Contrastes de hipótesis y selección de modelos

	H_0 es verdadera	H_0 es falsa
No rechazar H_0	1. Decisión correcta Nivel de confianza Probabilidad $p = 1 - \alpha$	3. Error tipo II Probabilidad $p = \beta$
Rechazar H_0	2. Error tipo I Nivel de significación Probabilidad $p = \alpha$	4. Decisión correcta Poder de prueba Probabilidad $p = 1 - \beta$

Mike Wiper
Departamento de Estadística
Universidad Carlos III de Madrid

Grado en Estadística y Empresa

Objetivo

Bayes factor BF_{12}			Interpretation
	>	100	Extreme evidence for M_1
30	-	100	Very Strong evidence for M_1
10	-	30	Strong evidence for M_1
3	-	10	Moderate evidence for M_1
1	-	3	Anecdotal evidence for M_1
	1		No evidence
1/3	-	1	Anecdotal evidence for M_2
1/10	-	1/3	Moderate evidence for M_2
1/30	-	1/10	Strong evidence for M_2
1/100	-	1/30	Very Strong evidence for M_2
	<	1/100	Extreme evidence for M_2

Ilustrar como hacer contrastes de hipótesis y selección de modelos desde el punto de vista bayesiano.

Contrastes de hipótesis

En principio, hacer contrastes a la bayesiana es fácil. Dadas las distribuciones a priori, $P(H_0)$ y $P(H_1) = 1 - P(H_0)$, se calculan las probabilidades a posteriori:

$$P(H_0|\text{datos}) \propto f(\text{datos}|H_0)P(H_0) \propto P(H_0) \int f(\text{datos}|H_0, \theta) f(\theta|H_0)\, d\theta$$

Luego, dada una *regla de decisión*, se decide rechazar H_0 o no.

Contrastes unilaterales

Para contrastes unilaterales del tipo $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$, podemos definir directamente una distribución a priori $f(\theta) \, \forall \, \theta$.

Entonces, implicitamente:

$$P(H_0) = \int_{-\infty}^{\theta_0} f(\theta) \, d\theta.$$

Dada la muestra, tenemos una distribución a posteriori $f(\theta|\text{datos})$ y luego:

$$P(H_0|\text{datos}) = \int_{-\infty}^{\theta_0} f(\theta|\text{datos}) \, d\theta.$$

A menudo, los resultados son parecidas a los análisis frecuentistas.

Ejemplo: contrastando la media de una normal

Sea $Y|\mu \sim \text{Normal}\left(\mu, \sigma^2\right)$. Queremos contrastar $H_0 : \mu \leq 0$ frente a $H_1 : \mu > 0$.

Dada la distribución a priori $\mu \sim \text{Normal}(m, v)$. Luego, $P(H_0) = \Phi\left(-\frac{m}{\sqrt{v}}\right)$.

Dada la muestra, la distribución a posteriori es

$$\mu|\text{datos} \sim \text{Normal}\left(\frac{m/v + n\bar{y}/\sigma^2}{1/v + n/\sigma^2}, \frac{1}{1/v + n/\sigma^2}\right)$$

y entonces,

$$P(H_0|\text{datos}) = P(\mu \leq 0|\text{datos}) = \Phi\left(-\frac{m/v + n\bar{y}/\sigma^2}{\sqrt{1/v + n/\sigma^2}}\right).$$

Cuando $v \to \infty$, observamos que a priori, $P(H_0) \to \frac{1}{2}$, y a posteriori:

$$P(H_0|\text{datos}) \to \Phi\left(-\frac{\sqrt{n}\bar{y}}{\sigma}\right).$$

que es el p-valor frecuentista para esta contraste.

Contrastes bilaterales

Para contrastes bilaterales del tipo $H_0 : \theta = \theta_0$ frente a $H_1 : \theta \neq \theta_0$, tenemos que definir probabilidades a priori $P(H_0)$ y $P(H_1)$, y, además definir una distribución a priori para θ dada H_1, $f(\theta|H_1)$.

Dada la muestra, tenemos una distribución a posteriori, $f(\theta|\text{datos})$ y luego:

$$P(H_0|\text{datos}) = \frac{P(H_0)f(\text{datos}|\theta_0)}{f(\text{datos})}.$$

El denominador es:

$$f(\text{datos}) = P(H_0)f(\text{datos}|\theta_0) + P(H_1)\int f(\text{datos}|\theta)f(\theta|H_1)\,d\theta.$$

Los resultados bayesianos y frecuentistas pueden ser muy diferentes.

Ejemplo: contraste bilateral para la probabilidad de cruz

Sea $\theta = P(\text{cruz})$ y supongamos que queremos contrastar $H_0 : \theta = \frac{1}{2}$ frente a $H_1 : \theta \neq \frac{1}{2}$.

Bajo H_1 es necesario fijar una distribución a priori para θ. Supongamos $\theta | H_1 \sim \text{Uniforme}(0,1) \sim \text{Beta}(1,1)$.

Supongamos que observamos 49581 cruces y 48870 caras en un experimento binomial.

Luego, $f(\text{datos}|H_0) = \begin{pmatrix} 98451 \\ 49581 \end{pmatrix} 0{,}5^{98451} \approx 1{,}95 \times 10^{-4}$.

Ejemplo: contraste bilateral para la probabilidad de cruz

Bajo H_1, tenemos:

$$\begin{aligned} f(\text{datos}|H_1) &= \int_0^1 f(\text{datos}|\theta, H_1) f(\theta|H_1)\, d\theta \\ &= \int_0^1 \binom{98451}{49581} \theta^{49581}(1-\theta)^{48870} \times 1\, d\theta \\ &= \binom{98451}{49581} B(49582, 48871) \\ &\approx 1{,}02 \times 10^{-5} \end{aligned}$$

Entonces, tenemos:

$$\begin{aligned} P(H_0|\text{datos}) &= \frac{P(H_0) f(\text{datos}|H_0)}{P(H_0) f(\text{datos}|H_0) + P(H_1) f(\text{datos}|H_1)} \\ &= \frac{\frac{1}{2} \times 1{,}9510^{-4}}{\frac{1}{2} \times 1{,}9510^{-4} + \frac{1}{2} \times 1{,}0210^{-5}} \approx 0{,}95 \end{aligned}$$

Ejemplo: contraste bilateral para la probabilidad de cruz

Desde el enfoque bayesiano, hay mucha evidencia a favor de H_0.

¿Qué pasa si hacemos el contraste frecuentista?

Ejemplo: contraste bilateral para la probabilidad de cruz

Desde el enfoque bayesiano, hay mucha evidencia a favor de H_0.

¿Qué pasa si hacemos el contraste frecuentista?

Usando una aproximación normal, bajo H_0,

$$\begin{aligned} \text{cruces}|H_0 &\sim \text{Normal}(98451 \times 0{,}5, 98451 \times 0{,}5 \times (1 - 0{,}5)) \\ &\sim \text{Normal}(49225{,}5, 24621{,}75) \end{aligned}$$

y luego, el p-valor es igual a $2P(\text{cruces} \geq 49581|H_0) \approx 0{,}0235$.

¡Rechazamos H_0 a un nivel de 5 % de significación!

¡Un resultado paradojico!

Comparación de modelos

En principio, podemos emplear las mismas técnicas comparación o selección de modelos.

Entre modelos $\mathcal{M}_1, ..., \mathcal{M}_N$, fijamos probabilidades a priori $P(\mathcal{M}_i)$ y dada la muestra, calculamos la probabilidad a posteriori $P(\mathcal{M}_i|\text{datos})$ para $i = 1, ..., N$.

En la práctica no es tan fácil.

Comparación de modelos

En principio, podemos emplear las mismas técnicas comparación o selección de modelos.

Entre modelos $\mathcal{M}_1, ..., \mathcal{M}_N$, fijamos probabilidades a priori $P(\mathcal{M}_i)$ y dada la muestra, calculamos la probabilidad a posteriori $P(\mathcal{M}_i|\text{datos})$ para $i = 1, ..., N$.

En la práctica no es tan fácil.

Por ejemplo en un modelo de regresión múltiple con 6 posibles regresores, tenemos $2^6 = 64$ modelos posibles. ¿Cómo podemos fijar las probabilidades a priori para cada modelo?

Buscamos un criterio que no depende de las probabilidades $P(\mathcal{M}_i)$.

El factor Bayes

Para dos modelos, $\mathcal{M}_0$ y $\mathcal{M}_1$, el factor Bayes a favor del modelo $\mathcal{M}_1$ es:

$$B_0^1 = \frac{P(\mathcal{M}_1|\text{datos})}{P(\mathcal{M}_0|\text{datos})} \frac{P(\mathcal{M}_0)}{P(\mathcal{M}_1)}$$

es decir la razón de las posibilidades ("odds") a posteriori partido por la razón de las posibilidades a priori.

Si el modelo 1 es correcto, entonces, cuando el número de datos en la muestra crece, $P((\mathcal{M}_1|\text{datos}) \to \infty$ y luego $B_0^1 \to \infty$. Si el modelo 0 es correcto, $B_0^1 \to 0$.

Interpretando el factor Bayes

Bayes factor BF_{12}			Interpretation
	>	100	Extreme evidence for M_1
30	-	100	Very Strong evidence for M_1
10	-	30	Strong evidence for M_1
3	-	10	Moderate evidence for M_1
1	-	3	Anecdotal evidence for M_1
	1		No evidence
1/3	-	1	Anecdotal evidence for M_2
1/10	-	1/3	Moderate evidence for M_2
1/30	-	1/10	Strong evidence for M_2
1/100	-	1/30	Very Strong evidence for M_2
	<	1/100	Extreme evidence for M_2

Distintos valores del factor Bayes muestran distintos niveles de evidencias a favor o en contra de $\mathcal{M}_1$.

El factor Bayes y la razón de verosimilitudes

Supongamos que queremos comparar dos modelos, $\mathcal{M}_0$ y $\mathcal{M}_1$, sin parámetros.

Entonces, por el teorema de Bayes:

$$P(\mathcal{M}_1|\text{datos}) = \frac{P(\mathcal{M}_1)f(\text{datos}|\mathcal{M}_1)}{f(\text{datos})}$$

y luego,

$$\begin{aligned} B_0^1 &= \frac{\frac{P(\mathcal{M}_1)f(\text{datos}|\mathcal{M}_1)}{f(\text{datos})}}{\frac{P(\mathcal{M}_0)f(\text{datos}|\mathcal{M}_0)}{f(\text{datos})}} \frac{P(\mathcal{M}_0)}{P(\mathcal{M}_1)} \\ &= \frac{f(\text{datos}|\mathcal{M}_1)}{f(\text{datos}|\mathcal{M}_0)} \end{aligned}$$

que es igual a la razón de verosimilitudes, y no depende de las probabilidades a priori.

Ejemplo

Supongamos que tiramos una moneda 12 veces y observamos 9 cruces. Sea $\theta = P(\text{cruz})$. Queremos comparar los hipótesis $H_0 : \theta = 0{,}5$ y $H_1 : \theta = 0{,}75$.

El factor Bayes a favor de H_1 es

$$B_0^1 = \frac{\binom{12}{9} 0{,}75^9 0{,}25^3}{\binom{12}{9} 0{,}5^{12}} \approx 4{,}8.$$

Evidencia moderada a favor de H_1.

¿El factor Bayes es igual a la razón de verosimilitudes frecuentista?

Si $\mathcal{M}$ tiene parámetros $\boldsymbol{\theta}$, entonces, la verosimilitud marginal bajo $\mathcal{M}$ es:

$$f(\text{datos}|\mathcal{M}) = \int f(\text{datos}|\boldsymbol{\theta}, \mathcal{M}) f(\boldsymbol{\theta}|\mathcal{M})\, d\theta$$

En el caso frecuentista, dado el EMV $\hat{\boldsymbol{\theta}}$, la verosimilitud es $f(\text{datos}|\hat{\boldsymbol{\theta}}, \mathcal{M})$.

Entonces, por lo general, el factor Bayes y la razón de verosimilitudes clásica no son iguales.

Ejemplo

Volvemos al ejemplo anterior. Supongamos que queremos comparar los hipótesis $H_0 : \theta = 0{,}5$ frente a $H_1 : \theta \neq 0{,}5$ y que bajo H_1, supongamos una distribución uniforme para θ.

Luego, $f(\text{datos}|H_0) = \left(\begin{array}{c} 12 \\ 9 \end{array} \right) 0{,}5^{12} \approx 0{,}054$ y

$$\begin{aligned} f(\text{datos}|H_1) &= \int_0^1 \left(\begin{array}{c} 12 \\ 9 \end{array} \right) \theta^9 (1-\theta)^3 \, d\theta \\ &= \left(\begin{array}{c} 12 \\ 9 \end{array} \right) B(10,4) \approx 0{,}077 \end{aligned}$$

Entonces, el factor Bayes a favor de H_1 es $B_0^1 = 1{,}43$ (escasa evidencia a favor de H_1).

Ejemplo

Volvemos al ejemplo anterior. Supongamos que queremos comparar los hipótesis $H_0 : \theta = 0{,}5$ frente a $H_1 : \theta \neq 0{,}5$ y que bajo H_1, supongamos una distribución uniforme para θ.

Luego, $f(\text{datos}|H_0) = \binom{12}{9} 0{,}5^{12} \approx 0{,}054$ y

$$\begin{aligned} f(\text{datos}|H_1) &= \int_0^1 \binom{12}{9} \theta^9 (1-\theta)^3 \, d\theta \\ &= \binom{12}{9} B(10,4) \approx 0{,}077 \end{aligned}$$

Entonces, el factor Bayes a favor de H_1 es $B_0^1 = 1{,}43$ (escasa evidencia a favor de H_1).

El EMV de θ es $\hat{\theta} = 0{,}75$ y luego la razón de verosimilitudes clásica es 4,8.

A lo frecuentista, comparando con χ_1^2, rechazariamos el hipótesis H_0 a favor de H_1.

El BIC

EL "Bayesian information criterion" (BIC) es un criterio frecuentista de selección de modelos.

Para un modelo, $\mathcal{M}$, con parámteros $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)^T$ el BIC se define como

$$BIC = -2 \log f(\text{datos}|\hat{\boldsymbol{\theta}}, \mathcal{M}) + k \log n.$$

donde n es el tamaño muestral.

Se selecciona el modelo con el mínimo valor del BIC.

Cuando $n \to \infty$, se puede demostrar que la verosimilitud marginal bayesiana:

$$f(\text{datos}|\mathcal{M}) \approx \exp(-BIC/2) \quad \text{o igualmente} -2 \log f(\text{datos}|\mathcal{M}) \approx BIC.$$

Problemas con el factor Bayes

- Dificultad de cálculo
  Fuera de problemas conjugadas, es muy complicado calcular el factor Bayes. Existen varias aproximaciones (Chib, "bridge sampling") pero son difíciles de implementar.
- ¿Qué pasa si usas distribuciones a priori impropias? ▸ Ejemplo

Problemas con el factor Bayes

- Dificultad de cálculo
  Fuera de problemas conjugadas, es muy complicado calcular el factor Bayes. Existen varias aproximaciones (Chib, "bridge sampling") pero son difíciles de implementar.
- ¿Qué pasa si usas distribuciones a priori impropias? ▸ Ejemplo
  Existen varias modificaciones del factor Bayes pero ninguna es del todo adecuada.

El AIC

El "Akäike information criterion" (AIC) es uno de los criterios clásicos más empleados para selección de modelos. Penaliza la sobreparameterización algo menos que el BIC.

Para un modelo $\mathcal{M}$ con parámetros $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)^T$ es

$$AIC = -2 \log f(\text{datos}|\hat{\boldsymbol{\theta}}, \mathcal{M}) + 2k = D(\hat{\boldsymbol{\theta}}) + 2k$$

dónde $D(\boldsymbol{\theta}) = -2 \log f(\text{datos}|\boldsymbol{\theta}, \mathcal{M})$ se llama la desvianza.

Se selecciona el modelo con el mínimo valor del AIC.

El DIC

Una versión bayesiana del AIC es el "deviance information criterion" o DIC. Se define el DIC como

$$DIC = D(\bar{\boldsymbol{\theta}}) + 2p_D$$

done $\bar{\boldsymbol{\theta}} = E[\boldsymbol{\theta}|\text{datos}, \mathcal{M}]$ es la media a posteriori de $\boldsymbol{\theta}$ y el número efectivo de parámetros, p_D es

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$$

dónde $\overline{D(\boldsymbol{\theta})} = E[D(\boldsymbol{\theta})|\text{datos}, \mathcal{M}]$.

Una definición alternativa para el número efectivo de parámetros es $p_D = \frac{1}{2}V[D(\boldsymbol{\theta})|\text{datos}, \mathcal{M}]$.

Una gran ventaja del DIC es que es muy fácil estimar a través del output MCMC.

Ejemplo: problemas del factor Bayes

Volvemos al ejemplo anterior y supongamos que bajo H_1, empleamos una distribución de Haldane $f(\theta|H_1) \propto \frac{1}{\theta(1-\theta)}$.

Luego, la verosimilitud marginal bajo H_1 es:

$$\begin{aligned} f(\text{datos}|H_1) &\propto \int_0^1 \binom{12}{9} \theta^9 (1-\theta)^3 \frac{1}{\theta(1-\theta)}\, d\theta \\ &\propto \binom{12}{9} B(9,3) \\ &\propto 0{,}444 \end{aligned}$$

Ejemplo: problemas del factor Bayes

Volvemos al ejemplo anterior y supongamos que bajo H_1, empleamos una distribución de Haldane $f(\theta|H_1) \propto \frac{1}{\theta(1-\theta)}$.

Luego, la verosimilitud marginal bajo H_1 es:

$$\begin{aligned} f(\text{datos}|H_1) &\propto \int_0^1 \binom{12}{9} \theta^9 (1-\theta)^3 \frac{1}{\theta(1-\theta)}\, d\theta \\ &\propto \binom{12}{9} B(9,3) \\ &\propto 0{,}444 \\ &\propto 1 \\ &\propto 10000000. \end{aligned}$$

El problema es que en la verosimilitud, tenemos $\propto$ en lugar de $=$ y luego, la verosimilitud marginal no está definida y entonces, tampoco está definido el factor Bayes.