

Capítulo 12

TÉCNICAS ESTADÍSTICO-MATEMÁTICAS PARA LA PRONOSTICACIÓN

Cuando en los estudios regionales se enfrentan cuestiones acerca de la indagación en el futuro de una situación determinada son factibles de utilizar para su definición diferentes métodos; entre los más relevantes se cuenta con: métodos económicos, métodos heurísticos y métodos estadístico-matemáticos.

Generalmente, el método más conveniente resulta el primero; sin embargo, es muy exigente en cuanto a indicadores y definiciones se refiere. Entre las técnicas más sobresalientes del segundo método se cuenta con el de expertos. El método estadístico-matemático es uno de los que más se emplean para realizar indagaciones, en el futuro, en el quehacer de la práctica cotidiana.

Sin ninguna pretensión de formalidad, se exponen el método estadístico-matemático. La explicación de estas técnicas no se hace de forma exhaustiva, se exponen las más utilizadas: el análisis de series históricas y el análisis de regresión.

El primer método es bivariado i.e. Toma en cuenta la variable tiempo y la variable observada a la cual se le desea determinar su tendencia secular.

La tendencia en un movimiento que se observa en la serie, basada por lo general en una dirección: si se hace un estudio de la población desde principios de siglo se podrá apreciar de manera general una tendencia creciente.

El análisis de regresión puede ser multivariado. En esta ocasión utilizaremos dos variables, una dependiente y otra independiente, sin atrevernos a estudiar la causalidad que requiera otro tratamiento. De todas formas se muestra que el procedimiento para considerar tres variables, aunque estos casos no los ejemplificaremos a los efectos de no alargar la exposición; ejemplificaciones que acometeremos dentro del propio curso, así como el estudio de la regresión con k variables.

Se da por sentado un conocimiento básico de estadística elemental que posibilite la comprensión de la línea de mejor

ajuste por el método de los mínimos cuadrados y el reconocimiento de su aplicación.

En el análisis de series históricas para estudiar la tendencia secular, el método seleccionado es el de mínimos cuadrados, el que trata en definitiva de ajustar una función de tal modo que la distancia entre el valor observado y el valor calculado sea mínima.

Quizás lo que tenga de novedoso la exposición es la prueba de aleatoriedad y la prueba sobre la existencia de la tendencia.

En el análisis de tendencia tenemos dos casos para la codificación de la variable temporal, de acuerdo con la cantidad de años que conforma la serie, el cual puede ser impar o par. Hecha esta distinción, el método de cálculo es igual para los dos casos.

ANÁLISIS DE LAS SERIES HISTÓRICAS

Para una mejor interpretación, se define una serie histórica, en ocasiones llamada también cronológica, a un conjunto de observaciones o de valores realizados en períodos de tiempo específicos, generalmente y, ello es conveniente, iguales.

Ejemplo de una serie histórica, puede ser el coeficiente de ocupación del suelo a través de diferentes períodos, o también la población residente en una provincia determinada en diferentes años.

Matemáticamente, una serie histórica está definida por los valores Y_1, Y_2, \dots, Y_n de una variable en períodos T_1, T_2, \dots, T_n . De ello se sigue que Y es una función de T , la cual se puede denotar formalmente por $Y = f(T)$.

Una serie histórica de la variable Y se puede graficar construyendo un gráfico cartesiano tradicional con los pares ordenados (T_i, Y_i) , el cual puede mostrar la población residente en la provincia Ciego de Ávila durante el período 1973-1979.

En una serie se han mostrado históricamente algunos movimientos característicos: movimientos a largo plazo o tendencia secular; movimientos cíclicos; movimientos estacionales y movimientos irregulares o aleatorios. En esta ocasión, nos interesa destacar la determinación del primero de ellos; lo cual puede realizarse de distintas formas, a saber: método de la mano alzada, método de los medios móviles, métodos de los semipro-medios y método de los mínimos cuadrados.

Como indicamos anteriormente, en este trabajo utilizaremos los dos últimos, en los cuales de lo que se trata es de hallar la ecuación de una línea de tendencia o curva de tendencia apro-

piada. Mediante esa ecuación podemos calcular la tendencia secular.

El proceso de cálculo de la tendencia sigue el diagrama 1.

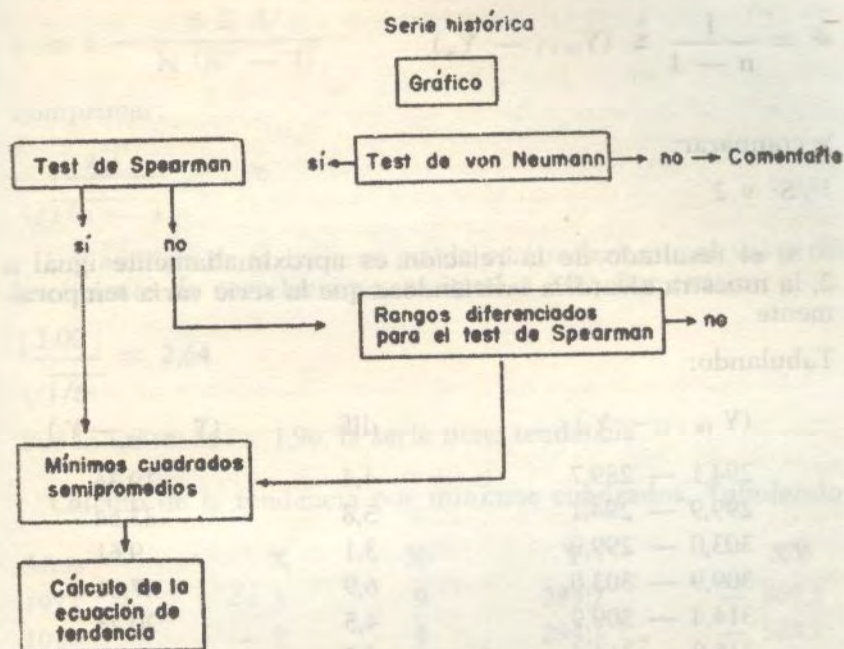


Diagrama 1

TENDENCIA LINEAL. MÍNIMOS CUADRADOS

Como ejemplo consideremos una sola unidad territorial de análisis y una sola variable: Ciego de Ávila y la población residente. Calcular la tendencia secular si es que existe:

Datos:

Años	Población en miles
1973	289,7
1974	294,1
1975	299,9
1976	303,0
1977	309,9
1978	314,4
1979	318,9

Determinación de la aleatoriedad de la muestra según el test de von Neumann

Calcular:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_{t+1} - Y_{t2})$$

y comparar:

$$s^2/S^2 \text{ v } 2$$

Si el resultado de la relación es aproximadamente igual a 2, la muestra aleatoria infiriéndose que la serie varía temporalmente

Tabulando:

$(Y_{(n+1)} - Y_t)$	dif.	$(Y_{(n+1)} - Y_t)$
294,1 — 289,7	4,4	19,36
299,9 — 294,1	5,8	33,64
303,0 — 299,9	3,1	9,61
309,9 — 303,0	6,9	47,61
314,4 — 309,9	4,5	20,25
318,9 — 314,4	4,5	20,25
Total	—	146,22

Sustituyendo:

$$s^2 = \frac{1}{6-1} (146,22) = 29,24$$

$$S^2 = \frac{(Y - \bar{Y})^2}{n} = 98,12$$

entonces:

$$\frac{s^2}{S^2} = \frac{29,24}{98,12} = 0,298$$

Como 0,2982 implica que la serie no es aleatoria.

Test de los rangos de Spearman. Comprobación si la serie es ordenada, es decir, tiene tendencia. Su cálculo se realiza mediante las ecuaciones siguientes:

$$\rho \equiv 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}$$

comprobar:

$$\frac{|\rho|}{\sqrt{1/n - 1}} \geq 1,96$$

Del análisis de la serie se puede comprobar que el valor del coeficiente de correlación ρ , es igual a 1.00. Por tanto:

$$\frac{|1.00|}{\sqrt{1/6}} = 2,64$$

Por lo tanto, $2,64 > 1,96$, la serie tiene tendencia.

Cálculo de la tendencia por mínimos cuadrados. Tabulando:

Años	X	X ²	Y	XY
1973	— 3	9	289,7	— 869,1
1974	— 2	4	294,1	— 588,2
1975	— 1	1	299,9	— 299,9
1976	0	0	303,0	0
1977	1	1	309,9	309,9
1978	2	4	314,4	628,8
1979	3	9	318,9	956,7
Total	0	28	2 129,9	138,1

La ecuación de la recta es:

$$Y = a + bx$$

el cálculo de los parámetros:

$$a = \frac{Y}{n} = \frac{2\,129,9}{7} = 304,27$$

$$b = \frac{XY}{X^2} = \frac{138,1}{28} = 4,932$$

Por lo tanto, la ecuación de la tendencia será:

$$Y = 304,27 + 4,932 X$$

TENDENCIA LINEAL. MÉTODO DE LOS SEMIPROMEDIOS

Un método que permite suavizar la tendencia secular lo constituye el de semipromedio. El método de semipromedios es una forma muy rápida de estimar una línea de tendencia recta. Los datos se dividen primero en dos partes, calculándose a cada parte los valores de tendencia central consecuentes, centrándolos en los puntos medios de los intervalos temporales. La recta que une ambas medias (o semipromedios) es la línea de tendencia estimada (Merrill).

Supongamos el ejemplo siguiente: Mortalidad infantil:

Años	X	Mort. Inf. Y (0/00)	Valor estimado de la tendencia	
1957	- 2	22,8	21,36	
1958	- 1	20,6	20,44	
1959	0	19,5	19,52	\bar{Y}_0
1960	1	17,8	18,60	
1961	2	16,9	17,68	
1962	3	16,0	16,76	
1963	4	15,6	15,84	
1964	5	15,4	14,92	
1965	6	13,8	14,00	\bar{Y}_6
1966	7	13,2	13,08	
1967	8	12,0	12,16	

Para ilustrar este método utilizaremos los datos de la tabla anterior acerca de la mortalidad infantil. Dado el que el período de 1957 a 1967 abarca un período impar de años continuos, el año central es 1962. De esta forma las medidas de ambas partes están basadas en igual cantidad de períodos y pueden centrarse fácilmente. El promedio durante el quinquenio 1957-1961, fue de:

$$Y_0 = \frac{22,8 + 20,6 + 19,5 + 17,8 + 16,9}{5} = 19,52$$

y durante 1963-1967, fue de:

$$Y_6 = \frac{15,6 + 15,4 + 13,8 + 13,2 + 12,0}{5} = 14,00$$

El promedio \bar{Y}_0 se centra en 1959 y el \bar{Y} en 1965. La línea de tendencia se obtiene uniendo ambos semipromedios.

La ecuación de la línea de tendencia recta es $Y' = a + bx$, donde Y' es el valor de tendencia estimado de la mortalidad infantil. Los valores de las constantes a y b dependen de los datos básicos y del año en que X es igual a cero. Si hacemos $X = 0$ en 1959, tenemos:

$$a = \bar{Y}_0 = 19,52$$

El coeficiente de pendiente o incremento anual de la tendencia b , es la variación anual. Despejando b en $Y' = a + bx$, obtendremos:

$$b = \frac{Y'_6 - a}{X}$$

Dado que $a = Y_0 = 19,52$ y $Y_6 = 14,00$, tenemos:

$$b = \frac{14,0 - 19,52}{6} \cong 0,92$$

y por tanto:

$$Y' \cong 19,52 - 0,92 X$$

Cuando una serie cronológica tenga un número par de años, el punto medio de la escala caerá entre los dos años centrales. En estos casos es conveniente utilizar una escala que mida a X en unidades semestrales. Los años se enumeran ... -5, -3, -1, 1, 2, 3, 5, ..., y la suma de X vuelve a ser cero (Merrill).

Veamos el ejemplo siguiente:

Años	X	Y	XY	X ²	Y'
1958	-5	4,2	-21,0	25	4,182
1959	-3	4,6	-13,8	9	4,676
1960	-1	5,2	-5,2	1	5,170
1961	1	5,7	5,7	1	5,664
1962	3	6,2	18,6	9	6,158
1963	5	6,6	33,0	25	6,652
Totales	0	32,5	17,3	70	—

Para obtener la línea de tendencia estimada calculamos:

$$a = \frac{\sum Y}{n} = \frac{32,5}{6} = 5,417$$

$$b = \frac{\sum XY}{\sum X^2} = \frac{17,3}{70} = 0,274$$

La línea de tendencia es entonces:

$$Y' = 5,417 + 0,274 X$$

Donde X está indicada en unidades semestrales con origen en 1960-1961.

TENDENCIA PARABÓLICA

La tendencia no siempre es posible representarla por una línea recta, en ocasiones se ajusta una curva del tipo de una parábola. Existen muchos tipos de parábolas, pero la pendiente secular se calcula siempre de igual forma y las distintas variantes lo que hacen es complicar los métodos de cálculo. La ecuación de una parábola es:

$$Y = a + bx + cx^2$$

los parámetros se calculan mediante:

$$a = \frac{\sum Y \sum X^4 - \sum X^2 Y \sum X^2}{n \sum X^4 - (\sum X^2)^2}$$

$$b = \frac{\sum XY}{\sum X^2}$$

$$c = \frac{n \sum X^2 Y - \sum Y \sum X^2}{n \sum X^4 - (\sum X^2)^2}$$

Supongamos que tenemos una serie histórica de la cantidad de asistentes a un determinado centro de servicios, necesitando establecer la ecuación de la tendencia para efectuar una estimación a largo plazo (Merrill).

Datos:

Años	Asistentes en miles	
1959	14,5	
1960	15,2	
1961	15,9	Si se grafican los datos se puede observar una línea de tendencia parabólica
1962	16,4	
1963	16,7	
1964	17,0	
1965	17,3	

Años	Asistentes en miles
1966	17,3
1967	17,2
1968	17,0
1969	16,7

Cálculo de la tendencia:

Años	X	Y	XY	X ²	X ² Y	X ⁴
1959	-5	14,5	-72,5	25	362,5	625
1960	-4	15,2	-60,8	16	243,2	256
1961	-3	15,9	-47,7	9	143,1	81
1962	-2	16,4	-32,8	4	65,6	16
1963	-1	16,7	-16,7	1	16,7	1
1964	0	17,0	0	0	0,0	0
1965	1	17,3	17,3	1	17,3	1
1966	2	17,3	34,6	4	69,2	16
1967	3	17,2	51,6	9	154,8	81
1968	4	17,0	68,0	16	272,0	256
1969	5	16,7	83,5	25	417,5	625
Total	0	181,2	24,5	110	1761,9	1958

Sustituyendo:

$$a = \frac{181,2 (1958) - (1\ 761,9) (110)}{11 (1958) - (110)^2} = 17,08$$

$$b = \frac{24,5}{110} = 0,227$$

$$c = \frac{11 (1\ 761,9) - (181,2) (110)}{11 (1958) - (110)^2} = 0,058$$

Consecuentemente la línea de tendencia es:

$$Y = 17,08 + 0,222 X - 0,058 X^2$$

TENDENCIA EXPONENCIAL O CURVA GEOMÉTRICA

Existen determinadas circunstancias donde ni la línea recta ni la parábola pueden expresar una tendencia secular, sobre todo cuando la tendencia sigue una ley de progresión geométrica, pudiendo entonces utilizar una función exponencial del tipo:

$$Y = ab^x$$

tomando logaritmos

$$\log Y = \log a + x \log b$$

los parámetros se calculan mediante:

$$\log a = \frac{\sum \log Y}{n}$$

$$\log b = \frac{\sum X \log Y}{\sum X^2}$$

Tomamos como ejemplo una situación similar a la del problema anterior únicamente varían los datos de la serie.

Datos:

Años	X	Y	log Y	X ²	X log Y
1960	-4	10	1,0000	16	-4,0000
1961	-3	17	1,2304	9	-3,6912

1962	— 2	19	1,2787	4	— 2,5574
1963	— 1	27	1,4314	1	— 1,4314
1964	0	39	1,5911	0	0
1962	1	55	1,7404	1	1,7404
1966	2	80	1,9031	4	3,8062
1967	3	100	2,0000	9	6,0000
1968	4	150	2,1761	16	8,7044
Total	0	497	14,3512	60	8,5710

Consecuentemente:

$$\log a = \frac{14,3512}{9} = 1,59458$$

$$\log b = \frac{8,5710}{60} = 0,14285$$

Siendo la ecuación de la tendencia:

$$\log Y = 1,59458 + 0,14285$$

Los antilogaritmos son los siguientes:

$$Y = 39,32 (1,39)^x$$

ANÁLISIS DE REGRESIÓN

El otro método que expondremos para su utilización en la pronosticación es el análisis de regresión. Dicho análisis es un instrumento que se emplea para valorar una relación existente entre dos o más variables.

Antes de comenzar un análisis de regresión es bueno señalar que debe existir una razón de carácter lógico o teórico que relacione las variables, lo cual sustenta la regresión y evita caer en graves errores. Los casos de regresión que expondremos son: regresión lineal simple y regresión lineal múltiple.

REGRESIÓN LINEAL SIMPLE

Otro modelo preferencial para el estudio de la pronosticación de la tendencia lo constituye el denominado análisis de regresión, instrumento harto poderoso de la Estadística Matemática. Este modelo y el análisis de correlación probablemente constituyen los métodos estadísticos más utilizados en la pronosticación.

Se denomina de esta manera al método que permite obtener ecuaciones donde solamente intervienen dos variables, una dependiente y otra independiente. Como habíamos dicho, un criterio general para encontrar la recta, es el llamado mínimos cuadrados.

La ecuación de la recta es:

$$Y = a + bx$$

el cálculo de los parámetros se realiza mediante

$$a = \bar{Y} - \bar{X}b$$

$$b = r \frac{S_y}{S_x}$$

el coeficiente de correlación se calcula mediante:

$$r = \frac{n \sum XY - (\sum X)(\sum Y)}{[(n \sum X^2 - (\sum X)^2)(n \sum Y^2 - (\sum Y)^2)]^{1/2}}$$

La significación de la correlación se determina a través de la experiencia siguiente:

$$t = r \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

para $n-2$ grados de libertad.

El valor crítico se obtiene de la tabla contenida en el apéndice de Siegel. El coeficiente de determinación es:

$$R^2 = r^2(100)$$

Ejemplo: Supongamos que tenemos las 14 provincias de Cuba en donde la variable dependiente es la densidad telefónica y la variable independiente es la densidad poblacional. Calcular la ecuación de regresión. Datos:

Provincia	Densidad poblacional	Densidad telefónica
1	59	15,8
2	103	24,0
3	2700	108,7

4	49	28,4
5	96	22,9
6	78	22,5
7	60	18,9
8	49	24,1
9	46	28,5
10	58	8,7
11	102	11,1
12	87	8,2
13	144	18,5
14	74	7,5

$$(\sum X)^2 = (3\ 750)^2 = 1\,389\,3750,0$$

$$(\sum Y)^2 = (347,8)^2 = 120\ 964,8$$

X	Y	X ²	Y ²	XY
59	15,8	3 481	249,64	932,2
103	24,0	10 609	576,00	2 472,0
2 700	108,7	7 290 000	11 815,69	293 490,0
49	28,4	2 401	284,00	1 391,6
96	22,9	9 216	524,41	2 198,4
78	22,5	6 084	506,25	1 755,0
60	18,9	3 600	357,21	1 134,0
49	24,1	2 401	580,81	1 180,9
46	28,5	2 116	812,25	1 311,0
58	8,7	3 364	75,69	504,6
102	11,1	10 404	123,21	1 132,2
87	8,2	7 569	67,24	713,4
144	18,5	20 736	342,25	2 664,0
74	7,5	5 476	56,25	555,0
3 750	347,8	7 377 457	16 370,90	331 219,9

El coeficiente de correlación es:

$$r = \frac{14 (331\ 219,9) - (3\ 750) (347,8)}{14 (7\ 373\ 457) - (3\ 750) 14 (347,8) - (16\ 370,9)}$$

$$r = 0,9548$$

La prueba de significación se calcula mediante

$$t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,9548 \sqrt{14-2}}{\sqrt{1-(0,9548)^2}} =$$

$$= \frac{0,9548 \sqrt{12}}{\sqrt{1-0,9116}} = \frac{3,307}{0,297} = 11,135$$

$$\frac{t}{0,95} = 178$$

El coeficiente de determinación es:

$$R = r^2 (100) = (0,9548)^2 (100) = 0,9116 (100) = 91,16 \%$$

La ecuación de la recta de regresión tiene la forma:
 $Y = a + bx$

Para el cálculo de los parámetros se utiliza:

$$a = \bar{X} - \bar{X}b = 24,84 - 264,64 (0,034) = 24,84 - 9,00 = 15,84$$

$$b = r \frac{SY}{SX} = 0,9548 \frac{25,20}{701,48} = 0,9548 (0,036) = 0,034$$

REGRESIÓN LINEAL MÚLTIPLE (TRES VARIABLES)

Los métodos de estimación de una variable por medio de otras, son similares a los métodos de regresión simple. Por ejemplo, si se quiere pronosticar el valor de Y (variable dependiente) en función de X_1 y X_2 (variable independiente), el problema se convierte en hallar el plano de mejor ajuste en sentido de los mínimos cuadrados.

Si la ecuación de regresión la escribimos como:

$$Y_{1,23} = a_{1,23} + b_{12,3} X^2 + b_{13,2} X_3$$

Donde $b_{12,3}$, $b_{13,2}$ son los coeficientes de regresión parcial, $a_{1,23}$ es una constante.

La obtención de los coeficientes de regresión se pueden determinar indirectamente a través de los denominados coeficientes B, los cuales se denominan también coeficientes de regre-

sión parcial standard. Se llaman standard porque se emplean cuando se utilizan medidas estandarizadas para todas las variables; se denominan parciales, porque los efectos de las otras variables permanecen constantes.

Las expresiones para su cálculo son:

$$b_{12,3} = \left(\frac{S_1}{S_2} \right) B_{12,3}$$

$$b_{13,2} = \left(\frac{S_1}{S_3} \right) B_{13,2}$$

Los coeficientes B a su vez se calculan mediante:

$$B_{12,3} = \frac{r_{12} - r_{13} \cdot r_{23}}{1 - r_{23}^2}$$

$$B_{13,2} = \frac{r_{13} - r_{12} \cdot r_{23}}{1 - r_{23}^2}$$

Para el cálculo de a:

$$a = \bar{X}_1 - b_{12,3} \cdot \bar{X}_2 - b_{13,2} \bar{X}_3$$

Error típico de estimación:

$$S_{1,23} = S_1 \sqrt{1 - R_{1,23}^2}$$

donde:

$$R_{12,3}^2 \equiv B_{12,3} r_{12} + B_{13,2} r_{13}$$

Ejemplo: tenemos las 14 provincias con las siguientes variables: Producción bruta (variable dependiente) Densidad poblacional y población (variables independientes). Se desea determinar la ecuación de regresión.

$$r_{12} = 0,900$$

$$r_{13} = 0,948$$

$$r_{23} = 0,842$$

$$X_1 = 110,407 \text{ (población bruta)}$$

$$X_2 = 264,64 \text{ (densidad poblacional)}$$

$$X_3 = 688,45 \quad (\text{población})$$

$$S_1 = 92,69$$

$$S_2 = 701,48$$

$$S_3 = 405,11$$

$$B_{12,3} = \frac{0,900 - 0,948 (0,842)}{1 - 0,842^2} = 0,351$$

$$B_{13,2} = \frac{0,948 - 0,900 (0,842)}{1 - 0,842^2} = 0,657$$

$$b_{12,3} = \frac{92,69}{701,48} (0,251) = 0,046$$

$$b_{13,2} = \frac{92,69}{405,11} = (0,653) = 0,149$$

$$a = 110,407 - 0,046 (264,44) - 0,149 (688,45) = 4,343$$

$$X_1 = 4,343 + 0,046 X_2 + 0,149 X_3$$

$$S_{12,3} = 92,69 (1 - 0,939) = 5,65$$

$$R_{12,3} = 0,351 (0,900) + 0,657 (0,948) = 0,939$$

BIBLIOGRAFIA DEL CAPÍTULO XII

1. MERRILL, W.: "Introducción a la estadística aplicada", en **Análisis de Varianza y Serie Cronológica**, Ministerio de Educación Superior, La Habana, 1977.
2. SPIEGEL, M. R.: **Theory and problems of statistics**, Edición Revolucionaria, La Habana, 1966 (hay edición en español).