UNIVERSIDAD DE EL SALVADOR

FACULTAD MULTIDISCIPLINARIA DE OCCIDENTE

DEPARTAMENTO DE MATEMÁTICAS

Universidad de El Salvador

Hacia la libertad por la cultura

PRACTICA SEMANA 12 DE R

CARRERA:

LICENCIATURA EN ESTADÍSTICA

DOCENTE:

LIC. JAIME ISAAC PEÑA

PRESENTADO POR:

CRISTIAN ALBERTO ZALDAÑA ALVARADO

Índice

Índice de tablas

Índice de figuras

1. REGRESIÓN LINEAL SIMPLE

Los modelos de regresión lineal son modelos probabilísticos basados en una función lineal, expresamos el valor de nuestra variable de estudio (interés), a la que también llamamos variable dependiente, en función de una o más variables a quienes llamamos variables independientes o explicativas, y las cuales suponemos tienen un efecto sobre nuestra variable de estudio. Los pasos básicos a seguir en el estudio de un modelo lineal son:

- Escribir el modelo matemático con todas sus hipótesis.
- Estimación de los parámetros del modelo.
- Inferencias sobre los parámetros.
- Diagnóstico del modelo.

1.1. EJEMPLO 1

En el archivo "costes.dat" se encuentra la información correspondiente a 34 fábricas de producción en el montaje de placas para ordenador, el archivo contiene la información sobre el costo total (primera columna) y el número de unidades fabricadas (segunda columna). Suponga que deseamos ajustar un modelo de regresión simple a los datos para estimar el costo total en función del número de unidades fabricadas.

Ejecutamos lo siguiente.

```
> Datos=read.table("costes.txt")
```

Renombrando a las variables

```
> names(Datos)=c("Costos","Unidades")
```

Realizando el diagrama de dispersion entre las dos variables

```
> plot(Datos$Unidades,Datos$Costos)
```

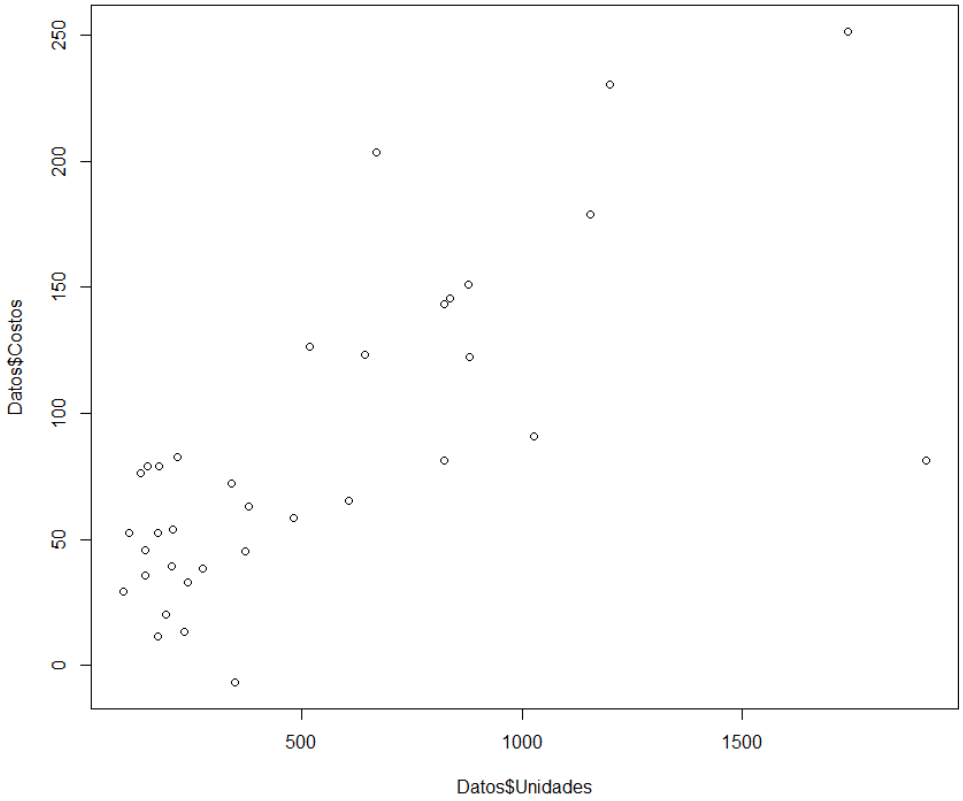

Figura 1: Gráfico de Dispersion

se aprecia una relación entre las variables por lo que se procede a ajustar el modelo de regresión

```
> regresion<-lm(Datos$Costos~Datos$Unidades)
> summary(regresion)

Call:
lm(formula = Datos$Costos ~ Datos$Unidades)

Residuals:
     Min       1Q   Median       3Q      Max
-137.386  -24.496   -0.117   29.848  105.028

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    33.92200   11.57500   2.931   0.0061 **
Datos$Unidades  0.09640    0.01665   5.789  1.8e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44.49 on 33 degrees of freedom
Multiple R-squared:  0.5039,        Adjusted R-squared:  0.4888
F-statistic: 33.51 on 1 and 33 DF,  p-value: 1.796e-06
```

Ejecutar lo siguiente:

```
> regresion2<-lm(Datos$Costos~Datos$Unidades-1)
> summary(regresion2)

Call:
lm(formula = Datos$Costos ~ Datos$Unidades - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-174.579   -4.844   19.527   35.812  114.095

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
Datos$Unidades  0.13350    0.01197   11.16 6.59e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49.21 on 34 degrees of freedom
Multiple R-squared:  0.7854,        Adjusted R-squared:  0.7791
F-statistic: 124.5 on 1 and 34 DF,  p-value: 6.591e-13
```

En este caso el modelo resultante sería: costos $= 0.1335$(unidades); el cual es un mejor modelo en términos de variabilidad explicada.

Una vez estimados los parámetros del modelo, el siguiente paso es validarlo, es decir verificar si se cumplen las cuatro hipótesis básicas del modelo (nulidad, normalidad, independencia y homocesdasticidad de los residuos). Para verificar esto, podríamos realizar los siguientes pasos:

Efectua un analisis grafico de bondad de ajuste del modelo

```
> par(mfrow = c(2, 2))
> plot(regresion2)
> par(oma=c(1,1,1,1), new=T, font=2, cex=0.5)
> mtext(outer=T, "Gráficos para validación del modelo: Costos en función de las unidades",
+ side=3)
```

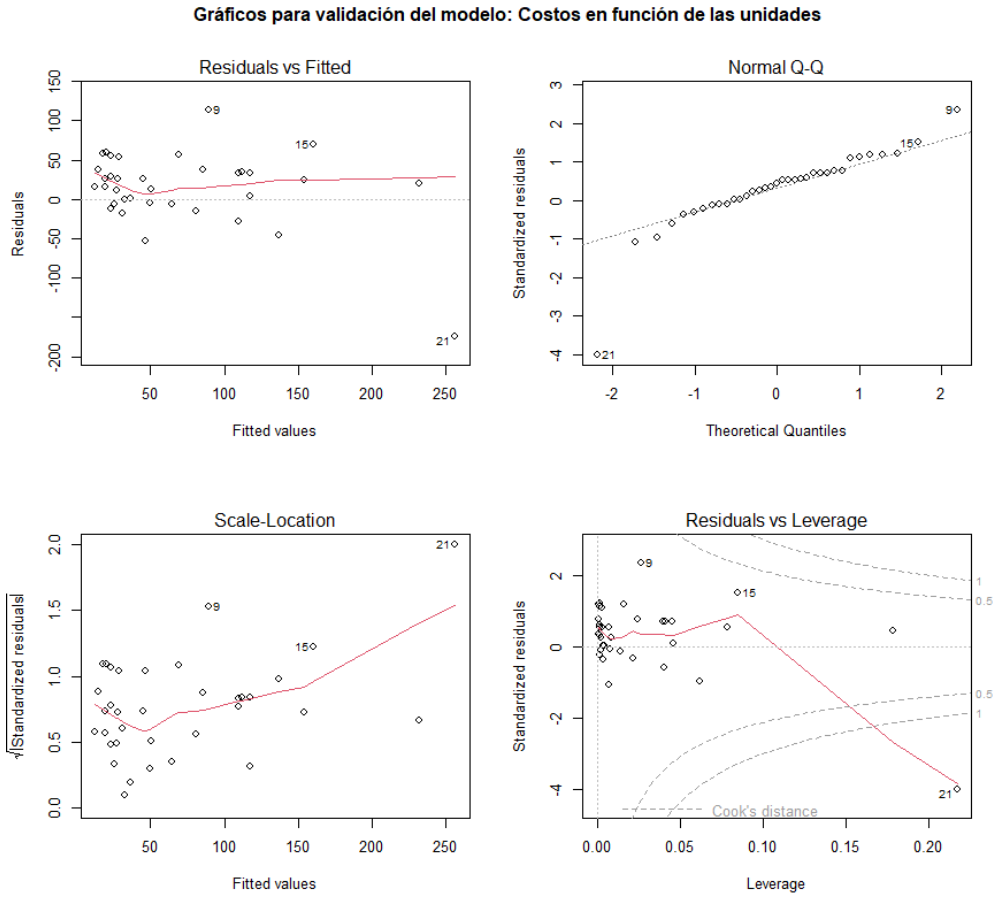

Figura 2: Gráficos para validación del modelo

En los gráficos que se muestra en la parte superior se contrasta los cuatro supuestos. En el de la izquierda se verifican: nulidad, independencia y homocedasticidad; a partir del gráfico mostrado parece existir indicios de falta de homocedasticidad, por su parte los residuos pueden considerarse constante pues no muestran ningún patrón; sin embargo, la media de los residuos no parece ser nula, lo cual indica falta de linealidad en el modelo (es decir, es necesario incorporar más variables o tal vez términos cuadráticos). En la figura de la derecha se contrasta la normalidad, y puede apreciarse que los residuos parecen seguir una distribución normal.

Por su parte, también es de mencionar que en el gráfico se muestran puntos que posiblemente sean observaciones atípicas, por lo que habría que estudiarlas.

Información sobre el modelo ajustado que proporciona la función lm()

```
> formula(regresion2) # Extrae la fórmula del modelo.

Datos$Costos ~ Datos$Unidades - 1

> coef(regresion2) # Extrae el vector de coeficientes de regresión.

Datos$Unidades
    0.1334998

> residuals(regresion2) # Extrae el vector de residuos.
```

```
         1          2          3          4          5          6
-5.9252241 24.9422285 57.2435787 16.1557654 29.5426287 55.6391079
         7          8          9         10         11         12
37.4643315 26.5128603 114.0951533 16.5660408 34.0763027 53.6636345
        13         14         15         16         17         18
26.6898859 -15.4751725 70.5629667 19.5271516 11.9533340 -5.3935511
        19         20         21         22         23         24
-11.3620767 4.7973082 -174.5790615 -45.9982846 -17.9578900 -4.2938112
        25         26         27         28         29         30
34.1602682 33.4958317 0.4420639 38.3225636 -53.0131662 26.0890974
        31         32         33         34         35
12.5185354 -28.6938846 58.5790516 1.7630220 59.1208689
```

```
> regresion2$fitted.values # Extrae un vector con los valores estimados.

       1        2        3        4        5        6        7        8
64.34689 154.05874 69.15288 19.35747 23.09546 23.36246 85.84035 19.09047
       9       10       11       12       13       14       15       16
89.44485 12.68248 117.07930 28.96945 45.38992 80.76736 159.93273 232.02260
      17       18       19       20       21       22       23       24
27.23395 25.63196 22.96196 117.47980 255.91906 136.97077 31.23895 49.39492
      25       26       27       28       29       30       31       32
111.60581 109.87031 32.30694 14.41798 46.32442 27.76795 50.72991 110.00381
      33       34       35
17.75547 36.71244 19.89147
```

```
> vcov(regresion2) # Extrae la matriz de covarianzas de los parámetros.

               Datos$Unidades
Datos$Unidades   0.0001432011
```

```
> ls.diag(regresion2) # Calcula los residuales, errores estándar de los parámetros, distancias Cook.

$std.dev
[1] 49.20928

$hat
 [1] 0.0137387236 0.0787524233 0.0158676214 0.0012433354 0.0017698828
 [6] 0.0018110415 0.0244497405 0.0012092731 0.0265461726 0.0005337028
[11] 0.0454832506 0.0027846574 0.0068361273 0.0216452725 0.0848722948
[16] 0.1786289532 0.0024610058 0.0021799913 0.0017494809 0.0457949567
[21] 0.2173184179 0.0622511001 0.0032380535 0.0080957252 0.0413299485
[26] 0.0400545613 0.0034632436 0.0006897629 0.0071205126 0.0025584620
[31] 0.0085392456 0.0401519584 0.0010460576 0.0044721637 0.0013128794

$std.res
 [1] -0.12124442  0.52808038  1.17260833  0.32851156  0.60087863  1.13168807
 [7]  0.77080784  0.53910368  2.34997086  0.33673450  0.70878358  1.09204005
[13]  0.54423845 -0.31793642  1.49895619  0.43784678  0.24320756 -0.10972401
[19] -0.23109519  0.09979981 -4.01007557 -0.96527557 -0.36552118 -0.08761148
[25]  0.70898925  0.69473704  0.00899894  0.77903567 -1.08115618  0.53084568
[31]  0.25548695 -0.59517005  1.19102962  0.03590740  1.20220642

$stud.res
```

```
 [1] -0.119473935  0.522403330  1.179328393  0.324159319  0.595144672
 [6]  1.136532408  0.766111172  0.533401168  2.529690985  0.332300135
[11]  0.703499199  1.095240031  0.538526063 -0.313692652  1.528103030
[16]  0.432581090  0.239812980 -0.108117520 -0.227850384  0.098335619
[21] -5.441881509 -0.964279034 -0.360815373 -0.086323206  0.703706396
[26]  0.689354550  0.008865626  0.774436698 -1.083933598  0.525161706
[31]  0.251943705 -0.589430743  1.198655383  0.035376083  1.210400702

$cooks
 [1] 2.047755e-04 2.383898e-02 2.216993e-02 1.343476e-04 6.401583e-04
 [6] 2.323641e-03 1.489076e-02 3.518799e-04 1.505953e-01 6.054895e-05
[11] 2.393841e-02 3.330121e-03 2.038767e-03 2.236389e-03 2.083829e-01
[16] 4.169239e-02 1.459274e-04 2.630304e-05 9.359475e-05 4.780083e-04
[21] 4.464949e+00 6.185333e-02 4.340279e-04 6.264812e-05 2.167080e-02
[26] 2.013939e-02 2.814313e-07 4.189037e-04 8.382848e-03 7.228166e-04
[31] 5.621878e-04 1.481789e-02 1.485440e-03 5.792050e-06 1.899999e-03

$dfits
 [1] -0.0141010149  0.1527389219  0.1497489923  0.0114372818  0.0250599080
 [6]  0.0484105087  0.1212841638  0.0185600398  0.4177445106  0.0076788503
[11]  0.1535668861  0.0578762950  0.0446787837 -0.0466592443  0.4653661151
[16]  0.2017315204  0.0119114345 -0.0050535591 -0.0095385991  0.0215426354
[21] -2.8675090121 -0.2484465426 -0.0205651234 -0.0077986703  0.1461131657
[26]  0.1408138463  0.0005226419  0.0203463104 -0.0917931003  0.0265973773
[31]  0.0233816685 -0.1205549113  0.0387881801  0.0023710564  0.0438860875

$correlation
               Datos$Unidades
Datos$Unidades              1

$std.err
                     [,1]
Datos$Unidades 0.01196667

$cov.scaled
               Datos$Unidades
Datos$Unidades   0.0001432011

$cov.unscaled
               Datos$Unidades
Datos$Unidades   5.913605e-08
```

> step(regresion2) # Permite obtener el mejor conjunto de regresión y proporciona la

```
Start:  AIC=273.71
Datos$Costos ~ Datos$Unidades - 1

                 Df Sum of Sq    RSS    AIC
<none>                         82333 273.71
- Datos$Unidades  1    301376 383709 325.58

Call:
lm(formula = Datos$Costos ~ Datos$Unidades - 1)
```

```
Coefficients:
Datos$Unidades
        0.1335

> #estimación de los coeficientes (válido únicamente en modelos de regresión múltiple).
```

de todos los resultados anteriores nos concentraremos en la instrucción: ls.diag(regresion2). Con esta instrucción obtenemos para cada observación en el conjunto de datos, medidas que nos ayudarán a identificar observación atípicas (tienen un impacto únicamente en las medidas resumen del modelo) y observaciones influyentes (tienen un efecto marcado en la estimación de los parámetros). Al digitar la instrucción anterior en R se mostrará los siguientes resultados (cada uno de ellos en un vector).