Clase 3: Modelos lineales

Justo Andrés Manrique Urbina

3 de septiembre de 2019

1. Leverage

Sea $h_{ii} = x_i^T(X^TX)^{-1}x_{ii}$ para el modelo lineal simple. Se tiene entonces que $h_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}}$. En el modelo lineal múltiple se tiene lo siguiente:

$$h_{ii} = \frac{1}{n} + (x_i - \bar{x})(\tilde{X}^T\tilde{X})^{-1}(x_i - \bar{x}).$$

en dónde:

$$X = \begin{pmatrix} (X_{11} - \bar{X}_1) & (X_{12} - \bar{X}_1) & \dots & (X_{1k} - \bar{X}_1) \\ \vdots & \vdots & \vdots & \vdots \\ (X_{n1} - \bar{X}_k) & (X_{n2} - \bar{X}_k) & \dots & (X_{nk} - \bar{X}_k) \end{pmatrix}$$

En este modelo se está excluyendo el intercepto.

2. Distancia de Cook

Recordemos que $\hat{B} \sim N(B, \sigma^2(X^TX)^{-1})$. Posteriormente lo estandarizamos a:

$$\frac{(\hat{B} - B)^T(X^TX)^{-1}(\hat{B} - B)}{\sigma^2} \sim X_p^2.$$

Para estimar el efecto de la i-ésima observación en la estimación de B. Definimos $\hat{B}$ como el estimador de mínimos cuadrados ordinarios y $\hat{B}_1$ como el estimador de mínimos cuadrados ordinarios eliminando la estimación i-ésima. La distancia entonces de define como:

$$D_i = \frac{(\hat{B}_i - \hat{B})^T(X^TX)^{-1}(\hat{B}_i - \hat{B})}{P\hat{\sigma}^2}.$$

$$D_i = \frac{(\hat{Y}_i - \hat{Y})^T(\hat{Y}_i - \hat{Y})}{P\hat{\sigma}^2}.$$

Por otro lado, también podemos definir la distancia de Cook como lo siguiente:

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1 - h_{ii}}.$$

3. Gráfico de variable adicional o residuos de regresión parcial

Definamos como:

$$Y = XB + \varepsilon.$$

Particionamos X en dos matrices: una que contenga X_j y otra que no contenga $X_j = X_{(j)}$. Lo mismo para B. Entonces tendremos:

$$Y = X_{(j)}B_{(j)} + X_jB_j + \varepsilon.$$

$$(1 - H_{(j)})Y = (1 - H_{(j)})(X_{(j)}B_{(j)} + X_jB_j + \varepsilon).$$

$$\varepsilon(y|X_{(j)}) = e(X_j|X_{(j)}) + \varepsilon^*.$$

Entonces se tiene que:

$$e(y|X_{(j)}) = B_je(X_j|X_{(j)}) + \varepsilon^*.$$

Si se grafica esto, la pendiente es B_j.

4. Diagnóstico

- **Supuesto de normalidad:** Revisión del qqplot de t_i (residuos estudentizados).
- **Homocedasticidad:** Gráfico de residuos estudentizados vs valores ajustados.

5. Criterios de información: Evaluación de modelos

$$AIC = -2l(\hat{B}, \hat{\sigma}^2) + 2(p + 1).$$

En dónde $p+1$ es el número de parámetros y $\hat{l}$ es la logverosimilitud evaluado en el EMV. La $\hat{l}$ corresponde al ajuste del modelo y $p + 1$ a la complejidad.