Evaluación de modelos de regresión

UCO ONLINE

Evaluación de modelos de regresión

UNIVERSIDAD Ð CÓRDOBA

1. Metodología de evaluación en regresión

En la lección anterior se ha estudiado un primer método de regresión, el método de los mínimos cuadrados, para ajustar una curva de regresión con una única variable. Antes de introducir otros métodos más avanzados (objetivo de la siguiente semana de curso), es conveniente conocer cómo podemos analizar el rendimiento de cualquier método de regresión. En esta lección se aborda la evaluación de los métodos de regresión desde dos perspectivas: 1) cómo de fiable es el método en cuanto a la estimación de los coeficientes, y 2) cómo de bueno es el modelo de regresión obtenido a la hora de predecir nuevos valores de la variable dependiente.

Antes de entrar en cada uno de los puntos anteriores, es importante recalcar una serie de aspectos metodológicos a tener en cuenta. Un método de regresión requiere de una muestra de los datos para ajustar la curva de regresión. La elección de la muestra es, por tanto, un punto crítico. Los modelos de regresión, al igual que otros modelos de aprendizaje supervisado, deben exhibir una buena capacidad de *generalización*. En regresión, este concepto alude a que el modelo sea capaz de realizar buenas predicciones con independencia de la muestra utilizada para realizar el ajuste. La generalización de un modelo está ligada a dos conceptos fundamentales:

a. Sobreajuste (*overfitting*): la curva de regresión es muy precisa a la hora de ajustar las observaciones que conforman la muestra, pero no predice correctamente valores fuera de esa muestra. Por ejemplo, si utilizamos solo una serie de puntos cercanos en el espacio (valores muy próximos de x), se corre el riesgo de que la curva ajustada se comporte muy bien en ese intervalo, pero no fuera de él.
b. Desajuste (*underfitting*): la curva de regresión no es capaz de predecir la muestra, y tampoco responde correctamente ante valores fuera de ella. En este caso, es posible que el tipo de curva que se esté ajustando no sea el adecuado para la distribución real que presentan los datos.

Existen diferentes estrategias para tratar de mitigar ambos fenómenos. En el caso del sobreajuste, es habitual separar el conjunto de datos disponibles en dos muestras: una de entrenamiento y otra de test. La partición de entrenamiento es utilizada para ajustar la curva, mientras que la de test se emplea para evaluar las predicciones del modelo sobre datos que no ha "visto" antes. Medir el grado de error en la predicción sobre el conjunto de test en lugar de sobre el conjunto de entrenamiento nos da una mejor visión de la capacidad de generalización del método. El tamaño de la partición de entrenamiento suele ser mayor que el de la partición test. Se pueden seguir proporciones como 2/3 – 1/3, 70%-30%, 75%-25%. Para el caso del desajuste, es importante elegir un método de regresión que ajuste el tipo de curva que mejor describe a los datos. En la segunda semana del curso se introducirán varios de ellos, tanto lineales como no lineales.

UCO ONLINE

Finalmente, si el método de regresión consta de parámetros, su elección también va a influir en su rendimiento. Algunos de los métodos de regresión más avanzados presentan este tipo de parámetros. No obstante, este fenómeno suele tener más influencia en los métodos de clasificación (que presentan muchos más parámetros), por lo que será durante su estudio cuando se profundizará en los mecanismos de ajuste de parámetros, así como en otras estrategias para el entrenamiento.

Aunque se sigan estas recomendaciones a la hora de aplicar cualquier método predictivo, seguirá siendo necesario comprobar si el modelo resultante sufre alguno de estos problemas. En el contexto de regresión, veremos a continuación cómo evaluar la calidad de la estimación y medir el error cometido a la hora de predecir.

2. Evaluación de la calidad de la estimación

Como vimos en la lección anterior, el método de los mínimos cuadrados calcula una estimación de los coeficientes $\hat{\beta}_0$ y $\hat{\beta}_1$. Al igual que cuando estimamos la media de la población en base a la media de una muestra, debemos ser conscientes de que la estimación implica cierto error que es conveniente calcular y analizar. En estadística, el análisis de la calidad de una estimación se basa en el concepto de error estándar (SE, *standard error*). Para calcular el error estándar asociado a la estimación de β_0 y β_1, se utilizan las siguientes fórmulas:

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \quad (1)$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (2)$$

En ambas, σ^2 representa la varianza de los errores. Este valor no es conocido, pero se puede estimar a su vez en base a los residuos. Esta estimación se denomina error residual estándar (RSE, *residual standard error*) y se calcula a partir del valor RSS (la suma de los residuos al cuadrado):

$$RSE = \sqrt{\frac{RSS}{(n-2)}} \quad (3)$$

Gracias a estas medidas de error, podemos calcular los *intervalos de confianza*. Un intervalo de confianza, estableciendo un nivel de probabilidad (por ejemplo, 95%), nos indica el rango de valores entre los cuales se encuentra el valor real del parámetro estimado con una certeza igual a dicha probabilidad. Al ser un rango, se expresa con un límite inferior y un límite superior, que en el caso de los coeficientes $\hat{\beta}_0$ y $\hat{\beta}_1$ se obtiene como sigue:

$$[\hat{\beta}_0 - 2 \cdot SE(\hat{\beta}_0), \hat{\beta}_0 + 2 \cdot SE(\hat{\beta}_0)\,] \quad (4)$$
$$[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)\,] \quad (5)$$

A la hora de interpretar un intervalo de confianza debemos considerar que cuanto más amplio es el intervalo, existe mayor probabilidad de que el valor real esté contenido en dicho intervalo. Un intervalo pequeño nos indica que la estimación es más precisa, pero la probabilidad de que el valor real esté contenido en dicho intervalo aumenta.

Los errores estándar (SE) también nos sirven para realizar test de hipótesis sobre los coeficientes estimados. Un test de hipótesis parte de una hipótesis (denominada hipótesis nula, H_0) que establece un fenómeno que queremos comprobar si es cierto o no. En regresión, nos interesa comprobar si realmente existe una relación de dependencia entre las variables x e y. Esta hipótesis está claramente relacionada con la estimación de la pendiente β_1. Si $\beta_1 = 0$, el valor de x carecería de sentido en la recta de regresión que determina el valor de y. Si $\beta_1 = 0$, entonces $y = \beta_0 + \varepsilon$. Por tanto, la hipótesis nula se puede expresar como:

$$H_0: \beta_1 = 0 \quad (6)$$

Para comprobar la veracidad de una hipótesis, se calcula un estadístico que nos indica un valor umbral a partir del cual podemos rechazar o no la hipótesis nula con un cierto grado de probabilidad. En el caso que nos ocupa, necesitamos saber si la estimación de β_1 es lo suficientemente distinta a cero, lo cual depende del error que cometemos al estimar su valor, esto es, $SE(\hat{\beta}_1)$. Si el error es alto, el valor de $\hat{\beta}_1$debería estar muy alejado de 0 para "confiar" en que su valor real (β_1) también lo está. Esta relación se puede expresar por medio del siguiente estadístico:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \quad (7)$$

Muchos paquetes estadísticos y librerías de regresión permiten calcular este valor para, a continuación, comprobar el cumplimiento de la hipótesis nula. El procedimiento consiste en determinar con qué probabilidad el valor del estadístico es igual o superior al valor obtenido, asumiendo que se cumple la hipótesis. A esa probabilidad se le conoce como *p-value*. En la práctica, queremos un *p-value* lo más pequeño posible, siendo habitual establecer umbrales como 0.05 o 0.01. Si el *p-value* es inferior a dicho umbral (denominado nivel de significancia, α), es muy improbable que estemos aceptando la hipótesis nula cuando en realidad es falsa, en nuestro caso, que no exista relación entre x e y. Un razonamiento similar se puede aplicar para el coeficiente β_0. En este caso, estaremos confirmando que, en ausencia de la variable x, la variable y toma valores distintos a cero.

3. Evaluación de la calidad de la predicción

Una vez que hemos obtenido una estimación de los coeficientes y, además, sabemos cómo de fiable es dicha estimación, podemos analizar la calidad del modelo de predicción en sí. En este apartado se estudian medidas que nos informan sobre el error cometido por el modelo de regresión a la hora de predecir, lo cual nos da cierta idea de cómo de bien se ajusta el modelo a los datos. Además, nos permiten comparar el rendimiento de varios modelos obtenidos por distintos métodos.

3.1. Medidas basadas en residuos

Un primer conjunto de medidas se basan en cuantificar el error "acumulado" que comete el modelo de regresión, utilizando para ello los residuos. En el apartado anterior se introdujo una primera medida de error: RSE. Nos interesa que esta medida tenga un valor cercano a cero, pues significa que los valores predichos ($\hat{y}$) son similares a los valores reales (y). La fórmula de RSE mostrada en la ecuación (3) puede reescribirse como:

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

Otra medida basada en la diferencia de residuos es MAE (*mean absolute error*). Nos informa del valor residual medio sobre una muestra de n observaciones, y también debe ser minimizada.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

La siguiente medida es MSE (*mean squared error*), que toma los residuos al cuadrado. De esta forma, los errores grandes son más penalizados que los errores pequeños.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{10}$$

Un inconveniente de MSE es que la unidad de medida se desvirtúa. Por este motivo, RMSE (*root mean squared error)* es más apropiada, ya que al tomar la raíz cuadrada, devuelve el valor a la escala original de la variable independiente. Al igual que las medidas anteriores, valores próximos a cero son mejores. Por sus propiedades, RMSE es la medida más recomendada para comparar diferentes modelos de regresión.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

3.2. El estadístico R²

Todas las medidas anteriores se interpretan de manera similar, pues se basan en un acumulado de los residuos individuales en la muestra. En contraste, el estadístico R^2 nos informa de la proporción de variabilidad de y que puede ser explicada por x. Al ser una medida de proporción, varía entre 0 y 1 y, por tanto, es independente a la escala de la variable y. Un valor cercano a 1 indica que una amplia parte del valor que toma y se debe al valor de x, esto es, el modelo de regresión realmente está explicando la relación existente entre x e y. Por el contrario, un valor próximo a 0 implica o bien que el modelo ajustado no es capaz de explicar adecuadamente esa relación o que el error que estamos cometiendo es grande (o ambas cosas). El estadístico R^2 se calcula según la siguiente fórmula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (12)$$

4. Ejemplo

Para finalizar la lección, vamos a retomar el ejemplo de la lección anterior, donde se aplicó el método de los mínimos cuadrados sobre dos muestras. En el primer caso, se consiguió un ajuste perfecto, por lo que es de esperar que las medidas basadas en residuo sean igual a cero:

Muestra 1: (1,3), (3,7), (4,9), (8,17), (10,21)
Estimación: $\hat{\beta}_0 = 1, \hat{\beta}_1 = 2$

Calculamos las medidas de error MAE, MSE y RMSE:

$$MAE = \frac{1}{5}[|3 - (1 + 2 \cdot 1)| + |7 - (1 + 2 \cdot 3)| + |9 - (1 + 2 \cdot 4)| + |17 - (1 + 2 \cdot 8)| + |21 - (1 + 2 \cdot 10)|] = 0$$

$$MSE = \frac{1}{5}[(3 - (1 + 2 \cdot 1))^2 + (7 - (1 + 2 \cdot 3))^2 + (9 - (1 + 2 \cdot 4))^2 + (17 - (1 + 2 \cdot 8))^2 + (21 - (1 + 2 \cdot 10))^2] = 0$$

$$RMSE = \sqrt{\frac{1}{5}[(3 - (1 + 2 \cdot 1))^2 + (7 - (1 + 2 \cdot 3))^2 + (9 - (1 + 2 \cdot 4))^2 + (17 - (1 + 2 \cdot 8))^2 + (21 - (1 + 2 \cdot 10))^2]} = 0$$

Comprobamos también que el valor del estadístico R^2 es el esperado (1 en este caso):

$$R^2 = 1 - \frac{\left(3-(1+2\cdot 1)\right)^2 + \left(7-(1+2\cdot 3)\right)^2 + \left(9-(1+2\cdot 4)\right)^2 + \left(17-(1+2\cdot 8)\right)^2 + \left(21-(1+2\cdot 10)\right)^2}{(3-11.4)^2 + (7-11.4)^2 + (9-11.4))^2 + (17-11.4))^2 + (21-11.4)^2} = 1$$

Vamos a repetir el proceso con la segunda muestra, donde el ajuste por medio del método de los mínimos cuadrados no era exacto:

Muestra 2: (1.1,3.1), (2.8,7.3), (4.2,8.7), (8.3,16.9), (10.2,20.8)
Estimación: $\hat{\beta}_0 = 1.23, \hat{\beta}_1 = 1.904$

Calculamos las medidas de error MAE, MSE y RMSE:

$$MAE = \frac{1}{5}[|3.1-(1.23+1.904\cdot 1.1)| + |7.3-(1.23+1.904\cdot 2.8)| + |8.7-(1.23+1.904\cdot 4.2)| + |16.9-(1.23+1.904\cdot 8.3)| + |20.8-(1.23+1.904\cdot 10.2)|] =$$

$$= \frac{0.2244+0.7388+0.5268+0.1332+0.1492}{5} = 0.3545$$

En este caso, los errores cometidos en el segundo y el tercer punto de la muestra son bastante superiores al resto. Veamos cómo afecta esto a las medidas MSE y RMSE:

$$MSE = \frac{1}{5}[(3.1-(1.23+1.904\cdot 1.1))^2 + (7.3-(1.23+1.904\cdot 2.8))^2 + (8.7-(1.23+1.904\cdot 4.2))^2 + (16.9-(1.23+1.904\cdot 8.3))^2 + (20.8-(1.23+1.904\cdot 10.2))^2] = 0.1827$$

$$RMSE = \sqrt{\frac{1}{5}[(3.1-(1.23+1.904\cdot 1.1))^2 + (7.3-(1.23+1.904\cdot 2.8))^2 + \cdots + (20.8-(1.23+1.904\cdot 10.2))^2]} = 0.4275$$

Finalmente, vamos a calcular el estadístico R^2 para poder interpretar más fácilmente la calidad de la predicción del modelo de regresión:

$$R^2 = 1 - \frac{\left(3.1-(1.23+1.904\cdot 1.1)\right)^2 + \left(7.3-(1.23+1.904\cdot 2.8)\right)^2 + \cdots + \left(20.8-(1.23+1.904\cdot 10.2)\right)^2}{(3.1-11.36)^2 + (7.3-11.36)^2 + \cdots + (20.8-11.36)^2} = 0.9957$$

Vemos que, a pesar de no ser un ajuste perfecto, la recta de regresión obtenida explica con bastante precisión la relación existente entre x e y, puesto que obtenemos un valor muy próximo a 1.

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.

Recursos en línea

Medidas de regresión en Scikit-Learn. Disponible en: https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics