Series de Tiempo

Clase 10 – Series de Tiempo

Luis Gutiérrez – Ricardo Olea

DATA SCIENCE UC
FACULTAD DE MATEMÁTICAS

FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

Contenido del curso

1. Introducción a la estadística y análisis descriptivo
2. Análisis descriptivo y gráfico
3. Probabilidad y Distribuciones
4. Muestreo
5. Inferencia estadística: Pruebas de hipótesis
6. Taller Práctico de inferencia
7. Introducción a los modelos estadísticos
8. Modelos predictivos I: Modelos de regresión lineal.
9. Modelos predictivos II: Regresión logística y otros modelos.
10. Modelos de Forecasting I.
11. Modelos de Forecasting II
12. Taller de Forecast

Series de Tiempo

Cuando la información es medida secuencialmente en el tiempo, la independencia de las observaciones suele perderse, es decir, el supuesto base de un modelo de regresión lineal se pierde.

Para solucionar este problema, la dependencia se representa mediante un modelo lineal con información pasada, presente y futura.

Los modelos utilizados para este tipo de datos se conocen como modelos para **Series de Tiempo.**

Series de Tiempo

Una serie de tiempo es conjunto de datos medidos en intervalos de tiempo constante (e.g., horas, días, semanas, años, etc).

Una característica distintiva de las series de tiempo, es el hecho que cada dato es ***dependiente*** de los datos previos de la serie.

Ejemplo:

Mes	Co2
Enero	315.42
Febrero	316.31
Marzo	316.50
Abril	317.56
Mayo	318.13
Junio	318.00
Julio	316.39
Agosto	314.65
Septiembre	313.68
Octubre	313.18
Noviembre	314.66
Diciembre	315.43

Series de Tiempo

Dada una variable $\boldsymbol{Y_t}$, se define el modelo aditivo de componentes como:

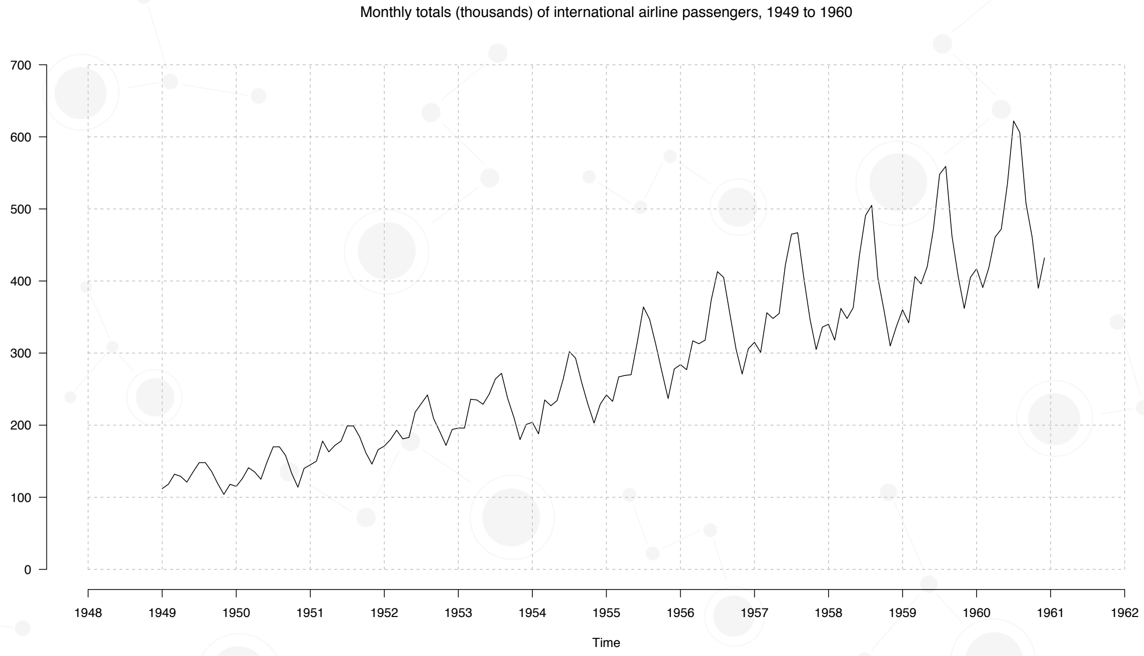
$$Y_t = T_t + S_t + C_t + \epsilon_t$$

Mientras que un modelo multiplicativo es de la forma

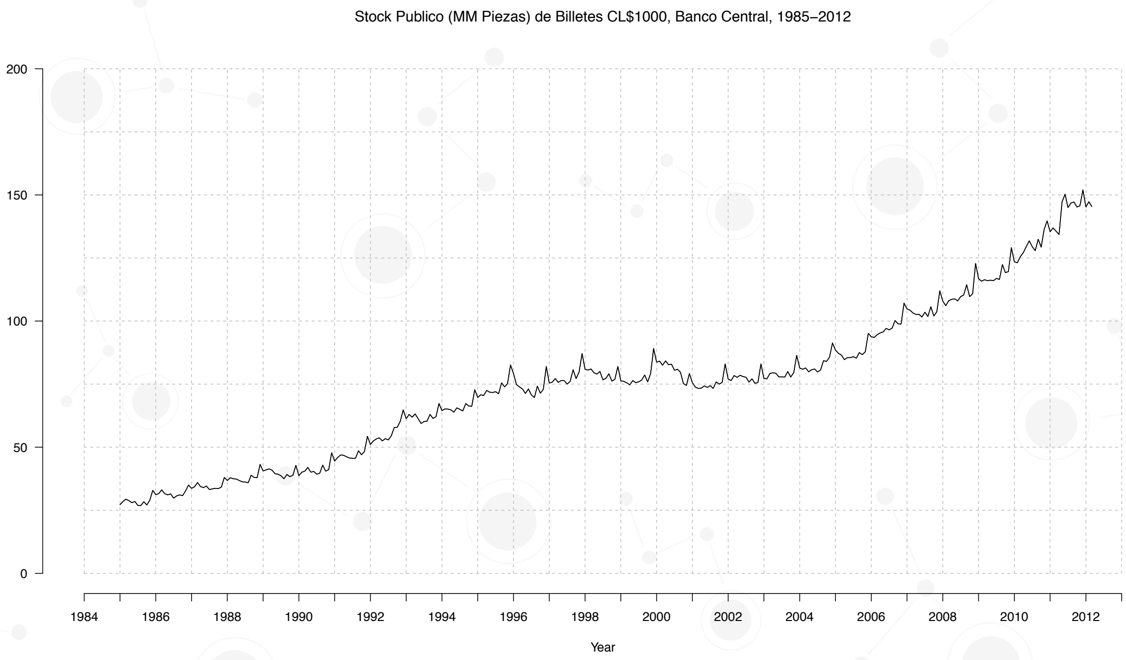
$$Y_t = T_t \cdot S_t \cdot C_t \cdot \epsilon_t$$

Donde $\boldsymbol{T}$ es la tendencia, $\boldsymbol{S}$ corresponde a la componente estacional, $\boldsymbol{C}$ son los posibles ciclos y $\boldsymbol{\epsilon}$ la componente del error. Las tres primeras son determinísticas, y es posible modelar su comportamiento, mientras que el error deberá ser asumido o estimado bajo otros criterios.

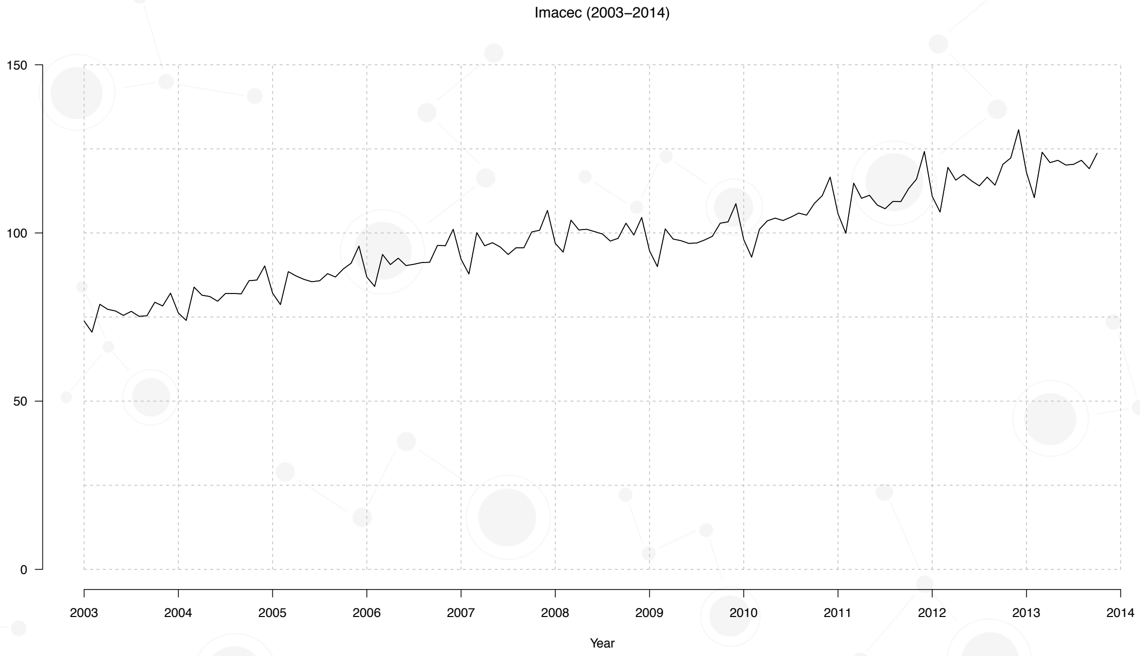
Series de Tiempo



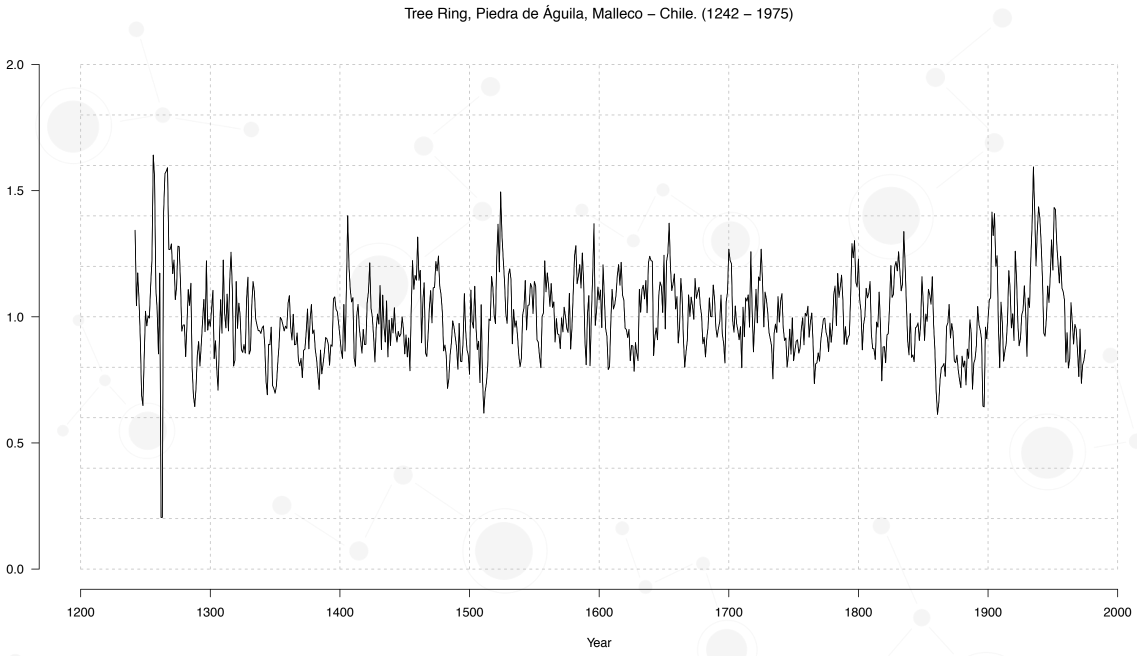
Series de Tiempo



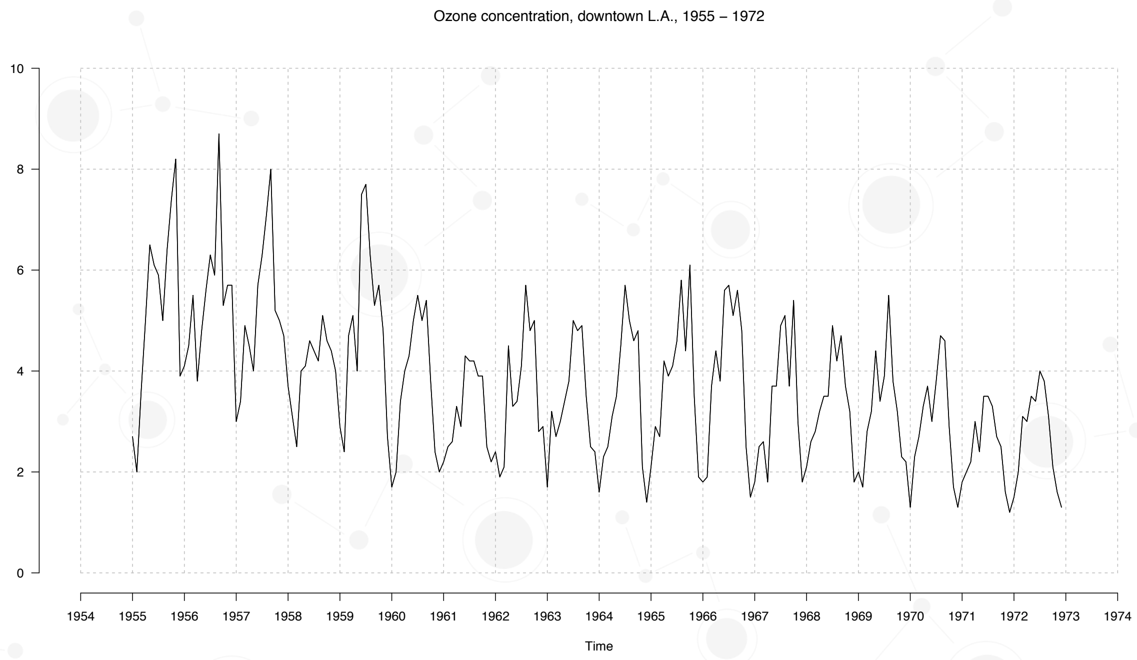
Series de Tiempo



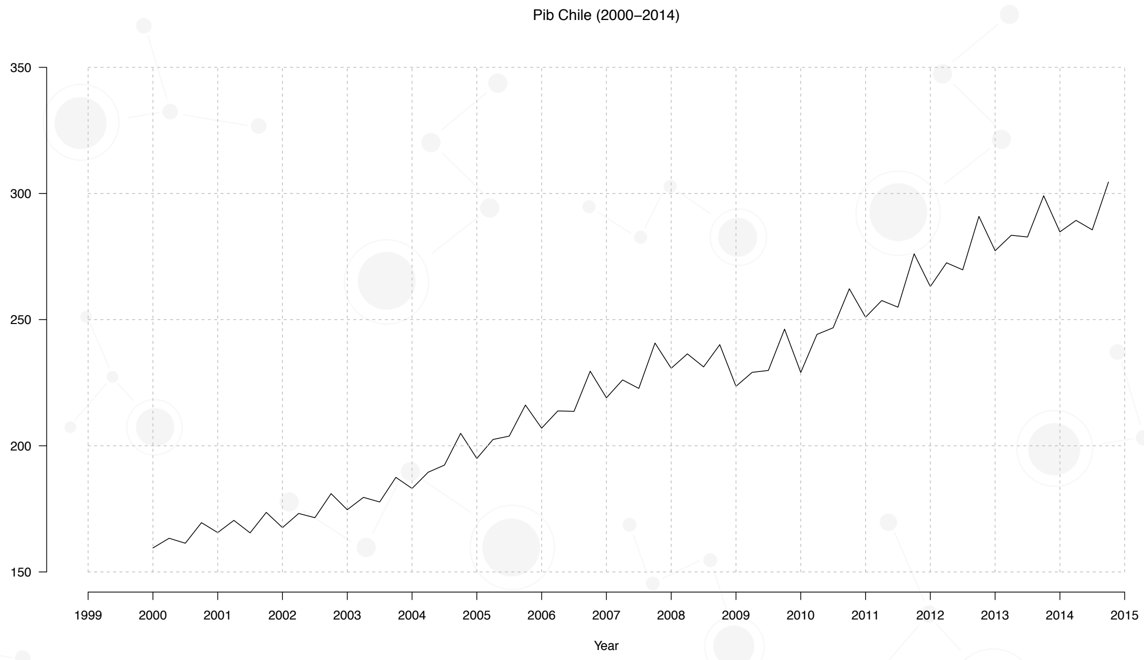
Series de Tiempo



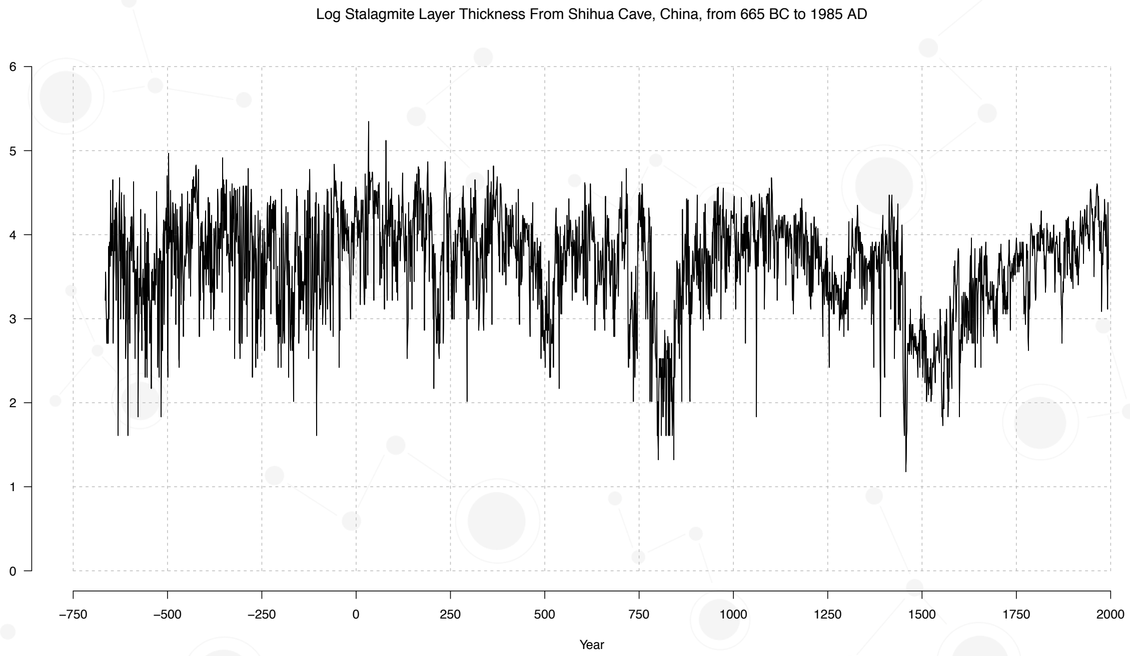
Series de Tiempo



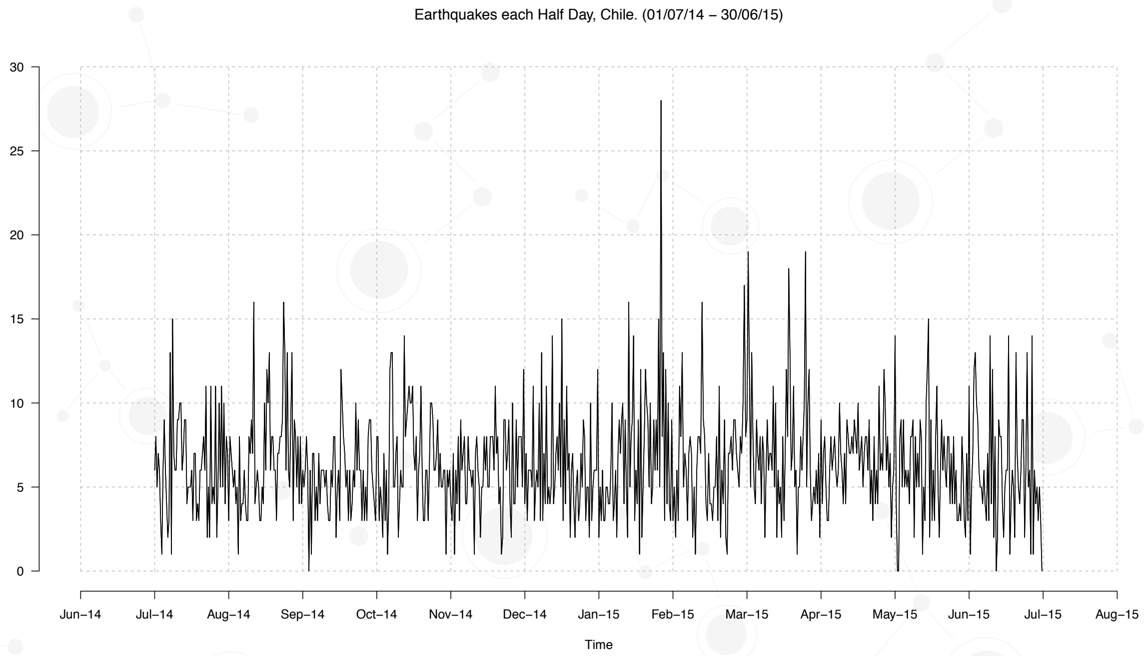
Series de Tiempo



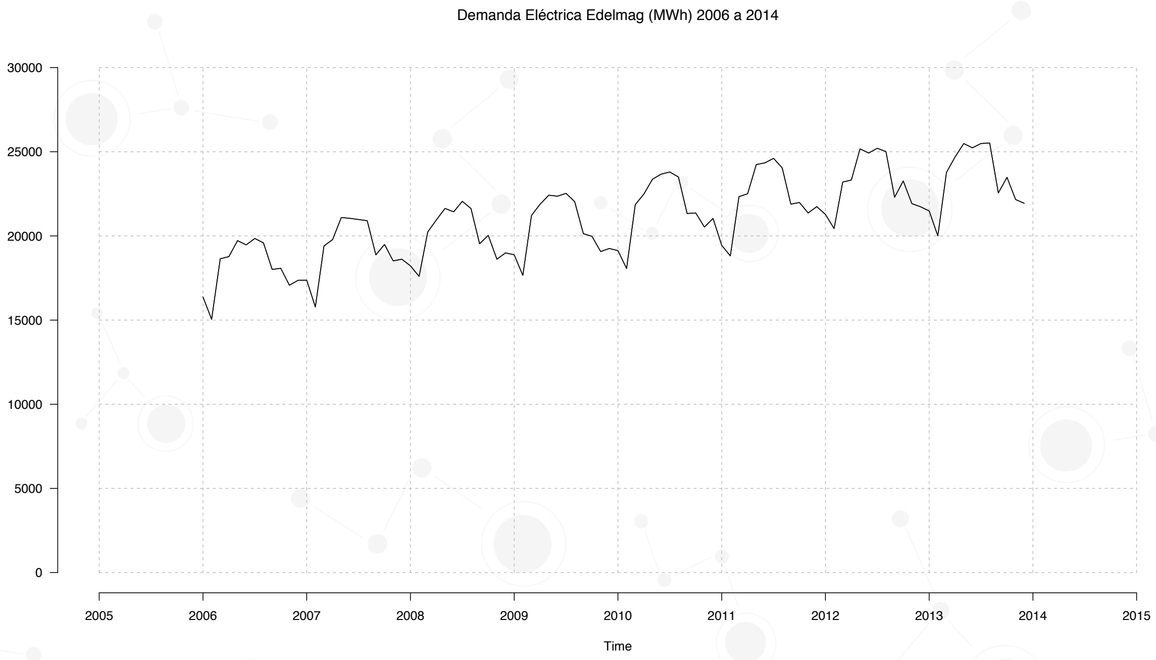
Series de Tiempo



Series de Tiempo



Series de Tiempo



Series de Tiempo

En R se debe indicar que una variable es una serie de tiempo, con el comando **ts()**. Además, **as.Date()** transforma un texto en cierto formato a un dato de fecha.

El comando **decompose()** determina un modelo aditivo, que posteriormente se puede gráficar.

Actividad

El archivo AirCanada.xlsx contiene datos recolectados en el aeropuerto de San Francisco, sobre la cantidad de pasajeros que llegan en vuelos de la aerolínea Air Canadá.

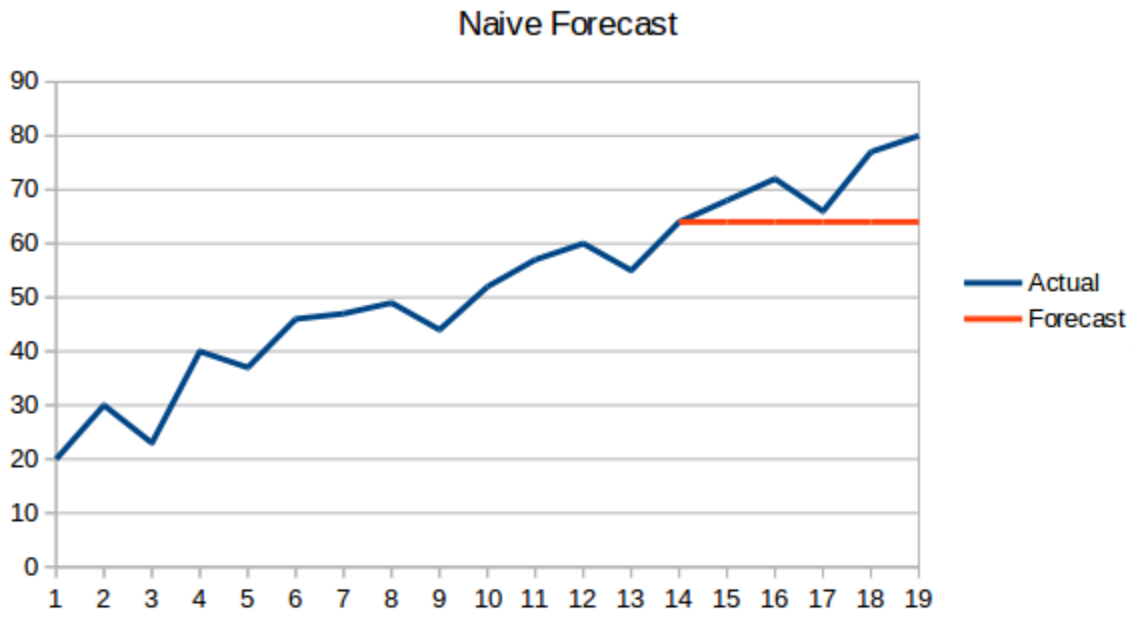
(Puede descargar los datos **aquí**)

Actividad I

- Cargue la base de datos e inspecciónela
- Utilizando el comando **ts()** transforme los datos a un formato "serie de tiempo“ que considere el período desde Enero de 2006 a Junio de 2018.
- Utilizando el comando **plot.ts()** grafique la serie de tiempo resultante. Comente.
- Utilizando la función **as.Date()** transforme la variable de tiempo a formato "date“. Realice un gráfico Y vs t para visualizar la serie.
- Utilice el comando **decompose()** y comente sobre la tendencia y estacionalidad de la serie.

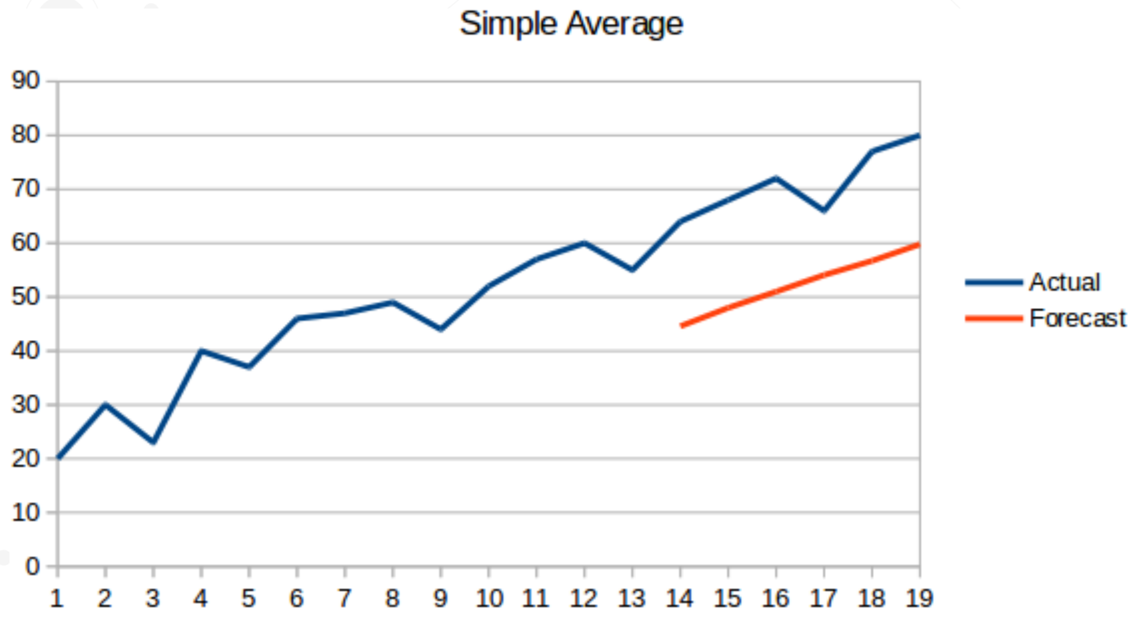
Modelamiento y predicción

1. Valor predicho es igual al dato previo



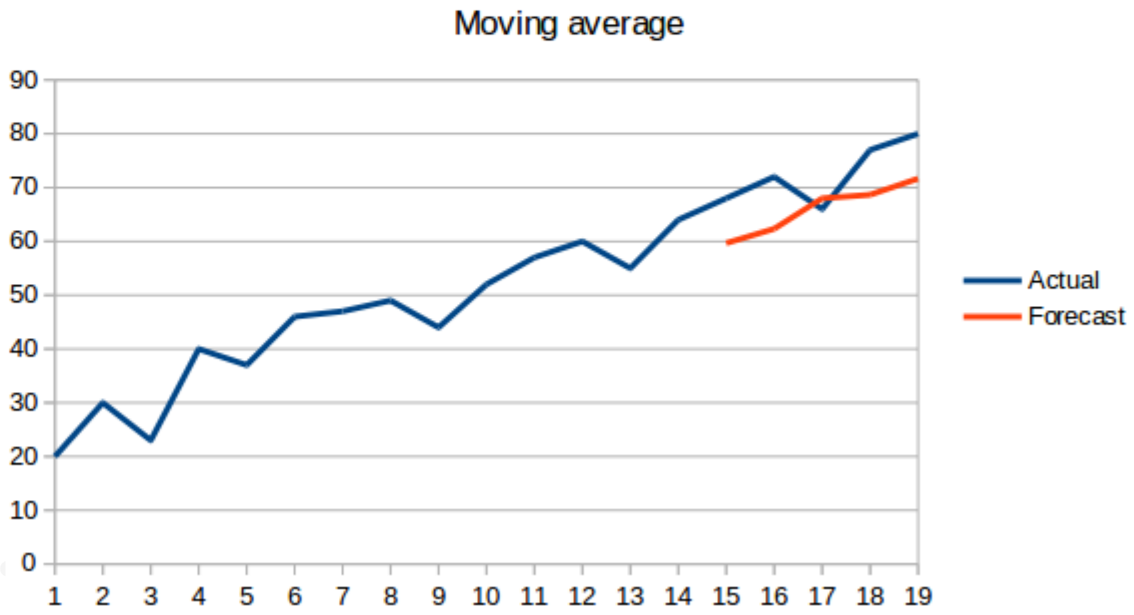
Modelamiento y predicción

2. Valor predicho es el promedio de **todos** los valores previos



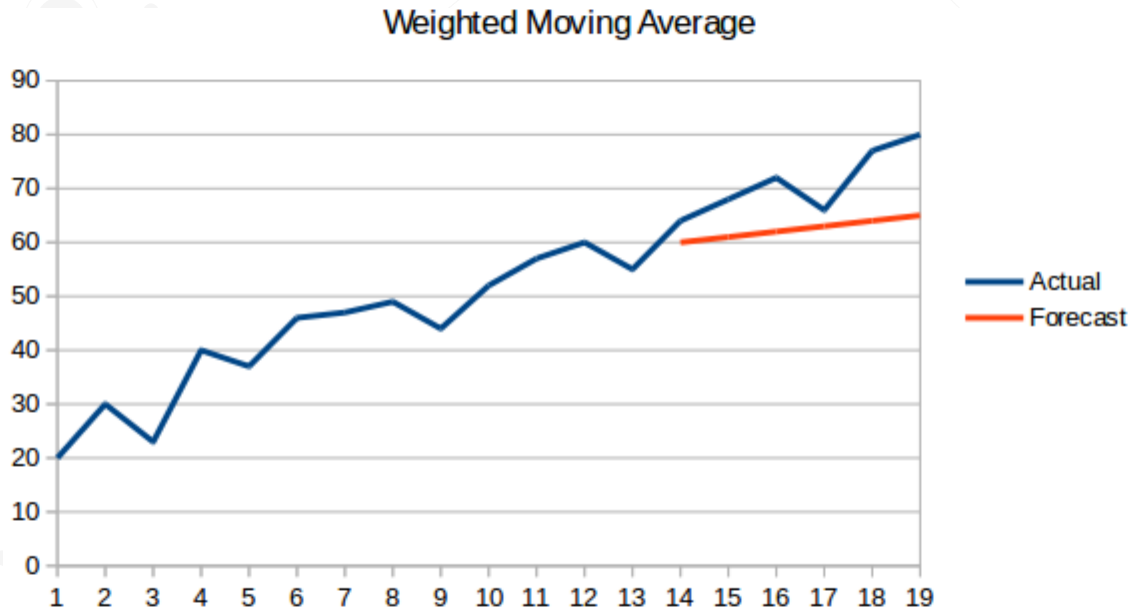
Modelamiento y predicción

3. Valor predicho es el promedio de **algunos** valores previos



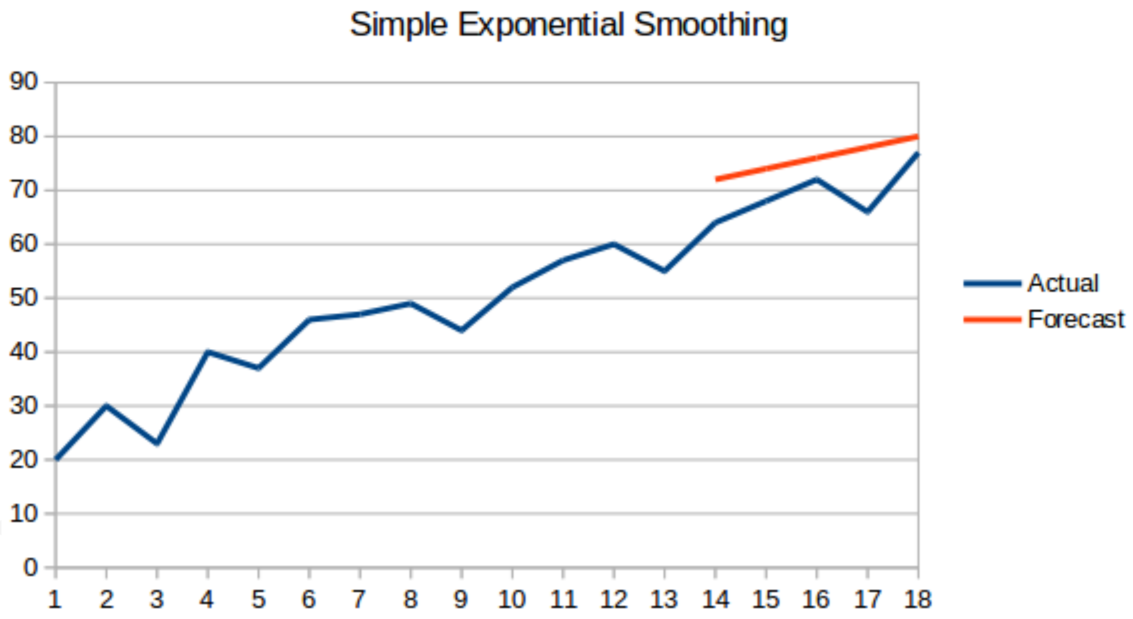
Modelamiento y predicción

4. Valor predicho es el promedio **ponderado** de **algunos** valores previos



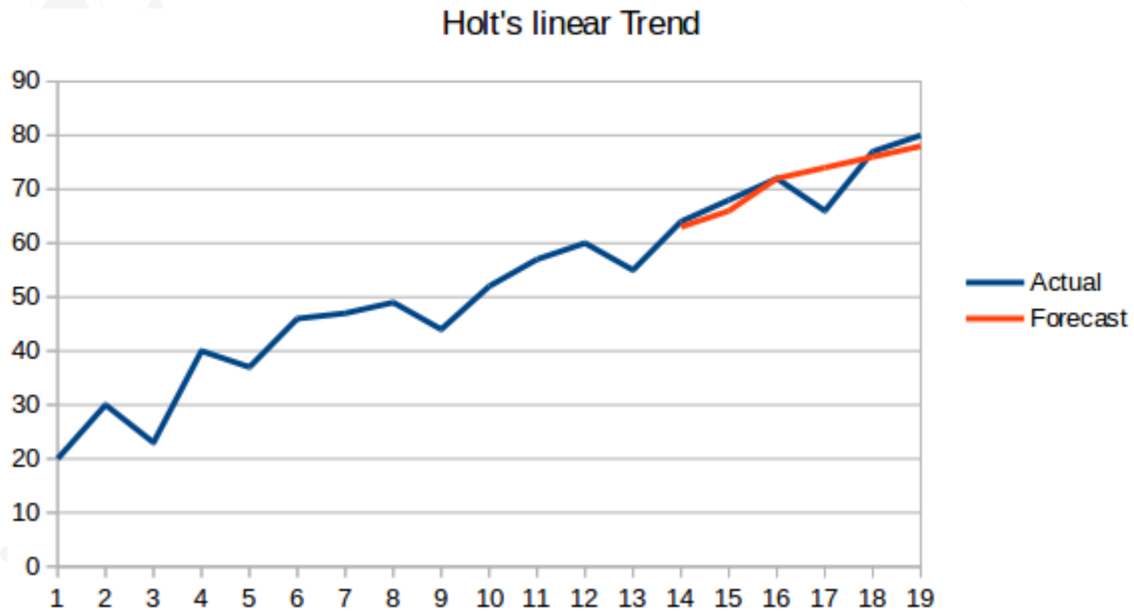
Modelamiento y predicción

5. Valor predicho es el promedio **ponderado** de **algunos** valores previos, dando mayor peso a observaciones más recientes



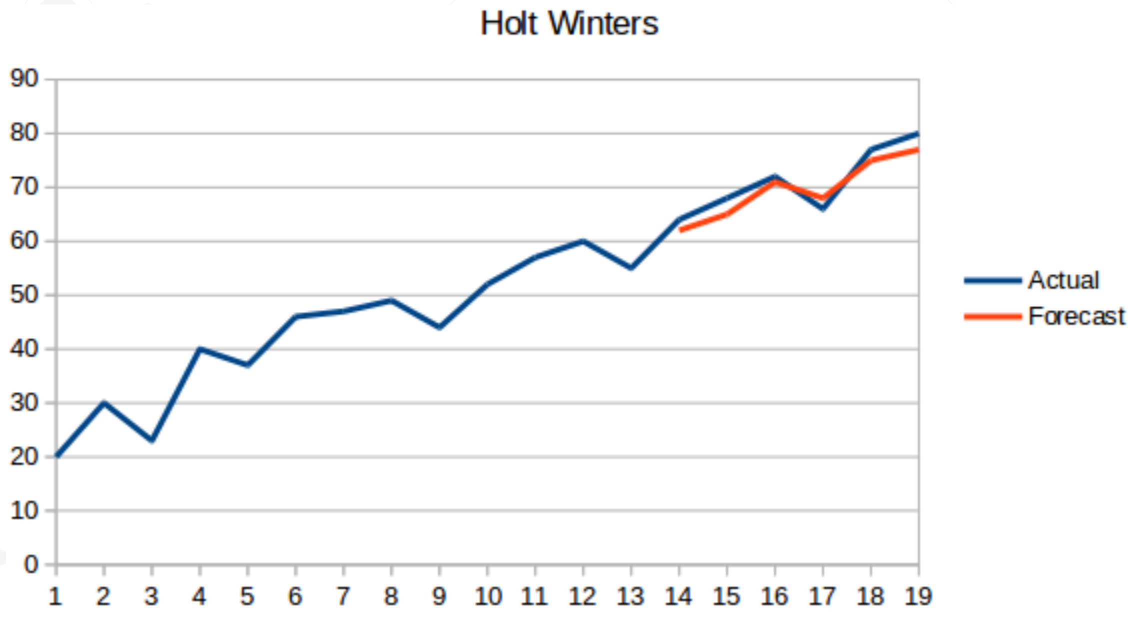
Modelamiento y predicción

6. La predicción toma en cuenta la **tendencia** de los datos



Modelamiento y predicción

7. La predicción toma en cuenta la **tendencia** y **estacionalidad** de los datos



Modelamiento y predicción

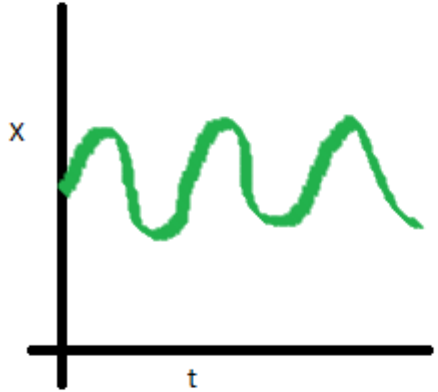
Importante: Para poder construir un modelo de series de tiempo, la serie debe cumplir con la condición de *estacionaridad*

Una serie se dice estacionaria si cumple lo siguiente:

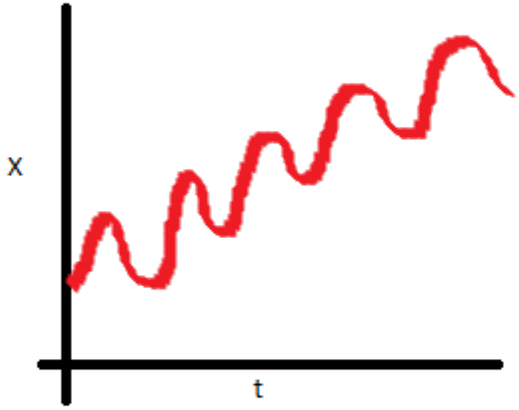
i. El valor esperado es el mismo en cualquier instante.
ii. La varianza es la misma en cualquier instante.
iii. La covarianza entre dos instantes es igual a si se toma en otros dos instantes.

Modelamiento y predicción

La media no es función del tiempo



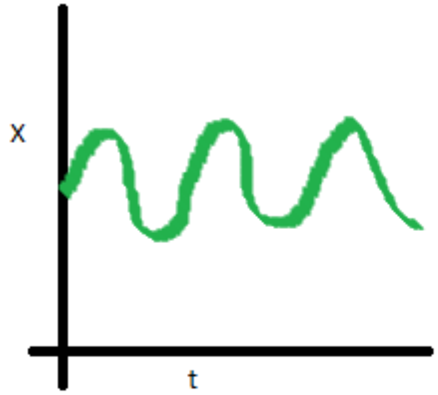
Serie estacionaria



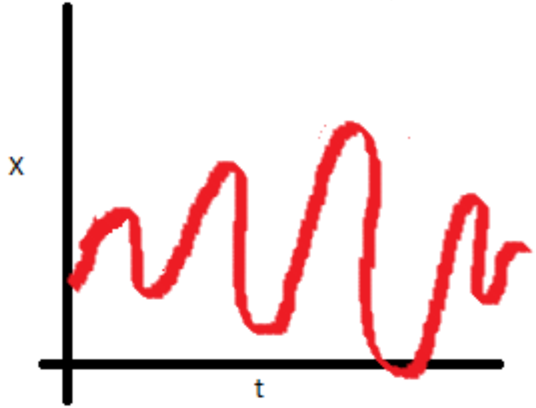
Serie NO estacionaria

Modelamiento y predicción

La varianza no es función del tiempo



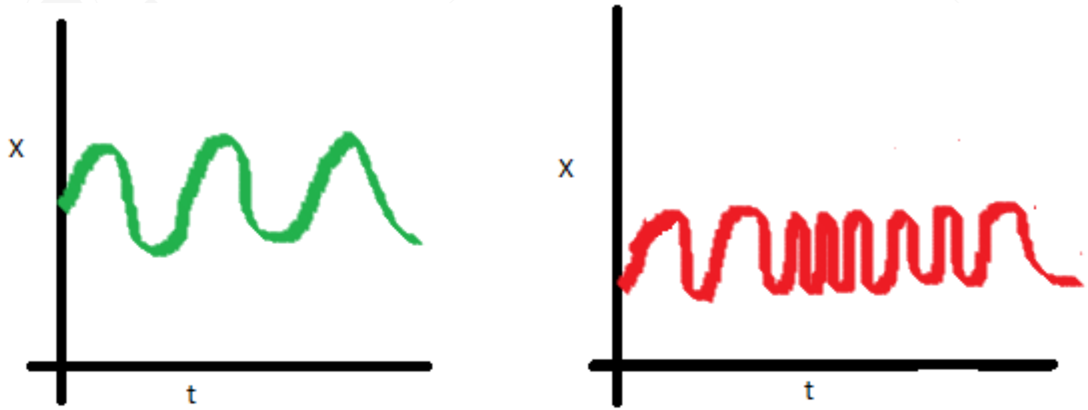
Serie estacionaria



Serie NO estacionaria

Modelamiento y predicción

La covarianzia entre términos no es función del tiempo



Serie estacionaria **Serie NO estacionaria**

Modelamiento y predicción

Una serie se puede transformar en estacionara si se aplica una **diferenciación**. Este método consiste en restar la variable con ella misma, un tiempo desfasado, es decir, $\boldsymbol{Z_t = Y_t - Y_{t-1}}$

En R, la función **diff()** realiza una diferenciación de la serie.

Modelamiento y predicción

La metodología propuesta por **Box-Jenkins** es muy popular, ya que propone manipular cualquier serie temporal, y está implementado en diversos software:

i. Verificar que la serie sea estacionaria.

ii. Identificar un modelo tentativo

iii. Estimar el modelo

iv. Verificar el diagnóstico

v. Usar el modelo para pronosticar

Modelamiento y predicción

El método de Holt-Winters es un método que utiliza un **suavizamiento exponencial**, basado en Media, Tendencia y Estacionalidad.

En **R**, el comando **HoltWinters()** recibe una serie de tiempo y retorna un posible ajuste

El ajuste obtenido puede utilizarse como input en la función **predict()**, con el argumento “n.ahead” para predecir n períodos.

Actividad II

Utilice nuevamente el archivo AirCanada.xlsx y realice lo siguiente:

- Grafique la serie nuevamente e indique si esta es estacionaria.
- Utilice la función **diff()** para diferenciar la serie ¿Qué puede concluir sobre la estacionaridad ahora?
- Utilice el método de Holt-Winters para ajustar la serie y realice una predicción para los 24 meses futuros (es decir, Julio 2018 a Junio 2020).
- Grafique la serie y la predicción. Comente.
- ¿Cuál es el número de pasajeros que es espera arriben en Julio de 2019?

Series de Tiempo

ARMA y **ARIMA** son modelos muy utilizados para analizar series de tiempo.

En un modelo ARMA, AR denota un proceso auto - regresivo mientras que MA denota un proceso con media móvil.

En un modelo ARIMA, la “I “ indica el número de diferencias usadas para hacer que la serie de tiempo sea estacionaria

Series de Tiempo

Modelo AR(1): asume que el resultado actual sólo depende del resultado anterior.

Puede generalizarse al caso AR(p)

Modelo MA(1): asume que el resultado actual es un promedio (ponderado) móvil de errores.

Puede generalizarse al caso MA(q)

Series de Tiempo

Una vez que la serie puede ser considerada como estacionaria, debe decidirse qué tipo de proceso la representa de mejor manera (i.e., AR, MA o ARMA)

Una decisión relacionada es la elección del orden p y/o q de los procesos.

El **ACF** y **PACF** son herramientas útiles para la elección de procesos adecuados.

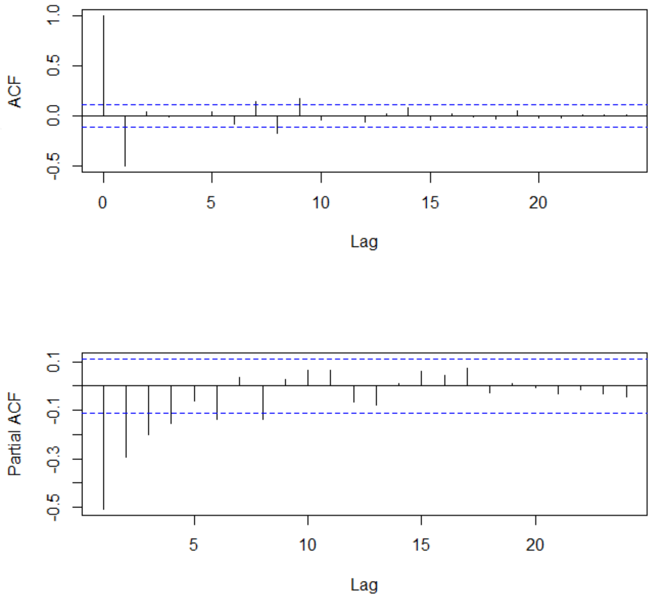
Series de Tiempo

	ACF	PACF
AR	Decae a 0	Cortan en 0 al rezago p
MA	Cortan en 0 al rezago q	Decae a 0
ARMA	Decae a 0	Decae a 0

En el caso de un AR, se verá un ACF con valores altos en los rezagos decayendo a 0, y en PACF con los primeros p rezagos distintos de 0.

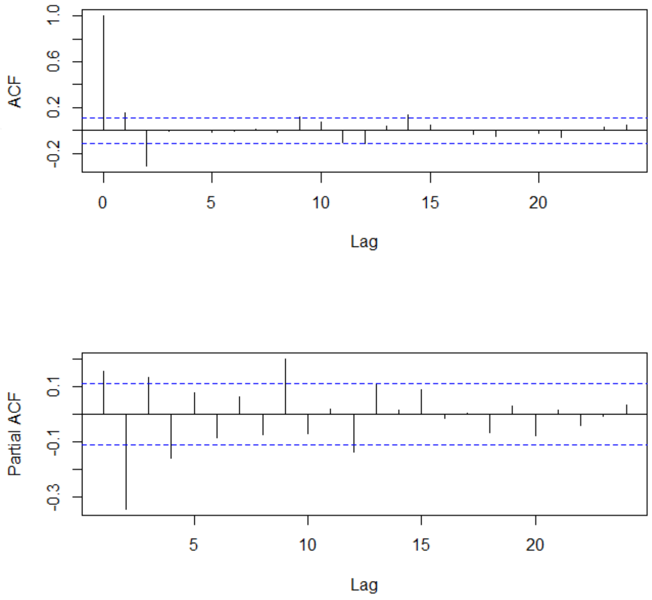
En el caso de un MA, el ACF tendrá los primeros q rezagos distintos de 0, y el PACF mostrará varios valores altos decayendo a 0.

Series de Tiempo



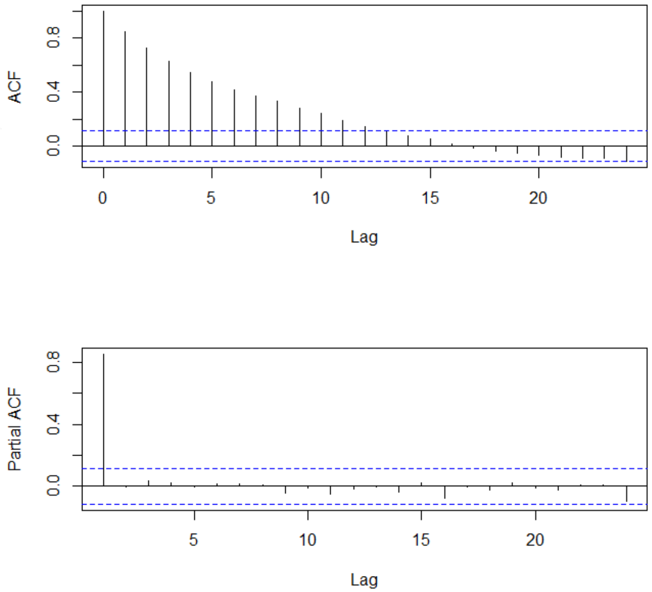
MA(1)

Series de Tiempo



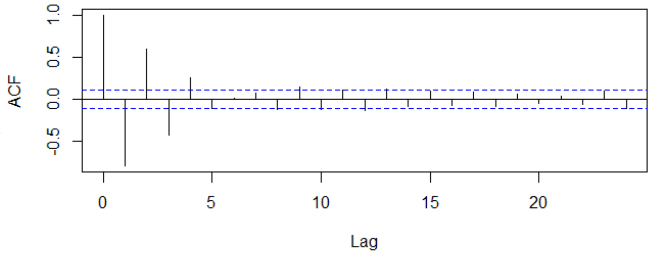
MA(2)

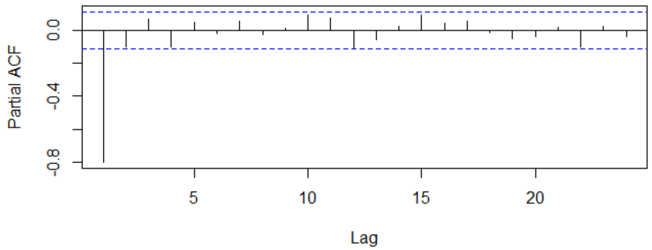
Series de Tiempo



AR(1)

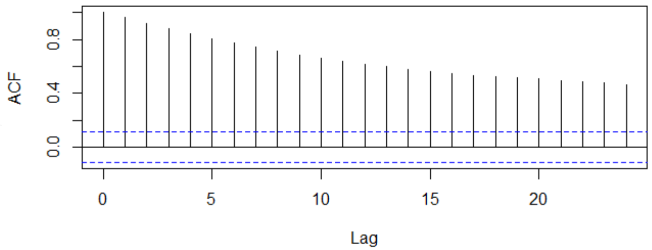
Series de Tiempo

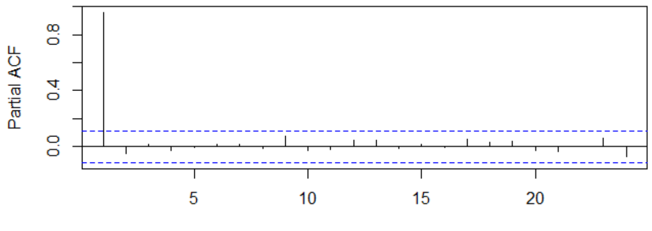




AR(1) (Coeficiente negativo)

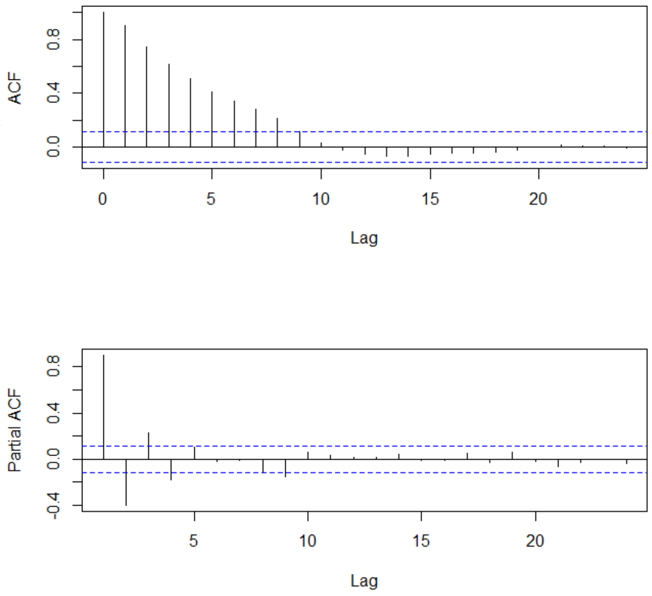
Series de Tiempo





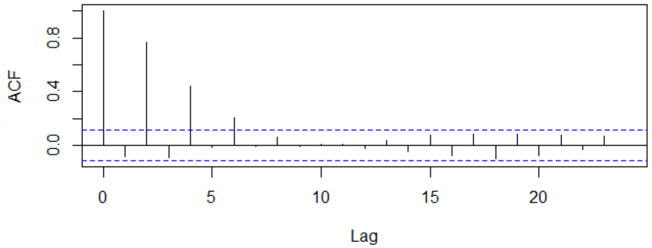
Serie No estacionaria

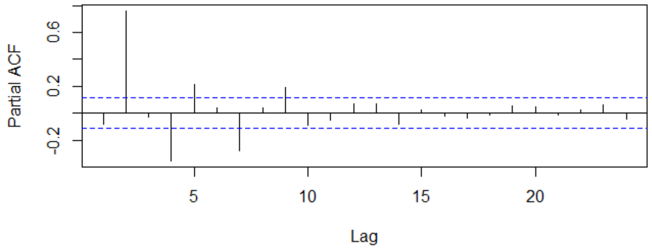
Series de Tiempo



ARMA(1,1)

Series de Tiempo





ARMA(3,2)

Análisis de homocedasticidad

¿Qué pasa si la serie no es estacionaria?

Con el comando **BoxCox.ar()** de la librería **TSA** nos muestra un gráfico que nos permite ver si es necesario realizar una transformación:

- Si el intervalo de confianza contiene al 0 se aplica logaritmo natural **log()**
- Si el intervalo de confianza contiene al 1 no se realiza transformación
- En otro caso se realiza la trasformación:

  $\frac{x_t^{\lambda}-1}{\lambda}$, donde λ es la línea central del intervalo, que se extrae con el argumento **$mle**

Análisis de homocedasticidad

- Cuando una serie de tiempo tiene una raíz unitaria se tiene que nuestra serie de tiempo NO es estacionaria, dado que su varianza y su media no se mantendría constante con el tiempo.

En **R** podemos testear la existencia de raíces unitarias con el Test de Dickey-Fuller con el comando **adf.test()** de la librería **tseries** se hace el siguiente test:

H0: La serie de tiempo tiene una raíz unitaria (no es estacionaria)

H1: La serie de tiempo NO tiene una raíz unitaria.

SARIMAX

Un enfoque moderno son los modelos SARIMAX. Esta propuesta se puede descomponer de la siguiente forma:

SARIMAX

Seasonal:
Permite modelar la estacionalidad

AutoRegressive:
Genera la relación de la variable consigo misma

Integrate:
Permite diferenciar (o integrar) la serie

Moving Average:
Genera la relación de la variable con un promedio variable

Covariable X:
Permite considerar una o más variables predictoras

SARIMAX

Los modelos SARIMAX pueden partir con submodelos más simples, como:

i. **AR(p)**: Modela la serie de forma autoregresiva a *p* rezagos de distancia.

ii. **MA(q)**: Modela la serie con promedios móviles a *q rezagos* de distancia.

iii. **ARMA(p,q)**: Modela la serie de forma autoregresiva y con promedios móviles.

iv. **ARIMA(p,d,q)**: Modela la serie con d diferenciaciones, y de forma autoregresiva con promedios móviles.

SARIMAX

En R, la función **arima()** ajusta modelos AR, MA, ARMA, ARIMA, SARIMA y la versión SARIMAX que mezcla estos modelos con una regresión lineal.

Para testear si los residuos son independientes, se utiliza el test de Ljung-Box, implementado en la función **checkresiduals()** de la librería **forecast.**

Por otro lado, la librería **forecast** contiene la función **auto.arima()** que entrega como output un posible modelo. ¡Cuidado! No necesariamente es el mejor, o incluso no cumple todos los supuestos.

Para hacer predicciones, se utiliza el comando **predict()** **o forecast()**

Actividad III

Utilice nuevamente el archivo AirCanada.xlsx y realice lo siguiente:

- Realice un análisis de homocedasticidad, tendencia y de estacionalidad.
- Evalúe la estacionaridad de la serie utilizando el test de Dickey-Fuller
- Utilizando ACF y PACF ajuste un modelo ARMA. Comente el ajuste.
- Realice un diagnóstico de los residuos.
- Compare ajustado con aquel que entrega como output la función auto.arima()
- Con el modelo seleccionado, realice una predicción a dos años del total de pasajeros.