



Proyecto 1: Series de Tiempo

Clemente Ferrer, Rodrigo Pizarro

OBSERVACIÓN PRELIMINAR: La parte computacional de la tarea fue realizada usando . Además, los códigos de todos los apartados fueron subidos al GitHub público siguiente:  MAT267.

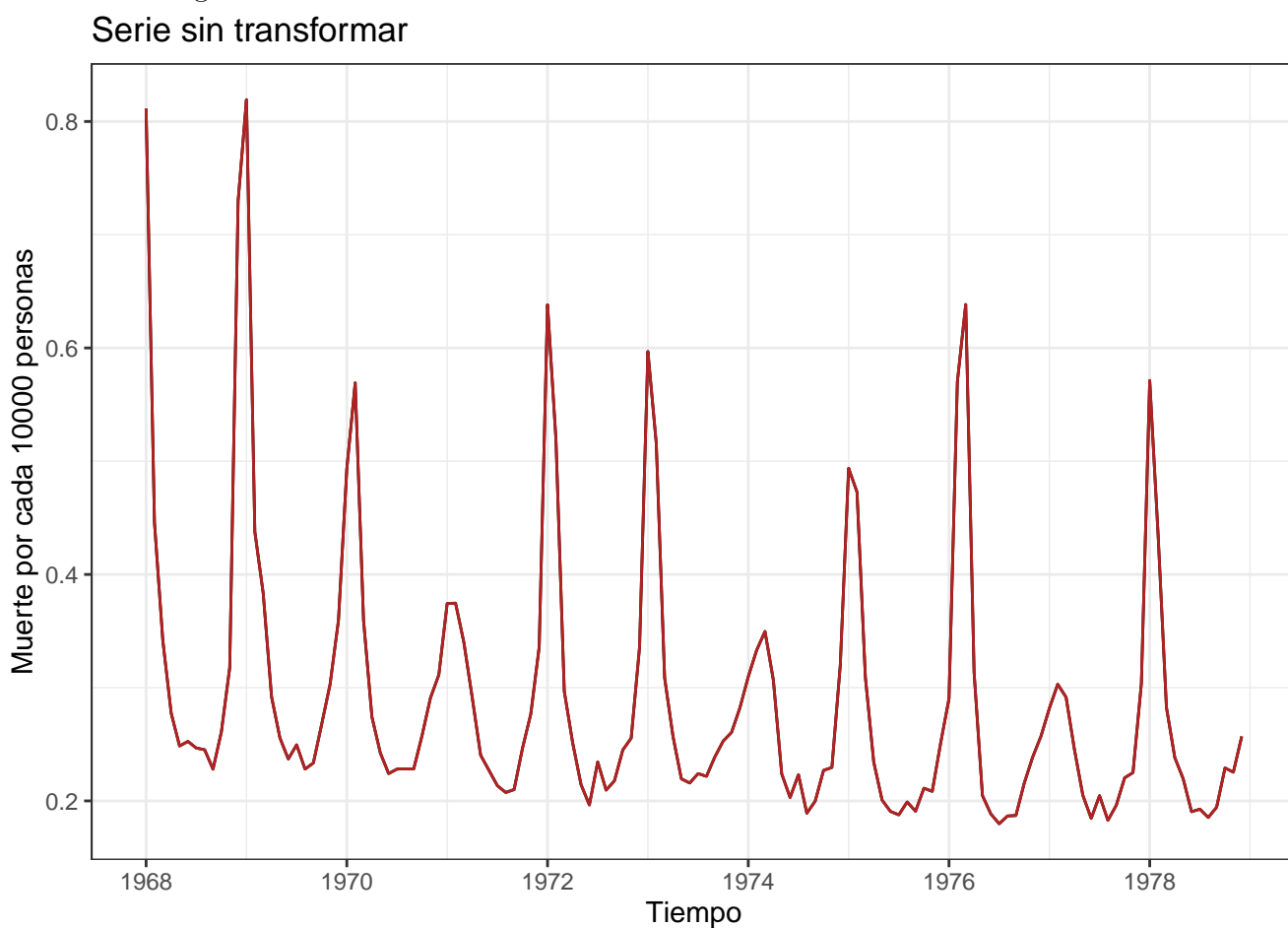
1. Considere la serie `flu.dat` que puede ser obtenida en el sitio

<http://www.stat.pitt.edu/stoffer/tsa2/tsa2.html>.

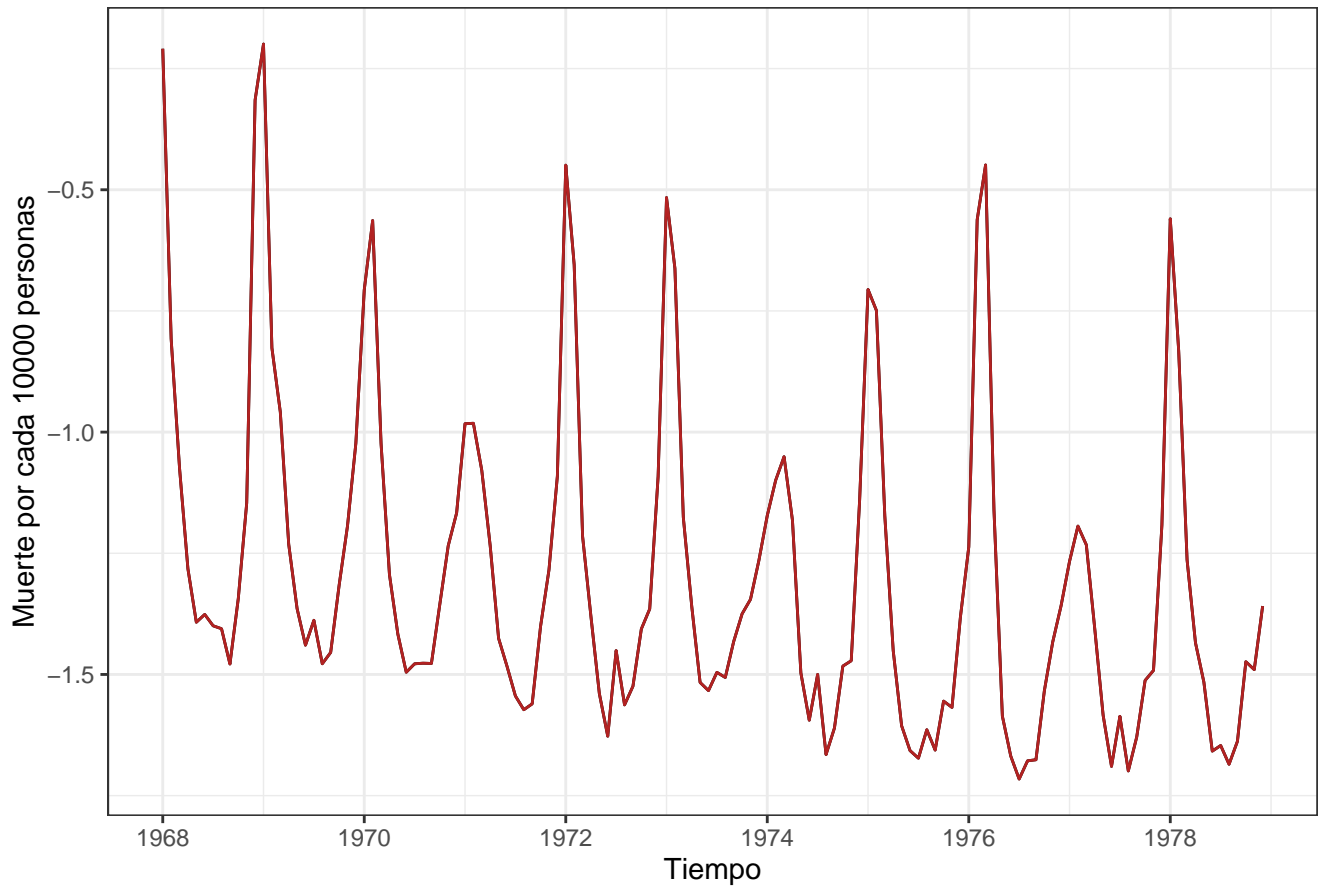
a. Transforme la serie adecuadamente para observar el efecto de la transformación en la media y la varianza.

Solución. Primero, calculamos la media y la varianza de la serie original, obteniendo $\mu = 0,2918623$ y $\sigma = 0,01578173$. Luego, podemos pasar los datos a una serie temporal, y calcular el λ apropiado para aplicar el filtro de BoxCox. En particular, el lambda calculado es $\lambda \approx 4,1 \cdot 10^{-5}$. Se sigue computando la media y la varianza de la transformada, que $\mu = -1,297194$ y $\sigma = 0,1141322$.

A continuación graficaremos ambas series:



Serie transformada



Notamos que hay clara diferencia en la varianza y en la media, tal como esperaríamos. Otro efecto claro es que como $Z_t < 1$ para todo t , al aplicar el filtro BoxCox todos nuestros datos quedan negativos y por ende la media también.

□

b. Se propone un modelo de la forma

$$Z_t = \sum_{j=0}^{\infty} \beta_j t^j + \epsilon_t \quad (1)$$

donde ϵ_t es un ruido blanco con varianza σ^2 .

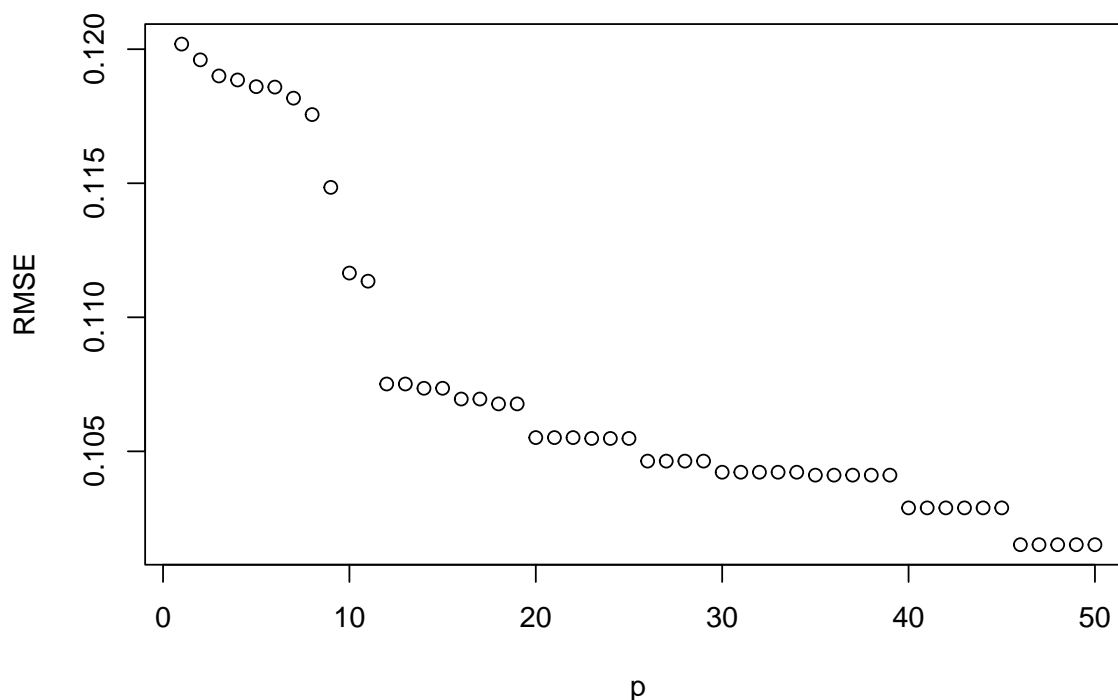
I. Proponga un método que permita truncar la serie infinita que define el modelo (1), de tal modo que el modelo resultante sea de la forma

$$Z_t = \sum_{j=0}^p \beta_j t^j + u_t$$

Solución. Ajustaremos modelos polinomiales de regresión desde $p = 1$ hasta $p = n$, donde n será un número tal que el error cuadrático medio esté cercano a 0,1 ya que esto indicaría que el error es pequeño. Note que es necesario ponerle un umbral al modelo, puesto que es obvio que mientras mas grande el grado del polinomio, menor será el error cuadrático medio. □

II. Estime p usando los datos de la serie `flu.dat`.

Solución. Calcularemos los errores cuadráticos medios desde $p = 1$ hasta $p = 50$ y veremos cuáles satisfacen o están cerca de satisfacer la condición umbral. A continuación, una gráfica de los RMSE respecto al valor p :



Notamos que a partir de $p = 46$, existe una tendencia a estabilizarse, por lo que estimaremos que $\hat{p} = 50$ es el valor adecuado. □

- III.** Estime los parámetros del modelo: β_j , $j = 0, 1, \dots, \hat{p}$, donde \hat{p} es la estimación de p propuesta en ii.

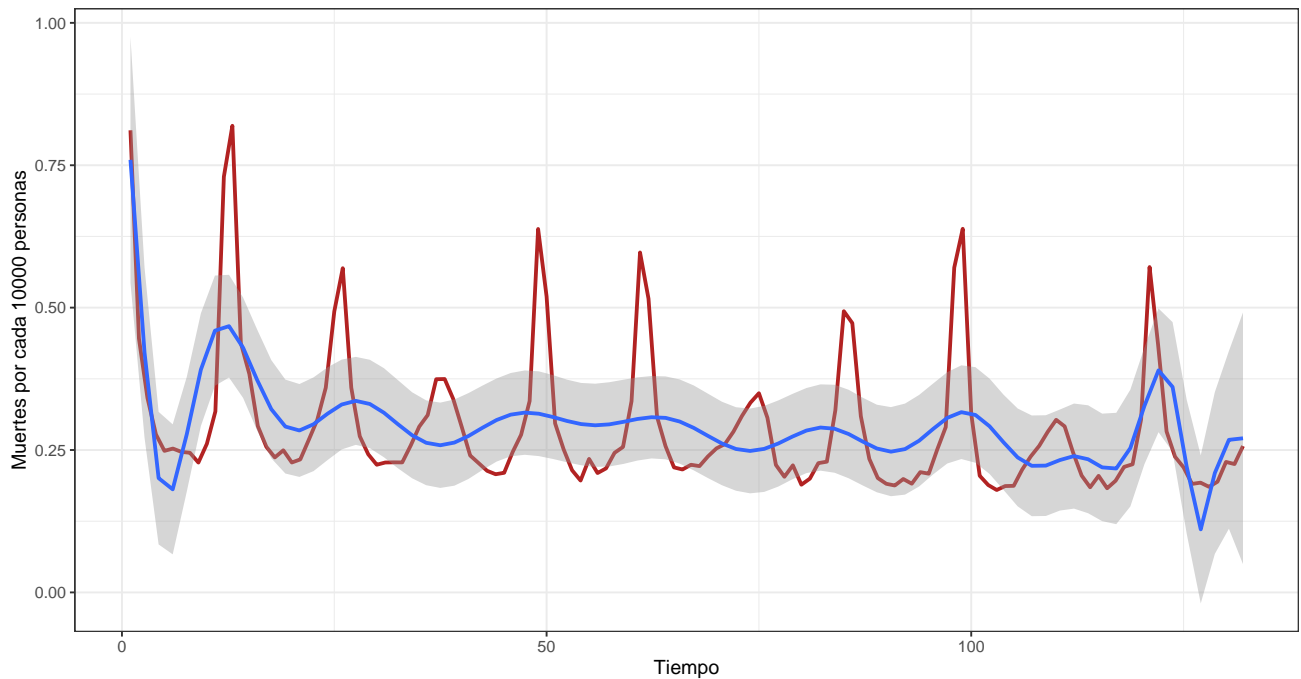
Solución. El comando `lm` nos entrega de manera inmediata los parámetros β_j :

	Estimate
1	8.786674e-02
2	-2.013942e-01
3	5.974867e-02
4	-8.154580e-03
5	6.363552e-04
6	-3.082743e-05
7	9.428177e-07
8	-1.704422e-08
9	1.184500e-10
10	1.952628e-12
11	-5.597656e-14
12	4.917439e-16
14	-2.440974e-20
16	1.619609e-24
18	-8.942500e-29
20	3.197432e-33
23	-5.604830e-40
26	9.564484e-47
30	-7.728711e-56
35	3.921105e-67
40	-1.768405e-78
46	2.576907e-92

A la izquierda es el valor j -ésimo y a la derecha el valor de β_j . Los índices que no aparecen están asociados a coeficiente 0. □

- c. Grafique la serie original y la serie ajustada en un mismo gráfico. ¿Hay evidencia para aseverar que el modelo estimado es una buena representación de los patrones de la serie original? Justifique.

Solución. Ajustamos las curvas en un mismo gráfico, considerando la curva roja como la original y la azul como la ajustada:

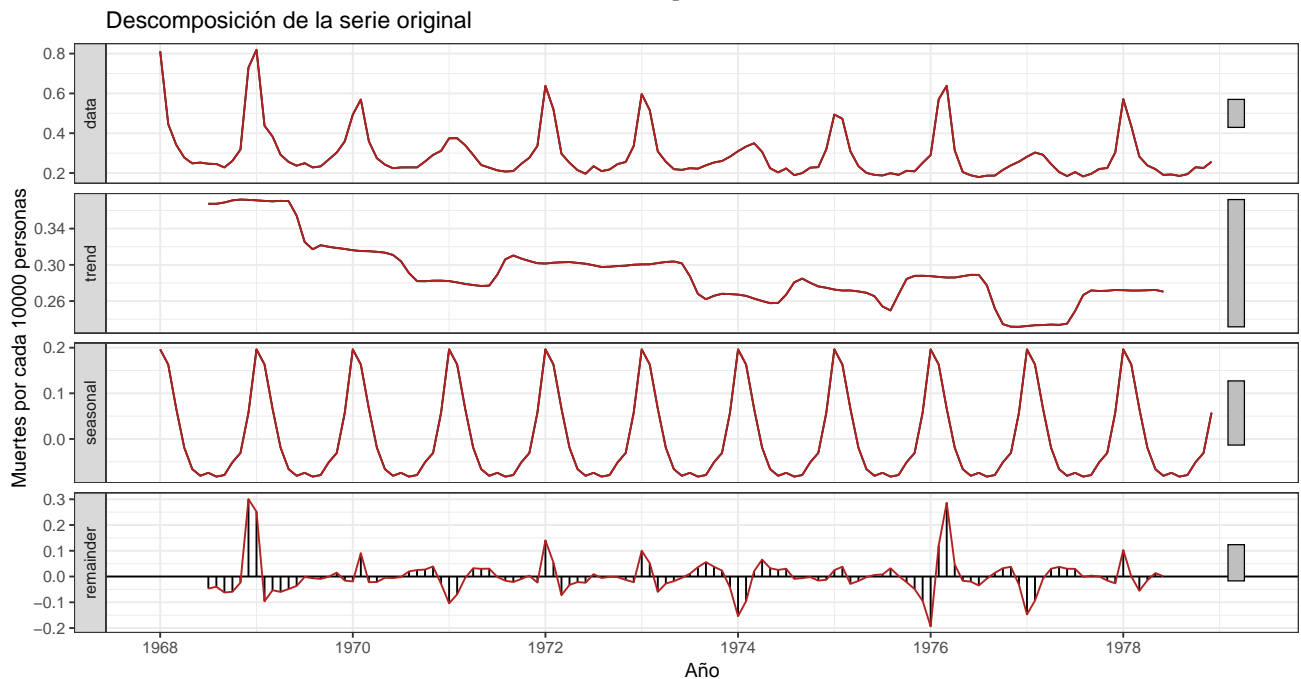


□

Notamos que si hay evidencia para justificar que el modelo estimado es una buena representación, ya que como primer parámetro, tenemos la minimización del RMSE y en segundo lugar, las gráficas obviamente difieren debido a un par de outliers, pero fuera de eso, la nube de puntos se concentra en el área gris, y por lo tanto, nuestra curva azul es un buen estimador para esta nube de puntos.

- d. Ajuste un modelo de descomposición a la serie `flu.dat`.

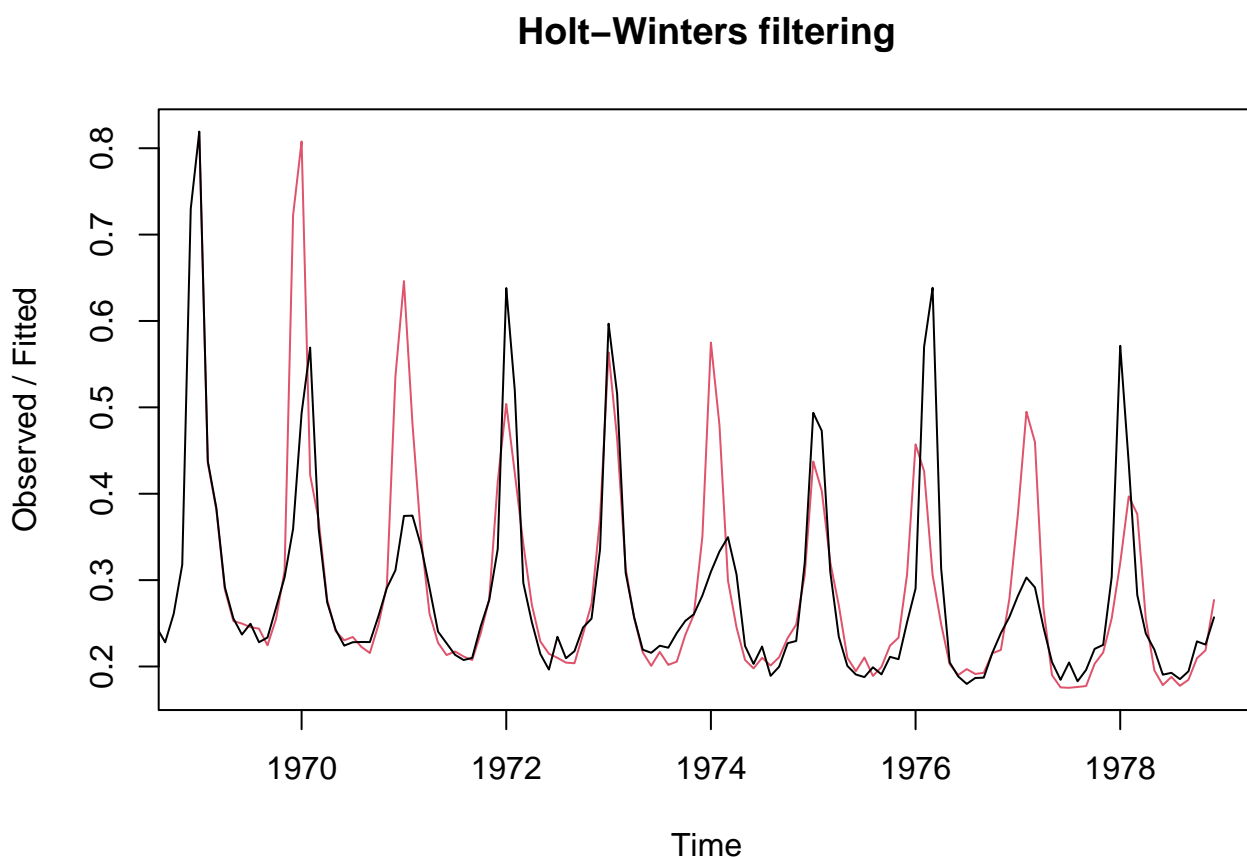
Solución. Descomponemos usando el comando `decompose` y graficamos:




□

- e. Ajuste un modelo de Holt-Winters a la serie `flu.dat`.

Solución. Ajustamos un modelo de Holt-Winters con la función `holt` y graficamos:



- f. De los dos modelos propuestos en la parte b), d) y e) ¿Cuál es el más apropiado?

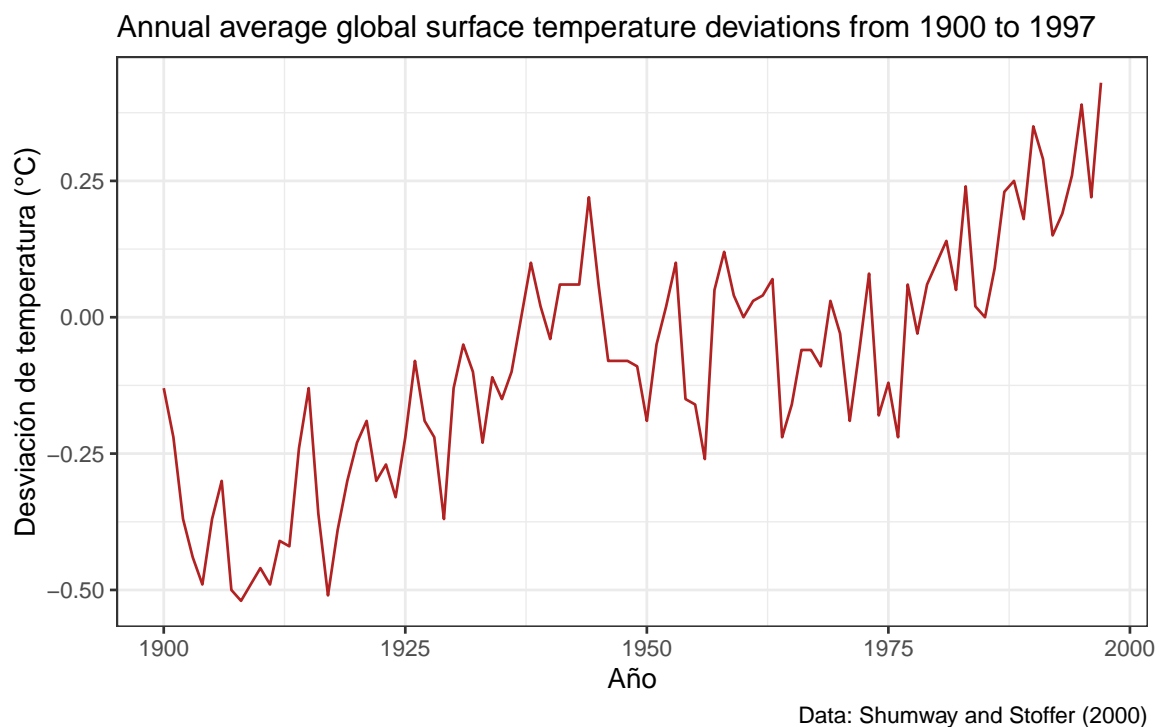
Solución. Analizando las gráficas y los valores de los parámetros de interés, podemos notar que el modelo que parece ser más apropiado es el modelo de Holt-Winters, ya que la comparativa de las gráficas que podemos observar en el ítem anterior, nos confirma que no existe una gran diferencia entre lo observado y lo ajustado. Más aún,  nos permite obtener el RMSE del modelo ajustado con Holt-Winters. Si lo calculamos, tenemos que $RMSE \approx 0,08130$, que es mejor que el modelo ajustado en b). □

2. En este ejercicio es necesario obtener la serie asociada al calentamiento de la tierra descrito en grados centígrados entre los años 1900-1997. Esta serie es presentada en Shumway and stoffer 2000, pagina 5. Para bajar el archivo `globtemp.dat` encuentre el sitio web

<http://www.stat.pitt.edu/stoffer/tsa2/tsa2.html>.

- a. Grafique la serie `globtemp.dat` en el tiempo.

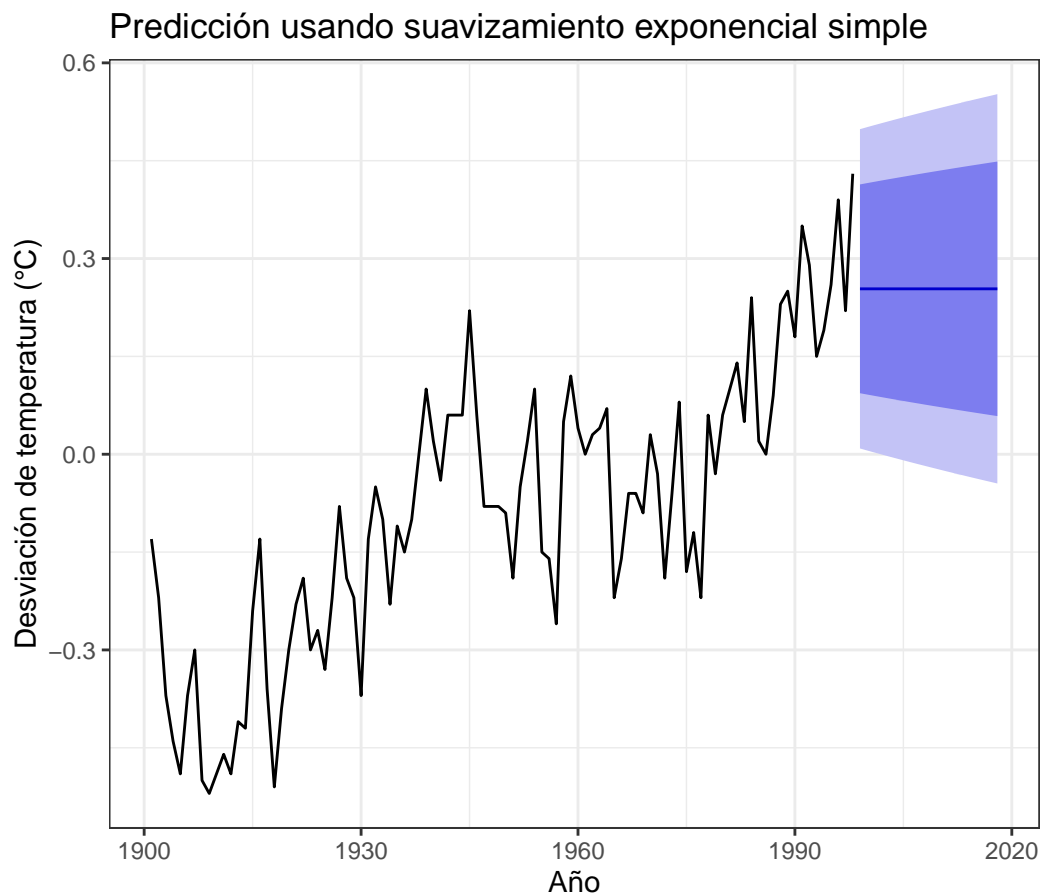
Solución. Convirtiendo la información entregada en el archivo a un data frame obtenemos el siguiente gráfico:



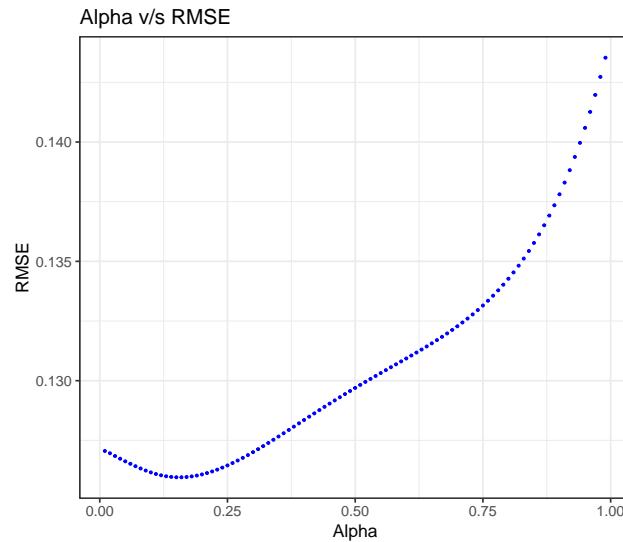
□

- b. Use un modelo de suavizamiento exponencial simple para predecir la serie hacia el futuro. Considere un valor apropiado para α .

Solución. Utilizando la función `ses`, y prediciendo para los próximos 15 años se obtiene el siguiente gráfico



En particular se usó un $\alpha = 0,16$ que resultó ser el que minimizaba el error cuadrático medio para $\alpha \in \{0, 0,01, \dots, 0,99, , 1,00\}$ como indica la figura inferior

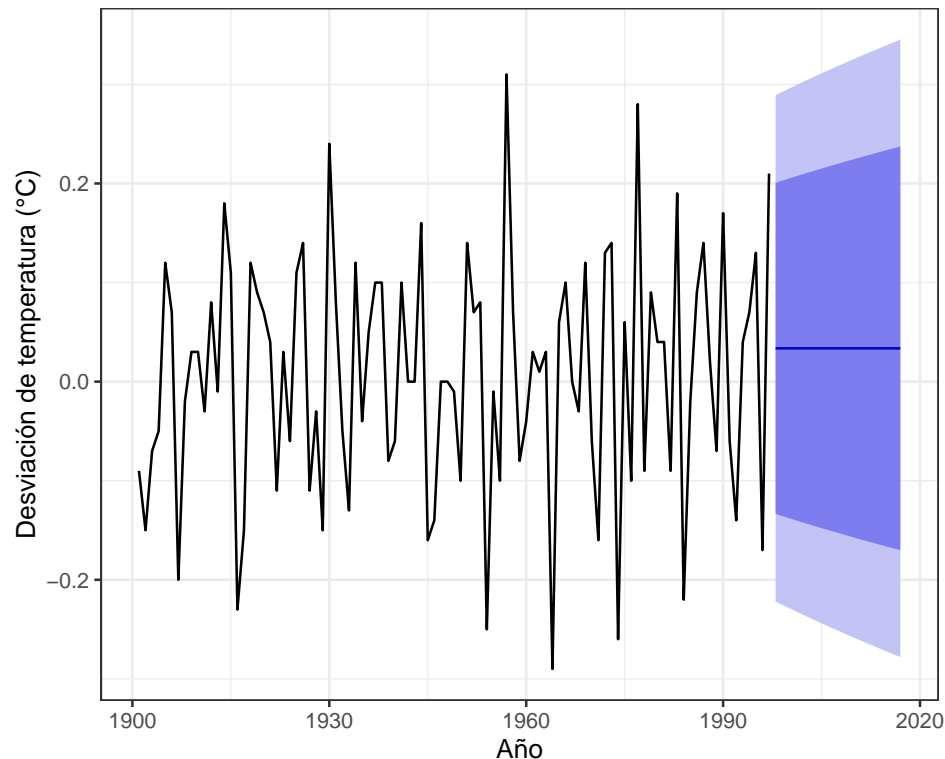


□

- c. Describa las bondades y limitaciones del modelo usado en los puntos anteriores.

Solución. El modelo de suavizamiento exponencial simple posee como ventaja ser lo suficientemente simple para realizar primeras inspecciones en la serie temporal. Ahora bien, esta misma ingenuidad denota limitaciones que pueden apreciarse al no proyectar tendencias. De hecho, en el gráfico del item anterior, se vislumbra una estimación plana proyectada hacia el futuro. Más aún, no reconoce ningún patron estacional. Estos dos últimas falencias pueden arreglarse usando un doble y triple suavizamiento respectivamente. En particular, para solucionar el reconocimiento de la tendencia, diferenciaremos la serie para eliminar la tendencia y aprovechar de mejor manera el predictor.

Predicción usando SES a la serie diferenciada

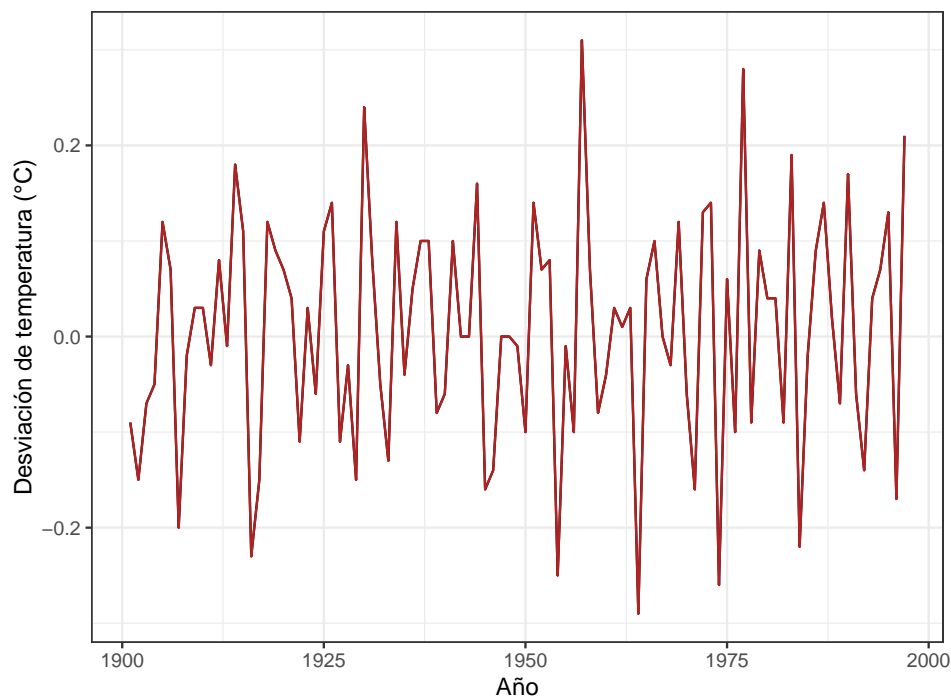


□

- d. A partir de la serie original obtenga una serie sin tendencia.

Solución. Usando el comando `diff` se obtiene lo pedido:

Serie sin tendencia



□

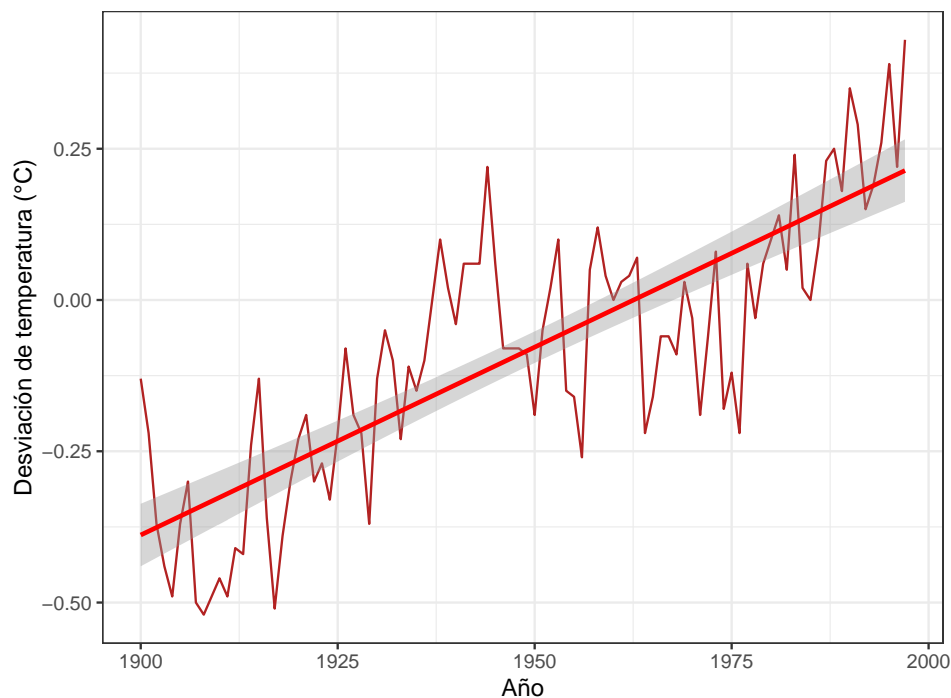
- e. Estime un modelo de regresión de la forma

$$Z_t = \beta_0 + \beta_1 t + \epsilon_t$$

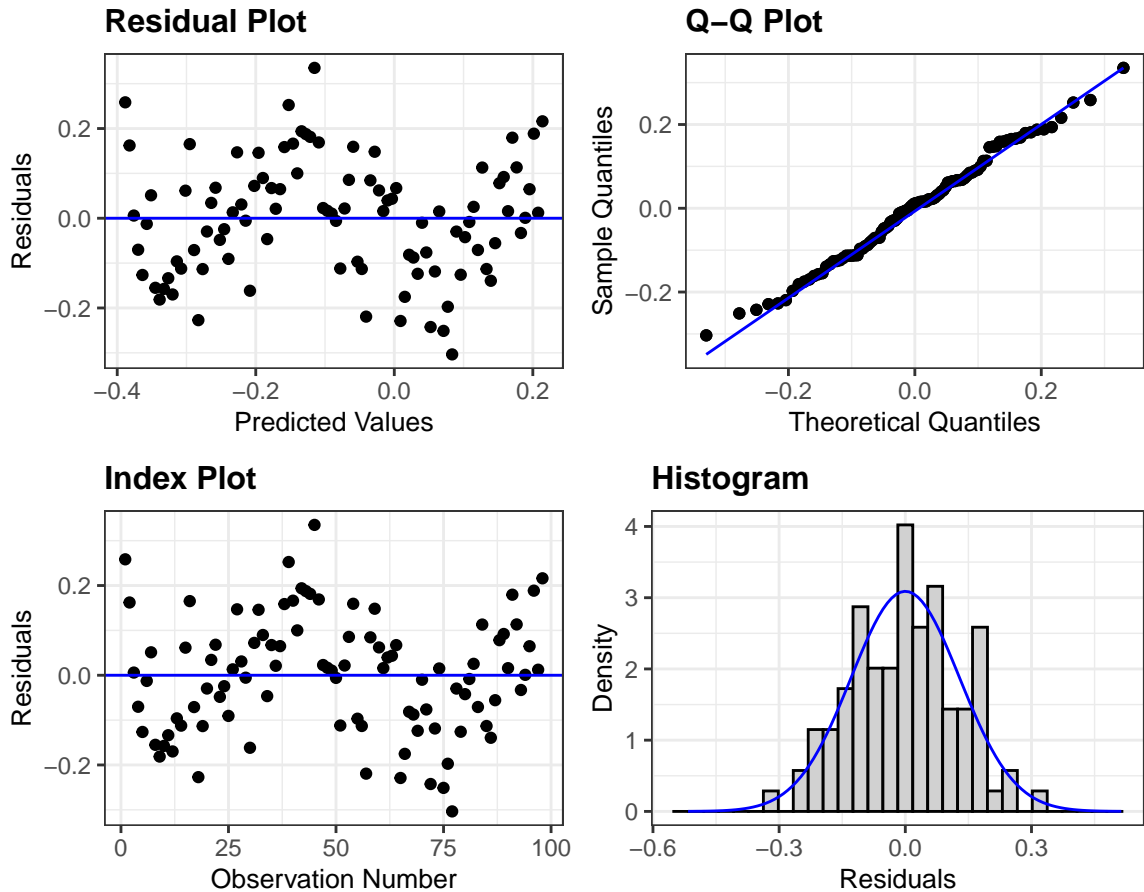
para la serie `globtemp.dat`, donde ϵ_t es una colección de variables aleatorias no correlacionadas con media cero y varianza σ^2 . ¿El modelo ajustado luce similar a la serie original?

Solución. Usando el comando `lm` se obtiene el modelo de regresión

Modelo de regresión lineal ajustado



donde $\beta_0 = -12,19$ y $\beta = 0,006$. Ahora bien, para responder la pregunta en cuestión, podemos analizar la colección ϵ_t a través de la función `resid_panel`, que nos proveerá de gráficos de diagnóstico de los residuos obtenidos.

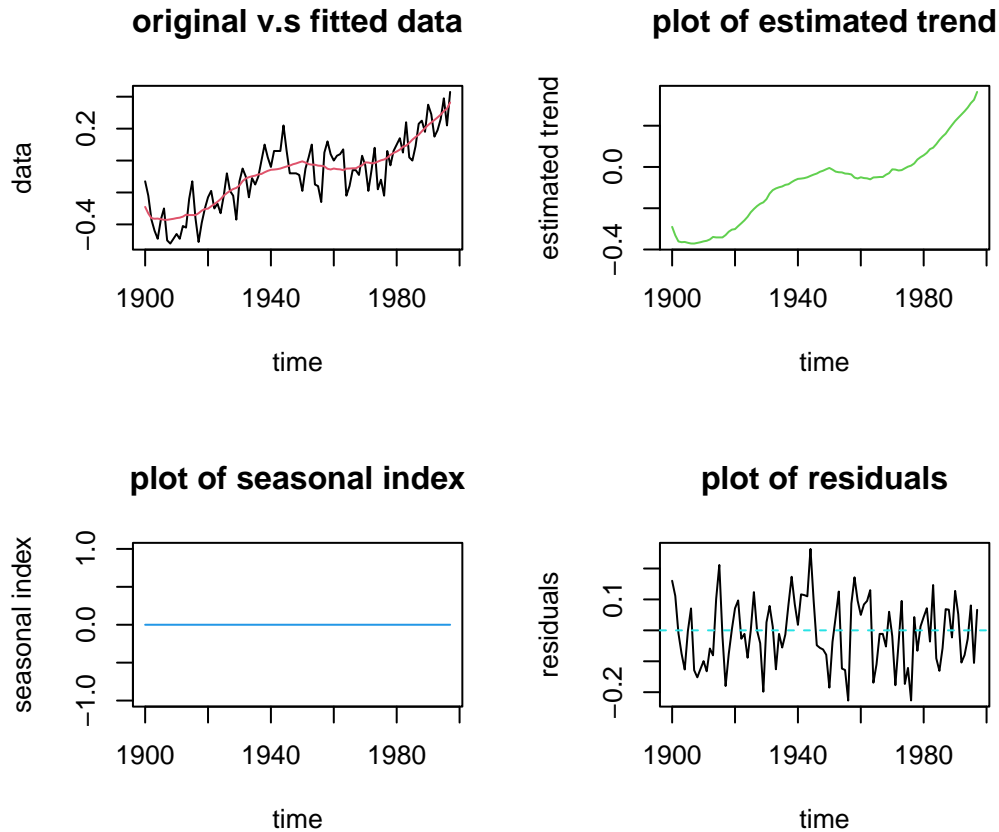


Inmediatamente notamos que la distribución de ϵ_t se asemeja a una normal con media cero, lo cual es de esperarse al tratarse de una regresión lineal. Ahora bien, cuando observamos el gráfico del modelo notamos que no luce similar. Más aún, el gráfico de los residuos (esquina superior izquierda) nos denota que no se ajusta bien la linealidad a los datos, por lo que será necesario considerar otro modelo para la serie temporal.

□

- f. Descomponga la serie `globtemp.dat` en tres partes: una tendencia, una parte estacional y una componente residual. Describa que observa.

Solución. Dada la forma del gráfico de la serie temporal, descompondremos considerando un modelo aditivo. Ahora bien, no podremos usar el comando `decompose`, pues los datos son anuales y tenemos solo una medición para cada año. Ahora bien, para solucionar esto, recurriremos al comando `ma.filter` que nos provee de la parte tendencial, estacional y residual para casos como este.



Observamos que la tendencia se ha clarificado como se esperaba. Además, el gráfico de la parte estacional es constante, dado que poseemos datos espaciados anualmente y por ende no nos aporta información relevante para el análisis.

Observación: para haber logrado una descomposición usando el comando `decompose` se debería haber forzado a considerar una frecuencia de dos veces por año, lo que alteraría la naturaleza de los datos. □

3. Al analizar cierta serie de tiempo trimestral se usó un método ingenuo obteniéndose:

† Ecuación de tendencia $T(t) = 84,65 + 4,71t$.

† Serie de residuos $W(t) = Y(t) - Z(t)$.

Serie de Residuos					
Trimestre	1997	1998	1999	2000	2001
1	-	20	15	13.75	8.75
2	-	-6.25	-7.50	-7.50	2.50
3	-11.25	-11.25	-21.50	-13.75	-
4	0.95	1.05	1.11	1.05	-

En base a los resultados de a) y b) dado que $t = 1$ corresponde al primer trimestre de 1997, prediga los valores de la serie en cada uno de los trimestres de 2002.

Solución. Dada la forma de T y W notamos que se ha usado un modelo de descomposición aditiva. Ahora bien, calculando e_h para el trimestre h se obtiene

$$e_1 = \frac{20 + 15 + 13,75 + 8,75}{4} = 14,38,$$

$$e_2 = \frac{-6,25 - 7,50 - 7,50 + 2,50}{4} = -4,69,$$

$$e_3 = \frac{-11,25 - 11,25 - 21,50 - 13,75}{4} = -14,44,$$

$$e_4 = \frac{0,95 + 1,05 + 1,11 + 1,05}{4} = 1,04.$$

Luego,

$$\bar{e} = \frac{1}{4} \sum_{h=1}^4 e_h = -0,93,$$

y por ende, las componentes estacionales son

$$\begin{aligned}\hat{E}_1 &= e_1 - \bar{e} = 15,31, \\ \hat{E}_2 &= e_2 - \bar{e} = -3,76, \\ \hat{E}_3 &= e_3 - \bar{e} = -13,51, \\ \hat{E}_4 &= e_4 - \bar{e} = 1,97.\end{aligned}$$

Finalmente, usando la ecuación de tendencia se obtienen las predicciones para los trimestres de 2002 como

$$\begin{aligned}\hat{Z}_{21} &= T(21) + \hat{E}_1 = 198,87, \\ \hat{Z}_{22} &= T(22) + \hat{E}_2 = 184,51, \\ \hat{Z}_{23} &= T(23) + \hat{E}_3 = 179,47, \\ \hat{Z}_{24} &= T(24) + \hat{E}_4 = 199,66.\end{aligned}$$

□

4. Considere una serie de tiempo $\{Z_t : t \in T\}$ descrita por la ecuación:

$$Z_t = \beta_0 + \beta_1 t + \beta_2 t^2 + S_t + \epsilon_t,$$

donde β_0 , β_1 y β_2 son parámetros desconocidos del modelo, S_t es un efecto estacional conocido y ϵ_t es un ruido aleatorio con media y varianza constante. Determine las ecuaciones que permiten estimar los parámetros β_0 , β_1 y β_2 .

Solución. Para obtener los parámetros se usará el método de mínimos cuadrados, para ello definiremos convenientemente $\tilde{\epsilon}_t = S_t + \epsilon_t$, dado que S_t es un conocido. Ahora bien

$$S(\beta_0, \beta_1, \beta_2) = \sum_{t \in T} (y_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \tilde{\epsilon}_t)^2.$$

Luego, diferenciando respecto a cada parámetro a estimar

$$\begin{aligned}\frac{\partial S}{\partial \beta_0} &= -2 \sum_{t \in T} (y_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \tilde{\epsilon}_t), \\ \frac{\partial S}{\partial \beta_1} &= -2 \sum_{t \in T} t(y_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \tilde{\epsilon}_t), \\ \frac{\partial S}{\partial \beta_2} &= -2 \sum_{t \in T} t^2(y_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \tilde{\epsilon}_t).\end{aligned}$$

Igualando a cero cada expresión se obtienen las siguientes ecuaciones

$$\begin{aligned}\sum_{t \in T} (y_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \tilde{\epsilon}_t) &= 0, \\ \sum_{t \in T} t(y_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \tilde{\epsilon}_t) &= 0, \\ \sum_{t \in T} t^2(y_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \tilde{\epsilon}_t) &= 0.\end{aligned}$$

□