3.3.2 *Intervalos de confianza para una y fija*

El cálculo, de los intervalos de confianza para una Y fija cuando se tienen múltiples variables explicativas es la misma que cuando se tenía una sola variable explicativa. Este intervalo es calculado primero para logit (π) y después es transformado para π.

$$\text{Intervalo de confianza para logit}(\pi_i) = \text{logit}(\hat{\pi}_i) \pm 1.96 \cdot ASE \qquad [3.12]$$

La dificultad con el cálculo de estos intervalos radica en determinar el ASE asociado con la predicción. Cuando los intervalos de confianza no

pueden ser obtenidos directamente a través de software, una de las dos siguientes técnicas pueden ser usadas para calcular el ASE para $\text{logit}(\pi_i)$.

Método 1

Si se puede obtener la matriz de varianza – covarianza, el ASE puede ser calculado usando:

$$\text{ASE para logit}(\pi_i) = \sqrt{\sum_{j=1}^{k} x_{j0}^2 \cdot Var(\hat{\beta}_j) + 2\sum_{j=1}^{k}\sum_{n=1}^{j} x_{n0} \cdot x_{j0} \cdot Cov(\hat{\beta}_n, \hat{\beta}_j)}$$

Donde $Var(\hat{\beta}_j)$ y $Cov(\hat{\beta}_n, \hat{\beta}_j)$ son valores obtenidos de la matriz varianza – covarianza, y x es el valor de la variable explicativa.

Como se puede ver, el cálculo del ASE en este método es complejo y necesita ser calculado para cada combinación de variables explicativas que interesen.

Método 2

Si no se puede obtener la matriz de varianza – covarianza, el ASE para $\text{logit}(\pi_i)$ puede ser calculado usando un procedimiento con el cual se transforma cada variable explicativa y recalcula el modelo. Usando este método se puede obtener una estimación del ASE para una combinación particular de valores de x. El procedimiento es el siguiente.

1. Crear nuevas variables explicativas, es decir, elegir un valor y restarlo a cada uno de los casos de las variables explicativas.
2. Recalcular el modelo usando las nuevas variables explicativas.
3. El ASE asociado con la constante da una estimación del ASE asociado para el logit (π_i).
4. Finalmente el intervalo de confianza para logit (π_i) puede ser calculado usando la ecuación [3.11]

3.4 *PRUEBAS DE SIGNIFICANCIA*

Las pruebas de bondad de ajuste para modelos de regresión logística pueden ser derivados del mismo estadístico usado para regresión logística simple. Usando estos estadísticos también es posible evaluar los efectos que individual o grupalmente las variables tienen en el modelo ajustado. A través del estadístico log-verosimilitud, de Wald y del análisis de tablas de clasificación pueden todas ser usadas para representar el ajuste del modelo.

3.4.1 *El estadístico log-verosimilitud*

El método más popular para determinar la bondad de ajuste de un modelo de regresión logística. El estadístico log-verosimilitud es usualmente indicado como menos dos veces la log-verosimilitud, es decir, -2LL; el cual se aproxima a la distribución Ji-cuadrada y permite determinar niveles de significancia.

3.4.2 El modelo total

Una medida del efecto que todas las variables explicativas en el modelo tienen sobre la variable respuesta puede ser obtenida comparando –2LL para un modelo sin variables explicativas llamado el modelo nulo, con –2LL para un modelo que incluye todas las variables explicativas.

-2LL para el modelo nulo es comúnmente llamado la función inicial log-verosimilitud. La diferencia entre estos dos modelos representa el efecto que las variables explicativas tienen. Como –2LL se aproxima a la distribución χ^2, la significancia del cambio en –2LL puede ser determinada usando los grados de libertad igual al número de parámetros en el modelo excluyendo la constante. Este estadístico en ocasiones es llamado como la Ji-cuadrada del modelo. Este estadístico puede ser visto como análogo al estadístico F en la regresión lineal.

La Ji-cuadrada del modelo puede ser derivada usando la siguiente ecuación:

$$\text{Ji-cuadrada del modelo} = (-2LL_0) - (-2LL_1) = 2LL_1 - 2LL_0 \qquad [3.13]$$

donde $-2LL_0$ es una medida de la devianza para el modelo nulo $\text{logit}(\pi_1) = \beta_0$

$-2LL_1$ es una medida de la devianza para el modelo $\text{logit}(\pi_1) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$.

3.4.3 Variables individuales

El efecto que variables individuales o variables en grupo tienen sobre el ajuste del modelo puede ser determinado comparando el ajuste de modelos anidados. La cantidad por la cual –2LL decrece cuando variables adicionales son añadidas al modelo, indica el tamaño del efecto que estas variables tienen. La significancia del cambio en –2LL está determinada por la prueba χ^2 con grados de libertad igual a la diferencia en el número de términos entre los dos modelos. Se hará referencia a este estadístico como $-2LL_{dif}$. Este estadístico puede ser visto como el análogo al estadístico parcial – F en la regresión lineal.

-2LL puede ser derivado usando la siguiente ecuación:

$$-2LL_{dif} = (-2LL_p) - (-2LL_{p+q}) \qquad [3.14]$$

donde p es el modelo anidado más pequeño y p+q es el modelo más grande.

-2LL evalúa el efecto individual y en grupo de variables sobre el modelo además de ser usado para seleccionar el modelo.

3.4.4 Selección del modelo

Como en la regresión lineal, la cantidad de devianza en el modelo puede ser minimizada incluyendo algunas variables explicativas. Maximizar el poder explicativo del modelo en esta forma no siempre es benéfico ya que se puede caer en la inclusión de variables irrelevantes que pueden dar muy poco poder

explicativo pero que pueden incrementar el error estándar asociado con la predicción. Por esta razón es útil reducir el número de variables en el modelo eliminando las que no tengan una influencia significativa.

3.4.5 Criterio para incluir o quitar variable

Para poder decidir que variables pueden ser incluidas o eliminadas en un modelo de regresión logística se recurre al estadístico $-2LL_{dif}$. Usando la ecuación [3.13] los modelos de regresión anidados pueden ser comparados con el cálculo del efecto que variables explicativas individuales o en grupo tienen en la variable respuesta.

3.5 BONDAD DE AJUSTE

En el análisis de regresión múltiple, el indicador de ajuste del modelo es R^2 ajustada, el cual mide la proporción de la variación en la variable respuesta que es explicada por las variables predictoras (explicativas). En el caso de la regresión logit, también se puede calcular la proporción de variación, pero en este caso es imposible calcularlo para los valores observados de la variable respuesta, la cual toma valores de 0 y 1, para ajustarse exactamente a los valores de π. El valor máximo de está proporción depende de la media y varianza de π.

Una medida alternativa de bondad de ajuste puede ser derivada del estadístico de probabilidad. Sea L_0 que denota la devianza para el modelo nulo

y L_1 es la devianza para el modelo $\text{logit}(\pi_i) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$.

Sea pseudo - R^2 como:

$$pseudo - R^2 = \frac{1 - \left(\frac{L_0}{L_1}\right)^{2/n}}{1 - L_0^{2/n}} = \frac{L_1^{2/n} - L_0^{2/n}}{L_1^{2/n}\left(1 - L_0^{2/n}\right)} \qquad [3.15]$$

donde n es el tamaño de la muestra. El valor mínimo que puede tomar pseudo $- R^2$ es cero cuando el ajuste es malo, esto sucede cuando $L_1 = L_0$, y el valor máximo es uno cuando el ajuste es bueno y esto sucede cuando $L_1 = 1$. La definición de pseudo $- R^2$ no es una prueba formal de significancia.

Otra definición fue sugerida en 1974 por McFadden, y es la siguiente:

$$pseudo - R^2 = 1 - \left(\frac{LL_0}{LL_1}\right)^{2/n} \qquad [3.16]$$

que también desafortunadamente no es una prueba formal de significancia.

Tal vez la mejor forma de calcular el pseudo $- R^2$ es tomar la definición que maneja SAS para la regresión logit

$$pseudo - R^2 = \frac{2LL_1 - 2LL_0 - 2k}{2LL_0} \qquad [3.17]$$

donde k es el número de parámetros (coeficientes) a ser estimados sin incluir el coeficiente de intersección. Como se mencionó anteriormente la cantidad – $2LL_0 - 2LL_1$ es el modelo Ji – cuadrada.

Hay muchas dificultades con estas medidas, entre las que se pueden mencionar:

1) Hay muchas formas de calcular el pseudo – R^2, lo cual puede dar diferentes resultados para un mismo conjunto de datos.
2) Hay pocas bases para elegir una medida sobre otra.
3) Las pruebas estadísticas que utiliza el pseudo – R^2 no son permitidas para alguna de las medidas.

Por estas razones muchos autores no presentan valores del pseudo – R^2 cuando reportan resultados del análisis de regresión logística.

3.6 DEVIANZA

La devianza en la regresión logística corresponde al SSE (error estándar) en la regresión lineal, y es definida como:

$$D = -2\sum\left\{y_i \cdot \log\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\cdot \log\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right\} \qquad [3.18]$$

Aunque este estadístico ha sido usado para determinar si una variable debería ser agregada a un modelo, y no debería ser usado como una medida absoluta de la calidad de bondad de ajuste.

Existen dos razones por las cuales la devianza podría ser inadecuada cuando es aplicada al modelo completo. La inadecuación potencial de la devianza puede ser mejor si se remplaza cada uno con m_i. Éste representa el número de repeticiones de la i-ésima combinación de los valores del regresor.

El valor D es generalmente comparado con el valor de χ^2 con n-p grados de libertad, quizá usando un nivel de significancia de 0.05, pero tal hipótesis es basada en el hecho de que $m_i \to \infty$ para cada i. Esto es, cada m_i debería tener un valor D grande, ya que D tiene una aproximación a una distribución Ji-cuadrada. D no debería ser usado cuando, no se tienen repeticiones de las combinaciones de los valores de los regresores.

Algunos conjuntos de datos de regresión logística tienen m_i grande, sin embargo, para estos conjuntos está devianza podría ser útil.

Capítulo 4

REGRESIÓN LOGÍSTICA POLITÓMICA

INTRODUCCIÓN

El modelo logit multinomial, también llamado el modelo logit politómico, es una generalización del modelo logit binario. Es este sentido, binario significa que la variable respuesta tiene dos categorías, y multinomial significa que la variable respuesta tiene tres o más categorías.

Se asume que las respuestas de cada combinación tiene una distribución multinomial, y las cantidades multinomiales de diferentes combinaciones son independientes. Por su modelo muestral y la identificación de una variable respuesta este modelo es llamado modelos de respuesta multinomial.

Una importante aplicación de los modelos logit es determinar los efectos de variables explicativas en una elección sometida a un conjunto de opciones, por ejemplo el partido de preferencia (PAN, PRD, VERDE ECOLOGISTA) o el tipo de transporte para ir al trabajo (coche propio, microbús, combi, metro, caminar, bicicleta, motocicleta).

Modelos para variables de respuesta que consisten de un conjunto discreto de elecciones son llamados modelos de elección discreta. Estos modelos son importantes herramientas para economistas y geógrafos.

El modelo logit multinomial es el método más usado para analizar variables de respuesta categórica. Entre las razones de su popularidad se encuentran:

1) Es una generalización natural del modelo logit multinomial
2) Es equivalente al modelo loglineal

4.1 MODELO LOGIT MULTINOMIAL GENERAL

La idea básica del modelo logit multinomial es comparar dos resultados o selecciones al mismo tiempo. La base principal para la construcción del modelo logit multinomial es el llamado "baseline" o riesgo.

Sin pérdida de generalidad, para un resultado (Y) con J categorías (j = 1, . . . , J), sea el logit-baseline la comparación de la j-ésima categoría con la primera (o baseline), es decir:

$$BL_j = \log\left(\frac{P_j(y=j)}{P_1(y=1)}\right) = \log\left(\frac{\pi_j}{\pi_1}\right) \qquad j = 2, \ldots, J \qquad [4.1]$$

donde P_j y P_1 denotan las probabilidades para la j-ésima y primer categoría respectivamente. El uso de la primera categoría como baseline es arbitrario, cualquier otra categoría pudo ser el baseline.

Considérese el caso en el que se tiene una variable independiente x con un número limitado de categorías, es decir., X = 1, . . .,I. para cada valor de x el logit-baseline es:

$$BL_{ij} = \log\left(\frac{P_j(y=j \mid x=i)}{P_1(y=1 \mid x=i)}\right) = \log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) \qquad [4.2]$$

La ecuación anterior representa un modelo saturado, la estimación de la ecuación [4.2] también puede ser obtenida como:

$$\log\left(\frac{F_{ij}}{F_{i1}}\right) = \log\left(\frac{f_{ij}}{f_{i1}}\right) \qquad [4.3]$$

son las observaciones y frecuencias esperadas en el i-ésimo renglón y j-ésima columna para la tabla de clasificación de $X \times Y$.

La ecuación [4.3] puede ser expresado como un modelo lineal generalizado:

$$BL_{ij} = \sum_{i=1}^{I} \log\left(\frac{F_{ij}}{F_{i1}}\right) \cdot I(x=i) \qquad [4.4]$$

donde I(*) es la función indicadora, i = 1 es verdad, cero en otro caso. Con una variable codificada como dummy y la primer categoria como referencia, la ecuación [4.4] usualmente es escrita como:

$$BL_{ij} = \alpha_j + \sum_{j=1}^{J} \beta_{ij} \cdot I(x=i) \qquad x > 1$$

donde α_j es el logit-baseline para x = 1, y β_{ij} es la diferencia entre el logit-baseline entre x = 1 y x = i. En este caso α_j y β_{ij} pueden ser estimados separadamente para todo i, j. Una estimación simultánea resultaria en un modelo equivalente en este caso. Para modelos no saturados las estimaciones simultáneas generalmente producen resultados diferentes.

4.2 *MODELO LOGIT MULTINOMIAL ESTÁNDAR*

En este apartado se hablará de una situación más general, ahora el i-ésimo renglón representa al i-ésimo individuo.

Sea y_i que denota la variable de resultado politómica con categorías codificadas de 1, . . . ,J. Asociada a cada categoría una probabilidad de respuesta (P_{i1}, P_{i2} , . . . ; P_{ij}), representando la i-ésima elección del elector de pertenecer a una categoría en particular.

Las probabilidades de respuesta dependen de una transformación no lineal de la función lineal:

$$x_i \cdot \beta_j = \sum_{k=0}^{K} \beta_{jk} x_{ik} \qquad [4.5]$$

donde k es el número del predictor. Es importante hacer notar que para modelos logit ordenados y binarios, los parámetros en los modelos logit multinomial varían de acuerdo a las categorías de respuesta.

Los modelos logit multinomial pueden ser vistos como una extensión del modelo logit binario.

Para una variable respuesta con J categorías la probabilidad P_{ij} puede ser calculada como:

$$\pi_{ij} = \frac{\exp(x_i' \cdot \beta_j)}{\sum_{j=1}^{J} \exp(x_i' \cdot \beta_j)} = \frac{\eta_{ij}}{\sum_{j=1}^{J} \eta_{ij}} \qquad [4.6]$$

Con $\sum_{j=1}^{J}\pi_{ij}=1$ para cualquier i con la usual normalización $\beta_1 = 0$ se tiene que $\eta_{i1} = 0$ lo que significa que en la ecuación [4.6] se tiene:

$$P[y_i = j \mid x_i] = \frac{\exp(x_i' \cdot \beta_j)}{1+\sum_{j=2}^{J}\exp(x_i' \cdot \beta_j)} \qquad \text{para j} > 1 \qquad [4.7]$$

$$P[y_i = 1 \mid x_i] = \frac{1}{1+\sum_{j=2}^{J}\exp(x_i' \cdot \beta_j)} \qquad [4.8]$$

Para un modelo con k variables independientes se tiene un total de (k+1)×(J-1) parámetros que deben ser estimados. Se puede ver que cuando J = 2, se estima un solo conjunto de parámetros correspondientes al resultado y = 2 con la primera categoría (y = 1) como la categoría de referencia, en este caso se habla del modelo logit binario. La relación cercana entre el modelo logit binario y el modelo logit multinomial puede ser obscurecido por el hecho de que la variable dependiente binaria es frecuentemente codificada como (0, 1) en lugar de (1, 2). Detrás de la diferencia de codificación, el modelo logit binario puede ser visto como un caso especial del modelo logit multinomial.

Una alternativa para codificar una variable de respuesta politómica es tomar los valores 0, . . . , J-1 en lugar de 1, . . . , J, esto hace que el modelo logit multinomial se parezca más al modelo logit binario. Con la codificación

de 0, . . . , J-1, se puede seguir la convención de tomar como primer categoria a (y = 0) como la categoria de referencia, asi que se tiene a β_0 en lugar de β_1.

4.3 ESTIMACIÓN DE LOS PARÁMETROS

La estimación de los parámetros es llevada a cabo usando el método de máxima verosimilitud iterativamente. Es conveniente definir un conjunto de J variables dummy, sea $d_{ij} = 1$ si $y_i = j$ y 0 en otro caso. Este resultado es uno y solamente uno para cada observación. El estimador de log-verosimilitud es:

$$\log L = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \cdot \log(\pi_{ij}) \qquad [4.9]$$

4.4 INTERPRETACIÓN DE RESULTADOS DE MODELOS LOGIT MULTINOMIAL

4.4.1 Proporciones y momios

Proporciones y momios juegan un importante papel en los modelos multinomiales. En el caso de los modelos logit multinomial, las proporciones entre las categorias j y 1 para una i dada son simplemente:

$$\frac{\pi_{ij}}{\pi_{i1}} = \frac{\eta_{ij}}{\eta_{i1}} = \exp(x_i' \cdot \beta_j) \qquad j = 2, \ldots, J \qquad [4.10]$$

Los momios o logit, es la función lineal de x_i:

$$\log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = x_i' \cdot \beta_j \qquad j = 2, \ldots, J \qquad [4.11]$$

Dada la proporción J-1 de la ecuación [4.10], la interpretación de los coeficientes del modelo logit multinomial no es tan sencilla como en la regresión logística binaria.

Supóngase, por ejemplo que β_j es positiva en [4.11]. Entonces una unidad de incremento en x_i' es causada porque se incrementa $\log\left(\frac{\pi_{ij}}{\pi_{i1}}\right)$, es decir, se incrementa en β_j unidades.

Cuando $\log\left(\frac{\pi_{ij}}{\pi_{i1}}\right)$ se incrementa la proporción $\left(\frac{\pi_{ij}}{\pi_{i1}}\right)$ también se incrementa, esto se debe a que $\log\left(\frac{\pi_{ij}}{\pi_{i1}}\right)$ es una función monótona creciente de $\left(\frac{\pi_{ij}}{\pi_{i1}}\right)$.

$\left(\frac{\pi_{ij}}{\pi_{i1}}\right)$ puede incrementarse cuando π_{ij} y π_{i1} decrecen. Un valor positivo de β_i no necesariamente significa que una unidad incrementada en x_i' incrementa π_{ij}.

Lo anterior no sucede en el caso de regresión logística binaria porque el numerador y denominador de $\frac{\pi}{1-\pi}$ no puede moverse en la misma dirección, es decir, si π se incrementa $1-\pi$ debe decrementarse en la misma cantidad. Por lo tanto si $\frac{\pi}{1-\pi}$ se incrementa, π debe incrementarse también.

En la regresión logística multinomial, los efectos de las variables predictoras en $\log\left(\frac{\pi_{ij}}{\pi_{i1}}\right)$ y $\left(\frac{\pi_{ij}}{\pi_{i1}}\right)$ pueden ser engañosos, ya que π_{ij} puede estar en la dirección opuesta.

La interpretación de β_{jk} como un log-momio cuando x_k es una variable continua requiere una comparación de $x_k = x_k^0$, el cual es un valor arbitrario de x_k:

$$\log\left[\frac{\frac{\left(\pi_j \mid x_k = x_k^0 + 1\right)}{\left(\pi_1 \mid x_k = x_k^0 + 1\right)}}{\frac{\pi_j \mid x_k = x_k^0}{\left(\pi_1 \mid x_k = x_k^0\right)}}\right] = \beta_{jk} \qquad [4.12]$$

La relación anterior interesa por el contraste entre las categorías j y k. Lo anterior se puede extender a contrastar cualesquiera dos categorías j y j' incluyendo la explicación de los coeficientes para las categorías j y j'. La ecuación [4.10], puede ser extendida a:

$$\frac{\pi_{ij}}{\pi_{ij'}} = \frac{\eta_{ij}}{\eta_{ij'}} = \exp\left[x_i'(\beta_j - \beta_{j'})\right] \qquad [4.13]$$

Para una variable explicativa x_k, la diferencia entre los coeficientes $(\beta_{jk} - \beta_{j'k})$ determina la dirección del cambio en las proporciones dentro de las categorías j y j'. Una diferencia positiva significa que cuando x_k incrementa, hay una proporción más grande de observar la alternativa j que la j'. Esto es equivalente a cambiar la categoría de referencia, así que se puede calcular el cambio relativo en las proporciones entre cualquiera dos categorías. Si se desea la información de cómo las probabilidades cambian en x_k, se necesita calcular los efectos marginales.

4.4.2 Efectos Marginales

Como el modelo logit multinomial es un modelo no lineal, el impacto de x_k sobre π_{ij} no es constante sobre el rango de x_k. En general, los efectos marginales de un cambio en x_k sobre π_{ij} es complicado pero puede ser calculado como:

$$\frac{\partial \pi_{ij}}{\partial x_{ik}} = \pi_{ij}\left(\beta_j - \sum_{j=2}^{J} \pi_{ij} \cdot \beta_j\right) \qquad [4.14]$$

Los efectos marginales son menos útiles para modelos logit multinomial que para los modelos logit binario. Los efectos marginales para modelos logit binarios no son ambiguos, cuando hay un coeficiente positivo implica un cambio positivo en la probabilidad cuando x_k incrementa. Sin embargo,

en modelos de respuesta multinomial un cambio en $P(y_i = i)$ no necesariamente tiene el mismo signo que β_{jk}. Se insiste tener precaución cuando se interpretan los efectos marginales de modelos de respuesta multinomial y se recomienda una interpretación más simple basada en proporciones y momios.

El estadístico Wald puede ser usado para probar limitaciones lineales en un subconjunto de parámetros. Esto es, coeficientes que pueden ser cero, mientras que otros no.

Para probar las diferencias en coeficientes de categorías j y j' se puede usar la fórmula general

$$z = \frac{\hat{\beta}_j - \hat{\beta}_{j'}}{\sqrt{Var(\hat{\beta}_j) + Var(\hat{\beta}_{j'}) - 2 \cdot Cov(\hat{\beta}_j, \hat{\beta}_{j'})}} \qquad [4.15]$$

4.5 DIFERENCIA ENTRE DOS MODELOS

El estadístico log-verosimilitud o log-L es matemáticamente el más conveniente para trabajar.

Supóngase que se tienen dos modelos de regresión logit, los cuales tienen las mismas variables de respuesta pero diferentes variables predictoras, donde el segundo modelo tiene todas las variables predictoras incluidas en el primer modelo más una más. Se dice que el primer modelo está contenido en el segundo.

Sea L_1 la probabilidad del primer modelo y L_2 para el segundo modelo.

Un estadistico para el cuál la distribución muestral es conocida, es $-\log\left(\frac{L_1}{L_2}\right)^2$, donde $L_1 < L_2$.

$$-\log\left(\frac{L_1}{L_2}\right)^2 = -2\log\left(\frac{L_1}{L_2}\right)$$

$$= -2[\log(L_1) - \log(L_2)] \qquad [4.16]$$

La condición $L_1 < L_2$ significa que el primer modelo está contenido en el segundo modelo. Cuando $L_1 < L_2$, $-\log\left(\frac{L_1}{L_2}\right)^2$ es positivo.

Se usa [4.16] para probar que el segundo modelo se ajusta a los datos significativamente mejor que el primer modelo. La prueba es una simple prueba de χ^2, es decir, $-\log\left(\frac{L_1}{L_2}\right)^2$ se distribuye como una χ^2 con grados de libertad igual a la diferencia entre el número de coeficientes a ser estimados en los dos modelos.

La pseudo-R^2 para probar la bondad de ajuste se calcula de la misma forma que para el modelo de regresión logistica múltiple.

$$pseudo - R^2 = \frac{2 \cdot LL_1 - 2 \cdot LL_0 - 2 \cdot k}{-2LL_0} \qquad [4.17]$$

donde k denota el número de parámetros (coeficientes) a ser estimado sin incluir el coeficiente de intersección.

4.6 EL MODELO LOGIT CONDICIONAL

Aunque el modelo estándar esbozado anteriormente es el más usado por las ciencias, una forma diferente del modelo logit es encontrado en la investigación económica. Este modelo ha llegado a ser conocido como el modelo de elección discreta o modelo logit condicional. Sin embargo, también es frecuentemente llamado modelo logit multinomial. El modelo logit multinomial difiere del modelo estándar ya que se toman en cuenta las características de elecciones y sus variaciones con el individuo como variable explicativa. La aplicación de modelos de elección discreta involucra la investigación en la elección del consumidor en la cual los "costos", "precios" u otras características de las elecciones fueron las principales variables explicativas.

En el modelo logit estándar, las variables explicativas son invariantes con respecto a las categorías de respuesta, pero sus parámetros varían con los resultados. En el modelo logit condicional, las variables explicativas varían por las respuestas así como por el individuo, por lo tanto sus parámetros son considerados constantes sobre todas las categorías de resultado.

En una perspectiva aleatoria de utilidad, la utilidad asociada con la j-ésima elección puede ser escrita como:

$$P(y_i = j \mid z_{ij}) = \pi_{ij} = \frac{\exp(z_{ij}' \cdot \alpha)}{\sum_{hc} \exp(z_{ih}' \cdot \alpha)} \qquad [4.19]$$

Este modelo permite el número de elecciones disponibles para variar según la elección de los individuos y su particular conjunto de elección C.

4.7.1 Interpretación

La interpretación de los resultados de los modelos logit condicional supone el uso de proporciones y momios. La proporción de elegir la alternativa j versus j' puede ser expresada como:

$$\frac{\pi_{ij}}{\pi_{ij'}} = \exp\left[(z_{ij} - z_{ij'})' \cdot \alpha\right] \qquad [4.20]$$

El cual implica el siguiente logit

$$\text{logit}\left(\frac{\pi_{ij}}{\pi_{ij'}}\right) = \left[(z_{ij} - z_{ij'})' \cdot \alpha\right] \qquad [4.21]$$

Esta expresión plantea que el logaritmo de proporciones entre las alternativas j y j' es proporcional al peso dado por la diferencia entre los valores de los individuos en la variable explicativa para las dos alternativas, con el peso dado por los parámetros estimados (coeficientes β). Si los valores de una variable explicativa son el mismo para dos alternativas (j y j'),

entonces las variables no influyen en la elección de los individuos entre las alternativas j y j'.

Esta interpretación es comparada con la del modelo logit multinomial, en la cual las diferencias en los coeficientes de categorías de respuesta cruzadas determina la dirección del cambio en el momio cuando la variable independiente cambia.

4.7 EL MODELO MIXTO

El modelo logit multinomial puede ser modificado para incluir características de los individuos con características de resultado. El modelo mixto combina características del modelo estándar con el modelo logit condicional. Esto puede ser fácilmente logrado incorporando niveles individuales en las variables independientes en el modelo logit condicional. El modelo logit puede ser más sencilla de implementar, esto en la práctica es crear un conjunto de variables dummy correspondientes a cada uno de las J alternativas y multiplicar cada nivel individual de las variables explicativas por su conjunto de dummies. El modelo resultante contiene constantes específicas de resultados además de los rasgos de nivel individual. La correspondiente función de utilidad asociada con la j-ésima elección es ahora:

$$u_{ij} = z_i' \cdot \alpha + x_i' \cdot \beta_j + \epsilon_{ij} \qquad [4.22]$$

donde z_{ij} y x_i denotan respectivamente la variación de resultados y la variación individual de las variables explicativas; α y β_j denotan los efectos de asociación.

El modelo general o mixto puede ser escrito como:

$$P(y_i = j \mid x_i, z_{ij}) = \pi_{ij} = \frac{\exp(z_{ij} \cdot \alpha + x_i' \cdot \beta_j)}{\sum_{h \in C} \exp(z_{ij} \cdot \alpha + x_i' \cdot \beta_j)} \qquad [4.23]$$

el cual combina las ecuaciones [4.6] y [4.19]. para identificar el modelo, se puede normalizar en alguna alternativa y el conjunto de β para la cual la alternativa es cero.

Por ejemplo sea $\beta_1 = 0$, la cual equivale a la normalización en la primera alternativa. Siguiendo esta propuesta, el logit del modelo multinomial puede ser expresado como:

$$\log\left(\frac{\pi_{ij}}{\pi_{ij'}}\right) = x_i'(\beta_j - \beta_{j'}) + (z_{ij} - z_{ij'})'\alpha \qquad [4.24]$$

Excluyendo la variación de los resultados de las variables explicativas del modelo da como resultado el modelo logit multinomial estándar.

Capítulo 5

APLICACIÓN

INTRODUCCIÓN

El objetivo de este apartado es aplicar la teoría expuesta anteriormente. Para lograrlo se consideran dos ejemplos.

Ejemplo 1. Macaco cola de muñón. Con este ejemplo se pretende identificar que características biológicas (sexo, edad, linaje y parentesco) y sociales (estatus) influyen en su comportamiento diferencial cuando los individuos detectan la presencia de un objeto novedoso, en este caso se trata de un cilindro metálico.

Ejemplo 2. Rata arrocera . Con este otro ejemplo lo que se pretende es identificar que especie de roedores dañan con mayor intensidad los tallos de caña.

5.1 Ejemplo 1. Macaco Cola de Muñón

Como se había mencionado anteriormente, el objetivo es identificar que características influyen en el comportamiento de los cola de muñón, mantenidos en cautiverio exterior del Instituto Nacional de Psiquiatría, División de Neurociencias, Laboratorio de Sicología y Etología.

La exposición de un objeto novedoso es una prueba comúnmente utilizada en estudios de laboratorio de exploración y motivación, de la cual se obtienen parámetros conductuales estandarizados, por lo que en este apartado se analizará la influencia existente entre las categorías de edad, sexo, linaje, parentesco y el estatus con la motivación para explorar y manipular un objeto novedoso, hasta lograr la habituación presentándolo en un contexto social. Esto se hizo ya que la motivación a explorar puede ser muy distinta si el actor se halla en compañía de animales conocidos o cuando se le aísla para manipularlo.

Para este experimento se registraron seis pautas conductuales, las cuales fueron codificadas y definidas como:

1 Latencia de contacto. Tiempo que permanece un sujeto sin hacer contacto físico con el objeto.

2 Aproximación. Es la frecuencia y el tiempo en que un actor se dirige hacia el objeto.

3 Contacto. Es la duración en que un individuo toca el objeto con cualquier parte del cuerpo excepto con las manos y los pies.

4 Manipulación. Es el tiempo que transcurre desde que un actor hace maniobrar el objeto con manos y pies hasta dejarlo completamente.

5 Cargar. Que el objeto sea levantado totalmente del suelo.

6 Rodar. Es el tiempo que dura el objeto rodando, producto de un impulso provocado o por la sujeción de los extremos.

La escala de esta variable es nominal, es decir, sólo categoriza las conductas.

Las demás variables fueron definidas y codificadas como sigue:

Sexo

1 Macho
2 Hembra

Edad

1 Infante
2 Juvenil
3 Adulto

Linaje. Se obtuvieron tres matrilíneas provenientes de:

1 Catrina con 10 sujetos, considerada como una hembra central, que es más activa socialmente que otras hembras, tiene un lazo cercano con el macho líder y es de alto rango.
2 Canela con 10 sujetos, considerada como concéntrica
3 Titania con 6 sujetos, considerada como periférica
4 Existe un solo sujeto que esta fuera de cualquier matrilínea

Parentesco: Se adquirieron tres generaciones y se consideran como:

1 Primera generación: abuelo(a)

2 Segunda generación: padre madre

3 Tercera generación: hijo(a)

Estatus: Determinado por la representatividad en el comportamiento social de cada uno de los sujetos, habiendo 3 categorías.

α = dominante

β = intermedio

γ = colateral

El software que se utilizó para obtener el modelo fue STATA. Ya introducida la base de datos se obtuvo lo siguiente:

Tabla 5.1 Modelo

Multinomial regression		Number of obs =	6245
		LR chi2(25) =	273.69
		Prob > chi2 =	0
Log likelihood = -8969.9071		Pseudo R2 =	0.015

cond	Var. Ind.	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
2	sexo	-0.27	0.13	-2.02	0.04	-0.53	-0.01
	edad	0.25	0.11	2.21	0.03	0.03	0.48
	estatus	0.05	0.06	0.80	0.43	-0.07	0.17
	linaje	0.11	0.07	1.72	0.09	-0.02	0.24
	parentesco	0.04	0.12	0.30	0.76	-0.19	0.26
	_cons	-1.41	0.54	-2.63	0.01	-2.46	-0.36
3	sexo	-0.55	0.13	-4.12	0.00	-0.80	-0.29
	edad	0.53	0.11	4.61	0.00	0.30	0.75
	estatus	0.10	0.06	1.61	0.11	-0.02	0.21

linaje	0.20	0.07	3.04	0.00	0.07	0.33
parentesco	0.17	0.11	1.52	0.13	-0.05	0.40
_cons	-2.24	0.53	-4.20	0.00	-3.28	-1.19
4 sexo	0.06	0.12	0.51	0.61	-0.17	0.29
edad	0.21	0.11	1.96	0.05	0.00	0.41
estatus	-0.28	0.06	-4.91	0.00	-0.39	-0.17
linaje	0.07	0.06	1.14	0.26	-0.05	0.19
parentesco	0.33	0.11	3.09	0.00	0.12	0.54
_cons	-1.54	0.49	-3.14	0.00	-2.50	-0.58
5 sexo	-0.02	0.21	-0.10	0.92	-0.44	0.40
edad	1.93	0.24	8.10	0.00	1.47	2.40
estatus	-0.28	0.13	-2.23	0.03	-0.53	-0.03
linaje	0.12	0.13	0.88	0.38	-0.15	0.38
parentesco	1.13	0.19	6.08	0.00	0.76	1.49
_cons	-9.46	0.89	-10.67	0.00	-11.20	-7.73
6 sexo	1.43	0.79	1.82	0.07	-0.11	2.97
edad	0.47	0.78	0.61	0.55	-1.06	2.01
estatus	-1.69	0.59	-2.87	0.00	-2.84	-0.54
linaje	0.12	0.56	0.22	0.83	-0.97	1.22
parentesco	1.86	0.66	2.81	0.01	0.56	3.16
_cons	-9.88	2.92	-3.39	0.00	-15.60	-4.17

(Outcome cond=1 is te comparison group)

El software toma como categoría de referencia a la conducta con mayor frecuencia, que en este caso es la latencia de contacto.

Esta regresión indica la probabilidad de ocurrencia de cada conducta de acuerdo a las variables independientes que se proponen, el modelo resultante es el siguiente:

$$P(Y=1)=\frac{1}{1+\eta_{i1}+\eta_{i1}+\eta_{i2}+\eta_{i3}+\eta_{i4}+\eta_{i5}+\eta_{i6}}$$

$$P(Y=j)=\frac{\eta_{ij}}{1+\eta_{i1}+\eta_{i1}+\eta_{i2}+\eta_{i3}+\eta_{i4}+\eta_{i5}+\eta_{i6}} \qquad j=2,\ldots,6$$

Donde

$$\eta_{i2} = \exp(-1.41 + 0.27 \cdot sexo + 0.25 \cdot edad + 0.05 \cdot estatus + 0.11 \cdot linaje + 0.04 \cdot parentesco)$$
$$\eta_{i3} = \exp(-2.24 - 0.55 \cdot sexo + 0.53 \cdot edad + 0.10 \cdot estatus + 0.20 \cdot linaje + 0.17 \cdot parentesco)$$
$$\eta_{i4} = \exp(-1.54 - 0.06 \cdot sexo + 0.21 \cdot edad - 0.28 \cdot estatus + 0.07 \cdot linaje + 0.33 \cdot parentesco)$$
$$\eta_{i5} = \exp(-9.46 - 0.02 \cdot sexo + 1.93 \cdot edad - 0.28 \cdot estatus + 0.12 \cdot linaje + 1.13 \cdot parentesco)$$
$$\eta_{i6} = \exp(-9.88 + 1.43 \cdot sexo + 0.47 \cdot edad - 1.69 \cdot estatus + 0.12 \cdot linaje + 1.86 \cdot parentesco)$$

Este modelo politómico tiene la posibilidad de dar respuesta a la probabilidad de manifestar cada conducta según cambien las variables independientes, indica por ejemplo que si aumenta una unidad cada una de las variables independientes las probabilidades de manifestar cada una de las conductas será.

Machos		Hembras	
Conducta	**Probabilidad**	**Conducta**	**Probabilidad**
1	0.56	1	0.60
2	0.16	2	0.14
3	0.09	3	0.06
4	0.18	4	0.20
5	0.00	5	0.00
6	0.00	6	0.00

Los resultados sugieren que de alguna manera las conductas predominantes son la latencia de contacto y la manipulación, tanto para hembras como para machos.

5.2 Ejemplo 2. Rata arrocera

La industria azucarera tiene una gran importancia en la economía nacional con aproximadamente una superficie cultivada de 600,000 hectáreas (Flores 1994). Este monocultivo siempre ha estado expuesto al ataque de diversas plagas y enfermedades, que en grado variable afectan los rendimientos de campo (Sánchez 1994).

Entre las plagas y enfermedades que afectan a los cultivos de caña se incluye a los roedores, como la segunda en importancia que causa graves pérdidas económicas (Flores 1974); cinco especies de éstos son los que afectan los cultivos de caña de azúcar, los cuales fueron codificados y definidos como:

1 Sigmodos hispidus
2 Oryzomys couesi
3 Oryzomys chapmani
4 Permoyscus leucopus
5 Reithrodonthomys sumichjasi

Pero a pesar de la gran importancia que estos mamíferos representan para la industria azucarera nacional, no se tiene información precisa de los daños y mucho menos se ha corroborado que todas las especies antes mencionadas causen daños a los cultivos, por lo que la finalidad de este apartado es identificar que especie de roedores dañan con mayor intensidad los tallos de caña.

Al igual que en el primer ejemplo, la escala de esta variable es nominal, es decir, sólo categoriza las especies.

Mes

1 Abril 1994
2 Mayo 1994
3 Junio 1994
4 Julio 1994
5 Agosto 1994
6 Septiembre 1994
7 Octubre 1994
8 Noviembre 1994
9 Diciembre 1994
10 Enero 1995
11 Febrero 1995
12 Marzo 1995
13 Abril 1995

Peso. Representa el peso de las malezas en el interior y áreas no cultivadas o baldías de los cultivos de caña. Esta variable es considerada como continua.

Cultivo.

1 Tipo permanente (caña de azúcar)
3 Tipo temporales (maíz, sorgo y arroz)

Edad. Representa las diversas edades de los cultivos de caña.También considerada como variable continua.

Sexo

1 Machos inmaduros

2 Machos adultos

3 Hembras adultas

4 Hembras gestantes

Altura caña. Representa la altura de la caña. Considerada como variable continua (variable de razón).

Caña. Peso mensual de las malezas en el interior de los cultivos de caña. Considerada como variable continua (variable de razón).

Baldíos. Peso mensual de las malezas en las áreas no cultivadas o baldías. Considerada como variable continua (variable de razón).

Al correr la regresión en STATA considerando a las cinco especies como variable dependiente de los meses del año, del peso, de la edad de los cultivos de caña de azúcar, del sexo, el peso medio de las malezas en los baldíos, peso medio de las malezas en los cultivos de caña y de la altura de la caña, se obtuvo lo siguiente:

Tabla 5.2 Modelo

Multinomial regresión	Number of obs	=	1604
	LR chi2(32)	=	508.72
	Prob > chi2	=	0
Log likelihood = -1068.596	Pseudo R2	=	0.1923

especie	Var. Ind.	Coef.	Std. Err.	Z	P>|z|	[95% Conf.	Interval]
2	mes	-0.06	0.04	-1.44	0.15	-0.14	0.02
	peso	-0.03	0.00	-12.34	0.00	-0.04	-0.03
	cultivo	-0.50	0.08	-6.42	0.00	-0.65	-0.35
	edad	0.00	0.00	-0.50	0.62	-0.01	0.00
	sexo	-0.04	0.01	-3.05	0.00	-0.06	-0.01
	altura caña	0.00	0.00	2.47	0.01	0.00	0.01
	caña	0.00	0.00	-1.08	0.28	0.00	0.00
	baldios	0.00	0.00	-1.78	0.08	0.00	0.00
	_cons	2.39	0.39	6.09	0.00	1.62	3.16
3	mes	0.06	0.11	0.57	0.57	-0.16	0.28
	peso	-0.10	0.02	-6.10	0.00	-0.13	-0.07
	cultivo	0.22	0.20	1.09	0.27	-0.17	0.61
	edad	0.01	0.01	1.07	0.29	-0.01	0.02
	sexo	0.06	0.04	1.52	0.13	-0.02	0.14
	altura caña	0.00	0.00	0.02	0.98	-0.01	0.01
	caña	0.00	0.00	-0.34	0.74	-0.01	0.01
	baldios	0.00	0.00	0.73	0.47	0.00	0.01
	_cons	-2.09	1.13	-1.84	0.07	-4.32	0.13
4	mes	-0.12	0.10	-1.23	0.22	-0.31	0.07
	peso	-0.04	0.01	-5.89	0.00	-0.05	-0.03
	cultivo	-0.23	0.17	-1.35	0.18	-0.56	0.10
	edad	0.00	0.01	-0.16	0.87	-0.02	0.01
	sexo	-0.02	0.03	-0.52	0.60	-0.08	0.04
	altura caña	0.01	0.00	2.00	0.05	0.00	0.01
	caña	-0.01	0.00	-1.79	0.07	-0.01	0.00
	baldios	-0.02	0.00	-4.57	0.00	-0.02	-0.01
	_cons	4.21	1.11	3.80	0.00	2.04	6.39
5	mes	-1.77	0.86	-2.06	0.04	-3.45	-0.09
	peso	-0.08	0.02	-4.77	0.00	-0.11	-0.05
	cultivo	-0.47	0.30	-1.61	0.11	-1.05	0.10
	edad	-0.27	0.16	-1.64	0.10	-0.59	0.05
	sexo	0.00	0.05	-0.08	0.94	-0.10	0.09
	altura caña	0.04	0.02	2.24	0.03	0.00	0.07
	caña	0.00	0.01	-0.03	0.97	-0.01	0.01
	baldios	0.02	0.01	1.60	0.11	0.00	0.03
	_cons	-3.25	2.56	-1.27	0.20	-8.26	1.77

(Outcome especie==1 is the comparison group)

El software toma como categoría de referencia a la especie con mayor frecuencia, que en este caso es Sigmodon hispidus, aunque se puede elegir cualquier otra especie como categoría de referencia.

Esta regresión indica la probabilidad de ocurrencia de cada especie de acuerdo a las variables independientes que se proponen, el modelo resultante fue el siguiente:

$$P(Y=1)=\frac{1}{1+\eta_{i1}+\eta_{i1}+\eta_{i2}+\eta_{i3}+\eta_{i4}+\eta_{i5}}$$

$$P(Y=j)=\frac{\eta_{ij}}{1+\eta_{i1}+\eta_{i1}+\eta_{i2}+\eta_{i3}+\eta_{i4}+\eta_{i5}} \quad j = 2, \ldots, 5$$

Donde

$$\eta_{i2}=\exp(2.389-0.059\cdot mes-0.033\cdot peso-0.500\cdot cultivo+0.001\cdot edad - 0.039\cdot sexo+0.003\cdot altcaña-0.001\cdot caña-0.001\cdot baldios)$$

$$\eta_{i3}=\exp(-2.092-0.063\cdot mes-0.099\cdot peso+0.218\cdot cultivo-0.007\cdot edad + 0.060\cdot sexo+0.000\cdot altcaña-0.001\cdot caña+0.002\cdot baldios)$$

$$\eta_{i4}=\exp(4.214-0.118\cdot mes-0.040\cdot peso-0.230\cdot cultivo-0.001\cdot edad - 0.016\cdot sexo+0.005\cdot altcaña-0.007\cdot caña-0.016\cdot baldios)$$

$$\eta_{is} = \exp(-3.245 - 1.769 \cdot mes - 0.079 \cdot peso - 0.475 \cdot cultivo - 0.271 \cdot edad$$
$$- 0.004 \cdot sexo + 0.039 \cdot altcaña - 0.000 \cdot caña - 0.015 \cdot baldios)$$

Este modelo politómico tienen la posibilidad de dar respuesta a la probabilidad de aparición por cada especie según cambien las variables independientes, indica por ejemplo que si aumenta una unidad cada una de las variables independientes las probabilidades de aparición por cada especie será.

Machos inmaduros	
Especie	**Probabilidad**
1	0.019
2	0.113
3	0.003
4	0.864
5	0.000

Machos maduros	
Especie	**Probabilidad**
1	0.020
2	0.111
3	0.003
4	0.866
5	0.000

Hebras adultas	
Especie	**Probabilidad**
1	0.020
2	0.108
3	0.004
4	0.868
5	0.000

Hembras gestantes	
Especie	**Probabilidad**
1	0.021
2	0.106
3	0.004
4	0.869
5	0.000

Las dos primeras especies fueron las de mayor abundancia, es decir fueron las más capturadas, por lo que tiene concordancia con las probabilidades; aunque resalta el resultado de la especie 4, ya que tiene una probabilidad muy alta de aparecer, lo cual indica que pudiera ser dominante en algún momento, si las condiciones le son propicias y causar graves daños a los cultivos de caña.

CONLCUSIONES

En el ejemplo 1, se observa en particular que las variables biológicas, preferentemente edad, linaje y parentesco son variables sobresalientes que influyen de manera significativa en el comportamiento diferencial cuando los individuos detectan la presencia de un objeto novedoso, siendo la latencia de contacto y la manipulación del objeto las conductas con mayor probabilidad, para el caso de machos y hembras.

En el ejemplo 2, aunque las dos primeras especies fueron las de mayor abundancia, hay que poner especial atención en la especie 4 ya que tienen una probabilidad muy alta de aparecer, lo cual indica que pudiera ser predominante en algún momento, si las condiciones le son propicias y causar graves daños a los cultivos de caña.

Es necesario tomar en cuenta algunas consideraciones de tipo práctico, para una mejor interpretación de los coeficientes β es necesario referirse al concepto de riesgo relativo, el $\exp(\beta)$ es la medida de influencia de la variable Xi sobre el riesgo de que ocurra ese hecho y suponiendo que el resto de las variables del modelo permanecen constantes.

Un intervalo de confianza para $\exp(\beta)$ que contenga al 1 indica que la variable no tiene influencia significativa en la ocurrencia del suceso, por el contrario, valores más alejados de este indican una mayor influencia de la variable.

Una de las ventajas de la regresión logística es que permite el uso de múltiples variables con relativamente pocos casos, sin embargo hay que tomar en cuenta algunas precauciones, se ha sugerido que el número de sujetos para poder usar esta técnica estadística sin problemas debe ser superior a $10(k+1)$ donde k es el número de variables explicativas; en el caso de que se introduzcan interacciones el número de elementos en la muestra debe aumentar.

En cuanto al número de variables independientes, la inclusión de un gran número de ellas en el modelo, puede indicar que no se ha reflexionado suficiente sobre el problema, al igual que en la regresión lineal es conveniente mantener la parsimonia en el modelo.

BIBLIOGRAFÍA

- Anderson T. W. / L. Sclove Stanle, An Introduction to the Statistical Analysis of Data.
  Houghton Miffin, USA 1978.

- M. Liebetrau, Measures of Association.
  Battele Pacific Northwest Laboratories Richland,Washington.
  Sage Publications, Inc., USA 1983.

- Agresti Alan, Analysis of Ordinal Categorical Data.
  John Wiley & Sons Inc., USA 1984.

- Agresti Alan / Finley Barbara, Statistical Methods for the Social Sciences.
  Dellen Publishing Company,
  Divisions of Macmillan, Inc., USA 1986.

- W. Hosmer David / Lemeshow Stanley, Applied Logistic Regression.
  John Wiley & Sons Inc., USA 1989.

- A. Powers Daniel / Xie Yu, Statistical Methods for Categorical Data Analysis.
  Academic Press., USA 1990.

- D. Retherford Robert / Choe Kim Minja, Statistical Models for Causal Analysis.
  John Wiley & Sons Inc., USA 1993.

- Lloyd J. Chris., Statistical Analysis of Categorical Data.
  John Wiley & Sons Inc,. USA 1999.

- Reference Manual, STATA. Volume 3.
  Computing Resource Center, Santa Mónica California 1992.