Cristina Cuesta[1], Luciano Mignini[2]
[1] *Instituto de Investigaciones Teóricas y Aplicadas – Escuela de Estadística - Facultad de Ciencias Económicas y Estadística – UNR*
[2] *Centro Rosarino de Estudios Perinatales*

ESTIMACIÓN DE LA INCIDENCIA DE PARTOS PRE-TÉRMINO EN FUNCIÓN DE LOS ESPACIAMIENTOS ENTRE EMBARAZOS. ANÁLISIS DE UNA REVISIÓN SISTEMÁTICA.

Introducción al problema

Desde el advenimiento de las revisiones sistemáticas, el paradigma de la medicina basada en la evidencia cobró mayor fuerza y expectativas. El análisis de publicaciones a nivel mundial sobre algún tema de interés para la comunidad médica científica es, aunque laborioso, simple de llevar a cabo a partir del uso de internet y otras herramientas de comunicación. Ni el idioma, ni la lejanía son un impedimento para conseguir publicaciones provenientes de estudios realizados en cualquier lugar del mundo. Aun resta por resolver la limitación provocada por posibles sesgos de publicación en los que se incurre cuando, por ejemplo, los resultados obtenidos no son publicados porque no son los esperados por el equipo de investigación que lo lleva a cabo. Otro tipo de sesgos pueden estar relacionados con las definiciones operacionales utilizadas en cada artículo, el periodo de tiempo considerado, el tipo de estudio, etc. Sin embargo, aun sabiendo el riesgo de cometer este tipo de sesgos, las revisiones sistemáticas continúan siendo una valiosa herramienta para evaluar cuestiones relacionadas a tratamientos, diagnósticos, estimaciones de incidencias, etc.

En el presente trabajo se muestra parte de un problema abordado en una revisión sistemática, ejecutada por el Centro Rosarino de Estudios Perinatales, cuyo objetivo general es estudiar la presencia de diversos eventos maternos y neonatales en función del espaciamiento entre embarazos de una misma mujer. En particular, el evento aquí abordado se refiere a que, luego de algún embarazo, el próximo embarazo sea pre-término, es decir que ocurra antes de las 37 semanas de gestación.

La particularidad que se presenta en el estudio de los intervalos intergenésicos (espaciamientos entre embarazos) es que los mismos se reportan agrupados de maneras muy disimiles. Así por ejemplo, en un artículo se informan estos intervalos agrupados de 0 a 24 meses vs. más de 24 meses, mientras que en otros el agrupamiento se hace de 0-18, 18-36 y más de 36 meses, etc. Es decir, la información no se reporta de forma consistente y por lo tanto, no se puede utilizar ninguna técnica tradicional de meta-análisis.

Con respecto a la asociación esperada, se tiene conocimiento de que cuando el espaciamiento entre embarazos es pequeño, la incidencia de partos pre-término es alta. Ésta disminuye acercándose a los 18-24 meses de espaciamiento y luego vuelve a aumentar. Sin embargo la forma funcional no es conocida, es decir no se puede suponer un modelo de relación funcional pre-especificado (como por ejemplo una relación de tipo cuadrática).

Se propone entonces para el análisis, la construcción de modelos que permitan utilizar la información disponible y a la vez intentar dar solución al objetivo planteado sin especificar una forma funcional a priori. Básicamente el modelo a utilizar es un modelo aditivo de regresión spline-penalizado donde la variable respuesta se refiere al evento (partos pre-término) mientras que las variables explicativas se refieren a datos sobre el espaciamiento registrado en los artículos.

Finalmente es necesario mencionar que hay al menos tres consideraciones que deben tenerse en cuenta en el análisis. La primera es que los datos deben ser ponderados a fin de tener en cuenta el tamaño del estudio; la segunda es que los datos que provienen del mismo artículo deben considerarse correlacionados y finalmente que se descartaran del análisis datos censurados.

Metodología

Como metodología de análisis se proponen modelos aditivos. Para dos variables explicativas su expresión es de la forma:

$$y_i = \beta_0 + f(x_{1i}) + g(x_{2i}) + \varepsilon_i \qquad (1)$$

donde f y g , son funciones de suavizado. La estimación de las funciones f y g se puede realizar mediante distintas técnicas de suavizado, dentro de las cuales se destacan las basadas en regresiones spline. Las mismas consisten en una regresión por tramos, donde cada uno de ellos es una región del campo de variación de la variable explicativa, en la que se ajusta un modelo de regresión polinómico (en general de bajo orden). Estos tramos se unen en los puntos extremos, conocidos con el nombre de “nodos”, para dar continuidad a la curva.

Los modelos de regresión spline dependen en gran medida de la cantidad de regiones que se considere y de la respectiva amplitud de las mismas. Para subsanar este inconveniente, se sugiere definir un número grande de regiones y luego ponderar la importancia de que las regiones sean consideradas diferentes. Esto se lleva a cabo con los modelos de regresión spline penalizados (o P-Splines). Las P-splines pueden ser utilizadas con facilidad

para ajustar modelos como (1).

De manera simple, un modelo aditivo de regresión spline lineal se expresa como sigue:

$$y_i = \beta_0 + \beta_{x_1} x_{1i} + \sum_{k=1}^{K_1} \beta_k^{*x_1} \left(x_{1i} - c_k^{x_1}\right)_+ + \beta_{x_2} x_{2i} + \sum_{k=1}^{K_2} \beta_k^{*x_2} \left(x_{2i} - c_k^{x_2}\right)_+ + \varepsilon_i \qquad (2)$$

donde:

$$\left(x_{1i} - c_k^{x_1}\right)_+ = \begin{cases} 0 & , \ x_{1i} \leq c_k^{x_1} \\ \left(x_{1i} - c_k^{x_1}\right) & , \ x_{1i} > c_k^{x_1} \end{cases} \quad \text{y} \quad \left(x_{2i} - c_k^{x_2}\right)_+ = \begin{cases} 0 & , \ x_{2i} \leq c_k^{x_2} \\ \left(x_{2i} - c_k^{x_2}\right) & , \ x_{2i} > c_k^{x_2} \end{cases}$$

se denominan funciones de base truncada y permiten el ajuste cuadrático por tramos siendo $c_1^{x_1}, \ldots, c_{K_1}^{x_1}$ y $c_1^{x_2}, \ldots, c_{K_2}^{x_2}$ los respectivos nodos en las direcciones de x_1 y x_2. Los mismos pueden ser elegidos utilizando distintas reglas, por ejemplo nodos equiespaciados dentro del rango de interés de cada variable explicativa, o, nodos ubicados en percentiles equiespaciados determinados por el conjunto de datos.

La representación matricial del modelo (2) resulta:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} \quad \text{con} \ E(\boldsymbol{\varepsilon}) = 0 \ \text{ y } \ Cov(\boldsymbol{\varepsilon}) = (\sigma_\varepsilon^2 \boldsymbol{I})$$

donde:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_{x_1} \\ \beta_{x_2} \end{bmatrix} \quad \boldsymbol{\beta}^* = \begin{bmatrix} \beta_1^{*x_1} \\ \vdots \\ \beta_{K_1}^{*x_1} \\ \beta_1^{*x_2} \\ \vdots \\ \beta_{K_2}^{*x_2} \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix} \ \text{ y}$$

$$\boldsymbol{Z} = \begin{bmatrix} \left(x_{11} - c_1^{x_1}\right)_+ & \cdots & \left(x_{11} - c_{K_1}^{x_1}\right)_+ & \left(x_{21} - c_1^{x_2}\right)_+ & \cdots & \left(x_{21} - c_{K_2}^{x_2}\right)_+ \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \left(x_{1n} - c_n^{x_1}\right)_+ & \cdots & \left(x_{1n} - c_{K_1}^{x_1}\right)_+ & \left(x_{2n} - c_n^{x_2}\right)_+ & \cdots & \left(x_{2n} - c_{K_2}^{x_2}\right)_+ \end{bmatrix}$$

El vector de los valores estimados, utilizando mínimos cuadrados y penalizando los coeficientes asociados a los nodos, está dado por:

$$\hat{Y} = \boldsymbol{A}\left(\boldsymbol{A}'\boldsymbol{A} + \boldsymbol{D}\right)^{-1}\boldsymbol{A}'\boldsymbol{Y} \qquad (3)$$

donde $\boldsymbol{A} = (\boldsymbol{X} \mid \boldsymbol{Z})$ y $\boldsymbol{D} = diag\left(0, 0, 0, \lambda_1^2 \mathbf{1}_{K_1 x 1}, \lambda_2^2 \mathbf{1}_{K_2 x 1}\right)$, siendo ésta última la matriz que controla la influencia de los nodos.

Los parámetros de suavizado, λ_1 y λ_2, inducen el suavizado en las correspondientes direcciones de x_1 y x_2, penalizando los coeficientes de los nodos. Dicha penalización se utiliza a los efectos de disminuir la "rugosidad" del ajuste controlando el peso de cada uno

de los nodos. Se pueden plantear distintos tipos de penalizaciones; la más usual es $\sum\sum \beta_k^{*x_i} < C$ y para un C determinado. Matricialmente, la restricción se expresa como $\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\beta}^*\end{pmatrix}' \boldsymbol{P} \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\beta}^*\end{pmatrix} \leq C$, siendo $P = \begin{bmatrix}\mathbf{0}_{3x3} & \mathbf{0}_{3x(K_1+K_2)}\\ \mathbf{0}_{(K_1+K_2)x3} & \boldsymbol{I}_{(K_1+K_2)x(K_1+K_2)}\end{bmatrix}$.

En particular, si se tratan los $\beta_k^{*x_1}$ y $\beta_k^{*x_2}$ como efectos aleatorios en un modelo mixto, más específicamente, si se define $\boldsymbol{U}' = [u_1^{x_1} \ \dots \ u_{K_1}^{x_1} \ u_1^{x_2} \ \dots \ u_{K_2}^{x_2}]$, un vector de efectos aleatorios que sigue una distribución Normal con:$E(\boldsymbol{U}) = \mathbf{0}$ y $Cov(\boldsymbol{U}) = \begin{bmatrix}\sigma_1^2 \boldsymbol{I}_{K_1 x K_1} & \mathbf{0}\\ \mathbf{0} & \sigma_2^2 \boldsymbol{I}_{K_2 x K_2}\end{bmatrix}$, se puede reescribir (2) como el siguiente modelo mixto:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{U} + \boldsymbol{\varepsilon} \qquad (4)$$

Ruppert (2003) y Ngo (2004) han mostrado que los estimadores obtenidos por mínimos cuadrados penalizados son equivalentes al mejor predictor lineal insesgado en los modelos mixtos (BLUP), siendo $\lambda_1 = \sigma_\varepsilon/\sigma_1$ y $\lambda_2 = \sigma_\varepsilon/\sigma_2$.

De esta forma, la representación de (4) como modelo mixto puede ser utilizada para facilitar el ajuste, la inferencia y la selección del modelo.

Los criterios de selección clásicos, tales como GVC y AIC pueden ser utilizados para buscar el modelo que mejor ajuste los datos. Estos modelos se basan en la verosimilitud pero, a su vez, realizan una penalización de acuerdo al número de parámetros en el modelo. El más utilizado en la práctica es el "Criterio de Información de Akaike" (AIC), el cual se calcula como:

$$AIC = -2\,L(\hat{\beta}, \hat{u}) + 2p$$

donde p es el número total de parámetros estimados. A tal efecto, el término $2p$ hace la función de penalización. Cuanto más pequeña resulte esta medida AIC, mejor es el ajuste del modelo.

La ventaja de este tipo de criterios radica en la posibilidad de comparar modelos que no se encuentran anidados.

Análisis de la información disponible en la revisión sistemática

La variable "tiempo que transcurre entre dos embarazos consecutivos" (TEE) es de tipo continua, sin embargo por tratarse de datos que se reportan agrupados habitualmente son categorizadas en intervalos de clases (por ejemplo 0-12 meses, 13-24 meses, más de 24 meses). Por otra parte, para cada intervalo categorizado de TEE se cuenta con la cantidad de partos y cuántos de ellos fueron pre-término.

En la revisión sistemática mencionada se detectaron 23 artículos que permitirían estimar la asociación entre intervalos entre embarazos y partos pre-término. Estos artículos proveen una totalidad de 74 registros. Cada registro contiene, entonces, la siguiente información:

* Código de identificación del artículo

* Límite inferior del intervalo de categorización de TEE

* Límite superior del intervalo de categorización de TEE

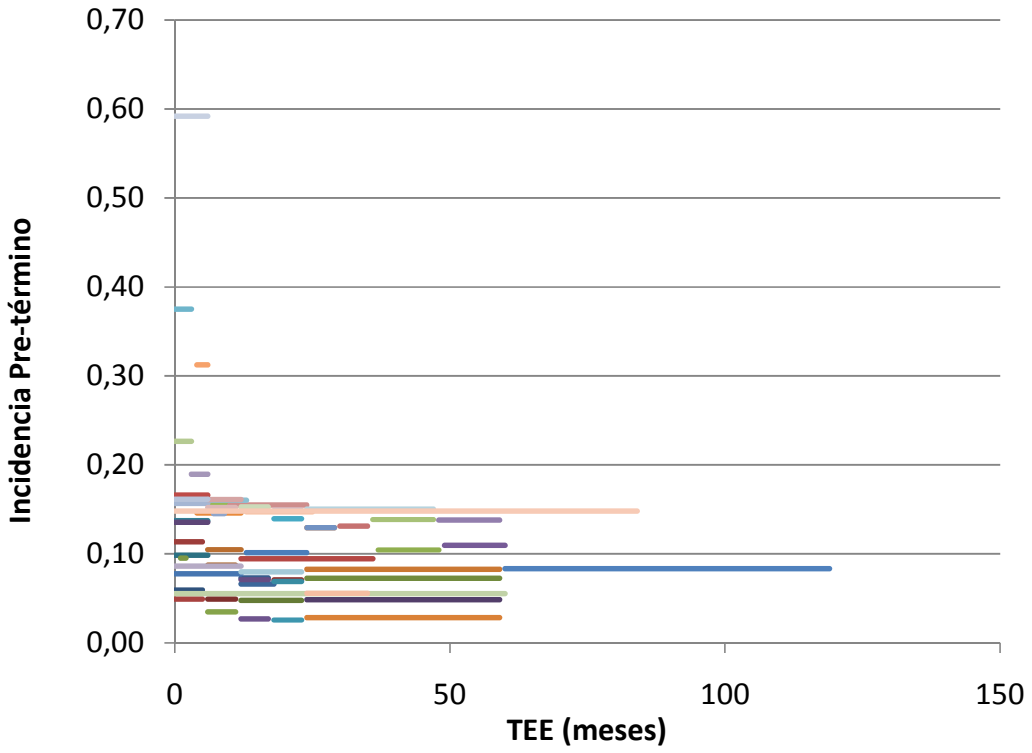
* Punto medio del intervalo de categorización de TEE

* Cantidad total de partos

* Cantidad de partos pre-término

En la Figura 1 puede verse que hay una gran discrepancia en la definición de los intervalos de clase. Hay una gran cantidad de intervalos de clase sobrepuestos lo que dificulta el análisis. Los intervalos censurados por derecha fueron excluidos del análisis.

Figura1. Incidencia Pre-término de acuerdo al tiempo transcurrido entre embarazos consecutivos.



Se construyeron modelos aditivos de regresión spline cuadrática con 10 nodos en cada variable explicativa. Para la estimación se ponderó la importancia de cada observación. Si no se tuviera en cuenta ninguna ponderación, se le estaría dando igual peso a cada parto, mientras que si se pondera por la cantidad total de partos en el estudio se estaría dando más ponderación a estudios más grande. Por ello se utiliza una alternativa intermedia ponderando cada dato por la raíz cuadrada de la cantidad total de partos (lo cual produce un efecto intermedio).

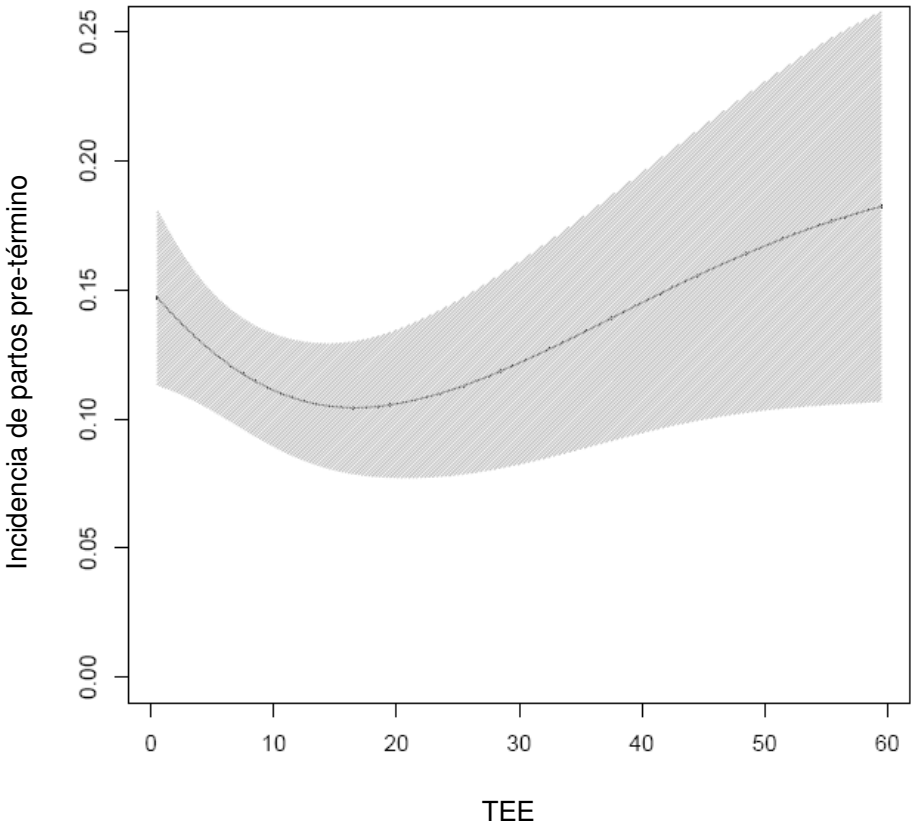
La variable respuesta considerada es el logaritmo de la incidencia, esto garantiza un mayor cumplimiento de los supuestos (que no pueden garantizarse utilizando directamente la incidencia). Se consideran dos conjuntos posibles de variables explicativas.

1) Límite inferior y limite superior de los intervalos de clase
2) Amplitud y punto medio del intervalo de clase.

Luego de comparar los criterios de información y analizar los residuos, se opta por el segundo conjunto de variables explicativas.

A partir de dicho modelo se estiman las incidencias para diferentes intervalos intergenésicos y se construyen intervalos de confianza (ver Figura 2). Estas estimaciones se condicen con lo esperado.

Figura2. Estimación de la incidencia (e intervalos de confianza del 95%) de nacimientos pre-termino en función del tiempo entre embarazos



Comentarios finales

Los modelos aditivos han permitido salvar la dificultad de utilizar conjuntamente información que a priori parecía difícil de unir, permitiendo así realizar estimaciones de las incidencias de parto pre-termino en función de los espaciamientos intergenesicos.

El análisis presentado está en período de evaluación. Otros modelos se han postulado, por ejemplo considerando splines lineales, considerando distintas ponderaciones, distinta cantidad de nodos, etc. También es necesario probar modelos que consideren que los errores no son independientes, sino que se debe considerar la correlación entre observaciones de un mismo artículo. Por otro lado, se está pensando en considerar a la variable respuesta como binaria y, en función de ello, planteando un modelo logit aditivo de regresiones splines penalizadas.

Finalmente, es necesario remarcar la importancia de utilizar metodologías estadísticas alternativas sobre nuevas áreas de aplicación, ampliando así el horizonte del conocimiento y abriendo paso a los nuevos métodos sobre los tradicionales.

REFERENCIAS BIBLIOGRÁFICAS

- Cai, W. (2008) Fitting Generalized Additive Models with the GAM Procedure in SAS 9.2. SAS Institute Inc., Cary NC (USA).

- Durbán, M. (2009). An introduction to smoothing with penalties: P-Splines. Boletín de Estadística e Investigación Operativa, Vol. 25, No. 3, pp. 195-205.

- Eilers, P., Marx, B. D. (2002). Generalized linear additive smooth structures. Journal of Computational and Graphical Statistics, Vol 11(4), 758-783.

- Hastie, T. J., Tibshirani, R. J. (1996) Generalized Additive Models. Encyclopaedia of Statistical Sciences.

- Marra, G., Radice, R. (2010). Penalised Regression splines: theory and application to medical research. Statistical Methods in Medical Research, 19: 107-125.

- Ngo, L., Wand, M. P. (2004). Smoothing with mixed model software. Journal of Statistical Software, Volume 09, Issue 01.

- Ruppert, D., Wand, M. P., Carroll, R. (2003). Semiparametric Regression, Cambridge University Press, Cambridge (UK).

- SAS Institute Inc. (1999). SAS/GRAPH® Software: Reference, Version 8. Chapter 30: The G3GRID Procedure. Cary, NC (USA).

- SAS Institute Inc. (2009). SAS/STAT® 9.2 User's Guide: Mixed Modeling (Book Excerpt). Cary, NC (USA).