Un modelo de riesgos aditivos para el análisis bayesiano de tiempos de desempleo [†]

Eduardo Beamonte

Departamento de Economía Aplicada
Universitat de València
beamonte@uv.es

Resumen

La función de azar, también llamada función de riesgo o de intensidad, se utiliza habitualmente para modelizar datos de supervivencia u otros tiempos de espera como tiempos de desempleo. Al contrario que el modelo de azares proporcionales de Cox, el modelo de riesgos aditivos asume que la función de azar es la suma, en lugar del producto, de una función de azar base y una función no negativa de las covariables. En este trabajo se propone introducir las covariables en el modelo mediante una función de azar Gamma, mientras que la función de azar base no está especificada. Siguiendo el paradigma bayesiano, se obtiene una aproximación a la distribución final utilizando técnicas de Monte Carlo basadas en cadenas de Markov y se estudian las distribuciones predictivas de nuevos individuos. Un banco de datos reales sobre tiempos de desempleo sirve de ejemplo para comprobar la bondad de los métodos propuestos.

Palabras clave: azares no proporcionales, datos de desempleo, datos de supervivencia, distribuciones predictivas, métodos MCMC, modelos de riesgos aditivos, modelos jerárquicos.
Clasificación JEL: C11, C14, C15.

1 Introducción

El análisis estadístico de tiempos de espera entre dos sucesos dados ha sido ampliamente utilizado en ciencias biomédicas, donde se conoce como análisis de supervivencia y se dirige principalmente a investigar los efectos de los factores de riesgo en el desarrollo de la enfermedad o en la muerte y a predecir tiempos de supervivencia. En los últimos años, este tipo de técnicas estadísticas se han utilizado profusamente en otras áreas de conocimiento y, especialmente, en las ciencias económicas y sociales (Follmann et al., 1990; Jaggia y Thosar, 1995; Beenstock, 1996).

En análisis de supervivencia, los modelos de riesgos aditivos y multiplicativos constituyen las dos principales herramientas para estudiar la relación entre factores de riesgo y el tiempo de supervivencia o espera. La función de riesgo, también llamada función de azar o de intensidad, de un tiempo de espera T asociado a un p-vector de covariables $\mathbf{x}$ se define como $h(t|\mathbf{x}) = f(t|\mathbf{x})/(1 - F(t|\mathbf{x}))$, donde $f(\cdot|\mathbf{x})$ y $F(\cdot|\mathbf{x})$

[†]Este trabajo ha sido financiado parcialmente por el Ministerio de Educación y Ciencia con cargo al proyecto PB96-0776.

son la función de densidad y de distribución, respectivamente, de la variable aleatoria T condicionada al vector de covariables $\mathbf{x}$. La función $S(t|\mathbf{x}) = 1 - F(t|\mathbf{x})$ recibe el nombre de función de supervivencia.

Bajo el modelo de riesgos aditivos (Aalen, 1980; Cox y Oakes, 1984; Lin y Ying, 1994; Lin et al., 1998), la función de riesgo es

$$h(t|\mathbf{x}) = h_0(t) + \boldsymbol{\beta_0}'\mathbf{x}, \tag{1.1}$$

y bajo el modelo de riesgos multiplicativos (Cox, 1972) tiene la forma

$$h(t|\mathbf{x}) = h_0(t)\exp(\boldsymbol{\alpha_0}'\mathbf{x}), \tag{1.2}$$

donde $h_0(t)$ es la función de azar base, $\boldsymbol{\beta_0}$ y $\boldsymbol{\alpha_0}$ son p-vectores de parámetros de regresión y las covariables $\mathbf{x}$ pueden depender del tiempo.

Frecuentemente, el tiempo de espera T está sujeto a censura por la derecha ya que algunos de los individuos pueden estar todavía en espera al final del estudio. Además, no se ha hallado una parametrización satisfactoria para la función de azar base y, consecuentemente, sólo se ha realizado inferencia semiparamétrica para los modelos (1.1) y (1.2).

La función de verosimilitud parcial introducida por Cox (1972, 1975) explica la extendida utilización del modelo (1.2), aunque es bien conocido que la situación de azares proporcionales no siempre se tiene en la práctica. Además, los modelos de azares aditivos describen un aspecto diferente de la relación entre el tiempo de espera y las covariables y han sido utilizados satisfactoriamente, en varias formas, por numerosos autores (Buckley, 1984; Aalen, 1989; McKeague y Sasieni, 1994; Kim y Lee, 1998, y referencias ahí citadas), aunque su análisis estadístico resulta más complicado que el del modelo de azares proporcionales (1.2). Desde una perspectiva bayesiana, el modelo de riesgos multiplicativos también ha sido el más utilizado, una excelente revisión de su tratamiento estadístico puede consultarse en Sinha y Dey (1997).

Los modelos de azares aditivos y multiplicativos, tal y como han sido definidos en las ecuaciones (1.1) and (1.2), modelizan los datos como si todos los individuos de la muestra (condicionalmente al vector de covariables) provinieran de una única población homogénea. Sin embargo, frecuentemente hay una heterogeneidad en la población que las covariables no pueden explicar apropiadamente. Con los recientes avances en materia de computación se puede recoger esta heterogeneidad añadiendo una estructura jerárquica en el modelo y convirtiéndolo entonces en un modelo de efectos aleatorios o en poblaciones.

En este trabajo proponemos un modelo jerárquico, descrito con detalle en la siguiente sección, donde la función de riesgo es la suma de una función de azar base, $h_0(t)$, y una función de azar Gamma específica para cada individuo y asociada a sus covariables mediante una relación no determinista incluida en el orden más alto de la jerarquía.

El análisis bayesiano completo de modelos jerárquicos es ahora posible utilizando técnicas de simulación. Desde la introducción del muestreo de Gibbs en el análisis bayesiano (Geman y Geman, 1984; Gelfand y Smith, 1990; Casella y George, 1992) se ha profundizado mucho sobre las propiedades matemáticas y en la metodología de

ésta y otras técnicas de Monte Carlo basadas en cadenas de Markov (Smith y Roberts, 1993; Tierney, 1994), y han sido ampliamente utilizadas con éxito (Chib y Greenberg, 1996; Gilks et al., 1996; Geweke et al., 1998). La implementación de los métodos MCMC no es directa. En la sección 3 se propone un algoritmo de Metropolis dentro de Gibbs que no es difícil de implementar y que ha funcionado bastante bien con los bancos de datos, tanto reales como simulados, analizados (Beamonte, 1998).

En la sección 4 se analiza un ejemplo real sobre los licenciados en Matemáticas por la Universitat de València. Se considera el tiempo transcurrido desde su licenciatura hasta la obtención del primer empleo y se relacionan estos tiempos de desempleo con covariables como el sexo, año de licenciatura y nota media. El tabajo concluye con unas consideraciones finales.

2 El modelo Gamma-poligonal aditivo

Sea T el tiempo de espera de un individuo con vector de covariables $\mathbf{x}$. La variable aleatoria T sigue un modelo Gamma-poligonal aditivo si su función de azar es de la forma

$$h\left(t\,|\,\mathbf{x}\right) = h_0(t) + h_1\left(t\,|\,\mathbf{x}\right) \text{ si } t > 0.$$

La parte no paramétrica del modelo, $h_0(t)$, se supone una función poligonal no negativa con vértices localizados en los tiempos $a_0 = 0 < a_1 < \cdots < a_g$, donde la poligonal toma los valores $\tau_0, \tau_1, \ldots, \tau_g$, respectivamente y es constante después del tiempo a_g.

$$h_0(t) = \begin{cases} \tau_{j-1} + \frac{(\tau_j - \tau_{j-1})(t - a_{j-1})}{a_j - a_{j-1}} & \text{si } a_{j-1} \le t \le a_j \ , \ j = 1, \ldots, g \\ \tau_g & \text{si } t \ge a_g \, . \end{cases} \tag{2.1}$$

La parte paramétrica, $h_1\left(t\,|\,\mathbf{x}\right)$, es la función de azar de una distribución Gamma con parámetros α y β (media α/β y varianza α/β^2).

$$h_1\left(t\,|\,\mathbf{x}\right) = \frac{t^{\alpha-1}\exp(-\beta t)}{\int_t^\infty s^{\alpha-1}\exp(-\beta s)\,ds} \text{ si } t > 0.$$

Los parámetros α y β son específicos para cada individuo de la población, pero relacionados con el vector $\mathbf{x}$ mediante un segundo nivel del modelo jerárquico:

$$\begin{aligned} \alpha \mid \beta, \mathbf{x} &\sim N\left(\log\frac{\alpha}{\beta}\,|\,\mathbf{b}'\mathbf{x}, \sigma_\alpha^2\right) \\ \beta &\sim N\left(\log\beta\,|\,\mu_\beta, \sigma_\beta^2\right). \end{aligned} \tag{2.2}$$

Esto es, dados β y $\mathbf{x}$, el logaritmo de la media, $\log(\alpha/\beta)$, es modelizado según una distribución Normal con media $\mathbf{b}'\mathbf{x}$, una combinación lineal de las covariables, y varianza σ_α^2. El logaritmo de β es modelizado también según una Normal con media μ_β y varianza σ_β^2. Los hiperparámetros $\mathbf{b}, \sigma_\alpha^2, \mu_\beta$ y σ_β^2 son constantes desconocidas comunes a todos los individuos de la población.

De este modo, el modelo Gamma-poligonal aditivo es un modelo en poblaciones que permite cierta heterogeneidad en la población. Esta es la principal diferencia con el habitual modelo aditivo (1.1), pero hay además dos diferencias que cabe destacar. La utilización de una función de azar base poligonal, en lugar de la habitual función escalonada, sólo añade un poco de complicación en el cálculo y tiene la ventaja de que $h(t|\mathbf{x})$ es continua. Por otra parte, si las covariables son independientes del tiempo entonces la parte paramétrica del modelo (1.1) es constante en t y entonces resulta la función de azar de un modelo exponencial, i.e., un modelo Gamma con parámetro α conocido e igual a uno. Así, el modelo Gamma-poligonal aditivo constituye una generalización del modelo (1.1).

En general, la función de supervivencia está relacionada con la de azar mediante la expresión $S(t) = \exp[-H(t)]$, donde $H(t) = \int_0^t h(s)\,ds$ es la función de azar acumulado. Entonces, la función de supervivencia del modelo Gamma-poligonal aditivo, dados los parámetros $\boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_g)$, α y β es

$$S(t|\boldsymbol{\tau}, \alpha, \beta) = S_0(t|\boldsymbol{\tau})\, S_1(t|\alpha, \beta)\,, \tag{2.3}$$

donde $S_0(t)$ y $S_1(t)$ son las funciones de supervivencia relativas al azar poligonal y al azar Gamma, respectivamente.

$$\begin{aligned} S_0(t|\boldsymbol{\tau}) &= \exp\left[-\int_0^t h_0(s)\,ds\right] = \exp\left(-\sum_{i=0}^{g} c_i(t)\,\tau_i\right) \\ S_1(t|\alpha, \beta) &= \int_t^{+\infty} \frac{\beta^\alpha}{\Gamma(\alpha)}\, s^{\alpha-1} \exp(-\beta s)\,ds, \end{aligned}$$

con $c_i(t)$ estadísticos positivos.

Entonces, la función de densidad es de la forma

$$\begin{aligned} f(t|\boldsymbol{\tau}, \alpha, \beta) &= S(t|\boldsymbol{\tau}, \alpha, \beta)\, h(t|\boldsymbol{\tau}, \alpha, \beta) \\ &= S_0(t|\boldsymbol{\tau})\, S_1(t|\alpha, \beta)\, [h_0(t|\boldsymbol{\tau}) + h_1(t|\alpha, \beta)]\,. \end{aligned} \tag{2.4}$$

3 Procedimiento inferencial

3.1 La función de verosimilitud

Supóngase que el tiempo de espera del individuo i-ésimo T_i es una variable aleatoria absolutamente continua condicionalmente independiente de un tiempo de censura (por la derecha) V_i dado el vector de covariables X_i. Sean $Y_i = \min(T_i, V_i)$ el tiempo de supervivencia y $\delta_i = I(T_i \leq V_i)$ el indicador de censura. Supóngase que las tripletas (Y_i, δ_i, X_i) son i.i.d., para $i = 1, \ldots, n$, y que la función de azar de T_i condicionada a X_i satisface el modelo Gamma-poligonal.

La función de verosimilitud es proporcional al producto de n factores, iguales a la densidad (2.4) para los datos no censurados y a la función de supervivencia (2.3) para las observaciones censuradas, y otros n factores iguales a la función de densidad

(2.2) de los parámetros α_i y β_i. De este modo, la función de verosimilitud resulta proporcional a

$$\prod_{i=1}^{n} S\left(t_i \,|\, \boldsymbol{\tau}, \alpha_i, \beta_i\right) \left[h\left(t_i \,|\, \boldsymbol{\tau}, \alpha_i, \beta_i\right)\right]^{\delta_i} f\left(\alpha_i, \beta_i \,|\, \mathbf{x_i}, \mathbf{b}, \sigma_\alpha^2, \mu_\beta, \sigma_\beta^2\right).$$

Además, una covariable categórica puede introducirse en el modelo de dos formas diferentes: como variables dummy dentro del vector $\mathbf{x}$ o particionando la población en estratos. En este caso, se supone que cada estrato tiene una función de azar base diferente -el vector $\boldsymbol{\tau}$ es distinto para cada uno de ellos-, pero los demás parámetros e hiperparámetros son comunes a todos los grupos de individuos.

Considerando k estratos, cada uno de ellos con una función de azar base poligonal como la definida en (2.1), con vector paramétrico $\boldsymbol{\tau}^{(j)}$ y n_j datos pertenecientes al estrato j-ésimo, $j = 1, \ldots, k$, la verosimilitud es proporcional a

$$\prod_{j=1}^{k} \prod_{i=1}^{n_j} S\left(t_{ij} \,|\, \boldsymbol{\tau}^{(j)}, \alpha_{ij}, \beta_{ij}\right) \left[h\left(t_{ij} \,|\, \boldsymbol{\tau}^{(j)}, \alpha_{ij}, \beta_{ij}\right)\right]^{\delta_{ij}} f\left(\alpha_{ij}, \beta_{ij} \,|\, \mathbf{x_{ij}}, \mathbf{b}, \sigma_\alpha^2, \mu_\beta, \sigma_\beta^2\right).$$

3.2 Distribuciones inicial y final

Para realizar un análisis bayesiano del modelo Gamma-poligonal aditivo deben especificarse distribuciones iniciales sobre los hiperparámetros $\mathbf{b}, \sigma_\alpha^2, \mu_\beta, \sigma_\beta^2$ y sobre el vector $\boldsymbol{\tau}$. Parece natural asumir independencia a priori entre $\boldsymbol{\tau}$, $(\mathbf{b}, \sigma_\alpha^2)$ y $(\mu_\beta, \sigma_\beta^2)$. Además, proponemos utilizar la habitual inicial conjugada Normal-Gamma inversa para los hiperparámetros, i.e.:

$$\begin{aligned}
\mu_\beta \mid \sigma_\beta^2 &\sim N\left(\mu_\beta \,|\, m_\beta, \sigma_\beta^2 v_\beta^2\right) \\
\sigma_\beta^2 &\sim Ga\left(1/\sigma_\beta^2 \,|\, a_\beta, b_\beta\right) \\
\mathbf{b} \mid \sigma_\alpha^2 &\sim N_p\left(\mathbf{b} \,|\, \mathbf{m}_\alpha, \sigma_\alpha^2 V_\alpha\right) \\
\sigma_\alpha^2 &\sim Ga\left(1/\sigma_\alpha^2 \,|\, a_\alpha, b_\alpha\right).
\end{aligned}$$

Como distribución inicial sobre el vector $\boldsymbol{\tau}$ utilizamos un proceso autocorrelado de primer orden, siguiendo las indicaciones de Gamerman (1991) que lo propuso en un contexto similar. Entonces,

$$\tau_i = \tau_{i-1} \exp(\epsilon_i),\ i = 1, \ldots, g,$$

donde $(\epsilon_1, \ldots, \epsilon_g)$ se suponen variables aleatorias Normales independientes de media cero y varianza σ_ϵ^2, y

$$\begin{aligned}
\tau_0 &\sim Ga\left(\tau_0 \,|\, a_\tau, b_\tau\right) \\
\sigma_\epsilon^2 &\sim Ga\left(1/\sigma_\epsilon^2 \,|\, a_\epsilon, b_\epsilon\right).
\end{aligned}$$

En el caso de que existan varios estratos, los vectores $\boldsymbol{\tau}$ asociados a cada uno de ellos pueden ser considerados independientes a priori con iniciales similares a la anteriormente especificada.

La distribución final resulta tan complicada que no parece posible realizar nigún estudio analítico, aunque la distribución inicial ha sido elegida lo más sencilla posible. Este problema es común a casi todo modelo en poblaciones. Sin embargo, es posible realizar un estudio de simulación utilizando el muestreo de Gibbs u otra técnica MCMC. De hecho, cualquier distribución condicional tiene una expresión más sencilla que la distribución conjunta, puesto que aquélla es proporcional a ésta.

La dsitribución final condicional completa de los hiperparámetros $(\mu_\beta, \sigma^2_\beta)$ es proporcional a

$$\left[\prod_{i=1}^{n} N\left(\log\beta_i \,|\, \mu_\beta, \sigma^2_\beta\right)\right] N\left(\mu_\beta \,|\, m_\beta, \sigma^2_\beta v^2_\beta\right) Ga\left(1/\sigma^2_\beta \,|\, a_\beta, b_\beta\right).$$

Los demás factores en la verosimilitud e inicial son parte de la constante de proporcionalidad. Esta expresión es la misma que aparece en el análisis conjugado habitual de datos Normales, de este modo (ver por ejemplo DeGroot, 1970, página 169) resulta proporcional a una distribución Normal-Gamma inversa.

$$\begin{aligned}
\mu_\beta \mid \sigma^2_\beta &\sim N\left(\mu_\beta \,|\, \frac{m_\beta + nv^2_\beta\,\overline{ln\beta}}{1+nv^2_\beta}, \frac{\sigma^2_\beta v^2_\beta}{1+nv^2_\beta}\right) \\
\sigma^2_\beta &\sim Ga\left(1/\sigma^2_\beta \,|\, a_\beta + \frac{n}{2}, b_\beta + \frac{ns^2_{ln\beta}}{2} + \frac{n\left(\overline{ln\beta} - m_\beta\right)^2}{2\left(1+nv^2_\beta\right)}\right), \qquad (3.1)
\end{aligned}$$

donde:

$$\begin{aligned}
\overline{ln\beta} &= \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\log\beta_{ij} \\
s^2_{ln\beta} &= \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\log\beta_{ij} - \overline{ln\beta}\right)^2.
\end{aligned}$$

De un modo similar, la distribución final condicional completa de $(\mathbf{b}, \sigma^2_\alpha)$ es proporcional a

$$\left[\prod_{i=1}^{n} N\left(\log\frac{\alpha_i}{\beta_i} \,|\, \mathbf{b}'\mathbf{x_i}, \sigma^2_\alpha\right)\right] N\left(\mathbf{b} \,|\, \mathbf{m}_\alpha, \sigma^2_\alpha V_\alpha\right) Ga\left(1/\sigma^2_\alpha \,|\, a_\alpha, b_\alpha\right),$$

la misma expresión que en el análisis conjugado habitual del modelo lineal Normal homocedástico (DeGroot, 1970, páginas 249-252). Así pues, resulta proporcional a una distribución Normal-Gamma inversa multivariante.

$$\begin{aligned}
\mathbf{b} \mid \sigma^2_\alpha &\sim N_p\left(\mathbf{b} \,|\, \hat{\mathbf{b}}, \sigma^2_\alpha\left(V^{-1}_\alpha + \mathbf{X}'\mathbf{X}\right)^{-1}\right) \\
\sigma^2_\alpha &\sim Ga\left(1/\sigma^2_\alpha \,|\, a_\alpha + \frac{n}{2}, b_\alpha + \frac{1}{2}\left[\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)'\mathbf{y} + \left(\mathbf{m}_\alpha - \hat{\mathbf{b}}\right)' V^{-1}_\alpha \mathbf{m}_\alpha\right]\right), \qquad (3.2)
\end{aligned}$$

donde $\mathbf{y} = \left(\log \frac{\alpha_{11}}{\beta_{11}}, \ldots, \log \frac{\alpha_{n_1 1}}{\beta_{n_1 1}}, \ldots, \log \frac{\alpha_{1k}}{\beta_{1k}}, \ldots, \log \frac{\alpha_{n_k k}}{\beta_{n_k k}}\right)'$, $\mathbf{X}$ es la matriz de covariables y $\hat{\mathbf{b}} = \left(V_\alpha^{-1} + \mathbf{X}'\mathbf{X}\right)^{-1} \left(V_\alpha^{-1}\mathbf{m}_\alpha + \mathbf{X}'\mathbf{y}\right)$.

Otro parámetro que tiene un análisis conjugado es σ_ϵ^2. La correspondiente distribución condicional completa es proporcional a

$$\left[\prod_{i=1}^{g} N\left(\log \tau_i \,|\, \log \tau_{i-1}, \sigma_\epsilon^2\right)\right] Ga\left(1/\sigma_\epsilon^2 \,|\, a_\epsilon, b_\epsilon\right),$$

y realizando sencillos cálculos,

$$\sigma_\epsilon^2 \sim Ga\left(1/\sigma_\epsilon^2 \,|\, a_\epsilon + \frac{g}{2}, b_\epsilon + \frac{1}{2}\sum_{i=1}^{g} \log \frac{\tau_i}{\tau_{i-1}}\right). \tag{3.3}$$

El resto de distribuciones condicionales no presentan un análisis conjugado. Para cada $i = 1, \ldots, n$, la condicional completa sobre el par (α_i, β_i) es proporcional a

$$S\left(t_i \,|\, \boldsymbol{\tau}, \alpha_i, \beta_i\right) \left[h\left(t_i \,|\, \boldsymbol{\tau}, \alpha_i, \beta_i\right)\right]^{\delta_i} f\left(\alpha_i, \beta_i \,|\, \mathbf{x_i}, \mathbf{b}, \sigma_\alpha^2, \mu_\beta, \sigma_\beta^2\right),$$

y no tiene una forma cerrada. De todas maneras es posible muestrear de ella utilizando un algoritmo de Metropolis con función importante $f\left(\alpha_i, \beta_i \,|\, \mathbf{x_i}, \mathbf{b}, \sigma_\alpha^2, \mu_\beta, \sigma_\beta^2\right)$, una distribución log-Normal bivariante.

Finalmente, la distribución final condicional completa de $\boldsymbol{\tau}$ tampoco presenta una forma cerrada y es proporcional a

$$\left[\prod_{i=1}^{n} S_0\left(t_i \,|\, \boldsymbol{\tau}\right) \left[h\left(t_i \,|\, \boldsymbol{\tau}, \alpha_i, \beta_i\right)\right]^{\delta_i}\right] \left[\prod_{j=1}^{g} N\left(\log \tau_j \,|\, \log \tau_{j-1}, \sigma_\epsilon^2\right)\right] Ga\left(\tau_0 \,|\, a_\tau, b_\tau\right).$$

Para muestrear de ella utilizamos el siguiente algoritmo de Metropolis: obtenemos cada ϵ_i de una distribución Normal de media $\epsilon_i^* \sigma^2 / C$ y varianza σ^2, donde ϵ_i^* es el valor obtenido en la etapa anterior, C es un parámetro de sintonización para Metropolis y $\sigma^{-2} = \frac{1}{C} + \frac{1}{\sigma_\epsilon^2}$. Una vez simulado el vector $(\epsilon_1, \ldots, \epsilon_g)$, obtenemos τ_0 de una distribución log-Normal de media τ_0^* (el valor de τ_0 en la etapa anterior), y cada τ_i como $\tau_{i-1} \exp(\epsilon_i)$, $i = 1, \ldots, g$.

Construimos el proceso de simulación de Metropolis dentro de Gibbs del siguiente modo. Partiendo de un punto inicial obtenemos (utilizando el algoritmo de Metropolis) nuevos valores para (α_i, β_i), $i = 1, \ldots, n$, y para el vector $\boldsymbol{\tau}$; entonces, obtenemos nuevos valores para $\sigma_\epsilon^2, \mathbf{b}, \sigma_\alpha^2, \mu_\beta$ y σ_β^2 muestreando de (3.3), (3.2) y (3.1), respectivamente. Siguiendo este proceso obtenemos una muestra de tamaño N, tan grande como se desee, proveniente de la distribución final.

Esta muestra puede utilizarse para la estimación de cualquier característica de interés de la distribución final, incluyendo momentos o densidades marginales. La dsitribución predictiva para un nuevo individuo con vector de covariables $\mathbf{x}$ puede estimarse utilizando Monte Carlo:

$$f(t \mid \mathbf{x}) \simeq \frac{1}{N}\sum_{i=1}^{N} S\left(t \,|\, \boldsymbol{\tau}^{(i)}, \alpha^{(i)}, \beta^{(i)}\right) h\left(t \,|\, \boldsymbol{\tau}^{(i)}, \alpha^{(i)}, \beta^{(i)}\right),$$

y, de un modo análogo, sus funciones de supervivencia y azar.

4 Aplicación a unos datos de desempleo

En 1994 se llevó a cabo una encuesta sobre la adecuación de los estudios de la licenciatura de Matemáticas a la actividad profesional, en virtud de un convenio firmado por la Universitat de València y la Conselleria de Educació i Ciència. Un millar de cuestionarios fueron remitidos a los licenciados en Ciencias Matemáticas por la Universitat de València durante los años 1978 a 1993 y fueron 559 los finalmente recibidos.

Uno de los items de la encuesta era el tiempo, en meses, desde la obtención de la licenciatura hasta la consecución del primer empleo. Esta es la variable a analizar en este trabajo, utilizando como covariables algunos otros items de la encuesta, aquellos posiblemente relacionados con el tiempo de desempleo: nota media, año de licenciatura, sexo y actitud ante el hipotético hecho de volver a cursar los mismos estudios. Utilizamos la nota media de licenciatura (aprobado, notable y sobresaliente) para formar tres estratos o grupos distintos, en lugar de tratarla como covariable.

En el primer estrato había 352 individuos con nota media aprobado y 37 de ellos censurados: todavía en situación de desempleo al cierre de la encuesta. Con nota media notable eran 171, 9 de ellos censurados. Finalmente, había 36 individuos con nota media sobresaliente y uno solo sin primer empleo.

Como no se disponía de información inicial sobre los hiperparámetros, utilizamos distribuciones iniciales pertenecientes a las familias comentadas en la sección 3.2, con varianzas razonablemente grandes. Los resultados que aquí se exponen fueron obtenidos con la siguiente distribución inicial: $m_\beta = 0$, $v_\beta^2 = 1$, $a_\beta = b_\beta = 1$, $m_\alpha = 0$, $v_\alpha^2 = 1$, $a_\alpha = b_\alpha = 1$; y, para los tres estratos, $a_\tau = b_\tau = 2$ y $a_\epsilon = b_\epsilon = 1$. Realizamos un análisis de sensibilidad utilizando otras distribuciones iniciales, todas ellas pertenecientes a la misma familia y con varianzas grandes, obteniendo resultados muy similares.

En un estudio preliminar utilizamos una función de azar base poligonal con un gran número de vértices, el mismo número y en las mismas posiciones temporales para los tres estratos. La figura 1 representa la estimación Monte Carlo de las tres funciones de azar base obtenidas en dicho estudio. Es conocido que la mayoría de los licenciados en Matemáticas en España encuentran su primer empleo como profesores de enseñanza secundaria y que las contrataciones se producen durante los dos meses anteriores al comienzo del curso académico. Este hecho puede explicar apropiadamente el comportamiento casi cíclico de las funciones de azar base para los estratos aprobado y notable observados en la figura 1.

Una vez observada la forma del azar base es posible reducir el número de parámetros del modelo. De hecho, con sólo unos pocos vértices pueden recogerse perfectamente los cambios de monotonicidad en las funciones de azar. En el estudio final sólo utilizamos vértices en los puntos dados por las componentes del vector $\mathbf{a} = (0, 2, 3, 4, 6, 12, 18, 24)'$. Elegimos dichos puntos para aproximar el ciclo anual en el comportamiento de las funciones de la figura 1, con vértices cada seis meses, pero añadimos algunos vértices durante el primer semestre porque gran parte de los datos

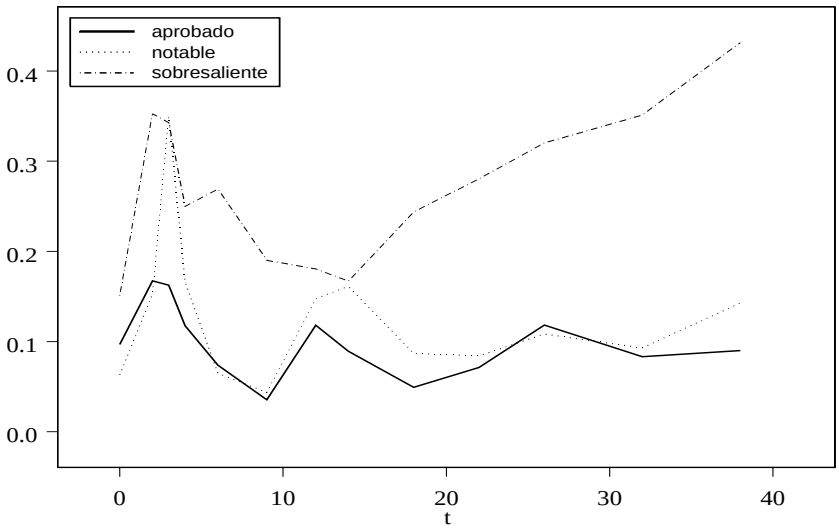

Figura 1: *Aproximación Monte Carlo al azar base de cada estrato.*

presentaban tiempos de supervivencia en dicho semestre.

Generamos una larga cadena de Markov desechando los 100000 primeros pasos (para alcanzar convergencia a la distribución final) y, registrando uno de cada 50 pasos (para reducir la autocorrelación de la cadena), hasta obtener una muestra de tamaño 10000. La figura 2 muestra la evolución de dicha cadena para los valores del coeficiente de la covariable año de licenciatura y la estimación Monte Carlo de su distribución marginal.

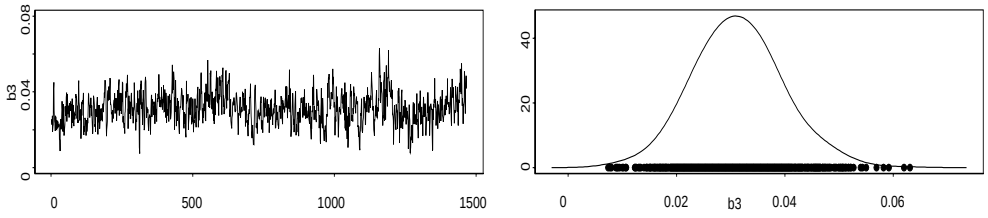

Figura 2: *Evolución de la cadena de Markov asociada al coeficiente de la covariable año de licenciatura, en la derecha, y su densidad predictiva estimada, en la izquierda.*

Realizamos los diagnósticos de convergencia utilizando el programa CODA (Best et al., 1995). Los resultados de estos análisis, tanto con tests gráficos (como el de la figura 2), como con tests estadísticos (test de Raftery y Lewis y el de Geweke), proporcionaron una aceptable convergencia a la distribución estacionaria de todos los

parámetros del modelo.

La tabla 1 recoge los intervalos de confianza Monte Carlo del 95% para algunos coeficientes del modelo. Algunos de los intervalos contienen al cero por lo que parece conveniente realizar una selección de covariables. Así, eliminamos secuencialmente las covariables sexo y actitud del modelo.

parámetro	2.5%	50%	97.5%
b_1	-0.982	0.308	1.6
b_2	-0.353	-0.0455	0.262
b_3	0.0159	0.0312	0.0481
b_4	-0.379	-0.0082	0.347
μ_β	-2.96	-2.48	-2.19

Tabla 1: *Intervalos intercuantílicos Monte Carlo para algunos hiperparámetros.*

Finalmente, de un modo similar, obtuvimos una muestra de tamaño 1000 de la distribución final con tres estratos y sólo la covariable año de licenciatura. Los análisis de esta muestra final también mostraron una rápida convergencia. En la tabla 2 se proporcionan las medias y desviaciones típicas predictivas obtenidas por Monte Carlo para algunos individuos, mientras que en la figura 3 se representan algunas funciones de supervivencia predictivas.

NOTA	Aprobado			Notable			Sobresaliente		
AÑO	1978	1988	1998	1978	1988	1998	1978	1988	1998
Media	8.19	10.41	12.81	6.87	8.41	9.97	3.55	3.91	4.11
Desviación	9.76	12.02	15.02	8.02	9.66	11.75	3.77	4.15	4.46

Tabla 2: *Media y desviación típica de la distribución predictiva de algunos individuos.*

Las supervivencias predichas para el año de licenciatura 1998 son muy similares para las notas medias aprobado y notable, siendo menores para el sobresaliente. Este comportamiento es análogo para todo año de licenciatura, como cabía esperar. La supervivencia en función del año de licenciatura también aparece ordenada en sentido directo al mismo, proporcionando mayores tiempos de espera hasta el primer empleo para los licenciados de las últimas promociones.

5 Conclusiones

El análisis de modelos en poblaciones, como el aquí estudiado, resultaba inviable sólo hace unos años. Ahora, con los grandes avances en materia de computación de los últimos tiempos, el analisis de esta clase de modelos no sólo es posible sino

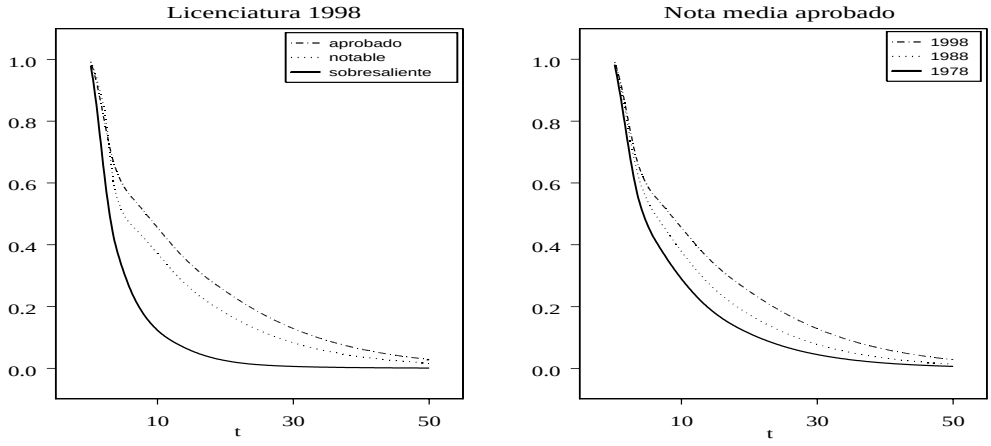

Figura 3: *Funciones de supervivencia predictivas para algunos individuos.*

recomendable ya que explican mejor la relación entre la variable dependiente y las covariables.

Se ha desarrollado un programa en Fortran, disponible por medio del primer autor, para el análisis del ejemplo aquí comentado y otros bancos de datos, tanto reales como simulados. En este trabajo se ha utilizado un ordenador personal con un microprocesador Pentium II a 266 Mhz.

References

Aalen, O. O. (1980). *A model for nonparametric regression analysis of counting processes*. New York: Springer-Verlag.

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, **8**, pp. 907–925.

Beamonte, E. (1998). *Aportaciones al análisis bayesiano semiparamétrico de datos de supervivencia*. Tesis Doctoral, Departamento de Estadística e Investigación Operativa. Universitat de València.

Beenstock, M. (1996). Training and the time to find a job in Israel. *Applied Economics*, **28**, pp. 935–946.

Best, N. G., Cowles, M. K. y Vines, S. K. (1995). *CODA manual version 0.30*. Cambridge: MRC Biostatistics Unit.

Buckley, J. (1984). Additive and multiplicative models for relative survival rates. *Biometrics*, **40**, pp. 51–62.

Casella, G. y George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, pp. 167–174.

Chib, S. y Greenberg, E. (1996). Markov chain Monte Carlo simulation methods in Econometrics. *Econometric Theory*, **12**, pp. 409–431.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, **34**, pp. 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, pp. 269–276.

Cox, D. R. y Oakes, D. A. (1984). *Analysis of survival data*. London: Chapman & Hall.

DeGroot, M. H. (1970). *Optimal statistical decisions*. New York: McGraw-Hill.

Follmann, D. A., Goldberg, M. S. y May, L. (1990). Personal characteristics, unemployment insurance, and the duration of unemployment. *Journal of Econometrics*, **45**, pp. 351–366.

Gamerman, D. (1991). Dynamic bayesian models for survival data. *Applied Statistics*, **40**, pp. 63–79.

Gelfand, A. E. y Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, pp. 398–409.

Geman, S. y Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, pp. 721–741.

Geweke, J., Keane, M. P. y Runkle, D. E. (1998). Statistical inference in the multinomial multiperiod probit model. *Journal of Econometrics*, **80**, pp. 125–165.

Gilks, W. R., Richardson, S. y Spiegelhalter, D. J., eds. (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Jaggia, S. y Thosar, S. (1995). Contested tender offers: an estimate of the hazard function. *Journal of Business & Economic Statistics*, **13**, pp. 113–119.

Kim, J. y Lee, S. Y. (1998). Two-sample goodness-of-fit tests for additive risk models with censored observations. *Biometrika*, **85**, pp. 593–603.

Lin, D., Oakes, D. y Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika*, **85**, pp. 289–298.

Lin, D. y Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, **81**, pp. 61–71.

McKeague, I. y Sasieni, P. (1994). A partly parametric additive risk model. *Biometrika*, **81**, pp. 501–514.

Sinha, D. y Dey, D. K. (1997). Semiparametric bayesian analysis of survival data. *Journal of the American Statistical Association*, **92**, pp. 1195–1212.

Smith, A. F. M. y Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B*, **55**, pp. 3–23.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, **22**, pp. 1701–1762.