Estadística e inferencia I

Segundo Cuatrimestre — 2024

Práctica 3: Regresión lineal

1. Consideremos la función $f : \mathbb{R} \to \mathbb{R}$ dada por $f(x) = 2x - 1$.

(*a*) Sean $X \sim \mathcal{U}(0, 1)$ y $Y \coloneqq f(X)$. Fijemos $n = 100$. Tomar muestras $x_1, \ldots, x_n$ de X y aplicar la función f a cada muestra para calcular $y_i = f(x_i)$, con $1 \leq i \leq n$. Realizar un ajuste lineal de la forma $Y = X\beta_1 + \beta_0 + \epsilon$ a partir de los samples $(x_1, y_1), \ldots, (x_n, y_n)$ generados y calcular, para cada $i \in \{1, \ldots, n\}$,

$$\varepsilon_i = y_i - (x_i\beta_1 + \beta_0).$$

Describir la distribución empírica dada por $\{\varepsilon_1, \ldots, \varepsilon_n\}$.

(*b*) Sean $X \sim \mathcal{U}(0, 1)$, $Z \sim \mathcal{N}(0, .25)$ y $Y \coloneqq f(X)+Z$. Fijemos $n = 100$. Tomar muestras $x_1, \ldots, x_n$ de X y $z_1, \ldots, z_n$ de Z y calcular $y_i = f(x_i) + z_i$, con $1 \leq i \leq n$. Realizar un ajuste lineal de la forma $Y = X\beta_1 + \beta_0 + \epsilon$ a partir de los samples $(x_1, y_1), \ldots, (x_n, y_n)$ generados y calcular, para cada $i \in \{1, \ldots, n\}$,

$$\varepsilon_i = y_i - (x_i\beta_1 + \beta_0).$$

Describir la distribución empírica dada por $\{\varepsilon_1, \ldots, \varepsilon_n\}$ haciendo un histograma.

(*c*) Repetir el paso anterior desde $n = 10$ hasta $n = 1000$. Graficar y estimar la media y la varianza de ε en función de n.

2. Consideremos la función $f : \mathbb{R} \to \mathbb{R}$ dada por $f(x) = 2x - 1$. Sean $X \sim \mathcal{U}(0, 1)$, $Z \sim \mathcal{N}(0, .25)$ e $Y \coloneqq f(X) + Z$.

(*a*) Fijemos $n = 100$. Tomar muestras $x_1, \ldots, x_n$ de X y $\varepsilon_1, \ldots, \varepsilon_n$ de Z y calcular $y_i = f(x_i) + z_i$, con $1 \leq i \leq n$. Realizar un ajuste lineal de la forma $Y = X\beta_1 + \beta_0 + \epsilon$ a partir de los samples $(x_1, y_1), \ldots, (x_n, y_n)$ generados. Hacer un diagrama de dispersión con los samples y la estimación del modelo de regresión lineal

(*b*) Repetir el paso anterior $m = 1000$ veces para obtener m pares de coeficientes $\hat{\beta}_0$ y $\hat{\beta}_1$, es decir, m samples de $\hat{\beta}_0$ y m samples de $\hat{\beta}_1$. Visualizar la distribución de cada $\hat{\beta}_j$ haciendo un histograma; visualizar la distribución conjunta haciendo un diagrama de dispersión con todas las estimaciones. Además, graficar todas las rectas de regresión estimadas.

(*c*) Repetir el paso anterior para $m = 100$ desde $n = 10$ hasta $n = 1000$. Graficar la media y la varianza de cada coeficiente $\hat{\beta}$ en función de n.

3. Supongamos que tenemos un dataset $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ para $n \in \mathbb{N}$ y hacemos un modelo de regresión lineal, donde $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

(*a*) Mostrar que los valores de β_0 y β_1 que minimizan la suma de los residuos $\sum_{i=1}^n \varepsilon_i^2$ son

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

(*b*) Podemos demostrar que un estimador no sesgado de σ es

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

¿Cómo se interpreta el 2 que aparece restando en el denominador?

(*c*) ¿Que pasaría si hicieramos un ajuste de la forma $Y \sim \beta X$?

4. En este ejercicio, utilizaremos el dataset de automóviles disponible en el siguiente enlace: https://archive.ics.uci.edu/static/public/9/auto+mpg.zip. Este conjunto de datos contiene información sobre varios atributos de automóviles, incluyendo Millas por galón de combustible (MPG) y Caballos de Fuerza (HP).

(*a*) Ajustar un modelo lineal que relacione MPG con HP utilizando todos los puntos del dataset. Calcular el R^2 para este modelo.
  (*i*) Ajustar logaritmo de MPG vs HP y calcular R^2.
  (*ii*) Ajustar logaritmo de MPG vs logaritmo HP y calcular R^2.

(*b*) Veamos qué pasa si no usamos todo el dataset sino sólo un porcentaje.
  (*i*) Ajustar un modelo lineal a MPG vs HP, pero esta vez utilizando solo el 80% de los puntos del dataset seleccionados al azar. ¿Cuánto vale R^2? ¿Y sobre el 20% restante de los puntos?
  (*ii*) Repetir el punto anterior para diferentes porcentajes de datos de entrenamiento.

(*c*) Utilizando el dataset completo, realizar ajustes lineales para relacionar MPG con cada una de las variables. Ordenar las variables de acuerdo al R^2 obtenido de la más importante a la menos importante.

5. Con el dataset del ejercicio anterior:

(*a*) Ajustar una regresión lineal múltiple de MPG en función de todas las otras variables. Hacer los gráficos pertinentes para analizar el comportamiento del modelo.

(*b*) Hacer *forward selection* para seleccionar el mejor modelo de regresión según el estadístico de Mallow. Este proceso consiste en agregar una variable a la vez al modelo inicial y comparar el valor del estadístico de Mallow para cada modelo, eligiendo el del valor más bajo.

(*c*) Realizar la regresión con todas las variables, pero con regularización de Ridge y de Lasso. Dividir el dataset en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%) para elegir el mejor valor de λ para cada uno.

6. Generar $n = 100$ samples de $X_1 \sim \mathcal{U}(0, 1)$, de $Z \sim \mathcal{N}(0, 0.1)$ y de $\varepsilon \sim \mathcal{N}(0, 1)$; a partir de ellas generar n samples de $X_2 \coloneqq 0.5X_1 + Z$ y de

$$Y \coloneqq 2 + 2X_1 + 0.3X_2 + \varepsilon.$$

(*a*) Calcular la correlación entre las muestras de X_1 y X_2, y graficar su distribución conjunta.

(*b*) Ajusta un modelo de regresión lineal por mínimos cuadrados para predecir Y utilizando tanto X_1 como X_2. Describir los resultados obtenidos, incluyendo los coeficientes de regresión $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$, y analizar la relación entre estos coeficientes y los verdaderos β_0, β_1 y β_2.

(*c*) Ajustar un modelo de regresión lineal por mínimos cuadrados para Y y analizar resultados
  (*i*) utilizando solo las muestras de X_1;
  (*ii*) utilizando solo las muestras de X_2.

(*d*) Graficar la distribución conjunta de $\hat{\beta}_1$ y $\hat{\beta}_2$.

7. Probar que las estimaciones del modelo $Y \sim X$ son las inversas del modelo $X \sim Y$ si y sólo si el ajuste del modelo es perfecto.