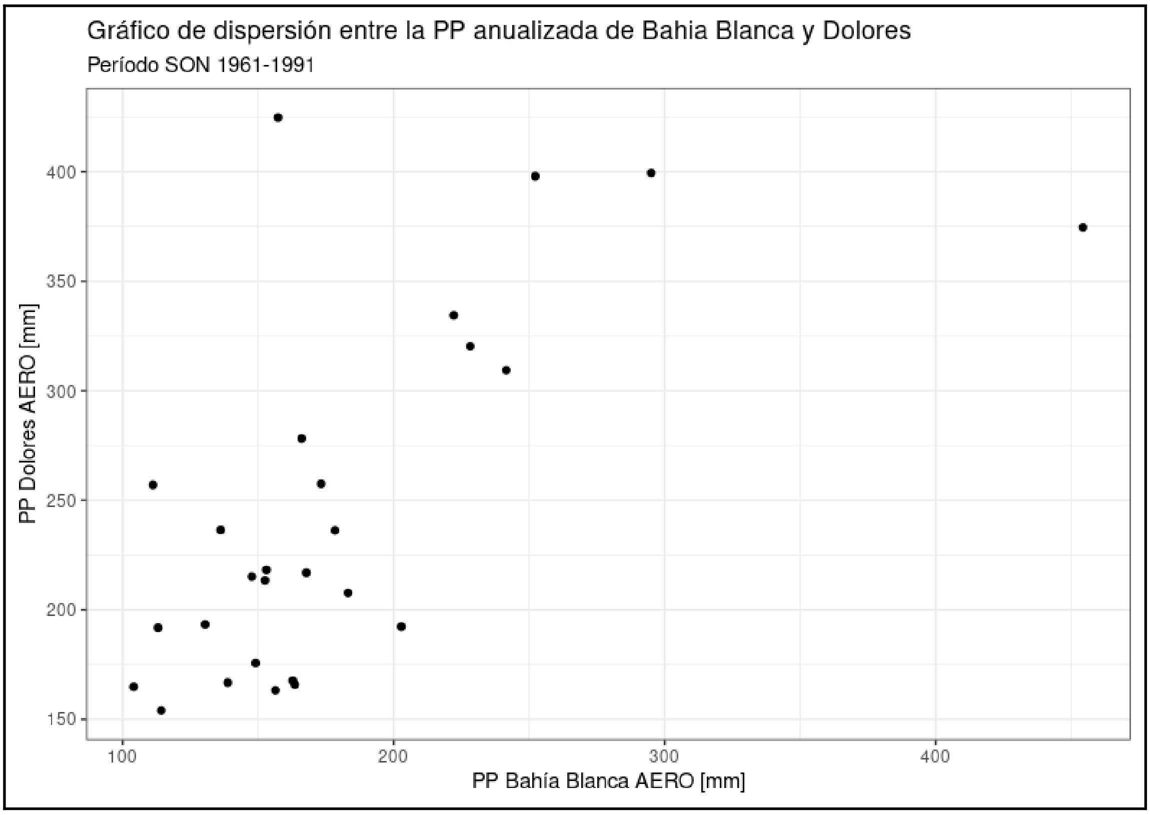
1er Parcial Clima II – Primera Fecha

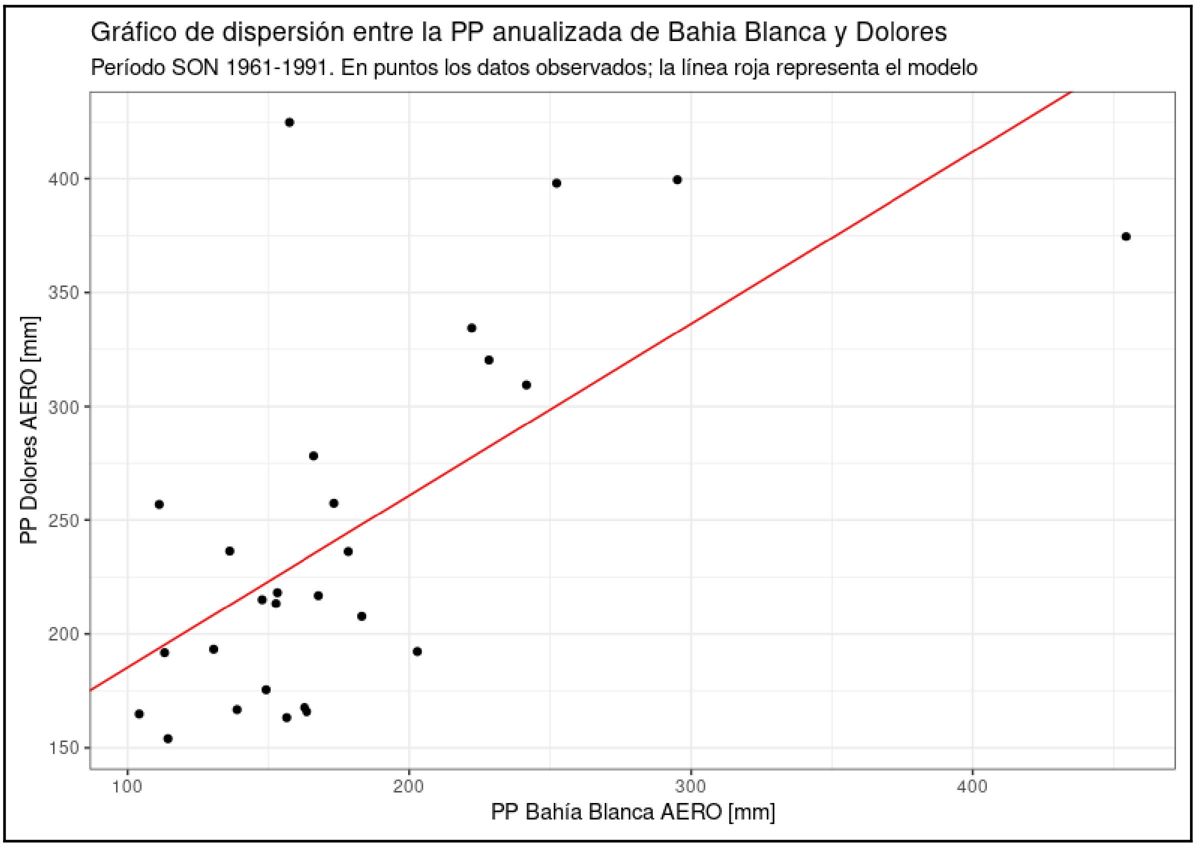
1.

   a. Al calcular la precipitación acumulada en ambas estaciones pluviométricas, detectamos valores faltantes en la estación de Dolores AERO para el período 1988-1991. Por lo tanto, es conveniente ubicar en qué momento ocurren los datos que faltan. Ya que es raro que durante 3 años no se registre precipitación. Filtrando por valores faltantes el *dataframe* correspondiente a los registros diarios de ambas estaciones para los meses de SON, construimos un nuevo *dataframe* de valores no disponibles (*NAs)* con el detalle de la fecha exacta y la estación responsable.
   Podemos obtener las fechas exactas y almacenarlas en un vector llamado "*fechas_nas".*
   Allí vemos que no es cierto que faltan datos durante todo el período de interés en esos años, sino que faltan en algunos días particulares; solo que *R* al calcular la suma y encontrarse con un *NA* le asigna *NA* como resultado.

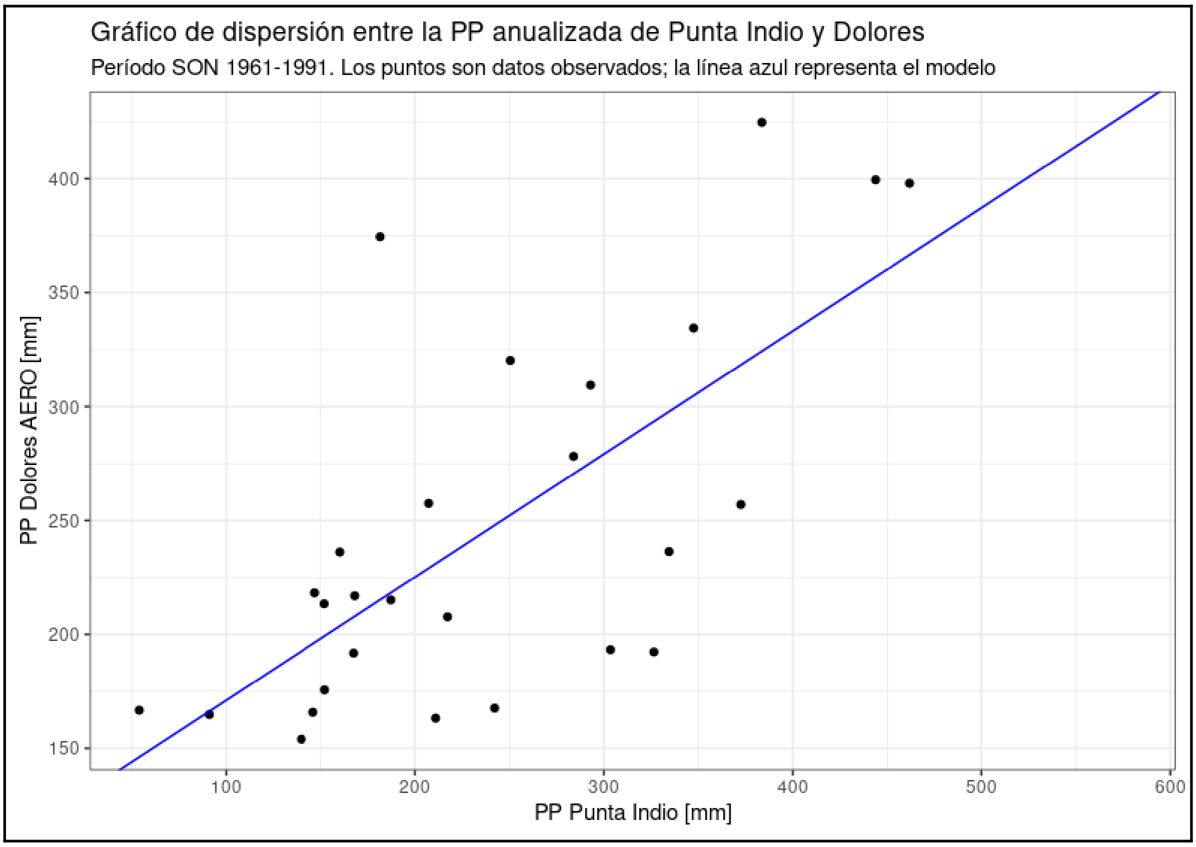
b. Para el gráfico de dispersión, voy a considerar como variable predictora a la precipitación de Bahía Blanca y como variable predictando a la de Dolores (ya que es la que presenta datos faltantes).



Observando el gráfico, podemos proponer un modelo de regresión lineal para ajustar los datos faltantes de Dolores.



c. Utilizando los datos de Punta Indio, obtenemos el siguiente gráfico de dispersión con su respectivo modelo de regresión lineal.

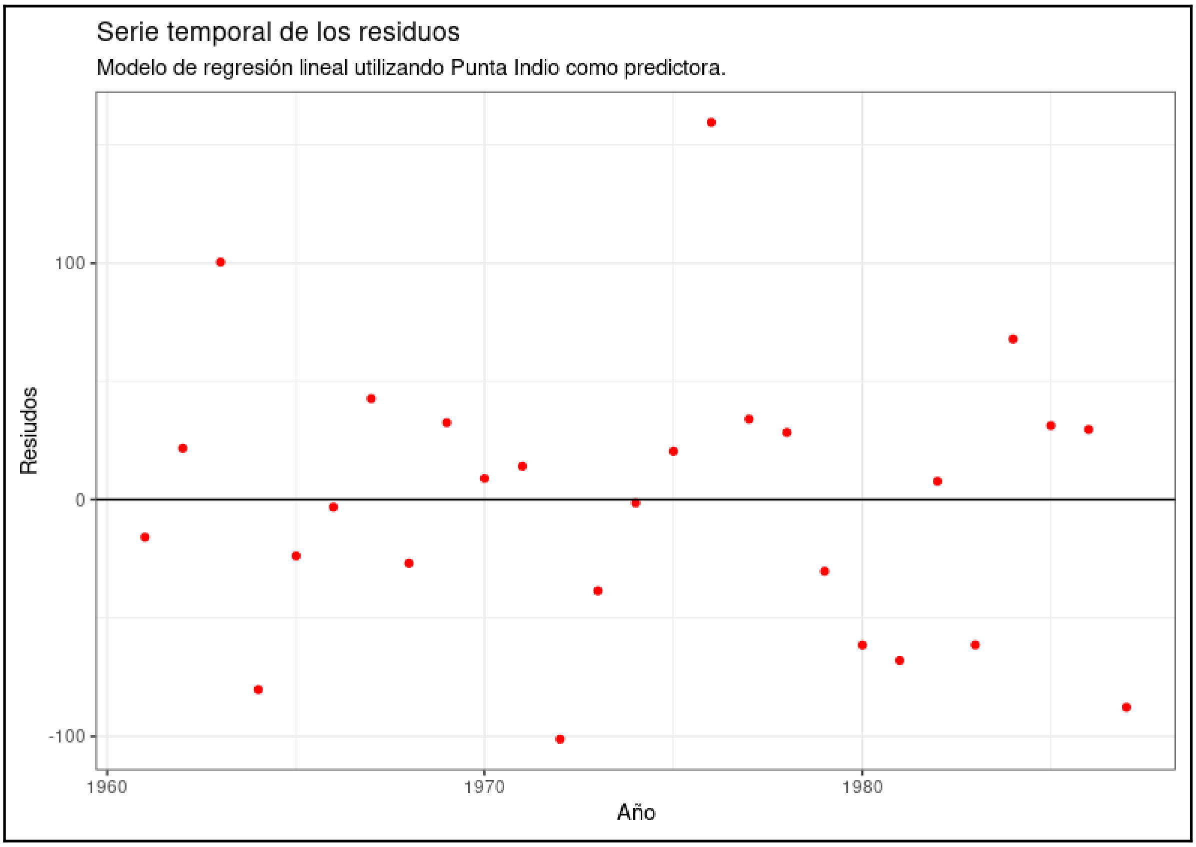


d. Para determinar cuál de los dos modelos es mejor. Voy a utilizar el coeficiente de determinación. Éste me permite ver qué porcentaje de la varianza original de mi variable predictando es representada por el modelo.

$$R^2 = \frac{SSR}{SST} \Rightarrow 0 \leq R^2 \leq 1$$

Donde *SSR* es la suma cuadrática de los residuos, es decir de los cuadrados de las diferencias entre el dato observado y la predicción del modelo; y *SST* es la suma cuadrática del desvío del dato con respecto a la media del mismo.
Por lo tanto, a mayor R^2, mejor modelo. En el sentido de cómo representa la varianza del dato original. Además, el coeficiente de determinación, es una medida para evaluar la *correlación* entre variables. Osea que mayor R^2, mayor correlación y por lo tanto, una mejor representación de la variable predictando.
Vemos que el coeficiente de determinación del modelo que utiliza a Punta Indio como variable predictora es de 0.49 aproximadamente. Eso quiere decir que casi el 50% de la varianza de la precipitación en Dolores es representada por este modelo; en contra partida, el modelo que relaciona a Bahía Blanca, sólo representa el 43%. Por ende, voy a utilizar el modelo de Punta Indio.
Tiene sentido que Punta Indio sea una mejor variable predictora, puesto que se encuentra mucho más cerca de Dolores y la precipitación es una variable muy sensible a variaciones espaciales.

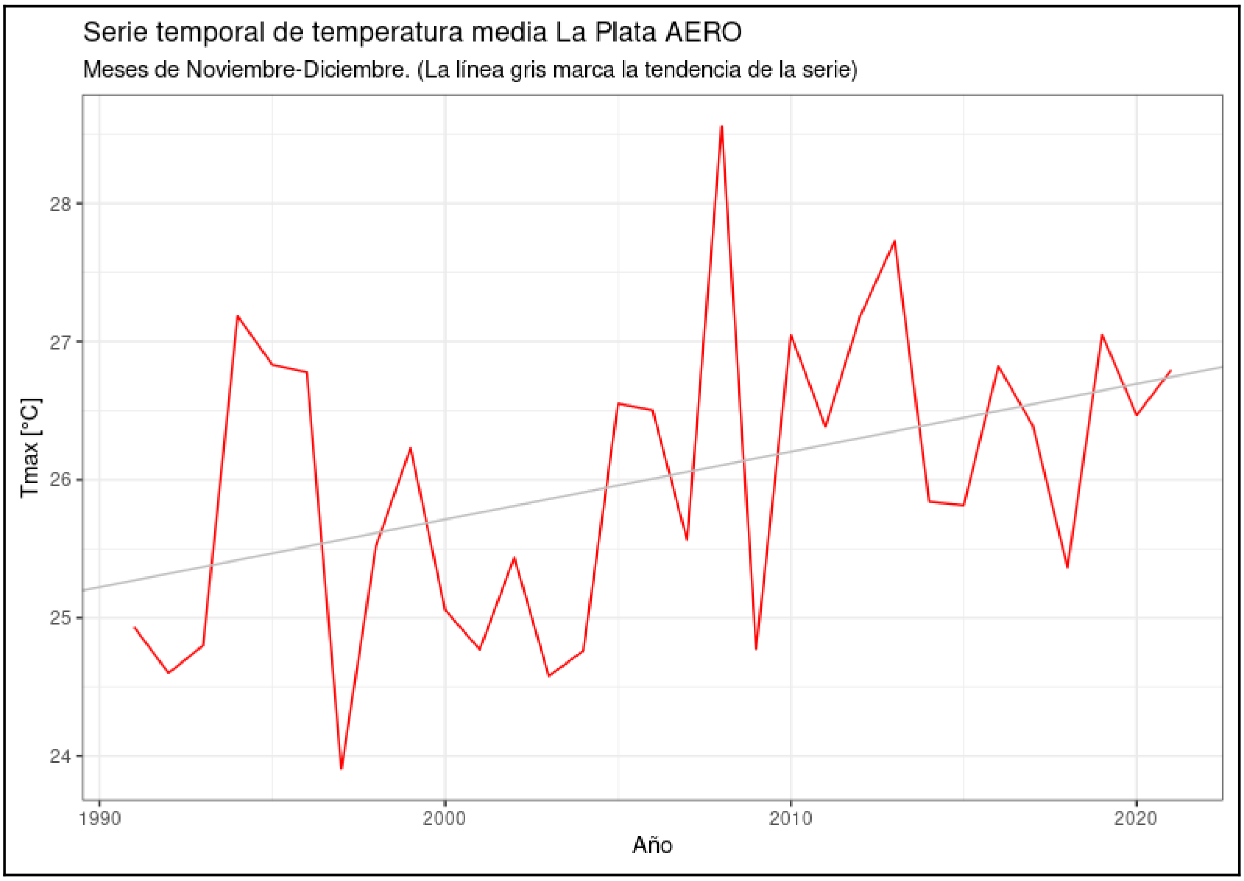
e. Planteo una serie temporal en donde grafico los residuos versus los años que presentan observación en Dolores.
   Observo que los residuos se distribuyen de manera homocedástica, es decir, que sus desvíos son prácticamente uniformes durante la serie temporal. Lo cual es esperado a la hora de plantear un modelo lineal.



f. La ecuación que dispongo es
   $Dolores(t) = Pendiente_{Punta\ Indio} * PP_{Punta\ Indio}(t) + Ordenda_{Punta\ Indio}$.
   Para rellenar los datos, *loopeo* por los índices de datos faltantes en mi *dataframe* guardado como *"pp_wide_anual"* que tiene toda la información relevante al ejercicio.

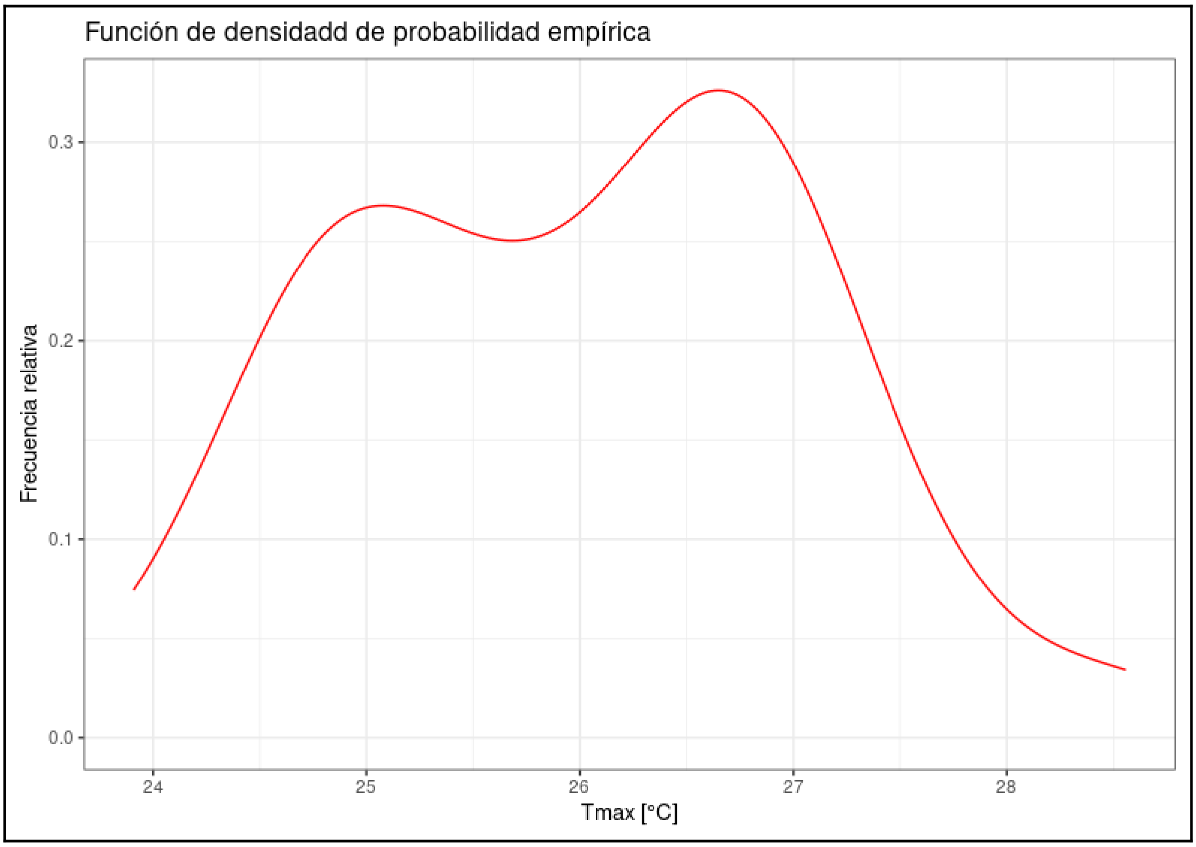
2.

a. Realizo un gráfico de línea para visualizar la serie temporal y evaluar tendencias.



Si realizo una regresión lineal, obtengo una pendiente de 0.05 aproximadamente, lo cual indica una tendencia, aunque positiva, pequeña. No recuerdo como filtrarla.
Sigo el análisis asumiendo que lo hice.

b. La función de densidad de probabilidad empírica se asemeja a una normal,



la única salvedad es que como no filtré la tendencia, mi variable no es estacionaria y por lo tanto tiene una combinación de distribuciones (sería como sumar dos normales). Por lo tanto adquiere una distribución bimodal.
Voy a ajustarla a una normal, que es el comportamiento teórico que esperaría de la variable.

c.