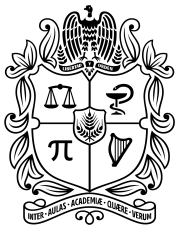
Series de tiempo univariadas - Presentación 4

Mauricio Alejandro Mazo Lopera

Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Proceso ruido blanco

Decimos que un proceso $\{w_t\}$ es **ruido blanco** si es una secuencia de variables aleatorias no correlacionadas provenientes de una distribución específica con:

- Media constante: Para todo t, $E(w_t) = \mu_w$ (usualmente $\mu_w = 0$).
- Varianza constante: Para todo t, $Var(w_t) = \sigma_w^2$.
- $\gamma(k) = Cov(w_t, w_{t+k}) = 0$, para todo $k \neq 0$.

Proceso ruido blanco

Decimos que un proceso $\{w_t\}$ es **ruido blanco** si es una secuencia de variables aleatorias no correlacionadas provenientes de una distribución específica con:

- Media constante: Para todo t, $E(w_t) = \mu_w$ (usualmente $\mu_w = 0$).
- Varianza constante: Para todo t, $Var(w_t) = \sigma_w^2$.
- $\gamma(k) = Cov(w_t, w_{t+k}) = 0$, para todo $k \neq 0$.

Note que esta definición implica que el proceso ruido blanco $\{w_t\}$ es estacionario con función de covarianza:

$$\gamma(k) = \begin{cases} \sigma_w^2, & k = 0 \\ 0, & k \neq 0 \end{cases}$$

Proceso ruido blanco

Además, la ACF de $\{w_t\}$ está dada por:

$$\rho_k = \frac{\gamma(k)}{\gamma(0)} = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases}$$

Proceso ruido blanco

Además, la ACF de $\{w_t\}$ está dada por:

$$\rho_k = \frac{\gamma(k)}{\gamma(0)} = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases}$$

De aquí, la PACF de $\{w_t\}$ está dada por:

$$\phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & 1 \end{vmatrix}}$$

Proceso ruido blanco

Además, la ACF de $\{w_t\}$ está dada por:

$$\rho_k = \frac{\gamma(k)}{\gamma(0)} = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases}$$

De aquí, la PACF de $\{w_t\}$ está dada por:

$$\phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & 1 \end{vmatrix}} = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases}$$

Estimación de la media, autocovarianza, ACF y PACF

Hasta ahora hemos visto teóricamente cómo calcular la esperanza o media, la autocovarianza, la ACF y la PACF de un proceso estacionario.

Estimación de la media, autocovarianza, ACF y PACF

Hasta ahora hemos visto teóricamente cómo calcular la esperanza o media, la autocovarianza, la ACF y la PACF de un proceso estacionario. A continuación veremos cómo estimar estos elementos basados en una muestra aleatoria.

Estimación de la media, autocovarianza, ACF y PACF

Hasta ahora hemos visto teóricamente cómo calcular la esperanza o media, la autocovarianza, la ACF y la PACF de un proceso estacionario. A continuación veremos cómo estimar estos elementos basados en una muestra aleatoria.

- **Media muestral:**

Si tenemos una serie de tiempo o realización de un proceso estocástico estacionario, $X_1, X_2, \ldots, X_n$, un estimador natural de la media $\mu = E(X_t)$ es la media muestral:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Estimación de la media, autocovarianza, ACF y PACF

Hasta ahora hemos visto teóricamente cómo calcular la esperanza o media, la autocovarianza, la ACF y la PACF de un proceso estacionario. A continuación veremos cómo estimar estos elementos basados en una muestra aleatoria.

- **Media muestral:**

Si tenemos una serie de tiempo o realización de un proceso estocástico estacionario, $X_1, X_2, \ldots, X_n$, un estimador natural de la media $\mu = E(X_t)$ es la media muestral:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Dos propiedades importantes para este estimador son:

$$E\left(\overline{X}\right) = \mu \quad \text{y} \quad \lim_{n\to\infty} Var\left(\overline{X}\right) = 0$$

- **Autocovarianza muestral:** Si tenemos una realización de un proceso estacionario, el estimador de $\gamma_k = \gamma(k)$ está dado por

$$\widehat{\gamma}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(X_t - \overline{X})(X_{t+k} - \overline{X})$$

con $\widehat{\gamma}(k) = \widehat{\gamma}(-k)$, para $h = 0, 1, 2, \ldots, n-1$.

- **Autocovarianza muestral:** Si tenemos una realización de un proceso estacionario, el estimador de $\gamma_k = \gamma(k)$ está dado por

$$\widehat{\gamma}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(X_t - \overline{X})(X_{t+k} - \overline{X})$$

con $\widehat{\gamma}(k) = \widehat{\gamma}(-k)$, para $h = 0, 1, 2, \ldots, n-1$. Una propiedad de gran interés en este caso es que para cada k, cuando n es muy grande, $\widehat{\gamma}_k$ es un estimador asintóticamente insesgado para γ_k; es decir, para k dado

$$\lim_{n\to\infty} E\left(\widehat{\gamma}_k\right) = \gamma_k$$

- **Autocorrelación muestral:** Si tenemos una realización de un proceso estacionario, el estimador de la ACF ρ_k está dado por

$$\widehat{\rho}_k = \frac{\widehat{\gamma}_k}{\widehat{\gamma}_0} = \frac{\sum_{t=1}^{n-k}(X_t - \overline{X})(X_{t+k} - \overline{X})}{\sum_{t=1}^{n}(X_t - \overline{X})^2}, \quad \text{para} \quad k = 0, 1, 2, \ldots$$

Al gráfico de k versus $\widehat{\rho}_k$ se le suele llamar **correlograma**.

- **Autocorrelación muestral:** Si tenemos una realización de un proceso estacionario, el estimador de la ACF ρ_k está dado por

$$\widehat{\rho}_k = \frac{\widehat{\gamma}_k}{\widehat{\gamma}_0} = \frac{\sum_{t=1}^{n-k}(X_t - \overline{X})(X_{t+k} - \overline{X})}{\sum_{t=1}^{n}(X_t - \overline{X})^2}, \quad \text{para} \quad k = 0, 1, 2, \ldots$$

  Al gráfico de k versus $\widehat{\rho}_k$ se le suele llamar **correlograma**.

Una propiedad de interés en este caso es que, para n grande, cuando $\{X_t\}$ es ruido blanco Gaussiano, se cumple que para cada k,

$$\widehat{\rho}_k \sim N\left(0, \frac{1}{n}\right)$$

- **Autocorrelación muestral:** Si tenemos una realización de un proceso estacionario, el estimador de la ACF ρ_k está dado por

$$\widehat{\rho}_k = \frac{\widehat{\gamma}_k}{\widehat{\gamma}_0} = \frac{\sum_{t=1}^{n-k}(X_t - \overline{X})(X_{t+k} - \overline{X})}{\sum_{t=1}^{n}(X_t - \overline{X})^2}, \quad \text{para} \quad k = 0, 1, 2, \ldots$$

  Al gráfico de k versus $\widehat{\rho}_k$ se le suele llamar **correlograma**.

Una propiedad de interés en este caso es que, para n grande, cuando $\{X_t\}$ es ruido blanco Gaussiano, se cumple que para cada k,

$$\widehat{\rho}_k \sim N\left(0, \frac{1}{n}\right)$$

Con base en este resultado, un IC aproximado al 95 % para saber si una correlación es significativa o no $\pm Z_{0.975}/\sqrt{n} = \pm 1.96/\sqrt{n}$.

- **Autocorrelación parcial muestral:** Si tenemos una realización de un proceso estacionario, el estimador de la PACF ϕ_{kk} está dado por

$$\widehat{\phi_{kk}} = \frac{\begin{vmatrix} 1 & \widehat{\rho}_1 & \widehat{\rho}_2 & \cdots & \widehat{\rho}_{k-2} & \widehat{\rho}_1 \\ \widehat{\rho}_1 & 1 & \widehat{\rho}_1 & \cdots & \widehat{\rho}_{k-3} & \widehat{\rho}_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \widehat{\rho}_{k-1} & \widehat{\rho}_{k-2} & \widehat{\rho}_{k-3} & \cdots & \widehat{\rho}_1 & \widehat{\rho}_k \end{vmatrix}}{\begin{vmatrix} 1 & \widehat{\rho}_1 & \widehat{\rho}_2 & \cdots & \widehat{\rho}_{k-2} & \widehat{\rho}_{k-1} \\ \widehat{\rho}_1 & 1 & \widehat{\rho}_1 & \cdots & \widehat{\rho}_{k-3} & \widehat{\rho}_{k-2} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \widehat{\rho}_{k-1} & \widehat{\rho}_{k-2} & \widehat{\rho}_{k-3} & \cdots & \widehat{\rho}_1 & 1 \end{vmatrix}}$$

Una propiedad de interés en este caso es que, para n grande, cuando $\{X_t\}$ es ruido blanco Gaussiano, se cumple que para cada k,

$$\widehat{\phi}_{kk} \sim N\left(0, \frac{1}{n}\right)$$

Una propiedad de interés en este caso es que, para n grande, cuando $\{X_t\}$ es ruido blanco Gaussiano, se cumple que para cada k,

$$\widehat{\phi}_{kk} \sim N\left(0, \frac{1}{n}\right)$$

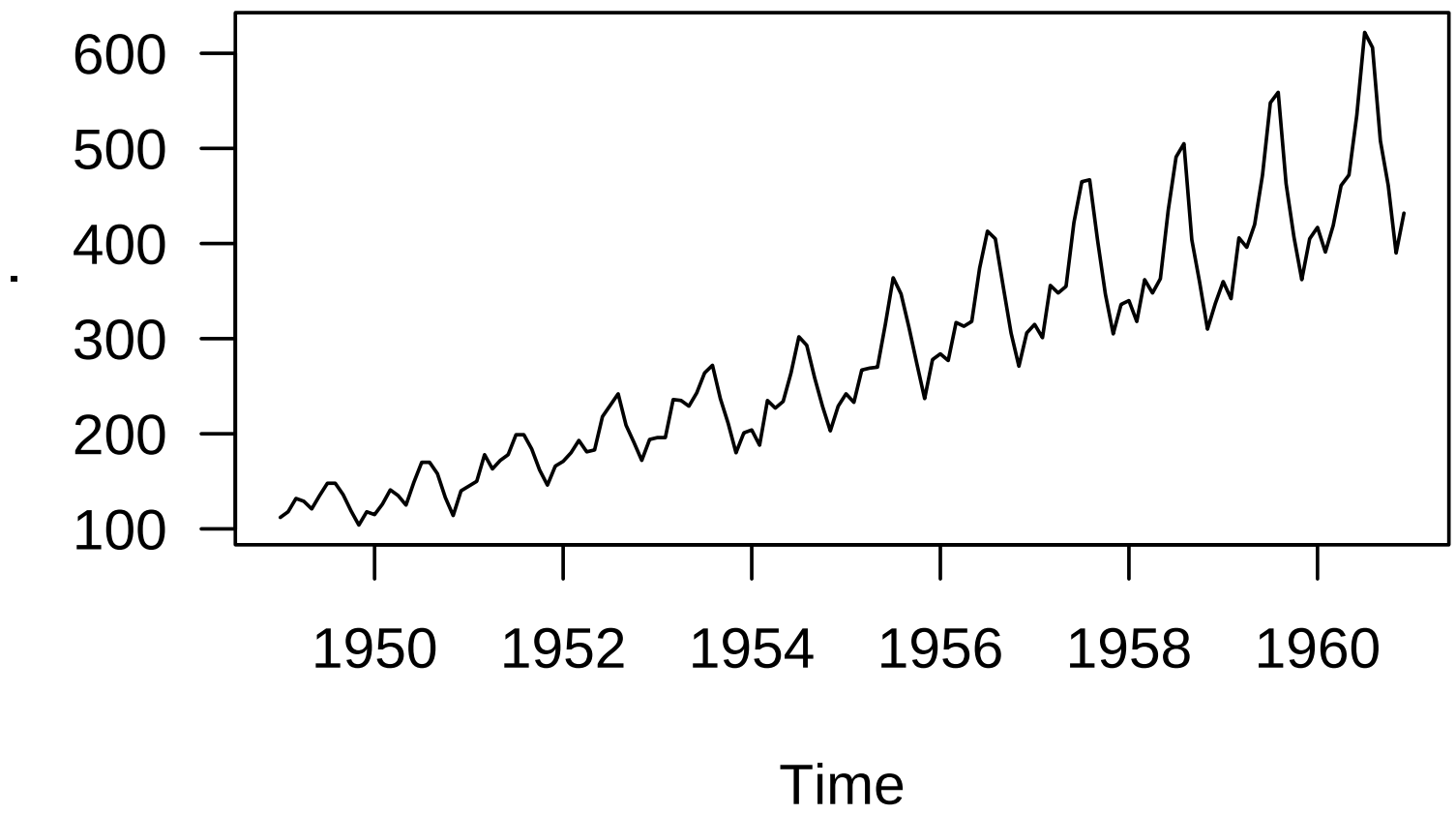
Con base en este resultado, un IC aproximado al 95 % para saber si una correlación parcial es significativa o no

$$\pm Z_{0.975}/\sqrt{n} = \pm 1.96/\sqrt{n}$$

Ejemplo 1 - Número de pasajeros anualmente

Consideremos el ejemplo del número de pasajeros aéreos:

```
require(magrittr)
AirPassengers %>% plot(las=1)
```



Ejemplo 1 - Número de pasajeros anualmente

```
require(lubridate)
require(zoo)
pasajeros <- data.frame(numero=as.matrix(AirPassengers),
  fecha = as.Date(as.yearmon(time(AirPassengers))))
```

Ejemplo 1 - Número de pasajeros anualmente

```
require(lubridate)
require(zoo)
pasajeros <- data.frame(numero=as.matrix(AirPassengers),
  fecha = as.Date(as.yearmon(time(AirPassengers))))
```

Los primeros seis valores de la ACF muestral se obtienen con los códigos:

```
vals.acf <- acf(pasajeros$numero, lag.max = 6,
                plot = FALSE)
vals.acf
```

```
##
## Autocorrelations of series 'pasajeros$numero', by lag
##
##     0     1     2     3     4     5     6
## 1.000 0.948 0.876 0.807 0.753 0.714 0.682
```

Ejemplo 1 - Número de pasajeros anualmente

Los primeros seis valores de la PACF muestral se obtienen con los códigos:

```
vals.pacf <- pacf(pasajeros$numero, lag.max = 6,
                  plot = FALSE)
vals.pacf
```

```
##
## Partial autocorrelations of series 'pasajeros$numero', by lag
##
##      1      2      3      4      5      6
##  0.948 -0.229  0.038  0.094  0.074  0.008
```

Ejemplo 1 - Número de pasajeros anualmente

Los primeros seis valores de la PACF muestral se obtienen con los códigos:

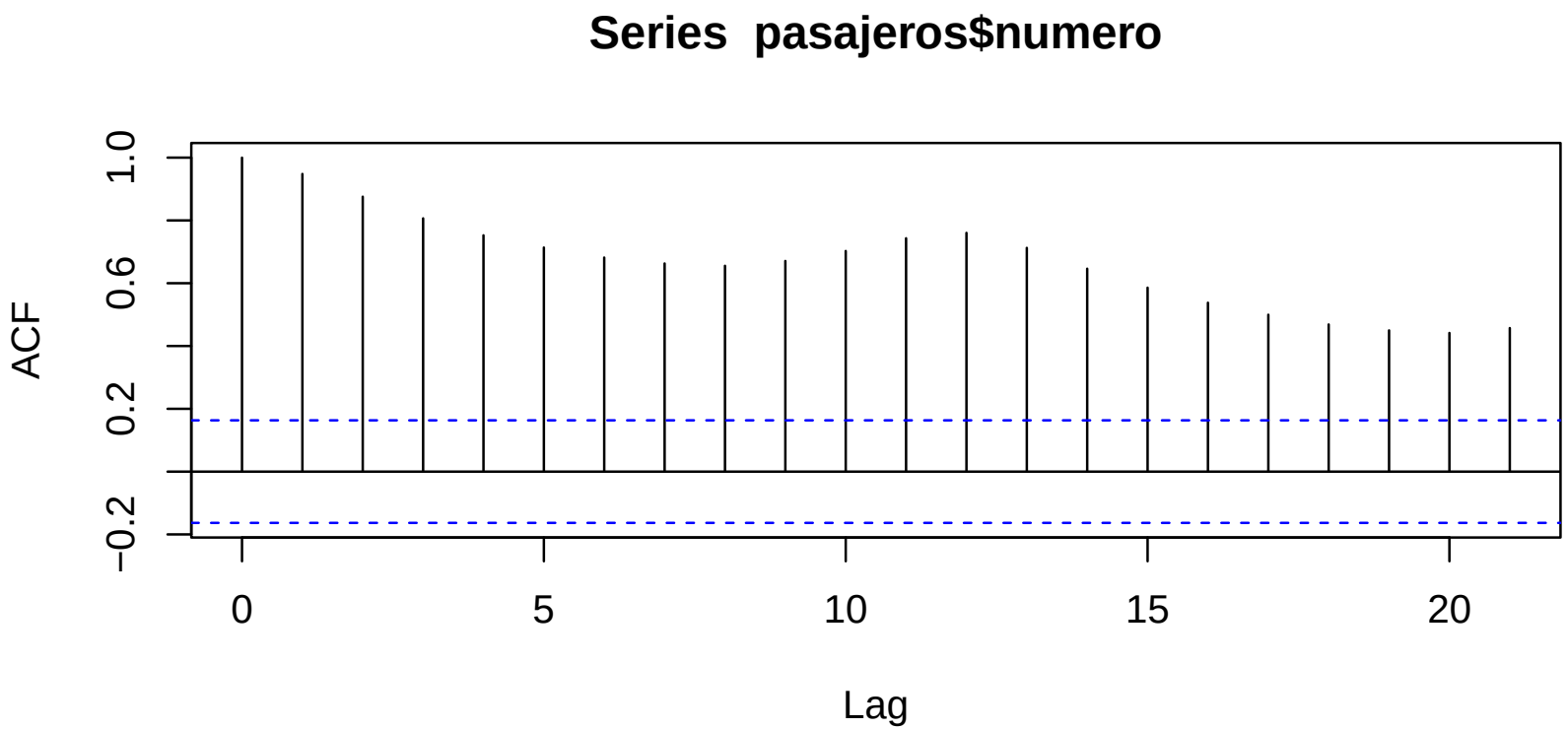
```
vals.pacf <- pacf(pasajeros$numero, lag.max = 6,
                  plot = FALSE)
vals.pacf
```

```
##
## Partial autocorrelations of series 'pasajeros$numero', by lag
##
##      1      2      3      4      5      6
##  0.948 -0.229  0.038  0.094  0.074  0.008
```

Resulta más fácil observar los gráficos de las ACF y PACF muestrales:

Ejemplo 1 - Número de pasajeros anualmente

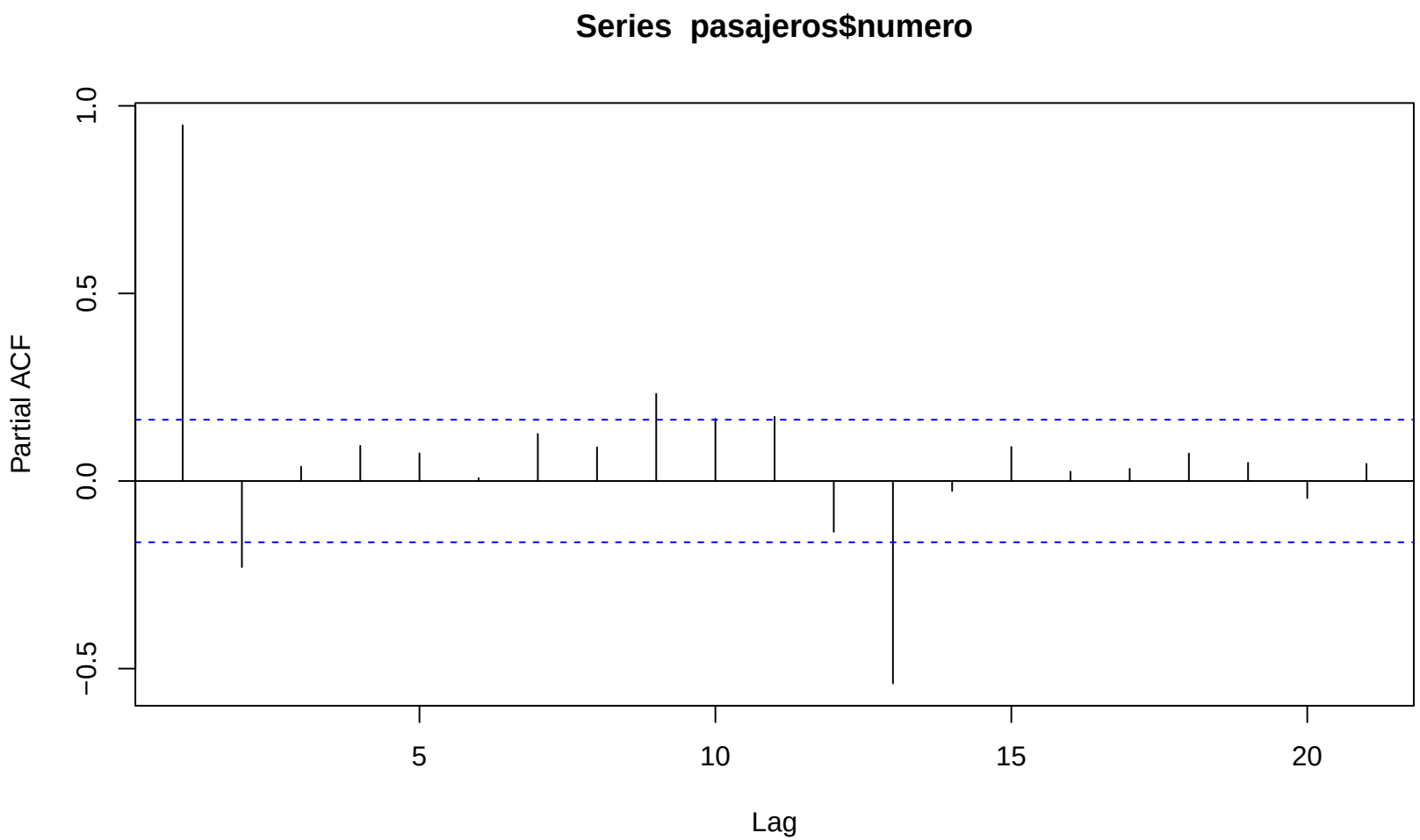
```
acf(pasajeros$numero)
```



Todos los valores están por fuera de la banda de confianza al 95 % ($\pm 1.96/\sqrt{n} = \pm 1.96/\sqrt{144}$). Esto indica que hay una alta correlación de cada dato con todos sus valores anteriores, lo cual se debe a la tendencia de la serie original.

Ejemplo 1 - Número de pasajeros anualmente

```
pacf(pasajeros$numero)
```



Ejemplo 1 - Número de pasajeros anualmente

Solo 6 valores están por fuera de la banda de confianza al 95 % ($\pm 1.96/\sqrt{n} = \pm 1.96/\sqrt{144}$). Los dos más relevantes se describen a continuación:

Ejemplo 1 - Número de pasajeros anualmente

Solo 6 valores están por fuera de la banda de confianza al 95 % ($\pm 1.96/\sqrt{n} = \pm 1.96/\sqrt{144}$). Los dos más relevantes se describen a continuación:

- La primera en el rezago (lag) 1, indica que cada dato está correlacionado fuertemente con el valor anterior de manera positiva. Por ejemplo, el número de pasajeros de diciembre aumenta si el número de pasajeros en noviembre también aumenta.

Ejemplo 1 - Número de pasajeros anualmente

Solo 6 valores están por fuera de la banda de confianza al 95 % ($\pm 1.96/\sqrt{n} = \pm 1.96/\sqrt{144}$). Los dos más relevantes se describen a continuación:

- La primera en el rezago (lag) 1, indica que cada dato está correlacionado fuertemente con el valor anterior de manera positiva. Por ejemplo, el número de pasajeros de diciembre aumenta si el número de pasajeros en noviembre también aumenta.
- La otra barra relevante es la que se presenta en el rezago (lag) 13, indicando que cada dato está relacionado fuertemente con su pasado en 13 periodos de manera negativa cuando eliminamos la influencia de los datos del medio. Por ejemplo, en diciembre de 1960 disminuyen los pasajeros si en noviembre de 1959 aumentaron, esto sin tener en cuenta lo que ocurrio entre diciembre de 1959 y noviembre de 1960. Dicho de otra forma, se "captura" un comportamiento estacional fuerte cada 13 meses.

Ejemplo 1 - Número de pasajeros anualmente

Una pregunta interesante que surge con lo anterior es: ¿qué ocurre con la AFC y la PACF muestrales si en lugar de analizar la serie original vemos la diferencia entre un valor y el inmediatamente anterior?

Ejemplo 1 - Número de pasajeros anualmente

Una pregunta interesante que surge con lo anterior es: ¿qué ocurre con la AFC y la PACF muestrales si en lugar de analizar la serie original vemos la diferencia entre un valor y el inmediatamente anterior?

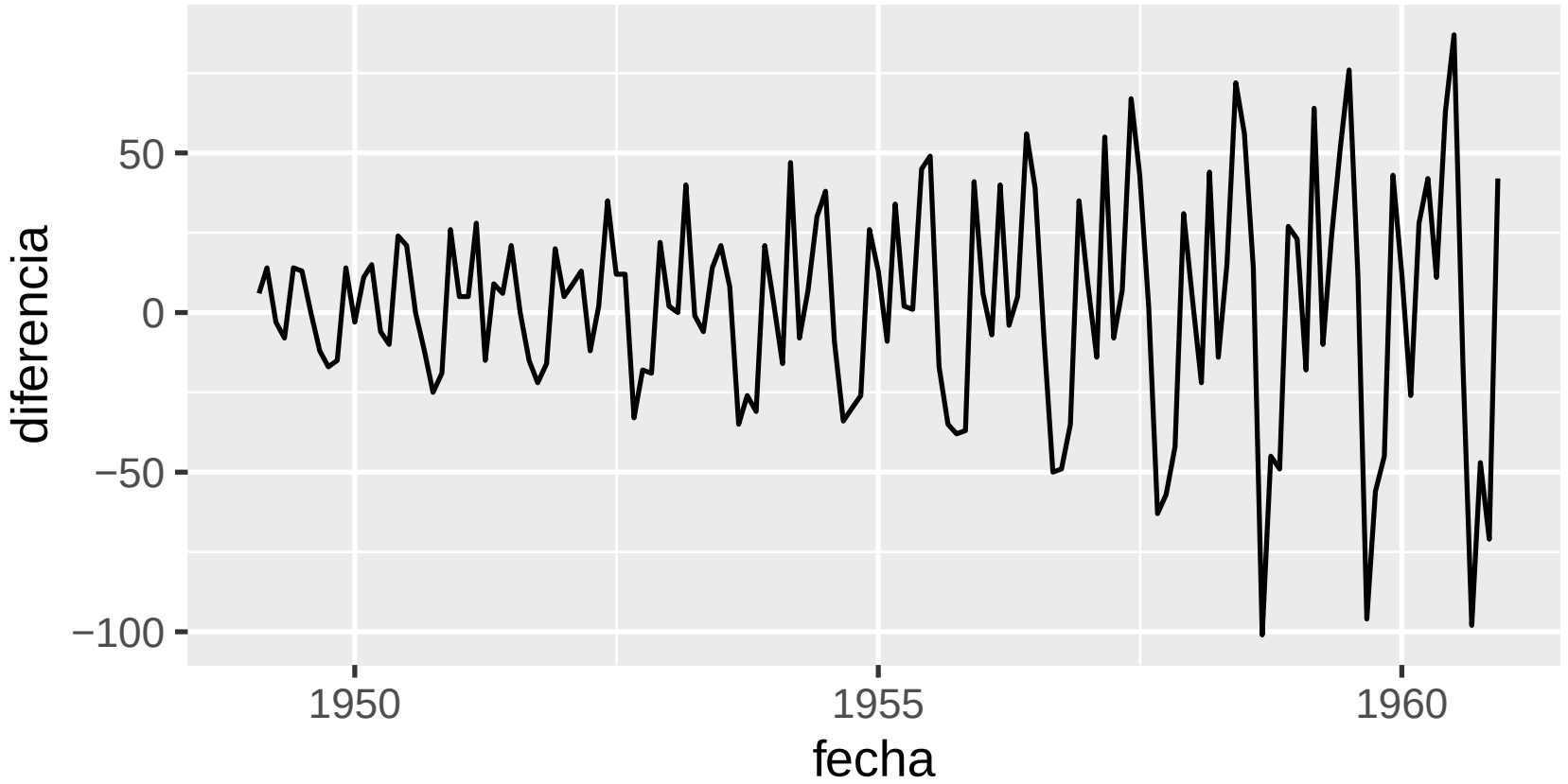
```
pasajeros$diferencia<-c(NA,diff(pasajeros$numero,lag=1))
pasajeros %>% head(6)
```

```
##   numero      fecha diferencia
## 1    112 1949-01-01         NA
## 2    118 1949-02-01          6
## 3    132 1949-03-01         14
## 4    129 1949-04-01         -3
## 5    121 1949-05-01         -8
## 6    135 1949-06-01         14
```

Ejemplo 1 - Número de pasajeros anualmente

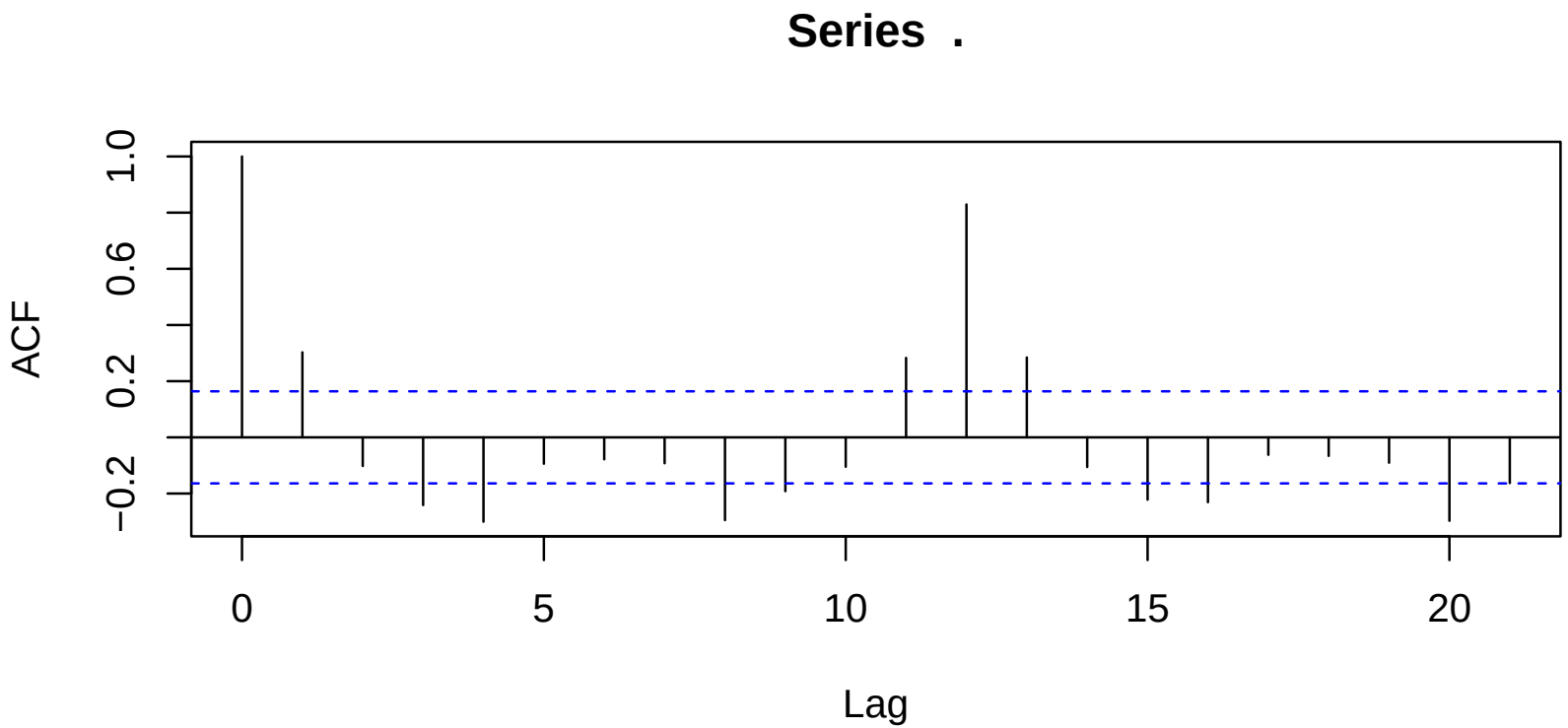
Consideremos el ejemplo del número de pasajeros aéreos:

```
require(tidyverse)
pasajeros %>% ggplot(aes(x=fecha, y=diferencia))+
  geom_line()
```



Ejemplo 1 - Número de pasajeros anualmente

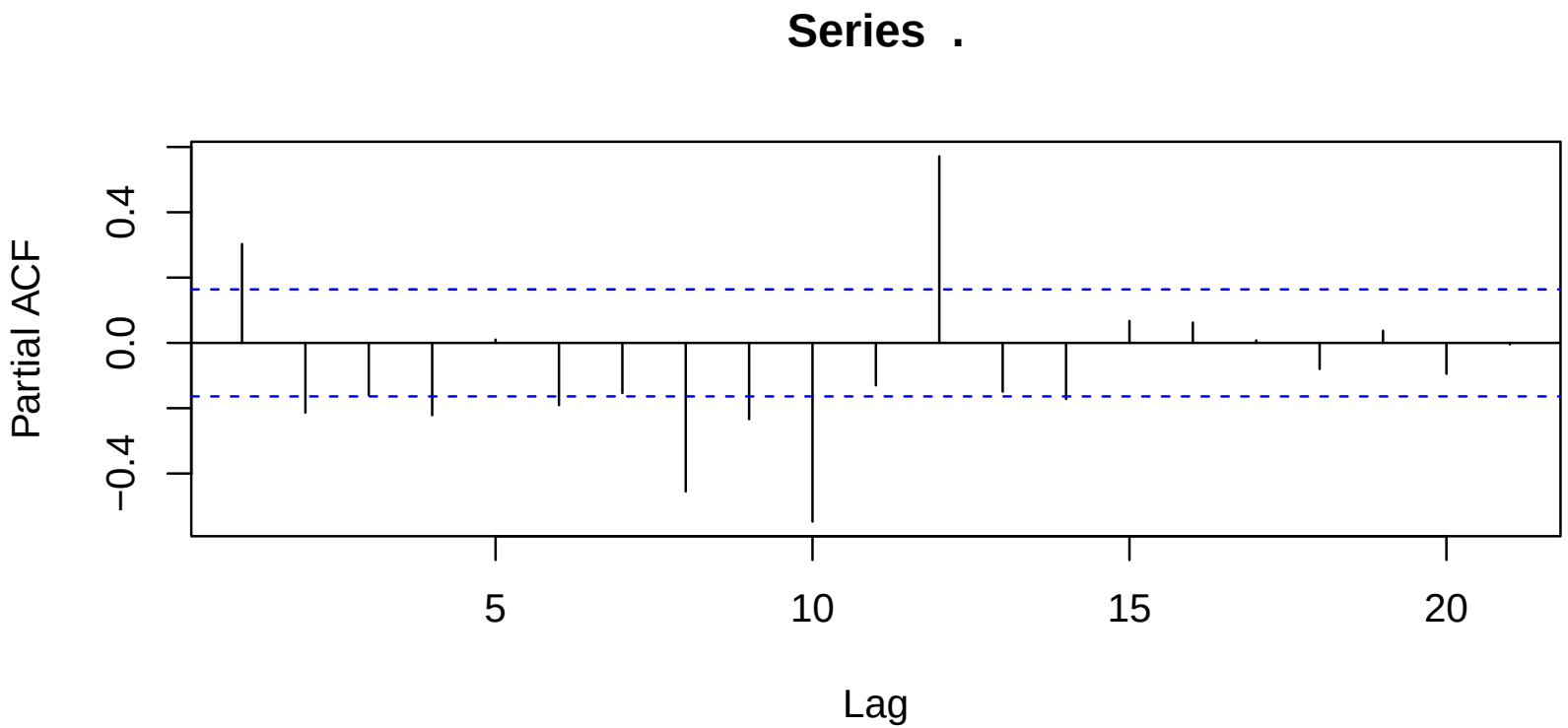
```
pasajeros$diferencia %>% na.omit() %>% acf()
```



Se observa un comportamiento estacional con rezagos (lags) que sobresalen cada 4 periodos, sobresaliendo el lag 1 y el lag 12.

Ejemplo 1 - Número de pasajeros anualmente

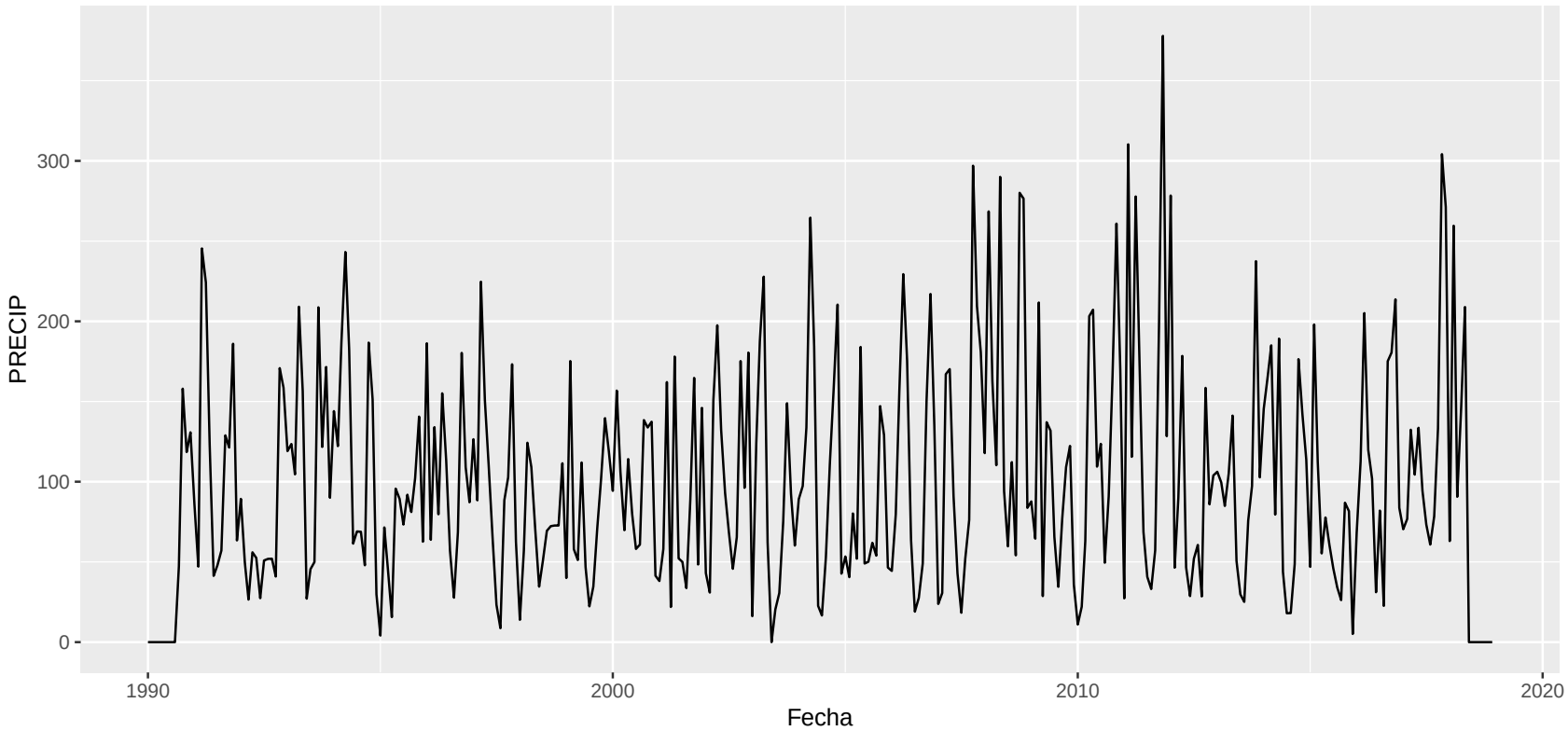
```
pasajeros$diferencia %>% na.omit() %>% pacf()
```



Lags 10 y 12 sobresalen.

Introducción: Ejemplo 2 - Precipitaciones - Mesitas

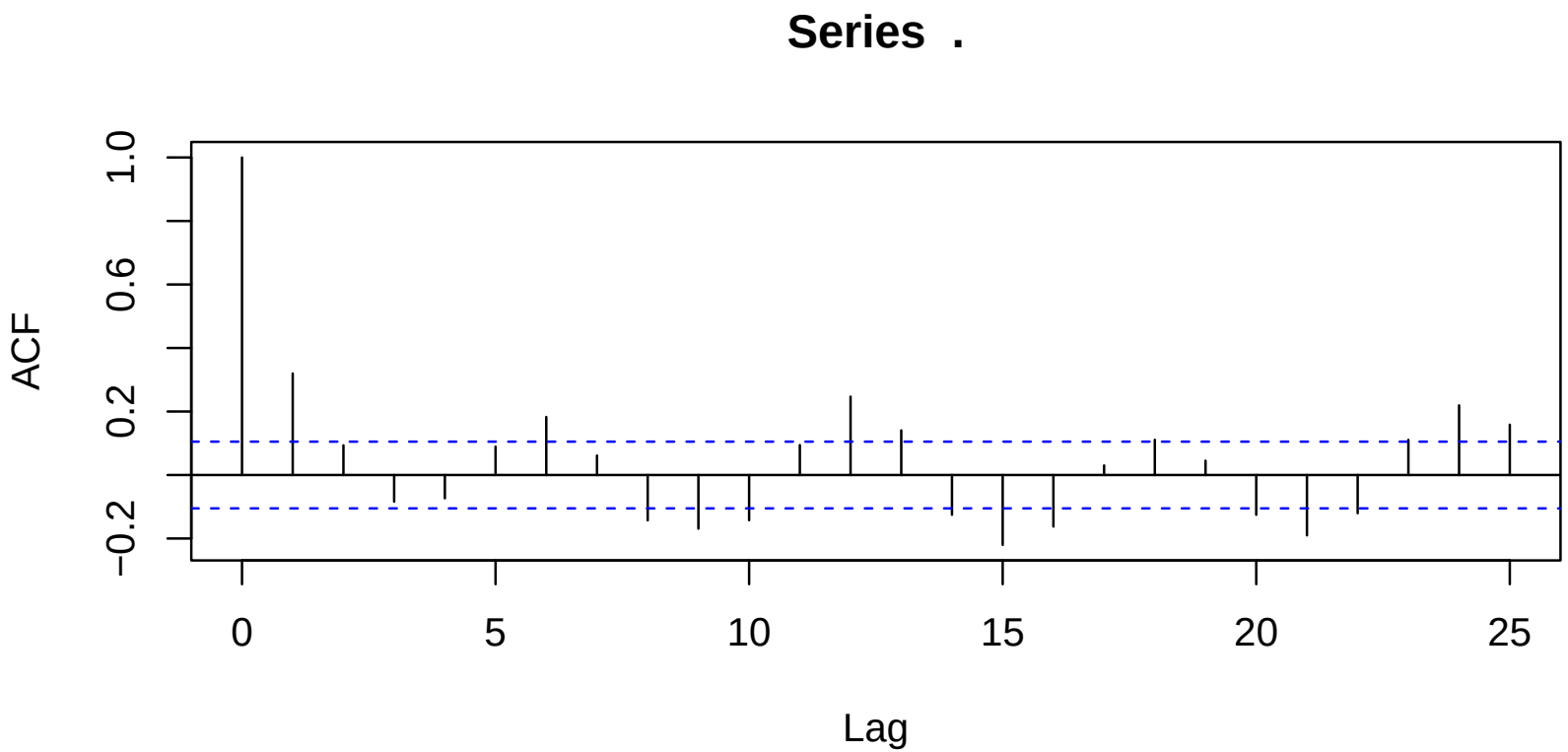
Gráfico 3



CÓDIGOS

Introducción: Ejemplo 2 - Precipitaciones - Mesitas

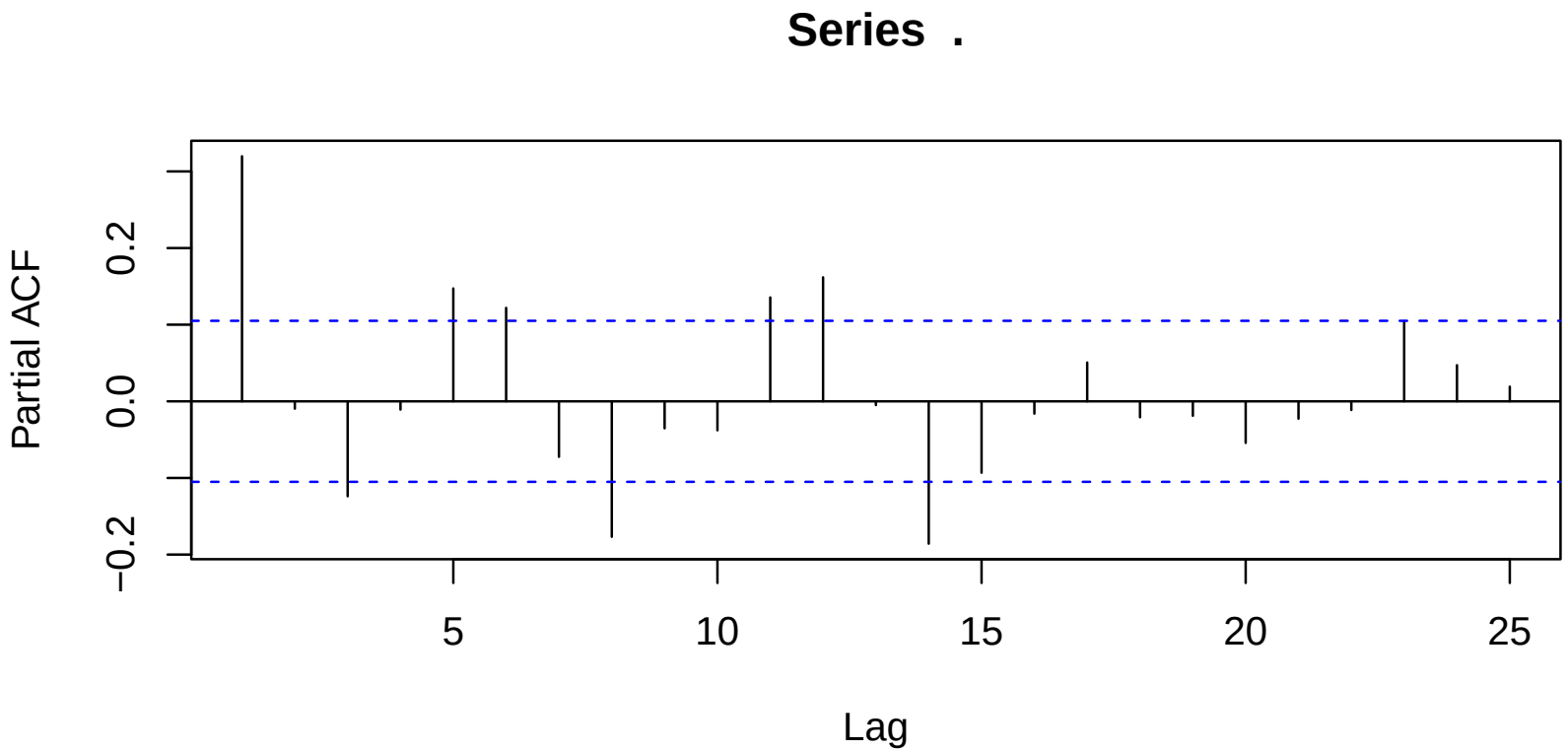
```
BD2_3$PRECIP %>% acf()
```



Se observan ciclos a lo largo del gráfico de correlaciones altas trimestrales, semestrales, anuales, etc. La segunda barra muestra una correlación alta entre cada mes y su antecesor inmediato.

Introducción: Ejemplo 2 - Precipitaciones - Mesitas

```
BD2_3$PRECIP %>% pacf()
```



También se observan ciclos a lo largo del gráfico de correlaciones altas trimestrales, semestrales, anuales, etc. La primera barra muestra una correlación alta entre cada mes y su antecesor inmediato.

Regresión clásica en el contexto de series de tiempo

Suponga que tenemos una serie se tiempo X_t para $t = 1, 2, \ldots, n$ y que queremos averiguar si está influenciada por un conjunto de variables explicativas conocidas (las cuales también pueden ser series de tiempo independientes entre ellas), $Z_{t1}, Z_{t2}, \ldots, Z_{tp}$.

Regresión clásica en el contexto de series de tiempo

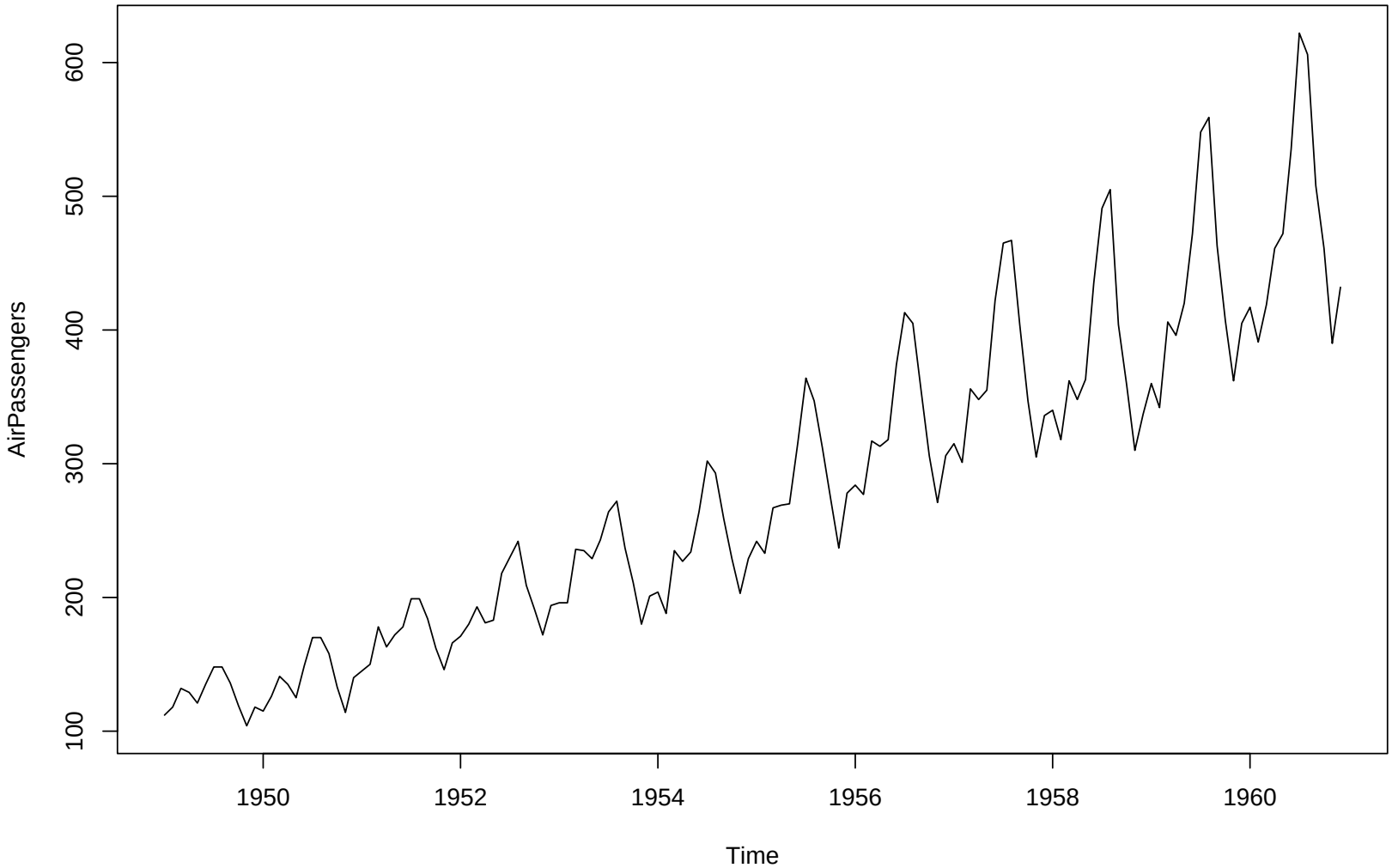
Suponga que tenemos una serie se tiempo X_t para $t = 1, 2, \ldots, n$ y que queremos averiguar si está influenciada por un conjunto de variables explicativas conocidas (las cuales también pueden ser series de tiempo independientes entre ellas), $Z_{t1}, Z_{t2}, \ldots, Z_{tp}$. El modelo de regresión lineal está dado por:

$$X_t = \beta_0 + \beta_1 Z_{t1} + \beta_2 Z_{t2} + \cdots + \beta_p Z_{tp} + w_t$$

donde $\beta_0, \beta_1, \ldots, \beta_p$ son los coeficientes de regresión a ser estimados y $\{w_t\}$ es un error aleatorio o proceso ruido blanco gaussiano con media 0 y varianza σ_w^2.

Regresión clásica en el contexto de series de tiempo

Consideremos el ejemplo 1 de esta presentación relacionado con el número de pasajeros:



Regresión clásica en el contexto de series de tiempo

En este caso podríamos modelar el número de pasajeros en el instante t, X_t, a través del modelo:

$$X_t = \beta_0 + \beta_1 t + w_t$$

Regresión clásica en el contexto de series de tiempo

En este caso podríamos modelar el número de pasajeros en el instante t, X_t, a través del modelo:

$$X_t = \beta_0 + \beta_1 t + w_t$$

En RStudio quedaría:

```
require(zoo)
pasajeros <- data.frame(numero=as.matrix(AirPassengers),
  fecha = as.Date(as.yearmon(time(AirPassengers))),
  t = c(1:length(AirPassengers)))
modelo1 <- lm( numero ~ t , data=pasajeros)
```

Extraemos los residuales del modelo: $residuales_t = X_t - \widehat{\beta}_0 - \widehat{\beta}_1 t$

```
mod1.residuales <- modelo1 %>% residuals()
```

Regresión clásica en el contexto de series de tiempo

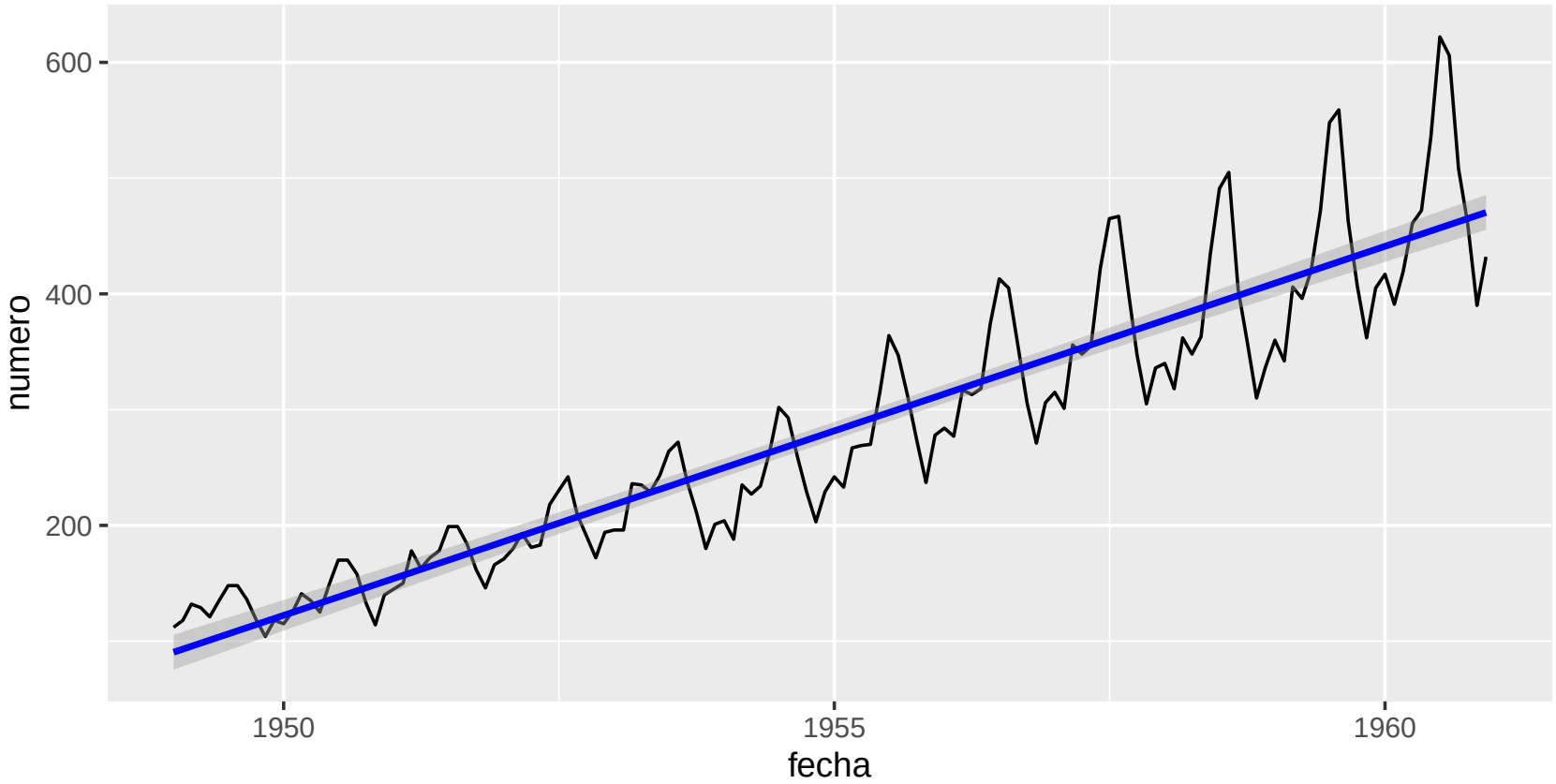
```
modelo1 %>% summary()
```

```
##
## Call:
## lm(formula = numero ~ t, data = pasajeros)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -93.858 -30.727  -5.757  24.489 164.999
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 87.65278    7.71635   11.36   <2e-16 ***
## t            2.65718    0.09233   28.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.06 on 142 degrees of freedom
## Multiple R-squared:  0.8536, Adjusted R-squared:  0.8526
## F-statistic: 828.2 on 1 and 142 DF,  p-value: < 2.2e-16
```

Regresión clásica en el contexto de series de tiempo

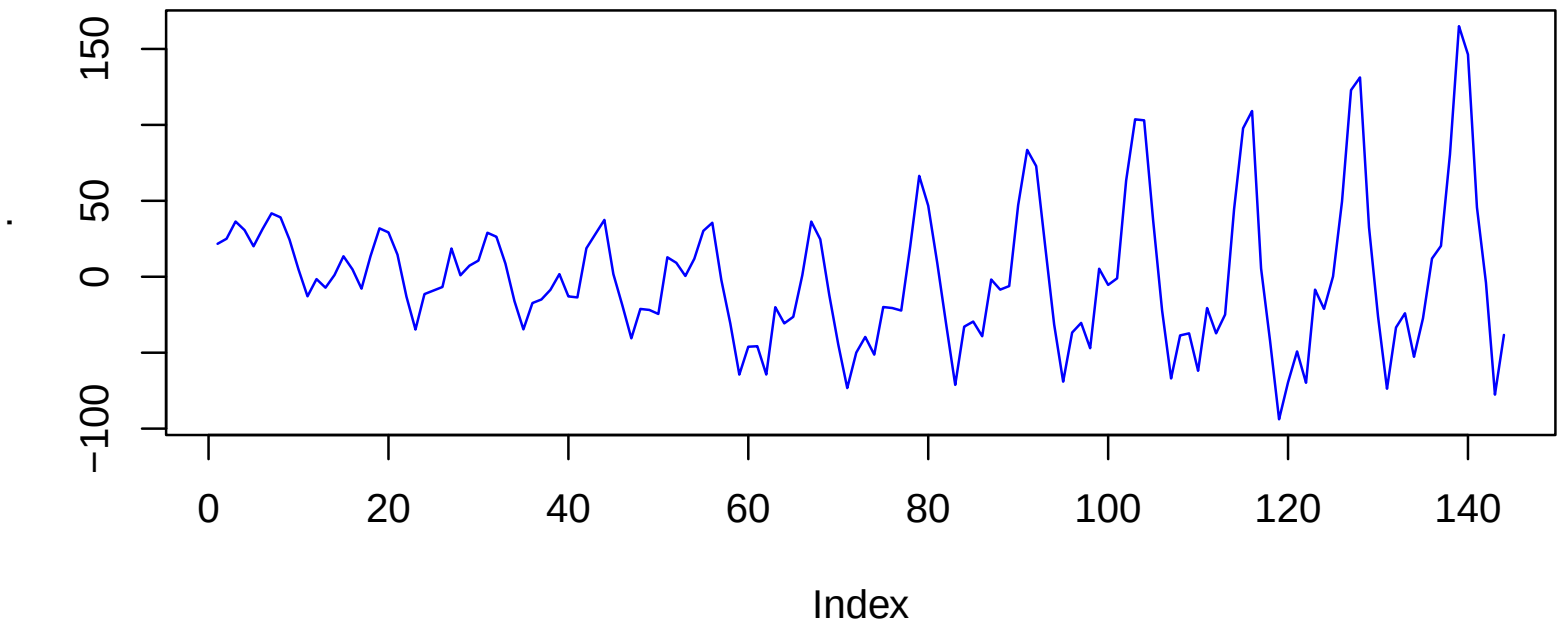
Vemos el modelo con el ajuste de **tendencia:**

```
pasajeros %>% ggplot(aes(x=fecha, y=numero)) +
  geom_line()+
  geom_smooth(method="lm", col="blue")
```



Regresión clásica en el contexto de series de tiempo

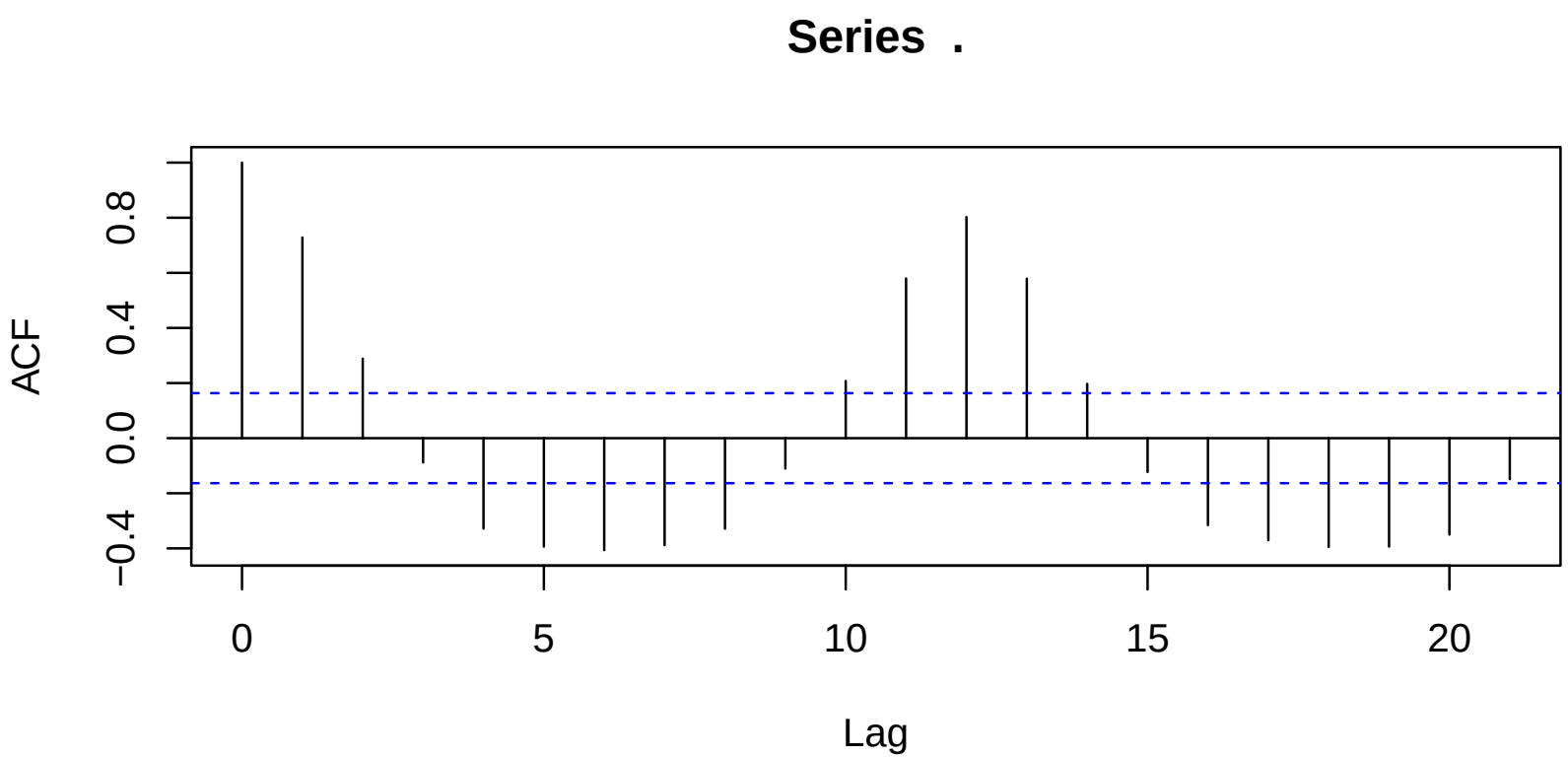
```
mod1.residuales %>% plot(type="l", col="blue")
```



Vemos que después de eliminar la tendencia, los residuales evidencian comportamiento estacional a lo largo del tiempo.

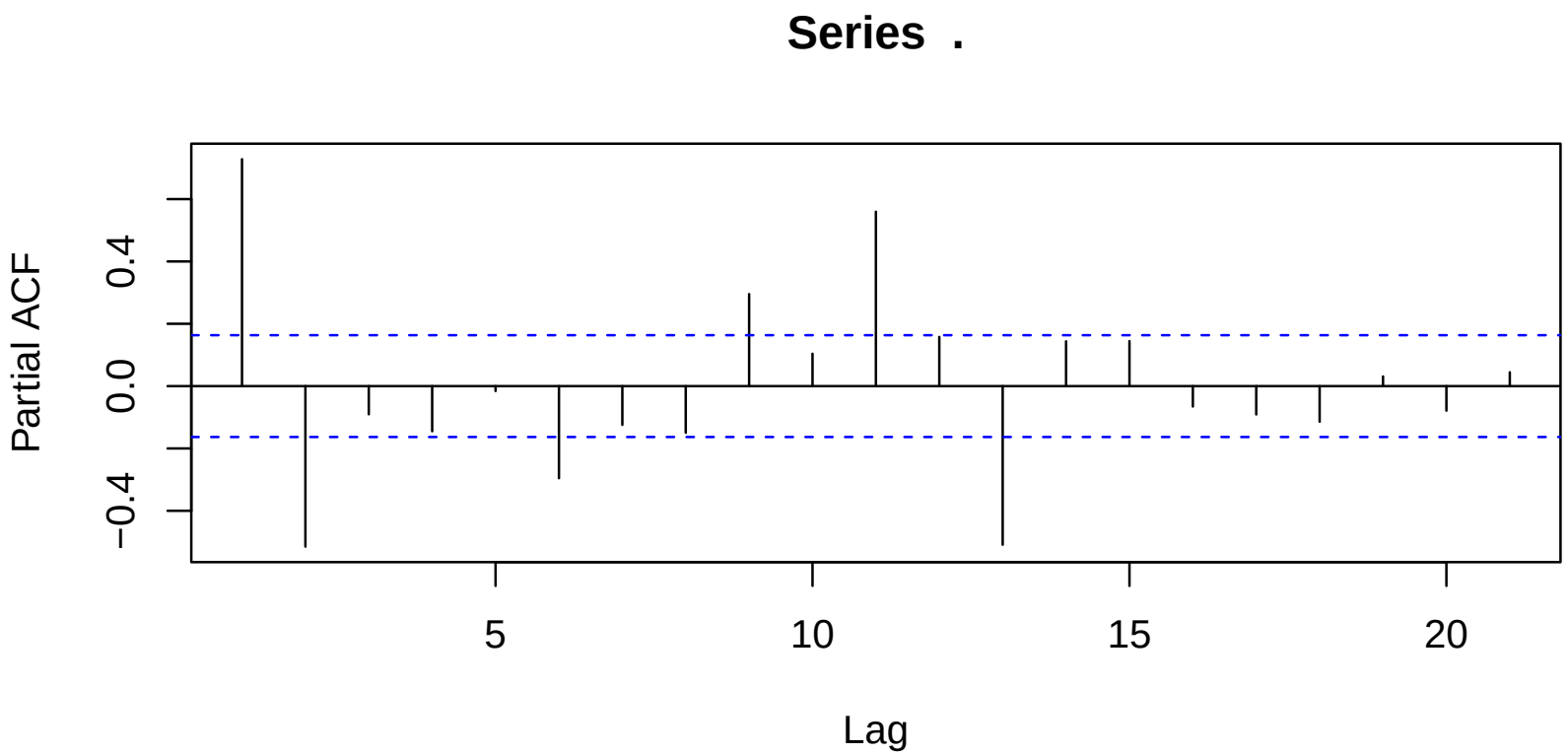
Regresión clásica en el contexto de series de tiempo

```
mod1.residuales %>% acf()
```



Están correlacionados evidenciando correlación estacional. Más adelante vemos cómo modelar esta correlación.

```
mod1.residuales %>% pacf()
```



Están correlacionados evidenciando correlación estacional. Más adelante vemos cómo modelar esta correlación.

Modelamiento general de la tendencia T_t:

En lugar de modelar la tendencia de una serie de tiempo X_t con regresión lineal, también es posible modelarla con curvas no lineales como:

- Polinomios.
- Regresión splines.
- Splines de suavizamiento.
- Regresión local.

Regresión polinomial

La relación polinomial de orden d está dada por:

$$X_t = \beta_0 + \beta_1 Z_t + \beta_2 Z_t^2 + \cdots + \beta_d Z_t^d + w_t$$

La ventaja es que es un modelo más flexible, pero se corre el riesgo de un sobreajuste. Por eso se recomienda usar un d menor o igual a 4.

Regresión polinomial

La relación polinomial de orden d está dada por:

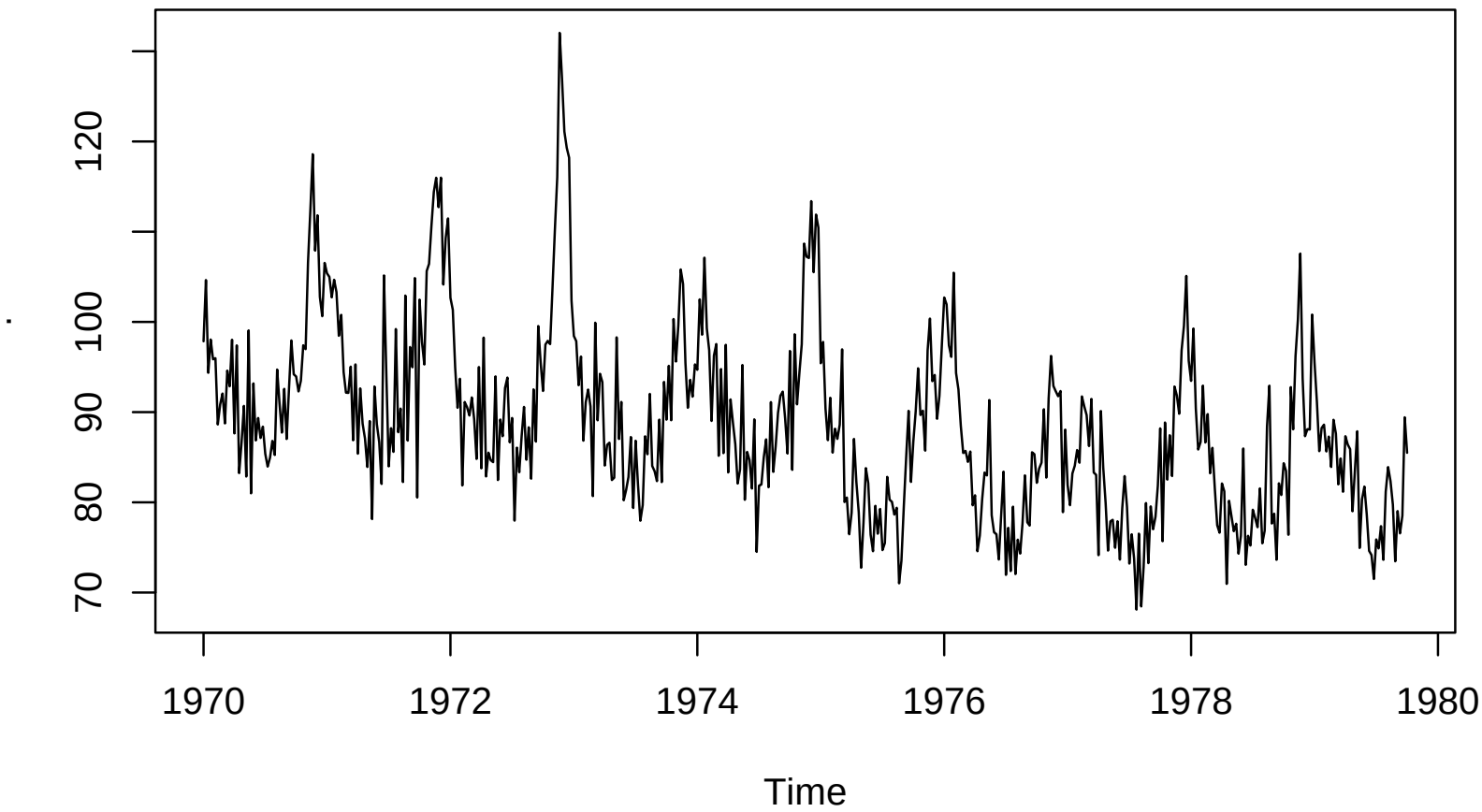
$$X_t = \beta_0 + \beta_1 Z_t + \beta_2 Z_t^2 + \cdots + \beta_d Z_t^d + w_t$$

La ventaja es que es un modelo más flexible, pero se corre el riesgo de un sobreajuste. Por eso se recomienda usar un d menor o igual a 4.

Como ejemplo, consideremos la serie de tiempo relacionada con la mortalidad cardiovascular semanal promedio en el condado de Los Ángeles; 508 promedios suavizados de seis días obtenidos al filtrar valores diarios durante el período de 10 años 1970-1979:

Regresión polinomial

```
require(astsa)
cmort %>% plot()
```



Regresión polinomial

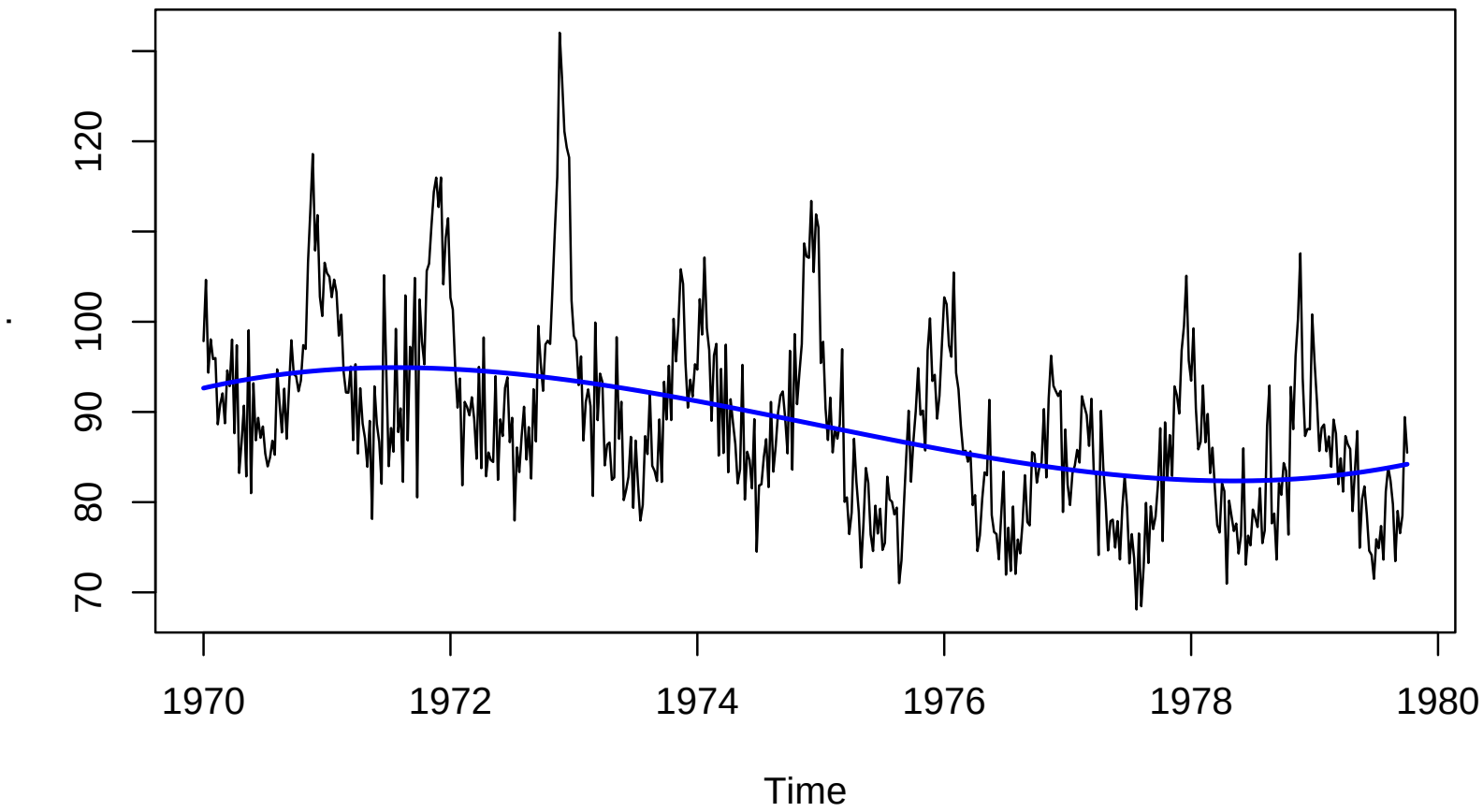
Ajustamos un polinomio de grado $d = 3$ con respecto a la semana:

```
semana <- cmort %>% time() %>% as.numeric()
modelo2 <- lm(cmort~poly(semana, degree=3))
modelo2 %>% summary()
```

```
##
## Call:
## lm(formula = cmort ~ poly(semana, degree = 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.750  -6.642  -1.023   5.302  38.382
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  88.6989     0.3903 227.284  < 2e-16 ***
## poly(semana, degree = 3)1 -103.2805     8.7959 -11.742  < 2e-16 ***
## poly(semana, degree = 3)2   -2.7720     8.7959  -0.315 0.752780
## poly(semana, degree = 3)3   31.9161     8.7959   3.629 0.000314 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.796 on 504 degrees of freedom
## Multiple R-squared:  0.2307, Adjusted R-squared:  0.2261
## F-statistic: 50.38 on 3 and 504 DF,  p-value: < 2.2e-16
```

Regresión polinomial

```
require(astsa)
cmort %>% plot()
lines(semana, modelo2$fitted.values, col="blue", lwd=2)
```



Splines cúbicos:

Si queremos ajustar un spline cúbico con k nodos, dados por $\xi_1, \xi_2, \ldots, \xi_k$, se debe plantear el modelo como:

$$X_t = \beta_0 + \beta_1 Z_t + \beta_2 Z_t^2 + \beta_3 Z_t^3 + \beta_{1h} h(Z_t, \xi_1) + \cdots + \beta_{kh} h(Z_t, \xi_k) + w_t$$

donde

$$h(z, \xi) = (z - \xi)_+^3 = \begin{cases} (z - \xi)^3, & \text{si } z > \xi \\ 0, & \text{e.o.c.} \end{cases}$$

donde ξ es el nodo.

Splines cúbicos:

Si queremos ajustar un spline cúbico con k nodos, dados por $\xi_1, \xi_2, \ldots, \xi_k$, se debe plantear el modelo como:

$$X_t = \beta_0 + \beta_1 Z_t + \beta_2 Z_t^2 + \beta_3 Z_t^3 + \beta_{1h} h(Z_t, \xi_1) + \cdots + \beta_{kh} h(Z_t, \xi_k) + w_t$$

donde

$$h(z, \xi) = (z - \xi)_+^3 = \begin{cases} (z - \xi)^3, & \text{si } z > \xi \\ 0, & \text{e.o.c.} \end{cases}$$

donde ξ es el nodo.

Note que en en este método se deben estimar $k + 4$ parámetros. Algunos paquetes estadísticos suelen designar estos con el nombre de grados de libertad y plantean que para un spline cúbico con k nodos, el número de grados de libertad es igual a $k + 4$.

Seleccionando el número de nodos y su ubicación:

Lo primero que se debe tener en cuenta es que entre mayor sea el número de nodos, mayor será la flexibilidad del modelo y entre menos nodos, más estabilidad. Ambas características en exceso pueden ser perjudiciales (¿Por qué?) y por tanto se debe buscar un equilibrio entre ambas.

Seleccionando el número de nodos y su ubicación:

La manera más utilizada de ubicar los nodos es considerando, de manera uniforme, los cuantiles de los datos en X. Así, por ejemplo, si se seleccionan 3 nodos y no se especifica donde ubicarlos, la manera más directa de ubicarlos es considerar:

$$\begin{aligned} \text{Primer nodo} &\longrightarrow \text{Percentil } 25\,\% \text{ de X} \\ \text{Segundo nodo} &\longrightarrow \text{Percentil } 50\,\% \text{ de X} \\ \text{Tercer nodo} &\longrightarrow \text{Percentil } 75\,\% \text{ de X} \end{aligned}$$

Relación entre **df** y el número de nodos en **bs()**

En el software R existe un paquete llamado **splines** que contiene la función **bs()**, la cual permite generar un spline definiendo el número de nodos o los grados de libertad, además del grado de los polinomios. Por defecto, ajusta splines cúbicos.

`df`	N° de nodos	Modelo resultante
3	0	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$ (modelo usual sin splines)
4	1	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - t_1)_+^3 + \epsilon$
5	2	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - t_1)_+^3 + \beta_5 (x - t_2)_+^3 + \epsilon$
6	3	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - t_1)_+^3 + \beta_5 (x - t_2)_+^3 + \beta_6 (x - t_3)_+^3 + \epsilon$
7	4	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - t_1)_+^3 + \beta_5 (x - t_2)_+^3 + \beta_6 (x - t_3)_+^3 + \beta_7 (x - t_4)_+^3 + \epsilon$
Este patrón continua así		

Por ejemplo, al usar `bs(x, df=6)`

- Se creará un spline básico con tres nodos t_1, t_2 y t_3,
- Los nodos estarán equiespaciados en los percentiles 25%, 50% y 75%.
- No es necesario decirle los nodos porque éstos se elijen automáticamente.
- Si el usuario quiere sus propios nodos se usa entonces el parámetro `knots`.

Spline cúbico

Ajustamos un spline cúbico con 3 nodos con respecto a la semana:

```
require(splines)
semana <- cmort %>% time() %>% as.numeric()
modelo3 <- lm(cmort~bs(semana, df=6, degree=3))
modelo3 %>% summary()
```

```
##
## Call:
## lm(formula = cmort ~ bs(semana, df = 6, degree = 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.368  -6.677  -0.872   5.015  38.604
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      90.8010     2.4677  36.796  < 2e-16 ***
## bs(semana, df = 6, degree = 3)1   6.6485     4.5743   1.453 0.146726
## bs(semana, df = 6, degree = 3)2   2.8011     2.9320   0.955 0.339857
## bs(semana, df = 6, degree = 3)3   0.3288     3.6354   0.090 0.927981
## bs(semana, df = 6, degree = 3)4 -13.7065     3.3503  -4.091    5e-05 ***
## bs(semana, df = 6, degree = 3)5  -1.6885     3.7978  -0.445 0.656789
## bs(semana, df = 6, degree = 3)6 -12.0285     3.4309  -3.506 0.000496 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.75 on 501 degrees of freedom
```

Spline cúbico

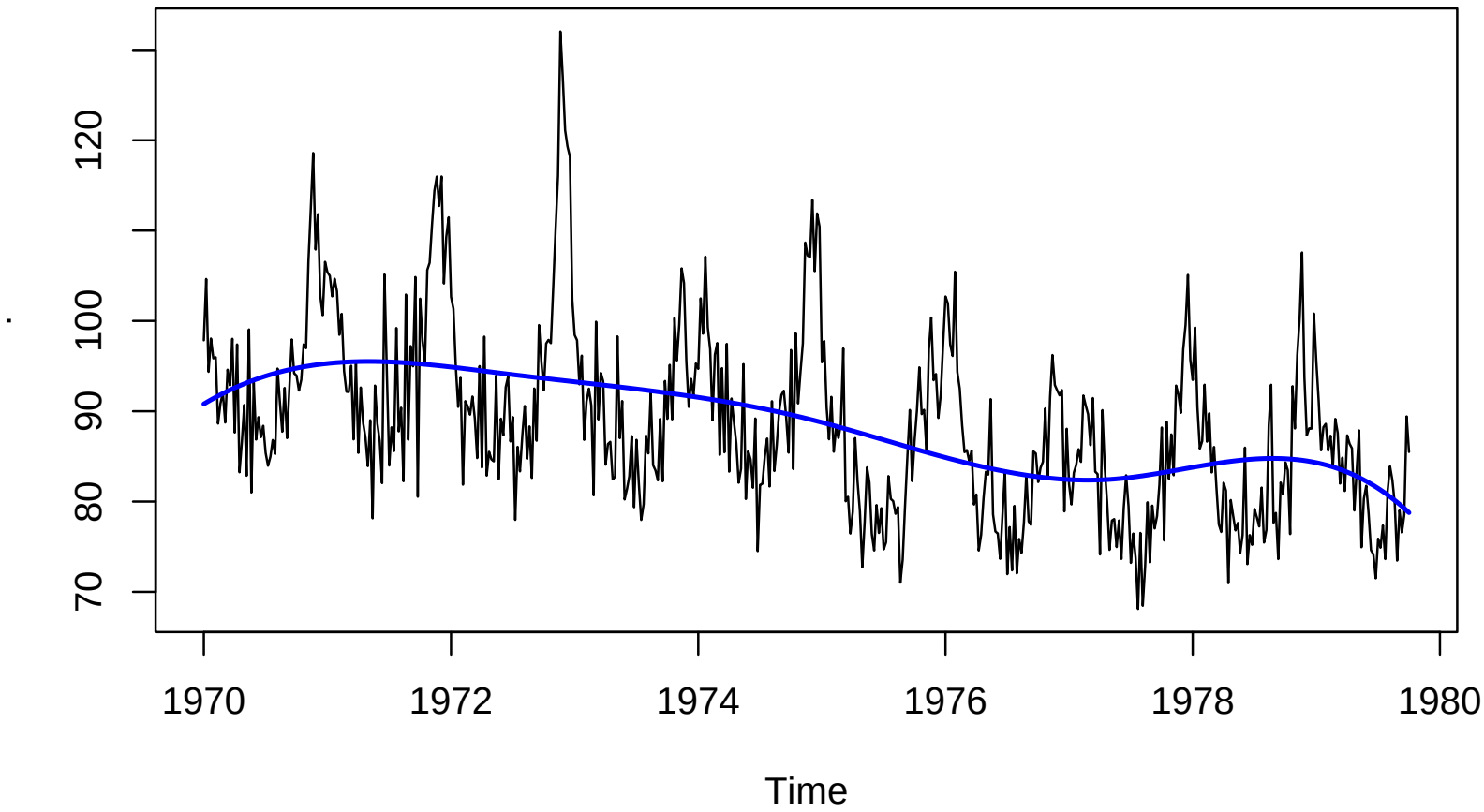
¿Cuáles fueron los nodos que seleccionó por defecto en este caso?

```
attr(bs(semana,df=6,degree=3),"knots")
```

```
##       25%       50%       75%
## 1972.438 1974.875 1977.312
```

Spline cúbico

```
cmort %>% plot()
lines(semana, modelo3$fitted.values, col="blue", lwd=2)
```




MUCHAS
Gracias!

Anexos - Mesitas

```
BD2<-read.csv("../../DATOS/Precipitaciones_Totales_Mensuales.csv",
                header=TRUE,fileEncoding = "utf8")
BD2_1 <-  filter(BD2,ESTACION=="Mesitas")
BD2_2 <- gather(BD2_1, "ENERO", "FEBRERO", "MARZO", "ABRIL",
                "MAYO", "JUNIO", "JULIO","AGOSTO",
                "SEPTIEMBRE","OCTUBRE", "NOVIEMBRE",
                "DICIEMBRE", key="MES",value="PRECIP")
BD2_3 <- BD2_2[order(BD2_2$ANIO),]
BD2_3$Fecha <- paste(rep(1,nrow(BD2_3)),BD2_3$MES,BD2_3$ANIO)
BD2_3$Fecha %<>% as.Date(format="%d%B%Y")
ggplot(BD2_3, aes(x=Fecha, y=PRECIP))+
  geom_line(col="black") + labs(title="Gráfico 3")
```

VOLVER