Trabajo 2 - Análisis de Regresión

Fabián Ramírez y Fabián Castellano

Ejercicio F

```
# Incluyo la libreria de los datos
library('carData')
library('car')
library('alr3')
# Una función  útil
library('model')
# Se instala con el siguiente codigo
# if (!require('devtools')) install.packages('devtools')
# devtools::install_github('fhernanb/model', force=TRUE)
library('miniUI')
library('webshot')
library('manipulateWidget')
```

```
#Escribo los datos que me solicitan
X_1 = c(17, 19, 19, 20, 20, 20, 20, 21, 21, 21 ,21, 25, 25,27, 28, 30, 30)
X_2 = c(42, 45, 45, 29, 29, 29, 29, 93, 93, 93, 93, 34,34, 98, 9, 73, 73)
Y   = c(90, 71, 76, 63, 63, 80, 80, 80, 64, 82, 66, 75,82, 99, 73, 67, 74)
```

Problema 1

Asumamos que el modelo que genera los datos es de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

Con U siguiendo la ley normal multivariada. Ajustaremos el modelo:

```
reg1 = lm(Y~X_1+X_2)
summary(reg1)
```

```
Call:
lm(formula = Y ~ X_1 + X_2)
Residuals:
     Min       1Q   Median       3Q      Max
-13.3764  -9.6597   0.3827   5.7693  21.2085
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.2462    14.9067   4.847 0.000259 ***
```

```
X_1          0.0286     0.6455   0.044 0.965279
X_2          0.0487     0.0876   0.556 0.586984
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 10.38 on 14 degrees of freedom
Multiple R-squared:  0.02267,Adjusted R-squared:  -0.1169
F-statistic: 0.1624 on 2 and 14 DF,  p-value: 0.8517
```

Por tanto el modelo ajustado viene dado por:

$$\widehat{Y} = 72{,}2462 + 0{,}0286 X_1 + 0{,}0487 X_2$$

Problema 2

Sabemos que:

$$\texttt{Residual Error} = \texttt{Lack of fit error} + \texttt{Pure Error}$$

Y el test que queremos realizar es:

$$\begin{cases} H_0 : \text{No hay falta de ajuste en el modelo} \\ v/s \\ H_1 : \text{Hay falta de ajuste en el modelo} \end{cases}$$

El cual se puede realizar mediante un test F de la forma:

$$F^* = \frac{\texttt{Mean Square due to Lack of fit}}{\texttt{Mean Square due to Pure error}}$$

```
pureErrorAnova(reg1)
```

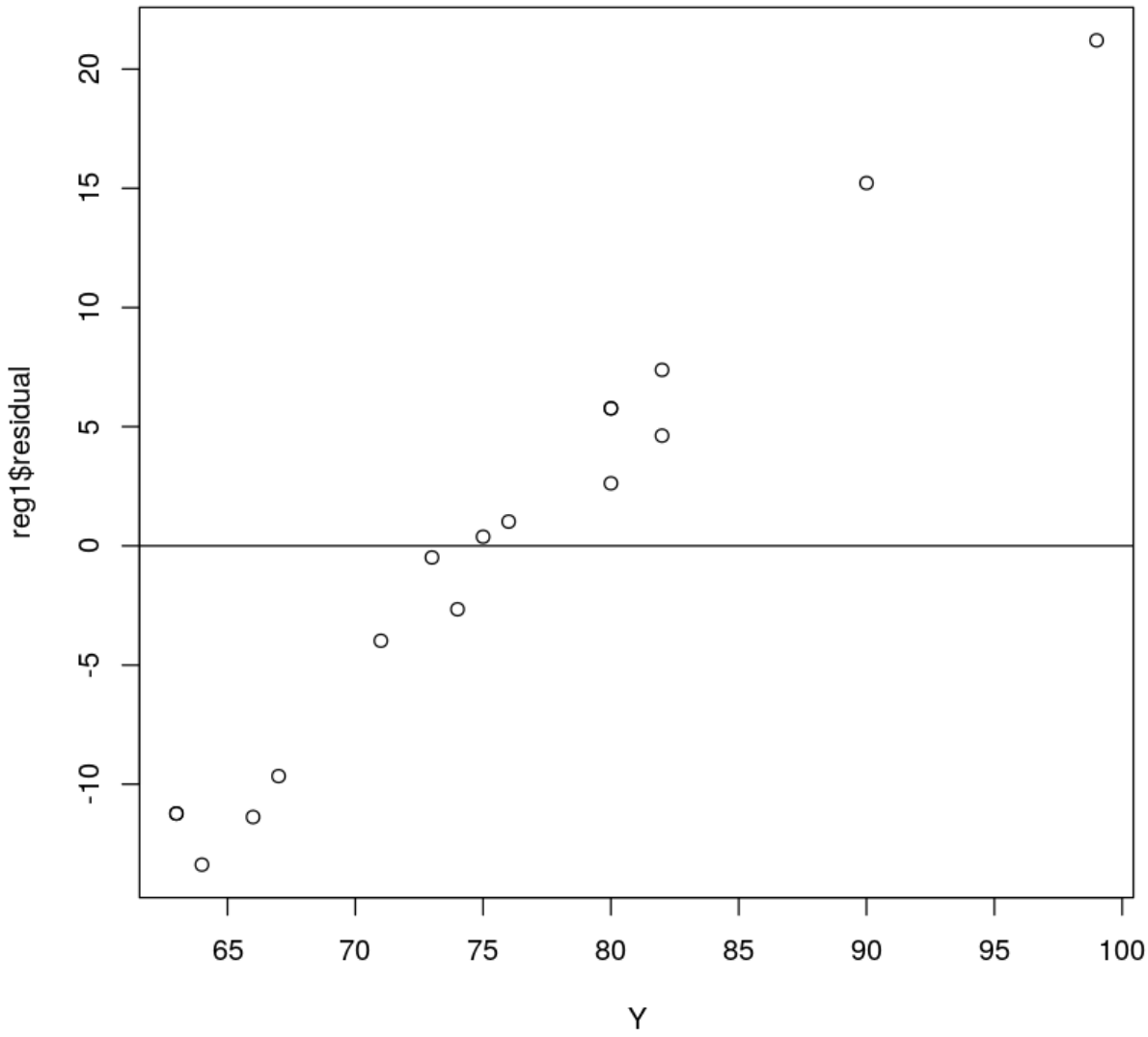
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X_1	1	1.688471	1.688471	0.02489147	0.8781208
X_2	1	33.323438	33.323438	0.49125462	0.5010860
Residuals	14	1509.105738	107.793267	NA	NA
Lack of fit	5	898.605738	179.721148	2.64945181	0.0967553
Pure Error	9	610.500000	67.833333	NA	NA

Entonces $F^* = 2{,}64945181$ y valor-p igual a 0,0967553. Por tanto la significancia mínima para rechazar el hecho de que no hay falta de ajuste en el modelo es de $0{,}096 = 9{,}6\,\%$, en otras palabras: * Si $F^* > F_\alpha$, debería buscar un modelo alternativo. * Si $F^* < F_\alpha$, no es necesario buscar un modelo mas complicado.
Por tanto en nuestro caso si pensamos en una significancia del 5 % podríamos pensar que el modelo no aparenta tener una falta de ajuste.

Problema 3

Realizaremos un plot de los residuales para poder sacar alguna conclusión:

```
plot(Y,reg1$residual)
abline(h=0)
```



Notemos que idealmente los residuos deberian estar alrededor del 0, por tanto para nuestro modelo tenemos que hay mucha discrepancia entre los valores reales y los valores ajustados, incluso tienen una tendencia lineal.

Problema 4

Notemos que el R^2 viene es 0,02267 por tanto X_1 y X_2 explican el 2,267 % de la variabilidad de Y, lo cual es demasiado bajo.

Ejercicio K

Problema 1

Definamos las variables.

```
X_1 = c(-1,-1,0,1,1)
X_2 = c(-1,0,0,0,1)
Y = c(7.2,8.1,9.8,12.3,12.9)
```

Notemos que:

```
X = cbind(rep(1,5),X_1,X_2)
print(t(X)%*%X)
print(t(X)%*%Y)
```

```
      X_1 X_2
    5   0   0
X_1 0   4   2
X_2 0   2   2
    [,1]
    50.3
X_1  9.9
X_2  5.7
```

Por tanto la ecuación solicitada viene dada por:

$$\begin{pmatrix} 5 & 0 & 0 \\ 0 & 4 & 2 \\ 0 & 2 & 2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 50{,}3 \\ 9{,}9 \\ 5{,}7 \end{pmatrix}$$

Problema 2

Recordemos que $b = (X^T X)^{-1} X^T Y$ entonces:

```
b = solve(t(X)%*%X)%*%t(X)%*%Y
print(b)
```

```
     [,1]
    10.06
X_1  2.10
X_2  0.75
```

Por tanto:

$$\begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 10{,}06 \\ 2{,}10 \\ 0{,}75 \end{pmatrix}$$

Problema 3

Recordemos que $SS(b) = \widehat{Y}^T Y = b^T X^T Y$ por tanto:

```
SS = t(b)%*%t(X)%*%Y
print(SS)
```

```
        [,1]
[1,] 531.083
```

Por lo tanto $SS(b) = 531{,}083$

Problema 4

Notemos que $SCE = (Y - Xb)^T (Y - Xb)$ y que $S^2 = \dfrac{1}{n-k-1}$ donde para nuestro caso $n = 5$ y $k = 2$ por tanto se tiene que:

```
SCE = t(Y -X%*%b)%*%(Y -X%*%b)
print(SCE)
S2 =SCE/(5-2-1)
print(S2)
```

```
      [,1]
[1,] 0.107
       [,1]
[1,] 0.0535
```

Por lo tanto la suma de cuadrados del error es 0.107 y S^2 es 0.0535

Problema 5

Calculamos la desviación estándar de cada parámetro b_i, la cual viene dada por:

$$\sqrt{(X^T X)^{-1}_{i+1,i+1} S^2}$$

(ver apunte página 19). Luego:

```
S_b0 = sqrt(  solve(t(X)%*%X)[0+1,0+1]*S2)
print(S_b0)
S_b1 = sqrt(  solve(t(X)%*%X)[1+1,1+1]*S2)
print(S_b1)
```

```
S_b2 = sqrt(  solve(t(X)%*%X)[2+1,2+1]*S2)
print(S_b2)
```

```
          [,1]
[1,] 0.1034408
          [,1]
[1,] 0.1635543
          [,1]
[1,] 0.2313007
```

Por lo tanto tenemos que: * se(b_0) = 0.1034408 * se(b_1) = 0.1635543 * se(b_2) = 0.2313007

Problema 6

Notemos que:

$$\widehat{y_0} = b_0 + 0{,}5b_1 + 0 * b_2$$

Por tanto:

```
y_0 = t(matrix(c(1,0.5,0)))%*%b
print(y_0)
```

```
      [,1]
[1,] 11.11
```

Por tanto $\widehat{y_0} = 11{,}11$.

Problema 7

Calculamos la desviación estandar de $\widehat{y_0}$ la cual viene dada por:

$$\begin{aligned}
\mathbb{V}[\widehat{y_0}] &= \mathbb{V}\left[\begin{pmatrix}1 & 0{,}5 & 0\end{pmatrix} b\right] \\
&= \begin{pmatrix}1 & 0{,}5 & 0\end{pmatrix} \mathbb{V}[b] \begin{pmatrix}1\\0{,}5\\0\end{pmatrix} \\
&= \begin{pmatrix}1 & 0{,}5 & 0\end{pmatrix} (X^T X)^{-1}\sigma^2 \begin{pmatrix}1\\0{,}5\\0\end{pmatrix} \\
&= \sigma^2 \begin{pmatrix}1 & 0{,}5 & 0\end{pmatrix} \begin{pmatrix}0{,}2 & 0{,}0 & 0{,}0\\0{,}0 & 0{,}5 & -0{,}5\\0{,}0 & -0{,}5 & 1{,}0\end{pmatrix} \begin{pmatrix}1\\0{,}5\\0\end{pmatrix} \\
&= \sigma^2 0{,}325
\end{aligned}$$

Luego aproximando σ^2 por su estimador S^2 tenemos que:

```
sqrt(c(1,0.5,0)%*%solve(t(X)%*%X)%*%c(1,0.5,0) *S2)
```

0.1318617

Por lo tanto tenemos que:

$$\begin{aligned} se\left(\widehat{y_0}\right) &= \sqrt{S^2 0{,}325} \\ &= 0{,}1318617 \end{aligned}$$

Problema 8

Si $b_2 = 0$ tenemos que el modelo se convierte en:

$$Y = b_0 + b_1 X_1 + \epsilon$$

Podemos reajustar el modelo obteniendo:

$$\widehat{Y_{aux}} = X_{aux} b_{aux}$$

con

$$X_{aux} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$$

y

$$b_{aux} = \begin{pmatrix} 10{,}060 \\ 2{,}475 \end{pmatrix}$$

Finalmente tenemos que:

$$SS(b_0, b_1) = b_{aux}^T X_{aux}^T Y = 530{,}5205$$

Fialmente:

$$SS(b_2 | b_1, b_0) = SS(b) - SS(b_0, b_1) = 0{,}5625$$

```
reg_aux=lm(Y~X_1)
reg_aux
b_aux = c(10.060,2.475 )
X_aux = cbind(rep(1,length(X_1)),X_1)
SSb_0b_1 = t(b_aux)%*%t(X_aux)%*%Y
print(SSb_0b_1)
SSb_2_dado_b_0b_1 = SS - SSb_0b_1
print(SSb_2_dado_b_0b_1 )
```

```
Call:
lm(formula = Y ~ X_1)
```

```
Coefficients:
(Intercept)          X_1
     10.060        2.475
```

```
         [,1]
[1,] 530.5205
       [,1]
[1,] 0.5625
```

Problema 9

Sea $V = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0{,}25 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$ entonces tenemos que: $V^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$ y nos dicen que el nuevo estimador es $b_V = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$ entonces:

```
inv.V = solve(diag(c(1,1,0.25,1,1)))
b_V = solve(t(X)%*%inv.V%*%X)%*%t(X)%*%inv.V%*%Y
print(b_V)
```

```
      [,1]
   9.9625
X_1 2.1000
X_2 0.7500
```

Por tanto $b_V = \begin{pmatrix} 9{,}9625 \\ 2{,}1 \\ 0{,}75 \end{pmatrix}$ es el nuevo estimador de mínimos cuadrados.

Ejercicio Z

Problema 1

```
#ingresamos los datos
Y = c(22.1,24.5,26.0,26.8,28.2,28.9,30.0,30.4,31.4,21.9,26.1,28.5,30.3,31.5,33.1,22.8,27.
 ↪3,29.8,31.8)
X = c(0,1,2,3,4,5,6,7,8,0,2,4,6,8,10,0,3,6,9)
X2=X^2
```

Realizamos la regresión:

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon$$

```
reg3 = lm(Y ~ X + X2)
```

```
summary(reg3)
```

```
Call:
lm(formula = Y ~ X + X2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66123 -0.28558 -0.05606  0.34252  0.65440

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.56123    0.19843 113.698  < 2e-16 ***
X            1.66802    0.09895  16.857 1.31e-11 ***
X2          -0.06796    0.01031  -6.591 6.21e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3942 on 16 degrees of freedom
Multiple R-squared:  0.9878,Adjusted R-squared:  0.9863
F-statistic: 649.9 on 2 and 16 DF,  p-value: 4.782e-16
```

Por tanto el modelo ajustado viene dado por:

$$\widehat{Y} = 22{,}56123 + 1{,}66802 X - 0{,}06796 X^2$$

Problema 2

Sea $\alpha = 0{,}05$, luego: * Sabemos que si

$$F^* = \frac{\frac{SCR}{k}}{\frac{SCE}{n-k-1}} > F_{1-\alpha}(k, n-k-1)$$

Entonces se rechaza la hipótesis de que el vector de estimadores sea todo nulo (significancia global del modelo). Como $F^* = 649{,}9$, y sabemos que:

$$F_{1-0{,}05}(2, 19-2-1) = 3{,}63372346759163$$

Entonces rechazamos la hipótesis nula y por ende el modelo es significativo.

```
qf(1-0.05,2,length(X)-2-1)
```

3.63372346759163

Problema 3

Análogo al **Ejercicio F, Problema 2** realizaremos el mismo test, por tanto necesitamos el nuevo F^* y los grados de libertad del Lack of fit y Pure Error, por lo tanto:

```
pureErrorAnova(reg3)
print('F(0.95,8,8) =' )
qf(1-0.05,8,8)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
X	1	195.2428882	195.24288822	2073.375804	5.977021e-11
X2	1	6.7515651	6.75156506	71.698036	2.895051e-05
Residuals	16	2.4865994	0.15541246	NA	NA
Lack of fit	8	1.7332660	0.21665825	2.300796	1.299026e-01
Pure Error	8	0.7533333	0.09416667	NA	NA

```
[1] "F(0.95,8,8) ="
```

3.43810123337316

Por lo tanto $F^* = 2{,}30 < F_{1-0{,}05}(8,8) = 3{,}44$ por tanto no es necesario buscar un modelo mas complicado, por tanto el modelo cuadratico es suficiente para explicar y predecir el fenomeno.

Problema 4

Notemos que si ajustamos el modelo:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Tenemos que:

```
reg4 = lm(Y~X)
summary(reg4)
```

```
anova(reg4)
pureErrorAnova(reg4)
```

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4464 -0.4282  0.1809  0.6127  0.9718

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.34638    0.29675   78.67  < 2e-16 ***
X            1.04546    0.05516   18.95 7.18e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7372 on 17 degrees of freedom
Multiple R-squared:  0.9548,Adjusted R-squared:  0.9522
F-statistic: 359.3 on 1 and 17 DF,  p-value: 7.182e-13
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
X	1	195.242888	195.2428882	359.2845	7.181631e-13
Residuals	17	9.238164	0.5434214	NA	NA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
X	1	195.2428882	195.24288822	2073.3758	5.977021e-11
Residuals	17	9.2381644	0.54342144	NA	NA
Lack of fit	9	8.4848311	0.94275901	10.0116	1.757069e-03
Pure Error	8	0.7533333	0.09416667	NA	NA

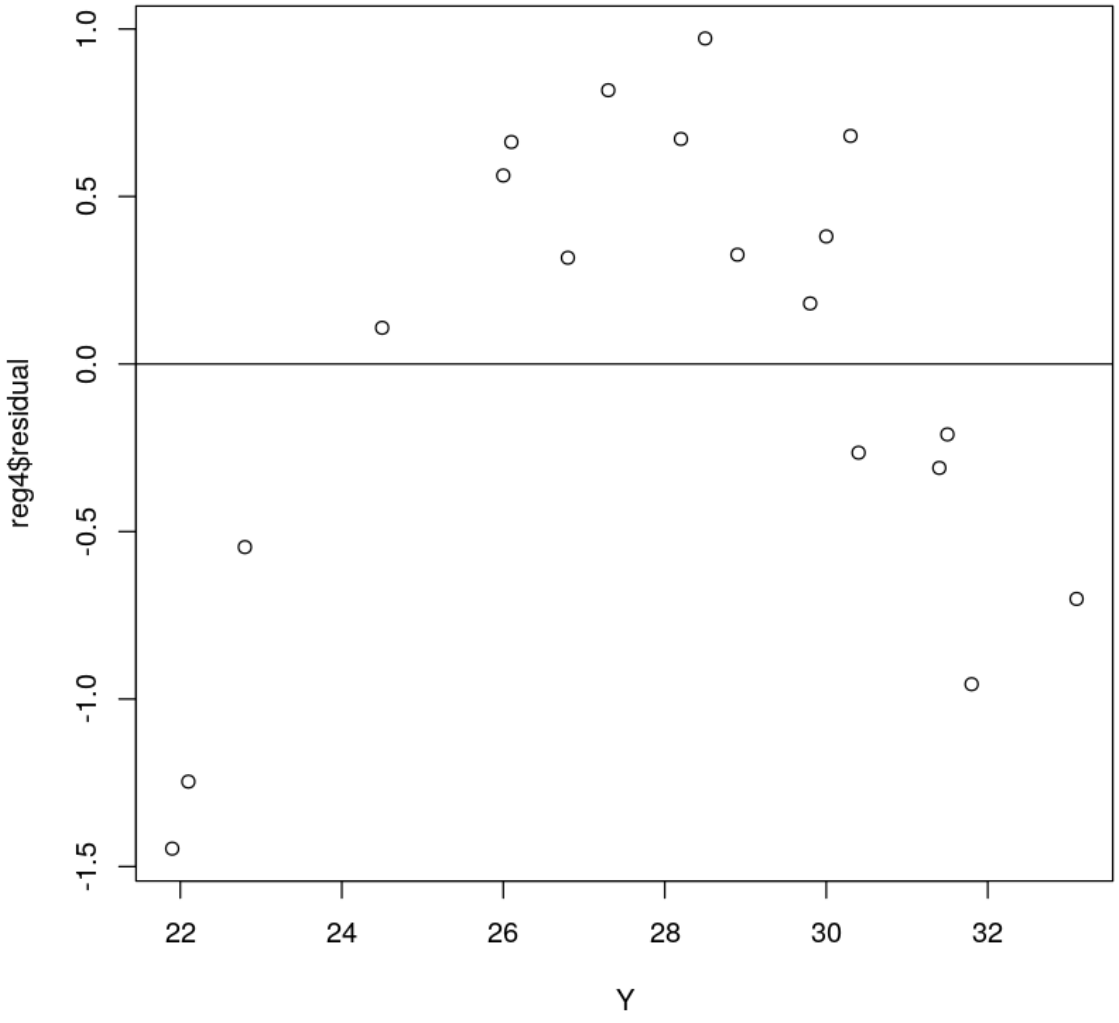
De toda esta información podemos sacar las siguientes conclusiones:

- El R^2 de este modelo (0.9548) es menor que el del modelo cuadrático (0.9878)
- El $R^2 - Ajustado$ de este modelo (0.9522) es menor que el del modelo cuadrático (0.9863) Con estas dos conclusiones tenemos que como disminuye el R^2 y **también el** $R^2 - Ajustado$ ya podríamos deducir que el modelo cuadrático es mejor, ya que el $R^2 - Ajustado$ penaliza la inclusión de nuevas variables al modelo.
- El test-F para la significancia global del modelo concluye rápidamente que el modelo es significativo ya que el

valor-p es muy pequeño (7.182e-13), por tanto la significancia mínima para rechazar que el vector de estimadores es el nulo es muy pequeña (del orden de 10^{-13})

- El test-F para el lack of fit (Ejercicio F, Problema 2) nos dice que tiene un valor-p relativamente pequeño, por tanto rechazo H_0 es decir **debería buscar un modelo alternativo**
- Si vemos el siguiente plot de los residuos podemos ver que el modelo tiene una tendencia cuadrática, lo cual nos dice que deberíamos pensar en un modelo cuadrático

```
plot(Y,reg4$residual)
abline(h=0)
```



Por lo tanto, todas las herramientas que conocemos nos dicen que debemos cambiar el modelo por uno cuadrático.

Problema 5

A modo de conclusión la nube de puntos se puede predecir mediante una función cuadrática dada por:

$$\widehat{Y} = 22{,}56123 + 1{,}66802X - 0{,}06796X^2$$

Y sabemos que el intervalo de confianza para la estimación viene dado por:

$$\left[\left(x^*\right)^T \widehat{\boldsymbol{\beta}} \pm S_e \sqrt{1 + (x^*)^T \left(X^T X\right)^{-1} (x^*)}\, t_{\frac{1+\gamma}{2}}(n-k-1) \right]$$

Donde podemos ir reemplazando para nuestro caso particular tomando:

- $$\left(x^*\right)^T \hat{\beta} = \begin{pmatrix} 1 \\ X^* \\ (X^*)^2 \end{pmatrix}^T \begin{pmatrix} 22{,}56123 \\ 1{,}66802 \\ -0{,}06796 \end{pmatrix} = 22{,}56123 + 1{,}66802X - 0{,}06796X^2$$

- $$S_e \sqrt{1 + (x^*)^T \left(X^T X\right)^{-1} (x^*)}\, t_{\frac{1+\gamma}{2}}(n-k-1) =$$
$$= 0{,}3942 \sqrt{1 + \begin{pmatrix} 1 & X^* & (X^*)^2 \end{pmatrix} \begin{pmatrix} 0{,}253355906 & -0{,}097147349 & 0{,}0079029365 \\ -0{,}097147349 & 0{,}063003809 & -0{,}0062664006 \\ 0{,}007902936 & -0{,}006266401 & 0{,}000684039 \end{pmatrix} \begin{pmatrix} 1 \\ X^* \\ (X^*)^2 \end{pmatrix}}\, t_{\frac{1+\gamma}{2}}(16)$$

Y además no se necesita un modelo mas complicado.

Bibliografía

[1] 'Applied Linear Regression', Sanford Weisberg, Wiley Interscience.

[2] 'The Elements of Statistical Learning', Hastie-Tibshirani-Friedman.

[3] 'Norman R. Draper, Harry Smith - Applied Regression Analysis, Third Edition (Wiley Series in Probability and Statistics) (1998)'