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

Jessica Quintero López
Yuberth Anderson Saavedra Coneo
Universidad Nacional de Colombia sede Medellín
jquinterol@unal.edu.co, yuasaavedraco@unal.edu.co

Resumen

Determinar la influencia de los factores ambientales y socioeconómicos sobre la tasa de natalidad, es el común de muchos artículos científicos; donde se aplican modelos estadísticos asumiendo hipotéticamente que la variable tasa de natalidad sigue una distribución normal univariada, hipótesis que no siempre se cumple. En este trabajo se usan los modelos GAMLSS, para estudiar la influencia de las variables temperatura, producto interno bruto y la contaminación por material particulado (PM2,5) sobre la tasa de natalidad 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 las observaciones se obtuvo que las variables temperatura, producto interno bruto y la contaminación por material particulado (PM2,5) influyen significativamente sobre explicación de la tasa de natalidad a nivel de país. En particular, los resultados encontrados en este trabajo sirven para describir la tasa de natalidad y para estimar la tasa de crecimiento poblacional de los países.

Introducción

A lo largo de la historia la humanidad se ha ido desarrollando en diferentes ámbitos, lo cual desde la segunda mitad del siglo XVIII ha generado que la población se duplique, esto ha sido un dato alarmante y ha llevado a algunos países a ofrecer medidas preventivas para evitar que las cifras sigan aumentando, por eso es importante y de nuestro interés determinar la influencia que tienen los factores ambientales y socioeconómicos sobre la tasa de natalidad; más áun, cuando este tema es el común de muchos artículos científicos donde aplican modelos estadísticos asumiendo que la variable tasa de natalidad sigue una distribución normal univariada, algo que en muchos casos no se cumple.

Así, el objetivo principal de este estudio es plantear modelos alternativos al aplicado por Mary Regina Boland (2018), quien usa 170 observaciones de distintos países obtenidos mediante un estudio observacional, y se abordan estadísticamente para cuantificar la tasa de natalidad, se hace mediante la metodología GAMLSS en términos de las covariables temperatura, contaminación por material particulado (PM2,5) y producto interno bruto. Para dicho propósito se usa el lenguaje de programación estadístico R.

Metodología

Se ajustaron ocho modelos a las observaciones por medio de la metodlogía GAMLSS con variable respuesta (tasa de natalidad) y covariables (temperatura, PM,25 y PIB); se modeló la varianza en términos de las covariables (temperatura y PIB), como cada uno de los parámetros de cada distribución considerada. Adicionalmente, se ajustó un modelo de regresión local (loess) para comparar con los otros modelos.

GAMLSS

Los modelos GAMLSS (Generalized Additive Model for Location Scale and Shape) propuestos por **Rigby y Stasinopoulos(2005)** son de gran utilidad ya que permiten modelar todos los parámetros de la variable de interés en función de las covariables, además, la ventaja que tienen sobre los modelos lineales usuales radica en que permiten elegir la distribución más adecuada para la variable respuesta y no se limitan al supuesto de normalidad.

LOESS

La regresión local fue propuesta originalmente por **Cleveland (1979)** y desarrollada por **Cleveland y Devlin (1988)** se caracterizan por permitir realizar ajuste de curvas y superficies a las observaciones mediante el suavizamientos; asi como, representar los datos punto a punto sin necesidad de especificar una función global al ajustar un modelo, lo cual impide a su vez, obtener una representación matemática para el modelo.

Análisis descriptivo de variables y datos

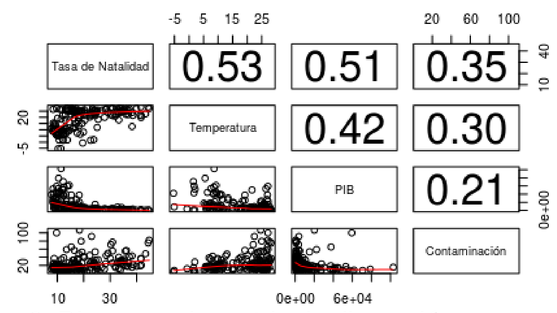

Figura 2. Diagrama de matriz de dispersión con correlaciones.

Con la Figura 2 se evita ajustar erróneamente el modelo, ya que se descarta la existencia de multicolinealidad entre las covariables. Además, no hay varianza constante ni se presenta relación lineal entre las variables explicativas.

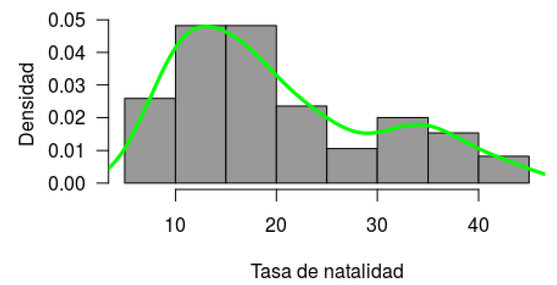

Figura 3. Histograma y curva de densidad para la tasa de natalidad.

En la Figura 3 se observa que la tasa de natalidad tiene una distribución asimétrica con sesgo a la derecha; por lo tanto, se descarta el supuesto de que la tasa de natalidad proviene de una distribución normal univariada.

Selección de variables

Se utilizó la metodología StepAIC *(Forward y Backward)* para encontrar las variables más significativas para la tasa de natalidad. Mediante ambos métodos se llegó al mismo modelo final.

Las variables más significativas son:
1. Temperatura (promedio desde 1961 a 1990)
2. PIB (producto interno bruto para el año 2016)
3. Contaminación (material particulado PM2,5)
4. TemperaturaProducto Interno Bruto (Temp $\times PIB$)

Resultados

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 R2 y la correlación entre los valores estimados y los verdaderos valores de la tasa de natalidad. También, 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 R2 como lo describen en la red de webs **Stack Exchange (2013).**

El modelo de regresión lineal múltiple usado por **Mary Regina Boland (2018),** tiene la siguiente forma:

$$Nat_i = \beta_0 + \beta_1 Temp_i + \beta_2 PIB_i + \beta_3 Cont_i + \beta_4 Temp_i PIB_i + \beta_5 Temp_i Cont_i + \beta_6 PIB_i Cont_i + \epsilon_i$$

Tabla 1: *Parámetros estimados para el modelo de Boland.*

	Estimado	Error Est.	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

En la Tabla 1 se puede observar que, a un nivel de significancia del 0,05 existen variables no significativas para el modelo de referencia; por lo cual, se realizó el proceso de selección de variables obteniendo un modelo de la forma:

$$Nat_i = \beta_0 + \beta_1 Temp_i + \beta_2 PIB_i + \beta_3 Cont_i + \beta_4 Temp_i PIB_i + \epsilon_i$$

Tabla 2: *Parámetros estimados para el modelo anterior.*

	Estimado	Error Est.	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

Observe en la Tabla 2 que, a excepción del PIB todas las covariables son significativas para el modelo. Además, la Tabla 3 muestra los valores de cada críterio de selección para los modelos considerados, observe:

Tabla 3: *Resultados para el modelo de referencia y el loess*

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
Mod9	Loess	0,83	0,68	241,26
Referencia	NO	0,73	0,54	1143,42

De la Tabla 3 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 supera por mucho a cualquier modelo GAMLSS considerado.

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 (s) en función de las covariables temperatura y producto interno bruto (PIB); no obstante, para efectos del propósito del trabajo (predecir la tasa de natalidad) el modelo de regresión local fue descartado como modelo final ya que este método sirve más para realizar inferencias sobre la muestra en particular.