Inferencia Estadística II

Solución

Junio de 2005

Problema I

1. El logaritmo de la verosimilitud de los datos en este modelo de Poisson se puede escribir:

$$\log L_\lambda(x) = \log \lambda \sum_{i=1}^{n} X_i - n\lambda + C$$

donde C es una constante que no depende de λ. De esta expresión se deduce que la razón de verosimilitudes es una función creciente del estadístico $T = \sum_{i=1}^{n} X_i$. Por lo tanto, el test UMP ϕ con nivel de significación (o error I) α para el contraste contrastar $H_0 : \lambda \leq \lambda_0$ frente a $H_1 : \lambda > \lambda_0$ se escribe

$$\phi = \begin{cases} 1 & \text{si } T > u \\ \gamma & \text{si } T = u \\ 0 & \text{si } T < u \end{cases}$$

donde la probabilidad γ y el umbral u verifican $\alpha = \gamma \mathbf{P}_{\lambda_0}(T = u) + \mathbf{P}_{\lambda_0}(T > u)$.

2. Por el Teorema Central del Límite se puede aproximar la distribución de T con una normal. Por lo tanto, si el tamaño muestral n es lo bastante grande podemos considerar el estadístico T como una variable continua y el test se escribe:

$$\phi = \begin{cases} 1 & \text{si } T > u \\ 0 & \text{si } T < u \end{cases}$$

Para calcular el umbral u cuando $\alpha = 5\%$, tipificamos T cuando $\lambda = \lambda_0$, que tendrá como media

$$\mathbf{E}_{\lambda_0}(T) = \sum_{i=1}^{n} \mathbf{E}(X_i) = n\lambda_0$$

y varianza

$$\mathbf{var}_{\lambda_0}(T) = \sum_{i=1}^{n} \mathbf{var}(X_i) = n\lambda_0$$

Obtenemos que

$$\begin{aligned} 5\% &= \mathbf{P}_{\lambda_0}(T > u) \\ &= \mathbf{P}\left(Z > \frac{u - n\lambda_0}{\sqrt{n\lambda_0}}\right) \end{aligned}$$

donde Z es una variable normal estándar. Por lo tanto, u verifica que

$$u_{5\%} = n\lambda_0 + z_{5\%}\sqrt{n\lambda_0}$$

Si $n = 200$ y $\lambda_0 = 2$, obtenemos que $u_{5\%} = 400 + 1.64 * 20 = 432.8$. Puesto que se observó $T = 450$, podemos afirmar (con un riesgo del $\alpha = 5\%$) que los datos evidencian que la especie de ave se está recuperando (rechazamos H_0).

Problema II El test basado en el estadístico de Pearson es el test adecuado para contrastar la independencia de las filas y columnas de esta tabla de contingencia. El estadístico de test se escribe aquí

$$\begin{aligned} K_n^* &= \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{\left(N_{ij} - \frac{N_{i.}N_{.j}}{n}\right)^2}{\frac{N_{i.}N_{.j}}{n}} \\ &= \left[\frac{\left(679 - \frac{836\cdot 784}{1074}\right)^2}{\frac{836\cdot 784}{1074}} + \ldots + \frac{\left(25 - \frac{85\cdot 159}{1074}\right)^2}{\frac{85\cdot 159}{1074}} \right] \end{aligned}$$

En realidad, no hace falta calcular esta suma para poder concluir. Basta con observar que por ejemplo, el ultimo termino de esta suma $\frac{\left(25-\frac{85\cdot 159}{1074}\right)^2}{\frac{85\cdot 159}{1074}} = 12.\,251$ es superior al cuantil 5% de la distribución asintótica de K_n^* (una χ^2 con $(3-1)(3-1) = 4$ grados de libertades), puesto que $\chi^2_{4,5\%} = 9.49$. Por lo tanto, podemos rechazar la hipótesis de independencia entre situación laboral y estado civil al menos en esta población.

Problema III

1. Si solo disponemos de la distribución *a priori* Q (una normal con media $\mu = 20$ y varianza $\tau^2 = 100$) la predicción de la renta θ es obvia: $d_Q = \mu = 20$ y el error de está predicción es $\overline{w}_Q(d_Q) = \tau^2 = 100$. Claramente, está predicción es muy imprecisa y no resultaría muy útil para luchar contra el fraude fiscal.
2. Si disponemos del consumo X del hogar, entonces la mejor predicción de la renta está basada en la distribución de θ condicionada por X.

   (a) Por la formula de Bayes, deducimos que esta distribución es una normal con varianza

   $$\mathbf{var}_Q(\theta|X) = \left[\frac{1}{\tau^2} + \frac{0.8^2}{\sigma^2}\right]^{-1} = 13.51$$

   y media

   $$\mathbf{E}_Q(\theta|X) = 13.51\left[\frac{\mu}{\tau^2} + \frac{0.8}{\sigma^2}(X-2)\right]$$

   (b) La mejor predicción de θ conociendo X es por lo tanto

   $$d_Q(X) = \mathbf{E}_Q(\theta|X) = 0.54 + 1.08X$$

   (c) Y el error promedio de esta predicción será

   $$\begin{aligned} W_Q(d_Q(X)) &= \mathbf{E}[\mathbf{var}_Q(\theta|X)] \\ &= \mathbf{E}[13.51] = 13.51 \end{aligned}$$

   que corresponde a un 10% del error anterior.