Regresión lineal simple

UCO ONLINE

Regresión lineal simple

1. Introducción a los problemas de regresión

En la lección anterior se presentaron varias situaciones en las que podemos aplicar métodos predictivos para intentar estimar o inferir la ocurrencia de un determinado evento a partir de un conjunto de datos. A continuación, se presentan algunos ejemplos adicionales que nos servirán para ilustrar los conceptos de esta lección:

- Una empresa inmobiliaria necesita tasar el precio de una vivienda, para lo cual decide analizar ciertas características de viviendas similares, tales como la superficie en metros, la antigüedad del inmueble, el número de habitaciones y baños, etc. ¿Podemos predecir el valor de la vivienda en base a alguno de estos factores?
- Una empresa que desarrolla proyectos de ingeniería tiene que elaborar un plan de proyecto que incluya la duración estimada del mismo. La empresa cuenta con un historial de proyectos sobre los que puede extraer la siguiente información: número de personas que conforman el equipo, años de experiencia del gestor del proyecto, número de proyectos anteriores de la misma temática y número de subempresas contratadas. ¿Podemos utilizar esta información para estimar la duración del nuevo proyecto?
- Un instituto de estudios demográficos desea analizar cómo afecta la tasa de paro a distintos grupos de población. Para ello, recopila mediante encuestas telefónicas datos personales y laborales de una muestra aleatoria de la población, tales como la edad, los años de experiencia laboral previa, etc. ¿Cuál de estos factores tiene mayor influencia a la hora de explicar la tasa de paro que sufre la población?

En todos estos ejemplos se pretende estimar un valor de una variable *continua* (el precio de la vivienda, la duración del proyecto, o la tasa de paro) en base a *una o más variables* conocidas (las características de las viviendas, de los proyectos o de las personas). En todos estos ejemplos, estas características son también variables numéricas, aunque también podrían incluirse otras no numéricas como el nivel de estudios o la ciudad de residencia en el caso del cálculo de la tasa de paro. No obstante, en esta primera aproximación a los problemas de regresión trabajaremos únicamente con variables de entrada numéricas.

De la lección anterior sabemos que los métodos de regresión se caracterizan por ser capaces de predecir el valor de una variable *continua*. En esta lección se define formalmente qué es la regresión lineal con una variable, comúnmente conocida como *regresión lineal simple*.

2. Regresión lineal con una variable

La regresión lineal simple consiste en encontrar una función lineal que describe cómo el valor de una única variable, *x,* influye sobre otra variable, *y*. Formalmente, la relación entre *x* e *y* se expresa:

$$y = f(x) = \beta_0 + \beta_1 x + \varepsilon \quad (1)$$

Cuando la relación se establece como una suma de términos, decimos que la relación es *lineal*. En este caso, además, la variable x no sufre ninguna transformación, sino que se toma su valor directamente para obtener y. No obstante, si sufriera alguna transformación, que podemos expresar como $\phi(x)$, el modelo seguiría siendo lineal, pues la propiedad de linealidad se define en base a los términos que componen la función.
El objetivo de un método de regresión es determinar los valores β_0 y β_1 que mejor aproximan la función dados pares de valores (x, y). Además, en la ecuación anterior se incluye otro término ε , que representa el error cometido. Es posible encontrar formulaciones en las que no se modela explícitamente ese error. En tal caso, la Ecuación 1 debe reformularse como sigue:

$$y \approx \beta_0 + \beta_1 x \quad (2)$$

o

$$\hat{y} = \beta_0 + \beta_1 x \quad (3)$$

En la Ecuación 2, expresamos que y es *modelada de forma aproximada* a partir de x. En la Ecuación 3, estamos representando que el resultado de aplicar la ecuación sobre un valor de x es en realidad una *estimación de y* ($\hat{y}$). Basándonos en los ejemplos presentados al comienzo del documento, podríamos definir algunos modelos de regresión como:

$$precio\_vivienda \approx \beta_0 + \beta_1 superficie\_vivienda \quad (4)$$

$$duración\_proyecto \approx \beta_0 - \beta_1 personas\_equipo \quad (5)$$

$$tasa\_paro \approx \beta_0 - \beta_1 edad \quad (6)$$

En cualquiera de los tres casos, es lógico pensar que el valor de la variable a estimar no depende de un único factor, sino de varios. En la segunda semana de curso se extenderá esta definición para incluir un número mayor de variables. De momento, en el siguiente apartado nos detendremos un poco más en qué representa cada uno de los elementos que componen la ecuación de regresión.

3. Elementos de un modelo lineal

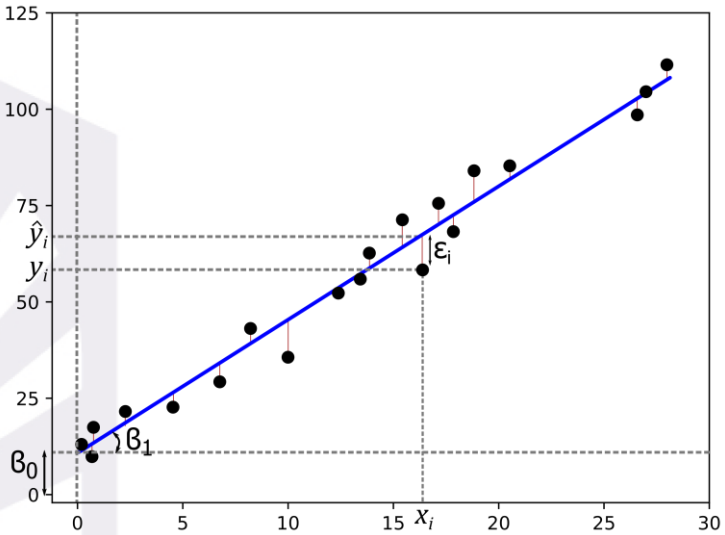

Figura 1. Ejemplo de un modelo de regresión lineal simple

La Figura 1 representa gráficamente el concepto de regresión lineal, y nos va a permitir analizar el significado de cada uno de los elementos que aparecen en la Ecuación 1. En la figura, tenemos un conjunto de observaciones (x, y), representadas como puntos. x es la variable *independiente* (también llamada explicativa), mientras que y es la variable *dependiente* (también conocida como variable respuesta), pues su valor se obtiene a partir de x. La recta trazada representa la función de regresión, caracterizada por dos *coeficientes:* β_0 es el *intercepto* y β_1 es la *pendiente*. β_0 representa el valor que alcanza y cuando $x = 0$, esto es, el punto en el que la recta corta al eje de ordenadas. β_1 es la pendiente de la recta respecto al eje de abscisas, e indica cuánto varía y cada vez que x se incrementa en una unidad. Cuando $\beta_1 > 0$, la relación lineal es positiva, es decir, a medida que aumenta x también aumenta y. Por el contrario, si $\beta_1 < 0$, la relación lineal es negativa, de forma que aumentar el valor de x implica decrementar el valor de y. Finalmente, cada uno de los puntos por los que pasa la recta nos da el valor $\hat{y}$, esto es, la estimación de y dado un valor de x. Idealmente, existiría una recta que pasase por todos los puntos, de forma que $\hat{y} = y$ para todo valor de x. Sin embargo, esto difícilmente ocurrirá en la realidad. Teóricamente, podemos trazar infinitas rectas que pasen por la nube de puntos, posiblemente dejando algunos puntos por encima y otros por debajo, como ocurre en la Figura 1. Como veremos más adelante, nos interesará encontrar aquella recta que se acerca lo más posible a todos los puntos, ya que es la que mejor se aproxima a la distribución de las observaciones (x, y).

Consideremos que la recta representada por la línea azul es la mejor función de regresión posible. Dado un punto (x_i, y_i), la distancia entre el punto y la recta se denomina *residuo*, el cual es obtenido como:

$$\varepsilon_i = y_i - \hat{y}_i \quad (7)$$

4. Hipótesis del modelo de regresión lineal simple

El modelo de regresión lineal es bastante simple en su formulación, pero eso no significa que su estudio o aplicación carezca de interés. Este tipo de modelos son muy utilizados en estadística y su capacidad de predicción puede ser superior a la de modelos más complejos, sobre todo cuando se dispone de pocos datos. La posibilidad de aplicar transformaciones a las entradas incrementa la capacidad de aplicación de los modelos lineales y, además, son esenciales para entender los modelos no lineales. No obstante, el modelo de regresión lineal simple parte de ciertas hipótesis respecto a la distribución de los datos que es necesario conocer. Si no se pueden asumir como ciertas, el modelo lineal posiblemente no sea el más adecuado para tratar de explicar los datos.

Linealidad

Como el propio nombre indica, el modelo de regresión lineal establece una relación lineal entre las variables x e y. Si tal relación no es la que existen en realidad entre los datos, el modelo resultante no podrá ser muy preciso.

Normalidad

Se asume que la distribución de los errores cometidos por el modelo lineal sigue una distribución normal: $\varepsilon_i \sim N(0, \sigma)$. σ es otro parámetro a estimar en el modelo lineal.

Homogeneidad

Relacionada con la anterior, esta propiedad establece que el valor promedio del error cometido por el modelo lineal es igual a cero. Como vimos antes, la recta de regresión debe "cruzar" la nube de puntos, cometiendo errores por exceso y por defecto en una magnitud similar.

Homocedasticidad

También relacionada con la distribución de los errores, esta propiedad nos dice que la varianza de los errores debe ser constante. Esta propiedad es difícil de cumplir, pues a menudo la varianza del error tiende a crecer conforme lo hace el valor de y.

Independencia
Los errores no están correlacionados, esto es, el residuo obtenido en un punto no guarda relación con los residuos de los puntos a su alrededor. El caso contrario ocurre, por ejemplo, cuando tenemos observaciones que corresponden a una serie temporal.

Finalmente, otro aspecto que influye notablemente en los modelos lineales es la presencia de valores atípicos (*outliers*), ya que su residuo asociado puede desviar notablemente el cálculo de las medidas de rendimiento del modelo lineal. Este tipo de medidas se estudiarán a final de esta semana.

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