

Uso del modelo lineal generalizado para el pronóstico del número de visitas a museos en México: comparación entre regresión lineal ordinaria y regresión de Poisson

Use of the generalized linear model for forecasting the number of visits to museums in Mexico: comparison between ordinary linear and Poisson regression

Ana Dinora Guzmán Chávez ¹  - Fernando Guerrero González ¹ - Everardo Vargas Rodríguez ¹  

¹ Universidad de Guanajuato
 Autor de correspondencia: evr@ugto.mx

Recepción: 08-09-2021 / Aceptación: 08-11-2021
© Nova Scientia, bajo licencia Creative Commons

Resumen

Es de gran interés para los establecimientos, que ofrecen un producto o servicio, por ejemplo los museos, conocer la cantidad de posibles visitas en función de otras variables que los clientes realizarán en un periodo de tiempo específico. Esto con la finalidad de evaluar el grado de demanda que se tiene y al mismo tiempo, evaluar si las estrategias de difusión están funcionando o de lo contrario, tomar decisiones más acertadas para mejorar la calidad de atención y satisfacción con los visitantes. Para implementar el modelo lineal generalizado para estimar el número de visitas anuales a museos en territorio mexicano, en función de variables predictoras se utilizó una base de datos de 110 museos tomados del Instituto Nacional de Estadística y Geografía (INEGI) de los años 2017 y 2018. Los modelos empleados fueron la regresión lineal ordinaria (RLO) y la regresión de Poisson (RP) y estos se aplicaron sobre todas las componentes principales de once variables predictoras (página web, capacidad de recepción y atención, temática principal, colección permanente, titularidad, entrada, cuota adulto, descuentos, tipo de visitas, días abiertos del año y exposiciones temporales) para contrarrestar la multiconlinealidad. Además, se midió el coeficiente de determinación (R^2) entre los datos observados y estimados para determinar el modelo con el mejor ajuste. Las funciones se usaron para estimar el número de visitas de los años 2019 y 2020. Con el modelo que mejor ajuste dio, se realizó un análisis de errores de estimación. Comparando los datos observados de los años 2017 y 2018 con los resultados estimados con las funciones características de cada modelo, los coeficientes de determinación fueron $R^2 = 0.61$ para la RLO y $R^2 = 0.86$ para la RP, respectivamente. Para los años 2019 y 2020, los coeficientes de determinación que se obtuvieron fueron: $R^2 = 0.68$ y $R^2 = 0.31$ para la RLO, y $R^2 = 0.87$ y $R^2 = 0.84$ para la RP, respectivamente. El máximo error de estimación que se registró fue de entre 10 001 y 20 000 visitas anuales. Se muestra que el modelo más adecuado para el pronóstico de futuras visitas anuales a cualquier museo en México es la regresión de Poisson. Se cree que el alto ajuste a los datos observados con este modelo se debe a que estos no contaban con un exceso de ceros y además cumplían con el supuesto de equidispersión. Finalmente, con la función característica propuesta, se puede estimar el número de visitas anuales con un error máximo del alrededor del 10 %, lo cual es bajo en comparación al número máximo de visitas anuales que recibe un museo.

Palabras clave: museos; pronóstico; modelo lineal generalizado; regresión lineal ordinaria; regresión de Poisson; Newton-Raphson; bondad de ajuste; cultura; oferta cultural; difusión; consumo cultural; visitantes; variables

Abstract

It is of great interest for establishments that offer a product or service, for example museums, to know the number of possible visits as a function of other variables that customers will make in a specific period. This to evaluate the degree of demand that exists and at the same time, evaluate if the dissemination strategies are working or otherwise, make more appropriate decisions to improve the quality of attention and satisfaction with visitors. To implement the generalized linear model to estimate the number of annual visits to museums in Mexican territory as a function of predictive variables, a database of 110 museums taken from the Instituto Nacional de Estadística y Geografía (INEGI) for the years 2017 and 2018 was used. The models used were regression ordinary linear (RLO) and Poisson regression

(RP) and these were applied to all the principal components of eleven predictor variables (web page, reception and service capacity, main theme, permanent collection, ownership, entry, adult fee, discounts, type of visits, open days of the year and temporary exhibitions) to counteract multicollinearity. Besides, the coefficient of determination (R^2) was measured between the observed and estimated data to determine the method with the best fit. The functions were used to estimate the number of visits for the years 2019 and 2020. With the model that best fit, an analysis of estimation errors was performed. Comparing the observed data for the years 2017 and 2018 with the results estimated with the characteristic functions of each model, the determination coefficients were $R^2 = 0.61$ for the RLO and $R^2 = 0.86$ for the PR, respectively. For the years 2019 and 2020, the coefficients of determination obtained were $R^2 = 0.68$ and $R^2 = 0.31$ for the RLO, and $R^2 = 0.87$ and $R^2 = 0.84$ for the RP, respectively. The maximum error of estimation registered was between 10 001 and 20 000 annual visits. It is shown that the most suitable model to forecast future annual visits to any museum in Mexico is the Poisson regression. It is believed that the high adjustment to the data observed with the PR method is because these did not have an excess of zeros and fulfilled the assumption of equidispersion. Finally, with the proposed characteristic function, the number of annual visits can be estimated with a maximum error of around 10 %, which is low compared to the maximum number of annual visits that a museum receives.

Keywords: museums; forecasting; generalized linear model; ordinary lineal regression; Poisson regression; Newton-Raphson; goodness of fit; culture; cultural offer; diffusion; cultural consumption; visitors; variables

1. Introducción

Es de gran interés para los establecimientos, que ofrecen un producto o servicio, conocer la cantidad de posibles visitas que los clientes realizarán en un periodo de tiempo específico. Del mismo modo, estudiar si un determinado conjunto de variables guarda una relación entre si y el número de visitas, lo que resulta muy útil, pues es posible explicar qué porcentaje de variación en la variable de visitas puede ser explicada por el conjunto de variables que se estudian. Por ejemplo, en los museos, es importante conocer el número de visitas la cual es una de las métricas más importantes para evaluar el grado de demanda que se tiene y al mismo tiempo, evaluar si las estrategias de difusión están funcionando o de lo contrario, tomar decisiones más acertadas para mejorar la calidad de atención y satisfacción con los visitantes.

Los modelos lineales generalizados permiten describir la relación que existe entre una variable dependiente o de respuesta y una o más variables independientes o predictoras (Uyanik y Güler, 2013) (Graefe, Armstrong, Jones, & Cuzán, 2014). Entre estos modelos se encuentra la regresión lineal ordinaria (RLO) (Glenn, 2002) y la regresión de Poisson (RP) (Montgomery, Peck, y Vining, 2021) (Gardner, Mulvey, & Shaw, 1995). Por ejemplo, en (Azhari, Widyaningsih, y Lestari, 2018) propusieron un modelo de regresión de Poisson para pronosticar el promedio de goles por cada equipo del foot ball. Los autores utilizaron cuatro variables predictoras y obtuvieron una proporción de predicción correcta del 80 %. Por otro lado, Casey A. G. y otros usaron el mismo método, utilizando dos variables predictoras para predecir la actividad diaria de incendios forestales en Alaska (Graff *et al.*, 2020). En el área de la salud, Maria I. V y otros utilizaron una regresión multivariante de cuasi-Poisson para modelar la evolución de los nuevos casos de covid-19 del año 2020 en Chile (Vicuña, Cristián, y Quiroga, 2021). Los autores también utilizaron dos variables indicadoras predictoras (festividad y día de la semana) y la bondad de ajuste del modelo se evaluó con el Pseudo R^2 Heinzl-Mittlböck que fue del 95.3 %.

Con respecto al modelo de regresión lineal múltiple, se han publicado varios trabajos en los que se utiliza dicho modelo. Por ejemplo, Z. Ismail y otros publicaron en (Ismail, Yahya, y Shabri, 2009) el pronóstico de los precios del oro basándose en ocho factores económicos. Las estimaciones de los parámetros para el modelo se llevaron a cabo utilizando un paquete estadístico y se usó el error cuadrático medio como medida para determinar la precisión del pronóstico. Otro ejemplo, es el pronóstico del número de pacientes asmáticos en Kota Kinabalu, Sabah (Gabda, Jubok, Budin, y Hassan, 2008). Los autores utilizaron como variables la calidad del aire y factores meteorológicos siendo en total cuatro variables.

En este trabajo se presenta como caso de estudio el número de visitas a los museos en México. Los datos utilizados para el estudio corresponden a la base de datos publicada por el Instituto Nacional de Estadística y Geografía (INEGI). Se implementó el modelo lineal generalizado para pronosticar el número de visitas por año. Se aplicó la

regresión lineal ordinaria y la regresión de Poisson. Las variables que fueron usadas son: página web (PW), capacidad de recepción y atención (CRA), temática principal (TP), colección permanente (CP), titularidad (T), entrada (E), cuota adulto (CA), descuentos (D), tipo de visitas (TV), días abiertos al año (DAA) y exposiciones temporales (ET). Estas fueron agrupadas mediante variables ficticias para mantener datos cuantitativos y sencillos de manejar. Para mitigar una posible colinealidad entre variables se obtuvo un nuevo conjunto de variables por medio del análisis de componentes principales sobre las cuales se aplicaron los modelos de regresión. Se tomó el número de visitas registradas en los años 2017 y 2018 para encontrar las funciones características de ambos modelos, obteniendo un $R^2 = 0.61$ para la RLO y $R^2 = 0.86$ para la RP. Estas funciones se usaron para estimar las visitas de los años 2019 y 2020; y comparando las visitas estimadas con las observadas se obtuvo un $R^2 = 0.68$ y $R^2 = 0.31$, utilizando la función obtenida con el modelo de RLO, respectivamente. Para el modelo de RP se obtuvieron $R^2 = 0.87$ y $R^2 = 0.84$, siendo el modelo más idóneo para el ajuste. El máximo error de pronóstico que se registró fue de entre 10 000-20 000 visitas lo que equivale alrededor de un 10 % de un máximo de 150 000 visitas observadas por año. Finalmente, los modelos de pronóstico se implementaron en el lenguaje de programación python manteniendo un bajo costo computacional y de programación con un grado de respuesta inmediato.

2. Métodos, técnicas e instrumentos

Modelo lineal generalizado

El modelo lineal generalizado, relaciona linealmente la variable dependiente con los factores y las variables independientes (covariables) mediante una función de enlace, definida de la siguiente manera (Montgomery *et al.*, 2021):

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (1)$$

Donde: $g(\mu)$ es la función enlace y $\mu = E(y)$ es el valor esperado de la variable dependiente y , x_k es el k -ésimo predictor y β_k es k -ésimo factor. Entonces para n observaciones, se tiene que el conjunto de ecuaciones está descrito como:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \quad (2)$$

Donde: $i = 1 \dots n$ y al resolver el conjunto de ecuaciones, se obtiene una función característica del modelo:

$$g(\mu_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (3)$$

Donde: $\mu_p = E(\hat{y})$ es el valor esperado pronosticado de la variable $\hat{y}$ y ε el error asociado al pronóstico.

En particular, la regresión lineal ordinaria y la regresión de Poisson son dos modelos muy comunes para el pronóstico de eventos que dependen de una o más variables. La función de enlace para el caso ordinario es de la forma $g(\mu_p) = \mu_p$, quedando la función característica del modelo como:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (4)$$

Por otro lado, la función de enlace para el caso de la regresión de Poisson es de la forma $g(\mu_p) = \ln(\mu_p)$. En este caso μ_p es el número promedio de veces que se espera que ocurra el mismo evento en un intervalo de tiempo. Esta se deriva del supuesto de que existe una distribución de Poisson (Montgomery *et al.*, 2021). Por tanto, la función característica de la regresión de Poisson queda de la siguiente forma:

$$\hat{y} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon} \quad (5)$$

Optimización de los factores β con el método Newton-Raphson

El sistema de ecuaciones de la forma del modelo lineal generalizado para encontrar la función característica es:

$$\mathbf{X}\beta + \varepsilon - G = 0$$

$$\begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} - \begin{bmatrix} g(\mu_0) \\ g(\mu_1) \\ \vdots \\ g(\mu_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (6)$$

Donde: $\mathbf{X}$ es una matriz de dimensiones $n \times (k + 1)$ determinada por las variables predictoras, β es un vector de parámetros desconocidos, ε un vector que contiene los errores aleatorios y G es un vector aleatorio que registra los valores de las observaciones. Los vectores β , ε y G tienen dimensiones de $1 \times (k + 1)$.

Este sistema de ecuaciones se puede resolver con el método Newton-Raphson que está basado en la convergencia iterativa y la solución se representa de la siguiente manera:

$$\beta_{p+1} = \beta_p - J_p^{-1} f(\beta_p), \quad p = 0, 1, 2, \dots \quad (7)$$

Donde: β_{p+1} es el valor de la $(p + 1)$ -ésima iteración representado mediante el vector de parámetros solución de dimensiones $(n \times 1)$, β_p es el vector de dimensiones $(n \times 1)$ que contiene el valor aproximado de la p -ésima iteración de la solución, J_p^{-1} representa la inversa de una matriz Jacobiana con dimensiones $(m \times n)$ y $f(\beta_{p+1})$ representa un vector con dimensiones $(n \times 1)$ para las ecuaciones evaluadas con los valores iniciales del vector β_p . La inicialización del método comienza asignando valores aleatorios al vector β_p .

Es importante mencionar que no se requiere de contar con una extensa cantidad de datos de muestra para obtener un modelo de regresión idóneo que ofrezca un buen ajuste, por ejemplo pueden ser veinte registros de muestra. Lo importante es que la muestra contenga un rango de entre el valor mínimo y el valor máximo para cada una de las variables predictoras en el conjunto de observaciones, pues esto genera una mejor variación de los datos permitiendo una mejor definición inicial. Siempre que se tengan nuevas observaciones se debe entrenar el conjunto de muestra y ajustar nuevamente los métodos de regresión y lograr así una mejor exactitud.

Bondad de ajuste

En el análisis de regresión es importante determinar qué tan bien nuestra función de regresión ajusta a los datos y para ello se emplea el test de bondad de ajuste denominado como coeficiente de determinación múltiple R^2 , que consiste en medir que tan bien se ajusta un conjunto de valores observados con los esperados. El R^2 se expresa de la siguiente manera:

$$R^2 = 1 - \frac{VNE}{VT}, \quad 0 \leq R^2 \leq 1 \quad (8)$$

Donde: $VNE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ es la suma de cuadrados de la diferencia entre el valor observado y_i y el valor estimado $\hat{y}_i$ para la i -ésima observación, $VT = \sum_{i=1}^n (y_i - \bar{y})^2$ es la suma de cuadrados de la diferencia entre y_i y la media de los datos observados $\bar{y}$.

Categorización de la base de datos

Los datos de conteo que se analizaron pertenecen al número de visitas anuales en los museos del territorio de México correspondientes a los años 2017-2020. La información se obtuvo de la base de datos del INEGI que contaba con información de 1156 museos registrados de los cuales se tomó una muestra de 110 museos para obtener un modelo de

regresión que permita el pronóstico de dichas visitas anuales por museo (*Visitas*). La información fue categorizada en once variables predictoras y una variable de respuesta con diferentes valores según distinto criterios (ver tabla 1). Es importante mencionar que la muestra de museos contempla toda la variabilidad de las variables predictoras según su criterio, por lo que fue suficiente para obtener una ecuación de regresión múltiple característica de todo el conjunto de datos.

Tabla 1. Categorización de las variables predictoras y de respuesta.

Table 1. Categorization of predictor and response variables.

Variable predictora Predictor variable	Valor según criterio Criterion value
Página web (PW) Web page	1 = Sí (Yes); 2 = No (No)
Capacidad de recepción y atención (CRA) Reception and service capacity	1 = de 1 a 25 visitantes (from 1 to 25 visitors); 2 = de 26 a 50 visitantes (from 26 to 50 visitors); ... 8 = de 10 001 visitantes en adelante (from 1000 onwards visitors)
Temática principal (TP) Main theme	1 = Arqueología (Archeology); 2 = Arte (Art); 3 = Paleontología (Paleontology); 4 = Historia (History); 5 = Industria (Industry); 6 = Ciencias (Science); 7 = Tecnología (Technology); 8 = Ambiental/ Ecológico (Environmental / Ecological)
Colección permanente (CP) Permanent collection	1 = Sí (Yes); 2 = No (No)
Titularidad (T) Ownership	1 = Federal (Federal); 2 = Estatal (State); 3 = Municipal (Municipal); 4 = Universidad (University); 5 = Asociación (Association); 6 = Fundación (Foundation); 7 = Negocios (Business); 8 = Eclesiástico (Ecclesiastical); 9 = Un solo hombre (One only man)
Entrada (E) Entrance	1 = Gratuita (Free); 2 = Algunos días gratis (Some days free); 3 = Sin costo (free of charge)
Cuota adulto en pesos mexicanos (CA) Adult fee	1 = 0; 2 = 1 a 25; 3 = 26 a 50
Descuentos en pesos mexicanos (D) Discounts	1 = 0, 2 = de 1 a 10; 3 = de 11 a 20; 4 = de 21 a 40; 5 = de 41 a 60 y 6 = más de 61 (more than 61)
Tipo de visitas (TV) Type of visits	0 = Sin guía (No guide); 1 = Con persona guía (With guide person); 2 = Con audio guía (Audio guide); 3 = Otros medios (Other means).
Días abiertos al año (DAA) Open days of the year	Desde 0 hasta 365 From 0 until 365
Exposiciones temporales (ET) Temporary exhibitions	1 = Sí (Yes); 2 = No (No)
Número de vistas en un año (Visitas) Annual visits per year	1 = 0 a 1000; 2 = 1001 a 5000; 3 = 5001 a 10 000; 4 = 10 001 a 20 000; 5 = 20 001 a 50 000; 6 = 50 001 a 100 000; 7 = más de 100 001 (more than 100 001)

Una vez categorizados los datos se procedió a realizar el análisis de componentes principales, esto con la finalidad de contrarrestar una posible colinealidad entre las variables predictoras. Con este método se obtiene un nuevo conjunto de variables $CP_i = \sum_{i=0}^{10} p_i v_i$ las cuales son combinaciones lineales de las variables originales v_i que están multiplicadas a unos coeficientes p_i (ver tabla 2). Por mostrar un ejemplo, se tiene que la primera componente principal se expresa de la siguiente forma: $CP_0 = 0.3582 * PW - 0.3481 * CRA - 0.0704 * TP + 0.0313 * CP - 0.1874 * T - 0.4104 * E - 0.4884 * CA - 0.4328 * D - 0.0634 * TV - 0.2432 * DAA - 0.2279 * ET$.

Tabla 2. Coeficientes para obtener los valores de las componentes principales.

Table 2. Coefficients to obtain the values of the principal components.

Coeficientes p											
v	CP0	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP8	CP9	CP10
PW	0.3582	0.0385	0.0274	0.1165	0.1922	0.2622	0.1758	-0.6193	0.5777	0.0396	0.0222
CRA	-0.3481	-0.2988	-0.0814	0.2231	-0.2429	-0.0803	0.0450	0.3634	0.6703	0.2852	-0.0533
TP	-0.0704	-0.1183	0.7909	0.0907	0.2779	-0.2680	-0.3414	0.0104	0.1578	-0.2333	-0.0480
CP	0.0313	0.4459	-0.3475	0.3171	0.5078	-0.5350	0.0285	0.1356	0.1169	-0.0193	-0.0394
T	-0.1874	0.1419	0.1173	-0.6516	-0.1471	-0.4805	0.4347	-0.1899	0.1663	-0.0203	-0.0291
E	-0.4104	0.2195	0.2305	0.1092	0.1850	0.1361	0.0642	-0.2400	-0.2307	0.6938	-0.2633
CA	-0.4884	0.2456	-0.0324	-0.0066	0.0417	0.1531	-0.1524	-0.1390	0.0805	-0.1057	0.7840
D	-0.4328	0.2597	-0.1735	-0.0863	-0.0127	0.3006	-0.1865	-0.0914	0.1584	-0.4938	-0.5523
TV	-0.0634	-0.4560	-0.3333	-0.4722	0.4830	-0.0297	-0.4167	-0.0922	0.0491	0.1876	-0.0113
DAA	-0.2432	-0.3795	0.0413	0.1139	0.4643	0.1956	0.6498	0.1288	-0.1348	-0.2648	0.0345
ET	-0.2279	-0.3850	-0.1887	0.3847	-0.2500	-0.4054	-0.0527	-0.5652	-0.2131	-0.1418	-0.0412

En la figura 1 se muestra el porcentaje de varianza individual y acumulada explicada de las componentes principales. Se puede observar que el número de componentes que se necesitan para explicar al menos el 90 % de la variabilidad de los datos son 7. Este número no implica una reducción de variables considerable, por lo que se decidió utilizar todas las componentes para aplicar las regresiones, las cuales explican el 100 % de la variabilidad de los datos.

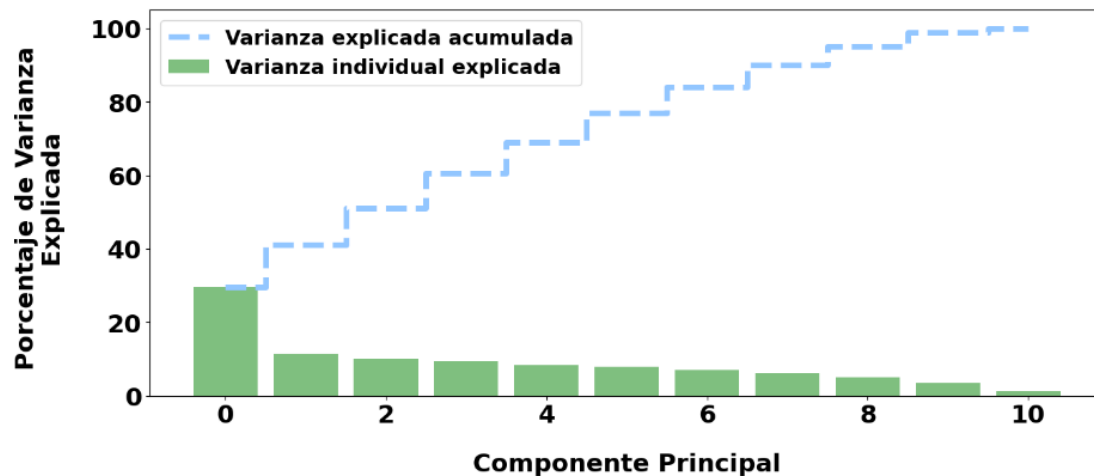


Figura 1. Varianza explicada acumulada e individual por cada componente principal.

Figure 1. Cumulative and individual variance explained for each principal component.

3. Resultados y discusión

Los dos modelos de regresión se aplicaron a los datos del 2017 y 2018, con los cuales se obtuvo las funciones características de ambos modelos. La función resultante que estima el número de visitas anuales en función de las componentes principales al aplicar el modelo de regresión lineal RLO es:

$$\text{Visitas} = 0.1723 - 0.6646 * PC0 - 0.0594 * PC1 - 0.0187 * PC2 - 0.0770 * PC3 - 0.2283 * PC4 - 0.3918 * PC5 - 0.3918 * PC6 + 0.5389 * PC7 - 0.0842 * PC8 + 0.1535 * PC9 + 0.0889 * PC10 \quad (9)$$

La función resultante de aplicar la RP es:

$$\ln(\text{Visitas}) = -0.112 - 0.1898 * PC0 - 0.1865 * PC1 - 0.0137 * PC2 - 0.0911 * PC3 - 0.0534 * PC4 - 0.0238 * PC5 - 0.1306 * PC6 - 0.1585 * PC7 - 0.0385 * PC8 - 0.0893 * PC9 - 0.0115 * PC10 \quad (10)$$

En la figura 2 se muestra el número de visitas anuales observadas y estimadas con las dos funciones obtenidas. En la figura 2a se muestran los resultados obtenidos con la RLO y en la figura 2b los obtenidos con la RP. Los coeficientes de determinación fueron de $R^2 = 0.61$ para la RLO y $R^2 = 0.86$ para la RP. Aquí, es importante mencionar que al redondear el valor de las visitas estimadas, no se registró un cambio considerable en los coeficientes de determinación; para la RLO se obtuvo un $R^2 = 0.59$ y para la RP un $R^2 = 0.83$.

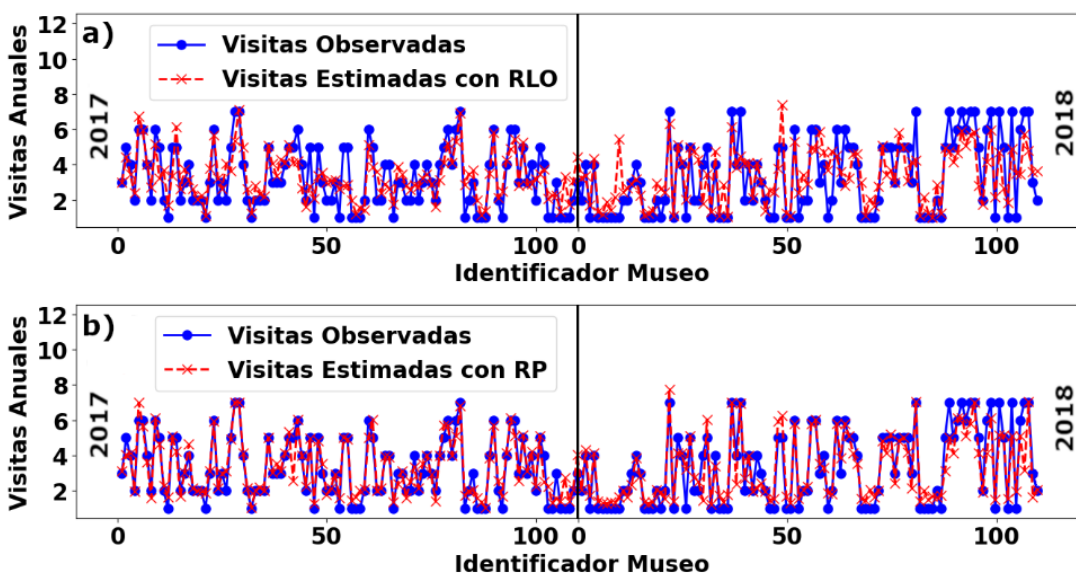


Figura 2. Número de visitas observadas durante los años 2017- 2018 y número de visitas esperadas obtenidas con: a) la regresión lineal ordinaria, y b) la regresión de Poisson.

Figure 2. Number of observed visits during the years 2017-2018 and the number of expected visits obtained with: a) the ordinary linear regression, and b) the Poisson regression.

La función característica obtenida con el modelo de RLO (ecuación 9) se utilizó para estimar el número de visitas anuales esperadas para los años 2019 y 2020 (ver figura 3). Comparando los resultados estimados con los datos observados, los coeficientes de determinación que se obtuvieron fueron: $R^2 = 0.68$ y $R^2 = 0.31$ para los años 2019 y 2020, respectivamente.

También la función característica obtenida con el modelo de RP (ecuación 10) se utilizó para obtener el número de visitas anuales esperadas para los años 2019 y 2020 (ver figura 4). El coeficiente de determinación que se obtuvo fue de $R^2 = 0.87$ y $R^2 = 0.84$ para los años 2019 y 2020, respectivamente. Es importante mencionar, que para aplicar los dos modelos de regresión utilizando los datos de los años 2019 y 2020, también se obtuvo el nuevo conjunto de variables aplicando análisis de componentes principales.

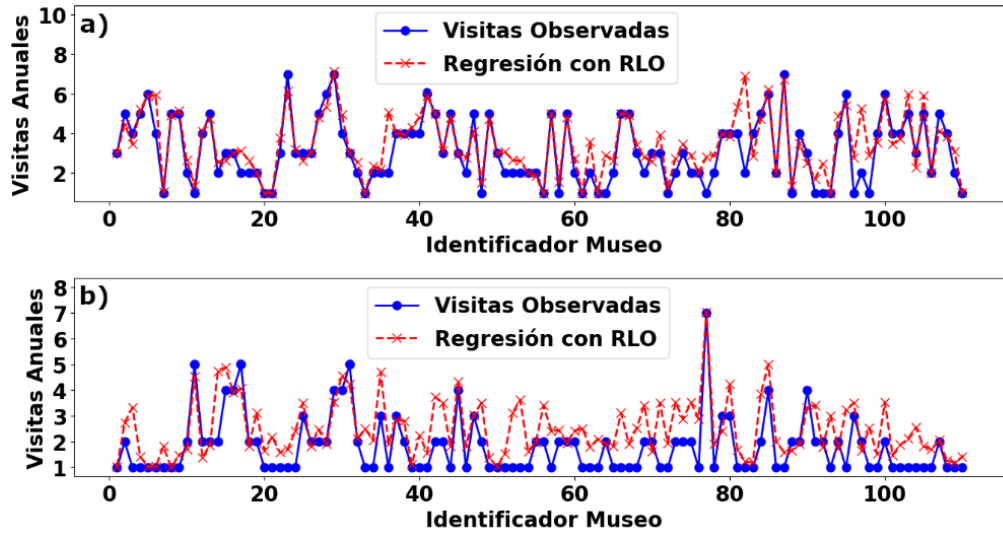


Figura 3. Número de visitas observadas y esperadas obtenidas con la RLO para los años: a) 2019, y b) 2020.
Figure 3. Number of observed and expected visits obtained with the RLO for the years: a) 2019, and b) 2020.

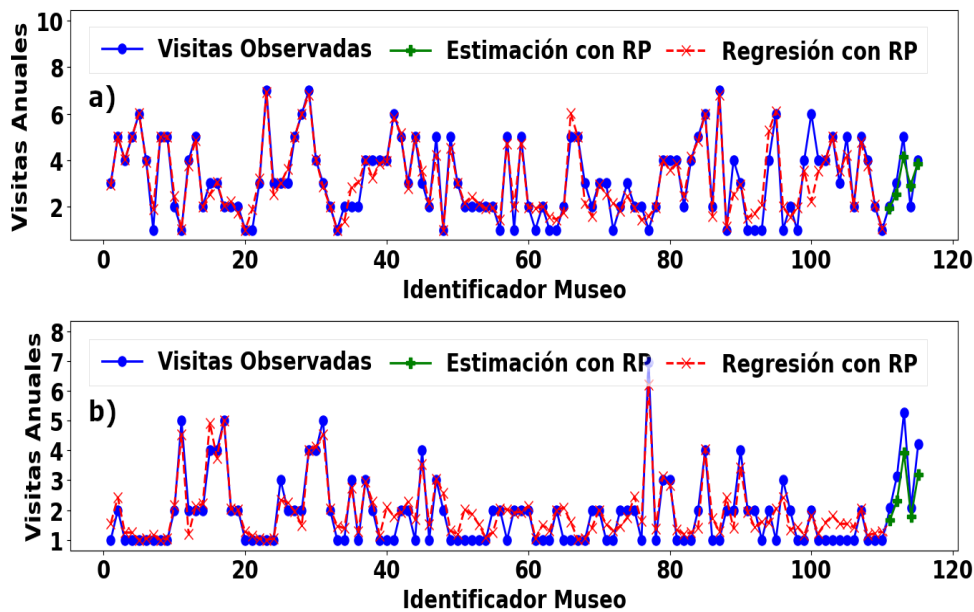


Figura 4. Número de visitas observadas y esperadas obtenidas con la RP para los años: a) 2019, y b) 2020.
Figure 4. Number of observed and expected visits obtained with the RP for years: a) 2019, and b) 2020.

Además, el modelo de RP, el cual fue el que mejor ajuste dio al número de visitas anuales observadas, se utilizó para estimar las vistas de cinco museos de prueba (museo 111 al 115) que no se tomaron de muestra para la obtención de dicho modelo. En la figura 4 se puede observar para cada uno de los años las visitas observadas y las visitas estimadas.

Para el análisis de errores, se utilizaron los datos obtenidos con el modelo de regresión de Poisson por ser el mejor método de ajuste. Para obtener el error, se restó el número de visitas anuales esperadas al número de visitas anuales observadas para los museos de muestra. Para cada año se tomaron los 5 museos que presentaron los mayores errores, obteniendo un total de 20 valores. De estos 20, se tomaron los cinco museos con los mayores errores. En la figura 5 se muestran los errores de estos cinco museos para los años de 2017-2020. Finalmente, con respecto a los errores que se presentan en las estimaciones del número de visitas anuales de los cinco museos de prueba, se tiene que para los años 2019 y 2020, el máximo error fue de ~ 1 , que en términos de visitas anuales es de 0 a 1000.

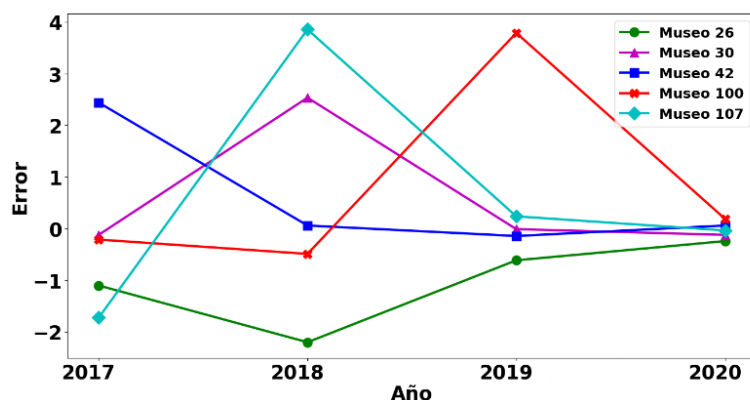


Figura 5. Error entre el número de visitas anuales esperadas y observadas en función del año para los cinco museos con los mayores errores.

Figure 5. Error between the number of expected and observed annual visits as a function of the year for the five museums with the biggest errors.

4. Conclusiones

Es notorio que con el modelo de RP se obtuvo el mejor ajuste con un coeficiente de determinación de ~85 % para el periodo de años de estudio, logrando una estimación muy similar a los datos observados. Este porcentaje alto se logró utilizando un nuevo conjunto de variables obtenidas mediante componentes principales. Además, también se cree que el alto ajuste a los datos observados se debe a que estos no contaban con un exceso de ceros y además cumplían con el supuesto de equidispersión, es decir, el promedio de visitas anuales $\mu = 3.42$ era similar a su varianza $\sigma^2 = 3.15$. Este supuesto es una asunción distribucional para el modelo de regresión Poisson requerida para una estimación con alta precisión. Finalmente, el bajo ajuste con el modelo de RLO, puede ser debido a que los datos no obedecían a una distribución normal y la variable dependiente no era continua, características tomadas en cuenta para un buen ajuste con este tipo de regresiones.

Tomando en cuenta que los resultados obtenidos con el modelo de RP son los que mejor ajuste dieron, fueron los que se utilizaron para el análisis de errores de estimación. El máximo error de estimación que se registró fue de alrededor de 4, en términos de visitas anuales, significa que fueron de entre 10 001 y 20 000; lo que equivale en promedio a un 10 % de un máximo de 150 000 visitas observadas por año. El análisis de componentes principales y los modelos de pronóstico se implementaron en el lenguaje de programación Python manteniendo un bajo costo computacional y de programación con un grado de respuesta inmediato.

Finalmente, en este trabajo se demostró que la implementación de un método de pronóstico basado en el modelo lineal generalizado es una estrategia viable para estimar el número de visitas anuales en algún establecimiento, específicamente en un museo, en el que se requiere un control del espacio, mejorar la experiencia de los visitantes o bien tomar decisiones más acertadas al percatarse de que el nivel de visitantes estimados será bajo.

5. Información adicional

No.

6. Agradecimientos

Agradecemos a la Universidad de Guanajuato por el financiamiento.

Información de los autores

Ana Dinora Guzmán Chávez ¹  0000-0002-5545-6852

Fernando Guerrero González ²

Everardo Vargas Rodríguez ³  0000-0001-5480-3384

Contribución de los autores en el desarrollo del trabajo

Los autores declaran que contribuyeron por igual para la realización de esta investigación.

Conflicto de interés

Los autores declaran que no existe conflicto de interés.

Referencias

- Azhari, H. R., Widyaningsih, Y., y Lestari, D. (2018). Predicting Final Result of Football Match Using Poisson Regression Model. *Journal of Physics: Conference Series* (pp. 1-5). Depok: IOPscience.
<https://doi.org/10.1088/1742-6596/1108/1/012066>
- Gabda, D., Jubok, Z. H., Budin, K., y Hassan, S. (2008, June). Multiple Linear Regression in Forecasting The Number of Asthmatics. *WSEAS Transactions on Information Science & Applications*, 5(6), 972-977.
- Gardner, W., Mulvey, E. P., y Shaw, E. C. (1995). Regression Analyses of Counts and Rates: Poisson, Overspersed Poisson, and Negative Binomial Modelos. *Psychological Bulletin*, 118(3), 392-404.
<https://doi.org/10.1037/0033-2909.118.3.392>
- Glenn, D. (2002). Multivariate Regression Trees: A new Technique for Modeling Species-Environment Relationships. *Ecology*, 83(4), 1105-1117. [https://doi.org/10.1890/0012-9658\(2002\)083\[1105:MRTANT\]2.0.CO;2](https://doi.org/10.1890/0012-9658(2002)083[1105:MRTANT]2.0.CO;2)
- Graefe, A., Armstrong, J. S., Jones, R. J., y Cuzán, A. G. (2014, March). Combining Forecasts: An Application to Elections. *International Journal of Forecasting*, 30(1), 43-54.
<https://doi.org/10.1016/j.ijforecast.2013.02.005>
- Graff, C. A., Coffield, S. R., Chen, Y., Foufoula-Georgiou, E., Randerson, J. T., y Smyth, P. (2020, July). Forecasting Daily Wildfire Activity Using Poisson Regression. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7), 4837-4851. <https://doi.org/10.1109/TGRS.2020.2968029>
- Ismail, Z., Yahya, A., y Shabri, A. (2009). Forecasting Gold Prices Using Multiple Linear Regression Method. *American Journal of Applied Sciences*, 6(8), 1509-1514.
- Montgomery, D. C., Peck, E. A., y Vining, G. G. (2021). *Introduction to Linear Regression Analysis*. New Jersey: John Wiley & Sons.
- Uyanik, G. K., y Güler, N. (2013, December 10). A Study on Multiple Linear Regression Analysis. *Precedia - Social and Behavioral Sciences*, 106, 234-240. <https://doi.org/10.1016/j.sbspro.2013.12.027>
- Vicuña, M. I., C. V., y Quiroga, B. F. (2021, April 23). Forecasting the 2020 COVID-19 Epidemic: A Multivariate Quasi-Poisson Regression to Model the Evolution of New Cases in Chile. *Frontiers in Public Health*, 9, 416-422. <https://doi.org/10.3389/fpubh.2021.610479>