

Séptima Clase de Análisis de Datos

Prof: Boris Panes
Universidad Del Desarrollo

Octubre 19, 2024

*Resumen general del trimestre de Análisis de Datos
Modelamiento matemático de datos estructurados y evaluación de resultados*

Durante la clase conversaremos sobre los proyectos T2 y T3

Asistencia por ZOOM: 28 alumnos promedio con asistencia sobre 120 minutos por clase

Asistencia presencial: 5 alumnos como mínimo todas las clases

33 alumnos efectivos por clase de un total de 38

Evaluación online: 3+1 (promedio de las mejores 3 notas)

Preparación de datos: 26 alumnos han completado la evaluación

Test estadístico: 17 alumnos han completado la evaluación

Regresión lineal simple: 18 alumnos han completado la evaluación

Regresión lineal múltiple y comparación de modelos: *en preparación*

Evaluaciones de los proyectos T1, T2 y T3: revisión general más detalles

Inicialmente se revisa el cumplimiento con el formato sugerido y la coherencia del trabajo en general. Por otro lado también se procede con una revisión de los puntos del algoritmo mostrado en clases. Las discrepancias relevantes son comunicadas como explicación de la nota final

Comentarios del proyecto T2

El requerimiento fundamental de T2 era escoger un par de variables numéricas a partir del conjunto de datos para realizar un análisis de regresión lineal simple

X: Variable independiente

Y: Variable dependiente

Se espera que las variables muestren una alta correlación

Además podríamos agregar que los residuos obtenidos luego del ajuste deberían distribuir normal

En términos de interpretación causal la variable X debería ser una variable cuyos valores podemos controlar, como por ejemplo el tamaño de las clases o el tiempo.

Es posible determinar con anticipación el tamaño de las clases

Es posible establecer con anticipación los tiempos a medir

Discusión causa-efecto de las variables consideradas

Entre las variables escogidas tenemos los siguientes ejemplos

X	Y
ausencias	notas de exámenes
Ozone	PM 2.5
shazam	streams
tiempo de estudio	notas de exámenes
GDP	calidad de vida
pH	solidos
escolaridad	expectativa de vida
mt ² y numero de piezas	precios de casas
notas de test 1	notas de test 2
propaganda televisiva	ventas

En la mayoría de los casos la variable X ocurre antes que la variable Y. También podemos notar que la variable X tiende a ser un factor mas controlable, mientras que la variable Y es un observable de carácter más espontaneo

Se ocuparon variadas estrategias para seleccionar X e Y

Bases para Proyecto T3

Objetivos: Analizar un nuevo conjunto de datos que contenga información sobre variables que evolucionan con la variable tiempo. Identificar variable relevante y secundarias. Revisar estacionariedad y aplicar substracción si es necesario. Modelar desde AR(1) hasta AR(4). Evaluación comparada y discusión de resultados.

Se aplican las mismas reglas generales de T1. En particular, la presentación debería tener entre 10 y 15 laminas y durar entre 15 y 30 minutos. El algoritmo del notebook, hilo conductor de la presentación y esquema del video debe considerar los siguientes pasos

- Seleccionar datos (csv de kaggle u otros)

 - Debe contener al menos 4 columnas dependientes del tiempo

- Descripción del contexto de la columna principal (variable dependiente)

- Preparación de datos

 - Discusión de elementos presentes en la serie: estacionariedad, tendencias, quiebres

 - Aplicación de transformación de variables para reducir tendencia

- Aplicación sistemática de AR(1) hasta AR(4) usando LinearRegresión

- Evaluación comparada de modelos usando métricas relevantes (R^2 ajustado, EP)

- Discusión de resultados

Calendario y Evaluaciones

TRIM.	FECHA	HORA
TRIM.2	sábado, 24 de agosto de 2024	11.20 - 12.30 12.30 - 13.40
TRIM.2	sábado, 31 de agosto de 2024	11.20 - 12.30 12.30 - 13.40
TRIM.2	sábado, 7 de septiembre de 2024	11.20 - 12.30 12.30 - 13.40
TRIM.2	sábado, 14 de septiembre de 2024	11.20 - 12.30 12.30 - 13.40
TRIM.2	sábado, 28 de septiembre de 2024	11.20 - 12.30 12.30 - 13.40
TRIM.2	sábado, 5 de octubre de 2024	11.20 - 12.30 12.30 - 13.40
TRIM.2	sábado, 19 de octubre de 2024	11.20 - 12.30 12.30 - 13.40

Más ejercicios con múltiples alternativas: 3+1 evaluaciones

50%

50%

Entrega T1: Limpieza y Estructura de Datos

Entrega T2: Regresión Lineal

Entrega T3: Series Temporales

Fecha de cierre de evaluaciones online y entrega de T3: **Domingo 27 de Octubre**

Resumen de contenidos

Contenidos de Series Temporales

Explicaciones generales

Definición y ejemplos

Diagnóstico del problema usando caso de estudio

Selección de caso práctico, relevante y autocontenido

Presentación de Notebook con caso de estudio (**Guía para T3**)

Exploración de los datos

 Limpieza, formato, ingeniería de datos

 Estacionariedad

Autocorrelación

 Modelo de Autoregresión de primer orden

Detalles técnicos y extensiones

Modelos autorregresivos (AR)

Análisis estadístico de los resultados

ARIMA y Autoregresión vectorial (VAR)

Recapitulación de resultados de AR(1)

Estacionariedad

Tendencias estocásticas y raíz unitaria

Diferenciación y substracción de tendencias

Modelo autorregresivo de orden p: AR(p)

Obtención de coeficientes para una regresión lineal múltiple

Modelo autorregresivo de retardos distribuidos: AR(p,q)

Evaluación comparada del ajuste de los modelos

ESR, R^2 y R^2 corregido

Discusión comparada entre los modelos AR(p,q)

Desarrollo de contenidos

Selección y exploración de datos para series temporales

En general, una serie temporal puede ser representada como una tabla con al menos 2 columnas y N filas. Cada fila representa un instante de tiempo, el cual es definido como la primera columna. El resto de las columnas son parámetros que evolucionan con el tiempo

Múltiples variables que dependen del tiempo

```
In [4]: data_inflacion[194:203]
```

```
Out[4]:
```

	freq	GDPC1	JAPAN_IP	PCECTPI	CPIAUCSL
194	7/1/2003	13372.357	99.142200	87.769	184.43
195	10/1/2003	13528.710	102.001053	88.124	184.80
196	1/1/2004	13606.509	103.271654	88.797	186.57
197	4/1/2004	13706.247	104.789317	89.421	188.60
198	7/1/2004	13830.828	105.636384	89.942	189.37
199	10/1/2004	13950.376	103.730482	90.652	191.03
200	1/1/2005	14099.081	105.777562	91.122	192.17
201	4/1/2005	14172.695	105.989329	91.728	193.67
202	7/1/2005	14291.757	105.636384	92.734	196.60

4 columnas que dependen del tiempo

Acrónimos de las columnas del dataset

En las paginas del Federal Reserve Bank of Sant Louis (FRED) podemos encontrar múltiples series de tiempo de tipo económico

<https://fred.stlouisfed.org/>

CPIAUCSL = Consumer Price Index for All Urban Consumers

<https://fred.stlouisfed.org/series/CPIAUCSL>

PCECTPI = Chain-type Price Index

<https://fred.stlouisfed.org/series/PCECTPI>

JAPAN_IP = Production, Sales, Work Started and Orders: Production Volume: Economic Activity: Industry (Except Construction) for Japan

<https://fred.stlouisfed.org/series/JPNPROINDQISMEI>

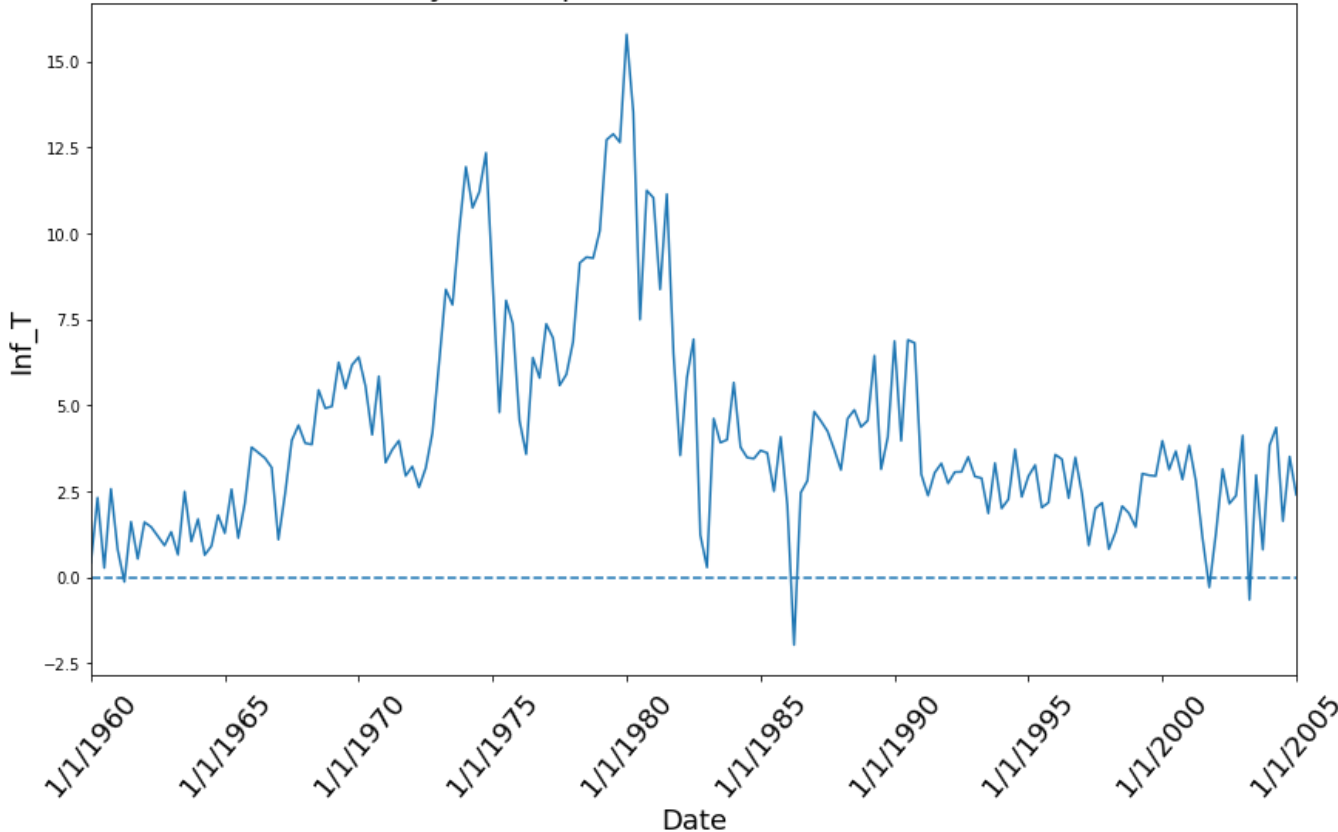
GDPC1 = Real Gross Domestic Product

<https://fred.stlouisfed.org/series/GDPC1>

Download, graph, and track [825,000 US and international time series](#) from [114 sources](#).

Presentación de datos iniciales

Inflación y desempleo en Estados Unidos, 1960-2004

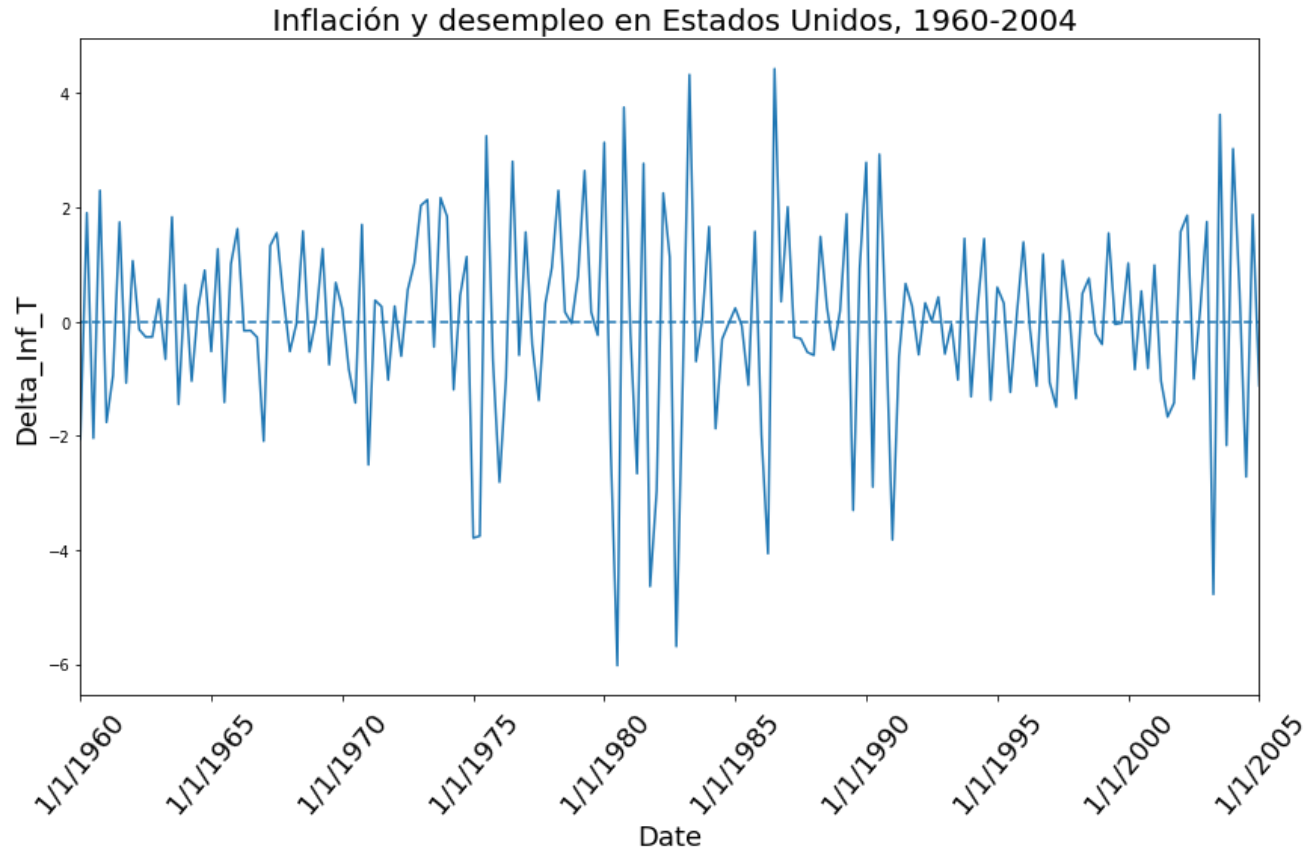


El IPC se calcula a partir de la razón entre los precios de una canasta base de productos

La inflación mide la variación del IPC. Además se incluye una actualización, es decir la variación de un trimestre se multiplica por 4 para obtener la variación proyectada de un año

$$\text{Inf}_t = 4 \times 100 \times \frac{\text{IPC}_t - \text{IPC}_{t-1}}{\text{IPC}_{t-1}}$$

Transformación de los datos iniciales



La variación de la tasa de inflación se obtiene de la substracción entre la tasa de inflación en t menos la tasa de inflación en $t-1$

$$\Delta \text{Inf}_t = \text{Inf}_t - \text{Inf}_{t-1}$$

Selección de Y_t para aplicar AR(1)

Cualitativamente podemos notar que Inf_t tiene un comportamiento donde se pueden observar oscilaciones, tendencias y quiebres. Por otro lado la serie de tiempo de Delta_Inf_t presenta un comportamiento mas similar a una función sinusoidal usual

Mas adelante veremos que en la práctica es posible verificar que Inf_t no es estacionaria mientras que Delta_Inf_t si lo es, considerando una definición cuantitativa de estacionariedad

Por esta razón, escogemos **Delta_Inf_t** como variabe dependiente del tiempo para aplicar AR(1)

Modelo genérico AR(1)

$$\longrightarrow \hat{Y}_t = \beta_0 + \beta_1 Y_{t-1}, \quad Y_{t,i} = \beta_0 + \beta_1 Y_{t-1,i} + u_i$$

Definición de Delta_Inf_t

$$\longrightarrow \Delta \text{Inf}_t = \text{Inf}_t - \text{Inf}_{t-1}$$

Selección de Y_t

$$\longrightarrow Y_t \leftarrow \Delta \text{Inf}_t$$

Regresión Lineal Simple

$$\longrightarrow X \leftarrow Y_{t-1}, \quad Y \leftarrow Y_t$$

Dado que vamos a aplicar AR(1) a la variable **Delta_Inf_t** es conveniente incorporar la variable desplazada en la tabla inicial

```
In [16]: df_mco
```

```
Out[16]:
```

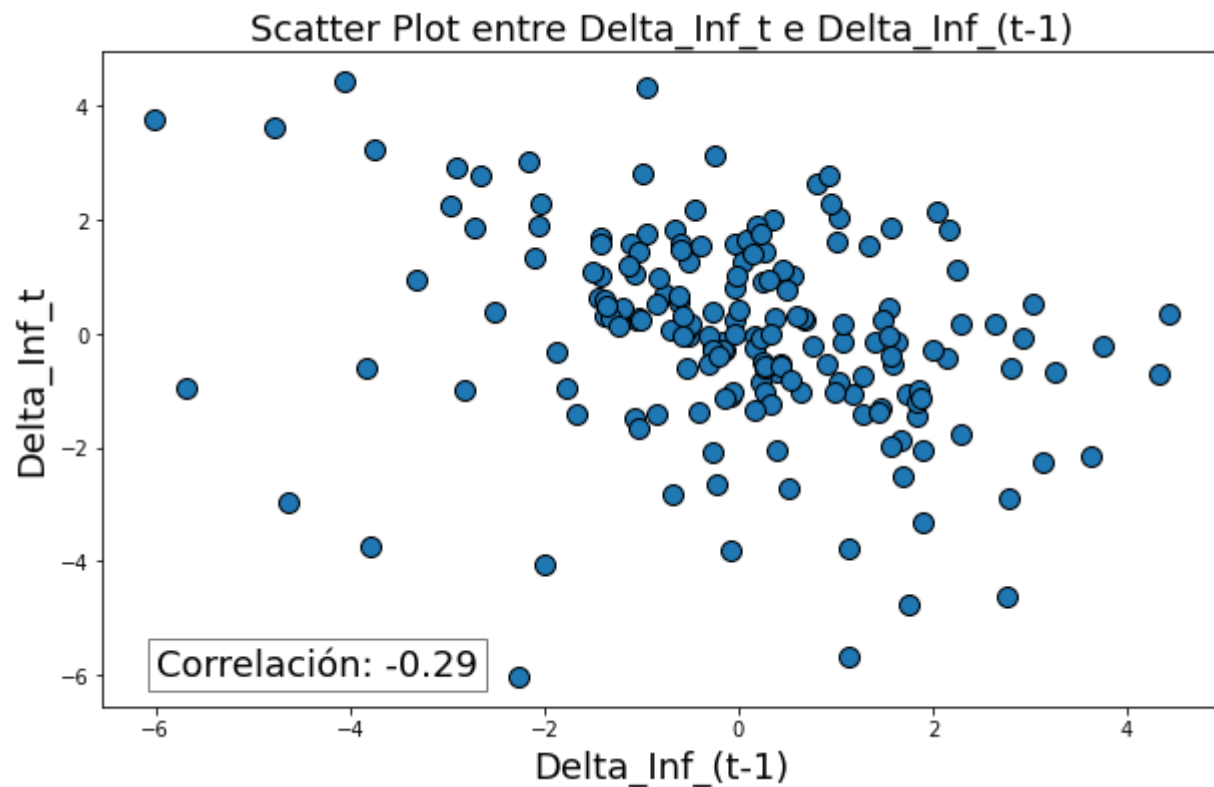
	freq	GDPC1	JAPAN_IP	PCECTPI	CPIAUCSL	Inf_T	Inf_T-1	Delta_Inf_T	Delta_Inf_T-1
0	1/1/1960	3123.162	12.184435	17.424	29.40	0.408580	2.466598	-2.058018	0.400482
1	4/1/1960	3111.310	12.676183	17.516	29.57	2.312925	0.408580	1.904345	-2.058018
2	7/1/1960	3119.057	13.222570	17.583	29.59	0.270544	2.312925	-2.042381	1.904345
3	10/1/1960	3081.300	13.850916	17.661	29.78	2.568435	0.270544	2.297891	-2.042381
4	1/1/1961	3102.251	14.615858	17.694	29.84	0.805910	2.568435	-1.762525	2.297891
...
175	10/1/2003	13528.710	102.001053	88.124	184.80	0.802472	2.971541	-2.169068	3.625956
176	1/1/2004	13606.509	103.271654	88.797	186.57	3.831169	0.802472	3.028696	-2.169068
177	4/1/2004	13706.247	104.789317	89.421	188.60	4.352254	3.831169	0.521085	3.028696
178	7/1/2004	13830.828	105.636384	89.942	189.37	1.633086	4.352254	-2.719168	0.521085
179	10/1/2004	13950.376	103.730482	90.652	191.03	3.506363	1.633086	1.873277	-2.719168

180 rows x 9 columns

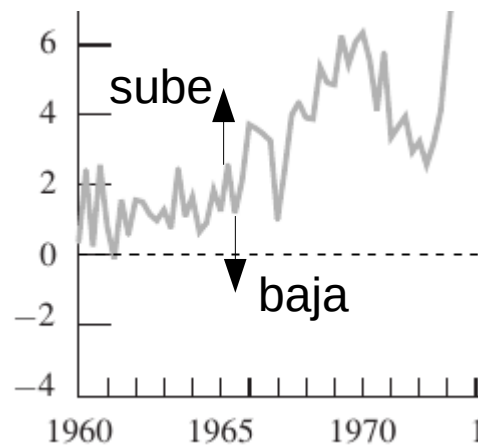
Notar que el shift realizado a la columna **Delta_Inf_t** puede generar valores NaN en las primeras componentes del intervalo de tiempo. **Chequear esta situación siempre**

Además debemos asegurar que el conjunto de datos para MCO sea consistente

Scatter plot entre Delta_Inf_t y Delta_Inf_t-1



La correlación negativa en la variación de la inflación tiende a cambiar de dirección en tiempos secuenciales



Modelo autorregresivo de primer orden o AR(1)

El modelo mas simple para intentar predecir el valor futuro de una variable conociendo sus valores pasados es considerar un modelo autorregresivo usando una regresión lineal simple con un retardo de una unidad

$$Y_{t,i} = \beta_0 + \beta_1 Y_{t-1,i} + u_i$$

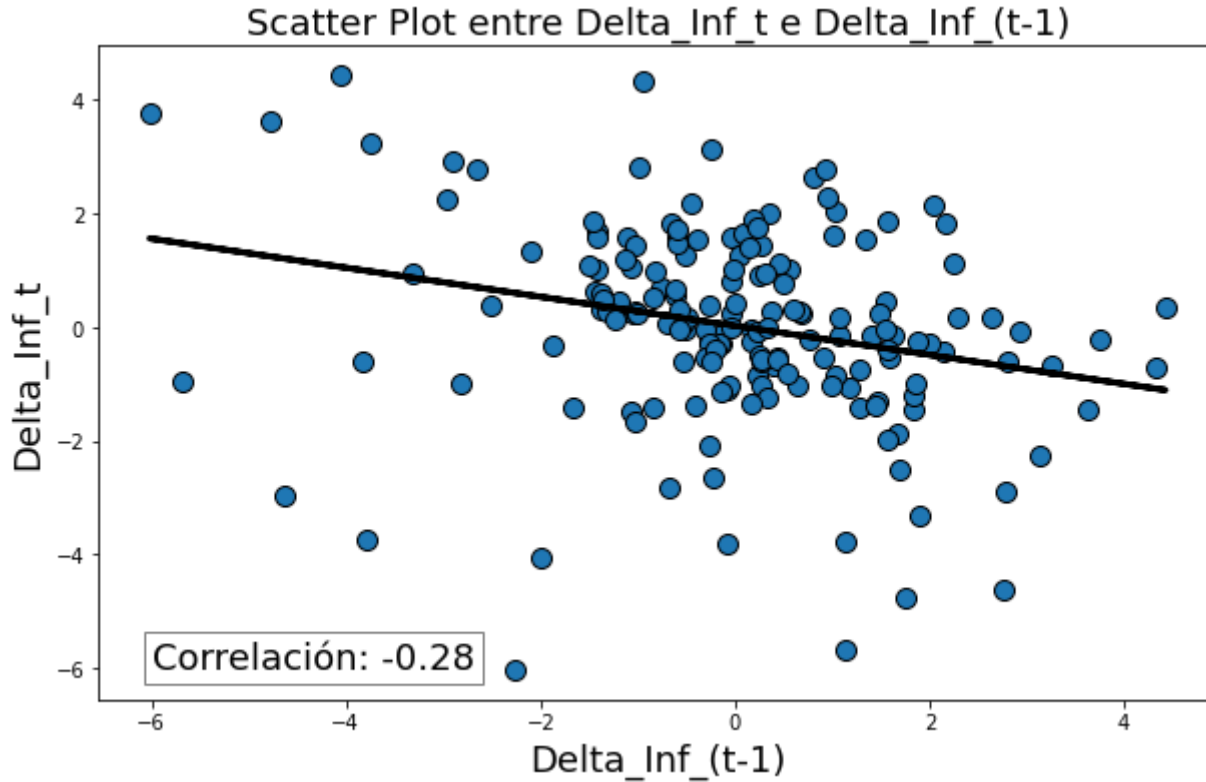
Por ejemplo, podemos considerar como variable de interés para modelar, la tasa de variación de la inflación, es decir nos interesa predecir el cambio en la inflación entre dos trimestres consecutivos, lo cual nos indicaría si los precios subirán o bajaran en el siguiente periodo

Para encontrar el valor de los coeficientes de la regresión lineal usamos nuevamente MCO

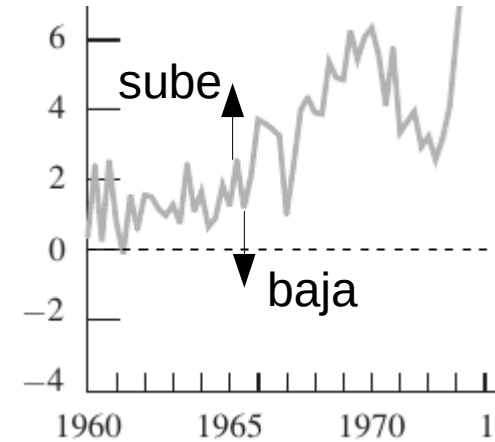
$$\widehat{\Delta \text{Inf}}_t = 0.02 - 0.25 \Delta \text{Inf}_{t-1}$$

Donde podemos observar los valores de la dispersión muestral de cada coeficiente calculados siguiendo las expresiones obtenidas en el capítulo de regresión lineal simple

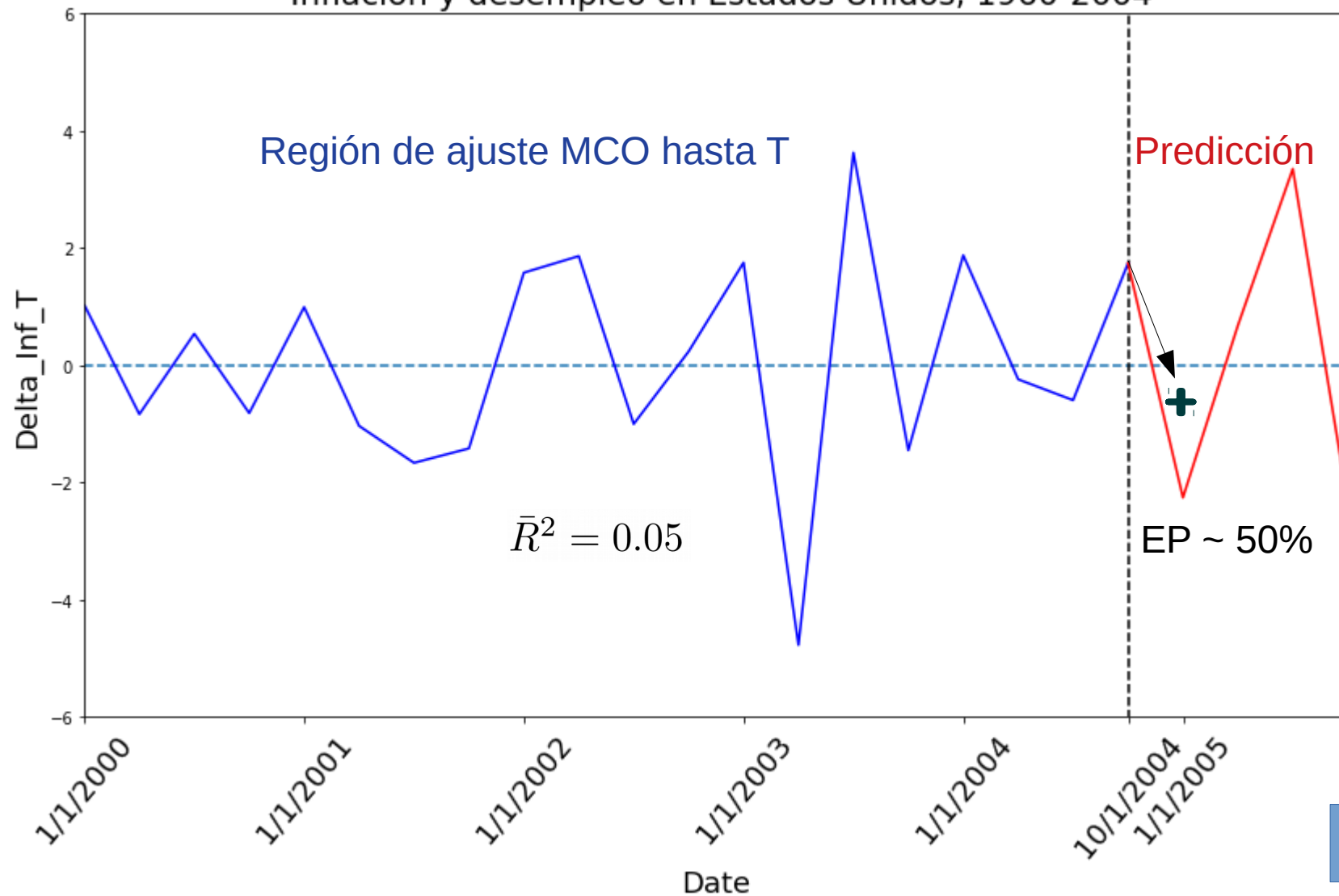
Regresión entre Delta_Inf_t y Delta_Inf_t-1



La correlación negativa en la variación de la inflación tiende a cambiar de dirección en tiempos secuenciales



Inflación y desempleo en Estados Unidos, 1960-2004



AR(1)

R² ajustado

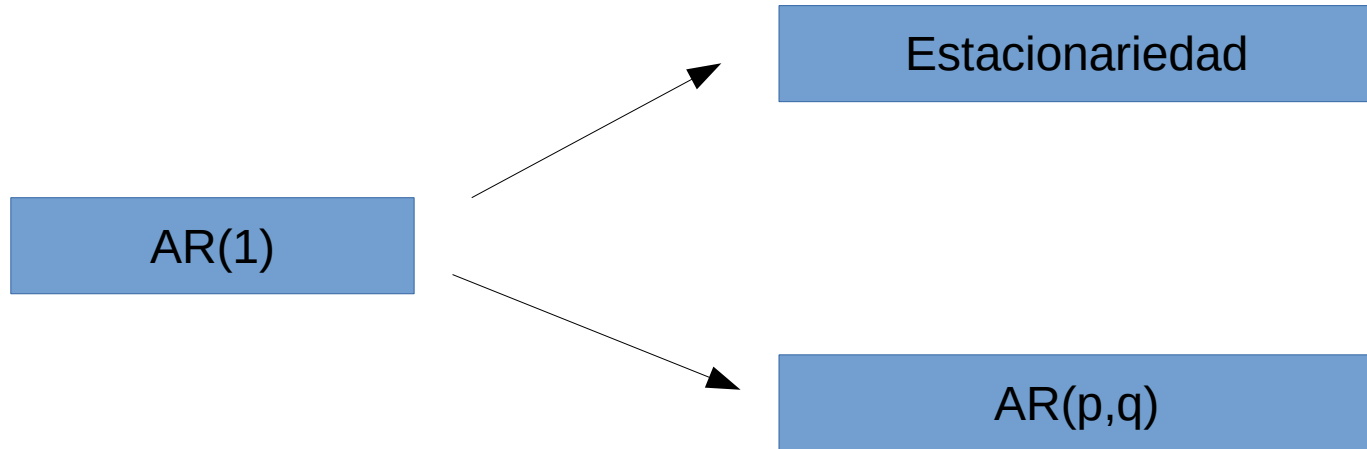
El estadístico definido como R² ajustado es una versión del estadístico R² que no aumenta necesariamente cuando se agrega una nueva variable independiente

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \frac{SR}{ST}$$

Donde n es el numero de puntos de MCO y **k+1 es el numero de coeficientes**. Para una regresión lineal simple k=1

Múltiples tipos de modelos pueden ser comparados usando R² ajustado

Esto aplica a los modelos de regresión lineal múltiple y por lo tanto a los modelos autorregresivos de múltiples retardos y variables independientes



Supuestos generales que permiten aplicación del método considerando datos que cumplen con ciertas condiciones

Reglas para extender el modelo básico utilizando RLM. Metodología para comparar modelos usando R^2 reducido

Estacionariedad y modelos de Autoregresión

Los modelos de autorregresión mas directos son literalmente modelos de regresión lineal. Por lo tanto se deben cumplir con las suposiciones de MCO y/o extensiones para garantizar la interpretación correcta de los resultados analíticos

La estacionariedad se plantea como una suposición sobre la distribución de probabilidad de los datos. Pero se puede verificar considerando los coeficientes de los modelos autorregresivos

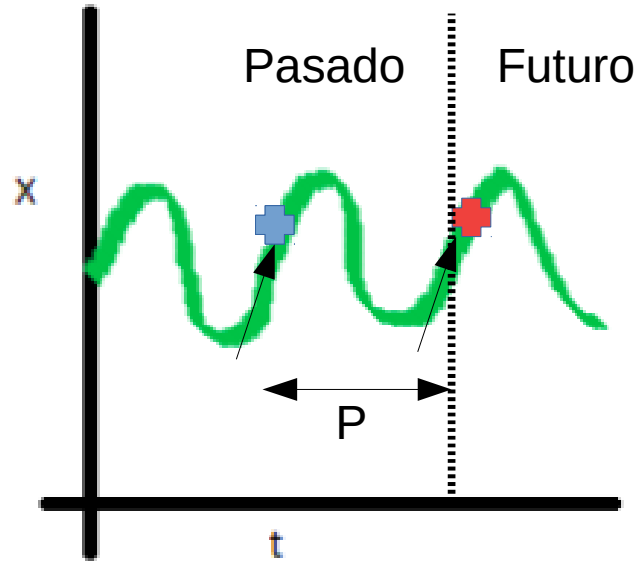
Estacionariedad de Y_t es la extensión de la suposición de i.i.d de MCO

Estacionariedad de Y_t requiere que los residuos sean estacionarios

Estacionariedad y modelos de Autoregresión

Considerando las características de estacionariedad es razonable asumir que un valor futuro de la serie temporal dependerá de los valores pasados

Serie temporal = Estacionariedad = Autorregresión y Fiabilidad Analítica



■ Evento conocido análogo al evento que mediremos

↗ Datos anteriores a la medición de un evento dado

■ Evento que queremos predecir

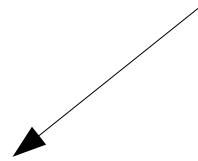
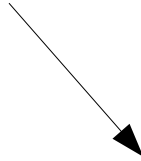
“Si los valores vienen subiendo con cierta pendiente entonces el siguiente valor debería ser mas alto”

$$x(t) = x(t - P)$$

Supuestos de MCO



Estacionariedad



Los datos observados representan una muestra de datos generado por fluctuaciones estadísticas alrededor de las predicciones de los modelos encontrados. Además permiten calcular intervalos de confianza y **test estadísticos** utilizando formulas simples y bien fundamentadas en demostraciones matemáticas en el contexto de estadística y probabilidades

De todas maneras, **siempre es posible aplicar las técnicas de modelamiento y predicción**, incluso cuando estos criterios no son cumplidos. De hecho hay algunos criterios que no son completamente obligatorios (homocedasticidad de los residuos).

En los casos donde los requerimientos no son cumplidos probablemente nos encontraremos con valores de R^2 muy bajos, así que en principio es factible realizar el análisis para mas tarde optar por una nueva opción dado los resultados obtenidos.

Supuestos de MCO: significancia de la tendencia observada

Bajo los siguientes supuestos es justificado plantear que los valores obtenidos de b_0 y b_1 a partir del método de mínimos cuadrados converjan a los valores reales de la regresión lineal que genera los datos observados

1- La media de los errores e_i condicionado sobre X_i es igual a cero. En la práctica esto quiere decir que se espera que los errores distribuyan normalmente en torno al valor Y predicho por la regresión lineal. X y e son incorrelacionados. Presencia de factores externos. **Este supuesto implica que los estimadores son insesgados**

2- Los valores de X_i e Y_i son independientes e idénticamente distribuidos (i.i.d). En la práctica esto significa que la muestra de datos es una representación aleatoria simple de los datos. Es una condición sobre la metodología de toma de datos. **Este supuesto se aplica para obtener la varianza muestral de los estimadores**

3- Los datos atípicos son improbables. Esto quiere decir que la distribución de probabilidad de las variables X_i e Y_i tiene una curtosis finita (decrece rápidamente para valores alejados de la media). **Aproximación utilizando durante la obtención de la distribución de los estimadores de la regresión**

Estacionariedad

La definición matemática del concepto de estacionariedad requiere que exista independencia en el tiempo de la distribución de probabilidad de la variable relevante. Es decir, existe al menos una propiedad que se mantiene inmutable en el tiempo

CONCEPTO CLAVE

14.5

Estacionariedad

Una serie temporal Y_t es *estacionaria* si su distribución de probabilidad no varía en el tiempo, es decir, si la distribución conjunta de $(Y_{s+1}, Y_{s+2}, \dots, Y_{s+T})$ no depende de s sea cual sea el valor de T ; de lo contrario, se dice que Y_t es *no estacionaria*. Dos series temporales, X_t e Y_t , se dice que son *conjuntamente estacionarias* si la distribución conjunta de $(X_{s+1}, Y_{s+1}, X_{s+2}, Y_{s+2}, \dots, X_{s+T}, Y_{s+T})$ no depende de s , independientemente del valor de T . La estacionariedad requiere que el futuro sea como el pasado, al menos en un sentido probabilístico.

Distribución conjunta = Distribución de probabilidad de múltiples variables aleatorias

Caso de ejemplo = Y_t distribuye de la misma forma para todo t

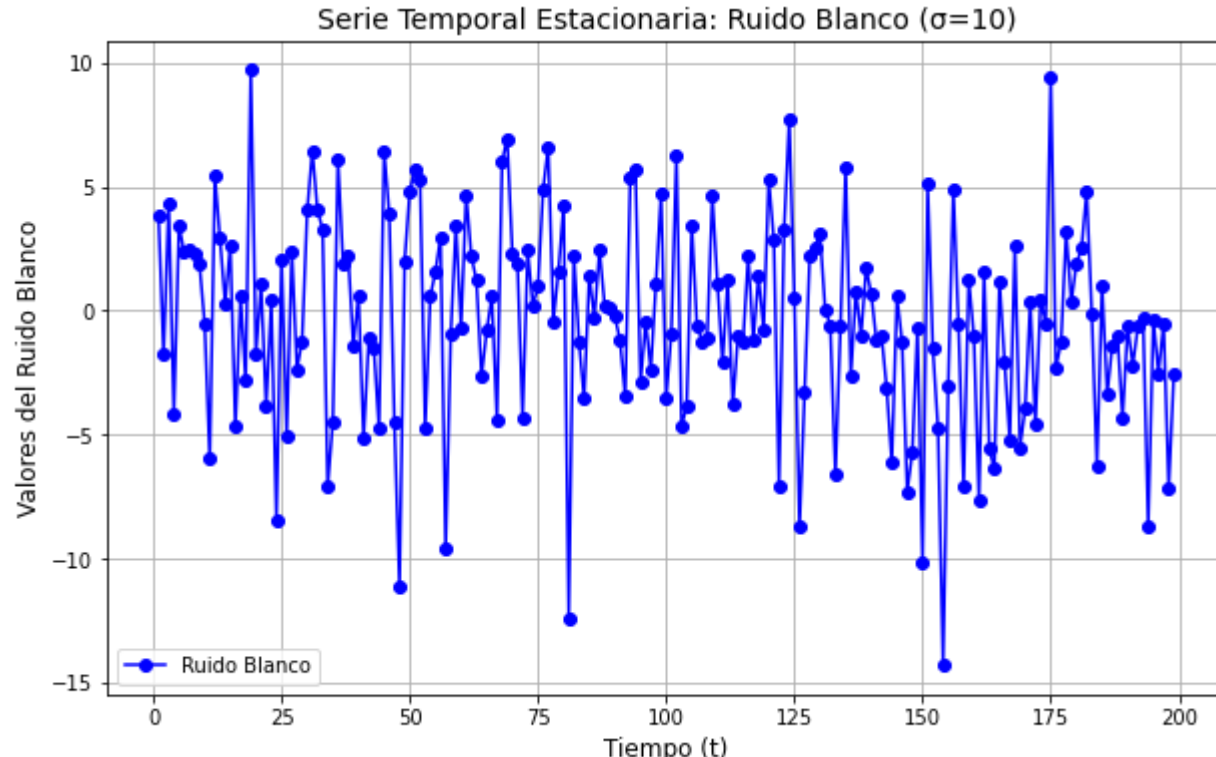
Estacionariedad: Ejemplo

Para clarificar ideas vamos a simular un proceso estacionario básico. Para esto vamos a considerar una variable aleatoria $Y_t = e_t$ con $e_t \sim N(0, \sigma^2)$ independiente del valor de t

En este caso $E(X_t) = E(e_t) = 0$ para todo t y de forma similar la desviación estándar también es independiente de t .

Este ejemplo también puede ser visualizado, considerando T puntos consecutivos de la variable t y generamos puntos con el valor que adquiriera e_t en ese instante dado. Claramente los puntos varían entre puntos consecutivos de t

Un estimador de $E(X_t)$ está dado por el promedio en un intervalo T , el cual nos permite evaluar estacionariedad



Tendencia: movimiento persistente a largo plazo de una variable en el tiempo

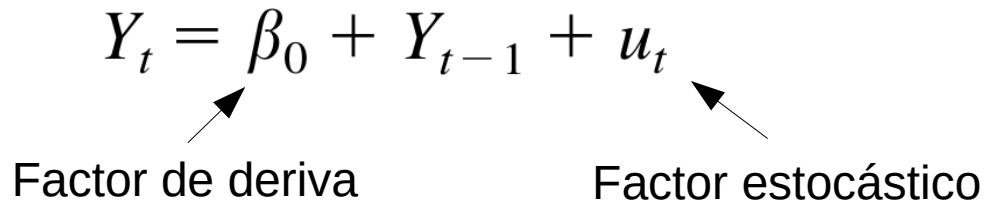
Tendencia determinista: movimiento persistente constante para todo tiempo

Tendencia estocástica: movimiento persistente oscilatorio en el tiempo, con componentes que se pueden interpretar como tendencias deterministas en plazos acotados (**e.g. IPC**). Es estocástica porque depende de factores aleatorios, como desastres naturales o malas decisiones monetaria

Dado que el segundo tipo contiene al primero, en general **es mas eficiente suponer** que los datos de series temporales pueden llegar a contener tendencias estocásticas. Esta discusión es análoga a la suposición de homocedasticidad y heterocedasticidad

Modelo de tendencia estocástica: paseo aleatorio con deriva

$$Y_t = \beta_0 + Y_{t-1} + u_t$$


Factor de deriva Factor estocástico

Este modelo **NO ES ESTACIONARIO**

$$\longrightarrow Y_t = \beta_0 + Y_{t-1} + u_t$$

El modelo de tendencia estocástica con deriva no es estacionario. Lo cual se puede verificar notando que la varianza de Y depende del tiempo

$$\text{var}(Y_t) = \text{var}(Y_{t-1}) + \text{var}(u_t)$$

Por otro lado podemos notar que el modelo de paseo aleatorio con deriva es un caso particular de la regresión lineal para $b_1=1$. Es notable verificar que en el caso mas general dado por $|b_1| < 1$ genera un modelo que SI ES ESTACIONARIO

Este modelo **SI ES ESTACIONARIO**

$$\longrightarrow Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$$

De hecho, para que el modelo sea estacionario es necesario requerir que u_t sea estacionario. Este es la suposición que respalda el proceso MCO, el cual se plantea en términos de la distribución de los residuos

Tratamiento de series no estacionarias

A partir de los datos potencialmente estacionarios se plantea un modelo $AR(p,q)$. Este modelo sera estacionario si cumple con la condición de **raíz cuadrada unitaria**. En caso contrario el conjunto de datos es no estacionario

Cuando se cumple el requerimiento de estacionariedad el modelo obtenido por la regresión lineal múltiple con residuos normales y MCO es el mas eficaz en cuanto a la generación de valores de los coeficientes con respecto al tamaño de la muestra

Por lo tanto corresponde realizar los procedimientos necesarios para intentar llevar la serie temporal a un comportamiento estacionario mediante transformaciones de variables

Detección de no
estacionariedad

$$\longrightarrow |\beta_p| = 1$$

La transformación de variables usando una transubstancia de tiempos consecutivos puede pasar de no estacionario a estacionario

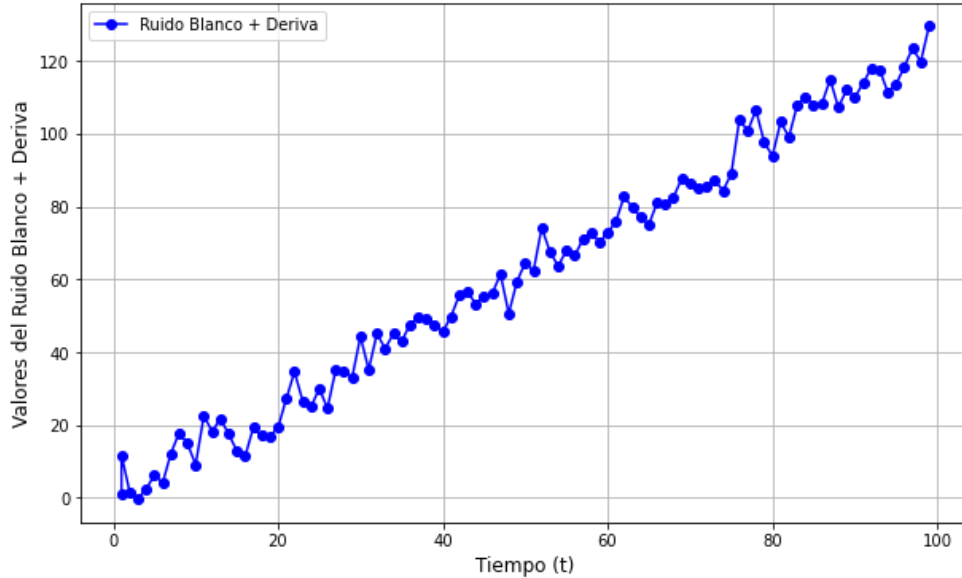
Detección de
estacionariedad

$$\longrightarrow |\beta_p| < 1$$

Recuperación de estacionariedad

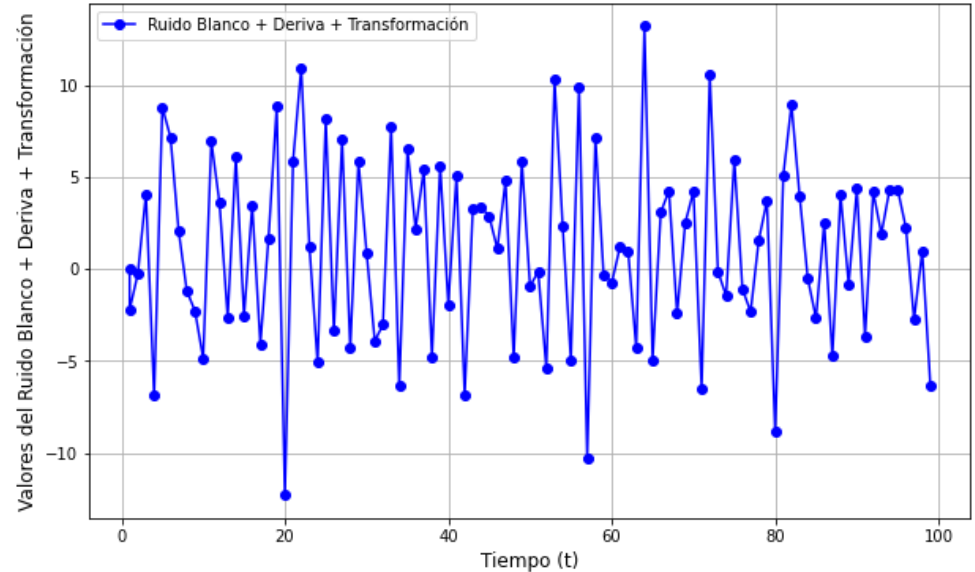
$$Y_t = \alpha t + \beta_1 Y_{t-1} + u_t \longrightarrow \begin{aligned} \Delta Y_t &= Y_t - Y_{t-1} \\ \Delta Y_t &= \alpha \Delta t + \beta_1 \Delta Y_{t-1} + \delta u_t \end{aligned}$$

Serie Temporal Estacionaria: Ruido Blanco ($\sigma=4$) + Deriva



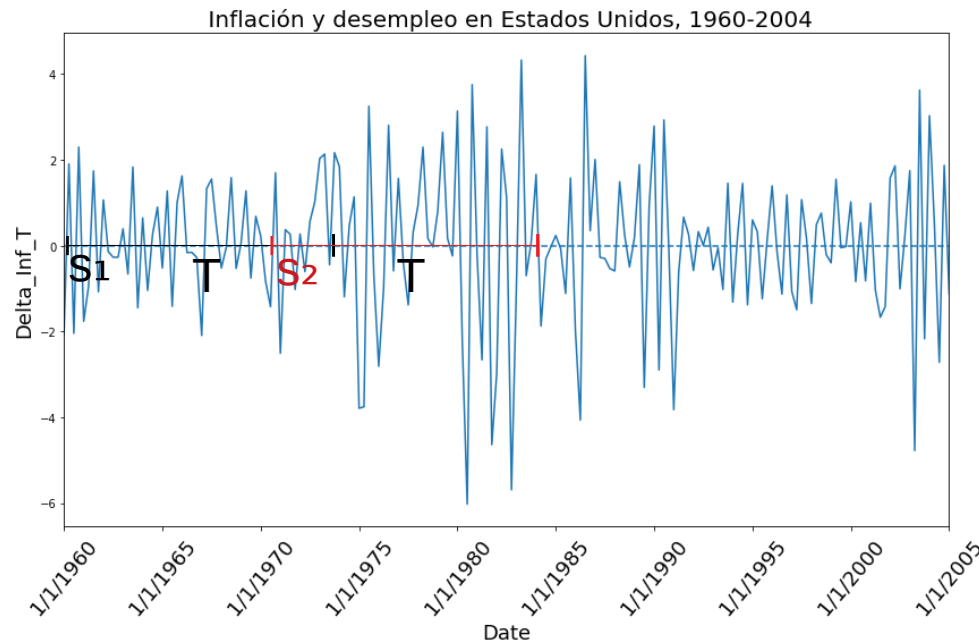
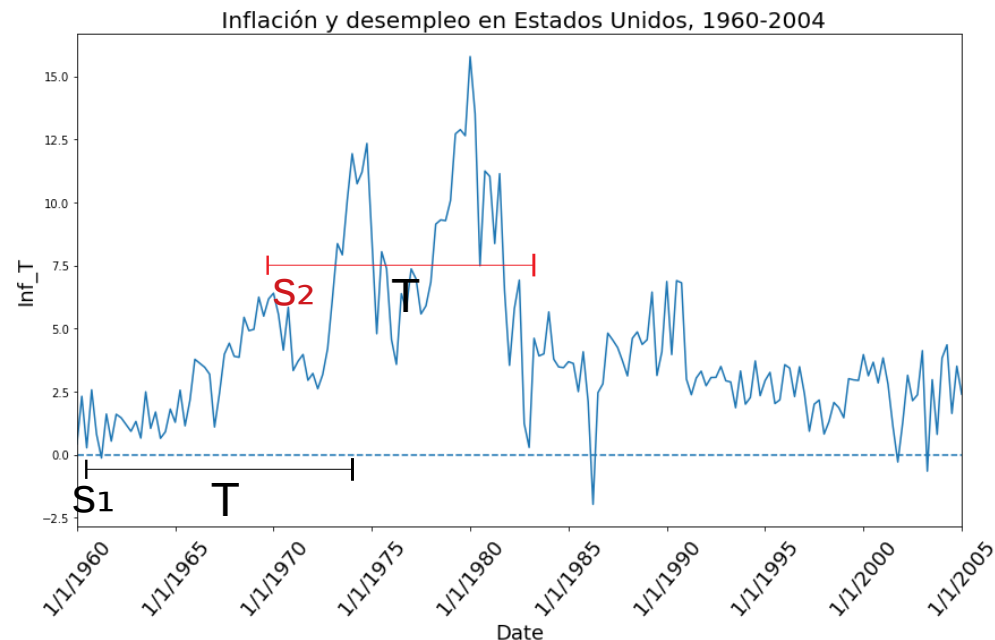
No estacionaria con tendencia

Serie Temporal Estacionaria: Ruido Blanco ($\sigma=4$) + Deriva + Transformación



Estacionario

Discusión de estacionariedad en el caso de estudio



$$\Delta \text{Inf}_t = \text{Inf}_t - \text{Inf}_{t-1}$$

La **operación de substracción** nos permite recuperar una serie temporal que muestra un **comportamiento más estacionario que la serie inicial (visualizar medias móviles)**

Modelo autorregresivo de orden (p,q): AR(p,q)

Modelo autorregresivo de orden p: AR(p)

La extensión desde un tiempo de retardo a p tiempos de retardo sigue la misma lógica de pasar de una regresión lineal a una regresión múltiple

$$\hat{Y}_t = \beta_0 + \beta_1 Y_{t-1} + \dots + \beta_p Y_{t-p}$$

Dado que vamos a considerar p retardos debemos asegurarnos que el conjunto de datos contenga estas nuevas columnas, tal que para cada tiempo podamos acceder a sus valores de Y hasta p tiempos atrás

Luego de esto debemos aplicar las técnicas de MCO para modelos de regresión lineal múltiple

Ejemplo: vamos a considerar el modelo para Delta_Inf_t con 4 retardos

$$\widehat{\Delta\text{Inf}}_t = \beta_0 + \beta_1 \Delta\text{Inf}_{t-1} + \beta_2 \Delta\text{Inf}_{t-2} + \beta_3 \Delta\text{Inf}_{t-3} + \beta_4 \Delta\text{Inf}_{t-4}$$

Mas allá de las posibles motivaciones para incluir este número de retardos nos fijaremos en los valores de R^2 ajustado

Modelo autorregresivo de orden p: AR(p)

In [166]: `df_mco_ar4.reset_index(drop=True)`

Out[166]:

	freq	GDPC1	JAPAN_IP	PCECTPI	CPIAUCSL	Inf_T	Inf_T-1	Delta_Inf_T	Delta_Inf_T-1	Delta_Inf_T-2	Delta_Inf_T-3	Delta_Inf_T-4
0	1/1/1962	3336.753	16.965323	17.856	30.11	1.600534	0.534224	1.066310	-1.074895	1.743167	-0.939958	-1.762525
1	4/1/1962	3372.706	16.965323	17.918	30.22	1.461309	1.600534	-0.139225	1.066310	-1.074895	1.743167	-0.939958
2	7/1/1962	3404.833	16.692129	17.965	30.31	1.191264	1.461309	-0.270044	-0.139225	1.066310	-1.074895	1.743167
3	10/1/1962	3418.046	16.582852	18.018	30.38	0.923788	1.191264	-0.267477	-0.270044	-0.139225	1.066310	-1.074895
4	1/1/1963	3456.080	17.238516	18.069	30.48	1.316656	0.923788	0.392868	-0.267477	-0.270044	-0.139225	1.066310

AR(1)

$$\widehat{\Delta \text{Inf}}_t = 0.02 - 0.25 \Delta \text{Inf}_{t-1}$$

AR(4)

$$\widehat{\Delta \text{Inf}}_t = 0.03 - 0.29 \Delta \text{Inf}_{t-1} - 0.29 \Delta \text{Inf}_{t-2} + 0.15 \Delta \text{Inf}_{t-3} - 0.01 \Delta \text{Inf}_{t-4}$$

Modelo	R ² ajustado	EP (Inf)
AR(1)	0.05	48%
AR(4)	0.19	48%

El ajuste AR(4) mejora el ajuste en términos de MCO pero el error en la predicción de la inflación es similar a AR(1)... Más variables

Modelo autorregresivo de retardos distribuidos: AR(p,q)

Otra extensión a considerar en el contexto de los modelos autorregresivos consiste en incorporar otras variables explicativas con su propio número de retardos

$$\hat{Y}_t = \beta_0 + \beta_1 Y_{t-1} + \dots + \beta_p Y_{t-p} \\ + \delta_1 X_{t-1} + \dots + \delta_q X_{t-q}$$

Una variable extra que se podría incluir en la predicción de la variación de la tasa de inflación corresponde a la tasa de desempleo en el periodo anterior

Concepto profundo y sobre los detalles de este tipo de modelamiento: La hipótesis que menciona que la media de los residuos condicionados sobre las variables libres para un modelo AR(p,q) es cero implica que los retardos p y q son los verdaderos. Por lo tanto si intentáramos incluir otras variables distribuidas o más retardos los coeficientes correspondientes serían nulos

Test de hipótesis usando test-t (AB) y test-F (ANOVA)

En general podemos notar que los coeficientes de los modelos AR(p,q) se comportan como estadísticos que dependen de la muestra considerada para su evaluación

Cuando los supuestos de MCO y extensiones, incluyendo estacionariedad, se cumplen entonces es posible aproximar las distribuciones muestrales de los coeficientes usando las mismas técnicas que las aplicadas a las diferencias de promedios

En particular, para AR(1) el test a utilizar es el test-t (test-AB) mientras que en el caso mas genérico, es decir AR(p,q) es test corresponde al test-F (test ANOVA).

En cualquier caso, la hipótesis nula queda definida por el escenario donde todos los coeficientes de la regresión son iguales a cero. Es decir, si el test de hipótesis genera un p-value menor a 0.5 entonces podemos descartar la hipótesis nula.

AR(1)

$$\longrightarrow t_{\text{obs}} = \frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_1}}$$

AR(p,q)

$$\longrightarrow F_{\text{obs}} = \frac{(ST-SR)/(k-1)}{SR/(n-k)}$$

Las significancias relativas también pueden ser relevantes para comparar modelos

Fronteras del modelamiento de los datos en la actualidad: AI



Mucha suerte y animo en su camino de aprendizaje!!

Coordinación de evaluaciones y proyectos

Nota promedio de proyectos (T1, T2 y T3) vale un 50% de la nota final

Nota promedio de los mejores 3 ejercicios online vale un 50%.