Ejemplo:

Significación de la Regresión. Tabla de Análisis de la Varianza

Supongamos que tenemos el modelo con intercept dado por

$$E(\mathbf{Y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1}$$

y queremos testear

$$H_o : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

de manera que $\omega = \Omega \cap H$. H impone $p-1$ restricciones l.i. Trataremos el caso en que $\text{rg}(\mathbf{X}) = p$

¿Quién es $\mathcal{V}_\omega$?

$dim(\mathcal{V}_\omega) = r - (p-1) = p - (p-1) = 1$ y tenemos que $\mathcal{V}_1 \in \mathcal{V}_p$

¿Quién es $\widehat{\boldsymbol{\eta}}_\omega$?

Bajo ω, $\beta_1 = \ldots = \beta_{p-1} = 0, \quad E(\mathbf{Y}) = \beta_0$.

Tenemos que:

$$\mathbf{X}_\omega = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix} \rightarrow \widehat{\beta_0} = (\mathbf{X}'_\omega \mathbf{X}_\omega)^{-1} \mathbf{X}'_\omega \mathbf{Y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{\mathbf{Y}}.$$

Luego:

$$\widehat{\boldsymbol{\eta}}_\omega = \mathbf{X}'_\omega \widehat{\beta_0} = \begin{pmatrix} \overline{\mathbf{Y}} \\ \overline{\mathbf{Y}} \\ . \\ . \\ \overline{\mathbf{Y}} \end{pmatrix}$$

Además:

$$\|\mathbf{Y}\|^2 = \|\mathbf{Y} - \widehat{\boldsymbol{\eta}}\|^2 + \|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2 + \|\widehat{\boldsymbol{\eta}}_\omega\|^2$$

Bajo Ω si $\text{rg}(\mathbf{X}) = p$

$$\widehat{\beta_\Omega} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \rightarrow \widehat{\boldsymbol{\eta}} = \mathbf{P}\mathbf{Y} \text{ donde } \mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

En efecto, $\widehat{\boldsymbol{\eta}}_\omega$ es la proyección ortogonal de $\widehat{\boldsymbol{\eta}}$ sobre $\mathcal{V}_\omega = \mathcal{V}_1$. Si fuera así, entonce $\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega \perp \widehat{\boldsymbol{\eta}}_\omega$.

$$\widehat{\boldsymbol{\eta}} = \mathbf{PY} \text{ y } \widehat{\boldsymbol{\eta}}_\omega = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-\mathbf{1}}\mathbf{1}'\mathbf{Y} = \mathbf{P_1 Y}$$

luego,

$$(\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega)'\widehat{\boldsymbol{\eta}}_\omega = \mathbf{Y}'(\mathbf{P} - \mathbf{P}_\omega)\mathbf{P}_\omega\mathbf{Y} = \mathbf{Y}'(\mathbf{P}\mathbf{P}_\omega - \mathbf{P}'_\omega\mathbf{P}_\omega)\mathbf{Y} = \mathbf{Y}'(\mathbf{P}_\omega - \mathbf{P}_\omega)'\mathbf{P}_\omega\mathbf{Y} = \mathbf{0}$$

$$\|\mathbf{Y}\|^2 = \|\mathbf{Y} - \widehat{\boldsymbol{\eta}}\|^2 + \|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2 + \|\widehat{\boldsymbol{\eta}}_\omega\|^2$$

Llamaremos

$\|\mathbf{Y}\|^2$: suma de cuadrados total

$\|\mathbf{Y} - \widehat{\boldsymbol{\eta}}\|^2$: suma de cuadrados residual

$\|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2$: suma de cuadrados de la regresión

$\|\mathbf{Y} - \widehat{\boldsymbol{\eta}}_\omega\|^2$: suma de cuadrados total corregida

Tenemos las siguientes igualdades

$$\begin{aligned}
\|\mathbf{Y}\|^2 &= \mathbf{Y}'\mathbf{Y} && \text{g.l}=n \\
\|\mathbf{Y} - \widehat{\boldsymbol{\eta}}\|^2 &= \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}\mathbf{Y} && \text{g.l}=n-p \\
\|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2 &= \widehat{\boldsymbol{\beta}}'\mathbf{X}\mathbf{Y} - n(\overline{\mathbf{Y}})^2 && \text{g.l}=p-1 \\
\|\mathbf{Y} - \widehat{\boldsymbol{\eta}}_\omega\|^2 &= \mathbf{Y}'\mathbf{Y} - n(\overline{\mathbf{Y}})^2 && \text{g.l}=n-1
\end{aligned}$$

Si quisiéramos verificar la significación de la regresión, haríamos

$$F = \frac{\|\widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_\omega\|^2/p - 1}{\|\mathbf{Y} - \widehat{\boldsymbol{\eta}}\|^2/n - p}$$

Muchos programas ofrecen en su salida una tabla como la que sigue

Fuente		g.l.	M.S.	F	p-valor
Regresión	$\|\widehat{\boldsymbol{\eta}}\|^2 - n(\overline{\mathbf{Y}})^2$	$p-1$	$(1) = \frac{\|\widehat{\boldsymbol{\eta}}\|^2 - n(\overline{\mathbf{Y}})^2}{p-1}$		
Residual	$\|\mathbf{Y} - \widehat{\boldsymbol{\eta}}\|^2$	$n-p$	$(2) = \frac{\|\mathbf{Y} - \widehat{\boldsymbol{\eta}}\|^2}{n-p}$	$(1)/(2)$	
Tot. Cor.	$\|\mathbf{Y}\|^2 - n(\overline{\mathbf{Y}})^2$	$n-1$			

Cuadro 1: Tabla de ANOVA

Datos de Biomasa

Producción de biomasa en el estuario de Cape Fear: los datos corresponden a un estudio de la Universidad de North Carolina en el que se muestrearon 3 tipos de vegetación en tres localidades. En cada una se muestreó al azar 5 lugares con un total de 45 observaciones. Analizaremos las variables del sustrato:

x_1=SAL: Salinidad

x_2=pH: Acidez

x_3= K: Potasio

x_4=Naa: Sodio

x_5=Zn: Zinc

y: Biomasa Aérea

En esta etapa nos concentraremos en identificar aquellas variables que muestran mayor relación con y. Ajustaremos el modelo

$$E(y) = \beta_0 + \beta_1 SAL + \beta_2 pH + \beta_3 K + \beta_4 Naa + \beta_5 Zn$$

SALIDA DE S-PLUS

DATOS DE BIOMASA

```
> sal.lm

Call:
lm(formula = BIO ~ ., data = bio)

Coefficients:
 (Intercept)          K          NAA       PH       SAL        ZN
    1252.589 -0.2853166 -0.008662343 305.4821 -30.28808 -20.67844

Degrees of freedom: 45 total; 39 residual
Residual standard error: 398.2671


> summary(sal.lm)

Call: lm(formula = BIO ~ ., data = bio)
Residuals:
    Min     1Q Median    3Q  Max
 -748.1 -223.7 -85.22 139.1 1072
```

```
Coefficients:
                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


Correlation of Coefficients:
    (Intercept)       K     NAA      PH     SAL
  K -0.3122
NAA  0.3767     -0.8103
 PH -0.8406      0.1212 -0.2442
SAL -0.9180      0.3047 -0.4324  0.6045
 ZN -0.8809      0.1908 -0.3386  0.8350  0.7113
```

SALIDA DE SX

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P	VIF
CONSTANT	1252.59	1234.73	1.01	0.3166	
K	-0.28532	0.34832	-0.82	0.4177	3.0
NAA	-0.00866	0.01593	-0.54	0.5897	3.3
PH	305.482	87.8831	3.48	0.0013	3.3
SAL	-30.2881	24.0298	-1.26	0.2150	2.2
ZN	-20.6784	15.0544	-1.37	0.1774	4.3

R-SQUARED	0.6773	RESID. MEAN SQUARE (MSE)	158617
ADJUSTED R-SQUARED	0.6360	STANDARD DEVIATION	398.267

SOURCE	DF	SS	MS	F	P
REGRESSION	5	1.298E+07	2596983	16.37	0.0000
RESIDUAL	39	6186050	158617		
TOTAL	44	1.917E+07			

CASES INCLUDED 45 MISSING CASES 0

Hipótesis Anidadas

La interpretación del test de F en términos de las hipótesis anidadas.

Supongamos que tenemos $H_1, H_2, \ldots, H_k$ un conjunto de hipótesis que imponen $q_1, q_2, \ldots, q_k$ restricciones independientes, respectivamente. Luego, las $q_1+q_2+\ldots+q_k$ funciones estimables son l.i. La secuencia de hipótesis anidadas estará dada por

$$\Omega, \omega_1 = \Omega \cap H_1, \omega_2 = \Omega \cap H_1 \cap H_2, \ldots, \omega_k = \Omega \cap H_1 \cap H_2 \ldots \cap H_k$$

Si lamamos $\mathcal{V}^{(j)}$ a los espacios asociados cada uno de dimensión $r - q_1 - q_2 - \ldots - q_j$

$$\mathcal{V}^{(r)} \supset \mathcal{V}^{(r-q_1)} \supset \ldots \mathcal{V}^{(r-q_1-q_2-\ldots-q_k)}$$

Sea $\widehat{\boldsymbol{\eta}}_{\omega_j}$ la proyección ortogonal de Y sobre $\mathcal{V}^{(j)}$, por lo tanto tenemos que

$$Y = Y - \widehat{\boldsymbol{\eta}} + \widehat{\boldsymbol{\eta}} - \widehat{\boldsymbol{\eta}}_{\omega_1} + \widehat{\boldsymbol{\eta}}_{\omega_1} - \widehat{\boldsymbol{\eta}}_{\omega_2} + \ldots + \widehat{\boldsymbol{\eta}}_{\omega_{k-1}} - \widehat{\boldsymbol{\eta}}_{\omega_k} + \widehat{\boldsymbol{\eta}}_{\omega_k}$$

y en consecuencia

$$\|Y\|^2 = \|Y-\widehat{\boldsymbol{\eta}}\|^2+\|\widehat{\boldsymbol{\eta}}-\widehat{\boldsymbol{\eta}}_{\omega_1}\|^2+\|\widehat{\boldsymbol{\eta}}_{\omega_1}-\widehat{\boldsymbol{\eta}}_{\omega_2}\|^2+\ldots+\|\widehat{\boldsymbol{\eta}}_{\omega_{k-1}}-\widehat{\boldsymbol{\eta}}_{\omega_k}\|^2+\|\widehat{\boldsymbol{\eta}}_{\omega_k}\|^2$$

donde cada suma tiene una distribución χ^2 no central bajo Ω con $n-r$, q_1, q_2, $\ldots$, q_k, $r - q_1 - q_2 - \ldots - q_k$ grados de libertad.

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,

$$1 - \alpha \; = \; P(\text{todos los intervalos son correctos}) = P(\cap_{i=1}^{q} E_i)$$

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

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

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

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

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

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

$$\widehat{\beta}_i \pm t_{42,0.025}\widehat{\sigma}_{\widehat{\beta}_i} \quad \text{siendo } t_{42,0.025} = 2.018$$

En este caso resultan:

$$\begin{aligned} \text{-1.072} &< \beta_0 < 58 \\ 314 &< \beta_{PH} < 510 \\ \text{-0.898} &< \beta_K < \text{-0.077} \end{aligned}$$

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

$$\begin{aligned} \text{-}1.206 &< \beta_0 < 192 \\ 291 &< \beta_{PH} < 533 \\ \text{-}0.995 &< \beta_K < 0.021 \end{aligned}$$

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

Joint Confidence Region

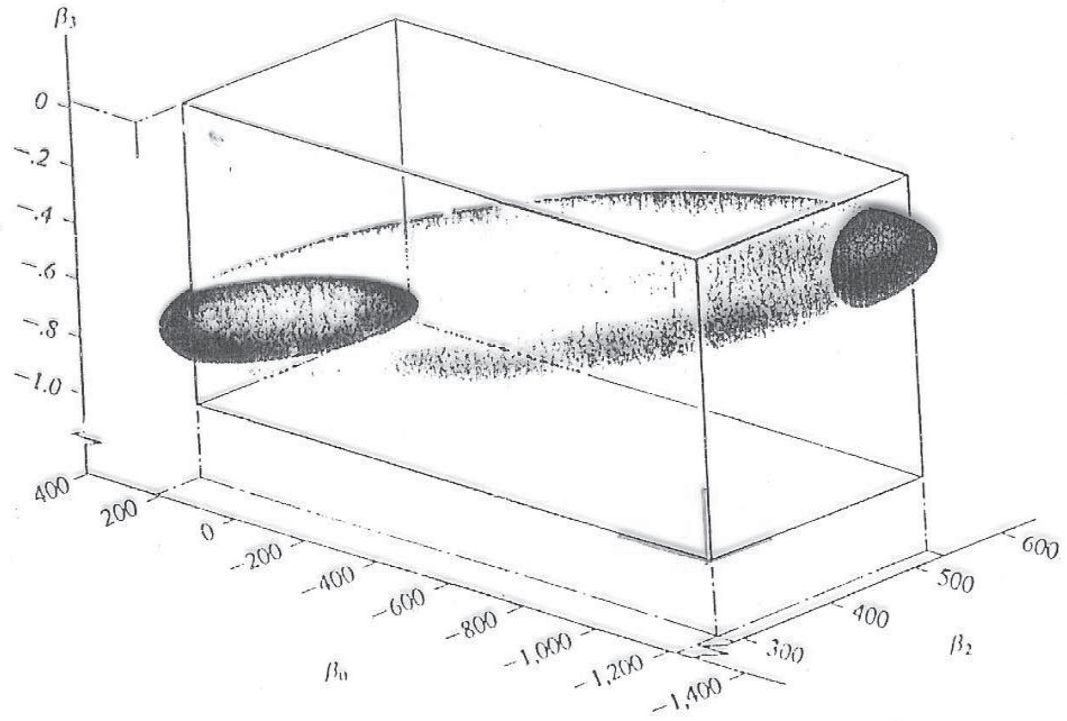

Figure 5.1 Three-dimensional 95% joint confidence region for β_0, β_2, and β_3. The intersection of the Bonferroni confidence intervals is shown as the box.

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 muestral 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

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

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 y productos internos, tendríamos

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

Como

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

obtenemos que

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

es decir

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

La relación entre el estadístico F y el coeficiente de correlación múltiple está dada por el siguiente resultado

Teorema: Supongamos que deseamos testear $H : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, que no involucra al intercept β_0, es decir $\mathbf{C}$ es de la forma $[0, \mathbf{C}_1]$. Consideremos $\omega_1 = \Omega \cap H$. Sea

$$R^2_{\omega_1} = \frac{\sum_{i=1}^{n} (\widehat{y}_{i\omega_1} - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

entonces el estadístico F para testear H será

$$F = \frac{(R^2 - R^2_{\omega_1})(n-p)}{(1-R^2)q}$$

Como corolario de este teorema obtenemos que $R^2 - R^2_{\omega_1} \geq 0$ pues $F \geq 0$ y por lo tanto, el coeficiente de correlación múltiple o de determinación R^2 nunca decrece al agregar una variable regresora extra.

Esta es una deventaja de R^2 si uno lo quiere usar para comparar el ajuste de modelos de distinto número de covariables, y por esta razón se suele utilizar el coeficiente de determinación ajustado definido por

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

que no crece necesariamente con p y de hecho se puede demostrar que R^2_{adj} aumenta al agregar una covariable sólo si el estadístico F que testea que los parámetros agregados son 0 es mayor a 1.