UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Inferencia
Soluciones a los ejercicios opcionales

Francesc Carmona

11 de abril de 2018

Ejercicios del libro de Faraway

Ejercicio 3.5

Find a formula relating R^2 and the F-test for the regression.

El contraste de significación de la regresión se basa en el estadístico

$$F = \frac{(TSS - RSS)/(m-1)}{RSS/(n-m)}$$

donde $m = k+1$ es el número de parámetros, n el número de observaciones y RSS la suma de los residuos al cuadrado. $TSS = S_y$ es la suma de las desviaciones al cuadrado de la variable respuesta, ya que cuando todos los coeficientes de las variables son cero, la estimación de β_0 es la media $\bar{y}$.

Por otra parte, el coeficiente de determinación es

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{TSS - RSS}{TSS} \qquad \text{de modo que} \qquad 1 - R^2 = \frac{RSS}{TSS}$$

entonces

$$F = \frac{TSS - RSS}{RSS} \times \frac{n-m}{m-1} = \frac{(TSS - RSS)/TSS}{RSS/TSS} \times \frac{n-m}{m-1} = \frac{R^2}{1-R^2} \times \frac{m-1}{n-m}$$

Ejercicio 3.6

Thirty-nine MBA students were asked about happiness and how this related to their income and social life. The data are found in `happy`*. Fit a regression model with* `happy` *as the response and the other four variables as predictors.*

(a) *Which predictors were statistically significant at the 1 % level?*

```
> library(faraway)
> data(happy)
> fit <- lm(happy ~ ., data=happy)
> summary(fit)


Call:
lm(formula = happy ~ ., data = happy)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-2.7186 -0.5779 -0.1172  0.6340  2.0651

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.072081   0.852543  -0.085   0.9331
money        0.009578   0.005213   1.837   0.0749 .
sex         -0.149008   0.418525  -0.356   0.7240
love         1.919279   0.295451   6.496 1.97e-07 ***
work         0.476079   0.199389   2.388   0.0227 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.058 on 34 degrees of freedom
Multiple R-squared:  0.7102,Adjusted R-squared:  0.6761
F-statistic: 20.83 on 4 and 34 DF,  p-value: 9.364e-09
```

La única variable predictora significativa al nivel $\alpha = 0.01$ es `love`.

(b) *Use the* `table` *function to produce a numerical summary of the response. What assumption used to perform the t-tests seems questionable in light of this summary?*

```
> table(happy$happy)


 2  3  4  5  6  7  8  9 10
 1  1  4  5  2  8 14  3  1
```

Los valores de la variable respuesta **happy** no parecen para nada de una variable con distribución normal.

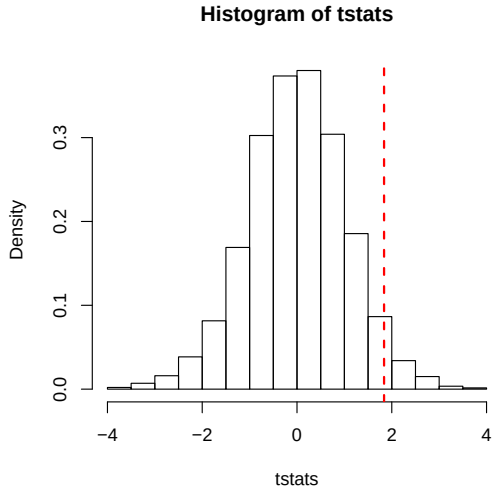
(c) *Use the permutation procedure described in Section 3.3 to test the significance of the money predictor.*

```
> nreps <- 4000
> tstats <- numeric(nreps)
> set.seed(123)
> for(i in 1:nreps){
+    lmods <- lm(happy ~ sample(money) + sex + love + work, data=happy)
+    tstats[i] <- summary(lmods)$coef[2,3]
+ }
> mean(abs(tstats) > abs(summary(fit)$coef[2,3]))

[1] 0.0795
```

La variable predictora **money** no es significativa.

(d) *Plot a histogram of the permutation t-statistics. Make sure you use the the probability rather than frequency version of the histogram.*

```
> hist(tstats, freq = FALSE)
> abline(v=summary(fit)$coef[2,3], lty=2, lwd=2, col="red")
```

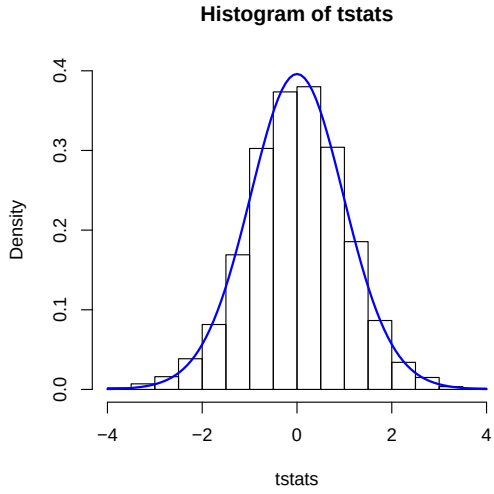


(e) *Overlay an appropriate t-density over the histogram.*

Hint: Use `grid <- seq(-3, 3, length = 300)` *to create a grid of values, then use the* `dt` *function to compute the t-density on this grid and the lines function to superimpose the result.*

Más sencillo con la función `curve()`:

```
> hist(tstats, freq = FALSE, ylim=c(0,0.4))
> curve(dt(x,df=34), col="blue", lwd=2, add=T)
```



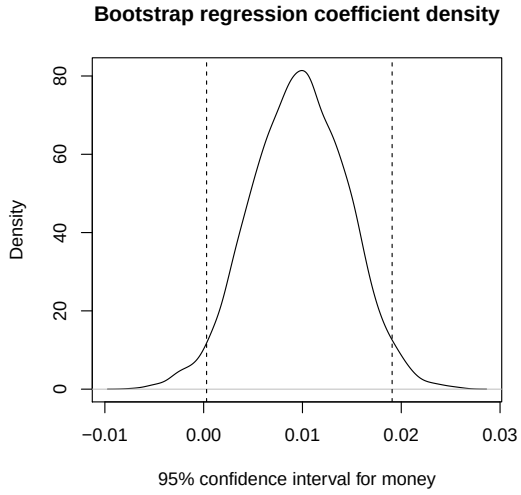
(f) *Use the bootstrap procedure from Section 3.6 to compute 90 % and 95 % confidence intervals for* β_{money}*. Does zero fall within these confidence intervals? Are these results consistent with previous tests?*

```
> set.seed(123)
> nb <- 4000
> coefmat <- matrix(NA,nb,5)
> resids <- residuals(fit)
```

```
> preds <- fitted(fit)
> for(i in 1:nb){
+ booty <- preds + sample(resids, rep = TRUE)
+ bmod <- update(fit, booty ~ .)
+ coefmat[i,] <- coef(bmod)
+ }
> colnames(coefmat) <- c("Intercept",colnames(happy[,2:5]))
> coefmat <- data.frame(coefmat)
> quantile(coefmat[,2],c(0.05,0.95))

        5%        95%
0.00187984 0.01748788

> confint(fit, parm = 2, level=0.9) # Intervalo de confianza con la t de Student

               5 %      95 %
money 0.000763349 0.0183928

> quantile(coefmat[,2],c(0.025,0.975))

       2.5%       97.5%
0.000297174 0.019075282

> confint(fit, parm = 2, level=0.95) # Intervalo de confianza con la t de Student

              2.5 %     97.5 %
money -0.001015941 0.02017209
```

Los intervalos por el método bootstrap no atrapan al cero, de forma que el coeficiente es significativo. Los intervalos calculados con la t de Student son más anchos y para un nivel del 95 % contiene al cero.

```
> plot(density(coefmat[,2]), type="l", xlab="95% confidence interval for money",
+            main="Bootstrap regression coefficient density")
> abline(v=quantile(coefmat[,2],0.025), lty=2)
> abline(v=quantile(coefmat[,2],0.975), lty=2)
```



Ejercicio 3.7

In the `punting` *data, we find the average distance punted and hang times of 10 punts of an American football as related to various measures of leg strength for 13 volunteers.*

(a) *Fit a regression model with* `Distance` *as the response and the right and left leg strengths and flexibilities as predictors. Which predictors are significant at the 5 % level?*

```
> data(punting)
> g <- lm(Distance ~ RStr + LStr + RFlex + LFlex, data=punting)
> (sg <- summary(g))


Call:
lm(formula = Distance ~ RStr + LStr + RFlex + LFlex, data = punting)

Residuals:
    Min      1Q  Median      3Q     Max
-23.941  -8.958  -4.441  13.523  17.016

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -79.6236    65.5935  -1.214    0.259
RStr          0.5116     0.4856   1.054    0.323
LStr         -0.1862     0.5130  -0.363    0.726
RFlex         2.3745     1.4374   1.652    0.137
LFlex        -0.5277     0.8255  -0.639    0.541

Residual standard error: 16.33 on 8 degrees of freedom
Multiple R-squared:  0.7365,Adjusted R-squared:  0.6047
F-statistic:  5.59 on 4 and 8 DF,  p-value: 0.01902
```

Aparentemente, ninguna de las variables predictoras es significativa (significación parcial).

(b) *Use an F-test to determine whether collectively these four predictors have a relationship to the response.*

El test F de significación colectiva de las variables regresoras se puede ver en la última línea del `summary()`. El contraste es significativo.

```
> c(sg$fstatistic,
+   p=pf(sg$fstatistic[1],sg$fstatistic[2],sg$fstatistic[3],lower.tail = F))

     value      numdf      dendf    p.value
5.58994089 4.00000000 8.00000000 0.01902482
```

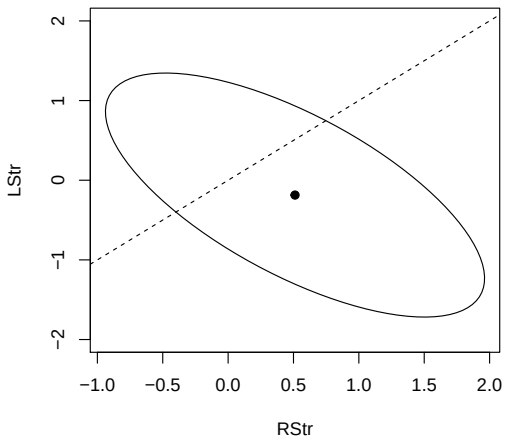
(c) *Relative to the model in (a), test whether the right and left leg strengths have the same effect.*

```
> g1 <- lm(Distance ~ I(RStr + LStr) + RFlex + LFlex, data=punting)
> anova(g1,g)

Analysis of Variance Table

Model 1: Distance ~ I(RStr + LStr) + RFlex + LFlex
Model 2: Distance ~ RStr + LStr + RFlex + LFlex
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      9 2287.4
2      8 2132.6  1    154.72 0.5804  0.468
```

Aceptamos la hipótesis de igualdad.

(d) *Construct a 95 % confidence region for (β_{RStr} , β_{LStr}). Explain how the test in (c) relates to this region.*

```
> require(ellipse)
> plot(ellipse(g,2:3),type="l",ylim=c(-2,2))
> points(coef(g)[2], coef(g)[3], pch=19)
> abline(0,1,lty=2)
```



El punto central es la estimación de los dos coeficientes.

La recta $\beta_{RStr} = \beta_{LStr}$ corta a la elipse, de modo que no podemos rechazar la igualdad de los coeficientes.

(e) *Fit a model to test the hypothesis that it is total leg strength defined by adding the right and left leg strengths that is sufficient to predict the response in comparison to using individual left and right leg strengths.*

```
> g3 <- lm(Distance ~ RStr + LStr, data=punting)
> g4 <- lm(Distance ~ I(RStr + LStr), data=punting)
> anova(g4,g3)

Analysis of Variance Table

Model 1: Distance ~ I(RStr + LStr)
Model 2: Distance ~ RStr + LStr
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     11 3061.3
2     10 2973.1  1    88.281 0.2969 0.5978
```

Aceptamos que el modelo con la suma total es suficiente.

(f) *Relative to the model in (a), test whether the right and left leg flexibilities have the same effect.*

```
> g5 <- lm(Distance ~ RStr + LStr + I(RFlex + LFlex), data=punting)
> anova(g5,g)

Analysis of Variance Table

Model 1: Distance ~ RStr + LStr + I(RFlex + LFlex)
Model 2: Distance ~ RStr + LStr + RFlex + LFlex
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      9 2648.4
2      8 2132.6  1    515.72 1.9346 0.2017
```

Aceptamos la hipótesis de igualdad.

(g) *Test for left-right symmetry by performing the tests in (c) and (f) simultaneously.*

```
> g6 <- lm(Distance ~ I(RStr + LStr) + I(RFlex + LFlex), data=punting)
> anova(g6,g)

Analysis of Variance Table

Model 1: Distance ~ I(RStr + LStr) + I(RFlex + LFlex)
Model 2: Distance ~ RStr + LStr + RFlex + LFlex
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1     10 2799.1
2      8 2132.6  2    666.43 1.25  0.337
```

Aceptamos la simetría.

(h) *Fit a model with* `Hang` *as the response and the same four predictors. Can we make a test to compare this model to that used in (a)? Explain.*

```
> g7 <- lm(Hang ~ RStr + LStr + RFlex + LFlex, data=punting)
> anova(g7,g)
```

```
Warning in anova.lmlist(object, ...): models with response '"Distance"' removed because
response differs from model 1

Analysis of Variance Table

Response: Hang
          Df  Sum Sq Mean Sq F value    Pr(>F)
RStr       1 1.98540 1.98540 30.0416 0.0005867 ***
LStr       1 0.19827 0.19827  3.0001 0.1214978
RFlex      1 0.14699 0.14699  2.2241 0.1742114
LFlex      1 0.00833 0.00833  0.1260 0.7317905
Residuals  8 0.52871 0.06609
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El aviso es debido a que los modelos tienen diferente variable respuesta y NO son comparables.

Ejercicios del libro de Carmona

Ejercicio 5.1

Sean $X \sim N(\mu_1, \sigma), Y \sim N(\mu_2, \sigma)$ variables independientes. En muestras de extensión n_1 de X, n_2 de Y, plantear la hipótesis nula

$$H_0 : \mu_1 = \mu_2$$

mediante el concepto de hipótesis lineal contrastable y deducir el test t de Student de comparación de medias como una consecuencia del test F.

El modelo lineal para contrastar esta hipótesis es

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n_1} \\ y_1 \\ y_2 \\ \vdots \\ y_{n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{n_1} \\ \epsilon_{n_1+1} \\ \epsilon_{n_1+2} \\ \vdots \\ \epsilon_{n_1+n_2} \end{pmatrix}$$

donde los errores verifican las condiciones de Gauss-Markov, ya que son independientes (muestras independientes y poblaciones independientes) y con la misma varianza (σ^2 para las dos poblaciones).

La matriz del diseño es de rango 2 que coincide con el número de parámetros, luego el modelo es de rango máximo y toda función lineal paramétrica es estimable. En nuestro caso

$$H_0 : \mu_1 - \mu_2 = 0$$

Los cálculos son sencillos

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix} \qquad \Rightarrow \qquad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 1/n_1 & 0 \\ 0 & 1/n_2 \end{pmatrix}$$

y las estimaciones son

$$\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1/n_1 & 0 \\ 0 & 1/n_2 \end{pmatrix} \begin{pmatrix} \sum x_i \\ \sum y_i \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$$

de manera que

$$RSS = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \sum x_i^2 + \sum y_i^2 - n_1\bar{x}^2 - n_2\bar{y}^2$$
$$= \sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2 = n_1 s_x^2 + n_2 s_y^2$$

En cuanto a la matriz de diseño cuando la hipótesis sea cierta, ésta se reduce a una columna de $n_1 + n_2$ unos y entonces

$$RSS_H = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{y} = \sum x_i^2 + \sum y_i^2 - (\sum x_i + \sum y_i)^2/(n_1 + n_2)$$
$$= \sum x_i^2 + \sum y_i^2 - (n_1\bar{x} + n_2\bar{y})^2/(n_1 + n_2)$$

de modo que el numerador del test F es

$$\begin{aligned} RSS_H - RSS &= n_1\bar{x}^2 + n_2\bar{y}^2 - (n_1\bar{x} + n_2\bar{y})^2/(n_1 + n_2) \\ &= \frac{1}{n_1 + n_2}((n_1 + n_2)n_1\bar{x}^2 + (n_1 + n_2)n_2\bar{y}^2 - n_1^2\bar{x}^2 - n_2^2\bar{y}^2 - 2n_1n_2\bar{x}\bar{y}) \\ &= \frac{1}{n_1 + n_2}(n_1n_2\bar{x}^2 + n_1n_2\bar{y}^2 - 2n_1n_2\bar{x}\bar{y}) \\ &= \frac{n_1n_2}{n_1 + n_2}(\bar{x} - \bar{y})^2 \\ &= \frac{1}{\frac{1}{n_1} + \frac{1}{n_2}}(\bar{x} - \bar{y})^2 \end{aligned}$$

Como los grados de libertad del numerador son $q = 1$ y los del denominador $n - r = n_1 + n_2 - 2$, el estadístico es

$$F = \frac{(RSS_H - RSS)/q}{RSS/(n - r)} = \frac{\frac{1}{\frac{1}{n_1} + \frac{1}{n_2}}(\bar{x} - \bar{y})^2}{(n_1 s_x^2 + n_2 s_y^2)/(n_1 + n_2 - 2)} = \frac{(\bar{x} - \bar{y})^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{\sigma}^2}$$

y como $F_{1,n_1+n_2-2} = t^2_{n_1+n_2-2}$, la raíz cuadrada del estadístico anterior coincide con el test t de Student de comparación de medias en poblaciones con igual varianza.

Ejercicio 5.2

Una variable Y depende de otra x (variable control no aleatoria) que toma los valores $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$ de acuerdo con el modelo lineal normal

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

Encontrar la expresión del estadístico F para la hipótesis

$$H_0 : \beta_2 = 0$$

La matriz de diseño de este modelo es

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{pmatrix}$$

Sabemos que es de rango máximo, de modo que la hipótesis planteada es contrastable.

La suma de cuadrados de los residuos es $RSS = \mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}$, donde $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ es

```
> X <- matrix(c(1,1,1,
+               1,2,4,
+               1,3,9,
+               1,4,16), byrow = T, ncol=3)
> ImP <- diag(rep(1,4)) - X %*% solve(t(X) %*% X) %*% t(X)
> ImP

      [,1]  [,2]  [,3]  [,4]
[1,]  0.05 -0.15  0.15 -0.05
[2,] -0.15  0.45 -0.45  0.15
[3,]  0.15 -0.45  0.45 -0.15
[4,] -0.05  0.15 -0.15  0.05
```

Por otra parte, la matriz de diseño bajo la hipótesis nula coincide con la del modelo de regresión simple.

```
> X0 <- matrix(c(1,1,
+                1,2,
+                1,3,
+                1,4), byrow = T, ncol=2)
> ImPH <- diag(rep(1,4)) - X0 %*% solve(t(X0) %*% X0) %*% t(X0)
> ImPH

     [,1] [,2] [,3] [,4]
[1,]  0.3 -0.4 -0.1  0.2
[2,] -0.4  0.7 -0.2 -0.1
[3,] -0.1 -0.2  0.7 -0.4
[4,]  0.2 -0.1 -0.4  0.3
```

Así, el numerador del test F para este contraste es

$$RSS_H - RSS = \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\boldsymbol{\omega}})\mathbf{y} - \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\boldsymbol{\Omega}})\mathbf{y} = \mathbf{y}'(\mathbf{P}_{\boldsymbol{\Omega}} - \mathbf{P}_{\boldsymbol{\omega}})\mathbf{y}$$

donde $\mathbf{I} - \mathbf{P}_{\boldsymbol{\omega}} - (\mathbf{I} - \mathbf{P}_{\boldsymbol{\Omega}}) = \mathbf{P}_{\boldsymbol{\Omega}} - \mathbf{P}_{\boldsymbol{\omega}}$ es

```
> ImPH - ImP

      [,1]  [,2]  [,3]  [,4]
[1,]  0.25 -0.25 -0.25  0.25
[2,] -0.25  0.25  0.25 -0.25
[3,] -0.25  0.25  0.25 -0.25
[4,]  0.25 -0.25 -0.25  0.25
```

es decir

$$\frac{1}{4}\begin{pmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

de forma que

$$RSS_H - RSS = \frac{1}{4}\begin{pmatrix} y_1 & y_2 & y_3 & y_4 \end{pmatrix}\begin{pmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}$$
$$= \frac{1}{4}(y_1^2 - 2y_1y_2 - 2y_1y_3 + 2y_1y_4 + y_2^2 + 2y_2y_3 - 2y_2y_4 + y_3^2 - 2y_3y_4 + y_4^2)$$
$$= \frac{1}{4}(y_1 - y_2 - y_3 + y_4)^2$$

con $q = r - r_H = 3 - 2 = 1$ grados de libertad.

En cuanto al denominador del test F, su valor es $RSS/(n-r) = RSS/(4-3) = RSS$.

Finalmente el estadístico F con $1, 1$ grados de libertad es

$$F = \frac{\frac{1}{4}(y_1 - y_2 - y_3 + y_4)^2}{RSS}$$

Por otra parte y dado que los grados de libertad del numerador son $q = 1$, el numerador es mucho más sencillo de hallar si directamente calculamos la t de Student equivalente.

Con $\mathbf{a}' = (0, 0, 1)$, tenemos que $\mathbf{a}'\boldsymbol{\beta} = (0, 0, 1)\boldsymbol{\beta} = \beta_2$. Entonces

$$\mathbf{a}'\hat{\boldsymbol{\beta}} = (0, 0, 1)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

donde la matriz $(0, 0, 1)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ es

```
> a <- c(0,0,1)
> t(a) %*% solve(t(X) %*% X) %*% t(X)

     [,1]  [,2]  [,3] [,4]
[1,] 0.25 -0.25 -0.25 0.25
```

de modo que

$$\hat{\beta}_2 = \mathbf{a}'\hat{\boldsymbol{\beta}} = \frac{1}{4}(y_1 - y_2 - y_3 + y_4)$$

luego

$$t = \frac{\mathbf{a}'\hat{\boldsymbol{\beta}}}{se(\mathbf{a}'\hat{\boldsymbol{\beta}})} = \frac{\hat{\beta}_2}{\sqrt{MSE \cdot \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}} = \frac{\frac{1}{4}(y_1 - y_2 - y_3 + y_4)}{\sqrt{RSS \cdot \frac{1}{4}}}$$

ya que $MSE = RSS/1$ y

```
> t(a) %*% solve(t(X) %*% X) %*% a

     [,1]
[1,] 0.25
```

Finalmente comprobamos que $t^2 = F$.

Ejercicio 5.5

Dado el siguiente modelo lineal normal

$$\begin{aligned} \beta_1 + \beta_2 &= 6.6 \\ 2\beta_1 + \beta_2 &= 7.8 \\ -\beta_1 + \beta_2 &= 2.1 \\ 2\beta_1 - \beta_2 &= 0.4 \end{aligned}$$

Estudiar si se puede aceptar la hipótesis $H_0 : \beta_2 = 2\beta_1$.

La matriz del diseño de este modelo es

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \\ -1 & 1 \\ 2 & -1 \end{pmatrix}$$

y no hay duda que su rango es 2, de modo que es de rango máximo. Eso significa que cualquier función lineal paramétrica es estimable, en particular $2\beta_1 - \beta_2$. Así pues, la hipótesis propuesta es contrastable. La estimación de los parámetros es

```
> X <- matrix(c(1,1,
+               2,1,
+               -1,1,
+               2,-1), byrow = TRUE, ncol=2)
> y <- c(6.6,7.8,2.1,0.4)
> betas <- solve(t(X) %*% X) %*% t(X) %*% y
> betas

      [,1]
[1,] 2.090
[2,] 4.025
```

La estimación de la varianza de los errores es

```
> r <- ncol(X)  # rango máximo
> n <- nrow(X)
> ee <- y - X %*% betas
> RSS <- sum(ee^2)
> (MSE <- RSS/(n-r))

[1] 0.24325
```

Como se trata de una hipótesis con una única fpe, el contraste con la t de Student es

```
> a <- c(2,-1)   # coeficientes de la fpe
> numerador <- t(a) %*% betas
> denominador <- sqrt(MSE * t(a) %*% solve(t(X) %*% X) %*% a)
> t.est <- numerador/denominador
> p.value <- pt(abs(t.est), df = n-r, lower.tail = F) * 2
> c(t.est,p.value)

[1] 0.3898061 0.7342749
```

de modo que el test no es significativo con $\alpha = 0.05$.

También podemos utilizar la función `lm()`

```
> x1 <- X[,1]
> x2 <- X[,2]
> g <- lm(y ~ 0 + x1 + x2)
> g0 <- lm(y ~ 0 + I(x1 + 2*x2))
> anova(g0,g)

Analysis of Variance Table
```

```
Model 1: y ~ 0 + I(x1 + 2 * x2)
Model 2: y ~ 0 + x1 + x2
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      3 0.52346
2      2 0.48650  1  0.036962 0.1519 0.7343
```

Observemos que bajo la hipótesis nula $\beta_2 = 2\beta_1$, de modo que en el modelo de la hipótesis nula solo queda un parámetro β_1 y sus valores son `x1 + 2*x2`.
En este caso el estadístico es la F con $1, 2$ grados de libertad y se comprueba que los estadísticos verifican $t^2 = F$ y que los p-valores son los mismos.

Ejercicio 5.6

Continuación del ejercicio 3.10:

El transportista discute con un amigo que afirma que el doble de la distancia entre A y B es equivalente a la distancia del trayecto $A \to C \to B$. ¿Podemos aclarar en términos estadísticos su discusión?

En primer lugar, vamos a traducir la pregunta en términos paramétricos. La distancia entre A y B es α y el trayecto $A \to C \to B$ es $\beta + \gamma$. Así pues, la hipótesis que plantea el transportista es

$$H_0 : 2\alpha = \beta + \gamma \qquad \Rightarrow \qquad H_0 : 2\alpha - \beta - \gamma = 0$$

Ahora veamos si es contrastable. Para ello la función paramétrica debe ser estimable.
En este caso las fpe deben verificar (ver solución del ejercicio 3.10) $2a_2 = a_1 + 4a_3$. En nuestra hipótesis

$$(a_1, a_2, a_3) = (2, -1, -1) \Rightarrow 2 \cdot (-1) = 2 + 4 \cdot (-1) \qquad \surd$$

de modo que podemos proceder a su contraste.

En el ejercicio 3.10 ya calculamos una estimación de los parámetros:

```
> require(MASS)
> y <- c(533,583,1111,1069)
> X <- matrix(c(2,1,0,
+               0,2,1,
+               2,3,1,
+               4,2,0), byrow=T, ncol=3)
> betas <- ginv(t(X) %*% X) %*% t(X) %*% y
```

Además necesitamos la estimación de la varianza de los errores $\hat{\sigma}^2$.

```
> n <- length(y)
> r <- 2
> residuos <- y - X %*% betas
> sigma2 <- sum(residuos^2)/(n-r)
```

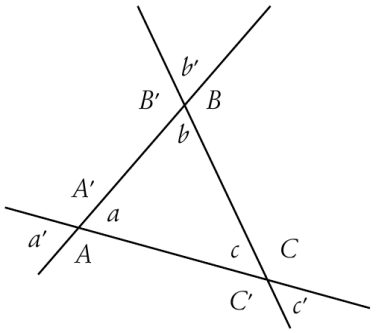
El estadístico t de Student es

```
> a <- c(2,-1,-1)
> numerador <- t(a) %*% betas
> denominador <- sqrt(sigma2 * t(a) %*% ginv(t(X) %*% X) %*% a)
> t.est <- numerador/denominador
> p.value <- pt(abs(t.est), df = n-r, lower.tail = F) * 2
> c(t.est,p.value)

[1] -15.725954991   0.004019217
```

El resultado estadístico es que no podemos dar la razón al transportista.

Ejercicio 5.10

Supongamos que cada uno de los valores $x_1, x_2, \ldots, x_{12}$ son las observaciones de los ángulos a, a',A, A',b, b', B, B',c, c',C, C' del triángulo del gráfico adjunto. Los errores de las observaciones $\epsilon_1, \ldots, \epsilon_{12}$ se asume que son independientes y con distribución $N(0, \sigma)$.



Antes de escribir el modelo asociado a estos datos observemos que, aunque aparentemente hay 12 parámetros $a, a', \ldots$, éstos están ligados por las conocidas propiedades de un triángulo, es decir

$$a = a' \qquad A = A' \qquad a + A = 180 \qquad a + b + c = 180$$

y de forma similar para b, b', B, B' y c, c', C, C'. El conjunto de estas relaciones nos conduce a que, realmente, sólo hay dos parámetros independientes, les llamaremos α y β. Si trasladamos a la izquierda las cantidades 180 y con estos parámetros, el modelo es

$$\begin{array}{llll} y_1 = \alpha + \epsilon_1 & y_2 = \alpha + \epsilon_2 & y_3 = -\alpha + \epsilon_3 & y_4 = -\alpha + \epsilon_4 \\ y_5 = \beta + \epsilon_5 & y_6 = \beta + \epsilon_6 & y_7 = -\beta + \epsilon_7 & y_8 = -\beta + \epsilon_8 \\ y_9 = -\alpha - \beta + \epsilon_9 & y_{10} = -\alpha - \beta + \epsilon_{10} & y_{11} = \alpha + \beta + \epsilon_{11} & y_{12} = \alpha + \beta + \epsilon_{12} \end{array}$$

donde

$$\begin{array}{llll} y_1 = x_1 & y_2 = x_2 & y_3 = x_3 - 180 & y_4 = x_4 - 180 \\ y_5 = x_5 & y_6 = x_6 & y_7 = x_7 - 180 & y_8 = x_8 - 180 \\ y_9 = x_9 - 180 & y_{10} = x_{10} - 180 & y_{11} = x_{11} & y_{12} = x_{12} \end{array}$$

Deseamos contrastar la hipótesis de que el triángulo es equilátero, es decir, que $a = b = c = 60$. Pero si $a = 60, b = 60$, c es automáticamente 60, luego la hipótesis es

$$H_0 : \alpha = \beta = 60$$

con 2 grados de libertad, no 3.
Resolver el contraste.

El modelo lineal que se plantea tiene $n = 12$ observaciones con matriz de diseño

$$\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ -1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & -1 \\ 0 & -1 \\ -1 & -1 \\ -1 & -1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$$

de rango $r = 2$.

En forma matricial la hipótesis es

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c} \qquad \Rightarrow \qquad H_0 : \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 60 \\ 60 \end{pmatrix}$$

Entonces tenemos

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 8 & 4 \\ 4 & 8 \end{pmatrix} = 4 \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \qquad \Rightarrow \qquad (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{4} \begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix} = \frac{1}{12} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \frac{1}{12} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix}$$

y la estimación de los parámetros

$$\hat{\alpha} = \frac{1}{12}[2(y_1 + y_2 - y_3 - y_4) - (y_5 + y_6 - y_7 - y_8) - (y_9 + y_{10} - y_{11} - y_{12})]$$
$$\hat{\beta} = \frac{1}{12}[-(y_1 + y_2 - y_3 - y_4) + 2(y_5 + y_6 - y_7 - y_8) - (y_9 + y_{10} - y_{11} - y_{12})]$$

Con estas estimaciones y como $\mathbf{A} = \mathbf{I}$ y $q = 2$, el numerador del test F es

$$\begin{pmatrix} \hat{\alpha} - 60 & \hat{\beta} - 60 \end{pmatrix} 4 \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \hat{\alpha} - 60 \\ \hat{\beta} - 60 \end{pmatrix} \frac{1}{2}$$

$$= 2[(\hat{\alpha} - 60)(2\hat{\alpha} + \hat{\beta} - 180) + (\hat{\beta} - 60)(\hat{\alpha} + 2\hat{\beta} - 180)]$$
$$= 4[\hat{\alpha}^2 + \hat{\beta}^2 + \hat{\alpha}\hat{\beta} - 180\hat{\alpha} - 180\hat{\beta} + 60 \cdot 180]$$

El denominador es $RSS/(n - r) = RSS/10$.

Ejemplo:

```
> X <- matrix(c(1,0,1,0,-1,0,-1,0,
+               0,1,0,1,0,-1,0,-1,
+               -1,-1,-1,-1,1,1,1,1), byrow = T, ncol=2)
> betas <- c(58,61)
> set.seed(123)
> y <- X %*% betas + rnorm(12,mean=0,sd=2)
> x1 <- X[,1]
> x2 <- X[,2]
> g <- lm(y ~ 0 + x1 + x2)
> cc <- coef(g)
> a <- c(cc[1]-60,cc[2]-60)
> XtX <- t(X) %*% X
> numerador <- t(a) %*% XtX %*% a / 2
> # numerador <- 4*(cc[1]^2 + cc[2]^2 + cc[1]*cc[2] - 180*cc[1]- 180*cc[2] + 60*180)
> denominador <- deviance(g)/10
> F <- numerador/denominador
> pf(F,2,10,lower.tail = F)

           [,1]
[1,] 0.9988585
```