CAPE
CUERPO ACADÉMICO DE PROBABILIDAD Y ESTADÍSTICA BUAP

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

Análisis de la incertidumbre en la estimación de un modelo demográfico mediante técnicas de computación bayesiana aproximada

ALVARO VALIÑO [a], LUCÍA COUDET [a], DANIEL CIGANDA [a]

[a] *Instituto de Estadística, Departamento de Métodos Cuantitativos, Facultad de Ciencias Económicas y de Administración, Universidad de la República, Montevideo, Uruguay*
[a] alvaro3432@gmail.com, [b] luciacoudet@gmail.com, [c] ciganda@demogr.mpg.de

Resumen

En este trabajo se presenta un estudio de la incertidumbre en la predicción de un modelo demográfico que simula las tasas específicas de fecundidad por edad (*ASFR*) para una población en la cual no existe control sobre los nacimientos.

El modelo se estima mediante técnicas de Computación Bayesiana Aproximada (*ABC*) sobre datos de una comunidad religiosa conocida por su rechazo de los métodos anticonceptivos. En particular, se considera el efecto de la tolerancia utilizada para aceptar valores de los parámetros en la incertidumbre resultante en la estimación.

Para ello, se realizó primero una estimación puntual considerada óptima como aquella que minimiza el Error Cuadrático Medio (*ECM*) y luego se construyeron diferentes intervalos de credibilidad. Si bien todos se encuentran construidos al 95% de credibilidad, la incertidumbre varía notablemente en función de los niveles de tolerancia considerados.

Como principal resultado es posible mencionar que se observó una relación positiva entre el nivel de incertidumbre y el incremento en la tolerancia del algoritmo. Por último, se presentan pasos a seguir para futuras investigaciones.

Keywords: fecundidad, proceso reproductivo, simulación, enfoque bayesiano.

1. Introduction

El objetivo de éste trabajo es representar la incertidumbre en la estimación de las tasas específicas de fecundidad por edad (*ASFR*) provenientes de una población en la que no existe control sobre los nacimientos. Es decir, en régimen de fecundidad natural.

Para ello, se considera la incertidumbre sobre los parámetros que controlan el descenso en el riesgo de concebir con la edad: α y κ.

- α edad en la que decae la fecundidad
- κ tasa para dicho α

Con el fin de lograr la reproducibilidad de los resultados obtenidos, se consideró apropiado utilizar un repositorio remoto público. [1].

[1] https://github.com/alvarovalinio/ABC_bayes

Se destaca que los resultados obtenidos fueron a través del lenguaje y entorno de programación para análisis estadístico y gráfico, R.

2. Marco metodológico

En primer lugar, las tasas específicas de fecundidad se definen como:

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

Dónde

- ${}_nB_x[0,T]$ es la cantidad de nacimientos de mujeres entre las edades x y x+n en el periodo de tiempo 0 a T.
- ${}_nL_x[0,T]$ es la cantidad de mujeres entre el periodo de tiempo 0 a T.

Con el fin de modelar las *ASFR*, se utilizó el modelo *Comfert*, un modelo computacional a nivel micro, que representa las trayectorias reproductivas de una cohorte de mujeres en un contexto de fecunidad "natural", es decir, en ausencia de control de los nacimientos.

Es importante destacar que uno de los supuestos base del modelo es que las mujeres están expuestas al riesgo de concebir inmediatamente después del inicio de una unión con cohabitación. De esta forma, la formación de la unión marca el inicio del proceso reproductivo. El tiempo de espera a este evento se modela de la siguiente manera:

$$ln(U) \sim N(\mu, \sigma)$$

El modelo también considera la heterogeneidad con respecto a la capacidad de concebir tanto entre mujeres como a través de tiempo. Éste último, incorporando al análisis el declive en la capacidad biológica de concebir debido al efecto del paso del tiempo. La heterogeneidad en la tasa de fecundidad se modela entonces como:

$$\phi \sim \Gamma(shape, rate)$$

A su vez, como se mencionó anteriormente, se asume un régimen de fecundidad natural, el cual implica la no utilización de ningún tipo de métodos anticonceptivos. Por otra parte, el modelo *Comfert* incluye periodos de no suceptibilidad de la madre luego del nacimiento.

De este forma, los componentes principales del modelo son:

1. Inicio $ln(U) \sim N(\mu, \sigma)$,
2. Heterogeneidad de la Tasa de fecundidad $\phi \sim \Gamma(shape, rate) \rightarrow h(x, \alpha, \kappa) = \frac{1}{1+e\{-\kappa*(x-\alpha)\}}$ y
3. Periodo de no-suceptibilidad

Sin embargo, a la hora de estimar los parámetros y debido a la incapacidad de estimar la función de verosimilitud, se procedió a trabajar con métodos de estimación que no dependen de dicha función (*Likelihood free*). En particular desde un enfoque bayesiano, utilizando la Computación bayesiana aproximada (*ABC*).

A continuación se presenta una breve descripción del algoritmo *ABC*:

Sean: 1) y_0 un dato observado, 2) $p(\theta)$ distribución a priori, 3) $p(y|\theta)$ la función de verosimilitud de $y|\theta$, 4) una medida de discrepancia (Δ), y 5) un valor de tolerancia ε.

Algoritmo:

1. Obtener θ^* de $P(\theta)$
2. Simular $P(y|\theta*)$

   if $y_{sim} = y_0$ (discrete data) $\Delta(\eta_0, \eta_{sim}) < \varepsilon$ (continuos data)

   $\rightarrow \theta^*$ forma parte de la posterior;

   else

   descartamos θ^*
3. Repetir 1 y 2 hasta que se tenga un número suficiente de valores para θ.

Debido al elevado costo computacional que implica obtener estimaciones a partir del modelo *Comfert*, se procedió a trabajar con una versión adaptada del algoritmo *ABC*.

Para ello, se ajusta un meta modelo (en particular un proceso gaussiano) que es utilizado como sustituto del modelo *Comfert*. A su vez, con el fin de optimizar la búsqueda de combinaciones de parámetros pero sin condicionar sustantivamente los resultados (zonas sin explorar) se procedió a utilizar la función denominada *acquisition function*.

De esta manera, los puntos pertenecientes al subconjunto del espacio paramétrico obtenido son evaluados como posibles candidatos de la distribución a posteriori $p(\theta|y)$.

Una vez seleccionados los candidatos, se procedió a estimar la densidad de la predictiva posterior ($p(\tilde{y}|y)$), obteniendo así para cada edad, estimaciones de la distribución de probabilidad a posteriori de la *ASFR*.

Como estimación puntual se toma la estimación resultante de elegir el conjunto de parámetros que minimiza el error cuadrático medio (*ECM*):

$$\text{ECM} = \frac{\sum_{i=1}^{n}(y_s - y_{obs})^2}{n}$$

Dónde:

- y_s son los valores simulados de las *ASFR* para cada edad, utilizando la metodología *ABC* (*Comfert abc*).
- y_{obs} son los valores de la *ASFR* para cada edad observados en la población.
- n son la cantidad de edades consideradas.

Por otro lado, con el fin de medir la incertidumbre en la estimación de las *ASFR* se calculan los intervalos de credibilidad al 95% de la predictiva posterior.

3. Resultados

A continuación se presentan gráficamente la evolución de las *ASFR* observadas en función de la edad:

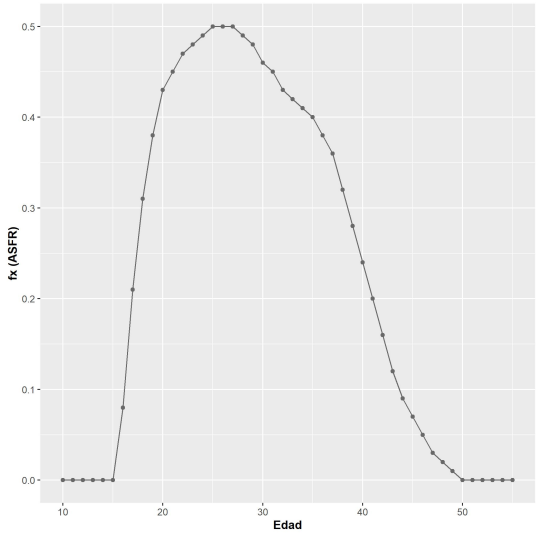

Fig. 1. Gráfico de líneas de las tasas de fecundidad por edad en la población

Por otra parte, en la figura 2 se presenta el gráfico de la estimación puntual para cada *ASFR* según el criterio del *ECM* definido en la sección marco metodológico.

No obstante, es importante destacar que el gráfico de la figura 2 fue construido a partir de una sola realización del proceso. Por lo tanto, no permite visualizar la aleatoriedad en las *ASFR*.

Con el fin de obtener una primera noción de dicha aleatoriedad, se presenta en la figura 3 el gráfico de las estimaciones de las *ASFR* por edad utilizando valores de α y κ seleccionados al azar:

Ahora bien, para obtener una medida de la incertidumbre mencionada, se procedió a calcular los intervalos de credibilidad y la estimación de la mediana (percentil 0.5 de la estimación de la distribución predictiva posterior para cada edad).

Con el fin de observar cómo varía la incertidumbre con respecto a la estimación puntual, se procedió a computar el *ECM* en cada combinación de α y κ seleccionando como candidatos de valores pertenecientes a la estimación de la distribución a posteriori aquellos que tienen error menor o igual al percentil que acumula un 10%, 50%, y 75% de la probabilidad de la distribución del error respectivamente. En la figura 4 se presenta el gráfico con los resultados.

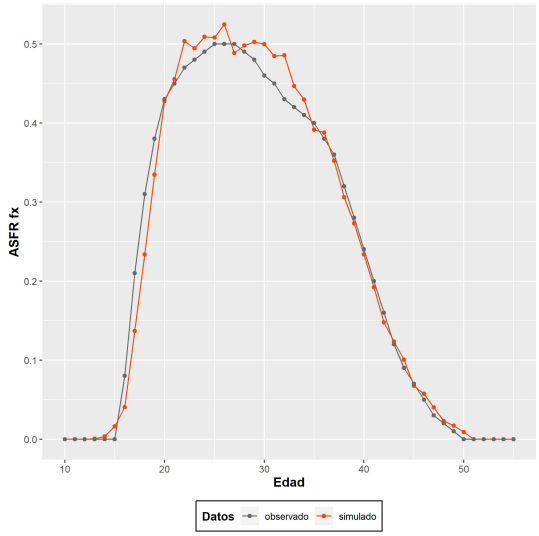

Fig. 2. Gráfico de líneas de las tasas de fecundidad por edad observada en la población y estimaciones según el criterio del mínimo *ECM*. Se observa que el modelo subestima las *ASFR* para edades tempranas (entre 10 y 20 años de edad). Por otro lado, el modelo tiende a sobrestimar para las edades entre 20 y 35 aproximadamente.

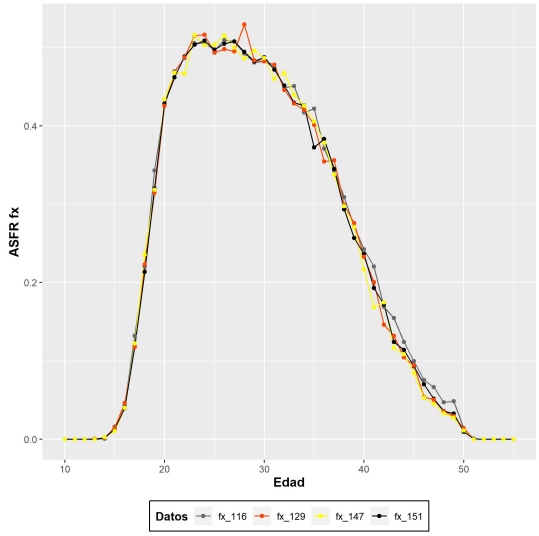

Fig. 3. Gráfico de líneas de las tasas de fecundidad por edad observada en la población y estimaciones para valores del *ECM* seleccionados aleatoriamente.

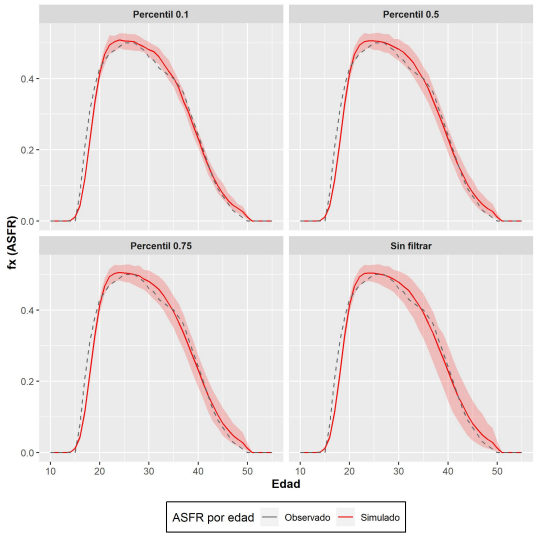

Fig. 4. Gráfico de líneas de las *ASFR*, valores observados, mediana e intervalo de confianza al 95%. Se observa que la amplitud de los intervalos de confianza aumenta conforme se incrementa el valor de ε seleccionado y la edad de la madre.

4. Conclusiones

Si bien el modelo *Comfert* nos permite obtener estimaciones para las tasas de fecundidad por edad, el enfoque bayesiano implementado (*ABC Bayes*) nos permite observar y cuantificar la incertidumbre subyacente en el fenómeno de estudio, pudiendo observar la existencia de aleatoriedad en las tasas de fecundidad por edad. Esto último en función de la parametrización seleccionada.

Si bien para valores elevados de ε el modelo parece lograr un ajuste satisfactorio en las edades avanzadas, se detecta un problema en todos los casos para edades tempranas. Una posible explicación está en que el modelo considera como parámetros la edad en la que decae la fecundidad y la tasa para dicha edad, sin considerar algún punto de inflexión para edades tempranas. Esto puede estar vinculado a los supuestos establecidos respecto a la edad de la unión (matrimonio), en particular a la modelización seleccionada para el tiempo de espera hasta el matrimonio.

Por otra parte, las estimaciones sobre la *ASFR* dependen del valor de ε seleccionado. Para valores de ε pequeños (en particular percentil 10%) existen valores observados que quedan por fuera del intervalo de credibilidad al 95%.

De esta manera, el modelo implementado no solamente nos pérmite observar y cuantificar la aleatoriedad en las tasas de fecundidad por edad, sino también que la incertidumbre aumenta significativamente cuanto mayor es el nivel de tolerancia utilizado.

5. Futuros trabajos

Para futuras investigaciones se considera de interés incorporar al modelo la salida de unidades por fallecimiento y los fenómenos asociados a las migraciones.

Adicionalmente, como extensión del modelo a poblaciones más avanzadas en la transición demográfica incluir una parametrización que considere la aleatoriedad proveniente de la utilización de métodos anticonceptivos y de la planificación o no del embarazo. Es decir, poblaciones bajo un régimen de fecundidad regulada.

En lo que respecta a la metodología utilizada, existen por lo menos dos formas de mejorar el enfoque utilizado.

En primer lugar, con respecto al elevado costo computacional inherente al modelo uno de los algoritmos más utilizados para reducirlo es el algoritmo denominado ABC regression adjustment (Beaumont et al., 2002). La idea principal de dicho algoritmo es correr un ABC estándar y considerar un margen de error amplio con el fin de ajustar la muestra obtenida mediante una regresión.

En segundo lugar, la obtención de los candidatos para la distribución a posteriori en función de una indicatriz implica una pérdida de información. Esto en el sentido que no permite cuantificar la distancia relativa entre el punto observado y simulado. Como posible alternativa se considera apropiado sustituir la función indicatriz por una función kernel.

Referencias

[1] Overview of Approximate Bayesian Computation S. A. Sisson Y. Fan and M. A. Beaumonty February 28, 2018.

[2] A review of Approximate Bayesian Computation methods via density estimation: inference for simulator-models Clara Grazian and Yanan Fan, September 2019.

[3] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[4] Demographic Models of the Reproductive Process:Past, Interlude, and Future. Daniel Ciganda, Nicolas Todd.

[5] Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686.

[6] Kirill Müller (2017). here: A Simpler Way to Find Your Files. R package version 0.1. https://CRAN.R-project.org/package=here.

[7] Matt Dowle and Arun Srinivasan (2020). data.table: Extension of data.frame. R package version 1.13.0. https://CRAN.Rproject. org/package=data.table.