Proyecto final HIGI: base de datos elegida por los alumnos.

Daniela Jiménez Pano
NUA:427292

June 17, 2021

Abstract

En el presente documento se llevará a cabo la recopilación del análisis estadístico de la base de datos elegido por lo alumnos ya que fue una asignación que la profesora del curso pidió que se realizara. Aunque el proyecto final consiste en el análisis de dos bases de datos, en este documento solamente se presentará la base de datos ya mencionada. Es necesario comentar que estos trabajos fueron realizados con ayuda del programa python, el cual fue un nuevo aprendizaje del último semestre.

1 Introducción

Dentro de los estudios demográficos, obtener la tasa de natalidad o conocer cómo es que aumentan o incrementan los nacimientos en un cierto país ayuda a tener una relación entre el número de nacimientos ocurridos en un cierto periodo de tiempo y la cantidad de efectivos de ese mismo periodo. Normalemente este lapso es de un año, y se puede leer como el número de nacimientos de una población por cada cierto número de habitantes en un año. [1]
Por lo tanto, como este era una información que llamaba la atención, se decidió investigar acerca de una base de datos con esta información, por lo que, gracias a los datos de la ONU se encontró una base de datos llamada UNdata, en la que se encontró el conjunto de datos: Live births by sex and urban/rural residence.[2] Con estos datos se buscaron hacer ciertas predicciones, el número total de personas en la lista nominal pasados 12 meses más el número total de casillas a necesitar tomando en cuenta la predicción y el número de nacimientos en México para el año 2018 y 2019, respectivamente.

Es necesario mencionar que la base de datos de donde se obtuvo la información anterior venían con datos de todos los países del mundo, además de que el total de nacimientos estaban acomodados con respecto al sexo y residencia (urbana o rural), sin embargo, para facilitar el análisis se decidió tomar los nacimientos del apartado 'Both sexes' (es decir, el total de la sumatoria de los dos sexos).

Por otro lado, los datos que se tomaron fueron de 2015 a 2017 ya que estos sí estaban completos con relacion a la zona de residencia, mientras que del año 2018 al 2019 solamente contaban con los nacimentos totales, sin tomar en cuenta el sexo o la zona de residencia, por lo cual el propósito de este análisis fue hacer una predicción del número de nacimientos en la zona urbana y rural para el año de 2018 y 2019, y con esto observar si la suma de estas cantidades resultaba siendo el total que viene en la página de UNdata.

2 Desarrollo

Para poder comenzar a realizar el análisis estadístico correspondiente, se subieron al entorno de python los documentos de cada una de las bases de datos y que el programa las pudiera leer:

```
import glob
files=glob.glob("./archivos//*.csv")
```

Llevando a cabo lo anterior, fue necesario acomodar las bases de datos con respecto al país de nuestro interés, en este caso México, por lo que cada uno de los documentos tenía que ser acomodado de esta manera:

```
data1=pd.read_csv(files[0],usecols=[0,1,2,3,6])
 data1_MX=data1[data1['Country']==' Mexico  ']
```

Además, como se podrá observar en el código anterior, se escogieron ciertas columnas ya que estas eran las que contenían la información importante para hacer el análisis, en este caso, el total de nacimientos, la zona, etcétera.

[9]:

	Country	Year	Area	Sex	Value
431	Mexico	2015	Total	Both Sexes	2096274
432	Mexico	2015	Total	Male	1068348
433	Mexico	2015	Total	Female	1027913
434	Mexico	2015	Urban	Both Sexes	1540464
435	Mexico	2015	Urban	Male	785867
436	Mexico	2015	Urban	Female	754585
437	Mexico	2015	Rural	Both Sexes	442539
438	Mexico	2015	Rural	Male	225103
439	Mexico	2015	Rural	Female	217436

Figure 1: Muestra del data frame generado para filtrar la información solamente para el país México

Lo siguiente a realizar fue concatenar los datos de 2015 hasta 2017 ya que estos 3 contenían la misma información, en este caso fue creado un data frame en los que se mostraba el total de personas nacidas en residencia urbana para así continuar con su regresión lineal y observar cómo es que los datos se comportaban:

```
total_urban=pd.concat([archivo,archivo2,archivo3], axis=0)
from IPython.display import HTML, display_html, display
display(HTML('<h2>Total de natalicios en zonas urbanas</h2>'))
display_html(total_urban)
```

Total de natalicios en zonas urbanas

	Country	Year	Area	Value
0	Mexico	2015	Urban	1540464
0	Mexico	2016	Urban	1476247
0	Mexico	2017	Urban	1468199

Figure 2: Data frame de la unión de los datos de 2015 a 2017 para las residencias urbanas

Teniendo esto, se decidió realizar una gráfica que nos permitiera conocer el comportamiento de los nacimientos para darnos una idea de cómo sería la predicción para 2018 y 2019, si incrementaban o disminuían:

Figure 3: Gráfica que muestra el comportamientos de los nacimientos de los años 2015, 2016 y 2017 en la residencia urbana.

Como se puede observar en la gráfica el comportamiento de los datos realiza una correlación negativa. Y a manera de poder predecir el valor de los nacimientos, se realizó una regresión lineal, es decir obtener los datos que satisfagan a la ecuación de la recta:

$$y = mx + b \tag{1}$$

Por lo tanto, gracias a la biblioteca sklearn fue posible obtener este modelo y aplicarlo para los años en los que se quería hacer una predicción:

```
from sklearn import linear_model
from sklearn.metrics import r2_score
regr=linear_model.LinearRegression()

x=total_urban['Year']
y=total_urban['Value']

X=x[:,np.newaxis]
print(X)
print(regr.fit(X,y))
print(regr.coef_)
m=regr.coef_[0]
b=regr.intercept_
y_p=m*X+b
print('y={0}*x+{1}'.format(m,b))
print(regr.predict(X)[0:3])
plt.scatter(x,y,color='blue')
plt.plot(x,y_p,color='red')
plt.title('Regresion lineal zonas urbanas', fontsize=16)
plt.xlabel('Aos',fontsize=13)
plt.ylabel('Valor',fontsize=13)
```

El cual nos presentó la siguiente gráfica de regresión:

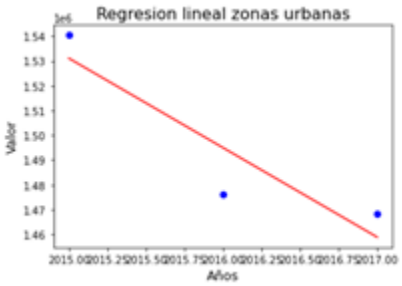

Figure 4: Gráfica de regresión lineal para los valores de los nacimientos en zonas urbanas

Por lo tanto, como era de esperar, la ecuación de la recta nos arrojó un resultado con pendiente negativa debido a que el comportamiento de los valores iba en decremento:

$$y = -36132.499999999985x + 74338089.99999997 \tag{2}$$

Ahora bien, teniendo la ecuación y regresión por mínimos cuadrados de los nacimientos en México para las residencias urbanas, se optó por hacer el procedimiento para las residencias rurales, entonces de igual manera se concatenaron los datos de nacimientos de 2015 a 2017 totales para las zonas rurales: Y se realizó la gráfica que nos ayudara a ver el comportamiento de los datos:
Asimismo, se ingresó el código (con variables diferentes) de la biblioteca sklearn para poder realizar regresiones lineales, y como se mencionó, esta regresión aplica para los nacimientos totales en zonas rurales y poder aplicar la ecuación de la recta a los años 2018 y 2019 que son los que buscamos predecir.

Figure 5: Data frame de la unión de los datos de 2015 a 2017 para las residencias rurales

Figure 6: Gráfica que muestra el comportamientos de los nacimientos de los años 2015, 2016 y 2017 en la residencia rural.

Por lo tanto se obtuvo la siguiente ecuación:

$$y = -1959.9999999999993x + 4391818.999999998 \tag{2}$$

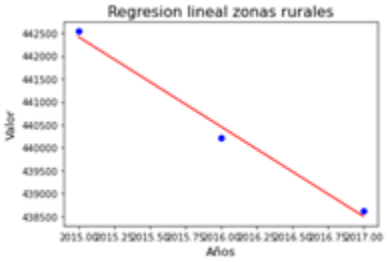

Figure 7: Gráfica de regresión lineal para los valores de los nacimientos en zonas urbanas

De manera similar que en la ecuación (2) la pendiente de la regresión lineal obtenida es negativa debido a que el comportamiento mostrado es un decremento en los datos de nacimientos.

3 Resultados y discusión

Gracias al análisis estadísitico realizado se pudo predecir los nacimientos en la zona rural y en la zona urbana para los años 2018 y 2019:

- Zonas urbanas:
  -Año 2018: 1422705.0 nacimientos.
  -Año 2019: 1386572.5 nacimientos.

- Zonas rurales:
  -Año 2018: 436538.9 nacimientos.
  -Año 2019: 434578.9 nacimientos.

Con estos resultados, se realizaron dos data frame, uno de las zonas urbanas y otro de las zonas rurales, para así poder hacer dos gráficas y observar cómo es que tendrían que verse los datos junto con las predicciones hechas:

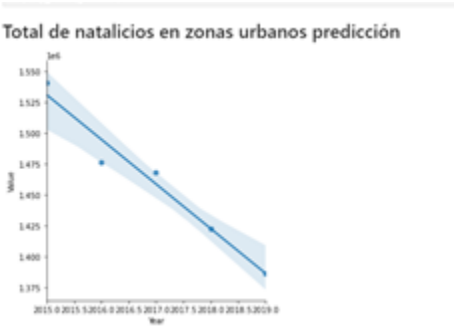

Figure 8: Gráfica que permite visualizar los datos de los nacimientos desde 2015 hasta 2019 en las zonas urbanas.

Figure 9: Gráfica que permite visualizar los datos de los nacimientos desde 2015 hasta 2019 en las zonas rurales.

Además se sumaron las predicciones de las zonas rurales y urbanas para 2018 y 2019 para realizar otro data frame en el cual se pudiera hacer una comparación entre los datos predecidos y los datos reales.

Comparación entre los datos de UnData y las predicciones de la regresión

	Country	Year	Area	Value
0	Mexico	2018	Total	2162535
0	Mexico	2019	Total	2092214
0	Mexico	2018	Total	1859243.99
0	Mexico	2019	Total	1821151.49

Figure 10: Data frame con la comparación de los datos reales y las predicciones de 2018 y 2019

Dando a ver que estos datos tenían ciertas diferencias, es decir que las predicciones seguían el patrón lineal, es decir, iban disminuyendo conforme el paso del tiempo, mientras que los reales no, estos aumentaban.
Haciendo dos data frames donde se unan los nacimientos totales de 2015 a 2017 junto con las predicciones y otro con los nacimientos totales de 2015 a 2017 con los datos reales, se realizaron dos gráficas para observar si seguian comportamientos parecidos o no.

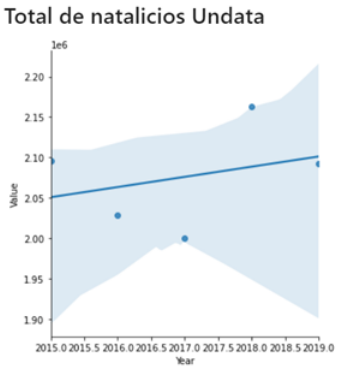

Figure 11: Gráfica que muestra el comportamiento de los nacimientos de 2015 hasta 2019 con los datos reales.

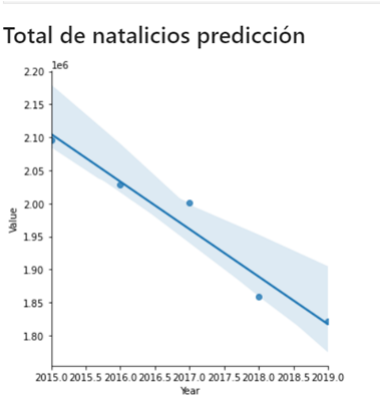

Figure 12: Gráfica que muestra el comportamiento de los nacimientos de 2015 hasta 2019 con los datos de la predicción.

4 Conclusiones

A manera de concluir este análisis, observando el comportamiento de las gráficas y con los valores obtenidos, podemos decir que los nacimientos o mas bien, la vida, no son aspectos que puedan ser predecidos, es decir, que la manera en la que crecen o decrecen estos datos en específico no siguen una tendencia lineal, por lo tanto, no es tan factible utilizar este método para poder decir qué es lo que iba a suceder en los años 2018 y 2019.

En lo personal, considero que este proyecto fue de mucha ayuda para mi desempeño como estudiante ya que python es una herramienta/programa que sé que tendré que usarla mucho para el desarrollo de distintos análisis, entonces, llevar a cabo un proyecto como este fue muy enriquecedor hablando del aspecto de adquirir nuevos conocimientos.

References

[1] Instituto Vasco de Estadística. Ficha metodológica: Tasa de natalidad. urlhttps://www.eustat.eus/documentos/datos/PI$_m etod/INDES - DEM02_c.asp, s.f.$

[2] UNdata. Live births by sex and urban/rural residence. url-http://data.un.org/Data.aspx?d=POPf=tableCode