

Índice general

1. Modelos estadísticos clásicos	3
1.1. Modelo de Regresión Lineal General	3
1.1.1. Componentes del Modelo Lineal Generalizado	4
1.1.1.1. Componente aleatoria	4
1.1.1.2. Componente sistemática	5
1.1.1.3. Función link o función enlace	5
1.1.2. Estimación de parámetros	5
1.1.3. Modelo de Regresión Poisson	6
1.1.4. Modelo de Regresión de Binomial Negativa	7
1.2. Análisis de series temporales	8
1.2.1. Metodología Box-Jenkis	10
Bibliografía	11

Capítulo 1

Modelos estadísticos clásicos

A continuación se exponen los modelos estadísticos clásicos de cara a predecir la demanda de los productos. Esta variable depende de varios factores, como el período del año, el precio del producto, el precio de los productos competidores o los gustos de cada consumidor, entre otros. Se trata de una variable cuantitativa discreta, ya que el número de productos que se venden será un valor entero no negativo $y = 0, 1, \dots$

1.1. Modelo de Regresión Lineal General

El objetivo es encontrar un modelo estadístico que describa la situación real de ventas de productos a través de un Modelo Lineal General (MLG), donde una variable de interés (variable objetivo) pueda ser descrita por un conjunto de variables explicativas (variables independientes).

Para ello, debemos estimar los parámetros que caracterizan al modelo, es decir, medir el efecto de cada variable explicativa sobre la variable objetivo, y con este fin, es necesario definir una serie de hipótesis del modelo de regresión lineal general:

- Independencia lineal entre las variables explicativas: Esto significa que cada variable explicativa contiene información adicional sobre la variable objetivo, ya que si hubiera información repetida, habría variables explicativas dependientes linealmente de otras.
- Los MLG suponen que existe una función g , llamada función link, que relaciona la media de la variable respuesta, μ con el resto de variables explicativas de la siguiente forma:

$$E[Y] = \mu = g^{-1}(\eta) = g^{-1}(X^t\beta)$$

Siendo:

- Y la variable objetivo
- $E(Y)$ es el valor esperado de la variable Y
- $\eta = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p = X^t\beta$ es el predictor lineal, se trata de una combinación lineal de parámetros desconocidos
 - X_1, \dots, X_p son las variables explicativas

- $\beta = (\beta_o, \beta_1, \dots, \beta_p)$ representan el efecto de cada variable independiente sobre la variable objetivo
- g es la función link, monótona y diferenciable

1.1.1. Componentes del Modelo Lineal Generalizado

En este tipo de modelización estadística podemos diferenciar tres componentes: la componente aleatoria, la sistemática y la función link o enlace. Será la combinación de estas tres componentes la que defina por completo un Modelo Lineal Generalizado.

1.1.1.1. Componente aleatoria

Esta componente es la que identifica la variable respuesta y su distribución de probabilidad.

Sea Y la variable aleatoria objetivo o variable respuesta objeto de estudio y sean las n variables aleatorias independientes e idénticamente distribuidas Y_1, \dots, Y_n la muestra aleatoria procedente de Y . Siendo Y la componente aleatoria cuya distribución pertenece a la familia exponencial de distribuciones.

En los MLG se supone que la variable respuesta Y se distribuye de tal forma que su función de probabilidad, en el caso de estar modelizando una variable discreta o función de densidad para el caso continuo viene dada por la siguiente expresión general, conocida como forma canónica:

$$f(y; \theta, \phi) = a(y, \phi) \cdot e^{\left(\frac{y\theta - k(\theta)}{\phi} \right)}$$

Donde

- θ es el parámetro canónico
- $k(\theta)$ es la función cumulante
- $\phi > 0$ se trata del parámetro de dispersión
- $a(y, \phi)$ es una constante normalizadora
- El soporte no depende de θ ni de ϕ

Además, la media de y es función del parámetro canónico θ , por tanto, se tiene que:

$$E(Y) = \mu = \frac{\partial}{\partial \theta} k(\theta), \text{Var}(Y) = \sigma^2 = \phi \frac{\partial^2}{\partial \theta^2} k(\theta) = \phi \frac{\partial}{\partial \theta} \left(\frac{\partial}{\partial \theta} k(\theta) \right) = \phi \frac{\partial}{\partial \theta} \mu > 0.$$

Es decir, μ es una función estrictamente creciente de θ , por lo que estos dos parámetros mantienen una relación biyectiva.

A

$$V(\mu) = \frac{\partial \mu}{\partial \theta}$$

se le denomina función varianza, por lo que se tiene que:

$$V(Y) = \phi \text{Var}(\mu)$$

1.1.1.2. Componente sistemática

Se trata de la coomponente que especifica las variables predictoras utilizadas en la función predictora lineal en forma de efectos fijos de un modelo lineal y recoge la variabilidad de la respuesta Y expresada a través de p variables explicativas X_1, \dots, X_p , que denotamos por X , y de los correspondientes parámetros desconocidos $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$.

Esta componente, también conocida como predictor lineal, viene representada por η y es una combinación lineal de las variables explicativas, que viene dada por: $\eta = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p = X^t \beta = X^t \beta$

1.1.1.3. Función link o función enlace

La *función link*: se trata de una función del valor esperado de la variable respuesta $E[Y]$, como una combinación lineal de las variables predictoras. Sin embargo, en muchos casos reales esta relación no es adecuada, por lo que es necesario la introducción de una función con el objetivo de relacionar el valor esperado con las variables explicativas. Por ello, introducimos la función link o función enlace, $g(\cdot)$ que relaciona μ con el predictor lineal de la siguiente forma:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

En problemas reales, pueden existir varias funciones link, por lo que se elegirá aquella que facilite la interpretación del modelo óptimo obtenido. En particular, para cada elemento de la familia exponencial existe una función enlace denominada función canónica, que permite relacionar el parámetro canónico con el predictor lineal.

$$\theta_i = \theta(\mu_i) = \eta_i = X_i^t \beta \quad g(\mu_i) = \theta(\mu_i)$$

1.1.2. Estimación de parámetros

Trás la construcción de los modelos, se estiman los parámetros desconocidos del predictor lineal, $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ por $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ y el valor del parámetro de dispersión ϕ por $\hat{\phi}$. Posteriormente, se valora la precisión de las estimaciones con el objetivo de seleccionar un modelo óptimo.

Generalmente, la estimación de los parámetros se lleva a cabo por el método de la *Máxima Verosimilitud* o el método de *Mínimos Cuadrados Ordinarios*. Una vez desarrollados los modelos, se realizará una comparación de los mismos con el objetivo de seleccionar el mejor. En el caso del modelado con fines predictivos, se seleccionará el modelo que explique el mayor porcentaje de variabilidad de la respuesta.

Para ello, emplearemos el **Criterio de información de Akaike (AIC)**, medida relativa de un modelo estadístico.

Dado un conjunto de modelos candidatos para los datos, el modelo preferido es aquel que tiene mínimo valor del AIC, que trata de proporcionar una compensación entre la bondad de ajuste del modelo y la complejidad del mismo. Es decir, el criterio penaliza al número de parámetros.

En el caso general, el AIC viene dado por la siguiente expresión:

$$AIC = 2k - 2\ln(\hat{L})$$

Siendo:

- k el número de parámetros del modelo
- \hat{L} es el máximo valor de la función de verosimilitud para el modelo estimado

Otro criterio en el que nos basaremos es en el criterio de bondad de ajuste, destacando el cálculo del **coeficiente de determinación** R^2 , que es una medida del grado de fiabilidad o bondad de ajuste del modelo ajustado a un conjunto de datos. Se trata de una medida acotada por definición, siendo sus límites $0 \leq R^2 \leq 1$. Un coeficiente de determinación igual a 1 indica un ajuste lineal perfecto, y por tanto, la variación total de la variable Y es explicada por el modelo de regresión. Por el contrario, el valor 0 indica que el modelo no explica nada de la variación total de la variable Y .

Para la bondad de ajuste, otra medida interesante es el RMSE, raíz del error cuadrático medio. Representa la raíz cuadrada de la distancia promedio entre el valor real y el pronosticado e indica el ajuste absoluto del modelo a los datos, es decir, cómo de cerca están los puntos observados de los valores predichos del modelo. Valores más bajos de RMSE indican un mejor ajuste.

En muchos casos la variable respuesta es de tipo conteo, como lo es la variable que queremos modelizar, demanda de productos. Se denominan variables de recuento o variables de tipo conteo, a aquellas que determinan el número de sucesos que ocurren en una misma unidad de observación en un intervalo espacial o temporal definido. Esta variable Y , puede tomar infinitos números de valores y su probabilidad va en descenso a medida que sea mayor el valor de la variable.

Para este caso, los modelos que tienen especial interés y que podemos formalizar a través de modelización lineal son el modelo de *Poisson* y el modelo de *Binomial negativa*. Estos modelos permiten analizar el comportamiento de variables de conteo frente a los valores del conjunto de variables explicativas.

1.1.3. Modelo de Regresión Poisson

Se trata del modelo más simple y es el modelo de referencia para variables respuesta de tipo conteo. Este modelo asume que la variable respuesta Y sigue una distribución de Poisson, por lo que en el caso de la modelización de ventas, se define como el número de ventas que ocurren en un intervalo de tiempo, cuya ocurrencia es aleatoria. Esta distribución se caracteriza por que la media y varianza coinciden:

$$E(Y) = Var(Y) = \mu$$

Se tiene que la distribución de probabilidad de Poisson, y en nuestro caso, la probabilidad de observar y ventas es:

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, \dots; \mu > 0$$

Y por tanto, la forma canónica o componente aleatoria para esta distribución es la siguiente:

$$f(y; \mu) = e^{-\mu} \cdot \frac{\mu^y}{y!} = \frac{1}{y!} e^{y \log(\mu) - \mu}, \quad y \in \{0, 1, \dots\}$$

Es decir, el modelo de Posición se obtiene tomando como función enlace el parámetro canónico.

donde:

- $\theta = \log(\mu)$ es el parámetro canónico
- $k(\theta) = \mu = e^\theta$ es la función cumulante
- $\phi = 1$ el parámetro de dispersión
- $a(y, \phi) = 1/y!$ la constante normalizadora

En este caso se tiene que: $g(\mu_i) = X_i^t \beta$ y una elección usual de la función link g el parámetro canónico, $g(x) = \log(x)$, lo que equivale a:

$$\mu_i = \exp(\beta_0) \cdot \exp(x_{i1}\beta_1) \dots \exp(\beta_p x_{ip}) = \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) \quad \text{ó} \quad \log(\mu_i) = \eta_i = X_i \beta$$

,

así si x_i se incrementa en una unidad, entonces μ_i se multiplica por $\exp(\beta_i)$. Por tanto, si $\beta_i > 0$, μ_i crece cuando x_i aumenta y si $\beta_i < 0$, μ_i decrece cuando x_i aumenta.

Este modelo se ha desarrollado suponiendo que la media y la varianza de los datos coinciden (equidispersión). Sin embargo, suele ocurrir lo que se conoce como sobredispersión, es decir, que la varianza es mayor que la media. Lo habitual es que esta situación se de debido a la existencia de heterogeneidad entre las observaciones. Cuando esto ocurra, recurriremos al modelo binomial negativo.

1.1.4. Modelo de Regresión de Binomial Negativa

Este modelo es empleado para variables de tipo conteo cuándo existe sobredispersión, es decir, la media condicional es menor que la varianza condicional (no coinciden). Existen diferentes modelos binomiales negativos en función de la variable que se trate de modelar, pero en este trabajo nos centraremos en el caso de datos de tipo conteo.

La distribución binomial negativa estudia la probabilidad de observar un número determinado de fracasos y (no se producen ventas), antes del r -ésimo éxito (se venden r unidades) en una serie de experimentos Bernoulli independientes, siendo r un número positivo. Se tiene que esta distribución pertenece a la familia exponencial si el parámetro de dispersión ϕ es una constante.

Se dice que la variable aleatoria de conteo (número de ventas) Y_i , con $i = 1, \dots, n$ sigue una distribución Binomial Negativa de parámetros r y p , y se representa como $Y_i \sim BN(r, p)$ si su función de probabilidad viene dada por:

$$P[Y_i = y_i] = \binom{y_i + r - 1}{r - 1} p^r (1 - p)^{y_i}$$

donde

- $0 < p < 1$

- $r > 0$
- $y_i = 0, 1, 2, \dots$

Y en este caso, la forma canónica o componente aleatoria para esta distribución es la siguiente:

$$f(y; \mu) = \exp \left\{ y \cdot \ln(1 - p) + r \ln(p) + \ln \left(\frac{y_i + r - 1}{r - 1} \right) \right\}$$

donde

- $0 < p < 1$
- $r = 0, 1, 2, \dots$
- $y_i = 0, 1, 2, \dots$
- $\theta = \log(1 - p)$ es el parámetro canónico
- $k(\theta) = -r \ln(p) = -r \ln(1 - e^\theta)$ es la función cumulante

En este caso la función link es de tipo logarítmico y viene dada por:

$$g(\mu_i) = \theta(\mu_i) = \ln \left(\frac{\alpha \mu_i}{1 + \alpha \mu_i} \right) = X_i^t \beta = \eta_i$$

1.2. Análisis de series temporales

Aplicaremos este modelo de predicción para tratar de identificar los patrones de la demanda anterior a lo largo del tiempo y luego proyectar (predecir) los patrones en el futuro.

Se define una serie temporal como una sucesión de datos ordenados en el tiempo que corresponden a una misma variable. Los datos son suelen ser tomados en intervalos regulares de tiempo.

Nuestro objetivo dentro del análisis de series temporales será identificar el proceso estocástico que ha sido capaz de generar la serie de estudio.

(No se si añadirlo) Se dice proceso estocástico a una colección o familia de variables aleatorias $\{X_t, \text{ con } t \in T\}$ que siguen la misma ley de distribución y están relacionadas entre sí, pudiendo por este motivo, describir la información de estas variables en términos de medias, variaciones y covarianzas.

A continuación encontramos las cuatro etapas en un análisis descriptivo de series temporales para elegir un modelo que se adecue a nuestros datos:

- **Representación gráfica de la serie.** Para tener así una primera aproximación del comportamiento de la serie y la existencia de posibles tendencias.
- **Modelización:** Se trata de encontrar el modelo que mejor se ajuste a los datos.
- **Validación de los modelos:** Es necesario saber si el modelo ajustado es adecuado o no, por lo que es muy importante el estudio de los residuos.
- **Predicciones:** Una vez construido y validado un modelo, realizaremos estimaciones del futuro con nuevas observaciones.

En un enfoque clásico de series temporales, asumiremos que el comportamiento de la variable con respecto al tiempo se compone de cuatro componentes:

1. **Tendencia:** Se trata del movimiento suave y regular de la serie a largo plazo. La tendencia existe cuando hay un aumento o disminución a largo plazo de los datos. Puede ser lineal (ajuste mediante una recta) o no lineal (aproximación mediante una curva, como por ejemplo logarítmica o exponencial)
2. **Ciclo:** Componente de tipo oscilante caracterizada por movimientos recurrentes en torno a la tendencia de la serie y que se repiten cada año pero sin una frecuencia fija.
3. **Componente estacional:** Se trata de movimientos regulares dentro de la serie con una periodicidad menor a un año, es decir, aquello que ocurre generalmente y con la misma intensidad año tras año en los mismos períodos, por ejemplo, en la misma época del año o día de la semana. Vamos a denotar por L al número de estaciones.
4. **Componente irregular:** Se trata de las variaciones de la serie sin un comportamiento sistemático y que no son explicadas por las otras tres componentes

Existen diferentes modelos de combinación de las componentes. Para describir los modelos necesitamos primero una nomenclatura básica. Denotando por X_t al valor de la variable en el instante t , se tiene:

$$X_t = f(T_t, E_t, I_t)$$

donde:

- T_t : Valor de la tendencia en el instante t
- E_t : Valor de la componente estacional en el instante t
- I_t : Valor de la componente irregular en el instante t (ruido).

Por tanto, los modelos que puede adoptar la función f son los siguientes:

- **Modelo multiplicativo:** La composición de la serie se realiza mediante el producto de sus componentes.

$$X_t = T_t \times E_t \times I_t$$

- **Modelo aditivo:** Las componentes se agregan para formar la serie temporal.

$$X_t = T_t + E_t + I_t$$

- **Modelo mixto:** La composición de la serie de la parte irregular viene de forma aditiva y la parte regular de forma multiplicativa.

$$X_t = T_t \times E_t + I_t$$

<http://www5.uva.es/estadmed/datos/series/series2.htm>

Tras haber detectado el modelo mas adecuado, podremos conocer el comportamiento de la serie a largo plazo.

El siguiente paso realizar una estimación de la tendencia, T_t , habiendo eliminado previamente la componente estacional para impedir que estas oscilaciones perturben la identificación de la tendencia.

Para estimar T_t , debemos hacer una hipótesis sobre su forma:

- **Tendencia determinista:** Se supone que la tendencia es una función determinística del tiempo:

$$T_t = a + bt \quad a, b \in \mathbb{R}$$

Siendo a y b constantes, que se estimarán mediante un modelo de regresión lineal.

Sin embargo, el método que aplicaremos será el que exponemos a continuación:

- **Tendencia evolutiva (método de medias móviles):** Este método consiste en definir la tendencia como una serie suavizada. Supondremos que la tendencia de la serie es una función que evoluciona lentamente y que podremos aproximar función simple del tiempo, suponiendo así una recta.

Una vez identificada la tendencia, procedemos a hacer un análisis de la estacionalidad de la serie, con el objetivo de:

- **Desestacionalizar la serie**, es decir, eliminar las oscilaciones periódicas que se repiten a lo largo de los años, haciendo así que los datos de distintas estaciones sean comparables. La serie desestacionalizada la conseguimos diferenciando la serie.
- **Realizar predicciones**, ya que si nuestros datos están afectados por una componente estacional, necesitaremos una estimación de esta de cara a realizar una predicción

Para desestacionalizar la serie, emplearemos los índices de variación estacional asociados a cada estación, ya que se suponen constantes año a año. Con esta técnica, se evidencian las diferencias en cada período, por ejemplo, podemos ver la diferencia del volumen de ventas en función de la época del año (mes, día de la semana, estación,...) Estos índices reflejan la cantidad fija o proporción en la que se modifica la tendencia en cada estación.

Una vez calculados estos índices, se desestacionaliza la serie, eliminando así el efecto de cada estación.

Por último, procedemos a realizar las predicciones. Para ello, necesitamos que se cumpla la condición de estacionariedad, es decir, la media y la varianza permanecen constantes en el tiempo (no tiene raíces unitarias). En el caso de no imponer esta condición de estacionariedad, predeciríamos características que no serán las mismas en el futuro que en el pasado.

Se tiene que todo proceso lineal es estacionario, por tanto, obtendremos trabajaremos con series estacionarias, y de lo contrario, podremos aplicar los mismos métodos a series no estacionarias realizando las transformaciones pertinentes para conseguir la estacionariedad.

En nuestro caso, aplicaremos la metodología Box-Jenkis como método predictivo.

1.2.1. Metodología Box-Jenkis

Esta metodología tiene en cuenta la dependencia existente entre los datos, es decir, cada observación en el instante t será modelada a partir de los valores pasados. Los modelos se conocen con el nombre de ARIMA (modelos integrados autorregresivos de medias móviles), que deriva de las siguientes componentes: AR (Autorregresivo) , I (integrado), MA(Medias móviles)

El siguiente paso es identificar el modelo más adecuado a través del estudio de la función de autocorrelación (FAC) y la función de autocorrelación parcial (FAP).

Nota: el método recomienda como mínimo 50 observaciones en la serie temporal.

Fases de la metodología Box-Jenkis:

1. Identificar el la estructura ARIMA que sigue la serie a través del estudio de la función de autocorrelación simple (FAS) y la función de autocorrelación parcial (FAP). Determinar el modelo arima consiste en identidicar los órdenes p y q de su estructura autoregresiva y de medias móviles
2. Estiamción de parámetros: Una vez tenemos identificado el modelo, estimamos los parámetros AR y MA del modelo por el método de máxima verosimilitud, obteniendo el error estándar y los residuos del modelo

Nota: Es muy importante comprobar que las estimaciones son significativamente no nulas.

3. Diagnósis del modelo: Comprobamos que los residuos sigan un proceso de ruido blanco mediante el Test de Ljung-Box.

Si hemos identificado varios modelos y todos ellos pasan la diagnósis, nos quedaremos con uno de ellos según el criterio del menor AIC

4. Predicción: una vez identificado y validado el mejor modelo, se realizan las predicciones con éste.