Tarea 3 Pablo Soria Garcia

Parte teórica 1

Supongamos que tenemos un vector con p observaciones de las n variables explicativas y con su respectiva variable dependiente y los expresamos en forma matricial así como nuestro vector de coeficientes de la regresión:

$$Y_{nx1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \ \beta_{p\,x\,1} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \ X_{n\,x\,p} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Supongamos que las n variables explicativas para cada una de las p personas le quitamos la media de la muestra es decir las centramos, asumimos el mismo supuesto con el vector de respuestas, el objetivo de esto es deshacernos de la β_0 sin perder generalidad, ya que la minimización no depende de β y lleva a que $\hat{\beta}_0 = 0$ entonces podemos plantear el problema de minimización de errores cuadráticos como sigue:

$$\hat{\beta} = argmin_{\beta} \, \|y - X\beta\|^2$$

Es decir el problema es encontrar la β que minimiza la norma euclidiana al cuadrado, esto se puede expresar como:

$$argmin_{\beta} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2$$

Que también podemos expresar como:

$$argmin_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2$$

Es decir tratamos de minimizar la diferencia al cuadrado entre nuestra variable respuesta y la respuesta que obtendríamos de multiplicar los coeficientes de regresión β's multiplicados por su respectiva variable explicativa, en otras palabras el error de predicción.

Para realizar el cálculo específico, sacamos el gradiente de con respecto a β he igualamos a cero, un paso previo para hacer la operación de derivar la norma es mostrarla como:

$$\nabla\|Y - X\beta\|^2 = \nabla(Y^TY + \beta^TX^TX\beta - 2\beta^TX^TY)$$

Ya que esto nos permite hacer uso de algunas propiedades cool del cálculo de matrices, tomando el gradiente con respecto a β, suponemos que $A = X^T X$ por lo tanto $\beta^T X^T X \beta$ es de la forma $x^T A x$, ya que A es simétrica porque $X^T X$ tiene dimensión nxn y su derivada es $2Ax$ (caso especial que vimos en clase), por otro lado el término $2\beta^T X^T Y$ lo podemos ver como $b^T x$ con $x = 2X^T Y$ por lo que su derivada se calcula trivialmente, el primer término es independiente de β por lo que el gradiente final queda de la forma:

$$= 2X^T X\beta - 2X^T Y$$

Ahora igualamos a cero y despejamos, para esto vamos a suponer que la matriz $X^T X$ es invertible.

$$2X^T X\beta - 2X^T Y = 0 \overset{1}{\Rightarrow} X^T X\beta = X^T Y$$

Cancelamos el escalar y reordenamos, entonces multiplicando por la izquierda por la inversa de $X^T X$:

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

Estas ecuaciones que tienen como variables las $\beta's$ se conocen como ecuaciones normales y coinciden con las ecuaciones que encontramos en el problema planteado desde el punto de vista estadístico.

La linealidad de este ajuste se debe a que el problema plantea encontrar la combinación lineal de coeficientes β que mejor representa linealmente a nuestra variable de respuesta Y mediante las variables explicativas X.

Por otro lado podríamos utilizar el mismo enfoque para ajustar linealmente a $y = x^2$ ya que este enfoque asume linealidad sobre los coeficientes de regresión β no sobre las variables explicativas X.

Parte teórica 2

Ahora vamos a incorporar una columna de 1 a la matriz de variables explicativas, esto se hace con el objetivo de quitar el supuesto de que dichas variables se encuentran centradas, es decir que a cada una se le resto su media. Esto implica que ahora la matriz se ve de la siguiente forma:

$$Y_{nx1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \; \beta_{p+1\,x\,1} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} \; X_{n\,x\,p+1} = \begin{bmatrix} 1 & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{np} \end{bmatrix}$$

Notemos que para que el problema siga teniendo sentido hemos incluido un β_0 (el coeficiente de intersección), ahora el vector $X\beta$ se ve de la siguiente forma gracias al ajuste de 1's en la primera columna de la matriz de variables explicativas:

$$X\beta_{nx1} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} \\ \vdots \\ \beta_0 + \beta_1 x_{1n} + \cdots + \beta_p x_{np} \end{bmatrix}$$

Esta forma es mucho más similar al planteamiento estadístico tradicional.

Parte teórica 3

Enfoquemos ahora el problema desde el punto de vista estadístico, tenemos la variable respuesta y las variables explicativas además de un error de estimación que se distribuye normal con media 0 y varianza σ^2 es decir:

$$Y_i = \beta_0 + \beta_1 X_i^1 + \cdots + \beta_p X_i^p + \epsilon_i$$

Dónde $\epsilon_i \sim N(0, \sigma^2)$

Esto lo podemos escribir de forma compacta como:

$$Y_i = \beta X_i + \in_i \quad \text{O de forma general} \quad Y = \beta X + \epsilon$$

Dónde $\beta = [\beta_1, \ldots\ldots, \beta_n]\ y\ X_i = \left[X_i^1, \ldots\ldots, X_i^p\right]$

Ahora bien, asumiendo que los errores no están correlacionados y que se distribuyen de forma normal, esto quiere decir que:

$$\epsilon_i \sim N(0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(y_i - x_i\beta)^2\right\}$$

Por lo tanto si tomaos el vector X_i como datos, las variables de respuesta también se distribuyen de forma normal y por tanto la función de verosimilitud de basado en la muestra es:

$$L = \prod_{i=1}^{n} N(x_i\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{\frac{-1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right\}$$

Si tomamos la log verosimilitud nos deja con el siguiente resultado:

$$L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)$$

Si derivamos con respecto a $\beta\ y\ \sigma^2$:

$$\frac{dL}{d\beta} = \frac{1}{\sigma^2}(y - X\beta)^T X$$

$$\frac{dL}{d\sigma^2} = \frac{n}{2\sigma^2} - \frac{1}{2\sigma^4}(y - X\beta)^T(y - X\beta)$$

Por lo tanto si igualamos a cero la primera derivada con respecto ha β:

$$\frac{1}{\sigma^2}(y - X\beta)^T = 0 \overset{1}{\Rightarrow} \hat{\beta} = \ (X^T X)^{-1} X^T Y$$

Que es la misma solución que encontramos cuando planteamos el problema de forma matricial.

Teorema de Gauss Markov

Teorema de Gauss-Márkov

En estadística, el **Teorema de Gauss-Márkov**, formulado por Carl Friedrich Gauss y Andréi Márkov, establece que en un modelo lineal general (MLG) en el que se establezcan los siguientes supuestos:

- Correcta especificación: el MLG ha de ser una combinación lineal de los parámetros (β) y no necesariamente de las variables: $Y = X\beta + u$
- Muestreo aleatorio simple: la muestra de observaciones del vector $(y_i,\ x_{2i},\ x_{3i},\ \dots,\ x_{ki})$ es una muestra aleatoria simple y, por lo tanto, el vector $(y_i,\ X_i')$ es independiente del vector $(y_i,\ X_j')$
- Esperanza condicionada de las perturbaciones nula: $E(u_i|X_i') = 0$
- Correcta identificación: la matriz de regresoras (X) ha de tener rango completo: rg(X)=K<=N
- Homocedasticidad: $Var(U|X) = \sigma^2 I$

el estimador mínimo cuadrático ordinario (MCO) de B es el estimador lineal e insesgado óptimo (ELIO o BLUE: best linear unbiased estimator), es decir, el estimador MCO es el estimador eficiente dentro de la clase de estimadores lineales e insesgados.