Nota: 4.2

UNIVERSIDAD NACIONAL DE COLOMBIA

SERIES DE TIEMPO UNIVARIADAS

Autor:

Felipe Lopera Angel

Jhonatan Smith Garcia

Profesor:

Mauricio Mazp Mazo

2022-02

EJERCICIO 1

20%

Se tiene que:

$X_t = W_{t-2} + 0.5w_{t-1} + 2\ \mathrm{W}_1 + 0,5\ \mathrm{W}_{t+1} + w_{t+2}$ con $\sigma_\omega^2 = 4,8$

$E(x_t) = E(\omega_{t-2}) + 0,5E((\omega_{t-1}) + 2E(\omega_t) + 0,5E(\omega_{t+1}) + E(\omega_{t+2})$

$E(X_t) = 0$

$\operatorname{Var}(X_t) = \operatorname{Var}(\omega_{t-2}) + 0,5^2 \operatorname{Var}(\omega_{t-1}) + 2^2 \operatorname{Var}(\omega_t) + 0,5^2 v_{ar}(w_{t+1}) + \operatorname{Var}(w_{t+2})$

$= \sigma_w^2 \left(\theta_1^2 + \theta_2^2 + \theta_3^2 + \theta_4^2 + \theta_5^2\right)$

Como ACF

$\rho_k = \frac{\gamma(k)}{\gamma(0)}, \gamma(k) = \operatorname{cov}(X_t, X_{t-k})$

Para K=1 se tiene que: $\gamma(k=1) = \operatorname{cov}(X_t, X_{t-1})$

$= \operatorname{cov}(\omega_{t-2} + 0.5\omega_{t-1} + 2\omega_t + 0,5\omega_{t+1} + \omega_{t+2}; w_{t-3} + 0,5\omega_{t-2} + 2\omega_{t-1} + 0.5\omega_t + \omega_{t+1})$

$\theta_1 = 1, \theta_2 = 0,5, \theta_3 = 2, \theta_4 = 0,5, \theta_5 = 0,5$

$\operatorname{Cov}(X_t, X_{t-1}) = (0,5)(1)\operatorname{Cov}(\omega_{t-2}, \omega_{t-2}) + (0,5)(2)\operatorname{Cov}(\omega_{t-1}, \omega_{t-1}) + (2)(0,5)\operatorname{Cov}(\omega_t, \omega_t) + (0,5)(1)\operatorname{Cov}(\omega_{t+1}, \omega_{t+1})$

Los demas terminos se anulan. Por independencia, de esta manera y de manera sistematica:

$$\begin{cases} \operatorname{cov}(\omega_A, \omega_B) \\ A \neq B \rightarrow 0 \\ A = B \rightarrow \sigma_w^2 = 4,8 \end{cases}$$

$$\begin{cases} \sigma_w^2 \left(\theta_1^2 + \theta_2^2 + \theta_3^2 + \theta_4^2 + \theta_5^2\right), k = 0 \\ 6w^2 \left(\theta_1\theta_2 + \theta_2\theta_3 + \theta_3\theta_4 + \theta_4\theta_5\right), k = 1 \\ \sigma_w^2 \left(\theta_1\theta_3 + \theta_2\theta_4 + \theta_3\theta_s\right), k = 2 \end{cases}$$

s?

$$\begin{cases} \sigma_w^2 \left(\theta_1\theta_n + \theta_2\theta_5\right), k = 3 \\ \sigma_w^2 \left(\theta_1\theta_5\right), k = 4 \\ 0, k \geq 5 \end{cases}$$

¿y los valores numéricos?

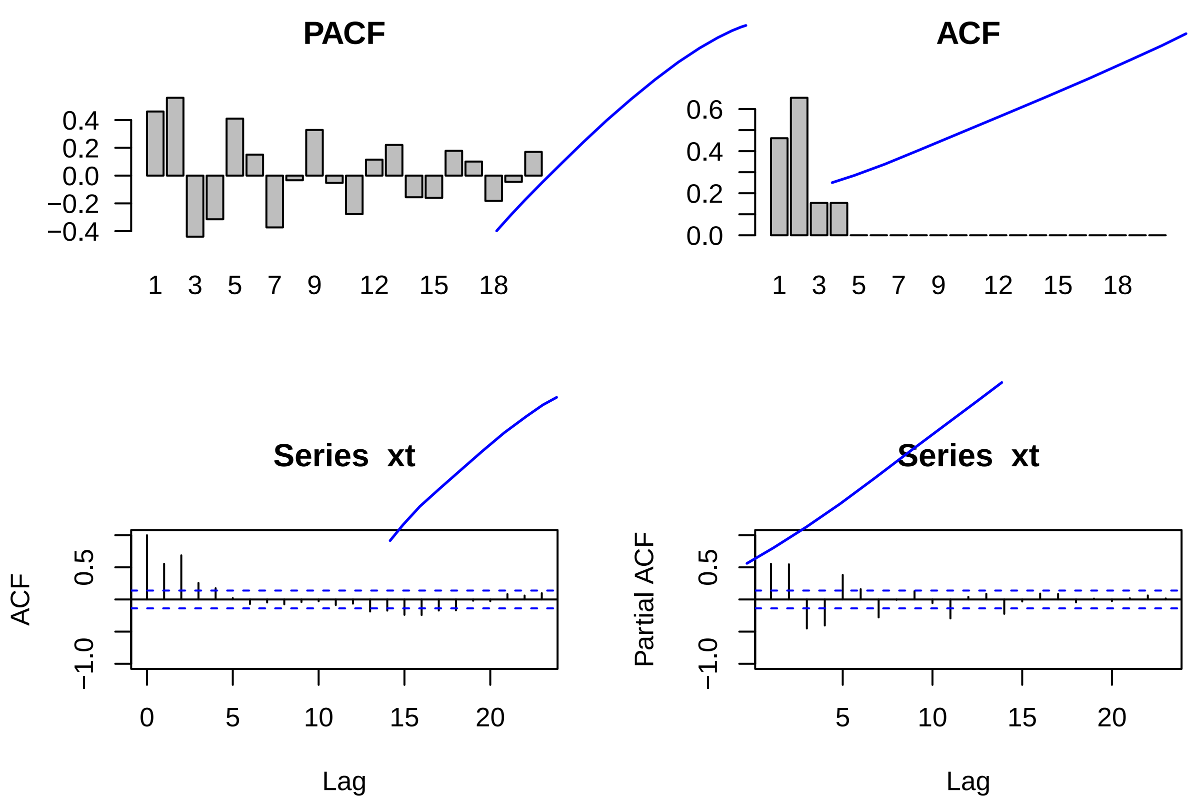
n?

$$
\begin{cases}
1 \rightarrow k = 0 \\
\frac{(\theta_1\theta_2+\theta_2\theta_3+\theta_3\theta_4+\theta_4\theta_5)}{(\theta_1^2+\theta_2^2+\theta_3^2+\theta_1^2+\theta_5^2)}, u = 1 \\
\frac{(\theta_1\theta_3+\theta_2\theta_4+\theta_3\theta_5)}{(\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_5^2)}, k = 2 \\
\frac{(\theta_1\theta_4+\theta_2\theta_5)}{(\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_5^2)}, k = 3 \\
\frac{(\theta_1\theta_5)}{(\theta_1^2+\theta_2^2+\theta_3^2+\theta_n^2+\theta_5^2)}, k = 4 \\
0, k \geq 5
\end{cases}
$$

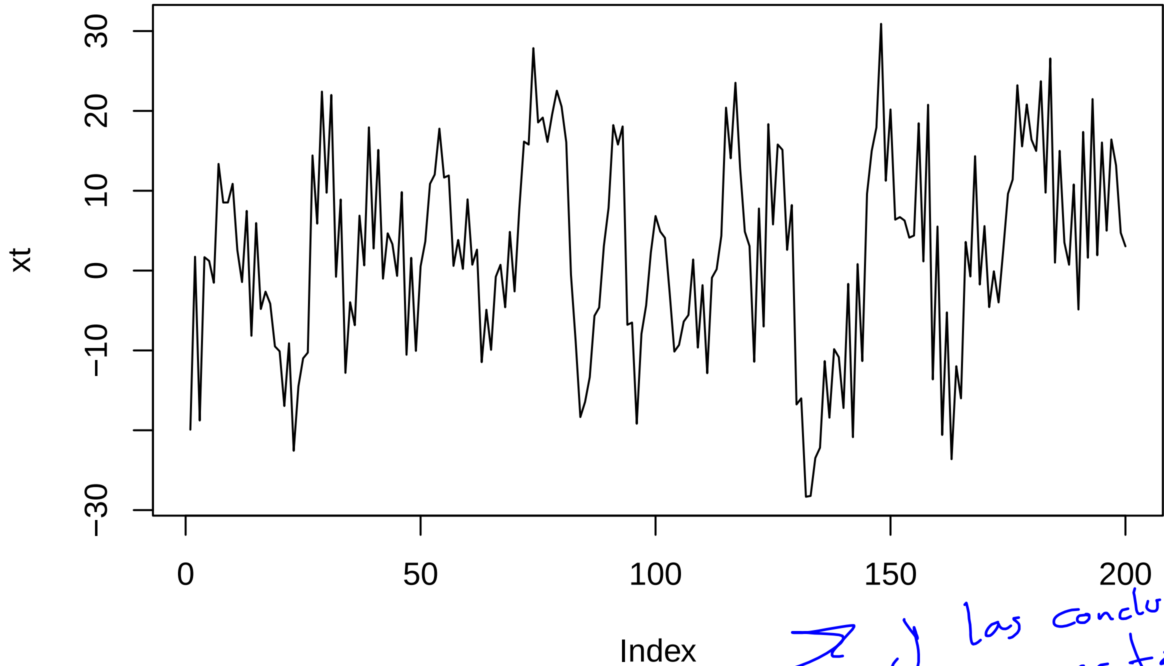
Simulacion

```
##
## Autocorrelations of series 'xt', by lag
##
##      0      1      2      3      4      5      6
##  1.000  0.555  0.687  0.257  0.175  0.021 -0.071

##
## Partial autocorrelations of series 'xt', by lag
##
##      1      2      3      4      5      6
##  0.555  0.547 -0.452 -0.405  0.382  0.161
```



Note que la ACF tiene corte y es consistente con lo que se esperaba dado el modelo teorico analizado. La PACF dado el modelo AR da indicios de un modelo 2 lo cual era esperable.



¿las conclusiones con respecto a lo teórico?

La serie en efecto parece tener un comportamiento ciclico y, con un lag de 2 segun lo estimado por la PACF.

EJERCICIO 2

22%

Se tiene $X_t = 3.1 + 0,9X_{t-1} \cancel{- 0,6X_{t-2}} + w_t$ por tanto, se procede a calcular el polinomio de rezago. $1 = 0,9B - 0,6B^2$, $1 - 0,9B + 0,6B = 0$

$E\left(X_t\right) = 3,1 + 0,9u - 0,6u + 0$ $u = \frac{3,1}{1-0,9+0,6}$ $\quad u = 4,428$

$\operatorname{Var}\left(X_t\right) = \operatorname{var}\left(3,1 + 0,9X_{t-1} - 0,6X_{t-2} + W_t\right)$

$$\begin{aligned}\operatorname{Var}\left(X_t\right) = & 0.9^2 \operatorname{var}\left(X_{t-1}\right) + (-0,6)^2 \operatorname{var}\left(X_{t-2}\right) + \operatorname{Var}\left(X_t\right) + (0,9)(-0,6)\operatorname{Cov}\left(X_{t-1}, X_{t-2}\right) \\ & + 0,9\operatorname{Cov}\left(X_{t-1}, \omega_t\right) - 0,6\operatorname{Cov}\left(X_{t-2}, w_t\right)\end{aligned}$$

$\sigma_x^2 = 0,9^2\sigma_x^2 + (-0,6)^2\sigma_x^2 + \sigma_w^2 + (0,9)(-0,6)\gamma(1)$ $\gamma(0) = \sigma_x^2 = \frac{\sigma_\omega^2 + (0,9)(-0,6)\gamma(1)}{(1-0,9^2-0,6^2)}$

Simulacion

-0.9

```
abs(polyroot(c(1,0.9,-0.6)))
```

El polinomio es

-0.6

```
## [1] 0.7430394 2.2430394
```

$1 - 0.9B + 0.6B^2$

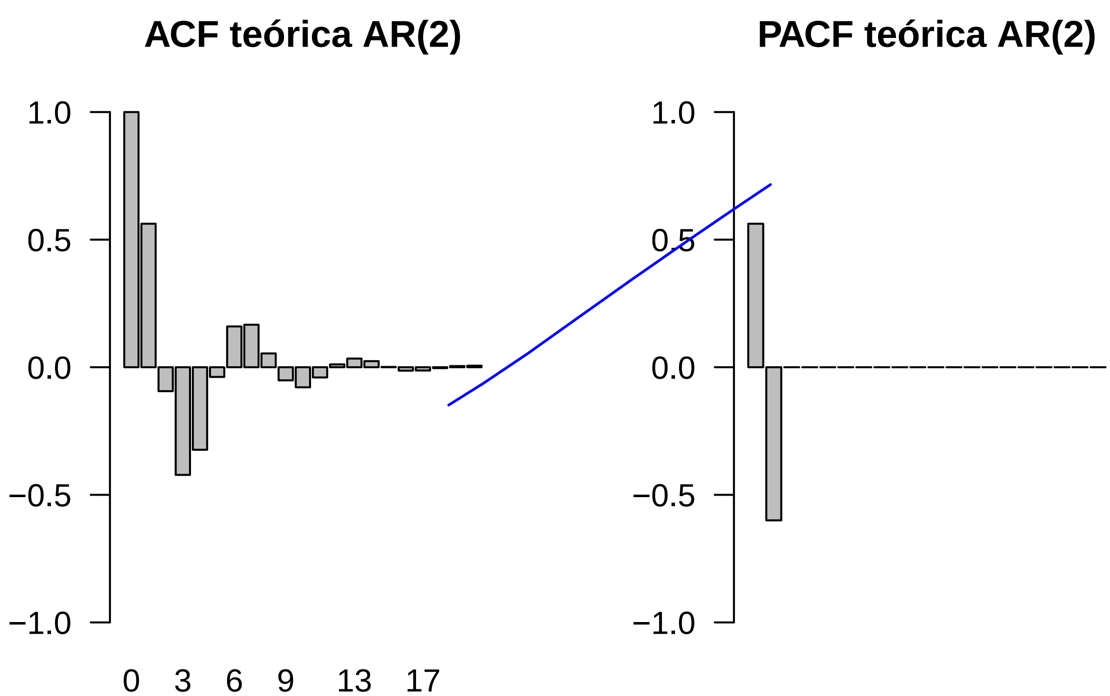
```
options(scipen = 100)
acf1<-ARMAacf(ar=c(0.9,-0.6) , lag.max = 20)
round(acf1,4)
```

```
##       0       1       2       3       4       5       6       7       8       9
##  1.0000  0.5625 -0.0938 -0.4219 -0.3234 -0.0380  0.1599  0.1667  0.0541 -0.0513
##      10      11      12      13      14      15      16      17      18      19
## -0.0787 -0.0400  0.0112  0.0341  0.0239  0.0011 -0.0134 -0.0127 -0.0034  0.0046
##      20
##  0.0061
```

```
pacf1<-ARMAacf(ar=c(0.9,-0.6), pacf=TRUE, lag.max = 20)
round(pacf1,4)
```

```
##  [1]  0.5625 -0.6000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
## [10]  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
## [19]  0.0000  0.0000
```

```
par(mfrow=c(1,2))
barplot(acf1, main="ACF teórica AR(2)", las=1, ylim=c(-1,1))
barplot(pacf1, main="PACF teórica AR(2)", las=1, ylim=c(-1,1))
```

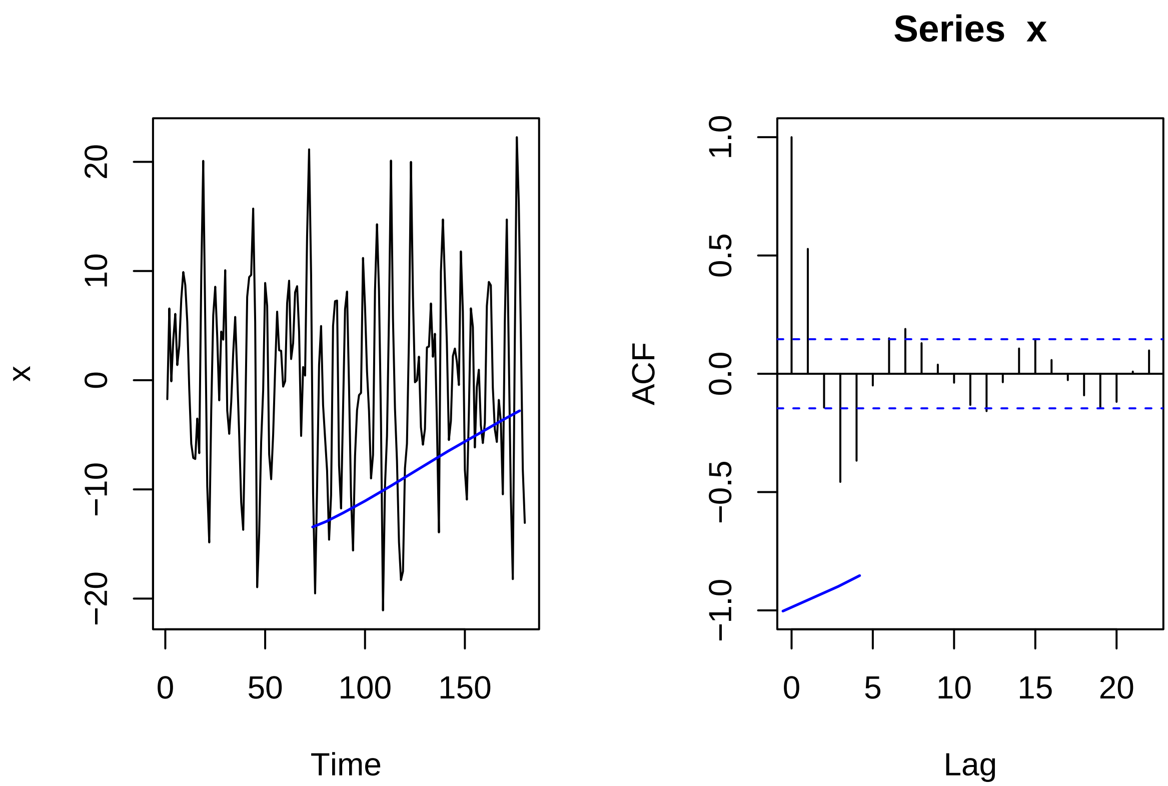



```
set.seed(123)
x <- arima.sim(model=list(ar=c(0.9,-0.6)),n=180,sd=6.2)
plot(x)

acf(x, lag.max = 6, plot=FALSE)
```

```
##
## Autocorrelations of series 'x', by lag
##
##      0      1      2      3      4      5      6
##  1.000  0.528 -0.143 -0.457 -0.367 -0.049  0.150
```

```r
acf(x, ylim=c(-1,1))
```




```r
pacf(x, lag.max = 6, plot=FALSE)
```

```
##
## Partial autocorrelations of series 'x', by lag
##
##      1      2      3      4      5      6
##  0.528 -0.584 -0.062 -0.095  0.060 -0.116
```

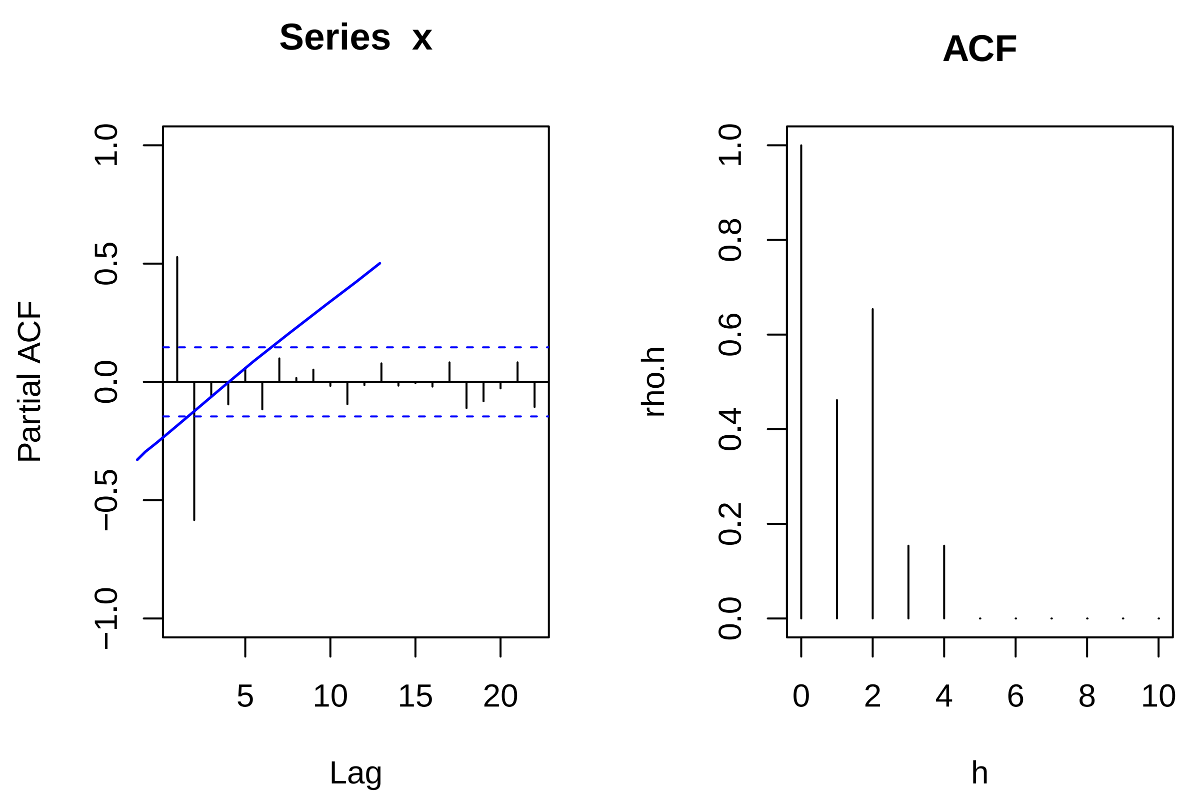
```r
pacf(x, ylim=c(-1,1))


theta <- matrix(c(1, 0.5, 2, 0.5, 1), ncol = 1)

rho <- function(h, sigma.sq) {
  base <- 2:(-2)
  lagged <- base - h
```

```
  aux.matrix <- matrix(0, length(base), length(base))
  row.counter <- 1; col.counter <- 1
  for (i in base) {
    for (j in lagged) {
      aux.matrix[row.counter, col.counter] <- ifelse(i == j, sigma.sq, 0)
      col.counter <- col.counter + 1
    }
    col.counter <- 1
    row.counter <- row.counter + 1
  }
  ans <- t(theta) %*% aux.matrix %*% theta
  return(as.numeric(ans))
}

h <- 0:10
rho.h <- sapply(0:10, rho, sigma.sq = 4.8)/rho(0, 4.8)
plot(h, rho.h, main = "ACF", type = "h")
```



Note que en este caso, se tiee un corte y un decaimiento esponencial. Se puede sospechar que se tabaja con un AR2 dado la teoria y graficamente se constata.

#EJERCICIO 3

A2?

Inciso a

Las dimensiones para cada conjunto de datos son, 3563,3764 y 4122 con 23 variables cada uno.

Inciso b

Al unir las tres bases de datos (sin hacer la transformacion por la variable hora) se tienew 11449 datos con 23 variables

Inciso c

```
datos.juntos$linea= toupper(datos.juntos$linea)

datos.juntos$linea = stri_trans_general(datos.juntos$linea,"Latin-ASCII" ) # se eliminan acentos


datos.juntos$linea=as.factor(datos.juntos$linea)

datos.juntos <- gather(datos.juntos,key="hora",value = "total.hora",3:22)

#Nota para Mauricio: Si se usa la fn pivot longer, las funciones de arreglos tipo unique o distinct dej

datos.juntos$diasem <- weekdays.POSIXt(datos.juntos$fecha)

datos.juntos$mes <- month(datos.juntos$fecha)

datos.juntos$anio <- year(datos.juntos$fecha)

datos.juntos$dia <- day(datos.juntos$fecha) # dia del mes

datos.juntos$diasem <- wday(datos.juntos$fecha,label = T,abbr = F)

datos.juntos$semana <- week(datos.juntos$fecha)

datos.juntos$dia.sem.num <- as.numeric(datos.juntos$diasem)

datos.juntos = datos.juntos %>% distinct()




names(datos.juntos)
```

```
##  [1] "fecha"       "linea"       "total.dia"   "hora"        "total.hora"
##  [6] "diasem"      "mes"         "anio"        "dia"         "semana"
## [11] "dia.sem.num"
```

```
dim(datos.juntos)
```

```
## [1] 228980     11
```

Este es el resultado despues de unir y cerar las variables nuevas. FInalmente se tiene 228980 datos con 11 variables.

#Inciso d

El dataframe del inciso d cumple las condiciones del inciso f, en consecuencia se usan ambos como solucion del problema

```
## [1] 21860    11
```

```
## [1] 21860    11
```

Ya se encuentran ordenados

```
dat_lin_A %>% head(3)
```

```
##        fecha   linea total.dia hora total.hora    diasem mes anio dia semana
## 1 2019-01-01 LINEA A    183664    4         14    martes   1 2019   1      1
## 2 2019-01-02 LINEA A    520286    4       8890 miércoles   1 2019   2      1
## 3 2019-01-03 LINEA A    563849    4       9879    jueves   1 2019   3      1
##   dia.sem.num
## 1           3
## 2           4
## 3           5
```

Inciso e

```
# Fechas pedidas
linea.a.antes <- filter(dat_lin_A, fecha <="2020-03-23")
linea.a.despues <- filter(dat_lin_A, fecha >"2020-03-23")
linea.b.antes <- filter(dat_lin_B, fecha <="2020-03-23")
linea.b.despues <-filter(dat_lin_A, fecha >"2020-03-23")
```

```
# promedios pedidos

prom.l.a.antes <-linea.a.antes %>% group_by(diasem,hora) %>%
summarise(prom_dia_hora = mean(`total.hora`))
prom.l.a.antes %>% head()
```

```
## # A tibble: 6 x 3
## # Groups:   diasem [1]
##   diasem  hora  prom_dia_hora
##   <ord>   <chr>         <dbl>
## 1 domingo 10           15953.
## 2 domingo 11           15996.
```

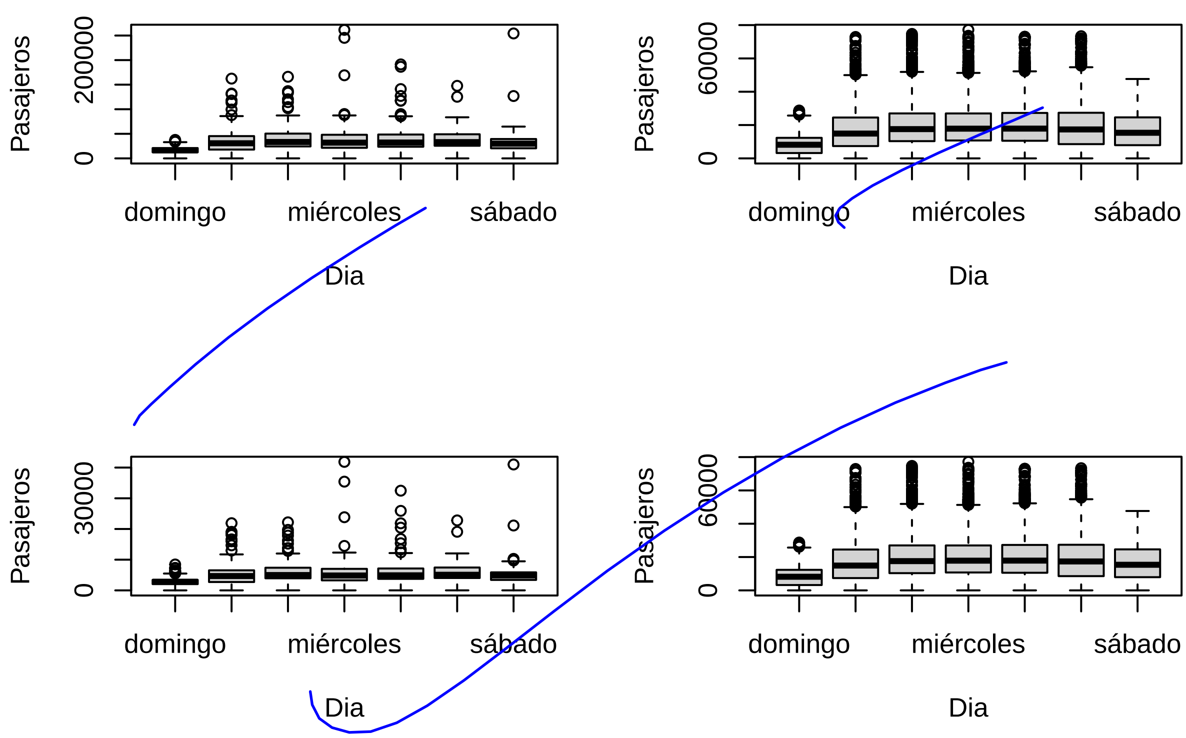
```
## 3 domingo 12           17786.
## 4 domingo 13           20108.
## 5 domingo 14           21689.
## 6 domingo 15           21031.
```

Este proceso se replica para cada linea dado el tiempo pedido.

Analisis descriptivo

Como era de esperarse, el numero de pasajeros que usan el metro depende por supuesto, del dia especifico de la semana. Es por esto que, El domingo tiene un boxplot más pequeño, siendo consistente con la informacion que da la intuicion.

```
par(mfrow=c(2,2))
boxplot(linea.a.antes$`total.hora`~linea.a.antes$diasem, xlab = "Dia", ylab = "Pasajeros")
boxplot(linea.a.despues$`total.hora`~linea.a.despues$diasem, xlab = "Dia", ylab = "Pasajeros")
boxplot(linea.b.antes$`total.hora`~linea.b.antes$diasem, xlab = "Dia", ylab = "Pasajeros")
boxplot(linea.b.despues$`total.hora`~linea.b.despues$diasem, xlab = "Dia", ylab = "Pasajeros")
```



#Inciso f

```
l.A <- filter(datos.juntos, linea == "LINEA A")
l.A %>% names
```

```
##  [1] "fecha"        "linea"        "total.dia"    "hora"         "total.hora"
##  [6] "diasem"       "mes"          "anio"         "dia"          "semana"
## [11] "dia.sem.num"
```

```
l.B <- filter(datos.juntos, linea == "LINEA A")
l.B %>% names
```

```
##  [1] "fecha"        "linea"        "total.dia"    "hora"         "total.hora"
##  [6] "diasem"       "mes"          "anio"         "dia"          "semana"
## [11] "dia.sem.num"
```

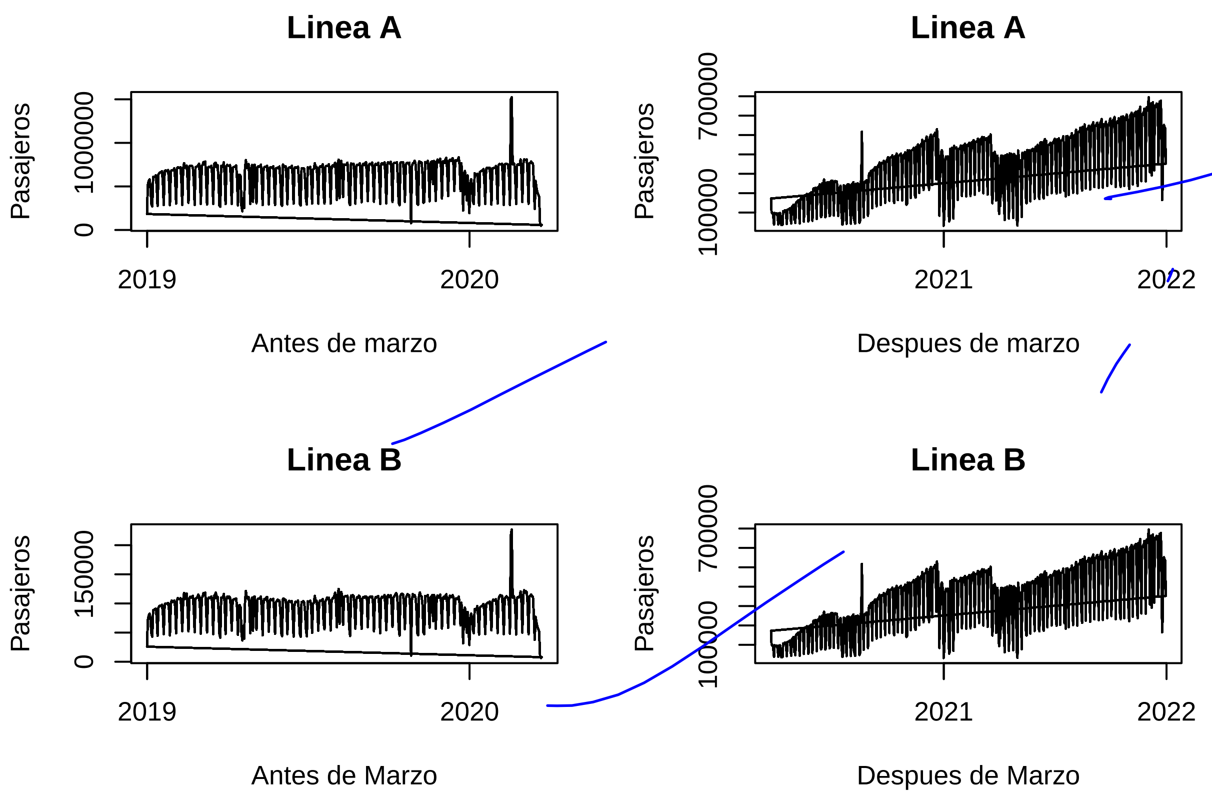
Inciso g

```
par(mfrow=c(2,2))
plot(linea.a.antes$fecha, linea.a.antes$total.dia, type ="l", xlab = "Antes de marzo", main = "Linea A"

plot(linea.a.despues$fecha, linea.a.despues$total.dia, type ="l", xlab="Despues de marzo", main = "Linea

plot(linea.b.antes$fecha, linea.b.antes$total.dia, type ="l", xlab = "Antes de Marzo", main = "Linea B"

plot(linea.b.despues$fecha, linea.b.despues$total.dia, type ="l", xlab = "Despues de Marzo", main = "Li
```



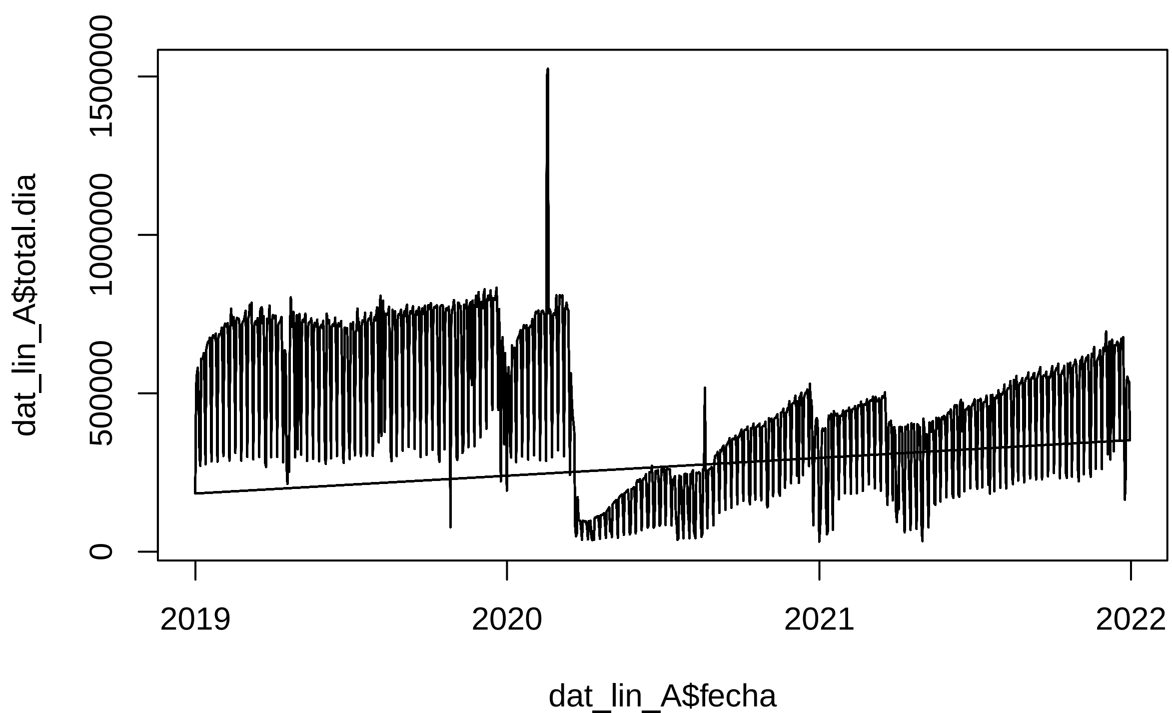
Para ambas lineas el comportamiento tanto en volumen como en tendencia y varianza parece similar. De hecho, ambas parecen replicas de la misma serie. Por esto se dospecha que la linea de metro no influye en el comportamiento del numero total de pasajeros que ingresan a el sistema.

Ahora bien, hay un dato en particular de una fecha que tiene un valor muy alto. ¿Este outlier es representativo? ¿De cuando es? Corresponde a las fechas de alerta ambiental de febrero del 2019. Por esta razón se decide no eliminar el dato.

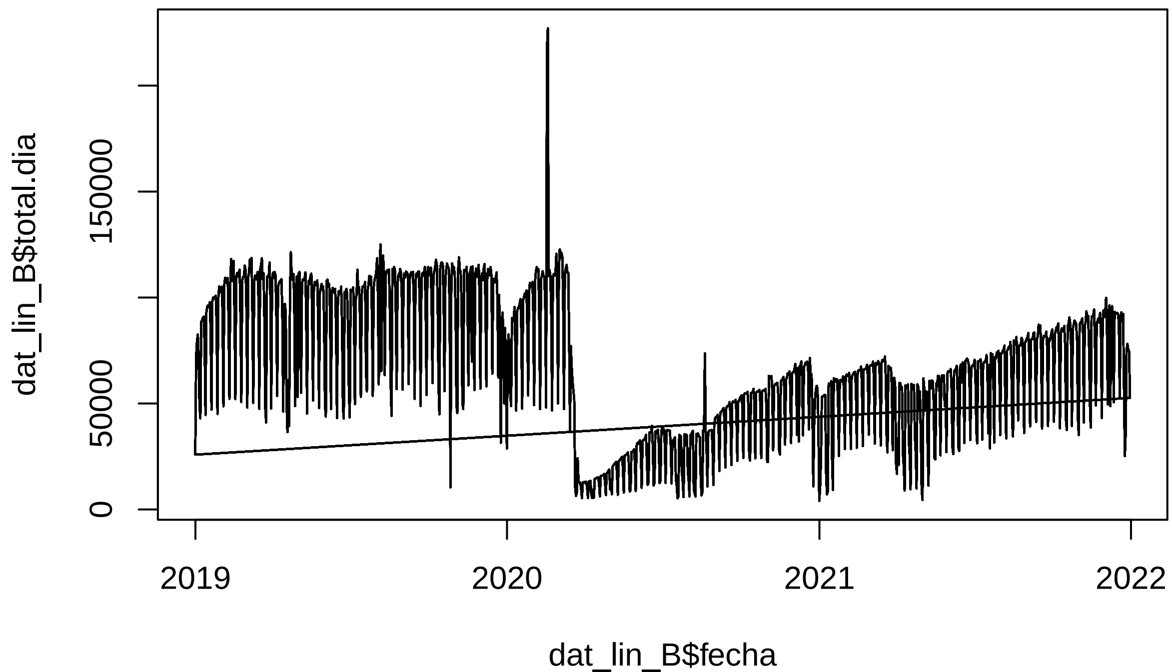
Antes del 23 de marzo del 2020 y despues del 23 de marzo del 2020 hay un punto de quiebre. Esto es dado que para esta fecha se decreta a nivel nacional estado de alerta por la pandemia de COVID-19. Se nota como el volumen de pasajeros cayó drasticamente a causa la pandemia. Luego a partir de dicha fecha la serie toma una tendencia creciente. ¿por qué? Tenga presente que la media de pasajeros existentes antes de pandemia era constante, oscilaba alrededor de un mismo valor pero,a causa de la pandemia, dicho valor decayó. Así, la idea del sistema metro ha sido refidelizar a los clientes y lo ha hecho, de manera progresiva, esto explica los patrones de tendencia creciente.

De alguna manera se podria ver un comportamiento estacional, sin embargo se requiere pruebas analiticas. Se sopecha dicho comportamiento es semanal, entre dias.

```
#par(mfrow=c(1,2))
plot(dat_lin_A$fecha, dat_lin_A$total.dia, type ="l")
```




```
plot(dat_lin_B$fecha, dat_lin_B$total.dia, type ="l")
```



#Inciso h

Se debe analizar el comportamiento estacional y de tendencia. Se procede a realizar analisis de un mes particular.

```
linea.a.antes <- filter(dat_lin_A, fecha <="2020-03-23")
linea.a.despues <- filter(dat_lin_A, fecha >"2020-03-23")
linea.b.antes <- filter(dat_lin_B, fecha <="2020-03-23")
linea.b.despues <-filter(dat_lin_A, fecha >"2020-03-23")
```

Note que las series antes de pandemia tienen un comportamiento lineal por tanto, se intenta ajustar un modelo del total de pasajeros por dia en funcion del dia de la semana y el mes. Esto se hace con una recta, nuevamente por lo ya mencionado.

```
mod1 <- lm(total.dia~diasem+mes, data= linea.a.antes)

summary(mod1)
```

```
##
## Call:
## lm(formula = total.dia ~ diasem + mes, data = linea.a.antes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -581529   -5021   18340   45104  866919
##
```

```
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  595585.0     2379.8 250.271 < 0.0000000000000002 ***
## diasem.L     187301.2     3381.2  55.395 < 0.0000000000000002 ***
## diasem.Q    -293435.9     3381.1 -86.786 < 0.0000000000000002 ***
## diasem.C      92696.1     3381.2  27.416 < 0.0000000000000002 ***
## diasem^4    -113475.6     3381.1 -33.562 < 0.0000000000000002 ***
## diasem^5     -11900.8     3381.1  -3.520            0.000434 ***
## diasem^6       9149.5     3381.1   2.706            0.006821 **
## mes            6453.7      354.1  18.227 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121000 on 8952 degrees of freedom
## Multiple R-squared:  0.5885, Adjusted R-squared:  0.5882
## F-statistic:  1829 on 7 and 8952 DF,  p-value: < 0.00000000000000022
```

Para cada uno de los parametros del modelo, se tiene que son significativos. Ademas, el R^2 ajustado es de 0.58 lo que implica que este modelo explica aproximadamente un 58% de la variabilidad total de estos datos (Linea A antes de pandemia)

```
mod2 <- lm(total.dia~diasem+mes, data= linea.a.despues)
summary(mod2)
```

```
##
## Call:
## lm(formula = total.dia ~ diasem + mes, data = linea.a.despues)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397562 -113619   27194  117116  214991
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   211073       2830  74.585 <0.0000000000000002 ***
## diasem.L      108333       3090  35.055 <0.0000000000000002 ***
## diasem.Q     -171989       3088 -55.702 <0.0000000000000002 ***
## diasem.C       52960       3082  17.181 <0.0000000000000002 ***
## diasem^4      -41252       3075 -13.415 <0.0000000000000002 ***
## diasem^5        4046       3075   1.316               0.188
## diasem^6        3184       3070   1.037               0.300
## mes            19151        363  52.753 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 132200 on 12892 degrees of freedom
## Multiple R-squared:  0.3726, Adjusted R-squared:  0.3723
## F-statistic:  1094 on 7 and 12892 DF,  p-value: < 0.00000000000000022
```

Revisar esta codificación

Note que, el nivel de significancia de dos parametros decae, lo que implica que este modelo ha decaido en efectividad para captar la tendencia de los datos.

Esto es cierto, dado que se desempeña mucho peor, pues al analizar el R^2 ajustado se ve que apenas abarca cerca de 37% de la variabilidad de los datos (Linea A despues de pandemia).

¿donde están
los de la línea
B?

Esto era esperable puesto que, de alguna manera en un primer analisis descriptiv era necesario constatar una tendencia creciente que, una linea recta puede ser pobre a la hora de explicarla. Estos analisis se extrapolan exactamente iguales para la linea B.

Las covariables seleccionadas para la linea A y B fueron: dia de la semana y mes; el dia de la semana fue seleccionado ya que hay dias puntuales en que cambia de manera constante la cantidad de pasajeros en el metro. Es esperable segun los datos y los analisis descriptivos hasta el momento que el numero total de pasajeros depende en gran medida del dia de la semana. Por este motivo se espera que el dia de la semana sea una variable importante a la hora de tomar deciciones.

Por este motivo se selecciona este modelo para la implementacion de modelamientos de tendencia.