Problemas de regresión múltiple

UCO ONLINE

Problemas de regresión múltiple

1. Introducción

Hasta el momento hemos estudiado la regresión lineal simple, donde la variable dependiente es estimada en base a una única variable independiente. No obstante, existen situaciones en las que la variable de salida no responde únicamente a variaciones de una variable de entrada, sino que puede verse influenciada por varias de ellas. De hecho, es lo habitual si pensamos en los ejemplos planteados al comienzo de la semana anterior:

- El precio de una vivienda suele crecer a medida que lo hace su superficie, pero no es el único factor que influye en el precio. Otros aspectos que pueden influir son la antigüedad del inmueble, la distribución de habitaciones, la ciudad o barrio donde se encuentra, etc.
- Tendemos a pensar que la duración de un proyecto de ingeniería va a reducirse conforme más personas participan en él. Sin embargo, la falta de experiencia de esas personas puede repercutir negativamente, igual que el hecho de que esas personas pertenezcan o no a una empresa subcontratada, ya que implica más dificultad en la comunicación, tiempos de respuesta, etc.
- La tasa de paro varía según los segmentos de edad, pero también pueden encontrarse diferencias según el nivel de estudios, la procedencia geográfica o el sector productivo al que nos refiramos.

Intentar explicar la variación en el precio de una vivienda, la duración de un proyecto o la tasa de paro solo en base a un criterio no parece del todo realista. Si disponemos de más variables de interés, podemos construir modelos de *regresión múltiple*, donde la variable dependiente responda a variaciones en más de una variable de entrada. Cuando consideramos más variables de entrada, aparecen nuevos conceptos como la *correlación* entre variables. Por otro lado, incrementar el número de variables de entrada hace que el problema de regresión se vuelva más complejo, ya que a cada variable se le debe asociar un coeficiente y, posiblemente, necesitemos expresiones más complejas para ajustar la relación entre las variables. Esto también repercute en la interpretación del modelo de regresión resultante, ya que tenemos más coeficientes que analizar.
A lo largo de esta semana, se estudiará la definición de un problema de regresión múltiple y los métodos de regresión que pueden utilizarse para su resolución. En concreto, en esta lección se extiende la definición del modelo de regresión lineal a varias variables y se presenta la generalización del método de los mínimos cuadrados para estimar los coeficientes.

2. Regresión lineal múltiple

El modelo lineal para regresión múltiple, también llamado regresión multivariable, es una extensión natural al modelo de regresión lineal simple. En lugar de tener dos coeficientes (intercepto y pendiente), tendremos $p + 1$ coeficientes, siendo p el número de variables de entrada (también llamados predictores). Formalmente, el modelo se expresa como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \quad (1)$$

En la ecuación anterior, la combinación entre las variables de entrada sigue siendo lineal, ya que se expresa como una suma de términos. Sin embargo, este modelo lineal no se corresponde con una recta, sino con un plano de tantas dimensiones como variables de entrada se tengan.
La interpretación del modelo de regresión múltiple es equivalente a la del modelo de regresión simple. El coeficiente β_j simboliza el efecto medio que tiene la variable x_j sobre y, suponiendo que todas las demás variables son fijas. Los coeficientes pueden ser positivos o negativos, indicando que influyen en el crecimiento o decrecimiento de la variable y. Igualmente, existe un valor constante (el intercepto, β_0) y un error (ε) asociados al modelo.
Algunos posibles modelos de regresión múltiple para los ejemplos que acompañan a la lección serían los siguientes:

$$precio\_vivienda \approx \beta_0 + \beta_1 superficie\_vivienda + \beta_2 número\_habitaciones \quad (2)$$

$$duración\_proyecto \approx \beta_0 - \beta_1 personas\_equipo + \beta_2 número\_subempresas \quad (3)$$

$$tasa\_paro \approx \beta_0 - \beta_1 edad - \beta_2 nivel\_estudios \quad (4)$$

En el primer caso, estamos expresando que el precio de la vivienda será mayor cuanta más superficie ocupe la vivienda y más habitaciones tenga. En el segundo ejemplo, la duración del proyecto se reduce conforme más personas forman el equipo, pero a su vez aumenta según el número de subempresas involucradas. Finalmente, la tasa de paro decrece a medida que aumenta la edad y el nivel de estudios.

3. Extensión del método de los mínimos cuadrados

El método de los mínimos cuadrados nos permitía estimar los valores de los dos coeficientes para un problema de regresión simple, minimizando para ello la suma de los residuos al cuadrado. Esta misma idea es aplicable al problema de regresión múltiple, ya que se compone de los mismos elementos: coeficientes a estimar y un error que queremos minimizar. Podemos expresar la relación entre los coeficientes a estimar y la variable a estimar de la siguiente manera:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{5}$$

En este caso, la suma de los residuos al cuadrado (RSS) se formula como sigue:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \tag{6}$$

Para el caso $p > 1$, el desarrollo matemático para obtener las fórmulas de estimación de los coeficientes con el método de los mínimos cuadrados se vuelve más complejo. Para explicarlo, es habitual utilizar una notación matricial. Por tanto, primero se debe expresar la ecuación del problema de regresión de la siguiente manera:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{7}$$

De forma abreviada, escribimos la ecuación anterior como:

$$Y = X\hat{\beta} + E \tag{8}$$

Buscamos que el error cuadrático sea mínimo, lo cual es equivalente a:

$$\min \sum_{i=1}^{n} \varepsilon^2 = \min \|E\|^2 \tag{9}$$

Para minimizar $\|E\|^2$, el vector E debe ser perpendicular al espacio vectorial que generan las columnas de la matriz X, lo cual se expresa:

$$X^T E = 0 \rightarrow \begin{cases} \sum_i^n \varepsilon_i = 0 \\ \sum_i^n \varepsilon_i x_{1i} = 0 \\ \vdots \\ \sum_i^n \varepsilon_i x_{pi} = 0 \end{cases} \tag{10}$$

Por simplicidad, se omite aquí el desarrollo completo del método. A partir del supuesto anterior, el vector de coeficientes se obtiene por medio de la siguiente expresión:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{11}$$

Donde T representa la operación de transposición de una matriz, y -1 indica la operación de inversión.

Una diferencia importante respecto a la regresión simple, es que puede no existir una solución única al problema de minimización. Esto es, no existe un único vector de coeficientes que minimice la suma de los residuos al cuadrado. Este caso sucede cuando la matriz resultante de $X^T X$ es singular. El caso más habitual ocurre cuando las variables están correlacionadas, aspecto que se estudiará en más detalle en el apartado siguiente.

Finalmente, conviene destacar que, al igual que en regresión lineal simple, la estimación de los coeficientes en regresión múltiple lleva asociado un error estándar a cada coeficiente. A partir de ellos, se puede calcular el intervalo de confianza, el estadístico t y los *p-values*.

4. Correlación entre variables

Ahora que se ha estudiado una formulación más general del problema de regresión, es conveniente detenerse en el concepto de *correlación* y en sus implicaciones. Decimos que existe correlación entre dos variables, a y b, cuando es posible encontrar una relación de dependencia estadística entre esas dos variables. Aunque dicha relación puede ser de cualquier tipo, es habitual asumir que nos referimos a que existe una relación lineal entre las variables. Si existe correlación entre dos variables, es fácil imaginar que podremos utilizar una variable para predecir el valor de la otra. No obstante, hay que tener cuidado a la hora de interpretar un coeficiente de correlación. Que exista correlación entre variables no nos asegura que exista una relación de causa-efecto entre ellas. En otras palabras: *"Correlación no implica causalidad"*. La relación causal puede deberse a otros fenómenos que no estamos midiendo.

En el ámbito de la regresión, podemos pensar en dos tipos de relaciones entre variables. Por un lado, entre la variable dependiente y la variable independiente. En este caso, que exista correlación entre ellas es deseable, ya que tenemos más garantías de que el modelo de regresión es fiable. De hecho, el estadístico R^2 se basa en el concepto de correlación. Por otro lado, puede existir correlación entre las variables de entrada de un problema de regresión múltiple. Por ejemplo, imaginemos que, en el ejemplo de la predicción del precio de la vivienda, la superficie de la vivienda estuviese dividida en un número de habitaciones de igual dimensión. Podríamos expresar dicha relación como: $superficie\_vivienda = 25m^2 \cdot número\_habitaciones$. En este caso, las variables están perfectamente correlacionadas, ya que una se expresa en base a la otra. A la hora de obtener un modelo de regresión, incluir ambas variables carece de sentido. Algunos métodos de regresión son capaces de detectar este tipo de correlaciones y excluir variables "redundantes" del modelo de regresión.

La forma más habitual de comprobar si dos variables están correlacionadas es calcular el *coeficiente de correlación de Pearson*. Este coeficiente mide la magnitud de la relación lineal entre las dos variables dada una muestra, y se calcula en base a la covarianza y la varianza de dicha muestra:

$$\rho(a,b) = \frac{cov(a,b)}{\sigma_a \sigma_b} = \frac{\sum_{i=1}^{n}(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i-\bar{a})^2 \sum_{i=1}^{n}(b_i-\bar{b})^2}} \quad (12)$$

El coeficiente ρ varía entre -1 y 1. Valores absolutos más próximos a 1 indican un mayor grado de correlación. Si $\rho = 0$, no existe relación lineal entre las variables. Además, el signo nos informa de cómo es la relación entre variables. Si el coeficiente es positivo, el crecimiento de una variable implica el crecimiento de la otra. Si es negativo, una variable decrece a medida que la otra crece.

Aunque debe valorarse en cada contexto de aplicación, la interpretación del coeficiente ρ suele hacerse en base a rangos (equivalentes para valores negativos) como los siguientes:

- Valores entre 0.1 y 0.3 indican una correlación débil.
- Valores entre 0.3 y 0.5 indican una correlación moderada.
- Valores superiores a 0.5 indican una correlación fuerte.

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.

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