MODELOS SEMIPARAMÉTRICOS EN EL ANÁLISIS DE SUPERVIVENCIA APLICADO A LA MEJORA GENÉTICA

J. P. Sánchez
Dep. de Ciencia Animal, UPV, Camino de Vera s/n 46022 Valencia
juansan@dca.upv.es

Modelos de fragilidad de riesgos proporcionales

Un concepto básico en el estudio de la supervivencia es el del riesgo que se produzca el fenómeno de interés en un momento t, dado que no se ha producido hasta ese momento y la modelización del mismo es esencial. Los modelos de fragilidad de riesgos proporcionales son una forma muy general de modelizar el riesgo. Su ecuación puede escribirse así:

$$h_i(t) = h_{o,n}(t) \cdot z_i \cdot \exp\{x_i'(t)\beta\} \qquad [1]$$

donde $h_i(t)$ es el riesgo de que se produzca el acontecimiento de interés para cada tiempo t y para el individuo i, y se modela como el producto de una función de riesgo base $h_{o,n}(t)$, que puede ser única o pude quedar definida de forma distinta para n subgrupos de la población, y el efecto de unas covariables (β) que pueden ser dependientes del tiempo, indicadas por $x_i'(t)$ y un término de fragilidad (z_i), que es un término de naturaleza aleatoria que en el contexto de la mejora genética es el que recoge el efecto genético que determina el carácter, aunque puede tener más componentes. Una forma alternativa de presentar este modelo es incluir el término de fragilidad en la parte exponencial ($z_i = \exp\{u_i\}$). Este modelo se dice que es de riesgos proporcionales pues se asume que dentro de un subgrupo n de la población y en el espacio de tiempo en el que no hay cambios de nivel en la covariables dependientes del tiempo, el ratio de riesgo entre dos individuos afectados por distintos niveles de las covariables se mantendrá constante.

La diferencia entre modelos paramétricos y semi-paramétricos viene dada porque en los primeros la o las funciones de riesgo base son descritas por algún modelo probabilístico que depende de pocos parámetros (Weibull, exponencial, etc) y en los segundos no es así. La flexibilidad de los modelos semi-paramétricos es mayor que la de los modelos totalmente paramétricos pues no están sujetos a ninguna restricción sobre cómo se ha de distribuir la función de riesgo base. Las demandas computacionales para hacer las inferencias a partir de ellos son mayores, de ahí que su uso no esté tan generalizado, y cuando se definen los caracteres a estudiar se hace de manera que algún modelo paramétrico sea adecuado.

En este trabajo se hace una revisión sucinta de los modelos semi-paramétricos aplicados a la mejora genética

Modelo de COX

El modelo de COX es el método semiparamétrico implementado en el Survival Kit (Ducrocq y Sölkner, 1998). En él la función de riesgo base se mantiene indefinida, esto es posible pues se define una función de verosimilitud que se conoce como de verosimilitud parcial, que no depende de la función de riesgo base, su ecuación es:

$$L_c(y \mid \beta, u) = \prod_{k \in \{unc.\}} \frac{\exp\{x_k'(t)\beta + u_k\}}{\sum_{j \in R(T_k)} \exp\{x_j'(t)\beta + u_j\}}$$

Donde k representa a los individuos no censurados, y si no hay repeticiones en los tiempos de muerte darán lugar a k diferentes tiempos de muerte y $R(T_k)$ es el grupo de individuos a riesgo en ese tiempo k. Cuando hay repeticiones en los tiempos de muerte la función de verosimilitud parcial se modifica ligeramente para tener en consideración este hecho.

Esta función de verosimilitud parcial tiene todas las propiedades de cualquier otra función de verosimilitud. La principal justificación para su uso es que si estamos interesados en la estimación del efecto de las covariables o en la predicción de los términos de fragilidad, la información que para ello hay entre dos tiempos de muerte consecutivos es despreciable (Ducrocq, 2001).

Esta función de verosimilitud parcial se puede emplear en enfoques bayesianos para la estimación. En concreto el Survival Kit implementa este enfoque, asumiendo *priors* planos para β y los parámetros de dispersión del término de fragilidad y diferentes *priors* para el término de fragilidad: Gamma, Normal o Normal Multivariante. Sólo es el último el que permite incluir matriz de relaciones de parentesco, por lo que es el más comúnmente empleado cuando se quieren estimar parámetros genéticos o hacer evaluaciones genéticas.
Con la implementación del Survival Kit, si se conocen los parámetros de dispersión, la estimación de los efectos fijos y la predicción de los efectos aleatorios (términos de fragilidad) se hace como la moda de la distribución posterior conjunta condicionada a los parámetros de dispersión del término de fragilidad, el error de estas estimas y predicciones se calcula como el término de la diagonal de la inversa de menos la matriz de segundas derivadas de la distribución posterior conjunta. Cuando no se conocen los parámetros de dispersión del término de fragilidad, este se puede estimar solamente para uno de los componentes del término de fragilidad. La estimación de este parámetro de dispersión se hace sobre la distribución marginal del mismo. La marginalización se hace vía Integración Laplacina (Ducrocq, 2001) y finalmente la estima suele ser la moda de la distribución marginal posterior. Cuando el volumen de datos no es excesivamente grande se puede llegar mediante técnicas de integración numérica a los tres primeros momentos que describen la distribución marginal posterior del parámetro de dispersión, para tener una estima de la incertidumbre de la estimación del parámetro de dispersión y del grado de simetría.
La integración Laplaciana requiere de información concentrada, es decir que es necesario que varios individuos estén afectados por el mismo nivel del efecto del término de fragilidad del que se quiere estimar el parámetro de dispersión. Por lo tanto los modelos animales no son aplicables. En un estudio de simulación (Ducrocq y Casella, 1996) demuestran que incluso usando el modelo macho cuando hay cinco o menos hijas por macho la integración Laplaciana no funciona bien.
En resumen, la principal ventaja de la implementación del modelo de COX en el survival kit es su relativa facilidad y rapidez de cómputo, mientras que su principal inconveniente es que sólo se puede estimar el parámetro de dispersión para un único componente del término de fragilidad y que para esta estimación se recurre a un método aproximado como es la integración Laplaciana que no permite modelos animales.

Piecewise Constant Hazard Model

En estos modelos inicialmente hay que dividir el eje del tiempo en diferentes segmentos, generalmente esta división se basa en los propios tiempos de muerte, y para cada uno de estos segmentos se asume que el riesgo base es constante, de manera que la función de riesgo base estará definida por una función de escalera con riesgo constante entre dos tiempos de muerte consecutivos. En modelos de este tipo se define una función de verosimilitud completa a la que los diferentes datos contribuirán de manera diferente en función de que sean completos o censurados. En concreto los datos no censurados lo harán con la función de densidad y los censurados por la derecha con la de supervivencia, que se pueden calcular dadas conexiones entre ellas y la función de riesgo que define el modelo [1].

$$\mathrm{L}(\beta,u\mid y)\alpha \prod_{i\in\{unc.\}} f(y_i)\cdot \prod_{i\in\{cen.\}} S(y_i)$$

Existe una conexión entre este modelo y el modelo de COX, si la función de verosimilitud completa se factoriza en una parte que dependa de la función de riesgo base, y otra que no lo haga, esta segunda parte corresponderá precisamente a la función de verosimilitud parcial.
Bajo este tipo de modelos la estimación se hace por metodología bayesiana, pero en las aplicaciones que de este modelo hay en mejora genética las marginalizaciones se hacen por métodos MCMC, en concreto se emplea muestreo de Gibbs, y cuando

las distribuciones condicionales tienen formas desconocidas se emplea el Adaptative Rejection Sampling (Gilks y Wild, 1992), para muestrear de ellas.
Existen dos aplicaciones de este modelo en el campo de la mejora genética. En la primera de ellas se emplea este modelo para estudiar el tiempo desde el momento de la entrada en control de crecimiento de terneros de carne hasta que sufren una enfermedad respiratoria. En este primer trabajo los efectos fijos fueron covariables continuas y siempre no dependientes del tiempo (Korsgaard y col., 1998), los *priors* de esta aplicación se asumieron uniformes para los efectos "fijos", Gamma para los componentes de varianza de los cuatro términos del elemento de fragilidad (efecto de rebaño-grupo de origen, efecto genético aditivo, efecto de año-estación y efecto residual), log-normal para el efecto de estos cuatro términos y se asumió el *prior* de Jeffreys para cada uno de los peldaños de la función de riesgo base. En la segunda aplicación de este modelo se estudia la longevidad en conejas de producción cárnica, y se incluyen variables categóricas dependientes del tiempo (Sánchez y col., 2005), los *priors* para los efectos "fijos" fueron uniformes, al igual que para los componentes de varianza de los dos elementos del término de fragilidad (efecto genético aditivo y efecto residual), log-normal para el efecto de estos dos términos e igualmente se asumió el *prior* de Jeffreys para cada uno de los peldaños de la función de riesgo base.
En estas aplicaciones no hay restricciones en el número de componentes de los términos de fragilidad (salvo las marcadas por las posibilidades de los datos de que se disponga) y de hecho siempre se incluye además del efecto genético un efecto residual para modelar efectos individuales de naturaleza no genética que afecten al riesgo de eliminación de los animales. Además la inclusión de este término residual es de vital importancia cuando el modelo no es el modelo animal, pues se fuerza a que sea éste el término que recoja explícitamente la parte de la variación genética no tenida en consideración por los otros efectos del modelo. Por ejemplo en el modelo padre ¾ de la varianza aditiva queda no explicada por el modelo y hay que forzar a que vaya al término residual, por lo tanto la varianza residual debe de ser al menos 3 veces la varianza de macho. Si no existe este término residual y no se impone esta restricción, el modelo macho o cualquier otro modelo que no sea el modelo animal, no tendrá un modelo animal equivalente acorde con las premisas del modelo infinitesimal. Además se pueden obtener resultados sesgados en las evaluaciones genéticas. El efecto de la no inclusión del término residual ha sido estudiado para modelos macho paramétricos por Damgaard y col. (2003), observando sesgos en las estimaciones de los parámetros del modelo y comportamientos inestables en las evaluaciones genéticas de los animales. Ducrocq y Casella (1996) postulan que es un término extra que aparece cuando se hace una transformación log-lineal del modelo, y que se distribuye como una distribución de valor extremo el que implícitamente recoge la parte de la variación genética no considerada por los otros factores del modelo.
El principal inconveniente de estas aplicaciones es su elevada demanda computacional, mientras que sus ventajas son por un lado que se permite el uso del modelo animal y por otro que no hay restricciones en el número de componentes del término de fragilidad a la hora de estimar componentes de varianza.

Referencias

Damgaard L.H., Korsgaard I.R., Simonsen J., Dalsgaard O., Andersen A.H., 2003. Books of Abstracts of the 54th Annual Meeting of the EAAP, page 73. Roma, Italy.
Ducrocq, 2001. Survival Analysis applied to animal breeding and epidemiology. Course notes.
Ducrocq V. y Casella G., 1996. Genet. Sel. Evol. 28:505-529.
Ducrocq V. y Sölkner J.,1998. 6th WCGALP 27:447:450.
Gilks W.R., Wild P., 1992.Appl. Stat. 41 (1992) 337-348
Korsgaard I.R., Madsen P., Jensen J.,1998. Genet. Sel. Evol. 30:241-256.
Sánchez J.P., Korsgaard I.R., Damgaard L.H., Baselga M., 2005 (enviado a Genet. Sel. Evol.).