INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE FÍSICA Y MATEMÁTICAS

MATERIA: ECONOMETRÍA

Modelo de Regresión Lineal Múltiple:

Medallero de los Juegos Olímpicos.

PRESENTA

Semiramís García de la Cruz

Julio 2020

ÍNDICE

INTRODUCCIÓN

Cada cuatro años, los ojos del mundo se centran en un evento por dos semanas: Los Juegos Olímpicos de Verano[i]. Actualmente, participan más de doscientas naciones, por lo que se consideran como la mayor competición del mundo deportivo.

Con excepción de sucesos históricos como guerras mundiales, o ahora una pandemia, los Juegos Olímpicos de Verano se han llevado a cabo cada cuatro años desde su primera emisión en la Era Moderna (Atenas, 1896), en consecuencia han evolucionado junto con la humanidad.

Su realización, nos hace voltear a ver el desempeño de nuestro país y nos hace preguntarnos, ¿qué factores hacen que un país gane más o menos medallas? Aquí es cuando entra en acción la econometría, definida como la aplicación de métodos estadísticos y matemáticos al análisis de datos económicos, con el propósito de dar un contenido empírico a las teorías económicas y verificarlas o refutarlas[ii]. Con estos conocimientos, es posible plantear modelos con datos sobre individuos, empresas o segmentos de la economía que no sería posible obtener en algún experimento, como en este caso.

En este proyecto, se plantea un modelo con el cual sea posible pronosticar la cantidad de medallas olímpicas que un país podría ganar considerando factores económicos, políticos y sociales.

PLANTEAMIENTO DEL PROBLEMA

En la investigación realizada para este proyecto, se trataron de analizar distintos factores que afectaran el medallero olímpico. Modelos remotos proponen dos variables como los mayores determinantes: el Producto Interno Bruto per cápita (PIB per cápita) y el tamaño de la población, con la premisa de que a mayor población y mayor potencial económico un país obtendría más medallas. Sin embargo, estos dos no son aspectos que afecten la competencia real, por lo que ninguno es directamente responsable por las medallas de los ganadores.

Ahora bien, como dijo Joseph Stiglitz: "Lo que medimos afecta las decisiones que tomamos". Debemos tomar dejar claro que un mayor potencial económico no nos indica mejoras en el bienestar de las personas. Por lo que para este modelo se propondrá también el análisis del Índice Desarrollo Humano (IDH), como un factor determinante ya que en él se toman en cuenta tres variables en una: el índice de salud (esperanza de vida), el índice de educación (alfabetización de adultos y tasa bruta de matrícula) y el PIB per cápita.

Si suponemos que el talento deportivo se reparte equitativamente en los habitantes, el tamaño de la población de cada país también influiría la posibilidad de tener un atleta y por ende, tener un medallista olímpico.

A mayor cantidad de participantes por país, habrá mayor probabilidad de obtener una medalla. La hipótesis en este caso es que una participación equitativa entre hombres y mujeres, incrementaría o incluso doblaría las oportunidades de tener atletas medallistas. Por lo tanto, la siguiente variable a considerar será la paridad de género, es decir, el número de mujeres participantes por cada país. Cabe mencionar que, se ha demostrado que al incrementarse la participación de mujeres, se decrementó el porcentaje de países con cero medallas[iii].

También tomaremos en cuenta si el país tuvo o tiene un gobierno soviético-comunista, debido a que en estos regímenes las Olimpiadas eran un instrumento político donde los atletas se entrenan desde muy jóvenes para llegar a un nivel de élite y reflejar el poderío del país.

Otro efecto interesante, en eventos como este, es la confianza que los equipos tienen al haber logrado obtener una medalla en emisiones anteriores. En otras palabras, ganar una medalla en los juegos anteriores aumenta la probabilidad de ganar en la siguiente emisión. Del mismo modo, los países que no ganan medallas pueden repetir ese resultado en un futuro.

Por último, tomaremos en cuenta el fenómeno "host". Luego de analizar los medalleros a lo largo del tiempo, se puede afirmar que el país anfitrión tiene ventajas: el equipo no viaja lejos, entonces está menos fatigado; el orgullo nacional y el hecho de estar familiarizado con las condiciones e instalaciones. Estas dos últimas variables serán evaluadas con variables *dummy*.

Para este análisis, centraremos nuestra atención en los juegos a partir de la década de los noventas, puesto que fue a partir de ese año que el IDH se comenzó a calcular por el Programa de las Naciones Unidas para el Desarrollo (PNUD). De modo que, centraremos la atención en los juegos de 1992 a 2016.

En la última competencia (Río 2016), se convocaron 207 naciones. Para fines prácticos tomaremos una muestra de 28 países que se escogerán con dos criterios. El primer punto es la constancia en los juegos, esto es, países con más participaciones en los juegos o que al menos cubran las que se analizarán. En segundo lugar, para tener variación en la muestra tenemos al menos dos participantes por continente; además se tomó en cuenta la accesibilidad a la información de cada país.

La fuente de los datos no olímpicos, como el IDH, provienen de las bases de datos del Banco Mundial (BM) y Naciones Unidas (UN). Por otro lado, toda información relacionada con las Olimpiadas, proviene de las páginas del Comité Olímpico Internacional (COI) o en algunos casos del comité de cada nación.

PLANTEAMIENTO DEL MODELO

En el curso de econometría, aprendimos que una herramienta para predecir es la aplicación del Modelo de Regresión Lineal. En este caso, se quiere entender la relación funcional entre la variable dependiente y las variables independientes, además de estudias las posibles causas de la variación que se pueda dar. En el Modelo de Regresión Lineal Múltiple se define

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

Donde

y: Es la variable dependiente

$x_1, x_2, \dots, x_k$: Son las variables independientes

β_0: Es el coeficiente constante o intercepto

$\beta_1, \beta_2, \dots, \beta_k$: Son los parámetros desconocidos que estarán bajo el efecto *ceteris paribus*

u: Es el término de error

En este trabajo, se propone analizar las variables mencionadas anteriormente para el MRLM, que se estructuran de la siguiente forma

$$medals = \beta_0 + \beta_1\, IDH + \beta_2\, logpop + \beta_3\, women + \beta_4\, host + \beta_5\, gob + \beta_6\, last + u$$

Donde

$medals$: Es el total de medallas obtenidas por una nación

IDH: Es el Índice Desarrollo Humano [0,1]

$logpop$: Es el logaritmo del total del tamaño de la población

$women$: Indica el número de mujeres participantes

$host$: Indica si el país es anfitrión *dummy* {1,0}

gob: Indica si el país tiene o tuvo un gobierno comunista *dummy* {1,0}

$last$: Es la cantidad de medallas obtenidas en los JJOO anteriores

Dependent Variable: MEDALS
Method: Least Squares
Date: 07/09/20 Time: 12:48
Sample: 1 196
Included observations: 196

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LOGPOP	2.221878	1.366067	1.626478	0.1055
WOMEN	0.095980	0.018287	5.248465	0.0000
IDH	-5.151023	6.231055	-0.826670	0.4095
HOST	4.519869	3.504927	1.289576	0.1988
GOBIERNO	2.102654	1.817070	1.157167	0.2487
LAST	0.645936	0.036917	17.49707	0.0000
C	-13.46192	12.70458	-1.059611	0.2907

R-squared	0.910389	Mean dependent var	25.15816
Adjusted R-squared	0.907544	S.D. dependent var	25.63332
S.E. of regression	7.794207	Akaike info criterion	6.979699
Sum squared resid	11481.68	Schwarz criterion	7.096775
Log likelihood	-677.0105	Hannan-Quinn criter.	7.027097
F-statistic	320.0194	Durbin-Watson stat	1.631739
Prob(F-statistic)	0.000000		

Sin embargo, no todas las variables mencionadas tienen el mismo nivel de significancia[iv].

Por lo tanto, el modelo se expresaría

$$medals = \beta_0\ logpop + \beta_1\ women\ + \beta_2\ host + \beta_3\ last\ + u$$

Donde

$medals$: Es el total de medallas obtenidas por una nación

$logpop$: Es el logaritmo del total del tamaño de la población

$women$: Indica el número de mujeres participantes

$host$: Indica si el país es anfitrión

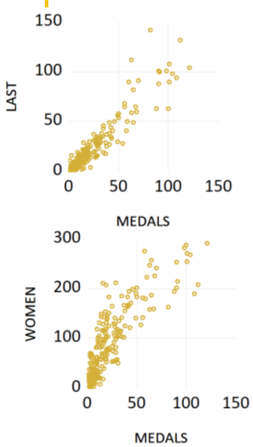
$last$: Es la cantidad de medallas obtenidas en los JJOO anteriores

SUPUESTOS DE GAUSS-MÁRKOV

En un modelo de regresión múltiple cada uno de los coeficientes de las variables independientes, explican un efecto parcial sobre la variable dependiente. Esa variación puede ser analizada individualmente bajo el efecto *ceteris paribus*, donde hacemos constantes a los demás factores. En este trabajo se presenta un modelo multivariable, que para garantizar su veracidad debe cumplir ciertas condiciones.

Los supuestos de Gauss-Márkov establecen lo que debe cumplir un estimador MCO (Mínimos Cuadrados Ordinarios) para que se considere ELIO (Estimador Lineal Insesgado Óptimo) o BLUE (por sus siglas en inglés Best Linear Unbiased Estimator). En este caso, se presenta una extensión de los cuatro supuestos de una regresión lineal simple (MRLS).

- Supuesto I: *Linealidad de los parámetros*



Este supuesto verifica que el modelo este compuesto por una relación lineal entre los parámetros beta, con respecto a la variable a explicar. Esta condición es bastante flexible, puesto que es posible tener relaciones no lineales, como en el caso de la variable logpop. Las demás variables cumplen con una relación lineal con la variable dependiente, por lo que este modelo cumple con el supuesto I.

- Supuesto II: *Muestreo aleatorio*

  Los datos recabados, son de tipo panel. Se recabó la información de 28 países a lo largo de siete olimpiadas, dando un total de 196 observaciones. Para seleccionar esos 28 países, se tomaron aquellos que tenían una participación en las últimas siete ediciones y tenían información disponible, hay al menos dos países por continente. Estos datos provienen de páginas oficiales del Banco Mundial y del Comité Olímpico Internacional.

- Supuesto III: *No hay colinealidad perfecta*

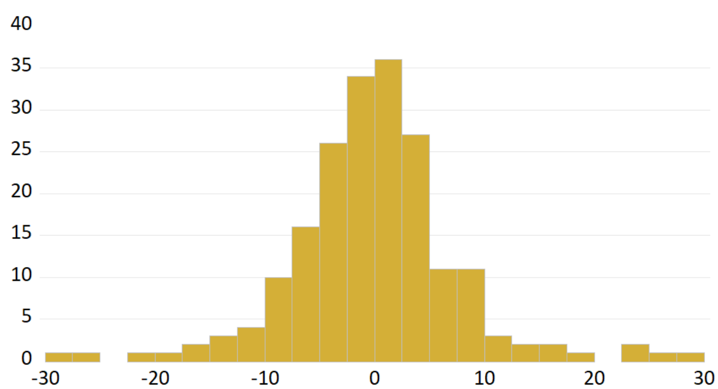
  En la muestra ninguna de las variables explicativas es constante. El supuesto, permite cierta correlación, pero en este caso no hay correlación perfecta.

Para comprobar que se cumpla esta condición, se pueden realizar regresiones lineales entre las variables independientes.

Covariance Analysis: Ordinary
Date: 07/09/20 Time: 13:03
Sample: 1 196
Included observations: 196

Covariance Correlation	MEDALS	LOGPOP	WOMEN	HOST	LAST	GOBIERNO	IDH
MEDALS	653.7148 1.000000						
LOGPOP	7.879182 0.536654	0.329750 1.000000					
WOMEN	1575.522 0.842507	20.69123 0.492648	5349.508 1.000000				
HOST	1.045372 0.220319	0.011810 0.110822	4.938958 0.363877	0.034439 1.000000			
LAST	652.3040 0.934722	7.658991 0.488659	1567.807 0.785348	0.617529 0.121916	744.9832 1.000000		
GOBIERNO	2.696793 0.301423	0.044494 0.221428	3.623178 0.141565	1.13E-18 1.74E-17	2.985423 0.312576	0.122449 1.000000	
IDH	1.035184 0.292292	-0.012895 -0.162116	5.152600 0.508584	0.001835 0.071381	1.062158 0.280937	-0.012487 -0.257614	0.019187 1.000000

- Supuesto IV: *Media condicional cero*
  En este supuesto requiere que el valor esperado del error u sea cero, dados cualquiera de los valores de las variables independientes. Este punto es esencial, para afirmar si los estimadores son insesgados.
- Supuesto V: *Homocedasticidad*
  En este supuesto, suponemos que la perturbación u cumple
  $$Var[u\ |logpop\ women\ host\ last] = \sigma^2$$
  Para comprobar este aspecto, se analizaron las series de datos para notar si existía una apariencia homocedástica.
- Supuesto VI: *Normalidad*
  Por último, se debe cumplir que el error poblacional u es independiente de las variables explicativas y tiene una distribución normal con media cero y varianza σ^2.
  Así aseguraremos que los estimadores tienen una varianza mínima.



Después de estudiar cada aspecto, podremos decir que nuestros estimadores son lineales, insesgados y óptimos, es decir ELIO.

ALCANCE

Con el modelo propuesto, se pretende estimar la cantidad de medallas que una nación obtendrá en los Juegos Olímpicos de Verano como una función lineal del tamaño de la población, el número de mujeres en el contingente olímpico nacional, su situación de anfitrión y los resultados obtenidos en la emisión anterior.

Los datos recabados para esta regresión son *datos de panel,* puesto que las observaciones son sobre múltiples fenómenos a lo largo de periodos de 4 años.

CÁLCULO DE LOS ESTIMADORES

En este trabajo, se optó por usar el software Eviews para calcular los estimadores.

Dependent Variable: MEDALS
Method: Least Squares
Date: 07/09/20 Time: 12:49
Sample: 1 196
Included observations: 196

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LOGPOP	3.054327	1.137990	2.683966	0.0079
WOMEN	0.082210	0.013970	5.884594	0.0000
HOST	5.564651	3.369587	1.651434	0.1003
LAST	0.666571	0.034978	19.05699	0.0000
C	-22.68285	8.307481	-2.730413	0.0069

R-squared	0.908981	Mean dependent var	25.15816
Adjusted R-squared	0.907075	S.D. dependent var	25.63332
S.E. of regression	7.813967	Akaike info criterion	6.974882
Sum squared resid	11662.09	Schwarz criterion	7.058507
Log likelihood	-678.5384	Hannan-Quinn criter.	7.008737
F-statistic	476.8656	Durbin-Watson stat	1.662992
Prob(F-statistic)	0.000000		

INTERPRETACIÓN

Variable	Coeficiente (beta)	Interpretación
logpop	3.054327	Un incremento del 1% en el tamaño de la población, incrementa las medallas ganadas en 0.03.
women	0.082210	Una mujer adicional en el contingente olímpico del país, incrementa 0.08 medallas.
host	5.564651	El ser anfitrión de los JJOO, incrementa 5.5 medallas ganadas para el país.
last	0.666571	Haber ganado una medalla en los JJOO anteriores, incrementa 0.66 medallas.
c	-22.6828	Este es el intercepto. Sin significado útil, puesto que ningún país podría "perder" 23 medallas.

ESTADÍSTICOS

Debido a la cantidad de observaciones, la hoja de cálculo de los estimadores se anexará en el trabajo.

- Estadístico R^2

  Existen dos formas de calcular este valor

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Donde

$$SCT = \sum_{i=1}^{n}(medals_i - \overline{medals})^2 \qquad SCR = \sum_{i=1}^{n}\hat{u}_i^2 = \sum_{i}^{n}(medals_i - \widehat{medals_i})^2 \qquad SCE = \sum_{i=1}^{n}(\widehat{medals_i} - \overline{medals})^2$$

$$SCT = 128128.097 \qquad SCR = 11662.095 \qquad SCE = 116465.621$$

Calculando de ambas formas se tiene

$$R^2 = 0.9089$$

Este coeficiente indica que, el modelo explica en un 90.89% a la variable real, es decir, las estimaciones del modelo se ajustan bastante bien al valor real.

- Varianza del error

  La varianza de los residuos nos ayuda a medir la variabilidad; sin embargo, para calcularlo debemos tomar en cuenta los grados de libertad.

  Este estadístico viene dado por

$$\hat{\sigma}_u^2 = \frac{SCR}{n-k-1}$$

  De donde obtenemos el valor

$$\hat{\sigma}_u^2 = 61.058$$

  La varianza es de 61.058, lo que describe que tan variada es la dispersión de las perturbaciones. Entre mayor sea este valor, nos dice que las mediciones son inciertas.

- Varianza de los estimadores beta

  Los cálculos de estos valores vienen dados por

$$var[\hat{\beta}_i] = \frac{\sum_{i=1}^{n} SCR}{(n-k-1)\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\boxed{\begin{aligned} var[\beta_1] &= 0.944716 \\ var[\beta_2] &= 0.000058 \\ var[\beta_3] &= 9.045642 \\ var[\beta_4] &= 0.00041 \end{aligned}}$$

- Error estándar de la regresión

  Este valor muestra la diferencia entre los valores reales y los estimados. Se calcula con la fórmula

$$e.e.r = \sqrt{\hat{\sigma}_u^2}$$

$$e.e.r = 7.8140$$

  En este caso, el modelo obtuvo un e.e.r de 7.8140, por lo que tenemos una falla en la predicción de aproximadamente 8 medallas. La posible razón de este resultado se deba en la gran diferencia en los resultados de los países, debido a que países como EEUU obtienen un promedio de 103 medallas por emisión, mientras la mayoría de los países tiene una media de alrededor de 15 medallas.

- Error estándar de los estimadores beta

  Estos valores se obtienen de la fórmula

$$e.e\left(\hat{\beta}_i\right) = \frac{\hat{\sigma}\sqrt{n^{-1}\sum_{i=1}^{n} x_i^2}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$\boxed{\begin{aligned} e.e.[\beta_1] &= 792.53 \\ e.e.[\beta_2] &= 1541.4 \\ e.e.[\beta_3] &= 1.5 \\ e.e[\beta_4] &= 390.09 \end{aligned}}$$

CONCLUSIONES

En este trabajo se construyó un modelo, para pronosticar el medallero olímpico de los países, en base a ciertas características: el tamaño de la población, la cantidad de mujeres participantes, su situación de ser o no anfitriones y su desempeño en la emisión anterior.

Inicialmente se propusieron también el IDH y si el país había tenido (o tiene) un régimen comunista; aunque, estas variables fueron desechadas por tener bajo nivel de significancia.

En el caso del IDH, se debe a que los datos no tienen variedad en su distribución. La razón es que no tenemos gran cantidad de naciones con índices bajos que estén participando constantemente. No obstante, existen países con IDH bajo, que han superado las expectativas. Esto se debe a un aspecto no observable: el talento nato. Aunque en este caso, los países obtienen sus medallas en disciplinas que no requieren mucha infraestructura como el atletismo. Además, está situación desencadena un efecto de “esperanza” en la juventud y niñez de dichos lugares, los cuales son inspirados por esos atletas sobresalientes.

Por otra parte, la hipótesis de que el tipo de gobierno influye en el medallero de un país, ha perdido importancia. Un ejemplo de esto lo tenemos con países como Rusia (antes Unión Soviética) que, si bien sigue teniendo altos resultados, disminuyó la cantidad de medallas que en promedio ganaba al perder algunos territorios al disolverse la URSS.

En cuento a las variables que sí se tomaron en cuenta, se muestra que sí existe un efecto de estos aspectos sobre la cantidad de medallas que un país obtiene.

Tanto en el caso de la variable "host" como "last", juegan un papel psicológico en los deportistas, eleva su confianza y se nota en los resultados. El elemento "logpop" y "women", miden más un aspecto demográfico y social. El primero, nos dice que entre mayor población tiene un país hay más probabilidades de tener un atleta de élite; el segundo, comprueba la hipótesis que diversos estudios han mencionado de que al incrementar la participación equitativa los países tienen mejores resultados. Aunque como pasó con la variable "gobierno" está variable irá perdiendo significancia, pues en la última emisión la mayoría de los equipos deportivos tenían paridad de género en sus contingentes, alcanzando en Río 2016 una participación femenina del 45%.

En conclusión, este modelo se ajusta bastante bien en la mayoría de los países. Los fallos más notables en la predicción, están en casos específicos como EEUU, Alemania y Japón; que tienen en común una educación deportiva de calidad, donde el grueso de su población ha practicado algún deporte y tiene la capacidad para intentar volverse un atleta de alto rendimiento si así lo deseara.

REFERENCIAS

Garduño, B. A. L. (2019, 6 noviembre). Análisis de datos para la toma de decisiones [Diapositivas]. Recuperado de https://www.slideshare.net/BlancaAzucenaLpezGarduo/anlisis-de-datos-para-la-toma-de-decisiones

Comité Olímpico Internacional. (2016). Historia y Datos de los JJOO y Mujeres. Recuperado de http://www.juegosolimpicosygenero.com/p/historia-de-los-jjoo.html

Clemente, Y. (2016, 3 agosto). Las mujeres en los Juegos Olímpicos. Recuperado de https://elpais.com/elpais/2016/07/21/media/1469128595_695055.html

Arancón, F. (2018, 25 mayo). La geopolítica de los Juegos Olímpicos. Recuperado de https://elordenmundial.com/la-geopolitica-de-los-juegos-olimpicos/

Instituto Valenciano de Investigaciones Económicas. (2010). *Desarrollo Humano: más allá del PIB*. Bancaja. Recuperado de http://web2011.ivie.es/downloads/docs/ch/ch114.pdf

Trivedi, Pravin & Zimmer, David. (2014). *Success at the Summer Olympics: How Much Do Economic Factors Explain?* Econometrics. 2. 10.3390/econometrics2040169.

Krishna, Anirudh & Haglund, Eric. (2008). *Why Do Some Countries Win More Olympic Medals? Lessons for Social Mobility and Poverty Reduction. Economic and Political Weekly*. 43. 143-151. 10.2307/40277720.

Indicators | Data. (2018). Recuperado de https://datos.bancomundial.org/indicator

Country Index - Olympics at Sports-Reference.com. (s. f.). Recuperado de https://web.archive.org/web/20090422081835/http://www.sports-reference.com/olympics/countries/

[i] Sólo se toman en cuenta los JJOO de verano porque los JJOO de invierno, están muy sesgados por tener menor cantidad de participantes, además de la ventaja que tienen aquellos países ubicados en las latitudes superiores.

[iii] Esta afirmación se da luego de leer una investigación publicada en el Instituto de Publicación Digital Multidisciplinario (MDPI), de Pravin K. Trivedi y David M. Zimmer.

[iv] Se llegó a esta conclusión al observar el contraste de la t. Rechazamos la hipótesis nula de no significatividad para las variables IDH y Gobierno.