Modelos de regresión parcialmente lineales:

Aplicación al estudio de la influencia de la vacuna neumocócica conjugada sobre los ingresos hospitalarios por neumonía

Autora: Gael Naveira Barbeito
Director: Germán Aneiros Pérez

Resumen

Los modelos de regresión parcialmente lineales fueron propuestos por Engle et al. en el año 1986, y estudiados extensamente en la literatura. Estos modelos permiten describir una variable respuesta como la suma de una componente lineal y una componente no paramétrica. Se caracterizan por su flexibilidad, por su fácil interpretación y por el hecho de que eluden el problema de la maldición de la dimensionalidad, lo que los convierte, en muchas situaciones reales, en modelos más adecuados que, por ejemplo, los modelos de regresión lineal.

En este trabajo nos proponemos revisar brevemente algunos de los artículos publicados en la literatura que tratan estos modelos en detalle, centrándonos en aquellos que suponen cierta estructura de dependencia para los errores. También nos proponemos comparar, a través de un estudio de simulación, dos aproximaciones (normal y bootstrap) con el objetivo de realizar contrastes sobre la componente lineal del modelo.

Por último se evaluará el impacto de la introducción de la vacuna neumocócica conjugada heptavalente (octubre del año 2001) sobre los ingresos por neumonía neumocócica en niños menores de dos años, para quienes está indicada la vacuna. Esta vacuna confiere inmunidad contra algunos de los serotipos más frecuentes de neumococo. Dicho microorganismo patógeno causa diversas infecciones y procesos invasivos severos, que afectan en mayor medida a los niños menores de 2 años y adultos mayores de 65, siendo ambos quienes portan la carga principal de la enfermedad. Una de las enfermedades producida por el neumococo es la neumonía neumocócica, considerada como una de las más prevalentes y serias, tanto en países desarrollados como en vías de desarrollo. El estudio de esta serie se llevará a cabo por medio de la aplicación de un modelo parcialmente lineal; así como de un modelo basado en procesos Box Jenkins y de un modelo de regresión segmentada. Este último se basa en la regresión lineal y permite ajustar dos rectas de regresión que dan idea de la tendencia; en nuestro caso, una para el período previo a la introducción de la vacuna y otra para el período posterior. Las principales ventajas del modelo de regresión segmentada son la sencillez y la fácil interpretación de los resultados; sin embargo su principal inconveniente es su falta de flexibilidad.

Finalmente se contrastará, para cada uno de los modelos, la hipótesis nula de que la intervención (introducción de la vacuna) no ha influido en la serie frente a la alternativa de que ha provocado un descenso en los ingresos.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

El Streptococcus pneumoniae (neumococo) es un microorganismo patógeno capaz de causar diversas infecciones y procesos invasivos severos, que reciben el nombre de enfermedad neumocócica, y afectan en mayor medida a niños menores de 2 años y adultos mayores de 65, quienes portan la carga principal de la enfermedad. En el caso de los niños menores de 2 años da lugar a enfermedades habituales en la niñez, como infecciones de oído; pero también puede dar lugar a enfermedades de mayor gravedad como meningitis, septicemia y neumonía neumocócica. En el caso de los adultos mayores de 65 años es causa frecuente de enfermedades pulmonares, como la bronquitis y la neumonía. La neumonía neumocócica es considerada como una de las enfermedades más prevalentes y serias tanto en los países desarrollados como en vías de desarrollo. [1, 2]

En junio del año 2001 se autorizó en España la vacuna neumocócica conjugada heptavalente (VNC-7v), indicada para niños menores de 2 años. En octubre de ese mismo año fue introducida en Galicia, sin ser incluida en el calendario de vacunación infantil. Esta vacuna confiere inmunidad contra algunos de los serotipos más frecuentes de neumococo. Como pauta de administración, se recomiendan tres dosis el primer año de vida (a los 2, 4 y 6 meses de edad) y una cuarta dosis de recordatorio en el segundo año (entre los 12 y 24 meses de edad). [2]

Desde su comercialización la distribución de esta vacuna aumentó de manera continuada en Galicia, siendo 10.045 el número de dosis distribuidas en farmacias en el año 2002, y 61.769 las distribuidas en el 2009. Esto supuso pasar de 0,5 dosis/niño a 2,7 dosis/niño.

En el año 2007, Grijalva et al. [3] publicaron un artículo en el que evaluaron el impacto del programa de vacunación de la VNC-7v (año de introducción de la vacuna 2000) en los Estados Unidos, no sólo en el caso de los ingresos por neumonía neumocócica sino también en el caso de las neumonías totales. Para ello realizaron un análisis de regresión lineal segmentada en el que se incluyeron términos para la

intervención y la tendencia secular, para los períodos de antes y después de la implementación del programa de vacunación. De esta manera se ajustan dos rectas de regresión que dan idea de la tendencia, una para cada período (pre y post intervención). Los resultados obtenidos para ambas series fueron comparados con una serie control (ingresos por deshidratación).

El análisis de regresión segmentada es una técnica usualmente utilizada en el estudio de series temporales interrumpidas, debido a su sencillez a la hora de estimar el efecto de una intervención sobre una serie. El modelo de regresión lineal segmentada (MLS) más común se podría describir de la siguiente manera [4]:

$$Y_i = \beta_0 + \beta_{\text{tiempo}}\text{tiempo}_i + \beta_{\text{interv}}\text{interv}_i + \beta_{\text{postinterv}}\text{postinterv}_i + \epsilon_i, \quad i = 1, ..., n \quad (1.1)$$

donde Y_i es la variable respuesta, la variable *tiempo* indica el número de meses transcurridos a instante i desde el comienzo del estudio, la variable *interv* indica si se ha producido ya la intervención (intervención=1) o no (intervención=0), la variable *postinterv* indica el número de meses transcurridos después de la intervención (codificada como 0 antes de la intervención) y el término de error (ϵ_i) representa la variabilidad aleatoria no explicada por el modelo. La constante β_0 estima el nivel basal de la serie a instante 0; la constante β_{tiempo} estima la tendencia basal (es decir, el cambio en el valor medio de la variable dependiente, que ocurre entre los meses anteriores a la intervención); β_{interv} estima el cambio de nivel de la serie justo en el momento de la intervención y por último, $\beta_{\text{postinterv}}$ estima el cambio en la tendencia en el período post-intervención, comparado con el período pre-intervención. La suma de β_{tiempo} y $\beta_{\text{postinterv}}$ se corresponde con la estimación de la pendiente después de la intervención. En el ajuste de este tipo de modelos se suele emplear el método de mínimos cuadrados ordinarios.

Como ya comentamos, los MLS se caracterizan por su sencillez y la fácil interpretación de los resultados. Pero el mayor inconveniente de estos modelos es la falta de flexibilidad, ya que se exige linealidad en cada segmento ajustado. Para superar este escollo nos planteamos el empleo de modelos de regresión parcialmente lineales. Estos modelos fueron propuestos en 1986 por Engle et al. [5], y permiten describir la variable respuesta como la suma de una componente lineal y una componente no paramétrica; esta última caracterizada por una función suave, $m(\cdot)$. En la actualidad existe un gran número de trabajos que estudian este tipo de modelos y sus propiedades, aplicando los resultados obtenidos a datos reales o realizando estudios de simulación.
Estos modelos suelen expresarse como sigue:

$$Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p + m(t_i) + \epsilon_i, \quad i = 1, ..., n$$

o bien, de la siguiente manera:

$$Y_i = X_i^T \beta + m(t_i) + \epsilon_i, \quad i = 1, ..., n \tag{1.2}$$

donde $(X_i^T, t_i) = ((x_{i1}, ..., x_{ip}), t_i)$ son puntos del diseño con t_i perteneciente a un dominio acotado, $\beta = (\beta_1, ..., \beta_p)$ es un vector de parámetros desconocido, $m(\cdot)$ es una función desconocida y, en último lugar, ϵ_i son los términos de error aleatorio.

Otra manera de abordar el problema que nos ocupa en este trabajo es mediante el empleo de modelos de intervención, basados en modelos Box Jenkins (MBJ) [6]. Este tipo de análisis estudia aquellas series temporales caracterizadas por un suceso (intervención) que afecta al transcurso de la serie de forma puntual o durante un período de tiempo. Para modelizar estas series se incluyen las variables ficticias, escalón o impulso ($S_i^{(h)}$ o $I_i^{(h)}$, siendo h el instante en el que se produce la intervención), en función de si el suceso dio lugar a un efecto permanente o transitorio, respectivamente. La variable escalón se define como cero antes de la intervención ($i < h$) y como uno a partir de ella ($i \geq h$). Mientras que la variable impulso se define como uno durante la intervención ($i = h$) y cero en otro caso ($i \neq h$).

Este modelo de intervención puede expresarse de la siguiente manera:

$$Y_i = w(B)S_i^{(h)} + X_i \quad o \quad Y_i = w(B)I_i^{(h)} + X_i$$

donde la función $w(B)$ se denomina función de transferencia y describe el efecto que la variable escalón o impulso ejerce sobre la variable respuesta.

En el proceso de construcción y estimación del modelo se deben proponer una función de transferencia $w(B)$ y un modelo ARIMA para la serie $\{X_i\}$, ya que en este tipo de modelos se parte del supuesto de que si no hubiera habido intervención la serie podría haber sido modelizada a través de un proceso ARIMA.

En este trabajo nos proponemos revisar los modelos de regresión parcialmente lineales, bajo el supuesto de que los errores mantengan una cierta estructura de dependencia. También compararemos, mediante un estudio de simulación, las aproximaciones normal y bootstrap en la realización del contraste de hipótesis sobre la componente paramétrica del modelo. Y, por último, se aplicarán los tres modelos planteados en el estudio de la serie de ingresos por neumonía neumocócica.

Capítulo 2

Modelo de regresión parcialmente lineal

Como ya se comentó en la introducción, los modelos de regresión parcialmente lineales (MPL) se caracterizan por su flexibilidad, además de por su fácil interpretación y por el hecho de que eluden el problema de la maldición de la dimensionalidad. Estas ventajas motivan su estudio y lo convierten en un modelo más adecuado que el modelo de regresión lineal y, en muchas ocasiones, que el MLS. En este trabajo estudiaremos el MPL (1.2) teniendo en cuenta errores con cierta estructura de dependencia.

2.1. Estimación

En la literatura existen diferentes propuestas a la hora de estimar β y $m(\cdot)$, ampliamente discutidas y estudiadas. Uno de los métodos más empleados se basa en la combinación de la estimación por mínimos cuadrados ordinarios y la estimación tipo núcleo (que denotaremos por MCO-TN) [7, 8, 9, 10]. Veamos como se estiman las componentes del modelo (1.2) por medio de este método:

En primer lugar se obtiene el estimador no paramétrico de la función $m(\cdot)$, bajo el supuesto de que β sea conocido. Esto sería

$$\hat{m}_{\beta,h}(t) = \sum_{j=1}^{n} w_{n,h}(t, t_j)(Y_j - X_j^T \beta) \tag{2.1}$$

donde $w_{n,h}(\cdot,\cdot)$ es una función de pesos, caracterizada por una función kernel $K(\cdot)$ y un parámetro de suavizado h>0.

A continuación, se estima β por mínimos cuadrados ordinarios, a partir del modelo $Y_i = X_i^T\beta + \hat{m}_{\beta,h}(t_i) + \epsilon_i$, de la siguiente manera,

$$\hat{\beta}_h = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y} \tag{2.2}$$

donde $\tilde{X} = (\tilde{X}_1, ..., \tilde{X}_n)$ e $\tilde{Y} = (\tilde{Y}_1, ..., \tilde{Y}_n)$, con $\tilde{X}_i = X_i - \sum_{j=1}^n w_{n,h}(t_i, t_j)X_j$ e $\tilde{Y}_i = Y_i - \sum_{j=1}^n w_{n,h}(t_i, t_j)Y_j$.

Finalmente, a partir de las expresiones (2.1) y (2.2), se obtiene el estimador de $m(\cdot)$

$$\hat{m}_h(t) = \sum_{j=1}^{n} w_{n,h}(t, t_j)(Y_j - X_j^T\hat{\beta}_h) \tag{2.3}$$

Speckman [7], en el año 1988, estudia las propiedades asintóticas de estos estimadores comparándolas con las de los estimadores del método basado en la estimación spline y mínimos cuadrados penalizados (denominados estimadores Green-Jennison-Seheult (GJS)). El autor llegó a la conclusión de que (asintóticamente) el MCO-TN es mejor a la hora de estimar β en términos de sesgo y que en términos de varianza no se aprecian diferencias. Otra cuestión importante, que hace notar Speckman en su artículo, es el hecho de que la estimación de β por mínimos cuadrados penalizados podría estar seriamente sesgada si se emplea validación cruzada para elegir el parámetro de suavizado, algo que no ocurre con MCO-TN.

Detallamos a continuación como se definen los estimadores GJS:

Estimador tipo spline:

$$\hat{m}_{GJS} = \tilde{K}(Y - X\hat{\beta}_{GJS})$$

donde $X = (X_1, ..., X_n)$, $Y = (Y_1, ..., Y_n)$ y $\tilde{K}$ es la matriz de pesos.

Estimador mínimos cuadráticos penalizados:

$$\hat{\beta}_{GJS} = (X^T(I - \tilde{K})X)^{-1}X^T(I - \tilde{K})Y$$

donde I es la matriz identidad.

En el artículo de Speckman, así como en el de Liang et al. (2000) [8], se asumen errores independientes e idénticamente distribuidos para el modelo (1.2). En muchas situaciones esta asunción no es plausible y se debe suponer que los errores presentan cierta estructura de dependencia. En esta línea nos encontramos con los trabajos de Aneiros y Quintela (2001a, 2001b) [11, 12], González y Aneiros (2003) [10] y Gao (1995) [9], entre otros.

En concreto, en los trabajos de Aneiros y Quintela, se asumen errores α-mixing y se propone el método de estimación basado en una combinación del estimador mínimo

cuadrático generalizado (MCG) y la estimación tipo núcleo (2.3), para estimar β y m en el modelo (1.2). La expresión del estimador mínimo cuadrático generalizado es

$$\hat{\beta}_{MCG} = (\tilde{X}^T \Psi^{-1} \tilde{X})^{-1} \tilde{X}^T \Psi^{-1} \tilde{Y}$$

donde Ψ es la matriz de correlaciones (generalmente desconocida y que debe ser estimada).

Después de este breve repaso por algunos de los métodos empleados para estimar las componentes del modelo 1.2, en este trabajo nos decantamos por el MCO-TN.

2.2. Propiedades asintóticas

Una vez centrados en el método de estimación MCO-TN podemos decir que el estimador mínimo cuadrático de β es asintóticamente normal [9]. Aneiros y Quintela llegan a la misma conclusión cuando se trata del estimador basado en mínimos cuadrados generalizados [11].

De modo que se tiene que

$$n^{\frac{1}{2}}(\hat{\beta}_h - \beta) \approx N(0, \sigma_\epsilon^2 (n^{-1} \tilde{X}^T \tilde{X})^{-1}) \tag{2.4}$$

Otra forma de llegar a la distribución asintótica del estimador de β es por medio de la técnica bootstrap. Esta técnica es planteada en el artículo de Liang et al. (2000) [8], en donde se llega a la conclusión de que la aproximación bootstrap funciona igual de bien que la aproximación normal e incluso da lugar a una mejor aproximación cuando se trata de los primeros cuatro momentos de los estimadores bootstrap.

2.3. Selección del parámetro de suavizado

Un aspecto clave en la estimación tipo núcleo es la elección del parámetro de suavizado h. La importancia de esta elección radica en el hecho de que si el parámetro de suavizado es pequeño, el estimador tiende a infrasuavizar, es decir, a

interpolar los datos; y, por lo contrario, si el parámetro es grande, el estimador tiende a sobresuavizar, es decir, tiende a una función constante. Por este motivo es necesario prestarle especial atención a dicha elección, y para ello existen diferentes métodos que permiten seleccionar el parámetro de ventana óptimo, por medio de la minimización de algún criterio de error.

En el artículo de Aneiros y Quintela [12] se propone aplicar el método de validación cruzada modificado en la estimación de β y m, tanto para el método MCO-TN como para el método basado en mínimos cuadrados generalizados. Este método de validación cruzada consiste en minimizar la siguiente expresión

$$CV_{l_n}(h) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - X_j^T\hat{\beta}_h - \hat{m}_{h,j,l_n}(t_j))^2 \tag{2.5}$$

donde $\hat{m}_{h,j,l_n}(\cdot)$ se corresponde con la estimación de $m(\cdot)$, una vez se hayan eliminado de la muestra aquellos datos Y_i que están más cercanos en el tiempo a Y_j (es decir, aquellos Y_i tal que $|j-i| \leq l_n$), y por tanto altamente correlacionados con Y_j.

En el trabajo de Ghement et al. (2007) [13] se trata el tema de la elección del parámetro de ventana por dos vías diferentes: el método plug-in y el método empírico global. Ambos métodos se basan en la minimización del error cuadrático medio.

2.4. Contraste basado en la aproximación normal

Una vez establecidos los estimadores del modelo (1.2), podemos estar interesados en realizar contrastes sobre las componentes del modelo. En esta línea nos encontramos con el artículo de González-Manteiga y Aneiros-Pérez (2003) [10], donde se plantean contrastar: la hipótesis paramétrica $H_{0\beta} : \beta = \beta_0$, la hipótesis no paramétrica $H_{0m} : m = m_0$ y la hipótesis de linealidad $H_{0m}^l : m \in U_l$ donde $U_l = gen\{f_1, ...; f_l\}$, siendo f_j (j=1,...l) funciones linealmente independientes. Los autores se basan en el método MCO-TN.

Los estadísticos que ellos proponen están basados en la distancia entre los estimadores de la función de regresión y los estimadores bajo la hipótesis nula, y presentan las siguientes expresiones:

$$d_\beta^2(\hat{r}_n, H_{0\beta}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{r}_n(X_i^T, t_i) - \hat{r}_{0,n}(X_i^T, t_i))^2 = (\hat{\beta}_b - \beta_0)^T(n^{-1}\tilde{X}^T\tilde{X})(\hat{\beta}_b - \beta_0) \tag{2.6}$$

$$d_m^2(\hat{r}_n^*, H_{0m}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{r}_n^*(X_i^T, t_i) - \hat{r}_{0,n}^*(X_i^T, t_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{m}_h(t_i, \hat{\beta}_b) - m_0(t_i))^2 \quad (2.7)$$

$$d_{m,l}^2(\hat{r}_n^*, H_{0m}^l) = min_{\theta\in\Theta}\int_0^1 (\hat{m}_h(t, \hat{\beta}_b) - F^T(t)\theta)^2 d\Omega_n(t) = \frac{1}{n}\sum_{i=1}^{n}(\hat{m}_h(t_i, \hat{\beta}_b) - F^T(t_i)\hat{\theta}_n)^2 \quad (2.8)$$

donde $\hat{r}_n(X_i^T, t_i) = X_i^T\hat{\beta}_b + \hat{m}_b(t_i, \hat{\beta}_b)$ y $\hat{r}_n^*(X_i^T, t_i) = X_i^T\hat{\beta}_b + \hat{m}_h(t_i, \hat{\beta}_b)$ son estimadores consistentes para la función de regresión $X_i^T\beta + m(t_i)$. Bajo las hipótesis nulas $H_{0\beta}$ y H_{0m}, otros estimadores consistentes para la función de regresión son $\hat{r}_{0,n}(X_i^T, t_i) = X_i^T\beta_0 + \hat{m}_b(t_i, \beta_0)$ y $\hat{r}_{0,n}^*(X_i^T, t_i) = X_i^T\beta_b + \hat{m}_0(t_i)$, respectivamente. En último lugar, $F(t) = (f_1(t), ..., f_l(t))^T$, $\theta = (\theta_1, ..., \theta_l)^T$ y Ω_n es la función de distribución empírica de los puntos del diseño $\{t_i\}_{i=1}^n$. Los subíndices b y h se corresponden con los parámetros de suavizado empleados en cada caso.

Los autores trabajan con dos parámetros de suavizado (b y h) por el hecho de que un parámetro común no permite verificar ciertas suposiciones necesarias, entre otras, para la obtención de resultados asintóticos para la distancia $d_m^2(\hat{r}_n^*, H_{0m})$.

Al estudiar las distribuciones asintóticas de los estadísticos para las distintas hipótesis, obtuvieron los siguientes resultados.

En el caso del contraste paramétrico:

- Bajo la hipótesis $H_{0\beta} : \beta = \beta_0$ se tiene que

$$F(b) \equiv \frac{nd_\beta^2(\hat{r}_n, H_{0\beta})}{\hat{\sigma}_\epsilon^2} = \frac{(\hat{\beta}_b - \beta_0)^T(\tilde{X}_b^T\tilde{X}_b)(\hat{\beta}_b - \beta_0)}{\hat{\sigma}_\epsilon^2} \xrightarrow{d} \chi_p^2 \quad (2.9)$$

  donde χ_p^2 denota la distribución ji-cuadrado con p grados de libertad y $\hat{\sigma}_\epsilon^2$ es un estimador consistente de σ_ϵ^2.

- Bajo la hipótesis alternativa local $H_{1\beta}^c : \beta = \beta_0 + cn^{-1/2}$ se tiene que

$$F(b) \xrightarrow{d} \chi_p^2(\theta) \quad (2.10)$$

  donde $\chi_p^2(\theta)$ denota la distribución ji-cuadrado descentrada con p grados de libertad, θ parámetro de centralización y $c \neq 0$ $(p \times 1)$ es un vector fijo arbitrario.

- Bajo la alternativa general $H_{1\beta} : \beta \neq \beta_0$ se tiene que $F(b) \to \infty$ cuando $n \to \infty$.

En el caso del contraste no paramétrico:

- Bajo la hipótesis $H_{0m} : m = m_0$ se tiene que

$$\sqrt{n^2h}\Big(d_m^2(\hat{r}_n^*, H_{0m}) - \frac{\sum_{s=-\infty}^{\infty}\gamma(s)\int K^2}{nh}\Big) \xrightarrow{d} N(0, \sigma_d^2) \tag{2.11}$$

  donde $\gamma(k) = E(\epsilon_1\epsilon_{1+k})$, K función kernel, $\sigma_d^2 = 2(\sum_{k=-\infty}^{\infty}\gamma(k))^2\int(K*K)^2$ y $K*K$ denota la convolución de K consigo mismo.

- Bajo la hipótesis alternativa $H_{1m} : m(t_i) = m_0(t_i) + (n^2h)^{-1/4}m^*(t_i)$ (i=1,...,n) y suponiendo que m_0 y m^* tienen $\nu \geq 2$ derivadas continuas en [0,1], se tiene que

$$\sqrt{n^2h}\Big(d_m^2(\hat{r}_n^*, H_{0m}) - \frac{\sum_{s=-\infty}^{\infty}\gamma(s)\int K^2}{nh}\Big) \xrightarrow{d} N(\int m^*(u)^2du, \sigma_d^2) \tag{2.12}$$

Y, por último, en el caso del contraste de linealidad:

- Bajo la hipótesis $H_{0m}^l : m \in U_l$ se tiene que

$$\sqrt{n^2h}\Big(d_{m,l}^2(\hat{r}_n^*, H_{0m}^l) - \frac{\sum_{s=-\infty}^{\infty}\gamma(s)\int K^2}{nh}\Big) \xrightarrow{d} N(0, \sigma_d^2) \tag{2.13}$$

- Bajo la hipótesis alternativa $H_{1m}^l : m(t_i) = F^T(t_i)\theta_0 + (n^2h)^{-1/4}m^*(t_i)$ (i=1,...,n), para algún $\theta_0 \in \Theta$, se tiene que

$$\sqrt{n^2h}\Big(d_{m,l}^2(\hat{r}_n^*, H_{0m}^l) - \frac{\sum_{s=-\infty}^{\infty}\gamma(s)\int K^2}{nh}\Big) \xrightarrow{d} N(\int m^*(u)^2du, \sigma_d^2) \tag{2.14}$$

  donde $F = (f_1, ..., f_l)^T$ con $f_j : [0,1] \to \Re$ $(j = 1, ..., l))$ funciones linealmente independientes y la función m^* tiene $\nu \geq 2$ derivadas continuas en [0,1], m^* es ortogonal a F.

2.5. Intervalos de confianza basados en bootstrap

Existen otras técnicas que permiten aproximar la distribución del estadístico del contraste, como por ejemplo la técnica bootstrap, que You y Zhou [14] describen con detalle pero orientada hacia la obtención de intervalos de confianza.

A continuación describimos los pasos que You y Zhou siguen en la obtención del intervalo de confianza para β, pero adaptándolos al método MCO-TN (ellos realmente trabajan con estimador mínimo cuadrático generalizado semiparamétrico para la componente paramétrica del modelo). Se modificó el supuesto realizado para los errores del modelo, ya que suponen que los errores siguen un proceso autorregresivo de orden 1, y, en los pasos que se describen a continuación, se va a suponer que los errores siguen un proceso ARMA.

Paso 1: Se parte de la muestra inicial (X_i^T, t_i, Y_i) con $i = 1, ..., n$ y se calculan los estimadores $\hat{\beta}_h$ y $\hat{m}_h(t)$.

Paso 2: Se estiman los residuos del modelo de la siguiente manera $\hat{\epsilon}_i = Y_i - X_i^T\hat{\beta}_h - \hat{m}_h(t_i)$ con $i = 1, ..., n$.

Paso 3: Ya que se parte del supuesto de que los residuos han sido generados por un modelo ARMA(p,q), se procede a ajustar el modelo para así obtener las innovaciones de los residuos, e_i.

Paso 4: A partir de las innovaciones de los residuos, una vez eliminadas las p primeras, se genera la muestra de innovaciones bootstrap $e_i^\star$, para $-N \leq i \leq n$ (con N suficientemente grande), como una muestra aleatoria de la distribución empírica de las innovaciones.

Paso 5: Se obtiene la muestra de residuos bootstrap $\{\epsilon_i^\star, i = 1, ..., n\}$ de la siguiente manera $\epsilon_i^\star = \sum_{j=0}^{\infty} \theta_j e_{i-j}^\star$ donde θ_j representa los coeficientes de la expresión causal del ARMA. En la práctica $\epsilon_i^\star$ se puede aproximar por $\sum_{j=0}^{N} \theta_j e_{i-j}^\star$ (para N suficientemente grande).

Paso 6: Se obtiene la muestra bootstrap $\{Y_i^\star, i = 1, ..., n\}$ por medio de la expresión $Y_i^\star = X_i^T\hat{\beta}_h + \hat{m}_h(t_i) + \epsilon_i^\star$, para $i = 1, ..., n$.

Paso 7: A partir de $(X_i^T, t_i, Y_i^\star)$, se construye $\beta_h^\star$.

Paso 8: Se repiten los pasos 4-7 un número grande de veces (M), de forma que se obtienen las M réplicas bootstrap del estimador $\{\hat{\beta}_1^\star, ..., \hat{\beta}_M^\star\}$.

Una vez obtenidas las M réplicas bootstrap del estimador, el intervalo de confianza del 100(1-α) % para $a^T\beta$ es

$$\left[a^T\hat{\beta}_h - \frac{1}{\sqrt{n}}z(1-\frac{\alpha}{2}), a^T\hat{\beta}_h - \frac{1}{\sqrt{n}}z(\frac{\alpha}{2})\right]$$

donde a es un vector de dimesión $p \times 1$ no nulo, $1-\alpha$ es el nivel de confianza y $P^\star(\sqrt{n}(a^T\hat{\beta}^\star - a^T\hat{\beta}) \leq z(\alpha)) = \alpha$, donde $P^\star$ denota la distribución de probabilidad bajo el remuestreo bootstrap.

You y Zhou comentan en su trabajo que $\sqrt{n}(\hat{\beta}_h - \beta)$ se puede aproximar (asintóticamente) por una distribución normal de media cero y varianza $\sigma^2(n^{-1}\tilde{X}^T\tilde{X})^{-1}$, y demuestran que tal aproximación se mantiene para $\sqrt{n}(\hat{\beta}_h^\star - \hat{\beta}_h)$.

Este resultado junto con el de la normalidad asintótica del estadístico $\hat{\beta}_h$ les hace llegar a que

$$sup_{x\in\Re^p}\left|(P^\star(\sqrt{n}(\hat{\beta}_h^\star - \hat{\beta}_h) \leq x) - P(\sqrt{n}(\hat{\beta}_h - \beta) \leq x))\right| \tag{2.15}$$

converge a cero en probabilidad, cuando n tiende a infinito.

Este último resultado será esencial en la realización del contraste bootstrap que se describirá en detalle en el capítulo siguiente.

Capítulo 3

Estudio de simulación

Antes de comenzar con el estudio de simulación conviene realizar algunos comentarios previos.

En primer lugar, en este estudio de simulación nos planteamos llevar a cabo el contraste unilateral izquierdo sobre uno de los coeficientes del vector β, ($H_{0\beta} : \beta_i = 0$), por medio de las aproximaciones normal y bootstrap. De forma más general, la hipótesis nula planteada sería $H_{0\beta} : a^T\beta = 0$, donde a es un vector $(p \times 1)$ que permite obtener una combinación lineal de los parámetros que componen β. En nuestro caso a será un vector con todas sus componentes cero excepto en la posición i donde el valor será uno.

Para la realización del contraste basado en la aproximación normal tendremos en cuenta el resultado 2.4:

$$n^{\frac{1}{2}}(\hat{\beta}_h - \beta) \approx N(0, \sigma_\epsilon^2(n^{-1}\tilde{X}^T\tilde{X})^{-1})$$

lo que nos lleva al siguiente estadístico para el contraste ($H_{0\beta} : a^T\beta = 0$ versus $H_{1\beta} : a^T\beta < 0$)

$$z = \frac{n^{\frac{1}{2}}a^T\hat{\beta}_h}{\sqrt{a^T[\hat{\sigma}_\epsilon^2(n^{-1}\tilde{X}^T\tilde{X})^{-1}]a}} \rightarrow N(0,1)$$

Para la realización del contraste bootstrap nos basamos en la técnica descrita por You y Zhou [14], pero orientada hacia los contrastes. Este nuevo enfoque requiere algunas modificaciones en algunos de los pasos del proceso bootstrap, que procedemos a describir nuevamente para facilitar su lectura.
Es necesario comentar que se emplearán dos procedimientos distintos en función de que β sea unidimensional o no. De esta manera se pretende poner a prueba los métodos bootstrap.

- En el caso de que β sea unidimensional este contraste se reduce a $H_{0\beta} : \beta = 0$. En este caso los pasos del remuestreo a llevar a cabo son (con un asterisco se indicarán aquellos pasos que resultan modificados):

  Paso 1: Se parte de la muestra inicial (X_i^T, t_i, Y_i) con $i = 1, ..., n$ y se calculan los estimadores $\hat{\beta}_h$ y $\hat{m}_h(t)$.

  Paso 2: Se estiman los residuos del modelo como $\hat{\epsilon}_i = Y_i - X_i^T\hat{\beta}_h - \hat{m}_h(t_i)$ con $i = 1, ..., n$.

  Paso 3: Ya que se parte del supuesto de que los residuos han sido generados por un modelo ARMA(p,q), se procede a ajustar el modelo y así obtener las innovaciones de los residuos, e_i.

  Paso 4: A partir de las innovaciones de los residuos, una vez eliminadas las p primeras, se genera la muestra de innovaciones bootstrap $e_i^\star$, para $-N \leq i \leq n$ (con N suficientemente grande), como una muestra aleatoria de la distribución empírica de las innovaciones.

  Paso 5: Se obtiene la muestra de residuos bootstrap $\{\epsilon_i^\star, i = 1, ..., n\}$ de la siguiente manera $\epsilon_i^\star = \sum_{j=0}^{\infty} \theta_j e_{i-j}^\star$ donde θ_j representa los coeficientes de la expresión causal del ARMA (en la práctica se puede aproximar $\epsilon_i^\star$ por $\sum_{j=0}^{N} \theta_j e_{i-j}^\star$, para N suficientemente grande). Se calcula la varianza de la muestra de residuos bootstrap, $\hat{\sigma}_\epsilon^{2\star}$.

  Paso 6*: Se obtiene la muestra bootstrap $\{Y_i^\star, i = 1, ..., n\}$ bajo la hipótesis nula por medio de la expresión $Y_i^\star = \hat{m}_h(t_i) + \epsilon_i^\star$, donde $\hat{m}_h$ se corresponde con la estimación no paramétrica de la función de regresión de Y frente a t.

  Paso 7: A partir de $(X_i^T, t_i, Y_i^\star)$, se construye $\beta_h^\star$.

  Paso 8*: A continuación se calcula $z^\star$ como $\frac{n^{\frac{1}{2}} a^T \hat{\beta}_h^\star}{\sqrt{a^T[\hat{\sigma}_\epsilon^{2\star}(n^{-1}\tilde{X}^T\tilde{X})^{-1}]a}}$.

  Paso 9: Se repiten los pasos 4-8 un número grande de veces (M), de forma que se obtienen las M réplicas bootstrap del estimador $\{z_1^\star, ..., z_M^\star\}$.

  A continuación se rechaza la hipótesis nula si el valor de z es menor que el cuantil α de la muestra bootstrap del estimador, siendo α el nivel de significación.

En este procedimiento bootstrap, el cambio más importante, al orientar el remuestreo hacia la realización de contrastes, se da en el paso 6*. Este paso nos permite obtener la muestra bootstrap del estadístico z bajo la hipótesis nula, y así obtener la ditribución asintótica del estadístico del contraste.

- En caso de que β no sea unidimensional se planteará otro procedimiento bootstrap para contrastar la hipótesis nula de que $H_{0\beta} : \beta_i = 0$ para algún valor de i fijado, sin realizar ninguna hipótesis sobre el resto de componentes de β, que pueden ser cero o no. En este caso, el paso 6 a utilizar vuelve a ser el propuesto por You y Zhou, pero se debe modificar el paso 8* descrito para el caso unilateral.

  Paso 8**: Se calcula $z^\star$ como $\frac{n^{\frac{1}{2}} a^T(\hat{\beta}_h^\star - \hat{\beta}_h)}{\sqrt{a^T[\hat{\sigma}_\epsilon^{2\star}(n^{-1}\tilde{X}^T\tilde{X})^{-1}]a}}$.

Para realizar el cálculo de z, se consideró el estimador consistente de σ_ϵ^2,

$$n^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

donde $\hat{\epsilon}_i$ son los residuos del MPL tomando como parámetro de ventana h=0,25.

Otro comentario que se debe hacer antes de comenzar con el estudio de simulación es que, en lo que a la estimación de β y m se refiere, se empleó el método de validación cruzada propuesto por Aneiros y Quintela [11] para obtener el parámetro de ventana óptimo, considerando como l_n el valor cero. Se trabajó con la función de pesos Nadaraya-Watson,

$$w_{n,h}(t,t_j) = \frac{K(\frac{t-t_j}{h})}{\sum_{i=1}^{n} K(\frac{t-t_i}{h})}$$

donde K es el kernel Epanechnikov, $K(u) = \frac{3}{4}(1-u^2)I_{\{|u|<1\}}$.

Una vez aclarados estos puntos, procedemos a explicar las simulaciones realizadas.

En ellas se tuvieron en cuenta dos tamaños muestrales (n=100 y n=200) y se consideraron dos modelos para los errores (AR(1) y ARMA(1,1)). Primeramente para el caso AR(1) se consideró el modelo $\epsilon_i = 0,7\epsilon_{i-1} + e_i$, mientras que para los errores ARMA(1,1) se consideró $\epsilon_i = 0,7\epsilon_{i-1} + e_i + 0,3e_{i-1}$.

Se realizaron 100 simulaciones y $M = 100$ réplicas bootstrap, y se consideró $\alpha = 0,05$ como nivel de significación.

3.1. Primera simulación

En primer lugar se estudió el modelo $Y_i = x_i\beta + m(t_i) + \epsilon_i$ $(i = 1, ..., n)$ donde $m(t_i) = \sin(2\pi t_i)$ (figura 3.1) y $t_i = \frac{i-0,5}{n}$. La variable x_i se genera según una distribución uniforme (0,1), una vez para cada valor de n considerado.

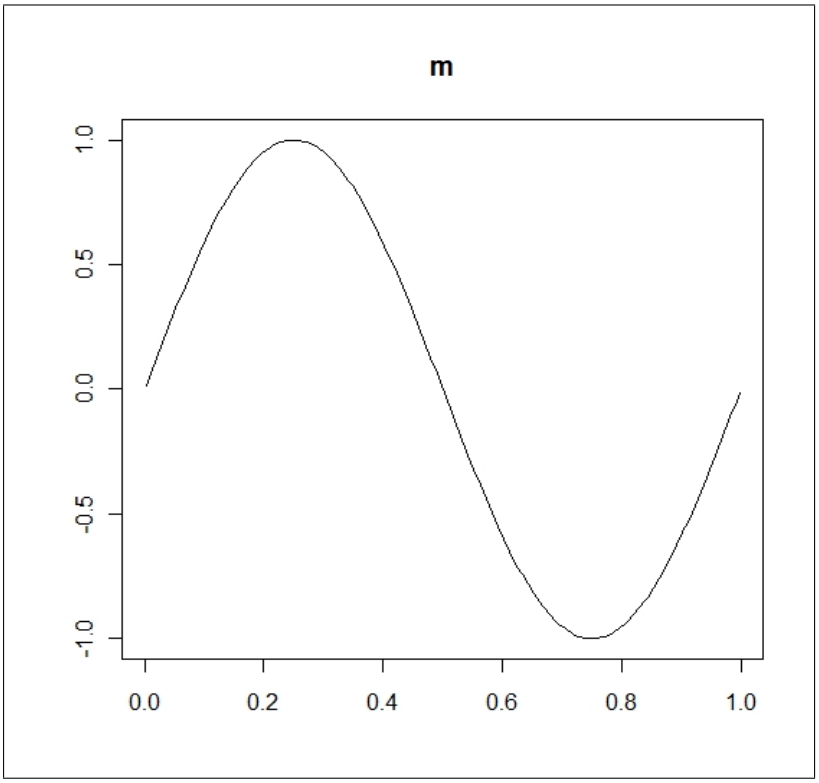

Figura 3.1: Función m para la primera simulación.

Como resultados de este primer estudio de simulación se obtuvieron los porcentajes de rechazo para ambos contrastes (normal y bootstrap en el caso unidimensional) (tabla 3.1), en función de las diferentes combinaciones posibles entre los valores considerados para n y los modelos ARMA para los errores, tomando $\beta = 0$, $\beta = -0,5$ y $\beta = -1$. Recordemos que la hipótesis nula es $H_0 : \beta = 0$ y la alternativa es $H_1 : \beta < 0$.

Cuando generamos valores de Y bajo la hipótesis de que β es igual a cero, la proporción de rechazos se corresponde con el error de tipo I. En este aspecto se podría decir que, con muestras pequeñas, funciona mejor boostrap aunque ambos contrastes se comportan de manera similar. Además, a medida que aumentamos el tamaño de la muestra la probabilidad del error de tipo I tiende a 0,05 (nivel de significación).

ϵ_i	n	**Aproximación**	$\beta = 0$	$\beta = -0,5$	$\beta = -1$
AR(1)	100	Normal	1	23	72
		Bootstrap	2	23	79
	200	Normal	3	53	95
		Bootstrap	3	54	95
ARMA(1,1)	100	Normal	1	16	58
		Bootstrap	2	22	65
	200	Normal	3	30	89
		Bootstrap	7	60	98

Cuadro 3.1: Porcentaje de rechazos para la primera simulación

Cuando generamos valores de Y bajo la hipótesis de que β es distinta de cero ($\beta = -0,5$ o $\beta = -1$), la proporción de rechazos está midiendo la potencia del contraste (probabilidad de rechazar la hipótesis nula cuando ésta es falsa). En ambos contrastes a medida que se aumenta el tamaño de la muestra aumenta el porcentaje de veces que se rechaza la hipótesis nula, pero el contraste bootstrap detecta en mayor medida la existencia de componente lineal en el modelo.

Si nos fijamos en los errores empleados, se observan mayores diferencias (a favor del bootstrap) cuando el error es ARMA(1,1).

3.2. Segunda simulación

En esta segunda simulación se planteó el modelo:

$$Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4 + m(t_i) + \epsilon_i \ (i = 1, ..., n)$$

donde la variable $Y = (Y_1, ..., Y_n)$ es estacional de período 3, de forma que $x_1 = (x_{11}, ..., x_{n1})$, $x_2 = (x_{12}, ..., x_{n2})$ y $x_3 = (x_{13}, ..., x_{n3})$ se corresponden con las variables indicadoras de cada cuatrimestre y $x_4 = (x_{14}, ..., x_{n4})$ representa una supuesta intervención ocurrida en la observación $\frac{n}{2} + 1$ y que continúa hasta la última de las observaciones. Como función suave se empleó $m(t) = t^3(1-t)^3$ (figura 3.2) y, al igual que en el ejemplo anterior, se consideró $t_i = \frac{i-0,5}{n}$. El contraste a realizar es $H_0 : \beta_4 = 0$ frente a $H_1 : \beta_4 < 0$.

Al incluir en el modelo las tres variables indicadoras que identifican cada cuatrimestre, el modelo se vuelve redundante ya que $x_1 + x_2 + x_3$ es igual a 1. Es decir,

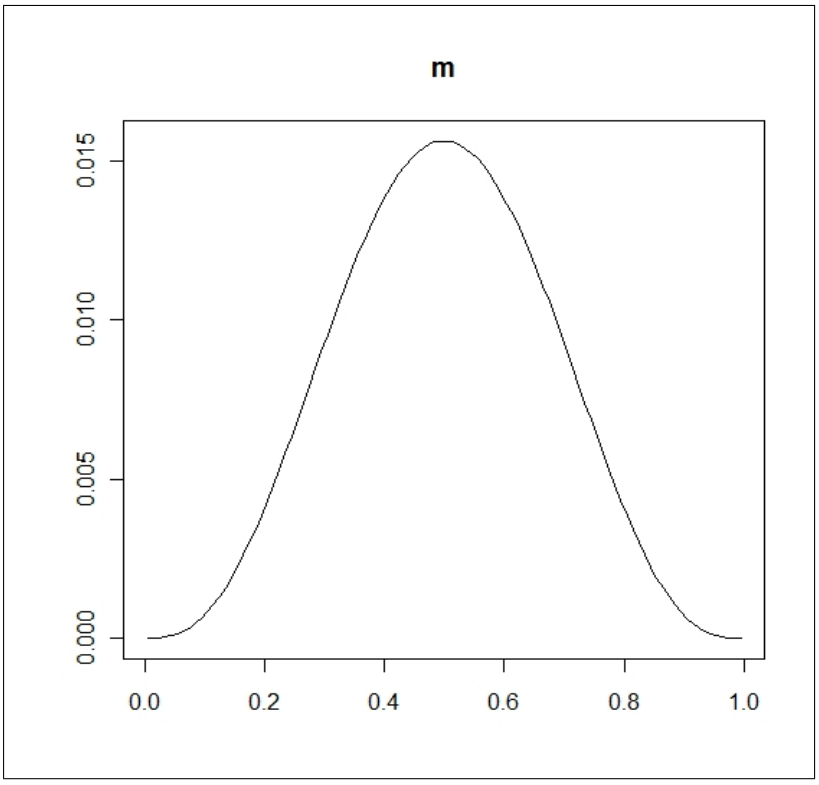

Figura 3.2: Función m para la segunda simulación.

una de las variables indicadoras se podría expresar en función del resto de variables indicadoras, por ejemplo x_1 se podría expresar como $1 - x_2 - x_3$. De esta manera el modelo anterior se puede reescribir como sigue

$$Y = (1 - x_2 - x_3)\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + m(t) + \epsilon =$$
$$= (\beta_2 - \beta_1)x_2 + (\beta_3 - \beta_1)x_3 + x_4\beta_4 + (m(t) + \beta_1) + \epsilon$$

Es decir, $Y = x_2\tilde{\beta}_2 + x_3\tilde{\beta}_3 + x_4\beta_4 + \tilde{m}(t) + \epsilon$, donde $\epsilon = (\epsilon_1, ..., \epsilon_n)$ es el término de error del modelo. Por tanto, éste sería el modelo a ajustar.

A partir de la estimación de $\tilde{\beta}_2$, $\tilde{\beta}_3$, β_4 y $\tilde{m}(t)$ se pueden obtener la estimación correspondiente al modelo original de la siguiente manera

$$\beta_2 = \tilde{\beta}_2 + \beta_1$$

$$\beta_3 = \tilde{\beta}_3 + \beta_1$$

$$m(t) = \tilde{m}(t) - \beta_1$$

Ahora quedaría por obtener la estimación de β_1. Para ello se impone la condición de que $\beta_1 + \beta_2 + \beta_3 = 0$. Esta condición nos lleva al siguiente resultado

$$\beta_1 + \beta_2 + \beta_3 = \beta_1 + (\tilde{\beta}_2 + \beta_1) + (\tilde{\beta}_3 + \beta_1) = 3\beta_1 + \tilde{\beta}_2 + \tilde{\beta}_3 = 0$$

por tanto $\beta_1 = \frac{-(\tilde{\beta}_2 + \tilde{\beta}_3)}{3}$.
Esta condición también nos garantiza que la elección de la variable indicadora, que no se incluye en el modelo, no influya en las estimaciones.

Al igual que en la simulación anterior, en la tabla 3.2 se presentan los porcentajes de rechazo para ambos contrastes, en función de las diferentes combinaciones posibles entre los valores considerados para n y los modelos ARMA para los errores. En este caso se tomó como $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$ los vectores $(-1; 3; -2; \mathbf{0})$, $(-1; 3; -2; \mathbf{-0{,}5})$ y $(-1; 3; -2; \mathbf{-1})$.

ϵ_i	n	**Aproximación**	$\beta_4 = 0$	$\beta_4 = -0,5$	$\beta_4 = -1$
AR(1)	100	Normal	16	27	37
		Bootstrap	29	39	54
	200	Normal	34	42	59
		Bootstrap	33	42	58
ARMA(1,1)	100	Normal	14	28	33
		Bootstrap	28	36	53
	200	Normal	23	31	44
		Bootstrap	27	39	43

Cuadro 3.2: Porcentaje de rechazos para la segunda simulación

Llama la atención que la proporción de rechazos cuando estamos bajo la hipótesis nula es elevada en ambos contrastes. Esto, que no ocurre en la primera simulación, podría estar motivado por el diferente procedimiento bootstrap empleado.
A la vista de los resultados se podría decir que ninguno de los contrastes (normal y bootstrap para el caso multidimensional) detecta muy bien la falta de componente lineal en el modelo, aunque es el bootstrap el que da peores resultados. No ocurre lo mismo cuando se está bajo la hipótesis alternativa, en ese caso el porcentaje de rechazos es superior en el método bootstrap en, prácticamente, todos los casos.

Por otro lado se aprecia que al aumentar el tamaño de muestra los contrastes son prácticamente iguales. Y, a diferencia de lo que ocurría en la simulación anterior, no se observa diferencias entre emplear errores AR(1) o ARMA(1,1).

Se podría decir, siempre teniendo en cuenta que los modelos de las simulaciones son distintos, que el primero de los procedimientos bootstrap planteados va mejor que el planteado en el caso multidimensional.

Capítulo 4

Aplicación a datos reales

En este capítulo nos planteamos la tarea de evaluar el impacto de la introducción de la VNC-7v sobre los ingresos por neumonía neumocócica en Galicia. Para ello modelizaremos la serie mensual de las tasas de ingresos hospitalarios por neumonía neumocócica a través de un MPL, así como un MLS y un MBJ.
Se consideraron todos los ingresos por neumonía neumocócica, es decir aquellos ingresos con diagnóstico principal de neumonía neumocócica (código 481 de la CIE 9-MC[1]), ocurridos en los hospitales gallegos del Servizo Galego de Saúde (SERGAS) y POVISA, durante el período de estudio 1997-2009.

La información correspondiente a los ingresos por neumonía neumocócica se obtuvo a partir de la base de datos del Conjunto Mínimo Básico de Datos de Altas Hospitalarias (CMBD-AH).

A la hora de calcular las tasas, la información correspondiente a la población de Galicia para el período 1998-2009, por sexo y grupos de edad, se obtuvo del Instituto Galego de Estatística. Y para el año 1997, se obtuvo de estimaciones intercensales llevadas a cabo por la Dirección Xeral de Innovación e Xestión da Saúde Pública.

Se calcularon las tasas de ingreso mensuales por 100.000 habitantes y se tuvieron en cuenta todos aquellos ingresos ocurridos en niños menores de 2 años. El motivo de realizar el estudio sólo en este grupo de edad se debe a que se trata de la población diana de la intervención, es decir, es el grupo de edad para el que está destinada la vacuna. No se estudiarán al grupo de mayores de 65 años ya que a ellos no se les administra esta vacuna, sino la vacuna antineumocócica polisacárida 23-valente.

Como ya se destacó en la introducción, la VNC-7v fue introducida en Galicia en octubre del año 2001. El período que va desde ese mes hasta diciembre de 2002

[1] Clasificación Internacional de Enfermedades 9, Modificación Clínica

se consideró como un período de transición, durante el que la vacuna fue introduciéndose progresivamente. Por ese motivo las observaciones correspondientes a dicho período fueron eliminadas de la serie (período de transición: desde la observación 58 a la 72, ambas incluidas). En la figura 4.1 se representa la serie completa de los datos, donde las líneas verticales delimitan el período considerado como transición.

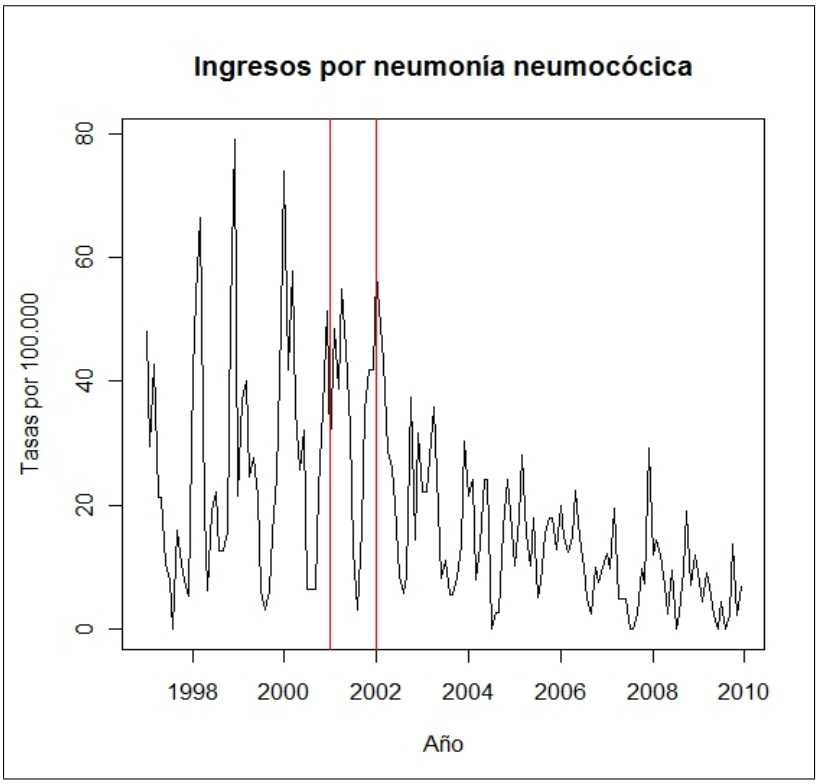

Figura 4.1: Serie de datos reales (completa). Período 1997-2009.

La serie resultante de eliminar dicho período de transición (figura 4.2) será con la que se trabajará a la hora de aplicar los tres métodos estudiados en este trabajo.

Si observamos la serie (figura 4.2) se aprecia claramente la presencia de heterocedasticidad, además de observaciones donde la tasa vale cero. Para tratar la heterocedasticidad de la serie se empleó la familia de transformaciones de Box-Cox que permite estabilizar su varianza. La transformación aplicada fue el logaritmo neperiano de las tasas aumentadas en 10 unidades (figura 4.3). El sumarle esta constante está justificado por el hecho de la no existencia del logaritmo neperiano de cero.

Además de ajustar los modelos propuestos también se llevará a cabo el contraste unilateral de la hipótesis nula $H_0 : \beta_{interv} = 0$ frente a la hipótesis alternativa $H_1 : \beta_{interv} < 0$. En el caso del MPL se aplicará el contraste basado en las aproximaciones normal y bootstrap, explicados en el capítulo 2.

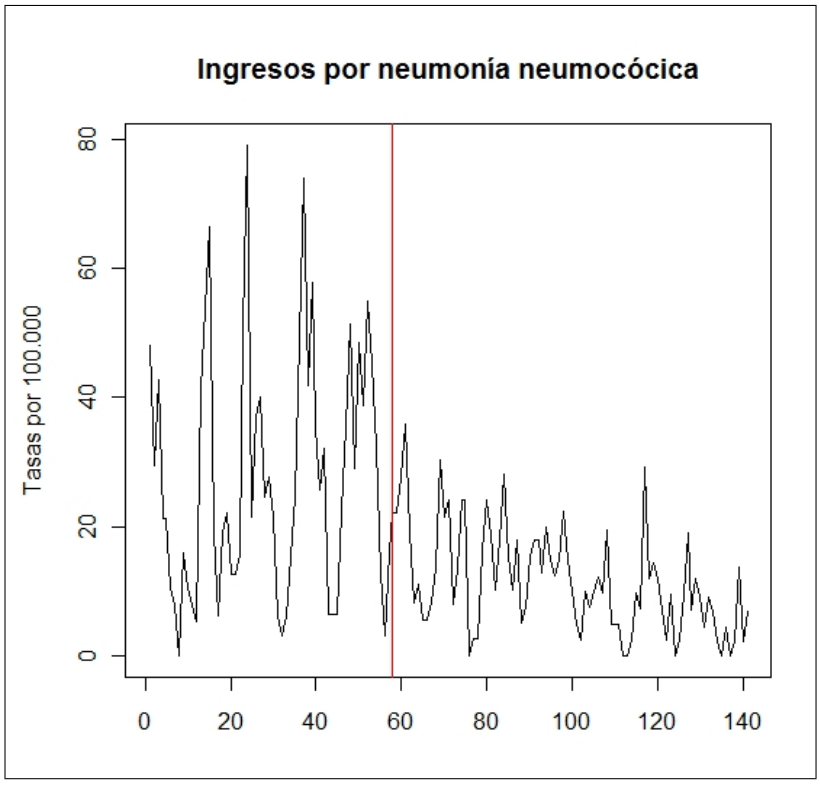

Figura 4.2: Serie de datos reales a estudio.

4.1. Aplicación del MLS

Como se comentó en la introducción, los MLS permiten estimar, de forma sencilla, el efecto de la introducción de la vacuna (intervención) sobre la serie. De esta manera se ajustan dos rectas de regresión que dan idea de la tendencia de la serie, una en el período previo a la intervención y otra en el período posterior.

Se llevó a cabo el ajuste del MLS (modelo (1.1)) en el que se incluyeron variables indicadoras de cada mes del año para poder controlar la estacionalidad, ya que se observaron tasas más altas en los meses de invierno y primavera. Tales variables indicadoras se incluyeron de forma independiente para el período pre y post intervención, dado que, aún después de aplicar la transformación de Box Cox, existe mayor variabilidad en el primer período. El modelo propuesto fue

$$Y_i = \beta_0 + \beta_{\text{tiempo}}\text{tiempo}_i + \beta_{\text{interv}}\text{interv}_i + \beta_{\text{postinterv}}\text{postinterv}_i + \sum_{j=1}^{24} \beta_j x_{ij} + \epsilon_i, \quad i = 1, ..., n$$

donde $(x_{ij}, j = 1, ..., 12)$ representan las variables indicadoras de cada mes del año para el período preintervención y $(x_{ij}, j = 13, ..., 24)$ las correspodientes para el

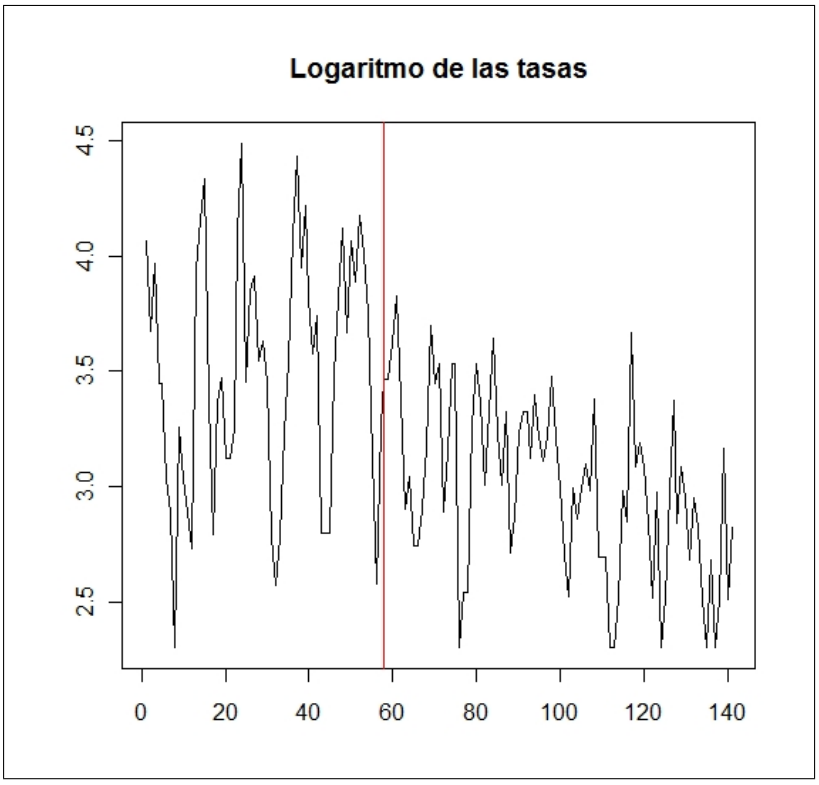

Figura 4.3: Serie a estudio transfomada.

período postintervención.

Al igual que ocurría en el ejemplo planteado en la segunda simulación, este modelo es redundante por lo que se debe proceder de un modo similar al empleado en ese caso. Por tanto, se eliminarán las variables indicadoras del mes enero tanto para antes como para después de la intervención.

De esta manera el modelo ajustado fue

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_{\text{tiempo}}\text{tiempo}_i + \tilde{\beta}_{\text{interv}}\text{interv}_i + \tilde{\beta}_{\text{postinterv}}\text{postinterv}_i + \\ + \sum_{j=2}^{12} \tilde{\beta}_j x_{ij} + \sum_{j=14}^{24} \tilde{\beta}_j x_{ij} + \epsilon_i, \quad i = 1, ..., n$$

A partir del ajuste de este modelo se puede obtener las estimaciones del modelo original de la siguiente manera

$$\hat{\beta}_0 = \hat{\tilde{\beta}}_0 - \hat{\beta}_1$$

$$\hat{\beta}_1 = \frac{-\sum_{j=2}^{12} \hat{\tilde{\beta}}_j}{12}$$

$$\hat{\beta}_i = \hat{\tilde{\beta}}_i + \hat{\beta}_1, \text{ con } i = 2, ..., 12$$

$$\hat{\beta}_{13} = \frac{-\sum_{j=14}^{24} \hat{\tilde{\beta}}_j}{12}$$

$$\hat{\beta}_i = \hat{\tilde{\beta}}_i + \hat{\beta}_{13}, \text{ con } i = 14, ..., 24$$

$$\hat{\beta}_{\text{interv}} = \hat{\tilde{\beta}}_{\text{interv}} + \hat{\beta}_1 - \hat{\beta}_{13}$$

$$\hat{\beta}_{\text{tiempo}} = \hat{\tilde{\beta}}_{\text{tiempo}}$$

$$\hat{\beta}_{\text{postinterv}} = \hat{\tilde{\beta}}_{\text{postinterv}}$$

En la tabla 4.1 se presentan la estimación de los coeficientes del modelo y en la figura 4.4 se representa la serie, junto con los valores ajustados (línea discontinua azul) y la tendencia (líneas rojas) para el período pre y post intervención.

A la vista del la figura 4.4 se observa una tendencia estimada creciente en el período preintervención y decreciente en el período postintervención, caracterizadas por los valores de la pendiente de 0,0071 ($\hat{\beta}_{\text{tiempo}}$) y -0,0075 ($\hat{\beta}_{\text{tiempo}} + \hat{\beta}_{\text{postinterv}}$), respectivamente.

El valor estimado para el coeficiente de la variable intervención fue -0,3737. Este valor se interpreta como que la intervención provocó una reducción de 0,3737 unidades en la serie del logaritmo neperiano de las tasas incrementadas en 10 unidades. Por tanto, en la serie original esto se interpreta como un descenso del 31,18 % ($(1 - e^{-0,3737}) * 100\,\%$). En cuanto a la estimación de los coeficientes para los meses preintervención y postintervención, se presenta un diagrama de barras comparando ambos períodos (Apéndice: figura 1). En él se observa que los valores estimados para el período postintervención son inferiores a los correspondientes del período preintervención, lo que indica que la variabilidad de la serie (logaritmos de las tasas incrementadas en 10 unidades) se redujo a raíz de la introducción de la vacuna.

Contrastemos ahora la hipótesis nula de que β_{interv} es igual a cero. Para ello nos basaremos en la teoría de regresión lineal y en la normalidad de los residuos del modelo, de forma que el estadítico empleado fue

$$t = \frac{\beta_i}{\hat{\sigma}\sqrt{(X^T X)^{-1}_{ii}}}$$

Coeficiente	**Estimación**	**Error estándar**
β_0	3,2902	0,0729
β_{t}	0,0071	0,0024
β_{interv}	-0,3737	0,0928
$\beta_{\text{postinterv}}$	-0,0146	0,0028
β_1	0,4401	0,1159
β_2	0,4629	0,1157
β_3	0,5809	0,1156
β_4	0,1789	0,1156
β_5	-0,0061	0,1156
β_6	-0,0376	0,1156
β_7	-0,4965	0,1156
β_8	-0,8436	0,1157
β_9	-0,4808	0,1159
β_{10}	-0,2408	0,1280
β_{11}	0,0917	0,1280
β_{12}	0,3508	0,1280
β_{13}	0,1604	0,0976
β_{14}	0,1585	0,0975
β_{15}	0,2086	0,0975
β_{16}	0,0874	0,0974
β_{17}	-0,0019	0,0974
β_{18}	-0,0178	0,0974
β_{19}	-0,3838	0,0974
β_{20}	-0,4279	0,0974
β_{21}	-0,2692	0,0974
β_{22}	0,1626	0,0975
β_{23}	0,0386	0,0975
β_{24}	0,2844	0,0976

Cuadro 4.1: Estimación de los coeficientes del MLS

donde β_i representa el coeficiente sobre el que se pretende hacer inferencia, $\hat{\sigma}^2$ es la estimación de la varianza del error, X se trata de la matriz del diseño (conteniendo las variables explicativas del modelo) y $(X^TX)^{-1}_{ii}$ representa el elemento ii de la matriz $(X^TX)^{-1}$.

Este estadístico sigue una distribución t de Student con n-p grados de libertad, siendo p el número de parámetros del modelo.

Como resultado del contraste se obtuvo un valor del estadístico de -4,03, con

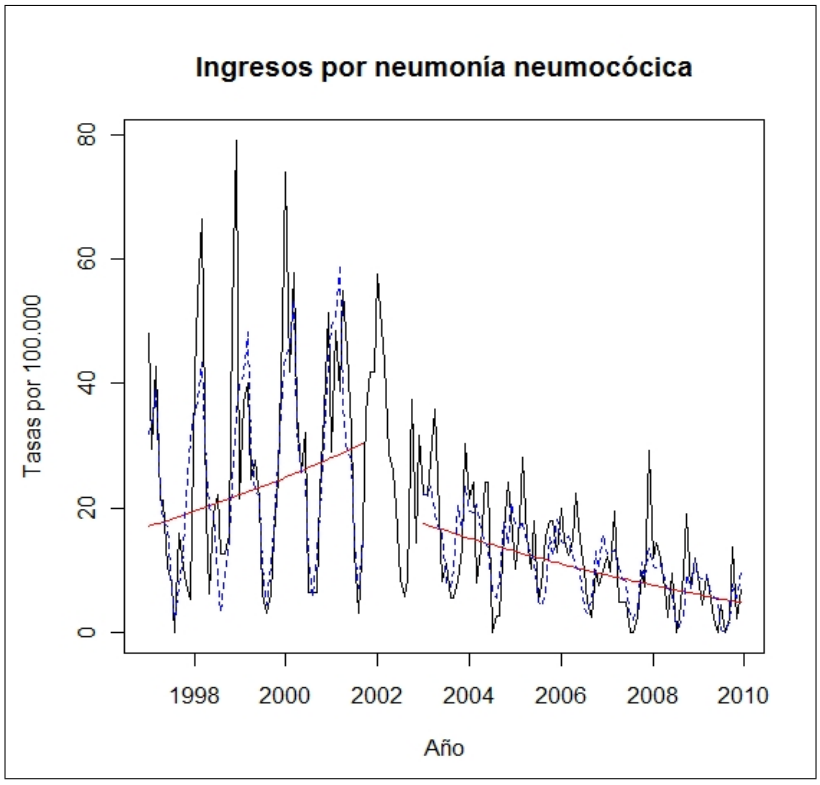

Figura 4.4: Serie original, tendencia y ajuste del MLS.

p valor menor que 0,0001. De forma que se rechaza la hipótesis nula de que la intervención no haya influido en la serie de los logaritmos. Por lo que se puede concluir que el descenso de 0,3737 unidades en esta serie se puede considerar menor que cero, significativamente.

4.2. Aplicación del MBJ

En esta sección aplicaremos un MBJ a nuestra serie a estudio (logaritmo de las tasas incrementadas en 10 unidades). Este modelo incluye una variable ficticia que representará el efecto provocado por la intervención (introducción de la vacuna en Galicia). En nuestro caso, dicho efecto se caracteriza por producir un cambio de nivel en la serie a partir del momento de inicio de la intervención, de forma que el efecto será permanente.

La variable ficticia a introducir en el modelo será una variable tipo escalón definida de la siguiente manera:

$$S_i^{(h)} = \begin{cases} 0 & \text{si } i < h \\ 1 & \text{si } i \geq h \end{cases}$$

donde h representa el instante en el que se produce la intervención (en nuestro caso, en la observación 58 (enero del año 2003)).

La serie fue modelizada de la siguiente manera:

$$Y_i = w(B)S_i^{(58)} + X_i$$

Puesto que en este tipo de modelos se parte del supuesto de que si no hubiera habido intervención la serie podría haber sido modelizada por un proceso Box Jenkins, se analizará la serie preintervención para proponer el modelo adecuado que podría estar generando la serie $\{X_i\}$.

A la vista de la serie y de la función de autocorrelación simple (FAS), (Apéndice: figura 3) nos indica que la serie preintervención presenta componente estacional (período 12), por lo que se procede a diferenciar estacionalmente.

Al observar de nuevo la FAS y la FAP (función de autocorrelación parcial) de la serie diferenciada (Apéndice: figura 4), se podría pensar que un modelo

$$ARIMA(0,0,0) \times (1,1,0)_{12} \quad o \quad ARIMA(0,0,0) \times (0,1,1)_{12}$$

han podido generar esta serie.
Un estudio pormenorizado de los residuos del primero de los modelos nos lleva a proponer el modelo $ARIMA(0,0,1) \times (1,1,0)_{12}$.

A la hora de proponer una función de transferencia se tuvo en cuenta que la introducción de la vacuna parece estar provocando un cambio en el nivel de la serie, reduciéndola en una cantidad constante. La función de transferencia considerada fue, por tanto, $w(B) = w_0$.

Una vez propuesto el modelo ARIMA de la serie preintervención y la función de transferencia, se procede a obtener los estimadores de máxima verosimilitud para el modelo de intervención. El modelo a ajustar es

$$Y_i = w_0 S_i^{(58)} + (1 + \phi_1)X_{i-12} - \phi_1 X_{i-24} + a_i + \theta_1 a_{i-1}$$

Las estimaciones obtenidas son

$$\hat{w}_0 = -0,4120 \text{ (error estándar} = 0,1349)$$

$$\hat{\phi}_1 = -0,4701 \text{ (error estándar} = 0,0859)$$

$$\hat{\theta}_1 = 0,3533 \text{ (error estándar} = 0,0709)$$

Si estudiamos los residuos del modelo se identifica un valor atípico aditivo en la observación 24 (Apéndice: figura 5), por lo que será incorporado al modelo. Las estimaciones del nuevo modelo se presentan en la tabla 4.2.

Coeficiente	**Estimación**	**Error estándar**
w_0	-0,4105	0,1343
ϕ_1	-0,4036	0,0868
θ_1	0,3519	0,0709
w_A	0,8932	0,3103

Cuadro 4.2: Estimación de los coeficientes del MBJ

Al chequear el nuevo modelo se llegó a que las innovaciones tienen media cero, varianza constante y están incorreladas. Además de aceptar su normalidad.

Por tanto, se puede decir que, la serie (en unidades logarítmicas) puede modelizarse como un

$$ARIMA(0,0,1) \times (1,1,0)_{12} \text{ sin constante}$$

que ha sufrido un cambio de nivel permanente en enero 2003 equivalente a un descenso de 0,4105 unidades. En unidades originales, esto supone un descenso del 33,67 % ($(1 - e^{-0,4105}) * 100\,\%$).

En la figura 4.5 se presentan los valores ajustados (línea discontinua roja) junto con la serie original.

Al realizar el contraste sobre la hipótesis nula $H_0 : w_0 = 0$ (es decir, $H_0 : \beta_{\text{interv}} = 0$) se llega a su rechazo, obteniendo el valor -3,06 como valor del estadístico y un p valor menor que 0,001. De manera que el descenso de 0,4105 unidades en la serie del logaritmo se puede considerar un descenso estadísticamente significativo.

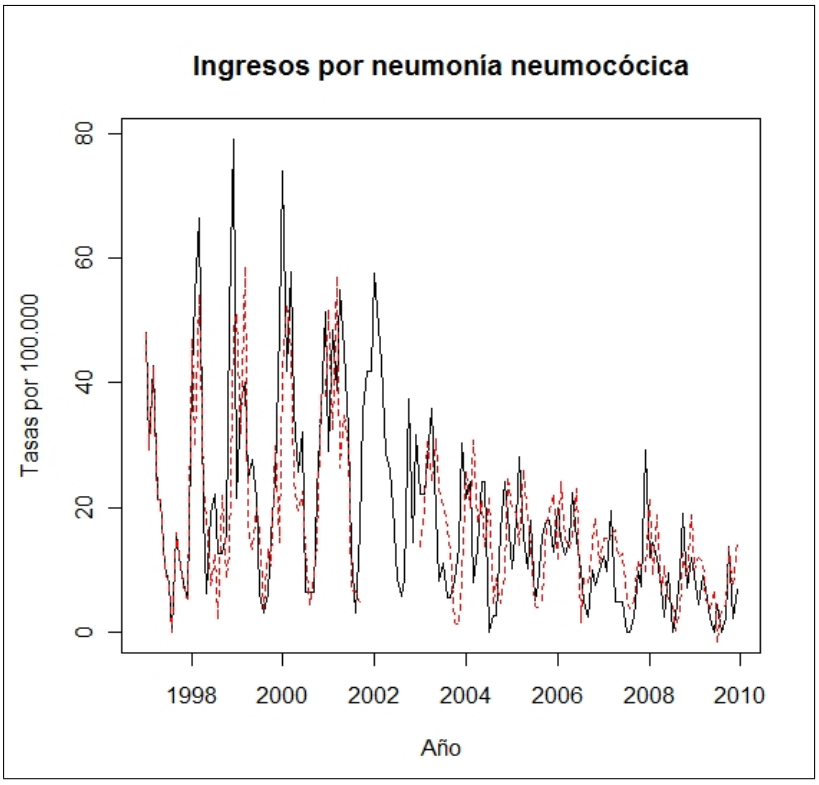

Figura 4.5: Serie original y ajuste del MBJ.

4.3. Aplicación del MPL

En esta sección aplicaremos un MPL al estudio de nuestros datos reales, de forma que el modelo estudiado será

$$Y_i = \beta_{\text{interv}}\text{interv}_i + \sum_{j=1}^{24} \beta_j x_{ij} + m(t_i) + \epsilon_i, \quad i = 1, ..., n$$

donde $(x_{ij}, j = 1, ..., 12)$ representan las variables indicadoras de cada mes del año para el período pre intervención y $(x_{ij}, j = 13, ..., 24)$ las correspodientes para el período post intervención. Y los puntos del diseño son $((\text{interv}_i, x_{i1}, .., x_{i24}), t_i)$, donde $t_i = \frac{i-0,5}{n} \in [0, 1]$.

Al igual que ocurría en la regresión segmentada, el modelo es redundante por lo que se ajustará sin incluir las variables indicadoras del mes enero, tanto para antes como para después de la intervención.

De esta manera el modelo a ajustar sería

$$Y_i = \tilde{\beta}_{\text{interv}}\text{interv}_i + \sum_{j=2}^{12}\tilde{\beta}_j x_{ij} + \sum_{j=14}^{24}\tilde{\beta}_j x_{ij} + \tilde{m}(t_i) + \epsilon_i, \quad i = 1, ..., n \tag{4.1}$$

A partir del ajuste de este modelo se pueden obtener las estimaciones del modelo original de la siguiente manera

$$\hat{\beta}_1 = \frac{-\sum_{j=2}^{12}\hat{\tilde{\beta}}_j}{12}$$

$$\hat{\beta}_i = \hat{\tilde{\beta}}_i + \hat{\beta}_1, \text{ con } i = 2, ..., 12$$

$$\hat{\beta}_{13} = \frac{-\sum_{j=14}^{24}\hat{\tilde{\beta}}_j}{12}$$

$$\hat{\beta}_i = \hat{\tilde{\beta}}_i + \hat{\beta}_{13}, \text{ con } i = 14, ..., 24$$

$$\hat{\beta}_{\text{interv}} = \hat{\tilde{\beta}}_{\text{interv}} + \hat{\beta}_1 - \hat{\beta}_{13}$$

$$\hat{m}(t) = \hat{\tilde{m}}(t) - \hat{\beta}_1$$

A la hora de ajustar el modelo (4.1) lo primero que debemos hacer es elegir el parámetro de ventana óptimo para obtener las estimaciones. Para ello aplicamos el método de validación cruzada que nos proporciona el valor 0,08 como ventana óptima.

A continuación estimamos los coeficientes de la parte lineal del modelo y la función suave (ver tabla 4.3 y figura 6 (en el Apéndice)).

En la figura 4.6 se presentan la serie original, los valores ajustados (línea discontinua azul) y la tendencia (líneas rojas) para el período pre y post intervención

De la información contenida en la tabla 4.3 se llega a que el coeficiente de la intervención estimado es -0,2559. Por lo que la serie del logaritmo de las tasas incrementadas en 10 unidades, presenta un cambio nivel en enero de 2003, disminuyendo en 0,2559 unidades.
Al trasladar este resultado a nuestra serie original (tasas de ingresos) se podría decir que la disminución fue del 22,58 %.

El diagrama de barras con las estimaciones de los coeficientes para los meses (Apéndice: figura 2) muestra, al igual que en el MLS, una disminución en el efecto

Coeficiente	Estimación	Error estándar
β_{interv}	-0,2559	0,2476
β_1	0,4302	0,1165
β_2	0,4562	0,1161
β_3	0,5741	0,1158
β_4	0,1718	0,1156
β_5	-0,0111	0,1156
β_6	-0,0397	0,1157
β_7	-0,4945	0,1159
β_8	-0,8376	0,1163
β_9	-0,4691	0,1170
β_{10}	-0,2376	0,1289
β_{11}	0,0977	0,1289
β_{12}	0,3597	0,1289
β_{13}	0,1519	0,0984
β_{14}	0,1517	0,0979
β_{15}	0,2025	0,0977
β_{16}	0,0828	0,0975
β_{17}	-0,0044	0,0974
β_{18}	-0,0176	0,0973
β_{19}	-0,3810	0,0974
β_{20}	-0,4239	0,0974
β_{21}	-0,2642	0,0975
β_{22}	0,1686	0,0976
β_{23}	0,0448	0,0978
β_{24}	0,2888	0,0981

Cuadro 4.3: Estimación de los coeficientes del MPL

provocado por cada mes del año tras la intervención, comparado con el período previo.

Por último se aplicaron los contrastes descritos en el capítulo 3 para contrastar la hipótesis nula $H_0 : \beta_{\text{interv}} = 0$ frente a la alternativa $H_1 : \beta_{\text{interv}} < 0$. El contraste normal nos lleva a aceptar que $\beta_{\text{interv}} = 0$ (p valor= $0,1597$), a diferencia de lo que ocurre con el contraste bootstrap (p valor= $0,05$). El valor del estadístico z fue -1,0334.

Cabe destacar del proceso bootstrap que, en el paso 3 de dicho proceso, se identificó un MA(3) como generador de la serie de los residuos (modelo ajustado para los residuos: $\epsilon_i = -0,4725e_{i-2} - 0,4655e_{i-3}$). A partir de este modelo se obtienen

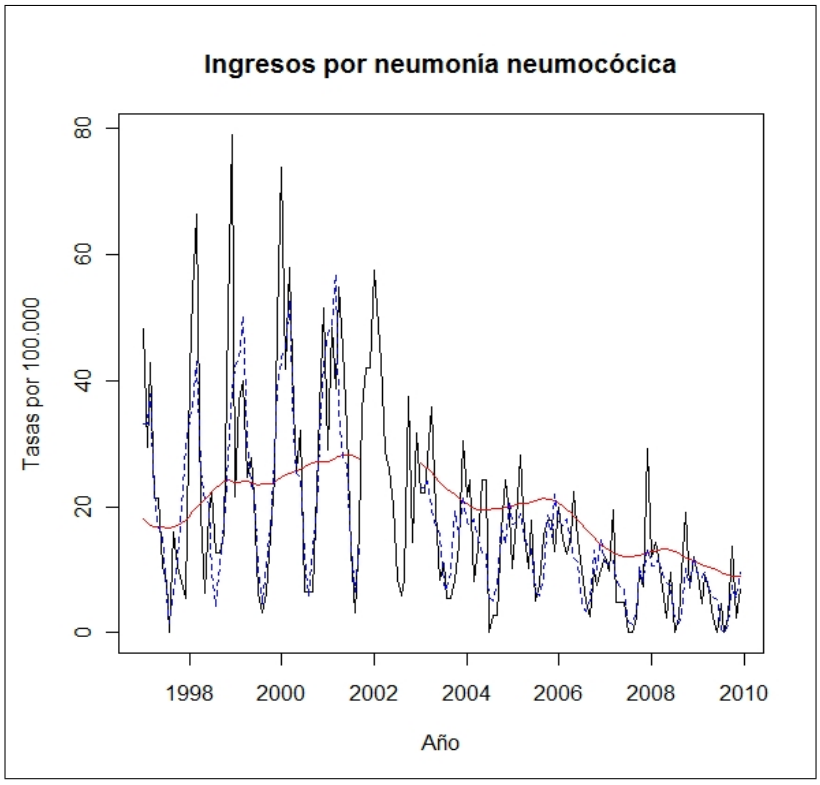

Figura 4.6: Serie original, ajuste y tendencia del MPL.

las innovaciones bootstrap y se continua con el resto de los pasos.

4.4. Comparación y comentarios

Una vez ajustados los tres modelos, nos planteamos analizar cúal de ellos da lugar a un mejor ajuste. Para ello representaremos en un diagrama de cajas conjuntamente los residuos de cada uno de los ajustes (Apéndice: figura 7).
A la vista de la gráfica podemos decir que el MBJ es el que da peores resultados ya que el recorrido intercuartílico de los residuos es más amplio que los obtenidos con los otros dos modelos. Esta conclusión se ve reflejada en la media de sus residuos al cuadrado que toma el valor más elevado (117,11) en comparación con el obtenido con los otros modelos.

Al comparar los MLS y MPL la cosa no está tan clara como en el MBJ, por eso recurrimos a la media de los residuos al cuadrado. Esta media nos proporciona un

valor de 77,96 para el MLS y un valor de 74,34 para el MPL, lo que nos permiten concluir que el MPL es mejor que MLS.

Por último comentar que, los tres modelos detectan un cambio de nivel, a excepción del MPL con aproximación normal. Además, a la vista de la figuras 4.4 y 4.6, se aprecia un cambio de tendencia, siendo esta creciente para el período preintervención y decreciente para el período postintervención.

Capítulo 5

Conclusiones

En este trabajo se ha presentado una pequeña revisión acerca de los MPL, en la que se describen algunos de los métodos aplicados en la estimación de las componentes del modelo, así como sus propiedades asintóticas y el problema de la selección del parámetro de ventana. Se han puesto a prueba dos procedimientos bootstrap a la hora de realizar contrastes sobre la componente lineal del modelo, que a su vez se compararon, por medio de un estudio de simulación, con la aproximación normal. Este estudio simulación nos sugiere que, como cabría de esperar, el primero de los procedimientos bootstrap va mejor que el segundo ya que tiene en cuenta explícitamente la hipótesis nula para generar respuestas bootstrap. En la comporación con la aproximación normal también sale ganando el bootstrap.
Por último, se aplicaron tres modelos (MLS, MBJ, MPL) a una serie de datos reales caracterizada por una variable intervención. Todos ellos detectaron un cambio de nivel en la serie a raíz de la intervención, a excepción del MPL con aproximación normal, lo que era de esperar ya que en las simulaciones hemos visto que funciona mejor el bootstrap. Además, tanto la regresión segmentada como el parcialmente lineal parecen sugerir un cambio de tendencia (primero crece y luego decrece).

Bibliografía

[1] WHO position paper (2007). Pneumococcal conjugate vaccine for childhood immunization. *Weekly epidemiological record*, **82(12)**, 93-104.

[2] Guevara, M., Barricarte, A., Pérez, B., Arriazu, M., García Cenoz, M. and Castillo, J. (2008). La vacuna neumocócica conjugada heptavalente (Prevenar): diferencias es su efectividad en distintas poblaciones. *An. Sist. Sanit. Navar*, **31(2)**, 171-192.

[3] Grijalva, C. G., Nuorti, J. P., Arbogast, P. G., Martin, S. W., Edwards, K. M. and Griffin M. R. (2007). Decline in pneumonia admissions after routine childhood immunisation with pneumococcal conjugate vaccine in the USA: a time-series analysis. *The Lancet*, **369**, 1179-1186.

[4] Wagner, A. K., Soumerai, S. B., Zhang, F. and Ross-Degnan, D. (2002). Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*, **27**, 299-309.

[5] Engle, R. F., Granger, W. J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.*, **80**, 310-319.

[6] Peña, D. (2005). *Análisis de Series Temporales*. Alianza Editorial.

[7] Speckman, P. (1988). Kernel smoothing in partial linear models. *J. R. Statist. Soc. B*, **50**, 413-436.

[8] Liang, H., Härdle, W. and Sommerfeld, V. (2000). Bootstrap approximation in a partially linear regression model. *J. Statist. Plann. Inference*, **91**, 413-426.

[9] Gao, J. (1995). Asymptotic theory for partly linear models. *Commun. Statist. Theory Method*, **24**, 1985-2009.

[10] González-Manteiga, W. and Aneiros-Pérez, G. (2003). Testing in partial linear regression models with dependent errors. *Nonparametric Statistics*, **15**, 93-111.

[11] Aneiros-Pérez, G. and Quintela-del-Río, A. (2001a). Asymptotic properties in partial linear models under dependence. *Test*, **10**, 333-355.

[12] Aneiros-Pérez, G. and Quintela-del-Río, A. (2001b). Modified cross-validation in semiparametric regression models with dependent errors. *Commun. Statist. Theory Method*, **30**, 289-307.

[13] Ghement, I. R., Heckman, N. E. and Petkau, A. J. (2007). Seasonal confounding and residual correlation in analyses of health effects of air pollution. *Environmetrics*, **18**, 375-394.

[14] You, J. and Zhou, X. (2005). Bootstrap of a semiparametric partially linear model with autoregressive errors. *Statistica Sinica*, **15**, 117-133.

Apéndice: Gráficos

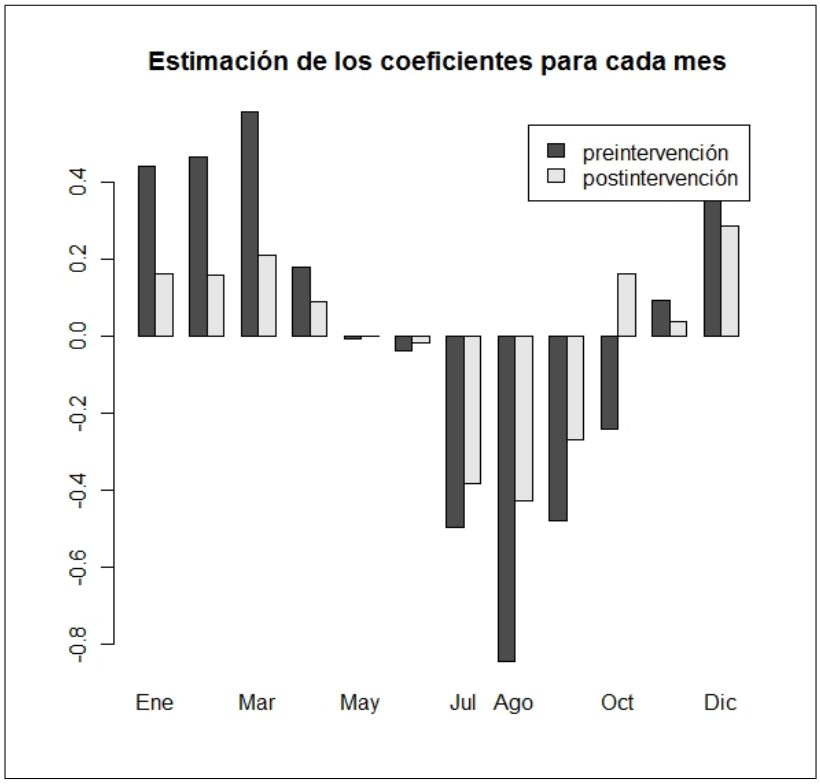

Figura 1: Diagrama de barras (MLS).

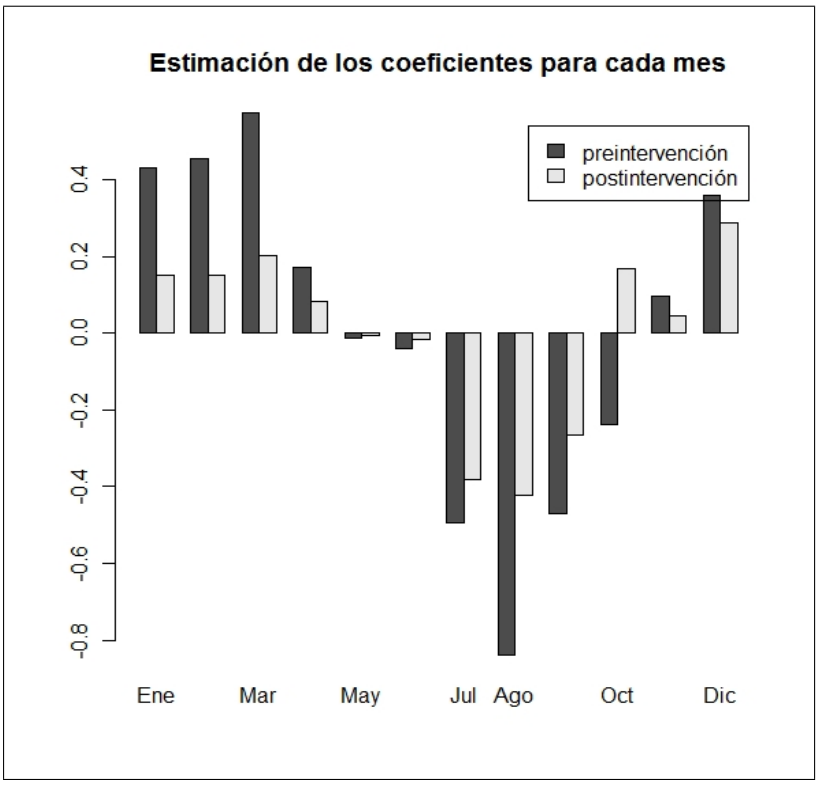

Figura 2: Diagrama de barras (MPL).

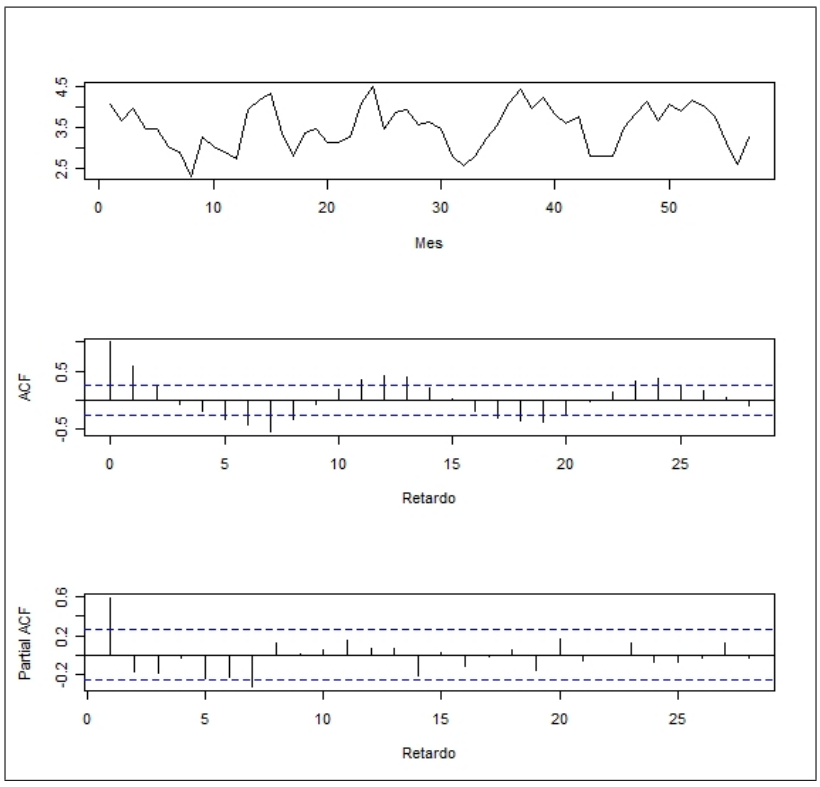

Figura 3: Serie preintervención junto con su FAS y su FAP.

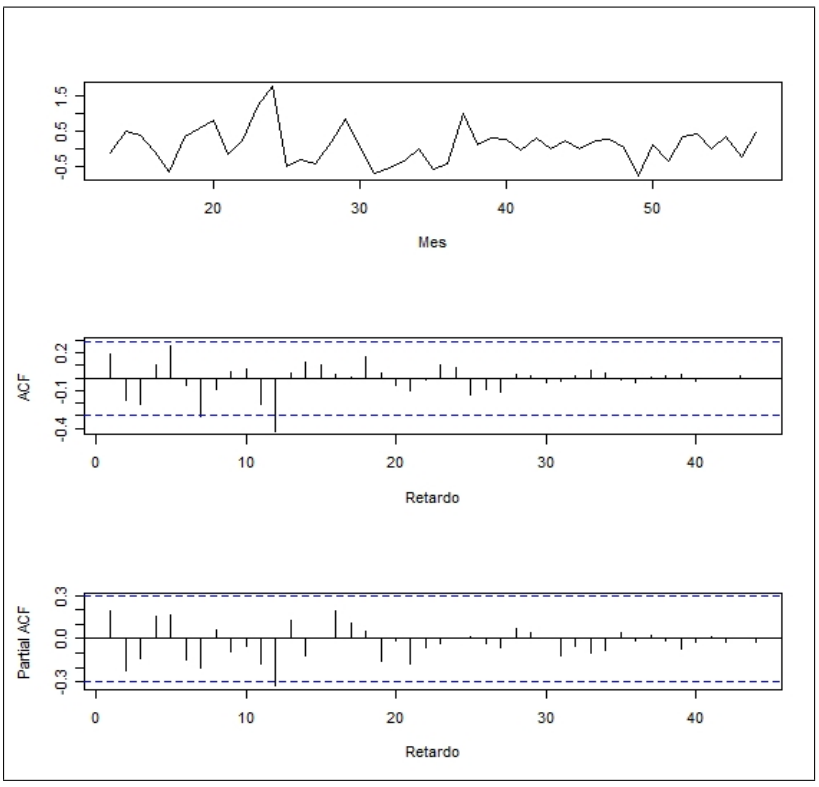

Figura 4: Serie diferenciada junto con su FAS y su FAP.

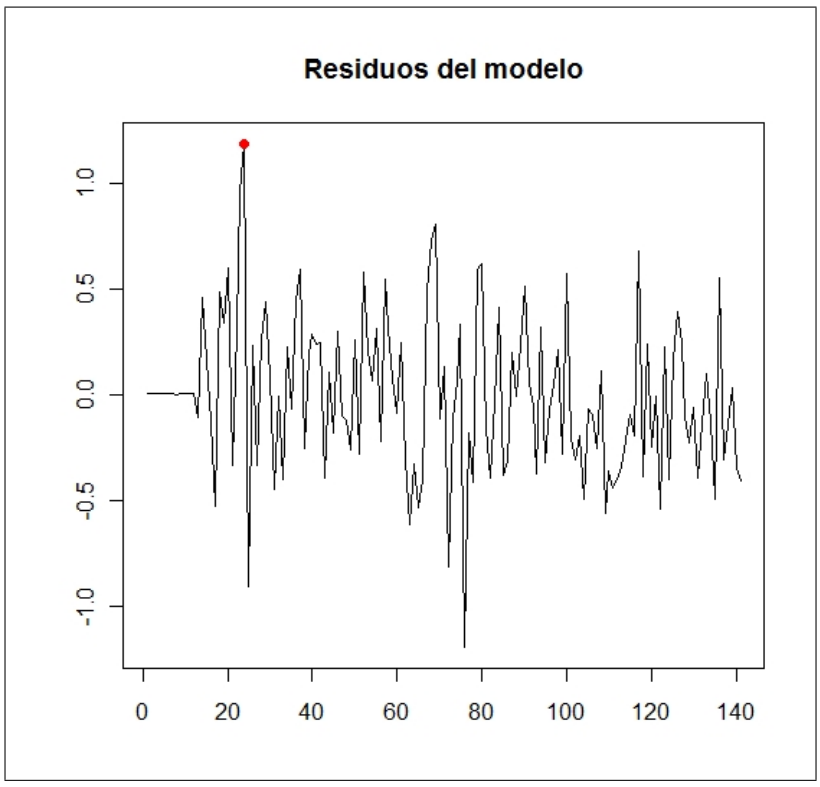

Figura 5: Residuos del MBJ y valor atípico aditivo.

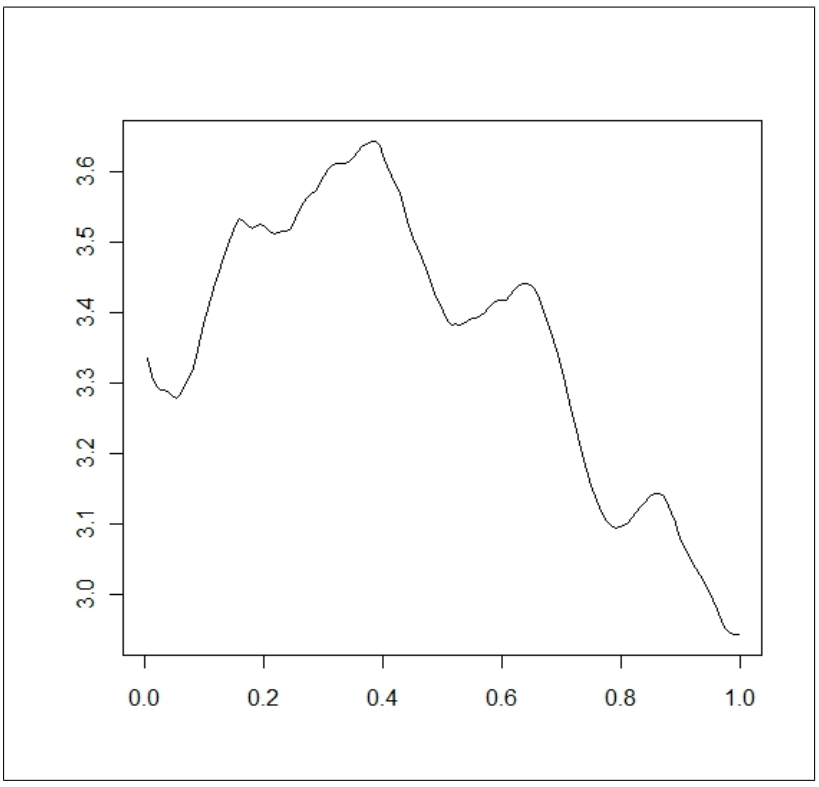

Figura 6: Estimación de la componente no paramétrica del MPL.

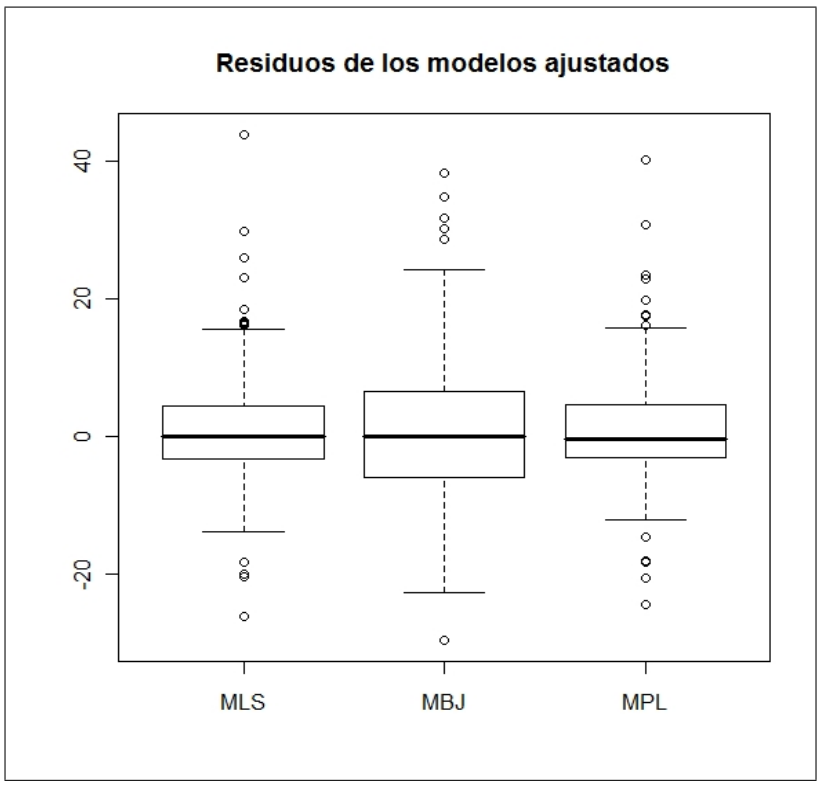

Figura 7: Diagrama de cajas de los residuos de cada uno de los modelos.