Modelo de Regresión Lineal: Tarea 2

Profesor: M. en C. Iván Toledano

Septiembre 2024

En una regresión lineal, el objetivo es modelar la relación entre una variable dependiente (Y) y una o más variables independientes ($X_1, X_2, \ldots, X_k$). La ecuación de una regresión lineal múltiple es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon \tag{1}$$

donde:

- β_0 es el término de intercepción.
- $\beta_1, \beta_2, \ldots, \beta_k$ son los coeficientes de regresión que representan el efecto de cada variable independiente X_i sobre la variable dependiente Y.

El conjunto de datos **mtcars** proviene de la revista *Motor Trend US* 1974, ampliamente utilizado para enseñar y demostrar técnicas de análisis de datos, y se encuentra incluido en el lenguaje de programación R. Este dataset contiene información sobre el consumo de combustible y otras características de 32 modelos de automóviles de diferentes marcas y tipos, principalmente de los años 1973-1974. Las variables registradas son las siguientes:

- **mpg**: Indica la eficiencia del consumo de combustible del vehículo en millas por galón.
- **cyl**: Número de cilindros del motor (4,6 u 8 cilindros).
- **disp**: La cilindrada o desplazamiento del motor en pulgadas cúbicas.
- **hp**: Potencia del motor en caballos de fuerza.
- **drat**: Relación entre el eje trasero (rear axle ratio).
- **wt**: Peso del automóvil en miles de libras (1=1000 libras).
- **qsec**: Tiempo en segundos que toma recorrer un cuarto de milla.
- **vs**: Tipo de motor (0=tipo V, 1=recto/línea).
- **am**: Tipo de transmisión (0 = automática, 1 = manual).
- **gear**: Número de marchas o velocidades en la transmisión (3,4, ó 5).
- **carb**: Número de carburadores.

En la siguiente actividad utilizaremos este dataset para crear varios modelos de regresión lineal y elegir el mejor de ellos según ciertas condiciones.

1. Ejercicios

Una prueba de hipótesis t (de dos colas) es utilizada para determinar si cada coeficiente β_i es significativamente diferente de cero, es decir, si la variable independiente asociada tiene un efecto significativo en la variable dependiente. Un coeficiente $\beta_i = 0$ significaría que la variable independiente asociada realmente no tiene una correlación con la variable dependiente. La formulación de la prueba de hipótesis es la siguiente:

- **Hipótesis nula** (H_0): $\beta_i = 0$. Esto implica que la variable independiente X_i no tiene un efecto significativo sobre Y.
- **Hipótesis nula** (H_1): $\beta_i \neq 0$. Esto implica que la variable independiente X_i tiene un efecto significativo sobre Y.

Una prueba de hipótesis consiste en calcular un estadístico (en este caso estadístico t) y compararlo con una distribución de probabilidad correspondiente (en este caso, una distribución Student t). De esta comparación se obtiene un **p-valor** asociado, se elige un nivel de significancia α adecuado, y se realiza una conclusión de la prueba de hipótesis,

- Si el p-valor es menor que el nivel de significancia (α, típicamente 0.05), se rechaza la hipótesis nula H_0, concluyendo que el coeficiente β_i es significativamente diferente de cero.
- Si el p-valor es mayor o igual a α, no se rechaza la hipótesis nula, lo que sugiere que no hay evidencia estadística suficiente para afirmar que β_i es diferente de cero.

Esta prueba de hipótesis ayuda a interpretar la relevancia de cada predictor en el modelo. Denominaremos como un **modelo válido** aquel en el cual **todas** sus variables predictoras son significativamente diferentes de cero.

Ejercicio 1:
Encontrar el mejor modelo válido para explicar el rendimiento (mpg) de un auto, haciendo uso de 3 de las variables continuas originales del conjunto de datos: *disp, hp, drat, wt, qsec*. Para este ejercicio se consideran como válidos coeficientes β_i cuyo p-valor asociado de la prueba de hipótesis se mayor a 0.1, es decir, se considera un nivel de significancia de 10 %.

Ejercicio 2:
Escribir la ecuación de la recta del mejor modelo encontrado, escribiendo explícitamente sus variables y coeficientes. Interpretar con sus palabras el efecto de cada variable independiente sobre la variable dependiente.

Los **residuos** son las diferencias entre los valores observados (i.e., de los datos) de la variable independiente Y y los valores predichos $\hat{Y}$ por el modelo de regresión. Estos valores predichos provienen de la ecuación del modelo, como la que escribieron en el ejercicio 2, haciendo la sustitución de los valores numéricos de las variables independientes de los datos. Matemáticamente, para cada observación i,

$$e_i = Y_i - \hat{Y}_i \tag{2}$$

Estos residuos representan el "error" del modelo, es decir, la parte de la variación en Y que no es explicada por las variables independientes X. Recordar que las variables recolectadas en el dataset tienen también efectos aleatorios que no es explicado por un modelo.

El **error estándar residual** (RMSE), también conocido como el error estándar de la estimación, mide la dispersión de los residuos de una regresión lineal (piénsenlo como la desviación estándar de los residuos). Es una medida de cuán lejos están, en promedio, los valores de los datos a los valores predichos por la recta ajustada. El error estándar residual se calcula mediante la siguiente ecuación,

$$\text{Error estandar residual} = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{e}_{\mathrm{i}}^2}{\mathrm{n}-\mathrm{k}-1}} \tag{3}$$

donde e_i son los residuos, n es el número de observaciones y k es el número de variables independientes en el modelo. El error estándar residual se expresa en las mismas unidades de la variable dependiente Y, y por lo tanto es fácil de interpretar. Un valor más bajo indica que los puntos de los datos se encuentran más cerca de la recta ajustada de la regresión, y por lo mismo, en una comparación entre diferentes modelos, un valor más bajo de éste generalmente muestra un mejor modelo.

Por otro lado, tenemos el coeficiente de determinación R^2, que es una medida estadística que indica la proporción de la varianza total de la variable dependiente Y que es explicada por el modelo de regresión. Se calcula mediante la siguiente expresión,

$$R^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} \tag{4}$$

donde $\mathrm{SSR} = \sum_{\mathrm{i}}^{\mathrm{n}} (\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i})^2$ es la suma de los cuadrados de los residuos y $\mathrm{SST} = \sum_{\mathrm{i}}^{\mathrm{n}} (\mathrm{Y_i} - \bar{\mathrm{Y}}_\mathrm{i})^2$ es la suma total de los cuadrados que representa la variación total de los datos. El valor de R^2 puede estar entre 0 y 1 e indica el porcentaje de la variación de Y explicada por el modelo. Ejemplo: Un valor de $R^2 = 0{,}75$ significa que el 75 % de la variación en Y es. explicada por las variables independientes del modelo. Aún así, un R^2 alto, aunque indique un buen ajuste, no garantiza que el modelo sea el mejor o el más apropiado. Un R^2 puede resultar de un sobreajuste (incluir demasiadas variables en el modelo), lo que podría afectar la capacidad predictiva en datos nuevos.

Mientras que el RMSE mide la precisión del modelo en términos absolutos (además de que tiene las mismas unidades que la variable independiente), el factor R^2 mide la proporción de la variabilidad explicada en términos relativos. Un valor del RMSE bajo y un R^2 alto son indicativos de un buen ajuste del modelo. Sin embargo, es importante usar ambos en conjunto, ya que un R^2 por sí solo no implica un RMSE bajo, especialmente en datasets con una gran cantidad de datos.

Ejercicio 3:
Comparar el valor de R^2 y RMSE del mejor modelo válido obtenido del ejercicio 1 contra aquel de un modelo que sólo considera el cilindraje (disp), para explicar el rendimiento (mpg) del auto. ¿Cuál de éstos consideras que es el mejor modelo y por qué?. Posteriormente, en el modelo de una sóla variable (disp), reemplazarla por la transformación $(\mathrm{disp})^{-0{,}46}$, es decir, que la relación entre (disp) y (mpg) es una ley de potencias. Hacer una regresión lineal sobre esta nueva variable. Comparar el modelo de tres variables válido obtenido del ejercicio 1 con el modelo de 1 sola variable $(\mathrm{disp})^{-0{,}46}$. ¿Cuál de estos modelos consideras que es el mejor y por qué?, ¿Cambia la respuesta del punto anterior?, ¿Por qué sucede esto?.

En una regresión lineal se hace la suposición de que los residuos siguen una **distribución normal** centrada alrededor de cero, y esto es fundamental por varias razones estadísticas y prácticas:

- **Validez de las inferencias estadísticas:** Para realizar pruebas de hipótesis y construir intervalos de confianza para los coeficientes del modelo de la regresión, como aquellos del ejercicio 1.
- **Optimización del estimador de mínimos cuadrados ordinarios (OLS):** El método utilizado para tener la mejor recta ajustada asume que:
  - Los residuos tienen una varianza constante (homocedasticidad).
  - Los residuos son independientes entre sí.
  - Los residuos son normalmente distribuidos alrededor de cero.

  Cuando los residuos son normalmente distribuidos, los estimadores de la OLS son estimadores insesgados y de varianza mínima, i.e., lo más eficientes posible.
- **Identificación de problemas de modelado:** Si los residuos no se encuentras normalmente distribuidos, esto puede ser una señal de que hay problemas con el modelado.
  - La relación entre la variable dependiente e independiente podría ser no lineal, y un modelo de regresión lineal simple no es adecuado.
  - Podría haber variables relevantes que no están siendo incluidas en el modelo.
  - Valores atípicos podrían haber distorsionado la distribución de los residuos.

Si los residuos no siguen una distribución normal, existen diferentes estrategias para abordar este problema:

- **Transformaciones de las variables:** Aplicar transformaciones (como logarítmica, ley de potencia, etc.) a la variable dependiente o a las independientes. para hacer que los residuos se aproximen más a una distribución normal.
- **Usar modelos no lineales:** Considerar modelos de regresión no lineal o más complejos.

Existen diferentes métodos para evaluar si ciertos datos siguen una distribución normal. Los más usuales son los que se presentan a continuación:

Gráfico Q-Q
Un **gráfico Q-Q** (cuantil-cuantil) es una representación visual que compara los cuantiles de los datos observados con los cuantiles teóricos de una distribución de referencia. Este gráfico es utilizado comúnmente para evaluar la normalidad de los datos, pero también puede usarse para comparar los datos con otras distribuciones (e.g., exponencial, Student t, etc.). Si los puntos en el gráfico Q-Q siguen de cerca la línea diagonal, esto indica que los datos observados tienen una distribución similar a la distribución teórica de referencia. Si se compara con una distribución normal, un buen ajuste a la línea sugiere que los datos son aproximadamente normales.

Prueba de normalidad de Shapiro-Wilk
La prueba de Shapiro-Wilk es una prueba estadística diseñada específicamente para evaluar la normalidad de una muestra de datos. Es una prueba de hipótesis basado en un estadístico W que sigue la siguientes hipótesis nula y alternativa:

- H_0 : Los datos siguen una distribución normal.
- H_1 : Los datos no siguen una distribución normal.

Prueba de normalidad de Kolmogorov-Smirnov
La prueba de Kolmogorov-Smirnov es una prueba no paramétrica que compara la distribución de una muestra con una distribución de referencia (como la normal, exponencial, etc.). También puede usarse para comparar dos muestras entre sí para determinar si provienen de la misma distribución. En el contexto de la normalidad, se usa para comparar la distribución de los datos con una distribución normal teórica. Se basa en un estadístico D para determinar una diferencia suficientemente grande como para rechazar la hipótesis de normalidad. Esta prueba sigue la siguiente hipótesis nula y alternativa,

- H_0 : Los datos siguen una distribución normal específica (o una de referencia).
- H_1 : Los datos no siguen una distribución la distribución normal especificada (o de referencia).

Ambas pruebas de hipótesis son útiles para evaluar la normalidad de los datos, pero cada una tiene sus fortalezas y debilidades. La elección de cuál usar dependerá del contexto, el tamaño de la muestra, y el propósito del análisis. La prueba de Shapiro-Wilk es generalmente más potente para evaluar la normalidad en muestras pequeñas (<50), mientras que la prueba de Kolmogorov-Smirnov ofrece más flexibilidad para comparar distribuciones en general.

Ejercicio 4:
Genera una gráfica Q-Q de los residuos del mejor modelo válido de tres variables identificado. Además, realiza la prueba de hipótesis de normalidad utilizando los métodos de Shapiro-Wilk y Kolmogorov-Smirnov considerando un nivel de significancia de $\alpha = 0.05$. investiga cuál de estas pruebas es más adecuada para evaluar la normalidad en este conjunto de datos. interpreta los resultados de las gráficas y de las pruebas de hipótesis para determinar si los residuos del modelo siguen una distribución normal. Posteriormente, analiza si los residuos de un modelo con una sola variable $(\text{disp})^{-0,46}$ también siguen una distribución normal, y verifica si los residuos son normales sin aplicar esta transformación. Finalmente, si cambiamos el nivel de significancia a 0.1 ¿Se mantienen las mismas conclusiones?

Ejercicio 5:
Concluye si el modelo de tres variables obtenido es adecuado para explicar la variable de rendimiento del auto (mpg). ¿Qué ajustes propondrías para mejorar el modelo en términos de precisión y reducción de varianza? Utiliza toda la información obtenida en los ejercicios anteriores, resumiéndola en tus propias palabras. **OPCIONAL:** Experimenta con todas las variables disponibles en el conjunto de datos para intentar encontrar un modelo mejor. Puedes emplear pruebas adicionales a las mencionadas para respaldar tus conclusiones.