Profesor: Graciela Martínez Sánchez, graciela.mtz@ciencias.unam.mx
Ayudante: José de Jesús Ojeda Gónzalez, jesusojeda@ciencias.unam.mx
Alumno: Cristobal Bautista Hernández.
Fecha de entrega: lunes 25 de mayo de 2020.

Instrucciones: La tarea se entrega de manera individual, ordenada y bien escrita. La solución de la tarea debe enviarse en un archivo `.tex`

1. `(R)` Considere los datos en el archivo `coronary.csv`, la cual reporta la incidencia de muertes por enfermedades coronarias, de acuerdo con el grupo de edad y uso del cigarro. Ajuste modelos Poisson usando y sin usar un "offset" e interprete en cada caso.

2. Considere 3 variables: X y Y, de 2 categorías, y Z de K categorías. Para esta tabla de contingencia de $2 \times 2 \times K$ se define el cociente de momios condicional

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{21k}\mu_{12k}},$$

el cual describe la asociación entre las variables X y Y en cada nivel de Z.

(i) Muestre que para el modelo $\log(\mu_{ij}) = \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$, el log cociente de momios condicional es igual a cero para cada nivel de Z. ¿Qué se puede concluir en este caso?.

(ii) Para el modelo de asociación homógenea, encuentre el cociente de momios condicional $\theta_{XY(k)}$ para cada nivel de Z.

(i) *Solución:* Tome k una categoría de Z. Entonces:

$$log(\theta_{XY(k)}) = log\left(\frac{\mu_{11k}\mu_{22k}}{\mu_{21k}\mu_{12k}}\right) = log(\mu_{11k}) + log(\mu_{22k}) - log(\mu_{21k}) - log(\mu_{12k})$$

Observe que:

$$log(\mu_{11k}) - log(\mu_{21k}) = (\lambda + \lambda_1^X + \lambda_1^Y + \lambda_k^Z + \lambda_{1k}^{XZ} + \lambda_{1k}^{YZ}) - (\lambda + \lambda_2^X + \lambda_1^Y + \lambda_k^Z + \lambda_{2k}^{XZ} + \lambda_{1k}^{YZ})$$

$$\Rightarrow log(\mu_{11k}) - log(\mu_{21k}) = (\lambda_1^X + \lambda_{1k}^{XZ}) - (\lambda_2^X + \lambda_{2k}^{XZ})$$

Por otro lado:

$$log(\mu_{22k}) - log(\mu_{12k}) = (\lambda + \lambda_2^X + \lambda_2^Y + \lambda_k^Z + \lambda_{2k}^{XZ} + \lambda_{2k}^{YZ}) - (\lambda + \lambda_1^X + \lambda_2^Y + \lambda_k^Z + \lambda_{1k}^{XZ} + \lambda_{2k}^{YZ})$$

$$\Rightarrow log(\mu_{22k}) - log(\mu_{12k}) = (\lambda_2^X + \lambda_{2k}^{XZ}) - (\lambda_1^X + \lambda_{1k}^{XZ})$$

Esto implica que:

$$log(\theta_{XY(k)}) = log(\mu_{11k}) + log(\mu_{22k}) - log(\mu_{21k}) - log(\mu_{12k}) = (\lambda_1^X + \lambda_{1k}^{XZ}) - (\lambda_2^X + \lambda_{2k}^{XZ}) + (\lambda_2^X + \lambda_{2k}^{XZ}) - (\lambda_1^X + \lambda_{1k}^{XZ})$$

Con lo cual, se puede concluir que $log(\theta_{XY(k)}) = 0$ para cada nivel de Z.
Por otro lado, □

(ii) *Solución:* Para el encontrar el cociente de momios con el modelo de asociación homogéneo $\log(\mu_{ij}) = \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY}$. De esta manera, para cada nivel de Z se tiene que:

$$log(\theta_{XY(k)}) = log\left(\frac{\mu_{11k}\mu_{22k}}{\mu_{21k}\mu_{12k}}\right) = log(\mu_{11k}) + log(\mu_{22k}) - log(\mu_{21k}) - log(\mu_{12k})$$

Haciendo algo análogo al inciso anterior se tiene:

$$log(\mu_{11k}) - log(\mu_{21k}) = (\lambda + \lambda_1^X + \lambda_1^Y + \lambda_k^Z + \lambda_{1k}^{XZ} + \lambda_{1k}^{YZ} + \lambda_{11}^{XY}) - (\lambda + \lambda_2^X + \lambda_1^Y + \lambda_k^Z + \lambda_{2k}^{XZ} + \lambda_{1k}^{YZ} + \lambda_{21}^{XY})$$

$$\Rightarrow log(\mu_{11k}) - log(\mu_{21k}) = (\lambda_1^X + \lambda_{1k}^{XZ} + \lambda_{11}^{XY}) - (\lambda_2^X + \lambda_{2k}^{XZ} + \lambda_{21}^{XY})$$

$$log(\mu_{22k}) - log(\mu_{12k}) = (\lambda + \lambda_2^X + \lambda_2^Y + \lambda_k^Z + \lambda_{2k}^{XZ} + \lambda_{2k}^{YZ} + \lambda_{22}^{XY}) - (\lambda + \lambda_1^X + \lambda_2^Y + \lambda_k^Z + \lambda_{1k}^{XZ} + \lambda_{2k}^{YZ} + \lambda_{12}^{XY})$$

$$\Rightarrow log(\mu_{22k}) - log(\mu_{12k}) = (\lambda_2^X + \lambda_{2k}^{XZ} + \lambda_{22}^{XY}) - (\lambda_1^X + \lambda_{1k}^{XZ} + \lambda_{12}^{XY})$$

De esta manera:

$$log(\theta_{XY(k)}) = (\lambda_1^X + \lambda_{1k}^{XZ} + \lambda_{11}^{XY}) - (\lambda_2^X + \lambda_{2k}^{XZ} + \lambda_{21}^{XY}) + (\lambda_2^X + \lambda_{2k}^{XZ} + \lambda_{22}^{XY}) - (\lambda_1^X + \lambda_{1k}^{XZ} + \lambda_{12}^{XY})$$

$$\therefore log(\theta_{XY(k)}) = \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$$

□

3. Pruebe que para el modelo $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$. Las estadísticas de Pearson X^2 y de log-verosimilitud G^2 puede ser expresadas como $X^2 = \sum_k X_k^2$ y $G^2 = \sum_k G_k^2$, donde X_k^2 y G_k^2 corresponden a las estadísticas para probar la independencia entre X y Y en cada nivel de Z.

Demostración: Sean X_k^2 y G_k^2 estadísticas que prueban la independencia entre X y Y para la categoría k de Z. Entonces se pueden definir como:

$$X_j^2 = \sum_j \frac{(y_{ij} - \widehat{\mu}_{ij})^2}{\widehat{\mu}_{ij}} \qquad ; \qquad G_j^2 = 2\sum_i n_{ij} log(\frac{n_{ij}}{\widehat{\mu}_{ij}})$$

Entonces:

$$\sum_j X_j^2 = \sum_j \sum_i \frac{(y_{ij} - \widehat{\mu}_{ij})^2}{\widehat{\mu}_{ij}} = \sum_j \sum_i \frac{(y_{ij} - \frac{n_{i\cdot k}n_{j\cdot k}}{n_{\cdot\cdot k}})^2}{\frac{n_{i\cdot k}n_{j\cdot k}}{n_{\cdot\cdot k}}} = X^2$$

Por otra parte:

$$\sum_j G_j^2 = 2\sum_j \sum_i n_{ij} log(\frac{n_{ij}}{\widehat{\mu}_{ij}}) = 2\sum_j \sum_i n_{ij} log(\frac{n_{ij}}{\frac{n_{i\cdot k}n_{j\cdot k}}{n_{\cdot\cdot k}}}) = G^2$$

□

4. (R) Utiliza los datos de byosinossis, pero ahora utiliza un modelo log-lineal para modelar la frecuencia de la tabla generada por : años de empleado, fumador, sexo, raza, lugar y byossinosis.

5. Dado u, sea Y una v.a. Poisson con media condicional $E(Y|u) = u\mu$, donde u es una v.a. positiva con $E(u) = 1$ y $V(u) = \tau$. Muestre que $E(Y) = \mu$ y $V(Y) = \mu + \tau\mu^2$. Explique como puede formular una distribución binomial negativa y un GLM binomial negativo utilizando esta construcción.

Demostración: Sea Y una v.a Poisson con media condicional $E[Y|u] = u\mu]$ donde una es v.a positiva con $E(u) = 1$ y $V(u) = \tau$. Entonces, la esperanza de dicha variable Poisson es:

$$E[Y] = E[E[Y|u]] = E[u\mu] = \mu E[u] = \mu$$

Por otra parte, la varianza es:

$$Var(Y) = E[Var(Y|u)] + Var(E[Y|u]) = E[u\mu] + Var(u\mu) = \mu E[u] + \mu^2 Var(u) = \mu + \tau\mu^2$$

Finalmente, para encontrar una distribución con dicha información, se toma $\tau = 1/k$ de tal manera que:

$$f(\lambda; k, \mu) = \frac{(k/\mu)^k}{\Gamma(k)} e^{-\lambda k/\mu} \lambda^{k-1}, \qquad \lambda > 0$$

Con lo cual, se llega a una distribución de probabilidad conjunta que se distribuye Binomial Negativa:

$$p(y; \mu, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{k}{\mu+k}\right)^k \left(1 + \frac{k}{\mu+k}\right)^y, \qquad y = 0, 1, 2, \ldots$$

□

6. (R) Los datos en el archivo tarea3.xls representan

LOS	Length of stay in hospital (in days)
Hospital	The hospital
Insurer	The insurer, either 0 or 1
Age	The age of the patient
Gender	The gender of the patient: 1 means Female, 0 means male
Race	The race of the patient: 1 means white, 2 means hispanic, 3 means black, 4 means Asian/Pacific Islander, 5 means Other
BedSize	The number of beds in the hospital: 1 means 1 to 99, 2 means 100 to 249, 3 means 250 to 400, 4 means 401 to 650
Owner	The hospital owner: 1 means public, 2 means private
Compl	If there were any treatment complication: 0 means no complications, 1 means complications

La variable respuesta LOS no puede tomar el valor cero, ajusta modelos para conteos (poisson o binomial negativo) que consideren este hecho y modelos que no lo consideren y compara tus resultados. Realiza análisis post ajuste y comenta.

7. Sea Y_i el número de éxitos en n_i ensayos con $Y_i \sim Bin(n_i, \pi_i)$, donde las probabilidades tienen una distribución Beta $\pi_i \sim B(a, b)$. La función de densidad para una v.a. con disitribución Beta es

$$f(\pi_i; a, b) = \frac{1}{B(a,b)} \pi_i^{a-1} (1 - \pi_i)^{b-1}, \qquad 0 \le \pi_i \le 1,$$

donde $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ es la función beta. Defina $\theta = a/(a + b)$ y pruebe lo siguiente.

(i) $E(\pi_i) = \theta$.
(ii) $V(\pi_i) = \theta(1 - \theta)/(a + b + 1) = \phi\theta(1 - \theta)$.
(iii) $E(Y_i) = n_i\theta$.
(iv) $V(Y_i) = n_i\theta(1-\theta)[1 + (n_i - 1)\phi]$, por lo que la varianza de Y_i es más grande que la varianza del modelo binomial (excepto si $n_i = 1$ o $\phi = 0$.)

(i) *Demostración:* Dado que $\pi_i \sim B(a, b)$, entonces:

$$E[\pi_i] = \int_0^1 \pi_i f(\pi_i; a, b)\, d\pi_i = \int_0^1 \pi_i \frac{1}{B(a,b)} \pi_i^{a-1}(1-\pi_i)^{b-1}\, d\pi_i = \frac{1}{B(a,b)} \int_0^1 \pi_i^{a-1+1}(1-\pi_i)^{b-1}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{\Gamma(a+b)a\Gamma(a)\Gamma(b)}{\Gamma(a)\Gamma(b)(a+b)\Gamma(a+b)} = \frac{a}{a+b} = \theta \Rightarrow E[\pi_i] = \theta$$

□

(ii) *Demostración:* Dado que por definición se llega a que $V(\pi_i) = E[\pi_i^2] - E[\pi_i]^2$. Para esto, primero se calcula:

$$E[\pi_i^2] = \int_0^1 \pi_i^2 f(\pi_i; a, b)\, d\pi_i = \int_0^1 \pi_i^2 \frac{1}{B(a,b)} \pi_i^{a-1}(1-\pi_i)^{b-1}\, d\pi_i = \frac{1}{B(a,b)} \int_0^1 \pi_i^{a-1+2}(1-\pi_i)^{b-1}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} = \frac{\Gamma(a+b)\,(a+1)\,a\,\Gamma(a)\Gamma(b)}{\Gamma(a)\Gamma(b)(a+b+1)(a+b)\Gamma(a+b)} = \frac{(a+1)a}{(a+b+1)(a+b)}$$

De esta manera, utilizando el resultado del inciso anterior y el presente resultado:

$$V(\pi_i) = \frac{(a+1)a}{(a+b+1)(a+b)} - \frac{a^2}{(a+b)^2} = \left(\frac{a}{a+b}\right)\left(\frac{(a+1)(a+b) - a(a+b+1)}{(a+b)(a+b+1)}\right)$$

$$= \left(\frac{a}{a+b}\right)\left(\frac{a^2+ab+a+b-a^2-ab-a}{(a+b)(a+b+1)}\right) = \left(\frac{a}{a+b}\right)\left(\frac{b}{(a+b)(a+b+1)}\right) = \theta\left(\frac{b}{a+b}\right)\left(\frac{1}{a+b+1}\right)$$

$$= \theta\left(\frac{a-a+b}{a+b}\right)\left(\frac{1}{a+b+1}\right) = \theta\left(\frac{a+b}{a+b} - \frac{a}{a+b}\right)\left(\frac{1}{a+b+1}\right) = \theta(1-\theta)/(a+b+1)$$

$$\therefore V(\pi_i) = \theta(1-\theta)/(a+b+1)$$

□

(iii) *Demostración:* Sea Y_i una v.a con distribución $Bin(n_i, \pi_i)$, entonces:

$$E[Y_i] = E[E[Y_i|\pi_i]] \sum_{j=0}^{n_i} y_{ij} f(Y_i|\pi_i) = \sum_{j=0}^{n_i} y_i \int_0^1 f(Y_i, \pi_i) f(\pi_i)\, d\pi_i = \int_0^1 \left[\sum_{j=0}^{n_i} y_i f(Y_i, \pi_i)\right] f(\pi_i)\, d\pi_i$$

La función $f(Y_i, \pi_i)$ es de distribución $Bin(n_i, \pi_i)$, entonces:

$$E[Y_i] = \int_0^1 n_i \pi_i f(\pi_i)\, d\pi_i = n_i \int_0^1 \pi_i f(\pi_i)\, d\pi_i = n_i\theta$$

En dicha parte se uso, el inciso (i), por lo tanto $E[Y_i] = n_i\theta$

□

(iv) *Demostración:* Para calcular la varianza, al igual que en el inciso (ii) se tiene que calcular $E[Y_i^2]$

$$E[Y_i^2] = E[E[Y^2|\pi_i]] = \sum y_i^2 f(Y_i|\pi_i) = \sum_{j=0}^{n_i} y_i^2 \int_0^1 f(Y_i, \pi_i) f(\pi_i)\, d\pi_i = \int_0^1 \left[\sum_{j=0}^{n_i} y_i^2 f(Y_i, \pi_i)\right] f(\pi_i)\, d\pi_i$$

Usando el valor del segundo momento para una v.a binomial con parametros (n_i, π_i), se llega a que:

$$E[Y_i^2] = \int_0^1 [n_i\pi_i(1-\pi_i) + n_i^2\pi_i^2] f(\pi_i)\, d\pi_i = \int_0^1 (n_i\pi_i - n_i\pi_i^2 + n_i^2\pi_i^2) f(\pi_i)\, d\pi_i$$

Aplicando los incisos (i) y (ii)

$$= n_i\left[\int_0^1 \pi_i f(\pi_i)\, d\pi_i\right] + (n_i^2 - n_i)\left[\int_0^1 \pi_i^2 f(\pi_i)\, d\pi_i\right] = n_i\theta + (n_i^2 - n_i)\left(\frac{(a+1)a}{(a+b+1)(a+b)}\right)$$

Usando esto y el inciso anterior:

$$Var(Y_i^2) = E[Y_i^2]-E[Y_i]^2 = n_i\theta+(n_i^2-n_i)\left(\frac{(a+1)a}{(a+b+1)(a+b)}\right)-(n_i\theta)^2 = n_i\theta\left[1+(n_i-1)\left(\frac{a+1}{a+b+1}\right)-n_i\theta\right]$$

$$= n_i\theta\left[1+(n_i-1)\left(\frac{a+1}{a+b+1}\right)-n_i\frac{a}{a+b}\right]$$

$$\Rightarrow Var(Y_i) = n_i\theta\left[\frac{(a+b)(a+b+1)}{(a+b)(a+b+1)}+(n_i-1)\left(\frac{(a+1)(a+b)}{(a+b+1)(a+b)}\right)-n_i\frac{a(a+b+1)}{(a+b)(a+b+1)}\right]$$

$$\Rightarrow Var(Y_i) = \left(\frac{n_i\theta}{a+b+1}\right)\left[\frac{(a^2+2ab+b^2+a+b)}{(a+b)}+(n_i-1)\left(\frac{a^2+ab+a+b}{(a+b)}\right)-n_i\frac{a^2+ab+a}{(a+b)}\right]$$

$$\Rightarrow Var(Y_i) = \left(\frac{n_i\theta}{a+b+1}\right)\left[\frac{a^2+2ab+b^2+a+b+n_i(a^2+ab+a+b)-(a^2+ab+a+b)-n_i(a^2+ab+a)}{a+b}\right.$$

$$\Rightarrow Var(Y_i) = \left(\frac{n_i\theta}{a+b+1}\right)\left[\frac{ab+b^2+n_ib}{a+b}\right] = \left(\frac{n_i\theta}{a+b+1}\right)\left[\left(\frac{b}{a+b}\right)(a+b+n_i)\right] = n_i\theta(1-\theta)\left(\frac{a+b+n_i}{a+b+1}\right)$$

$$\Rightarrow var(Y_i) = n_i\theta(1-\theta)\left(\frac{a+b+1}{a+b+1}+\frac{n_i-1}{a+b+1}\right) = n_i\theta(1-\theta)(1+(n_i-1)/\phi)$$

$\square$

8. En modelos de regresión ordinarios, uno de los supuestos es varianza constante $v(\mu_i)=\sigma^2$, sin embargo, suponga que la verdadera función varianza es $v(\mu_i)=\mu_i$. Sea el modelo nulo $\mu_i=\beta$, para $i=1,\dots,n$.

(i) Muestre que $u(\beta)=(1/\sigma^2)\sum_{i=1}^n(y_i-\beta)$, por lo que el estimador cuasi-verosímil de β es $\hat{\beta}=\bar{y}$.
(ii) Encuentre la varianza *naive* para $\hat{\beta}$.
(iii) Encuentre la varianza robusta para $\hat{\beta}$.
(i) *Demostración:* Sea $v(\mu_i)=\mu_i$ entonces:

$$u(\beta) = X'DV^{-1}(y_i-\mu_i) = \sum_{i=1}^{n}\frac{y_i-\beta}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\beta)$$

$\square$

(iii) *Demostración:*

$$\widehat{V}(\widehat{\beta}) = \widehat{G}_0^{-1}\,\widehat{G}_1\,\widehat{G}_0 = E\left[-\frac{\partial u(\widehat{\beta})}{\partial\widehat{\beta}}\right]^{-1}\widehat{G}_1\,E\left[-\frac{\partial u(\widehat{\beta})}{\partial\widehat{\beta}}\right] = E\left[-\frac{1}{\sigma^2}\right]^{-1}\widehat{G}_1\,E\left[-\frac{1}{\sigma^2}\right] = \widehat{G}_1$$

$\square$