Practica Propuesta 1

Pablo Marcos Parra

EJERCICIO 1

La hoja de datos **usair026(library(eftar))** contiene datos para estudiar la concentración de dióxido de azufre en el aire (medida en microgramos por metro cúbico) de 41 ciudades de los EEUU de América como función de una serie de variables climáticas y de población. Los datos están generalmente calculados como las medias entre los años 1969 y 1971.

Se tienen 41 observaciones de las variables:

y: Concentración de dióxido de azufre (SO_2) en el aire

x1: Media anual de temperatura (ºF)

x2: Número de fábricas con más de 20 empleados

x3: Tamaño de la población en miles de habitantes en el censo de 1970

x4: Velocidad de viento media anual en millas por hora

x5: Precipitación media anual en millas por hora

x6: Número medio de días con precipitación

Apartado A

Estima los coeficientes del modelo lineal en el que la variable respuesta es la concentración de dióxidode azufre (SO_2) en el aire y las variables explicativas son el resto de variables. Calcula los p-valores delos tests parciales de todos los coeficientes (salvo el término independiente). Estima los coeficientes del modelo lineal en el que se elimina la variable explicativa con un p-valor mayor y calcula los p-valores delos test parciales para este nuevo modelo. Con estos p-valores y el coeficiente de determinación decidecuál de los dos modelos es preferible utilizar.

```
library(eftar)
attach(usair026)
mod1<-lm(y ~ ., data=usair026); mod1
```

```
##
## Call:
## lm(formula = y ~ ., data = usair026)
##
## Coefficients:
## (Intercept)           x1           x2           x3           x4           x5
##   111.72848     -1.26794      0.06492     -0.03928     -3.18137      0.51236
##          x6
##    -0.05205
```

```
summary(mod1)
```

```
##
## Call:
## lm(formula = y ~ ., data = usair026)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.004  -8.542  -0.991   5.758  48.758
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 111.72848   47.31810   2.361 0.024087 *
## x1           -1.26794    0.62118  -2.041 0.049056 *
## x2            0.06492    0.01575   4.122 0.000228 ***
## x3           -0.03928    0.01513  -2.595 0.013846 *
## x4           -3.18137    1.81502  -1.753 0.088650 .
## x5            0.51236    0.36276   1.412 0.166918
## x6           -0.05205    0.16201  -0.321 0.749972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.64 on 34 degrees of freedom
## Multiple R-squared:  0.6695, Adjusted R-squared:  0.6112
## F-statistic: 11.48 on 6 and 34 DF,  p-value: 5.419e-07
```

Al hacer *summary* estamos obteniendo los p-valores de los test parciales. Como se puede observar, la variable que eliminaremos es la variable **x6** ya que es la que tiene un p-valor mayor (0.749972). Planteo ahora el modelo nuevo sin esa variable:

```
mod2<-update(mod1,. ~ . - x6); mod2
```

```
##
## Call:
```

```
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = usair026)
##
## Coefficients:
## (Intercept)           x1           x2           x3           x4           x5
##   100.15245     -1.12129      0.06489     -0.03933     -3.08240      0.41947
```

```
summary(mod2) #En este caso no hay ninguno que sea no significativo
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = usair026)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.253  -7.655  -0.581   6.059  49.438
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 100.15245   30.27521   3.308 0.002182 **
## x1           -1.12129    0.41586  -2.696 0.010707 *
## x2            0.06489    0.01554   4.174 0.000188 ***
## x3           -0.03933    0.01494  -2.633 0.012499 *
## x4           -3.08240    1.76562  -1.746 0.089622 .
## x5            0.41947    0.21624   1.940 0.060498 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.45 on 35 degrees of freedom
## Multiple R-squared:  0.6685, Adjusted R-squared:  0.6212
## F-statistic: 14.12 on 5 and 35 DF,  p-value: 1.409e-07
```

Ahora debemos elegir cuál de los dos modelos usaremos de aquí en adelante. Para ello nos fijaremos en el R cuadrado ajustado (información que también nos da *summary*). Como el R cuadrado del segundo modelo es mayor que el del primer modelo, **nos quedaremos y trabajaremos con ese segundo modelo (el que no tiene x6).**

Apartado B

Calcular los p-valores de los test secuenciales para ver si las variables climáticas (x1,x4,x5 y x6) que se incluyen en el modelo y las variables sobre la población (x2 y x3) para contrastar si aportan algo al conocimiento de la variable respuesta. Realizar los mismos contrastes en orden inverso.

```
m0<-lm(y ~ 0)
m1<-lm(y ~ 1)
m2<-lm(y ~ x1 + x4 + x5)
m3<-lm(y ~ x1 + x3 + x5 + x2 + x3)
anova(m0,m1,m2,m3)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 0
## Model 2: y ~ 1
## Model 3: y ~ x1 + x4 + x5
## Model 4: y ~ x1 + x3 + x5 + x2 + x3
##   Res.Df   RSS Df Sum of Sq        F    Pr(>F)
## 1     41 59058
## 2     40 22038  1     37020 167.8171 4.052e-15 ***
## 3     37 16419  3      5619   8.4909 0.0002139 ***
## 4     36  7942  1      8477  38.4280 3.770e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como vemos, como los p-valores son muy pequeños podemos decir que todas las varables aportan conocimiento a la variable respuesta.

Ahora veremos el mismo contraste en orden inverso:

```
m2<-lm(y ~ x2 + x3)
m3<-lm(y ~ x2 + x3 + x1 + x4 + x5)
anova(m0,m1,m2,m3)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 0
## Model 2: y ~ 1
## Model 3: y ~ x2 + x3
## Model 4: y ~ x2 + x3 + x1 + x4 + x5
##   Res.Df   RSS Df Sum of Sq        F   Pr(>F)
## 1     41 59058
## 2     40 22038  1     37020 177.3630 2.89e-15 ***
## 3     38  9117  2     12921  30.9528 1.82e-08 ***
## 4     35  7305  3      1811   2.8926  0.04901 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De manera inversa también obtenemos p-valores muy pequeños.

Apartado C

Calcula, según el modelo ajustado, cual sería la predicción para la concentración de dióxido de azufre yun intervalo de predicción del 99 por ciento en una ciudad con los mismos valores de las variables explicativas que Cincinnati.

Usaremos para este apartado, el segundo modelo (aunque con el primer modelo se haría de manera análoga):

```
datos<-cbind(usair026[26,]) #bindeamos la fila 26 (cincinnati) a la variable datos
predict(mod2,newdata = datos,interval="confidence",level = 0.99)
```

```
##                 fit      lwr      upr
## Cincinnati 46.25316 32.03423 60.47209
```

EJERCICIO 2

La hoja de datos esparragos (*library(eftar)*) contiene 40 observaciones de las variables:

clase:Clase de espárrago.

peso:Peso del espárrago en gramos.

fibra.sensorial:Peso de la cantidad de fibra sensorial del espárrago en gramos.

La fibra sensorial es la cantidad de fibra del espárrago blanco que el consumidor es capaz de detectar sensorialmente. Para realizar esta medida se diseña un procedimiento que simula el proceso de masticación. Como el peso de la cantidad de fibra sensorial está relacionado con el peso del espárrago dentro de cada una de las clases, resulta más conveniente utilizar como **variable respuesta la proporción (o tanto por ciento) del peso del espárrago que corresponde a la fibra sensorial**. El interés de los datos está en estudiar como varía la cantidad de fibra sensorial según la clase de los espárragos.

Se toman 8 observaciones de cada una de las siguientes clases de espárragos blancos. Los niveles c.china y c.peru corresponden a espárragos blancos extra en conserva originarios de China y Perú, mientras que el resto de los niveles corresponden a espárragos frescos. Los espárragos del nivel navarra son espárragos blancos frescos denominación de origen Navarra. Por último, las dos variedades de espárrago blanco fresco que se producen en Tudela de Duero Jacques Marionnet 2001 y thielim corresponden a los niveles td.jm2001y td.thielim.

Todos los p-valores pedidos se deben realizar dos veces una: con la función linearHypothesis de library(car)y otra con la función anova.

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.4
```

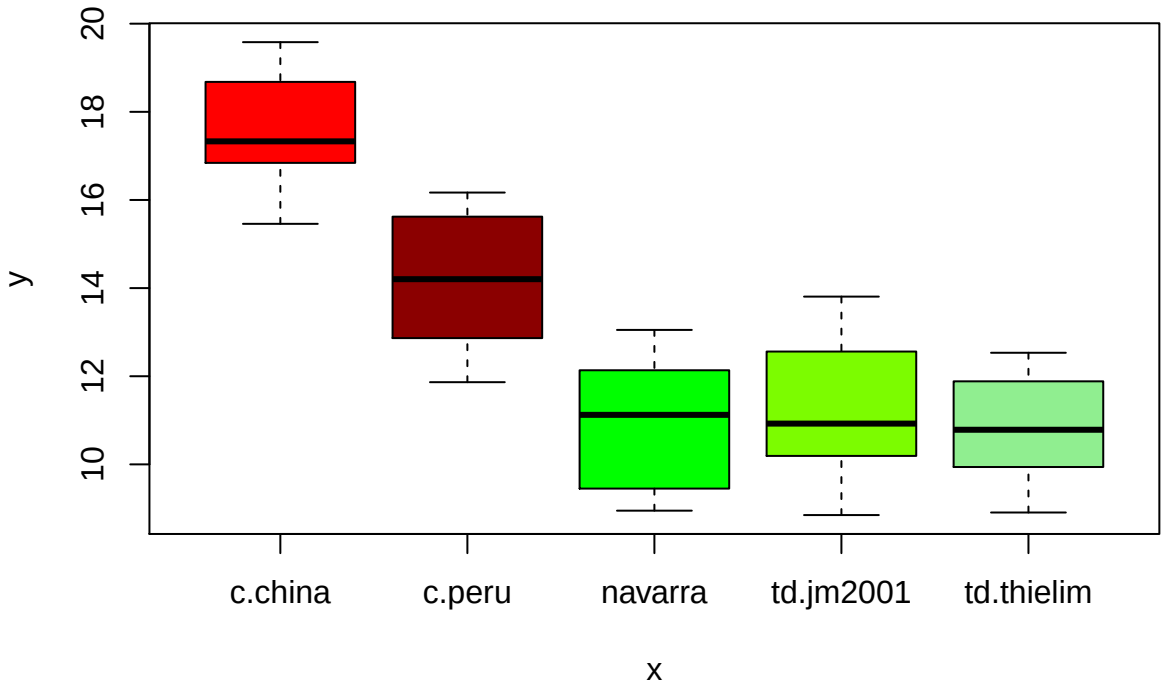
```
## Loading required package: carData
```

```
attach(esparragos)
```

Apartado A

Calcula los valores observados de la variable respuesta y representala en un diagrama de cajas múltiple agrupando por la clase del espárrago.

```
y<-fibra.sensorial*100/peso
esparragos<-cbind(esparragos,y)
plot(clase,y,col= c("red", "red4", "green","lawngreen","lightgreen"))
```



A la hora de realizar cualquier procedimiento anova usaremos lo siguiente:

```
Z<-diag(nlevels(clase)) [clase,]
f<- clase
Z<-diag(nlevels(f))[f,]
modelo1 <- lm(y ~ Z-1,);
modelo1
```

```
##
## Call:
## lm(formula = y ~ Z - 1)
##
## Coefficients:
##    Z1     Z2     Z3     Z4     Z5
## 17.59  14.18  10.93  11.25  10.83
```

Este *modelo1* es el modelo de la regresión lineal inicial para estos datos.

Apartado B

Utiliza el modelo lineal en el que los coeficientes vienen dados por la media de la variable respuesta dentro de cada clase de espárrago para estimar estos parámetros. Comprueba que estas estimaciones coinciden con las medias muestrales.

Ya tenemos el modelo, ahora veremos si coincide:

```
modelo1
```

```
##
## Call:
## lm(formula = y ~ Z - 1)
##
## Coefficients:
##    Z1     Z2     Z3     Z4     Z5
## 17.59  14.18  10.93  11.25  10.83
```

```
tapply(y,clase,mean)
```

```
##    c.china     c.peru    navarra  td.jm2001 td.thielim
##   17.59263   14.17633   10.92762   11.25147   10.83255
```

Como vemos en los resultados, efectivamente coinciden las estimaciones con las medias muestrales (truncadas a dos decimales).

Apartado C

Calcula el p-valor del test que contrasta si existe diferencia en cuanto a la fibra sensorial en las cinco clases de espárragos.

El contraste pedido es:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

```
H0 <-cbind(1,-diag(4));


R <- rbind(c(1,0,0,0,0),H0); R
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]    1   -1    0    0    0
## [3,]    1    0   -1    0    0
## [4,]    1    0    0   -1    0
## [5,]    1    0    0    0   -1
```

```
R1 <- solve(R); R1
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]    1   -1    0    0    0
## [3,]    1    0   -1    0    0
## [4,]    1    0    0   -1    0
## [5,]    1    0    0    0   -1
```

```
X <- Z %*% R1;
modeloA <- lm(y ~ 0 +X[,1]);
linearHypothesis(modelo1,H0)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Z1 - Z2 = 0
## Z1 - Z3 = 0
## Z1 - Z4 = 0
## Z1 - Z5 = 0
##
## Model 1: restricted model
## Model 2: y ~ Z - 1
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     39 353.35
## 2     35  77.22  4    276.13 31.289 4.058e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modeloA, modelo1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 0 + X[, 1]
## Model 2: y ~ Z - 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     39 353.35
```

```
## 2     35  77.22  4    276.13 31.289 4.058e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor obtenido es muy pequeño por lo que que podemos rechazar H_0, rechazar la igualdad de medias.

Apartado D

Calcula el p-valor del test que contrasta si existe diferencia en cuanto a la fibra sensorial entre los espárragos frescos y los espárragos en conserva.

El contraste pedido es:

$$H_0 : \frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4 + \mu_5}{3}$$

```
H1 <-c(3,3,-2,-2,-2)

R <- rbind(H1, cbind(diag(4),0)); R
```

```
##    [,1] [,2] [,3] [,4] [,5]
## H1    3    3   -2   -2   -2
##       1    0    0    0    0
##       0    1    0    0    0
##       0    0    1    0    0
##       0    0    0    1    0
```

```
R1 <- solve(R); R1
```

```
##        H1
## [1,]  0.0 1.0 0.0  0  0
## [2,]  0.0 0.0 1.0  0  0
## [3,]  0.0 0.0 0.0  1  0
## [4,]  0.0 0.0 0.0  0  1
## [5,] -0.5 1.5 1.5 -1 -1
```

```
X <- Z %*% R1;
modeloD <- lm(y ~ X[,-1])
linearHypothesis(modelo1,H1)
```

```
## Linear hypothesis test
##
## Hypothesis:
## 3 Z1  + 3 Z2 - 2 Z3 - 2 Z4 - 2 Z5 = 0
##
## Model 1: restricted model
## Model 2: y ~ Z - 1
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     36 305.89
## 2     35  77.22  1    228.68 103.65 5.295e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modeloD, modelo1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ X[, -1]
## Model 2: y ~ Z - 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     36 305.89
## 2     35  77.22  1    228.68 103.65 5.295e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor obtenido es muy pequeño por lo que podemos rechazar H_0, podemos rechazar igualdad de medias entre los espárragos frescos y de conserva.

Apartado E

Calcula el p-valor del test que contrasta si existe diferencia en cuanto a la fibra sensorial entre las dos clases de espárragos en conserva.

El contraste pedido es:

$$H_0 : \mu_1 = \mu_2$$

```
H2 <-c(1,-1,0,0,0)

R <- rbind(H2,c(3,3,-2,-2,-2),c(0,1,-1,-1,0),c(0,0,1,-1,-1),c(1,1,1,1,1))
R1 <- solve(R); R1
```

```
##        H2
## [1,]  0.5  0.1  0  0.0  0.2
## [2,] -0.5  0.1  0  0.0  0.2
## [3,]  0.0 -0.1  0  0.5  0.3
## [4,] -0.5  0.2 -1 -0.5 -0.1
## [5,]  0.5 -0.3  1  0.0  0.4
```

```
X <- Z %*% R1;
modeloE <- lm(y ~ X[,-1])
linearHypothesis(modelo1,H2)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Z1 - Z2 = 0
##
## Model 1: restricted model
## Model 2: y ~ Z - 1
##
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     36 123.91
## 2     35  77.22  1    46.684 21.16 5.337e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modeloE, modelo1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ X[, -1]
## Model 2: y ~ Z - 1
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     36 123.91
## 2     35  77.22  1    46.684 21.16 5.337e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor obtenido es muy pequeño por lo que podemos rechazar H_0, podemos rechazar la igualdad de medias entre los espárragos en conserva.

Apartado F

Calcula el p-valor del test que contrasta si existe diferencia en cuanto a la fibra sensorial entre las tres clases de espárragos frescos.

El contraste pedido es:

$$H_0 : \mu_3 = \mu_4 = \mu_5$$

$$H_1 : Alguna \;\; es \;\; distinta.$$

Para realizar este contraste lo haremos en dos etapas, es decir, primero contrastaremos $\mu_3 = \mu_4$ y luego contrastaremos $\mu_4 = \mu_5$. Primero contrastaremos:

$$H_0 : \mu_3 = \mu_4$$

```
H30<-c(0,0,1,-1,0) #mu3=mu4
H31<-c(0,0,0,1,-1) #mu4=mu5

R <- rbind(H30,H31,c(0,1,-1,-1,0),c(0,0,1,-1,-1),c(1,1,1,1,1));R
```

```
##     [,1] [,2] [,3] [,4] [,5]
## H30    0    0    1   -1    0
## H31    0    0    0    1   -1
##        0    1   -1   -1    0
##        0    0    1   -1   -1
##        1    1    1    1    1
```

```
R1 <- solve(R); R1
```

```
##      H30 H31
## [1,]  -7  -4 -1  5 1
## [2,]   3   2  1 -2 0
## [3,]   2   1  0 -1 0
## [4,]   1   1  0 -1 0
## [5,]   1   0  0 -1 0
```

```
X <- Z %*% R1;
modeloF0 <- lm(y ~ X[,-1])

linearHypothesis(modelo1,H30)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Z3 - Z4 = 0
##
## Model 1: restricted model
## Model 2: y ~ Z - 1
##
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     36 77.64
## 2     35 77.22  1   0.41952 0.1901 0.6655
```

```
anova(modeloF0, modelo1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ X[, -1]
## Model 2: y ~ Z - 1
```

```
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     36 77.64
## 2     35 77.22  1   0.41952 0.1901 0.6655
```

Como se puede observar, para la matriz R hemos usado ambas hipótesis para así luego no repetir calculos. El p-valor que obtenemos para este primer contraste es bastante alto, por lo que no podemos rechazar la igualdad $\mu_3 = \mu_4$.

Como no podemos rechazar la hipótesis, entonces si ahora contrastamos $\mu_4 = \mu_5$ y no lo rechazamos, podremos decir que no rechazamos la hipótesis de igualdad inicial $\mu_3 = \mu_4 = \mu_5$.

$$H_0 : \mu_4 = \mu_5$$

:

```
modeloF1 <- lm(y ~ X[,-2]) #Reutilizamos la X del contraste anterior

linearHypothesis(modelo1,H31)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Z4 - Z5 = 0
##
## Model 1: restricted model
## Model 2: y ~ Z - 1
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     36 77.922
## 2     35 77.220  1   0.70197 0.3182 0.5763
```

```
anova(modeloF1, modelo1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ X[, -2]
## Model 2: y ~ Z - 1
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     36 77.922
## 2     35 77.220  1   0.70197 0.3182 0.5763
```

El p-valor en este caso es bastante alto, por lo que no podemos rechazar la hipótesis de igualdad $\mu_4 = \mu_5$. Como no rechazamos este contraste ni el anterior, entonces podemos decir que NO podemos rechazar el H_0 inicial que era igualdad entre la media de los espárragos frescos.

Apartado G

Calcula el p-valor del test que contrasta si existe diferencia en cuanto a la fibra sensorial entre losespárragos frescos de Navarra y los espárragos frescos de Tudela de Duero.

El contraste pedido es:

$$H_0 : \mu_3 = \frac{\mu_4 + \mu_5}{2}$$

```
H4<-c(0,0,1,-1/2,-1/2)

R <- rbind(H4,c(3,3,-2,-2,-2),c(0,1,-1,-1,0),c(0,0,0,1,-1),c(1,1,1,1,1))
R1 <- solve(R);
X <- Z %*%R1
modeloG<-lm(y ~ 0 + X[,-1])
linearHypothesis(modelo1,H4)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Z3 - 0.5 Z4 - 0.5 Z5 = 0
##
## Model 1: restricted model
## Model 2: y ~ Z - 1
##
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     36 77.29
## 2     35 77.22  1  0.069789 0.0316 0.8599
```

```
anova(modeloG, modelo1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 0 + X[, -1]
## Model 2: y ~ Z - 1
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     36 77.29
## 2     35 77.22  1  0.069789 0.0316 0.8599
```

Como el p-valor obtenido es bastante alto, no podemos rechazar H_0 que era la igualdad de medias entre los esparragos de Navarra y los de Tudela de Duero.

Apartado H

Calcula el p-valor del test que contrasta si existe diferencia en cuanto a la fibra sensorial entre las dosclases de espárragos frescos de Tudela de Duero.

Este contraste lo hemos hecho en el apartado F, pero vuelvo a realizarlo aquí por separado.

El contraste pedido es:

$$H_0 : \mu_4 = \mu_5$$

```
H5<-c(0,0,0,1,-1)
```

Si nos fijamos, en el apartado anterior en la matriz R, la cuarta fila que hay es la que corresponde a este contraste por lo que volvemos a usar esa matriz.

```
modeloH<-lm(y ~ 0 + X[,-4])
linearHypothesis(modelo1,H5)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Z4 - Z5 = 0
##
## Model 1: restricted model
## Model 2: y ~ Z - 1
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     36 77.922
## 2     35 77.220  1   0.70197 0.3182 0.5763
```

```
anova(modeloH, modelo1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 0 + X[, -4]
## Model 2: y ~ Z - 1
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     36 77.922
## 2     35 77.220  1   0.70197 0.3182 0.5763
```

El p-valor en este caso también es bastante alto, por lo que no podemos rechazar H_0, es decir la igualdad de medias entre los espárragos de Tudela de Duero.