Intervalos de confianza para el modelo de regresión lineal múltple

En los siguiente suponemos que $\boldsymbol{y} = N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$.

Regiones de confianza para $\boldsymbol{\beta}$

Consideramos el estadístico de prueba, para $H_0 \ : \ \boldsymbol{C\beta} = \boldsymbol{t}$ como en el notebook de pruebas de hipótesis, $F = \frac{SSH/q}{SSE/(n-k-1)} \sim F(q, n-k-1, \lambda_1)$ con $(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t})'\left(\boldsymbol{C}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{C}'\right)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t})$ y $SSE = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. Si tomamos $\boldsymbol{C} = \boldsymbol{I}$ y $\boldsymbol{t} = \boldsymbol{\beta}$ obtenemos que

$$\mathbb{P}\left[\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{s^2(k+1)} \leq F_{\alpha, k+1, n-k-1}\right] = 1 - \alpha$$

con $s^2 = SSE/(n-k-1)$. Obtenemos la región de confianza

$$\left\{\boldsymbol{\beta}\,;\ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq (k+1)s^2 F_{\alpha, k+1, n-k-1}\right\}.$$

Para $k = 1$ está región puede ser evaluada en dos dimensiones como una elipse pero para $k > 1$ la región elipsoidal es difícil de evaluar; por lo que alternativamente consideramos intervalos de confianza para los $\beta_j' s$ en $\boldsymbol{\beta}$.

Intervalos de confianza para β_j

Recordamos que al final del notebook sobre pruebas de hipótesis vimos que con $g_{j,j}$ el j-ésimo elemento diagonal de $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$, entonces

$$\frac{\hat{\beta}_j}{s\sqrt{g_{j,j}}} \sim t(n-k-1, \beta_j/g_{j,j}).$$

Por lo que

$$t_j = \frac{\left(\hat{\beta}_j - \beta_j\right)}{s\sqrt{g_{j,j}}} \sim t(n-k-1).$$

De lo cual obtenemos que

$$\mathbb{P}\left[-t_{\alpha/2, n-k-1} \leq \frac{\left(\hat{\beta}_j - \beta_j\right)}{s\sqrt{g_{j,j}}} \leq t_{\alpha/2, n-k-1}\right] = 1 - \alpha,$$

equivalentemente

$$\mathbb{P}\left[\hat{\beta}_j - t_{\alpha/2, n-k-1}\, s\sqrt{g_{j,j}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2, n-k-1}\, s\sqrt{g_{j,j}}\right] = 1 - \alpha,$$

Por lo que obtenemos el intervalo de confianza al nivel $1 - \alpha$ para β_j

$$\left(\hat{\beta}_j - t_{\alpha/2, n-k-1}\, s\sqrt{g_{j,j}}\ ,\ \hat{\beta}_j + t_{\alpha/2, n-k-1}\, s\sqrt{g_{j,j}}\right)$$

Intervalos de confianza para $\boldsymbol{\alpha}'\boldsymbol{\beta}$

En el notebook para pruebas de hipotesis vimos que

$$\frac{\left(\boldsymbol{a}'\hat{\boldsymbol{\beta}}\right)^2}{s^2\boldsymbol{a}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{a}} \sim F\left(1, n-k-1, \frac{\left(\boldsymbol{a}'\boldsymbol{\beta}\right)^2}{2\sigma^2\boldsymbol{a}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{a}}\right)$$

Por lo que

$$F = \frac{\left(\boldsymbol{a}'\hat{\boldsymbol{\beta}} - \boldsymbol{a}'\boldsymbol{\beta}\right)^2}{s^2\boldsymbol{a}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{a}} \sim F\left(1, n-k-1\right)$$

y al tener 1 y $n-k-1$ grados de libertad podemos considerar alternativamente

$$t = \frac{\boldsymbol{a}'\hat{\boldsymbol{\beta}} - \boldsymbol{a}'\boldsymbol{\beta}}{s\sqrt{\boldsymbol{a}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{a}}} \sim t(n-k-1).$$

De lo cual, procediendo como en el caso anterior, obtenemos el intervalo de confianza al nivel $1-\alpha$ para $\boldsymbol{\alpha}'\boldsymbol{\beta}$

$$\left(\boldsymbol{\alpha}'\hat{\boldsymbol{\beta}} - t_{\alpha/2,n-k-1}\, s\sqrt{\boldsymbol{a}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{a}}\,,\ \boldsymbol{\alpha}'\hat{\boldsymbol{\beta}} + t_{\alpha/2,n-k-1}\, s\sqrt{\boldsymbol{a}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{a}}\right)$$

Intervalos de confianza para $\mathbb{E}[y]$

Consideramos $\boldsymbol{x}_0 = (1, x_{0,1}, \ldots, x_{0,k})$ el vector de covariables asociada a una nueva observación $y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + \epsilon_0$. Queremos encontrar un intervalo de confianza para $\mathbb{E}[y_0] = \boldsymbol{x}_0'\boldsymbol{\beta}$. Tenemos que $\boldsymbol{x}_0'\boldsymbol{\beta}$ es de la forma considerada anteriormente por lo que el intervalo de confianza al nivel $1-\alpha$ para $\mathbb{E}[y_0]$ está dado por

$$\left(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - t_{\alpha/2,n-k-1}\, s\sqrt{\boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0}\,,\ \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + t_{\alpha/2,n-k-1}\, s\sqrt{\boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0}\right)$$

Podemos expresar lo anterior en términos del modelo centrado:

$$\mathbb{E}[y_0] = \alpha + \boldsymbol{\beta}_1'(\boldsymbol{x}_{0_1} - \bar{\boldsymbol{x}}_1)$$
$$\hat{\mathbb{E}[y_0]} = \bar{y} + \hat{\boldsymbol{\beta}}_1'(\boldsymbol{x}_{0_1} - \bar{\boldsymbol{x}}_1)$$
$$\left(\bar{y} + \hat{\boldsymbol{\beta}}_1'(\boldsymbol{x}_{0_1} - \bar{\boldsymbol{x}}_1) - t_{\alpha/2,n-k-1}\, s\sqrt{\frac{1}{n} + (\boldsymbol{x}_{0_1} - \bar{\boldsymbol{x}}_1)'\left(\boldsymbol{X}_c'\boldsymbol{X}_c\right)^{-1}(\boldsymbol{x}_{0_1} - \bar{\boldsymbol{x}}_1)}\,,\ \bar{y} + \hat{\boldsymbol{\beta}}_1'(\boldsymbol{x}_{0_1} + \bar{\boldsymbol{x}}_1) + t_{\alpha/2,n-}\right.$$

En el caso de regresión lineal simple tenemos

$$\mathbb{E}[y_0] = \beta_0 + \beta_1 x_0$$
$$\hat{\mathbb{E}[y_0]} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$
$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{\alpha/2,n-2}\, s\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}\,,\ \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{\alpha/2,n-2}\, s\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}\right)$$

Intervalos de predicción

Anteriormente calculamos intervalos de confianza para el valor esperado de una observación nueva y_0 con covariable x_0, esto es distinto a querer calcular un intervalo de confianza para la cantidad aleatoria y_0 asociada a un covariable x_0. Ahora nos interesamos en esta última cuestión, la cual involucra un problema de predicción. Ya que y_0 es una variable aleatoria y no un parámetro del modelo, hablamos de intervalos de predicción en vez de intervalos de confianza.

Otra vez podemos considerar el estadístico $\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$, y un estadístico de prueba basado en $y_0 - \hat{y}_0$. Observamos que y_0 y $\hat{y}_0$ se distribuyen normal, $\mathbb{E}\left[\$y_0 - \hat{y}_0\right] = 0$ y

$$\begin{aligned} \operatorname{var}\left(y_0 - \hat{y}_0\right) &= \operatorname{var}\left(\boldsymbol{x}_0'\boldsymbol{\beta} + \epsilon - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}\right) \\ &= \operatorname{var}(\epsilon) + \operatorname{var}\left(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}\right) \\ &= \sigma^2 + \operatorname{var}\left(\boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}\right) \\ &= \sigma^2 + \sigma^2\boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X})^{-1}\boldsymbol{x}_0 \\ &= \sigma^2\left(1 + \boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0\right) \end{aligned}$$

Por tener observaciones i.i.d. en el modelo sabemos que SSE es independiente de y_0 y vimos que SSE es independiente de $\hat{\boldsymbol{\beta}}$, vea por ejemplo el teorema de pruba de hipóitesis más general en el notebook anterior, por lo que s^2 es independiente de $\left(y_0 - \hat{y}_0\right)$ y podemos considerar el siguiente estadístico de prueba con distribución t:

$$t = \frac{\left(y_0 - \hat{y}_0\right)}{s\sqrt{1 + \boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0}} \sim t\left(n - k - 1\right)$$

El intervalo de predicción al nivel $1 - \alpha$ entonces está dado por

$$\left(\hat{y}_0 - t_{\alpha/2,n-k-1}s\sqrt{1 + \boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0}\hat{y}_0 + t_{\alpha/2,n-k-1}s\sqrt{1 + \boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0}\right),$$

en términos del modelo centrado el intervalo es

$$\left(\bar{y} + \hat{\boldsymbol{\beta}}_1'(\boldsymbol{x}_{0_1} - \bar{\boldsymbol{x}}_1) - t_{\alpha/2,n-k-1}s\sqrt{1 + \frac{1}{n} + (\boldsymbol{x}_{0_1} - \bar{\boldsymbol{x}}_1)'\left(\boldsymbol{X}_c'\boldsymbol{X}_c\right)^{-1}(\boldsymbol{x}_{0_1} - \bar{\boldsymbol{x}}_1)}, \bar{y} + \hat{\boldsymbol{\beta}}_1'(\boldsymbol{x}_{0_1} + \bar{\boldsymbol{x}}_1) - t_{\alpha/}\right.$$

Lo cual en el modelo de regresión lineal se reduce a

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{\alpha/2,n-2}s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{\alpha/2,n-2}s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$$

Intervalo de confianza para σ^2

Sabemos que $\frac{(n-k-1)s^2}{\sigma^2} \sim \chi^2(n-k-1)$ por lo que para $\chi^2_{\alpha,\nu}$ el quantil superior a nivel α de la distribución Ji-cuadrada con ν grados de libertad

$$\mathbb{P}\left[\chi^2_{1-\alpha/2,n-k-1} \leq \frac{(n-k-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-k-1}\right] = 1-\alpha;$$

de lo cual obtenemos el intervalo de confianza al nivel $1-\alpha$ para σ^2

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

In [ ]: