Estimación de población de centros de atención infantil por medio de regresiones lineales

William Mendoza, Jefe del Departamento de Matemática
Eduardo Navas, Catedrático del Departamento de Electrónica e Informática
Universidad Centroamericana "José Simeón Cañas"

***Resumen*—Este artículo surge como una alternativa de solución al problema de estimar la población futura de algunos centros de atención de jóvenes y niños en riesgo social. La población de estos centros fluctúa mes a mes debido a diversos factores sociales, económicos y políticos, por lo que no hay forma de calcular exactamente cuántos jóvenes ingresarán y cuántos egresarán. Para realizar la estimación, se desarrolló un modelo matemático para proyectar la población de los centros en el siguiente mes basándose en regresiones lineales.**

I. Introducción

LAS características de la población de los centros de atención infantil, según se indagó en la institución encargada de la administración de estos centros son[1]:

1. Hay datos de población mensual atendida durante los últimos cinco años, a partir del año 2005.
2. A veces suceden eventualidades de orden político, social o económico que hacen variar la población de manera anómala durante normalmente un mes más o menos[2].
3. Los eventuales cambios en la legislación hacen que las tendencias cambien completamente.

Estas poblaciones no responden a modelos de población general como los modelos exponencial o logísticos, ya que tanto el aumento de la población como su disminución dependen de condiciones sociales, económicas y políticas variables y fluctuantes. El crecimiento de estas poblaciones no está relacionado con el tamaño actual de la población —característica básica del crecimiento exponencial[1]— y su decrecimiento no es debido a muertes naturales, por enfermedades o por agotamiento del ecosistema —característica del modelo logístico[1]—.

Por otro lado, no aplican los métodos de pronóstico como las medias móviles, ya que estos se basan en que el fenómeno es estacionario[3] (y los datos de población registrados no muestran patrones repetitivos); y finalmente las técnicas de suavización exponencial no siempre serían una buena alternativa ya que estas siempre le dan más ponderación a los datos más recientes[3].

Esto hace difícil poder tener una idea de cómo fluctuará la población en los siguientes meses; sin embargo, eso es fundamental para poder calcular el presupuesto de alimentación (y de otros rubros) de la institución encargada de los niños y las niñas.

A pesar de esta dificultad, es altamente prioritario tener una base objetiva para poder hacer un estimado del presupuesto de alimentación (y de otros rubros). Por ello, es necesario diseñar y aplicar un modelo matemático que permita aproximar la población de los centros, y a partir de esta estimación, calcular el presupuesto de alimentación, educación, vestuario, etc.

Así, este artículo expone un modelo de estimación de población que se espera contribuya a objetivizar el cálculo de la población siguiente, mes a mes, de los centros de atención de jóvenes en riezgo y así ayudar a calcular más apropiadamente el presupuesto para darles mejor atención.

[1] Sólo se mencionan las relevantes al volumen de la población

[2] Sucesos como redadas masivas, rescate de grupos de niños y niñas víctimas de trata de personas, etc.

II. Metodología

A continuación se describe el modelo diseñado para responder a las condiciones particulares del problema de predecir la población de los centros mencionados.

II-A. *Descripción del Modelo*

El modelo elegido consiste en dos rectas de regresión que se complementan. Una que considere datos de largo plazo disponibles, que aporte la información de la tendencia general de la población en el tiempo. Y otra que considere datos de corto plazo disponibles, que aporte información de los acontecimientos más recientes que modifican la tendencia general. Este planteamiento se ilustra en la figura 1.

En dicha figura, los círculos azules representan los datos de población registrados en cada período de tiempo (meses, años, etc.) y el círculo rojo representa la estimación puntual de la población para el período de tiempo que se desea estimar.

Si suponemos que los datos de población se irán registrando de manera consecutiva y sin borrar los datos antiguos, hay que definir desde cuándo se tomarán datos para la estimación de largo plazo (que en la figura 1 es el período de tiempo etiquetado como $t_{[I_L]}$). También hay que definir a partir de cuándo se considerán datos para la estimación de corto plazo (que en la figura 1 es el período de tiempo etiquetado como $t_{[I_C]}$).

Se asume que las dos rectas de regresión se calcularán hasta el último dato disponible (que en la figura 1 es el período de tiempo etiquetado como $t_{[I_L+N_L-1]}$).

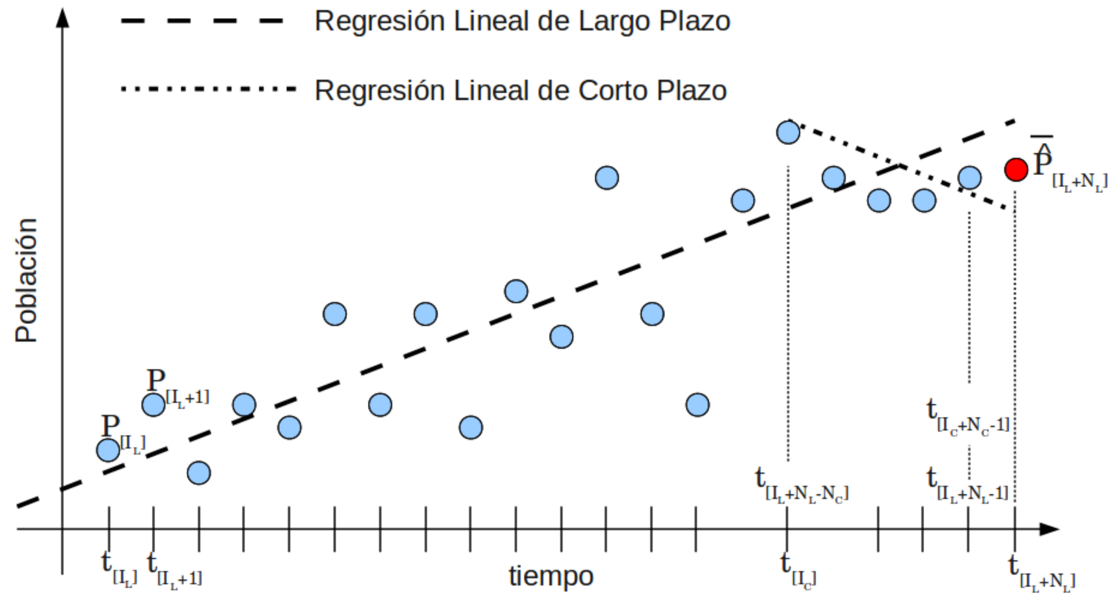

Figura 1. Diagrama de las rectas de regresión de largo y corto plazo

A continuación se presenta la nomenclatura utilizada en el modelo:

N_L: Número de períodos de largo plazo a considerar.

I_L: Índice del período inicial a considerar a largo plazo.

N_C: Número de períodos de corto plazo a considerar.

I_C: Índice del primer período de corto plazo a considerar.

Asumiendo que el último período será el mismo en la estimación de corto y largo plazo, se tiene que $I_L + N_L - 1 = I_C + N_C - 1$ (véase la figura 1) de lo que se concluye que:

$$I_C = I_L + N_L - N_C \tag{1}$$

$P_{[i]}$: El valor de la Población en el período i-ésimo.

$\overline{\hat{P}}_{[I_L+N_L]}$: Valor puntual estimado de la población en el período $I_L + N_L$ (ver figura 1), considerando la estimación de corto plazo y la de largo plazo. Nótese que $I_L + N_L = I_C + N_C$.

Adicionalmente, se requiere poder dar mayor peso a la estimación de corto plazo si los últimos períodos reflejan una tendencia fuerte que se espera perdure para el período a estimar, y a la inversa, poder dar mayor peso a la estimación de largo plazo si alguno de los últimos períodos reflejan una tendencia anómala que no se espera que perdure para el período a estimar. Para ello, se introduce el siguiente factor:

α_C: Peso ponderado para la estimación de corto plazo.

II-B. *Estimación puntual*

La forma elegida de estimar $\overline{\hat{P}}_{[I_L+N_L]}$ es utilizar un promedio ponderado de las estimaciones obtenidas con las dos regresiones lineales, la de corto plazo y la de largo plazo. Así:

$$\begin{aligned}\overline{\hat{P}}_{[I_L+N_L]} = {} & \left(A_L + B_L \cdot t_{[I_L+N_L]}\right)(1-\alpha_C) + \\ & \left(A_C + B_C \cdot t_{[I_L+N_L]}\right)\alpha_C \end{aligned} \tag{2}$$

Donde los parámetros A_L y B_L son los parámetros de la recta de regresión de largo plazo[3], y los parámetros A_C y B_C son los parámetros de la recta de regresión de corto plazo.

Dichos parámetros de regresión se calculan siguiendo las reglas de regresión lineal, tal como se explica en [4]. Las ecuaciones resultantes, para calcular los parámetros mencionados son (3)–(6).

II-C. *Precisión de la estimación*

La estimación puntual en sí misma no es muy útil, puesto que la probabilidad de que la predicción coincida con el valor real, es casi cero. Para ello, es necesario definir un *intervalo de confianza* alrededor de la estimación puntual, es decir un valor máximo y uno mínimo entre los cuales se espera que se encuentre el valor real con una cierta probabilidad.

Un propósito adicional de definir un intervalo de confianza es que en lugar de utilizar la estimación puntual para los cálculos del presupuesto, se utilice el límite superior del intervalo. Mejor que sobre y no que falte.

De [2] sabemos que el intervalo de confianza para la estimación de Y_{n+1} por medio de una recta de regresión del tipo $Y = A + B \cdot X$, está dado por (7).

[3] Estos parámetros están basados en la forma pendiente-intersecto de la recta: $Y = A + Bx$

Parámetros de la recta de regresión de Largo Plazo:

$$B_L = \frac{N_L \sum_{j=0}^{N_L-1} t_{[I_L+j]} \cdot P_{[I_L+j]} - \left(\sum_{j=0}^{N_L-1} t_{[I_L+j]}\right)\left(\sum_{j=0}^{N_L-1} P_{[I_L+j]}\right)}{N_L \sum_{j=0}^{N_L-1} \left(t_{[I_L+j]}\right)^2 - \left(\sum_{j=0}^{N_L-1} t_{[I_L+j]}\right)^2} \tag{3}$$

$$A_L = \frac{\sum_{j=0}^{N_L-1} P_{[I_L+j]}}{N_L} - B_L \frac{\sum_{j=0}^{N_L-1} t_{[I_L+j]}}{N_L} \tag{4}$$

Parámetros de la recta de regresión de Corto Plazo:

$$B_C = \frac{N_C \sum_{j=0}^{N_C-1} t_{[I_C+j]} \cdot P_{[I_C+j]} - \left(\sum_{j=0}^{N_C-1} t_{[I_C+j]}\right)\left(\sum_{j=0}^{N_C-1} P_{[I_C+j]}\right)}{N_C \sum_{j=0}^{N_C-1} \left(t_{[I_C+j]}\right)^2 - \left(\sum_{j=0}^{N_C-1} t_{[I_C+j]}\right)^2} \tag{5}$$

$$A_C = \frac{\sum_{j=0}^{N_C-1} P_{[I_C+j]}}{N_C} - B_C \frac{\sum_{j=0}^{N_C-1} t_{[I_C+j]}}{N_C} \tag{6}$$

$$\hat{Y}_{n+1} \pm t^*_{n-2,\alpha/2} \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right] s_e^2} \tag{7}$$

En donde $\hat{Y}_{n+1}$ es la estimación puntual de Y_{n+1}, $t^*_{n-2,\alpha/2}$ es el valor[4] de distribución acumulada de probabilidad *t de Student* con $n-2$ grados de libertad y α porcentaje de error en la estimación, n es el número de datos incluídos en la estimación, $\bar{x}$ es $\frac{\sum_{i=1}^{n} x_i}{n}$, y finalmente el término s_e^2 es el estimador insesgado de la varianza de los términos de error (véase [2] y [4]), calculado como:

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{SCE}{n-2} \tag{8}$$

Al numerador de esta expresión se le conoce como la Suma de los Cuadrados de los Errores (*SCE*), y el denominador se refiere a los grados de libertad.[2]

[4]Denotaremos aquí a la distribución *t de Student* como t^* para evitar confusión con la variable independiente de período de tiempo que estamos utilizando (t).

Puesto que son dos rectas de regresión en nuestro caso, hay dos radios para cada intervalo de confianza. Basándose en (7) y (8) estos radios son (10) y (11).

Finalmente, el intervalo de confianza seleccionado, tomando (2), (10) y (11), es:

$$\bar{\hat{P}}_{[I_L+N_L]} \pm [\rho_L \cdot (1 - \alpha_C) + \rho_C \cdot \alpha_C] \tag{9}$$

Es un intervalo centrado en la estimación puntual que se definió en (2) y con un radio promedio[5] entre ρ_L y ρ_C definidos en (10) y (11).

II-D. *Implementación del modelo*

El modelo se implementó como una hoja de cálculo dinámica.

Los datos se colócan de la siguiente manera en alguna hoja:

	A	B	C
1	**t**	**Período**	**Población**
2	1	2005-enero	$P_{[1]}$
⋮	⋮	⋮	⋮
12	11	2005-noviembre	$P_{[11]}$
13	12	2005-diciembre	$P_{[12]}$
⋮	⋮	⋮	⋮

[5]Es un promedio ponderado

Radio del intervalo de confianza de Largo Plazo:

$$\rho_L = t^*_{N_L-2,\alpha/2} \sqrt{\left[1 + \frac{1}{N_L} + \frac{\left(t_{[I_L+N_L]} - \frac{\sum_{i=0}^{N_L-1} t_{[I_L+i]}}{N_L} \right)^2}{\sum_{i=0}^{N_L-1} \left(t_{[I_L+i]}\right)^2 - N_L \left(\frac{\sum_{i=0}^{N_L-1} t_{[I_L+i]}}{N_L} \right)^2} \right] \cdot \frac{\sum_{i=0}^{N_L-1} \left(P_{[I_L+i]} - \left(A_L + B_L \cdot t_{[I_L+i]}\right)\right)^2}{N_L - 2}} \tag{10}$$

Radio del intervalo de confianza de Corto Plazo:

$$\rho_C = t^*_{N_C-2,\alpha/2} \sqrt{\left[1 + \frac{1}{N_C} + \frac{\left(t_{[I_L+N_L]} - \frac{\sum_{i=0}^{N_C-1} t_{[I_C+i]}}{N_C} \right)^2}{\sum_{i=0}^{N_C-1} \left(t_{[I_C+i]}\right)^2 - N_C \left(\frac{\sum_{i=0}^{N_C-1} t_{[I_C+i]}}{N_C} \right)^2} \right] \cdot \frac{\sum_{i=0}^{N_C-1} \left(P_{[I_C+i]} - \left(A_C + B_C \cdot t_{[I_C+i]}\right)\right)^2}{N_C - 2}} \tag{11}$$

En otra hoja, se especifican los parámetros de entrada:

	A	B
⋮	⋮	⋮
26	**Parámetros Modificables**	**Valor**
27	Período inicial a considerar a Largo Plazo	I_L
28	Número de períodos a considerar a Largo Plazo	N_L
29	Número de períodos a considerar a Corto Plazo	N_C
30	Peso ponderado para la estimación de Corto Plazo	α_C
31	Porcentaje de confianza del estimador	α

Para realizar las sumas selectivas y los demás cálculos dinámicos, se requiere de las siguientes funciones de hoja de cálculo (descripción adaptada de [5]):

- `FECHA(Año, Mes, Día)`
  Recibe tres números y devuelve un valor de fecha en función de ellos.
- `DIRECCIÓN(Fila, Columna, [Modo], [Tipo], [Hoja])`
  Devuelve una referencia de celda en forma de texto.
  Los parámetros son:
  - Fila: Representa el número de fila de la referencia de celda.
  - Columna: Representa el número de columna de la referencia de la celda (el número, no la letra).
  - Modo: Un número que determina el tipo de referencia devuelta:
    - 1: Absoluta (`$A$1`)
    - 2: Fila absoluta; Columna relativa (`A$1`)
    - 3: Fila relativa; Columna absoluta (`$A1`)
    - 4: Relativa (`A1`)
  - Tipo: Si se define en `0`, se utiliza la notación `R1C1` [6]. Si falta este parámetro o se define en otro valor distinto a `0`, se utiliza la notación `A1` (que es la tradicional).
  - Hoja: Representa el nombre de la hoja en forma de cadena. Si se omite, se asume que es una refencia a la misma hoja.
- `INDIRECTO(Referencia, [Tipo])`
  Devuelve la referencia especificada por una cadena de texto.
  Los parámetros son:
  - Referencia: Representa una referencia a una celda o a

un área (con formato de texto) para la que se devuelve el contenido.

Tipo: Si se define en `0`, se utiliza la notación `R1C1` [6]. Si falta este parámetro o se define en otro valor distinto a `0`, se utiliza la notación `A1` (que es la tradicional).

- `DESREF(Referencia, Filas, [Columnas], [Alto], [Ancho])`
  Devuelve el valor de una celda (o un área), desplazada una determinada cantidad de filas y columnas desde un punto de referencia concreto.
  Pueden crearse expresiones dinámicas como: `SUMA(DESREF(...))` para calcular sumas de áreas completas.
  Los parámetros son:
  Referencia: Es la referencia desde la que la función busca una nueva referencia.
  Filas: Es el número de filas en que la referencia se desplaza hacia arriba (valor negativo) o hacia abajo (valor positivo).
  Columnas: Es el número de columnas en que la referencia se desplaza hacia la izquierda (valor negativo) o la derecha (valor positivo).
  Alto: Es el alto vertical del área que comienza en la nueva posición de referencia, es decir, el número de filas que contendrá.
  Ancho: Es el ancho horizontal del área que comienza en la nueva posición de referencia, es decir, el número de columnas que contendrá.
- `DIST.T.INV(Probabilidad, GradosdeLibertad)`
  Calcula el inverso de la distribución t^* (La distribución t de Student).
  Los parámetros son:
  Probabilidad: Es la probabilidad asociada con la distribución t de dos colas. En nuestro caso, para estimar $t^*_{N_L-2,\alpha/2}$ y $t^*_{N_C-2,\alpha/2}$, este parámetro es $1-\alpha$.
  GradosdeLibertad: Es el número de grados de libertad de la distribución t. En nuestro caso, son $N_L - 2$ y $N_C - 2$.
- `NOD()`
  Esta función devuelve el valor especial `#N/A` (valor no disponible) que no aparece en los gráficos. Este valor es más apropiado que *cero* cuando no queremos que un valor aparezca en un gráfico.

II-E. *Validación del modelo*

Para verificar la validez del modelo, se aplicó a una serie de datos históricos de poblaciones de 16 centros.

Para cada centro, se tienen datos históricos de la población mensual atendida desde enero de 2005, denominado período 1, hasta marzo de 2010, denominado período 63, es decir, 63 períodos.

Para cada centro, se aplicó el modelo para predecir la población, desde el período 11 (noviembre de 2005) hasta el período 63 (marzo de 2010), es decir, 53 pruebas por cada centro. En total fueron aplicadas 848 pruebas (53×16).

Los parámetros de entrada establecidos en la prueba fueron los siguientes: $I_L = 1$, $N_C = 6$, $\alpha_C = 0{,}5$, $\alpha = 0{,}9$ y N_L se hizo iterar desde 10 hasta 62 para estimar la población de los períodos desde el 11 hasta el 63.

El resultado verificado fue que de las 848 ejecuciones, sólo 112 resultaron fallidas (13.2 %). Que una ejecución resultara fallida significa que el valor histórico real de la población resultó fuera del intervalo de confianza calculado en esa ejecución. Lo que en otras palabras significa que en esos meses la población sufrió drásticos cambios más allá de la tendencia registrada, con un incremento (o decremento) muy por encima de lo normal.

III. Conclusiones

1. Las poblaciones de centros de atención infantil, así como los centros de reinserción social y otros similares son difíciles de predecir, debido a que su fluctuación no depende de la población en sí, no tienen “estaciones” claramente definidas, y además, su “ecosistema” es controlado de forma artificial. Por lo que no responde a los modelos exponencial, logístico o similares.
2. El modelo propuesto tiene una certeza experimental aceptable de 86.8 % para estimar la población que se espera atender el siguiente mes, por lo que puede utilizarse para calcular los presupuestos para los centros.

Referencias

[1] Albert J. Jovell, *Análisis de regresión logística*, Centro de Investigaciones Sociológicas de Madrid, 1995.

[2] Paul Newbold, *Estadística para los negocios y la economía*, Prentice Hall, cuarta edición, pp. 403, 409.

[3] Monks Joseph G., *Administración de operaciones*, serie Schaum 1ª edición, México D.F., Mc. Graw Hill.

[4] José Hernández Salguero, *Elementos de Probabilidad y Estadística*, UCA Editores, 2002, pp. 288-289.

[5] Comunidad de usuarios de OpenOffice.org, *Wiki de Documentación de OpenOffice.org – Lista de funciones por Categoría*, revisado por última vez el domingo 5 de septiembre de 2010.
Enlaces permanentes:
http://wiki.services.openoffice.org/.../Calc:_Functions_listed_by_category
(http://wiki.services.openoffice.org/w/index.php?title=Documentation/How_Tos/Calc:_Functions_listed_by_category&oldid=149417)
http://wiki.services.openoffice.org/.../Calc/Funciones
(http://wiki.services.openoffice.org/w/index.php?title=ES/Traduccion/Calc/Funciones&oldid=100816)

[6] Comunidad de usuarios de OpenOffice.org, *Wiki de Documentación de OpenOffice.org – Notación R1C1*, revisado por última vez el domingo 5 de septiembre de 2010.
Enlaces permanentes:
http://wiki.services.openoffice.org/.../Calc:_R1C1_notation
(http://wiki.services.openoffice.org/w/index.php?title=Documentation/How_Tos/Calc:_R1C1_notation&oldid=176954)