Contents

Econometría Aplicada. Lección 1

Marcos Bujosa

June 14, 2024

Carga de algunos módulos de python

```python
# Para trabajar con los datos y dibujarlos necesitamos cargar algunos módulos de python
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib as mpl
import matplotlib.pyplot as plt   # data visualization
mpl.rc('text', usetex=True)
mpl.rc('text.latex', preamble=r'\usepackage{amsmath}')
```

1 Procesos estocásticos y datos de series temporales

Proceso estocástico es una secuencia de variables aleatorias X_t

$$\boldsymbol{X} = \{X_t \mid t = 0, \pm 1, \pm 2, \ldots\};$$

donde el conjunto de índices t recorre el conjunto de números enteros ($\mathbb{Z}$).

Serie temporal es una secuencia de datos tomados a lo largo del tiempo

$$\boldsymbol{x} = (x_1, x_2, \ldots x_n)$$

- Consideraremos cada dato x_t como una *realización de una variable aleatoria* X_t.
- Consecuentemente, consideraremos que una *serie temporal* es una *realización de un tramo* de un proceso estocástico:

$$(x_1, x_2, \ldots x_n) \text{ es una realización de } \{X_t \mid t = 1, 2, \ldots n\}.$$

1.1 El desafío

El análisis de *series temporales* trata sobre la inferencia estadística de muestras que **frecuentemente NO podemos asumir que sean realizaciones** de variables aleatorias *i.i.d.* (*independientes e idénticamente distribuidas*).

Además,

- Aunque el marco ideal es que la serie temporal analizada "**sea estacionaria**"
(*abuso del lenguaje que expresa que podemos asumir que la serie es una realización de un proceso estocástico estacionario*)

- lo habitual es que, por distintos motivos, **NO lo sea**

El desafío para el analista es

primero transformar los datos para lograr que sean "***estacionarios***"

y después transformar los datos estacionarios en una secuencia de "**datos *i.i.d***"
(*nuevo abuso del lenguaje que expresa que podemos asumir que los datos son realizaciones de variables aleatorias i.i.d.*)

2 Estacionariedad

El mayor objetivo del *análisis de series temporales* es inferir la distribución de un proceso $\boldsymbol{X} = \{X_t\}$ usando una muestra (serie temporal) $\boldsymbol{x} = (x_1, x_2, \dots x_n)$. Así podremos

Predecir datos futuros

Controlar datos futuros

Pero esto es casi imposible si los datos son inestables o caóticos a lo largo del tiempo
Por tanto, algún tipo de <u>estabilidad</u> o <u>estacionariedad</u> es necesaria.

2.1 Estacionariedad en sentido débil

Un proceso estocástico $\boldsymbol{X} = \{X_t\}$ se dice **estacionario** (*en sentido débil*) si para todo t y k de $\mathbb{Z}$

$$E(X_t) = \mu \tag{1}$$

$$Cov(X_t, X_{t-k}) = \gamma_k \tag{2}$$

- La primera igualdad sugiere que las realizaciones de $\{X_t\}$ generalmente oscilan entorno a μ.
- La segunda sugiere que la variabilidad de las realizaciones de $\{X_t\}$ entorno a μ es constante, pues para el caso particular $k = 0$

$$Cov(X_t, X_{t-0}) = Var(X_t) = \gamma_0 \quad \text{para todo } t$$

  Es decir, γ_0 es la varianza común a todas las variables aleatorias del proceso.

2.2 Función de autocovarianzas y función de autocorrelación

- La secuencia $\{\gamma_k\}$ con $k \in \mathbb{Z}$ se denomina *función de autocovarianzas*
- La secuencia $\{\rho_k\}$ con $k \in \mathbb{Z}$, donde

$$\rho_k = \frac{Cov(X_t, X_{t-k})}{\sqrt{Var(X_t)Var(X_{t-k})}} = \frac{\gamma_k}{\gamma_0}$$

  se denomina *función de autocorrelación* (ACF).

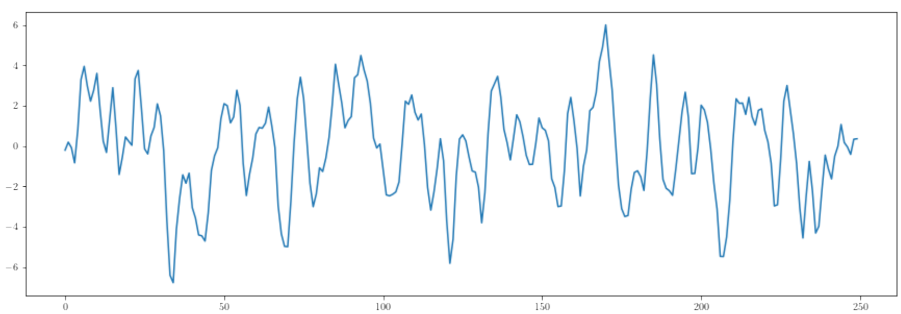
Debido a la estacionariedad, la correlación entre X_t y X_{t+k} no depende de t; tan solo depende de la distancia temporal k entre ambas variables.

Es más, la desigualdad de Chebyshev

$$P\left(|X_t - \mu| \geq c\sigma\right) \leq \frac{1}{c^2}, \qquad \text{donde } \sigma = \sqrt{\gamma_0}$$

sugiere que para cualquier proceso estacionario (y un c grande), al pintar una realización, tan solo un pequeño porcentaje de los datos caerán fuera de la franja $(\mu - c\sigma, \mu + c\sigma)$.

```
import statsmodels.api as sm
np.random.seed(12345)
arparams = np.array([.75, -.25])
maparams = np.array([.65, .35])
ar = np.r_[1, -arparams] # add zero-lag and negate
ma = np.r_[1, maparams] # add zero-lag
y = sm.tsa.arma_generate_sample(ar, ma, 250)
plt.figure(figsize=(15,5))
plt.plot(y)
plt.savefig(image)    # "image" no definido. Comentar esta línea al ejecutar el notebook
```



3 Transformaciones de realizaciones de procesos estocásticos NO estacionarios

Un proceso estocástico $\mathbf{X} = \{X_t\}$ puede ser

NO estacionario en media porque $E(X_t)$ depende de t.

NO estacionario en covarianza porque $Cov(X_t, X_{t-k})$ depende de t.

Separar o distinguir ambos tipos de no estacionariedad no es sencillo.

Veamos ejemplos de series temporales para los que

- no podemos asumir que son realizaciones de procesos estocásticos estacionarios
- y algunos intentos de transformación para obtener datos "**estacionarios**" (*) (*recuerde que esta expresión, aunque extendida, es un abuso del lenguaje*).

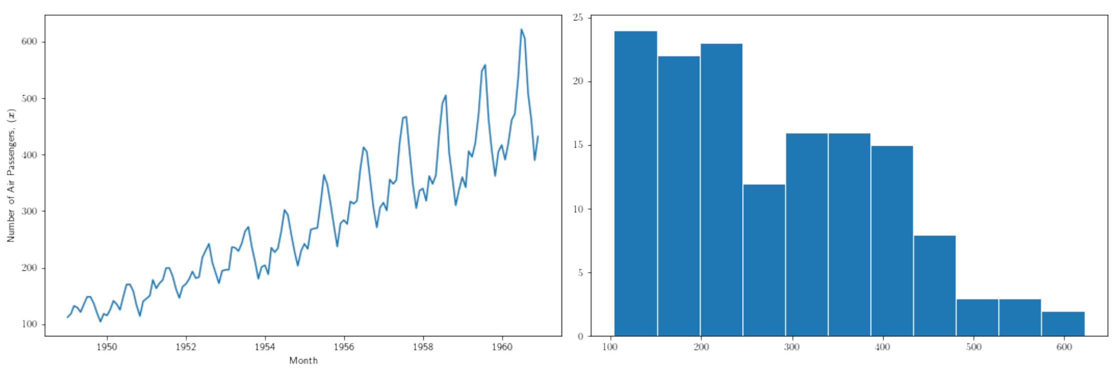
3.1 Internat. airline passengers: monthly totals in thousands. Jan 49 – Dec 60

```
path = './database/Datasets-master/airline-passengers.csv'
data = pd.read_csv(path)
data['Month']=pd.to_datetime(data['Month'])
data=data.set_index(['Month'])
print(data.head())
```

```
data['data_log'] = np.log(data)
data['data_log_diff'] = data['data_log'].diff(1)
data['data_log_diff_diff12'] = data['data_log_diff'].diff(12)
print(data.head())
print(data.tail())
```

$$\boldsymbol{x} = (x_1, \dots x_{114})$$

```
plt.figure(figsize=(15,5))
plt.subplot(1, 2, 1)
plt.plot(data['Passengers'])
plt.xlabel("Month")
plt.ylabel(r"Number of Air Passengers, ($\boldsymbol{x}$)")
plt.subplot(1, 2, 2)
plt.hist(data['Passengers'], edgecolor='white', bins=11)
plt.tight_layout()
plt.savefig(image)   # "image" no definido. Comentar esta línea al ejecutar el notebook
```



Serie *"no estacionaria"* (*):

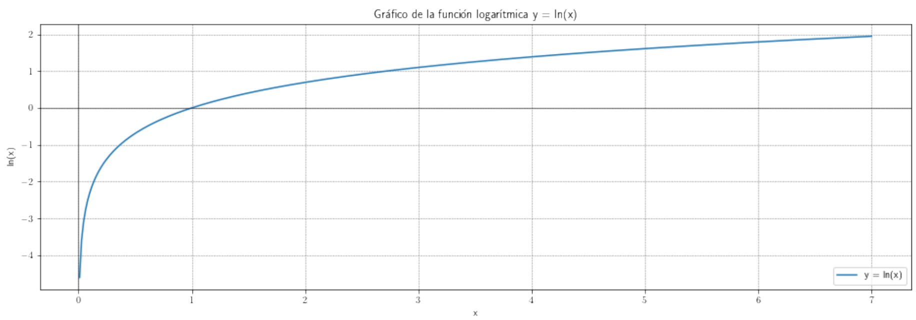
- La media crece de año en año
- La variabilidad estacional crece de año en año (fíjese en la diferencia entre el verano y el otoño de cada año)

3.1.1 Trasformación logarítmica de los datos

- Al aplicar la función logarítmica transformamos **monótonamente** los datos estabilizando la varianza cuando los valores son mayores que 0.567 (aprox.)
- ¡Pero ocurre lo contrario, pues se amplifica el valor absoluto, cuando los valores están entre 0 y 0.567! De hecho, $\lim_{x \to \infty} \ln(x) = -\infty$

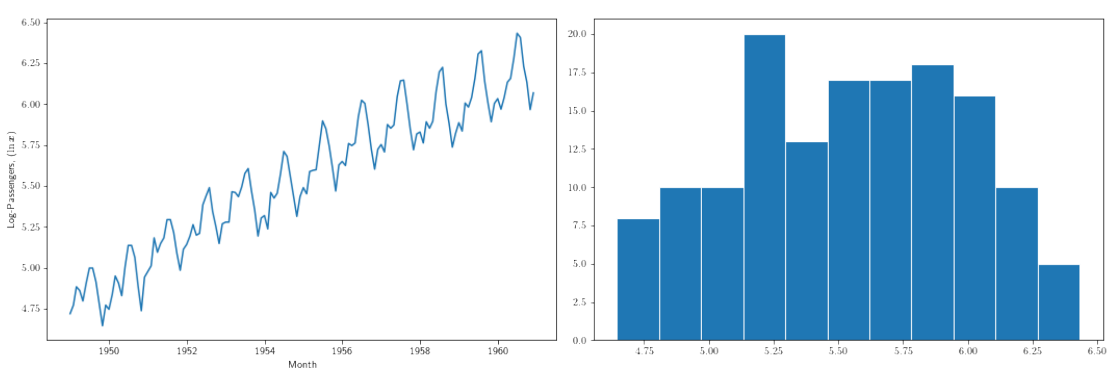
- Además, *el logaritmo no está definido para valores negativos.*

```python
# Definir el rango de valores para x (empezando desde un número positivo ya que log(0) no está definido)
x = np.linspace(0.01, 7, 400)  # Valores de 0.1 a 10

# Calcular y = log(x)
y = np.log(x)

# Crear el gráfico
plt.figure(figsize=(16, 5))
plt.plot(x, y, label='y = ln(x)')

# Añadir etiquetas y título
plt.xlabel('x')
plt.ylabel('ln(x)')
plt.title('Gráfico de la función logarítmica y = ln(x)')
plt.axhline(0, color='black',linewidth=0.5)
plt.axvline(0, color='black',linewidth=0.5)
plt.grid(color = 'gray', linestyle = '--', linewidth = 0.5)
plt.legend()

plt.savefig(image)   # "image" no definido. Comentar esta línea al ejecutar el notebook
```



$$\ln \boldsymbol{x} = \Big(\ln(x_1), \ldots \ln(x_{114})\Big)$$

```python
plt.figure(figsize=(15,5))
plt.subplot(1, 2, 1)
plt.plot(data['data_log'])
plt.xlabel("Month")
plt.ylabel(r"Log-Passengers, ($\ln\boldsymbol{x}$) ")
plt.subplot(1, 2, 2)
plt.hist(data['data_log'], edgecolor='white', bins=11)
plt.tight_layout()
plt.savefig(image)   # "image" no definido. Comentar esta línea al ejecutar el notebook
```



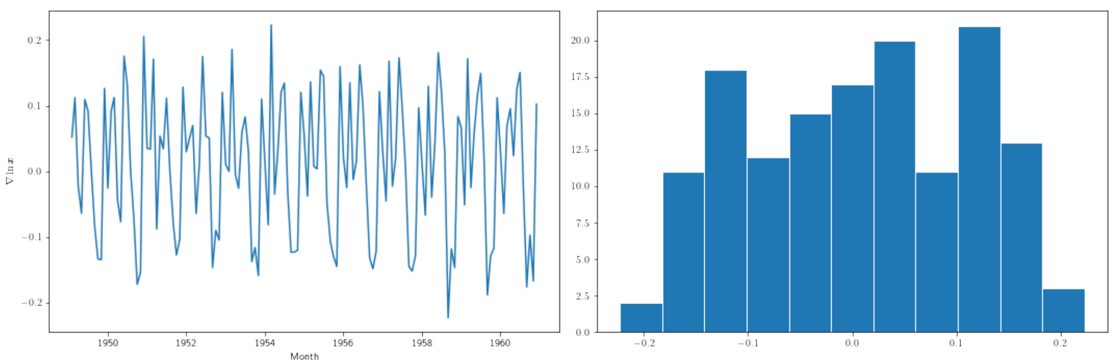
Ésta tampoco parece la realización de un proceso estocástico *estacionario*

- Ahora la variabilidad estacional parece mantenerse de año en año
- Pero la media sigue creciendo de año en año

3.1.2 Primera diferencia de la trasformación logarítmica de los datos

$$\boldsymbol{y} = \nabla \ln \boldsymbol{x} = \Big(\big[\ln(x_2) - \ln(x_1)\big], \ldots \; \big[\ln(x_{114}) - \ln(x_{113})\big] \Big)$$

```
plt.figure(figsize=(15,5))
plt.subplot(1, 2, 1)
plt.plot(data['data_log_diff'])
plt.xlabel("Month")
plt.ylabel(r"$\nabla\ln\boldsymbol{x}$")
plt.subplot(1, 2, 2)
plt.hist(data['data_log_diff'], edgecolor='white', bins=11)
plt.tight_layout()
plt.savefig(image)   # "image" no definido. Comentar esta línea al ejecutar el notebook
```



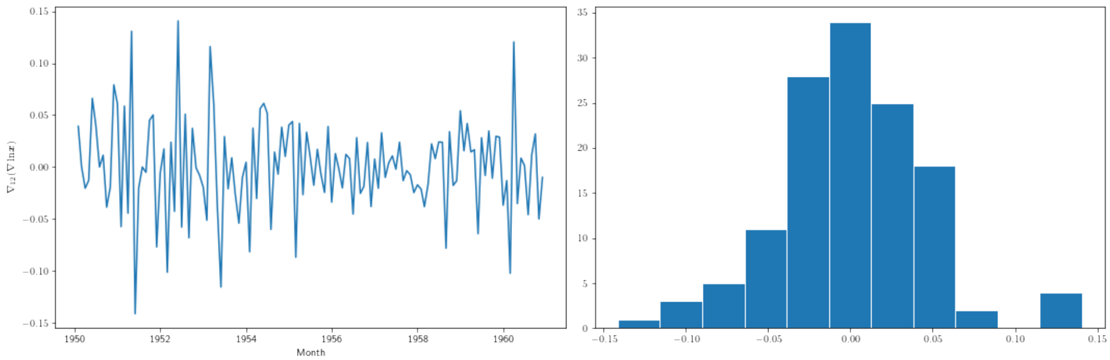
Esta serie tampoco parece *"estacionaria"* (*)

- Hay un componente periódico (de naturaleza estacional), debido a que hay pocos viajes en otoño y muchos en Navidad, Semana Santa y verano (i.e., el número esperado de viajeros parece cambiar en función del mes o estación del año).

3.1.3 Diferencia estacional de la primera diferencia de la trasformación logarítmica de los datos

$$\boldsymbol{z} = \nabla_{12}(\nabla \ln \boldsymbol{x}) = \nabla_{12}(\boldsymbol{y}) = \Big((y_{13} - y_1), \ldots \; (y_{113} - y_{101}) \Big)$$

```
plt.figure(figsize=(15,5))
plt.subplot(1, 2, 1)
plt.plot(data['data_log_diff_diff12'])
plt.xlabel("Month")
plt.ylabel(r"$\nabla_{12}(\nabla\ln\boldsymbol{x})$")
plt.subplot(1, 2, 2)
plt.hist(data['data_log_diff_diff12'], edgecolor='white', bins=11)
plt.tight_layout()
plt.savefig(image)   # "image" no definido. Comentar esta línea al ejecutar el notebook
```



Esta serie se aproxima más al aspecto de la realización de un proceso *estacionario*

- Aunque parece haber más varianza a principios de los 50 que a finales
- De propina, el histograma sugiere una distribución aproximadamente Gaussiana

3.2 Tasa logarítmica de crecimiento

La tasa logarítmica de variación de $\boldsymbol{y}$ se define como

$$\boldsymbol{z} = \nabla \ln \boldsymbol{y} = \Big(\big[\ln(y_2) - \ln(y_1)\big], \dots \; \big[\ln(y_n) - \ln(y_{n-1})\big] \Big)$$

es decir: $z_t = \ln y_t - \ln y_{t-1}$

Es una aceptable *aproximación* de la tasa de crecimiento (en tanto por uno) si los incrementos son pequeños.

t	y_t	Incremento en tanto por uno	$\ln y_t$	Primera diferencia de $\ln \boldsymbol{y}$	Incremento en tanto por uno desde $t = 1$	$\ln y_t - \ln y_1$
1	100.		4.605170			
2	101.00000	0.01	4.615120	0.0100	0.0100	0.0100
3	102.01000	0.01	4.625071	0.0100	0.0201	0.0199
4	103.03010	0.01	4.635021	0.0100	0.0303	0.0299
5	104.06040	0.01	4.644971	0.0100	0.0406	0.0398
6	105.10100	0.01	4.654922	0.0100	0.0510	0.0498
7	106.15201	0.01	4.664872	0.0100	0.0615	0.0597
8	107.21353	0.01	4.674823	0.0100	0.0721	0.0697
9	108.28567	0.01	4.684773	0.0100	0.0829	0.0796
10	109.36853	0.01	4.694723	0.0100	0.0937	0.0896

3.2.1 Observaciones sobre los datos transformados

Transformación de la serie temporal $\boldsymbol{y} = \{y_t\},\ t = 1:n$	Observaciones
$\boldsymbol{z} = \ln \boldsymbol{y} = \{\ln y_t\}$	A veces independiza la volatilidad del nivel e induce normalidad.
$\boldsymbol{z} = \nabla \boldsymbol{y} = \{y_t - y_{t-1}\}$	Indica al crecimiento absoluto entre periodos consecutivos.
$\boldsymbol{z} = \nabla \ln \boldsymbol{y}$	Tasa logarítmica de crecimiento. Aproximación del crecimiento relativo entre periodos consecutivos.
$\boldsymbol{z} = \nabla\nabla \ln \boldsymbol{y} = \nabla^2 \ln \boldsymbol{y}$	Cambio en la tasa log, de crecimiento. Indica la "aceleración" en el crecimiento relativo.
$\boldsymbol{z} = \nabla_s \ln \boldsymbol{y} = \{\ln y_t - \ln y_{t-s}\}$	Tasa de crecimiento acumulada en un ciclo estacional completo (s períodos). Cuando el período estacional es de un año, se conoce como "tasa anual" o "tasa interanual".
$\boldsymbol{z} = \nabla\nabla_s \ln \boldsymbol{y}$	Cambio en la tasa de crecimiento acumulada en un ciclo estacional completo. Es un indicador de aceleración en el crecimiento acumulado.