TRABAJOS DE COLABORACION

Determinación de un intervalo de confianza para la predicción del déficit técnico en el S. O. A.

POR

Dr. JESUS M. VEGAS ASENSIO

I. Introducción

Se plantea con frecuencia en la realidad económica el siguiente problema: Se han observado los valores que toma una determinada variable estadística durante un cierto período de tiempo y en base a esta información se trata de predecir el comportamiento de dicha variable o magnitud en el futuro a corto o a largo plazo.

Este problema está comprendido dentro de la teoría de Procesos Estocásticos en la Estadística Matemática. Sin embargo es frecuente en la práctica efectuar extrapolaciones lineales, algunas veces sin suficiente fundamento estadístico, y es por ello por lo que vamos a abordar esta cuestión tomando como variable de aplicación el déficit técnico del Seguro Obligatorio del Automóvil en nuestro país.

De la Estadística elaborada por el Fondo Nacional de Garantía de R. C. se obtiene la siguiente serie cronológica, en donde Y_i expresa el porcentaje de la prima de riesgo consumido por la siniestralidad en cada año:

Año (t)	$Y_{(i)}$ (% consumido)
1965	71,3
1966	69,5
1967	73,9
1968	85,9
1969	92,8
1970	93,1
1971	97,5
1972	105,2
1973	106,8

En primer lugar hay que advertir que este Seguro entró en vigor el 1 de junio de 1965. Además, en los tres primeros años de su implantación, debido en muchos casos al desconocimento parcial de su funcionamiento (especialmente por parte de las víctimas cubiertas por el Seguro), no era aplicado con los actuales baremos, en cuanto a indemnizaciones, lo cual justifica el comportamiento estable de la serie en estos tres primeros años y el brusco incremento (de un 12 por 100) al pasar al año siguiente. De este examen previo se deduce la necesidad de eliminar estos tres primeros valores antes de aplicar técnica estadística alguna que nos permita prever el déficit técnico a partir de 1973 en base a la serie observada.

Con la nueva serie así obtenida vamos a proceder a ajustar una recta por el procedimiento de los mínimos cuadrados. Las ecuaciones normales del ajuste lineal, como es suficientemente conocido, son:

$$\Sigma Y_i = an + b\Sigma t_i$$
$$\Sigma Y_i t_i = a\Sigma t_i + b\Sigma t_i^2$$

En nuestro caso, operando, resulta $\Sigma t_i = 21$; $\Sigma Y_i = 581{,}3$; $\Sigma Y_i t_i = 2.107{,}6$; $\Sigma t_i^2 = 91$ y $N=6$ (la escala en años se ha tomado de 1 a 6), es decir:

$$\left.\begin{aligned} 581{,}3 &= a6 + b21 \\ 2.107{,}6 &= a21 + b91 \end{aligned}\right\}$$

de donde se obtienen los valores

$$a = 82{,}17 \qquad \text{y} \qquad b = 4{,}19$$

con lo que la recta mínimo cuadrática es:

$$Y_t = 82{,}17 + 4{,}19\, t$$

lo que indica que el incremento medio anual no acumulativo del déficit técnico es el 4,19 %, cifra que expresa la tendencia anual de esta serie estadística.

Con objeto de pronosticar el valor de la variable Y_t en los dos próximos años, damos a t los valores $t=7$ y $t=8$, resultándonos $Y_7 = 111{,}6$ e $Y_8 = 115{,}7$, es decir, un déficit técnico del 11,6 % para 1974 y del 15,7 % para 1975, en el supuesto de que no variaran las situaciones vigentes que han condicionado las prestaciones de las Entidades aseguradoras en el ramo (costes sanatoriales y farmcéuticos, honorarios médicos, indemnizaciones, etc.) durante los años que han servido de base para el ajuste lineal. En el supuesto de que varíen estas condiciones, por ejemplo, ante un nuevo convenio médico-hospitalario, habría entonces que elevar las anteriores previsiones en la cuantía que represente la repercusión de las citadas variaciones en el porcentaje cosumido de la siniestralidad.

Se nos plantean ahora dos cuestiones fundamentales:

a) ¿Es correcto hacer una predicción lineal en este caso?

b) En el supuesto de que el ajuste lineal fuese estadísticamente correcto como medio de extrapolar los valores de Y_t, ¿cómo obtendríamos un intervalo de confianza para el valor extrapolado con un tamaño determinado?

Para resolver ambas cuestiones haremos uso de la Teoría de la Regresión lineal simple, como vamos a mostrar a continuación.

II. Propiedades de los estimadores mínimo-cuadráticos

El modelo lineal simple en que una variable aleatoria η depende de una variable matemática t se caracteriza porque:

1) Para un t fijo es $\eta_t = \alpha + \beta t + \varepsilon(t)$ donde $\varepsilon(t)$ es una variable que expresa las perturbaciones de carácter aleatorio que influyen también en el comportamiento de η.

2) La distribución de $\varepsilon(t)$ se supone normal con media nula y varianza constante σ^2 (Axiomas de normalidad y de homocedasticidad).

3) El valor medio de η para un t fijo es una función lineal de t y de los parámetros desconocidos α y β, es decir,

$$E(\eta_t) = \mu_t = \alpha + \beta t$$

4) Las perturbaciones $\varepsilon(t_i)$ y $\varepsilon(t_j)$, si t_i es distinto de t_j, son variables aleatorias independientes (Axioma de no autocorrelación).

Tomando una muestra de valores de η_t, $Y_1, Y_2, Y_3, \ldots, Y_n$, que se obtendrá dando n valores a t, se pueden determinar las constantes a y b, estimaciones de α y β, que nos dan la recta de regresión empírica.

Los estimadores mínimo-cuadráticos a y b corresponden a los valores de estos parámetros que hacen mínima la suma

$$\sum_{i=1}^{n} (Y_i - a - b\,t_i)^2 = \varphi(a, b)$$

La condición necesaria de mínimo, aplicada a la función $\varphi(a, b)$ origina las igualdades:

$$\left.\begin{aligned} \frac{\partial \varphi}{\partial a} &= -2 \sum_{i=1}^{n} (Y_i - a - b\,t_i) = 0 \\ \frac{\partial \varphi}{\partial b} &= -2 \sum_{i=1}^{n} (Y_i - a - b\,t_i)\,t_i = 0 \end{aligned}\right\}$$

que permiten obtener las llamadas ecuaciones normales

$$\Sigma Y_i = an + b\Sigma t_i$$

$$\Sigma Y_i t_i = a\Sigma t_i + b\Sigma t_i^2$$

o $\bar{Y} = a + b\bar{t}$, siendo $\bar{Y}$ y $\bar{t}$ las medias aritméticas de las variables Y_i y t, respectivamente.

A partir de estas ecuaciones se obtienen las siguientes expresiones para a y b:

$$b = \frac{\Sigma (Y_i - \bar{Y})(t_i - \bar{t})}{\Sigma (t_i - t)^2} \qquad [1]$$

$$a = \bar{Y} - b\,\bar{t} \qquad [2]$$

La Teoría formal de la regresión lineal simple se origina al intentar conocer propiedades de estos estimadores mínimo-cuadráticos, los cuales, como cualquer otro estadístico, son variables aleatorias con una determinada distribución de probabilidad. Basta pensar en la posibilidad de extraer muchas muestras de tamaño n de pares de valores (Y_i, t_i), en cada una de ellas se puede ajustar una recta por mínimos cuadrados a la correspondiente nube de puntos, y con el conjunto de pares de valores a y b formar una tabla de doble entrada que será la imagen empírica de la distribución en el muestreo de los estimadores mínimo-cuadráticos.

Veamos a continuación las propiedades de estos estimadores:

1) *Insesgados.*

En efecto de [1] se obtiene

$$E(b) = \frac{1}{\Sigma (t_i - t)^2} \Sigma (t_i - \bar{t}) E(Y_i - \bar{Y}) = \frac{1}{\Sigma (t_i - t_i)^2} \Sigma (t_i - \bar{t}) \beta \Sigma (t_i - \bar{t}) = \beta$$

A su vez de la expresión [2] se obtiene:

$$E(a) = E(\bar{Y} - b\,\bar{t}) = E(\bar{Y}) - \bar{t}\,E(b) = \alpha + \beta\,\bar{t} - \beta\,\bar{t} = \alpha$$

2) *Consistentes.*

En efecto, aplicando el teorema de Tchebycheff a los estimadores b y a resulta:

$$P[\,|b - \beta| \leqslant \varepsilon] \leqslant \frac{\mathrm{Var}\,(b)}{\varepsilon^2} \qquad [3]$$

Como a su vez la varianza de b es

$$\mathrm{Var}(b) = E[b - \beta]^2 =$$

$$= E\left[\frac{\Sigma(Y_i - \overline{Y})(t_i - \overline{t})}{\Sigma(t_i - \overline{t})^2} - \beta\right]^2 = E\left[\frac{\Sigma(Y_i - \overline{Y})(t_i - \overline{t}) - \beta\,\Sigma(t_i - \overline{t}^2)}{\Sigma(t_i - \overline{t})^2}\right]^2 =$$

$$= E\left[\frac{\Sigma(t_i - \overline{t})[\varepsilon(t_i) - \overline{\varepsilon}]}{\Sigma(t_i - \overline{t})^2}\right]^2 = E\left[\frac{\Sigma(t_i - \overline{t})\,\varepsilon(t_i)}{\Sigma(t_i - \overline{t})^2}\right]^2 = \frac{E[\Sigma(t_i - \overline{t})\,\varepsilon(t_i)]^2}{[\Sigma(t_i - \overline{t})^2]^2} =$$

$$= \frac{\Sigma(t_i - t)^2\,E[\varepsilon(\overline{t}_i)^2] + \Sigma(t_i - \overline{t})(t_j - \overline{t})\,E[\varepsilon(t_i)\,\varepsilon(t_j)]}{[\Sigma(t_i - \overline{t})^2]^2} =$$

$$= \frac{\Sigma(t_i - \overline{t})^2\,E[\varepsilon(t_i)^2]}{(\Sigma(t_i - \overline{t})^2]^2} = \frac{\sigma^2}{\Sigma(t_i - \overline{t})^2}.$$

Sustituyendo en [3] queda

$$P[\,|b - \beta| \geqslant \varepsilon] \leqslant \frac{\sigma^2}{\varepsilon^2\,\Sigma(t_i - \overline{t})^2}$$

expresión que en virtud del Axioma de convergencia (*) tiende, cuando $n \to \infty$, a $\sigma^2/\varepsilon^2 D_0 n$, expresión que a su vez tiende a cero al tender n a infinito.

Procediendo análogamente con el estimador a, resulta

$$P[\,|a - \alpha| \geqslant \varepsilon] \leqslant \frac{\mathrm{Var}(a)}{\varepsilon^2} \qquad [4]$$

ahora bien, poniendo

$$\overline{\varepsilon(t)} = \frac{\Sigma\,\varepsilon(t_i)}{n}$$

resulta

$$\mathrm{Var}(a) = E[a - \alpha]^2 = E[(\overline{Y} - b\,\overline{t}) - (\overline{Y} - \beta\,\overline{t} - \overline{\varepsilon(t)})]^2 =$$

$$= E[\overline{t}(\beta - b) + \overline{\varepsilon(t)}]^2 = E[\overline{t}^2(\beta - b)^2 + \overline{\varepsilon(t)}^2 + 2\,\overline{t}\,\overline{\varepsilon(t)}(\beta - b)] =$$

$$= \overline{t}^2\,E[\beta - b]^2 + \frac{E[\Sigma\,\varepsilon(t_i)]^2}{n^2} + 2\,\overline{t}\,E[\overline{\varepsilon}(t)(\beta - b)] =$$

(*) Este Axioma dice que la sucesión de medias aritméticas y la sucesión de varianzas de t_i cuando $n \to \infty$, son convergentes, es decir,

$$\lim_{n\to\infty} \frac{\Sigma\,t_i}{n} = T_0 \quad \text{y} \quad \lim_{n\to\infty} \frac{\Sigma(t_i - \overline{t})^2}{n} = D_0$$

$$= \bar{t}^2 \operatorname{Var}(b) + \frac{E[\Sigma\, \varepsilon(t_i)]^2}{n^2} + 2\,\bar{t}\, E\left[\frac{\varepsilon(\bar{t})\, \Sigma\,(t_i - \bar{t})\,\varepsilon(t_i)}{\Sigma\,(t_i - \bar{t})^2}\right] =$$

$$= \bar{t}^2 \operatorname{Var}(b) + \frac{n\,\sigma^2}{n^2} + 2\bar{t}\,\frac{\Sigma\,(t_i - \bar{t})\, E\,[\overline{\varepsilon(t)}\,\varepsilon(t_i)]}{\Sigma\,(t_i - \bar{t})^2} =$$

$$= \bar{t}^2 \frac{\sigma^2}{\Sigma\,(t_i - \bar{t})^2} + \frac{\sigma^2}{n} + 2\bar{t}\,\frac{\sigma^2\, \Sigma\,(t_i - t)}{n\, \Sigma\,(t_i - \bar{t})^2} = \frac{\sigma^2}{n}\left[1 + \frac{n\, t^2}{\Sigma\,(t_i - t)^2}\right]$$

ya que $\Sigma\,(t_i - \bar{t}) = 0$.

Sustituyendo este valor de Var (a) en [4] resulta:

$$P\,[\,|\,a - \alpha\,| \geqslant \varepsilon] < \frac{\operatorname{Var}(a)}{\varepsilon^2} = \frac{\sigma^2}{n\,\varepsilon^2}\left[1 + \frac{n\, t^2}{\Sigma\,(t_i - t)^2}\right]$$

expresión esta última que tiende sucesivamente a

$$\frac{\sigma^2}{n\,\varepsilon^2} \cdot \left(1 + \frac{T_0^{\;2}}{n\, D_0}\right)$$

y a cero al crecer infinitamente n. Por tanto hemos probado la convergencia en probabilidad de a y b a los parámetros α y β, respectivamente.

3) *Eficientes.*

Partiendo de los Axiomas de homocedasticidad y de no autocorrelación se demuestra (*) que los estimadores a y b son los de varianza mínima entre todos los estimadores insesgados que vengan expresados como funciones lneales de las Y_i; esta condición de ser *lineales insesgados óptimos permite* calificar de eficientes a los estimadores mínimo-cuadráticos.

4) *Suficientes.*

El teorema de Fisher-Neymann dice que "la condición necesaria y suficiente para que el estimador $T = T(x_1, x_2, \ldots, x_n)$ sea suficiente para el parámetro θ es que se pueda hacer una descomposición factorial del tipo $L(x, \theta) = g[T(x), \theta] \cdot h(x)$ en que $L(x, \theta)$ es la función de verosimilitud de la muestra y $h(x)$ no depende de θ.

(*) Arnaiz, G., "Estimación por el método de los mínimos cuadrados". Estadística Española 1960.

Para probar el carácter de suficiente de los estimadores mínimo-cuadráticos estableceremos la función de densidad conjunta de las n perturbaciones aleatorias $\varepsilon(t)$, teniendo en cuenta que las citadas variantes son normales, independientes y de varianza constante σ^2, por lo que tendrá la expresión

$$f[\varepsilon(t_1)\,\varepsilon(t_2)\,\ldots\,\varepsilon(t_n)] = f[\varepsilon(t_1)]\cdot f[\varepsilon(t_2)]\,\ldots\,f[\varepsilon(t_n)] = f[\varepsilon(t)]^n$$

es decir,

$$f[\varepsilon(t_1)\ldots\varepsilon(t_n)] = \left\{\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{\varepsilon(t)^2}{2\sigma^2}}\right\}^n = k\cdot e^{-\frac{1}{2\sigma^2}(\Sigma(Y_i-\alpha-\beta t_i)^2} \qquad [5]$$

Podemos hacer la siguiente transformación

$$\begin{aligned}(Y_i-\alpha-\beta t_i)^2 &= [(Y_i-a-b\,t_i)+t_i(b-\beta)+(a-\alpha)]^2 =\\ &= (Y_i-a-b\,t_i)^2+[t_i(b-\beta)+(a-\alpha)]^2+\\ &\quad +2\,[(b-\beta)(Y_i-a-b\,t_i)\,t_i+(a-\alpha)(Y_i-a-b\,t_i)]\end{aligned}$$

Al efectuar el sumatorio desde $i=1$ hasta n se anula el último término anterior por lo que la función exponencial [5] puede escribirse como el producto de dos exponenciales, en una de las cuales no figuran los parámetros α y β y en la otra sí figuran tales parámetros, así como sus estimadores a y b, pero dicha función es independiente de Y_i, lo que prueba el carácter suficiente de los estimadores mínimo-cuadráticos.

III. Intervalo de confianza del valor extrapolado

Si en la recta de regresión empírica $Y_t=a+bt$, damos a t un valor $t=t_0$, obtendremos el correspondiente $Y_0=a+bt_0$; pues bien, se demuestra (*) que dicho valor Y_0 es un estimador lineal, insesgado óptimo del pronóstico de η_t para un valor dado t_0 de la variable t.

Esto justifica que en la práctica, como hemos hecho en el primer punto de este trabajo, se tome como predictor de los futuros valores que pueda tomar la variable aleatoria η_t, el correspondiente valor de Y_t en la recta de regresión empírica.

La varianza del predictor Y_0 es igual a (**)

$$\mathrm{Var}(Y_0) = \sigma^2\left[\frac{1}{n}+\frac{(t_0-t)}{\Sigma(t_i-\overline{t})^2}\right]$$

(*) Johnston, J., "Econometric Methods", New York, 1963.

(**) Johnston, J., op. cit.

lo que permite asegurar una mayor precisión del pronóstico cuanto más próximo se encuentre el valor t_0 a la media aritmética $\bar{t}$.

El problema que se nos plantea ahora, una vez obtenida una muestra determinada (Y_1, t_1) (Y_2, t_2) (Y_3, t_3) ... (Y_n, t_n) que nos ha permitido calcular los valores correspondientes de a y b para dicha muestra, es hallar un intervalo de confianza del valor concreto Y_0, asociado con el t_0, para lo cual calcularemos la nueva varianza

$$V(Y_0) = E\left[(\alpha+\beta t_0+\varepsilon_0) - (a+b\,t_0)\right]^2 = E\left[\varepsilon_0(t)^2\right] + E\left[(\alpha - a) + (\beta - b)\,t_0\right]$$

pero este segundo sumando es precisamente Var (Y_0), luego queda

$$V(Y_0) = \sigma^2 + \mathrm{Var}\,(Y_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(t_0 - \bar{t})^2}{\Sigma\,(t_i - \bar{t})^2}\right]$$

También puede probarse que el estadístico

$$t_{n-2} = \frac{\alpha + \beta\,t_0 + \varepsilon_0(t) - (a + b\,t_0)}{\dfrac{V(Y_0)}{\sigma}\sqrt{\dfrac{n\,s^2}{n-2}}}$$

tiene una distribución t de Student con $n-2$ grados de libertad, siendo

$$n\,s^2 = \sum_{i=1}^{n}(Y_i - a - b\,t_i)^2$$

lo que permite obtener el siguiente intervalo de confianza, por ejemplo,, del 90 % par Y_0

$$P\left\{(a + b\,t_0) - t_s\sqrt{\frac{n\,s^2}{n-2}\left[\frac{n+1}{n} + \frac{(t_0 - \bar{t})^2}{\Sigma\,(t_i - \bar{t})^2}\right]} < Y_0 < \right.$$

$$\left. < (a + b\,t_0 + t_s\sqrt{\frac{n\,s^2}{n-2}\left[\frac{n+1}{n} + \frac{(t_0 - \bar{t})^2}{\Sigma\,(t_i - \bar{t})^2}\right]}\right\} = 0{,}90 \qquad [6]$$

tal que

$$P\left[-t_s < t_{n-2} < t_s\right] = 0{,}90$$

Haremos uso de este resultado en el apartado V de este trabajo.

IV. Test de linealidad

En el modelo general de la regresión lineal a cada valor t_i de la variable t le corresponde una distribución condicionada de η para dicho t_i de forma que obtenida una muestra de tamaño n podemos escribir: n_i como número de observaciones de t_i.

Y_{i1}, Y_{i2}, Y_{i3}, ..., Y_{in}, son los valores de η correspondientes al valor t_i, y $n=n_1+n_2+...+n_k$ es el supuesto de que en la muestra t tome los valores t_i, t_2, ..., t_k.

Si hallamos las expresiones:

$$S_1^2 = \frac{\sum_{j=1}^{n_i} \sum_{i=1}^{k} (Y_{ij}-\bar{Y}_i)^2}{n-k} \quad \text{y} \quad S_2^2 = \frac{\sum_{i=1}^{k} n_i(\bar{Y}_i - Y_{i}')^2}{k-2}$$

S_1^2 es independiente de la curva de regresión; siempre es un estimador centrado de σ^2 (varianza constante de la distribución condicionada de η para cada valor t_i). En cambio S_2^2 no es independiente de la curva de regresión; sólo será un estimador centrado de σ^2 si es cierta la hipótesis de linealidad. En caso contrario, cuanto más lejos se halle dicha hipótesis, mayor diferencia habrá entre S_2^2 y σ^2, es decir, entre S_2^2 y S_1^2.

Generalmente se hace uso del test de la F de Snedecor,

$$F_{\substack{K-2 \\ n-K}} = \frac{S_2^2}{S_1^2}$$

siendo $K-2$ y $n-K$ los grados de libertad.

Para que sea cierta la hipótesis de linealidad, el valor de F debe estar, en virtud de lo dicho anteriormente, próximo a 1.

Sin embargo en el caso que nos ocupa, como a cada valor de t le corresponde un único valor de η, no es posible obtener S_1^2. En este caso podemos acudir a la variable

$$Z_i = \frac{Y_t - Y_i}{S_2}$$

donde Y_i es el valor muestral observado y Y_t es el correspondiente valor en la recta de regresión $Y_t = a + bt$.

En condiciones generales esta nueva variable se distribuirá aproximadamente normal con parámetros 0,1, luego para que la hipótesis de linealidad sea admisible la media muestral de Z deberá estar muy próxima a acero y su desviación típica ser $\leqslant 1$, siendo igualmente de gran importancia atender a los cambios de signo de Z_i. Así, por ejemplo, en el caso expuesto en la figura 1 el modelo lineal no sería válido (regresión parabólica).

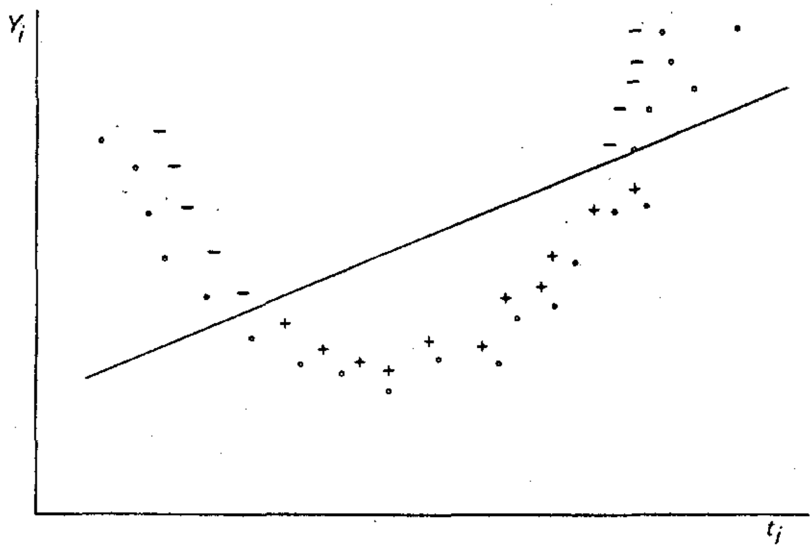

Figura 1

Signo de los valores Z_i.

V. Aplicación al Seguro Obligatorio del Automóvil

En el primer epígrafe de este trabajo habíamos obtenido la recta de regresión lineal:

$$Y_t = 82{,}17 + 4{,}19\, t$$

En base a lo expuesto en los apartados II, II y IV estamos ahora en condiciones de responder a las dos cuestiones que nos planteamos al final del citado primer apartado.

a) *Test de linealidad.*

Para ello elaboramos la siguiente tabla:

Y_i	Y_t	$(Y_t - Y_i)$	$(Y_t - Y_i)^2$	$Z_i = \frac{Y_t - Y_i}{S_2}$	Z_i^2
85,9	86,3	+ 0,4	0,16	+ 0,21	0,044
92,8	90,6	— 2,2	4,84	— 1,16	1,345
93,1	94,8	+ 1,7	2,89	+ 0,89	0,792
97,5	98,9	+ 1,4	1,96	+ 0,75	0,562
105,2	103,1	— 2,1	4,41	— 1,10	1,210
106,8	107,3	+ 0,5	0,25	+ 0,26	0,068

siendo

$$S_2^2 = \frac{\sum_{i=1}^{6} (Y_t - Y_i)^2}{n-2} = \frac{14,51}{4} = 3,627$$

luego

$$S_2 = \sqrt{3,627} \simeq 1,9$$

Los parámetros $\overline{Z}$ y $\sigma^2(z)$ serán

$$\overline{Z} = \frac{\Sigma Z_i}{n} = \frac{-0,15}{6} = -0,025$$

$$\sigma^2(z) = \frac{\Sigma Z_i^2}{n} - (0,025)^2 \simeq 0,67; \qquad \sigma(z) = 0,815$$

Como se puede observar, $\overline{Z} = -0,025$ está muy próxima a cero, $\sigma(Z) = 0,815$ es $\leqslant 1$. Y el signo de los valores Z_i es prácticamente alternativo, por lo que en virtud de lo expuesto anteriormente podemos aceptar en este caso la hipótesis de linealidad, que nos permite hacer uso de la recta mínimo-cuadrática para predecir el valor del déficit técnico de la tarifa del S. O. A. en función de los respectivos valores de t (años).

b) *Determinación del intervalo de confianza de los valores extrapolados.*

Al comienzo de este trabajo habíamos obtenido los valores de Y_t correspondientes a los años $t=7$ (1974) y $t=8$ (1975), dándonos, respectivamente, $Y_7 = 111,6$ e $Y_8 = 115,7$.

Para obtener un intervalo de confianza del 90 % para estos valores aplicaremos la expresión [4], con las siguientes cifras para $t_0=7$.

$$n = 6; \qquad S^2 = 2{,}41; \qquad (t_0 - \bar{t})^2 = 12{,}25; \qquad \Sigma(t_i - \bar{t})^2 = 17{,}50$$

y

$$P[-t_\epsilon < t_4 < t_\epsilon] = 0{,}90$$

se verifica para $t_\epsilon=2{,}132$, luego sustituyendo resulta

$$111{,}6 - 2{,}132 \sqrt{\frac{14{,}5}{4}\left[\frac{7}{6} + \frac{12{,}25}{17{,}50}\right]} = 104{,}08$$

$$111{,}6 + 2{,}132 \sqrt{\frac{14{,}5}{4}\left[\frac{7}{6} + \frac{12{,}25}{17{,}50}\right]} = 119{,}12$$

de donde el intervalo de confianza del 90 % para Y_7 será:

$$P[104{,}08 < Y_7 < 119{,}12] = 0{,}90$$

El mismo intervalo para Y_8 será

$$111{,}7 - 2{,}132 \sqrt{\frac{14{,}5}{4}\left[\frac{7}{6} + \frac{20{,}25}{17{,}50}\right]} = 106{,}36$$

$$111{,}7 + 2{,}132 \sqrt{\frac{14{,}5}{4}\left[\frac{7}{6} + \frac{20{,}25}{17{,}50}\right]} = 125{,}04$$

es decir,

$$P[106{,}36 < Y_8 < 125{,}04] = 0{,}90$$

La longitud de este segundo intervalo es sensiblemente superior a la del primero, lo cual es consecuencia de la expresión [6], que recoge la ecuación de dos ramas parabólicas que se van alejando de la recta de regresión lineal a medida que crece t.

Por tanto la prediccón basada en este método sólo es válida muy a corto plazo. Es decir, para valores de t muy próximos a $\bar{t}$.

En nuestro caso podemos afirmar que tenemos una confianza del 90 % de que el déficit técnico de la tarifa del S. O. A. está comprendido entre el 11,6-7,52 y el 11,6+7,52 % en 1974. Lógicamente una reducción del intervalo de confianza lleva consigo una reducción paralela de la confianza del mismo, por ejemplo, sólo tendremos una esperanza del 30 % de que Y_7 esté comprendido en el intervalo 110,14 y 113,06 % y del 60 % de que esté comprendido en el intervalo 108,28 y 114,92, es decir, un déficit técnico situado entre el 8,2 % y el 14,92 %. En el supuesto de una elevación de la tarifa vigente este déficit quedará reducido en la misma cuantía en que se hubiera elevado dicha tarifa (*).

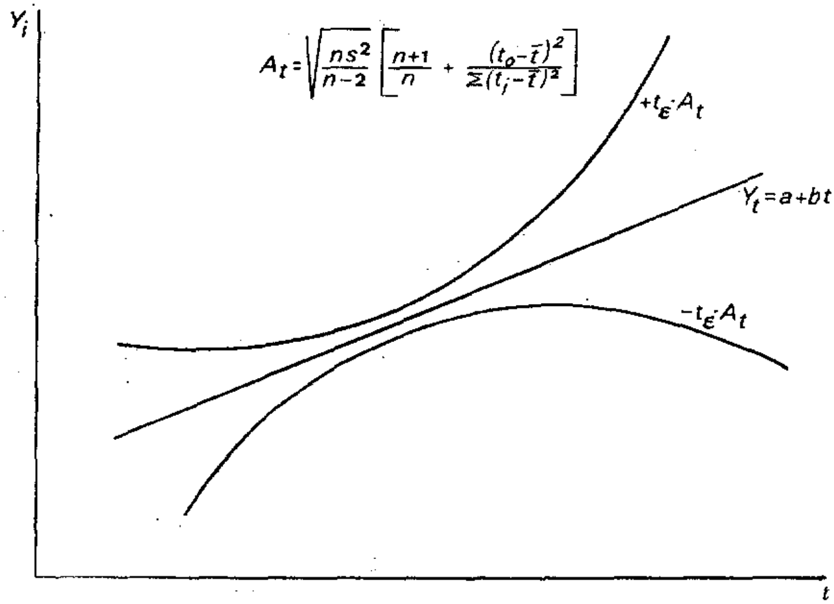

Figura 2

Bandas de confianza.

VI. Conclusión.

El procedimento de los mínimos cuadrados como método de extrapolar valores de la variable Y_t sólo debe aplicarse cuando se desconozca la estructura del proceso estocástico η_t. En el problema que nos ha ocupado este proceso está integrado por dos subprocesos, el subproceso de llegada $Pn(t)$ que nos da la probabilidad que ocurran n siniestros en $(0, t)$ y el subproceso de cambio $V(X)$, función que nos indica la probabilidad de que ocurrido un siniestro su cuantía sea menor o igual a X. Conocidas ambas distribuciones y bajo ciertas condiciones, la distribución del daño total en un período fijo $(0, t)$ será:

$$F(X, t) = P[X(t) \leqslant X] = \sum_0^\infty Pn(t) \cdot V^*(X)$$

sendo $X(t)$ la siniestralidad total pagada en $(0, t)$ y $V^*(X)$ la convolución n-sima de $V(X)$. Lo que nos permitiría predecir $X(t)$ con mucha mayor exactitud que la obtenida por el método de los mínimos cuadrados.

(*) Este artículo está escrito con anterioridad a la elevación de la Tarifa del S.O.A. en un 15 por 100 según dispone la O. M. de 13-XI-74.

BIBLIOGRAFIA

NIETO DE ALBA, UBALDO: "Introducción a la Estadística" (III tomo), Aguilar, 1974.

ARNAIZ, GONZALO: "Estimación por el método de los mínimos cuadrados". Estadística Española, 1960.

JOHNSTON, J.: "Econometric Methods", New York, 1963.

ALCAIDE, ANGEL: "Teoría formal de la regresión lineal". Anales de Economía, 1964.

RÍOS, SIXTO: "Métodos Estadísticos", 1970.