# Parcial 1

Julián Úsuga Ortiz

## 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 $\sigma_w^2 = 4.8$.

## 2.

Sea $X_t$ el proceso estocastico 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 $\sigma_w^2 = 6.2$.

## 3.

```
library(readxl)
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6     v purrr   0.3.4
v tibble  3.1.8     v dplyr   1.0.9
v tidyr   1.2.0     v stringr 1.4.0
v readr   2.1.2     v forcats 0.5.1
```

```
-- Conflicts --------------------------------------- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```r
library(lubridate)
```

```
Attaching package: 'lubridate'

The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```r
data_19 <- read_excel("data/Afluencia_Metro_2019.xlsx",
                      range = "A4:W3566",
                      col_names = FALSE)
```

```
New names:
* `` -> `...1`
* `` -> `...2`
* `` -> `...3`
* `` -> `...4`
* `` -> `...5`
* `` -> `...6`
* `` -> `...7`
* `` -> `...8`
* `` -> `...9`
* `` -> `...10`
* `` -> `...11`
* `` -> `...12`
* `` -> `...13`
* `` -> `...14`
* `` -> `...15`
* `` -> `...16`
* `` -> `...17`
* `` -> `...18`
* `` -> `...19`
* `` -> `...20`
* `` -> `...21`
* `` -> `...22`
* `` -> `...23`
```

```
data_20 <- read_excel("data/Afluencia_Metro.xlsx",
                      range = "A4:W3767",
                      col_names = FALSE)
```

```
New names:
* `` -> `...1`
* `` -> `...2`
* `` -> `...3`
* `` -> `...4`
* `` -> `...5`
* `` -> `...6`
* `` -> `...7`
* `` -> `...8`
* `` -> `...9`
* `` -> `...10`
* `` -> `...11`
* `` -> `...12`
* `` -> `...13`
* `` -> `...14`
* `` -> `...15`
* `` -> `...16`
* `` -> `...17`
* `` -> `...18`
* `` -> `...19`
* `` -> `...20`
* `` -> `...21`
* `` -> `...22`
* `` -> `...23`
```

```
data_21 <- read_excel("data/Afluencia_2021.xlsx",
                      range = "A4:W4125",
                      col_names = FALSE)
```

```
New names:
* `` -> `...1`
* `` -> `...2`
* `` -> `...3`
* `` -> `...4`
* `` -> `...5`
* `` -> `...6`
```

* `` -> `...7`
* `` -> `...8`
* `` -> `...9`
* `` -> `...10`
* `` -> `...11`
* `` -> `...12`
* `` -> `...13`
* `` -> `...14`
* `` -> `...15`
* `` -> `...16`
* `` -> `...17`
* `` -> `...18`
* `` -> `...19`
* `` -> `...20`
* `` -> `...21`
* `` -> `...22`
* `` -> `...23`

```
metro.cols <- c(
  "fecha",
  "linea",
  "04:00",
  "05:00",
  "06:00",
  "07:00",
  "08:00",
  "09:00",
  "10:00",
  "11:00",
  "12:00",

  "13:00",
  "14:00",
  "15:00",
  "16:00",
  "17:00",
  "18:00",
  "19:00",
  "20:00",
  "21:00",
  "22:00",
  "23:00",
```

```
  "per.dia.tot"
)

colnames(data_19) <- metro.cols
colnames(data_20) <- metro.cols
colnames(data_21) <- metro.cols
```

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).

```
datos_juntos <- bind_rows(data_19, data_20, data_21)
rm(data_19, data_20, data_21)
```

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

c).

```
datos_juntos$fecha <- datos_juntos$fecha |> as_date()
```

```
data <- pivot_longer(datos_juntos, cols = ends_with(":00"), names_to = "hora", values_to =

data$hora <- hm(data$hora)
data$dia <- day(data$fecha)
data$dia.sem <- weekdays(data$fecha)
data$sem <- week(data$fecha)
data$mes <- month(data$fecha)
data$anio <- year(data$fecha)
```

Dimensiones son 228980 filas y 10 columnas.

d).

```
data$linea <- data$linea |> tolower()
dat_lin_A <- data |> filter(linea == "línea a")
dat_lin_B <- data |> filter(linea == "línea b")
```

```
dat_lin_A |> arrange(fecha, hora) |> head(5)
```

```
# A tibble: 5 x 10
  fecha      linea   per.dia.~1 hora     per.num   dia dia.sem   sem   mes  anio
  <date>     <chr>        <dbl> <Period>   <dbl> <int> <chr>   <dbl> <dbl> <dbl>
1 2019-01-01 línea a     183664 4H 0M 0S      14     1 Tuesday     1     1  2019
2 2019-01-01 línea a     183664 5H 0M 0S    4554     1 Tuesday     1     1  2019
3 2019-01-01 línea a     183664 6H 0M 0S    5696     1 Tuesday     1     1  2019
4 2019-01-01 línea a     183664 7H 0M 0S    4893     1 Tuesday     1     1  2019
5 2019-01-01 línea a     183664 8H 0M 0S    4399     1 Tuesday     1     1  2019
# ... with abbreviated variable name 1: per.dia.tot
```

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

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

e).

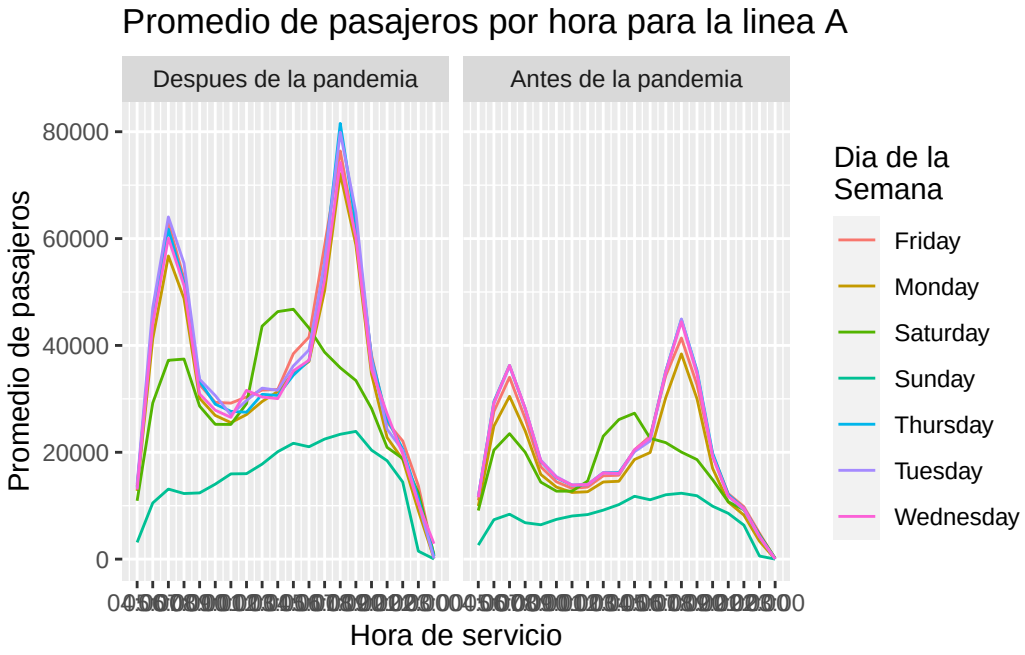
El 23 de Marzo de 2020 el Gobierno de Colombia expidió el Decreto 457 para el período de aislamiento preventivo obligatorio a causa del virus del COVID-19.

```
dat_lin_A <- dat_lin_A |>
  mutate(pandemia = if_else(fecha < dmy("23-03-2020"), "no", "si"))
dat_lin_B <- dat_lin_B |>
  mutate(pandemia = if_else(fecha < dmy("23-03-2020"), "no", "si"))
```

```
facet.names <- c(
  'si'="Antes de la pandemia",
  'no'="Despues de la pandemia"
)

dat_lin_A |>
  group_by(dia.sem, hora, pandemia) |>
  summarise(promedio = mean(per.num, na.rm = TRUE)) |>
  ggplot() +
  geom_line(aes(x = as_datetime(hora), y = promedio, color = dia.sem)) +
  scale_x_datetime(breaks = "1 hour", date_labels =  "%H:00") +
  facet_grid(cols = vars(pandemia), labeller=as_labeller(facet.names)) +
  labs(title = "Promedio de pasajeros por hora para la linea A",
       x = "Hora de servicio",
       y = "Promedio de pasajeros") +
  scale_color_discrete(name = "Dia de la\nSemana")
```

```
`summarise()` has grouped output by 'dia.sem', 'hora'. You can override using
the `.groups` argument.
```




se puede ver que es la mitad ... etc.

f).

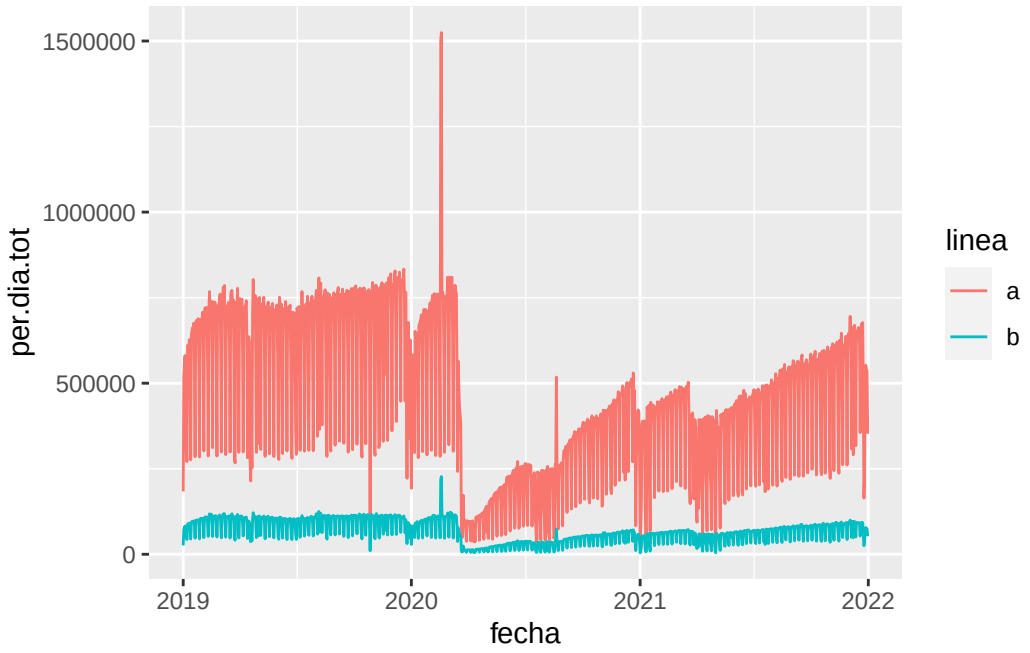
```
lin.a <- dat_lin_A |>
  group_by(fecha) |>
  summarise(per.dia.tot = sum(per.num, na.rm=TRUE))
lin.b <- dat_lin_B |>
  group_by(fecha) |>
  summarise(per.dia.tot = sum(per.num, na.rm=TRUE))

# falta añadir día, día de la semana, semana, mes, año
# para lin.a y lin.b
```

g).

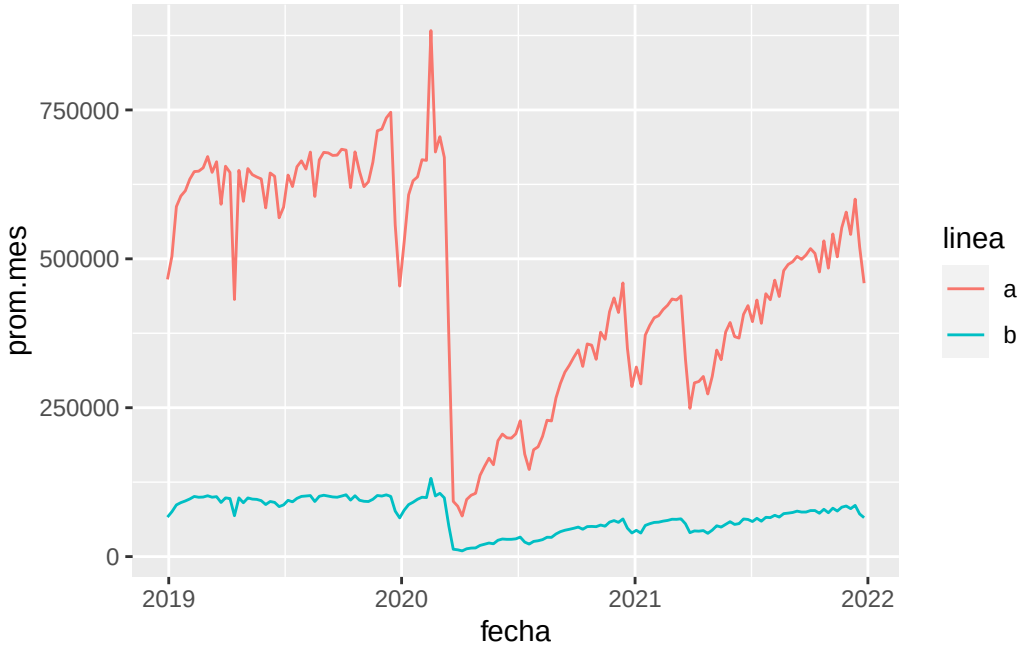
```
lin.a$linea <- "a"
lin.b$linea <- "b"

bind_rows(lin.a, lin.b) |>
  ggplot(aes(x = fecha, y = per.dia.tot, color = linea)) +
  geom_line()
```




```
# Esto es promedio por semana
bind_rows(lin.a, lin.b) |>
  group_by(fecha = floor_date(fecha, unit = "week") ,linea) |>
  summarise(prom.mes = mean(per.dia.tot)) |>
  ggplot(aes(x = fecha, y = prom.mes, color = linea)) +
  geom_line()
```

```
`summarise()` has grouped output by 'fecha'. You can override using the
`.groups` argument.
```

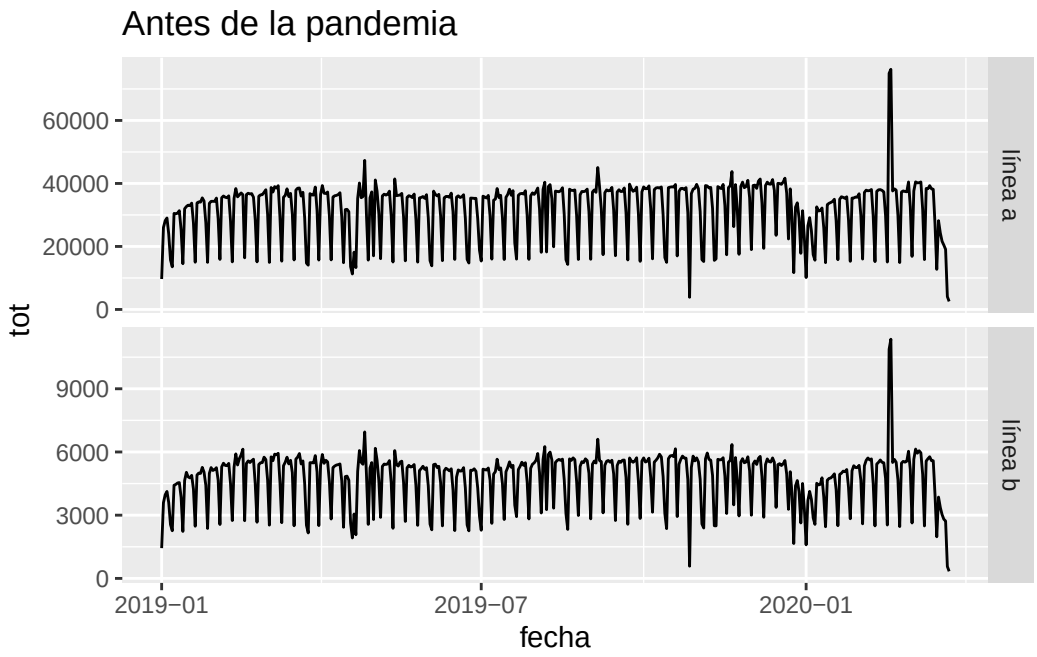



Se ve una caida en marzo de 2020 y una lenta recuperacion, no se ha llegado a niveles de 2019 otra grafica o tabla y conclusiones..

h).

```
bind_rows(dat_lin_A, dat_lin_B) |>
  filter(pandemia == "no") |>
  group_by(fecha, linea) |>
  summarise(tot = mean(per.num, na.rm = TRUE)) |>
  ggplot(aes(x= fecha, y = tot)) +
  geom_line() +
  facet_grid(rows = vars(linea), scales = "free") +
  labs(title = "Antes de la pandemia")
```

```
`summarise()` has grouped output by 'fecha'. You can override using the
`.groups` argument.
```

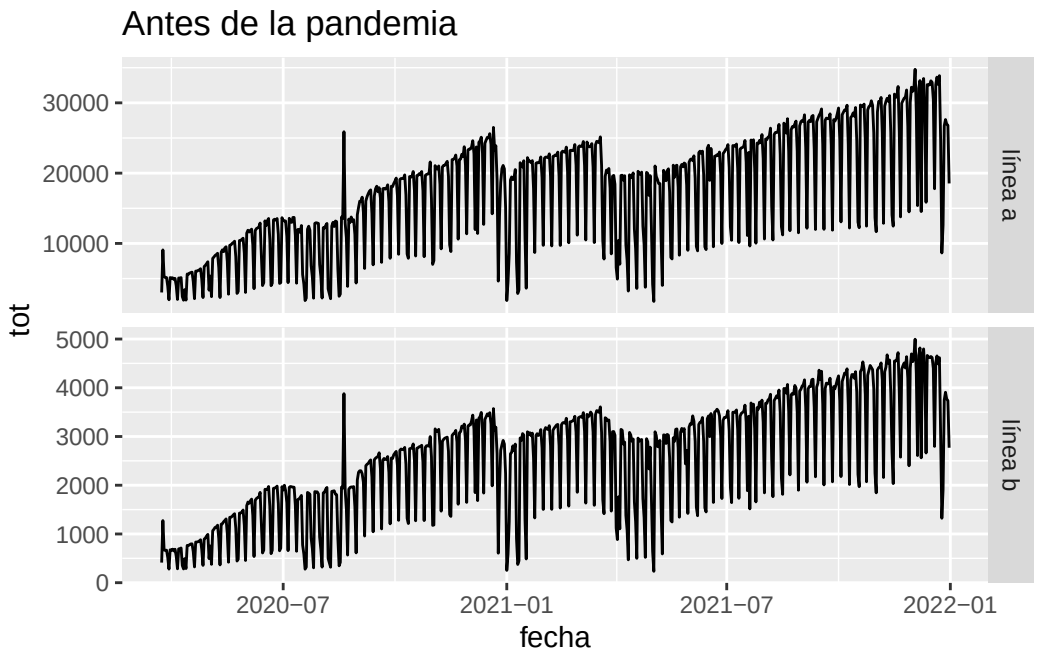



para antes de la pandemia la linea a y b tuvieron un comportamiento muy similar

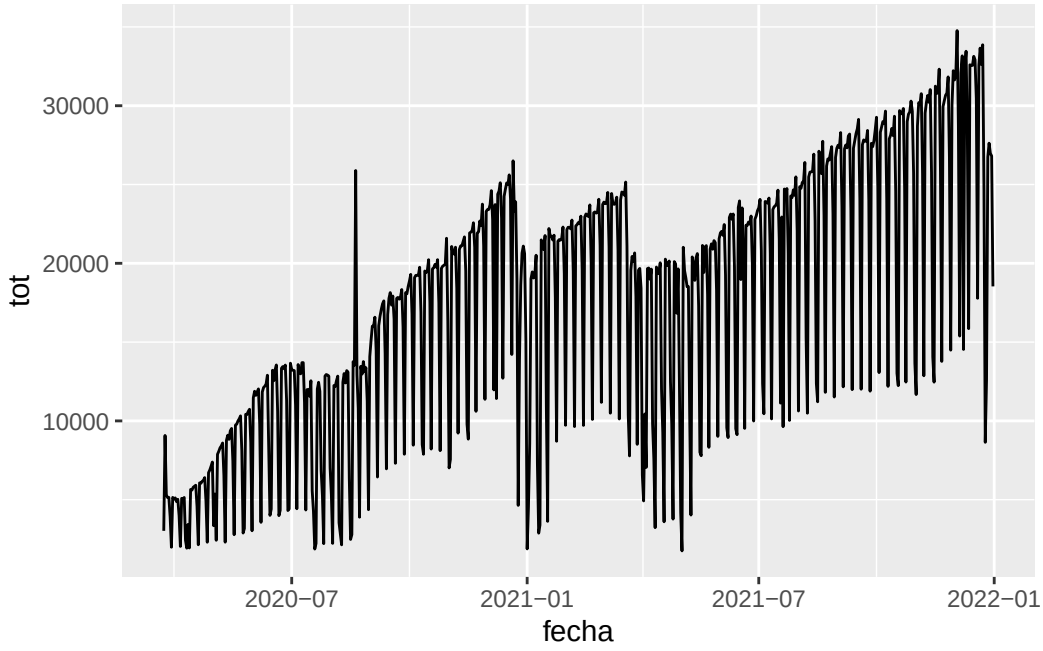
se ve que es estacionaria alrededor de una media, ademas, en la ultima semana hay una baja en la afluencia muy probablemente a causa del covid

```
bind_rows(dat_lin_A, dat_lin_B) |>
  filter(pandemia == "si") |>
  group_by(fecha, linea) |>
  summarise(tot = mean(per.num, na.rm = TRUE)) |>
  ggplot(aes(x= fecha, y = tot)) +
  geom_line() +
  facet_grid(rows = vars(linea), scales = "free") +
  labs(title = "Antes de la pandemia")
```

```
`summarise()` has grouped output by 'fecha'. You can override using the
`.groups` argument.
```




```
dat_lin_A |>
  filter(pandemia == "si") |>
  group_by(fecha) |>
  summarise(tot = mean(per.num, na.rm = TRUE)) |>
  ggplot(aes(x= fecha, y = tot)) + geom_line()
```




en esta serie de tiempo se ve una recuperacion en ambas lineas (a y b), no es estacionaria ya que se ve una clara tendencia de crecimiento positivo no lineal, …. .

i).

regresion con lm para dos dataset: antes y despues de la pandemia

puede ser con las covariables: indicadora dia de la semana (lunes, martes etc), mes, hora?

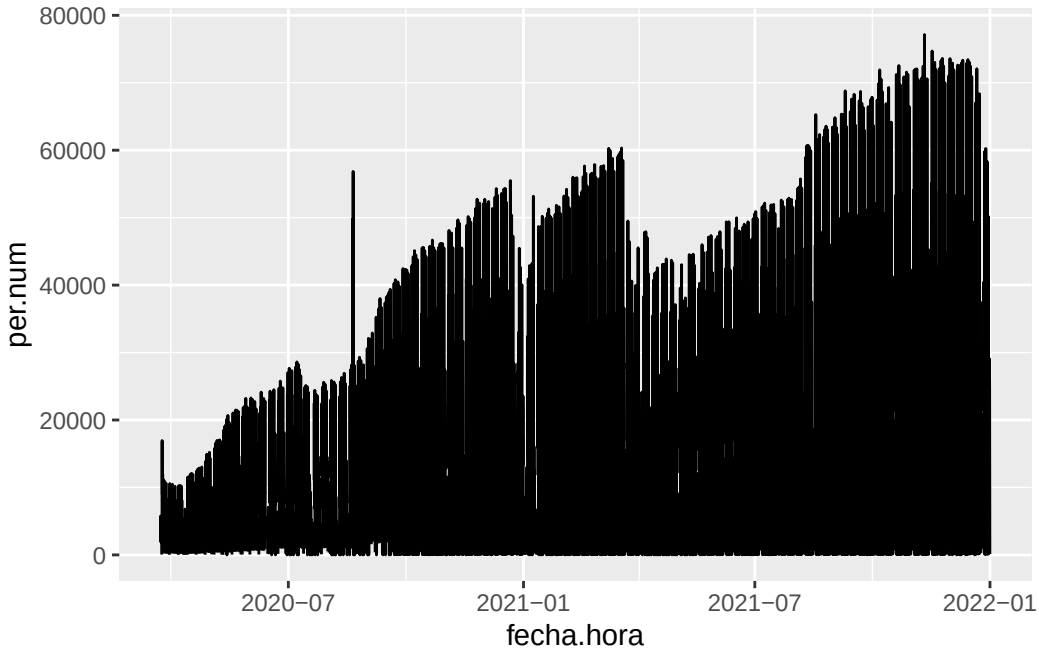
son 4 modelos: antes pandemia linea a antes pandemia linea a despues pandemia linea b despues pandemia linea b

contrastar, dar una interpretación del summary del model

graficar el modelo vs real

```
dat_lin_A |>
  filter(pandemia == "si") |>
  mutate(fecha.hora = fecha + hora) |>
  ggplot(aes(x = fecha.hora, y = per.num)) + geom_line()
```

```
Warning: Removed 1 row(s) containing missing values (geom_path).
```


80000
60000
40000
20000
0
per.num
2020−07
2021−01
2021−07
2022−01
fecha.hora