GLMS y series temporales © EDICIONES ROBLE, S.L.

Indice

GLMS y series temporales

I. Introducción

En esta unidad, se mostrarán modelos que pueden usarse para estudiar variables objetivo de distinta naturaleza: los GLMs o modelos lineales generalizados. Conociéndolos se dispondrá de una batería de posibilidades en función de los datos que se quieran estudiar.

En la segunda parte, se tratarán series temporales y su modelado. El estudio de series temporales es muy interesante para elaborar presupuestos y construir sistemas logísticos, ya que un buen método predictivo que anticipe en el tiempo las necesidades reducirá costes considerablemente y las desviaciones de presupuesto también serán menores.

II. Objetivos específicos

- Entender los GLMs y los tipos de variables objetivo en los que se aplica cada caso.
- Comprender el concepto de serie temporal.
- Entender lo que es una serie estacionaria.
- Comprender la estacionalidad y la tendencia de una serie.
- Desarrollar modelos de predicción SARIMA sobre una serie.

III. GLMs

Los modelos lineales generalizados o GLMs suponen una ampliación de los que se han visto hasta ahora (regresión lineal y logística). Para cada distribución probabilística de la familia exponencial, se puede generar un modelo análogo a los que se han construido que servirá para explicar variables objetivo de distinta naturaleza.

Hay multitud de modelos basados en la familia exponencial, aquí se verán:

- **Gaussiano**: regresión lineal (ya estudiado).
- **Binomial**: regresión logística (ya estudiado).
- **Multinomial**: para clasificación múltiple. Se usa cuando hay que afrontar un problema de clasificación en el que se tienen más de dos categorías (para este caso se usa el GLM Binomial visto). Ejemplo: clasificar una imagen con una cifra manuscrita a uno de los diez dígitos.
- **Poisson o binomial negativa**: para variables de conteo. Se usa cuando la variable objetivo está compuesta por números enteros no negativos (0,1,2,3,4...).
  Ejemplo: estimar la cantidad de llamadas que recibe un *call center* en un día en función de una colección de predictores.
- **Gamma**: para variables positivas. Se usa para variables positivas continuas y se adapta a multitud de formatos de función de densidad, lo que la convierte en una regresión muy versátil. Ejemplo: estimar el ROA (número mayor que 0) que una inversión va a tener tomando como predictores características de la misma.

Descarga: Consulta el notebook UD6 N01

Descárgate el archivo UD6_N01 en R y este csv. También puedes verlo en .html

IV. Series temporales

4.1. Conceptos

Serie temporal

Una serie temporal o cronológica es una secuencia de datos, observaciones o valores, medidos en determinados momentos y ordenados cronológicamente.

Las series se representan como secuencias X1, X2, X3... Xt que recogen un valor numérico cada instante de tiempo, con una cierta periodicidad fijada: segundos, días, semanas, meses, años...

En las siguientes imágenes, se puede observar la serie temporal de afluencia de pasajeros según el mes y año desde 1949 a 1960:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	264	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	284	277	317	313	318	374	413	405	355	306	271	306
1957	315	301	356	348	355	422	465	467	404	347	305	336
1958	340	318	362	348	363	435	491	505	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	405
1960	417	391	419	461	472	535	622	606	508	461	390	432

Imagen 6.1. Dataframe con serie temporal. *Fuente:* elaboración propia.

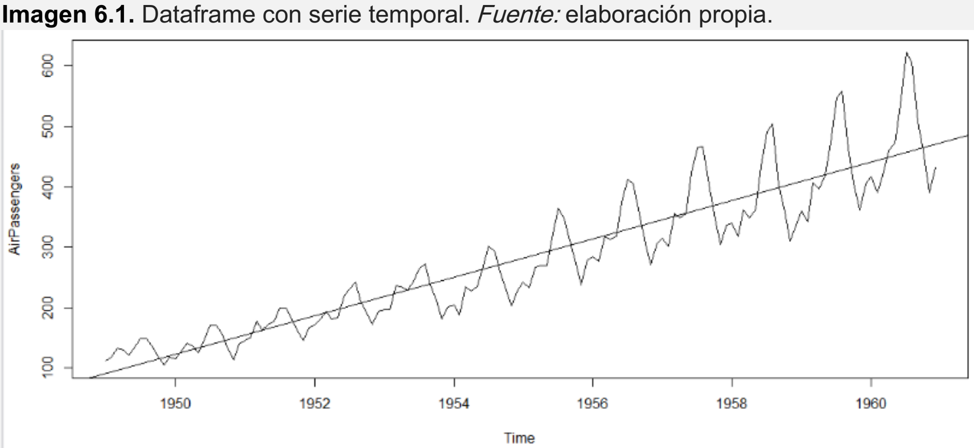

Imagen 6.2. Serie temporal. *Fuente:* elaboración propia con R.

Se estudia una serie temporal para recoger patrones de comportamiento a lo largo del tiempo. Esta información se puede utilizar en un modelo matemático que permita realizar predicciones de la serie temporal.

Realizar predicciones de una serie temporal es importante, ya que permite realizar presupuestos ajustados, estimar necesidades de personal y de logística, predecir la demanda de consumo de bienes para suplir las necesidades sin pasarse ni quedarse corto, etc.

Todas estas capacidades son convenientes y deseables en el seno de cualquier negocio, pero también se pueden usar para predecir series de carácter meramente científico como series sismológicas que indiquen si un terremoto sucederá, encefalogramas que indican si un epiléptico sufrirá un ataque, la predicción del tiempo, etc.

Se señalan, a continuación, algunos conceptos importantes de una serie temporal:

Frecuencia

Intervalo temporal sobre el cual es esperable que la serie repita un cierto patrón.

Ejemplo: la temperatura en una zona cada 24 horas (ciclo debido a la rotación de la Tierra) o cada 365 días (ciclo solar).

Estacionalidad

Es la repetición de patrones a lo largo de distintas frecuencias.

Ejemplo: en España, en verano, se producen muchas más contrataciones que en primavera u otoño debido a la afluencia de turismo. Esto es una manifestación de estacionalidad en la serie “contrataciones de empleo”. En la imagen 6.2. se puede observar un mismo dibujo año tras año, que corresponde a la estacionalidad cada 12 meses.

Tendencia

Es la manifestación de la evolución del comportamiento de la serie. Puede haber tendencias de crecimiento, decrecimiento o mantenimiento.

Ejemplo: en la serie temporal de la imagen 6.2. se puede observar una tendencia creciente en la cantidad de pasajeros año tras año.

Ruido

Las series temporales tienen una componente formada por variaciones que no siguen ningún patrón. Está representada por las desviaciones numéricas que no tienen estructura.

Ejemplo: en las ventas diarias de un supermercado influye si cierran inesperadamente algunos supermercados de la zona, esto producirá mayor afluencia al mismo sin ningún tipo de patrón, por lo que se puede considerar ruido.

Serie estacionaria

Una serie se dice que es **estacionaria** (no confundir con estacionareidad o con estacionalidad) si se verifica que su media y su varianza son constantes a lo largo del tiempo y que las dependencias temporales del valor actual respecto a los anteriores son fijas. Es decir, lo que pasa hoy y depende de días pasados no cambia conforme la serie avanza, sino que es una regla fija.

Para poder ajustar modelos matemáticos sobre una serie temporal, esta debe ser estacionaria. Esto es imprescindible, ya que los modelos recogen el patrón de dependencia de un día respecto a los *n* días pasados en forma de fórmula algebraica, y esta fórmula carece de sentido si no se tiene estacionareidad.

Se verán dos ejemplos, uno con una serie estacionaria y otro en el que la serie no lo es:

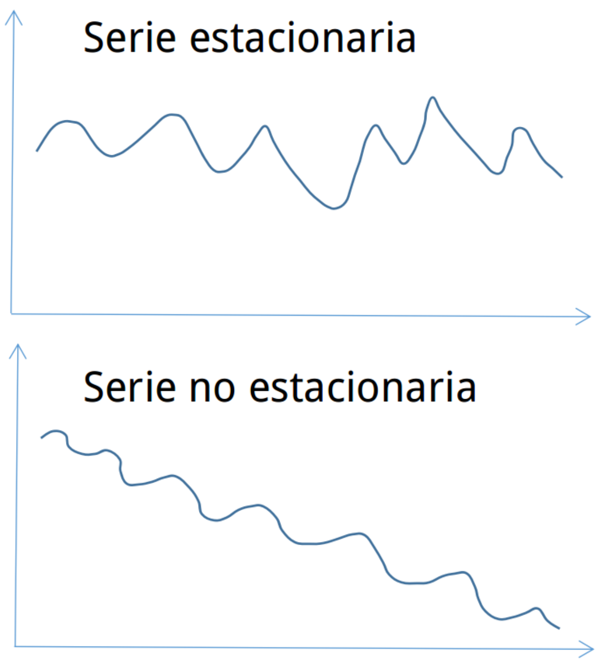

Imagen 6.3. Estacionaria *vs* no estacionaria. *Fuente:* elaboración propia.

Descarga: Consulta el notebook UD6 N02

Descárgate el archivo UD6_N02 y ejecútalo en R. También puedes verlo en .html

4.2. Modelos lineales de Box-Jenkins

Las estructuras matemáticas clásicas que se aplican para realizar predicciones de las series temporales son los modelos ARIMA y SARIMA de Box-Jenkins.

Estos son modelos que se basan en dos ideas:

Medias móviles
Dar una predicción basada en la media de los ruidos que se han sucedido en los últimos valores de la serie.

Autorregresión
Dar una predicción basada en una combinación lineal de los últimos valores de la serie.

Si se resumen los tipos, existen:

- **Modelos Autorregresivos AR:** La variable se explica por los valores de la variable en periodos anteriores.
- **Modelos de Media Móvil MA:** La variable se explica por los valores de los residuos anteriores.
- **Modelos Autorregresivso de Media Móvil ARMA:** Combinación de modelos AR y MA.
- **Modelos Autorregresivo integrado de Media Móvil ARIMA:** Modelos ARMA con integración de variables.
- **Modelos SARIMA:** Modelos ARIMA con estacionalidad.

4.3. Correlogramas y diagramas de ajuste de modelos

Una vez se confirma que una serie es estacionaria, tiene sentido analizar las correlaciones del valor de la serie en cada instante respecto a los instantes anteriores. Esto se puede representar gráficamente mediante los **autocorrelogramas ACF** y los **autocorrelogramas parciales PACF**

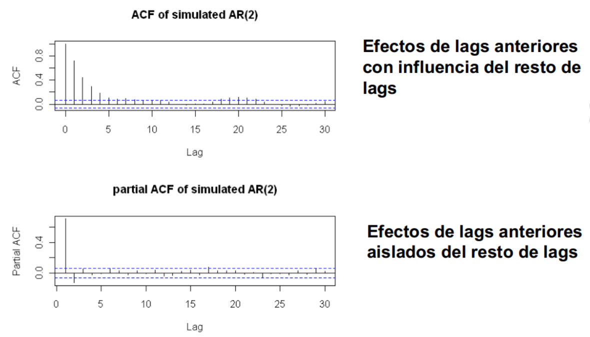

Imagen 6.4. Autocorrelogramas. *Fuente:* elaboración propia con R.

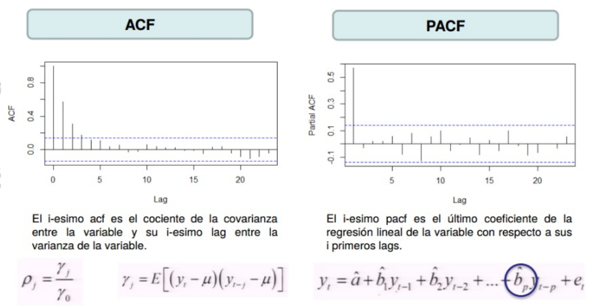

Imagen 6.5. Autocorrelogramas. *Fuente:* elaboración propia con R.

Para ajustar un modelo SARIMA, debemos fijar los parámetros (p,d,q, P,D,Q) y la frecuencia de la serie. Estos parámetros se ajustan según los valores observados en las gráficas ACF y PACF:

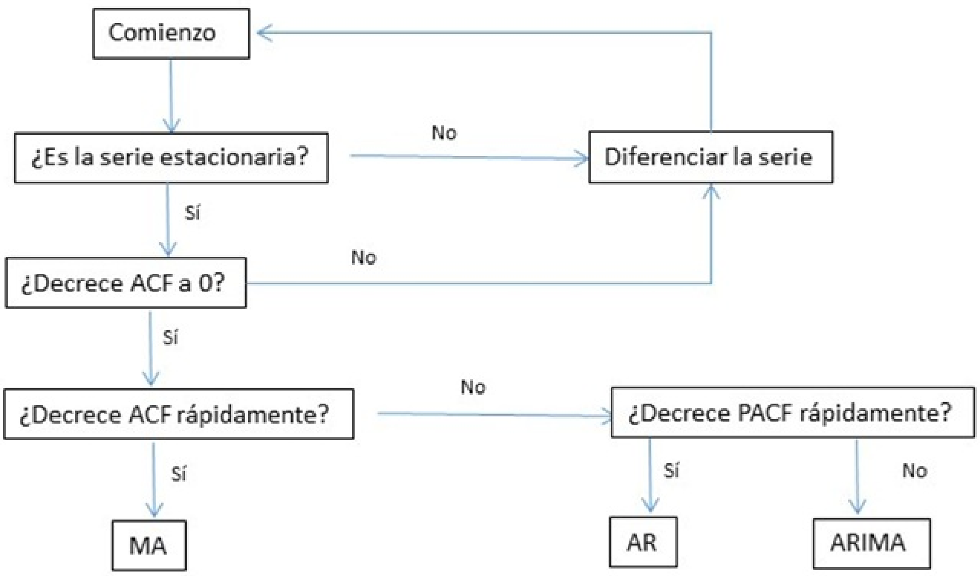

Imagen 6.6. Metodología Box-Jenkins. *Fuente:* elaboración propia.

Finalmente, los modelos de series temporales para predictivo se suelen evaluar usando la métrica MAPE o error porcentual medio. Esto se debe a que las predicciones habitualmente sirven para presupuestar y el objetivo es reducir el porcentaje de error en los presupuestos.

A pesar de ver la metodología por encima, la selección del modelo óptimo se realizará usando el criterio AIC de modo sistematizado en librerías específicas de R diseñadas explícitamente para series temporales.

Descarga: Consulta el notebook UD6 N03

Descárgate el archivo UD6_N03 y ejecútalo en R. También puedes verlo en .html

V. Resumen final

En esta última unidad, se completa el estudio de los modelos generalizados exponiendo casos de aplicación.

También se han expuesto los fundamentos del análisis de series temporales y la creación de modelos predictivos sobre las mismas. Si bien es un campo complejo, aquí se han puesto a disposición del alumno los elementos esenciales para poder comenzar su estudio.

Ejercicios

Caso práctico

Como repaso del tema y preparación para el Caso práctico final, se presenta el siguiente caso práctico.

Descárgate el archivo ACTIVIDAD6_UD6 en R y la csv del caso. También puedes verlo en .html

Cuando lo hayas realizado, puedes descargar su solución y comprobar tus resultados.

Solución

En los siguientes archivos dispones de la solución de la actividad propuesta:

- Solución en .html
- Solución en R.

Recursos

Enlaces de Interés

- **http://UD6_N03_corr.Rmd**
  http://UD6_N03_corr.Rmd
  Notebook UD6 N03

- **http://UD6_N03_corr.Rmd**
  http://UD6_N03_corr.Rmd

- **http://UD6_N03_corr.html**
  http://UD6_N03_corr.html

- **http://UD6_N01.Rmd**
  http://UD6_N01.Rmd
  UD6_N01.Rmd

- **http://UD6_N01.html**
  http://UD6_N01.html
  UD6_N01.html

Glosario.

- **Estacionalidad**: Comportamiento de la serie que tiene que ver con la repetición de un patrón según la frecuencia de la misma.

- **Frecuencia de una serie temporal**: Unidad de repetición elemental de los ciclos de la serie. Puede ser diaria, anual, mensual...

- **GLM**: Modelo lineal generalizado. Supone una ampliación del concepto de la regresión lineal y logística en un marco de distribuciones en la familia exponencial que permite más aplicaciones.

- **Serie temporal**: Secuencia numérica de valores tomados con una cierta periodicidad fija.

- **Tendencia**: Es la evolución observada de la serie.