CAPÍTULO 3

Modelo Aditivo de Componentes

Una serie de tiempo es una sucesión finita de variables aleatorias, ordenadas de acuerdo a una unidad de tiempo: $\{Y_1, \ldots, Y_T\}$. Por ejemplo, la concentración en el aire de dióxido de azufre SO_2 en ppm (ppm=partes por millón, $100ppm = 262mg/m^3$), medida semanalmente en una estación de monitoreo, es importante para vigilar la calidad del aire en una ciudad.

En este capítulo se introduce el Modelo Aditivo de Componentes y se analizan modelos específicos para las componentes de la tendencia y la estacionalidad, definidas a continuación.

3.1. El Modelo Aditivo de Componentes de series de tiempo

Dada una serie de tiempo $\{Y_t, t = 1, \ldots, T\}$, el Modelo Aditivo de Componentes consiste en asumir que Y_t se puede descomponer en tres componentes:

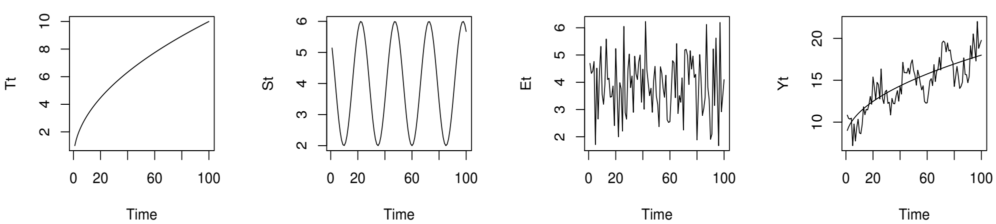
$$Y_t = T_t + S_t + \varepsilon_t, \tag{3.1}$$

donde T_t es la tendencia, S_t es la componente estacional y ε_t es la componente de errores. Las componentes T_t y S_t son funciones determinísticas de t. Su evolución es completamente predecible. El modelo (3.1) se denomina también Modelo Estructural Básico, ver Harvey [2006].

El modelo multiplicativo consiste en asumir que Y_t se puede descomponer en tres componentes:

$$Y_t = T_t S_t e^{\varepsilon_t}. \tag{3.2}$$

En lo que sigue solamente se tratará el modelo aditivo. En la Figura (3.1) siguiente se muestra la idea de la descomposición. Al superponer las series en los paneles (a), (b) y (c) se obtiene la serie en el panel (d).



(a) Tendencia (b) Patrón Estacional (c) Ruido (d) Serie

Figura 3.1: Descomposición de la Serie Y_t

El análisis consiste en modelar y estimar T_t y S_t y luego extraerlas de Y_t para obtener

$$\widehat{\varepsilon}_t = Y_t - \widehat{T}_t - \widehat{S}_t.$$

La serie $\hat{\varepsilon}_t$ se modela y estima para finalmente reconstruír Y_t, $\widehat{Y}_t = \widehat{T}_t + \widehat{S}_t + \widehat{\varepsilon}_t$, y poder realizar el pronóstico $\hat{Y}_{T+h} = \widehat{T}_{T+h} + \widehat{S}_{T+h} + \widehat{\varepsilon}_{T+h}$, utilizando la información disponible $Y_1, \ldots, Y_T$, con $h = 1, 2, \ldots, m$ un horizonte de pronóstico. Sin embargo, puede suceder que la serie $\widehat{\varepsilon}_t$ sea incorrelacionada, es decir, $Corr(\hat{\varepsilon}_t, \widehat{\varepsilon}_{t+s}) = 0$, para $s \neq 0$. En este caso $\widehat{\varepsilon}_{T+h} = 0, \forall h > 0$.

3.2. Modelos para la Tendencia

Definición 3.2.1 (Tendencia). *Se define como una función T_t de t que describe la evolución lenta y a largo plazo del nivel medio de la serie.*

La función T_t puede depender de parámetros, que deben estimarse. Pero también es posible utilizar modelos no-paramétricos, como medias móviles y regresiones polinómicas locales. Los modelos paramétricos se pueden denominar "modelos globales", en contraposición a los no-paramétricos, denominados "modelos locales".

Modelos: Una lista de posibles modelos paramétricos para T_t es:

$$\text{Lineal:} \quad T_t = \beta_0 + \beta_1 t, \tag{3.3a}$$

$$\text{Cuadrático:} \quad T_t = \beta_0 + \beta_1 t + \beta_2 t^2, \tag{3.3b}$$

$$\text{Cúbico:} \quad T_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3, \tag{3.3c}$$

$$\text{Exponencial-lineal:} \quad T_t = \exp(\beta_0 + \beta_1 t). \tag{3.3d}$$

$$\text{Exponencial-cuadrático :} \quad T_t = \exp(\beta_0 + \beta_1 t + \beta_2 t^2). \tag{3.3e}$$

$$\text{Logístico:} \quad T_t = \frac{\beta_2}{1 + \exp(\beta_0 + \beta_1 t)}, \tag{3.3f}$$

La forma de la tendencia depende de los rangos de los parámetros, por ejemplo, en la tendencia cuadrática (3.3b) se cumple:

1. Si $\beta_1, \beta_2 > 0$, T_t es monótona creciente.
2. Si $\beta_1, \beta_2 < 0$, T_t es monótona decreciente.
3. Si $\beta_1 > 0$ y $\beta_2 < 0$, T_t es cóncava hacia abajo (o cóncava).
4. Si $\beta_1 < 0$ y $\beta_2 > 0$, T_t es cóncava hacia arriba (o convexa).

La componente T_t en algunos casos puede ser estacional, es decir, periódica, de período largo. Por ejemplo, en una serie diaria, S_t es una función periódica de período 30 días, y T_t también periódica de período 360 días.

Ejemplo 3.2.1. *Un análisis de tendencia puede ser útil para obtener conclusiones sobre el efecto de políticas generales.*

Por ejemplo, los datos mensuales de la tasa de empleo en Australia entre 02/1978 y 08/2104 se pueden ver en la Figura 3.2. Este indicador se ajusta estacionalmente y se mide en términos de miles de personas de 15 años y más como porcentaje de la población en edad laboral. Un análisis de la tendencia requiere considerar el cambio de tendencia que parece darse desde finales de 2008.

Se pueden consultar datos sobre tasa de empleo en varios países de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), ver ([1]).

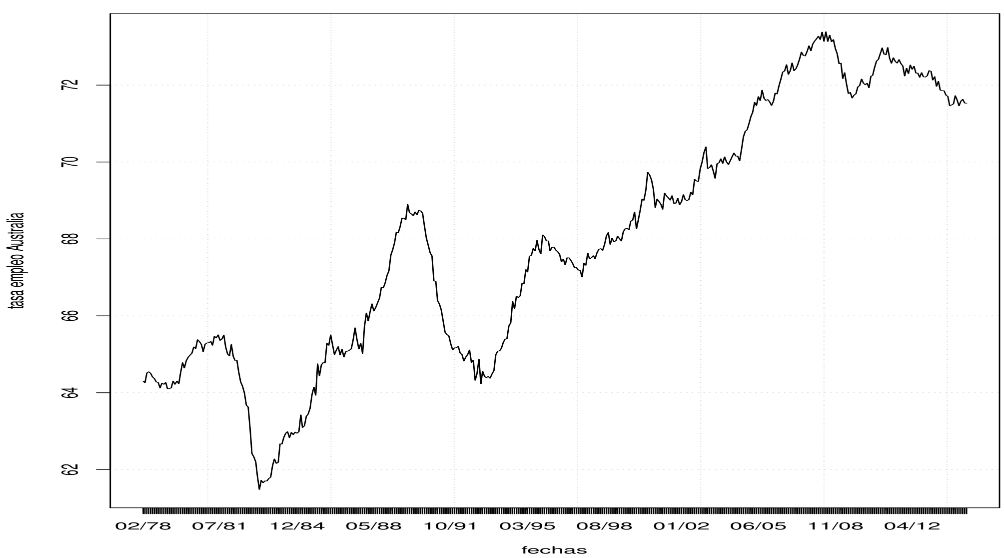


Figura 3.2: Tasa de Empleo, Australia, 02/1978 - 09/2014

Ejemplo 3.2.2. *Establecer la tendencia de una serie puede proporcionar pronósticos a corto y mediano plazo. Por ejemplo, en la Figura 3.3, donde se muestra la evolución del log-precio del kwh de energía eléctrica en el mercado MEN, un pronóstico puede ser la base para la negociación de contratos de futuros: compra a plazo a generadores.*

[1] https://data.oecd.org/emp/employment-rate.htm

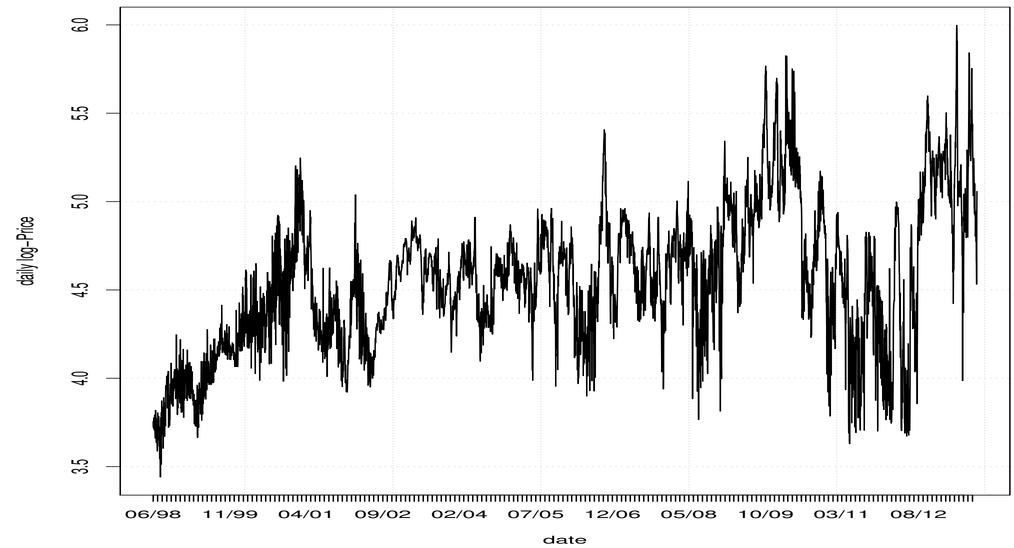


Figura 3.3: Logaritmo del precio del kwh en la bolsa energía Colombia

3.3. Estimación de la Tendencia

En este capítulo se introduce la estimación de la tendencia mediante modelos de regresión lineal y no lineal. En un capítulo posterior se introducen modelos no paramétricos. Son modelos que incluyen regresión polinómica local, filtros lineales y no lineales y medias móviles. Hay otros métodos que no se consideran en este curso, por ejemplo, wavelets (onditas). En ocasiones la estimación de la tendencia se denomina también "suavizar una serie" ó "extraer de la tendencia de una serie".

Para la estimación de los parámetros $\underline{\beta} = (\beta_0, \beta_1, \beta_2)'$ en los modelos lineales (3.3a), (3.3b) y (3.3c) se utiliza el método de mínimos cuadrados ordinarios (MCO). En este método el vector de parámetros estimados $\underline{\hat{\beta}}$ es el vector que produce el valor mínimo de la suma de errores cuadrados. Es decir $\underline{\hat{\beta}}$ es el valor en el cual $G(\underline{\beta}) = \sum_{t=1}^{T} \big(Y_t - T_t(\underline{\beta})\big)^2$ toma el valor mínimo.

$$\underline{\hat{\beta}} = \underset{\underline{\beta}}{\operatorname{argmin}}\, G(\underline{\beta}). \tag{3.4}$$

Para los modelos (3.3d) y (3.3f) se usa el método de mínimos cuadrados no lineales, que también minimiza la suma de errores cuadrados $G(\underline{\beta}) = \sum_{t=1}^{T} \big(Y_t - T_t(\underline{\beta})\big)^2$, pero $T_t(\underline{\beta})$ es una función no lineal de $\underline{\beta}$.

Aunque la serie tenga una componente estacional S_t, $Y_t = T_t + S_t + \varepsilon_t$, solamente consideramos un modelo de regresión entre Y_t y T_t, tal que $Y_t = T_t + \eta_t$, donde η_t es el término de error, de forma que $\eta_t = S_t + \varepsilon_t$. Por ejemplo.

1. En el caso lineal $T_t = \beta_0 + \beta_1 t$ con $\underline{\beta} = (\beta_0, \beta_1)'$, se ajusta el modelo de regresión lineal: $Y_t = \beta_0 + \beta_1 t + \eta_t$.
2. En el caso cuadrático $T_t = \beta_0 + \beta_1 t + \beta_2 t^2$ con $\underline{\beta} = (\beta_0, \beta_1, \beta_2)'$, se ajusta el modelo de regresión lineal $Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \eta_t$. Nótese que en este caso hay que definir la variable explicativa adicional t^2.
3. El modelo con tendencia exponencial-lineal (3.3d),

$$Y_t = \exp(\beta_0 + \beta_1 t) + \varepsilon_t$$

es no lineal y está relacionado, aunque no es equivalente, con un modelo con tendencia lineal para el logaritmo de Y_t, denominado modelo Logarítmico Lineal ó Log-Lineal, definido como

$$\ln Y_t = \beta_0 + \beta_1 t + \varepsilon_t. \tag{3.5}$$

Es posible estimar por mínimos cuadrados ordinarios el modelo Log-Lineal (3.5) y utilizar los parámetros estimados $\hat{\beta}_0, \hat{\beta}_1$ como valores iniciales en la estimación del modelo con tendencia exponencial por mínimos cuadrados no lineales. Pero los parámetros estimados en ambos modelos no necesariamente coinciden. Y los valores estimados $\widehat{Y}_t$ en el modelo exponencial-lineal (3.3d) no necesariamente coinciden con los valores estimados $\exp(\widehat{\ln Y_t})$ con (3.5).

3.4. Pronósticos con base en la Tendencia

Definición 3.4.1. ***(Pronósticos)****. Suponga la serie con tendencia*

$$Y_t = T_t + \eta_t, \ t = 1, \ldots, T,$$

con $(\eta_t, t \geq 1)$ *una sucesión* $iid(0, \sigma^2)$*. Los pronósticos de* Y_t *en los tiempos* $T + 1, T + 2, \ldots, T + h, \ h \geq 1$ *se definen como*

$$\widehat{Y}_{T+j} = \widehat{T}_{T+j}, j = 1, \ldots, h. \tag{3.6}$$

donde $\widehat{T}_t$ es la función estimada de la tendencia. Por ejemplo, en el modelo lineal

$$Y_t = \beta_0 + \beta_1 t + \eta_t,$$

$\widehat{T}_t = \hat{\beta}_0 + \hat{\beta}_1 t$. Al reemplazar t por $T+h$ se obtiene $\widehat{T}_{T+h} = \widehat{\beta}_0 + \widehat{\beta}_1(T+h) + \hat{\eta}_{T+h}$. Los pronósticos $\widehat{\eta}_{T+h}$, pueden o no ser cero. Son cero si los residuos $\widehat{\eta}_t,\ t = 1, \ldots, T$ son incorrelacionados. Para decidir esto se realiza una prueba de incorrelación y una prueba de normalidad. Por el contrario, puede suceder que los residuos estimados $\hat{\eta}_t$ sean autocorrelacionados y por tanto, diferentes de cero. En este caso el pronóstico es

$$\widehat{Y}_{T+j} = \widehat{T}_{T+j} + \hat{\eta}_{T+j}.$$

La definición general de pronóstico, para una serie $(Y_t, t \in \mathbb{Z})$, con base en la información $Y_1, \ldots, Y_T$, es una esperanza condicional,

$$\widehat{Y}_{T+j} = \mathbb{E}(Y_{T+j}|Y_1, \ldots, Y_T), j = 1, \ldots, h. \tag{3.7}$$

Otros tipos de pronósticos, en general, son

- El pronóstico por intervalo se obtiene si se asume que $\eta_t \sim iid\, N(0, \sigma^2)$. Entonces se cumple que $\widehat{Y}_{T+h} \pm 1.96\widehat{\sigma}$ es un intervalo de confianza del 95 %. Por ejemplo, en el modelo de tendencia lineal, el IC es

$$\widehat{\beta}_0 + \widehat{\beta}_1(T+h) \pm 1.96\widehat{\sigma}, \tag{3.8}$$

  donde $\widehat{\sigma}^2 = MSE = \frac{1}{T-2}\sum_{t=1}^{T} e_t^2$

- El pronóstico de densidad es $Y_{T+h} \sim N(\widehat{Y}_{T+h}, \widehat{\sigma}^2)$.

Los pronósticos se pueden calcular en R mediante la función `predict(objeto,...)` de la librería `forecast`. De la ayuda:

> "predict(objeto,...) es una función genérica para pronósticos a partir de los resultados de varias funciones de ajuste de modelos. La función invoca métodos particulares que dependen de la clase del primer argumento".

Por ejemplo, con un modelo de tendencia cuadrática $Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \eta_t$, para $t = 1, 2, \ldots, T$, si se genera un modelo de regresión lineal y se quiere calcular los pronósticos fuera de la muestra, en lo tiempos `T+1,...,T+m`, se programa

```
m2 = lm(y ~ t + t2}
tt = seq(T+1,T+m}
tt2 = tp*tp
y.pron = predict(m2,data.frame(t=tt,t2=tt2))
```

3.5. Medidas de precisión en los pronósticos

Definición 3.5.1. *El procedimiento de validación cruzada consiste en reservar los últimos m datos de una muestra para contrastarlos con los pronósticos realizados a partir del modelo estimado con los $T-m$ primeros datos, utilizando ciertas medidas de calidad de pronósticos.*

Asuma $Y_t, t = T-m+1, \ldots T$ y que se calcularon con estos datos el valor de los pronósticos $\widehat{Y}_t$ y los residuos estimados $\widehat{\eta}_t = Y_t - \widehat{Y}_t$. Se definen varias medidas para la precisión del pronóstico, por ejemplo

1. El error absoluto porcentual medio ó MAPE:

$$MAPE = \frac{1}{m} \sum_{t=T-m+1}^{T} \left| \frac{Y_t - \widehat{Y}_t}{Y_t} \right|. \tag{3.9}$$

2. La raíz cuadrada del error cuadrático medio, RMSE:

$$RMSE = \sqrt{\frac{1}{m} \sum_{t=T-m+1}^{T} (Y_t - \widehat{Y}_t)^2}. \tag{3.10}$$

3. El estadístico U de Theil,

$$U = \sqrt{\frac{\sum_{t=T-m+1}^{T} (Y_t - \widehat{Y}_t)^2}{\sum_{t=T-m+1}^{T} Y_t^2}} \tag{3.11}$$

> En Bliemel [1973] se argumenta que si $U \geq 1$ entonces el modelo de pronósticos es equivalente a los pronósticos "ingenuos", que consisten en pronosticar con el valor inmediatamente anterior en la serie. Sin embargo, en un blog reciente ([2]), R. Hyndman alega que no hay razón para que el U tenga que ser menor de uno porque es un cociente de valores medios cuadrados. Se excluyó el U de Theil de las versiones recientes de la función `accuracy`. Sin embargo, el U de Theil se puede calcular con la función `TheilU` de la librería `DescTools`.

Éstas y otras medidas se pueden calcular con la función `accuracy` de la librería `forecast`, con la instrucción `B=accuracy(y, yhat)`. Ver más detalles en Hyndman and Koehler [2006]. Devuelve un vector B, con componentes:

```
ME: Mean Error
RMSE: Root Mean Squared Error
MAE: Mean Absolute Error
MPE: Mean Percentage Error
MAPE: Mean Absolute Percentage Error
MASE: Mean Absolute Scaled Error
ACF1: Autocorrelation of errors at lag 1.
```

Ejemplo 3.5.1. *Suponga que se calcularon pronósticos para tres modelos: lineal, cuadrático y exponencial-lineal, que están como columnas en la matriz* `pr`*. El código siguiente calcula RMSE, MAE, MAPE, U-Theil. Note que se excluyeron ME y MPE porque producen valores negativos. En la Tabla 3.3 están los resultados.*

```
#----------Ejemplo código para calidad de pronósticos
R = rbind(accuracy(yf,pr[,1]),
accuracy(yf,pr[,2]),
accuracy(yf,pr[,3]))
rownames(R) = c("M.lin","M.cuad","M.exp")
require(DescTools)
Utheil=c(TheilU(yf,pr[,1], type=2),
TheilU(yf,pr[,2], type=2),
```

[2] https://stats.stackexchange.com/questions/8396/how-to-produce-theils-u-with-package-forecast-2-16-in-r

```
TheilU(yf,pr[,3], type=2))
R = cbind(R,Utheil)
R = R[,-c(1,4)]
(R)
```

3.5.1. Caso de Estudio: Pronóstico de Ventas al Menudeo

Este caso está analizado en Diebold [1999, sección 4.5]. El objetivo aquí es repetir el análisis utilizando R. Los modelos a utilizar son los modelos lineal, cuadrático, cúbico y exponencial. Se aplicarán los criterios AIC y BIC para escoger el más adecuado.

Descripción de los Datos

Ventas al menudeo en USD a precios del año 1999. Periodicidad: Mensual. Fechas: 01/1955 - 12/1994 Número de observaciones: 468. Datos ajustados estacionalmente. Es decir, si se tiene $Z_t = T_t + S_t + \varepsilon_t$, se dice que Y_t está ajustada estacionalmente o desestacionalizada si $Y_t = Z_t - S_t = T_t + \varepsilon_t$. Es decir, es una serie en la que se eliminó la componente estacional.

Validación cruzada. Usar el período 01/1955 - 12/1993 para estimar los modelos (período de entrenamiento) y el periodo 01/1994 - 12/1994 para examinar la eficiencia de los pronósticos fuera de la muestra.

Lectura de Datos, validación cruzada

```
D = read.table("ventas_al_menudeo.dat",header=T)
attach(D) # utiliza el nombre de las columnas como variables
# RTTR es el volumen de ventas en grandes almacenes en USD de 1999
# La variable RTRR del archivo tiene datos faltantes NA

y = na.omit(RTRR)/10000
# Convertir los datos en un objeto tipo ts
y = ts(y,frequency=12,start=c(1955,01))
```

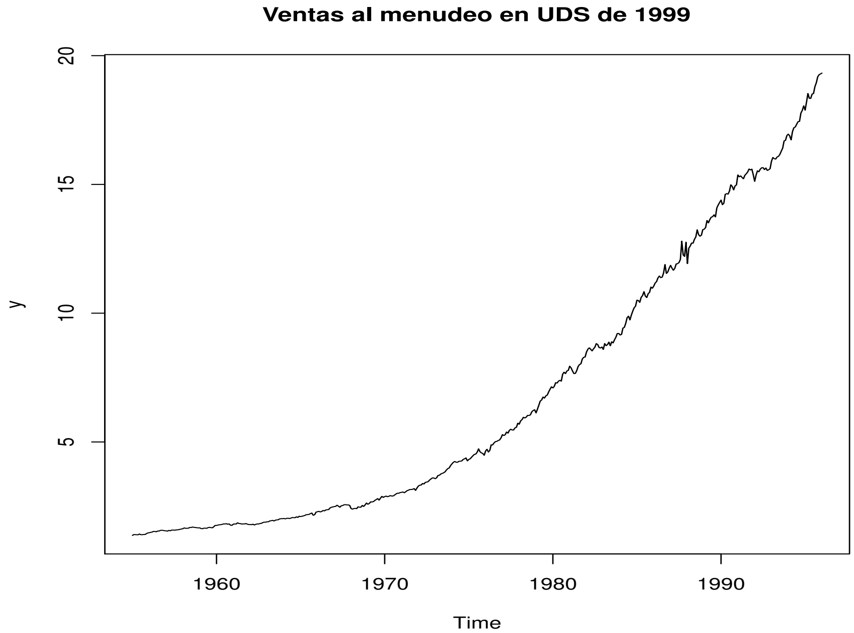


Figura 3.4: Ventas al menudeo, del texto Diebold(1999)

```
# Generar datos para validacion cruzada: dejar el
# ultimo año
T = length(y)
yi = y[1:(T-12)]
yf = y[(T-12+1):T]
```

Ajustar modelos:

```
# Ajustar 4 modelos: lineal, cuadrático, cúbico, log-lin
t = seq(1:(T-12))
t2 = t^2; t3 = t^3; lyi = log(yi)
# estimacion por minimos cuadrados ordinarios
mod.lin = lm(yi~t)
mod.cuad = lm(yi~t+t2)
mod.cub = lm(yi~t+t2+t3)
mod.llin = lm(lyi~t) # auxiliar para el exponencial
summary(mod.lin)
```

```
summary(mod.cuad)
summary(mod.cub)
```

Ajuste del Modelo Exponencial Lineal

```
# paso 1) estimar el modelo auxiliar log - linear
mod.llin = lm(lyi~t)
# paso 2) guardar los parametros del log-lineal
b0.est = mod.llin$coefficient[1]
b1.est = mod.llin$coefficient[2]
# paso 3) guardar los datos en un data.frame
Ds = data.frame(yi,t)
# paso 4) usar la funcion nls
mod.exp = nls(yi~exp(beta0+beta1*t),
data=Ds,start=list(beta0=b0.est, beta1=b1.est))
# paso 5) resultados
summary(mod.exp)
```

Resultados del ajuste de los Modelos

Nótese de los resultados en la Tabla 3.1 que el modelo cúbico no ajusta ya que el coeficiente de t^3 no dá significativo al nivel de $5\,\%$.

Calcular los Estadísticos de Selección del Modelo

Se procedió a calcular los estadísticos de calidad de ajuste, y los resultados está en la Tabla 3.2. Para calcular los estadísticos se usa la función `medidas()`, que los calcula para cada modelo.

```
M.lin = medidas(mod.lin,yi,2)
M.cuad = medidas(mod.cuad,yi,3)
M.cub = medidas(mod.cub,yi,4)
M.exp =  medidas(mod.exp,yi,2)
M = cbind(M.lin,M.cuad,M.cub,M.exp)
(M)
```

Tabla 3.1: Ajuste de los Modelos Lineal, Cuadrático, Cúbico y Exponencial

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-17537.673441	1503.384304	-11.665463	0.000000
t	340.738727	5.405165	63.039465	0.000000
(Intercept)	18886.034127	383.923338	49.192201	0.000000
t	-111.729690	3.678499	-30.373717	0.000000
t2	0.938731	0.007390	127.020988	0.000000
(Intercept)	19386.588312	513.151618	37.779455	0.000000
t	-124.127295	9.210396	-13.476868	0.000000
t2	1.002967	0.044379	22.600006	0.000000
t3	-0.000089	0.000061	-1.467890	0.142793
(Intercept)	9.390e+00	1.432e-02	655.8	0.000000
t	5.769e-03	3.536e-05	163.1	0.000000

	M.lin	M.cuad	M.cub	M.exp
R2-ajus	0.892	0.997	0.997	0.989
MSE	2.709	0.078	0.078	0.264
AIC	1.001	-2.543	-2.544	-1.327
BIC	1.018	-2.517	-2.509	-1.309

Tabla 3.2: Estadísticos de Selección de Modelos

De la Tabla 3.2 se puede concluír que el modelo que mejor ajusta los datos es el modelo cuadrático (3.3b) ya que presenta el menor BIC.

Cálculo de Pronósticos con los Modelos

```
# Pronósticos sin incluir el cubico
tt=seq((T-12+1),T,1)
tt2 = tt*tt
pr = mat.or.vec(length(tt),3)
pr[,1] = predict(mod.lin,data.frame(t=tt))
pr[,2] = predict(mod.cuad,data.frame(t=tt,t2=tt2))
```

```
pr[,3] = predict(mod.exp,data.frame(t=tt))
# graficas de los pronosticos
par(mfrow=c(1,1))
plot(tt,yf,type='b',lwd = 2,
ylim = c(10,22))
lines(tt,pr[,2],col='blue',lwd=2)
lines(tt,pr[,1],col='red')
lines(tt,pr[,3],col='magenta')
```

Los pronósticos están en las columnas de la matriz `pr`, generada en el programa anterior. A partir del examen de la Figura (3.5) se puede concluír que el modelo cuadrático genera pronósticos confiables a corto plazo.

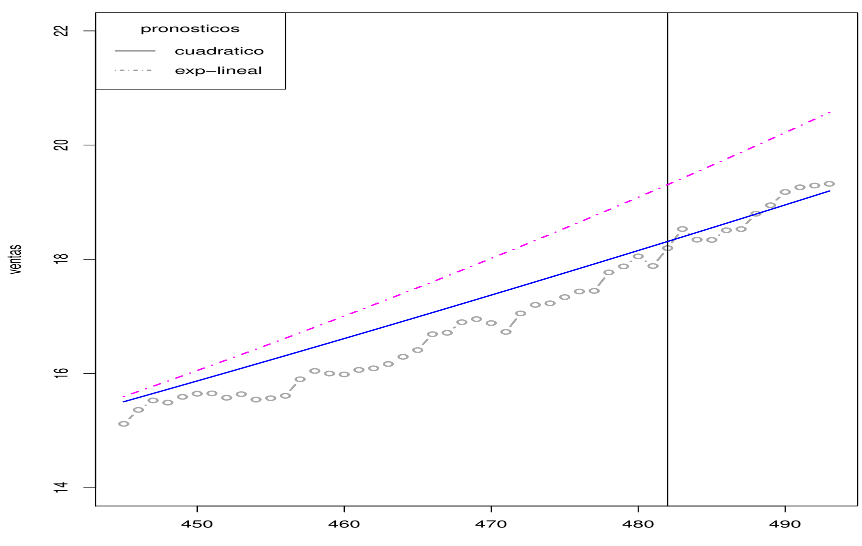


Figura 3.5: Pronósticos (—-) versus Observados (-o-o-)

Cálculo de estadísticos de calidad de pronósticos

```
# Estadísticos de calidad
R = rbind(accuracy(yf,pr[,1]),
accuracy(yf,pr[,2]),
```

```
accuracy(yf,pr[,3]))
rownames(R) = c("M.lin","M.cuad","M.exp")
require(DescTools)
Utheil=c(TheilU(yf,pr[,1], type=2),
TheilU(yf,pr[,2], type=2),
TheilU(yf,pr[,3], type=2))
R = cbind(R,Utheil)
R = R[,-c(1,4)] # excluye ME y MPE
(R)
```

Tabla 3.3: Estadísticos de calidad de pronósticos para tres modelos

	RMSE	MAE	MAPE	U-Theil
M.lin	3.924	3.913	26.325	0.209
M.cuad	0.157	0.145	0.771	0.008
M.exp	1.169	1.163	5.836	0.062

3.6. Modelos para la Componente Estacional

En esta sección se introducen modelos para la componente estacional S_t, en el modelo de componentes: $Y_t = T_t + S_t + \varepsilon_t$, con T_t la tendencia y ε_t la componente aleatoria. Además se exponen procedimientos de estimación de S_t con base en regresión lineal múltiple.

Una definición de Estacionalidad está en Hylleberg et al. [1990]:

> "La estacionalidad es el movimiento intra-anual, sistemático, aunque no necesariamente regular, causado por cambios en el clima, el calendario, y los momentos de las decisiones, hechas por los agentes de la economía, directa ó indirectamente, en sus decisiones de producción y consumo. Estas decisiones están influenciadas por las dotaciones, las expectativas y preferencias de los agentes, y las técnicas de producción disponibles en la economía".

Definición 3.6.1 (Componente Estacional). *La componente estacional S_t se define como una función periódica, de período $s = 2, 3, \ldots$ que describe la estacionalidad* *de una serie. El conjunto de valores de S_t, para $t = 1, \ldots, s$, se denomina el patrón estacional. El período estacional s es el número mínimo de períodos que tarda el patrón estacional en repertirse.*

Suponga que Y_t es una serie con componente estacional. La unidad de tiempo t de Y_t es día, semana, mes, trimestre, semestre, año. El período estacional de $s = 12$, significa que el patrón estacional es anual, consiste de doce meses, para una serie mensual. De manera similar con $s = 4$, serie trimestral, el patrón estacional es anual y consiste de cuatro meses.

Para observar un patrón estacional semanal es necesario definir la serie con frecuencia diaria. El patrón estacional se observa cada $s = 7$ días.

Una serie con frecuencia semanal se muestrea 52 veces al año. Se podría observa un patrón estacional por ejemplo, cada $s = 13$ semanas, que equivale, aproximadamente, a 4 trimestres, es decir, a un patrón estacional trimestral.

Para una serie con frecuencia diaria es posible definir un patrón estacional mensual colocando $s = 30$. También es posible un período de $s = 365$ días.

Ejemplo 3.6.1. *Una ejemplo de serie con componente estacional es la serie **nottem**, en la librería stats de R. Corresponde a la serie mensual "Average Monthly Temperatures at Nottingham, 1920-1939", que contiene las temperaturas promedio de grados Fahrenheit tomadas mensualmente en el Castillo de Nottingham (UK), por 20 años, entre 1920 y 1939, ver ([3]). Datos del libro de Anderson, O. D. (1976) Time Series Analysis and Forecasting: The Box-Jenkins Approach. La componente estacional de período $s = 12$ se observa claramente.*

[3] http://www.astrostatistics.psu.edu/datasets/R/html/datasets/html/00Index.html

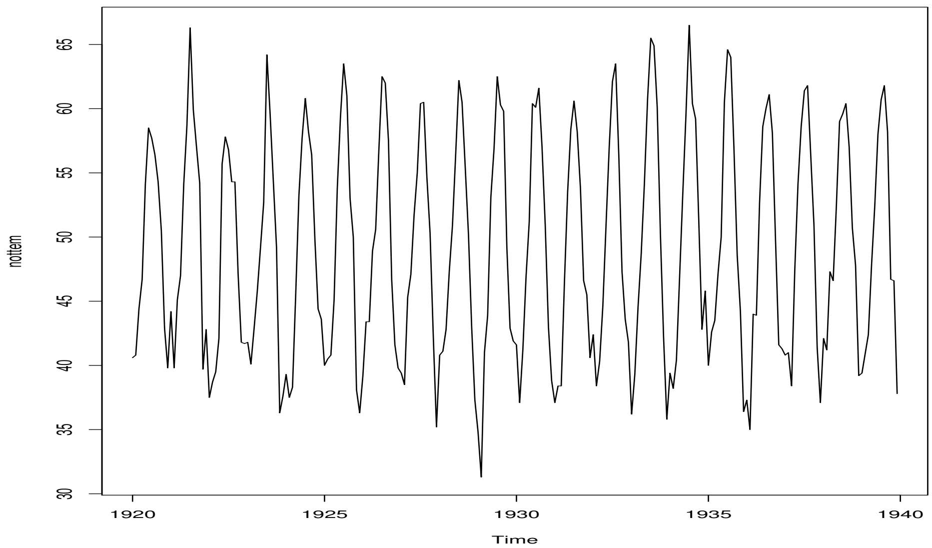


Figura 3.6: Average Monthly Temperatures at Nottingham, 1920-1939

Propiedades de $\mathrm{S_t}$

En todo lo que sigue $S_t,\ t = 1, 2, \ldots$ es una sucesión real, que cumple las propiedades siguientes. El período $s > 1$ es un entero positivo.

1. S_t es periódica con periodo s, es decir cumple: $S_{t+s} = S_t$ para $t = 1, 2, \ldots$. Por tanto, sólo se requiere definir S_t en los primeros s valores, $S_t,\ t = 1, 2, \ldots, s$. Es decir, basta con definir el patrón estacional.

2. Si $S1_t$ y $S2_t$ son funciones periódicas con periodo s entonces $aS1_t+bS2_t$, para $a, b \in \mathbb{R}$, es también una función periódica de período s. Luego, el conjunto de funciones periódicas de período s es un espacio vectorial de dimensión s.

3. Una base para este espacio vectorial está conformada por las s variables indicadoras estacionales, $I_j(t) \in \{0, 1\}$, para $j = 1, \ldots, s, t = 1, 2, \ldots$, definidas por
$$I_j(t) = \begin{cases} 1\,, & t = j, j+s, j+2s, \ldots \\ 0\,, & \text{en otro caso} \end{cases} \tag{3.12}$$

Por tanto, si S_t es periódica de período s, existen constantes $\delta_j, j = 1, \ldots, s$ tales que

$$S_t = \sum_{j=1}^{s} \delta_j I_j(t), \ t = 1, 2, \ldots. \tag{3.13}$$

La condición para $I_j(t) = 1$ se puede dar también de la forma

$$t \equiv j(\operatorname{mod} s) \Leftrightarrow t - j = s\lfloor t/s \rfloor. \tag{3.14}$$

donde $\lfloor x \rfloor$ es la parte entera de x, p.ej. $\lfloor 2.34 \rfloor = 2$. La expresión a la izquierda en (3.14) se lee " t congruente con j, módulo s", lo cual significa que $t - j$ es divisible por s, es decir que el cociente $\frac{t-j}{s}$ es un entero, que se expresa como $t - j = s\lfloor t/s \rfloor$.

Por ejemplo, con $s = 4$, es $I_3(11) = 1$?. Si $s = 4, j = 3, t = 11$, utilizando (3.14),

$$\begin{aligned} (t-j)/s &= \lfloor t/s \rfloor \\ (11-3)/4 &= \lfloor 11/4 \rfloor \Leftrightarrow 8/4 = \lfloor 2.75 \rfloor \Leftrightarrow 2 = 2. \end{aligned}$$

por tanto, $I_3(11) = 1$. Más simple, $11 - 3 = 8$ es divisible por 4, luego $I_3(11) = 1$. En cambio, $I_3(10) = 0$, que se verifica de igual manera.

4. También se puede utilizar el conjunto

$$\{sen(2\pi jt/s), cos(2\pi jt/s), j = 1, \ldots, s\}.$$

Para S_t periódica de período s, escogemos $m \geq 1$ y constantes $\beta_{1,j}, \beta_{2,j}, j = 1, \ldots, m$ tales que,

$$S_t \approx \sum_{j=1}^{m} \beta_{1,j} sin(2\pi jt/s) + \beta_{2,j} cos(2\pi jt/s),$$

para $t \geq 1$. El valor de m se escoge en los rangos siguientes

$$\text{rango } m = \begin{cases} 1 \leq m \leq \frac{s}{2} - 1, & \text{si } s \text{ es par} \\ 1 \leq m \leq \frac{s-1}{2}, & \text{si } s \text{ es impar} \end{cases} \tag{3.15}$$

Valores $m = 1, 2$ son recomendables por parsimonia.

5. Nótese que en el modelo con variables indicadoras el número de parámetros aumenta con el período s. Para períodos grandes, por ejemplo $s = 360$, el modelo (3.13) no sería práctico. Una alternativa sería un modelo con variables trigonométricas

$$\{sen(2\pi jt/360), cos(2\pi jt/360), j = 1, \ldots, m\}.$$

3.7. Modelo de componentes aditivas

Si se define un modelo para la componente de tendencia T_t, por ejemplo,

$$T_t = \sum_{j=0}^{k} \beta_j t^j, k = 1, 2, 3,$$

y el modelo para la componente estacional con variables indicadoras (3.13), entonces el modelo de componentes consiste en el modelo de regresión lineal múltiple para Y_t está dado por:

$$Y_t = \beta_0 + \sum_{j=1}^{k} \beta_j t^j + \sum_{j=1}^{s} \delta_j I_j(t) + \varepsilon_t, \tag{3.16}$$

donde $k = 1, 2, 3$, $\beta_i,\ i = 1, \ldots, k$ y $\delta_j,\ j = 1, \ldots, s$ son parámetros a estimar. Es decir, se puede escribir Y_t de la forma $Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i X_{i,t} + \varepsilon_t$, donde $X_{i,t}$ es de la forma t_i ó $I_i(t)$, y se pueden utilizar todos los recursos de regresión lineal múltiple para la estimación y pronóstico con este modelo.

A continuación se enumeran los pasos para estimar el modelo con tendencia y estacionalidad S_t con variables indicadoras, (3.16).

1. Identificar el período de la serie Y_t, s, si es que existe. Al declarar en R la serie como objeto “ts” con frecuencia, por ejemplo 12, se está identificado el posible período como $s = 12$. Sin embargo, el caso $s = 360$ no se incluye con indicadoras. En este caso sería preferible usar funciones trigonométricas, como se explica más adelante.

2. Generar las s variables indicadoras $I_j(t)$ para $t = 1, 2, \ldots, T$. Es decir, generar una matriz con entrada (t, j) dada por $I_j(t)$, donde al variar $j = 1, \ldots, s$ varían las columnas y al variar $t = 1, \ldots, T$ varían las filas.

3. Estimar los parámetros en el modelo (3.16) mediante regresión lineal. Nótese que se trata de estimar los parámetros de las componentes de tendencia y estacional, conjuntamente.

La función en (3.16)

$$T_t + S_t = \beta_0 + \sum_{j=1}^{k} \beta_j t^j + \sum_{j=1}^{s} \delta_j I_j(t) \tag{3.17}$$

se denominará “componente estructural” , y los residuales ε_t en (3.16) se denominarán “residuos estructurales” .

El Problema de la Multicolinealidad con Variables Indicadoras

Se deben incluír solamente $s-1$ variable indicadoras en el modelo en lugar del conjunto completo de s, para evitar problemas de multicolinealidad, ya que las variables indicadoras cumplen

$$\sum_{j=1}^{s} I_j(t) = 1, \forall t,$$

por lo que estarían perfectamente correlacionadas con el intercepto β_0, pues, aunque no se hace explícito, esta constante está asociada con una variable explicativa que es constante e igual a 1 (ver Diebold [1999, pág. 90]). En consecuencia, el modelo (3.16) se modifica y queda

$$Y_t = \beta_0 + \sum_{j=1}^{k} \beta_j t^j + \sum_{j=1}^{s-1} \delta_j I_j(t) + \varepsilon_t, \tag{3.18}$$

Implementación en R del modelo de componentes aditivas

Suponga que Y_t tiene una componente estacional de período s. Se declara como objeto `ts` en R, con frecuencia s con el comando `ts()` al incluír la opción, por ejemplo, `start=c(1995,2)`, que indica que la serie mensual inicia en febrero de 1995. Ejemplo: `y = ts(y,frequency=12, start=c(1995,2))`.

El primer paso consiste en generar las variables indicadoras $I_j(t), j = 1, \ldots, s, t = 1, \ldots, T$. Luego calcular una matriz de ceros y unos, por ejemplo, It, de dimensión $T \times (s-1)$, a partir de la información sobre la serie Y_t de frecuencia s.

Si el dato de Y_1 corresponde a la estación j entonces la primer fila de la matriz It será un vector de ceros excepto en la posición j, que corresponde a 1. De acuerdo a este valor inicial se genera la matriz It.

El siguiente código en R calcula la matriz $I_j(t)$ de dimensión $T \times s$.

```
#funcion para calcular la matriz It
#--------------------------
dummy=function(s,n,i){
# s: estaciones, n: numero de datos, i: estación de inicio
A=diag(s)
for(j in 1:(floor(n/s))){
A=rbind(A,diag(s))}
A=A[i:(n+i-1),]
return(A)}
```

En la librería `forecast` hay varios recursos y funciones para estimar y pronosticar con el modelo (3.18). Con la función `seasonaldummy` de la librería `forecast` se genera la matriz It. Adicionalmente, el comando `seasonaldummyf` genera otra matriz de ceros y unos, por ejemplo, Itp, con $s-1$ columnas y con un número de filas igual al número de pronósticos, a partir del primer tiempo de pronósticos.

Adicionalmente, en `forecast` está implementada la función `tslm` que calcula el modelo de componentes pero con la tendencia lineal solamente.

```
fit <- tslm(y ~ trend + season)
plot(forecast(fit, h=20))
```

Ejemplo 3.7.1. *Considere la serie de producción de cemento Portland, trimestral, en toneladas por mil, entre Q1 1956 y Q3 1994, en Australia. El modelo de componentes que se asume para esta serie es de la forma*

$$Y_t = \beta_0 + \beta_1 t + \sum_{j=1}^{4-1} \delta_j I_t(t) + \varepsilon_t. \tag{3.19}$$

Es decir, el modelo a estimar es lineal para la tendencia, más una componente estacional con base en variables indicadoras.

Los resultados del modelo lineal con variables estacionales (3.19) está en el Cuadro 3.4 siguiente. Una comparación entre la serie ajustada y la estimada se puede ver en la Figura (3.7). Con un R-cuadrado ajustado $\overline{R}^2 = 0.8966$, el ajuste resulta aceptable. También se observa que los coeficientes de las variables indicadoras ItQ2, ItQ3 no son significativos, sin embargo, ItQ1 sí, y por tanto, concluímos existe una componente estacional en la serie.

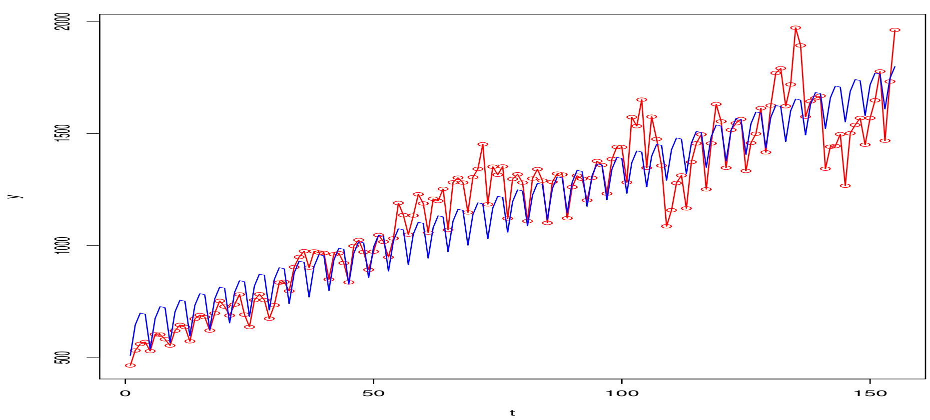


Figura 3.7: Ajuste con Variables Indicadoras y tendencia lineal. La componente estacional parece cambiar con el tiempo, es posiblemente un cambio estructural.

Tabla 3.4: Resultados del Ajuste del Modelo con Variables Estacionales

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	665.2149	24.1428	27.55	0.0000
t	7.2364	0.2023	35.78	0.0000
ItQ1	-163.0881	25.6796	-6.35	0.0000
ItQ2	-33.5553	25.6788	-1.31	0.1933
ItQ3	12.0287	25.6796	0.47	0.6402

Código R 3.7.1.

```
# -----Ejemplo con producción de cemento
#------------cargar datos
#  fpp2: Data for "Forecasting: Principles and Practice" (2nd Edition)
require(fpp2)
# Total quarterly production of Portland cement in Australia
# (in millions of tonnes) from 1956:Q1 to 2014:Q1.
y = qcement
# y = ts(y,frequency=4,start=c(1956,1),end=c(1994,3))
frequency(y)
[1] 4
#----modelo con tendencia lineal y estacionalidad
#----con variables indicadoras estacionales
library(forecast)
t = seq(1,length(y))
It = seasonaldummy(y)
mod1 = lm(y ~ t + It)
summary(mod1)
#----
library(xtable)
print(xtable(mod1),digits=4)
r1 = mod1$residuals
yhat1 = mod1$fitted.values
#------------------------------
```

Ejercicio 3.7.1. *Cambiar el modelo (3.19) por uno con tendencia cuadrática. Cómo se compara con el modelo (3.19) utilizando las medidas de ajuste (AIC, BIC, R2 ajustado, MSE), y las pruebas F parciales?.*

Ejemplo 3.7.2. *Continuando con el Ejemplo (3.7.1), ahora se estima el modelo (3.19), pero utilizando funciones trigonométricas. El procedimiento de estimación empieza con la generación de las variables explicativas $sen(2\pi it/s)$, $cos(2\pi it/s)$. Para calcular estas variables se utiliza la función* `It.trig = fourier(y,k)`, *la cual calcula los armónicos $sen(2\pi jt/s)$, $cos(2\pi jt/s)$, para cada $j = 1, 2, \ldots, k$.*

Nótese que al ser $s = 4$ para, entonces el número de armónicos es como máximo $s/2 - 1 = 1$.

$$Y_t = \beta_0 + \beta_1 t + \beta_1 sin(2\pi t/4) + \beta_2 cos(2\pi t/4) + \varepsilon_t. \tag{3.20}$$

Se incluye el cálculo de pronósticos a 8 trimestres con los modelos de variables indicadoras y de funciones trigonométricas. Los resultados están en la Tabla 3.5.

Código R 3.7.2.

```
#----modelo con tendencia lineal y estacionalidad
#----con variables trigonometricas
It.trig = fourier(y,1)
mod2 = lm(y ~ t + It.trig)
summary(mod2)
yhat2 = mod2$fitted.values
```

Tabla 3.5: Coeficientes estimados en el Modelo Lineal con variables trigonométricas

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6104	0.0184	33.25	0.0000
ti	0.0074	0.0002	34.43	0.0000
It.trigS1-4	-0.0825	0.0129	-6.41	0.0000
It.trigC1-4	0.0156	0.0130	1.21	0.2294
It.trigC2-4	0.0303	0.0091	3.32	0.0012

En la Tabla 3.6 aparecen los valores estimados dentro de la muestra $\widehat{Y}_t$, para los modelos lineal con indicadoras y lineal con trigonométricas. La conclusión que puede obtenerse es que en este caso es indiferente utilizar variables indicadoras ó funciones trigonométricas, dado que los pronósticos dentro de la muestra son muy similares.

Tabla 3.6: Seis primeros valores ajustados con variables Indicadoras y con Variables Trigonométricas

	lineal+indicadoras	lineal+trigonom
1	0.51	0.51
2	0.64	0.64
3	0.68	0.68
4	0.69	0.69
5	0.53	0.53
6	0.67	0.67

Ejemplo 3.7.3. *Un modelo para la demanda diaria de energía del mercado Nord Pool (países escandinavos) es el siguiente. La serie de precios es diaria para un período de 30/12/1996-26/04/2000.*

Esta serie diaria presenta dos componentes estacionales: una con período semanal y otra con período anual. Además se asume una tendencia lineal. Denotando la componente estacional semanal por s_t y la componente estacional anual por S_t, se escribe el modelo como sigue.

$$
\begin{aligned}
Y_t &= a + bt + s_t + S_t + e_t, \qquad (3.21)\\
s_t &= \sum_{j=1}^{6} \delta_j I_j(t),\\
S_t &= \alpha\, cos\left(\frac{2\pi t}{365}\right) + \beta \operatorname{sen}\left(\frac{2\pi t}{365}\right).
\end{aligned}
$$

En este modelo la componente con período anual se modela con funciones trigonométricas y la componente con período semanal con variables indicadoras. El siguiente código en R implementa el modelo (3.21).

Código R 3.7.3.

```
# Código en R --------------------
library(forecast)
D = read.table("serie.dat",header=T)
attach(D)
y = ts(y, frequency = 7)
n = length(y)
t = 1:n
St = seasonaldummy(y)
x1 = cos(2*pi*t/365)
x2 = sin(2*pi*t/365)
modelo1 = ln(y ~ t + St + x1 + x2)
summary(modelo1)
```

3.8. Problemas

1. Con base en los datos de la tasa de empleo en Australia, en la Figura 3.2, ajuste varios modelos de tendencia, dejando como período de validación cruzada los 3 últimos años (36 meses). Los datos se pueden descargar de la página web mencionada en el Ejemplo 3.2.1.

2. En el Cuadro 3.7 se muestran los datos del total de habitantes en Medellín, según los Censos desde 1905 hasta 2005, según el DANE ([4]). En Poveda

Tabla 3.7: Población de Medellín censos 1905 - 2005

	Año	Población
1	1905	59.810
2	1912	70.550
3	1918	79.150
4	1928	120.040
5	1938	168.270
6	1951	358.190
7	1964	772.890
8	1973	1.077.250
9	1985	1.468.090
10	1993	1.630.010
11	2005	2.223.080

[1982] se planteó ajustar los datos de desde 1938 hasta 1973 mediante un modelo de componentes con tendencia logística, (3.3f).

$$Y_t = \frac{\beta_2}{1 + \beta_1 \exp(-\beta_0 t)} + \varepsilon_t. \tag{3.22}$$

Realice el ajuste de la tendencia utilizando un modelo de tendencia logístico y pronostique a 10 años.

[4]http://es.wikipedia.org/wiki/Demografía_de_Medellín

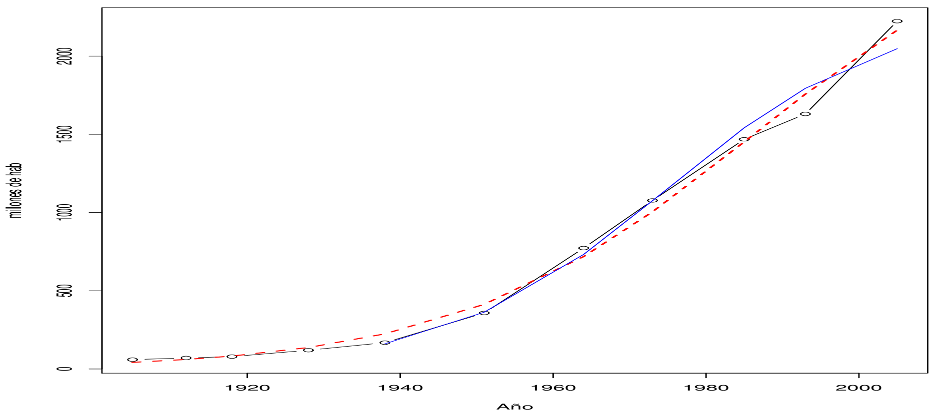


Figura 3.8: Ajuste Logístico a Población de Medellín

3. Consumo de agua mensual en el área metropolitana en metros cúbicos, por estrato, para los estratos 1,2,3,4,5,6, incluyendo el consumo total. Desde enero de 2002. En total son 7 series que se identificarán con los nombres : consumo 1, consumo 2,..., consumo 6, consumo total. Los datos están en el archivo consumo.dat, en la página del curso. Diseñe un plan inicial de análisis y descríbalo la introducción del trabajo. Por ejemplo, analizar las tendencias de todos los estratos, determinar si hay estacionalidad en todos o algunos.

    a) Lea los datos a partir del archivo asignado que se encuentra en la página web del curso. Grafique la serie. Desarrolle los siguientes puntos y reporte los resultados de cada uno. Con estos resultados elabore el reporte del trabajo.

    b) Use la función de descomposición stl para analizar lo siguiente:

        1) Hay una tendencia global que puede describirse por una función lineal, cuadrática ó cúbica?

        2) La tendencia se puede describir mejor por secciones? Es decir, no hay tendencia global?

        3) Se puede observar una componente estacional?. Cuál es el período?.

    c) Si observa que es factible que existan tendencia y estacionalidad, estímelas conjuntamente, con la restricción que se definión para la estacionalidad de considerar solamente $s - 1$ componentes estacionales (variables indicadoras), donde s es el período estacional. Reporte una tabla con los parámetros estimados, errores estándar, estadísticos t y valores p. Reporte el AIC y el R-cuadrado ajustado.

d) Si observa solamente tendencia ó estacionalidad, estímelas y reporte los estadísticos del punto anterior. En el caso de estacionalidad reporte el gráfico el patrón estacional, es decir, los coeficientes $\widehat{\delta}_i$ versus $i = 1, 2, \ldots, s$.

e) Reporte qqplot normal e histograma de los residuos, así como la prueba Jarque-Bera de normalidad. Comente sobre estos resultados.

f) Calcule pronósticos con base en el modelo estimado. Para esto escoja un horizonte de pronósticos adecuado a su serie.

g) Reporte las conclusiones del trabajo: un resumen de los resultados que encontró, y de los problemas que se presentaron en la elaboración del modelo. Por ejemplo, un comentario acerca de lo que Usted crea que logró el modelo: capturó la dinámica de la serie?, su tendencia?, los pronósticos parecen realistas y confiables?. Qué otras alternativas podrían haberse propuesto?. Cuál es una de las críticas al modelo de descomposición?.