Análisis de Series Temporales (Continuación)

Juan A. Botía
juanbot@um.es

Departamento de Ingeniería de la Información y las Comunicaciones
Universidad de Murcia

TIIA, Primer Cuatrimestre, 2009/2010

El modelo aditivo

Podemos representar una serie como dependiente del tiempo

$$Y = F(t)$$

, siendo Y la variable y F dicha función

Hay cuatro factores que caracterizan una serie temporal

- Movimientos a largo plazo
- Movimientos cíclicos
- Variaciones estacionales
- Variaciones irregulares o aleatorias

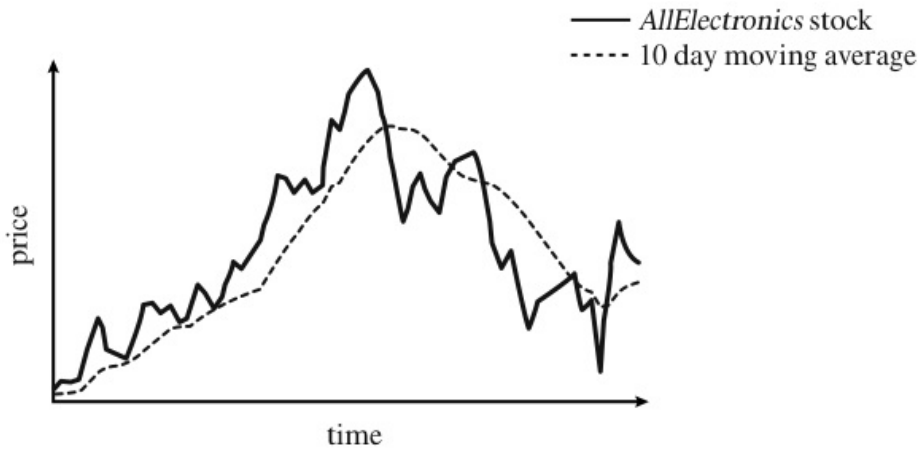
Movimientos a largo plazo

Definición

Se refiere a la tendencia, medida en intervalos de tiempo de gran tamaño, que sigue la serie temporal.

Representación: curva de tendencia (i.e monótona).

Ejemplo: mediante medias móviles ponderadas representamos la tendencia de la correspondiente serie como en el siguiente gráfico



Movimientos cíclicos o variaciones cíclicas

Definición

se refiere a repeticiones mediante ciclos de las tendencias que acabamos mencionar de una forma no aleatoria.

Estas no tienen por qué ser periódicas ni exactamente iguales los ciclos a lo largo del tiempo.

Ejemplo: Un ejemplo de estos ciclos puede ser el económico, caracterizado por las cuatro fases de recesión, recuperación, crecimiento y caída.

Variaciones estacionales y aleatorias

Estacionales

Este tipo de variaciones ocurren anualmente, como los picos de aumento de las ventas de flores el día de San Valentín o en las ventas de regalos en Navidad.

Irregulares o aleatorios

Consisten en movimientos esporádicos de las series temporales, debidos a sucesos inesperados como una huelga, o una inundación, por ejemplo.

Descomposición del modelo aditivo

- Análisis de series es la descomposición de la señal en esos cuatro componentes
- Podemos expresar la señal como

$$Y = T \oplus Z \oplus S \oplus R,$$

en donde $\oplus$ será el operador de agregación (suele ser una suma o un producto)

- Si $\oplus = +$, modelo aditivo.
  - T, Z, S y R son variables aleatorias y la suma de sus componentes da lugar a la serie temporal según

$$Y_t = T_t + Z_t + S_t + R_t, t = 1, 2, \ldots, n.$$

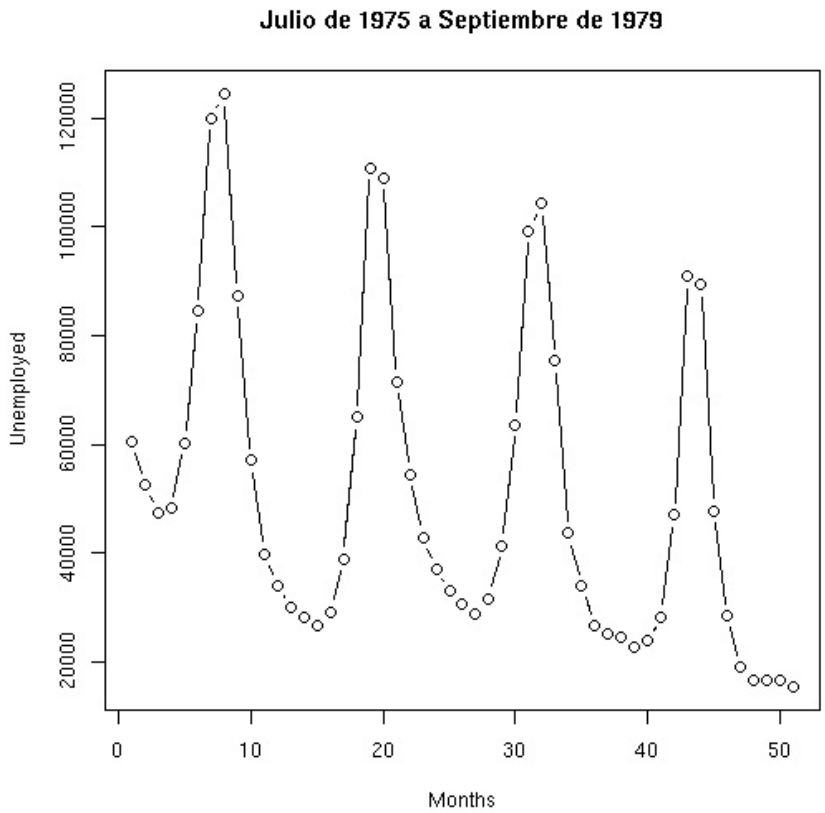
Descomposición del modelo aditivo (y II)

- Los componentes T_t y Z_t resumen el comportamiento de la serie temporal a largo plazo.
- R_t puede verse como el error al predecir el modelo no estocástico $y_t = T_t + Z_t + S_t$
- La esperanza del error, $E(R_t)$ existe y es cero al generarse errores por encima o debajo del modelo no estocástico (i.e. y_t), con lo cual se equilibran.

Ejemplo

Obsérvemos la figura que aparece a continuación generada a partir de R, con

```
> emp $<$- read.table("unemployed1.txt") \\
> plot(emp[[2]], emp[[3]],"b", xlab="Months",
    ylab="Unemployed",
    main="Julio de 1975 a Septiembre de 1979")}
```



Obsérvese como dicha figura muestra un componente de temporada y una tendencia descendente. El período que ahí aparece (i.e. de 1975 a 1979), puede ser excesivamente corto para que en él pueda observarse un ciclo a largo plazo.

1

[1] Serie obtenida en http://statistik.mathematik.uni-wuerzburg.de/timeseries/data/rawdata/unemployed1.txt

Averiguando la tendencia en una serie

- Lo más común es utilizar una media móvil (seq. medias aritméticas) como esta

$$\frac{y_1 + y_2 + \ldots + y_n}{n}, \frac{y_2 + y_3 + \ldots + y_{n+1}}{n}, \frac{y_3 + y_4 + \ldots + y_{n+2}}{n}, \ldots$$

- Tiende a reducir la cantidad de variación presente en los datos
- O bien una media móvil ponderada
  - Una posibilidad: los valores se ponderan, (más importancia a valores centrales y menos a más extremos, equilibra el efecto suavizante entre media y media)
  - Segunda posibilidad: más importancia a valores más recientes que a más antiguos, utilizando un factor del tipo δ^{t-i} variable, siendo t el tiempo e i la posición del valor en la serie, con $0 \leq \delta \leq 1$ (si $\delta = 1$, media móvil convencional)

Averiguando la tendencia en una serie (y II)

- Alternativa par obtener tendencia: utilizar ajuste por mínimos cuadrados, aplicándolo por intervalos
- Estacionalidad (i.e. fluctuaciones)
  - ej. aumento en la compra de abrigos en invierno
  - Estas fluctuaciones se han de identificar y, posteriormente, eliminar (después, hacemos análisis de tendencias y ciclos)
  - Para eliminarlas utilizamos el seasonal index
  - Ejemplo: si las ventas de los meses de octubre, noviembre y diciembre son un 80%, 120% y 140% de las ventas mensuales en promedio de todo el año, 80, 120 y 140 son los índices de temporada para esos meses del año
  - Dividimos el valor real de cada mes por su correspondiente índice (i.e. destemporizamos)
  - la tendencia, ciclos y los movimientos irregulares siguen estando ahí

Averiguando la tendencia en una serie (y III)

Una vez destemporizada la serie

1. suavizamos con medias móviles los datos
2. calculamos índices cíclicos y aplicamos similarmente a destemporización
3. Con lo que al final obtenemos la tendencia
4. Esta nos permitirá realizar predicciones en la serie.
5. La ventana de tiempo de estas predicciones tendrá que ver con a qué plazo hemos modelado la tendencia en la serie temporal.

Regresión en series temporales

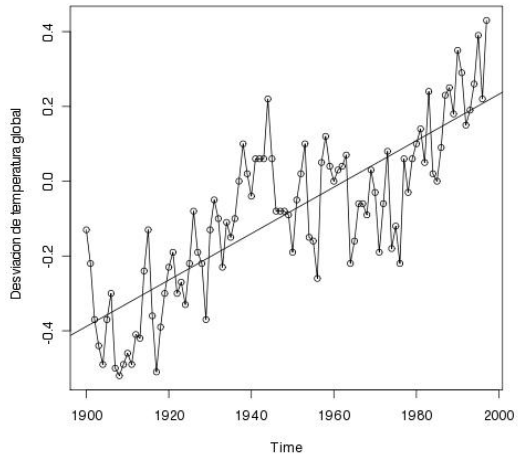
En la regresión en series temporales, todo se reduce a expresar la serie x_t en base, como siempre, a la combinación lineal de unas determinadas entradas $z_{t1}, z_{t2}, z_{t3}, \ldots$ independiente y ordenadas en el tiempo, mediante determinados coeficientes $\beta_1, \beta_2, \ldots, \beta_1$ y la relación es la de un modelo de regresión lineal

$$x_t = \beta_1 z_{t1} + \beta_2 z_{t2} + \cdots + \beta_q z_{tq} + w_t,$$

siendo w_t un error gausiano.

Ejemplo de uso

- Utilizaremos el ejemplo 2.1. de [1]
- En este ejemplo se hace uso de una serie que registra temperaturas globales como aparecen en la figura



de tal forma que se representan desviaciones de temperaturas desde el promedio, para los años de 1900-1997.

Ejemplo de uso

- Nótese tendencia creciente (justif. cambio climático)
- Podríamos utilizar R para ajustar el modelo de regresión a

$$x_t = \beta_1 + \beta_2 t + w_t, \;\; t = 1900, 1901, \ldots, 1997.$$

- En este caso el número de coeficientes (i.e. q) 2, $z_{t1} = 1$ y $z_{t2} = t$.
- Con reg. lineal simple, $\hat{\beta}_1 = -12.186$ y $\hat{\beta}_2 = .006$, es decir

$$\hat{x_t} = -12.186 + .006t.$$

- El incremento en grados cada 100 años es de 0.6 grados
- Con R

```
> gtemp = scan("/mydata/globtemp.dat")
> x = gtemp[45:142]
> t = 1900:1997
> fit=lm(x~t) # regress x on t
> summary(fit) # regression output
> plot(t,x, type="o", xlab="year", ylab="temp deviation")
> abline(fit) # add regression line to the plot
```

Teoría de regresión en el contexto de series temporales

- Teoría de regresión lineal está muy desarrollada como teoría
- Varios métodos para obtener los coeficientes de manera adecuada
- Varias alternativas para seleccionar, de entre posibles modelos que podamos generar, que expliquen los datos, aquellos que estadísticamente tengan una aproximación más significativa.

Teoría de regresión en el contexto de series temporales

Qué es un análisis de varianza (ANOVA)

La técnica del análisis de varianza es usada comúnmente para realizar un test estadístico que comprueba si la media de dos muestras es igual.

- Idea básica: la suma de desviaciones cuadráticas se puede descomponer en dos sumandos
  1. El primer sumando se refiere a la contribución al error, es decir a RSS, del modelo como tal
  2. El segundo se refiere a la contribución a RSS del ruido blanco
- De esta forma (i.e. SS),

$$\sum_{t=1}^{n}(x_t - \bar{x})^2 = \sum_{t=1}^{n}(\hat{x}_t - \bar{x})^2 + \sum_{t=1}^{n}\hat{w}_t^2,$$

- $\hat{x}_t$ se refiere a la salida del modelo, sin considerar en él, el término w_t

Fundamentos de ANOVA

- ANOVA intenta comparar poblaciones para detectar el efecto de un tratamiento (nosotros comparamos modelos)
- Idea básica (y ii): comparan las medias de las poblaciones de datos (i.e. poblaciones compuestas errores del modelo lineal) mediante comparación de varianzas
- Test de hipótesis:

$$H_0 = \bar{x} = \bar{x}'.$$

- ¿Qué varianzas se estudian para ello?
  - han de ser forzosamente estimadores (hay dos que se obtienen independientemente)
  - (1) error debido al modelo de regresión exclusivamente (i.e. sin considerar w_t), (2) w_t (i.e. lo que se denomina error del estimador intramuestral)
  - Ambos combinados en un ratio tal que si mayo que 1, hay significancia estadística.
  - Resulta que el ratio de estos dos estimadores sigue una distribución F, cuando H_0 es verdad.

Fundamentos de ANOVA (y II)

- Mediante vectores rescribimos el modelo de regresión

$$x_t = \beta' z_t + w_t,$$

β' es el vec. de coef., z_t es el vector de entradas y w_t $iid(o, \sigma_w^2)$.

- Para estimar los β minimizamos

$$RSS = \sum_{t=1}^{n} (x_t - \beta' z_t)^2,$$

- Para aplicar un ANOVA, necesitamos el estimador de la varianza intramuestral

$$s_w^2 = \frac{RSS}{n-q}.$$

Aplicando el ANOVA

- Supongamos que queremos nuestro modelo con otro modelo que hace uso de menos parámetros, i.e. q_1 parámetros, con $q_1 < q$ y el conjunto de variables es $z_{1t} = (z_{t1}, z_{t2}, \ldots, z_{tq1})'$ de tal forma que el nuevo modelo es

$$x_t = \beta_1' z_{1t} + w_t,$$

  a comparar con

$$x_t = \beta' z_t + w_t.$$

- Utilizaremos una tabla
  - Filas: la contribución a la varianza de cada uno de estos errores (modelo y ruido) en una fila
  - Columnas: los grados de libertad usados en cada error, y dos columnas más para el valor sin escalar y escalado (compensamos en SS el tamaño muestral)

Tabla ANOVA

Fuente	df	SS	SS medio
$z_{t,q_1+1}, \ldots, z_{t,q}$	$q - q_1$	$SS_{reg} = RSS_1 - RSS$	$MS_{reg} = SS_{reg}/(q - q_1)$
Error	$n - q_1$	RSS	$s_w^2 = RSS/(n - q)$
Total	$n - q_1$	RSS_1	

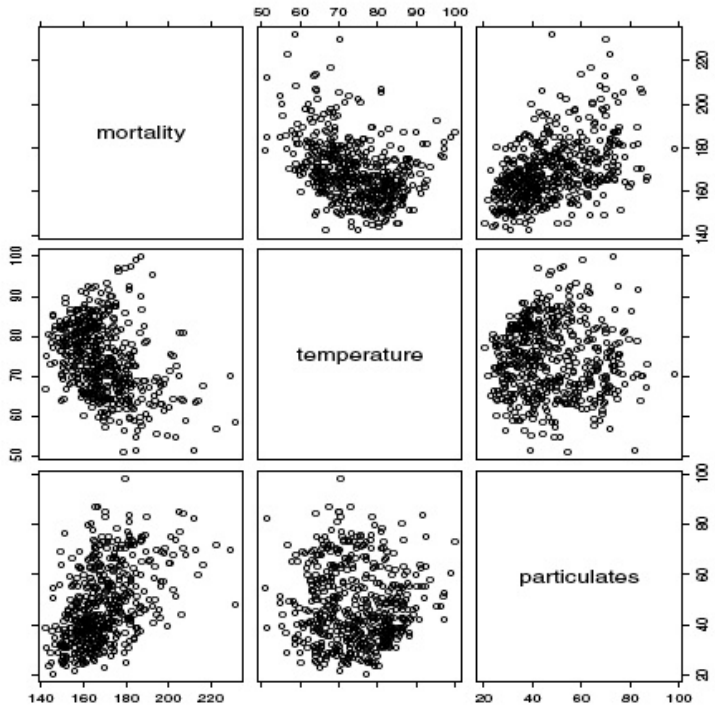
De esta forma, el estadístico se forma según

$$F_{q-q_1,n-q} = \frac{MS_{reg}}{s_w^2},$$

en donde n es el número de ejemplares de la muestra, q es el número de parámetros del modelo con más parámetros de entre los que se están comparando, y q_1 en este caso es el número de parámetros del otro modelo.

Un ejemplo

- Intentamos relacionar la cantidad de partículas en suspensión en el aire, la temperatura ambiente y la mortalidad en Los Ángeles
- ¿Hay algún indicio de correlación? Utilizamos un scatter plot



- Correlación positiva, por pares, entre las tres variables

Un ejemplo (y II)

Para ello, vamos a definir los siguientes modelos

$$\begin{aligned} M_t &= \beta_0 + \beta_1 t + w_t && A \\ M_t &= \beta_0 + \beta_1 t + \beta_2 (T_t - T.) + w_t && B \\ M_t &= \beta_0 + \beta_1 t + \beta_2 (T_t - T.) + \beta_3 (T_t - T.)^2 + w_t && C \\ M_t &= \beta_0 + \beta_1 t + \beta_2 (T_t - T.) + \beta_3 (T_t - T.)^2 + \beta_4 P_t + w_t && D, \end{aligned}$$

- $T.$ es la media de la temperatura (evita problemas de escala con los β)
- A solamente va a marcar la tendencia. El modelo B es lineal con respecto a la temperatura, el C es cuadrático, y D es cuadrático con respecto a temperatura y polución.

Un ejemplo (y III)

- Podemos utilizar un análisis de varianza (i.e. un ANOVA) comparando, por pares, las distintas variables que podríamos usar en nuestros modelos
- Usando el estadístico F, a partir de RSS

Modelo	RSS	s_w^2	AIC
A	40.020	79.09	5.38
B	31.413	62.20	5.14
C	27.985	55.52	5.03
D	20.509	40.77	4.72

- Si queremos comparar los modelos A y D, tenemos que $q = 5$, $q_1 = 2$, $n = 508$, tenemos que

$$F_{3,503} = \frac{(40.020 - 20.509)}{20.509} \frac{503}{3} = 160,$$

y como $F_{3,\infty}(.001) = 5.42$, no podemos aceptar la hipótesis nula. Así que damos por bueno el RSS del modelo D con respecto del A.

Si aplicamos R

Por tanto, nos quedamos con este último. La serie de comandos en R para computar la regresión es la siguiente

```
> mort = scan("/mydata/cmort.dat")
> temp = scan("/mydata/temp.dat")
> part = scan("/mydata/part.dat")
> temp = temp - mean(temp)
> temp2 = temp^2
> t = 1:length(mort)
> fit = lm(mort~t + temp + temp2 + part)
> summary(fit) # Results
> AIC(fit)/508 # R gives n*AIC
> pairs(cbind(mort, temp, part)) # scatterplot matrix
```

Conclusiones

- Lo que hemos visto pretende ser solamente una introducción
- La filosofía es similar, las técnicas son diferentes
- Es importante reparar en que no todos los datos se presentan en forma de casos sin asociación
- Cada vez más es necesario manejar otras visiones

Bibliografía


Robert H. Shumway and David S. Stoffer.
Time Series Analysis and Its Applications with R Examples.
Springer, 2nd edition edition, 2006.