Estudio de las Tasas específicas de fecundidad mediante técnicas ABC

Taller de modelos computacionales con aplicaciones en demografía

Alvaro Valiño - Lucia Coudet

Abril de 2021

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Resumen ejecutivo

El siguiente trabajo consiste en la presentación de una metodología para modelar las tasas de fecundidad por edad en una población de Huteritas. En primera instancia se utilizó un modelo basado en agentes denominado *Comfert*. Luego, con el fin de obtener estimaciones de los parámetros de este, se utilizó un enfoque bayesiano. En particular, el enfoque ABC el cual no necesita una expresión analítica de la función de verosimilitud.

A su vez, debido al elevado costo computacional que esto implica, se utilizó un metamodelo, en particular un proceso gaussiano.

Luego, se consideró de interés obtener una medida de incertidumbre en cuanto a las estimaciones de los parámetros y por consecuente de los valores simulados. Para ello se realizó primero una estimación puntual considerada óptima y luego se construyeron diferentes intervalos de credibilidad. Si bien todos se encuentran construidos al 95% de credibilidad, su estimación varía en función de los niveles de tolerancia considerados en el algoritmo ABC.

Como principal resultado se observó una relación positiva entre el nivel de incertidumbre y el incremento en la tolerancia del algoritmo. Asimismo, se observó que el modelo no realiza un ajuste del todo adecuado para edades tempranas. Por último, se presentan pasos a seguir para futuras investigaciones.

Palabras claves: fecundidad, fertilidad, proceso reproductivo, simulación, enfoque bayesiano, máxima verosimilitud aproximada.

Introducción

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 de Huteritas [1].

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 ese α.

Con el fin de lograr la reproducibilidad de los resultados obtenidos, se consideró apropiado utilizar un repositorio remoto público.[2]

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

[1] https://es.wikipedia.org/wiki/Huteritas
[2] https://github.com/alvarovalinio/ABC_bayes

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 denominado *Trayectorias reproductivas completas* (Comfert), el cual es una extensión del modelo de primer nacimiento. Dónde éste modela la frecuencia mensual de nacimientos después del casamiento.

Es importante destacar que uno de los supuestos base del modelo es que las mujeres están expuestas al riesgo de concebir inmediatamente luego del matrimonio. A su vez, se asume la no utilización de ningún tipo de métodos anticonceptivos.

Por otro lado, el modelo Comfert incluye periodos de no suceptibilidad de la madre luego del nacimiento. De este forma, los componentes principales 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 del modelo, debido a la incapacidad de estimar la función de verosimilitud se procedió a trabajar con métodos aproximados (Likelihood free). En particular desde un enfoque bayesiano, utilizando la Computación bayesiana aproximada (*Bayesian ABC*).

A constinuación 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 posterior, 4) una medida de discrepancia (distancia euclídea), 4) un valor de tolerancia ϵ

Algoritmo:

1. Simular θ^* $P(\theta)$

2. Simular $P(y|\theta*)$

   if $y_{sim} = y_0$ (discrete data) $\Delta(\eta_0, \eta_{sim}) < \epsilon$ (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 y a través de una función denominada *acquisition function*, la cual permite definir un subconjunto del espacio paramétrico. De esta manera, los puntos pertenecientes a dicho subconjunto 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 distribucion 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 (MSE):

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

Dónde:

- y_s son los valores simulados de las ASFR para cada edad, utilizando *Comfert_abc*
- y_{obs} son los valores observados en la población de Huteritas de la ASFR para cada edad
- n son la cantidad de edades consideradas.

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

Resultados

Como fue mencionado en la introducción, los datos utilizados son observaciones de las tasas de fecundidad por edad en una población de Huteritas.

A continuación se presenta gráficamente la evolución de la ASFR observada en función de la edad:

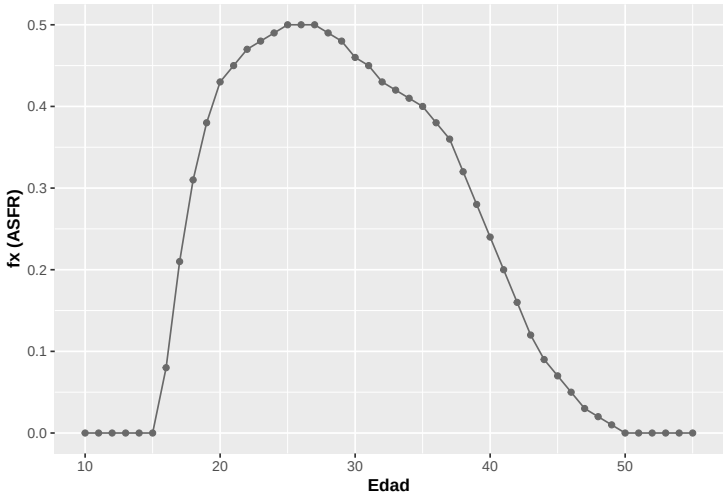

Figure 1: Scatter plot de las tasas de fecundidad por edad en una población de Huteritas

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

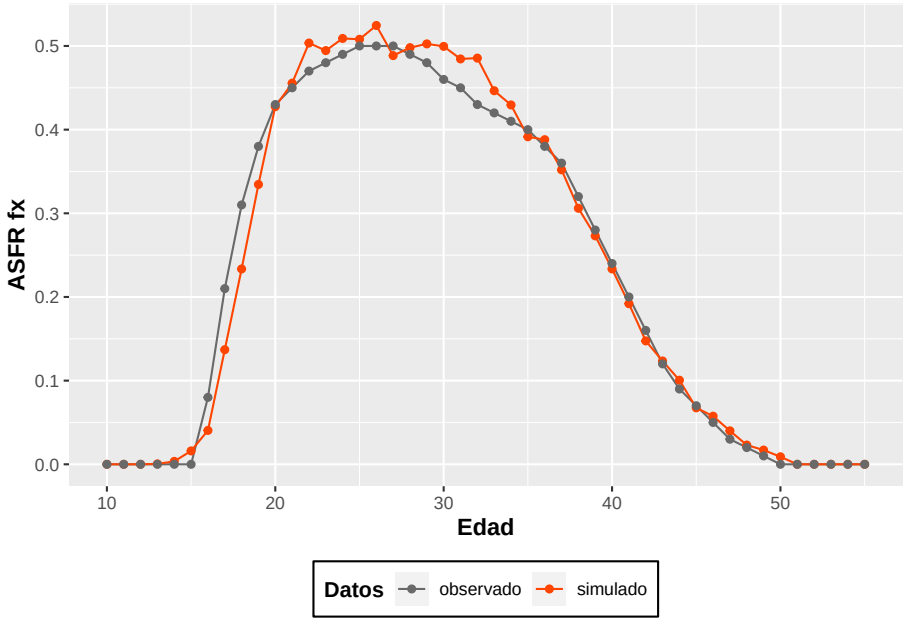

Figure 2: Tasa de fecundidad por edad observada en una población de Huteritas y estimaciones según el criterio del mínimo mse. Se observa que el modelo subestima las ASFR para edades tempranas (entre 10 y 20 años de edad). Por otro lado, el modelo tiende a sobrestimar para las edades entre 20 y 35 aproximadamente.

Es importante destacar que el gráfico anterior 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, a continuación se presenta el gráfico de las estimaciones de las ASFR por edad utilizando valores de α y κ seleccionados al azar:

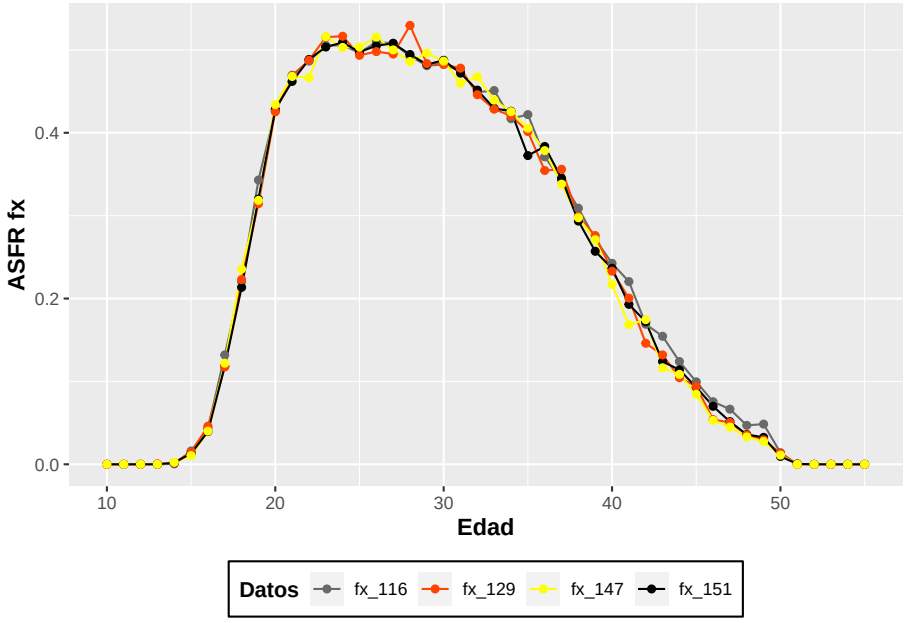

Figure 3: Tasa de fecundidad por edad observada en una población de Huteritas y estimaciones para valores del MSE seleccionados aleatoriamente

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 MSE 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.

A continuación se presenta el gráfico con los resultados:

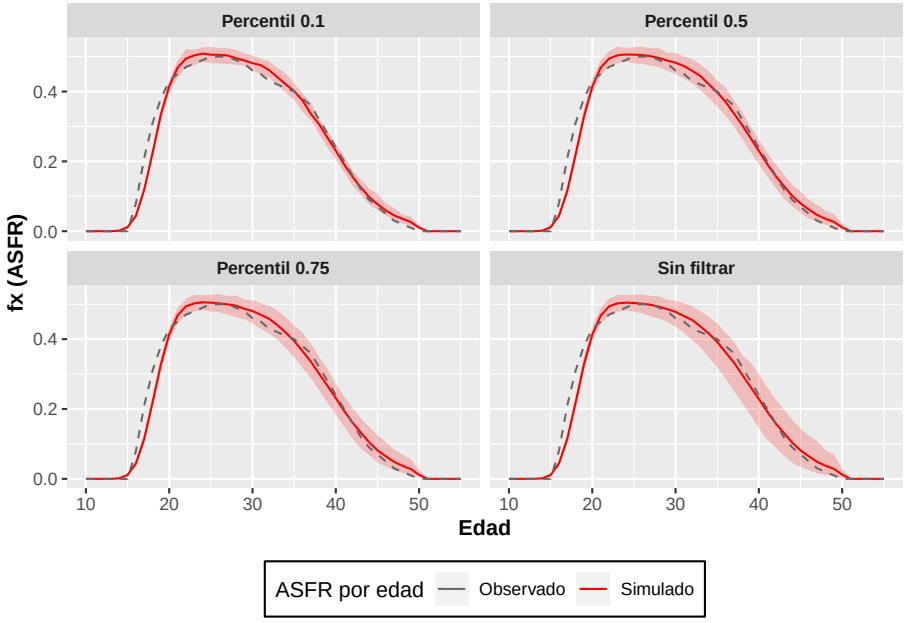

Figure 4: Gráfico de líneas de la ASFR por edad, valores observados, mediana e intervalo de confianza al 95%. Se observa que la amplitud de los intervalos de confianza aumenta conforme se incrementa el valor de epsilon seleccionado y la edad de la madre.

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.

Como podemos observar, el modelo no presenta un ajuste adecuado en algunas edades, en particular para edades tempranas (menos de 20 años).

Si bien para valores elevados de épsilon el modelo parece lograr un ajuste satisfactorio en las edades avanzadas, el problema persiste en todos los casos para las 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 al supuesto de que la concepción comienza a partir del matrimonio o también a la modelización seleccionada para el tiempo de espera.

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

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.

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.

Biblografía

- Overview of Approximate Bayesian Computation S. A. Sisson Y. Fan and M. A. Beaumonty February 28, 2018.
- A review of Approximate Bayesian Computation methods via density estimation: inference for simulator-models Clara Grazian and Yanan Fan, September 2019.
- 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/.
- Demographic Models of the Reproductive Process:Past, Interlude, and Future. Daniel Ciganda, Nicolas Todd.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686.

- Kirill Müller (2017). here: A Simpler Way to Find Your Files. R package version 0.1. https://CRAN.R-project.org/package=here.
- Matt Dowle and Arun Srinivasan (2020). data.table: Extension of `data.frame`. R package version 1.13.0. https://CRAN.R-project.org/package=data.table.