Intervalos Simultáneos y Regiones de Confianza

Método de Bonferroni

Queremos hallar intervalos de confianza para q combinaciones lineales de la forma $\mathbf{c}_i'\boldsymbol{\beta}$ $i = 1, 2, \ldots, q$.

Bajo normalidad, para cada combinación lineal el intervalo de la forma

$$\mathbf{c}_i'\widehat{\boldsymbol{\beta}} \pm t_{n-r,\delta/2}\widehat{\sigma}_{\mathbf{c}_i'\widehat{\boldsymbol{\beta}}}$$

tiene nivel $1 - \delta$.

Definamos los eventos

$$E_i : \mathbf{c}_i'\boldsymbol{\beta} \text{ pertenece al intervalo } i$$

tenemos que $P(E_i) = 1 - \delta$

Luego,

$$\begin{aligned} 1 - \alpha &= P(\text{todos los intervalos son correctos}) = P(\cap_{i=1}^{q} E_i) \\ &= 1 - P((\cap_{i=1}^{q} E_i)^c) = 1 - P(\cup_{i=1}^{q} E_i^c) \end{aligned}$$

$$\geq 1 - \sum_{i=1}^{q} P(E_i^c) = 1 - q\delta$$

Así, por ejemplo si cada intervalo tiene nivel 0.95 ($\delta = 0.05$) y $q = 10$ tendríamos que

$$1 - \alpha \geq 1 - q\delta = 1 - 10 * 0.05 = 0.50$$

¿ Cómo podríamos mejorar esto?

Si cada $\delta = \dfrac{\alpha}{q}$, entonces preservaríamos un nivel global superior a $1 - \alpha$.

Una clara desventaja de este método es que si q es grande al exigir que cada intervalo tenga nivel $1 - \frac{\alpha}{q}$, podemos obtener intervalos muy anchos y por lo tanto, de escaso valor práctico.

Método de Scheffé

Supondremos s.p.g. que $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_q$ son l.i. Sea $\Psi = \mathbf{C}\boldsymbol{\beta}$, donde $\mathbf{C} \in I\!R^{q\times p}$. Inicialmente supondremos que $\text{rg}(\mathbf{X}) = p$. En este caso, sabemos que

$$\frac{(\widehat{\Psi} - \Psi)'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\widehat{\Psi} - \Psi)}{qs^2} \sim \mathcal{F}_{q,n-p}$$

entonces

$$\begin{aligned}
1-\alpha &= P(\mathcal{F}_{q,n-p} \leq F_{q,n-p,\alpha}) \\
&= P((\widehat{\Psi} - \Psi)'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\widehat{\Psi} - \Psi) \leq qs^2 F_{q,n-p,\alpha}) \\
&= P((\widehat{\Psi} - \Psi)'\mathbf{L}^{-1}(\widehat{\Psi} - \Psi) \leq m) \\
&= P(\mathbf{b}\mathbf{L}^{-1}\mathbf{b} \leq m)
\end{aligned}$$

Recordemos que dada $\mathbf{L}$ una matriz definida positiva tenemos que

$$\sup_{\mathbf{h}\neq\mathbf{0}} \frac{(\mathbf{h}'\mathbf{b})^2}{\mathbf{h}'\mathbf{L}\mathbf{h}} = \mathbf{b}'\mathbf{L}^{-1}\mathbf{b}$$

con lo cual, tenemos

$$\begin{aligned} 1-\alpha &= P(\sup_{\mathbf{h}\neq\mathbf{0}} \frac{(\mathbf{h}'\mathbf{b})^2}{\mathbf{h}'\mathbf{L}\mathbf{h}} \leq m) \\ &= P(\frac{(\mathbf{h}'\mathbf{b})^2}{\mathbf{h}'\mathbf{L}\mathbf{h}} \leq m \quad \forall \mathbf{h}\neq\mathbf{0}) \\ &= P(\frac{|\mathbf{h}'\widehat{\Psi} - \mathbf{h}'\Psi|}{s(\mathbf{h}'\mathbf{L}\mathbf{h})^{1/2}} \leq \sqrt{qF_{q,n-p,\alpha}} \quad \forall \mathbf{h}\neq\mathbf{0}) \end{aligned}$$

Luego, para cualquier función lineal $\mathbf{h}'\Psi$ tenemos el intervalo de confianza

$$\mathbf{h}'\widehat{\Psi} \pm \sqrt{qF_{q,n-p,\alpha}}\, s(\mathbf{h}'\mathbf{L}\mathbf{h})^{1/2}$$

siendo la probabilidad total de la clase $1-\alpha$.

Supongamos que $r = p$ y $\mathbf{C} = I_p$, en ese caso tendríamos

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq ps^2 F_{p,n-p,\alpha}$$

que define lo que se conoce como el elipsoide de confianza.

¿Cómo es en el caso general en el que rg$(\mathbf{X}) = r$?

Tenemos que $\mathbf{c}_1'\boldsymbol{\beta}, \mathbf{c}_2'\boldsymbol{\beta}, \ldots, \mathbf{c}_q'\boldsymbol{\beta}$ son l.i. Sea $\Psi = \mathbf{C}\boldsymbol{\beta}$, donde $\mathbf{C} \in I\!R^{q\times p}$, rg$(\mathbf{C}) = q$.

Recordemos que

$$\frac{(\widehat{\Psi} - \Psi)'\mathbf{B}^{-1}(\widehat{\Psi} - \Psi)}{qs^2} \sim \mathcal{F}_{q,n-r}$$

donde $\widehat{\Psi} \sim N(\Psi, \Sigma_{\widehat{\Psi}})$, $\Sigma_{\widehat{\Psi}} = \sigma^2\mathbf{B} = \sigma^2\mathbf{A}^*\mathbf{A}^{*'}$.

Como rg$(\mathbf{C}) = q$, entonces $\mathbf{B}$ tiene inversa, por lo tanto

$$\begin{aligned}
1-\alpha &= P((\widehat{\Psi} - \Psi)'\mathbf{B}^{-1}(\widehat{\Psi} - \Psi) \leq qs^2F_{q,n-r,\alpha}) \\
&= P((\widehat{\Psi} - \Psi)'\mathbf{B}^{-1}(\widehat{\Psi} - \Psi) \leq m) \\
&= P(\sup_{\mathbf{h}\neq\mathbf{0}} \frac{(\mathbf{h}'\mathbf{b})^2}{\mathbf{h}'\mathbf{B}\mathbf{h}} \leq m) \\
&= P(\frac{|\mathbf{h}'\widehat{\Psi} - \mathbf{h}'\Psi|}{s(\mathbf{h}'\mathbf{B}\mathbf{h})^{1/2}} \leq \sqrt{qF_{q,n-r,\alpha}} \quad \forall \mathbf{h} \neq \mathbf{0})
\end{aligned}$$

De esta forma,

$$\mathbf{h}'\widehat{\Psi} \pm \sqrt{qF_{q,n-r,\alpha}}\; s(\mathbf{h}'\mathbf{B}\mathbf{h})^{1/2}$$

resulta un intervalo de confianza para la función lineal $\mathbf{h}'\Psi$ y la probabilidad total de la clase es $1-\alpha$. Observemos que este intervalo es de la forma:

$$\mathbf{h}'\widehat{\Psi} \pm \sqrt{qF_{q,n-r,\alpha}}\; \widehat{\sigma}_{\mathbf{h}'\widehat{\Psi}}$$

Volvamos al ejemplo de Biomasa

```
> cor(xx)
           BIO           K         NAA          PH         SAL          ZN
BIO  1.0000000 -0.20511626 -0.27206950  0.77418613 -0.10316780 -0.62440784
  K -0.2051163  1.00000000  0.79213460  0.01869352 -0.02049881  0.07396686
NAA -0.2720695  0.79213460  1.00000000 -0.03771997  0.16226567  0.11704693
 PH  0.7741861  0.01869352 -0.03771997  1.00000000 -0.05133280 -0.72216711
SAL -0.1031678 -0.02049881  0.16226567 -0.05133280  1.00000000 -0.42083353
 ZN -0.6244078  0.07396686  0.11704693 -0.72216711 -0.42083353  1.00000000

Análisis con todas las variables: lm(formula = BIO ~ K + NAA + PH + SAL + ZN)
                Value Std. Error    t value   Pr(>|t|)
(Intercept)  1252.5895  1234.7294     1.0145     0.3166
          K    -0.2853     0.3483    -0.8191     0.4177
        NAA    -0.0087     0.0159    -0.5438     0.5897
         PH   305.4821    87.8831     3.4760     0.0013
        SAL   -30.2881    24.0298    -1.2604     0.2150
         ZN   -20.6784    15.0544    -1.3736     0.1774

Residual standard error: 398.3 on 39 degrees of freedom
Multiple R-Squared: 0.6773
F-statistic: 16.37 on 5 and 39 degrees of freedom, the p-value is 1.082e-008
```

```
lm(formula = BIO ~ K + PH + SAL + ZN)
                Value Std. Error    t value   Pr(>|t|)
(Intercept)  1505.4479  1133.6647     1.3279     0.1917
          K    -0.4388     0.2023    -2.1688     0.0361
         PH   293.8169    84.4685     3.4784     0.0012
        SAL   -35.9374    21.4758    -1.6734     0.1021
         ZN   -23.4497    14.0396    -1.6703     0.1027

Residual standard error: 394.7 on 40 degrees of freedom
Multiple R-Squared: 0.6749
F-statistic: 20.76 on 4 and 40 degrees of freedom, the p-value is 2.525e-009

lm(formula = BIO ~ K + PH + SAL)
               Value Std. Error   t value  Pr(>|t|)
(Intercept) -131.1184  582.5120    -0.2251    0.8230
          K   -0.4900    0.2043    -2.3985    0.0211
         PH  410.1454   48.8253     8.4003    0.0000
        SAL  -12.0533   16.3687    -0.7364    0.4657

Residual standard error: 403.3 on 41 degrees of freedom
Multiple R-Squared: 0.6522
F-statistic: 25.63 on 3 and 41 degrees of freedom, the p-value is 1.682e-009
```

```
lm(formula = BIO ~ K + PH)
               Value Std. Error   t value  Pr(>|t|)
(Intercept) -506.7131  279.8016    -1.8110    0.0773
          K   -0.4871    0.2031    -2.3977    0.0210
         PH  411.9779   48.4954     8.4952    0.0000

Residual standard error: 401.1 on 42 degrees of freedom
Multiple R-Squared: 0.6476
F-statistic: 38.59 on 2 and 42 degrees of freedom, the p-value is 3.074e-010
```

Los intervalos de confianza de nivel individual 95 % obtenidos a partir del último modelo ajustado serían tal como vimos de la forma

$$\widehat{\beta}_i \pm t_{42,0.025} \quad t_{42,0.025} = 2.018$$

En este caso resultan:

$$\begin{aligned} -1.072 &< \beta_0 < 58 \\ 314 &< \beta_1 < 510 \\ -0.898 &< \beta_2 < 0.077 \end{aligned}$$

Si los calculamos con el método de Bonferroni como para que el nivel global resulte 95 % usaríamos $t_{42,0.025/3} = 2.50$ y estos resultan

$$\begin{aligned} \text{-}1.206 &< \beta_0 < 192 \\ 291 &< \beta_1 < 533 \\ \text{-}0.995 &< \beta_2 < 0.021 \end{aligned}$$

La región de confianza obtenida a partir de método de Scheffé sería

Joint Confidence Region

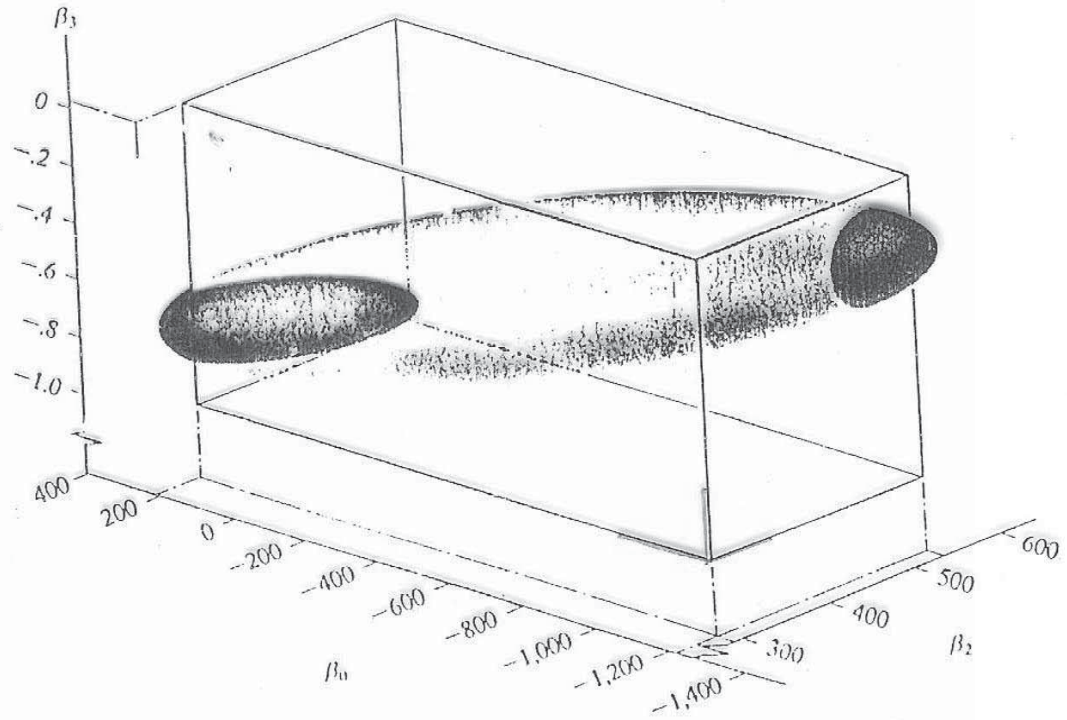

Figure 5.1 Three-dimensional 95% joint confidence region for β_0, β_2, and β_3. The intersection of the Bonferroni confidence intervals is shown as the box.

Comparación entre los métodos

Se puede ver que si las q combinaciones son l.i. entonces

$$t_{\nu,\frac{\alpha}{2q}} < \sqrt{qF_{q,\nu,\alpha}}$$

Por ejemplo, si $\alpha = 0{,}05$, $q = 5$ y $n = 26$, entonces

$$\sqrt{qF_{q,\nu,\alpha}} = 3{,}68 \quad t_{\nu,\frac{\alpha}{2q}} = 2{,}85$$

En general, si se quieren realizar intervalos simultáneos para k funciones paramétricas de las cuales q son l.i.,para $\alpha = 0{,}05$ se puede ver que si $q \leq k$ y k no mucho mas grande que q, entonces

$$t_{\nu,\frac{\alpha}{2k}} < \sqrt{qF_{q,\nu,\alpha}}$$

Cuando k es mucho más grande que q, entonces la desigualdad se invierte.

Relación entre el tests de F y el método de Scheffé

Los intervalos

$$\mathbf{h}'\widehat{\Psi} \pm \sqrt{qF_{q,n-r,\alpha}}\, s(\mathbf{hBh})^{1/2} \quad (*)$$

y el test de F para chequear $H : \Psi = \delta$ están relacionados.

El test de F **no** es significativo al nivel α si y sólo si

$$\frac{(\widehat{\Psi} - \delta)'\mathbf{B}^{-1}(\widehat{\Psi} - \delta)}{qs^2} \leq F_{q,n-r,\alpha}$$

que es cierto si y sólo si $\Psi = \delta$ está en la región $(\widehat{\Psi} - \Psi)'\mathbf{B}^{-1}(\widehat{\Psi} - \Psi) \leq m$, o sea si y sólo si $\mathbf{h}'\delta$ está contenido en $(*)$.

Dicho de otra forma, F es significativo si uno o más intervalos $(*)$ no contienen a $\mathbf{h}'\delta$, el problema es identificar cuál de las combinaciones lineales es la que no está contenida.

Coeficiente de Correlación Múltiple (o coeficiente de determinación)

Supongamos que tenemos el modelo

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{ip-1} + \epsilon_i$$

y nos interesa testear

$$H : \beta_1 = \ldots = \beta_{p-1} = 0$$

Consideremos Ω y $\omega = \Omega \cap H$. Llamaremos $\widehat{\boldsymbol{\eta}}$ a la proyección de Y sobre el subespacio asociado a Ω y $\widehat{\boldsymbol{\eta}}_\omega$ a la proyección sobre el subespacio asociado a ω.

¿Cuál es la correlación entre el vector de observaciones Y y el vector de predichos $\widehat{Y}$ (o $\widehat{\boldsymbol{\eta}}$) ?

$$R = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(\widehat{y}_i - \overline{\widehat{y}})}{\{\sum_{i=1}^{n} (y_i - \overline{y})^2 \sum_{i=1}^{n} (\widehat{y}_i - \overline{\widehat{y}})^2\}^{1/2}}$$

Recordemos que cuando hay ordenada al origen, tenemos que

$$\frac{\partial}{\partial \beta_0}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2 = \sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_{p-1}x_{ip-1}) = \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2 = 0$$

entonces

$$\overline{y} = \overline{\widehat{y}}$$

y en consecuencia

$$R = \frac{\sum_{i=1}^{n}(y_i-\overline{y})(\widehat{y}_i-\overline{y})}{\{\sum_{i=1}^{n}(y_i-\overline{y})^2 \sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2\}^{1/2}}$$

Visto en términos de proyecciones, tendríamos

$$R = \frac{\langle y-\widehat{\boldsymbol{\eta}}_\omega, \widehat{\boldsymbol{\eta}}-\widehat{\boldsymbol{\eta}}_\omega\rangle}{\|y-\widehat{\boldsymbol{\eta}}_\omega\|\|\widehat{\boldsymbol{\eta}}-\widehat{\boldsymbol{\eta}}_\omega\|}$$

Como

$$\begin{aligned}\langle Y-\widehat{\boldsymbol{\eta}}_\omega, \widehat{\boldsymbol{\eta}}-\widehat{\boldsymbol{\eta}}_\omega\rangle &= \langle Y-\widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\eta}}-\widehat{\boldsymbol{\eta}}_\omega\rangle + \langle \widehat{\boldsymbol{\eta}}-\widehat{\boldsymbol{\eta}}_\omega, \widehat{\boldsymbol{\eta}}-\widehat{\boldsymbol{\eta}}_\omega\rangle \\ &= \|\widehat{\boldsymbol{\eta}}-\widehat{\boldsymbol{\eta}}_\omega\|^2\end{aligned}$$

obtenemos que

$$R = \frac{\|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2}{\|Y - \widehat{\boldsymbol{\eta}}_\omega\|^2}$$
$$= \frac{\text{Suma Cuadrados Total Regresión}}{\text{Suma Cuadrados Total Corregida}}$$

es decir

$$R = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Obtendremos el siguiente resultado:

Teorema: Supongamos que deseamos testear $H : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, que no involucra

al intercept β_0. Sea

$$R_\omega = \frac{\sum_{i=1}^{n} (\widehat{y}_{i\omega} - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

entonces el estadístico F para testear H será

$$F = \frac{(R^2 - R^2_\omega)(n-p)}{(1-R^2)q}$$