Test

Daniel Carrasco

Proyecto Final Métodos y Herramientas de la investigación 1

Análisis estadístico poblacional sobre el efecto del gasto en publicidad sobre las ventas utilizando los datos de la Tabla 1 (Ventas-Publicidad.xlsx), que muestra el nivel de ventas de una población de 90 empresas condicional al gasto en publicidad.

Determine el valor esperado de las ventas y compare con el valor esperado de esta variable condicionada al gasto en publicidad

LLamando paqueteria

```
# install.packages("readxl")
library(readxl)

# install.packages("xlsx")
library(xlsx)

# install.packages("sampling")
library(sampling)

# install.packages("ggplot2")
library(ggplot2)
```

Definimos el directorio en cual se guardará el archivo.

```
setwd ("C:/Users/Daniel/Desktop")
getwd()
```

```
[1] "C:/Users/Daniel/Desktop"
```

Cargamos los datos que muestran los gastos en publicidad versus las ventas en millones.

```
# Insertamos la tabla original
VENTAS_PUBLICIDAD_MOD1 <- read_excel("Ventas-Publicidad.xlsx",sheet="Tabla_Mod_1")
VENTAS_PUBLICIDAD_MOD1
```

```
# A tibble: 9 x 10
  `10 Mill` `11 Mill` `12 Mill` `13 Mill` `14 Mill` `15 Mill` `16 Mill`
      <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
1     16000     18260     15000     15000     20000     20000     21912
2     32868     36520     40000     40000     50000     54780     60000
3     50000     54780     58000     60000     73040     80000     89000
4     50000     82170     90000     90000    100000    100500    120000
5    100000    109560    120000    120000    140000    160000    200000
6    180000    170000    182600    188973    219120    257880    300000
7    219120    273900    280000    328680    365200    400000    500000
8    300000    365200    380000    434120    500000    550000    650000
9    547800    730400    913000    821700   1064558   1460800   1500000
# i 3 more variables: `17 Mill` <dbl>, `18 Mill` <dbl>, `19 Mill` <dbl>
```

Observamos que la tabla VENTAS_PUBLICIDAD_MOD1 nos entrega columnas de 10 a 19 millones, suponemos que es la inversión en publicidad, en las cuales se listan observaciones de números de ventas para diferentes empresas.

Cargo la segunda tabla modificada y la formateo como dejo en data frame

```
VENTAS_PUBLICIDAD_MOD2 <- read_excel("Ventas-Publicidad.xlsx",sheet="Tabla_Mod_2")
VENTAS_PUBLICIDAD_MOD2 <- as.data.frame(VENTAS_PUBLICIDAD_MOD2)
VENTAS_PUBLICIDAD_MOD2
```

```
   Publicidad_Mill Ventas_Mill
1               10       16000
2               10       32868
3               10       50000
4               10       50000
5               10      100000
6               10      180000
7               10      219120
8               10      300000
9               10      547800
10              11       18260
11              11       36520
```

12	11	54780
13	11	82170
14	11	109560
15	11	170000
16	11	273900
17	11	365200
18	11	730400
19	12	15000
20	12	40000
21	12	58000
22	12	90000
23	12	120000
24	12	182600
25	12	280000
26	12	380000
27	12	913000
28	13	15000
29	13	40000
30	13	60000
31	13	90000
32	13	120000
33	13	188973
34	13	328680
35	13	434120
36	13	821700
37	14	20000
38	14	50000
39	14	73040
40	14	100000
41	14	140000
42	14	219120
43	14	365200
44	14	500000
45	14	1064558
46	15	20000
47	15	54780
48	15	80000
49	15	100500
50	15	160000
51	15	257880
52	15	400000
53	15	550000
54	15	1460800

55	16	21912
56	16	60000
57	16	89000
58	16	120000
59	16	200000
60	16	300000
61	16	500000
62	16	650000
63	16	1500000
64	17	35000
65	17	73040
66	17	100000
67	17	140000
68	17	230000
69	17	400000
70	17	600000
71	17	883085
72	17	1826000
73	18	40000
74	18	90000
75	18	105000
76	18	180000
77	18	280000
78	18	434686
79	18	730400
80	18	1000000
81	18	2487041
82	19	60000
83	19	120000
84	19	165784
85	19	250000
86	19	365200
87	19	600000
88	19	1095600
89	19	1643400
90	19	4000000

Buen se observa que es la misma información, pero ordenada de una forma diferente, en 2 columnas.

consulto las clases de ambas tablas cargadas

```
class(VENTAS_PUBLICIDAD_MOD1)
```

```
[1] "tbl_df"     "tbl"        "data.frame"
```

```
class(VENTAS_PUBLICIDAD_MOD2)
```

```
[1] "data.frame"
```

Se nos solicita determinar el valor esperado de las ventas y comparar con el valor esperado de esta variable condicionada al gasto en publicidad

Valor esperado de ventas considerando el gasto en publicidad

```
sapply(VENTAS_PUBLICIDAD_MOD1,mean)
```

```
 10 Mill  11 Mill  12 Mill  13 Mill  14 Mill  15 Mill  16 Mill  17 Mill
166198.7 204532.2 230955.6 233163.7 281324.2 342662.2 382323.6 476347.2
 18 Mill  19 Mill
594125.2 922220.4
```

El comando "sapply" nos permitio determinar las medias de las observaciones de cada columna, esto quiere decir que podemos decir por ejemplo: "Que el valor o la cantidad esperada de ventas si invertimoes 13 millones en publicidad es de 230955"

```
sapply(VENTAS_PUBLICIDAD_MOD2,mean)
```

```
Publicidad_Mill     Ventas_Mill
           14.5        383385.3
```

Al aplicar el "sapply" en la segunda tabla podemos obtener la media de la inversión en publicidad y la media de las ventas en general.

Podemos obtener esto de una forma alternativa

```
mean (VENTAS_PUBLICIDAD_MOD1$`10 Mill`)
```

```
[1] 166198.7
```

```
##
colMeans(VENTAS_PUBLICIDAD_MOD1)
```

```
 10 Mill  11 Mill  12 Mill  13 Mill  14 Mill  15 Mill  16 Mill  17 Mill
166198.7 204532.2 230955.6 233163.7 281324.2 342662.2 382323.6 476347.2
 18 Mill  19 Mill
594125.2 922220.4
```

Es interesante este metodo dado a que vemos que la media de las venta con el comando anterior fue aproximada.

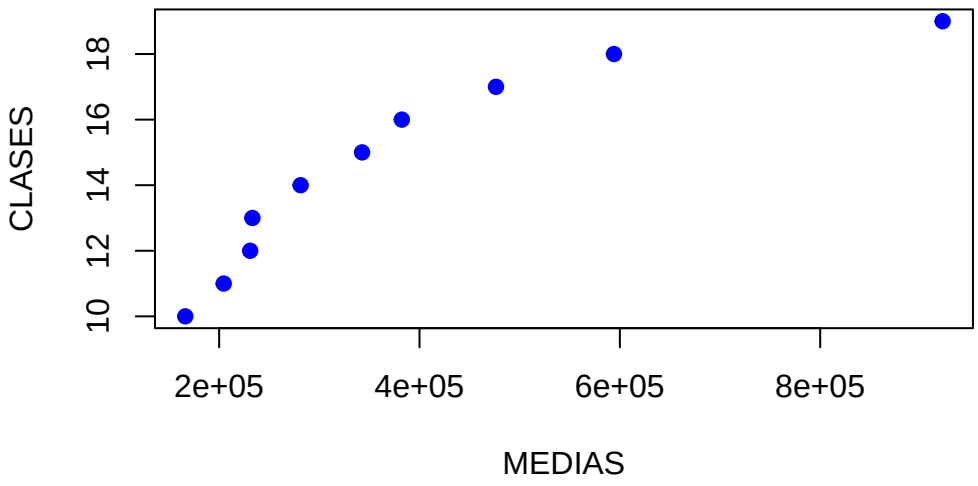
Para observarlos de mejor forma y poder trabajar con las medias, creamos un dataframe con las medias entregadas y con un vector que represente los millones de inversión

```
MEDIAS <- colMeans(VENTAS_PUBLICIDAD_MOD1)
CLASES <- c(10,11, 12, 13, 14, 15, 16, 17, 18, 19 )

BASE_MEDIAS_MILL    <- as.data.frame(cbind(MEDIAS,CLASES))

BASE_MEDIAS_MILL
```

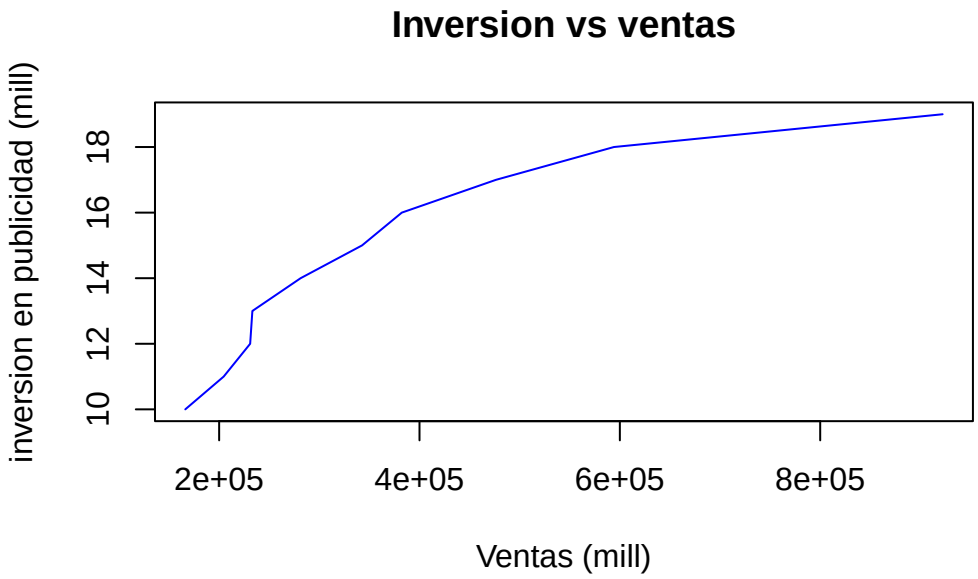
```
          MEDIAS CLASES
10 Mill 166198.7     10
11 Mill 204532.2     11
12 Mill 230955.6     12
13 Mill 233163.7     13
14 Mill 281324.2     14
15 Mill 342662.2     15
16 Mill 382323.6     16
17 Mill 476347.2     17
18 Mill 594125.2     18
19 Mill 922220.4     19
```

Grafique la proyección lineal encontrada en la pregunta anterior.

```
plot(BASE_MEDIAS_MILL, col="blue", pch=19)
```

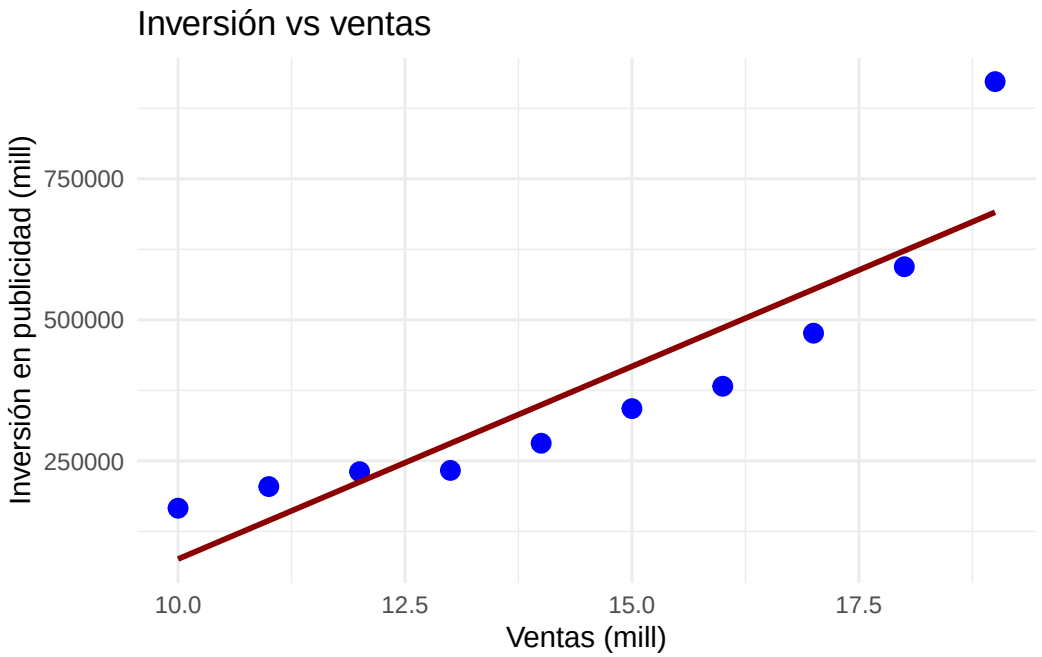



```
plot(BASE_MEDIAS_MILL,
     col="blue",
     pch=19,
     xlab="Ventas (mill)",
     ylab="inversion en publicidad (mill)",
     main="Inversion vs ventas",
     type="l")
     abline(lm(MEDIAS~CLASES,BASE_MEDIAS_MILL))
```

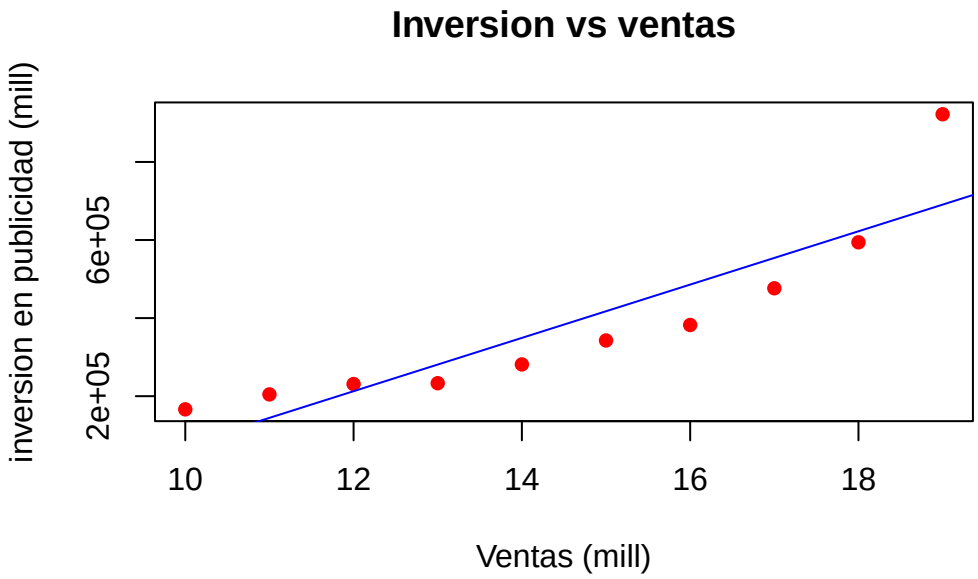



```
library(ggplot2)

ggplot(BASE_MEDIAS_MILL, aes(x = CLASES, y = MEDIAS)) +
  geom_point(color = "blue", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "darkred") +
  labs(
    title = "Inversión vs ventas",
    x = "Ventas (mill)",
    y = "Inversión en publicidad (mill)"
  ) +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

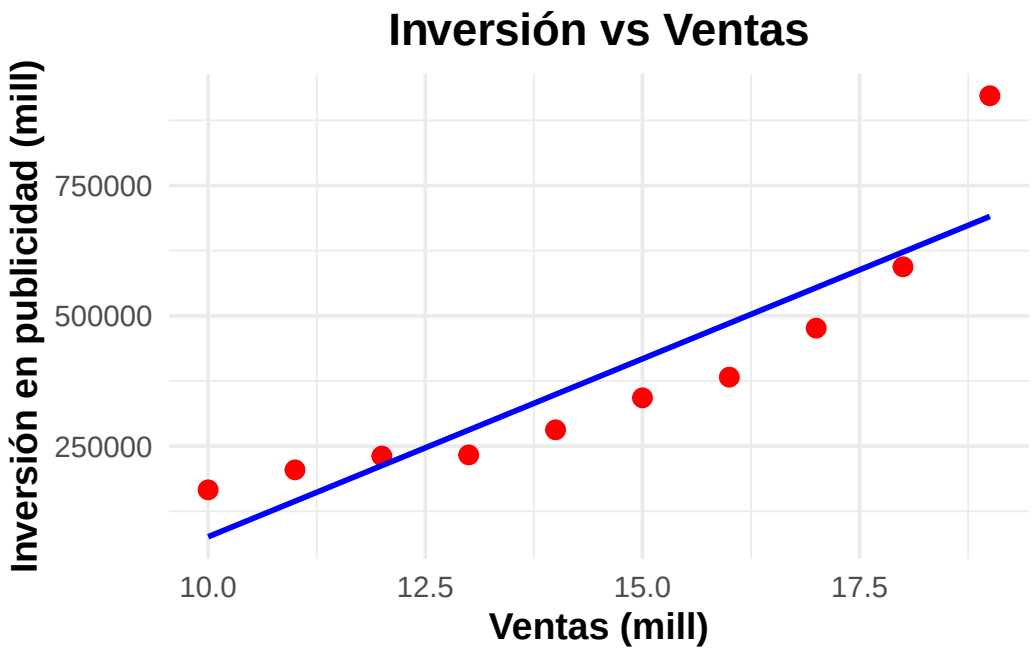



```
plot(MEDIAS~CLASES,
     col="red",
     pch=16,
     xlab="Ventas (mill)",
     ylab="inversion en publicidad (mill)",
     main="Inversion vs ventas")
abline(lm(MEDIAS~CLASES),col="blue")
```




```
ggplot(BASE_MEDIAS_MILL, aes(x = CLASES, y = MEDIAS)) +
  geom_point(color = "red", size = 3) +  # puntos rojos
  geom_smooth(method = "lm", se = FALSE, color = "blue", linewidth = 1) +  # línea de regresi
  labs(
    title = "Inversión vs Ventas",
    x = "Ventas (mill)",
    y = "Inversión en publicidad (mill)"
  ) +
  theme_minimal(base_size = 14) +  # estilo limpio
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold")
  )
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Como conclusión claramente vemos que existe una relación positiva la invertir en publicidad con el número de ventas.

Modelamieno econometrico

Generar un modelo econométrico en el cual analice la relación mencionada en la pregunta anterior. ¿Por qué tiene sentido definir un error aleatorio? ¿Qué propiedades debería tener este error?

Generamos una regesión lineal del data frame que creamos.

```
REG_MILL_PUBLICIDAD     <-  lm(MEDIAS~CLASES,BASE_MEDIAS_MILL)
REG_MILL_PUBLICIDAD
```

```
Call:
lm(formula = MEDIAS ~ CLASES, data = BASE_MEDIAS_MILL)

Coefficients:
(Intercept)       CLASES
    -606756        68286
```

```
summary (REG_MILL_PUBLICIDAD)
```

```
Call:
lm(formula = MEDIAS ~ CLASES, data = BASE_MEDIAS_MILL)

Residuals:
    Min      1Q  Median      3Q     Max
-103490  -73129  -38026   49681  231550

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -606756     178034  -3.408 0.009249 **
CLASES         68286      12044   5.670 0.000471 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109400 on 8 degrees of freedom
Multiple R-squared:  0.8007,    Adjusted R-squared:  0.7758
F-statistic: 32.14 on 1 and 8 DF,  p-value: 0.0004706
```

Como se observa en el coeficiente de clases el cual representa los millones de inversión en publicidad es positivo, además con un p value bajo, por ende, es muy significativo.

MEDIAS$=-606{,}756+68{,}286\times$CLASES

Ahora generamos la regresión lineal, con la tabla original.

```
REG_MILL_PUBLICIDAD2 <-   lm(Ventas_Mill~Publicidad_Mill,VENTAS_PUBLICIDAD_MOD2)
REG_MILL_PUBLICIDAD2
```

```
Call:
lm(formula = Ventas_Mill ~ Publicidad_Mill, data = VENTAS_PUBLICIDAD_MOD2)

Coefficients:
    (Intercept)  Publicidad_Mill
        -606756            68286
```

```
summary (REG_MILL_PUBLICIDAD2)
```

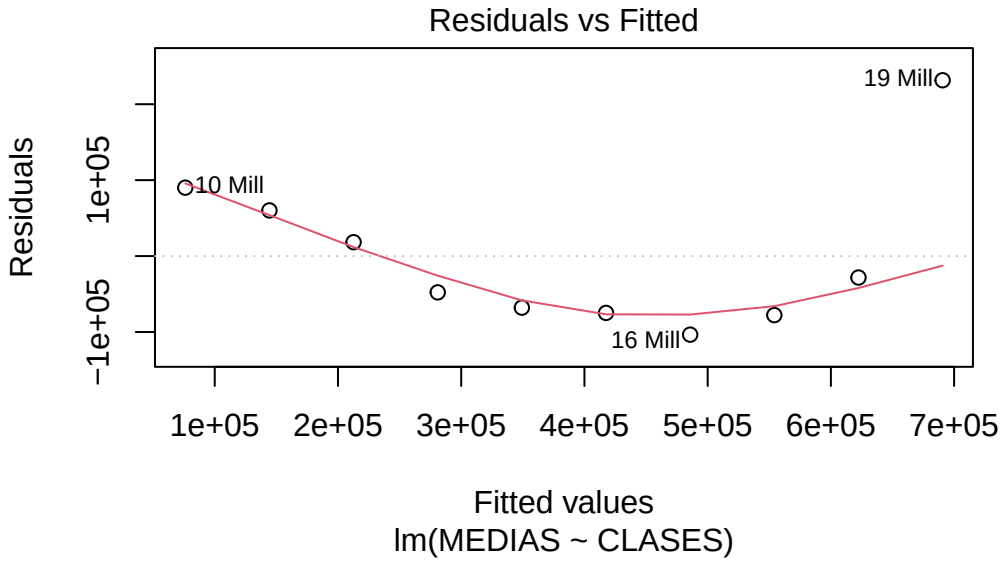
```
Call:
lm(formula = Ventas_Mill ~ Publicidad_Mill, data = VENTAS_PUBLICIDAD_MOD2)
```

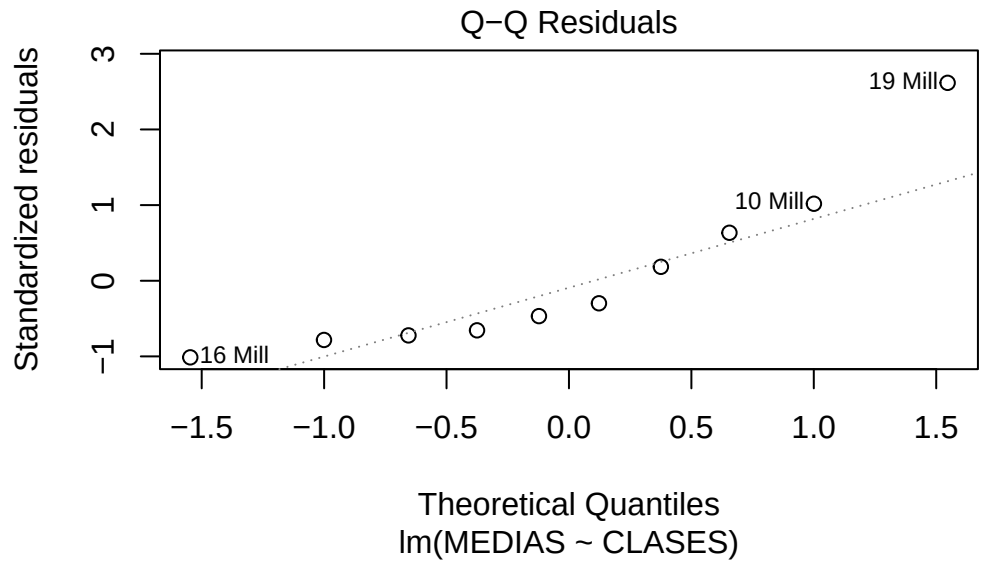
```
Residuals:
    Min      1Q  Median      3Q     Max
-630670 -322331 -115269  124139 3309330

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -606756     302687  -2.005  0.04808 *
Publicidad_Mill    68286      20477   3.335  0.00125 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 558000 on 88 degrees of freedom
Multiple R-squared:  0.1122,    Adjusted R-squared:  0.1021
F-statistic: 11.12 on 1 and 88 DF,  p-value: 0.001251
```

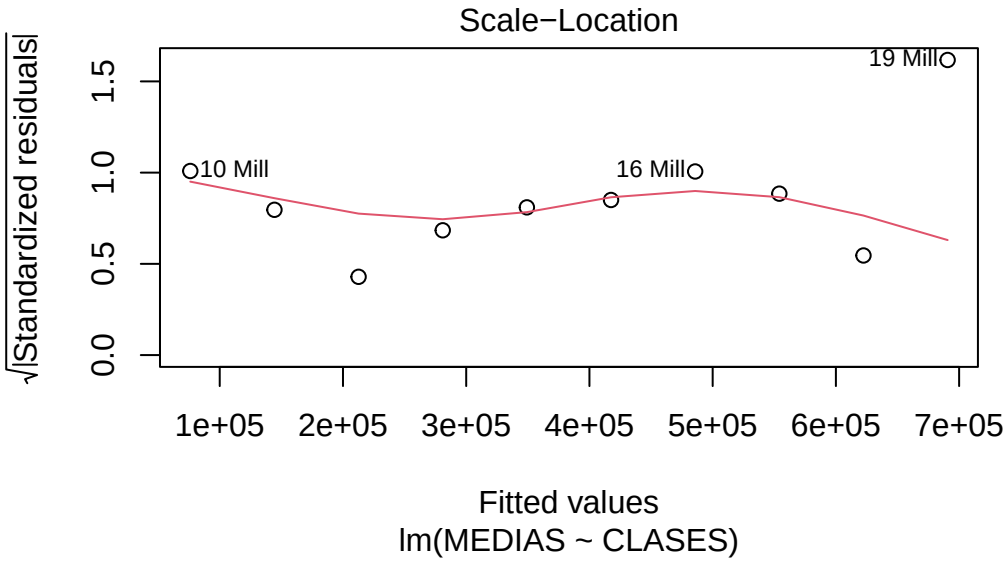
Se mantiene la conclusión.

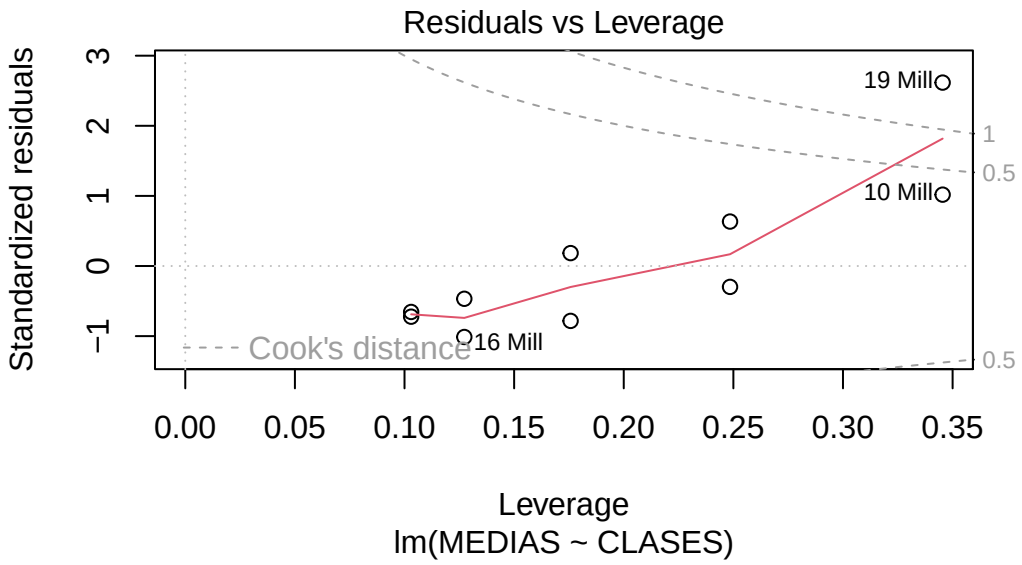
Graficamos

```
plot(REG_MILL_PUBLICIDAD)
```




Q−Q Residuals
Standardized residuals
19 Mill
10 Mill
16 Mill
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
lm(MEDIAS ~ CLASES)


Scale−Location
√|Standardized residuals|
19 Mill
10 Mill
16 Mill
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
Fitted values
lm(MEDIAS ~ CLASES)



Por cada millón extra invertido en publicidad por parte de las empresas encuestadas las ventas aumentan en promedio 68.286 millones. Tal como lo muestra la regresión planteada la variable inversión en publicidad si es estadísticamente relevante para explicar la cantidad de venta, esto con un nivel de certeza del 95%

¿Por qué tiene sentido definir un error aleatorio?

Resp: Tiene sentido definir un nivel de validación de error aleatorio, debido a que, cuando se estudian datos de corte transversal obtenidos de muestra aleatoria, es muy probable que algunos valores se presenten variaciones significativas con respecto a la media de la población, definir un nivel de confianza del 95% por ejemplo nos permite indicar que con un 95% de certeza la estimación encontrada representará la población estudiada, pero existe un 5% que puede no estar representada por la estimación.