Capítulo 3: Regresión Múltiple

Fernando A. Crespo R.

14 de Octubre de 2021

Índice

3.1 Matriz de Correlación

- Lo primero es estudiar la relación entre variables. Para ello se estudia la matriz de correlación. Donde se despliega los coeficientes de correlación para cada par de variables.
- Multicolinealidad: cuando dos variables presentan una alta correlación entre ellas. Ya que eso significa que hay una relación directa entre las dos variables. Por lo tanto, es díficil después, imputar cuanto aporta una variable para explicar la variación de otra. Dos variables con correlación alta no proporcionan información adicional.
- Regla para elegir:
  1. No debe haber correlación alta entre variables predictoras o explicativas (nosotros las denominamos como x).

3.1 Marco Teórico: Modelo Clásico de Regresión Múltiple Lineal

- El vector columna $\mathbf{x_j}$ contiene las n observaciones con $k = 1, \ldots, k$, añadamos el vector columna a la matriz $\mathbf{X} \in \mathbf{M_{n \times k+1}}(\mathbb{R})$, donde la primera columna es de unos ($\mathbf{x_1} = \mathbb{1}_\mathbf{n}$). Puede escribirse como:

$$\mathbf{y} = \mathbb{1}_\mathbf{n}\beta_\mathbf{0} + \mathbf{x_1}\beta_\mathbf{1} + \mathbf{x_2}\beta_\mathbf{2} \ldots + \mathbf{x_k}\beta_\mathbf{k} + \epsilon, \tag{1}$$

  o como:

$$\mathbf{y} = \beta_\mathbf{0} + \mathbf{x_1}\beta_\mathbf{1} + \mathbf{x_2}\beta_\mathbf{2} \ldots + \mathbf{x_k}\beta_\mathbf{k} + \epsilon. \tag{2}$$

ϵ es la perturbación aleatoria.

3.3 Supuestos

- Supuesto 1:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon. \tag{3}$$

con

$$\mathbf{X} = [\mathbf{1l_n}\ \mathbf{x_1}\ \mathbf{x_2}\ \ldots\ \mathbf{x_k}]. \tag{4}$$

- Supuesto 2: $\mathbf{X}$ es una matriz $\in M_{n\times k+1}(IR)$ de rango completo, esta es la condición de identificación.
- Supuesto 3:

$$\mathbb{E}[\epsilon|\mathbf{X}] = \begin{bmatrix} IE[\epsilon_1|X] \\ IE[\epsilon_2|X] \\ \vdots \\ IE[\epsilon_n|X] \end{bmatrix} = \mathbf{0}. \tag{5}$$

3.3 Supuestos

- Se pide además:

$$\mathbf{Var}[\epsilon_\mathbf{j}|\mathbf{X}] = \sigma^\mathbf{2}, \tag{6}$$

para $i = 1, \ldots, n$, y

$$\mathbf{Cov}[\epsilon_\mathbf{i}, \epsilon_\mathbf{j}|\mathbf{X}] = \mathbf{0}, \tag{7}$$

para $i \neq j$.

- La varianza constante se denomina homocedasticidad, y en caso de que varíe por variable, se denomina heterocedasticidad.
(7) se denomina la no autocorrelación, o de forma equivalente, que el experimento se hizo con elección aleatoria de los datos.

3.3 Supuestos

- Supuesto 4:

$$\mathbf{E}[\epsilon\epsilon^{\mathbf{t}}|\mathbf{X}] = \sigma^{\mathbf{2}}\mathbf{I}. \tag{8}$$

- Supuesto 5: $\mathbf{X}$ es una matriz conocida de $M_{n\times k+1}(IR)$ de constantes (no es estocástica).
- Supuesto 6:

$$\epsilon|\mathbf{X} \sim \mathbf{N}[\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I}]. \tag{9}$$

3.4 Estimadores de la Regresión

- Los estimadores se pueden obtener por dos vías:
  - Obteniendo el Estimador Máximo Verosímil dado los supuestos sobre los errores, u
  - Obteniendo los mínimos cuadráticos.
- Vamos a obtener los estimadores a partir de minimizar:

$$\mathbf{S}(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^{\mathbf{t}}(\mathbf{Y} - \mathbf{X}\beta) \in \mathbb{R}. \tag{10}$$

- Desarrollando (10), obtenemos:

$$\mathbf{S}(\beta) = \mathbf{Y}^{\mathbf{t}}\mathbf{Y} - \mathbf{2}\mathbf{Y}^{\mathbf{t}}\mathbf{X}\beta + \beta^{\mathbf{t}}(\mathbf{X}^{\mathbf{t}}\mathbf{X})\beta. \tag{11}$$

- Para minimizar necesitamos $\mathbf{S}'(\beta)$ e igualamos a $\mathbf{0}$:

$$\mathbf{S}'(\beta) = -\mathbf{2}\mathbf{Y}^{\mathbf{t}}\mathbf{X} + \mathbf{2}\beta^{\mathbf{t}}(\mathbf{X}^{\mathbf{t}}\mathbf{X}) = \mathbf{0}. \tag{12}$$

3.4 Estimadores de la Regresión

- Como $\mathbf{X} \in \mathbf{M_{n \times k+1}}(\mathbb{R})$ es de rango completo:

$$\hat{\beta} = (\mathbf{X^t X})^{-1} \mathbf{X^t Y}. \tag{13}$$

- El valor estimado $\hat{\mathbf{Y}}$, se calcula:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X^t X})^{-1} \mathbf{X^t Y} \in \mathbb{R}. \tag{14}$$

- Usando (14) podemos obtener la estimación de los errores $\hat{\epsilon}$:

$$\hat{\epsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}(\mathbf{X^t X})^{-1} \mathbf{X^t Y} = (\mathbf{I} - \mathbf{X}(\mathbf{X^t X})^{-1} \mathbf{X^t})\mathbf{Y}. \tag{15}$$

3.4 Estimadores de la Regresión

- Además de (15):

$$\mathbf{X^t}\hat{\epsilon} = \mathbf{X^t}(\mathbf{I} - \mathbf{X}(\mathbf{X^t}\mathbf{X})^{-1}\mathbf{X^t})\mathbf{Y} = \mathbf{0}, \tag{16}$$

- y :

$$\hat{\mathbf{Y}}^{\mathbf{t}}\hat{\epsilon} = \hat{\beta}\mathbf{X^t}\hat{\epsilon} = \mathbf{0}. \tag{17}$$

- La variación cuadrática del error $\hat{\epsilon}$:

$$\hat{\epsilon}^{\mathbf{t}}\hat{\epsilon} = \mathbf{Y^t}(\mathbf{I} - \mathbf{X}(\mathbf{X^t}\mathbf{X})^{-1}\mathbf{X^t})\mathbf{Y} = \mathbf{Y^t}\mathbf{Y} - \mathbf{Y^t}\mathbf{X}\hat{\beta}. \tag{18}$$

3.5 Coeficiente de Determinación

- De (18), tenemos:

$$\mathbf{Y^t Y} = (\hat{\mathbf{Y}} + \hat{\epsilon})^{\mathbf{t}}(\hat{\mathbf{Y}} + \hat{\epsilon}) = \hat{\mathbf{Y}}^{\mathbf{t}}\hat{\mathbf{Y}} + \hat{\epsilon}^{\mathbf{t}}\hat{\epsilon}, \tag{19}$$

- y de las varianzas:

$$\sum_{\mathbf{j=1}}^{\mathbf{n}}(\mathbf{Y_j} - \overline{\mathbf{Y}})^{\mathbf{2}} = \sum_{\mathbf{j=1}}^{\mathbf{n}}(\hat{\mathbf{Y}}_{\mathbf{j}} - \overline{\mathbf{Y}})^{\mathbf{2}} + \sum_{\mathbf{j=1}}^{\mathbf{n}}\hat{\epsilon}_{\mathbf{j}}^{\mathbf{2}}. \tag{20}$$

- El coeficiente de determinación:

$$\mathbf{R^2} = \mathbf{1} - \frac{\sum_{\mathbf{j=1}}^{\mathbf{n}}\hat{\epsilon}_{\mathbf{j}}^{\mathbf{2}}}{\sum_{\mathbf{j=1}}^{\mathbf{n}}(\mathbf{Y_j} - \overline{\mathbf{Y}})^{\mathbf{2}}} = \frac{\sum_{\mathbf{j=1}}^{\mathbf{n}}(\hat{\mathbf{Y}}_{\mathbf{j}} - \overline{\mathbf{Y}})^{\mathbf{2}}}{\sum_{\mathbf{j=1}}^{\mathbf{n}}(\mathbf{Y_j} - \overline{\mathbf{Y}})^{\mathbf{2}}}. \tag{21}$$

3.5 Coeficiente de Determinación

- De (21), el estadístico para el test de Hipótesis Nula:

$$\mathbf{H_0} : \beta_1 = \ldots = \beta_k = \mathbf{0}, \tag{22}$$

- Es una F de Fisher de k y $n-k-1$ grados de libertad:

$$\mathbf{F} = \frac{\frac{\mathbf{R^2}}{\mathbf{k}}}{\frac{\mathbf{1-R^2}}{\mathbf{n-k-1}}}. \tag{23}$$

- Rechazo Hipótesis Nula si, con α el nivel de significación:

$$\mathbb{P}(\mathbf{F} \geq \mathbf{f}) \leq \alpha. \tag{24}$$

3.5 Coeficiente de Determinación

- Si $\mathbf{R^2 = 0}$, significa que las variables no influyen en el modelo.
- Si $\hat{\beta} = (\mathbf{X^t X})^{-1}\mathbf{X^t Y}$ entonces tenemos:
  - $\mathbf{E}(\hat{\beta}) = \beta$, y
  - $\mathbf{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X^t X})^{-1}$.
- Para $\hat{\epsilon}$ tenemos:
  - $\mathbf{E}(\hat{\epsilon}) = \mathbf{0}$, y
  - $\mathbf{Cov}(\hat{\epsilon}) = \sigma^2[\mathbf{I} - \mathbf{X}(\mathbf{X^t X})^{-1}\mathbf{X^t}]$.
- Luego de $\mathbf{E}(\hat{\epsilon}^{\mathbf{t}}\hat{\epsilon}) = (\mathbf{n - k - 1})\sigma^2$ tenemos:

$$\mathbf{s^2} = \frac{\hat{\epsilon}^{\mathbf{t}}\hat{\epsilon}}{\mathbf{n - k - 1}} = \frac{\mathbf{Y^t}[\mathbf{I} - \mathbf{X}(\mathbf{X^t X})^{-1}\mathbf{X^t}]\mathbf{Y}}{\mathbf{n - k - 1}}. \tag{25}$$

3.5 Coeficiente de Determinación

- Para ver cuanto se ajusta el modelo, se puede ver el gráfico de distribución de los errores.
- Para ver el test de Hipótesis sobre los estimadores:

$$\begin{array}{ll} H_0 & \beta_j = \beta_j^* \\ H_1 & \beta_j \neq \beta_j^* \end{array} , \tag{26}$$

donde el valor β_j^*, es el valor particular que deseamos observar.

- Para ello observamos la variable aleatoria U_j, que tiene una distribución t_{n-k-1}:

$$\mathbf{U_j} = \frac{\hat{\beta}_\mathbf{j} - \beta_\mathbf{j}^*}{\widehat{\sigma_{\hat{\beta}_\mathbf{j}}}} = \frac{\hat{\beta}_\mathbf{j} - \beta_\mathbf{j}^*}{\sqrt{\widehat{\mathbf{Var}(\hat{\beta}_\mathbf{j})}}} = \frac{\hat{\beta}_\mathbf{j} - \beta_\mathbf{j}^*}{\sqrt{\mathbf{s^2}((\mathbf{X^t X})^{-1})_{\mathbf{jj}}}}. \tag{27}$$

3.5 Coeficiente de Determinación

- Rechazamos la hipótesis nula cuando $|U_j| > c$, con c elegido desde un nivel de significación α_0.
- Para un valor u_j observado, el valor de la cola esta dado por:

$$P(U_j > |u_j|) + P(U_j < -|u_j|) = p_{value}, \tag{28}$$

  por lo tanto, rechazamos la hipótesis nula cuando $p_{value} \leq \alpha_0$.
- Para obtener un intervalo de confianza usamos:

$$\hat{\beta}_{\mathbf{j}} \pm \left(\mathbf{t_{n-k-1}}\left(\frac{\alpha}{\mathbf{2}}\right)\right)\sqrt{\mathbf{s^2}((\mathbf{X^tX})^{-1})_{\mathbf{jj}}}. \tag{29}$$

3.5 Coeficiente de Determinación

- Intervalo de confianza para la varianza del error:

$$\left(\frac{(n-k-1)s^2}{\chi^2_{n-k-1}(1-\frac{\alpha}{2})}, \frac{(n-k-1)s^2}{\chi^2_{n-k-1}(\frac{\alpha}{2})} \right). \tag{30}$$

3.5 Coeficiente de Determinación

- Para estimar un valor, dado un vector x_0, incluye el valor 1 si el modelo incluye la constante, en caso contrario no.

$$\hat{y}_0 = x_0^t \hat{\beta}. \tag{31}$$

Tenemos el valor:

$$h_0 = x_0^t (X^t X)^{-1} x_0. \tag{32}$$

El intervalo de confianza es:

$$\hat{y}_0 \pm s\sqrt{1 + h_0} \cdot t_{n-k-1} \left(1 - \frac{\alpha}{2}\right). \tag{33}$$

3.5 Coeficiente de Determinación

- Estadístico de Dubin Watson para ver autocorrelación entre el error:

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_i^2}. \tag{34}$$

El valor debe ser cercano a 2, para saber si es simétrico, si esta cerca de 4 la autocorrelación es negativa, y si está cerca de 0, está correlacionado positivamente.

3.5 Coeficiente de Determinación

- Test de Breusch-Pagan :

$$H_0 : \textit{Los errores tienen varianza constante}. \tag{35}$$

El valor del test es mayor que 0.05, entonces podemos asumir que los datos tienen varianza constante.