temporales de interés farmacéutico. Aplicación al caso de la escarlatina

X. Tomás, L. G. Sabaté*, J. Cuadros, M. E. Gracia-Aso

Departament d'Estadística Aplicada. Facultat d'Economia IQS. Institut Químic de Sarrià. Universitat Ramon Llull
Via Augusta, 390. 08017 Barcelona (España)

Nonlinear forecasting models in time series analysis of pharmaceutical interest

Models no lineals de previsió per a sèries temporals amb interès farmacèutic

Recibido: 24 de mayo de 2007; aceptado: 27 de noviembre de 2007

RESUMEN

En este trabajo se presentan los resultados obtenidos tras el ajuste de un modelo no lineal y un modelo robusto de predicción a la serie temporal correspondiente a la incidencia de la escarlatina en Catalunya desde el año 2000 hasta mediados del 2005, según los datos publicados en el Butlletí Epidemiològic de Catalunya (BEC) por el Departament de Salut de la Generalitat de Catalunya. El ajuste se ha realizado tanto por mínimos cuadrados como mediante el método de la mínima mediana de residuales. Los intervalos de confianza, dada la no normalidad de las residuales, se han estimado mediante la técnica del bootstrap. El modelo se ha validado comparando las predicciones realizadas para las últimas 47 semanas frente a los casos declarados en el BEC. Salvo episodios de incidencia aguda, las previsiones resultan aceptables. El modelo sinusoidal ajustado muestra una tendencia con un periodo de 52 semanas, una incidencia máxima hacia la segunda quincena de marzo y una mínima incidencia sobre la segunda quincena de septiembre. Los puntos de inflexión se sitúan, aproximadamente, coincidiendo con los solsticios de verano e invierno.

Palabras clave: **Series temporales. Previsión. Modelos no lineales. Métodos robustos de estimación. Bootstrap. Escarlatina.**

SUMMARY

This paper presents the fit of both a nonlinear and a robust model to a time series data about the incidence of scarlet fever in Catalunya from 2000 to 2005. Data were published by the Departament de Salut of the Generalitat de Catalunya in Butlletí Epidemiològic de Catalunya (BEC).
Fitting was performed both by ordinary least squares and least median of residuals method. Since residuals are non-normally distributed, the confidence intervals were evaluated by a bootstrap procedure. The model was validated looking at the forecast for the last 47 weeks and the published values. Except for acute incidence periods, acceptable forecast are obtained.
The sinusoidal model fitted has a period of 52 weeks, maxima are located at the second half of march and minima at the second half of september. Inflexion points are located about summer and winter solstice.

Key words: **Time series. Forecasting. Nonlinear methods. Robust estimation methods. Bootstrap. Scarlet fever.**

* lgsab@iqs.edu

RESUM

Es presenten els resultats obtinguts en l'ajust d'un model no lineal i un model robust de predicció a la sèrie temporal corresponent a la incidència de l'escarlatina a Catalunya des de l'any 2000 fins a mitjans del 2005, segons les dades publicades pel Departament de Salut de la Generalitat de Catalunya al Butlletí Epidemiològic de Catalunya (BEC).
L'ajust s'ha realitzat tant per mínims quadrats ordinaris com pel mètode de la mínima mediana de residuals. Els intervals de confiança, comprovada la no normalitat de les residuals, s'han avaluat mitjançant el procediment del bootstrap. S'ha validat el model comparant les seves prediccions per les darreres 47 setmanes amb els casos declarats al BEC. A banda d'episodis d'incidència aguda, les previsions resulten acceptables.
El model sinusoïdal ajustat mostra una tendència amb un període de 52 setmanes, una incidència màxima cap a la segona quinzena de març, una incidència mínima al voltat de la segona quinzena de setembre. Els punts d'inflexió es situen, aproximadament, cap als solsticis de l'estiu i de l'hivern.

Paraules clau: **Sèries temporals. Previsió. Models no lineals. Mètodes robustos d'estimació. Bootstrap. Escarlatina.**

INTRODUCCIÓN

La incertidumbre sobre lo que acontecerá en el futuro acompaña permanentemente a la actividad económica pública y privada. Este desconocimiento obliga a las empresas y a la Administración, a tomar decisiones en un entorno incierto. En consecuencia la capacidad de efectuar buenos pronósticos es un elemento clave tanto en el desarrollo de la política económica y social de un país, como en la toma de decisiones empresariales (previsión de fabricación, logística, gestión de existencias, etc.) es evidente pues la necesidad y trascendencia de obtener previsiones cuantitativas útiles[1].
Se entiende por método de previsión aquel conjunto de técnicas que ofrecen la valoración cuantitativa de la evolución futura del fenómeno en estudio. El análisis de series temporales, uno de los múltiples métodos de previsión disponibles, se basa en el estudio de los datos históricos de una variable y presupone que la serie se comportará en el futuro de la misma forma que lo hizo en el pasado, salvo en una componente aleatoria suficientemente pequeña y por consiguiente será útil para hacer previsiones a un plazo más o menos corto.

Los métodos de análisis de series temporales para una variable más utilizados siguen una de dos estrategias generales, o bien utilizan la metodología de Box-Jenkins mediante la cual buscan un modelo generador de los datos construido sobre medias móviles y estructuras autorregresivas de los errores, o bien descomponen la serie en movimientos elementales: tendencia, ciclos y estacionalidad[2]; la tendencia acostumbra a ser una función sencilla, lineal en los coeficientes, la descripción del ciclo y la estacionalidad es más complicada y una opción posible es la utilización de funciones periódicas no lineales[3].
Desde el punto de vista estadístico el modelado no lineal tiene por objetivo estimar los coeficientes de la ecuación propuesta que conducen al "mejor" ajuste a los datos experimentales. Generalmente se utiliza como criterio de ajuste la minimización de la suma de cuadrados de los errores residuales (criterio de mínimos cuadrados), porque en determinadas condiciones es el método que proporciona las mejores estimaciones de los coeficientes del modelo. El hecho de que la función a ajustar sea no lineal conlleva que el método de mínimos cuadrados conduzca a un sistema de ecuaciones que solamente se pueda resolver utilizando un algoritmo iterativo de búsqueda de los coeficientes de la ecuación. Todo ello lleva a que al plantearse un ajuste no lineal se debe disponer de diferentes procedimientos para realizarlo según sea el algoritmo de búsqueda utilizado y el criterio de parada seleccionado, requiriendo todos ellos un primer valor de los parámetros para poder iniciar el proceso iterativo de búsqueda[4].
El objetivo de este trabajo es estudiar las posibilidades de aplicación de un modelo no lineal para describir la tendencia y la estructura periódica anual de los datos históricos relativos a la incidencia la escarlatina en Catalunya, una enfermedad de declaración obligatoria, así como su adecuación para realizar previsiones de la incidencia a corto plazo.

MATERIAL Y MÉTODOS

Todos los cálculos se han realizado con Microsoft Office Excel 2003 y Statgraphics 5.1. Se ha partido de los datos recopilados en el Butlletí Epidemiològic de Catalunya (BEC) publicado por el Departament de Salut de la Generalitat de Catalunya[5]. Los datos corresponden al número de casos semanales de escarlatina declarados en Catalunya desde la primera semana de 2000 hasta completar un periodo de 280 semanas (mediados de 2005) tal como se muestra en el gráfico 1; de su observación se constata un comportamiento periódico de manera que hacia la misma época (alrededor de la sema-

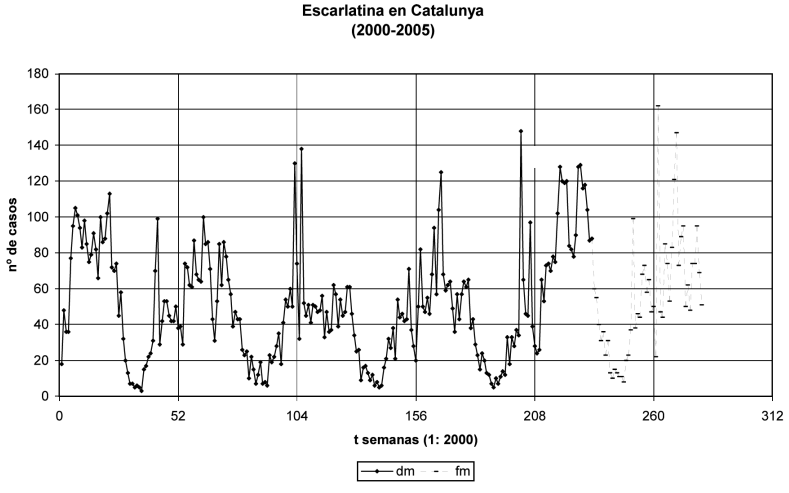

Gráfico 1. Casos de escarlatina declarados en Catalunya utilizados en este estudio.

na 37 de cada año, durante el mes de septiembre) se observa el valor mínimo anual de incidencias. Se observa también que los datos tienen una media de 51,8 casos semanales y no presentan tendencia ya que el ajuste de una tendencia lineal tiene un R^2 = 0,004 y su distribución es asimétrica con algunos valores altos atípicos que corresponden a máximos anuales. Existe pues un comportamiento cíclico para el que una función de tipo sinusoidal podría ser un buen modelo, tal como se cita en estudios anteriores[6 y 7].

Ajuste de un modelo sinusoidal

Para ajustar un modelo a una serie temporal se recomienda dividir los datos en dos partes[1], la primera de aproximadamente el 80% de los datos de más antigüedad, se utiliza para estimar los parámetros del modelo (recibe el nombre de datos "dentro de muestra" o "conjunto de calibración") y el 20% restante para estudiar la capacidad de previsión del modelo (que recibe el nombre de datos "fuera de muestra" o "conjunto de validación" del modelo y aparece en el gráfico 1 en trazo discontinuo). En este trabajo se utilizaron los datos de las 233 primeras semanas para el ajuste y los de las restantes 47 semanas para la previsión.

Se ajustó el modelo sinusoidal descrito por la ecuación (1)

$$y = b_0 + b_1 t + A\,seno\left(\frac{2\pi}{T}(t - F)\right) \qquad (1)$$

en el que las variables son:

y: número de casos semanales

t: semanas, siendo t = 1 la semana 1 de 2000 según el BEC

Y los parámetros a estimar son:

b_0: es la ordenada en el origen para la recta de tendencia

b_1: es pendiente de la recta de tendencia

A: es la amplitud de la onda sinusoidal

T: es el periodo de la onda en semanas

F: es la fase también en semanas

Criterio de la mínima suma de los cuadrados de las residuales

La estimación de los coeficientes se realizó por mínimos cuadrados no lineales con ayuda de la herramienta Solver de Microsoft Office Excel 2003, que utiliza un algoritmo basado en el método gradiente reducido generalizado[8]; se utilizaron los datos "dentro de muestra", para los que se buscó el mínimo de la función "suma de los cuadrados de los errores residuales" a partir de los valores iniciales estimados para los coeficientes en base al gráfico de la serie; en la tabla I se muestran los valores iniciales y finales (modelo 1) hallados así como los valores de la mediana de los cuadrados de los errores dentro de muestra (Med(e^2)), del EAM (error absoluto medio) y RECM (raiz cuadrada del error cuadrático medio), tanto para los datos de ajuste (dm) como los de previsión (fm).

TABLA I

Resultados del ajuste del «modelo 1».

	Valores iniciales	Valores finales
	Gráfico	Modelo 1 (dm)
b_0	52	45,6212874
b_1	0	0,02939759
A	30	30,0554741
T	52	52,1872119
F	36	49,9609938
SCR(dm)	371450,14	130535,98
Med(e^2)	1194,65	188,49
EAM(dm)	34,1	17,7
RECM(dm)	39,9	23,7
EAM(fm)	33,3	17,2
RECM(fm)	19,9	18,7

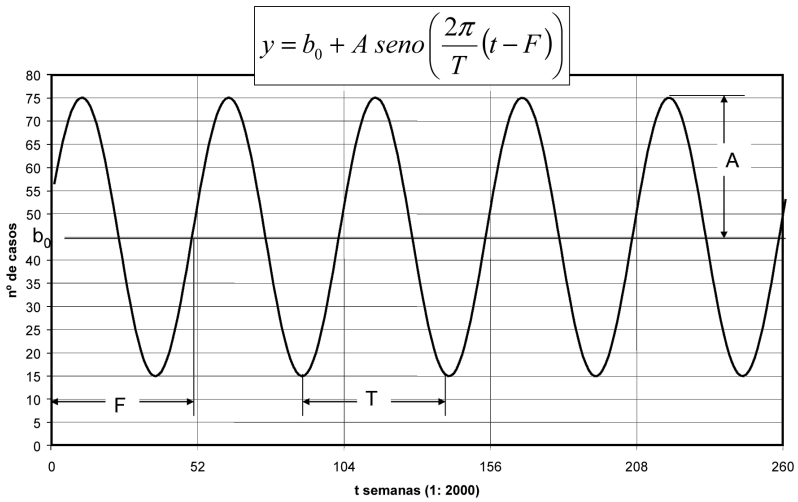

Gráfico 2. Parámetros del modelo sinusoidal.

El modelo ajustado se representa en el gráfico 3. El gráfico 4 muestra el diagrama de caja de las residuales obtenidas en el ajuste, es decir la diferencia entre el número de casos declarados cada semana y el calculado según el modelo para dicha semana. Se observa la presencia de puntos alejados entre los errores del ajuste del modelo 1. Esto es una indicación de la presencia de datos influyentes sobre el ajuste y corresponden a episodios de alta incidencia de la escarlatina.

En el gráfico 5 se muestran los dos modelos ajustados, se observan diferencias entre ellos, el "modelo 2" reproduce mejor los mínimos, se acerca más a los máximos que el "modelo 1" y ambos modelos presentan errores similares tanto en calibración como en predicción; en consecuencia, y tras el análisis expuesto, se adopta el "modelo 2" como el mejor modelo para describir el aspecto general de la serie.

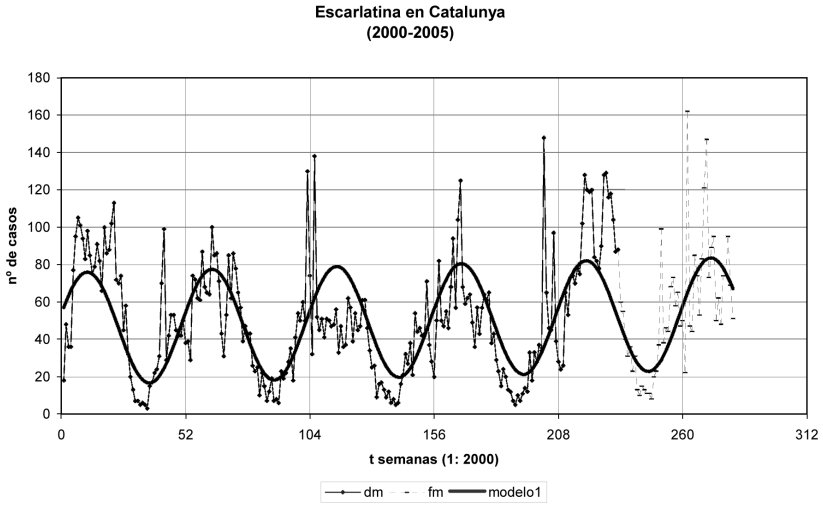

Gráfico 3. Representación del modelo 1 ajustado junto con los datos originales.

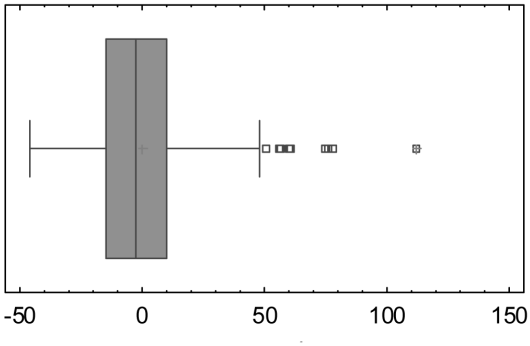

Gráfico 4. Diagrama de caja de las residuales en el que se observa la presencia de puntos alejados que influyen en el ajuste del modelo 1 por mínimos cuadrados.

Criterio de la mínima mediana de los cuadrados de las residuales

Para que la estimación de los coeficientes no se viera influida por la presencia de estos puntos influyentes se decidió hacer un segundo ajuste basado en un criterio robusto a la presencia de puntos alejados y consistente en hacer mínima la mediana de los cuadrados de los errores residuales dentro de muestra (LMS: Least Median of Squares)[9]. Los resultados obtenidos se muestran en la columna "modelo 2" de la tabla II; los cambios más importantes que se observan son una disminución de la pendiente de la tendencia lineal, un aumento en la amplitud de la función sinusoidal y una disminución de la fase. Respecto a la capacidad de previsión, tanto el RECM como el EAM fuera de muestra se pueden considerar iguales en ambos modelos.

TABLA II

Resultados del ajuste del «modelo 2».

	Valores finales	Valores finales
Criterio	**Modelo 1 (dm)**	**Modelo 2 (dm)**
b_0	45,6212874	46,0905281
b_1	0,02939759	0,01168476
A	30,0554741	36,9921686
T	52,1872119	52,186511
F	49,9609938	47,9846553
SCR(dm)	130535,98	144071,68
Med(e^2)	188,49	117,12
EAM(dm)	17,7	17,7
RECM(dm)	23,7	24,8
EAM(fm)	17,2	17,5
RECM(fm)	18,7	18,7

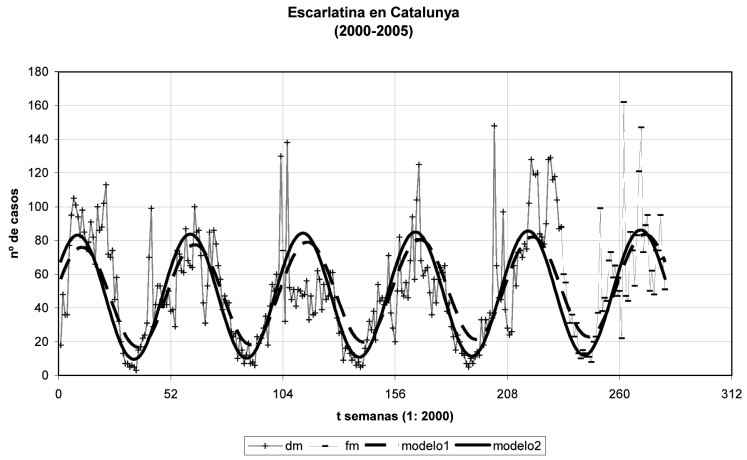

Gráfico 5. Modelos ajustados.

RESULTADOS

Una vez escogido el método de ajuste (la mínima mediana de los cuadrados de las residuales) para la ecuación 1 y comprobado que los valores del RECM y EAM eran parecidos dentro y fuera de muestra, se procedió al ajuste definitivo utilizando la serie entera. Los resultados obtenidos se muestran en la tabla III.

Sustituyendo los coeficientes estimados en la ecuación (1)

TABLA III

Resultados del ajuste del modelo 2 con todos los datos junto con los intervalos de confianza al 95% obtenidos por bootstrap.

	valor estimado	Intervalo de confianza al 95%	
		LI	LS
b_0	46,1256	44,2002	48,2802
b_1	0,0369	0,0111	0,0631
A	38,9922	37,6521	40,4267
T	52,1539	51,9172	52,4420
F	48,0344	47,4459	48,6693
MAE	18,1680	15,1071	21,4405
RECM	25,1942	21,2630	28,4277

resulta

$$y = 46{,}12557 + 0{,}03686t + 38{,}99\ seno\left(\frac{2\pi}{52{,}1539}(t - 48{,}0344)\right) \quad (2)$$

La tendencia lineal prácticamente no presenta pendiente (b_1 = 0,03686) ya que se incrementa dos casos por año aproximadamente; la amplitud de la onda sinusoidal es de 39 casos (A = 38,99), se reproduce bastante bien el valor y la posición de los mínimos anuales pero no tanto la posición ni el valor de los máximos ya que el valor pronosticado puede ser casi la mitad de máximo local. Estos valores de los máximos extremadamente altos en comparación con la tendencia de la serie podrían interpretarse como el reflejo de episodios excepcionales de la incidencia de la escarlatina.

El periodo estimado (T = 52,15) corresponde aproximadamente al número de semanas por año (considerando la existencia de años bisiestos y las fracciones de semanas en el cambio de año) y la fase (F = 48,03) corresponde a principios del mes de diciembre, momento en que la curva cruza en ascenso a la línea de tendencia.

Como se aprecia en el gráfico 5 hay discrepancias en el comportamiento correspondiente al año 2002 y en cierta medida al máximo del año 2004 que parecen indicar un adelanto de algún episodio agudo, mientras que en 2005 se observan oscilaciones mayores en el tramo de subida. En términos generales el modelo muestra que la tendencia sinusoidal se mantiene estable a lo largo de los años estudiados.

Análisis de las residuales

Analizados los errores del ajuste con Statgraphics 5.1, se observa que no se distribuyen de acuerdo a una ley normal (Gráfico 6a) y que no son completamente independientes ya que los 4 primeros coeficientes de autocorrelación son estadísticamente significativos al nivel $p \leq 0.05$ (Gráfico 6b).

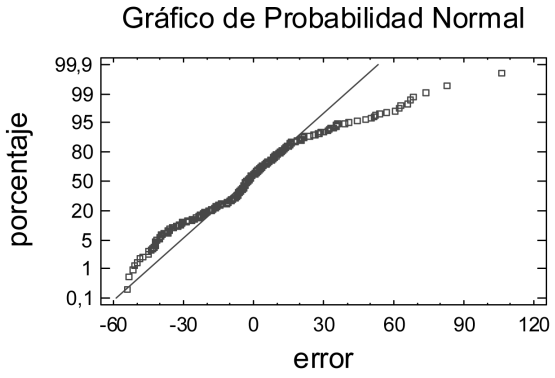

Gráfico 6a. Falta de normalidad en los errores.

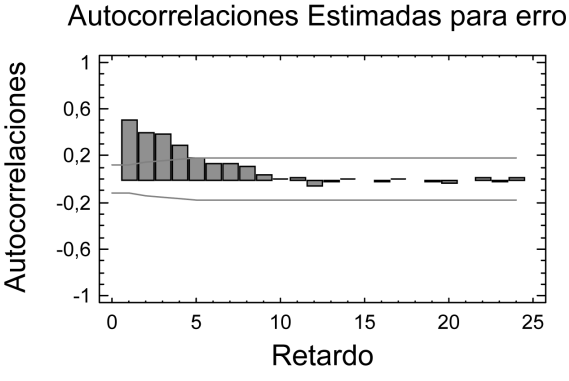

Gráfico 6b. Presencia de autocorrelación en los errores.

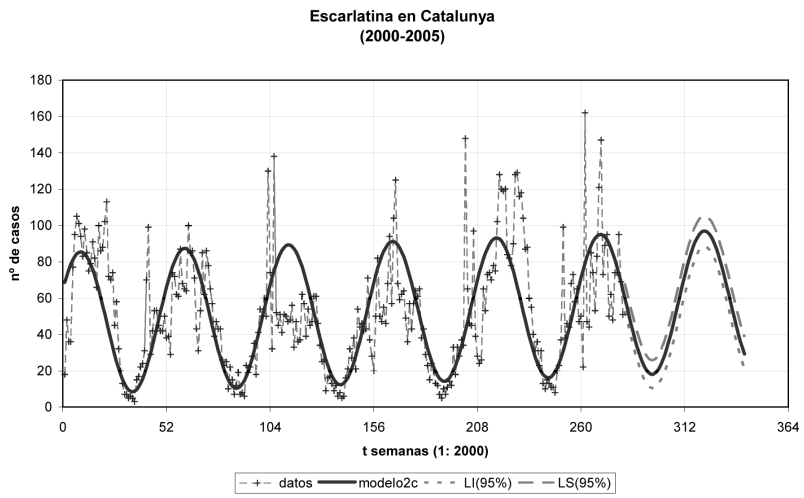

Gráfico 7. Modelo completo ajustado con los intervalos de confianza obtenidos para las previsiones.

Intervalos de confianza de las estimaciones

Dada la estructura de las residuales obtenidas no se pueden establecer fácilmente los intervalos de confianza de las estimaciones realizadas de forma que permitan realizar predicciones con indicación de su margen de error. En estas circunstancias se puede utilizar el método denominado "bootstrap"[10-13]. El procedimiento consiste en tomar al azar por muestreo aleatorio simple (con reposición) bloques de 4 errores consecutivos para reproducir la estructura de autocorrelación hasta completar un vector de 280 valores que se añaden al valor esperado según la ecuación (2) generando así una nueva serie a la que se le ajusta el mismo modelo según el criterio de la mínima mediana partiendo de los coeficientes conocidos. El procedimiento anterior se repite un número suficientemente grande de veces; en este trabajo se han generado 1000 muestras. Como consecuencia se obtienen estimaciones de las distribuciones muestrales de cada uno de los coeficientes y de las previsiones, a partir de las cuales se calculan los cuantiles que corresponde al 2,5% (Límite inferior LI) y al 97,5% (Límite superior LS) de los valores obtenidos mediante el proceso de remuestreo; los valores de estos cuantiles fijan los límites del intervalo de confianza al 95% de los correspondientes coeficientes (tabla III) y previsiones (gráfico 7).

CONCLUSIONES

Se ha estudiado la tendencia de la evolución de la escarlatina en Catalunya con los datos semanales publicados en el Butlletí Epidemiològic de Catalunya (BEC) desde el año 2000 hasta mediado el año 2005. Dada la apariencia de la serie temporal se ha ajustado una función sinusoidal. Se ha aplicado una metodología para el ajuste de una función no lineal en presencia de puntos influyentes basado en el método mínima mediana de los cuadrados de los errores residuales junto con estimación bootstrap de los intervalos de confianza en presencia de errores con autocorrelación.

El modelo describe la incidencia de la enfermedad a lo largo del año de manera aceptable, con excepción de los máximos que se comportan como puntos atípicos para la tendencia.

El análisis del modelo ajustado indica que la incidencia de la escarlatina presenta una periodicidad anual (T = 52,15 semanas), una máxima incidencia sobre la segunda quincena del mes de marzo y una incidencia mínima alrededor de la segunda quincena de septiembre. Los puntos de inflexión se sitúan sobre la segunda quincena de junio y la segunda quincena de diciembre de cada año.

BIBLIOGRAFIA

[1]. Valls Corral, L. (2003): «Diseño, programación y validación de una herramienta para la previsión cuantitativa a corto plazo en la empresa», Tesis Doctoral, Facultat d'Economia IQS, Institut Químic de Sarrià, Universitat Ramon Llull, Barcelona.

[2]. Holton, J. y Keating B. (1996): «Business Forecasting», Richard D. Irwin, Inc. New York.

[3]. Rizzo, C. *et al.* (2007): «Trends for influenza-related Deaths during Pandemic and Epidemia Seasons, Italy, 1969-2001»; *Emerging Infectious Diseases* **13(5)**: 694-699.

[4]. Tomás, X. (1996): «Modelització quimiomètrica: Tècniques d'optimització» en «Modelització macroscòpica en ciències experimentals». E. Casassas, M. Esteban Ed. Institut d'Estudis Catalans. Barcelona. ISBN 84-7283-339-9.

[5]. Generalitat De Catalunya. Departament De Salut (2000-2005). Butlletí Epidemiològic de Catalunya (BEC). http://www.gencat.net/salut/depsan/units/sanitat/html/ca/publicacions/spbec.htm (última consulta 9 de febrero de 2007).

[6]. Casaní Martínez C. *et al.* (2001): «Estudio epidemiológico de un brote de escarlatina»; *Revista Pediatría de Atención Primaria,* **3(9)**, 41-49.

[7]. Giner, E. *et al.* (2004): «Brotes de escarlatina en población escolar». *Boletín Epidemiológico Semanal,* **12(4)**, 37-44.

[8]. Kadson, L.S.; Waren, A.; Jain, A. Ratner, M. (1978): «Design and Testing of a Generalized Reduced Gradient Code for Nonlinear Programming», *ACM Transactions on Mathematical Software*, **4(1)**, 34-50.

[9]. Rousseeuw, P.J. y Leroy, A.M. (1987): «Robust Regression and Outlier Detection», John Wiley & Sons, New York.

[10]. Efron, B. y Tibshirani, R.J. (1993): «An introduction to the Bootstrap», Chapman & Hall / CRC.

[11]. Carpenter, J. y Bithell, J. (2000): «Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians», *Statistics in Medicine,* **19**, 1141-1164.

[12]. Grigoletto, M. (1998): «Bootstrap prediction intervals for autoregressions: some alternatives»; *International Journal of Forecasting,* **14**, 446-456.

[13]. Arnholt, A.T. (2007): «Resampling with R». *Teaching Statistics* **29(1)**: 21-26.