Guía 3. Modelación estadística del número de garrapatas por oveja

(Para entregar el viernes 11 de octubre de 2019)

Como han visto ya, la función de probabilidad Binomial Negativa es:

$$\begin{aligned} f(x;\theta,\beta) &= P[X=x;\theta,\beta] = \binom{x+\theta-1}{x}(1-\beta)^{\theta}\beta^{x}, \\ &= \left(\frac{\Gamma(x+\theta)}{x!\Gamma(\theta)}\right)(1-\beta)^{\theta}\beta^{x}, \quad \text{para } x=0,1,2,... \end{aligned} \tag{1}$$

donde $\theta > 0$ es un número real positivo, y β es la probabilidad de éxito en un lanzamiento Bernoulli. Sir Ronald Fisher propuso usarla para describir el número de garrapatas en ovejas. En la definición habitual de la distribución binomial negativa, el valor de θ se define como un entero positivo que representa el número de fracasos fijos y la variable aleatoria X es el número de éxitos vistos antes de que ocurra el último de los fracasos. Esta representación es una generalización de esa presentación, al igual que el uso de la función Gama reemplazando la expresión de combinaciones.

Los datos de Fisher se dan en la siguiente tabla:

$x = \#$ de garrapatas	0	1	2	3	4	5	6	7	8	9	10	Total
$frec(x) = \#$ de ovejas	7	9	8	13	8	5	4	3	0	1	2	60

.

Es decir, hubo 13 ovejas que tuvieron 3 garrapatas cada una, tres ovejas con siete garrapatas, etc. Esta es una manera resumida de dar los 60 valores $x_1, ..., x_{60}$, aprovechando que varios de ellos toman el mismo valor, se pueden resumir en una tabla de datos agrupados. Toma en cuenta que debido al clima frío inglés, históricamente la industria de la lana es de suma importancia y muy desarrollada.

1. Da la expresión del valor esperado μ y de la varianza δ de la binomial negativa, en términos de los parámetros θ, β. Da la expresión de estimadores de momentos para estos parámetros.

2. Para una muestra de $X_1, ..., X_n$ observaciones binomiales negativas, da la expresión de la función de log verosimilitud en términos de θ, β, simplificándola lo más que puedas. Identifica a través de esta función el vector T de estadísticas suficientes e indica la dimensión que tiene. Recuerda que para los datos de Fisher, su dimensión es menor o igual a $n = 60$ y mayor o igual a 2 (el número de parámetros desconocidos en la distribución que se desea estimar). Demuestra que efectivamente T es suficiente para θ, β.

3. Da la expresión de la log verosimilitud $l(\theta, \beta; x_1, ..., x_{60})$ y de sus dos primeras derivadas parciales, que constituyen la función Score (el vector de primeras derivadas parciales). Di si es posible encontrar los estimadores de máxima verosimilitud (emv) de θ,β de manera analítica o si hay que hacerlo numéricamente. Al igualar a cero las primeras derivadas paraciales, se puede dar una solución de un parámetro en términos del otro. A esto se le llama estimador de máxima verosimilitud restringido y se denotan como $\hat{\theta}(\beta)$ o $\hat{\beta}(\theta)$. En este caso ¿cuál de los dos es el que se puede dar con una expresión analítica cerrada? (Para el otro parámetro no es posible hacerlo).

4. Encuentra y presenta los emv de θ y β. Grafica los contornos de la verosimilitud relativa $R(\theta, \beta)$ de niveles $\{1, 0.5, 0.3, 0.15, 0.1, 0.05, 0.01\}$. Recuerda que

$$R(\theta, \beta; x_1, ..., x_{60}) = \exp\left[l(\theta, \beta) - l\left(\hat{\theta}, \hat{\beta}\right)\right].$$

5. Aprovechando la propiedad de invarianza de la función de verosimilitud frente a reparametrizaciones uno a uno, da la expresión de la logverosimilitud relativa en términos de los parámetros (μ, δ). Grafica los contornos de $R(\mu, \delta)$ de mismos niveles que en el inciso anterior. Compara con los contornos que obtuviste en el inciso anterior y comenta sobre cuál reparametrización se ve más simétrica con respecto a los emv.

6. **LA VEROSIMILITUD PERFIL DE UN PARAMETRO DE INTERES.** Para distribuciones estadísticas que tienen más de un parámetro $\theta_1, ..., \theta_k$ (y menos de seis, $k < 6$), la verosimilitud perfil de un parámetro de interés θ_1 es una buena manera de expresar la información contenida en los datos sobre ese parámetro, de manera unidimensional y en ausencia de conocimiento sobre los parámetros restantes $(\theta_2, ..., \theta_k)$ a los cuales se les llama de estorbo o incidentales. La log verosimilitud perfil para una muestra de n observaciones $\vec{x}$ evaluada en un valor del parámetro de interés θ_1 denotada como $l_p(\theta_1)$, se calcula como el máximo de la log verosimilitud global sobre los parámetros restantes para ese valor fijo del parámetro de interés θ_1. Formalmente, se define como

$$\begin{aligned} l_p(\theta_1; \vec{x}) &= \max_{\theta_2,...,\theta_k | \theta_1} l(\theta_1, \theta_2, ..., \theta_k; \vec{x}) \\ &= l\left[\theta_1, \hat{\theta}_2(\theta_1), \hat{\theta}_3(\theta_1), ..., \hat{\theta}_k(\theta_1); \vec{x}\right], \end{aligned}$$

donde $\hat{\theta}_j(\theta_1)$ es el estimador de máxima verosimilitud restringido del parámetro de estorbo θ_j como función del parámetro de interés θ_1. Es decir, para cada valor fijo del parámetro de interés θ_1 se maximiza sobre los parámetros restantes la logverosimilitud. Para el caso binomial negativo, se puede encontrar una expresión analítica cerrada para uno de los dos parámetros θ, pero no para el otro β, porque implica derivar una función Gama. Da la expresión de la logverosimilitud perfil de θ y grafícala para los datos de ovejas. La verosimilitud perfil de β se debe encontrar maximizando numéricamente sobre θ para cada valor fijo de β. Para mayores detalles sobre la verosimilitud perfil, revisa el Capítulo 3 del libro de Pawitan y la Sección 10.3 del libro de Kalbfleisch.

7. Sobre la gráfica de contornos de la verosimilitud relativa $R(\theta, \beta; x_1, ..., x_{60})$ marca los puntos de coordenadas

$$\left[\theta, \hat{\beta}(\theta)\right]$$

y únelos con una linea sólida. Esta trayectoria corresponde a los puntos donde se evalúa la logverosimilitud $l(\theta, \beta)$ para calcular la log verosimilitud perfil $l_p(\theta)$. Marca además sobre tus contornos con un asterisco el emv $\left(\hat{\theta}, \hat{\beta}\right)$.

8. Calcula numéricamente el intervalo de verosimilitud perfil de nivel $c = 0.1465$ para θ y grafica la verosimilitud relativa perfil de θ, $R_p(\theta; \vec{x})$ marcando este intervalo sobre su curva. Nota que se obtiene de cortar horizontalmente a $R_p(\theta; \vec{x})$ a la altura $c = 0.1465$, recordando que el intervalo de verosimilitud perfil de nivel c se define como

$$IV(c) = \{\theta : R_p(\theta; \vec{x})\}.$$

9. Grafica la función de probabilidad binomial negativa estimada con los emv, graficando con un asterisco la probabilidad de que $X = x$ para x en $\{0, 1, ..., 20\}$.

10. Ahora realiza una gráfica donde en el eje vertical se grafique el número de ovejas y en el horizontal el número de garrapatas x que le corresponden, para x en $\{0, 1, ..., 20\}$. Compara los valores observados para cada x, marcándolos con asteriscos, contra los valores esperados bajo el modelo estimado los cuales serían

$$n\ P\left[X = x; \hat{\theta}, \hat{\beta}\right] = 60\ P\left[X = x; \hat{\theta}, \hat{\beta}\right].$$

Es decir, se compararán los puntos observados, indicados con asterisco $[x, frec(x)]$ contra los puntos estimados con el modelo, indicados con el símbolo +, $\left(x, n\ P\left[X = x; \hat{\theta}, \hat{\beta}\right]\right)$.

11. Grafica la función de distribución binomial negativa estimada para estos datos de ovejas, para $0 \leq x \leq 20$. Indica en la misma gráfica la función de distribución empírica, $F_n(x)$.

12. Si el modelo estima bien a los datos observados, el histograma de frecuencias estimado debe ser cercano al observado. También la función de distribución estimada debe estar cercana a la empírica. ¿Así ocurre con estos datos?

13. Con base en los emv de θ, β, calcula la probabilidad estimada de que $X \geq 10$, que una oveja tenga 10 o más garrapatas. Di también cuál es la probabilidad estimada de que no tenga garrapatas?

14. Con base en tus resultados de los incisos anteriores, di si el modelo Binomial Negativo describe bien a estos datos. ¿Le recomendarías al granjero buscar a un fumigador que de tratamiento a las ovejas? Argumenta tu respuesta.

REFERENCIA:

Fisher, R. A. (1941). the Negative Binomial Distribution. Annals of Eugenics, V. 6 , 391-398. [La biblioteca del CIMAT tiene cinco libros con toda la obra de Fisher y allí encontrarán este artículo].