Tema 6: Modelización con datos de series temporales

Universidad Complutense de Madrid

2013

Introducción (I)

Una **característica** que distingue los datos de series temporales de los datos de sección cruzada, es que **los datos temporales están ordenados** de una forma natural **cronológicamente** (primero va enero de un año, después febrero de ese año, etc.)

Este hecho es muy importante, ya que **determina qué tipo de relaciones son posibles entre un tipo de datos y otros**. Así:

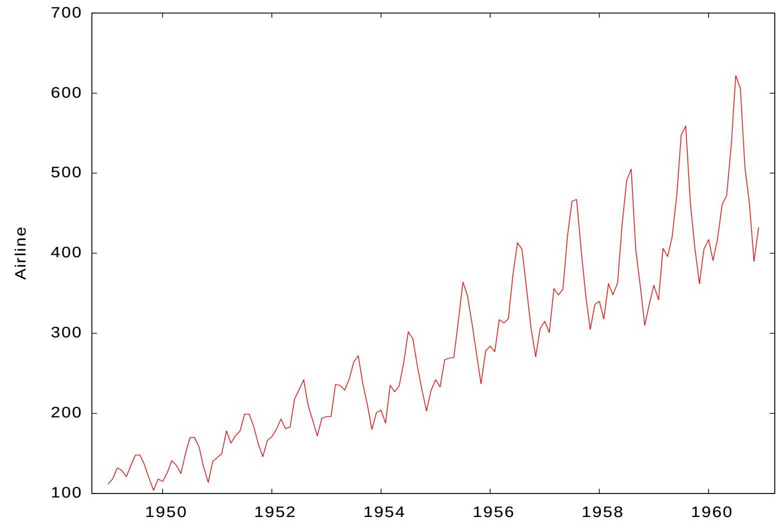
(1) El orden de las observaciones en una sección cruzada es irrelevante. Por ejemplo, en una sección de datos de salarios, primero podemos tener al individuo que más gana hasta el que menos gana o a la inversa. Ningún resultado de la estimación cambiaría al cambiar el orden de los datos. En cambio, en una serie temporal el orden es cronológico y único. El IPI trimestral de un país irá ordenado en el tiempo, por ejemplo, tendremos una muestra desde el primer trimestre de 1990 hasta el segundo de 2010.

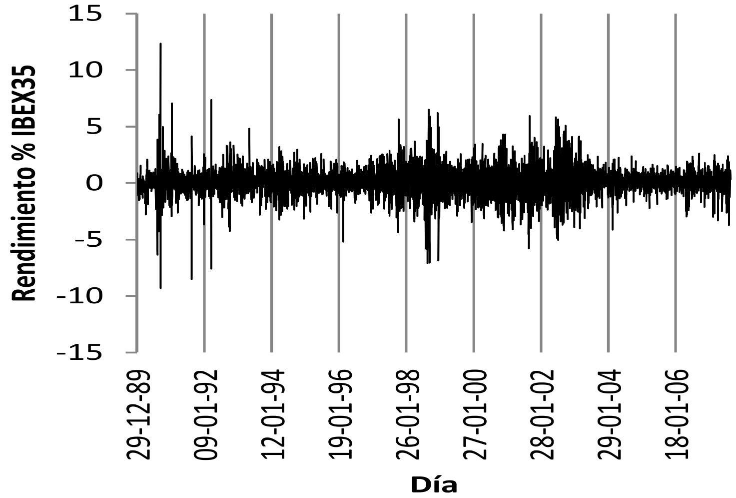
Introducción (II)

(2) En una serie temporal es más razonable suponer que existe correlación serial. Por ejemplo, el IPI de un trimestre de un año puede estar correlacionado con el valor de IPI del trimestre anterior; las ventas de una semana de una empresa tendrán relación con las ventas de la(s) semana(s) anterior(es). Es más difícil que el peso de un niño tenga correlación con el peso de otros niños de la muestra.

(3) Con datos temporales, existe un fundamento empírico a lo que entendemos por causalidad. Es decir, una variable ***X*** causa a otra variable ***Y***, si los valores pasados de la ***X*** están correlacionados con los valores presentes de la ***Y***. Con datos de sección cruzada, la causalidad entre variables puede existir, pero no tenemos una manera empírica de detectar la dirección de la causalidad (qué variable causa a qué otra variable).

Características comunes de las series temporales





En una serie temporal de baja frecuencia (es decir, con datos mensuales, trimestrales, etc) las características más habituales son: (a) una **tendencia** (en el caso del nº de pasajeros a crecer); (b) **estacionalidad** (en el caso de los pasajeros, vuelan más personas siempre en vacaciones de verano) y (c) una **varianza** (dispersión alrededor de la media) **que crece con la media**.

En una serie temporal de alta frecuencia (es decir, con datos diarios, horarios, etc.) se encuentra: (a) una **media estable** a lo largo del tiempo; (b) **no hay estacionalidad** y (c) **una varianza que cambia con el tiempo**, de modo que se alternan períodos de alta volatilidad (alta varianza) con períodos de baja volatilidad (baja varianza). La varianza cambia de forma no sistemática.

Objetivos del análisis de series temporales (I)

En este tema, nos centramos en datos temporales de baja frecuencia (series anuales, trimestrales o mensuales). Las variables medidas en alta frecuencia suelen ser financieras y su modelización es más complicada.

El objetivo es modelizar las características más habituales que hemos visto. Es decir:

- **Capturar la tendencia y el comportamiento estacional** observado
- **Tratar la varianza no constante** (heterocedasticidad)
- **Modelizar la autocorrelación serial.** Es decir, encontrar un **modelo estadístico que sea capaz de reproducir esa "inercia"** o autocorrelación que tienen muchas variables económicas temporales.

Objetivos del análisis de series temporales (II)

Este objetivo se puede conseguir usando **distintos enfoques:**

(1) Usar un modelo univariante: Es decir, intentamos **explicar la correlación de una variable temporal usando para ello sólo su propia historia pasada y reciente.** No incluimos variables explicativas adicionales. Puede parecer una restricción, pero si el objetivo es predecir a corto plazo el futuro de la variable, estos modelos funcionan muy bien o mejor que otras especificaciones alternativas.

(2) Usar un modelo de relación (en este curso, un modelo de regresión). En este caso, se pueden **usar ideas del análisis univariante para que el modelo de regresión esté bien construido.** Si el objetivo es predecir a medio y largo plazo, es evidente que hay que tener en cuenta la correlación contemporánea y dinámica de unas variables sobre otra (que es la variable de interés, la que hay que predecir).

Modelos deterministas de series temporales (I)

Dada **una serie temporal** denotada por y_t, **se supone que puede descomponerse de modo aditivo como:**

$$y_t = t_t + s_t + c_t + \varepsilon_t$$

donde t_t es **la tendencia**, s_t es el **componente estacional**, c_t es el **componente cíclico** (ó ciclo) y ε_t el **error**.

Es difícil definir cada una de estas componentes. No hay consenso en la literatura sobre el tema.

La **tendencia** debe recoger **el movimiento a largo plazo de una serie**, independientemente de otros componentes irregulares. Es decir, debe recoger el nivel subyacente y regular de la serie.

Modelos deterministas de series temporales (II)

El **componente estacional** debe recoger las oscilaciones que se producen con un período inferior o igual al año. Es decir, son **oscilaciones a corto plazo que se repiten en años sucesivos**. Las razones por las que una serie presenta estacionalidad pueden ser de tipo físico (el clima, etc.) o de tipo institucional (vacaciones, festividades varias, etc.).

El **error** debe recoger **movimientos transitorios e irregulares de la serie**. Esta componente puede descomponerse en una parte claramente aleatoria e imprevisible y en otra parte no siempre previsible, pero que se puede identificar *a posteriori* (como una huelga, una catástrofe natural, un cambio político, etc.)

Modelos deterministas de series temporales (III)

En la práctica siempre se supone que:

$$E[\varepsilon_t] = 0, \forall t \qquad E[\varepsilon_t^2] = \sigma^2, \forall t$$

$$E[\varepsilon_t \varepsilon_s] = 0, \forall t \neq s \qquad \varepsilon \sim N[0, \sigma^2 I]$$

El **ciclo** se define de diversas formas. Desde el punto de vista macroeconómico deben ser oscilaciones en torno a la tendencia que se deben a la alternancia entre períodos de crisis y de prosperidad. **Desde el punto de vista estadístico, el ciclo incluye cualquier característica que no sea tendencia, estacionalidad y ruido.**

Si se han tomado logaritmos a la serie, la descomposición de la variable original será de tipo multiplicativo, es decir:

$$y_t = e^{t_t} \cdot e^{s_t} \cdot e^{c_t} \cdot e^{\varepsilon_t}$$

Modelos deterministas de series temporales (IV)

Los **modelos más simples para la tendencia** son regresiones de la variable con respecto al tiempo. Un **modelo de tendencia lineal** sería (cuando la misma crece o decrece):

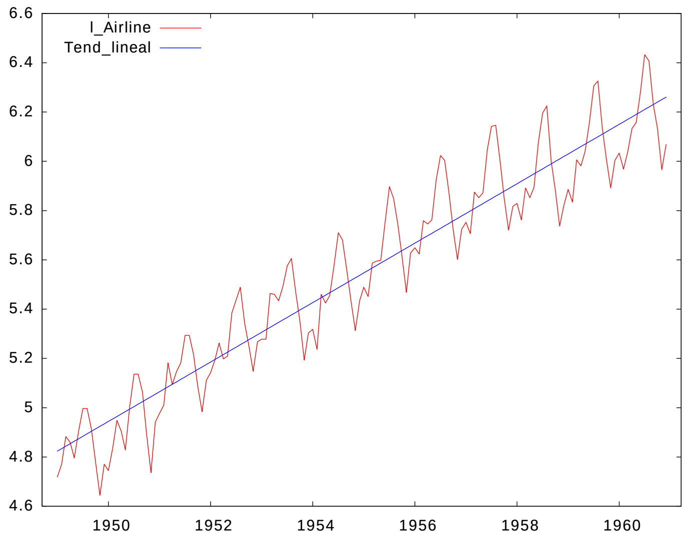
$$t_t = \alpha_0 + \alpha_1 t, \forall t = 1, 2, \ldots, n$$

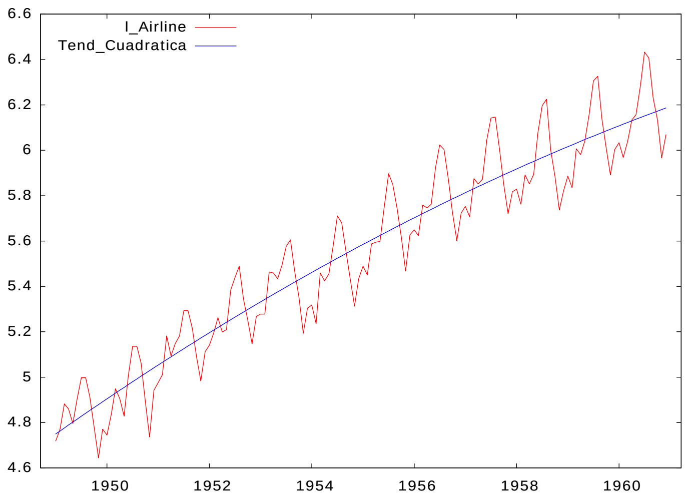
y un **modelo de tendencia cuadrático:**

$$t_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2, \forall t = 1, 2, \ldots, n$$

Los gráficos de la derecha muestran el resultado de ajustar una tendencia lineal (arriba) a la serie del nº de pasajeros (en logs) y una tendencia cuadrática (abajo) a la misma serie.

Estos modelos explican la evolución pasada de una serie en función de pautas simples, **pero tienen problemas** y limitaciones.





Modelos deterministas de series temporales (V)

Los **modelos de tendencia determinista** son una extensión inmediata de los métodos de regresión. Aunque son útiles para describir las pautas que sigue una serie temporal, **las predicciones que proporcionan suelen ser muy malas** (es decir, con un gran error asociado).

La **razón** de esto es que **en una serie temporal la observación más reciente depende, en general, de sus valores pasados, pero esta dependencia suele ser más fuerte con los datos más recientes y más débil con los más alejados.** Los modelos de tendencias deterministas proporcionan predicciones que no utilizan esta propiedad.

Si la serie no tiene estacionalidad, el modelo es: $y_t = \alpha_0 + \alpha_1 t + \varepsilon_t$ donde *t* es el tiempo. El parámetro α_1 representa la pendiente de la recta que describe la evolución de la serie (crecimiento esperado entre dos períodos). La previsión de la variable en el período *T+k* es: $\hat{y}_T(k) = \hat{\alpha}_0 + \hat{\alpha}_1 (T + k)$

Modelos deterministas de series temporales (VI)

Si se observara **una tendencia exponencial**, del tipo $t_t = \exp[\alpha_0 + \alpha_1 t]$, tendríamos tomando logaritmos:

$$\ln y_t = \alpha_0 + \alpha_1 t + \varepsilon_t$$

donde los parámetros α_0 y α_1 se estiman por MCO y *t* es el tiempo.

Estacionalidad: es un **cambio en la media de la serie que se repite periódicamente cada s estaciones.** Si la serie es mensual, s=12; si es trimestral, s=4; si es semanal, s=52 ó 53, etc.

Una forma determinista de captar la estacionalidad consiste en definir variables dummies (con valores 0 ó 1). Por ejemplo, para una serie mensual definir las doce dummies correspondientes a Enero (S1), Febrero (S2), …., hasta Diciembre (S12). Se puede escribir como:

$$s_t = \beta_1 S1_t + \beta_2 S2_t + \ldots + \beta_{12} S12_t$$

Modelos deterministas de series temporales (VII)

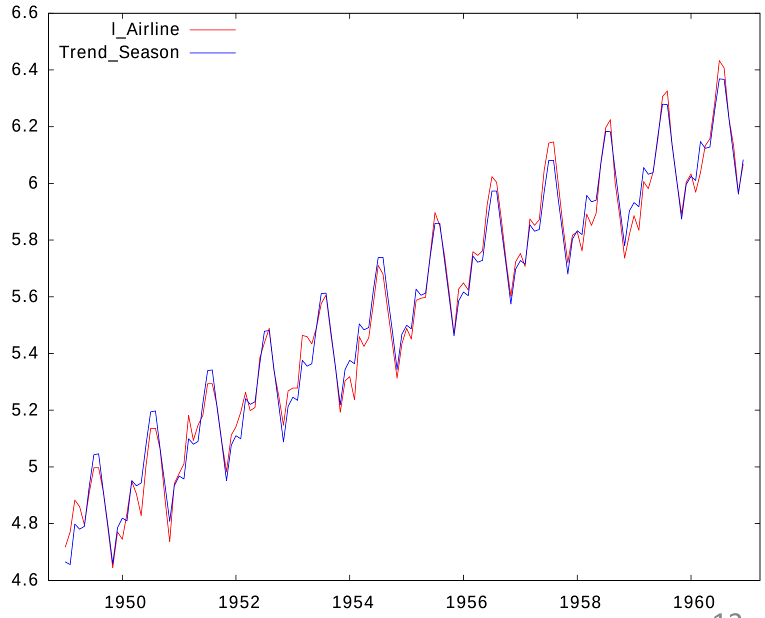
Si a la **serie mensual del nº de pasajeros (en logaritmos)** le ajustamos un modelo de tendencia cuadrática y una estacionalidad determinista mensual, se escribiría como:

$$\ln(Airline_t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \beta_1 S1_t + \beta_2 S2_t + \ldots + \beta_{11} S11_t + \varepsilon_t$$

La figura de la derecha muestra el resultado del ajuste del modelo de tendencia + estacionalidad anterior a los datos de pasajeros.

En **color azul** se muestra la **evolución de la serie real (en logaritmos)**

En **color rojo** la **evolución de la serie ajustada** al estimar por MCO el modelo

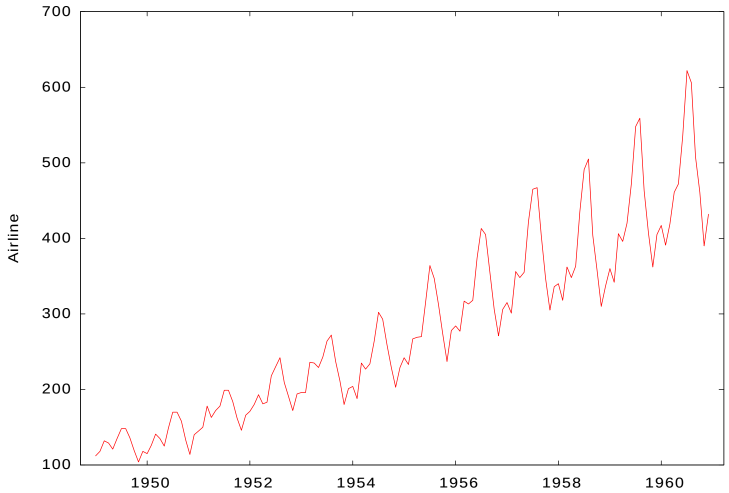


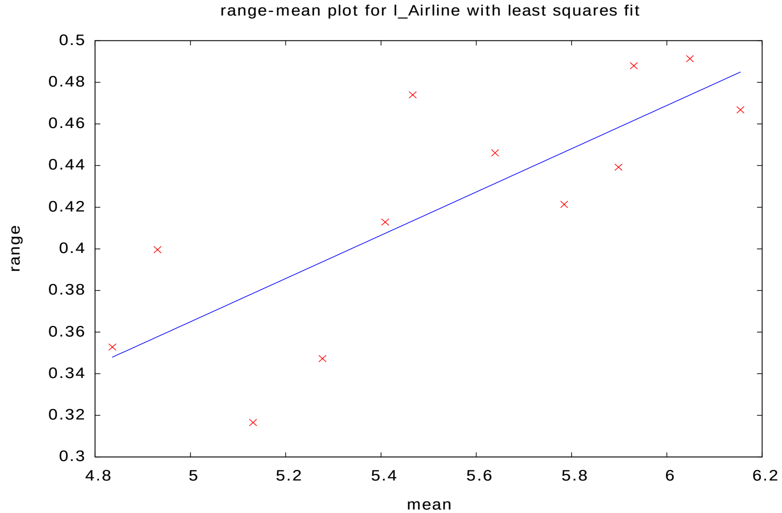
Otra aproximación (I)

Otra forma de capturar los cambios en la media (tendencia creciente, decreciente, etc) y los cambios en la varianza (crece la dispersión conforme crece la media) en una serie temporal, consiste en realizar una serie de transformaciones en los datos que eliminan estas características típicas.

Por ejemplo, si la varianza crece a medida que crece la media (ver la figura de la izquierda) ó bien, el **gráfico rango-media** de la derecha (se dibujan pares de valores de la media local y la desviación típica local, calculados para submuestras de igual tamaño de la serie).

La **transformación logarítmica** hace que la dispersión sea más o menos constante a medida que crece la media.

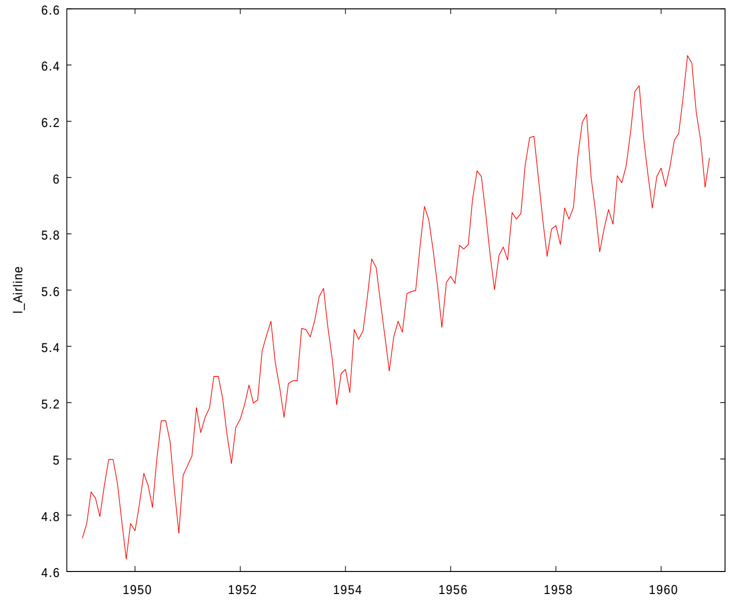




Otra aproximación (II)

Si tomamos logaritmos a la serie del nº de pasajeros (ver Figura de abajo) comprobamos que:

- La dispersión de la serie es más o menos constante a medida que crece la media
- La transformación logarítmica no consigue que la media de la serie sea constante (se sigue apreciando una tendencia creciente).



Si una serie temporal tiene una media constante a lo largo del tiempo, decimos que es estacionaria con respecto a la media. Si tiene varianza constante con respecto al tiempo, decimos que es estacionaria en varianza. Si una serie temporal es estacionaria (en media y en varianza) encontrar un modelo que explique su autocorrelación es mucho más fácil.

Otra aproximación (III)

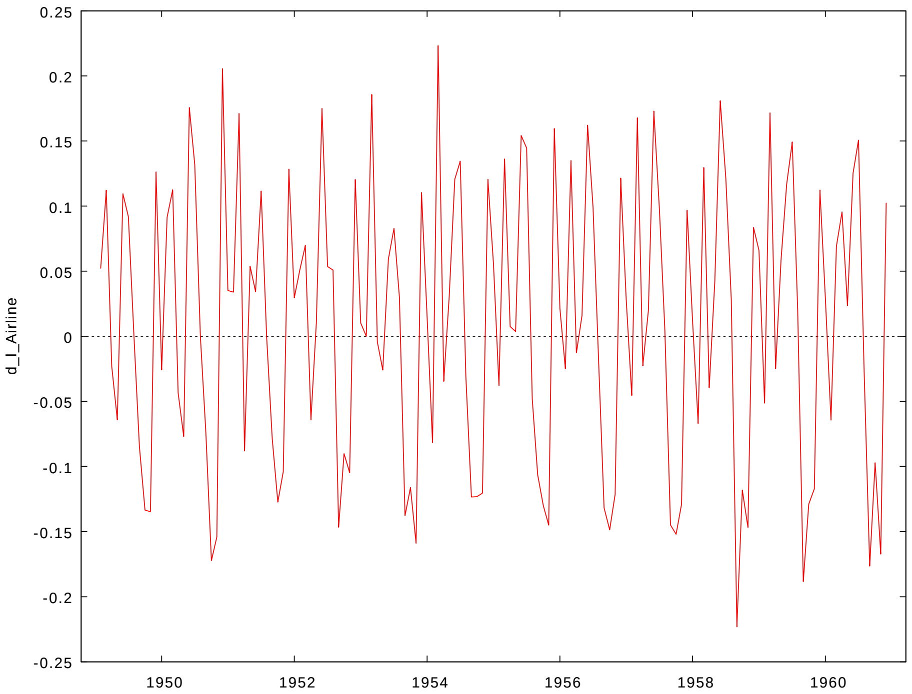
La transformación que elimina la tendencia (o lo que es lo mismo, induce estacionariedad en media) **es la diferenciación**. Tomar una diferencia regular consiste en calcular la diferencia entre cada dato (por ejemplo, mensual) y el anterior. Siempre se pierde el primer dato de la serie.

$$\nabla y_t = y_t - y_{t-1}$$

La primera diferencia de la serie de pasajeros (en log) se representa en la figura de la derecha.

Se observa que **la serie fluctúa alrededor de una media estable y finita.**

No obstante, todavía es estacional esta variable. Es decir, los picos altos (por encima de la media) son meses de verano y los bajos, son meses en donde se vuela mucho menos.



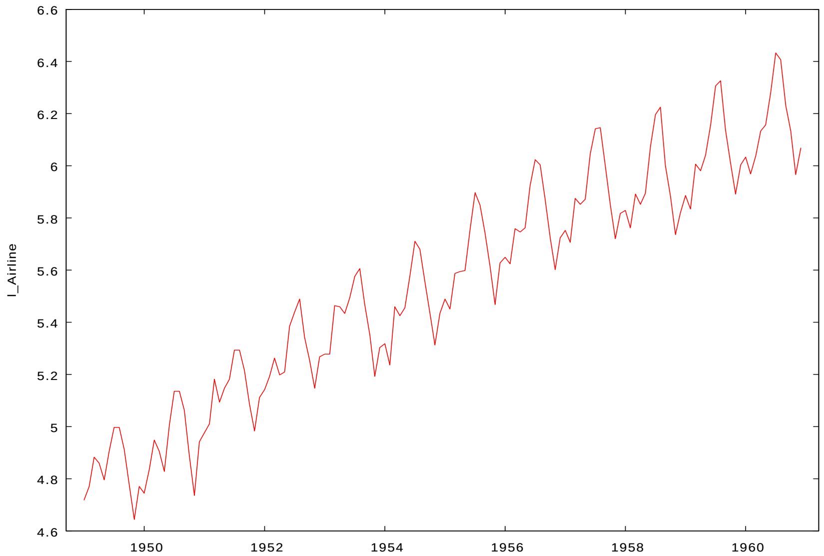
Otra aproximación (IV)

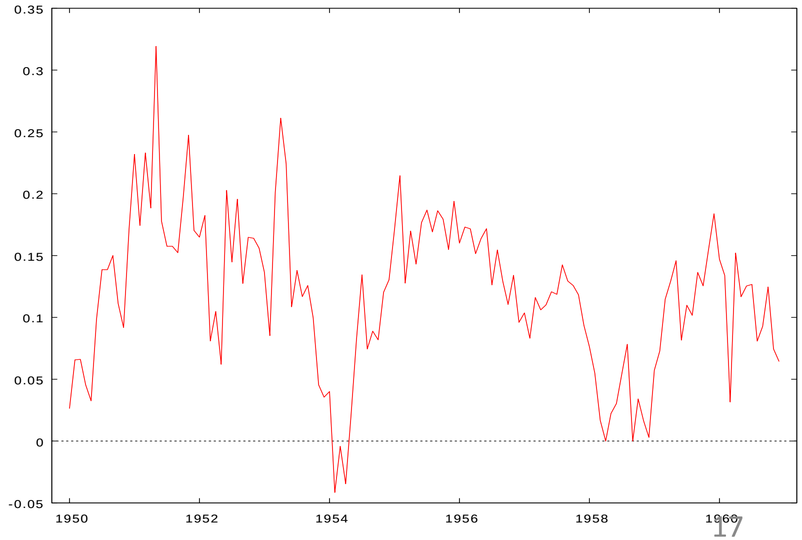
De nuevo, tenemos la serie en logaritmos, para mostrar el marcado comportamiento estacional de la misma (picos que se repiten cada 12 meses).

Una forma de desestacionalizar es tomar una diferencia estacional. Es decir, calcular la **diferencia entre el valor de la serie en un mes de un año con respecto al dato de ese mismo mes, pero del año anterior**. Se representa como:

$$\nabla_s y_t = y_t - y_{t-s}$$

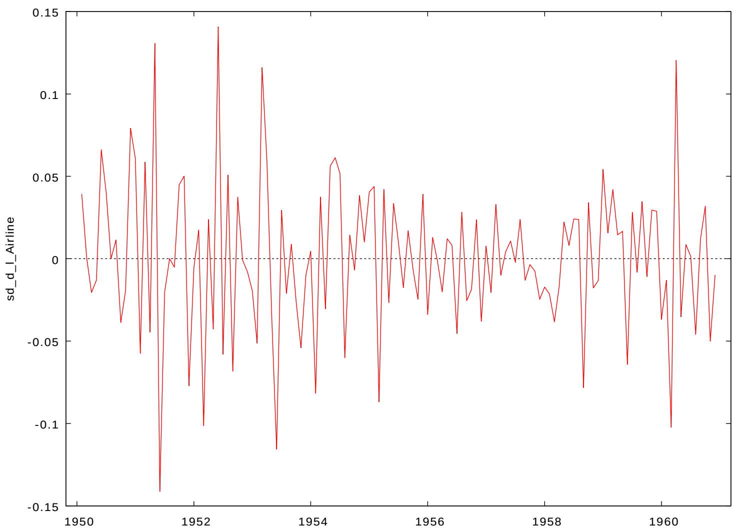
donde s es el período estacional (s=12). La figura de abajo representa la diferencia estacional de la serie. No se observa estacionalidad, aunque localmente la serie no es estacionaria sino que deambula.





Otra aproximación (V)

Al final, **hay que tomar todas las transformaciones que induzcan estacionariedad**, es decir, el **logaritmo** (para que la varianza sea constante), la **diferencia regular** (para eliminar la tendencia) y la **diferencia estacional** (para eliminar el componente estacional). Se presenta el gráfico de la serie con todas las transformaciones tomadas secuencialmente. Se escribiría como:



$$\nabla\nabla_{12} \ln NP_t = z_t$$

donde *NP* es el número de pasajeros.

Estas transformaciones son útiles para estabilizar la media y la varianza de una serie temporal económica

Otra aproximación (VI)

Además, estas **transformaciones tienen una interpretación económica sencilla**, que se resume en la siguiente tabla:

Transformación	Interpretación
$z_t = \nabla y_t = y_t - y_{t-1}$	Cambio en y_t Es un indicador de crecimiento absoluto.
$z_t = \ln y_t - \ln y_{t-1} \simeq \frac{y_t - y_{t-1}}{y_{t-1}}$	Es la tasa logarítmica de variación de una variable. Es un indicador de crecimiento relativo. Si se multiplica por 100 es la tasa de crecimiento porcentual de la variable.
$z_t = \nabla[\ln y_t - \ln y_{t-1}]$	Es el cambio en la tasa logarítmica de variación de una variable. Es un indicador de la aceleración de la tasa de crecimiento relativo de una variable.
$z_t = \ln y_t - \ln y_{t-s} \simeq \frac{y_t - y_{t-s}}{y_{t-s}}$	Es la tasa de crecimiento (en log) acumulada durante s períodos. Si el período estacional es un año, se interpreta como la tasa de crecimiento anual de una variable.

Modelos de autocorrelación (I)

Los modelos deterministas son útiles para descomponer una serie temporal, pero no sirven para predecir.

Si el objetivo es predecir a corto plazo, se adopta otra aproximación consistente en los siguientes pasos.

(1) Decidir **qué transformaciones tomar a los datos para inducir estacionariedad** en media y/o en varianza.

(2) Encontrar qué estructura de autocorrelación (modelo estocástico) **explica mejor esa serie ya transformada**. Para ello, se definen los **modelos ARMA** (familia de modelos estadísticos sencillos que son capaces de reproducir una gran variedad de series estacionarias).

Estacionariedad: Se dice que **una serie es estacionaria si sus propiedades estadísticas permanecen constantes a lo largo del tiempo**. Si y_t es una serie estacionaria (en media y varianza) se cumple que:

$$E[y_t] = \mu, \quad E[(y_t - \mu)^2] = \gamma_0, \qquad E[(y_t - \mu)(y_{t-k} - \mu)] = \gamma_k, \forall t, \forall k$$

donde μ, γ_0, γ_k son momentos finitos que no dependen del tiempo.

Modelos de autocorrelación (II)

El **modelo de autocorrelación más sencillo es un AR(1**), es decir, un modelo autorregresivo de orden 1 (el orden del modelo se escribe entre paréntesis). Si una serie temporal estacionaria sigue un modelo AR(1), entonces:

$$y_t = c + \phi_1 y_{t-1} + a_t$$

donde *c* es una constante, ϕ_1 es el parámetro autorregresivo y a_t es un error con esperanza cero, varianza constante (igual a σ_a^2) y ausencia de autocorrelación con cualquier otro error fechado en otro instante. La interpretación es como en una regresión, salvo que **a la variable dependiente le influye sólo su pasado más inmediato y no existen otras variables explicativas.**

Recuérdese que éstas eran las hipótesis habituales (y deseables) de las perturbaciones aleatorias en un modelo de regresión lineal.

Cuando una variable aleatoria tiene estas propiedades en el análisis de series temporales se le denomina **ruido blanco**. Si, además, la distribución que sigue es normal, entonces hablamos de un **ruido blanco gaussiano**.

Modelos de autocorrelación (III)

Los **momentos de un AR(1) son:** $y_t = c + \phi_1 y_{t-1} + a_t$

Media: $E[y_t] = c + \phi_1 E[y_{t-1}]$ y bajo estacionariedad, $E[y_t] = E[y_{t-1}]$ por lo que la media del proceso es:

$$\mu = E[y_t] = \frac{c}{1-\phi_1}$$

Varianza: $\mathrm{var}[y_t] = \phi_1^2 \mathrm{var}[y_{t-1}] + \sigma_a^2$ y bajo estacionariedad $\mathrm{var}[y_t] = \mathrm{var}[y_{t-1}]$ por lo que:

$$\gamma_0 = \mathrm{var}[y_t] = \frac{\sigma_a^2}{1-\phi_1^2}$$

Autocovarianzas: $\gamma_1 = \mathrm{cov}[y_t y_{t-1}] = E[\tilde{y}_t \tilde{y}_{t-1}] = E[(\phi_1 \tilde{y}_{t-1} + a_t)\tilde{y}_{t-1}] = \phi_1 \gamma_0$ donde la "tilde" en las variables indica que están desviadas con respecto a su media:

$$\gamma_2 = \mathrm{cov}[y_t y_{t-2}] = E[\tilde{y}_t \tilde{y}_{t-2}] = E[(\phi_1 \tilde{y}_{t-1} + a_t)\tilde{y}_{t-2}] = \phi_1 \gamma_1 = \phi_1^2 \gamma_0$$

y por inducción se tiene que $\gamma_j = \mathrm{cov}[y_t y_{t-j}] = \phi_1^j \gamma_0$ Se ha tenido en cuenta estacionariedad en covarianza, es decir, la covarianza entre dos variables sólo depende del desfase que hay entre ellas y no del tiempo. Por ello:

$E[\tilde{y}_t \tilde{y}_{t-1}] = E[\tilde{y}_{t-1} \tilde{y}_{t-2}] = \gamma_1$ o bien, $E[\tilde{y}_{t-1} \tilde{y}_{t-3}] = E[\tilde{y}_{t-2} \tilde{y}_{t-4}] = \gamma_2$

Modelos de autocorrelación (IV)

A partir de las autocovarianzas, se calcula lo que se denomina la **Función de Autocorrelación Simple o ACF de un modelo**, **cuyos valores no son más que coeficientes simples de autocorrelación de distinto orden:**

ACF de un AR(1):

$$r_1 = \frac{\gamma_1}{\gamma_0} = \frac{\phi_1 \gamma_0}{\gamma_0} = \phi_1 \qquad r_2 = \frac{\gamma_2}{\gamma_0} = \frac{\phi_1^2 \gamma_0}{\gamma_0} = \phi_1^2$$

y, por inducción, $r_j = \frac{\gamma_j}{\gamma_0} = \phi_1^j$

Además de la ACF, se suelen calcular **las autocorrelaciones parciales o valores de la PACF del modelo.** ¿Cómo se obtienen los valores de la PACF para cualquier modelo de autocorrelación? **El primero es el estimador MCO del coeficiente que relaciona el valor de la variable con su primer retardo**, es decir:

$$y_t = \phi_{10} + \phi_{11} y_{t-1} + \varepsilon_t$$

Modelos de autocorrelación (V)

El segundo valor de la PACF, mide la relación lineal entre la variable y su segundo retardo, pero teniendo en cuenta la influencia del primer retardo. Es decir, la regresión que hay que construir es:

$$y_t = \phi_{20} + \phi_{21} y_{t-1} + \phi_{22} y_{t-2} + \varepsilon_t$$

y el estimador MCO del parámetro ϕ_{22} es el segundo coeficiente de la PACF. Por tanto, **el coeficiente *j*-ésimo se calcularía a través de la regresión**:

$$y_t = \phi_{j0} + \phi_{j1} y_{t-1} + \phi_{j2} y_{t-2} + \ldots + \phi_{jj} y_{t-j} + \varepsilon_t$$

En el caso de **un AR(1), el único coeficiente de la PACF distinto de cero es el primero (y además coincide con el parámetro autorregresivo en cuantía y signo), siendo los demás nulos**. Esto es una "pista" importante a la hora de identificar si una serie temporal estacionaria sigue una estructura AR(1) o no.

Para facilitar el análisis, **se suelen dibujar los valores de la ACF y PACF teóricas** de cada modelo.

Modelos de autocorrelación (VI)

En la práctica, **cuando tenemos una serie estacionaria (ya transformada) sus autocorrelaciones simples y parciales son desconocidas y hay que estimarlas.**

Los **correlaciones teóricas de la ACF hay que sustituirlas por correlaciones muestrales.** Es decir:

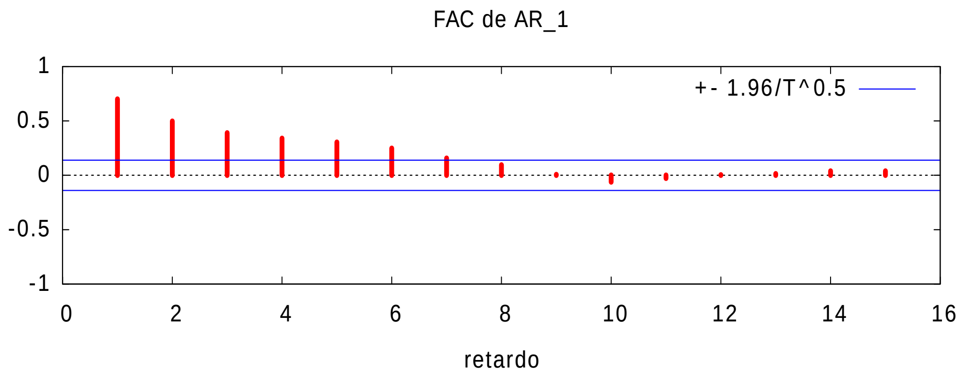
$$\hat{r}_j = \frac{\sum_{t=j+1}^{n}(y_t - \bar{y})(y_{t-j} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

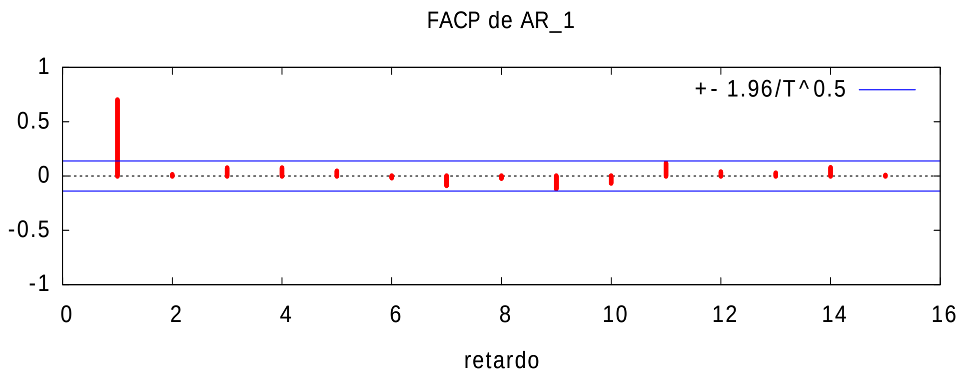
donde $\bar{y} = \sum_{t=1}^{n} y_t \big/ n$ es la media muestral de la serie.

Los valores muestrales de la PACF se obtienen estimando por MCO el correspondiente coeficiente (ϕ_{jj}) .Es necesario estimar cada coeficiente en un modelo diferente, en donde cada vez que se quiere estimar un coeficiente nuevo, se añade a la regresión un nuevo retardo de la serie.

$$y_t = \phi_{j0} + \phi_{j1} y_{t-1} + \phi_{j2} y_{t-2} + \ldots + \phi_{jj} y_{t-j} + \varepsilon_t$$

Modelos de autocorrelación (VII)





Cuando se tiene una serie temporal estacionaria se estiman los valores de la ACF y PACF y se dibujan. **Se calculan unas bandas de significación para saber qué coeficientes son distintos de cero** (fuera de bandas) y cúales no (dentro de bandas)

En estas figuras se muestra el perfil de la **ACF y PACF de un proceso AR(1) simulado donde el valor del parámetro autorregresivo es 0.7** (positivo y menor que uno).

Notad que **la ACF es positiva y va decayendo** porque el parámetro es positivo y menor que uno.

En **la PACF sólo el primer valor es "distinto" de cero estadísticamente**, indicando el orden del AR.

La identificación de un modelo consiste en comparar la ACF y PACF teórica de un modelo con la ACF y PACF estimada de una serie estacionaria.

Autocorrelación en el modelo de regresión (I)

En el **contexto del modelo de regresión lineal general:**

$$y_t = x_t^T \beta + \varepsilon_t \qquad \forall t = 1, 2, \ldots, n$$

decimos que las perturbaciones tienen autocorrelación si existen observaciones distintas $t \neq s$, tales que los errores asociados tienen una covarianza distinta de cero (y por tanto, correlaciones distintas de cero):

$$E[\varepsilon_t \varepsilon_s] \neq 0 \qquad \forall t \neq s$$

En este caso, **la matriz de varianzas y covarianzas de los errores no es diagonal.**

$$Y = X\beta + \varepsilon \qquad E[\varepsilon \varepsilon^T] = \Omega$$

Las **consecuencias sobre las propiedades del estimador MCO** de β son las mismas que si el problema es la heterocedasticidad. Es decir, **el estimador MCO sigue siendo lineal e insesgado, pero deja de ser eficiente.**

Autocorrelación en el modelo de regresión (II)

Al igual que en el caso de heteroscedasticidad, si existe autocorrelación en los errores de una regresión, sabemos que los contrastes habría que llevarlos a cabo usando una estimación de la matriz de varianzas siguiente:

$$\mathrm{var}[\hat{\beta}] = (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}$$

donde la matriz Ω no es diagonal y es desconocida. **Si la estructura de autocorrelación es desconocida, se puede usar una idea similar a la de la corrección de White cuando el problema es la heteroscedasticidad. Así, Newey y West (1987) proponen el siguiente estimador de la matriz anterior:**

$$\hat{\mathrm{var}}[\hat{\beta}] = (X^T X)^{-1} X^T \hat{V} X (X^T X)^{-1}$$

donde $$X^T \hat{V} X = \sum_{j=0}^{p} \sum_{t=j+1}^{n} w_j \hat{\varepsilon}_t \hat{\varepsilon}_{t-j} [x_t x_{t-j}^T + x_{t-j} x_t^T]$$

y $w_j = 1 - \frac{j}{p+1}$ Por último, p es el orden máximo de la autocorrelación en el error del modelo.

Autocorrelación en el modelo de regresión (III)

Obsérvese que decidir **el ordende autocorrelación *p* es a veces difícil**, ya que si este valor es alto la autocorrelación entre los errores es larga y si es bajo, la estructura de autocorrelación es más bajo.

Un procedimiento diferente para trabajar con autocorrelación, supone conocer el tipo de relación que hay entre el error de un modelo en un instante de tiempo y otro. Por ejemplo, supongamos que queremos estimar eficientemente los parámetros de la siguiente regresión simple:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

donde sabemos que el error sigue una estructura de autocorrelación AR(1). Es decir:

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + a_t$$

donde ϕ_1 es el parámetro autorregresivo y a_t es un proceso de ruido blanco (con esperanza nula, varianza constante y ausencia de autocorrelación serial)

Autocorrelación en el modelo de regresión (IV)

El **objetivo** es **encontrar una transformación del modelo original en el que en lugar del error con autocorrelación, aparezca un ruido blanco (que por definición no tiene autocorrelación).** En el ejemplo anterior, la transformación es fácil:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \qquad (1)$$

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1} \qquad (2)$$

$$\phi_1 y_{t-1} = \phi_1 \beta_0 + \phi_1 \beta_1 x_{t-1} + \phi_1 \varepsilon_{t-1} \qquad (3)$$

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + a_t \qquad (4)$$

Restando las expresiones (1) y (3) se tiene:

$$y_t - \phi_1 y_{t-1} = (1-\phi_1)\beta_0 + \beta_1 x_t - \phi_1 \beta_1 x_{t-1} + \varepsilon_t - \phi_1 \varepsilon_{t-1}$$

O bien:

$$y_t - \phi_1 y_{t-1} = (1-\phi_1)\beta_0 + \beta_1 (x_t - \phi_1 x_{t-1}) + a_t$$

El problema es que si el parámetro AR es desconocido, no podemos obtener los datos transformados de la variable dependiente y de la independiente.

Autocorrelación en el modelo de regresión (V)

En el caso de que el parámetro ϕ_1 sea desconocido, es necesario estimar conjuntamente el mismo con los parámetros β. Evidentemente, el criterio de estimación no es MCO. En el caso del que el parámetro ϕ_1 sea conocido (algo raro en la práctica) se pueden transformar los datos de la regresión anterior y estimar eficientemente por MCO. Por ejemplo, supongamos que $\phi_1 = 0.5$. El modelo transformado donde el error es ruido blanco es:

$$y_t - 0.5y_{t-1} = (1-0.5)\beta_0 + \beta_1(x_t - 0.5x_{t-1}) + a_t$$

Al aplicar MCO a la regresión anterior se obtiene una estimación de $\hat{\beta}_0^* = 0.5\hat{\beta}_0$ y de $\hat{\beta}_1$. En el caso de que $\phi_1 = 1$, en el modelo transformado trabajamos con datos diferenciados, es decir:

$$y_t - y_{t-1} = (1-1)\beta_0 + \beta_1(x_t - x_{t-1}) + a_t$$

Observad que el término constante desaparece en este caso y que $\nabla y_t = y_t - y_{t-1}$ $\nabla x_t = x_t - x_{t-1}$

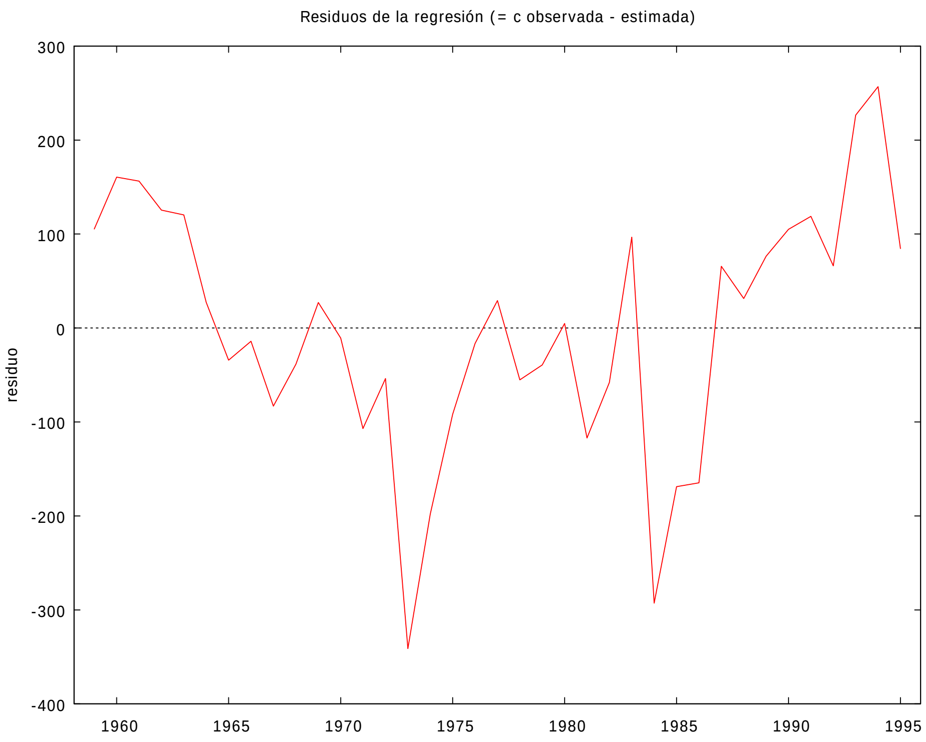
Ejemplo (I)

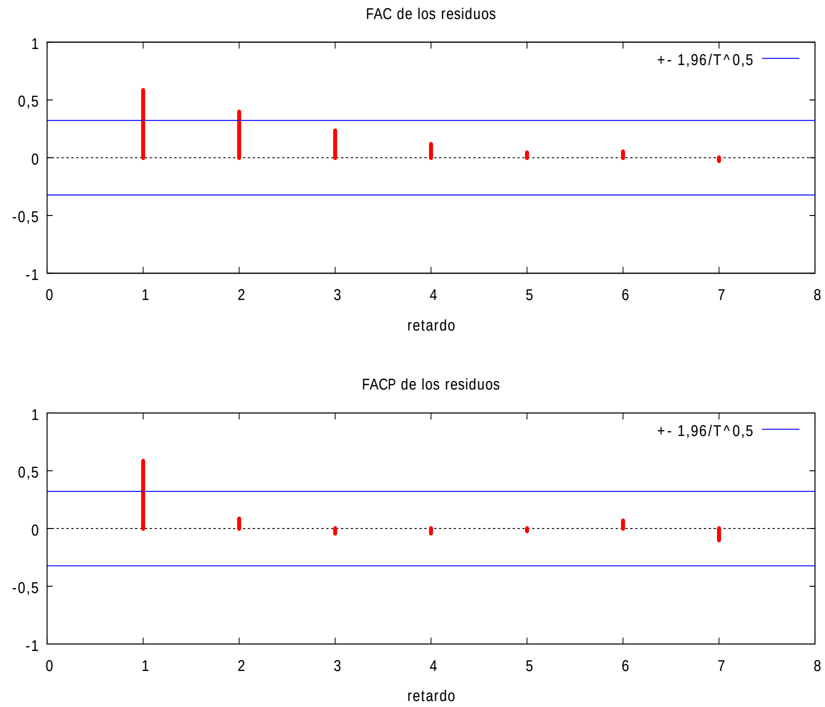
En el Tema 2 ya veíamos **la relación entre Consumo y Renta con datos temporales en la economía americana en términos per cápita**

```
Modelo 3: MCO, usando las observaciones 1959-1995 (T = 37)
Variable dependiente: c

             Coeficiente    Desv. Típica   Estadístico t    Valor p
  ----------------------------------------------------------------
  const      463,177        98,7912            4,688      4,10e-05  ***
  y            0,779419      0,00691064      112,8        1,99e-046 ***

Media de la vble. dep.  11328,65   D.T. de la vble. dep.   2505,241
Suma de cuad. residuos  619971,4   D.T. de la regresión    133,0920
R-cuadrado              0,997256   R-cuadrado corregido    0,997178
F(1, 35)                12720,51   Valor p (de F)          1,99e-46
Log-verosimilitud      -232,4412   Criterio de Akaike      468,8824
Criterio de Schwarz     472,1042   Crit. de Hannan-Quinn   470,0182
```

Ejemplo (II)

Los **residuos MCO resultantes** de dicho modelo tienen la siguiente evolución a lo largo del tiempo (gráfico de la izquierda) junto con la siguiente ACF y PACF (gráficos de la derecha). Parece claro **un modelo AR(1) con parámetro positivo.**





Ejemplo (III)

Si el término de error de la regresión de Consumo sobre Renta sigue una estructura autorregresiva de orden 1, **la estimación del modelo** que recoge ésta es:

```
Modelo: ARMAX, usando las observaciones 1959-1995 (T = 37)
Estimado usando el filtro de Kalman (MV exacta)
Variable dependiente: c
Desviaciones típicas basadas en el Hessiano

             Coeficiente   Desv. Típica     z       Valor p
  ---------------------------------------------------------
  const       2127,76      1031,52          2,063   0,0391     **
  phi_1          0,978174     0,0409301    23,90    3,17e-126 ***
  y              0,663438     0,0666839     9,949   2,55e-023 ***

Media de la vble. dep.  11328,65   D.T. de la vble. dep.   2505,241
media innovaciones      14,83765   D.T. innovaciones       104,7581
Log-verosimilitud      -226,1832   Criterio de Akaike      460,3663
Criterio de Schwarz     466,8100   Crit. de Hannan-Quinn   462,63
```

La estimación del parámetro autorregresivo es muy cercana a la unidad, lo que indica que se podrían tomar las primeras diferencias de Consumo y Renta.

Si además tomamos logaritmos, la conclusión es que relacionaríamos la Tasa log de variación del Consumo en función de la Tasa log de variación de la Renta.

Correlación espuria (I): Concepto

- Una correlación espuria es una relación empírica entre dos acontecimientos sin conexión lógica
- Las correlaciones espurias pueden producirse con datos de corte transversal o series temporales
  - **Ejemplo:** En 1952 J. Neyman analizó la relación entre la tasa de nacimientos y la población de cigüeñas en varias regiones, encontrando un elevado coeficiente de correlación entre ambas variables
  - **Ejemplo:** Utilizando datos anuales para el período 1866-1911, G. Udny Yule encontró que el coeficiente de correlación entre la tasa de mortalidad en Inglaterra-Gales y el porcentaje de matrimonios en la iglesia de Inglaterra era de 0.95
- **Al estimar regresiones entre series no estacionarias es muy fácil que la relación sea espuria, ya que basta con que ambas series tengan algo de tendencia para que surja una aparente relación entre ellas**
- **Al suprimir la tendencia, por ejemplo, diferenciando los datos, la relación espuria desaparece**

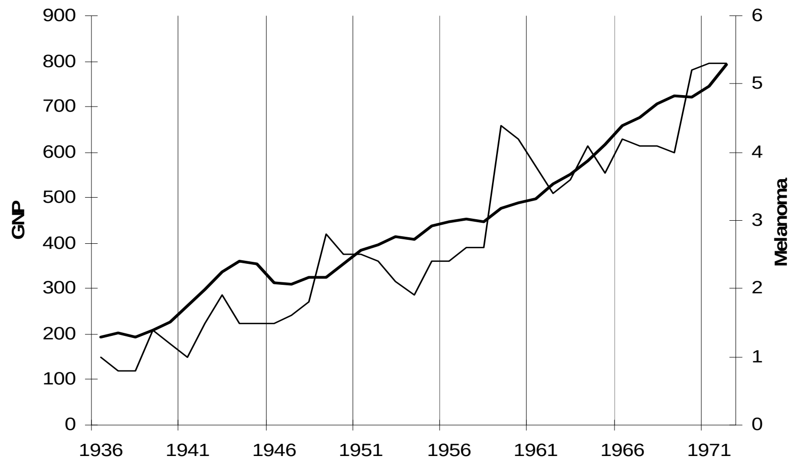
Correlación espuria (II): Los datos

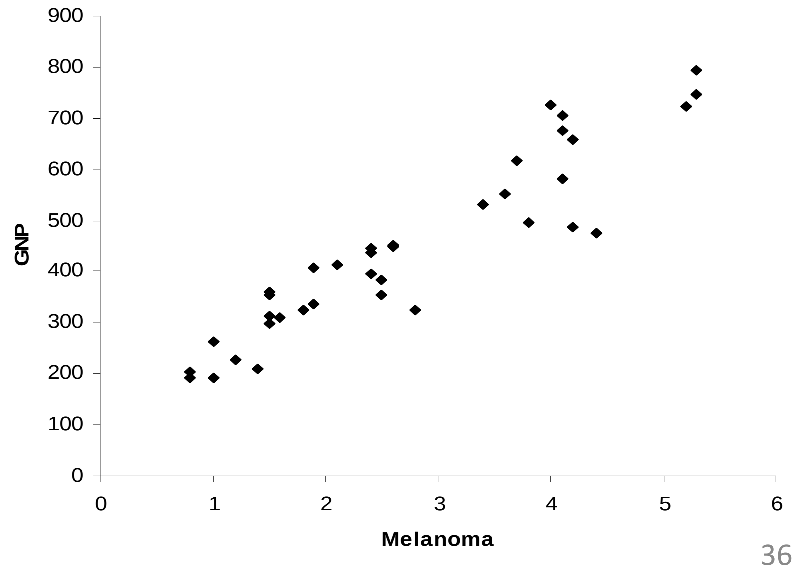
Los gráficos muestran las series anuales (1936-1972) de:

PNB nominal en Estados Unidos (datos en miles de millones de dólares), y la **Incidencia del melanoma en la población masculina** (datos ajustados de edad) en el estado de Connecticut.

Aparentemente, ambas series mantienen una fuerte y clara relación lineal, aunque conceptualmente resulte absurdo relacionarlas.

En el gráfico de abajo se muestra la nube de puntos real de PNB versus Incidencia del melanoma.





Correlación espuria (III): Relación estática en niveles

El cuadro muestra los **resultados de una regresión** en donde el PNB actúa como variable endógena y la incidencia de melanoma como variable explicativa

Resulta inmediato ver que:

Todos los coeficientes son estadísticamente significativos y

El R^2 es muy elevado, de cerca del 87%

El coeficiente estimado implica que, si aumentara la incidencia de melanoma en un caso, cabría esperar un aumento del PNB de 118.981 millones de dólares (???)

Modelo 1:
MCO, usando las observaciones 1936-1972 (T = 37)
Variable dependiente: GNP

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	118.566	23.7290	4.997	1.62e-05	***
MELANOMA	118.981	7.81415	15.23	5.22e-017	***

Media de la vble. dep.	443.6730	D.T. de la vble. dep.	171.4417
Suma de cuad. residuos	138787.6	D.T. de la regresión	62.97110
R-cuadrado	0.868836	R-cuadrado corregido	0.865088
F(1, 35)	231.8413	Valor p (de F)	5.22e-17
Log-verosimilitud	-204.7517	Criterio de Akaike	413.5034
Criterio de Schwarz	416.7252	Crit. de Hannan-Quinn	414.6392
rho	0.554021	Durbin-Watson	0.879122

Correlación espuria (IV): Relación en primeras diferencias

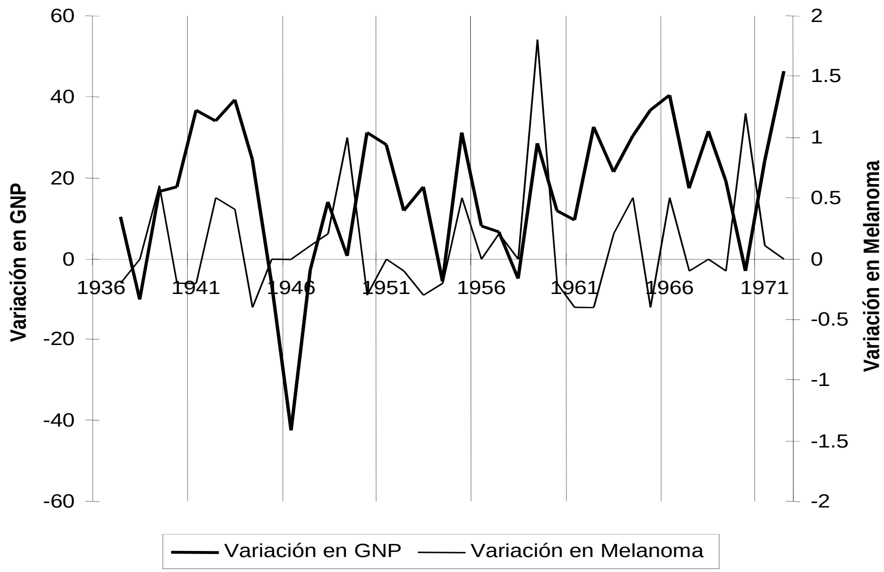
Si relacionamos las variables en primeras diferencias, esto es:

$$\nabla GNP_t = GNP_t - GNP_{t-1}$$

$$\nabla Melanoma_t = Melanoma_t - Melanoma_{t-1}$$

la tendencia suele desaparecer y, con ella, la relación espuria (véase el gráfico de arriba).

Consecuentemente, la relación de regresión entre las variables diferenciadas no resulta significativa y lógicamente, el coeficiente de bondad de ajuste es muy pequeño (del 0.0338%)



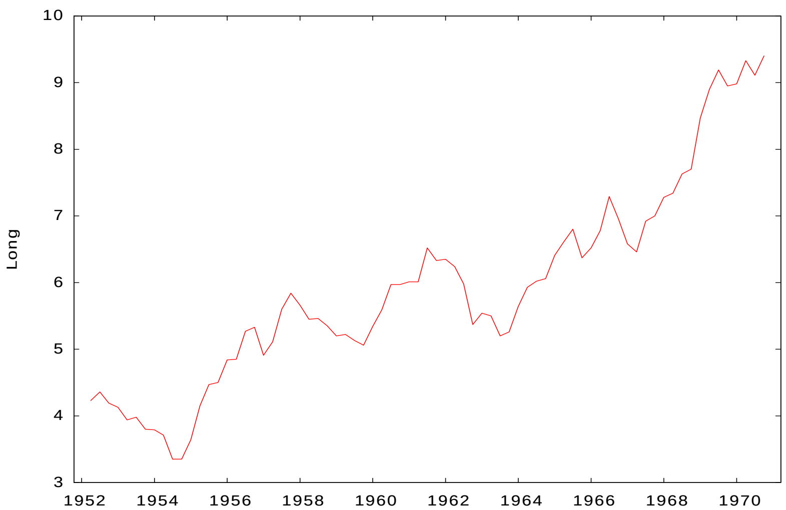

```
Modelo 3:
MCO, usando las observaciones 1937-1972 (T = 36)
Variable dependiente: d_GNP

               Coeficiente   Desv. Típica  Estadístico t   Valor p
  ---------------------------------------------------------------
  const          16.5684       3.17933         5.211       9.14e-06 ***
  d_MELANOMA      0.706295     6.58576         0.1072      0.9152

Media de la vble. dep.   16.65278   D.T. de la vble. dep.    18.22001
Suma de cuad. residuos   11614.98   D.T. de la regresión     18.48289
R-cuadrado               0.000338   R-cuadrado corregido    -0.029064
F(1, 34)                 0.011502   Valor p (de F)           0.915224
Log-verosimilitud       -155.0594   Criterio de Akaike       314.1187
Criterio de Schwarz      317.2858   Crit. de Hannan-Quinn    315.2241
rho                      0.356257   Durbin-Watson            1.262415
```

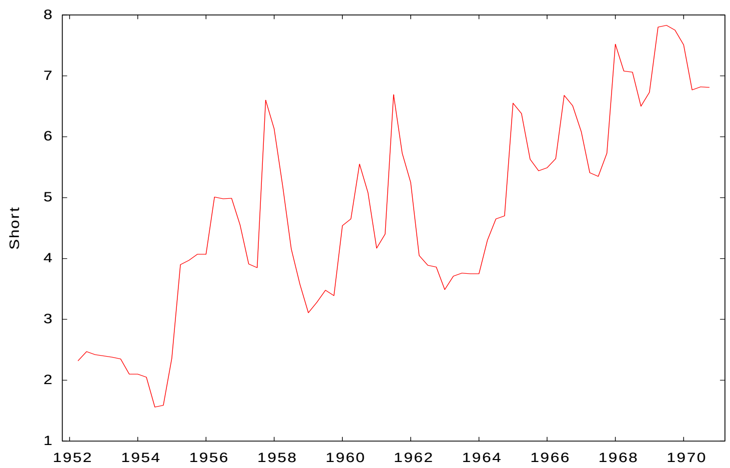
Cointegración (I): Concepto

- A diferencia de la correlación espuria, la **cointegración** es un concepto que caracteriza las relaciones válidas entre series no estacionarias.
- **Se dice que un conjunto de series está cointegrado si cada una de ellas necesita *d* diferencias para ser estacionaria,** esto es, son "integradas de orden "*d*" o I(*d*), **pero existe una combinación lineal de las mismas que es integrada de menor orden, es decir, I(*d*-*m*).**
- La situación más habitual es que dos series sean I(1) y su diferencia (o una combinación lineal de las mismas) sea I(0)
- La presencia **de cointegración supone que parte de la tendencia de las series es un componente común** (*cofeature*) y existe una combinación lineal de las series que carece de esta característica común.
- **Otra forma de entender la cointegración es que existe un equilibrio a largo plazo entre las series**, de manera que las desviaciones de este equilibrio tienden a desaparecer a corto plazo.

Cointegración (II): Detección visual de raíces unitarias

- **Las figuras muestran el perfil de los rendimientos porcentuales de activos de deuda pública a largo plazo (20 años) y corto plazo (91 días) en el Reino Unido entre 1952 (segundo trimestre) y 1979 (cuarto trimestre).**
- **Ambas series** muestran una tendencia creciente, por lo que **son, evidentemente, no estacionarias.**
- Es fácil comprobar que **su primera diferencia es estacionaria**, por lo que se trata de series I(1), es decir, integradas de orden uno.





Cointegración (III): *Spread*

La figura de la derecha muestra el diferencial o *spread* entre ambas series:

$$s_t = long_t - short_t$$

Como puede observarse, el *spread* muestra fluctuaciones amplias a corto plazo, pero retorna sistemáticamente a una media estable (de 1.18 puntos).

Asimismo, las funciones de autocorrelación simple (ACF) y parcial (PACF) muestran la pauta característica de un proceso AR(1) con parámetro positivo.