

Modelo lineal

Sea $\mathcal{L} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, con $x_i \in \mathbb{R}^p$
 $y_i \in \mathbb{R}$.

Queremos predecir y_i a partir de x_i .

Definimos \hat{y}_i como la predicción a partir de x_i .

Una solución es $\hat{y}_i = \beta_0 + \underline{x}_i^T \underline{\beta}$, el error es

$$e_i = y_i - \hat{y}_i.$$

En forma vectorial esto puede escribirse

como

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & \underline{x}_1^T \\ 1 & \underline{x}_2^T \\ \vdots & \vdots \\ 1 & \underline{x}_n^T \end{bmatrix} \begin{bmatrix} \beta_0 \\ \underline{\beta} \end{bmatrix}.$$

$$\underline{\hat{Y}} = \underline{X} \underline{\beta} \quad \text{con } \underline{X} \in \mathbb{R}^{n \times (p+1)}$$

Si $\underline{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, entonces el vector de error es

$\underline{e} = \underline{Y} - \underline{\hat{Y}}$. Una solución para encontrar $\underline{\beta}$ es minimizar

$\|\underline{e}\|$, pero esto es lo mismo que minimizar $\frac{\|\underline{e}\|^2}{2} = \frac{(\underline{Y} - \underline{\hat{Y}})^T (\underline{Y} - \underline{\hat{Y}})}{2}$.

$$\text{Así, } \hat{\underline{\beta}} = \arg \min_{\underline{\beta}} \frac{(\underline{Y} - \underline{\hat{Y}})^T (\underline{Y} - \underline{\hat{Y}})}{2} \quad (1)$$

Observemos que:

$$\begin{aligned} (\underline{y} - \hat{\underline{y}})^T (\underline{y} - \hat{\underline{y}}) &= \underline{y}^T \underline{y} - \underline{y}^T \hat{\underline{y}} - \hat{\underline{y}}^T \underline{y} + \hat{\underline{y}}^T \hat{\underline{y}} \\ &= \underline{y}^T \underline{y} - \underline{y}^T \underline{X} \beta - \hat{\beta}^T \underline{X}^{-T} \underline{y} + \hat{\beta}^T \underline{X}^T \underline{X} \beta \\ &= \underline{y}^T \underline{y} - 2 \hat{\beta}^T \underline{X}^{-T} \underline{y} + \hat{\beta}^T \underline{X}^T \underline{X} \beta \end{aligned}$$

luego

$$\frac{\partial}{\partial \beta} \frac{\|\underline{y}\|^2}{2} = \frac{\partial}{\partial \beta} \frac{1}{2} (\hat{\underline{y}} - \underline{y})^T (\hat{\underline{y}} - \underline{y})$$
$$= -\underline{X}^{-T} \underline{y} + \underline{X}^T \underline{X} \beta$$

Igualando esta derivada a cero se obtiene

$$\underline{X}^T \underline{X} \beta = \underline{X}^T \underline{y} \quad (2)$$

A (2) se le conoce como ecuación normal.

Observemos que $\frac{\partial}{\partial \beta^T \partial \beta} \frac{\|\underline{y}\|^2}{2} = \underline{X}^T \underline{X}$, luego

el β que satisface (2) es un mínimo sv $\underline{X}^T \underline{X}$ es definida positiva.

Cómo calcular $\hat{\beta}$:

- 1) Si $(\underline{X}^T \underline{X})^{-1}$ existe entonces $\hat{\beta} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{y}$
- 2) Si $(\underline{X}^T \underline{X})^{-1}$ no existe, se puede usar la pseudoinversa. Si $A \in \mathbb{R}^{m \times n}$, entonces su pseudoinversa es una matriz $A^+ \in \mathbb{R}^{n \times m}$ que satisface
 - i) $A A^+ A = A$
 - ii) $A^+ A A^+ = A^+$
 - iii) $(A^+ A)^T = A^+ A$

$(\underline{X}^T \underline{X}) \underline{\beta} = \underline{X}^T \underline{y}$ entonces $\hat{\beta} = (\underline{X}^T \underline{X})^+ \underline{X}^T \underline{y}$

es la solución de mínima norma. Observamos que $(\underline{X}^T \underline{X})^{-1}$ puede no existir si hay más variables que observaciones o si hay variables linearmente independientes.

Además se puede definir $\text{Cond}(A) = \|A\| \cdot \|A^+\|$ donde $\|\cdot\|$ es una norma matricial. Si $\text{Cond}(A)$ es muy grande, errores pequeños en los datos pueden tener a grandes errores en la solución.

3) Una alternativa numéricamente eficiente es la descomposición QR. Se puede mostrar que si $A \in \mathbb{R}^{n \times n}$ entonces $A = QR$, con $Q \in \mathbb{R}^{n \times n}$ orthonormal (i.e. $Q^T Q = I_{n \times n}$) y R matriz triangular superior.

Si $A \in \mathbb{R}^{m \times n}$, entonces $A = QR$, con $Q \in \mathbb{R}^{m \times m}$ y $R \in \mathbb{R}^{m \times n}$ triangular superior.

Se puede escribir $R = \begin{bmatrix} R_1 \\ \Theta \end{bmatrix}$, con $R_1 \in \mathbb{R}^{n \times n}$ triangular superior y $\Theta \in \mathbb{R}^{(m-n) \times n}$ matriz de ceros.

También se puede escribir $Q = [Q_1 \ Q_2]$, con $Q_1 \in \mathbb{R}^{m \times n}$ y $Q_2 \in \mathbb{R}^{m \times (m-n)}$. Así $QR = [Q_1 \ Q_2] \begin{bmatrix} R_1 \\ \Theta \end{bmatrix} = Q_1 R_1$.

Esta descomposición no es única. La descomposición QR se puede hallar usando el algoritmo de Gramm-Schmidt.

Para los modelos lineales se tiene $\underline{X} = QR$ con $Q \in \mathbb{R}^{n \times n}$ y $R = \begin{bmatrix} R_1 \\ \Theta \end{bmatrix}$ con $R_1 \in \mathbb{R}^{(p+1) \times (p+1)}$ y $\Theta \in \mathbb{R}^{(n-p-1) \times (p+1)}$ matriz de ceros.

Observemos que $\underline{X}^T \underline{X} \underline{\beta} = \underline{X}^T \underline{y}$ se puede escribir como $R^T Q^T Q R \underline{\beta} = R^T Q \underline{y}$.

Así $R^T R \underline{\beta} = R^T Q^T \underline{y}$, que se puede reescribir como

$$\begin{bmatrix} R_1^T & Q^T \end{bmatrix} \begin{bmatrix} R_1 \\ \Theta \end{bmatrix} \underline{\beta} = \begin{bmatrix} R_1^T & \Theta^T \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} \underline{y}$$

De esta forma $R_1^T R_1 \underline{\beta} = R_1^T Q_1^T \underline{y}$. Si se define

$c = R_1 \underline{\beta}$ y $b = R_1^T Q_1^T \underline{y}$ entonces el problema se puede resolver en dos pasos:

i) $R_1^T c = b$ se resuelve para c (fácilmente porque R_1^T es matriz triangular inferior).

ii) $R_1 \underline{\beta} = c$ se resuelve para $\underline{\beta}$ (fácilmente porque R_1 es matriz triangular superior)

Casos maravillosos:

1) Si $\hat{\underline{\beta}} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{y}$ entonces $\hat{\underline{y}} = \underline{X} \hat{\underline{\beta}}$ es $\hat{\underline{y}} = \underline{X} (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{y}$. Si se define $H = \underline{X} (\underline{X}^T \underline{X})^{-1} \underline{X}^T$, entonces $\hat{\underline{y}} = H \underline{y}$.

A H se le llama matriz sombra porque tiene un " $\hat{\cdot}$ " sobre \underline{y} ($\hat{\underline{y}} = H \underline{y}$). H es simétrica y $H^T H = H$.

$H \in \mathbb{R}^{n \times n}$. Además $I_{n \times n} - H$ también es simétrica e idempotente.

$$2) \underline{y} - \hat{\underline{y}} = \underline{y} - H \underline{y} = (I_{n \times n} - H) \underline{y} = \underline{0}$$

Luego $\underline{Y}^T \underline{Y} = \underline{Y}^T (I_{n \times n} - H) (I_{n \times n} - H) \underline{Y} = \underline{Y}^T (I_{n \times n} - H)^2 \underline{Y}$.
 Este segundo resultado quiere decir que se pueden conocer los errores del modelo sin conocer los coeficientes.

3) Si se reemplaza \underline{Y} por $\underline{X}\hat{\beta}$ y luego $\hat{\beta}$ por $(\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y}$ en la función de costo se tiene:

$$\begin{aligned}
 & \underline{Y}^T \underline{Y} - 2 \hat{\beta}^T \underline{X}^T \underline{Y} + \hat{\beta}^T \underline{X}^T \underline{X} \hat{\beta} = \\
 &= \underline{Y}^T \underline{Y} - 2 \underline{Y}^T \underline{X} (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y} + \\
 &\quad \underline{Y}^T \underline{X} (\underline{X}^T \underline{X})^{-1} \cancel{(\underline{X}^T \underline{X})} \cancel{(\underline{X}^T \underline{X})}^T \underline{X}^T \underline{Y} \\
 &= \underline{Y}^T \underline{Y} - 2 \underline{Y}^T H \underline{Y} + \underline{Y}^T H \underline{Y} = \underline{Y}^T \underline{Y} - \underline{Y}^T H \underline{Y} \\
 &= \underline{Y}^T (I_{n \times n} - H) \underline{Y} \\
 &= \underline{Y}^T (\underline{Y} - \hat{\underline{Y}}) = \underline{Y}^T \underline{e}
 \end{aligned}$$

Luego el valor mínimo del error cuadrático medro también se puede conocer antes de calcular $\hat{\beta}$.

Propiedades estadísticas:

Si $\underline{Y}_i = \beta_0 + \underline{X}_i^T \underline{\beta} + \epsilon_i$, con $\epsilon_i \sim \text{iid } N(0, \sigma^2)$

entonces $\underline{Y} = \underline{X} \underline{\beta} + \underline{\epsilon}$, con $\underline{\epsilon} \sim N(\underline{0}, \sigma^2 \underline{I}_{n \times n})$.

Luego $\underline{Y} \sim N(\underline{X} \underline{\beta}, \sigma^2 \underline{I}_{n \times n})$.

Como $\hat{\underline{\beta}} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y}$, entonces $\hat{\underline{\beta}}$ sigue una distribución normal p+1 variancia con

$$E[\hat{\underline{\beta}}] = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{X} \underline{\beta} = \underline{\beta}$$

$$\begin{aligned} D(\hat{\underline{\beta}}) &= (\underline{X}^T \underline{X})^{-1} \underline{X}^T \sigma^2 \underline{I}_{n \times n} \underline{X} (\underline{X}^T \underline{X})^{-1} \\ &= \sigma^2 (\underline{X}^T \underline{X})^{-1} = C \end{aligned}$$

$$\underline{\epsilon} = C^{-1/2} (\hat{\underline{\beta}} - \underline{\beta}) \sim N(\underline{0}, \underline{I}_{(p+1) \times (p+1)})$$

y $\underline{\epsilon}^T \underline{\epsilon} \sim \chi^2_{p+1}$. Para probar $H_0: \underline{\beta} = \underline{0}$ vs. $\underline{\beta} \neq \underline{0}$ se utliza $\chi^2_0 = C^{-1/2} \hat{\underline{\beta}}^T \hat{\underline{\beta}}$ que bajo H_0 se distribuye χ^2_{p+1} .

Además, como $\underline{r} = (\underline{I}_{n \times n} - H) \underline{Y}$

$$D(\underline{r}) = (\underline{I}_{n \times n} - H) \sigma^2 (\underline{I}_{n \times n} - H) = \sigma^2 (\underline{I}_{n \times n} - H)$$

$$\text{Luego } \text{Var}(r_i) = \text{Var}(Y_i - \hat{Y}_i) = \sigma^2 (1 - h_{ii}).$$

Un estimador de σ^2 es $\hat{\sigma}^2 = \frac{\underline{r}^T \underline{r}}{n-p-1}$

Leverage

$$\frac{\partial \hat{Y}_i}{\partial Y_i} = \frac{\partial}{\partial Y_i} H Y = H, \text{ bds } H = (h_{ij})_{ij} \text{ es tal}$$

que $\frac{\partial \hat{Y}_i}{\partial Y_i} = h_{ii}$. A h_{ii} se le denomina leverage.

Un valor de h_{ii} muy alto implica que pequeñas cambios en Y_i provocan grandes cambios en \hat{Y}_i .

Los puntos con alto leverage son sospechosos de ser atípicos. Un alto leverage indica un valor extremo para Y_i .

Puntos influyentes

Son aquellos que tienen un alto impacto en $\hat{\beta}$. Se pueden encontrar con la distancia de Cook D_i :

$$D_i = \frac{\sum_{j=1}^n (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p s^2}, \quad s^2 = \frac{\underline{y}^T \underline{y}}{n-p},$$

y $\hat{Y}_{j(i)}$ es la predicción para Y_j a partir de un modelo estimado con la muestra $\underline{Z} \setminus h(\underline{x}_i, Y_i)$.

Residuales

$$r_i = Y_i - \hat{Y}_i. \quad E(r_i) = \underline{X}^T \hat{\beta} - \underline{X}^T \beta = 0$$

$$\text{Var}(r_i) = \sigma^2(1-h_{ii}), \text{ bds } r_i \sim N(0, \sigma^2(1-h_{ii}))$$

Se define el residual estudentizado t_i como

$$t_i = \frac{e_i}{\hat{\sigma} \sqrt{1-h_{ii}}} , \quad \hat{\sigma}^2 = \frac{\underline{x}^T \underline{x}}{n-p-1} , \quad \hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

Usando t_i , puede escribirse D_i como $D_i = \frac{t_i^2}{p+1} \times \frac{h_{ii}}{1-h_{ii}}$

Diagnosticos

1. Residuales (y) vs. Ajustados (\bar{x})
2. Normal Q-Q plot: percentiles de los residuales estandarizados vs. los de la normal estandarizada
3. Localización - Escala: $\sqrt{\text{residuales estandarizados}} (y)$ vs. Ajustados (\bar{x})
4. Residuales (y) vs. Leverage (\bar{x}): t_i vs. h_{ii} estandarizados
5. Distancia de Cook

Pruebas de Hipótesis

1. Significancia de la regresión

Total Sum of Squares SS_T

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ &\quad + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^n t_i (\hat{y}_i - \bar{y}) = \sum_{i=1}^n t_i (\hat{\beta}_0 + \underline{x}_i^T \hat{\beta}_{-x} - \bar{y}) \\ &= \underline{x}^T (\hat{\underline{y}} - \bar{y} \underline{1}_n) = \underbrace{\underline{x}^T \underline{x}}_O(a) \hat{\beta} - \underbrace{\frac{\underline{x}^T \underline{1}_n}{n}}_O(b) \cdot \bar{y} \end{aligned}$$

a) De las ecuaciones normales se tiene:

$$0 = -\underline{X}^T \underline{y} + \underline{X}^T \underline{X} \hat{\beta} \Rightarrow -\underline{X}^T \underline{y} + \underline{X}^T \hat{\underline{y}} = 0$$

$$\Rightarrow \underline{X}^T (\underline{y} - \hat{\underline{y}}) = 0 \Rightarrow \boxed{\underline{X}^T \underline{1} = 0} \text{ ó } \boxed{\underline{1}^T \underline{X} = 0^T}$$

b) Como $\underline{X}^T \underline{1} = 0$, entonces $\underline{X}_i^T \underline{1} = 0$, donde \underline{X}_i es la i -ésima columna de \underline{X} . En particular $\underline{X}_1 = \underline{1}_n$, luego $\underline{1}_n^T \underline{1} = 0$

De esta manera la suma de cuadrados totales se puede escribir como

$$\begin{aligned} SST &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= SSR + SSE \end{aligned}$$

SST tiene $n-1$ grados de libertad

SSR (suma de cuadrados del modelo) tiene p grados de libertad

SSE (suma de cuadrados del error) tiene $n-p-1$ grados de libertad

$$\begin{aligned} \text{En forma vectorial } SST &= \sum_{i=1}^n (y_i - \bar{y})^2 = (\underline{y} - \bar{y} \underline{1}_n)^T (\underline{y} - \bar{y} \underline{1}_n) \\ &= \underline{y}^T \underline{y} - \bar{y} \underline{y}^T \underline{1}_n - \bar{y} \underline{1}_n^T \underline{y} + \bar{y}^2 \underline{1}_n^T \underline{1}_n \\ &= \underline{y}^T \underline{y} - \bar{y} n \bar{y} - \bar{y} n \bar{y} + \bar{y}^2 n \\ (\ast) \quad &= \underline{y}^T \underline{y} - n \bar{y}^2 = \underline{y}^T \underline{y} - \frac{1}{n} \underline{y}^T \underline{1}_n \underline{1}_n^T \underline{y} \\ &= \underline{y}^T \underline{y} - \frac{1}{n} \underline{y}^T J_{n \times n} \underline{y} = \underline{y}^T (I_{n \times n} - \frac{1}{n} J_{n \times n}) \underline{y}, \end{aligned}$$

con $J_{n \times n} = \underline{1}_n \underline{1}_n^T$.

(*) Observemos que $\bar{Y} = \frac{\underline{Y}^T \underline{1}_n}{n}$, luego

$$\bar{Y}^2 = \frac{\underline{Y}^T \underline{1}_n}{n} \frac{(\underline{1}_n^T \underline{Y})}{n} = \frac{\underline{Y}^T \underline{1}_n \underline{1}_n^T \underline{Y}}{n^2}$$

en forma vectorial SSE es

$$\begin{aligned} & \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \underline{Y}^T \underline{Y} \\ &= \underline{Y}^T (\mathbf{I}_{n \times n} - \mathbf{H}) (\mathbf{I} - \mathbf{H}) \underline{Y} \\ &= \underline{Y}^T (\mathbf{I}_{n \times n} - \mathbf{H}) \underline{Y} \end{aligned}$$

en forma vectorial SSR es

$$\begin{aligned} & \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = (\hat{\underline{Y}} - \bar{\hat{\underline{Y}}} \underline{1}_n)^T (\hat{\underline{Y}} - \bar{\hat{\underline{Y}}} \underline{1}_n) \\ &= \hat{\underline{Y}}^T \hat{\underline{Y}} - \bar{\hat{\underline{Y}}} \hat{\underline{Y}}^T \underline{1}_n - \bar{\hat{\underline{Y}}} \underline{1}_n^T \hat{\underline{Y}} + \bar{\hat{\underline{Y}}}^2 \underline{1}_n^T \underline{1}_n \\ &= \hat{\underline{Y}}^T \mathbf{H} \hat{\underline{Y}} - 2 \bar{\hat{\underline{Y}}} \hat{\underline{Y}}^T \underline{1}_n + n \bar{\hat{\underline{Y}}}^2 \\ &= \hat{\underline{Y}}^T \mathbf{H} \hat{\underline{Y}} - 2n \bar{\hat{\underline{Y}}}^2 + n^2 \bar{\hat{\underline{Y}}} \\ &= \hat{\underline{Y}}^T \left(\mathbf{H} - \frac{1}{n} \mathbf{J}_{n \times n} \right) \hat{\underline{Y}} \end{aligned}$$

(*)

(*) Se verifica el hecho de que $\underline{Y}^T \underline{1}_n = \hat{\underline{Y}}^T \underline{1}_n$, lo que se deduce del hecho de que $\underline{Y}^T \underline{1}_n = 0$

Así, $SST = SSR + SSE$ implica que

$$\underline{Y}^T \left(\mathbf{I}_{n \times n} - \frac{1}{n} \mathbf{J}_{n \times n} \right) \underline{Y} = \hat{\underline{Y}}^T \left(\mathbf{H} - \frac{1}{n} \mathbf{J}_{n \times n} \right) \hat{\underline{Y}} + \underline{Y}^T \left(\mathbf{I}_{n \times n} - \mathbf{H} \right) \underline{Y}$$

- SST se interpreta como la desviación de los Y_i de su valor promedio
- SSR se interpreta como la diferencia entre las predicciones del modelo oficial y aquellas del modelo nulo (solo tiene d intercepto $\hat{\beta}_0 = \bar{Y}$)

• SSE es la diferencia entre los valores predichos por el modelo y los observados.

Un valor de SSR pequeño indica que el modelo no hace mucho compromiso con el modelo nulo.

Se pueden definir $MST = \frac{SST}{n-1}$, $MSR = \frac{SSR}{p}$ y

$$MSE = \frac{SSE}{n-p-1}.$$

El MSE es un estimador de σ^2 .

Cuando $\beta_0 \approx 0$, entonces $\hat{Y}_i = \beta_0 = \bar{Y}$ y $SSR \approx 0$, luego $SST \approx SSE$. Se define el estadístico $F_0 = MSR/MSE$ para probar la hipótesis nula $H_0: \beta_0 = 0$ vs. $H_a: \beta_0 \neq 0$.

Se rechaza H_0 para valores grandes de F_0 . Bajo H_0 $F_0 \sim F_{p, n-p-1}$. Luego se rechaza H_0 a nivel α si $F_0 > F_{1-\alpha, p, n-p-1}$

Se pueden probar las p hipótesis $H_{0,i}: \beta_i = 0$ vs $H_{a,i}: \beta_i \neq 0$.

Para ello se define el estadístico $t_i = \frac{\hat{\beta}_i}{SD(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\sqrt{\frac{1}{n-2} (\hat{X}^T \hat{X})_{ii}}}$, con $t_i \sim t_{n-p-1}$.

la matriz H

$$H = \underline{X} (\underline{X}^T \underline{X})^{-1} \underline{X}^T \in \mathbb{R}^{n \times n}$$

$$\text{i) } H = H^T \quad \text{y} \quad \text{ii) } H^T H = H.$$

$$\begin{aligned} \text{iii) } \text{Tr}(H) &= \text{Tr}(\underline{X} (\underline{X}^T \underline{X})^{-1} \underline{X}^T) \\ &= \text{Tr}((\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{X}) = p+1 \\ \Rightarrow \sum_{i=1}^n h_{ii} &= p+1 \end{aligned}$$

Para $A \in \mathbb{R}^{n \times n}$, $A \times A = A^2 = B = (b_{ij})_{ij}$

con $b_{ij} = \sum_{k=1}^n a_{ik} a_{kj}$. Si $i=j$, entonces

$b_{ii} = \sum_{k=1}^n a_{ik} a_{ki}$. Si A es simétrica entonces

$$a_{ik} = a_{ki} \text{ y } b_{ii} = \sum_{k=1}^n a_{ik}^2 = a_{ii}^2 + \sum_{k \neq i} a_{ik}^2.$$

$$\text{Si } A = AA, \text{ entonces } a_{ii} = a_{ii}^2 + \sum_{k \neq i} a_{ik}^2.$$

Como $\sum_{k \neq i} a_{ik}^2 \geq 0$ entonces $a_{ii} \geq a_{ii}^2$, de donde $0 \leq a_{ii} \leq 1$.

De lo anterior se tiene que $0 \leq h_{ii} \leq 1$ y

$$\sum_{i=1}^n h_{ii} = p+1.$$

¿Cuál es el efecto de tener observaciones duplicadas?

$$\begin{aligned} x_i = \underline{x}_j &\Rightarrow \hat{y}_i = \hat{y}_j = \hat{\beta}_0 + \underline{x}_i^T \hat{\beta}_{\underline{x}} = \hat{\beta}_0 + \underline{x}_i^T \hat{\beta}_{\underline{x}} \\ &= [1 \ \underline{x}_i^T] \hat{\beta} = [1 \ \underline{x}_j^T] \hat{\beta}. \end{aligned}$$

Si $\underline{x}_{i,\perp} = \begin{bmatrix} 1 \\ \underline{x}_i \end{bmatrix}$ y $\underline{x}_{j,\perp} = \begin{bmatrix} 1 \\ \underline{x}_j \end{bmatrix}$, entonces

$$h_{ij} = \underline{x}_{i,\perp} (\underline{X}^T \underline{X})^{-1} \underline{x}_{j,\perp}.$$

Pero si $\underline{x}_{i,A} = \underline{x}_{j,A}$, entonces $h_{ij} = h_{ii} = h_{jj}$ y se tiene que $h_{ii} = h_{ii}^2 + \sum_{k \neq i} h_{ik} = 2h_{ii}^2 + \sum_{k \neq i, j} h_{ik}$, de donde $h_{ii} \leq 2h_{ii}^2$ y así $h_{ii}(1-2h_{ii}) \leq 0$, de donde $h_{ii} \leq \frac{1}{2}$. Es decir que al duplicar un vector de características su leverage se divide por 2. Observemos que se asume que $\underline{x}_i = \underline{x}_j$, pero no que $y_i = y_j$

El R^2 , C_p , AIC y BIC

Se define como $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$

El R^2 ajustado es $R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$

$$C_p = \frac{1}{n} (SSE + 2p \hat{\sigma}^2)$$

$$AIC = \frac{1}{n} (SSE + 2p \hat{\sigma}^2)$$

$$BIC = \frac{1}{n} (SSE + \ln(n)p \hat{\sigma}^2)$$

$$\hat{\sigma}^2 = \frac{SSE}{n-p-1}$$

Inflación de la varianza

$$VI(\hat{\beta}_j) = \frac{1}{1 - R^2_{x_j | X_{-j}}}$$

$R^2_{x_j | X_{-j}}$ es el R^2 de la regresión donde x_j es la variable respuesta y todos los otros predictores son variables explicativas