

ANÁLISIS BAYESIANO DE TIEMPOS DE DESEMPLEO CON COVARIABLES

EDUARDO BEAMONTE¹ y JOSÉ D. BERMÚDEZ²

This work deals with the bayesian analysis of unemployment times with covariates. We propose a Gamma hierarchical model and we study the posterior distribution trough Gibbs sampling, using Markov Chain Monte-Carlo methods. We apply this model to a real data set of people with a Math's degree. The main objective of the work is the prediction of the unemployment time of new individuals.

Key words: Unemployment data, Covariates, Gamma Hierarchical Model, Markov-Chain Monte-Carlo Methods, Prediction.

1.- INTRODUCCIÓN.

Gran parte de la dificultad en el análisis de tiempos de desempleo radica en el hecho de que algunos de ellos son tiempos censurados. Esta situación, típica del análisis de supervivencia, consiste en que las observaciones realizadas en algunos individuos no muestran su tiempo total de desempleo y tan sólo se sabe que dicho tiempo es mayor que el observado.

El análisis de datos de supervivencia dentro del ámbito socio-económico ha experimentado un gran desarrollo en los últimos tiempos (Follman et al., 1990; Albert y Chib, 1994; Chib y Greenberg, 1995; ...).

En el presente trabajo proponemos un modelo jerárquico Gamma para el análisis bayesiano de datos de supervivencia progresivamente censurados por la derecha. Las especiales características de este tipo de datos hacen que el estudio de cualquier modelo paramétrico sea realmente complicado. Nosotros proponemos estudiar la distribución final a través de una muestra generada a partir de la misma, en la línea de trabajo propuesta por Gelfand y Smith (1990), entre otros.

¹Eduardo Beamonte. Departamento de Economía Aplicada. Universitat de València. Avda. Blasco Ibáñez, 30. 46010 Valencia.

²José D. Bermúdez. Departamento de Estadística e Investigación Operativa. Universitat de València. C/ Dr. Moliner, 50. 46100 Burjasot. Valencia.

Para la obtención de una muestra de la distribución final utilizamos el muestreo de Gibbs (Chib, 1992; Bermúdez y Beamonte, 1993, ...). Básicamente, consiste en construir una función de transición que defina una cadena de Markov irreducible y para la que la distribución final sea estacionaria. Este proceso exige que sea fácil muestrear a partir de las distribuciones condicionales completas, como es el caso del presente estudio.

El objetivo principal de este trabajo es el cálculo de densidades predictivas y funciones de supervivencia de nuevos individuos. Para ello llevamos a cabo una selección de covariables influyentes en el modelo.

El modelo y los métodos desarrollados para su análisis se aplican a un banco de datos de 559 individuos, licenciados en Ciencias Matemáticas por la Universitat de València.

2.- MODELIZACIÓN Y ANÁLISIS.

El modelo jerárquico Gamma propuesto es:

$$t \sim \text{Ga} (t \mid \alpha, \beta)$$

$$(\log \alpha, \log \beta)' \sim N_2 ((\log \alpha, \log \beta)' \mid Bx, H),$$

donde t es el tiempo de supervivencia y x es el vector de covariables de cada individuo, con un primer elemento unidad como en los modelos lineales generalizados. Esto es, cada tiempo de supervivencia es $\text{Ga} (\alpha, \beta)$, siendo los parámetros α y β características propias no observables del individuo, dependientes de su vector de covariables y ciertos hiperparámetros B y H comunes a todos ellos. De este modo, $(\alpha, \beta)'$ sigue una distribución log-Normal bivalente con media Bx y matriz de precisión H (Beamonte y Bermúdez, 1995).

Si denotamos por $\{t_1, \dots, t_r\}$ a los tiempos de supervivencia correspondientes a los datos no censurados y por $\{T_{r+1}, \dots, T_n\}$ a los tiempos de censura correspondientes a los datos censurados, el vector paramétrico completo objeto del muestreo de Gibbs es el formado por los parámetros del modelo, $(\alpha_1, \beta_1, \dots, \alpha_n, \beta_n)$, los hiperparámetros, $(B \text{ y } H)$, y los tiempos de supervivencia no observados, (t_{r+1}, \dots, t_n) , sujetos a las restricciones $t_i > T_i$, $i=r+1, \dots, n$. La distribución final viene dada por:

$$f (\alpha_1, \beta_1, \dots, \alpha_n, \beta_n, B, H, t_{r+1}, \dots, t_n \mid t_1, \dots, t_r, T_{r+1}, \dots, T_n, X),$$

siendo $X_{n \times k}$ la matriz de covariables; esto es, una matriz cuya fila i -ésima coincide con el vector de covariables del i -ésimo individuo. Y a partir de ella, obtenemos la distribución marginal de interés:

$$f(B, H | t_1, \dots, t_r, T_{r+1}, \dots, T_n, X). \quad (1)$$

La implementación del algoritmo de Gibbs pasa por una adecuada descomposición de las distribuciones condicionales completas, de modo que sea relativamente sencillo muestrear a partir de las mismas. Para una detallada explicación del desarrollo particular para este modelo puede consultarse Beamonte y Bermúdez (1995).

Una vez obtenida una muestra $\{(B^{(i)}, H^{(i)}), i=1, \dots, M\}$ a partir de (1), el cálculo de la densidad predictiva de un nuevo individuo con vector de covariables x se realiza como sigue:

$$f(t | x, t_1, \dots, t_r, T_{r+1}, \dots, T_n, X) = \iint Ga(t | \alpha, \beta) \cdot f(\alpha, \beta | x, t_1, \dots, t_r, T_{r+1}, \dots, T_n, X) d\alpha d\beta \approx \frac{1}{m} \sum_{j=1}^m Ga(t | \alpha_{(j)}, \beta_{(j)}), \quad (2)$$

con $\{(\alpha_{(j)}, \beta_{(j)}), j=1, \dots, m\}$, una muestra obtenida a partir de:

$$f(\log \alpha, \log \beta | x, t_1, \dots, t_r, T_{r+1}, \dots, T_n, X) = \int N((\log \alpha, \log \beta)' | Bx, H) \cdot f(B, H | t_1, \dots, t_r, T_{r+1}, \dots, T_n, X) dB dH \approx \frac{1}{M} \sum_{j=1}^M N((\log \alpha, \log \beta)' | B^{(j)}x, H^{(j)}).$$

3.- ANÁLISIS DE UNOS DATOS DE DESEMPLEO.

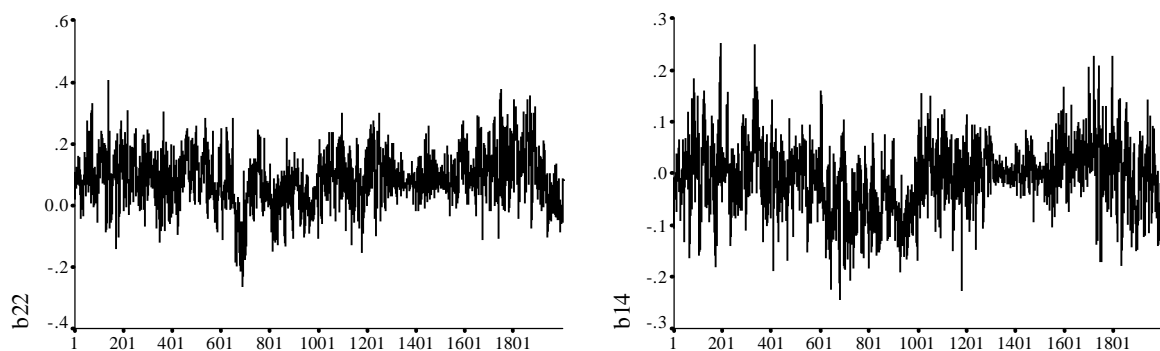
Los datos que analizamos en este trabajo proceden de una encuesta dirigida a los licenciados en Ciencias Matemáticas por la Universitat de València. La labor de campo con los cuestionarios cumplimentados se cerró en octubre de 1994 (Encuesta sobre la valoración de la adecuación de los estudios a la actividad profesional. Convenio de la Universitat de València con la Conselleria de Educació i Ciència).

El banco de datos finalmente considerado consta de 559 individuos, para todos y cada uno de los cuales aparece medida una variable indicando su tiempo de supervivencia (meses transcurridos desde la obtención de la licenciatura hasta el primer empleo), una variable indicadora de la censura (1, si ha encontrado primer empleo y 0, en caso contrario) y cinco

covariables: x_1 , constante e igual a 1, x_2 , sexo (1, si es hombre y 0, si es mujer), x_3 , año de obtención de la licenciatura, x_4 , nota media de la licenciatura (1, aprobado, 2, notable y 3, sobresaliente) y x_5 , actitud ante el hecho de volver a cursar la licenciatura (1, sí y 0, no).

Para el análisis de estos datos hemos utilizado el modelo comentado en el apartado anterior, analizando la distribución final mediante el algoritmo de Gibbs. Para ello, obtuvimos una realización de la cadena de Markov a partir de (1), de la siguiente forma:

Realizamos 100000 pasos en la misma y monitorizamos la evolución de las medias (calculadas cada 50 pasos) de los parámetros de la matriz B. Obtuvimos una rápida convergencia, además de muy poca variabilidad en todos ellos. En la figura 1 se muestra la evolución de las medias de los parámetros b_{22} y b_{14} .



-Figura 1. Evolución de las medias de los parámetros b_{22} y b_{14} .-

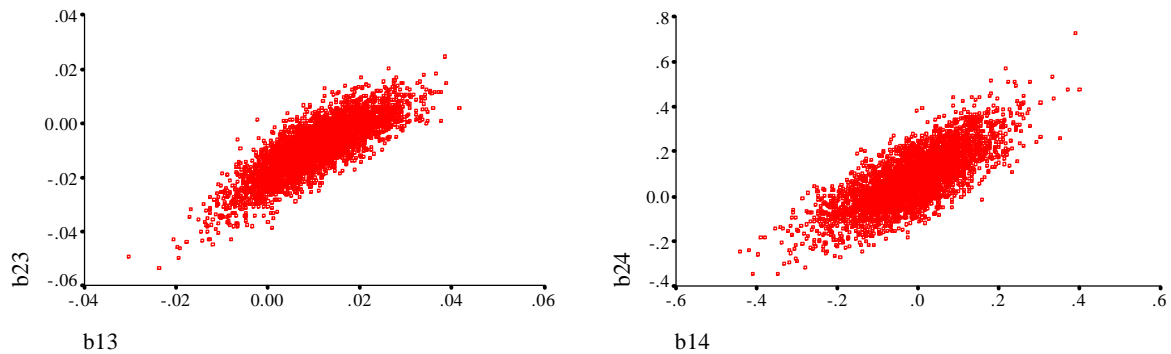
Para la obtención de una muestra de la distribución final decidimos desechar los 5000 primeros pasos de la cadena de Markov (convergencia) y, posteriormente, registrar uno de cada 20 (independencia) hasta un tamaño muestral igual a 4000. Es necesario un tamaño muestral tan elevado para la posterior aplicación del algoritmo SIR (Rubin, 1988).

La selección de variables a considerar en el modelo la llevamos a cabo mediante un procedimiento backward, combinando el propio muestreo de Gibbs con el algoritmo SIR. La utilización del algoritmo SIR permite reducir los tiempos de cálculo pero da tan sólo una aproximación y reduce el tamaño de la muestra generada previamente mediante el algoritmo Gibbs, por ello una utilización alternada de ambos procedimientos creemos que proporciona los mejores resultados.

Para seleccionar la variable a eliminar en cada paso, calculamos regiones de confianza de cada una de las columnas de parámetros de la matriz B (Wei y Tanner, 1990), así como la

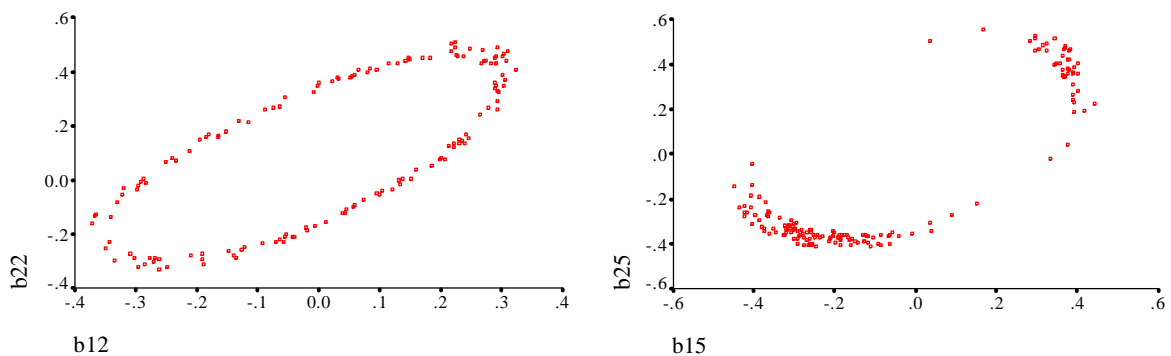
distancia de Mahalanobis de la muestra obtenida de cada una de esas columnas al vector origen.

En la figura 2 se representan las nubes de puntos muestrales correspondientes a dos columnas de parámetros de la matriz B.



-Figura 2. Nubes de puntos correspondientes a las muestras de $(b_{13}, b_{23})'$ y $(b_{14}, b_{24})'$.-

En la figura 3 observamos cómo el origen $(0, 0)'$ queda incluido claramente en las regiones de confianza 0.95 de $(b_{12}, b_{22})'$ y $(b_{15}, b_{25})'$.



-Figura 3. Regiones de confianza 0.95 de $(b_{12}, b_{22})'$ y $(b_{15}, b_{25})'$.-

También calculamos las distancias de Mahalanobis del vector $(0, 0)'$ a las nubes de puntos formadas por las muestras obtenidas de cada una de dichas columnas. En la tabla 1 aparecen tales distancias.

	x_2	x_3	x_4	x_5
D^2	0.8197	12.0731	1.3579	1.1256

-Tabla 1. Distancias de Mahalanobis del $(0, 0)'$ a las muestras de $(b_{1j}, b_{2j})'$.-

La menor de las distancias es la correspondiente a la covariable x_2 (sexo), por tanto sería ésta la covariable candidata a ser eliminada del modelo. Su valor, 0.8197, está asociado al cuantil 0.336 de una distribución F con 2 y 3998 grados de libertad, por lo que podemos considerar $b_{12} = b_{22} = 0$.

A continuación, obtenemos mediante el algoritmo SIR una muestra de tamaño 200 de los nuevos parámetros obtenidos una vez eliminada la covariable sexo del estudio. Repetimos sobre esta muestra el cálculo de distancias de Mahalanobis y regiones de confianza 0.95.

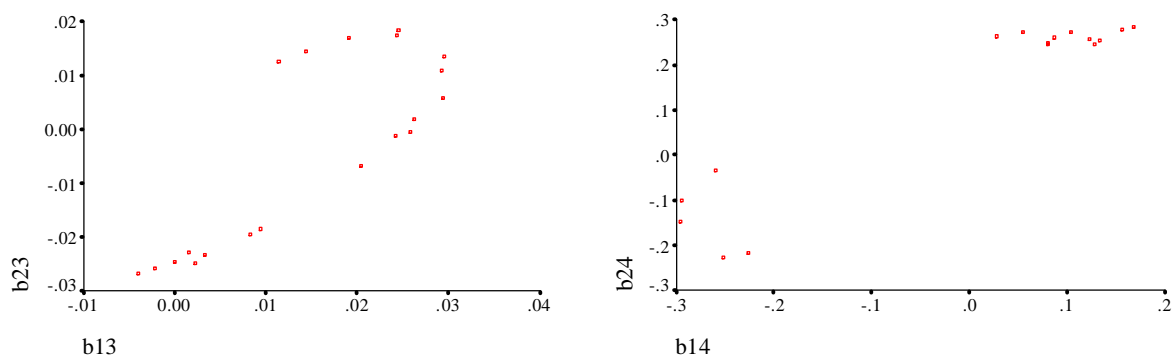
	x_3	x_4	x_5
D^2	12.6396	1.7641	0.02

-Tabla 2. Distancias de Mahalanobis del $(0, 0)'$ a las muestras de $(b_{1j}, b_{2j})'$.-

Eliminamos, por lo tanto, la covariable x_5 y proseguimos el análisis con las covariables x_3 (año de obtención de la licenciatura) y x_4 (nota media de la licenciatura). Mediante el algoritmo Gibbs obtuvimos una muestra de tamaño 500 de los nuevos parámetros siguiendo los mismos criterios acerca de la convergencia e independencia comentados con anterioridad. Las nuevas distancias de Mahalanobis y regiones de confianza 0.95 resultaron:

	x_3	x_4
D^2	7.806	0.7415

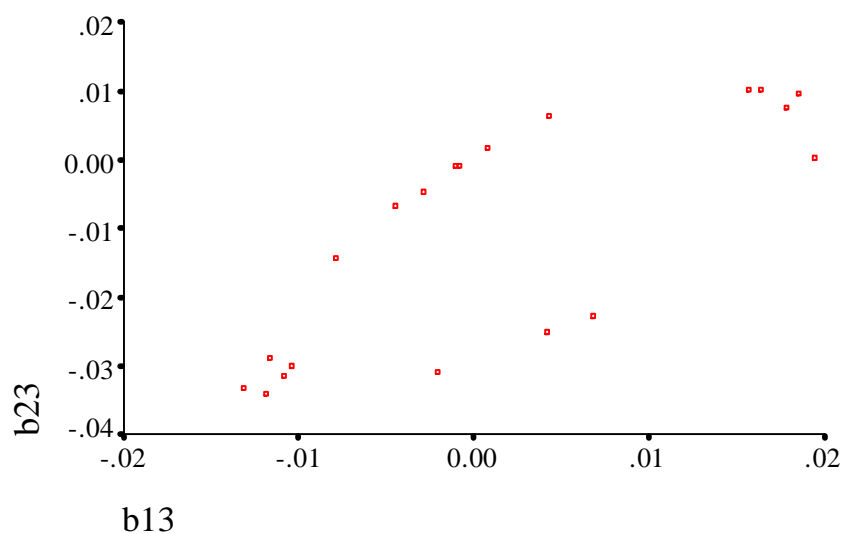
-Tabla 3. Distancias de Mahalanobis del $(0, 0)'$ a las muestras de $(b_{1j}, b_{2j})'$.-



-Figura 4. Regiones de confianza 0.95 de los parámetros correspondientes a x_3 y x_4 .-

Finalmente, obtenemos mediante muestreo de Gibbs una muestra de tamaño 500 de los parámetros B y H, considerando tan sólo dos covariables en el modelo, x_1 (constante e igual a

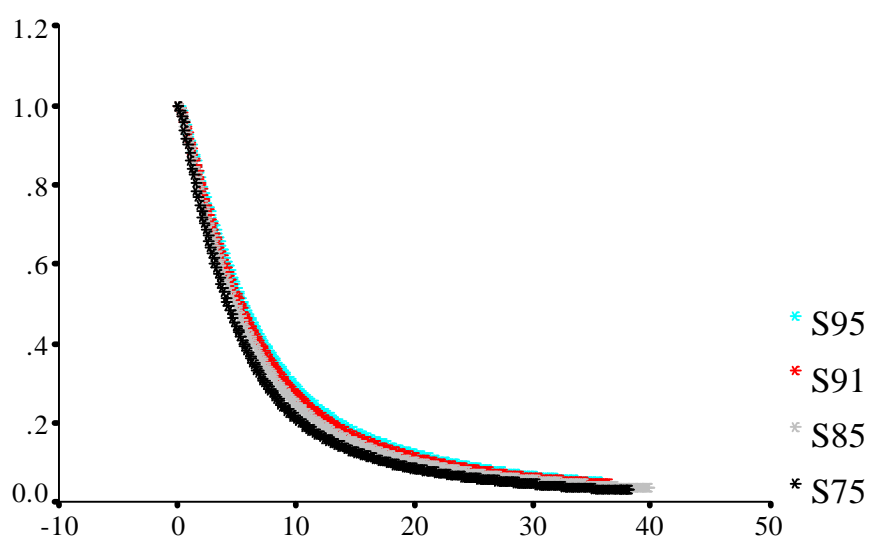
1) y x_3 . Comprobamos que esta última no es prescindible al obtener una distancia de Mahalanobis $D^2 = 5.5972$ y la siguiente región de confianza 0.95.



-Figura 5. Región de confianza 0.95 de los parámetros correspondientes a x_3 .-

Concluimos el trabajo con el cálculo de funciones de supervivencia predictivas ($S(t) = 1 - F(t)$, donde $F(t)$ es la función de distribución predictiva obtenida a partir de (2)) de nuevos individuos.

En la figura 6 se representan funciones de supervivencia para individuos con distintos años de obtención de la licenciatura, desde el más reciente hasta el más lejano.



-Figura 6. Funciones de supervivencia para $x_3=75$, $x_3=85$, $x_3=91$ y $x_3=95$.-

En la tabla 4 aparecen reflejadas la media y la varianza de las densidades predictivas comentadas con anterioridad.

	$x_3=75$	$x_3=85$	$x_3=91$	$x_3=95$
media	8.261	9.4279	10.2251	10.8021
varianza	193.1853	253.5909	301.9634	340.751

-Tabla 4. Media y varianza de las densidades predictivas.-

4.- CONSIDERACIONES FINALES.

En primer lugar cabe destacar la aplicabilidad y adecuación de este modelo a diferentes estudios de supervivencia. En concreto y para el que nos ocupa, es de resaltar la coherencia de los resultados obtenidos con los esperados, dada la difícil situación en la que se encuentran los licenciados de las últimas promociones para acceder al primer puesto de trabajo (congelación de plazas de oposición en enseñanza secundaria, complicado acceso a puestos docentes en la Universidad, situación económica a nivel nacional, ...). Del mismo modo, también parece lógica la exclusión de todas las covariables del modelo a excepción de x_1 y del año de obtención de la licenciatura. Básicamente, la situación anual del acceso a la docencia en enseñanza secundaria ha determinado el tiempo de obtención del primer empleo. Es presumible que en un futuro inmediato, dadas las restricciones en tales accesos, covariables como la nota media tengan una mayor y no prescindible influencia.

En la investigación acerca de la adecuación del modelo a los datos estudiados, introducimos una componente cuadrática en x_3 , como alternativa a la tendencia lineal para dicha variable. Se obtuvieron unos resultados bastante similares que confirmaron la validez del modelo finalmente considerado.

Asímismo, realizamos el cálculo de predictivas incorporando covariables inicialmente desechadas y las comparamos con las correspondientes predictivas obtenidas sólo con las covariables influyentes. El resultado de tales comparaciones fue bastante satisfactorio al comprobar que apenas existían diferencias entre ellas.

REFERENCIAS.

- Albert, J.H. y Chib, S.** (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *J. Bussiness & Econom. Statist.*, **11**, pp. 1-15.
- Beamonte, E y Bermúdez, J.D.** (1993). Comparación de curvas de supervivencia Gamma. *Qüestiiò*, **19**, pp. 171-186.
- Bermúdez, J.D. y Beamonte, E.** (1993). Análisis bayesiano de datos de supervivencia Gamma utilizando muestreo de Gibbs. *Estadística Española*, **35**, pp. 629-644.
- Chib, S.** (1992). Bayes inference in the Tobit censored regression model. *J. Econometrics.*, **51**, pp. 79-99.
- Chib, S. y Greenberg, E.** (1995). Hierarchical analysis of SUR models with extensions to correlated serial errors and time varying parameter models. *J. Econometrics.*, **68**, pp. 339-360.
- Follman, D.A., Goldberg, M.S. y May, L.** (1990). Personal characteristics, unemployment insurance, and the duration of unemployment. *J. Econometrics.*, **45**, pp. 351-366.
- Gelfand, A.E. and Smith, A.F.M.** (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, pp. 398-409.
- Rubin, D.B.** (1988). Using the SIR algorithm to simulate posterior distributions (with discussion). In *Bayesian Statistics 3* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith eds.), pp. 395-402. Oxford: Oxford University Press.
- Wei, G.C.G. and Tanner, M.A.** (1990). Calculating the content and boundary of the highest posterior density region via data augmentation. *Biometrika*, **77**, pp. 649-652.