Universidad de Puerto Rico en Aguadilla

Regresión Lineal con R

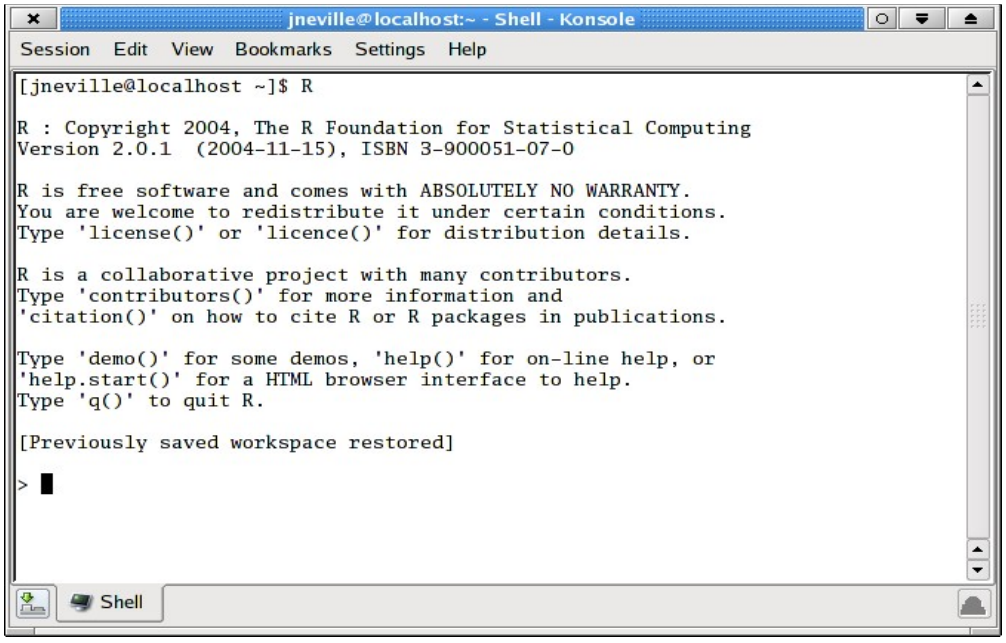
Prof. José Neville Díaz Caraballo

1. **¿Qué es R?**
   R es un sistema para la realización de cálculos estadísticos y la creación de gráficas. Este consiste en un lenguage, acceso a funciones en el sistema y la habilidad de correr programas guardados en archivos ¨script¨. R es más allá de un paquete estadístico, es un lenguage. Existen versiones para Windows, Mac, Unix y Linux. Este documento estará presentado en Linux Fedora Core 3.

2. **¿Cómo obtener R?**
   R es gratis, su licencia es GPL. Para obtener su copia visite http://www.r-project.org/. Este es uno de los proyectos más importantes del ¨open sorce¨.

3. **Pantalla Inicial**

   El Programa luce de esta manera:



   Los comandos deben ser escritos en la línea de comandos > . Para ejecutar debes oprimir ¨enter¨. Para comandos usaremos Bitstream Charter y para comentario el signo de # seguido del comentario todo en Bitstream Vera Sans Mono.

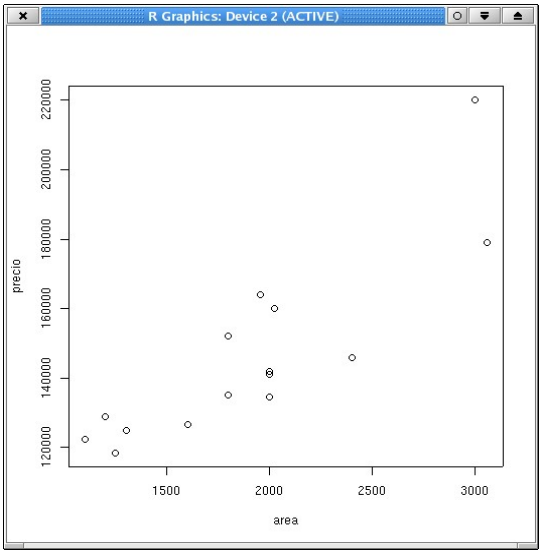
4. Regresión Lineal

> datos<-read.table("ftp://math.uprag.edu/pub/area-precio.dat",header=T)
#read.table commando para leer datos

> attach(datos) #attach comando que establece el conjunto de datos a
#utilizar

> ?plot #? el signo de pregunta junto al comando nos da la ayuda de ese
#comando en particular.

> plot(area,precio) #plot diagrama de dispersión de las variables área y
#precio



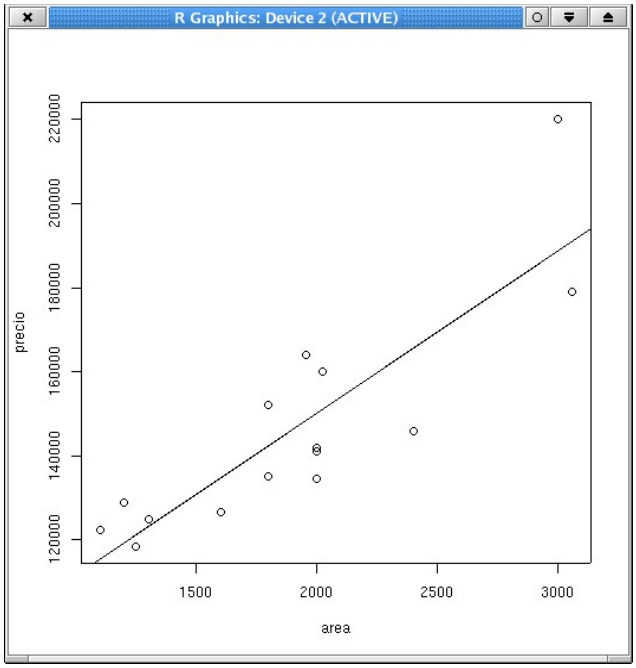
> cor(area,precio) #cor correlación Pearson
[1] 0.85818

> cor.test(area,precio) #cor.test correlación Pearson con prueba de hipótesis

Pearson's product-moment correlation

data: area and precio
t = 6.0275, df = 13, p-value = 4.251e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6172832 0.9519517
sample estimates:
   cor

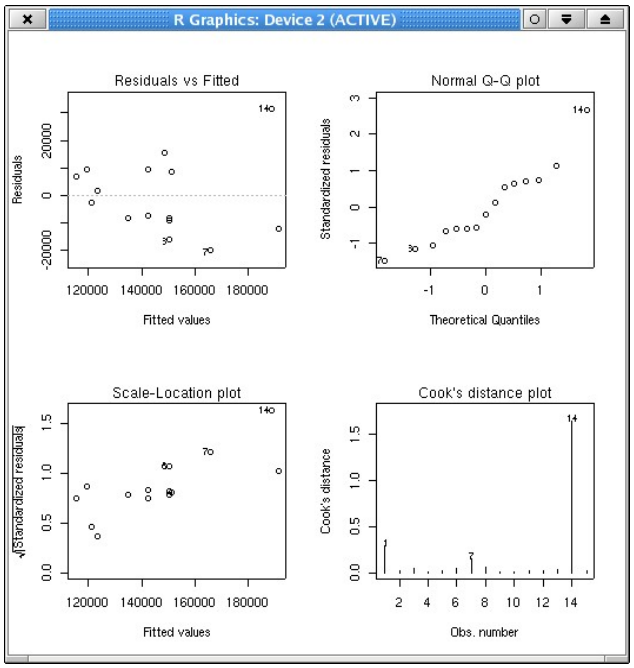
```
0.85818
> r1<-lm(precio~area)   #lm(y ~ x) regresión lineal
> r1

Call:
lm(formula = precio ~ area)

Coefficients:
(Intercept)        area
  73167.75       38.52


> r2<-lm(precio~0+area)   #lm(y ~ 0+x) regresión lineal a través del origen
> r2

Call:
lm(formula = precio ~ 0 + area)

Coefficients:
 area
73.86

> summary(r2)    #summary resumen de un objeto. En este caso de  la
                 #regresión sin intercepto

Call:
lm(formula = precio ~ 0 + area)

Residuals:
   Min     1Q Median     3Q    Max
-47013  -6221   8323  22925  41253

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
area   73.861      3.344   22.09 2.79e-12 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 25680 on 14 degrees of freedom
Multiple R-Squared: 0.9721,     Adjusted R-squared: 0.9701
F-statistic: 487.9 on 1 and 14 DF,  p-value: 2.786e-12



>  plot(area,precio)
> abline(r1)    #abline añade la línea de regresión al diagrama de puntos
```




```
> summary(r1)

Call:
lm(formula = precio ~ area)

Residuals:
   Min     1Q Median     3Q    Max
-19623  -8759  -2745   9157  31263

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 73167.748  12674.143   5.773 6.46e-05 ***
area           38.523      6.391   6.028 4.25e-05 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 14120 on 13 degrees of freedom
Multiple R-Squared: 0.7365,     Adjusted R-squared: 0.7162
F-statistic: 36.33 on 1 and 13 DF,  p-value: 4.251e-05
```

```
> resid(r1) #resid residuales de lm(precio~area)
      1       2       3       4       5       6       7
-12048.346  -8304.662 -15713.891   1752.259  -8213.891  15481.124 -19623.119
      8       9      10      11      12      13      14
 9604.566  -7509.276  -2744.541   8823.033   9490.724   6956.873  31263.038
     15
-9213.891

> coef(r1) #coef coeficientes de lm(precio~area)
(Intercept)        area
73167.74838    38.52307

> fitted(r1) #fitted  ŷ valor estimado que nos da el modelo lm(precio~area)
     1      2      3      4      5      6      7      8
191048.3 134804.7 150213.9 123247.7 150213.9 148518.9 165623.1 119395.4
     9     10     11     12     13     14     15
142509.3 121244.5 151177.0 142509.3 115543.1 188737.0 150213.9

> anova(r1) #anova tabla de análisis de varianza
Analysis of Variance Table

Response: precio
          Df     Sum Sq    Mean Sq F value    Pr(>F)
area       1 7241245891 7241245891  36.331 4.251e-05 ***
Residuals 13 2591087442  199314419
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1



> par(mfrow=c(2,2))    #par(mfrow=c(2,2)) divide la pantalla en dos filas
                        #y dos columnas

> attributes(r1)  #attributes demuestra los atributos de un objeto

$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

$class
[1] "lm"
```

```
> plot(r1)    #r1 objeto con los atributos de lm(precio ~ area)
              #plot(r1) gráficas de análisis de residuales
```

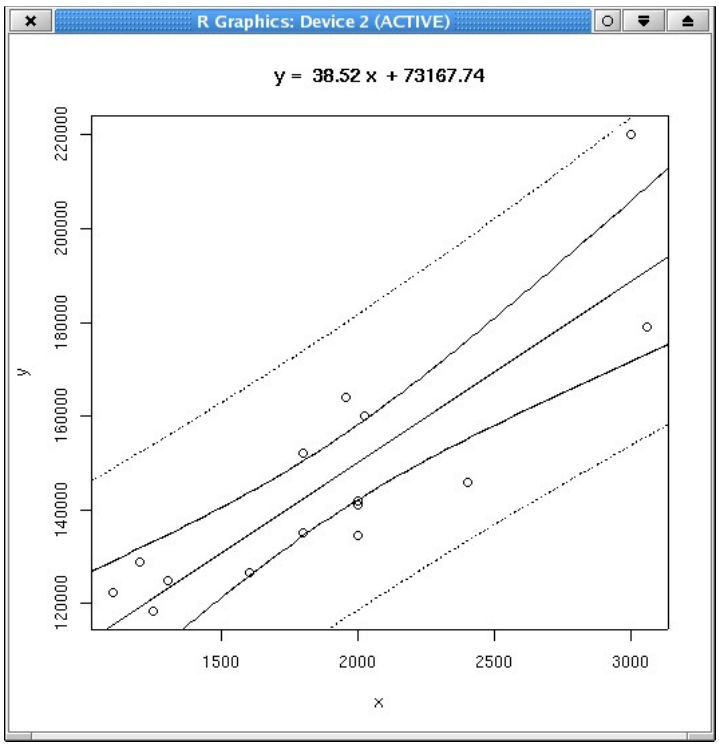


Warning message:
X11 used font size 8 when 7 was requested

```
> datos
   area precio
1  3060 179000
2  1600 126500
3  2000 134500
4  1300 125000
5  2000 142000
6  1956 164000
7  2400 146000
8  1200 129000
9  1800 135000
10 1248 118500
11 2025 160000
12 1800 152000
13 1100 122500
14 3000 220000
15 2000 141000

>area2<-area^2       #creando el vector de área al cuadrado
```

```
>precio2<-precio^2   #creando el vector de precio al cuadrado

> length(area)    #length calcula el tamaño del objeto, similar a count en
                  #minitab
[1] 15

> Sxx<-sum(area2)-(sum(area)^2)/length(area)
> Syy<-sum(precio2)-(sum(precio)^2)/length(precio)

> as.numeric(area)    #as.numeric se utiliza para eliminar el problema de
                       #¨overflow¨
 [1] 3060 1600 2000 1300 2000 1956 2400 1200 1800 1248 2025 1800 1100 3000 2000
> as.numeric(precio)
 [1] 179000 126500 134500 125000 142000 164000 146000 129000 135000 118500
[11] 160000 152000 122500 220000 141000

> Sxy<-sum(as.numeric(area*precio))-
sum(as.numeric(precio))*sum(as.numeric(area))/length(area)

> beta<-Sxy/Sxx    #calculando beta
> beta
[1] 38.52307
> alpha<-mean(precio)-beta*mean(area)    #calculando alpha
> alpha
[1] 73167.75

>install.packages("Simple",contriburl="http://www.math.csi.cuny.edu/Statistics/R/simpleR/")
#instalando un paquete desde una dirección de internet

>library("Simple")    #library llamando la librería Simple
```

```
>> simple.lm(area,precio,show.ci=TRUE,conf.level=.90)    #simple.lm excelente rutina
                                                         #creando bandas de
                                                         #confianzas
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)         x
  73167.75      38.52
```



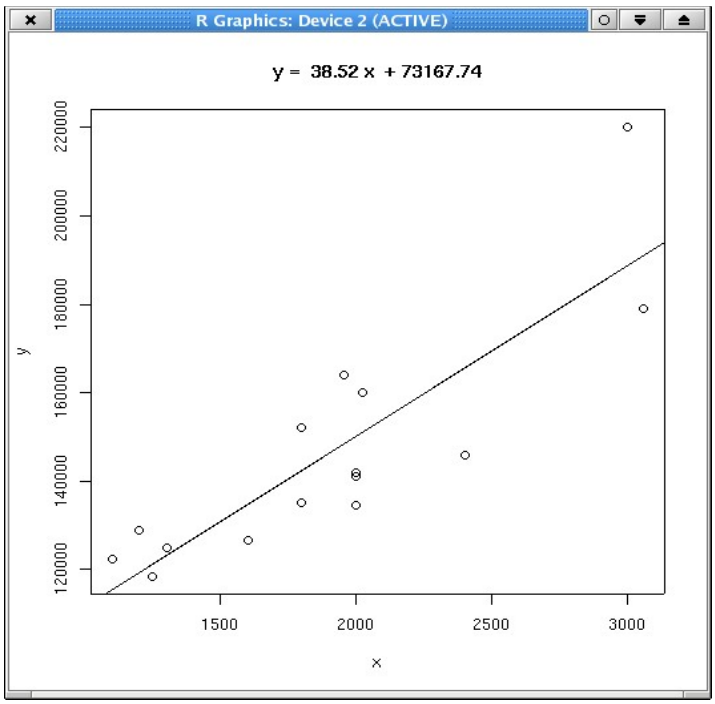
> ?simple.lm #?simple.lm solicitando ayuda sobre la función

> simple.lm(area,precio,pred=c(2500)) #prediciendo el valor de una casa
#con área de 2500

[1] 169475.4

Call:
lm(formula = y ~ x)

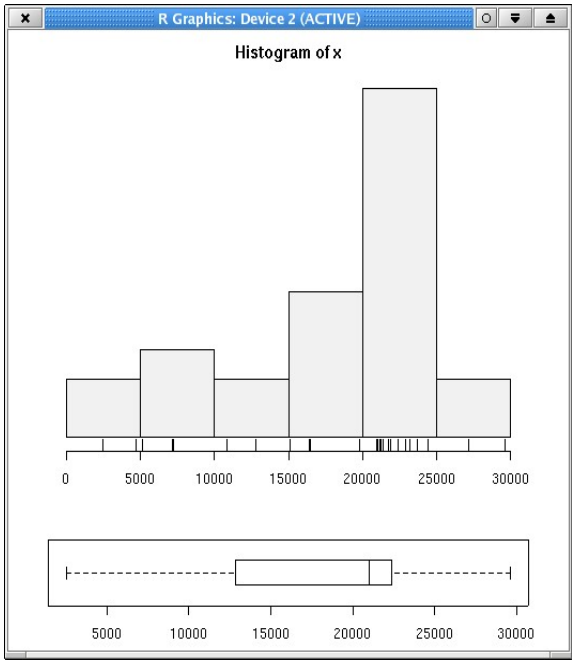
Coefficients:
(Intercept) x
73167.75 38.52



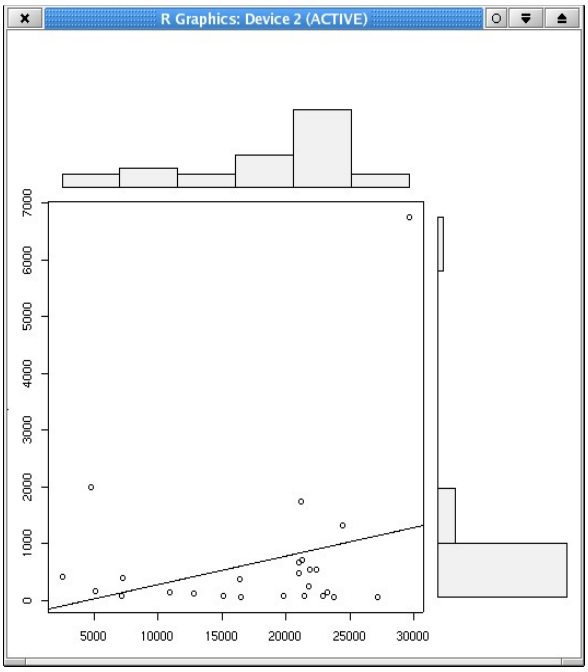
> data(emissions) #data se utiliza para llamar el conjunto de datos

> attach(emissions) #attach haciéndolo disponible

> simple.hist.and.boxplot(perCapita) #Esta rutina pertenece a la librería Simple



> simple.scatterplot(perCapita,CO2) #Esta rutina pertenece a la librería Simple

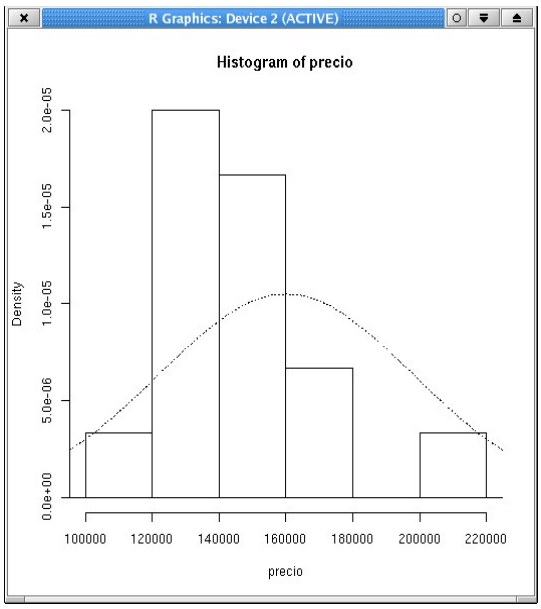



```
> summary(precio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 118500  127800  141000  146300  156000  220000

> sd(precio)
[1] 26501.12
```

> x<-rnorm(100,146300,26501) #Vector con media y sd igual a precio

> hist(precio,probability=TRUE) #hist histograma de la variable precio

> curve(dnorm(x,mean=mean(x),sd=sd(x)),lty=3,add=T) #Esta añade la curva normal




```
> rendimiento<-
read.table("/home/jneville/Desktop/Documentos/Regresion/rendimiento.dat",header=T)
#llamando el conjunto de datos

> rendimiento
   Y X1 X2 X3 X4 X5 X6
1  67 61 64 75 60 81 45
2  67 71 45 67 80 86 48
3  52 59 67 64 69 79 54
4  52 71 32 44 48 65 43
5  66 62 51 72 71 81 43
6  55 67 51 60 68 81 39
7  42 65 41 58 71 76 35
8  68 78 65 73 93 77 42
9  80 76 57 84 85 79 35
10 50 58 43 55 56 84 40
11 87 86 70 81 82 75 30
12 84 83 38 83 69 79 41
13 70 84 82 68 64 78 37
14 79 78 53 82 84 78 39
15 83 65 49 82 65 55 38
16 75 86 63 79 84 80 41

> library("MASS") #llamando la librería MASS que contiene la función stepAIC

> rendimiento.lm<-lm(Y~.,data=rendimiento) #creando el objeto que contiene la
                                           #regresión multivariada

> rendimiento.backward<-stepAIC(rendimiento.lm,k=log(length(rendimiento)))
```

```
#realizando el stepwise, importante usar k=log(length(rendimiento) este nos
#brinda la salida más similar a Minitab. Recuerde, este es un curso
#introductorio y no estaremos discutiendo las diferencias entre los
#métodos que utiliza Minitab y R.

Start:  AIC= 55.89
Y ~ X1 + X2 + X3 + X4 + X5 + X6

       Df Sum of Sq     RSS    AIC
- X6    1      3.54  228.15  54.19
- X2    1      7.28  231.88  54.45
- X5    1     26.96  251.56  55.76
<none>              224.60  55.89
- X4    1     59.75  284.36  57.72
- X1    1    226.19  450.79  65.09
- X3    1   1123.71 1348.31  82.62

Step:  AIC= 54.19
Y ~ X1 + X2 + X3 + X4 + X5

       Df Sum of Sq     RSS    AIC
- X2    1      6.64  234.78  52.71
- X5    1     23.91  252.05  53.84
<none>              228.15  54.19
- X4    1     59.26  287.41  55.94
- X1    1    243.94  472.09  63.88
- X3    1   1120.52 1348.67  80.68

Step:  AIC= 52.71
Y ~ X1 + X3 + X4 + X5

       Df Sum of Sq     RSS    AIC
- X5    1     29.25  264.04  52.64
<none>              234.78  52.71
- X4    1     61.35  296.14  54.48
- X1    1    237.31  472.09  61.94
- X3    1   1135.84 1370.62  78.99

Step:  AIC= 52.64
Y ~ X1 + X3 + X4

       Df Sum of Sq     RSS    AIC
<none>              264.04  52.64
- X4    1    115.92  379.96  56.52
- X1    1    261.92  525.96  61.72
- X3    1   1289.16 1553.20  79.05
```

Haciendo predicciones en regresión lineal simple

```
> new<-c(1800,2000,2500,3000) #vector con los valores que se desean predecir

> attributes(r1)
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

$class
[1] "lm"

> r1$coefficients[1] #intercepto r1<-lm(precio~area)
(Intercept)
  73167.75

> r1$coefficients[2] #pendiente
   area
38.52307

> predictions<-r1$coefficients[1]+r1$coefficients[2]*new

> predictions
[1] 142509.3 150213.9 169475.4 188737.0
```

Haciendo predicciones en regresión lineal múltiple

```
> attributes(rendimiento.backward)
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
[13] "anova"

$class
[1] "lm"

> rendimiento.backward$coefficients #observando los coeficientes
(Intercept)          X1          X3          X4
-20.3358187   0.5099809   1.0432851  -0.3132179

> rendimiento #observando el conjunto de datos
   Y X1 X2 X3 X4 X5 X6
1  67 61 64 75 60 81 45
2  67 71 45 67 80 86 48
3  52 59 67 64 69 79 54
4  52 71 32 44 48 65 43
5  66 62 51 72 71 81 43
6  55 67 51 60 68 81 39
7  42 65 41 58 71 76 35
```

```
8  68 78 65 73 93 77 42
9  80 76 57 84 85 79 35
10 50 58 43 55 56 84 40
11 87 86 70 81 82 75 30
12 84 83 38 83 69 79 41
13 70 84 82 68 64 78 37
14 79 78 53 82 84 78 39
15 83 65 49 82 65 55 38
16 75 86 63 79 84 80 41

> predictionsstep<-
rendimiento.backward$coefficients[1]+rendimiento.backward$coefficients[2]*rendimiento[2]+r
endimiento.backward$coefficients[3]*rendimiento[4]+rendimiento.backward$coefficients[4]*ren
dimiento[5]


> predictionsstep
       X1
1  70.22633
2  60.71550
3  54.91127
4  46.74292
5  64.16106
6  55.13120
7  51.08501
8  66.47325
9  79.43517
10 49.08356
11 82.34477
12 86.97323
13 73.40003
14 78.68177
15 78.00316
16 79.63177

> predictionsstep<-
rendimiento.backward$coefficients[1]+rendimiento.backward$coefficients[2]*rendimiento[2]+r
endimiento.backward$coefficients[3]*rendimiento[4]+rendimiento.backward$coefficients[4]*ren
dimiento[5]

> residuales<-predictionsstep-rendimiento[1]  #   e=ŷ−y

> residuales
        X1
1   3.2263315
2  -6.2844971
3   2.9112726
4  -5.2570835
5  -1.8389393
```

```
6   0.1311975
7   9.0850119
8  -1.5267519
9  -0.5648346
10 -0.9164423
11 -4.6552270
12  2.9732325
13  3.4000259
14 -0.3182251
15 -4.9968381
16  4.6317670

> mean(residuales) #siempre la media de los residuales es cero
       X1
-8.881784e-16

> S2<-1/(dimension[1]-(3)-1)*sum(residuales^2) #Estimación de la varianza
                                                #poblacional
> S2
[1] 22.00309

> sqrt(S2) #Estimación de la desviación estandar poblacional
[1] 4.690746

> z<-residuales/sqrt(S2)  #Residuales estudentizados residuales divididos
                          #entre la estimación de la desviación población.

> z
        X1
1   0.68780782
2  -1.33976507
3   0.62064176
4  -1.12073516
5  -0.39203561
6   0.02796944
7   1.93679483
8  -0.32548171
9  -0.12041468
10 -0.19537241
11 -0.99242794
12  0.63385073
13  0.72483699
14 -0.06784106
15 -1.06525456
16  0.98742661
```

```
> predic<-read.table("/home/jneville/Desktop/Documentos/Regresion/predic.dat",header=F)
> predic
  V1 V2 V3
1 70 68 59
2 71 66 69
3 68 67 56
4 54 71 85

> predictionsstep<-
rendimiento.backward$coefficients[1]+rendimiento.backward$coefficients[2]*predic[1]+rendimi
ento.backward$coefficients[3]*predic[2]+rendimiento.backward$coefficients[4]*predic[3]

> predictionsstep
       V1
1 67.82638
2 63.11761
3 66.70279
4 54.65288

> rendimiento.backward

Call:
lm(formula = Y ~ X1 + X3 + X4, data = rendimiento)

Coefficients:
(Intercept)         X1         X3         X4
   -20.3358     0.5100     1.0433    -0.3132

>> predict(rendimiento.backward,newdata=data.frame(X1=55,X3=69,X4=66))
[1] 59.02743

#De esta forma puedes predecir una observación

#Como exportar un conjunto de datos de R para ser usado en otro paquete
estadístico

> library("UsingR") #Librería la cual contiene el conjunto de datos
> data("babies")    #Conjunto de datos que deseamos exportar
> write.table(babies,"babies.txt",sep=" ",quote=FALSE,row.names=FALSE,col.names=TRUE)
#Comando para exportar. Note babies conjunto de datos, "babies.txt"
nombre del archivo a guardar, sep=" " separación.

> help.start() #pedir ayuda, saldrá en el ¨browser¨. Visite ¨packages¨ y
#seleccione el archivo que desea.
```