Estadística

Gianpaolo Luciano Rivera

2016

1. Conceptos Generales

- **Muestra Aleatoria**: Una muestra es cualquier conjunto de *variables aleatorias idénticamente distribuidas (v.a. iid)* sobre las que se hará inferencia. Representan un subconjunto de la población sobre la que se desea conocer un *parámetro* en especifico.
- **Datos observados**: son mi conjunto de v.a. $\mathbf{y} = (y_1, \ldots, y_n)$ usualmente independientes entre ellas.
- **Tamaño de la muestra**: Número de observaciones registradas n.
- **Parámetros**: Constantes[1] $\theta = (\theta_1, \ldots, \theta_k)$ conocidas o desconocidas que permiten conocer alguna característica de la población.
- **Espacio paramétrico**: Conjunto Θ al que pertenecen los parámetros. $\theta \in \Theta$

.

2. Inferencia Estadistica

Dado nuestros datos, nos gustaría poder sacar o sintetizar la información que estos contienen sobre el fenómeno que estemos estudiando. Para ello diseñamos modelos estadísticos que logran describir con cierta precision la realidad. El principio básico es que queremos estimar o hacer inferencia en algún *parámetro* de interés y ver si los datos lo respaldan y si es estadisticamente significativo.

- **Verosimilitud**: Función de distribución (teórica) de mis datos y parámetros. $L(\theta; \mathbf{y}) = f_{Y_1,\ldots,Y_n}(\mathbf{y}; \theta)$. Y en caso de que las observaciones sean independientes entre sí

$$L(\theta; \mathbf{y}) = f_{Y_1,\ldots,Y_n}(\mathbf{y}; \theta) = \prod_{i=1}^{n} f_{Y_i}(y_i; \theta)$$

- **Log-Verosimilitud**: A veces es más fácil trabajar con el logaritmo de la verosimilitud. $l(\theta; \mathbf{y}) = \log L(\theta; \mathbf{y})$ y en caso de independencia: $l(\theta; \mathbf{y}) = \sum_i^n \log[f_{Y_i}(y_i; \theta)]$.
- **Cociente de Verosimilitud**: Usado para normalizar la verosimilitud en su valor máximo $\hat{\theta}_{\text{MLE}}$ (Estimador de Máxima Verosimilitud). $L(\theta)/L(\hat{\theta})$
- **Desviación** Herramienta similar al cociente de verosimilitudes en escala logarítmica: $D(\hat{\theta}, \theta) = 2[l(\hat{\theta}) - l(\theta)]$
- **Estadístico**: Es una función $\mathbf{t} : \mathbb{R}^n \to \mathbb{R}^k$ de la muestra $\mathbf{t}(\mathbf{y})$ que usualmente "condensa" la información.[2]
- **Estadístico de Máxima Verosimilitud**: Es la variable que maximiza la verosimilitud: $\hat{\theta}_{\text{MLE}} = \hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} l(\theta)$ Hay muchas maneras de encontrarlo pero usualmente se maximiza de forma normal la función de verosimilitud.

[1] En estadística clásica

[2] La dimension de $\mathbf{t}$ puede ser mayor que la de θ sin embargo nunca puede ser menor pues si queremos usar $\mathbf{t}$ para hacer inferencia sobre θ nos haría falta información. Usualmente esperamos que sus dimensiones sean iguales.

- **Invarianza**: Si $\hat{\theta}$ es MLE y $g(\theta)$ es una función continua. Entonces $g(\hat{\theta})$ es MLE en la nueva reparametrización.

- **Suficiencia**: Un estadístico $\mathbf{t} = \mathbf{t}(\mathbf{y})$ es suficiente para el parámetro $\theta \in \Theta$ si la distribución condicional de mis datos $\mathbf{y}$ dado $\mathbf{t}$ es libre de θ. En otras palabras, cualquier resultado de $\mathbf{y}$ que tenga el mismo $\mathbf{t}(\mathbf{y}) = \mathbf{t}$ nos debe llevar a la misma conclusión sobre θ.

$$\text{Si } \mathbf{t}(\mathbf{y}) = \mathbf{t}(\mathbf{z}) \iff \frac{L(\mathbf{y};\theta)}{L(\mathbf{z};\theta)} \text{ no depende de } \theta \Rightarrow \mathbf{t} \text{ es suficiente}$$

- **Criterio de Factorización**: Un estadistico $\mathbf{t} = \mathbf{t}(\mathbf{y})$ es suficiente para θ si y solo si, podemos factorizar la densidad como:

$$f(\mathbf{y};\theta) = g(\mathbf{t}(\mathbf{y});\theta)h(\mathbf{y})$$

- **Suficiencia mínima** Como la representación de $\mathbf{t}(\mathbf{y})$ no es única, nos interesa la más pequeña posible. Un estadistico suficiente $\mathbf{t}(\mathbf{y})$ es **mínimo suficiente** si es una función de cualquier otro estadístico suficiente. Ie, tiene dimensión mínima

- **Parámetros independientes a variaciones**: Los parámetros componentes de $\theta = (\psi, \lambda)$ son independientes a variaciones si el espacio parametrico Θ es el producto cartesiano de Ψ y de Λ respectivamente. Ie: $\Theta = \Psi \times \Lambda$. Y significa que el conjunto de valores de ψ es el mismo para cada λ fija.

- **Ortogonalidad de la Verosimilitud**: Una parametrización dada: $\theta = (\psi, \lambda)$ donde los parametros son independientes a variaciones, se llama ortogonal en la verosimilitud si esta se puede factorizar: $L(\theta, \mathbf{y}) = L_1(\psi, t_1(\mathbf{y}))L_2(\lambda, t_2(\mathbf{y}))$

2.1. Familias Exponenciales

2.1.1. Regularidad y Propiedades Analíticas

Muchas distribuciones tienen características similares y se les llama **Familia Exponencial**. Haciendo esta generalización llegamos a muchas propiedades que se cumplen para todas ellas.
Un modelo estadístico para una muestra $\mathbf{y}$ es una familia exponencial con **parámetro canónico** $\theta = (\theta_1, \ldots, \theta_k)$ y **estadístico canónico** $\mathbf{t}(\mathbf{y}) = (t_1(\mathbf{y}), \ldots, t_k(\mathbf{y}))$ si su densidad (o verosimilitud) f tiene la forma

$$L(\theta;\mathbf{y}) = f(\mathbf{y};\theta) = a(\theta)h(\mathbf{y})e^{\theta^T \mathbf{t}(\mathbf{y})}$$

donde $\theta^T \mathbf{t}(\mathbf{y})$ es el producto punto de los dos vectores de tamaño k, y donde $a(\theta)$ y $h(\mathbf{y})$ son dos funciones medibles. Se define la **constante normalizadora** como la estructura que hace que la densidad integre a 1:

$$C(\theta) = \frac{1}{a(\theta)} = \int h(\mathbf{y})e^{\theta^T \mathbf{t}(\mathbf{y})}\, d\,\mathbf{y}$$

- Si $\mathbf{y}$ tiene una distribución exponencial, entonces la distribución de $\mathbf{t}$ es también exponencial y bajo ciertas condiciones de regularidad en $\mathbf{t}(\mathbf{y})$ la distribución de $\mathbf{t}$ es:

$$f(\mathbf{t};\theta) = a(\theta)g(\mathbf{t})e^{\theta^T \mathbf{t}}$$

  donde la **función de estructura**:

$$g(\mathbf{t}) = \sum_{\mathbf{t}(\mathbf{y})=\mathbf{t}} h(\mathbf{y}) = \int_{\mathbf{t}(\mathbf{y})=\mathbf{t}} h(\mathbf{y})\, d\,\mathbf{y}$$

- El **orden** de una familia exponencial es la dimensión más baja del estadístico canónico para la cual la familia de distribuciones se puede representar. La representación es **minima** si el estadístico canónico $\mathbf{t}$ tiene esa dimensión. Para revisar esto, solo hay que ver que no haya dependencia linear entre los elementos de $\mathbf{t}$ y θ.

- Una familia exponencial mininima es llamada **llena** si su espacio paramétrico es *maximal*, ie: es igual al espacio canónico Θ.

- Si $\mathbf{y}$ es exponencial y su representación es mínima, entonces el estadístico $\mathbf{t} = \mathbf{t}(\mathbf{y})$ es **mínimo suficiente** para θ.
- El espacio paramétrico Θ es un *conjunto convexo* en $\mathbb{R}^k$.
- **Familias regulares**: Una familia exponencial de orden k es llamada **regular** si el espacio paramétrico Θ es un conjunto abierto en $\mathbb{R}^k$. Nota: si solo hay un conjunto finito de valores de t entonces una familia llena es necesariamente regular y el espacio parámetro canónico $\Theta = R^{\dim t}$
- **Propiedades de** $C(\theta)$.
  - $C(\theta)$ es estrictamente convexa
  - Para una familia reglar arbitrariamente diferenciable tenemos el **vector de medias** y la **matriz de covarianza**:

$$\nabla \log C(\theta) = \mathrm{E}_\theta[\mathbf{t}] = \mu_t(\theta)$$
$$\nabla^2 \log C(\theta) = \mathrm{Var}_\theta[\mathbf{t}] = V_t(\theta)$$

  - **Momentos de orden mayor**:

$$\mathrm{E}_\theta[e^{\psi^T \mathbf{t}}] = \frac{C(\theta+\psi)}{C(\theta)}$$
$$\mathrm{E}_\theta[t_j^r] = \frac{\delta^r C(\theta)}{\delta \theta_j^r}\frac{1}{C(\theta)}$$

2.1.2. Verosimilitud y Máxima Verosimilitud

La función de *log-verosimilitud* para una familia exponencial

$$\log L(\theta) = \theta^T \mathbf{t} - \log C(\theta) + h(\mathbf{y})$$

con $h(\mathbf{y})$ una contante es una función suave y estrictamente cóncava. La función de **marcador** es el gradiente de esta

$$U(\theta) = \nabla \log L(\theta) = \mathbf{t} - \mu_t(\theta)$$

que si se iguala a cero y resolvemos el sistema para θ encontramos el $\hat{\theta}_{\mathrm{MLE}}$. Derivando nuevamente y cambiando el signo obtenemos la **información observada**:

$$J(\theta) = -\nabla U(\theta) = -\nabla^2 \log L(\theta)$$

que en $\hat{\theta}$ debe ser una cantidad positiva o en su defecto una matriz positiva definida. La **información de Fisher o información esperada** es el valor esperado de la información observada.

$$I(\theta) = \mathrm{E}_\theta[J(\theta)]$$

En particular para modelos exponenciales[3] tenemos:

$$I(\theta) = J(\theta) = V_t(\theta)$$

- $\mathrm{E}[U(\theta, \mathbf{y})] = 0$
- En una familia regular exponencial $\mu_t(\theta)$ es una función uno a uno de θ para $\theta \in \Theta$ con $\mu_t(\Theta)$ también abierto. La función de log verosimilitud tiene un máximo único en Θ si y solo si $\mathbf{t} \in \mu_t(\Theta)$ y entonces el $\hat{\theta}_{\mathrm{MLE}}(\mathbf{t})$ es la raíz única de la **ecuación de verosimilitud**: $\mu_{\mathrm{t}}(\theta) = \mathbf{t}$.
- $\hat{\theta}_{\mathrm{MLE}}(\mathbf{t}) = \mu_t^{-1}(\mathbf{t})$

[3]Esto no es cierto para otro tipo de distribuciones, sin embargo siempre tenemos que *para la parametrización canónica* $J(\hat{\theta}) = I(\hat{\theta})$

2.1.3. Parametrizaciones alternativas

A veces es más conveniente usar otra parametrización, es decir, poner todo en función de otra variable $\psi = \psi(\theta)$ o dada su inversa.

- **Lema de Reparametrización**: Si ψ o $\theta = \theta(\psi)$ son dos parametrizaciones equivaelentes del mismo modelo entonces tenemos que las funciones de marcador están relacionadas:

$$U_\psi(\psi; \mathbf{y}) = \left(\frac{\delta\theta}{\delta\psi}\right)^T U_\theta(\theta(\psi); \mathbf{y})$$

  . De manera análoga las correspondientes matrices de información:

$$I_\psi(\psi) = \left(\frac{\delta\theta}{\delta\psi}\right)^T I_\theta(\theta(\psi)) \left(\frac{\delta\theta}{\delta\psi}\right)$$
$$J_\psi(\hat{\psi}) = \left(\frac{\delta\theta}{\delta\psi}\right)^T J_\theta(\theta(\hat{\psi})) \left(\frac{\delta\theta}{\delta\psi}\right)$$

- **Parametrización de Valor Medio**: el parámetro de valor medio es $\mu_t(\theta)$ y es estimado insesgadamente por su MLE $\hat{\mu_t} = \mathbf{t}$. Su varianza es $\mathrm{Var}[\hat{\mu_t}] = \mathrm{Var}[\mathbf{t}]$ que es la inversa de la matriz de información de Fisher bajo parametrización con μ_t.

$$I(\mu_t) = \mathrm{Var}(\mathbf{t})^{-1}$$

  y en el MLE se cumple $J(\hat{\mu_t}) = I(\hat{\mu_t})$

- **Parametrización mixta**: Se usa cuando queremos hacer inferencia condicional. Si tenemos nuestro estadístico canónico particionado en: $\mathbf{t} = (u, v)^T$ y el correspondiente parámetro canónico: $\theta = (\theta_u, \theta_v)^T$ respectivamente entonces:

  - El modelo marginal para u es una familia regular exponencial para cada θ_v dada, dependiendo de θ_v pero con uno y el mismo espacio paramento para su parámetro de valor medio μ_u
  - E modelo condicional para $\mathbf{y}$ dado u y por lo tanto para v dado u es una familia regular exponencial con estadístico canónico v. El modelo condicional depende de u pero con uno y el mismo espacio para métrico θ_v como en el modelo conjunto.
  - Si una familia regular exponencial es dada una parametrización mixta (μ_u, θ_v) los dos componentes de los parámetros son independientes a variaciones y ortogonales en información. Además la info de Fisher esta dada por:

$$I(\mu_u, \theta_v) = \begin{bmatrix} \mathrm{Var}[u]^{-1} & 0 \\ 0 & (Var[t]^{vv})^{-1} \end{bmatrix}$$

- **Inferencia condicional para el parámetro canónico**: Supongamos que tenemos un parámentro de interés ψ que se puede expresar linearmente en el parámetro canónico θ y otro que es una molestia[4] λ o $\mu_u = \mathrm{E}_\theta[u]$ que es *independiente a variaciones de* ψ, ie: $\theta = (\psi, \lambda)$. Después de hacer la transformación linear correspondiente a mi estadistico t, tenemos que $t = (v, u)$.

  - **Principio de Condicionalidad**: Inferencia estadistica sobre el parámetro de interés ψ en presencia de la molestia deberá ser condicionando en u. Esto es, la distribución condicional de y o v dado u.
  - La verosimilitud para (ψ, μ_u) se factoriza:

$$L(\theta, \mathbf{t}) = L_1(\psi, v|u) L_2(\mu_u, \psi; u)$$

  - En algunos casos, L_2 solo depende de μ_u y la situación es llamada **S-auxilliar**[5] y se dice que hay un **corte**.

[4] nuisance
[5] S-ancillarity

- **Perfil de Verosimilitud** Cuando tenemos nuestro parámetro de interés ψ y la molestia λ. Encontramos $\hat{\lambda}(\psi)$ y construimos el *perfil de verosimilitud* $L_p(\psi) = L(\psi, \hat{\lambda}(\psi))$ para dejar todo en una sola variable. Esta función tiene la curvartura correcta para expresar información sobre ψ en comparación contra $L(\psi, \hat{\lambda})$. Esta curvatura está dada por: $-\nabla^2 \log L_p(\psi) = J_{\psi\psi} - J_{\psi\lambda} J_{\lambda\lambda}^{-1} J_{\lambda\psi}$ con las matrices calculadas en $(\psi, \hat{\lambda}(\psi))$

2.1.4. Completitud y Teorema de Basu

- **Completitud**: Un estadístico $\mathbf{t}$ es completo (o la familia de distribuciones del estadístico $\mathbf{t}$ es completa) si la propiedad.

$$\mathrm{E}[h(\mathbf{t})] = 0 \quad \forall \theta \in \Theta$$

  para alguna función $h(\mathbf{t})$, requiere que la función $h(\mathbf{t})$ sea cero. Que un estadístico sea completo significa que el modelo parametrico es suficientemente rico para alcanzar la dimensión de este, ie: son las mismas.

- En una familia exponencial llena, el estadístico canónico $\mathbf{t} = \mathbf{t}(\mathbf{y})$ es completo.

- **Teorema de Basu**: Si $\mathbf{t}$ es un estadístico suficiente y completo, tipicamente siendo el canónico en una familia exponencial llena, sea $\mathbf{u}$ otro estadístico, entonces $\mathbf{u}$ y $\mathbf{t}$ son mutuamente independientes precisamente cuando la distribución de $\mathbf{u}$ sea libre de parámetros.

2.2. Propiedades Asintoticas del MLE

2.2.1. Asintotas para muestras grandes

- **Existencia y consistencia del MLE de** θ: para una muestra de tamaño n de una familia regular exponencial y para cada $\theta \in \Theta$ la $P(\hat{\theta}(\mathbf{t}_n/n) \text{ exista}; \theta) \to 1 \quad \text{cuando} \quad n \to \infty$.Además $\hat{\theta}(\mathbf{t}_n/n) \to \theta$ en probabilidad si $n \to \infty$. Ambas convergencias son uniformes en subconjuntos compactos de Θ.

- Nota: el resultado aplica para cualquier otra reparametrización.

- **Normalidad asintotica del MLE**: Para una muestra de tamaño n de una familia regular exponencial, y para cualquier función del parámetro $\eta = \eta(\mu_{\mathbf{t}})$, el MLE $\hat{\eta}$ es asintoticamente normal si $n \to \infty$, ie:

$$\sqrt{n}(\hat{\eta} - \eta) \sim N\left(0, \left(\frac{\delta\eta}{\delta\mu_{\mathbf{t}}}\right)^T \mathrm{Var}(\mathbf{t}) \left(\frac{\delta\eta}{\delta\mu_{\mathbf{t}}}\right)\right)$$

  donde $\mathrm{Var}(\mathbf{t})$ es la matriz de covarianza para $\mathbf{t}$ expresado en una parametrizacion correcta y estimada en el MLE.Para el parámetro canonico tenemos que:

$$\sqrt{n}(\hat{\theta} - \theta) \sim N(0, V_t(\theta)^{-1})$$

- La familia regular exponencial implica que el MLE es unico y es un maximo global.

2.2.2. Aproximaciones de Punto-Silla

Técnicas para aproximar densidades cuando la normal (o χ^2) es demasiado cruda.

- **Aproximiación de Punto-Sila**: La aproximación de una densidad $f(t)$ en una familia exponencial es:

$$f(t; \theta_0) \approx \frac{1}{(2\pi)^k/2\sqrt{\det V_t(\hat{\theta}(t))}} \frac{C(\hat{\theta}(t)}{C(\theta_0)} e^{(\theta_0 - \hat{\theta}(t))^T t}$$

  con la función de estructura correspondiente:

$$g(t) \approx \frac{1}{(2\pi)^k/2\sqrt{\det V_t(\hat{\theta}(t))}} C(\hat{\theta}(t)) e^{\hat{\theta}(t)^T t}$$

- **Aproximiación de Punto-Sila para el MLE**: La aproximación de a densidad $f(hat\psi; \psi_0)$ de un MLE $\hat{\psi} = \hat{\psi}(t)$ en cualquier parametrización suave de una familia regular exponencial es:

$$f(\hat{\psi}; \psi_0) \approx \frac{\sqrt{\det I(\hat{\psi})}}{(2\pi)^{k/2}} \frac{L(\psi_0)}{L(\hat{\psi})}$$

  con $I(\hat{\psi})$ es la matriz de información de Fisher para el parametro ψ evaluada en el MLE.

2.3. Pruebas para reducir modelos

2.3.1. Pruebas Asintoticamente equivalentes

Sea $\theta = (\psi, \lambda)^T$, bajo mi hipótesis nula H_0: $\psi = \psi_0$ quiero probar si puedo reducir la dimensionalidad de mi modelo $\dim(\psi) + \dim(\lambda)$ a $\dim(\lambda)$. Tengo cuatro pruebas que bajo condiciones de regularidad y si $n \to \infty$ se distribuyen: $\chi^2(\dim \psi)$

- *Desviación*: $W = 2[l(\hat{\theta}) - l(\hat{\theta}_0)]$
- *Prueba de Wald*: $W_e = (\hat{\psi} - \psi_0)^T (I^{\psi\psi})^{-1}(\hat{\psi})(\hat{\psi} - \psi_0) =$
- *Prueba de Marcador de Rao*: $W_u = U(\hat{\theta}_0)^T I(\hat{\theta}_0)^{-1} U(\hat{\theta}_0)$
- *Forma Cuadrática*: $W_e^* = (\hat{\theta}_0 - \hat{\theta})^T I(\hat{\theta}_0)(\hat{\theta}_0 - \hat{\theta})$

3. Notas adicionales

- **Matriz Jacobiana**: Sean y y x funciones vectoriales se define su derivada como:

$$\frac{dy}{dx^T} = \begin{bmatrix} \frac{\delta y_1}{\delta x^T} \\ \vdots \\ \frac{\delta y_n}{\delta x^T} \end{bmatrix} = \begin{bmatrix} \left(\frac{\delta y_1}{\delta x}\right)^T \\ \vdots \\ \left(\frac{\delta y_n}{\delta x}\right)^T \end{bmatrix} = \frac{dy^T}{dx}$$