Analisis Multivariado 2

Ejercicio 1:

Demostrar que

* (x_i, y_i) $i=1 \dots n$ muestra de entrenamiento
* (x_i', y_i') $i=1 \dots m$ muestra de validación

Considerando un modelo de regresión lineal, con una muestra X iid $\sim D$ con una distribución D asignada para los dato

$\hat{\beta} = (X^T X)^{-1} X^T Y$ $\longrightarrow$ Con X la matriz de datos $n \times p$
n individuos y p variables

$$\underbrace{E\left[\frac{1}{n}\sum_{1}^{n}\left(Y_i - \hat{\beta}^T x_i\right)^2\right]}_{\text{Error de entrenamiento}} \leq \underbrace{E\left[\frac{1}{m}\sum_{1}^{m}\left(Y_i' - \hat{\beta}^T x_i'\right)^2\right]}_{\text{Error de validación}}$$

Esto implica que la esperanza del error de entrenamiento es optimista respecto al error de validación como vimos en clase

1er Paso: Argumentar que el error de validación es el mismo si tenemos m observaciones o solamente una, es decir:

$$\underbrace{E\left[\frac{1}{m}\sum_{1}^{m}\left(Y_i' - \hat{\beta}^T x_i'\right)^2\right]}_{A} = \underbrace{E\left[\left(Y_1' - \hat{\beta}^T x_1'\right)^2\right]}_{*}$$

Si desarrollamos (A)

$$E(z) = E_t\left\{E_v\left[\frac{1}{m}\sum_{1}^{m}\left(Y_i - \hat{\beta} x_i\right)^2 \,\Big|\, t\right]\right\}$$

Siendo t las posibles muestras de entrenamiento posibles la esperanza del interior refiere a

$$E_v\left[\frac{1}{m}\sum_{1}^{m}\left(Y_i - \hat{\beta}^T X_i\right)^2\right] \overset{\text{Linealidad Esp}}{=} \frac{1}{m}\sum_{1}^{m} E_v\left(Y_i - \hat{\beta}^T x_i\right)^2$$

$$= \frac{1}{m}\left\{\underbrace{E_v\left[\left(Y_1 - \hat{\beta}^T X_1\right)^2\right] + \dots E_v\left[\left(Y_m - \hat{\beta}^T X_m\right)^2\right]}_{m \text{ veces, dado que asumimos iid}}\right\}$$

$$= \frac{\cancel{m}}{\cancel{m}} E_v\left[\left(Y_1 - \hat{\beta}^T X_1\right)^2\right]$$

Considerando la primera observación como representante.

De esta manera es indiferente tener m puntos de prueba o 1, por lo tanto, si asumimos $m=n$ (El conjunto de entrenamiento tiene la misma cantidad de observaciones que el de validación no perdemos generalidad).

De esta manera,

$$E_T\left[E_v\left(\frac{1}{m}\sum_{1}^{m}\left\{Y_i - \hat{\beta}^T x_i\right\}^2\right)\right] = E_T\left\{E_v\left(\left(Y_1 - \hat{\beta} x_1\right)^2\right)\right\}$$

Parte 2: Considerando las siguientes variables aleatorias

$$A = \frac{1}{n}\sum_{1}^{n}\left(Y_i - \hat{\beta}^T x_i\right)^2 \quad y \quad B = \frac{1}{n}\sum_{1}^{n}\left(Y_i' - \tilde{\beta}^T x_i'\right)^2$$

Con $\hat{\beta}^T$ el coeficiente MCO en la muestra de entrenamiento, y $\tilde{\beta}^T$ el coeficiente MCO en la muestra de validación

Si seguimos asumiendo $m=n$ y utilizando que este supuesto no impacta en la demostración, podemos decir que:

$A \rightarrow$ MSE en la muestra de entrenamiento para $\hat{\beta}$
$B \rightarrow$ MSE en la muestra de validación para $\tilde{\beta}$

Además, lo que utilizamos en la demostración de la parte 1

$$\mathcal{E}(A) = \mathcal{E}\left[\frac{1}{n}\sum_{1}^{n}\left(Y_i - \hat{\beta}^T x_i\right)^2\right] = \mathcal{E}\left[\left(Y_1 - \hat{\beta}^T X_1\right)^2\right]$$

$$\mathcal{E}(B) = \mathcal{E}\left[\frac{1}{n}\sum_{1}^{n}\left(Y_i' - \tilde{\beta}^T x_i'\right)^2\right] = \mathcal{E}\left[\left(Y_1' - \tilde{\beta}^T X_1'\right)^2\right]$$

Para acotar notación tanto $\mathcal{E}(A)$ y $\mathcal{E}(B)$ fueron simplificado su notación pero no hay que olvidar que seguimos considerando la aleatoriedad proveniente a la muestra de entrenamiento y validación

Al que ambos conjuntos provienen del mismo mecanismo generador de datos D (por hipotesis).

$$\mathcal{E}(A) = \mathcal{E}(B) \longrightarrow \mathcal{E}\left[\left(Y_1 - \hat{\beta}^T x_1\right)^2\right] = \mathcal{E}\left[\left(Y_1' - \tilde{\beta}^T x_1'\right)^2\right]$$

Recordando que $B = \frac{1}{n}\sum_{1}^{n}(Y_i' - \tilde{\beta}^T x_i')^2$

Y que $\tilde{\beta}$ es un estimador proveniente de una estimación MCO
por el teorema de Gauss-Markov
es de mínima varianza, el que es insesgado
para estimar β el $MSE = Var(\hat{\beta})$

$$B = MSE(\tilde{\beta}) = \frac{1}{n}\sum_{1}^{n}\left(Y_i' - \tilde{\beta}^T x_i'\right)^2 \leq \frac{1}{n}\sum_{1}^{n}\left(Y_i' - \hat{\beta}^T x_i'\right)^2$$

Recordemos que $\hat{\beta}$ fue estimado con el conjunto de datos de entrenamiento y lo estamos utilizando con un nuevo conjunto, es decir, si proponemos un Nuevo estimador para β
CY

$$C = (X'X)^{-1}X + M$$

$$Var(\beta^*) = Var(CY) = C\,Var(Y)C'$$

$$= \sigma^2 CC'$$

$$= \sigma^2([X'X]^{-1}X + M)([X'X]^{-1}X + M)'$$

$$= \sigma^2([X'X]^{-1}X + M)(X[X'X]^{-1} + M') \quad \{MX = 0\}$$

$$= \sigma^2(X'X)^{-1} + \sigma^2 MM'$$

$$= Var(\hat{\beta}) + \sigma^2 MM' \rightarrow MM' > 0$$

siempre la varianza va a ser mayor

$$B \leq \frac{1}{n}\sum_{1}^{n}\left(Y_i' - \hat{\beta}x_i'\right)^2 \rightarrow \text{tomando Esperanza}$$

$$E(B) \leq E\left[\frac{1}{n}\sum_{1}^{n}\left(Y_i' - \hat{\beta}x_i'\right)^2\right] \quad \text{usando } E(A) = E(B)$$

$$\mathcal{E}(A) \leq \mathcal{E}\left[\frac{1}{n}\sum_{1}^{n}\left(y_i' - \hat{\beta}x_i'\right)^2\right]$$

$$\mathcal{E}\left[\frac{1}{n}\sum_{1}^{n}\left(y_i - \hat{\beta}^T x_i\right)^2\right] \leq \mathcal{E}\left[\frac{1}{n}\sum_{1}^{n}\left(y_i' - \hat{\beta}^T x_i'\right)^2\right]$$

Reescribiedo y recordado que $m = n$

$$\underbrace{\mathcal{E}\left[\frac{1}{n}\sum_{1}^{n}\left(y_i - \hat{\beta}^T x_i\right)^2\right]} \leq \mathcal{E}\left[\frac{1}{m}\sum_{1}^{n}\left(y_i' - \hat{\beta}x_i'\right)^2\right]$$

$\mathcal{E}(Err_T) \leq \mathcal{E}(Err_V)$ → El error de entrenamiento es optimisto