APROXIMACIONES PARAMÉTRICAS AL ANÁLISIS DE SUPERVIVENCIA MEDIANTE MODELOS DE RIESGOS PROPORCIONALES

Joaquim Casellas Vidal
Grup de Recerca en Remugants, Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona).

INTRODUCCIÓN

El análisis de supervivencia (Allison, 1995) es una herramienta estadística con importantes aplicaciones en el campo de la producción animal. Aunque la longevidad de los animales puede ser considerada como un carácter dicotómico (sobrevive o no a cierta edad), y analizada como tal mediante regresión logística, modelos umbral o una simple comparación de proporciones, la consideración de este carácter como variable continua incrementa la informatividad de nuestros análisis. En este sentido, no tienen las mismas repercusiones productivas ni económicas un cordero que muere al segundo día de vida que otro que muere al mes, mientras que el enfoque dicotómico vivo o muerto al destete, los considera a ambos iguales. Por otro lado, el análisis de supervivencia presenta ventajas destacables en cuanto al manejo de datos censurados, los cuales son muy frecuentes en los sistemas productivos actuales.

Aunque el análisis de supervivencia se desarrolló inicialmente mediante aproximaciones semiparamétricas (Cox, 1972; Prentice y Gloeckler, 1978), los modelos paramétricos (Ducrocq et al., 1998a,b) presentan ventajas importantes en cuanto a tiempo de computación (Ducrocq et al., 2000), si bien requieren que los datos analizados se ajusten a la distribución asumida. En este sentido, esta ponencia se basará en los principios y características del análisis paramétrico de supervivencia, así como sus aproximaciones frecuentistas y Bayesianas.

DISTRIBUCIONES DE PROBABILIDAD CARACTERÍSTICAS

Las técnicas de análisis de supervivencia se fundamentan en dos distribuciones de probabilidad específicas, las funciones de supervivencia y riesgo. La función de supervivencia $S(t)$ en el momento *t* se define como la probabilidad de que un individuo sobreviva más de *t* días (Cox, 1972). Esta función no está descrita para valores de *t* negativos, alcanza su máximo en $t = 0$ [$S(t = 0) = 1$] y decrece hasta 0 cuando $t \to \infty$. Por otro lado, la función de riesgo $h(t)$ representa la tasa instantánea de muerte en el momento *t* condicionada a que el individuo sobreviva más de *t* días ($T > t$). La multiplicación de ambas funciones origina la función de densidad $f(t) = S(t).h(t)$, de crucial importancia a la hora de analizar la longevidad mediante máxima verosimilitud o aproximaciones Bayesianas.

RIESGOS PROPORCIONALES Y EFECTOS DEPENDIENTES DE TIEMPO

Si consideramos que la longevidad de un individuo *i* está influida por un conjunto de efectos sistemáticos $\boldsymbol{\beta}$, podemos considerar nuestras funciones de riesgo y supervivencia según:

$$h(t|\mathbf{x}_i) = h_0(t).e^{\mathbf{x}_i'\boldsymbol{\beta}}; \quad S(t|\mathbf{x}_i) = S_0(t)^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}$$

siendo $\mathbf{x}_i$ el vector de incidencia de $\boldsymbol{\beta}$, y $h_0(t)$ y $S_0(t)$ las funciones de supervivencia y riesgo bases de nuestra población, las cuales definiremos posteriormente, en función de la distribución paramétrica asumida. Esta reparametrización nos introduce de lleno en los modelos de riesgos proporcionales. Si consideramos dos individuos, *i* y *j*, el cociente entre sus funciones de riesgo respectivas será:

$$\frac{h_i(t|\mathbf{x}_i)}{h_j(t|\mathbf{x}_j)} = \frac{h_0(t).e^{\mathbf{x}_i'\boldsymbol{\beta}}}{h_0(t).e^{\mathbf{x}_j'\boldsymbol{\beta}}} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{e^{\mathbf{x}_j'\boldsymbol{\beta}}} = e^{(\mathbf{x}_i' - \mathbf{x}_j')\boldsymbol{\beta}} = \text{constante}$$

manteniéndose por tanto la proporcionalidad para cualquier valor de *t*. En este sentido, el mismo cociente calculado entre dos efectos sistemáticos *A* y *B* ($e^{\beta_A-\beta_B}$) caracterizará la razón de riesgo de *A* respecto *B*, la cual nos indica cuantas veces es más probable que muera un individuo afectado por *A* que por *B*.

El modelo $h(t|\mathbf{x}) = h_0(t).e^{\mathbf{x}'\boldsymbol{\beta}}$ asume que los diferentes efectos fijos considerados influyen de la misma manera a lo largo de toda la vida de los animales. Esto puede ser cierto para efectos como el sexo del animal, la raza o el año de nacimiento, mientras que otras influencias solo afectan durante un periodo de tiempo concreto. Tomemos como ejemplo una camada de lechones, la supervivencia de los cuales se ve influida por un efecto ambiental permanente caracterizado en gran medida por la cerda. Cuando algunos lechones se cambian de camada con el objetivo de homogeneizarlas, el efecto ambiental recibido también cambia. En este sentido, deberemos definir $h(t)$ y $S(t)$ distintas para el periodo pre-homogenización (1) y post-homogenización (2), siendo t_h la edad a la que se intercambian los animales, y p_1 y p_2 los efectos ambientales recibidos durante el periodo 1 y 2 respectivamente (Ducrocq et al., 1988a):

$$h_1(t|\mathbf{x},p_1) = h_0(t).e^{(\mathbf{x}'\boldsymbol{\beta}+p_1)};\quad S_1(t|\mathbf{x},p_1) = S_0(t)^{\exp(\mathbf{x}'\boldsymbol{\beta}+p_1)} \qquad t\in(0,t_h)$$

$$h_2(t|\mathbf{x},p_2) = h_0(t).e^{(\mathbf{x}'\boldsymbol{\beta}+p_2)};\quad S_2(t|\mathbf{x},p_2) = S_0(t)^{\exp(\mathbf{x}'\boldsymbol{\beta}+p_2)} \qquad t\in(t_h,\infty)$$

DISTRIBUCIONES PARAMÉTRICAS UTILIZADAS

Algunas distribuciones paramétricas han sido empleadas repetidamente en la literatura para llevar a cabo análisis de supervivencia, siendo probablemente la distribución exponencial la más simple de todas ellas. Tomando como punto de partida una distribución exponencial de parámetro λ, podremos definir sus funciones de supervivencia y riesgo bases como:

$$h_0(t) = \lambda;\quad S_0(t) = e^{-\lambda t}$$

La característica más importante de esta distribución es que su función de riesgo se mantiene constante a lo largo del tiempo, caracterizando la exponencial como una distribución sin memoria de tiempo (un animal tiene la misma probabilidad de morir a cualquier edad). La distribución Weibull, de parámetros ρ y λ, es una generalización de la exponencial que ha recibido una atención especial durante los últimos años dada su simplicidad y flexibilidad (Ducrocq et al., 1988a). Sus funciones características son:

$$h_0(t) = \lambda\rho(\lambda t)^{\rho-1};\quad S_0(t) = e^{-(\lambda t)^{\rho}}$$

simplificándose a las de una distribución exponencial cuando $\rho = 1$. Además de la exponencial y la Weibull, también se han utilizado otras distribuciones como la log-normal, gamma, Rayleigh y Gompertz, aunque en menor medida (Ducrocq, 1987).

ANÁLISIS DE SUPERVIVENCIA MEDIANTE MÁXIMA VEROSIMILITUD

Dentro del marco de trabajo de los modelos de supervivencia, podemos definir la función de verosimilitud del conjunto de parámetros a estimar como (Cox, 1972):

$$L(\rho,\lambda,\boldsymbol{\beta}) = \prod_i f(y_i) = \prod_i h(y_i)^{\delta_i} S(y_i)$$

siendo $\delta_i = 0$ si se trata de un dato censurado o $\delta_i = 1$ si es un registro completo. Asumiendo que nuestros registros se distribuyen siguiendo una Weibull, el logaritmo de la función de verosimilitud tomara la forma (Ducrocq et al., 1988a):

$$\ln L(\rho,\lambda,\boldsymbol{\beta}) = N\ln(\rho) + (\rho-1)\sum_i \ln(y_i)^{\delta_i} + \sum_i \left((\mathbf{x}_i'\boldsymbol{\beta})^{\boldsymbol{\delta}_i} - y_i^{\rho}e^{\mathbf{x}_i'\boldsymbol{\beta}}\right)$$

donde *N* representa el número de registros a analizar. La maximización de esta función nos proporciona estimaciones máximo verosímiles de los diferentes parámetros del modelo. La ampliación de estos modelos para la estimación de

variables dependientes del tiempo así como efectos aleatorios podemos encontrarla en Ducrocq et al. (1988a,b).

APROXIMACIÓN BAYESIANA AL ANÁLISIS DE SUPERVIVENCIA

Existen múltiples trabajos que abordan la resolución de los modelos de supervivencia mediante aproximaciones Bayesianas (Ibrahim et al., 2001) aunque su aplicación en el campo de la mejora animal ha sido muy escasa. En este sentido, Ducrocq y Casella (1996) plantearon una primera aproximación para la estimación de componentes de varianza, la cual podemos encontrar implementada en el programa Survival Kit (Ducrocq y Sölkner, 1998).

Desde un punto de vista Bayesiano, podemos asumir que la distribución de nuestros datos de supervivencia dados los restantes parámetros del modelo sigue una distribución paramétrica, que asumiremos exponencial para simplificar la notación:

$$p(\mathbf{y}|\lambda,\boldsymbol{\theta}) = \prod_i \left(\lambda.\mathrm{e}^{\mathbf{W}\boldsymbol{\theta}}\right)^{\delta_i} \mathrm{e}^{(-\lambda y_i \exp(\mathbf{W}\boldsymbol{\theta}))}$$

donde **W** es la matriz de incidencias de los efectos sistemáticos (β) y genéticos aditivos (**a**) representados por **θ**. A partir de aquí, y siguiendo un planteamiento paralelo al de los modelos lineales, podemos definir las siguientes distribuciones *a priori*:

$$p\left(\mathbf{a}\middle|\mathbf{A},\sigma_a^2\right) \sim NMV\left(\mathbf{0},\mathbf{A}\sigma_a^2\right) \qquad p\left(\sigma_a^2\middle|\nu,S^2\right) \propto \sigma_a^{2^{-\left(\frac{\nu}{2}+1\right)}} \mathrm{e}^{-\frac{\nu S^2}{2\sigma_a^2}}$$

siendo **A** la matriz de parentescos aditivos, y σ_a^2 la varianza genética aditiva. Para los restantes parámetros del modelo podemos definir distribuciones *a priori* uniformes, al igual que para los componentes de varianza si fijamos $\nu = -2$ y $S^2 = 0$. La distribución posterior conjunta será:

$$p\left(\lambda,\theta,\sigma_a^2,\sigma_p^2|y,\nu,S^2\right) = \sigma_a^{2^{-\left(\frac{q+\nu}{2}+1\right)}} \exp\left(\frac{\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}+\nu S^2}{2\sigma_a^2}\right) \cdot \prod_i \left(\lambda.\mathrm{e}^{\mathbf{W}\boldsymbol{\theta}}\right)^{\delta_i} \mathrm{e}^{(-\lambda y_i \exp(\mathbf{W}\boldsymbol{\theta}))}$$

representando *q* el número de niveles del efecto genético aditivo. Deberemos recurrir a procedimientos de Metropolis-Hastings (Hastings, 1970) para obtener muestras aleatorias de las distribuciones marginales posteriores de λ y **θ** (y ρ en caso de analizar un modelo Weibull). Por otro lado, el componente de varianza genético se podrá muestrear a partir de una distribución χ^{-2} con $\tilde{\nu} = q + \nu$ grados de libertad y $\tilde{S}^2 = \left(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} + \nu S^2\right)/\tilde{\nu}$ como parámetro de escala. Para la inclusión de variables dependientes del tiempo deberemos modificar la distribución *a priori* de nuestros datos, tal como se ha definido anteriormente.

REFERENCIAS BIBLIOGRÁFICAS

Allison P.D. 1995. Survival analysis using the SAS® system. A practical guide. SAS Inst. Inc., Cary, NC.

Cox D.R. 1972. J. Royal Stat. Soc. Series B 34: 187-220.

Ducrocq V. 1987. Ph. D. dissertation, Cornell University, Ithaca, NY.

Ducrocq V., Besbes B., Protais M. 2000. Genet. Sel. Evol. 32: 23-40.

Ducrocq V., Casella G. 1996. Genet. Sel. Evol. 28: 505-529.

Ducrocq V., Quaas R.L., Pollak E.J., Casella G. 1988a. J. Dairy Sci. 71: 3061-3070.

Ducrocq V., Quaas R.L., Pollak E.J., Casella G. 1988b. J. Dairy Sci. 71: 3071-3079.

Ducrocq V., Sölkner J. 1998. Proc. 6th World Cong. Genet. Applied Livest. Prod., Armidale.

Hastings W.K. 1970. Biometrika 57: 97-109.

Ibrahim J., Chen M.H., Sinha D. 2001. Bayesian survival analysis. Springer, NY.

Prentice R., Gloeckler L. 1978. Biometrics 34: 57-67.