Parcial 1

Julián Úsuga Ortiz - Ivan Santiago Rojas Martinez

El código puede verse en el siguiente repositorio.

Punto 1.

Sea X_t el proceso estocástico dado por:

$$X_t = w_{t-2} + 0.5w_{t-1} + 2w_t + 0.5w_{t+1} + w_{t+2}$$

donde los w_t son independientes con media 0 y varianza $\sigma_w^2 = 4.8$.

a)

$$\begin{aligned} E[X_t] &= E[w_{t-2} + 0.5w_{t-1} + 2w_t + 0.5w_{t+1} + w_{t+2}] \\ &= 0 \end{aligned}$$

$$\begin{aligned} Var[X_t] &= Var[w_{t-2} + 0.5w_{t-1} + 2w_t + 0.5w_{t+1} + w_{t+2}] \\ &= Var[w_{t-2}] + 0.5^2 Var[w_{t-1}] + 2^2 Var[w_t] + 0.5^2 Var[w_{t+1}] + Var[w_{t+2}] \\ &= 4.8 + 0.25 * 4.8 + 4 * 4.8 + 0.25 * 4.8 + 4.8 \\ &= 31.2 \end{aligned}$$

b)

- k = 0

$$\begin{aligned}\gamma(0) &= Cov[X_t, X_t] \\ &= Var[X_t] = 31.2\end{aligned}$$

- k = 1

$$\begin{aligned}\gamma(1) &= Cov[X_t, X_{t+1}] \\ &= Cov[w_{t-2} + 0.5w_{t-1} + 2w_t + 0.5w_{t+1} + w_{t+2}, w_{t-1} + 0.5w_t + 2w_{t+1} + 0.5w_{t+2} + w_{t+3}] \\ &= Cov[w_{t-2}, X_{t+1}] + Cov[0.5w_{t-1}, w_{t-1}] + \\ &\quad Cov[2w_t, 0.5w_t] + Cov[0.5w_{t+1}, 2w_{t+1}] + \\ &\quad Cov[w_{t+2}, 0.5w_{t+2}] \\ &= 0 + 0.5 * 4.8 + \\ &\quad 2 * 0.5 * 4.8 + 2 * 0.5 * 4.8 + \\ &\quad 0.5 * 4.8 \\ &= 14.4\end{aligned}$$

- k = 2

$$\begin{aligned}\gamma(2) &= Cov[X_t, X_{t+2}] \\ &= Cov[w_{t-2} + 0.5w_{t-1} + 2w_t + 0.5w_{t+1} + w_{t+2}, w_t + 0.5w_{t+1} + 2w_{t+2} + 0.5w_{t+3} + w_{t+5}] \\ &= Cov[w_{t-2}, X_{t+2}] + Cov[0.5w_{t-1}, X_{t+2}] + \\ &\quad Cov[2w_t, w_t] + Cov[0.5w_{t+1}, 0.5w_{t+1}] + \\ &\quad Cov[w_{t+2}, 2w_{t+2}] \\ &= 0 + 0 + \\ &\quad 2 * 4.8 + 0.5^2 * 4.8 + \\ &\quad 2 * 4.8 \\ &= 20.4\end{aligned}$$

- k = 3

$$\begin{aligned}\gamma(-2) &= Cov[X_t, X_{t+3}] \\ &= Cov[w_{t-2} + 0.5w_{t-1} + 2w_t + 0.5w_{t+1} + w_{t+2}, w_{t+1} + 0.5w_{t+2} + 2w_{t+3} + 0.5w_{t+4} + w_{t+5}] \\ &= Cov[w_{t-2} + 0.5w_{t-1} + 2w_t, w_{t+1} + 0.5w_{t+2} + 2w_{t+3} + 0.5w_{t+4} + w_{t+5}] + Cov[0.5w_{t+1}, w_{t+1}] + \\ &\quad Cov[w_{t+2}, 0.5w_{t+2}] \\ &= 0 + 0.5 * 4.8 + \\ &\quad 0.5 * 4.8 \\ &= 4.8\end{aligned}$$

- k = 4

$$\begin{aligned}\gamma(4) &= Cov[X_t, X_{t+4}] \\ &= Cov[w_{t-2} + 0.5w_{t-1} + 2w_t + 0.5w_{t+1} + w_{t+2}, w_{t+2} + 0.5w_{t+3} + 2w_{t+4} + 0.5w_{t+5} + w_{t+6}] \\ &= Cov[w_{t-2} + 0.5w_{t-1} + 2w_t + 0.5w_{t+1}, w_{t+2} + 0.5w_{t+3} + 2w_{t+4} + 0.5w_{t+5} + w_{t+6}] + \\ &\quad Cov[w_{t+2}, w_{t+2}] \\ &= 0 + 4.8 \\ &= 4.8\end{aligned}$$

Se puede probar que $\gamma(i) = \gamma(-i)$ para $i = 1, 2, 3, 4$

Y que $\gamma(k) = 0$ para $|k| \geq 5$

Con los valores anteriores y reemplazando, la ACF está dada por

$$\rho(k) = \begin{cases} 1 & \text{si } k = 0 \\ 0.4615385 & \text{si } |k| = 1 \\ 0.6538462 & \text{si } |k| = 2 \\ 0.1538462 & \text{si } |k| = 3 \\ 0.1538462 & \text{si } |k| = 4 \\ 0 & \text{si } |k| \geq 5 \end{cases} \tag{1}$$

y la PACF por:

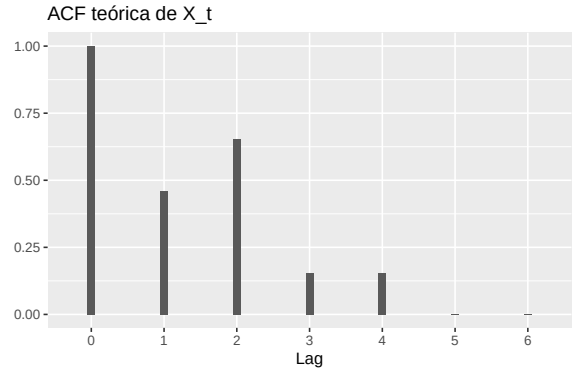
$$\begin{aligned} \Phi_{0,0} &= 0 & \Phi_{1,1} &= 0.4615385 & \Phi_{2,2} &= 0.5601504 \\ \Phi_{3,3} &= -0.4396252 & \Phi_{4,4} &= -0.3143931 & \Phi_{5,5} &= 0.4101445 \end{aligned}$$

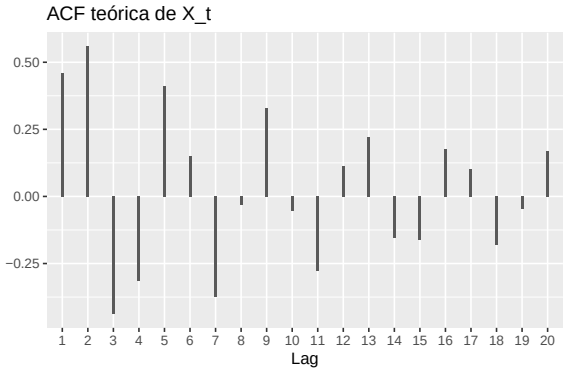
A continuación los valores de la PACF desde el Lag 6 hasta el 20:

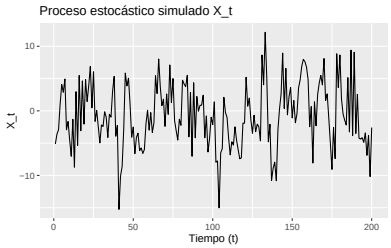
0.1509651, -0.3730572, -0.0335281, 0.3285716, -0.0527841, -0.2773402, 0.1148472, 0.2206910, -0.1560573, -0.1608494, 0.1782139, 0.1010639, -0.1825514, -0.0455188, 0.1705496

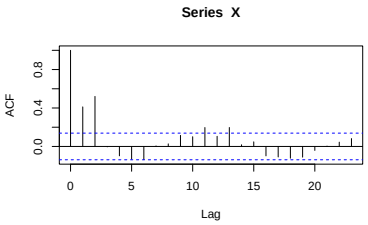
Como $\Phi_{1,1} > 0$ y el modelo es MA entonces el PACF va a oscilar alrededor del 0 de forma senoidal.

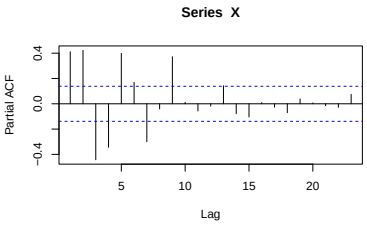
Las ACF y PACF se pueden ver en la siguiente gráfica:











c) y d)

Simulación del proceso estocástico X_t y sus ACF y PACF muestrales.

Al comparar las ACF como las PACF muestrales con las teóricas se puede ver una similitud grande, además, se puede ver que la serie se corta después del Lag 3 en la ACF y se reduce infinitamente de forma senoidal en la PACF, propiedades de un modelo MA(3).

2.

Sea X_t el proceso estocástico estacionario dado por:

$$X_t = 3.1 + 0.9X_{t-1} - 0.6X_{t-2} + w_t$$

donde w_t es ruido blanco gaussiano con media 0 y varianza $\sigma_w^2 = 6.2$.

a)

$$\begin{aligned}
X_t &= 3.1 + 0.9X_{t-1} - 0.6X_{t-2} + w_t \\
\mu_t &= 3.1 + 0.9\mu_t - 0.6\mu_t \\
\mu_t - 0.9\mu_t + 0.6\mu_t &= 3.1 \\
(1 - 0.9 + 0.6)\mu_t &= 3.1 \\
\mu_t &= \frac{3.1}{1 - 0.9 + 0.6} = 4.4285
\end{aligned}$$

b)

$$\begin{aligned}
var[X_t] = cov[X_t, X_t] &= cov[X_t, 3.1 + 0.9X_{t-1} - 0.6X_{t-2} + w_t] \\
&= 0.9cov[X_t, X_{t-1}] * 0.6cov[X_t, X_{t-2}] + cov[X_t, w_t] \\
&= 0.9\gamma(-1) - 0.6\gamma(-2) + \sigma_w^2
\end{aligned}$$

Como $var[X_t] = \gamma(0) = \sigma_x^2$

$$\begin{aligned}
&= 0.9\gamma(-1) - 0.6\gamma(-2) \\
\gamma(0) &= 0.9\gamma(-1) - 0.6\gamma(-2)
\end{aligned}$$

Dividiendo por $\gamma(0)$

$$
\begin{aligned}
1 &= 0.9\frac{\gamma(-1)}{\gamma(0)} - 0.6\frac{\gamma(-2)}{\gamma(0)} + \frac{\sigma_w^2}{\gamma(0)} \\
1 &= 0.9\rho(1) - 0.6\rho(2) + \frac{\sigma_w^2}{\gamma(0)} \\
1 - 0.9\rho(1) + 0.6\rho(2) &= \frac{\sigma_w^2}{\rho(0)} \\
\sigma_x^2 = \gamma(0) &= \frac{\sigma_w^2}{1 - 0.9\rho(1) + 0.6\rho(2)} \\
\gamma(1) &= cov[X_{t-1}, 3.1 + 0.9X_{t-1} - 0.6X_{t-2} + w_t] \\
&= cov[X_{t-1}, X_{t-1}]0.9 - 0.6cov[X_{t-1}, X_{t-2}] \\
&= \gamma(0)0.9 - 0.6\gamma(1) \\
1.6\gamma(1) &= \gamma(0) * 0.9
\end{aligned}
$$

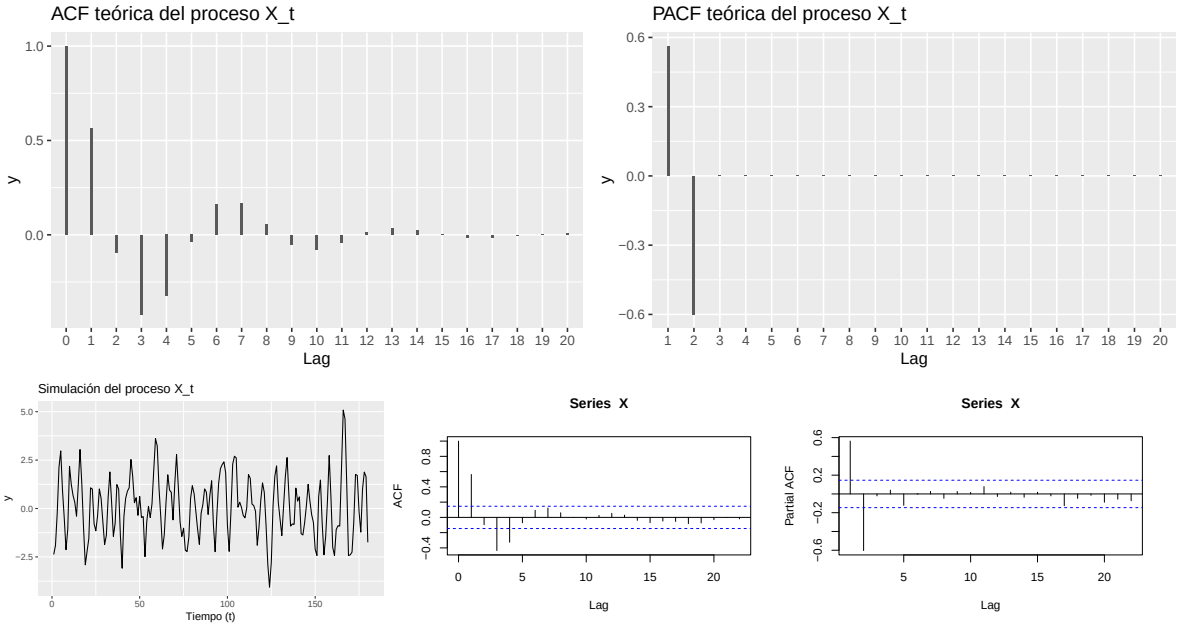
$$
\begin{aligned}
\gamma(1) &= \gamma(0)\frac{0.9}{1.6} \\
\rho(1) &= \frac{\gamma(0)0.9/1.6}{\gamma(0)} = \frac{0.9}{1.6} \\
\gamma(2) &= cov[X_{t-2}, 3.1 + 0.9X_{t-1} - 0.6X_{t-2} + w_t] \\
&= cov[X_{t-2}, X_{t-1}]0.9 - 0.6cov[X_{t-2}, X_{t-2}] \\
&= \gamma(1)0.9 - 0.6\gamma(0) \\
&= \gamma(0)\frac{0.9^2}{1.6} - 0.6\gamma(0) \\
\rho(2) &= \frac{\gamma(2)}{\gamma(0)} = \frac{0.9^2}{1.6} - 0.6 = -0.09375
\end{aligned}
$$

Por lo que:

$$
\begin{aligned}
\gamma(0) &= \frac{6.2}{1 - \frac{0.9^2}{1.6} + 0.6(-0.09375)} \\
\gamma(0) &= \sigma_x^2 \approx 14.1714
\end{aligned}
$$

c) y d)

Al comparar las ACF y PACF muestrales con las ACF y PACF teóricas se observa una gran similitud, además se observa una caída de forma senoidal en la ACF y un corte después del Lag p = 3, caracteristicas de un proceso autoregresivo de orden 2.



3.

a)

Las dimensiones de los datos de 2019 son 3563 filas y 23 columnas.

Las dimensiones de los datos de 2020 son 3764 filas y 23 columnas.

Las dimensiones de los datos de 2021 son 4122 filas y 23 columnas.

b)

Las dimensiones de los datos_juntos son 11449 filas y 23 columnas.

c)

Las dimensiones del data frame son 228980 filas y 10 columnas.

d)

Table 1: Línea A

fecha	linea	per.dia.tot	hora	per.num	dia	dia.sem	sem	mes	anio
2019-01-01	línea a	183664	4H 0M 0S	14	1	Mar	1	1	2019
2019-01-01	línea a	183664	5H 0M 0S	4554	1	Mar	1	1	2019
2019-01-01	línea a	183664	6H 0M 0S	5696	1	Mar	1	1	2019
2019-01-01	línea a	183664	7H 0M 0S	4893	1	Mar	1	1	2019
2019-01-01	línea a	183664	8H 0M 0S	4399	1	Mar	1	1	2019

Las dimensiones de los datos de la línea A son 21860 filas y 10 columnas.

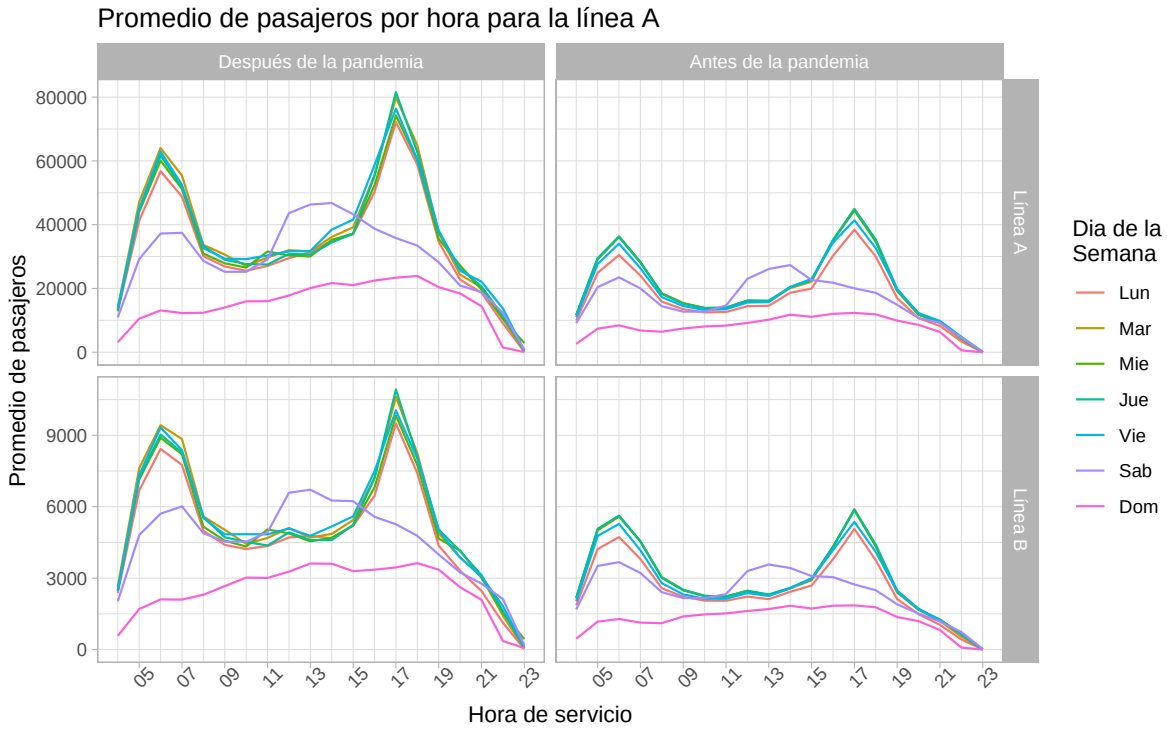
Table 2: Linea B

fecha	linea	per.dia.tot	hora	per.num	dia	dia.sem	sem	mes	anio
2019-01-01	línea b	25852	4H 0M 0S	NA	1	Mar	1	1	2019
2019-01-01	línea b	25852	5H 0M 0S	678	1	Mar	1	1	2019
2019-01-01	línea b	25852	6H 0M 0S	833	1	Mar	1	1	2019
2019-01-01	línea b	25852	7H 0M 0S	682	1	Mar	1	1	2019
2019-01-01	línea b	25852	8H 0M 0S	676	1	Mar	1	1	2019

Las dimensiones de los datos de la línea B son 21860 filas y 10 columnas.

e)

El 23 de Marzo de 2020 el Gobierno de Colombia expidió el Decreto 457 para el período de aislamiento preventivo obligatorio a causa del virus del COVID-19.



Se puede observar que después de la pandemia se disminuyó aproximadamente a la mitad la afluencia de pasajeros en la **línea A** y durante las semanas el comportamiento es muy similar para antes de la pandemia y después de la pandemia.

Los dias con menor afluencia de pasajeros son los dias **sábado** y **domingo**

Luego, se puede observar que después de la pandemia se disminuyó aproximadamente a la mitad la afluencia de pasajeros en la **línea B** y durante las semanas el comportamiento es muy similar para antes de la pandemia y después de la pandemia. Los dias con menor afluencia de pasajeros son los dias **sábado** y **domingo**

f)

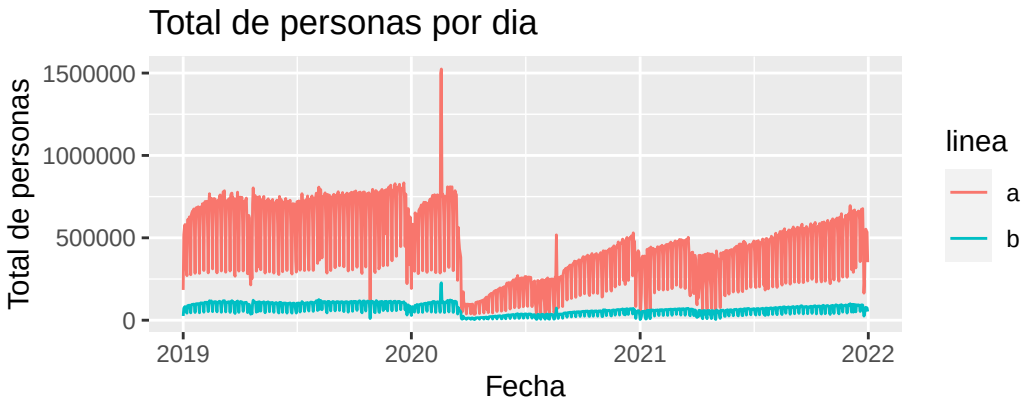
Table 3: Linea A

fecha	dia	dia.sem	sem	mes	anio	per.dia.tot
2019-01-01	1	Mar	1	1	2019	183664
2019-01-02	2	Mie	1	1	2019	520286
2019-01-03	3	Jue	1	1	2019	563849
2019-01-04	4	Vie	1	1	2019	580689
2019-01-05	5	Sab	1	1	2019	478900
2019-01-06	6	Dom	1	1	2019	298762

Table 4: Linea B

fecha	dia	dia.sem	sem	mes	anio	per.dia.tot
2019-01-01	1	Mar	1	1	2019	25852
2019-01-02	2	Mie	1	1	2019	71880
2019-01-03	3	Jue	1	1	2019	79431
2019-01-04	4	Vie	1	1	2019	82533
2019-01-05	5	Sab	1	1	2019	71938
2019-01-06	6	Dom	1	1	2019	48582

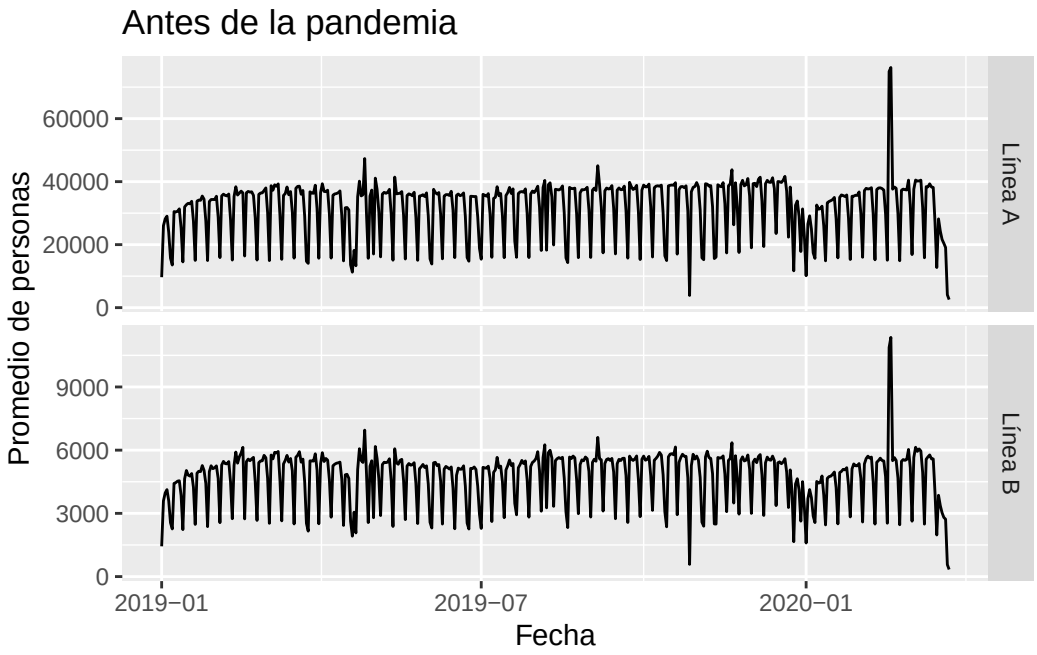
g)



Se ve una caída**(cisne negro)** en Marzo de 2020 y una lenta recuperación, no se ha llegado a niveles de 2019.

Se observa que la línea que cuenta con mayor afluencia de pasajeros es la **línea A**.

h)



Para antes de la pandemia la línea A y B tuvieron un comportamiento muy similar.

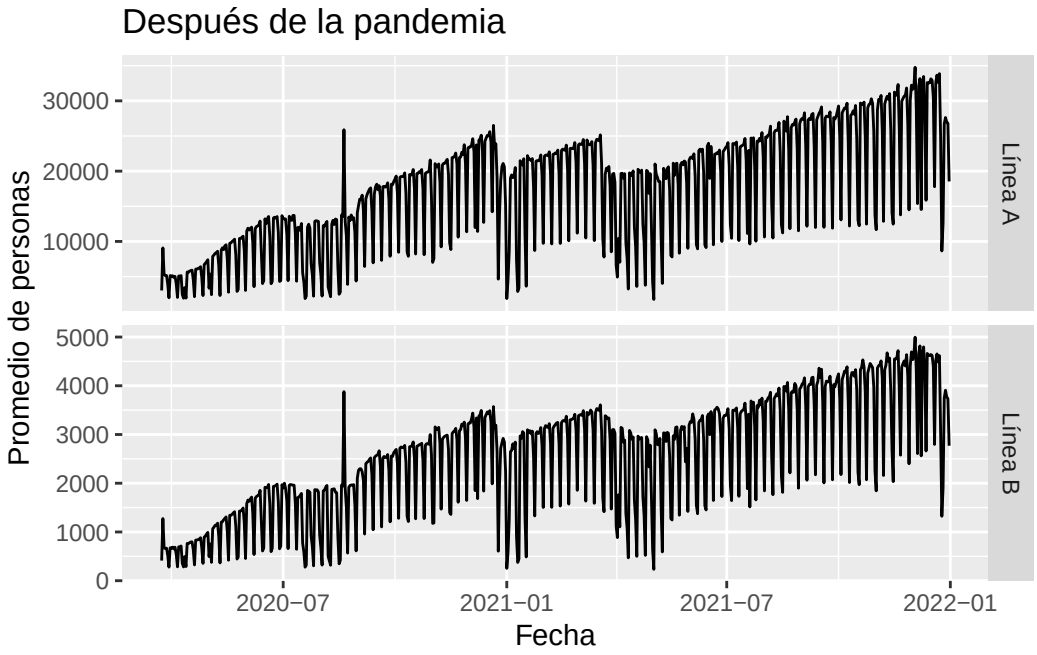
Se ve que es estacionaria alrededor de una media, además, en la ultima semana hay una baja afluencia de personas en ambas líneas muy probablemente a causa del COVID-19.

En esta serie de tiempo se ve una recuperación en ambas líneas (A y B), no es estacionaria ya que se ve una clara tendencia de crecimiento positivo no lineal.

i)

Se plantea el siguiente modelo ya que en el análisis descriptivo se puede ver que semana a semana hay una tendencia, cuando se acerca el fin de semana hay una reducción

Tambien se debe notar que las series para después de la pandemia no tiene una varianza constante, por lo que este comportamiento no es posible de capturar por un modelo de regresión lineal simple sin no antes hacer transformaciones a los datos.



$$
\begin{aligned}
\text{Numero Pasajeros} = \beta_0 &+ \beta_{dia.sem.Martes} I_{dia.sem.Martes} + \\
&\beta_{dia.sem.Miercoles} I_{dia.sem.Miercoles} + \beta_{dia.sem.Jueves} I_{dia.sem.Jueves} + \\
&\beta_{dia.sem.Viernes} I_{dia.sem.Viernes} + \beta_{dia.sem.Sabado} I_{dia.sem.Sabado} + \\
&\beta_{dia.sem.Domingo} I_{dia.sem.Domingo} + \beta_{mes.1} I_{mes.1} + \\
&\beta_{mes.2} I_{mes.2} + \beta_{mes.3} I_{mes.3} + \\
&\beta_{mes.4} I_{mes.4} + \beta_{mes.5} I_{mes.5} + \\
&\beta_{mes.6} I_{mes.6} + \beta_{mes.7} I_{mes.7} + \\
&\beta_{mes.8} I_{mes.8} + \beta_{mes.9} I_{mes.9} + \\
&\beta_{mes.10} I_{mes.10} + \beta_{mes.11} I_{mes.11} + \\
&\beta_{mes.12} I_{mes.12}
\end{aligned}
$$

Modelo lineal A antes de pandemia:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-182854.66738	192407.41977	-0.9503514	0.3419595
fecha	42.84251	10.63082	4.0300300	0.0000562
dia.semMar	68183.78701	4499.10909	15.1549531	0.0000000

	Estimate	Std. Error	t value	Pr(>|t|)
dia.semMie	35152.51580	4501.62336	7.8088532	0.0000000
dia.semJue	50143.92608	4502.77676	11.1362230	0.0000000
dia.semVie	67530.16820	4502.27813	14.9991107	0.0000000
dia.semSab	-78713.91861	4502.31195	-17.4829997	0.0000000
dia.semDom	-363435.52323	4500.67220	-80.7513871	0.0000000
mes2	119993.68328	4665.62799	25.7186564	0.0000000
mes3	52164.51476	4753.50257	10.9739111	0.0000000
mes4	42827.43037	5725.18045	7.4805381	0.0000000
mes5	72354.89280	5614.48196	12.8871895	0.0000000
mes6	53256.84476	5650.57049	9.4250386	0.0000000
mes7	67066.41269	5576.87085	12.0258142	0.0000000
mes8	90537.23440	5585.60014	16.2090433	0.0000000
mes9	112735.76968	5676.21193	19.8610924	0.0000000
mes10	96617.54912	5657.46591	17.0778845	0.0000000
mes11	91750.61734	5784.02243	15.8627700	0.0000000
mes12	106187.02571	5806.73606	18.2868697	0.0000000

Modelo lineal A después de pandemia:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-9554110.3295	70096.409337	-136.299568	0.0000000
fecha	530.0556	3.756171	141.115970	0.0000000
dia.semMar	49833.0414	2409.988948	20.677705	0.0000000
dia.semMie	50246.8418	2411.547117	20.835936	0.0000000
dia.semJue	53291.4238	2419.722319	22.023777	0.0000000
dia.semVie	36449.8094	2412.283996	15.110082	0.0000000
dia.semSab	-26103.0000	2432.228292	-10.732134	0.0000000
dia.semDom	-190527.8322	2417.187675	-78.822110	0.0000000
mes2	72090.3965	4285.490469	16.821971	0.0000000
mes3	16819.2618	3935.319332	4.273925	0.0000193
mes4	-89720.1166	3652.195561	-24.566077	0.0000000
mes5	-48634.0899	3621.826632	-13.428056	0.0000000
mes6	-15372.6123	3648.479044	-4.213430	0.0000253
mes7	-20806.3416	3615.081695	-5.755428	0.0000000
mes8	-5430.1168	3627.363082	-1.496987	0.1344210
mes9	45353.4258	3642.956703	12.449620	0.0000000
mes10	54347.4763	3630.170144	14.971055	0.0000000
mes11	62769.4011	3675.879773	17.076021	0.0000000
mes12	69903.3535	3660.542901	19.096444	0.0000000

Modelo lineal B antes de pandemia:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	103720.759917	27976.152110	3.7074705	0.0002106
fecha	-1.032924	1.545727	-0.6682449	0.5039945
dia.semMar	10955.060106	654.173110	16.7464237	0.0000000
dia.semMie	6549.693145	654.538686	10.0065791	0.0000000
dia.semJue	8591.937355	654.706392	13.1233442	0.0000000
dia.semVie	10527.680788	654.633891	16.0817839	0.0000000

	Estimate	Std. Error	t value	Pr(>|t|)
dia.semSab	-6501.032975	654.638808	-9.9307174	0.0000000
dia.semDom	-46630.426484	654.400388	-71.2567219	0.0000000
mes2	20761.384909	678.385057	30.6041306	0.0000000
mes3	9596.935131	691.162072	13.8852167	0.0000000
mes4	8668.436447	832.444609	10.4132291	0.0000000
mes5	12215.125769	816.348982	14.9631175	0.0000000
mes6	6512.528814	821.596276	7.9266776	0.0000000
mes7	10613.571507	810.880305	13.0889497	0.0000000
mes8	17254.049195	812.149549	21.2449163	0.0000000
mes9	18665.856750	825.324557	22.6163836	0.0000000
mes10	16655.570305	822.598873	20.2474995	0.0000000
mes11	13615.162655	841.000266	16.1892489	0.0000000
mes12	11132.096654	844.302841	13.1849570	0.0000000

Modelo lineal B después de pandemia:

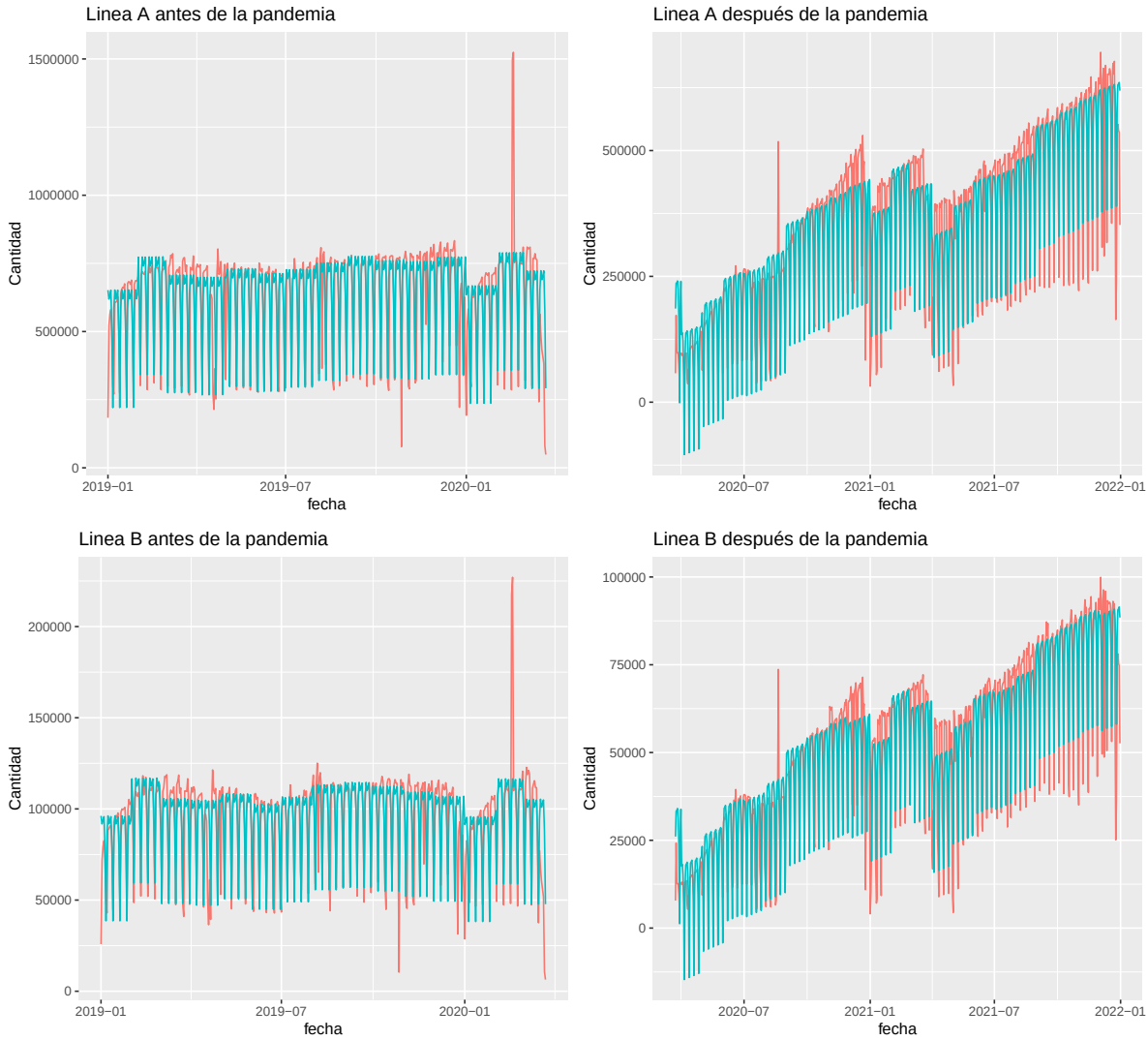
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.518194e+06	9846.064720	-154.192982	0.0000000
fecha	8.387547e+01	0.527609	158.972766	0.0000000
dia.semMar	7.085593e+03	338.518155	20.931205	0.0000000
dia.semMie	7.325147e+03	338.737023	21.624878	0.0000000
dia.semJue	7.753848e+03	339.885349	22.813129	0.0000000
dia.semVie	4.660667e+03	338.840528	13.754750	0.0000000
dia.semSab	-2.979133e+03	341.641996	-8.720043	0.0000000
dia.semDom	-2.534293e+04	339.529321	-74.641370	0.0000000
mes2	1.133311e+04	601.959742	18.827025	0.0000000
mes3	5.654132e+03	552.773091	10.228667	0.0000000
mes4	-1.083915e+04	513.004221	-21.128767	0.0000000
mes5	-5.080757e+03	508.738461	-9.986972	0.0000000
mes6	6.406999e+02	512.482182	1.250190	0.2112529
mes7	-5.469585e+02	507.791036	-1.077133	0.2814410
mes8	1.589926e+03	509.516137	3.120463	0.0018097
mes9	8.707187e+03	511.706488	17.015979	0.0000000
mes10	9.895838e+03	509.910429	19.407013	0.0000000
mes11	1.099722e+04	516.331015	21.298788	0.0000000
mes12	9.025963e+03	514.176727	17.554204	0.0000000

Se puede observar en los *summary* de los 4 modelos que:

- El nivel de referencia es el intercepto β_0 y representa el efecto del mes de Enero y el dia Lunes en el flujo de pasajeros.
- Para la serie *antes de la pandemia* y *línea A* se puede observar que todas las covariables son significativas menos el intercepto β_0, esto significa que todas sirven para modelar la serie, pero la tendencia explicada por el mes Enero y el dia Lunes está cerca de 0, osea, no aporta una tendencia significante.
- Para la serie *después de la pandemia* y *línea A* se puede observar que todas las covariables son significativas menos $\beta_{mes.agosto}$, esto significa que todas sirven para modelar la serie,

pero la tendencia explicada por la tendencia del mes de Agosto no difiere a la del mes de Enero y el dia Lunes, el cual es el mes y dia de referencia contenido en β_0, de otra se puede interpretar que $\beta_{mes.agosto}$ está muy cercano a 0 porque su tendencia ya fue explicada por el mes de Enero y el dia Lunes.

- Para la serie *antes de la pandemia* y *línea B* se puede observar que todas las covariables son significativas menos $fecha$, esto significa que todas sirven para modelar la serie, pero la tendencia explicada por la tendencia global de la serie, osea, la línea recta con menor distancia a *todos* los puntos es 0, esto podria interpretarse como que la serie es estacionaria, aunque faltaria verificar que la varianza sea constante y la correlación no dependa del tiempo.
- Para la serie *después de la pandemia* y *línea B* se puede observar que todas las covariables son significativas menos $\beta_{mes.junio}$ y $\beta_{mes.julio}$, esto significa que todas sirven para modelar la serie, pero la tendencia explicada por estos dos meses no difiere a la del mes de Enero y el dia Lunes, el cual es el mes y dia de referencia contenido en el intercepto β_0.
- Además, es de notar que las series para antes de la pandemia tenian mucha menos variablidad y un comportamiento deseado, ya que son mucho mas faciles de modelar que después de la pandemia.


Linea A antes de la pandemia
Cantidad
1500000
1000000
500000
0
2019−01
2019−07
2020−01
fecha
Linea A después de la pandemia
Cantidad
500000
250000
0
2020−07
2021−01
2021−07
2022−01
fecha
Linea B antes de la pandemia
Cantidad
200000
150000
100000
50000
0
2019−01
2019−07
2020−01
fecha
Linea B después de la pandemia
Cantidad
100000
75000
50000
25000
0
2020−07
2021−01
2021−07
2022−01
fecha