Modelos alternativos para predecir la tasa de natalidad en función de los factores ambientales y socioeconómicos de un país *

Jessica Quintero López *Universidad Nacional de Colombia - sede Medellín*
Freddy Hernández Barajas *Universidad Nacional de Colombia - sede Medellín*
Yuberth Anderson Saavedra Coneo *Universidad Nacional de Colombia - sede Medellín*

Resumen: Determinar la influencia de los factores ambientales y socioeconómicos sobre la tasa de natalidad o la fertilidad femenina, es el común de muchos artículos científicos; donde se aplican modelos estadísticos asumiendo hipotéticamente que la variable tasa de natalidad o fertilidad femenina sigue una distribución normal univariada, hipótesis que no siempre se cumple. En este documento se usan los modelos GAMLSS, para estudiar la influencia de las variables temperatura, producto interno bruto *(PIB)* y la contaminación por partículas finas de aire *(PM2,5)* sobre la tasa de natalidad o fertilidad femenina a nivel de país. Los modelos GAMLSS permiten que el investigador asuma distribuciones estadísticas para la variable respuesta diferentes a la normal y que se puedan modelar todos los parámetros en función de las covariables. Al aplicar GAMLSS a los datos se obtuvo que las variables temperatura, producto interno bruto *(PIB)* y la contaminación por partículas finas de aire *(PM2,5)* influyen significativamente sobre explicación de la tasa de natalidad o la fertilidad femenina a nivel de país.

Keywords: GAMLSS, Regresión, Loess, Tasa de natalidad, Temperatura, Contaminación, PIB.

Introducción

Este artículo de investigación aborda la problemática sobre la tasa de natalidad, que según la Organización Mundial de la Salud (2018) se define como: *"La razón entre el número de nacidos vivos en una población durante un año específico y la población total a mitad de año, para el mismo año, usualmente multiplicada por 1000"*.

En la segunda mitad del siglo XVIII, la revolución industrial causó cambios demográficos que se fundamentaron en la mejora de las condiciones higiénicas, sanitarias y alimenticias que disminuyeron de forma notoria las enfermedades y decesos por desnutrición. De esta manera, en los últimos 70 años la población mundial se ha duplicado, lo cual es un dato alarmante para las naciones. Algunos países han invertido en educación y en planificación familiar con el fin de evitar embarazos en hogares de bajos recursos, y de disminuir el índice de nacimientos en los países. Estudios científicos sobre la influencia de las altas y bajas temperaturas en algunos países han mostrado que la temperatura es un factor sobresaliente que influye en la determinación de la tasa de natalidad; asimismo, se ha encontrado relación entre la tasa de natalidad y las partículas finas de aire *(PM2,5)*, reportando que a nivel mundial las altas contaminaciones del aire se presentan en los países más pobres; además, los procesos de fertilidad femenina se han visto íntimamente perjudicados por las altas contaminaciones. Por consiguiente, es interesante para el interés de este

* **Autor de contacto**: yuasaavedraco@unal.edu.co.

estudio, analizar la tasa de natalidad en términos de factores ambientales como la temperatura y la contaminación por partículas finas de aire *(PM2,5)*; como también, del producto interno bruto *(PIB)* que representa el factor socioeconómico.

El objetivo principal de este estudio es plantear modelos alternativos al aplicado por Mary Regina Boland (2018), quien usa datos obtenidos mediante un estudio observacional, y se abordan estadísticamente para explicar y cuantificar la tasa de natalidad como proxy de la fecundación femenina a nivel de país, mediante la metodología GAMLSS en términos de las covariables temperatura *(representada mediante la temperatura media en grados Celsius desde el año 1961 hasta el año 1990)*, la contaminación por partículas finas de aire *(PM2,5)*, y el producto interno bruto *(PIB)* para el año 2016. Dicho propósito se llevará a cabo usando el lenguaje de programación R Core Team (2019), para dar respuesta a nivel de país de los efectos de los factores ambientales y socioeconómicos mencionados sobre la fecundidad femenina; es decir, la tasa de natalidad.

Este documento está dividido por las siguientes secciones; en la sección 1, se presenta la introducción del estudio; en la sección 2, se hace una breve descripción de los modelos GAMLSS; en la sección 3, se presenta un análisis descriptivo de los datos y se muestran algunos de los patrones y relaciones entre las variables mediante gráficos bidimensionales y tridimensionales; en la sección 4, se muestra los resultados con los diferentes modelos considerados, los criterios para la elección del mejor modelo y los resultados del mejor modelo; por último, en la sección 6 están las conclusiones del estudio.

Modelos GAMLSS

Los modelos GAMLSS propuestos por Rigby, R., Stasinopoulos, D., (2005) son de gran utilidad ya que permiten modelar los parámetros de la respuesta en función de las covariables; además, permiten elegir entre más de 100 distribuciones continuas, discretas y mixtas la distribución más adecuada para la variable respuesta, y no se limitan al supuesto de normalidad. En dichos modelos, las observaciones son independientes, y la función de masa o de densidad de probabilidad depende del vector de parámetros. Los modelos GAMLSS se puede aplicar fácilmente por medio del paquete *gamlss* disponible en R Core Team (2019). La función $fitDist$ del paquete *gamlss* suministra una lista de distribuciones que se ajustan mejor a la variable respuesta, estas están basadas sobre el criterio de información de Akaike generalizado *(GAIC)*, para una penalización dada por k, donde es k es por defecto igual a 2.

Análisis descriptivo de variables y datos

En este trabajo se utilizan los datos analizados por Mary Regina Boland (2018) para investigar el papel que juegan los factores ambientales y socioeconómicos en la fecundidad femenina; para esto, se realiza un resumen descriptivo de los datos.

En la figura 1 se realiza un análisis para descartar la posible existencia de multicolinealidad entre las covariables y evitar ajustes erróneos del modelo. Se puede observar una relación alta entre la variable respuesta tasa de natalidad y la temperatura, lo cual indica que esta covariable será significativa para el modelo.

En la figura 2 se realizan tres diagramas de dispersión por cada pareja de covariables; en el diagrama de la izquierda la tasa de natalidad aumenta para PIB e índices de contaminación relativamente bajos. Además, en el grafico del centro la tasa de natalidad aumenta para altas temperaturas e indices de contaminación moderadamente bajos. Por último, en el grafico de la derecha la tasa de natalidad aumenta para temperaturas medias y para los PIB moderadamente bajos; lo cual indica que existe mayor tasa de natalidad en los países más pobres y cálidos.

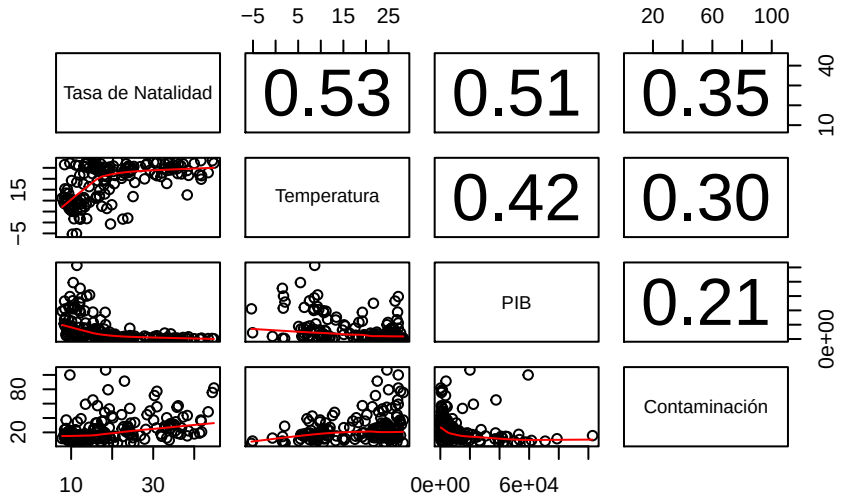

Figure 1: Matriz de diagrama de dispersión con correlaciones

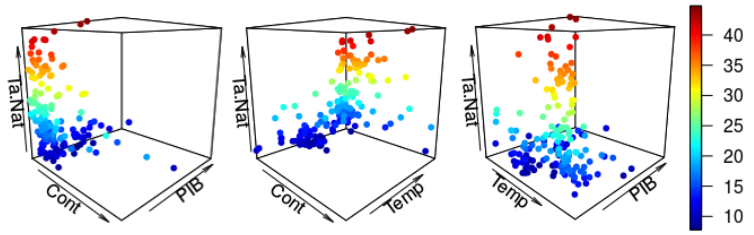

Figure 2: Tasa de natalidad en función de las covariables.

Resultados

En esta sección se presentan los resultados del modelo aplicado por Mary Regina Boland (2018) y los resultados de los modelos alternativos considerados. Como críterios de selección se tuvo en cuenta el AIC para el modelo lineal, GAIC para los modelos GAMLSS, el *Pseudo* R^2 y la correlación entre los valores estimados de la tasa de natalidad y los verdaderos valores de la variable respuesta. Adicionalmente, se consideró un modelo de regresión local al que se le obtiene el AIC con la función creada por Michael Friendly (2005), y la aproximación del *Pseudo* R^2 como lo describen en la red de webs Stack Exchange (2013). El modelo aplicado por Mary Regina Boland (2018), tiene una ecuación con la siguiente forma:

$$Natalidad_i = \beta_0 + \beta_1 Temp_i + \beta_2 PIB_i + \beta_3 Cont_i + \beta_4 Temp_i PIB_i + \beta_5 Temp_i Cont_i + \beta_6 PIB_i Cont_i + \varepsilon_i$$

En la tabla 1 se observa que, a un nivel de significancia del 0,05 existen variables no significativas para el modelo de referencia; por lo cual, se hace un proceso de selección de variables que permita quitar el enmascaramiento de unas variables sobre otras. Se obtiene un modelo de la forma:

$$Natalidad_i = \beta_0 + \beta_1 Temp_i + \beta_2 PIB_i + \beta_3 Cont_i + \beta_4 Temp_i PIB_i + \varepsilon_i$$

Por consiguiente, se uso la función $fitDist$ del paquete *gamlss* para obtener la familia de distribuciones que mejor se ajustan a la variable respuesta, de dicha lista, se seleccionaron ocho distribuciones y se le aplicaron al modelo anterior. La tabla 3 muestra los valores de cada críterio de selección para el modelo de regresión lineal múltiple de referencia y para el modelo de regresión local *(loess)*; en cambio, la tabla 4 muestra dichos valores para los modelos GAMLSS. De las tablas 3 y 4 se sigue que, el mejor modelo GAMLSS es el que tiene una distribución $IGAMMA$ en la variable respuesta; sin embargo, el modelo de regresión local *(loess)* supera por mucho a cualquier modelo GAMLSS considerado.

Table 1: Parámetros estimados para el modelo de referencia.

	Estimado	Error estándar	Valor t	Valor-P
Intercept	10,9400680	3,4869750	3,1374094	0,0020231
Temp	0,5117237	0,1555395	3,2899908	0,0012279
PIB	0,0000726	0,0000624	1,1631774	0,2464575
Cont	-0,0379574	0,1364888	-0,2780991	0,7812887
Temp:PIB	-0,0000205	0,0000048	-4,2630611	0,0000340
Temp:Cont	0,0074629	0,0057336	1,3016023	0,1948890
PIB:Cont	-0,0000015	0,0000018	-0,8140533	0,4168020

Table 2: Parámetros estimados para el modelo ajustado.

	Estimado	Error estándar	Valor t	Valor-P
Intercept	7,2851235	1,8633947	3,909598	1,348e-04
Temp	0,6900294	0,0839170	8,222760	5,619e-14
PIB	1,068e-04	5,529e-05	1,931717	0,0551080
Cont	0,1214172	0,0296942	4,088919	6,757e-05
Temp:PIB	-2,411e-05	3,643e-06	-6,617357	0,885e-10

Table 3: Resultados para el modelo de referencia y el loess.

Modelo	Distribución	Cor	Pseudo R^2	AIC
Mod9	Loess	0,83	0,68	241,26
Referencia	NO	0,73	0,54	1143,42

Table 4: Resultados para los modelos GAMLSS.

Modelo	Distribución	Correlación	Pseudo R^2	AIC
Mod1	IGAMMA	0,72	0,62	1063,56
Mod2	GIG	0,75	0,61	1069,57
Mod3	LOGNO2	0,74	0,61	1071,08
Mod4	IG	0,74	0,60	1072,27
Mod5	BCPE	0,72	0,50	1078,65
Mod6	exGAUS	0,71	0,47	1086,65
Mod7	WEI2	-0,57	0,57	1105,23
Mod8	NO	0,71	0,61	1114,05

Conclusiones

En este trabajo se analizaron diferentes modelos de regresión lineal múltiple, alternativos al de Mary Regina Boland (2018), donde se modela la tasa de natalidad como proxy de la fecundidad femenina en función de la temperatura, los grados de contaminación y el producto interno bruto *(PIB)* respectivo de cada uno de los 170 países de los se utilizó la información. Adicionalmente, para los modelos GAMLSS se consideraron ocho familias para la distribución de la variable respuesta, como también se tuvo en cuenta un modelo de regresión local.

En particular, cada parámetro de los modelos GAMLSS considerados fueron modelados en términos de las covariables, generando significancia de todas las covariables en el modelo final. Por último, se obtuvo que el mejor modelo GAMLSS es el que tiene una distribución $IGAMMA$ en la variable respuesta, con el parámetro de escala *sigma* (σ) en función de las covariables temperatura y producto interno bruto *(PIB)*; no obstante, el modelo de regresión local *(loess)* fue tomado como modelo final ya que presento menor medida en el AIC, mayor correlación entre los valores estimados $E[Y]$ y los verdaderos valores de la variable respuesta, y un mayor R^2.

References

Mary Regina Boland. 2018. *A model investigating environmental factors that play a role in female fecundity or birth rate*. San Francisco, California: PLOS ONE.
URL: *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0207932*

Michael Friendly. 2005. *loess: choose span to minimize AIC*. Hamilton, Ontario, Canadá: .
URL: *https://stat.ethz.ch/pipermail/r-help/2005-November/082849.html*

Organización Mundial de la Salud. 2018. *El embarazo en la adolescencia*. Ginebra, Suiza: OMS.
URL: *https://www.who.int/es/news-room/fact-sheets/detail/adolescent-pregnancy*

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org/*

Rigby, R., Stasinopoulos, D.,. 2005. *Generalized additive models for location, scale and Shape*. London, England: Royal Statistical Society.
URL: *https://doi.org/10.1111/j.1467-9876.2005.00510.x*

Stack Exchange. 2013. *How to get an R-squared for a loess fit?* New York, U.S: .
URL: *https://stats.stackexchange.com/questions/24993/how-to-get-an-r-squared-for-a-loess-fit*