MÉTODOS PARA EL ANÁLISIS DE DATOS BIOGRÁFICOS

COURGEAU Daniel[1]

Resumen

En este artículo se presentan tres enfoques para analizar datos biográficos. El enfoque no-paramétrico permite tratar las interacciones entre varios fenómenos demográficos y conduce a diversos tipos de dependencia entre ellos. El enfoque paramétrico busca introducir en el análisis la heterogeneidad observada en el seno de la población, relacionando los comportamientos temporales con diversas características dependientes o independientes del tiempo. El enfoque semi-paramétrico hace posible un tratamiento simultáneo de las interacciones y de la heterogeneidad. La cuestión de la heterogeneidad no observada en los datos se aborda también. Estos diversos métodos se ilustran por medio de ejemplos de aplicación con datos provenientes de encuestas.

*

* *

INTRODUCCIÓN

Mientras el análisis longitudinal clásico privilegió el estudio de un evento, los demás fueron considerados perturbadores. Ante esta situación, el análisis de datos biográficos trata de ver la manera cómo diversos eventos de la existencia influyen sobre el desarrollo posterior de la vida del individuo, y cómo algunas características llevan al individuo a comportarse de una manera diferente de los demás.

Este cambio de perspectiva conduce a formular las bases del análisis de datos biográficos en términos del análisis de procesos estocásticos complejos. Trabajos en probabilidad, realizados sobre todo en Francia, permitieron establecer firmemente este análisis sobre la teoría de las martingalas (Dellacherie y Meyer, 1980), la integración estocástica (Kunita y Watanabe, 1967 ; Dellacherie, 1980) y los procesos de conteo ("counting processes" en inglés, Bremaud y Jacod, 1977). No obstante, sólo presentamos aquí una visón simplificada de estos métodos y remitimos al lector interesado en sus fundamentos teóricos a la muy completa obra de Andersen et al. (1993).

[1] : Institut National d'Etudes Démographiques, 27 rue du Commandeur, 75675 Paris Cedex 14

Trataremos primero de ver cómo un evento familiar, económico o de otro tipo experimentado por un individuo va a modificar las probabilidades de ocurrencia de otros eventos de su existencia. Trataremos de examinar, por ejemplo, cómo el matrimonio puede influir sobre su carrera laboral, su vida familiar y su movilidad espacial. Esto nos llevará a generalizar el análisis demográfico clásico al estudio de las ***interacciones*** entre los eventos, gracias a los métodos ***de análisis no-paramétrico.***

Es preciso considerar luego ***la heterogeneidad*** en el seno de las poblaciones, la cual influye sobre las probabilidades de ocurrencia de un evento dado. Por ejemplo se puede pensar que el nivel educativo de un individuo o el orden de nacimiento entre hermanos influyen sobre la probabilidad de abandonar el agro, según la condición de empresario o de trabajador agrícola. Esto nos llevará a generalizar los métodos de regresión, de larga tradición en econometría, a los métodos de ***análisis paramétrico,*** que permiten incluir la heterogeneidad de las poblaciones de manera dinámica, en vez de estática.

Pero estos métodos conducen también a una representación paramétrica del tiempo de permanencia de los individuos en diferentes estados; por lo tanto sus resultados no son completamente satisfactorios. Preferimos un método que sintetiza los dos análisis anteriores: ***el análisis semi-paramétrico***. En efecto, esta técnica permite conservar la estimación no-paramétrica de los tiempos de permanencia, al tiempo que incluye el efecto paramétrico de diversas características. Veremos en particular que este método permite resolver en parte los problemas ligados a la heterogeneidad no observada. Ilustraremos esta presentación con aplicaciones a la encuesta «Biografía familiar, laboral y migratoria» (denominada más sencillamente «3B»), que se llevó a cabo en Francia durante 1981.

1 - EL ANÁLISIS NO-PARAMÉTRICO

Como lo indicamos antes, el desarrollo del análisis de datos biográficos considerados como un proceso estocástico conduce a una formulación probabilística, de la cual carecía completamente el análisis longitudinal clásico. Para presentar este enfoque, es preferible comenzar por el caso más simple, donde se estudia, al igual que en análisis clásico, la ocurrencia de un solo evento.

1.1 Formulación probabilística para un solo evento

Supongamos que estemos siguiendo una muestra de mujeres nacidas entre 1911 y 1935, como en la encuesta 3B, y que se observan las primeras uniones que ocurren durante el transcurso del tiempo[2]. Es evidente que se pueden extraer un gran número de muestras diferentes, que arrojan fechas de matrimonio distribuidas de diferentes

[2] : Aquí se supone que el único evento posible es el matrimonio; obviamente eliminaremos esta hipótesis más adelante.

maneras. Sin embargo, la distribución temporal de estas fechas no será completamente aleatoria, ya que proviene de una misma población. Se considera entonces la edad al matrimonio de estas mujeres cómo ***una variable aleatoria*** que puede tomar diversos valores (15, 16, 17....), desconocidos a priori, pero sobre los cuales la observación de una muestra nos suministra una información cuya precisión trataremos de medir.

El siguiente paso consiste en formalizar esta información con una terminología probabilista. La medición de la edad o de la duración en tiempo continuo o discreto distingue dos casos.

1.1.a Tiempo continuo

Sea T una variable aleatoria positiva o nula y continua (para este ejemplo: el momento del matrimonio, medido con gran precisión con respecto al nacimiento). Es posible definir varias funciones asociadas a esta variable aleatoria, algunas ya conocidas en demografía clásica.

La función de permanencia (en estado de celibato) generaliza la función de sobrevida o función de riesgo : es la probabilidad de que el individuo no haya experimentado el evento (el matrimonio en este caso) antes de una fecha dada t :

$$S(t) = P\,(T \geq t)$$

Se comprueba fácilmente que $S(0) = 1$ (todas las mujeres son solteras a la fecha de nacimiento) y que $S(\_) \geq 0$ (algunas mujeres permanecen solteras durante toda su vida). Sin embargo, siempre se puede colocar artificialmente una masa en el infinito (las mujeres solteras durante toda su vida), para que $S(\_) = 0$.

La ***densidad de probabilidad*** del evento estudiado es la derivada con respecto al tiempo de $S(t)$:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} = -\frac{d\,S(t)}{dt}$$

No existe ningún equivalente de esta densidad en demografía clásica.

Se prefiere la densidad condicional, más comúnmente llamada en demografía ***cociente instantáneo*** :

$$h(t) = \lim_{\Delta t \to 0} \frac{P(T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{d[\log S(t)]}{dt}$$

Por definición, este cociente no es necesariamente menor que uno.

Finalmente, introducimos el ***cociente acumulado,*** que tampoco tiene equivalente en demografía clásica, pero que es indispensable para trabajar con eventos múltiples (abandono del celibato por matrimonio o por unión libre, por ejemplo):

$$H(t) = \int_{x=0}^{t} h(x)dx = \int_{0}^{t} -\frac{d\log S(x)}{dx}dx = -\log S(t)$$

Se puede apreciar que una de estas diversas funciones basta para poder calcular las otras.

1.1.b Tiempo discreto

Sea T una variable aleatoria discreta que puede tomar los valores $t_1, t_2 \ldots$, en los cuales pueden ocurrir los eventos :

$$f(t_i) = P\,(T = t_i)$$

$$h(t_i) = P(T = t_i |\ T \geq t_i) = \frac{f(t_i)}{S(t_i)}$$

por lo tanto:

$$S(t_i) = \prod_{j=1}^{i-1}[1 - h(t_j)]$$

luego:

$$H(t_i) = -\log S(t_i)$$

una vez más, el cociente acumulado $H(t_i)$ no tiene equivalente en demografía clásica.

1.2 Estimación para un solo evento

Generalmente, la observación de la ocurrencia de un evento no es completa, y algunos individuos pueden dejar de ser observados antes de haber experimentado el evento. En el caso del matrimonio, por ejemplo, algunas mujeres solteras pueden salir de observación, y a pesar de ello, casarse después de la encuesta. En tales casos, hablaremos de datos ***censurados***. No discutiremos aquí los diversos tipos de censura, por lo que remitimos al lector interesado a la obra de Andersen et al., 1993, pp. 135-168.

Empecemos por suponer que se trabaja en tiempo discreto y que se observa en el momento t_i, a N_i individuos, de los cuales d_i experimentan el evento y m_i son

censurados. Aquí se supone que los eventos se producen antes de la salida de observación (censura).

Para estimar el cociente $h(t_i)$ correspondiente, se calcula la verosimilitud como función del cociente y se toma como estimación el valor del cociente que maximiza dicha verosimilitud. Su expresión es:

$$L_i = [h(t_i)]^{di}[1 - h(t_i)]^{Ni-di}$$

Se aprecia fácilmente que el máximo de L_i se presenta en los mismos valores de $h(t_i)$ que el máximo de log (L_i). Por lo tanto, es más sencillo maximizar la expresión:

$$\log L_i = d_i \log[h(t_i)] + (N_i - d_i)\log[1 - h(t_i)]$$

Se obtiene el máximo al anular la derivada con respecto a $h(t_i)$, lo que conduce al siguiente estimador:

$$\hat{h}(t_i) = \frac{d_i}{N_i}$$

Su varianza se calcula gracias a las propiedades asintóticas de $\sqrt{N_i}\,(\hat{h}(t_i) - h(t_i))$,

se trata del inverso de la matriz de información de Fisher, que en este caso es diagonal:

$$\text{var}[\hat{h}(t_i)] = \frac{d_i(N_i - d_i)}{N_i^3}$$

De allí se deduce el estimador $S(t_i)$, denominado de Kaplan y Meier :

$$\hat{S}(t_i) = \prod_{j=1}^{i}(1 - \frac{d_i}{N_i})$$

y el de $H(t_i)$:

$$\hat{H}(t_i) = -\log[\hat{S}(t_i)]$$

La varianza asintótica de $S(t_i)$ puede calcularse mediante la fórmula de Greenwood :

$$\text{var}[\hat{S}(t_i)] = (\hat{S}(t_i))^2 \sum_{j=1}^{i} \frac{d_j}{N_j(N_j - d_j)}$$

y la de $H(t_i)$ con:

$$\text{var}\left[\hat{H}(t_i)\right] = \sum_{j=1}^{i} \frac{d_j}{N_j(N_j - d_j)}$$

Si trabajamos en tiempo continuo, se requieren hipótesis suplementarias. Sigamos suponiendo que se observan d_i eventos y m_i salidas de observación en el intervalo anual $]t_{i-1}, t_i]$. Supongamos además que los cocientes instantáneos permanecen constantes en este intervalo, tanto para el evento estudiado $h(t_i)$, como para la salida de observación $C(t_i)$, y que los dos eventos sean independientes. Con esto podemos calcular las contribuciones a la verosimilitud de los grupos de individuos siguientes:

- los que pertenecen todavía a la muestra al final del intervalo, sin haber experimentado el evento;
- los que experimentan el evento;
- los que salen de observación.

La maximización del logaritmo de la verosimilitud con respecto a $h(t_i)$ y a $C(t_i)$ conduce a un sistema de ecuaciones de dos incógnitas, que permite estimar de la manera siguiente el cociente instantáneo (Courgeau y Lelièvre, 1989, pp 60-61 ; 1992 ; pp. 69-71).

$$\hat{h}(t_i) = -\frac{d_i}{d_i + m_i} \log\left(1 - \frac{d_i + m_i}{N_i}\right) \approx \frac{d_i}{N_i - \frac{1}{2}(d_i + m_i)}$$

Hay que tener en cuenta que este cociente instantáneo no es el mismo cociente demográfico clásico, el cual mide la probabilidad condicional de experimentar el evento a lo largo del intervalo en ausencia de censuras :

$$\hat{q}(t_i) = 1 - \exp(-\hat{h}(t_i)) \approx \frac{d_i}{N_i - \frac{1}{2} m_i}$$

Desde luego, es posible estimar $S(t_i)$ y $H(t_i)$, así como las varianzas respectivas de los diferentes estimadores.

Se puede eliminar las hipótesis que sustentan estas estimaciones, con el propósito de obtener estimadores más satisfactorios, que permitan analizar varios eventos en competición (riesgos múltiples).

1.3 Eventos competitivos

Supongamos ahora que un individuo expuesto al riesgo pueda salir del estado inicial con diferentes alternativas: mortalidad por causas, salida del estado de soltero hacia la unión libre o el matrimonio, etc.

Si se calculan los estimadores de Kaplan y Meier para cada causa, considerando a las otras causas como censuras, tenemos que suponer la independencia entre los diferentes riesgos; generalmente, esta hipótesis no se cumple. Para evitar este inconveniente, se puede reformular el problema en términos de procesos de conteo. Estos procesos permiten estudiar individuos expuestos a riesgos múltiples en tiempo continuo.

El proceso multivariado $N=((N_1(t), N_2(t), \ldots N_k(t) ; t)$ es una colección de k procesos de conteo que pueden ser mutuamente dependientes. Para cada proceso estocástico $N_i(t)$ se define un proceso de intensidad $\Lambda_i(t)$, como la probabilidad de ocurrencia del evento i en el intervalo $(t, t+\Delta t)$, que conoce el pasado del proceso, $F(t)$. Por lo tanto :

$$\Lambda_i(t) = \lim_{\Delta t \to 0} \frac{E\left[N_i(t+\Delta t) - N_i(t) \middle| F(t)\right]}{\Delta t}$$

Aalen [1978] desarrolló el modelo de intensidad multiplicativa, que expresa el proceso de intensidad de la siguiente manera :

$$\Lambda_i(t) = h_i(t)\, Y_i(t)$$

donde $Y_i(t)$ representa, por ejemplo, la población expuesta al riesgo de experimentar el evento i y $h_i(t)$ la intensidad de este evento para un individuo dado. Anotemos aquí que es posible definir $Y_i(t)$ de manera más compleja, por ejemplo cuando se estudia la migración entre áreas de un mismo país; en ese caso, $Y_i(t)$ puede representar el producto de las poblaciones de las dos áreas.

La intensidad acumulada está dada por la fórmula :

$$H_i(t) = \int_{x=0}^{t} h_i(x)dx$$

Se puede estimar esta intensidad acumulada bajo hipótesis muy poco restrictivas, utilizando la teoría de las martingalas y los procesos de conteo (para mayores detalles sobre esta estimación, ver : Andersen et al., pp. 176-331). Si $t_{i1} < t_{i2} < \ldots$son las fechas observadas para la ocurrencia del evento i en la población, el estimador de Nelson-Aalen de la intensidad acumulada es el siguiente :

$$\hat{H}_i(t) = \sum_{t_{ij} \le t} \frac{1}{\left[Y_i(t_{ij})\right]}$$

y el estimador de su varianza es :

$$Var(\hat{H}_i(t)) = \sum_{t_{ij} \le t} \frac{1}{\left[Y_i(t_{ij})\right]^2}$$

Si varios eventos ocurren en el instante t_{ij} (en esas condiciones ya no se tiene un proceso de conteo), se generalizan estas estimaciones reemplazando el uno del numerador por el número de eventos. Estas intensidades siempre están bien definidas, aun cuando los riesgos no sean independientes, y se puede comparar entre ellas las gráficas de los estimadores de Nelson-Aalen para los diversos riesgos (Aalen 1982, pp. 10-11). Sin embargo, se trata de riesgos en un contexto en el que las otras posibilidades de salida del estado inicial están presentes; generalmente es imposible calcular riesgos cuando las otras causas han desaparecido (Courgeau y Lelièvre, 1996, p 649).

Como veremos más adelante, estas gráficas también son útiles cuando se desea probar la validez de diversos modelos paramétricos.

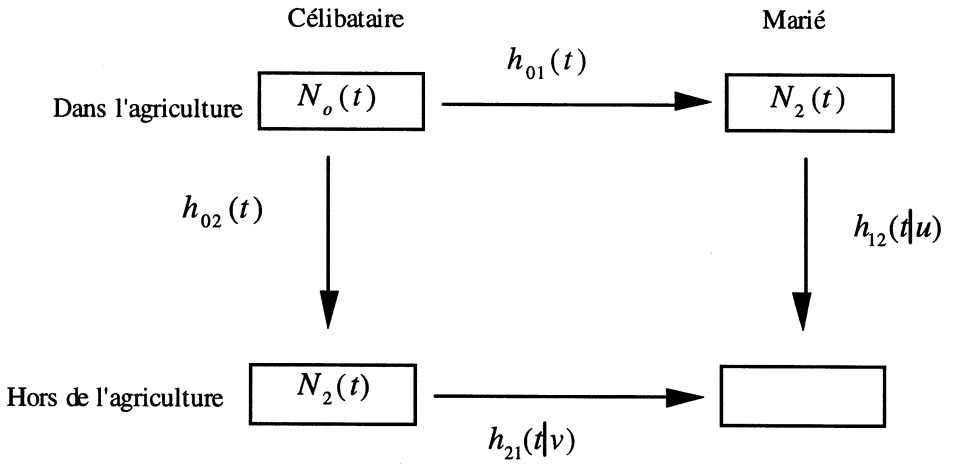
1.4 Interacciones entre eventos

Consideremos ahora el caso más general en el cual dos o más eventos pueden ocurrir e interactuar. Examinaremos el caso bivariado y remitimos al lector interesado en los casos más complejos a Courgeau y Lelièvre, 1989, pp. 85-91 ; 1992, pp. 99-106.

Tomemos como ejemplo el caso de las dependencias que pueden existir entre la salida del agro y el matrimonio de empresarios o trabajadores agrícolas. Otros estudios efectuados en Francia han mostrado que, con respecto a los demás, el celibato es mucho más común entre los hombres del agro. Cabe la duda de si los hombres que abandonan el medio agrícola conservan este fuerte celibato, o por el contrario, adoptan el comportamiento del resto de la población. Otra posible pregunta es si las mujeres casadas abandonan más frecuentemente el agro que las solteras. La respuesta a este tipo de preguntas requiere un estudio biográfico.

En todos los casos se van a observar dos fenómenos que intervienen en dominios diferentes de la vida del individuo, fenómenos cuyo orden de ocurrencia no está fijado de antemano: el matrimonio es posible antes de dejar el agro, o a lo contrario, se puede abandonar la agro antes de casarse. Por lo tanto, trataremos de ver si un cambio de estado matrimonial o residencial modifica la probabilidad de ocurrencia del otro evento.

Con el propósito de formalizar la situación, la ilustramos con la siguiente gráfica :



Gráfica 1 : Representación de la interacción entre el matrimonio y el abandono de la agricultura

Al comienzo, todos los individuos están en la situación 0 (solteros del agro); de acuerdo con su evolución, ellos pueden decidir pasar a la situación 1 (casados del agro) o a la situación 2 (solteros no pertenecientes al agro), para poder llegar a la situación final (casados no pertenecientes al agro), alcanzada sólo por algunos miembros de la población inicial[3].

Introduzcamos ahora dos variables T_1 y T_2, las edades de ocurrencia del matrimonio y de la salida del agro. Se observa que los cocientes se pueden expresar como las siguientes probabilidades condicionales :

$$h_{01}(t) = \lim_{\Delta t \to 0} \frac{P(T_1 < t + \Delta t \mid T_1 \geq t, T_2 \geq t)}{\Delta t}$$

$$h_{21}(t|v) = \lim_{\Delta t \to 0} \frac{P(T_1 < t + \Delta t \mid T_1 \geq t, T_2 = v)}{\Delta t} \quad donde \quad v < t$$

Para $h_{02}(t)$ et $h_{12}(t|u)$. tenemos fórmulas análogas.

Supongamos ahora que los cocientes h_{21} y h_{12} no dependen de la edad de ocurrencia del otro evento[4]. Se puede demostrar que si los eventos correspondientes a los diferentes cocientes son iguales a $n_{01}(t)$, $n_{02}(t)$, $n_{21}(t)$, $n_{12}(t)$, entonces se tienen los estimadores siguientes:

[3] : En este ejemplo, se descartan los casos de simultaneidad.

[4] : Es fácil eliminar esta hipótesis y calcular las series de cocientes para los diferentes edades de ocurrencia del otro evento.

$$\hat{h}_{01}(t) = \frac{n_{01}(t)}{N_0(t) - \frac{1}{2}(n_{01}(t) + n_{02}(t))} \; ; \; \hat{h}_{21}(t) = \frac{n_{21}(t)}{N_2(t) - \frac{1}{2}(n_{21}(t) - n_{02}(t))}$$

con fórmulas análogas para $\hat{h}_{02}(t)$ et $\hat{h}_{12}(t)$.

A partir de esta serie de diferentes cocientes, es interesante probar, por ejemplo, la igualdad de $\hat{h}_{01}(t)$ et de $\hat{h}_{21}(t)$, o la de los cocientes acumulados correspondientes. Si se cumple para todo t, podremos concluir que la salida del agro no influye para nada en las probabilidades de matrimonio de los individuos; en caso contrario, se podrá probar el signo de la diferencia. Para ello, podemos calcular la cantidad:

$$D = \frac{\hat{h}_{01}(t) - \hat{h}_{21}(t)}{\sqrt{\frac{\hat{h}_{01}(t)}{\hat{Y}_0(t)} + \frac{\hat{h}_{21}(t)}{\hat{Y}_2(t)}}}$$

Donde $\hat{Y}_0(t)$ et $\hat{Y}_2(t)$ son los denominadores de $\hat{h}_{01}(t)$ et $\hat{h}_{21}(t)$ dados anteriormente.

Si $\hat{h}_{01}(t)$ no difiere significativamente de $\hat{h}_{21}(t)$, entonces la variable D tendrá una distribución normal de media 0 y desviación estándar 1. Se puede operar de la misma manera con los cocientes acumulados.

Como ejemplo, incluimos en las figuras 1 y 2 las curvas que representan los cocientes acumulados de nupcialidad de las mujeres según su situación laboral (en el agro o fuera de él) y los cocientes acumulados de salida del agro, según su estado civil. Se aprecia claramente, y las pruebas lo confirman, que la salida del agro no tiene ninguna incidencia sobre la nupcialidad, sino que por el contrario, las mujeres casadas en el medio agrícola tienen mayores probabilidades de permanecer en él. Se trata, por lo tanto, de una *dependencia unidireccional* que evidencia una estrategia : el matrimonio con un agricultor va a determinar la permanencia en el mundo agrícola. También es interesante observar (Courgeau y Lelièvre, 1986) que para los hombres, la dependencia unilateral es opuesta a la de las mujeres: la probabilidad de casarse se duplica cuando dejan el agro; por el contrario, el hecho de ser o no casado no influye sobre la salida del agro.

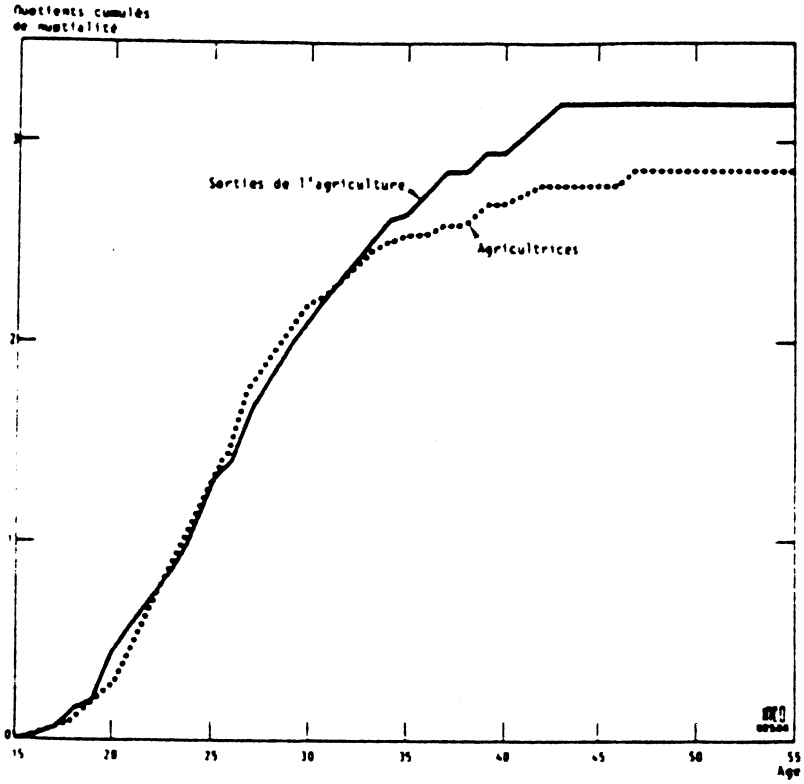


figura 1 : cocientes acumulados de nupcialidad de las mujeres según su situación laboral (en el agro : .... , fuera del agro : —)

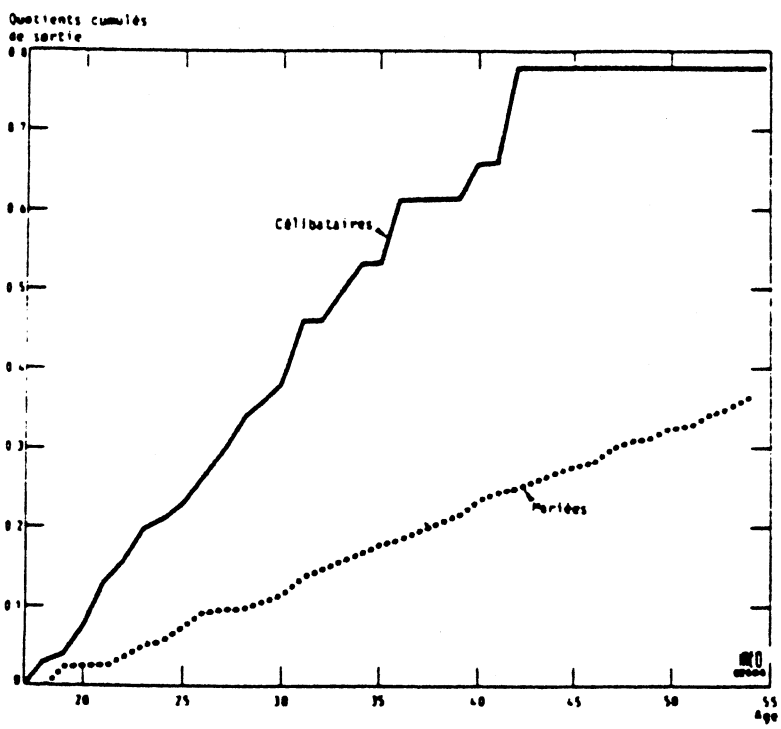


figura 2: cocientes acumulados de salida del agro de las mujeres según su estado civil (solteras : — , casadas : ....)

Otros tipos de dependencia pueden aparecer en estos análisis. La más frecuente es la *dependencia bidireccional:* el haber experimentado un evento afecta la probabilidad de experimentar otro y viceversa. Esto es lo que ocurre con la dependencia entre fecundidad y migración hacia las metrópolis (Courgeau, 1987, 1989). Para los nacimientos de orden mayor que 2, la encuesta 3B muestra que la migración hacia la metrópoli va a reducir la probabilidad de tener un hijo más, mientras que la migración en sentido contrario la incrementa. Una *dependencia a priori* también puede aparecer en algunos casos. Cuando se aísla a la población que va a emigrar hacia la metrópoli antes de los 50 años, aparece una baja fecundidad, tanto antes como después de la migración. Por lo tanto, estamos ante un fenómeno de *selección,* en la población no metropolitana, de personas con baja fecundidad que emigran hacia las grandes ciudades sin cambiar su patrón de fecundidad. A la inversa, cuando se aísla en una metrópoli a la población femenina que va a emigrar, se encuentra que, antes de emigrar ,estas mujeres tienen una fecundidad igual a la de las que permanecen, pero después de la migración, la fecundidad aumenta. Esta vez estamos frente a un fenómeno de *adaptación* provocada por la migración, ya que estas mujeres adoptan el comportamiento de las sedentarias del medio no metropolitano.

Las numerosas interacciones estudiadas en la encuesta 3B no arrojaron ningún caso de *independencia total* entre eventos. Esto demuestra el peligro que se corre cuando se tratan fenómenos demográficos separadamente, bajo la hipótesis de independencia mutua.

2. EL ANÁLISIS PARAMÉTRICO

La heterogeneidad de las poblaciones va a introducir una propensión diferencial a experimentar los diferentes eventos demográficos. Según el nivel educativo, el orden de nacimiento entre hermanos, el tamaño de su familia de origen, etc., los individuos pueden tener probabilidades diferentes de casarse, migrar, cambiar de empleo, etc. Por lo tanto, se podría tratar de resolver este problema descomponiendo la población en grupos cada vez más homogéneos (por sexo, nivel educativo y región de origen, etc.) pero, en contrapartida, el número de casos en cada grupo se irá reduciendo. Este procedimiento conduce rápidamente a unas poblaciones expuestas al riesgo muy pequeñas , desde luego, a resultados no significativos.

Lo anterior plantea la necesidad de introducir diversas características individuales en un modelo de tipo regresivo, sin dejar por ello de trabajar con el conjunto de la población o con un subgrupo lo suficientemente numeroso. Ahora bien, en este modelo se debe también introducir el tiempo, por medio de una función base paramétrica que aproxime correctamente la distribución no paramétrica estimada para el conjunto de la población. Un método semejante permite incluir un gran número de características, pero será necesario verificar las hipótesis en las cuales se fundamenta el modelo.

2.1 Algunos modelos paramétricos

Entre los numerosos modelos paramétricos que permiten describir correctamente el comportamiento de una población o de algún subgrupo de ella, conviene escoger aquellos que dependen de un mínimo de parámetros, pero permiten al mismo tiempo un buen ajuste a la distribución no-paramétrica que se puede estimar. Presentaremos aquí algunos de ellos y remitimos al lector interesado en una presentación más detallada a Courgeau y Lelièvre, 1989, pp. 95-125 ; 1992, pp. 109-144.

2.1.a Distribución exponencial

Esta es la distribución más simple que se obtiene cuando el cociente instantáneo permanece constante a lo largo del tiempo. En ese caso, tenemos:

$$h(t) = \rho;\ \ S(t) = \exp(-\rho t)$$
$$f(t) = \rho \exp(-\rho t);\ \ H(t) = \rho t$$

Una prueba fácil de esta distribución consiste en representar el cociente acumulado no-paramétrico en función del tiempo. Esta curva debe aproximarse a una recta si la distribución es exponencial.

Es muy frecuente que una distribución exponencial se ajuste bien a un número reducido de periodos, a pesar de que no sea válida para el conjunto de periodos. En estos casos es posible descomponer la distribución y estimar los parámetros ρ correspondientes.

2.1.b Distribución de Gompertz

Esta distribución se aplica frecuentemente en demografía, tanto para la movilidad espacial y laboral, como para la mortalidad (en edades avanzadas). El cociente instantáneo es una función exponencial del tiempo. Se expresa como sigue:

$$h(t) = \lambda\rho \exp(\rho t)\ \ ;\ S(t) = \exp(\lambda[1 - \exp(\rho t)]$$
$$f(t) = \lambda\rho \exp\{\rho t + \lambda[1 - \exp(\rho t)]\}\ ;\ H(t) = \lambda[\exp(\rho t) - 1]$$

Los cocientes son uniformemente crecientes $(\rho > 0)$, o uniformemente decrecientes $(\rho < 0)$. Una prueba de esta distribución se obtiene con la representación del logaritmo del cociente en función de t. Cuando se trabaja con migraciones de diversos rangos o con la movilidad laboral, el parámetro ρ es generalmente negativo. En este caso una parte de la población $(\exp \lambda)$ sigue siendo sedentaria.

El parámetro ρ es positivo cuando se considera la mortalidad de las personas de edad. En este caso toda la población desaparece.

2.1.c Distribución log-logística

Frecuentemente, esta distribución es una mejor alternativa que la distribución log-normal , cuyos parámetros son más difíciles de estimar. En estos casos se tiene:

$$h(t) = \lambda\rho(\rho t)^{\lambda-1}\left[1+(\rho t)^{\lambda}\right]^{1};\ S(t) = \left[1+(\rho t)^{\lambda}\right]^{1}$$

$$f(t) = \lambda\rho(\rho t)^{\lambda-1}\left[1+(\rho t)^{\lambda}\right]^{2};\ H(t) = \log\left[1+(\rho t)^{\lambda}\right]$$

Por lo tanto, una prueba de esta distribución consiste en representar log(exp *(H(t))* - *1)* en función de log *t*. Una relación lineal comprueba que la distribución log-logística es válida. Este tipo de distribución se utiliza a menudo para la nupcialidad y la fecundidad según el rango de nacimiento, ya que si $\lambda>1$ la curva de nacimientos tiene un máximo.

2.2 Modelos de regresión

Incluyamos ahora diversas características de los individuos, que podemos representar con un vector Z :

$$Z = (Z_1, ..., Z_n)$$

Frecuentemente, estas características son variables cualitativas con valores dicotómicos (1 si el individuo posee la característica, 0 de lo contrario), pero también pueden ser variables cuantitativas (número de hermanos y hermanas, número de migraciones realizadas durante la infancia, etc.). Estas características pueden depender del tiempo (variable igual a 0 mientras que el individuo no posee la característica, para convertirse en 1 desde el momento de su adquisición)

Dos grandes grupos de modelos describen el efecto que ejercen estas características sobre los cocientes instantáneos.

2.2.a Modelos de riesgos proporcionales

La hipótesis que fundamenta estos modelos es la siguiente: las diferentes características individuales actúan en forma multiplicativa sobre un cociente base, el mismo para toda la población, y al cual se le dará una forma paramétrica. Esto conduce a un cociente instantáneo de la forma:

$$h(t;Z) = h_0(t)\varphi(Z,\beta)$$

donde $h_0(t)$ es una función paramétrica de t, que depende de un cierto número de parámetros aunque independiente de las características Z, y $\varphi(Z,\beta)$ es una función de las características Z cuyos parámetros β son independientes de t. La función utilizada con mayor frecuencia es exponencial:

$$\varphi(Z,\beta) = \exp(Z\beta) = \exp(Z_1\beta_1 + Z_2\beta_2 + ... + Z_n\beta_n)$$

Se observa fácilmente que cuando todas las variables Z son nulas, el modelo se reduce a la expresión paramétrica del cociente base:

$$h(t;0) = h_0(t)$$

Se observa fácilmente que cuando todas las variables Z son nulas, el modelo se reduce a la expresión paramétrica del cociente base:

$h(t;0) = h_0(t)$

si sólo la variable Z_l es igual a 1 siendo todas las demás 0, se observa que:

$h(t;0) = h_0(t)\exp\beta_1$

De allí se deduce la relación siguiente:

$$\frac{h(t;Z)}{h(t;0)} = \exp\beta_1,$$

que es por lo tanto independiente de la duración. Así se puede comprobar la validez del modelo de riesgos proporcionales. Basta dividir la población en dos subpoblaciones, una que posea la característica Z_1 y otra que no la posea, para luego estimar de manera no-paramétrica para cada una de ellas *h(t)* o mejor aun *H(t)* . Las gráficas de log *(H(t))* en función de *t* deben ser paralelas para que el modelo se cumpla. De esta manera se puede probar el modelo de riesgos proporcionales para cada una de las características consideradas por separado.

La figura 3 representa los logaritmos de los cocientes acumulados del primer cambio de residencia después de la salida del hogar de los padres, según que la vivienda inicial sea o no gratuita (encuesta 3B). La comprobación exitosa del paralelismo de las curvas permite utilizar el modelo de riesgos proporcionales.

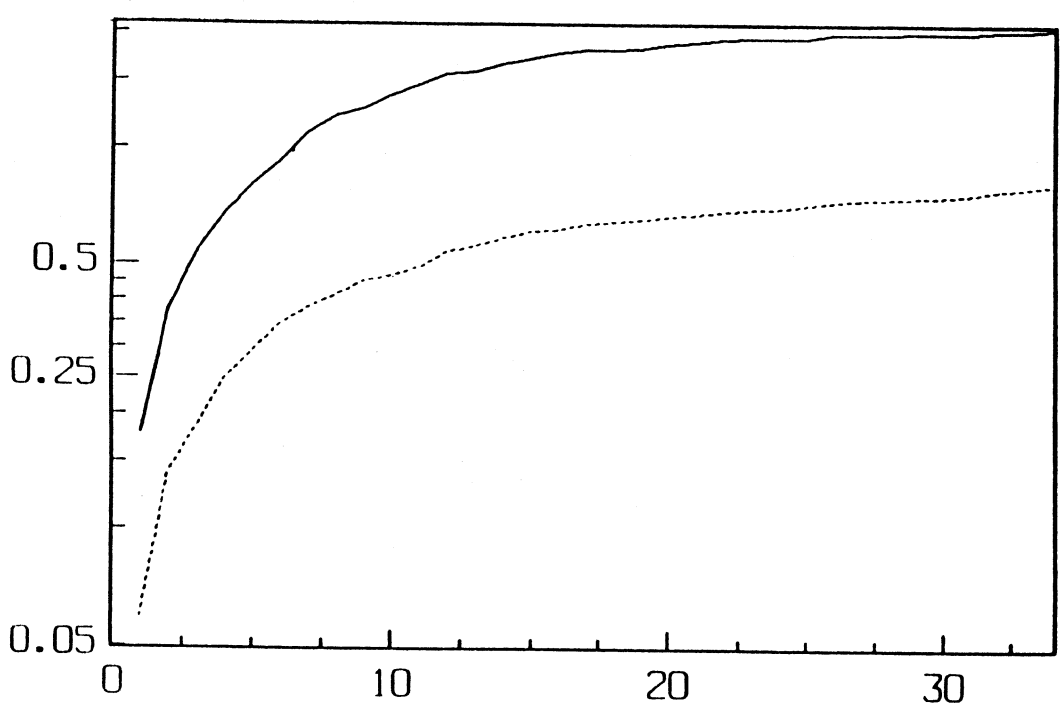


Figura 3 : logaritmos de los cocientes acumulados del primer cambio de residencia después de la salida del hogar de los padres (viviendas gratuitas : —, otras viviendas : ...)

2.2.b Modelos con tiempo de salida acelerado

Supongamos ahora que las características influyen directamente sobre la función de permanencia, en vez de sobre el cociente instantáneo. En ese caso, un individuo tipo con todas sus variables Z nulas tiene una función de permanencia igual a $S_o(t)$; el individuo que posea todas las características Z tendrá la siguiente función de permanencia:

$$S(t;Z) = S_0(t \exp Z\beta)$$

Se puede demostrar que en este caso la expresión del cociente instantáneo es:

$$h(t;Z) = h_0(t \exp Z\beta) \exp Z\beta$$

Si $h_0(t)$ es una distribución de Gompertz se tiene:

$$h(t;Z) = \lambda\rho \exp(\rho t \exp Z\beta) \exp Z\beta = \lambda\rho \exp(Z\beta + \rho t \exp Z\beta)$$

mientras que para un modelo de riesgos proporcionales la expresión será:

$$h(t;Z) = \lambda\rho \exp(Z\beta + \rho t)$$

Por lo consiguiente, los dos cocientes no tienen la misma forma.

Este modelo puede expresarse también en términos de variables aleatorias:

$$T = T_0 \exp(-Z\beta)$$

donde T_0 es la duración de la permanencia de un individuo con características todas nulas. Esta relación puede expresarse también así:

$$\log T = \log T_0 - Z\beta$$

2.3 Estimación de los modelos paramétricos

Para estimar correctamente el modelo, es necesario incluir de nuevo los datos censurados. Con este propósito representamos el conjunto de los datos recolectados bajo forma de una tripla:

$$(t_i^o, \delta_i, Z_i) \quad i = 1, \ldots, n$$

donde t_i^o es la duración de la observación, ya sea hasta la ocurrencia del evento, en caso de éste se observe (δ_i =1), o bien hasta el momento de la censura, antes de que el evento ocurra (δ_i = 0).

2.3.a Cálculo de la verosimilitud

Consideremos el caso en que la censura es independiente de que el individuo haya experimentado o no el evento: esto es lo que de hecho sucede cuando se realiza una encuesta retrospectiva, ya que la fecha de la encuesta es independiente de la historia de vida individual. Introduzcamos la variable aleatoria T^s, que corresponde a la censura, cuya función de permanencia es $0_i(t)$ y densidad de probabilidad $q_i(t)$. Esta variable es independiente de la fecha de ocurrencia del evento estudiado T, cuya función de permanencia es $S(t;Z_i)$ y densidad de probabilidad $f(t;Z_i)$. La variable observada es $T^0 = \min(T^s, T)$.

Calculemos la probabilidad siguiente para los individuos que han experimentado el evento:

$$P(t \le T_i^o < t + \Delta t, \delta_i = 1; Z_i) = 0_i(t) f(t;Z_i) \Delta t$$

para los individuos que no lo hayan hecho tenemos:

$$P(t \le T_i^o < t + \Delta t, \delta_i = 0\ ; Z_i) = q_i(t) S(t;Z_i) \Delta t$$

Como $0_i(t)$ y $q_i(t)$ no suministran ninguna información sobre β, se puede considerar que la verosimilitud es proporcional al valor:

$$L(\beta) = \prod_{i=1}^{n} f(t_i;Z_i)^{\delta_i} S(t_i;Z_i)^{1-\delta_i} = \prod_{i=1}^{n} h(t_i;Z_i)^{\delta_i} S(t_i;Z_i),$$

donde $h(t\ ;Z)$ es el cociente instantáneo de ocurrencia del evento.

En estas condiciones se vuelve posible, gracias al método de máxima verosimilitud, estimar los diversos parámetros, tanto los de la distribución paramétrica supuesta, como los efectos de las diferentes características individuales.

2.3.b Estimación y prueba de los parámetros

Expresemos ahora la verosimilitud de las observaciones en su forma logarítmica:

$$\log L(\beta) = \sum_{i=1}^{n} (\delta_i \log h(t_i;Z_i) + \log S(t_i;Z_i))$$

El método consiste en dar a los parámetros β los valores que maximizan la verosimilitud. Para ello, basta calcular las derivadas de $\log L(\beta)$ con respecto a los diferentes parámetros e igualarlas a 0. Se tendrá entonces un sistema con tantas incógnitas como ecuaciones. Se demuestra entonces que, bajo unas condiciones simples generalmente satisfechas ($L(\beta)$ debe ser tres veces diferenciable y deben cumplirse ciertas condiciones límites en la tercera derivada), este sistema tiene solución. La estimación de parámetros que se obtiene de esta manera es asintóticamente insesgada y de varianza mínima. La distribución asintótica del estimador es una normal con tantas

variables como parámetros por estimar. La media de esta ley es el verdadero valor de β. Si se calcula el negativo de la matriz de las segundas derivadas de Log $L(\beta)$, llamada matriz de información de Fisher, se demuestra que dicha matriz es un estimador de la matriz de covarianza de los parámetros β.

La solución del sistema :

$$\frac{d \log L(\beta)}{d\beta} = U(\beta) = 0$$

se complica rápidamente con el aumento del número de parámetros, y puede carecer de una solución analítica simple. Por esta razón se utilizan métodos numéricos para obtener soluciones aproximadas. El algoritmo de Newton-Raphson[5] es entonces un método de uso común.

Habiendo resuelto el sistema, se pueden realizar diferentes pruebas sobre los parámetros estimados $\hat{\beta}$. Se puede probar, por ejemplo, que los parámetros estimados son diferentes de los valores β_o que se habían escogido a priori. Si se designa a la matriz de covarianza como $V(\beta)$, podemos plantear el estadístico siguiente (el superíndice T representa el vector transpuesto) :

$$(\hat{\beta} - \beta_o)^T V(\beta)^{-1} (\hat{\beta} - \beta_o)$$

Cuando los parámetros $\hat{\beta}$ no son significativamente diferentes de β_o, este estadístico tiene distribución χ^2 con tantos grados de libertad como parámetros por estimar.

También se puede utilizar directamente la verosimilitud al calcular la fracción :

$$R(\beta_o) = \frac{L(\beta_o)}{L(\hat{\beta})}$$

Se demuestra en efecto que la distribución asintótica de $(-2 \log R(\beta_o))$ es una χ^2 con tantos grados de libertad como parámetros. Finalmente, también es posible emplear $U(\beta_o)$. Cuando $\beta=\beta_o$, el estadístico $U(\beta_o)$ es asintóticamente normal, de media 0 y varianza $V(\beta_o)$. En estas condiciones, el estadístico :

$$U^T(\beta_o) V(\beta_o)^{-1} U(\beta_o)$$

es asintóticamente distribuido como una χ^2 con tantos grados de libertad como parámetros.

Traigamos como ejemplo el estudio de la duración de permanencia en la vivienda de hombres nacidos entre 1931 y 1935, proveniente también de la encuesta 3B (Courgeau, 1985). En este caso se estiman diferentes modelos de Gompertz, que

[5] Para mayores detalles sobre este método, ver Courgeau y Lelièvre, 1989, p. 137 ; 1992, pp. 157-158

incluyen un número creciente de características. El número de duraciones de permanencia es de 2 523, de las cuales 493 se mantienen todavía en el momento de la encuesta en 1981.

Comparemos con el modelo exponencial (modelo de Gompertz, en el cual $\rho=0$ y $\lambda\rho \rightarrow C^{te}$). Un modelo de este tipo conduce a un cociente constante, estimado en 0,1237, con -6 273,15 como valor máximo de verosimilitud. Incluyamos ahora los diferentes grupos de edad, así como la duración de permanencia. Esto conduce a un nuevo máximo igual a -5 991,65. Utilizando la fracción de las verosimilitudes, se obtiene - 2 log R = 562,96, lo cual muestra que los efectos de la edad y de la duración de permanencia son efectivamente significativos.

El cuadro 1 contiene la estimación de los diferentes parámetros β, con sus desviaciones estándar, el estadístico que permite probar los efectos de diferentes parámetros $\left(\frac{\beta_i^2}{V(\beta_i)}\right)$ y los valores de exp(β_i).

Cuadro 1 - Estimación de los parámetros β, de la desviación estándar, prueba del χ^2 con un grado de libertad de la nulidad de los β y estimación de exp (β_i).

Característica considerada	$\hat{\beta}_i$	$\sigma(\hat{\beta}_i)$	$\left[\frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)}\right]^2$	exp β_i
Constante	- 2,727	0,2880	89,187	0,0654
Menos de 20 años	1,131	0,2920	15,008	3,100
20-24 años	1,410	0,2912	23,460	4,097
25-29 años	0,892	0,2940	9,203	2,440
30-34 años	0,626	0,2978	4,417	1,870
35-39 años	0,135	0,3064	0,194	1,145
40-44 años	- 0,177	0,3292	0,127	0,889

El cuadro muestra un efecto muy significativo de todos los grupos de edad inferiores a 35 años sobre la probabilidad de migrar. La curva que da el efecto multiplicativo de cada edad es idéntica a la curva clásica de migración por edad.

También existe un efecto significativo de la duración de permanencia con un parámetro $\hat{\rho}$ = -0.0629 y una desviación estándar igual a 0,0045. Así, la probabilidad de migrar se reduce casi a la mitad después de 10 años.

La adición de nuevas características mejora la calidad del modelo: características familiares (log L = -5963,17), estatus de ocupación de la vivienda (log L = -5755,45), características laborales (log L = -5685,59), acontecimientos políticos y origen de los padres (log L = -5637,67). De esta manera se llega a un modelo que recurre a 32 características diferentes para explicar el comportamiento migratorio de esta población.

Se podría pensar que algunas características incorporadas en el camino, correlacionadas con la edad, explican mejor el comportamiento que la edad de por sí. En este caso, el efecto de la edad se vería atenuado en los modelos donde se incluyen de manera más precisa las demás características. En efecto, la movilidad de un individuo casado se reduce al 80% de lo que es para un individuo soltero, y más joven en promedio; la movilidad de un individuo que posee su vivienda se reduce al 20% de lo que era para un arrendatario, también más joven en promedio, etc.

Cuando el modelo incluye todas las características, el efecto de la edad ya no es significativo, como lo muestra el cuadro 2 :

Cuadro 2 - Efecto de los grupos de edad al comienzo de la permanencia, cuando el modelo incluye simultáneamente todas las variables

Grupos de edad	$\hat{\beta}_i$	$\sigma(\hat{\beta}_i)$	$\left[\frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)}\right]^2$	exp β_i
Menos de 20 años	0,234	0,298	0,619	1,264
20-24 años	0,201	0,295	0,466	1,223
25-29 años	0,014	0,294	0,002	1,014
30-34 años	- 0,160	0,297	0,293	0,852
35-39 años	- 0,399	0,303	1,741	0,671

Se constata que el factor verdaderamente explicativo de los cambios migratorios va a ser la situación familiar, laboral, etc. en vez de la edad misma. Se puede afirmar que la edad influiría sólo como una variable de reemplazo, cuyo efecto desaparece completamente desde el momento en que se incluyen las características actuantes, a las cuales está ligado.

2.4 La heterogeneidad no observada

Los modelos presentados hasta ahora sólo pueden incluir características observadas en la encuesta. Así se haya tenido el mayor cuidado posible para tratar de

captar todas las causas de variación de los cocientes, se puede pensar que otras características, más difíciles de observar o de medir en una encuesta, o aun características que a juicio del investigador no pueden influir sobre el evento estudiado, tengan de hecho un efecto no despreciable sobre los cocientes.

En ciertos casos puede existir una dependencia entre las características observadas y la heterogeneidad no observada, como lo vimos en el estudio anterior sobre las migraciones: los eventos que influyen sobre la migración están correlacionados con la edad. Mientras que no se los observe, el efecto de la edad sigue siendo significativo debido a esta correlación. Una vez incluidos se puede eliminar la edad del modelo, sin que esto afecte su validez. Por lo consiguiente, son cuestiones de pertinencia de las características observadas y de especificación correcta del modelo las que influyen en este caso.

En otros casos, las características observadas son independientes de la heterogeneidad no observada, la cual puede entonces afectar la estimación de los parámetros relativos a los efectos de las características observadas. Si bien es cierto que, para los modelos lineales con varianza homosedástica, la omisión de variables no correlacionadas con las variables incluidas no tiene consecuencia alguna sobre las estimaciones de mínimos cuadrados, esta característica no se mantiene en el caso general. Mientras que no se incluyan en el modelo todas las variables explicativas, los métodos de máxima verosimilitud no suministran generalmente una estimación correcta , aun si estas variables no observadas son totalmente no correlacionadas con las variables observadas.

A pesar de la falta de información sobre estas variables omitidas, algunos investigadores han tratado de introducir una distribución paramétrica, o aun, no paramétrica de la heterogeneidad no observada. Se puede demostrar la viabilidad de considerar el efecto de esta heterogeneidad de manera multiplicativa y de estimar los valores de nuevos parámetros correspondientes a las características observadas, las cuales tienen en cuenta la heterogeneidad no observada. Se encuentra que según la distribución supuesta de esta heterogeneidad (Heckman y Singer, 1984), o aun según la distribución paramétrica utilizada para estimar el efecto de las características observadas (Trussel y Richards, 1985), los parámetros estimados pueden variar enormemente, al punto de ser de signos opuestos.

Estos resultados llevan a preferir el enfoque semi-paramétrico, para el que se dispone de resultados más precisos con respecto al efecto de la heterogeneidad no observada sobre la estimación de los parámetros.

3. EL ANÁLISIS SEMI-PARAMÉTRICO

El enfoque semi-paramétrico busca liberarse de la hipótesis paramétrica relativa a la forma de la distribución de la función de riesgos, pero conservando la del efecto multiplicativo de las características individuales (modelos de riesgo proporcional). La expresión del modelo propuesto es:

$$h(t;Z) = h_0(t)\exp(Z\beta)$$

donde $h_o(t)$ es una función no-paramétrica de t, llamada cociente instantáneo base. De hecho, es el cociente instantáneo del individuo que posee todas sus características Z iguales a 0.

3.1 Estimación de los parámetros

Los parámetros β se estiman por medio del cálculo de una verosimilitud parcial, empleada por Cox (1972). Condicionalmente al conocimiento de la población expuesta al riesgo y al hecho que una ocurrencia del evento se produjo en t_i,, la probabilidad de que el individuo i experimente el evento es igual a :

$$\frac{h_0(t_i)\exp(Z_i\beta)}{\underset{\ell\in R_{ii}}{Z}\ h_0(t_i)\exp\,(Z_\ell\beta)} = \frac{\exp(Z_i\beta)}{\sum_{\ell\in R_i}\exp(Z_\ell\beta)}$$

donde R_i es el conjunto de los individuos expuestos al riesgo en t_i *-0*. Es evidente que esta verosimilitud parcial no incluye ya más el cociente instantáneo base. Al efectuar el producto de estas probabilidades sobre toda la población se obtiene la verosimilitud parcial por maximizar :

$$PL(\beta) = \frac{\exp\sum_{i=1}^{n} Z_i\beta}{\prod_{i=1}^{n}\left[\sum_{\ell\in R_i}\exp(Z_\ell\beta)\right]}$$

Es conveniente anotar que esta verosimilitud parcial no es una verosimilitud en el sentido habitual, puesto que no es proporcional a la probabilidad condicional (o marginal) de los eventos observados. A pesar de ello, su maximización conduce a estimadores asintóticamente insesgados y normalmente distribuidos. Asintóticamente, la matriz de covarianza es el inverso del negativo de la matriz de las segundas derivadas de la verosimilitud. Estas estimaciones, que han sido objeto de numerosas controversias entre estadísticos, están ahora bien sustentadas en el marco de la teoría de martingalas y de los procesos de conteo (Andersen et al., 1993, pp. 476-591).

3.2 Estimación de la función base

Ahora queda por estimar la componente no-paramétrica del modelo. Observemos primero que la función de permanencia puede expresarse así:

$$S(t;Z) = S_o(t)^{\exp(Z\beta)},$$

donde $S_0(t)$ es la función base no-paramétrica.

Sean $t_1 < t_2 < ... < t_R$ las ocurrencias de eventos observados. En el intervalo $[t_i, t_{i+1}[$, algunos individuos experimentan el evento t. Su contribución a la verosimilitud es por lo tanto:

$$S_0(t_i)^{\exp(Z\beta)} - S_0(t_i + 0)^{\exp(Z\beta)}$$

y notamos T_i el conjunto de estos individuos. La contribución a la verosimilitud de un individuo censurado durante el intervalo es:

$$S_0(t_\ell + 0)^{\exp(Z\beta)}$$

y notamos M_i el conjunto de ellos.

La expresión de la verosimilitud es entonces:

$$L = \prod_{i=1}^{R} \left\{ \prod_{\ell \in T_i} (S_0(t_i)^{\exp Z\ell\beta} - S_0(t_i + 0)^{\exp(Z_\ell\beta)}) \prod_{\ell \in M_i} S_0(t_\ell + 0)^{\exp(Z\ell\beta)} \right\}$$

La maximización de esta verosimilitud, con β conocido, permite estimar la función base $S_0(t)$[6]

3.3 Un ejemplo de estimación

Prosigamos con el análisis no-paramétrico iniciado en 1.4, de las interacciones entre el matrimonio y la salida del agro para las mujeres. Sólo consideraremos aquí el abandono de medio agrícola antes o después del matrimonio e incluiremos diferentes características individuales. Con este propósito, apelaremos al modelo semi-paramétrico siguiente:

$$h(t; Z; Z') = h_0(t)\exp(Z\beta_1 + H(t-u)(\beta_0 + Z\beta_2 + Z'\beta_2')$$

donde

$$H(x) = \begin{cases} 0 & Si\ x<0 \\ 1 & Si\ x \geq 0 \end{cases}$$

u es la fecha de matrimonio, Z el vector de características que influyen antes y después del matrimonio, Z' el de las características que influyen sólo después del matrimonio (características del cónyuge, por ejemplo). Esta formulación supone que el cociente base se multiplica por una constante $\exp\beta_0$ cuando la mujer contrae matrimonio: en 1.4 comprobamos que esto era correcto. El cuadro 3 muestra el modelo óptimo obtenido con las características observadas.

[6] : Para mayores detalles sobre esta estimación, ver Courgeau y Lelièvre, 1989, pp. 159-161 ; 1992, pp. 185-187.

Cuadro 3 - Modelo óptimo de salida del agro para las mujeres (Valor de los parámetros)

Conjunto de variables	Efecto principal β1	Perturbación β0	Interacción β2
Número de hermanos/hermanas	0,012 **		0,000
Hermana mayor	- 0,320 **		0,296
Padre agricultor	- 0,928 **		0,806 *
Matrimonio		**- 0,228**	
mujer del agro al momento del matrimonio			- 1,040
Esposo agricultor			- 0,359 **
Suegro agricultor			- 0,126
* Resultado significativo al 10 %	** Resultado significativo al 5%		

El examen de este cuadro permite precisar las diferentes características de las mujeres propensas o no a abandonar el agro. Así por ejemplo, la mayor de dos hijos de un agricultor, tiene un factor multiplicativo del riesgo base de salida del agro de exp (0,012 - 0,32 - 0,928) = 0,290, mientras que la menor de los cinco hijos de un trabajador agrícola tiene un factor multiplicativo de exp (4 x 0,012) = 1,048. La probabilidad de que ésta última deje el agro es por lo tanto 3,85 veces mayor que la de la primera. Después del matrimonio, las mujeres que siguen permaneciendo en el agro tienen características muy marcadas. Son las hijas mayores de familias poco numerosas, de padres agricultoras y casadas con un agricultor. Es evidente que estas condiciones van a favorecer la reunión de las tierras de las dos familias. La estrategia de las otras mujeres, que constituyen la mayoría, será abandonar el agro.

3.4 La heterogeneidad no observada

Como el modelo semi-paramétrico no contiene ninguna especificación paramétrica del cociente base, es necesario volver a examinar el efecto de la heterogeneidad no observada.

Un trabajo teórico de Bretagnole y Huber-Carol (1988) estudió de qué manera la omisión de características afectaba los parámetros estimados para las características observadas en un modelo de este tipo. Cuando se trabaja sobre datos con posibles censuras a la derecha y si la heterogeneidad no observada es independiente de las características observadas, esta omisión no afecta el signo de los parámetros estimados, pero provoca una disminución de sus valores absolutos. Por lo tanto, si el efecto de una característica parecía importante cuando se omitían otras características. la inclusión de éstas últimas en el modelo sólo puede reforzar el efecto de la primera. Por el contrario, algunas características sin efectos importantes a primera vista pueden volverse muy significativas al incluir otras características no observadas en principio.

Este resultado permite precisar mejor el efecto de la heterogeneidad no observada. Si la heterogeneidad y las características observadas son independientes, también nos asegura de la validez del signo de los parámetros estimados.

CONCLUSIONES

Los métodos de análisis de datos biográficos presentados nos permitieron alcanzar los principales objetivos mencionados en la introducción.

Los modelos no-paramétricos permiten estimar probabilidades de transición de un estado a otro, cuando los diferentes fenómenos demográficos estudiados interactúan. Esto permite poner en evidencia dependencias de diverso tipo: unilaterales, bilaterales, totales, a priori, etc.

Los modelos paramétricos permiten apreciar cómo influyen diferentes características individuales sobre las probabilidades de transición de un estado a otro. Estos modelos requieren hipótesis más restrictivas que las de los modelos no-paramétricos. Por lo consiguiente, es indispensable verificar su validez para los datos estudiados. Los modelos de riesgos proporcionales o de tiempos de salida acelerados que presentamos aquí no son la única alternativa posible. Se les puede reemplazar por otros que se adapten mejor a unas situaciones particulares. La heterogeneidad no observada también plantea otros problemas difíciles de resolver.

Los modelos semi-paramétricos constituyen la mejor solución cuando se desea incluir simultáneamente las interacciones entre fenómenos y la heterogeneidad de la población. En esa situación, se han podido resolver algunos de los problemas planteados por la heterogeneidad no observada.

Se puede afirmar que la recolección y el análisis de biografías abren un amplísimo campo de investigación, inscrito a su vez dentro de una corriente más general que reagrupa a la totalidad de las ciencias sociales. Generalizaciones sencillas de estos métodos permiten el estudio de situaciones más complejas. El análisis de modelos multi-niveles (Courgeau, 1994, 1995b, 1996) o el análisis de estructuras sociales complejas como la familia o el grupo doméstico (Courgeau, 1995a) pueden ser abordados con métodos derivados de los aquí presentados. En este sentido, dichos métodos trascienden el nivel del individuo para avanzar hacia una comprensión más profunda de las sociedades humanas.

BIBLIOGRAFÍA

AALEN O., 1978, *Nonparametric inference for a family of counting processes,* The Annals of Statistics, 6,4, pp 701-726.

AALEN O., 1982, *Practical applications of the nonparametric statistical theory for counting processes,* Statistical Research Report, n° 2, Institute of Mathematics, University of Oslo, 60p.

ANDERSEN P.K., BORGAN Ø., GILL R., KEIDING N., 1993, *Statistical Models Based on Counting Processes,* Springer Verlag, New-York, VII + 768p.

BREMAUD P., JACOD J., 1977, *Processus poncuels et martingales : résultats récents sur la modélisation et le filtrage*, Advanced Applied Probabilities, 9, pp 362-416.

BRETAGNOLLE J., HUBER-CAROL C., 1988, *Effects of omitting covariates in Cox's model for survival data,* Scandinavian Journal of Statistics, 15, pp 125-138.

COURGEAU D., 1985 c, *Interaction between spatial mobility, family and career life-cycle : A French survey*, European Sociological Review, 1, n° 2, pp 139-162.

COURGEAU D., 1987, *Constitution de la famille et urbanisation*, Population, 42, n° 1; pp 57-82.

COURGEAU D., 1989, *Family formation and urbanization*, Population, English selection, n° 1, pp 123-146.

COURGEAU D., 1994, *Du groupe à l'individu : l'exemple des comportements migratoires,* Population, 49, 1, pp 7-26.

COURGEAU D., 1995 a, *Event history analysis of household formation and dissolution,* in Household Demography and Household Modelling, van Imhoff, Kuijsten, Hooimeijer, van Wissen eds, Plenum Press, New-York, pp 185-202.

COURGEAU D., 1995 b, *From the group to the individual : what can be learned from migratory behaviour,* Population : an English Selection, 7, pp 145-162.

COURGEAU D., 1996, *Towards a multilevel analysis in social science/vers une anlyse multi-niveaux en sciences sociales,* in Spatial Analysis of Biodemographic Data, John Libbey - INED, pp 9-22.

COURGEAU D., LELIEVRE E., 1986, *Nuptialité et agriculture,* Population, 41, n° 2, pp 303-326.

COURGEAU D., LELIEVRE E., 1989, *Analyse démographique des biographies,* Editions de l'INED, Paris, 270 p.

COURGEAU D., LELIEVRE E., 1992, *Event history analysis in demography,* Clarendon Press, Oxford, 226 p.

COURGEAU D., LELIEVRE E., 1996, *Changement de paradigme en démographie,* Population, 51, n° 3, pp 645-654.

COX D., 1972, *Regression models and life tables,* Journal of the Royal Statistical Society, B 34, pp 187-220.

DELLACHERIE C., 1980, *Un survol de la théorie de l'intégrale stochastique,* Stochastic Processes Applications, 10, pp 115-144.

DELLACHERIE C., MEYER P.A., 1980, *Probabilités et potentiels : Théorie des martingales,* Hernan, Paris, 476 p.

HECKMAN J., SINGER B., 1984, *A method for minimizing the impact of distributional assumptions in ecometric models for duration data,* Econometrica, 52, 2, pp 271-320.

KUNITA H., WATANABE S., 1967, *On square integrable martingales,* Nagoya Mathematic Journal, 30, pp 209-245.

TRUSSELL J., RICHARDS T., 1985, *Correcting for immeasured heterogenetic in hazard models using the Heckman-Singer procedure,* in Sociological Methodology, Tuma ed., Jossey Bass, pp 242-278.