Contrastes de hipótesis

Contents

1 Contrastes para las β_i usando la distribución *t-student*

1.1 Teoría

Queremos resolver los contrastes:

$$H_0 : \beta_i = 0 H_1 : \beta_i \neq 0$$

para el modelo $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i$. Hemos visto que

$$\frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} \to t_{n-k-1}$$

El procedimiento se puede resumir en:

- Partimos de un estadístico con distribución conocida.
- Suponemos que H_0 es cierta y obtenemos la distribución del estadístico.
- Comprobamos si los datos que tenemos (la muestra) son un valor probable para la distribución obtenida o no.

Por tanto, si H_0 es cierta:

$$t_0 = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \to t_{n-k-1}$$

Sea $t_{n-k-1;\alpha/2}$ el valor de una t-student con (n-k-1) grados de libertad tal que

$$P(t_{n-k-1} \geq t_{n-k-1;\alpha/2}) = \alpha/2$$

- si $|t_0| \geq t_{n-k-1;\alpha/2}$: se rechaza H_0
- si $|t_0| \leq t_{n-k-1;\alpha/2}$: no se rechaza H_0

Se define el *pvalor* como:

$$pvalor = 2P(t_{n-k-1} \geq |t_0|)$$

Por tanto

- si $pvalor \leq \alpha$: se rechaza H_0
- si $pvalor \geq \alpha$: no se rechaza H_0

1.2 Ejemplo

Veamos por ejemplo el modelo *kid_score* ~ *mom_hs* + *mom_iq* + *mom_age*:

```
load("datos/kidiq.Rdata")
str(d)
```

```
## 'data.frame':    434 obs. of  5 variables:
##  $ kid_score: int  65 98 85 83 115 98 69 106 102 95 ...
##  $ mom_hs   : Factor w/ 2 levels "no","si": 2 2 2 2 2 1 2 2 2 2 ...
##  $ mom_iq   : num  121.1 89.4 115.4 99.4 92.7 ...
##  $ mom_work : Factor w/ 4 levels "notrabaja","trabaja23",..: 4 4 4 3 4 1 4 3 1 1 ...
##  $ mom_age  : int  27 25 27 25 27 18 20 23 24 19 ...
```

```
m1 = lm(kid_score ~ . - mom_work, data = d)
summary(m1)
```

```
##
## Call:
## lm(formula = kid_score ~ . - mom_work, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.289 -12.421   2.399  11.223  50.169
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.98466    9.13013   2.298   0.0220 *
## mom_hssi     5.64715    2.25766   2.501   0.0127 *
## mom_iq       0.56254    0.06065   9.276   <2e-16 ***
## mom_age      0.22475    0.33075   0.680   0.4972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.15 on 430 degrees of freedom
## Multiple R-squared:  0.215,  Adjusted R-squared:  0.2095
## F-statistic: 39.25 on 3 and 430 DF,  p-value: < 2.2e-16
```

Veamos de donde salen los valores de la tabla anterior:

- Estimate:

```r
(beta_e = coefficients(m1))
```

```
## (Intercept)    mom_hssi      mom_iq     mom_age 
##  20.9846620   5.6471512   0.5625443   0.2247505
```

- Std. Error:

```r
(beta_se = sqrt(diag(vcov(m1))))
```

```
## (Intercept)    mom_hssi      mom_iq     mom_age 
##  9.13012544  2.25765592  0.06064506  0.33074520
```

- t value:

```r
(t_value = beta_e/beta_se)
```

```
## (Intercept)    mom_hssi      mom_iq     mom_age 
##   2.2983980   2.5013339   9.2760110   0.6795276
```

- Pr(>|t|) (es decir, p-valores):

```r
n = nrow(d)
k = 3
(pvalores = 2*pt(abs(t_value), df = n - k -1, lower.tail = F))
```

```
##  (Intercept)     mom_hssi       mom_iq      mom_age 
## 2.201813e-02 1.274346e-02 8.650677e-19 4.971693e-01
```

- Si juntamos todo en una tabla:

```r
data.frame(beta_e, beta_se, t_value, pvalores)
```

```
##                 beta_e    beta_se   t_value     pvalores
## (Intercept) 20.9846620 9.13012544 2.2983980 2.201813e-02
## mom_hssi     5.6471512 2.25765592 2.5013339 1.274346e-02
## mom_iq       0.5625443 0.06064506 9.2760110 8.650677e-19
## mom_age      0.2247505 0.33074520 0.6795276 4.971693e-01
```

2 Relación entre intervalos de confianza y contrastes

Sean los contrastes bilaterales:

$$H_0 : \beta_i = 0 H_1 : \beta_i \neq 0$$

Y los intervalos de confianza de los mismos parámetros:

$$\beta_i \in (a_i, b_i)$$

Existe una relación entre ambos:

- Si $0 \in (a_i, b_i) \Rightarrow$ no se rechaza H_0.
- Si $0 \notin (a_i, b_i) \Rightarrow$ se rechaza H_0.

En el caso del ejemplo, si miramos pvalores e intervalos:

```r
confint(m1)
```

```
##                  2.5 %     97.5 %
## (Intercept)  3.0394352 38.9298887
## mom_hssi     1.2097371 10.0845653
## mom_iq       0.4433466  0.6817419
## mom_age     -0.4253280  0.8748289
```

```
summary(m1)
```

```
## 
## Call:
## lm(formula = kid_score ~ . - mom_work, data = d)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -53.289 -12.421   2.399  11.223  50.169 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 20.98466    9.13013   2.298   0.0220 *  
## mom_hssi     5.64715    2.25766   2.501   0.0127 *  
## mom_iq       0.56254    0.06065   9.276   <2e-16 ***
## mom_age      0.22475    0.33075   0.680   0.4972    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 18.15 on 430 degrees of freedom
## Multiple R-squared:  0.215,  Adjusted R-squared:  0.2095 
## F-statistic: 39.25 on 3 and 430 DF,  p-value: < 2.2e-16
```

3 Contraste para σ^2

El contraste es:

$$H_0 : \sigma^2 = \sigma_0^2 H_1 : \sigma^2 \neq \sigma_0^2$$

El estadístico del contraste que vamos a utilizar es:

$$\frac{(n-k-1)\hat{s}_R^2}{\sigma^2} \rightarrow \chi^2_{n-k-1}$$

Por tanto, si la hipótesis nula es cierta,

$$\chi_0^2 = \frac{(n-k-1)\hat{s}_R^2}{\sigma_0^2} \rightarrow \chi^2_{n-k-1}$$

Como ejemplo, vamos a contrastar

$$H_0 : \sigma^2 = 20^2 H_1 : \sigma^2 \neq 20^2$$

```
(chisq_0 = sum(resid(m1)^2)/20^2)
```

```
## [1] 354.0126
```

```r
# limites del contraste bilateral
c(qchisq(0.05/2,df = n-k-1), qchisq(1-0.05/2,df = n-k-1))
```

```
## [1] 374.4397 489.3477
```

Por tanto se rechaza la hipótesis nula. El mismo resultado se obtiene mirando el intervalo de confianza.

4 Contraste de regresión múltiple

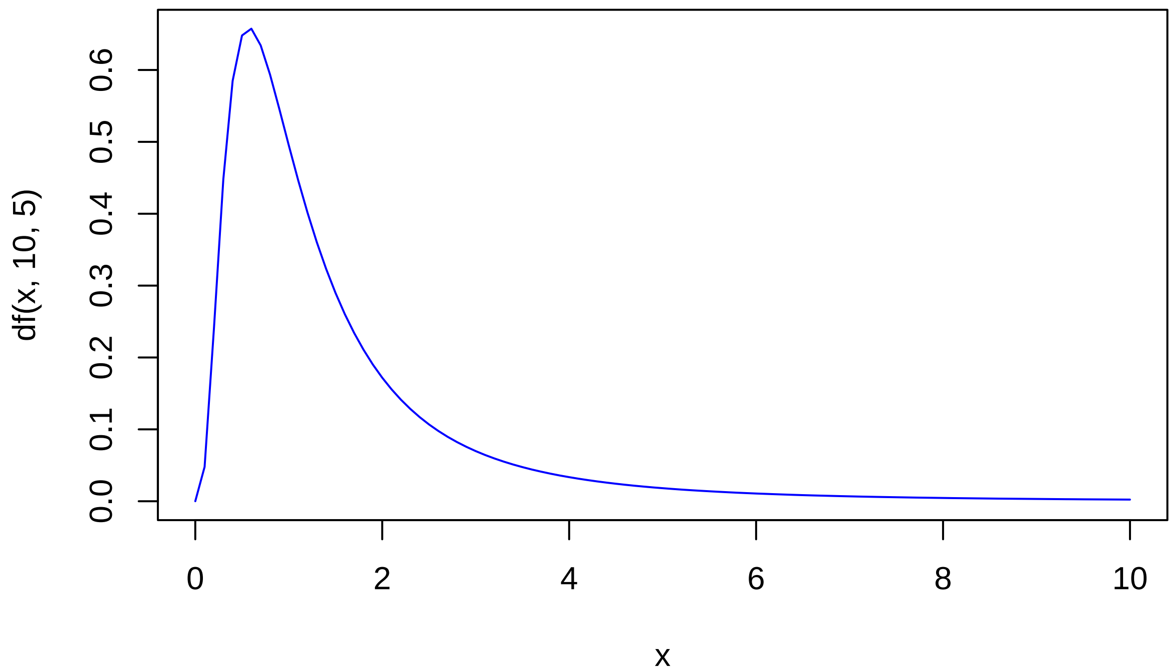
4.1 La distribución F

Sean una χ^2_m y una χ^2_n, ambas independientes. La distribución F se define como

$$\frac{\chi^2_m/m}{\chi^2_n/n} \sim F_{m,n}$$

La distribución F es similar a la χ^2:

```r
curve(df(x,10,5), from = 0, to = 10, col = "blue")
```



4.2 Descomposición de la suma de cuadrados

Tenemos el modelo

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki} + e_i = \hat{y}_i + e_i$$

Restando la media $\bar{y} = \sum y_i/n$:

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + e_i$$

Elevando al cuadrado y sumando se tiene:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2$$

ya que $\sum (\hat{y}_i - \bar{y}) e_i = 0$. Se denominan:

- Suma de cuadrados total:

$$SCT = \sum (y_i - \bar{y})^2$$

- Suma de cuadrados del modelo:

$$SCM = \sum (\hat{y}_i - \bar{y})^2$$

- Suma de cuadrados de los residuos:

$$SCR = \sum e_i^2$$

Por tanto, se cumple que

$$SCT = SCM + SCR$$

4.3 Expresiones alternativas para la suma de cuadrados

- Suma de cuadrados total:

$$SCT = (n-1)\hat{s}_y^2$$

ya que la varianza $\hat{s}_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$.

- Suma de cuadrados estimados por el modelo:

$$SCM = (n-1)\hat{\beta}_a^T S_{xx} \hat{\beta}_a = (n-1)\hat{\beta}_a^T S_{xy}$$

Para probar esa relación se tiene que:

$$\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_{1i} - \bar{x}_1) + \cdots + \hat{\beta}_k (x_{ki} - \bar{x}_k)$$

Por tanto:

$$\begin{bmatrix} \hat{y}_1 - \bar{y} \\ \hat{y}_2 - \bar{y} \\ \cdots \\ \hat{y}_n - \bar{y} \end{bmatrix} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \cdots & x_{k1} - \bar{x}_k \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{k2} - \bar{x}_k \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \cdots & x_{kn} - \bar{x}_k \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \cdots \\ \hat{\beta}_k \end{bmatrix}$$

Es decir

$$\hat{y} - \bar{y} = X_a \hat{\beta}_a$$

$$SCM = \sum (\hat{y}_i - \bar{y})^2 = (\hat{y} - \bar{y})^T (\hat{y} - \bar{y}) = \hat{\beta}_a^T X_a^T X_a \hat{\beta}_a$$

$$= (n-1)\hat{\beta}_a^T S_{xx} \hat{\beta}_a = (n-1)\hat{\beta}_a^T S_{xx} S_{xx}^{-1} S_{xy} = (n-1)\hat{\beta}_a^T S_{xy}$$

- Suma de cuadrados de los residuos:

$$SCR = (n-k-1)\hat{s}_R^2$$

ya que la varianza resiudal es $\hat{s}_R^2 = \frac{\sum e_i^2}{n-k-1}$.

4.4 Contraste

Este contraste establece, de manera conjunta, si alguno de los regresores influye en la respuesta. Es decir, en el modelo $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i$ se constrasta si

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 H_1 : \text{Algún } \beta_i \neq 0$$

Para resolver este contraste, se puede demostrar que:

- Si $\beta_1 = \beta_2 = \cdots = \beta_k = 0 \Rightarrow SCM/\sigma^2 \sim \chi_k^2$
- $SCR/\sigma^2 \sim \chi_{n-k-1}^2$
- SCM y SCR son independientes.

Por lo tanto es razonable utilizar el estadístico:

$$\frac{\frac{SCM/\sigma^2}{k}}{\frac{SCR/\sigma^2}{n-k-1}} \sim F_{k,n-k-1} \Rightarrow F_0 = \frac{SCM/k}{SCR/(n-k-1)} \sim F_{k,n-k-1}$$

Si $\beta_1 = \beta_2 = \cdots = \beta_k = 0$, SCM ≈ 0 y el estadístico tomará valores pequeños; cuando algún β sea distinto de cero, SCM $>$ SCR y el estadístico irá tomando cada vez valores más altos. Por tanto se rechazará la hipótesis nula para valores grandes del estadístico:

- si $F_0 > F_\alpha$: se rechaza H_0
- si $F_0 \leq F_\alpha$: no se rechaza H_0

4.5 Ejemplo

Queremos contrastar si $\beta_{mom\_age} = \beta_{mom\_hs} = \beta_{mom\_iq} = 0$ en el modelo *kid_score ~ mom_hs + mom_iq + mom_age*, es decir, si estos regresores influyen en la puntuación obtenida en el test.

```
(SCT = sum((d$kid_score - mean(d$kid_score))^2) )
```

```
## [1] 180386.2
```

```
(SCR = sum(resid(m1)^2))
```

```
## [1] 141605
```

```
(SCM = SCT - SCR)
```

```
## [1] 38781.13
```

```
(F_0 = SCM/k/(SCR/(n-k-1)))
```

```
## [1] 39.25446
```

```
(F_alfa = qf(0.05, df1 = k, df2 =n-5-1))
```

```
## [1] 0.1171935
```

```
# pvalor
1 - pf(F_0, k, n-k-1)
```

```
## [1] 0
```

- En R:

```
summary(m1)
```

```
##
## Call:
## lm(formula = kid_score ~ . - mom_work, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.289 -12.421   2.399  11.223  50.169
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.98466    9.13013   2.298   0.0220 *
## mom_hssi     5.64715    2.25766   2.501   0.0127 *
## mom_iq       0.56254    0.06065   9.276   <2e-16 ***
## mom_age      0.22475    0.33075   0.680   0.4972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.15 on 430 degrees of freedom
## Multiple R-squared:  0.215,  Adjusted R-squared:  0.2095
## F-statistic: 39.25 on 3 and 430 DF,  p-value: < 2.2e-16
```

Luego se rechaza la hipótesis nula, y al menos uno de los regresores influye en *kid_score*.

5 Contraste para un grupo de coeficientes

Consideremos el modelo de regresión con k regresores:

$$y = X\beta + u, \ dim(\beta) = k \times 1$$

Y consideremos otro modelo de regresión en el que se utilizan m de los k regresores ($m < k$):

$$y = X'\beta' + u', \ dim(\beta') = m \times 1$$

Sea *SCR(k)* la suma de cuadrados residual del primer modelo, y *SCR(m)* la suma de cuadrados residual del segundo modelo. Se puede demostrar que:

$$F_0 = \frac{(SCR(m) - SCR(k))/(k-m)}{SCR(k)/(n-k-1)} \sim F_{k-m,n-k-1}$$

Con este estadístico podemos resolver el contraste

$$H_0 : \text{Los modelos son iguales} H_1 : \text{Los modelos NO son iguales}$$

Si el estadístico toma valores pequeños quiere decir que la suma de cuadrados residual es parecida en ambos modelos, luego se considera que los modelos son equivalentes. Por tanto, se rechazará la hipótesis nula para valores grandes del estadístico:

- si $F_0 > F_\alpha$: se rechaza H_0
- si $F_0 \leq F_\alpha$: no se rechaza H_0

5.1 Ejemplo: contraste para un regresor

Vamos a analizar si el regresor *mom_age* puede eliminarse de la lista. El contraste que resolvemos es $H_0 : \beta_{mom\_age} = 0$ en el modelo *kid_score ~ mom_hs + mom_iq + mom_age*. Para ello lo comparamos con el modelo *kid_score ~ mom_hs + mom_iq*. Si los modelos son equivalentes quiere decir que $\beta_{mom\_age} = 0$:

```
m2 = lm(kid_score ~ mom_hs + mom_iq, data = d)
(SCR2 = sum(resid(m2)^2))
```

```
## [1] 141757.1
```

```
m = 2
(F_0 = ((SCR2 - SCR)/(k-m))/(SCR/(n-k-1)))
```

```
## [1] 0.4617577
```

```
# F_alfa
qf(0.95, k-m, n-k-1)
```

```
## [1] 3.863175
```

```
# pvalor
1-pf(F_0, k-m, n-k-1)
```

```
## [1] 0.4971693
```

Luego no se puede rechazar la hipótesis nula (los modelos son iguales), luego el regresor *mom_age* se puede eliminar del modelo Se obtiene el mismo resultado que con el contraste de la t-student.

- Con R:

```
anova(m2, m1)
```

```
## Analysis of Variance Table
##
## Model 1: kid_score ~ mom_hs + mom_iq
## Model 2: kid_score ~ (mom_hs + mom_iq + mom_work + mom_age) - mom_work
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    431 141757
## 2    430 141605  1    152.06 0.4618 0.4972
```

5.2 Ejemplo: el contraste de regresión múltiple

El contraste de regresión múltiple ($H_0 : \beta_{mom\_hs} = \beta_{mom\_iq} = \beta_{mom\_age} = 0$) también se puede resolver utilizando este estadístico. Los dos modelos a comparar son: *kid_score ~ 1 + mom_hs + mom_iq + mom_age* y *kid_score ~ 1*. El 1 hace referencia al β_0, y se estima por defecto si no se indica explicitamente:

```
m3 = lm(kid_score ~ 1, d)
```

Por tanto

```
anova(m3, m1)
```

```
## Analysis of Variance Table
##
## Model 1: kid_score ~ 1
## Model 2: kid_score ~ (mom_hs + mom_iq + mom_work + mom_age) - mom_work
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    433 180386
## 2    430 141605  3     38781 39.255 < 2.2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.3 Ejemplo: contraste sobre una pareja de regresores

El contraste que resolvemos es $H_0 : \beta_{mom\_iq} = \beta_{mom\_age} = 0$ en el modelo *kid_score ~ mom_hs + mom_iq + mom_age*. Para ello lo comparamos con el modelo:

```
m4 = lm(kid_score ~ mom_hs, data = d)
(SCR4 = sum(resid(m4)^2))
```

```
## [1] 170261.2
```

```
m = 1
(F_0 = ((SCR4 - SCR)/(k-m))/(SCR/(n-k-1)))
```

```
## [1] 43.50887
```

```
# F_alfa
qf(0.05, k-m, n-k-1)
```

```
## [1] 0.05129941
```

```
# pvalor
1-pf(F_0, k-m, n-k-1)
```

```
## [1] 0
```

Luego se rechaza la hipótesis nula.

- Con R:

```
anova(m4, m1)
```

```
## Analysis of Variance Table
##
## Model 1: kid_score ~ mom_hs
## Model 2: kid_score ~ (mom_hs + mom_iq + mom_work + mom_age) - mom_work
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    432 170261
## 2    430 141605  2     28656 43.509 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.4 Ejemplo: contraste de igualdad de regresores

El contraste que resolvemos es $H_0 : \beta_{mom\_iq} = \beta_{mom\_age}$ en el modelo *kid_score ~ mom_hs + mom_iq + mom_age*. Hacemos la comparación con el modelo:

```
m5 = lm(kid_score ~ mom_hs + I(mom_iq + mom_age), data = d)
(SCR5 = sum(resid(m5)^2))
```

```
## [1] 141933.5
```

```
m = 2
(F_0 = ((SCR5 - SCR)/(k-m))/(SCR/(n-k-1)))
```

```
## [1] 0.9974542
```

```
# pvalor
1-pf(F_0, k-m, n-k-1)
```

```
## [1] 0.318489
```

OJO, el modelo m5 es:

$$kid\_score_i = \beta_0 + \beta_1 mom\_hs_i + \beta_2 mom\_iq_i + \beta_2 mom\_age_i + u_i$$

- En R:

```
anova(m5,m1)
```

```
## Analysis of Variance Table
##
## Model 1: kid_score ~ mom_hs + I(mom_iq + mom_age)
## Model 2: kid_score ~ (mom_hs + mom_iq + mom_work + mom_age) - mom_work
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    431 141934
## 2    430 141605  1    328.48 0.9975 0.3185
```

Como vemos no se puede rechazar la hipótesis nula *Los modelos son iguales*, luego no se puede rechazar $H_0 : \beta_{mom\_iq} = \beta_{mom\_age}$.

Por último vamos a resolver este contraste con la t-student. El modelo m1 es

$$kid\_score = \beta_0 + \beta_1 mom\_hssi + \beta_2 mom\_iq + \beta_3 mom\_age + u$$

La hipótesis nula del contraste es: $H_0 : \beta_2 = \beta_3$. Para encontrar el estadístico del contraste tenemos que:

$$\hat{\beta}_2 \sim N(\beta_2, \sigma^2 Q_{3,3}), \quad \hat{\beta}_3 \sim N(\beta_3, \sigma^2 Q_{4,4})$$

Por tanto:

$$\hat{\beta}_2 - \hat{\beta}_3 \sim N(\beta_2 - \beta_3, \sigma^2(Q_{3,3} + Q_{4,4} - 2Q_{3,4}))$$

Es decir:

$$\frac{(\hat{\beta}_2 - \hat{\beta}_3) - (\beta_2 - \beta_3)}{\sqrt{\sigma^2(Q_{3,3} + Q_{4,4} - 2Q_{3,4}))}} \sim N(0,1)$$

Reemplazando σ^2 por $\hat{s}_R^2$ se tiene:

$$\frac{(\hat{\beta}_2 - \hat{\beta}_3) - (\beta_2 - \beta_3)}{\sqrt{\hat{s}_R^2(Q_{3,3} + Q_{4,4} - 2Q_{3,4}))}} \sim t_{n-k-1}$$

Finalmente el estadístico del contraste cuando H_0 es cierta es:

$$t_0 = \frac{\hat{\beta}_2 - \hat{\beta}_3}{\sqrt{\hat{s}_R^2(Q_{3,3} + Q_{4,4} - 2Q_{3,4}))}} \sim t_{n-k-1}$$

```
beta_var = vcov(m1)
(t0 = (coef(m1)[3] - coef(m1)[4])/sqrt(beta_var[3,3] + beta_var[4,4] - 2*beta_var[3,4]))
```

```
##    mom_iq
## 0.9987263
```

```r
# pvalor
2*pt(abs(t0), n - k - 1, lower.tail = F)
```

```
##   mom_iq
## 0.318489
```