Regresión

natetas

2025-03-28

5. Regresión

5.1 Propósito

Las tendencias en series de tiempo pueden clasificarse como *estocásticas* o *determinísticas*. Consideramos que una tendencia es estocástica cuando muestra cambios inexplicables en su dirección y atribuimos las tendencias transitorias aparentes a una alta correlación serial con error aleatorio. Este tipo de tendencias, comunes en las series financieras, pueden simularse en R utilizando modelos como el paseo aleatorio o el proceso autorregresivo (Capítulo 4). En contraste, cuando existe una explicación física plausible para una tendencia, usualmente buscamos modelarla de manera determinista. Por ejemplo, una tendencia determinista creciente en los datos puede estar relacionada con un aumento poblacional, o un ciclo regular puede estar asociado a una frecuencia estacional conocida. Las tendencias deterministas y la variación estacional pueden modelarse mediante regresión.

La diferencia práctica entre tendencias estocásticas y deterministas radica en que extrapolamos las segundas al hacer pronósticos. Justificamos la extrapolación a corto plazo al afirmar que las tendencias subyacentes generalmente cambian lentamente en comparación con el horizonte de pronóstico. Por la misma razón, la extrapolación a corto plazo debe basarse en una línea ajustada a los datos más recientes en lugar de un polinomio de alto orden.

En este capítulo, se estudian modelos de regresión adecuados para el análisis de series de tiempo que contienen tendencias deterministas y variación estacional regular. Comenzamos con modelos lineales para tendencias y luego consideramos modelos que incorporan variables indicadoras y armónicas para capturar tendencias y variaciones estacionales. También se exploran modelos de regresión con variables explicativas y la transformación de Box-Cox, que se utiliza para estabilizar la varianza.

El análisis de regresión en series de tiempo suele diferir del análisis estándar de regresión debido a que los errores tienden a estar correlacionados en el tiempo. Esto puede hacer que los errores estándar de los coeficientes estén subestimados y que los valores p sean más pequeños de lo que deberían, lo que puede llevar a atribuir significancia estadística excesiva a las pruebas en el software estadístico estándar. Es crucial presentar evidencia estadística correcta. Por ejemplo, un grupo de protección ambiental podría verse afectado por acusaciones de que está afirmando falsamente tendencias estadísticamente significativas. En este capítulo, se emplean los mínimos cuadrados generalizados para obtener estimaciones mejoradas del error estándar y corregir la autocorrelación en la serie residual.

5.2 Modelos lineales

5.2.1 Definición

Un modelo para una serie de tiempo $\{x_t : t = 1, \ldots, n\}$ es *lineal* si se puede expresar como:

$$x_t = \alpha_0 + \alpha_1 u_{1,t} + \alpha_2 u_{2,t} + \cdots + \alpha_m u_{m,t} + z_t$$

donde $u_{i,t}$ es el valor del predictor i-ésimo (o variable explicativa) en el tiempo t y z_t es el error en el tiempo t. Los parámetros del modelo, $\alpha_0, \alpha_1, \ldots, \alpha_m$, pueden estimarse mediante mínimos cuadrados. Nótese que los errores z_t pueden tener media 0 pero no necesitan ser ruido blanco gaussiano.

Un ejemplo de un modelo lineal es el polinomio de orden p de la forma:

$$x_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \cdots + \alpha_p t^p + z_t$$

Las variables predictoras pueden escribirse como $u_{i,t} = t^i$ para $i = 1, \ldots, p$. El término *lineal* hace referencia a la suma de los parámetros del modelo, cada uno multiplicado por una variable predictora única.

Un caso especial de un modelo lineal es la tendencia lineal simple obtenida al poner $p = 1$ en la ecuación anterior:

$$x_t = \alpha_0 + \alpha_1 t + z_t.$$

En este caso, el valor de la línea en el tiempo t representa la tendencia m_t. Para el polinomio general, la tendencia en t es simplemente el valor del polinomio subyacente evaluado en t, es decir:

$$m_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \cdots + \alpha_p t^p.$$

Muchos modelos no lineales pueden transformarse en modelos lineales. Por ejemplo, un modelo de la forma $x_t = e^{\alpha_0 + \alpha_1 t + z_t}$ puede transformarse tomando logaritmos naturales para obtener un modelo lineal para la serie y_t:

$$y_t = \log x_t = \alpha_0 + \alpha_1 t + z_t.$$

En la ecuación anterior, la regresión por mínimos cuadrados podría utilizarse para ajustar el modelo, es decir, estimar los parámetros α_0 y α_1, y hacer predicciones para x_t. Para obtener una predicción de x_t, la transformación inversa necesita aplicarse (es decir, la exponencial). Sin embargo, esto suele afectar los valores esperados de x_t, por lo que se utilizan factores de corrección para obtener mejores resultados.

Los modelos de series de tiempo reales no suelen ajustarse exactamente a los modelos lineales. Un ejemplo de modelo no lineal es el modelo de Bass (Sección 5.3), que se ajusta utilizando la función de mínimos cuadrados no lineales `nls` en R.

5.2.2 Estacionariedad

Los modelos lineales para series de tiempo no son estacionarios cuando incluyen funciones del tiempo. La diferenciación a menudo puede transformar una serie no estacionaria con una tendencia determinista en una serie estacionaria. Por ejemplo, si la serie de tiempo $\{x_t\}$ está dada por una función lineal más un ruido blanco $x_t = \alpha_0 + \alpha_1 t + z_t$, entonces las diferencias de primer orden están dadas por:

$$\nabla x_t = x_t - x_{t-1} = z_t - z_{t-1} + \alpha_1$$

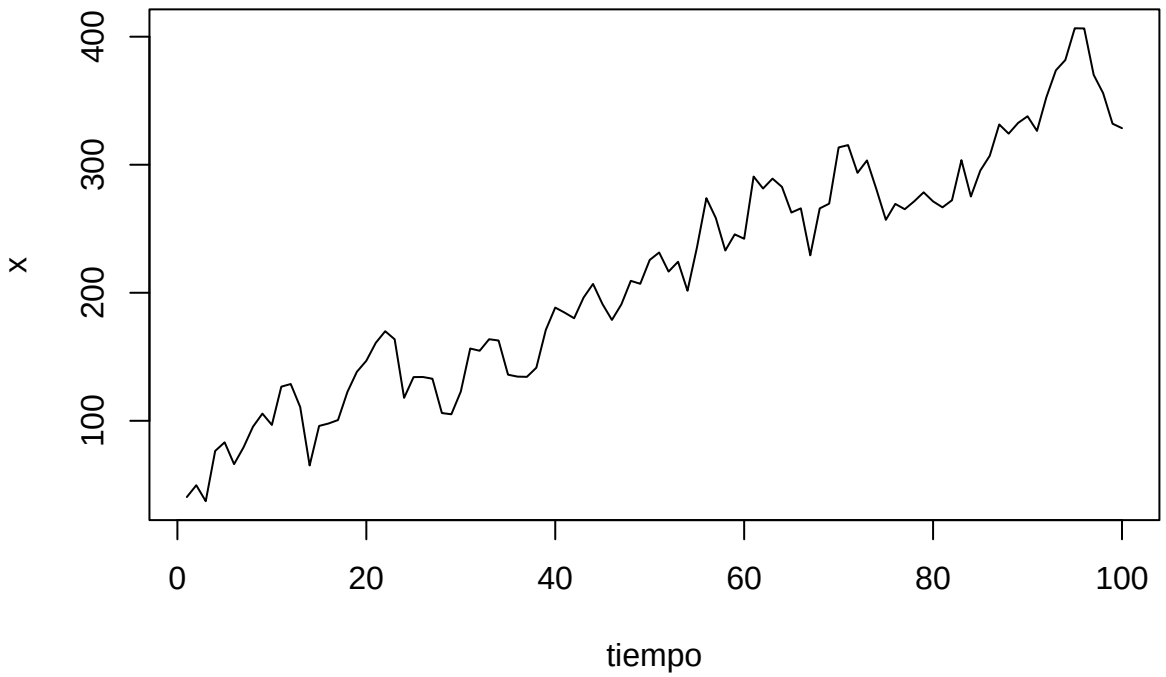
Suponiendo que la serie de errores $\{z_t\}$ es estacionaria, la serie $\{\nabla x_t\}$ también será estacionaria, ya que no es función de t. En la Sección 4.3.6 encontramos que la diferenciación de primer orden puede transformar una serie no estacionaria con tendencia estocástica (como un paseo aleatorio) en una serie estacionaria. Por lo tanto, la diferenciación puede eliminar tanto tendencias estocásticas como deterministas en las series de tiempo. Si la tendencia subyacente es un polinomio de orden m, se requiere diferenciación de orden m para eliminar la tendencia.

Nótese que diferenciar una función lineal más ruido blanco produce una serie estacionaria diferente que simplemente restar la tendencia. La segunda opción genera ruido blanco, mientras que la diferenciación genera una serie de términos de ruido blanco consecutivos (un proceso MA, descrito en el Capítulo 6).

5.2.3 Simulación

En la regresión de series de tiempo, es común que la serie de errores $\{z_t\}$ en la Ecuación (5.1) esté autocorrelacionada. En el siguiente código se simula y grafica una serie de tiempo con una tendencia lineal creciente $(50 + 3t)$ con errores autocorrelacionados:

```
set.seed(1)
z <- w <- rnorm(100, sd = 20)
for (t in 2:100) z[t] <- 0.8 * z[t - 1] + w[t]
Time <- 1:100
x <- 50 + 3 * Time + z
plot(x, xlab = "tiempo", type = "l")
```



El modelo correspondiente al código anterior se puede expresar como:

$$x_t = 50 + 3t + z_t,$$

donde $\{z_t\}$ sigue un proceso AR(1):

$$z_t = 0.8z_{t-1} + w_t,$$

y $\{w_t\}$ es un ruido blanco gaussiano con desviación estándar $\sigma = 20$. Una gráfica temporal de una realización de $\{x_t\}$ se muestra en la Figura 5.1.

5.3 Modelos ajustados

5.3.1 Modelo ajustado a datos simulados

Los modelos lineales suelen ajustarse minimizando la suma de los errores cuadrados,

$$\sum z_t^2 = \sum (x_t - \alpha_0 - \alpha_1 u_{1,t} - \cdots - \alpha_m u_{m,t})^2,$$

lo cual se logra en R utilizando la función `lm`:

```
x.lm <- lm(x ~ Time)
coef(x.lm)
```

```
## (Intercept)        Time
##   58.551218    3.063275
```

```
sqrt(diag(vcov(x.lm)))
```

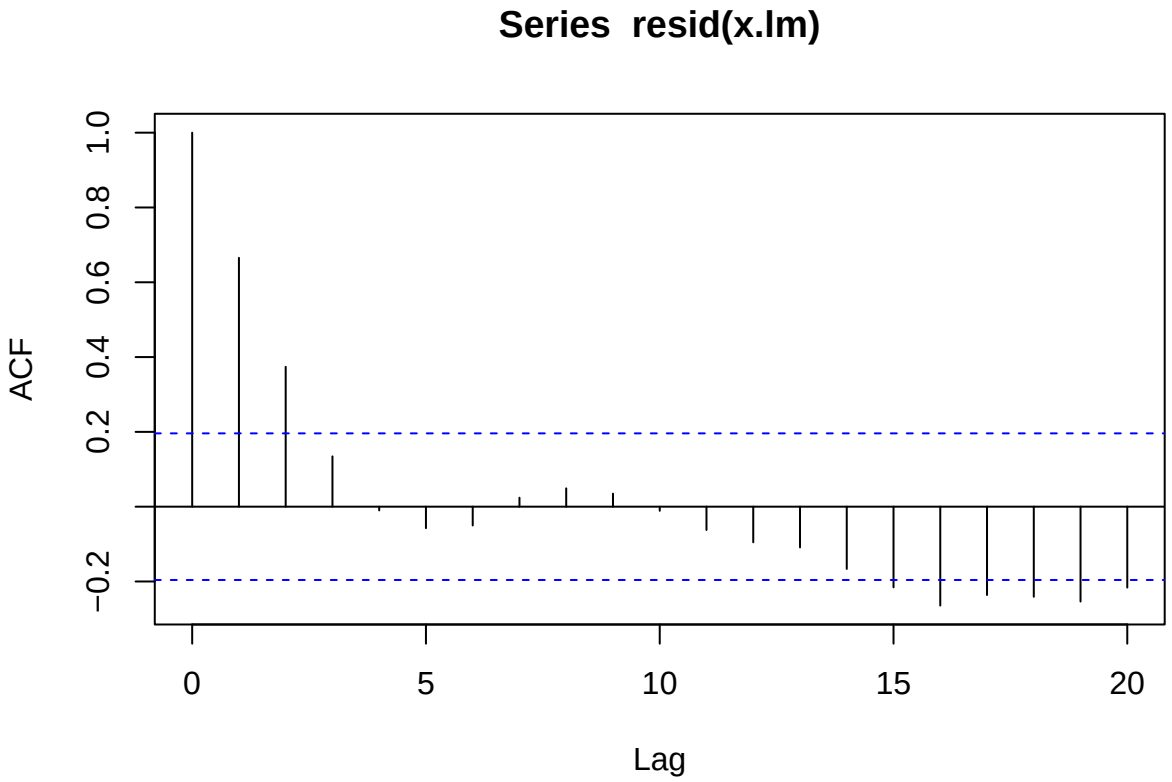
```
## (Intercept)        Time
##  4.88006278  0.08389621
```

En el código anterior, los parámetros estimados del modelo lineal se extraen usando `coef`. Como era de esperarse, las estimaciones están cercanas a los valores subyacentes reales de 50 para la intersección y 3 para la pendiente. Los errores estándar se extraen utilizando la raíz cuadrada de los elementos diagonales obtenidos de `vcov`. Sin embargo, estos errores estándar tienden a estar subestimados debido a la autocorrelación en los residuos.

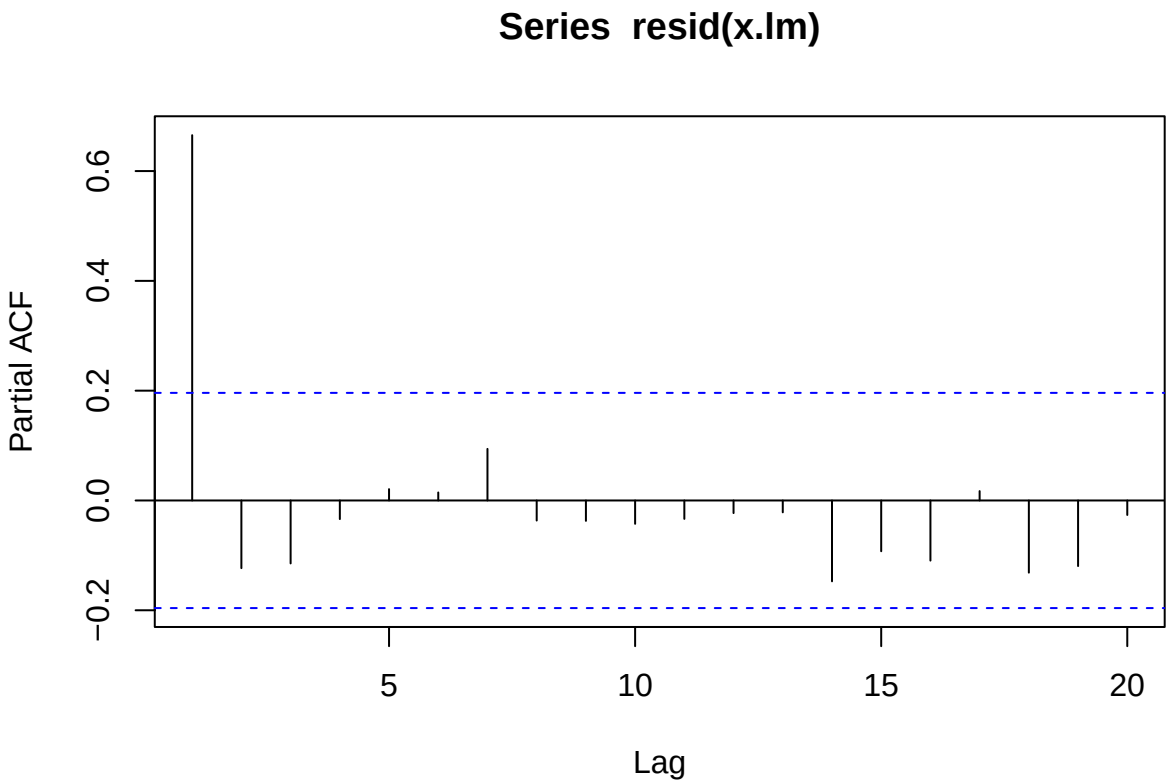
La función `summary` también puede usarse para obtener esta información, pero tiende a proporcionar detalles adicionales, como pruebas t, que pueden no ser adecuadas para un análisis de regresión en series de tiempo debido a la autocorrelación en los residuos.

Después de ajustar un modelo de regresión, es importante considerar diversos gráficos de diagnóstico. En el caso de la regresión de series de tiempo, un gráfico de diagnóstico clave es el **correlograma de los residuos**:

```
acf(resid(x.lm))
```




```
pacf(resid(x.lm))
```



Como era de esperarse, la serie de residuos está autocorrelacionada (Figura 5.2). En la Figura 5.3, solo la autocorrelación parcial en el rezago 1 es significativa, lo que sugiere que la serie de residuos sigue un proceso AR(1). Esto es consistente con la simulación, ya que se usó un proceso AR(1) para generar estos residuos.

5.3.2 Modelo ajustado a la serie de temperatura (1970–2005)

En la Sección 1.4.5, se extrajeron datos de temperatura para el período 1970–2005. El siguiente modelo de regresión se ajusta a la temperatura global durante este período, y se presentan intervalos de confianza aproximados al 95% para los parámetros utilizando `confint`. La variable explicativa es el tiempo, por lo que la función `time` se usa para extraer los valores de tiempo del objeto de serie de temperatura.

```
# Realizado con el método del libro
Global <- read.table("global.dat",header=T)
Global.ts <- ts(Global, st = c(1856, 1), end = c(2005, 12), fr = 12)
temp <- window(Global.ts, start = 1970)
temp.lm <- lm(temp ~ time(temp))
coef(temp.lm)
```

```
##                  X.0.384      X.0.457      X.0.673      X.0.344      X.0.311
## (Intercept) -8.922734243 -9.167606412 -9.538610925 -9.043841763 -7.811130290
## time(temp)   0.004421211  0.004540247  0.004706661  0.004479054  0.003859277
##                  X.0.071      X.0.246      X.0.235      X.0.380      X.0.418
## (Intercept) -8.00438799 -7.491607892 -7.323539581 -6.956392716 -6.977911795
## time(temp)   0.00397746  0.003723689  0.003637966  0.003450994  0.003449414
##                  X.0.670      X.0.386
## (Intercept) -7.185270934 -7.750298250
## time(temp)   0.003524615  0.003816307
```

```
confint(temp.lm)
```

```
##                              2.5 %       97.5 %
## X.0.384:(Intercept) -1.447858e+01 -3.366893176
## X.0.384:time(temp)   1.626502e-03  0.007215920
## X.0.457:(Intercept) -1.507256e+01 -3.262651383
## X.0.457:time(temp)   1.569926e-03  0.007510568
## X.0.673:(Intercept) -1.501023e+01 -4.066990581
## X.0.673:time(temp)   1.954317e-03  0.007459005
## X.0.344:(Intercept) -1.389175e+01 -4.195932766
## X.0.344:time(temp)   2.040450e-03  0.006917658
## X.0.311:(Intercept) -1.246514e+01 -3.157119439
## X.0.311:time(temp)   1.518208e-03  0.006200346
## X.0.071:(Intercept) -1.213977e+01 -3.869010233
## X.0.071:time(temp)   1.897275e-03  0.006057645
## X.0.246:(Intercept) -1.156977e+01 -3.413444823
## X.0.246:time(temp)   1.672284e-03  0.005775094
## X.0.235:(Intercept) -1.148437e+01 -3.162709673
## X.0.235:time(temp)   1.544978e-03  0.005730954
## X.0.380:(Intercept) -1.102992e+01 -2.882868174
## X.0.380:time(temp)   1.401922e-03  0.005500065
## X.0.418:(Intercept) -1.164231e+01 -2.313516452
## X.0.418:time(temp)   1.103122e-03  0.005795707
## X.0.670:(Intercept) -1.243652e+01 -1.934022917
## X.0.670:time(temp)   8.831226e-04  0.006166107
## X.0.386:(Intercept) -1.304717e+01 -2.453425060
## X.0.386:time(temp)   1.151864e-03  0.006480750
```

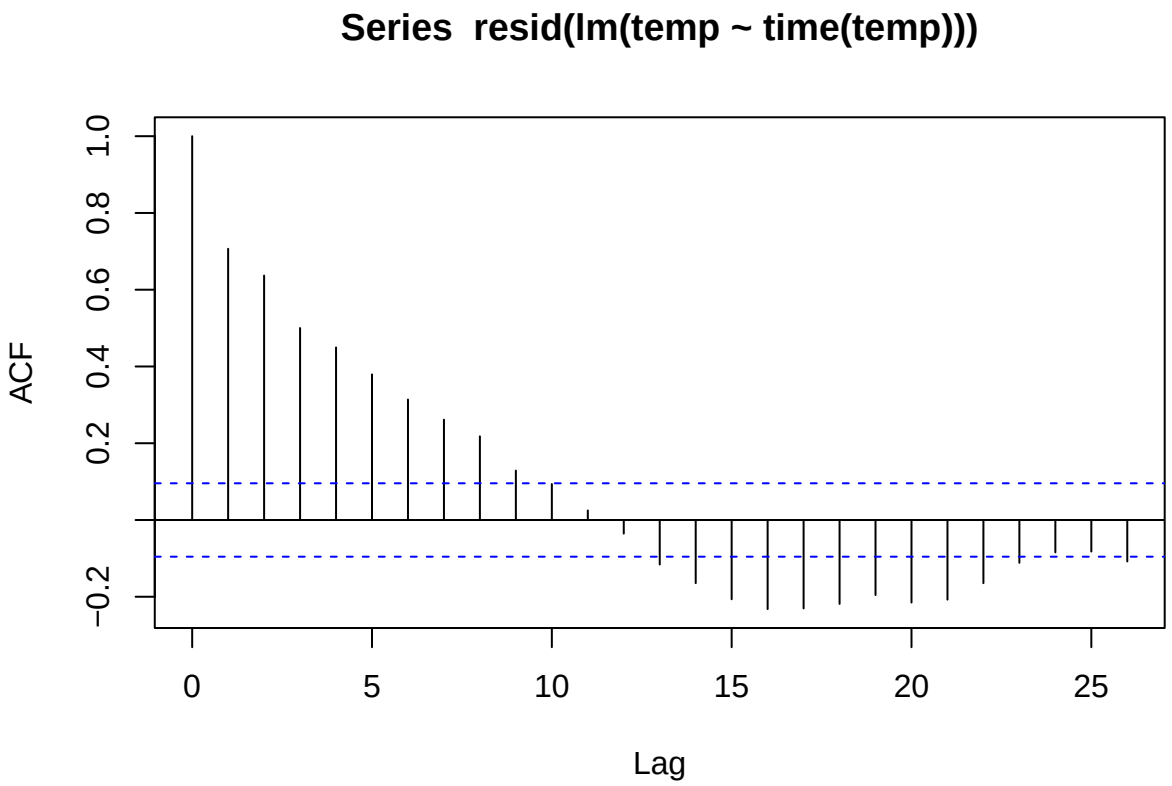
```r
# Realizado con el método visto en clase
Global <- read.table("global.dat",header=T)
aux2 = as.numeric(Global[1,])
for(i in 2:149)
{
    aux1=as.numeric(Global[i,])
    aux2=c(aux2,aux1)
}
Global.ts <- ts(aux2,start=c(1856,1),end=c(2004,12),frequency=12)
temp <- window(Global.ts, start = 1970)
temp.lm <- lm(temp ~ time(temp))
coef(temp.lm)
```

```
##  (Intercept)   time(temp)
## -35.93428900   0.01817157
```

```r
confint(temp.lm)
```

```
##                     2.5 %       97.5 %
## (Intercept) -38.31222357 -33.55635443
## time(temp)    0.01697511   0.01936802
```

```
acf(resid(lm(temp ~ time(temp))))
```



El intervalo de confianza para la pendiente **no contiene el cero**, lo que proporciona evidencia estadística de una tendencia creciente en las temperaturas globales si la autocorrelación en los residuos es despreciable. Sin embargo, la serie de residuos muestra autocorrelación positiva en rezagos cortos (Figura 5.4), lo que lleva a una subestimación del error estándar y un intervalo de confianza demasiado estrecho para la pendiente.

Intuitivamente, la correlación positiva entre valores consecutivos reduce la longitud efectiva del registro porque valores similares tienden a ocurrir juntos. La siguiente sección ilustra esta problemática, aunque puede omitirse para los lectores que no requieran detalles matemáticos.

5.4 Mínimos Cuadrados Generalizados

Hemos visto que en la regresión de series de tiempo es común y esperado que la serie de residuos esté autocorrelacionada. Si existe correlación serial positiva en los residuos, esto implica que los errores estándar de los parámetros de regresión estimados probablemente estén subestimados (Ecuación 5.5), y por lo tanto, deben corregirse.

Un procedimiento de ajuste conocido como **mínimos cuadrados generalizados** (GLS, por sus siglas en inglés) puede utilizarse para obtener mejores estimaciones de los errores estándar de los parámetros de regresión y así corregir la autocorrelación en la serie de residuos. El procedimiento se basa esencialmente en maximizar la verosimilitud considerando la autocorrelación en los datos y está implementado en R mediante la función `gls` (dentro de la librería `nlme`, la cual deberá cargarse previamente).

5.4.1 Ajuste de GLS a datos simulados

El siguiente ejemplo ilustra cómo ajustar un modelo de regresión a la serie simulada de la Sección 5.2.3 utilizando mínimos cuadrados generalizados:

```
library(nlme)
x.gls <- gls(x ~ Time, cor = corAR1(0.8))
coef(x.gls)
```

```
## (Intercept)        Time
##   58.233018    3.042245
```

Salida esperada:

```
(Intercept)    Time
  58.23       3.04
```

```
sqrt(diag(vcov(x.gls)))
```

```
## (Intercept)        Time
##  11.9245679   0.2024447
```

Salida esperada:

```
(Intercept)    Time
  11.925       0.202
```

En este caso, se utiliza una autocorrelación de rezago 1 de 0.8 porque este fue el valor utilizado para simular los datos (Sección 5.2.3). Para series históricas, la autocorrelación de rezago 1 debe estimarse a partir del correlograma de los residuos de un modelo lineal ajustado previamente.

En el ejemplo anterior, los errores estándar de los parámetros son considerablemente mayores que los obtenidos mediante OLS en la Sección 5.3.3 y son más precisos, ya que toman en cuenta la autocorrelación. En general, las estimaciones de los parámetros mediante GLS suelen diferir ligeramente de las obtenidas con OLS debido al proceso de ponderación. Por ejemplo, la pendiente se estima como 3.06 usando `lm`, pero 3.04 usando `gls`. En principio, los estimadores GLS son preferibles porque tienen errores estándar más pequeños.

5.4.2 Intervalo de confianza para la tendencia en la serie de temperatura

Para calcular un intervalo de confianza aproximado del 95% para la tendencia en la serie de temperatura global (1970–2005), se utiliza el método GLS para estimar el error estándar considerando la autocorrelación en los residuos (Figura 5.4). En la función `gls`, la serie de residuos se aproxima mediante un proceso AR(1) con una autocorrelación de rezago 1 de 0.7, leída de la Figura 5.4. Este valor se utiliza como parámetro en la función `gls`:

```
temp.gls <- gls(temp ~ time(temp), cor = corAR1(0.7))
confint(temp.gls)
```

```
##                     2.5 %       97.5 %
## (Intercept) -41.24486006 -29.61067134
## time(temp)    0.01498944   0.02084316
```

Salida esperada:

```
               2.5 %      97.5 %
(Intercept) -39.8057   -28.4966
time(temp)   0.0144     0.0201
```

Aunque los intervalos de confianza anteriores son más amplios que los de la Sección 5.3, **el cero no está contenido en los intervalos**, lo que implica que las estimaciones son estadísticamente significativas, y en particular, que la tendencia lo es. Por tanto, existe evidencia estadística de una **tendencia creciente** en las temperaturas globales durante el período 1970–2005. Esto sugiere que, si las condiciones actuales persisten, es probable que las temperaturas continúen aumentando en el futuro.

5.5 Modelos lineales con variables estacionales

5.5.1 Introducción

Dado que las series de tiempo son observaciones medidas secuencialmente en el tiempo, los efectos estacionales suelen estar presentes en los datos, especialmente los ciclos anuales causados directa o indirectamente por el movimiento de la Tierra alrededor del Sol. Ya se han observado efectos estacionales en varias de las series que hemos analizado, incluyendo la serie aérea (§1.4.1), la serie de temperatura (§1.4.5) y la serie de producción eléctrica (§1.4.3). En esta sección se consideran modelos de regresión lineal con variables predictoras que representan efectos estacionales.

5.5.2 Variables indicadoras estacionales aditivas

Supongamos que una serie de tiempo contiene s estaciones. Por ejemplo, si se mide cada mes del calendario, $s = 12$; si se mide cada seis meses (verano e invierno), entonces $s = 2$.

Un modelo de indicadores estacionales para una serie $\{x_t : t = 1, \ldots, n\}$ que contiene s estaciones y una tendencia m_t, se expresa como:

$$x_t = m_t + s_t + z_t \tag{5.6}$$

donde $s_t = \beta_i$ cuando t cae en la i-ésima estación ($i = 1, \ldots, n$; $i = 1, \ldots, s$), y $\{z_t\}$ es la serie de errores residuales, que puede estar autocorrelacionada.

Este modelo tiene la misma forma que el modelo de descomposición aditiva (Ecuación 1.2), pero difiere en que la tendencia m_t se formula con parámetros. En la ecuación (5.6), m_t no es un término constante (intercepto), sino que puede ser un polinomio de orden p con parámetros $\alpha_1, \ldots, \alpha_p$. Entonces, la ecuación (5.6) se convierte en una tendencia polinómica donde el término constante depende de la estación, y los s parámetros estacionales $(\beta_1, \ldots, \beta_s)$ corresponden a s posibles términos constantes en la Ecuación (5.2). Por lo tanto, la ecuación (5.6) puede escribirse como:

$$x_t = m_t + \beta_{1+(t-1) \bmod s} + z_t \tag{5.7}$$

Por ejemplo, con una serie $\{x_t\}$ observada cada mes del calendario comenzando con $t = 1$ en enero, un modelo de indicador estacional con tendencia lineal se da por:

$$x_t = \alpha_1 t + s_t + z_t = \begin{cases} \alpha_1 t + \beta_1 + z_t & t = 1, 13, \ldots \\ \alpha_1 t + \beta_2 + z_t & t = 2, 14, \ldots \\ \vdots & \\ \alpha_1 t + \beta_{12} + z_t & t = 12, 24, \ldots \end{cases} \tag{5.8}$$

Los parámetros del modelo en la ecuación (5.8) pueden estimarse mediante OLS o GLS, tratando la estación estacional s_t como un 'factor'. En R, el factor mensual puede aplicarse usando variables indicadoras extraídas con la función `cycle` (§1.4.1).