Contraste de hipótesis

Regresión: modelos y métodos

Francesc Carmona Pontaque

PID_00298352

Universitat Oberta de Catalunya

Francesc Carmona Pontaque

Cómo citar este recurso de aprendizaje con el estilo Harvard:

Carmona Pontaque, F. (2024) *Contraste de hipótesis. Regresión: modelos y métodos*. [Recurso de aprendizaje textual]. 1.ª ed. Barcelona: Fundació Universitat Oberta de Catalunya (FUOC).

Primera edición: febrero 2024

Av. Tibidabo, 39-43, 08035 Barcelona
Autoría: Francesc Carmona Pontaque
Producción: FUOC

Planteamiento y objetivos

En los módulos anteriores se ha introducido el concepto de modelo lineal, el cual permite tratar de forma unificada métodos como la regresión lineal simple, la regresión lineal múltiple, la regresión polinómica o los diseños experimentales. Hemos aprendido a construir el modelo y estimar sus parámetros mediante el método de los mínimos cuadrados. Con las condiciones de Gauss-Markov garantizamos el buen funcionamiento del método de los mínimos cuadrados y hemos visto que no es necesario que los errores (o la variable respuesta) tengan distribución normal para que la estimación funcione bien. Ahora bien, cuando esto sucede –es decir, si además de estar centrados en cero, ser incorrelacionados (o independientes) entre ellos y tener la varianza constante, los errores se distribuyen según una ley normal– entonces es posible obtener propiedades de los estimadores que nos permiten hacer inferencia sobre los parámetros. Es decir, que podemos, por ejemplo, construir intervalos de confianza o realizar contrastes de hipótesis sobre los parámetros del modelo. Todas estas cuestiones se tratarán en este módulo.

Para realizar los contrastes o calcular los intervalos, bajo la suposición de normalidad de los errores, necesitaremos trabajar con las distribuciones t de Student y la F de Fisher-Snedecor. Por ello, es muy conveniente, por no decir imprescindible, dominar los conceptos básicos de inferencia estadística: intervalos de confianza, contraste de hipótesis clásico, test t de Student y el test F.

En muchos casos, el contraste se enuncia mediante hipótesis con los coeficientes de regresión, es decir, hipótesis lineales paramétricas. Una hipótesis paramétrica no es más que una restricción sobre el modelo general y que se puede asociar a un modelo más simple. Así pues, el contraste de hipótesis paramétricas es equivalente al contraste de modelos. Para ello utilizaremos el test F en general. Sin embargo, en el caso de una hipótesis paramétrica con una única ecuación, el test F se puede simplificar a un test t de Student, puesto que el numerador del estadístico F tiene un solo grado de libertad.

Como sabemos, cuando la matriz de diseño del modelo general no es de rango máximo estamos ante una situación delicada. En los diseños de regresión esto no suele ocurrir, pero en otros diseños sí. En este caso, las hipótesis paramétricas tienen que ser *demostrables*, lo que significa que las funciones lineales de los parámetros implicadas tienen que ser *estimables*.

También se estudiará la situación en la que la hipótesis de normalidad del error falla y tenemos que utilizar un contraste de permutaciones. Otro aspecto que se trabaja en este módulo es la estimación de los coeficientes mediante intervalos de confianza, bien bajo la suposición de normalidad de los errores (modelo lineal normal) o, de forma más general, utilizando los métodos *bootstrap*.

1. Contraste de un parámetro

Consideremos un modelo lineal normal de la forma $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, de manera que $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Supongamos que es un modelo de regresión como el que hemos planteado con los datos de las islas Galápagos y donde el rango de $\mathbf{X}$ es máximo. En este caso, cualquier parámetro del modelo (coeficiente de regresión β_i) es estimable por el método MC y conocemos su distribución

Rango de X

En un modelo de regresión suponemos que el rango de la matriz de diseño $\mathbf{X}$ es máximo y coincide con el número de columnas $r = m$.

$$\widehat{\beta}_i \sim N(\beta_i, \text{var}(\widehat{\beta}_i))$$

donde $\text{var}(\widehat{\beta}_i) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{[i+1,i+1]}$ y el subíndice $[i+1, i+1]$ indica el elemento $i+1$ de la diagonal de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$.

Nota

El subíndice se entiende si tenemos en cuenta que el modelo de regresión tiene intercepción β_0.

Para contrastar las hipótesis $H_0 : \beta_i = 0;\ H_1 : \beta_i \neq 0$ debemos estimar σ^2 con su estimador MC y utilizar el estadístico t de Student, tal y como se muestra a continuación:

$$t = \frac{\widehat{\beta}_i}{ee(\widehat{\beta}_i)} \sim t_{n-m}$$

donde

$$ee(\widehat{\beta}_i) = \sqrt{\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{[i+1,i+1]}}$$

Por ejemplo, vamos a contrastar si el coeficiente de regresión de la variable `Area` es cero.

Lenguaje

En un lenguaje más coloquial, se dice que el coeficiente puede ser *significativo* o no. Este último caso se corresponde con la hipótesis nula $H_0 : \beta_{\text{Area}} = 0$. La significación se da cuando rechazamos la hipótesis nula y admitimos la alternativa $H_1 : \beta_{\text{Area}} \neq 0$.

```
data(gala, package="faraway")
lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
           data = gala)
ss <- summary(lmod)
ee.Area <- ss$sigma * sqrt(ss$cov.unscaled[2,2])
t.est <- coef(lmod)[2] / ee.Area
t.est
```

```
     Area
-1.067611
```

Podemos comprobar que este valor es el que figura en la tercera columna del `summary(lmod)`, cociente de la primera y segunda columnas.

```
ss$coef[2,3]
```

```
[1] -1.067611
```

Lo mismo ocurre con los otros coeficientes. En el `summary(lmod)`, la primera columna son las estimaciones de los coeficientes, la segunda sus errores estándar, la tercera el estadístico t y la última su p-valor.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
```

Un p-valor inferior al nivel de significación (usualmente 0.05) indica que rechazamos la hipótesis nula.

¡Atención!

Esto no significa que debamos eliminar la variable `Area` del modelo. Aún es necesario evaluar la calidad del modelo. También podría ser que no tuviéramos suficientes datos para mostrar su efecto real. Además, este contraste se hace en presencia de las otras variables regresoras.

```
pt(abs(t.est), df = 30-6, lower.tail = FALSE) * 2
```

```
    Area
0.296318
```

En este caso no podemos rechazar la hipótesis nula.

La distribución t de Student del estadístico también nos permite calcular el **intervalo de confianza** para el coeficiente β_i

$$\widehat{\beta}_i \pm t_{n-m}(\alpha/2) \cdot ee(\widehat{\beta}_i)$$

Por ejemplo

```
prob <- c(0.05/2, 1-0.05/2)
coef(lmod)[2] + qt(prob, df=30-6) * ee.Area
```

```
[1] -0.07021580  0.02233912
```

Aunque es mucho más sencillo utilizar la función `confint()`.

```
confint(lmod)[2,]
```

```
      2.5 %      97.5 %
-0.07021580  0.02233912
```

Nivel de confianza

La función `confint()` dispone del parámetro `level` que controla el nivel de confianza.

Contrastar una hipótesis del tipo $H_0 : \beta_i = b$, donde b es una constante conocida, también se puede hacer con un estadístico t. Basta con cambiar el numerador del estadístico por $\widehat{\beta}_i - b$, ya que su desviación estándar es la misma que antes.

2. Contraste de una función paramétrica estimable

En los modelos lineales normales con matriz de diseño de rango no máximo, las únicas hipótesis que podemos contrastar son las que se forman con funciones paramétricas estimables. Por ejemplo, para contrastar

$$H_0 : \mathbf{a}'\boldsymbol{\beta} = 0$$

debemos asegurar que $\psi = \mathbf{a}'\boldsymbol{\beta}$ es una FPE que tiene, en consecuencia, un estimador lineal único.

Supongamos que en el diseño *crossover* visto en el módulo anterior, la distribución de los errores es normal y se verifican las condiciones de Gauss-Markov. Veamos cómo contrastar la hipótesis $H_0 : \alpha = \beta$, ya que esta es la principal pregunta: ¿los efectos de los fármacos son iguales?
Es evidente que esta hipótesis se puede escribir $H_0 : \alpha - \beta = 0$ y sabemos que $\psi = \alpha - \beta$ es estimable.

El contraste $H_0 : \mathbf{a}'\boldsymbol{\beta} = 0,\ H_1 : \mathbf{a}'\boldsymbol{\beta} \neq 0$ se resuelve también con un estadístico t de Student

$$t = \frac{\mathbf{a}'\boldsymbol{\beta}}{ee(\mathbf{a}'\boldsymbol{\beta})} \sim t_{n-r}$$

Grados de libertad

En este estadístico los grados de libertad son $n - r$, donde r es el rango de $\mathbf{X}$.

donde

$$ee(\mathbf{a}'\widehat{\boldsymbol{\beta}}) = \sqrt{\widehat{\sigma}^2 \cdot \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{a}}$$

Para contrastar la hipótesis $H_0 : \alpha - \beta = 0$ del modelo *crossover* debemos recuperar los datos del módulo anterior y algunos de sus elementos.

Test F

Más adelante veremos un test F equivalente.

```
library(MASS)
cmod <- lm(y ~ x1 + x2 + x3)
ss <- summary(cmod)
X.co <- model.matrix(cmod)
XtXginv <- ginv(t(X.co) %*% X.co)
coef.co <- XtXginv %*% t(X.co) %*% y
a <- c(0,1,-1,0)
ee.a <- ss$sigma * sqrt(t(a) %*% XtXginv %*% a)
t.est <- sum(a*coef.co) / ee.a
```

```
t.est
```

```
       [,1]
[1,] 2.1711
```

```
pt(abs(t.est), df=40-3, lower.tail = FALSE) * 2
```

```
           [,1]
[1,] 0.03641071
```

Luego rechazamos la hipótesis nula y admitimos la diferencia entre los efectos de los fármacos.

El intervalo de confianza para la función paramétrica $\psi = \mathbf{a}'\boldsymbol{\beta}$ es

$$\mathbf{a}'\widehat{\boldsymbol{\beta}} \pm t_{n-r}(\alpha/2) \cdot ee(\mathbf{a}'\widehat{\boldsymbol{\beta}})$$

donde habitualmente $\alpha = 0.05$ y el nivel de confianza es del $95\,\%$.

En el caso de la FPE $\psi = \alpha - \beta$ del *crossover* tenemos

```
prob <- c(0.05/2, 1-0.05/2)
sum(a*coef.co) + qt(prob, df=40-3) * as.vector(ee.a)
```

```
[1]  0.5906834 17.1093166
```

Aviso

Cuando en **R** se suman o multiplican vectores de longitudes distintas suele aparecer un mensaje de aviso, aunque el cálculo sea correcto. En este caso, se ha evitado el aviso con la instrucción `as.vector()` aplicada al número `ee.a`.

3. Contraste de modelos

En una situación experimental disponemos de un modelo lineal normal $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$ que se supone válido y con rango $\mathbf{X} = r$. Además, el interés radica en contrastar un modelo *más simple* como $\mathbf{Y} = \widetilde{\mathbf{X}}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ con rango $\widetilde{\mathbf{X}} = r - q$.

Más simple

Por un modelo más simple se entiende que las columnas de $\widetilde{\mathbf{X}}$ son combinaciones lineales de las de $\mathbf{X}$, de modo que los subespacios vectoriales respectivos verifican $\langle\widetilde{\mathbf{X}}\rangle \subset \langle\mathbf{X}\rangle \subset \mathbb{R}^n$.

Habitualmente, el modelo más simple consiste en imponer restricciones lineales a los parámetros del modelo general. Estas restricciones son del tipo $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, donde $\mathbf{A}$ es una matriz $q \times m$ y las q filas representan funciones paramétricas estimables (FPE) del modelo con matriz de diseño $\mathbf{X}$.

En resumen, el contraste de modelos

$$H_0 : \mathbf{Y} = \widetilde{\mathbf{X}}\boldsymbol{\theta} + \boldsymbol{\epsilon} \qquad H_1 : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

es equivalente a

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{0} \qquad H_1 : \mathbf{A}\boldsymbol{\beta} \neq \mathbf{0} \qquad \text{en el modelo } \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

El contraste de estas hipótesis se resuelve con un estadístico con distribución F de Fisher-Snedecor basado en la comparación de las sumas de cuadrados residuales de los dos modelos.

Teorema fundamental del análisis de la varianza

Si $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ es un modelo lineal normal, de manera que $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, y consideramos una hipótesis lineal contrastable $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, con rango $\mathbf{A} = q$, entonces con los estadísticos

$$\text{SCR} = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \qquad \text{SCR}_H = (\mathbf{Y} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}})'(\mathbf{Y} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}})$$

podemos calcular el estadístico

$$F = \frac{(\text{SCR}_H - \text{SCR})/q}{\text{SCR}/(n-r)}$$

Si H_0 es cierta, este estadístico sigue la distribución F de Fisher-Snedecor con q y $n - r$ grados de libertad.

Demostración

La demostración puede verse en el teorema 5.3.1. de Carmona (2005).

El cálculo del numerador y del denominador se suele expresar en forma de tabla general del análisis de la varianza (ver tabla 1).

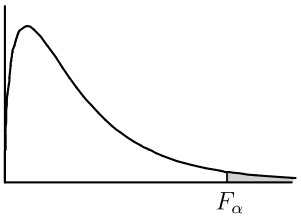
	Grados de libertad	Suma de cuadrados	Cuadrados medios	Cociente
Desviación	q	$\text{SCR}_H - \text{SCR}$	$(\text{SCR}_H - \text{SCR})/q$	F
Residuo	$n-r$	SCR	$\text{SCR}/(n-r)$	

Tabla 1: Tabla general del análisis de la varianza

Criterio de decisión

Si $F > F_\alpha$ se rechaza H_0; si $F \leq F_\alpha$ se acepta H_0.

Donde, para un nivel de significación α, F_α se elige de forma que $P(F_{q,n-r} > F_\alpha) = \alpha$.



Este estadístico F también se puede escribir en términos de las FPE con la matriz **A**.

$$F = \frac{(\mathbf{A}\widehat{\boldsymbol{\beta}})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}})}{q\,\widehat{\sigma}^2}$$

donde $\widehat{\sigma}^2 = \text{SCR}/(n-r)$.

En la práctica es más sencillo utilizar las sumas de cuadrados de los dos modelos.

La función `anova()`

En **R** disponemos de la función `anova()` que justamente contrasta dos modelos u objetos `lm`.

Veamos como primer ejemplo lo que en regresión se llama **significación de la regresión**. Se trata de contrastar si todos los coeficientes de regresión son cero (excepto el de intercepción). En términos paramétricos, la hipótesis nula es

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

Observemos que, en realidad, son p condiciones del tipo $\beta_i = 0$. Todas son FPE, ya que suponemos rango máximo, y en total son $q = p$.

Significación de la regresión

Este contraste es muy importante en regresión. Suponiendo que el modelo sea válido, si no rechazamos la hipótesis nula, el modelo sería completamente inútil.

En forma matricial las condiciones son

$$\mathbf{A}\boldsymbol{\beta} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

El rango de $\mathbf{A}$ es p.

Sin embargo, para contrastar esta hipótesis es mucho más sencillo si lo planteamos como un contraste de modelos.

La significación del modelo de regresión con los datos de las islas Galápagos se resuelve en **R** así

```
lmod <- lm(Species ~ ., data = gala[,-2])
lmod0 <- lm(Species ~ 1, data = gala[,-2])
anova(lmod0,lmod)
```

```
Analysis of Variance Table

Model 1: Species ~ 1
Model 2: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     29 381081
2     24  89231  5    291850 15.699 6.838e-07
```

La función `summary()`

El mismo resultado se puede leer en la última linea del `summary(lmod)`.

En el caso de que el contraste de modelos se haga con una única restricción, es decir $q = 1$, el test F es equivalente al test t, ya que $F = t^2$.

Por ejemplo, el contraste de la hipótesis $\beta_{\text{Area}} = 0$ se puede resolver con un contraste de modelos.

```
lmod <- lm(Species ~ ., data=gala[,-2])
lmod0 <- lm(Species ~ Elevation + Nearest + Scruz + Adjacent,
                        data = gala[,-2])
anova(lmod0,lmod)
```

La función `update()`

La función `update()` permite modificar un modelo `lm` para quitar o poner variables del *data.frame*

```
Analysis of Variance Table

Model 1: Species ~ Elevation + Nearest + Scruz + Adjacent
Model 2: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     25 93469
2     24 89231  1    4237.7 1.1398 0.2963
```

Observemos que el p-valor es el mismo que obtuvimos con el estadístico t y además $t^2 = F$.

```
ss <- summary(lmod)
ss$coef[2,4]       # p-valor


[1] 0.296318


ss$coef[2,3]^2     # t^2


[1] 1.139792
```

Otra hipótesis que se puede resolver como un contraste de modelos es H_0 : $\beta_{\text{Area}} = \beta_{\text{Adjacent}}$. Si los dos coeficientes son iguales, podemos considerar que son uno solo y sumar las dos variables.

```
lmod0 <- lm(Species ~ I(Area + Adjacent) + Elevation + Nearest +
                          Scruz, data = gala[,-2])
anova(lmod0, lmod)


Analysis of Variance Table

Model 1: Species ~ I(Area + Adjacent) + Elevation + Nearest + Scruz
Model 2: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
  Res.Df    RSS Df Sum of Sq     F  Pr(>F)
1     25 109591
2     24  89231  1     20360 5.476 0.02793
```

En este caso, rechazamos la hipótesis considerada.

Una hipótesis del tipo $H_0 : \beta_{\text{Elevation}} = 0.5$ también se puede contrastar así

```
lmod0 <- lm(Species ~ Area + offset(0.5 * Elevation) + Nearest +
                          Scruz + Adjacent, data = gala[,-2])
anova(lmod0, lmod)


Analysis of Variance Table

Model 1: Species ~ Area + offset(0.5 * Elevation) + Nearest + Scruz +
    Adjacent
Model 2: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     25 131312
2     24  89231  1     42081 11.318 0.002574
```

La función `offset()`

La función `offset()` permite realizar un calculo e incorporar el resultado como una variable regresora sin coeficiente a estimar.

En este caso también rechazamos la hipótesis considerada.

En el diseño *crossover* la principal hipótesis $H_0 : \alpha = \beta$ se puede contrastar con un test F. Observemos que si los dos efectos son iguales, el parámetro común es el mismo en las cuatro situaciones experimentales y se confunde con la media general μ.

```
cmod <- lm(y ~ x1 + x2 + x3)
cmod0 <- lm(y ~ x3)
anova(cmod0,cmod)


Analysis of Variance Table

Model 1: y ~ x3
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     38 6931.2
2     37 6147.9  1    783.23 4.7137 0.03641
```

El resultado es el mismo que cuando aplicamos el test t.

4. Contrastes de permutaciones

Lectura complementaria

Leer el apartado 3.3. del libro de Faraway (2014).

Para calcular los estadísticos de los apartados anteriores hemos supuesto la normalidad de los errores. En muchas circunstancias, especialmente cuando el tamaño muestral es grande, los contrastes explicados son bastante robustos gracias al teorema del límite central, que implica la aproximación de los errores a la distribución normal. Sin embargo, no siempre podemos garantizar una buena aproximación. Los contrastes de permutaciones no necesitan la suposición de normalidad.

La idea de un contraste de permutaciones es elegir un estadístico, suponiendo cierta la hipótesis nula, y calcularlo para un gran número de permutaciones de los datos. La distribución de los valores del estadístico resultantes se utiliza para obtener un p-valor. Diremos que el contraste de permutaciones es *exacto* cuando se calculan todas las permutaciones posibles. Cuando el número de permutaciones es muy elevado, se puede seleccionar un número alto pero al azar.

Por ejemplo, el contraste de significación de la regresión con los datos de las islas Galápagos se ha calculado con el estadístico F. La suposición de normalidad hace que la distribución del estadístico sea una F de Fisher-Snedecor. El mismo estadístico mide bien si la hipótesis nula de no significación de la regresión es cierta o no. Sin embargo, si no hacemos la suposición de normalidad de los errores, entonces desconocemos la distribución real de F. Si la hipótesis nula es cierta, es decir, las variables regresoras no sirven para predecir la respuesta ya que sus coeficientes son cero, entonces podemos permutar los valores de la variable respuesta y el resultado del estadístico debe ser similar. Esos resultados proporcionan una distribución de valores F, de forma que, si el estadístico F observado (sin permutar nada) es un valor extremo de esa distribución, entonces podemos rechazar la hipótesis nula como hacemos habitualmente en la inferencia estadística.

Tal y como se ve en el apartado 3.3. del libro de Faraway (2014), vamos a contrastar la significación de la regresión con las variables regresoras `Nearest` y `Scruz` únicamente.

Lectura complementaria

Ver el ejemplo del apartado 5.3.1. de Cao Abad y Fernández Casal (2022).

Para no repetir la solución de Faraway, utilizaremos la función `boot()` del paquete `boot`.

```
set.seed(123)
library(boot)
statistic <- function(data, i){
  g <- lm(Species[i] ~ Nearest + Scruz, data)
  summary(g)$fstatistic[1]
}
```

```
res.boot <- boot(gala, statistic, R = 4000, sim = "permutation")
# hist(res.boot$t)
pval <- mean(res.boot$t > res.boot$t0)
pval


[1] 0.55825
```

Del mismo modo, se pueden contrastar las hipótesis sobre los coeficientes de las variables regresoras uno a uno. También para no repetir la solución de Faraway, vamos a utilizar la función `lmp()` del paquete `lmPerm` que calcula los contrastes de permutaciones para cada coeficiente de forma directa.

```
library(lmPerm)
summary(lmp(Species ~ Nearest + Scruz, data=gala))


[1] "Settings:  unique SS : numeric variables centered"

Call:
lmp(formula = Species ~ Nearest + Scruz, data = gala)

Residuals:
   Min     1Q Median     3Q    Max
-97.88 -73.54 -46.30  18.34 344.82

Coefficients:
        Estimate Iter Pr(Prob)
Nearest   1.1792   51    0.804
Scruz    -0.4406  313    0.243

Residual standard error: 116.2 on 27 degrees of freedom
Multiple R-Squared: 0.04269,Adjusted R-squared: -0.02823
F-statistic: 0.602 on 2 and 27 DF,  p-value: 0.5549
```

5. Intervalos y regiones de confianza para los coeficientes en regresión

Como ya se ha visto en el apartado 1 de este módulo, es posible dar intervalos de confianza para cada coeficiente de regresión gracias a la distribución t de Student.

Por ejemplo, para la regresión explicada en el apartado 1, los intervalos de confianza de los coeficientes son

```
confint(lmod)
```

```
                 2.5 %      97.5 %
(Intercept) -32.4641006 46.60054205
Area         -0.0702158  0.02233912
Elevation     0.2087102  0.43021935
Nearest      -2.1664857  2.18477363
Scruz        -0.6850926  0.20404416
Adjacent     -0.1113362 -0.03827344
```

Sin embargo, si deseamos una región de confianza del 95 % para dos de los coeficientes, no será suficiente con el rectángulo que forman los dos intervalos hallados por separado. Por ejemplo, para β_{Area} y β_{Adjacent}, el rectángulo que forman los dos intervalos calculados no tiene probabilidad 0.95.

Consideremos los sucesos

$$A_1 : \text{el IC no cubre a } \beta_{\text{Area}}$$

$$A_2 : \text{el IC no cubre a } \beta_{\text{Adjacent}}$$

con $P(A_1) = 0.05$ y $P(A_2) = 0.05$.

Según la desigualdad de Bonferroni, la probabilidad de que los dos intervalos cubran a su coeficiente es

$$P(\bar{A}_1 \cap \bar{A}_2) \geq 1 - P(A_1) - P(A_2) = 1 - 2 \cdot 0.05 = 0.9$$

de forma que, si los intervalos de confianza se estiman por separado al $95\,\%$, la desigualdad de Bonferroni garantiza que ambos intervalos conjuntamente calculados sobre la misma muestra tienen una confianza superior al $90\,\%$.

En resumen, si queremos hallar regiones de confianza debemos utilizar un método de cálculo de intervalos de confianza simultáneos o calcular una región de confianza con nivel de confianza 0.95.

El **método de Bonferroni** permite calcular intervalos de confianza simultáneos con nivel de confianza del $95\,\%$. Simplemente, cada uno de los intervalos debe tener una confianza del $(1 - 0.05/s) \cdot 100\,\%$ con s igual al número de coeficientes.

```
confint(lmod, parm=c(2,6), level=1-0.05/2)
```

```
              1.25 %     98.75 %
Area     -0.07754904  0.02967237
Adjacent -0.11712508 -0.03248458
```

Además, es posible construir una región de confianza con base en el siguiente resultado:

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

Esta región tiene forma de elipsoide y en el caso de dos parámetros podemos dibujar la elipse correspondiente.

Demostración

La demostración se puede ver en el apartado 6.3.1. (regresión simple) y en la página 140 del apartado 8.3. (regresión múltiple) de Carmona (2005).

Para los dos parámetros β_{Area} y β_{Adjacent} del modelo de regresión tenemos

```
library(ellipse)
plot(ellipse(lmod, c(2,6)), type="l", ylim=c(-0.13,0),
                            lwd=1.5, col="blue")
points(coef(lmod)[2], coef(lmod)[6], pch=19)
```

Al gráfico le podemos añadir los intervalos de confianza por separado y los simultáneos

```
abline(v=confint(lmod)[2,], lty=2)
abline(h=confint(lmod)[6,], lty=2)
abline(v=confint(lmod, level=1-0.05/4)[2,], lty=2, lwd=1.5,
                                            col="blue")
abline(h=confint(lmod, level=1-0.05/4)[2,], lty=2, lwd=1.5,
                                            col="blue")
```

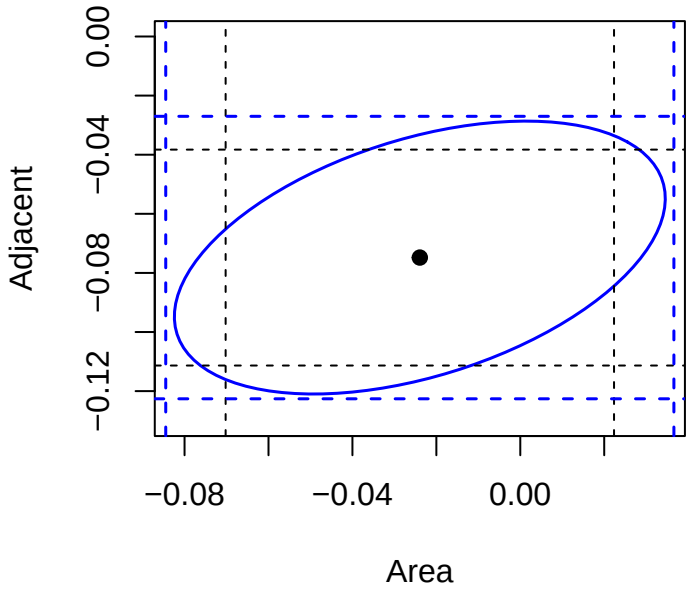


Figura 1

Región de confianza al 95 % e intervalos de confianza individuales (en negro) y simultáneos (en azul) para los coeficientes β_{Area} y β_{Adjacent}.

El método de Bonferroni no es el único para hallar intervalos de confianza simultáneos. El **método de Scheffé** se basa en maximizar la forma cuadrática del estadístico utilizado en los intervalos individuales. Esto se traduce en cambiar el cuantil de la t por un múltiplo del cuantil de la F.

$$\hat{\beta}_i \pm (sF_{s,n-m}(\alpha))^{1/2} \cdot ee(\hat{\beta}_i)$$

¿Qué es mejor, utilizar la región de confianza o los intervalos simultáneos?

- Ambos sistemas definen áreas o volúmenes con formas distintas donde es altamente probable que se encuentren los parámetros desconocidos.
- La región de confianza se basa en la distribución conjunta de los estimadores.
- Tiene la ventaja de tener en cuenta la estructura de covarianzas, por lo que el conjunto de valores plausibles que define tiene el menor tamaño posible.
- Tiene la desventaja que para dimensiones mayores de 3 es de difícil visualización.

Los intervalos simultáneos se fundamentan en algún método de ajuste de los intervalos individuales. El método de Bonferroni utiliza el principio de inclusión-exclusión. El método de Scheffé se basa en un argumento de optimización. Tienen la ventaja que definen regiones fáciles de interpretar (son rectángulos, cajas...). Tienen la desventaja que definen regiones plausibles muy amplias. Como los métodos de ajuste no son exactos resultan muy conservadores en algunos casos, lo que puede llevar a contradicciones si se comparan con la región de confianza.

6. Intervalos *bootstrap*

Las regiones de confianza y los intervalos del apartado anterior se basan en la suposición de normalidad para los errores del modelo. El método *bootstrap* proporciona un algoritmo para hallar intervalos de confianza sin esa suposición.

Lectura complementaria

En el documento web de Carmona (2023) se puede ver en qué consiste este método y sus diversas aplicaciones.

La idea es calcular el estimador del coeficiente en una regresión con submuestras con reemplazamiento de la muestra original. Si se calculan para un gran número de submuestras, el resultado es una distribución aproximada del coeficiente que utilizaremos para obtener un intervalo de confianza sobre ella.

Lectura complementaria

Ver el apartado 3.6. de Faraway (2014).

Con las funciones apropiadas de los paquetes `car` y `boot`, hallar los intervalos de confianza para los coeficientes del modelo de regresión de las islas Galápagos es muy sencillo. La función `Boot()` del paquete `car` hace un *bootstrap* sobre los coeficientes de un modelo de regresión con `R` submuestras de los residuos.

```
library(car)
B <- Boot(lmod, f=coef, R=4000, method = "residual")
```

El siguiente paso es calcular los intervalos con la función `boot.ci()` del paquete `boot` para cada uno de los coeficientes según su posición (`index`).

```
boot.ci(B, type="perc", index=2)  # Area


BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 4000 bootstrap replicates

CALL :
boot.ci(boot.out = B, type = "perc", index = 2)

Intervals :
Level     Percentile
95%   (-0.0811,  0.0309 )
Calculations and Intervals on Original Scale
```

```
boot.ci(B, type="perc", index=6)  # Adjacent


BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 4000 bootstrap replicates

CALL :
boot.ci(boot.out = B, type = "perc", index = 6)

Intervals :
Level     Percentile
95%   (-0.1196, -0.0341 )
Calculations and Intervals on Original Scale
```

También podemos dibujar la densidad estimada del coeficiente de la variable `Area` y el intervalo de confianza obtenido con este método, tal y como se observa en la figura 2.

```
library(ggplot2)
coefmat <- data.frame(B$t)
ggplot(coefmat, aes(x=Area)) + geom_density() +
   geom_vline(xintercept = c(-0.0811, 0.0309), lty=2) +
   theme_light()
```

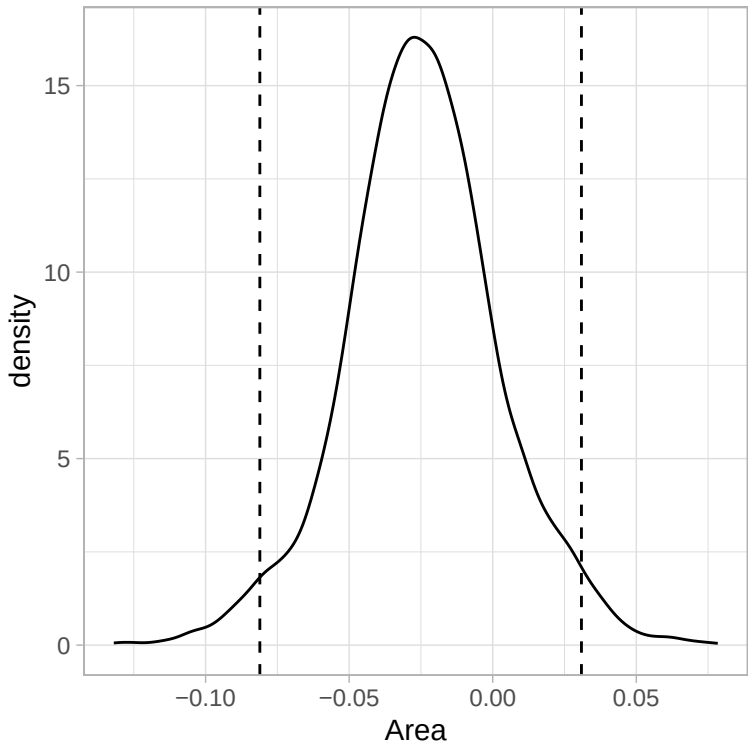


Figura 2

Densidad de los coeficientes de regresión *bootstrap* para β_{Area} y su intervalo de confianza al 95 %.

7. Intervalo de confianza para la varianza del modelo

Como ya se ha explicado en el apartado 6. del módulo "Estimación",

$$\mathrm{SCR}/\sigma^2 \sim \chi^2_{n-r}$$

Con este resultado, además, es posible construir un intervalo de confianza para σ^2. Basta con hallar los valores a, b de la distribución χ^2_{n-r} tales que

$$P[a < \mathrm{SCR}/\sigma^2 < b] = 1 - \alpha$$

entonces

$$P[\mathrm{SCR}/b < \sigma^2 < \mathrm{SCR}/a] = 1 - \alpha$$

Por ejemplo, en el caso del diseño *crossover* tenemos

```
alpha <- 0.05
a <- qchisq(alpha/2, df=40-3)
b <- qchisq(1-alpha/2, df=40-3)
deviance(cmod) * c(1/b,1/a)
```

```
[1] 110.4392 278.1158
```

Bibliografía

Cao Abad, R. y Fernández Casal, R. (2022) *Técnicas de Remuestreo*. Universidade da Coruña. Disponible en: `https://rubenfcasal.github.io/book_remuestreo/`

Carmona, F. (2005) *Modelos lineales*. e-UMAB, Universitat de Barcelona.

Carmona, F. (2023) Bootstrap. Cursos R, Universitat de Barcelona. Disponible en: `https://www.ub.edu/cursosR/files/bootstrap.html`

Faraway, J.J. (2014) *Linear Models with R*. 2.ª ed. Chapman and Hall/CRC.

Seber, G.A.F. and Lee, A.J. (2003) *Linear Regression Analysis*. John Wiley & Sons.