Modelo Lineal Generalizado

1er. Cuatrimestre de 2010

Nuestro ojetivo será estudiar la relación entre dos o más variables, que podrán ser tanto continuas como categóricas.

En algunos casos diferenciaremos entre variable de respuesta y variables explicativas, mientras que en otros, simplemente, nos interesará estudiar la asociación entre las variables presentes sin hacer esta distinción.

A diferencia de lo que se trata habitualmente en Modelo Lineal, la variable de respuesta podrá ser categórica.

Abordaremos tres grandes temas:

- Tablas de Contingencia
- Modelo Lineal Generalizado
- Modelos Log-lineales

Veremos algunos ejemplos que introduzcan estos temas.

Consideremos el caso en el una muestra de 980 norteamericanos fue clasificada de acuerdo con el sexo y su identificación político-partidaria. En esta situación nos interesa estudiar si hay asociación o no entre las variables categóricas G: Género y C:Identificación partidaria.

	C: Identificación partidaria			
G: Género	Demócrata	Independiente	Republicano	Total
Femenino	279	73	225	577
Masculino	165	47	191	403
Total	444	120	416	980

Cuadro 1: General Social Survey, 1991

Esta es una tablas de contingencia bastante sencilla.

Para responder a esto, en primera instancia, veremos los test de independencia o de homogeneidad basados en la distribución χ^2 que fueron introducidos por Pearson.

Sin embargo, estos tests, como muchos otros, tienen algunas limitaciones. Una de ellas es que si bien nos indican cuanta evidencia de asociación entre las variables de interés existe, no nos dicen nada sobre la naturaleza de esta relación.

Para comprender más profundamente la asociación entre variables nos ayudarán los modelos log-lineales y los modelos lineales generalizados, siendo estos últimos una generalización del modelo lineal habitual.

Ejemplo: Datos del Titanic

Estos datos se pueden encontrar en

http://www-m4.ma.tum.de/courses/WS08-09/glm/titanic.txt

corresponden a un curso dictado por la Prof. Claudia Czado de la Technishe Universität München y tienen como fuente

http://www.encyclopedia-titanica.org/

Las variables consideradas son

Name: Nombre del Pasajero

PClass: Clase del Pasajero

Age: Edad del Pasajero

Sex: Género del Pasajero

Survived: Survived=1 el Pasajero sobrevivió

Survived=0 el Pasajero no sobrevivió

```
   titanic <- read.table("c:\\users\\ana\\glm\\titanic.txt", header = T)
attach(titanic)
names(titanic)
"Name"     "PClass"   "Age"      "Sex"      "Survived"

length(Age)
[1] 1313

table(PClass)
PClass
1st 2nd 3rd
322 280 711

table(Sex)
Sex
female   male
   462    851

 table(Survived)
Survived
  0   1
863 450
```

```
table(Survived,PClass)
        PClass
Survived 1st 2nd 3rd
       0 129 161 573
       1 193 119 138

table(Survived,Sex)
        Sex
Survived female male
       0    154  709
       1    308  142
```

El objetivo es describir la asociació entre las variables presentes. Por ejemplo, de estas dos últimas tablas podríamos decir que los pasajeros hombres y los de tercera clase sobrevivieron menos.

En este ejemplo consideramos a *Survival* como variable de respuesta y a las demás como predictoras. En este caso, las variables predictoras pueden ser variables categóricas, ordinales o no, o biencuantitativas, ya sea continuas o discretas.

Survival	Sex	
	Female	Male
0	154	709
1	308	142

Cuadro 2: Tabla de 2×2

En forma genérica

Y	X	
	0	1
0	$1 - \pi(0)$	$1 - \pi(1)$
1	$\pi(0)$	$\pi(1)$

Cuadro 3: Tabla de 2×2

En esta situación podríamos intentar modelar la variable de respuesta en función de las explicativas.

Recordemos que en el modelo lineal simple habitual, si Y es nuestra variable de respuesta y x una variable explicativa podemos formular el modelo como:

$$E(Y) = \beta_0 + \beta_1 x \tag{1}$$

Si, como en el ejemplo, nuestra variable de respuesta es binomial, entonces $E(Y) = \pi$, por lo tanto la generalización inmediata de (1) sería:

$$E(Y) = \pi = \pi(x) = \beta_0 + \beta_1 x \tag{2}$$

Sin embargo, (2) no parece ser un modelo adecuado, pues $\beta_0 + \beta_1 x$ podría tomar valores fuera del intervalo $(0, 1)$.

Un problema evidente de este modelo es que la probabilidad π es acotada, mientras que las $\mathbf{x}'\boldsymbol{\beta}$ (en el caso de un vector de covariables) pueden tomar cualquier valor real. Si bien esto podría controlarse imponiendo complicadas restricciones a los coeficientes, esta solución no resulta muy natural.

Una solución sencilla es *transformar* la probabilidad mediante una función que mapee el intervalo $(0, 1)$ sobre la recta real y luego modelar esta transformación como una función lineal de las variables independientes.

Una elección muy frecuente es:

$$\text{logit}(\pi) = \log\left[\frac{\pi}{1-\pi}\right] = \beta_o + \beta_1 x$$

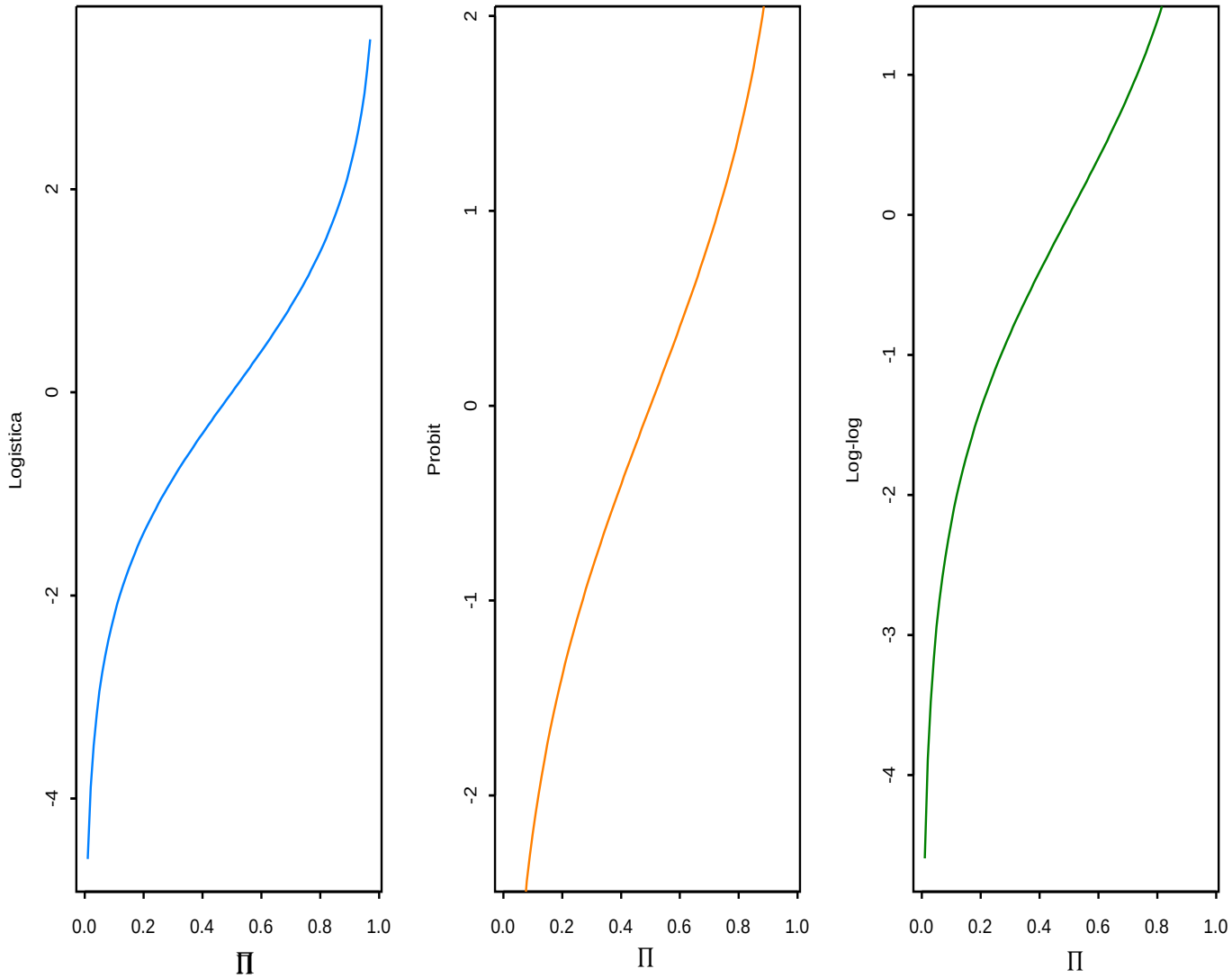
que da origen al modelo de regresión logística. Otra forma de escribirlo es

$$\pi(x) = \frac{e^{\beta_o+\beta_1 x}}{1 + e^{\beta_o+\beta_1 x}}$$

Esta es sólo una elección posible y veremos más adelante porque es una elección razonable.

Un punto a destacar es que en este modelo es natural la heteroscedasticidad, pues $V(Y) = \pi(1-\pi)$, que será función de x.

El modelo definido se conoce como modelo logístico, es un caso del **modelo**


Logistica
Probit
Log-log
Π
Π
Π
0.0
0.2
0.4
0.6
0.8
1.0
2
0
-2
-4
1
-1
-3

lineal generalizado con respuesta binomial y función de enlace logit.

Si bien los coeficientes $\boldsymbol{\beta}$ tienen una interpretación similar a la que tienen en el modelo lineal, debemos tener en cuenta que el miembro de la derecha es un logit y no una media, por lo que deberemos precisar cuál es su significado en este caso.

Estos temas los desarrollaremos en el contexto más general del **modelo lineal generalizado**.

Este modelo es una extensión que comprende al modelo lineal que aplicamos cuando el supuesto de normalidad es razonable y que abarca también el caso de una respuesta Poisson, Binomial Negativa, Gamma, Exponencial, etc.

Una vez establecido el modelo que queremos ajustar deberemos estimar los parámetros, hallar intervalos de confianza para los mismos, evaluar la bondad del ajuste y es probable que nos interese realizar algún test que involucre a los parámetros. También tendremos que evaluar la influencia de las observaciones en la determinación de los valores estimados.

En nuestro último ejemplo consideramos de nuevo el caso de una tabla de contingencia. Supongamos F es la variable que identificamos en las filas y C la de las columnas y que nos interesa estudiar la asociación de las variables categóricas.

	C						
F	1	2	.	.	j	.	. J
1	.	.			...		.
2	.	.			...		.
.	.	.			...		.
i	.	.			π_{ij}		.
.	.	.			...		.
I	.	.			...		.

Cuadro 4: Tabla Genérica

Sabemos que si F y C son independientes, las π_{ij} se pueden escribir en términos

de las marginales como

$$\pi_{ij} = \pi_i^F \, \pi_j^C \ .$$

¿Qué ocurre en el caso general cuando no suponemos independencia?

Si pensamos en los valores esperados, $m_{ij} = n\pi_{ij}$ también podremos expresar a m_{ij} usando un modelo multiplicativo:

$$m_{ij} = \tau \tau_i^F \, \tau_j^C \tau_{ij}^{FC} \ , \tag{3}$$

donde, como en ANOVA, los τ deberán satisfacer ciertas restricciones.

Si tomamos logaritmo en (3) queda:

$$\begin{aligned} \log m_{ij} &= \log \tau + \log \tau_i^F + \log \tau_j^C + \log \tau_{ij}^{FC} \\ \log m_{ij} &= \mu + \mu_i^F + \mu_j^C + \mu_{ij}^{FC} \end{aligned}$$

que resulta un modelo aditivo, al que estamos más acostumbrados.

Este tipo de modelos recibe el nombre de log-lineal. Una diferencia con el modelo lineal habitual es que aquí las dos variables tienen un rol simétrico. El investigador deducirá una asociación entre las variables interpretando los parámetros. Esta tarea puede ser más o menos compleja si la cantidad de parámetros es muy elevada, como ocurre cuando aumenta el número de variables en el problema.

Bibliografía:

- Agresti, A. (1990). Categorical Data Analysis. Wiley, New York.
- Christensen, R. (1997). Log-linear Models and Logistic Regression. 2da. Edición. New York: Springer Verlag.
- Bishop, I., Fienberg, S. y Holland, P. (1976). Discrete Multivariater Analysis: Theory and Practice.
- Dobson, A. (2001). An Introduction to Generalized Linear Models. 2da. Edición. Londres: Chapman and Hall.
- Lindsey, J. (1997). Applying Generalized Linear Models. New York: Springer Verlag .
- Mc. Cullagh y Nelder, J. A. (1989). Generalized Linear Models. 2da. Edición. Londres: Chapman and Hall.
- Santner, T. y Duffy, D. (1989). The Statistical Analysis of Discrete Data. New York: Springer Verlag.
- Rao, C. R. (1965). Linear Statistical Inference and Its Applications. New York: Wiley.

Tablas de Contingencia

En la primera parte del curso estudiaremos la relación entre 2 ó 3 variables categóricas. Introduciremos parámetros que describan la asociación entre variables categóricas y luego haremos inferencia sobre estos parámetro

Sean X e Y dos variables categóricas de respuesta, de manera que X tiene I niveles e Y tiene J niveles posibles. Cuando clasificamos sujetos de acuerdo a las dos variables tenemos IJ combinaciones posibles.

Las casillas de la tabla representan los IJ resultados posibles. La probabilidad de que (X, Y) tome el valor (i, j) será π_{ij}. Cuando las celdas contienen la frecuencia de cada resultado ij tenemos una **tabla de contingencia**, término que introdujo Pearson en 1904. También suele llamársela **tabla de clasificación cruzada**. Una tabla de contingencia con I filas y J columnas se dice una tabla de $I \times J$.

				Y				
X	1	2	.	.	j	.	.	J
1	.	.			...			.
2	.	.			...			.
.	.	.			...			.
.	.	.			...			.
i	.	.			π_{ij}			.
.		.			...			.
.	.	.			...			.
I	.	.			...			.

Cuadro 5: Distribución Conjunta

La distribución de probabilidad π_{ij} es la distribución conjunta de X e Y, mientras que las marginales de ambas variables las obtendremos sumando filas y columnas respectivamente: π_{i+} y π_{+j}, donde

$$\pi_{i+} = \sum_{j=1}^{J} \pi_{ij} \qquad \pi_{+j} = \sum_{i=1}^{I} \pi_{ij}$$

En muchos casos una de las variables, digamos Y es una variable de respuesta y la otra, X, es una variable explicativa. En general, es de interés estudiar cómo cambia la distribución de Y cuando pasamos de un nivel de X a otro. Dado que un sujeto está clasificado en la fila i de X, $\pi_{j|i}$ es la probabilidad de que clasifique en la columna j de Y, es decir $\{\pi_{1|i}, \ldots, \pi_{J|i}\}$ es la **probabilidad condicional** de Y dado que $X = i$. En términos de las probabilidades definidas, tenemos que

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} \quad \forall\, i, j\,.$$

Diremos que X e Y son **independientes** si

$$\pi_{ij} = \pi_{i+}\pi_{+j} \quad \forall\, i, j\,,$$

y cuando vale la independencia

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} = \frac{\pi_{i+}\pi_{+j}}{\pi_{i+}} = \pi_{+j} \quad \forall\, i, j\,.$$

En una tabla de 2×2 tendríamos:

	Columnas		
Filas	1	2	Total
1	π_{11} $(\pi_{1\|1})$	π_{12} $(\pi_{2\|1})$	π_{1+} (1)
2	π_{21} $(\pi_{1\|2})$	π_{22} $(\pi_{2\|2})$	π_{2+} (1)
Total	π_{+1}	π_{+2}	1

Cuadro 6: Distribución 2×2

Supongamos que en n individuos observamos (X, Y) y volcamos esta información en una tabla de contingencia: n_{ij} el número de individuos que tienen $X = i$ e $Y = j$, de manera que

$$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \, .$$

En el **caso muestral** la información también suele disponerse sobre una tabla como la que sigue:

	$Y=1$	$Y=2$	$\cdot$	$Y=j$	$\cdot$	$Y=J$
$X=1$	n_{11}	n_{12}				n_{1J}
$X=2$	n_{21}	n_{22}				n_{2J}
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$X=i$	$\cdot$	$\cdot$	$\cdot$	n_{ij}	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$X=I$	n_{I1}	n_{I2}				n_{IJ}

Cuadro 7: Tabla de Contingencia Genérica de $I \times J$

Ejemplo de 2×2

Comencemos por considerar un ejemplo de los más sencillos de tablas de contingencia en el que tenemos dos variables.

En general tendremos:

Podría interesarnos testear la hipótesis hipótesis de independencia entre las variables.

Podemos escribir esta hipótesis como

G:Género	**C**: Cree en la vida postmortem		
	Si	No	Total
Fem	435	147	582
Masc	375	134	509
Total	810	281	1091

Cuadro 8: General Social Survey, 1991

$$H_o : \pi_{ij} = \pi_{i+}\,\pi_{+j} \quad \forall i\,\forall j$$

Para resolver este problema necesitamos estimar las probabilidades. Podríamos estimar las probabilidades bajo el supuesto de independencia y comparar los valores observadosn con los valores esperados bajo independencia mediante un estadístico conocido como χ^2 de Pearson cuya distribución asintótica necesitaremos estudiar.

Podríamos estimar por máxima verosimilitud las probabilidades y realizar un test de cociente de verosimilitud. Para para esto necesitamos asumir una distribución subyacente.

	C		
G	Si	No	Total
Fem	n_{11}	n_{12}	n_{1+}
Masc	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	$n_{++} = n$

Cuadro 9: Tabla de 2×2

Si cada una de las n observaciones es clasificada en forma independiente en una de las $I \times J$ celdas de la tabla con probabilidad π_{ij}, entonces el vector aleatorio que representa el número de individuos clasificados en la celda (i, j), $\mathbf{n}$, tiene distribución multinomial. La frecuencias esperadas en cada casilla son $\mu_{ij} = n\pi_{ij}$. Para $I = J = 2$ sería

$$P(\mathbf{n} = (n_{11}^*, \ldots, n_{22}^*)) = \frac{n!}{n_{11}^*!\, n_{12}^*!\, n_{21}^*!\, n_{22}^*!}\, \pi_{11}^{n_{11}^*}\, \pi_{12}^{n_{12}^*}\, \pi_{21}^{n_{21}^*}\, \pi_{22}^{n_2^*}\,.$$

Salvo constantes, el log–likelihood en el caso general queda:

$$\ell = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \pi_{ij}$$

La maximización de ℓ debe contemplar que $\Sigma_{i=1}^{I} \Sigma_{j=1}^{J} \pi_{ij} = 1$.

La pregunta aquí es cómo llegaron los datos a la tabla.

Tipos de Muestreo

Dada una tabla de contingencia hay varios esquemas de muestreo que pueden conducir a los datos tal como los hemos observado y que podrían influir en el modelo de probabilidad a utilizar. En este caso tenemos dos factores F y C cada uno con dos niveles. En general, tendremos un factor *fila* con I niveles y un factor *columna* con J niveles que corresponde a una tabla de $I \times J$.

El número total de celdas es $N = I \times J$. Los totales marginales muestrales son

$$
\begin{aligned}
n_{i+} &= \sum_{j=1}^{J} n_{ij} && \text{total fila} \\
n_{+j} &= \sum_{i=1}^{I} n_{ij} && \text{total columna} \\
n_{++} &= n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} && \text{gran total}
\end{aligned}
$$

Usaremos una notación similar a la anterior, por ejemplo, p_{ij} será proporción

muestral de la casilla (i, j) definida por

$$p_{ij} = \frac{n_{ij}}{n}\,.$$

Las condicionales muestrales quedarán definidas análogamente, por ejemplo,

$$p_{j|i} = \frac{p_{ij}}{p_{i+}} = \frac{n_{ij}}{n_{i+}} \quad \forall\, i, j\,.$$

En el ejemplo anterior asumimos que todos los datos han sido recolectados muestreando 1091 individuos que fueron clasificados de acuerdo con el **sexo** y la **creencia en la vida postmortem**. Vemos las dos variables como respuesta y nos interesa su distribución conjunta.

En los experimentos que responden a este esquema de muestreo, seleccionamos una muestra de n individuos de una población y registramos los valores (X, Y) para cada individuo. La distribución conjunta de $\{n_{ij}\}$ es multinomial de parámetros n y $\boldsymbol{\pi} = \{\pi_{ij}\}$: $M(n, \boldsymbol{\pi})$, donde

$$\pi_{ij} = P(X = i, Y = j).$$

En este caso el gran total n es conocido y fijo. A veces, se expresa los parámetros como medias de las celdas:

$$\mu_{ij} = E(n_{ij}) = n\pi_{ij} \, .$$

Esto se conoce como **muestreo multinomial**.

Dado que la distribucón multinomial aparecerá con frecuencia en el análisis de datos categóricos, repasaremos algunas de sus propiedades.

Distribución Multinomial

Supongamos que realizamos n ensayos independientes, de manera que cada ensayo puede resultar en uno de los eventos $E_1, \ldots, E_N$ (los E_j's son mutuamente excluyentes y exhaustivos). En cada ensayo, el evento E_j puede ocurrir con probabilidad π_j y por lo tanto $\pi_1 + \ldots + \pi_N = 1$.

Si llamamos

$$X_j = \text{ número de veces que el evento } E_j \text{ ocurre ,}$$

entonces

$$X_1 + \ldots + X_N = n$$

y la distribución de $\mathbf{X} = (X_1, \ldots, X_N)'$ es multinomial de parámetros n y $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N)'$, es decir

$$\mathbf{X} = (X_1, \ldots, X_N)' \sim M(n, \pi_1, \ldots, \pi_N)$$
$$\mathbf{X} = (X_1, \ldots, X_N)' \sim M(n, \Pi)\,.$$

De manera que:

$$\begin{cases} P(X_1 = x_1, \ldots, X_N = x_N) = \dfrac{n!}{x_1! \ldots x_N!} \pi_1^{x_1} \ldots \pi_n^{x_N} & \text{si } \sum_{i=1}^{N} x_i = n \\ \\ 0 & \text{caso contrario (c.c.)} \end{cases}$$

Como en el caso de la distribución binomial, algunas propiedades más elementales de esta distribución, como el cálculo de esperanza o matriz de covarianza, se deducen fácilmente pensando a $\mathbf{X}$ como una suma de vectores de 0's y 1's. Más precisamente, podemos escribir

$$\mathbf{X} = \mathbf{Y}_1 + \ldots + \mathbf{Y}_n$$

donde las $\mathbf{Y}_i$ son independientes y cada una es $M(1, \pi_1, \ldots, \pi_N) = M(1, \Pi)$.

$$\mathbf{Y}_i' = (0, \ldots, \underset{\substack{\downarrow \\ j}}{1}, \ldots, 0)$$

$\mathbf{Y}_i$ es un vector con un 1 en la coordenada j si E_j ocurrió en el i-ésimo ensayo y en el resto de las posiciones 0's.

Los elementos de $\mathbf{Y}_i$ son Bernoulli correlacionadas.

Esperanza y Varianza

Por ejemplo, supongamos que $N = 2$ y que $\mathbf{Y} \sim M(1, \pi_1, \pi_2)$. Los resultados posibles son son

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{con probabilidad } \pi_1$$

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{con probabilidad } \pi_2 = 1 - \pi_1$$

La media de $\mathbf{Y} = (Y_1, Y_2)'$ es

$$E(\mathbf{Y}) = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix}$$

El segundo momento de $\mathbf{Y}$ es:

$$E(\mathbf{YY}') = E \begin{pmatrix} Y_1^2 & Y_1 Y_2 \\ Y_1 Y_2 & Y_2^2 \end{pmatrix}$$

$$= \begin{pmatrix} \pi_1 & 0 \\ \\ 0 & \pi_2 \end{pmatrix}$$

Por lo tanto, la matriz de covarianza de $\mathbf{Y}$ es:

$$\begin{aligned} \Sigma_{\mathbf{Y}} &= E(\mathbf{Y}\mathbf{Y}') - E(\mathbf{Y})E(\mathbf{Y}') \\ &= \begin{pmatrix} \pi_1 & 0 \\ \\ 0 & \pi_2 \end{pmatrix} - \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} (\pi_1 \quad , \quad \pi_2) \end{aligned}$$

$$= \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{pmatrix}$$

Vamos a demostrar que si hay N resultados posibles e $\mathbf{Y} \sim M(1, \pi_1, \ldots, \pi_N)$

$$\begin{aligned} E(\mathbf{Y}) &= \Pi = (\pi_1, \ldots, \pi_N)' \\ \Sigma_{\mathbf{Y}} &= \Delta(\Pi) - \Pi\Pi' \end{aligned}$$

Probaremos que si $\pi_j \neq 0, \forall j$

$$rg(\Sigma_{\mathbf{Y}}) = N - 1\,.$$

Luego, si $\mathbf{X} \sim M(n, \pi_1, \ldots, \pi_N)$

$$\begin{aligned} E(\mathbf{X}) &= n\Pi \\ \Sigma_{\mathbf{X}} &= \Delta(n\Pi) - n\Pi\Pi' \end{aligned}$$

Otras Propiedades

Enumeraremos algunas propiedades de la distribución multinomial que nos resultarán útiles, algunas serán ejercicio de la práctica.

1. El espacio paramétrico natural de la distribución multinomial es el simplex, en R^N definido por

$$\mathcal{S} = \{\boldsymbol{\pi} : \pi_j > 0, \pi_1 + \ldots + \pi_N = 1\}.$$

2. Si $\mathbf{X} = (X_1, \ldots, X_N)' \sim M(n, \pi_1, \ldots, \pi_N)$, entonces

$$\begin{aligned} X_j &\sim Bi(n, \pi_j) \\ Cov(X_i, X_j) &= -n\,\pi_i\,\pi_j \quad i \neq j\,, \end{aligned}$$

   es decir las X_i están negativamente correlacionadas.

3. Si $\mathbf{X} = (X_1, \ldots, X_N)' \sim M(n, \pi_1, \ldots, \pi_N)$, entonces

$$X^* = (X_1 + X_2, X_3, \ldots, X_N)' \sim M(n, \pi_1 + \pi_2, \pi_3, \ldots, \pi_N)$$

Es decir, si se colapsa una multinomial sumando celdas la distribución sigue siendo multinomial.

4. Sea $\mathbf{X} = (X_1, \ldots, X_N)' \sim M(n, \pi_1, \ldots, \pi_N)$. Consideremos la distribución condicional de

$$\mathbf{X}|_{\substack{X_1 + X_2 = z \\ X_3 + \ldots + X_N = n - z}}$$

Los vectores $(X_1, X_2)'$ y $(X_3, \ldots, X_N)'$ son condicionalmente independientes y multinomiales:

$$(X_1, X_2)' \sim M(z, \frac{\pi_1}{\pi_1 + \pi_2}, \frac{\pi_2}{\pi_1 + \pi_2})$$

$$(X_3, \ldots, X_N)' \sim M(n - z, \frac{\pi_3}{\pi_3 + \ldots + \pi_N}, \ldots, \frac{\pi_N}{\pi_3 + \ldots + \pi_N})$$

5. Si $X_1, \ldots, X_N$ son variables independientes tales que $X_j \sim \mathcal{P}(\lambda_j)$, entonces

$$(X_1, \ldots, X_N)'|_{\sum_{j=1}^{N} X_j = n} \sim M(n, \pi_1, \ldots, \pi_N)$$

donde

$$\pi_j = \frac{\lambda_j}{\lambda_1 + \ldots + \lambda_N}$$

Por lo tanto, la distribución de $X_1, \ldots, X_N$ puede ser factorizada en el producto de

$$n = \sum_{j=1}^{N} X_j \sim \mathcal{P}(\sum_{j=1}^{N} \lambda_j)$$

y

$$(X_1, \ldots, X_N)'|_{n=n^*} \sim M(n^*, \pi_1, \ldots, \pi_N)$$

Esto será especialmente útil a la hora de calcular la función de verosimilitud bajo ciertas condiciones.

En nuestro ejemplo hemos hablado de la función de verosimilitud, recordemos algunas propiedades.

Propiedades de los Estimadores de Máxima Verosimilitud

Recordemos que si la variable aleatoria Y tiene función de densidad (f.d.)o probabilidad puntual (f.p.p.)$f(y, \boldsymbol{\theta})$, la verosimiltud $L(\boldsymbol{\theta}, y)$ es simplemente $f(y, \boldsymbol{\theta})$ mirada como función de $\boldsymbol{\theta}$ con y fijo.

La función de probabilidad puntual o densidad es definida sobre el soporte $y \in \mathcal{Y}$, mientras que la verosimilitud es definida sobre un espacio paramétrico Θ.

En muchos casos es conveniente trbajar con el logaritmo de la verosimilitud (log-likelihood)

$$l(\boldsymbol{\theta}, y) = \log L(\boldsymbol{\theta}, y)$$

En general, tendremos una muestra aleatoria $Y_1, \ldots, Y_n$ con f.d. o f.p.p. $f(y, \boldsymbol{\theta})$, de manera que la verosimilitud será:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta})$$

y la log-versosimilitud

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i, \boldsymbol{\theta})$$

Una propiedad útil de los EMV es la de **invariancia** que dice que si g es una función con inversa g^{-1}, de manera que $\phi = g(\boldsymbol{\theta})$ implica que $\boldsymbol{\theta} = g^{-1}(\phi)$, entonces el EMV de ϕ, $\widehat{\phi}$, se calcula como

$$\widehat{\phi} = g(\widehat{\boldsymbol{\theta}}) \,,$$

siendo $\widehat{\boldsymbol{\theta}}$ el EMV de $\boldsymbol{\theta}$.

Como ya sabemos, podemos maximizar $L(\boldsymbol{\theta})$ o bien maximizar $l(\boldsymbol{\theta})$. En problemas regulares, el EMV puede hallarse igualando a 0 las derivadas pimeras de $l(\boldsymbol{\theta})$ respecto de $\boldsymbol{\theta}$.

La derivada primera de $l(\boldsymbol{\theta})$ respecto de $\boldsymbol{\theta}$ se llama score. En el caso univariado tenemos:

$$l'(\theta) = \sum_{i=1}^{n} u_i(\theta)$$

donde

$$u_i(\theta) = \frac{\partial}{\partial \theta} \log f(y_i, \theta)$$

Si tenemos q parámetros, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$, el vector de scores es

$$l'(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial l}{\partial \theta_1} \\ \frac{\partial l}{\partial \theta_2} \\ . \\ . \\ \frac{\partial l}{\partial \theta_q} \end{bmatrix} = \begin{bmatrix} \Sigma_{i=1}^{n} \frac{\partial}{\partial \theta_1} \log f(y_i, \theta) \\ \Sigma_{i=1}^{n} \frac{\partial}{\partial \theta_2} \log f(y_i, \theta) \\ . \\ . \\ \Sigma_{i=1}^{n} \frac{\partial}{\partial \theta_q} \log f(y_i, \theta) \end{bmatrix}$$

Una propiedad bien conocida del score es que su esperanza es nula:

$$E(l'(\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \int l'((\boldsymbol{\theta}_0))f(y,\boldsymbol{\theta}_0)dy = \int \frac{f'(y,\boldsymbol{\theta}_0)}{f(y,\boldsymbol{\theta}_0)} f(y,\boldsymbol{\theta}_0)dy = 0$$

La varianza de los score es conocida como la **información de Fisher**. En el caso univariado, la información de Fisher es:

$$i(\theta) \;=\; V(u(\theta)) = V(l'(\theta)) = E\left[(l'(\theta))^2\right]$$

Recordemos que

$$i(\theta) = E(-l''(\theta)) = -E(\frac{\partial^2}{\partial\theta^2}\log f(y,\theta))$$

En el caso multivariado, tendremos $\mathbf{I}(\theta)$ es una matriz de $q \times q$ tal que:

$$\{\mathbf{I}(\theta)\}_{ij} \;=\; -E(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(y,\theta))$$

En Estadística se probó que, bajo condiciones de regularidad, los EMV son asintóticamente normales, de manera que

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{\mathcal{D}} N(0, \mathbf{I}^{-1}(\theta))$$

¿Cómo sería en el caso general de una multinomial cualquiera?

Para simplificar la notación, indicaremos $\{n_1, \ldots, n_N\}$ las observaciones de cada casilla, con $n = \Sigma_{i=1}^{N} n_i$ y siendo $\{\pi_1, \ldots, \pi_N\}$ las probabilidades de cada celda.

Luego la función de verosimilitud será:

$$L = L(\pi_1, \ldots, \pi_N) = \frac{n!}{\prod_{i=1}^{N} n_i!} \prod_{i=1}^{N} \pi_i^{n_i} \qquad \text{donde } \sum_{i=1}^{N} \pi_i = 1$$

Como el ln es una función estrictamente creciente, hallar el máximo de L equivale a hallar el máximo de

$$l = \ln L = \ln \left(\frac{n!}{\prod_{i=1}^{N} n_i!} \right) + \sum_{i=1}^{N} n_i \ln \pi_i \qquad \text{donde } \sum_{i=1}^{N} \pi_i = 1$$

Como $\Sigma_{i=1}^{N} \pi_i = 1$, entonces $\pi_N = 1 - \Sigma_{i=1}^{N-1} \pi_i$ y $n! / \prod_{i=1}^{N} n_i!$ es constante, buscamos el máximo de

$$l = \ln L = cte + \sum_{i=1}^{N-1} n_i \ln \pi_i + n_N \ln(1 - \sum_{i=1}^{N-1} \pi_i)$$

Para buscar los puntos críticos planteamos:

$$\frac{\partial l}{\partial \pi_i} = \frac{n_i}{\pi_i} - \frac{n_N}{1 - \sum_{i=1}^{N-1} \pi_i} = \frac{n_i}{\pi_i} - \frac{n_N}{\pi_N} = 0$$

Luego,

$$\frac{n_i}{n_N} = \frac{\pi_i}{\pi_N} \Rightarrow \frac{n}{n_N} = \frac{1}{\pi_N} \Rightarrow \widehat{\pi}_N = \frac{n_N}{n}$$

$$\Rightarrow \widehat{\pi}_i = \frac{n_i \widehat{\pi}_N}{n_N} = \frac{n_i}{n}$$

Por lo tanto, tal como es de esperar

$$\widehat{\pi}_i = \frac{n_i}{n} = p_i \quad i = 1, \ldots, N$$

Muestreo Poisson

Otra posibilidad es que los 4 datos de la tabla del ejemplo sean realizaciones independientes de v.a. con distribución Poison.

Un proceso que sustentaría este modelo es que en cada celda los individuos lleguen aleatoriamente al lugar donde se los clasifica. En este caso n no está prefijado y todos los valores de la tabla son aleatorios.

En el **muestreo de Poisson** tenemos que

$$n_{ij} \sim \mathcal{P}(\lambda_{ij}) \quad \begin{array}{l} i = 1, \ldots, I \\ j = 1, \ldots, J \end{array}$$

independientes. En este esquema el gran total no está fijo, sino que es aleatorio.

Ejemplo:

Supongamos que se realiza en un control de velocidad durante una hora. Para ello se cuenta con un radar que registra la velocidad de cada auto que pasa por el puesto de observación. Supongamos que de cada auto que pasa se registra

la velocidad y la marca del auto. Así se obtienen

$$\begin{aligned} X &= \text{marca del auto (1 = Ford, 2 = Fiat, 3 = Chevrolet, 4 = Otros)} \\ Y &= \text{si el auto excede el límite de velocidad (1 = Si, 0 = No).} \end{aligned}$$

Es claro que, bajo independencia, $n \sim \mathcal{P}(\lambda_{++})$.

Si tenemos un muestreo de Poisson, la distribución de los n_{ij} condicional a que n está fijo en un valor, digamos n^*, ya no es más Poisson, más aún ya no son más independientes.

La distribución condicional de los n_{ij} dado que $n = n^*$ es multinomial. Para simplificar la notación, indicaremos $\{n_1, \ldots, n_N\}$, entonces si $\Sigma_{i=1}^{N} k_i = n^*$

$$\begin{aligned} P(n_1 = k_1, \ldots, n_N = k_N \mid \sum_{i=1}^{N} n_i = n^*) & \\ = \frac{P\left(n_1 = k_1, \ldots, n_N = k_N \cap \Sigma_{i=1}^{N} n_i = n^*\right)}{P(n = n^*)} & \end{aligned}$$

$$= \prod_{i=1}^{N} \frac{e^{-\lambda_i}\lambda_i^{k_i}}{k_i!} \quad \frac{n^*!}{e^{-\sum_{i=1}^{N}\lambda_i}\left\{\sum_{i=1}^{N}\lambda_i\right\}^{n^*}}$$

$$= \left\{\prod_{i=1}^{N}\lambda_i^{k_i}\right\}\frac{n^*!}{\prod_{i=1}^{N}k_i!}\frac{1}{\left\{\sum_{i=1}^{N}\lambda_i\right\}^{n^*}}$$

$$= \frac{n^*!}{\prod_{i=1}^{N}k_i!} \quad \prod_{i=1}^{N}\pi_i^{k_i}$$

donde $\pi_i = \frac{\lambda_i}{\sum_{j=1}^{N}\lambda_j}$ o volviendo a la notación original:

$$\pi_{ij} = \frac{\lambda_{ij}}{\lambda_{++}}$$

Entonces:

$$(n_1, \ldots, n_N)|_{n=n^*} \sim M(n^*, \pi_1, \ldots, \pi_N)\,.$$

Esto nos servirá a la hora de plantear la verosimilitud cuando deseemos estimar.

En efecto, podemos factorizar la verosimilitud como el producto del likelihood de la Poisson n ($n \sim \mathcal{P}(\lambda_{++})$) y el likelihood de una multinomial correspondiente a $\{n_{ij}\}$ dado n, con parámetros

$$\pi_{ij} = \frac{\lambda_{ij}}{\lambda_{++}}$$

El total n no da información acerca de las π_{ij}.

Es interesante observar, que desde el punto de vista de la verosimilitud, la inferencia sobre $\boldsymbol{\pi}$ es la misma si n es considerado fijo o aleatorio.

Muestreo Multinomial Independiente

Otra alternativa podría ser que sea razonable asumir que se tomó una muestra de 582 mujeres y otra muestra independiente de 509 hombres a los que se clasificó según su creencia. En este caso nos centramos en la distribución condicional de la creencia dado cada nivel de género.

En este esquema los totales por fila están fijos y tenemos n_1 individuos de género femenino y n_2 individuos de género masculino.

Si π_i es la probabilidad de que el individuo crea en la vida postmortem para el nivel i de género tendremos:

$$\pi_i = P(C = 1|G = i) = \frac{\pi_{i1}}{\pi_{i+}}$$

Las variables de interés serán: $Y_{i1} \sim Bi(n_i, \pi_i)$, $i = 1, 2$, que cuentan el número de individuos que sí creen en cada género.

La distribución conjunta de (Y_{11}, Y_{21}) es

$$P\left((Y_{11}, Y_{21}) = (y_{11}, y_{21})\right) = \frac{n_1!}{y_{11}!y_{12}!}\pi_1^{y_{11}}(1-\pi_1)^{y_{12}}\frac{n_2!}{y_{21}!y_{22}!}\pi_2^{y_{21}}(1-\pi_2)^{y_{22}}$$

La hipótesis de interés es la de **homogeneidad**, es decir que la probabilidad de creencia es la misma en ambos género:

$$H_0 : \pi_1 = \pi_2$$

Notemos que bajo independencia $\pi_{ij} = \pi_{i+}\pi_{+j}$, entonces la probabilidad condicional

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} = \pi_{+j}$$

es decir no depende de la fila i, con lo cual homogeneidad e independencia son equivalentes.

Supongamos que decidimos de antemano que vamos a muestrear n_{i+} individuos con $X = i\,(i = 1, \ldots, I)$ y que para cada uno de ellos registramos el valor de Y.

En este esquema cada fila de tabla $(n_{i1}, n_{i1}, \ldots, n_{iJ})'$ es multinomial con probabilidades

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}}$$

y las filas son independientes.

Este tipo de muestreo es razonable de aplicar cuando los datos provienen de un muestreo aleatorio estratificado (estratos definidos por X) o en un experimento en el X =grupo de tratamiento.

También es adecuado cuando no tenemos totales por filas fijos, pero estamos interesados en $P(Y|X)$ y no en $P(X)$, lo que corresponde a que Y es el resultado de interés y no deseamos modelar a X.

Por lo que vimos, en estos tres tipos de muestreo el núcleo de la verosimilitud es el mismo.

La importancia de estos resultados es que el análisis que hagamos es independiente del esquema de muestreo y depende de los parámetros de interés.