Métodos de regresión lineal generalizada

UCO ONLINE

Métodos de regresión lineal generalizada

UNIVERSIDAD D CÓRDOBA

1. Introducción

Una vez extendido el modelo de regresión lineal simple a múltiples variables de entrada, cabe preguntarse si es posible extender otros aspectos del modelo lineal. En esta lección, se estudian otras formas de relacionar las variables de entrada para lograr un mejor ajuste a los datos. Además, el modelo lineal simple asume ciertas propiedades en la distribución de los datos y de las propias variables modeladas, restricciones que también pueden "relajarse" para cubrir situaciones más realistas.

Como primera forma de extensión del modelo lineal, se estudiarán dos posibilidades. En primer lugar, se explicará la eliminación de la componente aditiva del modelo. En segundo lugar, se abordará la incorporación de funciones de transformación no lineal en las variables de entrada, ejemplificado en el caso de la regresión polinómica.

A continuación, la lección se centrará en la *regresión lineal generalizada* (GLM, ***generalized linear model***). Se trata de una ampliación del modelo lineal múltiple, donde la variable dependiente continúa respondiendo a una combinación lineal de las variables de entrada, pero variando la función que las relaciona y permitiendo que los errores no tengan por qué seguir una distribución normal. En este caso, podemos obtener modelos basados en otras distribuciones de probabilidad (Poisson, Gamma, etc.). Realmente, los métodos vistos hasta ahora -englobados bajo el término "regresión lineal ordinaria"- son casos particulares de GML, pues pueden formularse bajo sus mismos supuestos. Desarrollada por J.A. Nelder, W.M. Wedderburn y P. McCullagh, GLM no es más que una unificación de los distintos métodos de regresión que existían hasta el momento.

Finalmente, también se estudiará el caso particular de la regresión logística, donde la variable dependiente es binaria. Este tipo de modelo puede verse como una primera forma básica de clasificación y, de hecho, existen métodos de clasificación que internamente se basan en ajustar este tipo de modelo de regresión.

UCO ONLINE

2. Extensiones básicas del modelo lineal

2.1. Dependencia entre variables de entrada

En el modelo lineal múltiple ordinario, cada variable de entrada tiene asociado un coeficiente propio. Esto se traduce en que cada coeficiente β_i nos indica cómo varía y (en media) ante el aumento de una unidad de la variable asociada a dicho coeficiente (x_i). Dicha variación solo depende de x_i, en otras palabras, a la hora de explicar la variación en la variable de salida estamos asumiendo una independencia entre las variables de entrada. No obstante, puede que un aumento en una variable x_i implique que otra variable x_j deje de comportarse igual que si x_i no estuviera en el modelo. Este concepto se conoce como *efecto de interacción* entre variables. Una forma de incluir este fenómeno en el modelo de regresión es permitir que exista un término que vincule a las variables, con su correspondiente coeficiente. Por ejemplo, en un modelo con dos variables de entrada, un tercer término puede ser el producto de ambas:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon \quad (1)$$

La idoneidad de introducir este tipo de término de interacción puede depender del problema que estemos resolviendo. Una forma de saber si es necesario o no es obtener los *p-values* asociados a cada coeficiente. Recordemos que dichos valores nos indican si debemos o no aceptar la hipótesis nula de que el coeficiente es igual a cero. Es decir, si la variable de entrada asociada a él tiene relevancia o no. Habitualmente, introducir este tipo de término no debe implicar la desaparición de los coeficientes asociados a cada variable por separado. También podemos estudiar el valor del estadístico R^2 para comprobar si el modelo con el nuevo término nos proporciona un mejor ajuste que el modelo sin dicho término.

2.2. Regresión polinómica

Una segunda modificación que podemos plantear en el modelo de regresión lineal es utilizar funciones no lineales sobre las variables de entrada, pero manteniendo la linealidad respecto a los coeficientes. El caso más habitual es transformar la recta de regresión en una curva mediante una expresión polinómica como la siguiente:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon \quad (2)$$

En este caso, conseguimos que la curva tenga una forma cuadrática, pero es extensible a polinomios de grado superior. A diferencia del apartado anterior, podemos suponer la idoneidad de este tipo de curva de ajuste inspeccionando visualmente la distribución de los datos. Si apreciamos que la tendencia es curva, posiblemente necesitemos utilizar regresión polinómica. Por otro lado, incrementar el número de términos puede suponer un sobreajuste, por lo que su uso debe analizarse en detalle.

3. Regresión lineal generalizada

3.1. Formulación y componentes

La suposición de partida de GLM es cada valor estimado de y se obtiene a partir de las variables de entrada siguiendo una distribución de probabilidad de una familia determinada. Concretamente, la media de la distribución es el factor que depende de la combinación lineal de los predictores. Entre las familias más habituales podemos encontrar la normal, binomial, de Poisson, Gamma, etc. Al utilizar un método de regresión, lo que estaremos estimando realmente son los parámetros de dicha distribución (su media, varianza, forma, etc.). De forma genérica, podemos expresar un modelo de regresión como se muestra a continuación:

$$E(Y) = g(\beta X) \tag{3}$$

Para evitar confusión con el modelo lineal ordinario, utilizaremos letras mayúsculas para expresar que: 1) Y no es el conjunto de valores estimados en la muestra, sino una variable que sigue una distribución de tipo exponencial, y 2) X es una transformación (lineal o no lineal) de los predictores originales. En esta nueva formulación, los componentes que participan son los siguientes:

- Componente aleatoria, $E(Y)$: representa la probabilidad de distribución de la variable de respuesta, esto es, cómo varía el error en la variable de salida.
- Componente sistemática o predictor lineal, βX: explican la forma en la que se relacionan las variables de entrada. Sigue siendo una combinación lineal de predictores.
- Función de enlace, g: es la que determina cómo se relaciona la variable de salida con las variables de entrada.

Para entender mejor esta nueva formulación, vamos a comprobar que el modelo lineal ordinario es un caso particular de GLM. Para ello, debemos descomponer el modelo lineal ordinario en cada uno de los componentes anteriores para comprobar que, efectivamente, puede expresarse de tal forma:

- La distribución normal o Gaussiana pertenece a la familia de funciones exponenciales, por tanto: $E(Y) \sim N(\mu, \sigma^2)$
- El predictor lineal es la suma de los coeficientes que multiplican a cada predictor: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$.
- La función de enlace es la función unidad: $g(E(Y)) = E(Y)$.

3.2. Distribución de Poisson

Esta distribución es adecuada cuando la variable de salida representa un conteo, esto es, el número de eventos que ocurre un fenómeno. Se trata de una distribución unimodal, donde la media y la varianza coinciden. Cuando se modela GLM con la distribución de Poisson, la función de enlace es el logaritmo, de forma que se expresa como:

$$\log\big(E(Y)\big) = \beta_0 + \sum_{i=1}^{p} \beta_p x_p \tag{4}$$

El hecho de que la función de enlace sea el logaritmo garantiza que la variable de salida tome valores positivos, lo cual es adecuado cuando queremos modelar un conteo o una ratio.

3.3. Distribución Gamma

Otra distribución de probabilidad que puede utilizarse en GML es la distribución Gamma. Es similar a la de Poisson en el sentido de que la variable de salida solo toma valores positivos (en este caso, excluyendo el cero). No obstante, dichos valores son continuos a diferencia de la distribución de Poisson, donde son discretos. Un ejemplo de variable que puede modelarse con una distribución Gamma es el coste de un producto o servicio. Otros ejemplos son el modelado del riesgo o de una inversión.

El modelo GML que sigue una distribución Gamma tiene la siguiente formulación:

$$\mathrm{E(Y)} = e^{\beta_0} x_1^{\beta_1} \ldots x_p^{\beta_p} \tag{5}$$

3.4. Regresión logística

La regresión logística es un caso especial de regresión con aplicaciones en clasificación. Su característica principal es que la salida modela la probabilidad de que la variable dependiente tome un valor entre un número limitado de opciones. Es decir, la variable de salida es categórica, lo cual nos recuerda a un problema de clasificación. El modelo más común permite diferenciar dos valores de salida, esto es, la variable es binaria. Algunos ejemplos de este tipo de problema son la probabilidad de éxito o fracaso de una solicitud, o la probabilidad de padecer o no una enfermedad.

Para modelar este tipo de problema, necesitamos una función que nos proporcione valores entre 0 y 1: la función logística (también conocida como sigmoide). Matemáticamente, el modelo de regresión logística toma la siguiente forma:

$$E(Y) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} \quad (6)$$

Este modelo es extensible a variables de salida que toman más de un valor (lo que asociaríamos a un problema multi-clase), así como a variables de salida que representan un orden o escala (por ejemplo, la puntuación de una película).

3.5. Estimación de máxima verosimilitud

En regresión lineal ordinaria podemos utilizamos el método de los mínimos cuadrados para estimar los coeficientes y ajustar así la recta de regresión. El equivalente en GML son los métodos de *estimación de máxima verosimilitud* (MLE, *maximum likelihood estimation*). Como se explicó anteriormente, el objetivo en GML es estimar los parámetros de la distribución de probabilidad, utilizando para ello la muestra de observaciones. En otras palabras, tratamos de encontrar los parámetros que hacen que la distribución represente lo más fielmente la muestra de datos disponible.

Un método MLE debe optimizar una función, la llamada función de verosimilitud, que es específica según el tipo de distribución que estemos modelando. El método de los mínimos cuadrados es realmente una particularización de MLE, donde la solución por lo general es única. Sin embargo, para otras distribuciones no es posible obtener la solución de manera exacta, por lo que muchos métodos MLE son procedimientos iterativos que tratan de optimizar la función de verosimilitud. Los dos métodos MLE más conocidos son el método de gradiente descendiente y el método Newton-Rapshon.

Referencias

B. Caffo. “Regression Models for Data Science in R”. Leanpub (CCA-NC 3.0), 129 páginas. 2015. Disponible en: https://github.com/bcaffo/regmodsbook

G. Hackeling. “Mastering Machine Learning with Scikit-Learn”. Pactkt Publishing, 221 páginas. 2014.

T. Hastie, R. Tibshirani, J. Friedman. “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”. Springer Series in Statistics, 2ª edición, 745 páginas. 2017.

G. James, D. Witten, R. Tibshirani, T. Hastie. “An Introduction to Statistical Learning with Applications in R”. Springer Texts in Statistics, 1ª edición (7ª impresión), 426 páginas. 2017. Disponible en: https://www.statlearning.com/

B. Lantz. “Machine Learning with R”. Packt Publising, 396 páginas. 2013.

P. McCullagh, J.A. Nelder. “Generalized Linear Models”. Chapman & Hall/CRC, 2ª edición, 532 páginas. 1989.

D. Peña. “Regresión y diseño de experimentos”. Alianza Editorial, 744 páginas. 2010.