Prueba de evaluación continua 1

Regression, modelos i métodos

Núria Mercadé Besora

03 - 12 - 2023

Ejercio 1

El conjunto de datos `possum.csv` consta de nueve mediciones morfométricas recogidas en cada una de las 104 zarigüeyas de cola de cepillo (*brushtail possum*, *Trichosurus caninus Ogilby*), atrapadas en siete localizaciones geográficas de Australia, desde el sur de Victoria hasta el centro de Queensland.

Descripción de las variables:

- **`case`: número de observación**
- **`site`: uno de los siete lugares donde quedaron atrapadas las zarigüeyas. Las localizaciones geográficas fueron, en orden, Cambarville, Bellbird, Whian Whian, Byrangery, Conondale, Allyn River y Bulburin**
- **`pop`: población geográfica clasificadas como `Vic` (Victoria) o `others` (Nueva Gales del Sur o Queensland)**
- **`sex`: sexo con niveles `f` (hembras) o `m` (machos)**
- **`age`: edad**
- **`hdlngth`: longitud de la cabeza (mm)**
- **`skullw`: ancho del cráneo (mm)**
- **`totlngth`: longitud total (cm)**
- **`taill`: longitud de la cola (cm)**
- **`footlgth`: longitud del pie**
- **`earconch`: longitud de la parte externa de la oreja**
- **`eye`: distancia desde el canto medial al canto lateral del ojo derecho**
- **`chest`: circunferencia del pecho (cm)**
- **`belly`: circunferencia del vientre (cm)**

Se cree que las características morfológicas que predicen linealmente la longitud total de las zarigüeyas son la longitud de la cola y la de la cabeza. Según los expertos, también parece que la relación entre la longitud de la cabeza y la longitud total difiere entre las dos poblaciones consideradas en el estudio, luego habrá que añadir como variables explicativas `pop` y la interacción con `hdlngth`.

(a) Ajustar el modelo correspondiente a la hipótesis de partida y que además incluya el sexo como variable regresora. ¿Es significativo el modelo obtenido? ¿Qué test estadístico se emplea para contestar a esta pregunta? Plantear la hipótesis nula y la alternativa del test.

Siguiendo la hipòtesis planteada en el enunciado, ajustaremos una regresión en la que la variable respuesta sea la longitud total, `totlngth`, y las variables predictoras `sex`, `hdlngth`, `taill`, y `pop`. Además, como se tiene conocimiento que la relación entre longitud de la caebza y longitud total difere entre las poblaciones del estudio, icluimos un termino de interacción entre la población, `pop`, y la longitud de la cabeza, `hdlngth`.

A continuación leemos los datos, los processamos para eliminar observaciones con datos perdidos, y construimos el modelo.

```
datos <- read.csv(here::here("possum.csv")) # leemos datos
# Eliminamos observaciones con missings
datos <- datos %>%
  filter(if_all(names(datos), ~ !(is.na(.x) | .x == "NA")))
# Contstuimos modelo regressión lineal, donde incluimos el termino de
# interacción entre la población y la longitud de la cabeza (pop*hdlngth)
reg1 <- lm(totlngth ~ sex + hdlngth + taill + pop + pop*hdlngth, data = datos)
# Mostramos summary
summary(reg1)
```

```
Call:
lm(formula = totlngth ~ sex + hdlngth + taill + pop + pop * hdlngth,
    data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7734 -1.4540  0.2365  1.3352  4.5573

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.9583     7.0279  -0.563 0.574605
sexm           -1.1158     0.4352  -2.564 0.011912 *
hdlngth         0.5103     0.0673   7.583 2.26e-11 ***
taill           1.1683     0.1321   8.843 4.89e-14 ***
popVic        -46.4016    13.7677  -3.370 0.001087 **
hdlngth:popVic  0.5316     0.1474   3.606 0.000498 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.036 on 95 degrees of freedom
Multiple R-squared:  0.7765,    Adjusted R-squared:  0.7647
F-statistic: 66.01 on 5 and 95 DF,  p-value: < 2.2e-16
```

Como se ve en el summary, el p-valor del modelo obtenido es menor a 0.05, lo que indica que es estadísticamente significativo al nivel de significación del 5%. El test estadístico que se utiliza para és el F-test, donde se contrasta la hipótesis nula (H_0) en la qual se asume que no hay ningúna relación de las variables con la variable respuesta, con la hipótesis alternativa (H_1) que al menos un variable si que tenga efecto:

$$H_0 : \beta_{sex} = \beta_{hdlngth} = \beta_{taill} = \beta_{pop} = \beta_{pop \cdot hdlngth} = 0$$

$$H_1 : \beta_i \neq 0$$

para algún \$j = \$ `sex`, `hdlngth`, `taill`, `pop`, o `pop:hdlngth`.

(b) ¿Se puede decir que la relación entre la longitud total y la de la cabeza de una de las dos poblaciones (Victoriana o Queensland/Nueva Gales) es significativamente más intensa que la otra? Argumentar en función de los resultados obtenidos del modelo aplicado en el apartado anterior. ¿Qué test estadístico se necesita para contestar a esta pregunta?

Si nos fijamos en el coeficiente del termino de interacción entre la población y la longitud de la cabeza, `hdlngth:popVic`, este es estadísticamente significativo (p-valor menor a 0.05). Por lo tanto, podemos considerar que en una de las dos poblaciones la relación entre la longitud total y la de la cabeza es significativamente más intensa que en la otra. Concretamente, como el coeficiente es positivo, la relación es más intensa en la población de Victoria.

Para contestar esta pregunta se utiliza el T-test, en el cual se compara la distribución de T de Student bajo la hipótesis nula $H_0 : \beta_{hdlngth*pop} = 0$, frente a la alternativa $H_1 : \beta_{hdlngth*pop} \neq 0$.

(c) Calcular los intervalos de confianza al 95% para los parámetros que acompañan a la variable taill (longitud de la cola) y sex. ¿Qué interpretación práctica tienen la estimación por intervalo de estos dos parámetros?

A continuación usamos R para obtener los intervalos de confianza del 95% para `taill` y `sex`.

```
confint(reg1, level = 0.95) [c("taill", "sexm"), ]
```

```
           2.5 %     97.5 %
taill  0.9060702  1.4306292
sexm  -1.9796793 -0.2518881
```

El hecho que para ninguna de las dos variables los intervalos cruzan el zero, nos indica que tienen significancia estadística al 5% (como se confirma por los p-valores en el summary). Además, que los intervalos sean positivos para la longitud de la cola, indica que cuanto más larga, más longitud del animal. Para la variable sexo, se usa como referencia el masculino y los valores de los intervalos son negativos. Esto significa que la longitud total de los machos es menor a la de las hembras.

(d) Calcular un intervalo de predicción al 95% para la longitud total de una zarigüeya hembra Victoriana de longitud de cabeza de 92.5 mm y de cola de 35.5 cm. Comprobar previamente que los valores observados no suponen una extrapolación.

En primer lugar veamos un resumen de los parámetros estadísticos que describen el conjunto de datos. Vemos que los datos que nos dan para la predicción no suponen una extrapolación, de hecho la longitud de la cabeza esta muy cerca de la media de los datos. Con respeto al sexo y la población, las dos categorías en cada una están representadas más o menos equitativamente.

```
# pasamos sex y pop a factor para poder obtener la frequencia absoluta de sus
# categorias en el summary
summary(datos %>%
          mutate(pop = as.factor(pop),
                 sex = as.factor(sex)))
```

```
      case             site          pop     sex         age
 Min.   :  1.00   Min.   :1.000   other:58   f:42   Min.   :1.000
```

```
 1st Qu.: 26.00   1st Qu.:1.000   Vic  :43   m:59   1st Qu.:2.000
 Median : 54.00   Median :4.000                     Median :3.000
 Mean   : 52.76   Mean   :3.673                     Mean   :3.822
 3rd Qu.: 79.00   3rd Qu.:6.000                     3rd Qu.:5.000
 Max.   :104.00   Max.   :7.000                     Max.   :9.000
    hdlngth          skullw         totlngth         taill
 Min.   : 82.50   Min.   :50.00   Min.   :75.00   Min.   :32.00
 1st Qu.: 90.70   1st Qu.:55.00   1st Qu.:84.50   1st Qu.:36.00
 Median : 92.90   Median :56.40   Median :88.00   Median :37.00
 Mean   : 92.73   Mean   :56.96   Mean   :87.27   Mean   :37.05
 3rd Qu.: 94.80   3rd Qu.:58.10   3rd Qu.:90.00   3rd Qu.:38.00
 Max.   :103.10   Max.   :68.60   Max.   :96.50   Max.   :43.00
    footlgth       earconch          eye            chest          belly
 Min.   :60.3   Min.   :41.30   Min.   :12.80   Min.   :22.00   Min.   :25.00
 1st Qu.:64.5   1st Qu.:44.80   1st Qu.:14.40   1st Qu.:25.50   1st Qu.:31.00
 Median :67.9   Median :46.80   Median :14.90   Median :27.00   Median :32.50
 Mean   :68.4   Mean   :48.13   Mean   :15.05   Mean   :27.06   Mean   :32.64
 3rd Qu.:72.5   3rd Qu.:52.00   3rd Qu.:15.70   3rd Qu.:28.00   3rd Qu.:34.00
 Max.   :77.9   Max.   :56.20   Max.   :17.80   Max.   :32.00   Max.   :40.00
```

A continuación hacemos la predicción y calculamos sus intervalos al 95%:

```r
# data frame con nuevos datos
nuevos_datos <- tibble(sex = "f", pop = "Vic", hdlngth = 92.5, taill = 35.5)
# prediccion
predict(reg1, newdata = nuevos_datos, interval = "prediction", level = .95)
```

```
      fit      lwr      upr
1 87.4934 83.38307 91.60373
```

(e) Considerar un modelo más general que incluye el resto de variables morfológicas y decidir si nos podemos quedar con el modelo reducido del apartado (a). Escribir en forma paramétrica las hipótesis H0 y H1 de este contraste. Comparar el ajuste de ambos modelos.

Ajustamos el modelo general con el resto de variables.

```r
# Contstuimos modelo regressión incluyendo el resto de variables
reg0 <- lm(totlngth ~ sex + hdlngth + taill + pop + skullw + footlgth +
             earconch + eye + chest + belly + pop*hdlngth,
           data = datos)
# Mostramos summary
summary(reg0)
```

```
Call:
lm(formula = totlngth ~ sex + hdlngth + taill + pop + skullw +
    footlgth + earconch + eye + chest + belly + pop * hdlngth,
    data = datos)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.9653 -1.4974  0.1364  1.2930  5.1007

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.24974    8.84837  -0.028 0.977546
sexm           -1.05473    0.46004  -2.293 0.024222 *
hdlngth         0.41646    0.10258   4.060 0.000105 ***
taill           1.13391    0.14573   7.781 1.22e-11 ***
popVic        -47.65036   14.32069  -3.327 0.001276 **
skullw          0.05336    0.10307   0.518 0.605952
footlgth        0.09575    0.10913   0.877 0.382644
earconch       -0.12323    0.14392  -0.856 0.394171
eye             0.04326    0.22197   0.195 0.845931
chest           0.16825    0.16397   1.026 0.307626
belly          -0.08193    0.10394  -0.788 0.432644
hdlngth:popVic  0.54627    0.15340   3.561 0.000595 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.055 on 89 degrees of freedom
Multiple R-squared:  0.7866,    Adjusted R-squared:  0.7602
F-statistic: 29.82 on 11 and 89 DF,  p-value: < 2.2e-16
```

Comparamos ambos modelos con F-test. La hipòtesis nula a contrastar es

$$H_0 : \beta_{skullw} = \beta_{footlgth} = \beta_{earconch} = \beta_{eye} = \beta_{chest} = \beta_{bely} = 0$$

y la hipotesi alternativa es

$$H_1 : \beta_i \neq 0$$

on i representa qualsevol de les variables `skullw`, `footlgth`, `earconch`, `eye`, `chest`, o `bely`.

A continuación se realiza el contraste de hipotesi. EL p-valor resultante es mayor a 0.05. Basandonos en el nivel de significación del 5%, no descartaremos la hipòtesis nul · la, por lo que se justifica la utilización del model más sencillo.

```r
# ANOVA
anova(reg1, reg0)
```

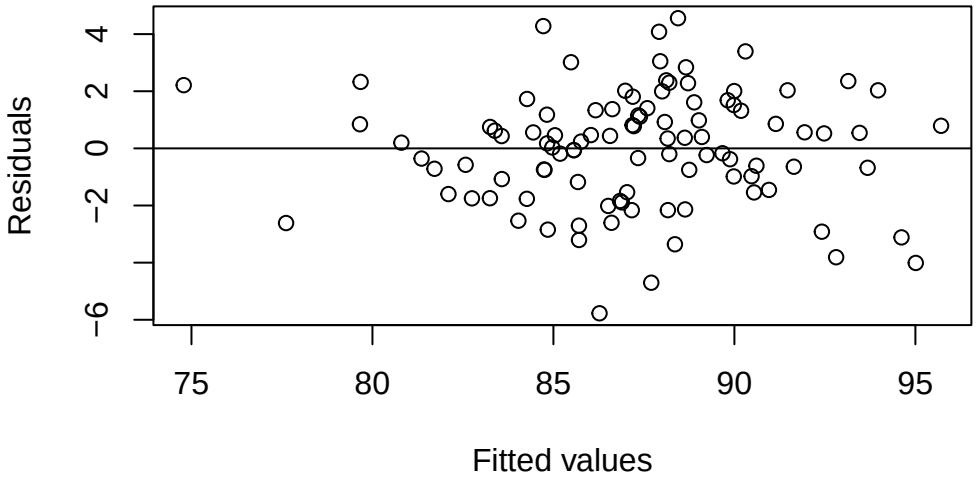
```
Analysis of Variance Table

Model 1: totlngth ~ sex + hdlngth + taill + pop + pop * hdlngth
Model 2: totlngth ~ sex + hdlngth + taill + pop + skullw + footlgth +
    earconch + eye + chest + belly + pop * hdlngth
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     95 393.64
2     89 375.87  6    17.774 0.7015 0.6491
```

(f) Verificar las hipótesis de Gauss-Markov y la normalidad de los residuos del modelo seleccionado en el apartado anterior. Realizar una completa diagnosis del modelo para ver si se cumplen las condiciones del modelo de regresión y estudiar la presencia de valores atípicos, de alto leverage y/o puntos influyentes. Construir los gráficos correspondientes y justificar su interpretación. ¿Podemos

considerar el modelo ajustado como fiable?

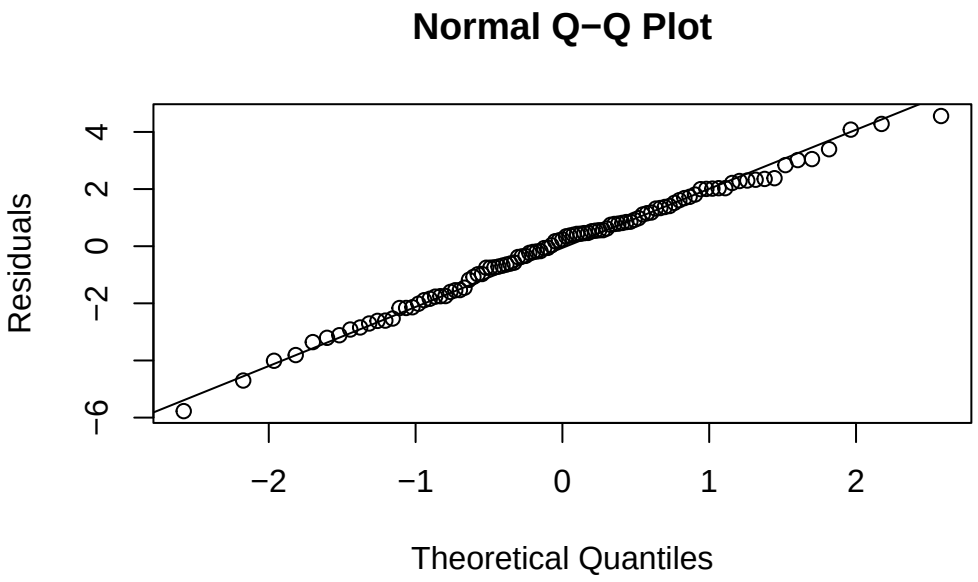
Empezamos comprobando si se verifica la suposición de varianza constante de los errores. Para ello representamos en un diagrama de puntos los valores ajustados frente a las residuos. Una inspección visual del gráfico nos alerta de una ligera no-linealidad ademas no constancia de la varianza de los errores.

```
# Scatter plot de valors ajustats vs. residus
plot(reg1$fitted.values, reg1$residuals,
     xlab = "Fitted values", ylab = "Residuals")
# recta a l'origen amb pendent 0
abline(h = 0)
```



A continuación comprobamos la suposición de normalidad. Empezamos por representar el gráfico QQ. Como se ve, los puntos se ajustan bien a la linea, lo que nos indica normalidad.

```
qqnorm(reg1$residuals, ylab = "Residuals") ## qq plot
qqline(reg1$residuals) # linea normal teòrica
```



Además, usamos el test de Shapiro-Wilk para contrastar la hipótesis nula de que los residuos son normales. Obtenemos un p-valor mayor a 0.05, por lo que podemos asumir normalidad de los residuos.

```
# Test shapiro-wilk
shapiro.test(reg1$residuals)
```

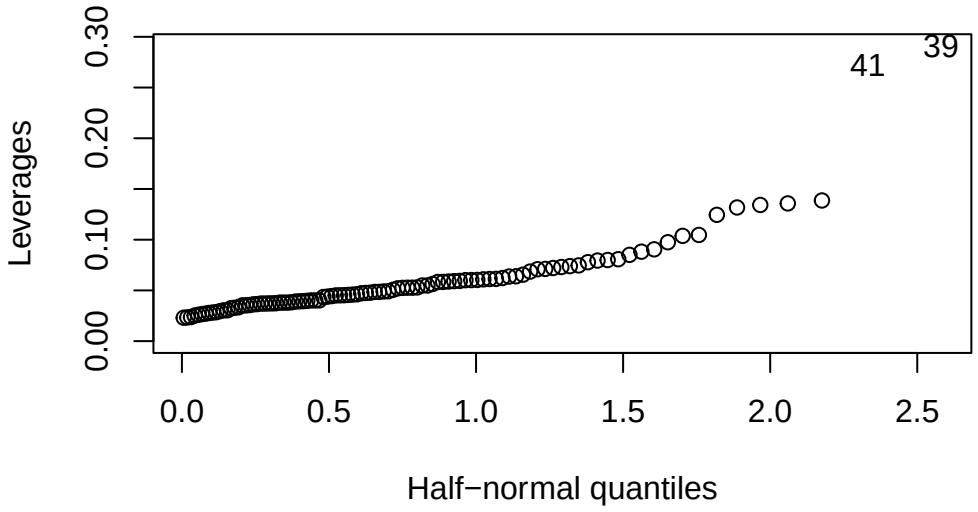
```
    Shapiro-Wilk normality test

data:  reg1$residuals
W = 0.99152, p-value = 0.7805
```

Ahora estudiamos si hay leverage. Para ello, empezamos caluclando estos valores y los representamos, en el cual vemos que las observaciones 39 y 41 presentan los valores mas grandes .

```
leverages <- hatvalues(reg1) # calcul leverages
head(leverages) # mostramos los 5 primeros leverages
```

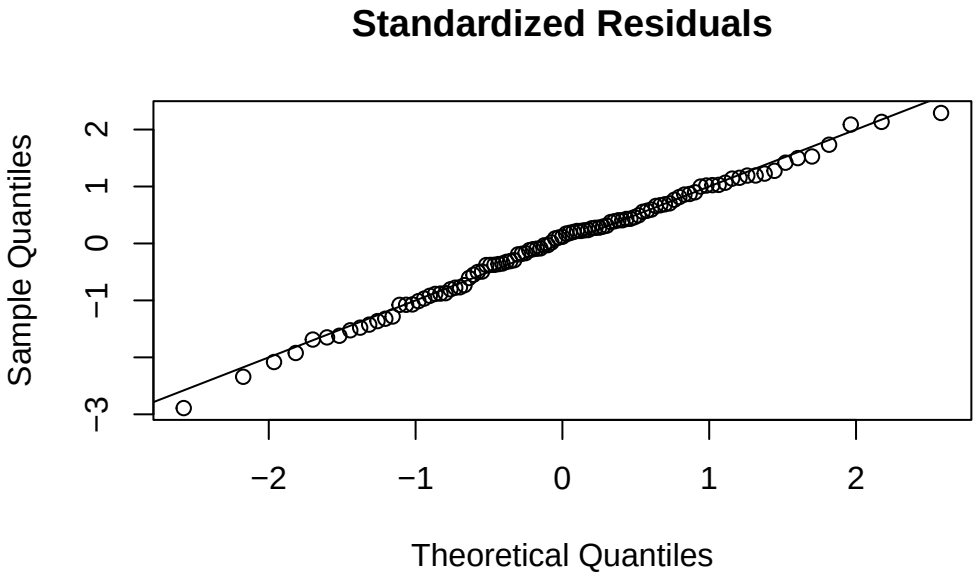
```
         1          2          3          4          5          6
0.03979353 0.03336862 0.06017778 0.04387512 0.03902275 0.03628208
```

```
# Plot de leverages
estats <- row.names(datos)
halfnorm(leverages, labs = estats, ylab = "Leverages")
```



Luego representamos el QQ plot de los residuos estandarizados. En en plot vemos las observaciones 39 y 41 no resultan ser tan extremas, y se ajustan bien a la linea.

```
# QQ plot residus standaritzats
qqnorm(rstandard(reg1), main="Standardized Residuals")
abline(0,1)
```



Para detectar possibles valores atípicos usaremos los residuos studentizados. Estos se calculan para cada observación con un modelo ajustando sin ella, lo que los hace una buena herramienta para detectar outliers. Los calculamos y mostramos los 6 valores mas grandes (en valor absoluto):

```
stud <- rstudent(reg1) # residuos studentitzats
# calculamos el valor absoluto i los ordenamos de mayor a menor
stud.sorted <- sort(abs(stud), decreasing = TRUE)
# mostramos los primeros 6 valores
head(stud.sorted, 6)
```

```
      38       93       31       44       82       22
3.010996 2.398725 2.344644 2.175969 2.126117 2.120040
```

Ahora calculamos el valor crítico de Bonferroni

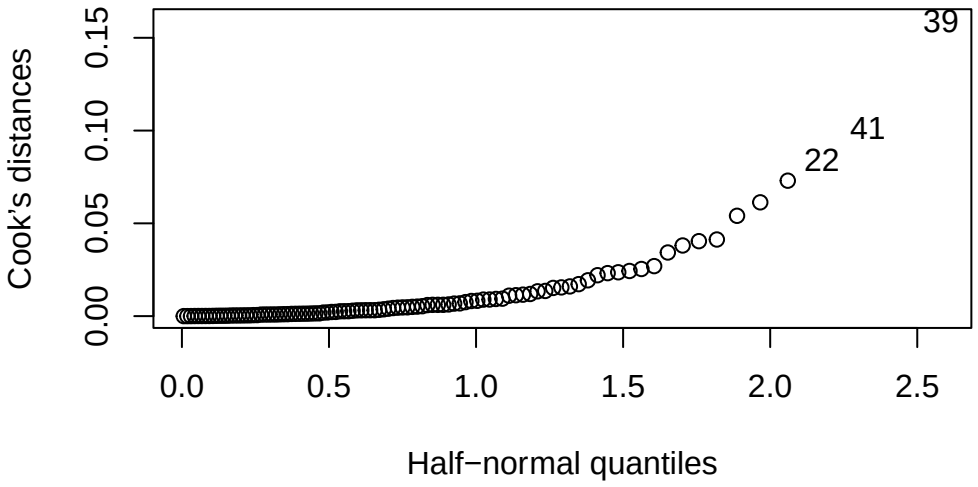
```
n  <- nrow(datos) # observaciones
df <- df.residual(reg1)# grados de libertad
alpha <- 0.05 # significacion--ón
qt(alpha/(n*2), df) # valor crítico bonferroni
```

```
[1] -3.607663
```

Como no hay ningún valor de los residuos studentizados que sea mayor que $|-3.61|$, podemos considerar que no hay outliers en nuestros datos.

Investigamos si hay puntos influyentes con la distancia de Cook.

```
cook <- cooks.distance(reg1) # calculamos la distancia de cook
# representamos los valores en  un grafico
halfnorm(cook, 3, labs = estats, ylab="Cook's distances")
```



Representado los valores en un gráfico, observamos que la observación 39 tiene el valor más alto de este parámetro. Procedemos a ajustar un modelo sin esta observación. Si bien algunos coeficientes han cambiando, todos siguen mostrando una asociación hacia la misma dirección, por lo que no consideramos la observacion 39 como influyente.

```
# ajustem regressió amb el subset
reg2 <- lm(totlngth ~ sex + hdlngth + taill + pop + pop*hdlngth,
           data = subset(datos, rownames(datos) != "39"))
# mostrem el resultat
sumary(reg2)
```

```
                Estimate Std. Error t value  Pr(>|t|)
(Intercept)     -4.402674   6.984147 -0.6304  0.529975
sexm            -1.142459   0.432427 -2.6420  0.009655
hdlngth          0.509074   0.066832  7.6172 2.013e-11
taill            1.183641   0.131556  8.9972 2.492e-14
popVic         -35.028241  15.547163 -2.2530  0.026583
hdlngth:popVic   0.410332   0.166302  2.4674  0.015420

n = 100, p = 6, Residual SE = 2.02118, R-Squared = 0.76
```

A pesar de haber detectado una ligera no-linealidad y no constancia de los errores, la resta de diagnósticos realizados han salido bién. En general, podemos considerar el modelo como fiable.

Ejercio 2

En el ejercicio 7.7 del libro de Afifi et al. (2012) se indica un procedimiento para generar unos datos que se utilizan para practicar algunos de los temas estudiados en esta primera parte de la asignatura. Se trata de obtener 100 casos independientes para cada una de 10 variables con distribución normal estándar (media = 0 y varianza = 1). Llamaremos X1,X2, . . . ,X9,Z a estas variables. Para generar los 100 datos utilizaremos las siguientes semillas:

X_1 `seed 36541`

X_2 `seed 43893`

X_3 `seed 45671`

X_4 `seed 65431`

X_5 `seed 98753`

X_6 `seed 78965`

X_7 `seed 67893`

X_8 `seed 34521`

X_9 `seed 98431`

Z `seed 67895`

Ahora tenemos 10 números independientes, aleatorios y normales para cada uno de los 100 casos. La esperanza o media poblacional es 0 y la desviación estándar poblacional es 1. El siguiente paso es transformar los datos de forma que las variables tengan algún tipo de intercorrelación. Las transformaciones se concretan de modo que algunas variables sean función lineal de otras. La propuesta es:

```
x1 <- 5*x1

x2 <- 3*x2

x3 <- x1 + x2 + 4*x3

x4 <- x4

x5 <- 4*x5

x6 <- x5 - x4 +6*x6

x7 <- 2*x7

x8 <- x7 + 2*x8

x9 <- 4*x9

y <- 5 + x1 + 2*x2 + x3 + 10*z
```

***Nota*: Observemos que esta transformación es secuencial. Es decir, que los datos que intervienen en una transformación particular no son los originales, sino los transformados en los pasos anteriores. Para entender esta cuestión, es mejor utilizar dos conjuntos de letras distintas. Llamaremos Z_1, Z_2, ... ,Z_9, Z a los datos originales con distribución $N(0,1)$. Entonces, las primeras transformaciones son:**

$$\begin{aligned} X_1 &= 5Z_1 \\ X_2 &= 3Z_2 \\ X_3 &= X_1 + X_2 + 4Z_3 \\ &\vdots \end{aligned}$$

El resultado final es una muestra aleatoria de 100 casos para 10 variables $X_1, X_2, ..., X_9, Y$ con distribución normal multivariante. Las medias y varianzas poblacionales se muestran en la tabla 1.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Y
Media	0	0	0	0	0	0	0	0	0	5
Var.	25	9	50	1	16	53	4	8	16	297

Tabla 1: Medias y varianzas poblacionales de las variables.

a) Probar que las medias y varianzas poblacionales de las variables transformadas son las de la tabla 1. Para simplificar nos limitaremos a las variables X1, X2, X3 y Y.

Para empezar, repasamos algunas propiedades de la esperanza (media poblacional), de la varianza y de la covarianza poblacionales de una combinación lineal de variables:

1. **Esperanza (Media Poblacional):**
   - La esperanza de una constante a es a.
   - La esperanza de una variable multiplicada por una constante a es a veces la esperanza de la variable.
   - La ezperanza de la suma de variables aleatorias es la suma de sus esperanzas individuales.
2. **Varianza:**
   - La varianza de una constante a es 0.
   - La varianza de una variable multiplicada por una constante a es a^2 veces la varianza de la variable.
   - La varianza de la suma de variables aleatorias es la suma de sus varianzas individuales más el doble de la suma de sus covarianzas:

$$\text{Var}(X_1 + X_2 + X_3 ...) = \text{Var}(X_1) + \text{Var}(X_2) + ... + 2 \cdot (\text{Cov}(X_1, X_2) + \text{Cov}(X_1, X_3) ...)$$

3. **Covarianza:**
   - La covarianza de una constante a con cualquier variable es 0.
   - La covarianza entre dos variables multiplicadas por una constante a es a veces la covarianza original.
   - La covarianza de dos variables aleatorias independientes es 0.
   - La covarianza de dos variables iguales, és su varianza $\text{Cov}(X, X) = \text{Var}(X)$.
   - La covarianza con una combinación lineal es $\text{Cov}(aX + bY, cZ) = ac\text{Cov}(X, Z) + bc\text{Cov}(Y, Z)$.

Las variables X_1 y X_2 se forman multiplicando una variable aleatoria por las constantes 5 y 3 respectivamente. Entonces, como la esperanza de estas variables se trata de multiplicar por la constante correspondiente las esperanzas de Z_1 y Z_2, las cuales son 0 (distribución $N(0,1)$. Para encontrar la varianza de X_1 y X_2, multplicamos la varianza de Z_1 y Z_2 por 5^2 y 3^2 respectivamente. En resumen,

$$\mathrm{E}(X_1) = 5 \cdot \mathrm{E}(Z_1) = 5 \cdot 0 = 0$$

$$\mathrm{Var}(X_1) = 5^2 \cdot \mathrm{Var}(Z_1) = 5^2 \cdot 1 = 25$$

$$\mathrm{E}(X_2) = 3 \cdot \mathrm{E}(Z_2) = 3 \cdot 0 = 0$$

$$\mathrm{Var}(X_2) = 3^2 \cdot \mathrm{Var}(Z_2) = 3^2 \cdot 1 = 9$$

La esperanza de X_3 y Y es la suma de esperanzas individuales, multiplicadas por la constante que acompaña a cada una de las variables en la combinación lineal. Su varianza, es la suma de las varianzas de las variables de la combinación lineal multiplicadas por sus respectivos coeficientes al cuadrado, más el doble de la suma de sus covarianzas. Para X_3, las 4 variables aleatorias que la forman son independientes, por lo que sus covarianzas son 0. Para Y, todas las variables de la combinación son independientes excepto por X_1 con X_3 y X_2 con X_3. Estas són:

$$\mathrm{Cov}(X_1, X_3) = \mathrm{Cov}(X_1, X_1) + \mathrm{Cov}(X_1, X_2) + 4\mathrm{Cov}(X_1, Z_3) = \mathrm{Var}(X_1) = 25$$

$$\mathrm{Cov}(X_2, X_3) = \mathrm{Cov}(X_2, Z_1) + \mathrm{Cov}(X_2, X_2) + 4\mathrm{Cov}(X_2, Z_3) = \mathrm{Var}(X_2) = 9$$

$$\mathrm{E}(X_3) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + 4 \cdot \mathrm{E}(Z_3) = 0 + 0 + 0 = 0$$

$$\mathrm{Var}(X_3) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 4^2 \cdot \mathrm{Var}(Z_3) = 25 + 9 + 16 = 50$$

$$\mathrm{E}(y) = \mathrm{E}(5) + \mathrm{E}(X_1) + 2 \cdot \mathrm{E}(X_2) + \mathrm{E}(X_3) + 10 \cdot \mathrm{E}(Z) = 5 + 0 + 0 + 0 + 0 = 0$$

$$\mathrm{Var}(X_3) = \mathrm{Var}(5) + \mathrm{Var}(X_1) + 2^2 \cdot \mathrm{Var}(X_2) + \mathrm{Var}(X_3) + 10^2 \cdot \mathrm{Var}(Z) + 2 \cdot (\mathrm{Cov}(X_1, X_3) + 2\mathrm{Cov}(X_2, X_3)) =$$
$$= 0 + 25 + 4 \cdot 9 + 50 + 100 \cdot 1 + 2(25 + 9 \cdot 2) = 297$$

Calcular las correlaciones poblacionales entre las variables $X_1, X_2, ..., X_9$, Y y comprobar que son las que tenemos en la tabla 2. Para simplificar nos limitaremos a las variables X_1, X_2 y X_3.

La correlación entre dos variables aleatorias X y Y es

$$\rho_{X,Y} = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Con la expresión ya deducimos que la correlación de una variable son si misma és 1:

$$\rho_{X,X} = \mathrm{Corr}(X, X) = \frac{\mathrm{Cov}(X, X)}{\sigma_X \sigma_X} = \frac{\mathrm{Var}(X)}{\mathrm{Var}(X)} = 1$$

Entonces, queda comprovado que $\rho_{X_1,X_1} = \rho_{X_2,X_2} = \rho_{X_3,X_3} = 1$ Calculamos ahora ρ_{X_1,X_2}, ρ_{X_1,X_3}, ρ_{X_2,X_3}, ya que $\rho_{X,Y} = \rho_{Y,X}$.

$$\rho_{X_1,X_2} = \mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_{X_1} \sigma_{X_2}} = \frac{0}{\sigma_{X_1} \sigma_{X_2}} = 0$$

$$\rho_{X_1,X_3} = \text{Corr}(X_1, X_3) = \frac{\text{Cov}(X_1, X_3)}{\sigma_{X_1}\sigma_{X_3}} = \frac{25}{\sqrt{\text{Var}(X_1)\text{Var}(X_3)}} = \frac{25}{\sqrt{25 \cdot 50}} = 0.707$$

$$\rho_{X_2,X_3} = \text{Corr}(X_2, X_3) = \frac{\text{Cov}(X_2, X_3)}{\sigma_{X_2}\sigma_{X_3}} = \frac{9}{\sqrt{\text{Var}(X_2)\text{Var}(X_3)}} = \frac{9}{\sqrt{9 \cdot 50}} = 0.424$$

c) Calcular el coeficiente poblacional de correlación múltiple al cuadrado entre Y y las variables X_1 y X_9 únicamente. Su valor es 0.34.

Calcular también el mismo coeficiente entre Y y las variables X_1, X_2 y X_3. Su valor es 0.66.

Para calcular el coeficiente poblacional de correlación múltiple al cuadrado (R^2) usamos la formula

$$R^2 = \mathbf{c}^\top R_{xx}^{-1}\mathbf{c}$$

donde $\mathbf{c}$ es el vector de los coeficientes de regressión, y R_{xx} esla matriz de covariancias de las variables independientes.

Para el primer caso, con los valores de la tabla 2 obtenemos el vector

$$\mathbf{c} = (0.580, 0)$$

y la matriz

$$R_{xx} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Hacemos el calculo en R:

```r
R <- matrix(c(1,0,0,1), nrow = 2)
c <- c(0.59, 0)

# coeficiente poblacional de correlación múltiple al cuadrado
t(c) %*% solve(R) %*% c
```

```
       [,1]
[1,] 0.3481
```

En el segundo apartado, el vector es

$$\mathbf{c} = (0.580, 0.522, 0.763)$$

y la matriz

$$R_{xx} = \begin{pmatrix} 1 & 0 & 0.707 \\ 0 & 1 & 0.424 \\ 0.707 & 0.424 & 1 \end{pmatrix}.$$

Hacemos el calculo en R:

```r
R <- matrix(c(1, 0, 0.707, 0, 1, 0.424, 0.707, 0.424, 1), nrow = 3)
c <- c(0.59, 0.522, 0.763)

# coeficiente poblacional de correlación múltiple al cuadrado
t(c) %*% solve(R) %*% c
```

```
          [,1]
[1,] 0.6689982
```

d) Generar los datos según el procedimiento de Afifi descrito al principio de este ejercicio.

Generamos los 100 casos independientes para las 10 variables aleatorias con las semillas establecidas en el enunciado:

```r
# Fijar las semillas y generar 100 casos para cada variable:
set.seed(36541)  # Semilla para X1
Z1 <- rnorm(100)

set.seed(43893)  # Semilla para X2
Z2 <- rnorm(100)

set.seed(45671)  # Semilla para X3
Z3 <- rnorm(100)

set.seed(65431)  # Semilla para X4
Z4 <- rnorm(100)

set.seed(98753)  # Semilla para X5
Z5 <- rnorm(100)

set.seed(78965)  # Semilla para X6
Z6 <- rnorm(100)

set.seed(67893)  # Semilla para X7
Z7 <- rnorm(100)

set.seed(34521)  # Semilla para X8
Z8 <- rnorm(100)

set.seed(98431)  # Semilla para X9
Z9 <- rnorm(100)

set.seed(67895)  # Semilla para Z
Z <- rnorm(100)
```

Creamos las variables aleatorias en R:

```r
# Calculamos todas la variables:
x1 <- 5*Z1
x2 <- 3*Z2
```

```r
x3 <- x1 + x2 + 4*Z3
x4 <- Z4
x5 <- 4*Z5
x6 <- x5 - x4 + 6*Z6
x7 <- 2*Z7
x8 <- x7 + 2*Z8
x9 <- 4*Z9
y <- 5 + x1 + 2*x2 + x3 + 10*Z
```

Calcular el modelo de regresión con los datos y como respuesta y los datos `x1,...,x9` como regresoras. ¿Cual es el coeficiente de determinación de este modelo? Comparar a simple vista los coeficientes estimados de los parámetros con los valores teóricos.

El coeficiente de determinación de este modelo es 0.77 (el valor R-squared del summary). A simple vista vemos que el intercept tiene un valor de 5.27, muy cerca del valor teórico, 5. Además, es estadísticamente significativo (p-valor menor a 0.05). Las variables `x1` y `x3` también son significativas y su coeficiente es 1.15 y 1.18 respectivamente, cerca del 1 teórico. Sin embargo, la variable `x2`, aun que también muestra significación estadística, su coeficiente esta más lejos del teórico (1.14 vs. 2). En cuanto a `x4`, esta también muestra significación al nivel del 5% y su coeficiente es del -2.2, muy lejos del 0 teòrico. Para las demàs, ninguna de ellas muestra un efecto (p-valores > 0.05) lo que encaja ya que su coeficiente teòrico es 0.

```r
# Creamos un dataframe con los datos:
datos <- tibble(x1 = x1, x2 = x2, x3 = x3, x4 = x4, x5 = x5,
                x6 = x6, x7 = x7, x8 = x8, x9 = x9, y = y)

# ajustamos regresión
reg <- lm(y ~ ., data = datos)

summary(reg)
```

```
Call:
lm(formula = y ~ ., data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-16.3001  -4.8237   0.2004   5.2356  14.9068

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.269015   0.787334   6.692 1.82e-09 ***
x1           1.151818   0.246577   4.671 1.04e-05 ***
x2           1.144896   0.293303   3.903 0.000183 ***
x3           1.180932   0.189273   6.239 1.41e-08 ***
x4          -2.201566   0.938712  -2.345 0.021210 *
x5          -0.072739   0.241813  -0.301 0.764255
x6          -0.034317   0.130469  -0.263 0.793128
x7          -0.515657   0.549052  -0.939 0.350153
x8          -0.124393   0.376325  -0.331 0.741756
```

```
x9          -0.005784   0.191521  -0.030 0.975973
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.62 on 90 degrees of freedom
Multiple R-squared:  0.7747,    Adjusted R-squared:  0.7522
F-statistic: 34.39 on 9 and 90 DF,  p-value: < 2.2e-16
```

Si repetimos el procedimiento, pero ahora generamos 10,000 casos, obtenemos coeficientes mucho más similares a los teóricos, y los resultados del T-test encajan con lo que sabemos. Esto se debe a que 100 observaciones son pocas para ajustar el modelo con presición.

```r
set.seed(36541)  # Semilla para X1
Z1_10000 <- rnorm(1000)

set.seed(43893)  # Semilla para X2
Z2_10000 <- rnorm(1000)

set.seed(45671)  # Semilla para X3
Z3_10000 <- rnorm(1000)

set.seed(65431)  # Semilla para X4
Z4_10000 <- rnorm(1000)

set.seed(98753)  # Semilla para X5
Z5_10000 <- rnorm(1000)

set.seed(78965)  # Semilla para X6
Z6_10000 <- rnorm(1000)

set.seed(67893)  # Semilla para X7
Z7_10000 <- rnorm(1000)

set.seed(34521)  # Semilla para X8
Z8_10000 <- rnorm(1000)

set.seed(98431)  # Semilla para X9
Z9_10000 <- rnorm(1000)

set.seed(67895)  # Semilla para Z
Z_10000 <- rnorm(1000)

# Calculamos todas la variables:
x1_10000 <- 5*Z1_10000
x2_10000 <- 3*Z2_10000
x3_10000 <- x1_10000 + x2_10000 + 4*Z3_10000
x4_10000 <- Z4_10000
x5_10000 <- 4*Z5_10000
x6_10000 <- x5_10000 - x4_10000 + 6*Z6_10000
x7_10000 <- 2*Z7_10000
x8_10000 <- x7_10000 + 2*Z8_10000
```

```r
x9_10000 <- 4*Z9_10000
y_10000 <- 5 + x1_10000 + 2*x2_10000 + x3_10000 + 10*Z_10000

# Creamos un dataframe con los datos:
datos_10000 <- tibble(x1 = x1_10000, x2 = x2_10000, x3 = x3_10000, x4 = x4_10000,
                      x5 = x5_10000, x6 = x6_10000, x7 = x7_10000, x8 = x8_10000,
                      x9 = x9_10000, y = y_10000)

# ajustamos regresión
reg_10000 <- lm(y ~ ., data = datos_10000)

summary(reg_10000)
```

```
Call:
lm(formula = y ~ ., data = datos_10000)

Residuals:
    Min      1Q  Median      3Q     Max
-25.437  -6.691  -0.249   6.427  36.425

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.30386    0.30429  17.430   <2e-16 ***
x1           0.96941    0.09549  10.152   <2e-16 ***
x2           2.03449    0.12852  15.830   <2e-16 ***
x3           1.04035    0.07453  13.959   <2e-16 ***
x4           0.33040    0.29792   1.109   0.2677
x5          -0.07433    0.09151  -0.812   0.4168
x6          -0.02159    0.05125  -0.421   0.6737
x7          -0.27967    0.21300  -1.313   0.1895
x8           0.23288    0.15015   1.551   0.1212
x9           0.14967    0.07814   1.915   0.0557 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.579 on 990 degrees of freedom
Multiple R-squared:  0.6827,    Adjusted R-squared:  0.6798
F-statistic: 236.7 on 9 and 990 DF,  p-value: < 2.2e-16
```

e) Con el modelo del apartado anterior, contrastar las siguientes hipótesis:

i) $H_0 : \beta_0 = 5$

Creamos un modelo donde la variable respuesta sea `y-5`. El resultado de la regresivo nos dará el resultado del T-test para la hipotiposis nula de que el intercept sea 0, que equivale a $H_0 : \beta_0 = 5$ cuando la variable respuesta es `y-5`. Obtenemos un p-valor mayor a 0.05, por lo que no podemos descartar la hipótesis nula. Es decir, justificamos $\beta_0 = 5$.

```r
summary(lm(I(y-5) ~ ., data = datos))
```

```
Call:
lm(formula = I(y - 5) ~ ., data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-16.3001  -4.8237   0.2004   5.2356  14.9068

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.269015   0.787334   0.342 0.733389
x1           1.151818   0.246577   4.671 1.04e-05 ***
x2           1.144896   0.293303   3.903 0.000183 ***
x3           1.180932   0.189273   6.239 1.41e-08 ***
x4          -2.201566   0.938712  -2.345 0.021210 *
x5          -0.072739   0.241813  -0.301 0.764255
x6          -0.034317   0.130469  -0.263 0.793128
x7          -0.515657   0.549052  -0.939 0.350153
x8          -0.124393   0.376325  -0.331 0.741756
x9          -0.005784   0.191521  -0.030 0.975973
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.62 on 90 degrees of freedom
Multiple R-squared:  0.7747,    Adjusted R-squared:  0.7522
F-statistic: 34.39 on 9 and 90 DF,  p-value: < 2.2e-16
```

ii) $H_0 : \beta_2 = 2$

Hacemos el T-test, $t = \frac{\hat{\beta}_0 - \beta_{H_0}}{se(\hat{\beta}_0)}$. y luego estimamos el p valor. Obtenemos significacia estadística (p-valor < 0.05) por lo que descartamos la hipótesis nula.

```r
# valor T
t.valor <- (reg$coefficients["x2"]-2)/summary(reg)$coef["x2", 2]
# Grados de libertad
df <- reg$df.residual
#p-valor:
2*pt(t.valor, df, lower.tail = TRUE)
```

```
        x2
0.004482339
```

Comprobamos lo mismo con las variables que contienen 10,000 casos. Esta vez, si que obtenemos un p-valor > 0.05 por lo que en este caso si que aceptaríamos la hipótesis nula.

```r
# valor T
t.valor <- (reg_10000$coefficients["x2"]-2)/summary(reg_10000)$coef["x2", 2]
# Grados de libertad
df <- reg_10000$df.residual
#p-valor:
```

```r
2*pt(t.valor, df, lower.tail = TRUE)
```

```
      x2
1.211543
```

iii) $H_0 : \beta_1 = \beta_3$

Para comprobar esta hipótesis, ajustamos un modelo donde agrupamos estas dos variables sacando factor común de su coeficiente (ya que la hipótesis nula es que es el mismo). Contrastamos este modelo con el generado en el aparado d). Obtenemos un p-valor superior a 0.05, por lo que está justificado no descartar la hipótesis nula.

```r
lm_iii <- lm(y ~ I(x1 + x3) + x2 + x4 +x5 + x6 + x7 + x8 + x9)

anova(lm_iii, reg)
```

```
Analysis of Variance Table

Model 1: y ~ I(x1 + x3) + x2 + x4 + x5 + x6 + x7 + x8 + x9
Model 2: y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     91 5225.9
2     90 5225.6  1   0.29784 0.0051 0.9431
```

iv) $H_0 : \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$

Ahora, ajustamos un modelo donde no usamos las variables `x4`, `x5`, `x6`, `x7`, `x8`, `x9` y lo contrastamos con el modelo en d). Obtenemos un p valor mayor a 0.05, por lo que no se descarta la hipótesis nula y se justifica la utilización del modelo reducido.

```r
lm_iv <- lm(y ~ I(x1 + x3) + x2 + x4)
anova(lm_iii, reg)
```

```
Analysis of Variance Table

Model 1: y ~ I(x1 + x3) + x2 + x4 + x5 + x6 + x7 + x8 + x9
Model 2: y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     91 5225.9
2     90 5225.6  1   0.29784 0.0051 0.9431
```

f) Con los datos generados, calcular la matriz de correlaciones muestrales dos a dos y contrastar cuando se puede aceptar que son cero. Utilizar la corrección por comparaciones múltiples de Benjamini & Hochberg (1995) (también conocida como false discovery rate) y comentar el resultado.

Hazemos un test de correlacion dos a dos entre todas las variables y luejo ajustamos el p.valor corrigiendo por Benjamini & Hochberg (BH). El resultado nos indica que `x3` esta correlacionada con

x1 y x2, también podemos aceptar correlación entre x5 y x6, entre x7 y x8, y entre y con x1, x2, y x3. Sin embargo, la correlación entre x4 y x6 no la hemos detectado. Los resultados del ajuste BH son iguales que los obtenidos sin este mètodo.

```
# matriz de correlación
cor.matrix <- cor(datos)

# data frame con los resultados de las comparaciones
resultados <- tibble(var1 = NA, var2 = NA, p.value = NA)

tt <- 0 # contador bucle 1
for (jj in names(datos)[1:9]) {
  tt <- tt + 1
  for (ii in names(datos)[(tt+1):10]) {

    test.result <- cor.test(datos[[ii]], datos[[jj]])
    resultados <- resultados %>%
      rbind(tibble(var1 = jj, var2 = ii, p.value = test.result$p.value))
  }
}

# mostramos los resultados que serian significaticos:
print(resultados %>% filter(p.value <= 0.05))
```

```
# A tibble: 7 x 3
  var1  var2   p.value
  <chr> <chr>    <dbl>
1 x1    x3    2.24e-14
2 x1    y     3.88e-14
3 x2    x3    4.07e- 4
4 x2    y     2.11e- 5
5 x3    y     2.99e-26
6 x5    x6    2.31e- 7
7 x7    x8    2.08e-16
```

```
# Aplicamos ahora la correcion de  Benjamini & Hochberg:
resultados <- resultados %>%
  mutate(p.adjusted = p.adjust(p.value, method = "BH"))

# mostramos los resultados que serian significaticos con este ajuste:
print(resultados %>% filter(p.adjusted <= 0.05))
```

```
# A tibble: 7 x 4
  var1  var2   p.value p.adjusted
  <chr> <chr>    <dbl>      <dbl>
1 x1    x3    2.24e-14   3.37e-13
2 x1    y     3.88e-14   4.36e-13
3 x2    x3    4.07e- 4   2.62e- 3
4 x2    y     2.11e- 5   1.59e- 4
5 x3    y     2.99e-26   1.35e-24
```

```
6 x5    x6    2.31e- 7   2.08e- 6
7 x7    x8    2.08e-16   4.68e-15
```