CUERPO ACADÉMICO DE PROBABILIDAD Y ESTADÍSTICA BUAP

Decimocuarta Semana Internacional de la Estadística y la Probabilidad
14-18 de junio de 2021

Estimación de un modelo de simulación medíante computación bayesiana aproximada

Juan Ignacio Baccino [a], **Mauro Loprete** [a], **Daniel Ciganda** [b]

[a] *Instituto de Estadística, Facultad de Ciencias Económicas y de Administración Universidad de la República (Udelar)*
[a] nacho.baccino.3323@gmail.com, [a] mauroloprete1@gmail.com

[b] Max Planck Institute for Demographic Research
[b] ciganda@demogr.mpg.de

Resumen

El objetivo de este trabajo es estimar los parámetros de un modelo computacional del comportamiento reproductivo, en ausencia de una expresión analítica para la función de verosimilitud. Para esto se utilizan técnicas de Computación Bayesiana Aproximada (ABC) y se analiza la incertidumbre asociada a las predicciones del modelo. Se trabaja con el modelo Comfert, un modelo de microsimulación que modela las trayectorias reproductivas de una cohorte de mujeres en un régimen de fecundidad natural, es decir, en ausencia de intentos dirigidos a prevenir nacimientos. Con estas trayectorias simuladas se obtienen las tasas específicas de fecundidad por edad para dicha cohorte y se utiliza esta información para ajustar el modelo a las tasas observadas en una población histórica. Los datos utilizados provienen de una cohorte de Huteritas, una comunidad anabaptista frecuentemente estudiada en demografía por su rechazo del uso de métodos anticonceptivos y el nivel elevado de su fecundidad. Los resultados obtenidos ilustran la utilidad del enfoque bayesiano en la realización de inferencia estadística sobre modelos computacionales.

Keywords: Modelos Computacionales, Computación Bayesiana Aproximada, Fecundidad, Huteritas.

1. Introducción

El desarrollo de métodos de estimación libres de verosimilitud es uno de los principales factores detrás del crecimiento en el uso de modelos computacionales en varias disciplinas. El objetivo de estos métodos es encontrar la combinación de valores de parámetros que logran el mejor ajuste de datos simulados con los observados.

Estos métodos son útiles en contextos de inferencia cuando no se tiene una forma explicita o es imposible trabajar con la función de verosimilutd, pero si es posible simular resultados utilizando el modelo en cuestión. En su forma más elemental el algoritmo ABC puede expresarse en el siguiente pseudocódigo

A simple ABC pseudocode

> Calculate the model on a sample from the prior distribution
>
> Obtain efficient samples in parameter space
> Compute Δ_{θ_j}
>
> Return:
> A fraction p of θ values with lowest Δ_θ
> and this is the approximate posteriori distribution

De esta forma es posible reconstruir una aproximación a la distribución a posteriori de los parámetros.

El presente trabajo tiene como objetivo utilizar el enfoque ABC para estimar un modelo demográfico de microsimulación para el que no es posble derivar un expresión de la función de verosimilitud, pero desde el

que es posible simular datos que pueden compararse con los datos provenientes de una población histórica.

El modelo con el que trabajamos, *Comfert*, simula las trayectorias reproductivas de una cohorte de mujeres en un régimen de fecundidad natural. La idea de fecundidad natural en Demografía refiere a una población en la que no existe control explícito de la natalidad.

Para ajustar el modelo se utilizan una serie de tasas específicas de fecundidad por edad de una cohorte de Huteritas, una comunidad religiosa Anabaptista con un estilo de vida tradicional que incluye el rechazo al uso de métodos anticonceptivos.

2. Metodología

2.1. Medición de la Fecundidad. Se entiende como proceso reproductivo a la secuencia de nacimientos de una mujer y edades de la madre al nacimiento. A partir de esta secuencia es posible calcular los indicadores del nivel de la fecundidad más frecuentemente utilizados, como las tasas específicas de fecundidad por edad, definidas como:

$$ {}_nF_x\,[0,T] = \frac{{}_nB_x\,[0,T]}{{}_nL_x\,[0,T]} $$

Siendo el numerador la cantidad de nacimientos de mujeres de edades x a x + n sobre la poblaciónón promedio (en T años) de este grupo de mujeres. Como se puede ver, esta tasa mide la cantidad promedio de hijos por mujer a la edad x.

En la figura 1 se puede ver las tasas especificas por edad para los Hutteritas, resultando en una Tasa Global de Fecundidad (suma de las tasas especificas) de 10,67.

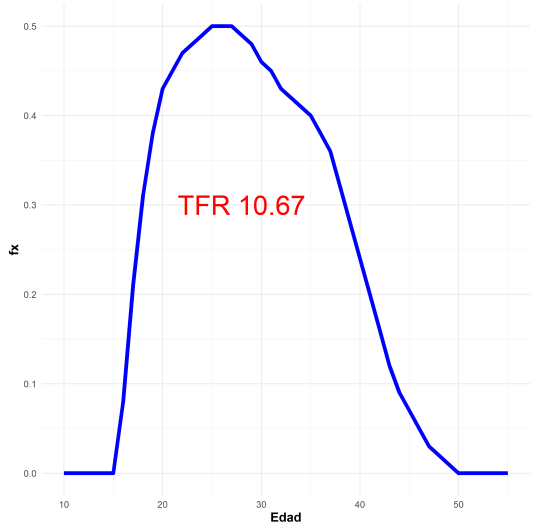

Fig. 1. Tasa de fecundidad especifica por edad y tasa global de fecundidad

A la hora de modelar las tasas de fecundidad por edad, algunos de los modelos más frecuentemente utilizados son el de *Coale-Trussell*, los *Splines cuadráticos* ó el modelo de *Gompertz*.

En el modelo de *Coale-Trussel* se plantea una función respecto a la edad en términos de la fecundidad natural y un descuento considerando un control sobre la fecundidad por una proporción de matrimonios a la edad x

$$ f(x) = cG_0\left(\frac{x-\mu}{\sigma}\right) Mn(x)e^{-mv(x)} $$

Mientras que el modelo de *Gompertz* propone una transformación de la fecundidad acumulada a una determinada edad usando una transformación $Log-Log$

$$ Y(x) = -log\left\{-log\left(\frac{F(x)}{F}\right)\right\} $$

Por último, se plantea el modelo de *Spline cuadrático* propuesto por Schmertmann (2003) el cual considera cuatro parámetros principales α(edad comienzo proceso reproductivo),P (Modo de la serie), H (fecundidad cae al menos la mitad del nivel máximo) y R (Nivel máximo de fecundidad).

$$ f(x) = R\sum \Theta_k\,(x-\varepsilon_k)^2 $$

Lo que tienen en común todos estos modelos es que intentan ajustar la curva delineada por las tasas específicas de fecundidad, es decir, son modelos macro, que ajustan cantidades a nivel agregado. El modelo que utilizamos para este trabajo, por otro lado, está definido a nivel individual, es decir, simula eventos en el curso de vida de las personas y los comportamientos que dan lugar a dichos eventos. Dicho de otro modo, las tasas son generadas a partir de la simulación de múltiples trayectoria reproductivas correspondientes a una cohorte ficticia de mujeres.

Al estar implementado a nivel individual el modelo tiene mayor flexibilidad y capacidad explicativa que los modelos macro, donde los parámmteros muchas veces no son directamente interpretables. Sin embargo, esto hace también que el modelo sea más complejo y que no sea posible recurrir a técnicas más establecidas de estimación como las basadas en la maximización de la función de verosimilitud.

La edad media de la unión es sumamente importante ya que se utiliza como inicio del proceso reproductivo para una mujer, esta es incluida en el modelo *Comfert* mediante una distribución $Log-Normal$ de parámetros la edad media de la unión y una varianza. Además de ser una distribución con asimetría positiva,tiene un soporte en los $\mathbb{R}^+$. Al ser un modelo de microsimulacion, todas las mujeres no tendrán el mismo riesgo de concebir a una determinada edad ni en el correr del tiempo, es por esto que se define heterogeneidad de fecundabilidad entre mujeres y en el tiempo.

La primera refiere a que el deseo de una mujer en edad fértil difiere a la de sus pares, para esto se utiliza una distribución de probabilidad si definimos a ϕ como la proporción de concepción en el primer mes, entonces $\phi \sim \Gamma(gshape\ ,\ grate)$.

En cambio la heterogeneidad en el tiempo es obtenida mediante la expresión

$$h(x,\alpha,\kappa) = (1+\exp\{-\kappa(x-\alpha)\})^{-1}$$

Siendo α la Edad de Inflexión del declive de la fecundidad y κ la tasa de decilve de la fecunidad.

Por último se define un periodo de no susceptibilidad que refiere al periodo de tiempo en los que la madre luego del nacimiento donde se vuelve a incluir el riesgo de concepción,en este caso es fijo y equivale a 6 meses.

2.2. Datos. Una de las dificultades que se presenta al utilizar el concepto de fecundidad natural es la recolección de datos y encontrar poblaciones que la practiquen. Una de las población en particular en la cual no se practica el control de la natalidad o que el control sea tan limitado que no modifique la fecunidad natural es la población conocida como los Hutteritas, más precisamente la población Hutteritas de los Estados Unidos y Cánada, la cual fue utilizada como datos observados para estimar los párametros de nuestro modelo. Las características que presentan los Hutteritas hacen un verdadero experimento demográfico que permiten establecer parámetros de reproducción natural como se utiliza a la hora de modelar la fecundidad en este trabajo. La colonia original de los Huteritas en EE.UU. partió con 443 habitantes (221 varones y 222 mujeres), en el censo de 1880, y llegó a formar 93 colonias con un total de 8.542 habitantes en 1950. En la década 1940–1950 la población norteamericana aumentó en un 14,5 % y los Hutteritas en un 52,1 %. Se calcula que, actualmente, los Hutteritas se duplican cada 16 años.

2.3. Método ABC. El método consiste en dar una distribución a priori a los parámetros de interés, en nuestro caso α y κ de manera uniforme en sus respectivos intervalos. En nuestro caso se utiliza un muestreo LHS (Latin Hipercube Sampling) de tal manera de conseguir una muestra para realizar una búsqueda secuencial eficiente del espacio de parámetros. Con el sub-espacio de parámetros obtenido se evalúa el modelo y se calcula el error cuadrático medio de los datos simulados respecto a los observados, para luego mediante un proceso de simulación gaussiano y un remuestreo de los parámetros originales, generar nuevas evaluaciones más eficientes, con un tamaño de simulaciones definida de antemano. La simulación de la salida del modelo es costosa computacionalmente y con la utilización de estos algoritmos podemos enfocarnos en la región del espacio de parámetros donde la distancia entre los datos simulados y observados tiende a ser menor.

Pseudocódigo ABC

Obtain an an initial sample of $\mathscr{X}$ of size n_0, $\theta_1,...,\theta_{n_0}$
Compute $\Delta_{\theta_1},...,\Delta_{\theta_{n_0}}$
Set $n = n_0$

While $n \leq N$ **do**
Map the relationship $\theta \to \Delta_\theta$ with a Gaussian Process emulator $G(\cdot)$
Obtain an new sample of $\mathscr{X}$ of size n^*
Obtain predictions for $G(\theta_k) \qquad k = 1,....,n^*$
Compute acquisition function $A(\cdot)$
Obtain the new locations to be explored θ_j
Compute Δ_{θ_j}
Increment n

end While

Return:
The value of θ that minimizes Δ_θ; or
A fraction p of θ values with lowest Δ_θ

Una vez que se obtienen las simulaciones, se consideran aquellas que no sobrepasan una tolerancia o distancia respecto a los datos observados, en nuestro caso se consideró un valor del 10 % mas pequeño respecto al *mse*. A modo de resumen se presentan las primeras 6 filas (sin ordenar) de las 52 simulaciones que pasaron esta regla.

α	κ	mse
36.05930	0.3251051	0.0008657
36.62916	0.3646193	0.0011001
36.21106	0.3531512	0.0008155
35.54010	0.3580796	0.0007798
35.88977	0.3371296	0.0008744
33.87930	0.2250504	0.0010580

Tabla 1: Tabla con primeras 6 de 52 filas de valores aceptados.

Una vez que se consiguen las combinaciones de parámetros aceptados estamos frente a la distribución marginal a posteriori de cada parámetro y con esto podemos hacer inferencias sobre los mismos. En nuestro caso la cantidad de simulaciones aceptadas fueron 52

(pares de parámetros). Una vez obtenida la distribución a posteriorí de cada parámetro volvimos a simular el modelo pero en los intervalos modales y ver la aleatoriedad de las estimaciones.

3. Resultados y discusión

Para representar la incertidumbre en la estimación se llevaron adelante los siguientes pasos: en primer lugar, se consideraron las 52 simulaciones que se generaron con los valores de α y κ aceptados. Con estos, se constuyó un intervalo de credibilidad para la Tasa fecundidad por Edad para lograr estudiar que con una probabilidad del 95 % se encuentre comprendido en él.

A continuación, se presenta una gráfica con la serie de la Tasa de Fecundidad por Edad observada (azul), la simulación con menor error cuadrático (lineas) y el intervalo anteriormente mencionado

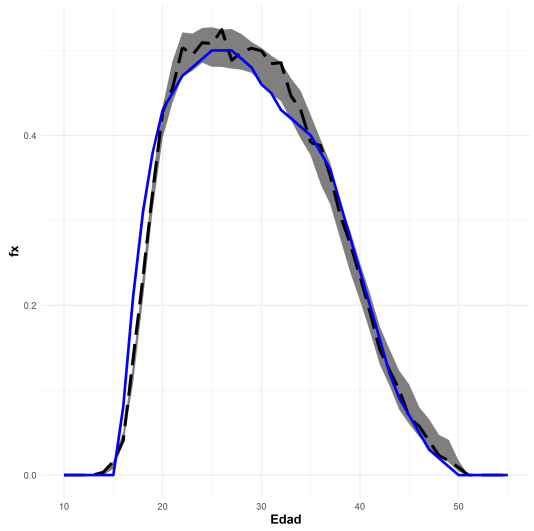

Fig. 2. Intervalo de Credibilidad al 95 % para los 52 pares de parámetros aceptados

Para hacer un análisis más profundo acerca de la incertidumbre se decidió volver a simular el modelo pero en el intervalo modal de cada una de las distribuciones marginales a posteriori de los parámetros. Observando la distribución a posteriori se aprecio que el intervalo modal de α está entre 35 y 35.5, mientras que para κ se encuentra entre 0.30 y 0.32. Se volvió a estudiar el modelo en los intervalos anteriormente mencionados modificando el intervalo a priori.

Con los resultados obtenidos se volvió a graficar la serie de las tasas específicas de fecundidad por edad observad.Se tomó la misma serie generada por la simulación mostrada anteriormente con su respectivo intervalo de credibilidad al 95 % y también se incluyó la nueva serie con menor error cuadrático medio y su intervalo de credibilidad asociado, que es el resultado de un volver a aplicar el algoritmo ABC, considerando como distribución a priori el intervalo modal de ambos parámetros de forma uniforme.

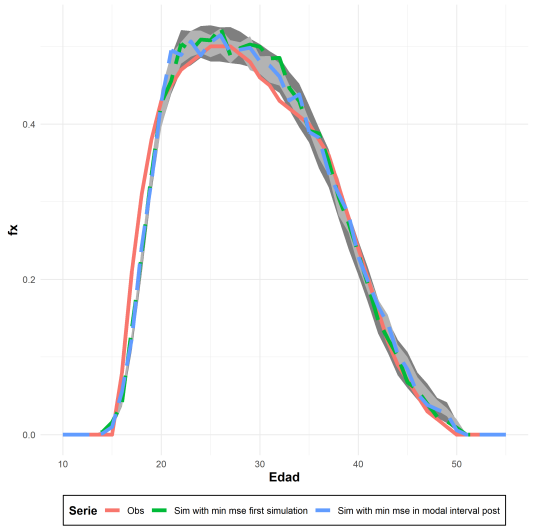

Fig. 3. Intervalo de Credibilidad al 95 % tomando como posteriori el intervalo modal y superponiendo con el I.Credibilidad anterior

4. Conclusiones y pasos a futuro

Como conclusión general, construyendo diferentes intervalos de credibilidad para la Tasa de Fecundidad por Edad, se logró evaluar de forma correcta el desempeño del método ABC para generar estimaciones en el modelo Comfert. La incertidumbre de las predicciones se encuentra contenida en los intervalos para la mayoría de las edades.

En próximos trabajos se buscara mejorar la aplicación de estas herramientas para lograr una estimación correcta para todo el rango de edades, en concreto tasas de 17 a 19 años, ya que en la sección anterior se puede observar que para dichas edades el modelo no ajusta bien.

También hay que considerar que en el presente trabajo solo dos de los parámetros se encuentran libres,el próximo paso sería estimar el modelo completo mediante ABC, para tener una mejor compresión de los determinantes de la Tasa de Fecundidad.

5. Referencias

Articulos

[2] Nicolas Todd Daniel Ciganda. "Demographic Models of the Reproductive Process: Past, Interlude, and Future". En: ().

[5] Michael U. Gutmann y Jukka Corander. "Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models". En: (2015). arXiv: 1501.03291 [stat.ML].

[7] Louis Henry. "Some data on natural fertility". En: *Eugenics Quarterly* 8.2 (1961), págs. 81-91. DOI: 10.1080/19485565.1961.9987465.

[8] Anastasia Kostaki y Peristera Paraskevi. "Modeling fertility in modern populations". En: *Demographic Research* 16 (2007), págs. 141-194. DOI: 10.4054/demres.2007.16.6.

[10] Carl Schmertmann. "A system of model fertility schedules with graphically intuitive parameters". En: *Demographic Research* 9 (2003), págs. 81-110. DOI: 10.4054/demres.2003.9.5.

[11] Mindel C. Sheps. "An Analysis of Reproductive Patterns in an American Isolate". En: *Population Studies* 19.1 (1965), pág. 65. DOI: 10.2307/2173165.

Paquetes de R

[1] Stefan Milton Bache y Hadley Wickham. *magrittr: A Forward-Pipe Operator for R*. R package version 2.0.1. 2020. URL: https://CRAN.R-project.org/package=magrittr.

[3] Matt Dowle y Arun Srinivasan. *data.table: Extension of 'data.frame'*. R package version 1.14.0. 2021. URL: https://CRAN.R-project.org/package=data.table.

[4] Mark Fairbanks. *tidytable: Tidy Interface to data.table*. R package version 0.6.1. 2021. URL: https://github.com/markfairbanks/tidytable.

[6] Lionel Henry y Hadley Wickham. *purrr: Functional Programming Tools*. R package version 0.3.4. 2020. URL: https://CRAN.R-project.org/package=purrr.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[12] Hadley Wickham. *forcats: Tools for Working with Categorical Variables (Factors)*. R package version 0.5.1. 2021. URL: https://CRAN.R-project.org/package=forcats.

[13] Hadley Wickham. *tidyr: Tidy Messy Data*. R package version 1.1.3. 2021. URL: https://CRAN.R-project.org/package=tidyr.

[14] Hadley Wickham y col. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.5. 2021. URL: https://CRAN.R-project.org/package=dplyr.

[15] Hadley Wickham y col. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.3.3. 2020. URL: https://CRAN.R-project.org/package=ggplot2.

[16] Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.31. 2021. URL: https://yihui.org/knitr/.