CAPÍTULO 6

SELECCIÓN DE VARIABLES EN REGRESIÓN

Selección de variables o también llamado selección de un subconjunto de predictoras es un procedimiento estadístico que es importante por diversas razones, entre estas están:

a) No todas las variables predictoras tienen igual importancia, por lo tanto es más eficiente trabajar con un modelo donde las variables importantes estén presentes y las que tienen poca importancia no aparezcan.
b) Algunas variables pueden perjudicar la confiabilidad del modelo, especialmente si están correlacionadas con otras, luego se hace necesario eliminarlas ya que son redundantes.
c) Computacionalmente es más fácil trabajar con un conjunto de variables predictoras pequeño.
d) Es más económico recolectar información para un modelo con pocas variables.
e) Si se reduce el número de variables entonces el modelo se hace más **parsimonioso**. Se dice que un modelo es **parsimonioso** si consigue ajustar bien los datos pero usando la menor cantidad de variables predictoras posibles. Es más conveniente porque sus predicciones son más confiables y además es más robusto que el modelo original.

Desde que empezó a trabajarse en esta área en los años 60 y gracias al desarrollo de las computadoras se han introducido muchos métodos de selección de variables. Aquí describiremos sólo algunos de ellos.

6.1 Metodos "Stepwise"

La idea de estos métodos es elegir el mejor modelo en forma secuencial pero incluyendo (o excluyendo) una sola variable predictora en cada paso de acuerdo a ciertos criterios. El proceso secuencial termina cuando una regla de parada se satisface.

Hay tres algoritmos más comúnmente usados, los cuales serán descritos a continuación.

6.1.1 "Backward Elimination" (Eliminación hacia atrás).

En este caso se comienza con el modelo completo y en cada paso se va eliminando una variable. Si resultara que todas las variables predictoras son importantes, es decir tienen "p-value" pequeños para la prueba t, entonces no se hace nada y el mejor modelo es el que tiene todas las variables predictoras disponibles. En caso contrario, en cada paso la variable que se elimina del modelo es aquella que satisface cualquiera de estos requisitos equivalents entre sí:

a) Aquella variable que tiene el estadístico de t, en valor abosluto, más pequeño entre las variables incluidas aún en el modelo. Es decir, aquella variable con el F parcial más pequeño. El F parcial está definido por:

$$F_p=[SSR_k-SSR_{k-1}]/MSE_k$$

donde SSR_k es la suma de cuadrados debido a la regresión con k variables y SSR_{k-1} es la misma suma con k-1 variables. $MSE_k=SSE_k/n-k-1)$ es el cuadrado medio del error del modelo que incluye k variables. Hay que calcular el F_p para cada una de las variables presentes aún en

el modelo y se elimina del modelo aquella variable que dá el F_p mas pequeño. Se puede mostrar que $t^2=F_p$. En realidad todo el proceso se entiende mucho mejor con la t que con la F.

b) Aquella variable que produce la menor disminución en el R^2 al ser eliminada del modelo. Es decir, aquella variable que produce el mas pequeño incremento en la suma de cuadrados del error.
c) Aquella variable que tiene la correlación parcial (en valor absoluto) más pequeña con la variable de respuesta, tomando en cuenta las variables que quedarían en el modelo. La correlación parcial de Y con la variable X_i se define como la correlación entre los residuales de la regresión de Y con todas las variables predictoras, excepto X_i y los residuales de la regresión de X_i con todas las otras restantes variables predictoras.

El método "Backward" padece del efecto de anidamiento ya que toda variable que es eliminada del modelo ya no vuelve a entrar a él.

El proceso termina cuando se cumple una de las siguientes condiciones:

a) Se llega a un modelo con un número prefijado p* de variables predictoras.
b) El valor de la prueba de F parcial para todas las variables incluidas en el modelo son mayores que un número prefijado F-out (por lo general este valor es 4). O en forma equivalente, se para cuando el valor absoluto del estadístico de t para cada variable es mayor que la raíz cuadrada de F-out (por lo general, |t|>2). Algunas veces se prefija de antemano un nivel de significación dado α* (digamos del 10%) para la prueba de t o de F parcial en cada paso y en este caso se termina el proceso cuando todos los p-values son menores que α*.

6.1.2 "Forward Selection" (Selección hacia adelante).

Aquí se empieza con la regresión lineal simple que considera como variable predictora a aquella que esta más altamente correlacionada (sin tomar en cuenta el signo) con la variable de respuesta. Si esta primera variable no es significativa entonces se considera el modelo $\hat{Y} = \overline{Y}$ y se para el proceso, de lo contrario se sigue y en el siguiente paso se añade al modelo la variable que reune cualquiera de estos requisitos equivalents entre sí:

a) Aquella variable que tiene el estadístico de t , en valor absolute, más grande entre las variables no incluidas aún en el modelo. Es decir, la variable con el F-parcial más grande.
b) Aquella variable que produce el mayor incremento en el R^2 al ser añadida al modelo. Es decir, aquella variable que produce la mayor reducción en la suma de cuadrados del error.
c) Aquella variable que tiene la correlación parcial más alta (en valor absoluto) con la variable de respuesta, tomando en cuenta las variables ya incluidas en el modelo.

Aquí también está presente el efecto de anidamiento ya que toda variable que es añadida al modelo ya no puede salir del mismo.

El proceso termina cuando se cunple una de las siguientes condiciones:

a) Se llega a un modelo con un número prefijado p* de variables predictoras.
b) El valor de la prueba de F parcial para cada una de las variables no incluidas aun en el modelo es menor que un número prefijado F-in (por lo general este valor es 4). O en forma equivalente se para cuando el valor absoluto del estadistico de t es menor que la raíz cuadrada de F-in (por lo general, |t|<2). Algunas veces se prefija de antemano un nivel de significación

dado α^* (digamos del 15%) para la prueba de t o de F parcial en cada paso y en este caso se termina el proceso cuando todos los p-values de la prueba t de las variables no incluidas aún son mayores que α^*.

6.1.3 "Stepwise Selección" (Selección Paso a Paso)

Fue introducido por Efroymson (1960) para subsanar el problema de anidamiento de los dos métodos anteriores. Se puede considerar como una modificación del método "Forward". Es decir, se empieza con un modelo de regresión simple y en cada paso se puede añadir una variable en forma similar al método forward, pero se coteja si alguna de las variables que ya están presentes en el modelo puede ser eliminada. Aqui se usan F-out y F-in con F-in ≤ F-out.
El proceso termina cuando ninguna de las variables, que no han entrado aún, tienen importancia suficiente como para entrar al modelo.

Prácticamente todos los programas estadísticos ejecutan los procedimientos "stepwise". En R se puede usar la función regsubsets de la librería **leaps** En S-Plus existe la función **stepwise** que tiene la opción **method**, la cual permite elegir entre el método backward, forward y stepwise (Efroymson). En MINITAB se sigue la secuencia **STAT ▸ Regression ▸ Stepwise**.

Ejemplo 1: Aplicar los métodos "stepwise" al conjunto de datos **grasa**. La variable de respuesta grasa: porcentaje de grasa en el cuerpo. En 252 sujetos se midieron las siguientes variables predictoras:

edad (en años)
peso (en libras)
altura (en pulgadas)
cuello (circunferencia en cms)
pecho (circunferencia en cms)
abdomen (circunferencia en cms)
cadera (circunferencia en cms)
muslo (circunferencia en cms)
rodilla (circunferencia en cms)
tobillo (circunferencia en cms)
biceps (circunferencia en cms)
antebrazo (circunferencia en cms)
muneca (circunferencia en cms)

para predecir su porcentaje de grasa en el cuerpo

A continuacion se muestra la salida en el R para el método de selección hacia adelante. Hemos creado dos funciones **selforw** y **backelim** que ejecutan el método de selección hacia adelante y el método de eliminación hacia atrás, respectivamente. Las funciones están disponibles en la página de internet del texto. Para el primer método hemos usado un nivel de significación del 15 porciento y para el segundo un nivel de significación del 10 porciento.

```
> grasa=read.table("http://math.uprm.edu/~edgar/grasa.txt",header=T)
> selforw(grasa[,2:14],grasa[,1],.15)
Loading required package: leaps
```

Seleccion Forward

p=numero de coeficientes en el modelo, p=1 es por el intercepto
nvar=p-1=numero de variables predictoras
add.var=la variable que ha sido anadida al modelo actual
pvmax=p-value de F-parcial correspondiente a la variable mas importante en cada paso

p	nvar	add.var	pvmax	s	r2	r2adj	Cp
2	1	abdomen	0.0000	4.877	0.662	0.660	72.869
3	2	peso	0.0000	4.456	0.719	0.717	20.691
4	3	muneca	0.0047	4.393	0.728	0.724	14.210
5	4	antebrazo	0.0098	4.343	0.735	0.731	9.314
6	5	cuello	0.1000	4.328	0.738	0.733	8.559
7	6	edad	0.1098	4.314	0.741	0.734	7.973
8	7	muslo	0.1098	4.291	0.744	0.737	6.338

La variable mas importante, seleccionada en el primer paso es abdomen, la cual da un valor del coeficiente de determinación del 66.2% al hacer la regresión simple con la variable de respuesta grasa. Despues son seleccionadas en estricto orden, peso, muneca, antebrazo, cuello, edad y muslo. Las restantes variables no son escogidas porque sus "p-values" correspondientes a la prueba de F-parcial son mayores del 15% que se había elegido de antemano. El coeficiente de determinación, R^2, del modelo de regresión múltiple incluyendo las siete variables predictoras es del 74.4%. Notar también que la desviación estándar estimada del error va disminuyendo a medida que se añaden mas variables en el modelo. Las otras estadísticas R^2 ajustado y el Cp de Mallows serán explicadas mas adelante.

La siguiente tabla lista los valores de t y los p-values correspondientes para cada una las variables no incluídas en el modelo, cuando se hace la regresión considerando las variables ya incluídas y cada una de las que falta incluir.

	t-value	P-value
altura	-0.48	0.629
pecho	0.20	0.840
cadera	-1.41	0.159
rodilla	-0.01	0.991
tobillo	0.83	0.406
biceps	1.18	0.240

Notar que todas las variables no incluídas aún en el modelo tiene "P-value" grande, mayor de 0.15, lo cual indica que ellas son no significativas. En consecuencia, el proceso termina. El valor de F critico es F(1,243,.15)=2.085 y el de t crítico correspondiente es 1.44 y se puede ver que todos los t-values en valor absoluto son menores que 1.44.

El método de eliminación hacia atrás da los siguientes resultados:

```
> backelim(grasa[,2:14],grasa[,1],.10)
Eliminacion hacia atras
```

p=numero de coeficientes en el modelo
nvar=p-1=numero de variables predictoras
rem.var=la variable a ser removida, el modelo actual no incluye

esta variable
pvmin=pvalue de la F parcial correspondiente a la variable menos importante en cada paso

	p	nvar	rem.var	pvmin	s	r2	r2adj	Cp
14	14	13	rodilla	0.9497	4.296	0.749	0.736	12.004
13	13	12	pecho	0.8045	4.288	0.749	0.737	10.065
12	12	11	altura	0.4928	4.283	0.748	0.738	8.533
11	11	10	tobillo	0.3957	4.281	0.748	0.738	7.250
10	10	9	biceps	0.2888	4.282	0.747	0.738	6.367
9	9	8	cadera	0.1594	4.291	0.744	0.737	6.338

La primera de las trece predictoras que es removida del modelo es rodilla, porque es la que da el "p-value" mas grande para la prueba de F-parcial. Luego se eliminan las variables: pecho, altura, tobillo, bíceps y cadera. Las restantes siete variables se quedan en el modelo, porque sus "p-values" deben ser menores que 0.10. Mas especificamente el valor crítico de F corresponde a una $F(1,n-k-1,\alpha)=F(1,244,.10)=2.72$, aqui k=7 número de variables presentes en el modelo en el paso 7, y el correspondiente valor critico de t es 1.65. Notar que las variables que se quedan en el modelo son las mismas que son elegidas con el método de selección hacia adelante, pero que el R^2 ha bajado muy poco. Observar también que la desviación estándar estimada del error inicialmente disminuye cuando se eleiminan variables de modelo, pero al final comienza a crecer.
Si se escoge un nivel de significación del 5% entonces el proceso termina en 10 pasos y solo quedan cuatro variables en el modelo: peso, abdomen, antebrazo y muñeca.

Las salidas en S-Plus para el método de eliminación hacia atrás, método de selección hacia delante, y método "stepwise" respectivamente son como sigue:

```
> grasa.y<-grasa[,1]
> grasa.x<-grasa[,2:14]

> breg<-stepwise(grasa.x,grasa.y,method="back")
> breg
$rss:
 [1]  4411.522  4412.655  4421.330  4434.613  4455.324  4491.849  4553.520  4619.874
4658.236  4786.054  4943.245  5947.463 17578.990

$size:
 [1] 12 11 10  9  8  7  6  5  4  3  2  1  0

$which:
      edad peso altura cuello pecho abdomen cadera muslo rodilla tobillo biceps antebrazo muneca
12(- 9)    T    T      T     T     T       T     T     T       F       T     T         T      T
11(- 5)    T    T      T     T     F       T     T     T       F       T     T         T      T
10(- 3)    T    T      F     T     F       T     T     T       F       T     T         T      T
 9(-10)    T    T      F     T     F       T     T     T       F       F     T         T      T
 8(-11)    T    T      F     T     F       T     T     T       F       F     F         T      T
 7(- 7)    T    T      F     T     F       T     F     T       F       F     F         T      T
 6(- 4)    T    T      F     F     F       T     F     T       F       F     F         T      T
 5(- 8)    T    T      F     F     F       T     F     F       F       F     F         T      T
 4(- 1)    F    T      F     F     F       T     F     F       F       F     F         T      T
 3(-12)    F    T      F     F     F       T     F     F       F       F     F         F      T
```

```
 2(-13)  F  T  F  F  F  T  F  F  F  F  F  F  F
 1(- 2)  F  F  F  F  F  T  F  F  F  F  F  F  F
 0(- 6)  F  F  F  F  F  F  F  F  F  F  F  F  F

$f.stat:
 [1]  0.003988103  0.061378633  0.471848985  0.723998939  1.130246734  1.992092066
3.350023992  3.570128594  4.042716405
[10]  6.777492645  8.145201631 50.584224182 488.928083209

$method:
[1] "backward"

> freg<-stepwise(grasa.x,grasa.y,method="forw")
> freg
$rss:
 [1] 5947.463 4943.245 4786.054 4658.236 4607.169 4559.235 4491.849 4455.324 4434.613
4421.330 4412.655 4411.522 4411.448

$size:
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13

$which:
      edad peso altura cuello pecho abdomen cadera muslo rodilla tobillo biceps antebrazo muneca
 1(+ 6)    F    F      F     F     F       T     F     F       F       F     F         F      F
 2(+ 2)    F    T      F     F     F       T     F     F       F       F     F         F      F
 3(+13)    F    T      F     F     F       T     F     F       F       F     F         F      T
 4(+12)    F    T      F     F     F       T     F     F       F       F     F         T      T
 5(+ 4)    F    T      F     T     F       T     F     F       F       F     F         T      T
 6(+ 1)    T    T      F     T     F       T     F     F       F       F     F         T      T
 7(+ 8)    T    T      F     T     F       T     F     T       F       F     F         T      T
 8(+ 7)    T    T      F     T     F       T     T     T       F       F     F         T      T
 9(+11)    T    T      F     T     F       T     T     T       F       F     T         T      T
10(+10)    T    T      F     T     F       T     T     T       F       T     T         T      T
11(+ 3)    T    T      T     T     F       T     T     T       F       T     T         T      T
12(+ 5)    T    T      T     T     T       T     T     T       F       T     T         T      T
13(+ 9)    T    T      T     T     T       T     T     T       T       T     T         T      T

$f.stat:
 [1] 488.928083209  50.584224182   8.145201631   6.777492645   2.726691325   2.575843689
3.660481076   1.992092066   1.130246734
[10]  0.723998939   0.471848985   0.061378633   0.003988103

$method:
[1] "forward"

> stepreg<-stepwise(grasa.x,grasa.y,method="efroymson")
> stepreg
$rss:
[1] 5947.463 4943.245 4786.054 4658.236 4607.169 4559.235 4491.849

$size:
```

[1] 1 2 3 4 5 6 7

$which:

	edad	peso	altura	cuello	pecho	abdomen	cadera	muslo	rodilla	tobillo	biceps	antebrazo	muneca
1(+ 6)	F	F	F	F	F	T	F	F	F	F	F	F	F
2(+ 2)	F	T	F	F	F	T	F	F	F	F	F	F	F
3(+13)	F	T	F	F	F	T	F	F	F	F	F	F	T
4(+12)	F	T	F	F	F	T	F	F	F	F	F	T	T
5(+ 4)	F	T	F	T	F	T	F	F	F	F	F	T	T
6(+ 1)	T	T	F	T	F	T	F	F	F	F	F	T	T
7(+ 8)	T	T	F	T	F	T	F	T	F	F	F	T	T

$f.stat:
[1] 488.928083 50.584224 8.145202 6.777493 2.726691 2.575844 3.660481

$method:
[1] "efroymson"

Notar que S-Plus da toda la secuencia de como todas las variables son removidas en el método "backward" y de como son añadidas en el método "forward" pero no selecciona las mejores variables. Sin embargo, en el método "stepwise" (Efroymson) si se reportan las mejores variables.

En MINITAB, aplicaremos el método "stepwise", con un nivel de significación del 15% para remover una variable y del 15% para que entre una variable. Este último porcentaje puede ser menor que el nivel de significación para remover variables. Los resultados son como siguen:

Stepwise Regression: grasa versus edad, peso, ...

```
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15

 Response is  grasa   on 13 predictors, with N =  252


    Step          1        2        3        4        5        6        7
Constant     -39.28   -45.95   -27.93   -34.85   -30.65   -25.89   -33.26

abdomen        0.631    0.990    0.975    0.996    1.008    0.945    0.918
T-Value        22.11    17.45    17.37    17.76    17.89    13.82    13.21
P-Value        0.000    0.000    0.000    0.000    0.000    0.000    0.000

peso                   -0.148   -0.114   -0.136   -0.123   -0.094   -0.119
T-Value                 -7.11    -4.84    -5.48    -4.75    -2.98    -3.51
P-Value                 0.000    0.000    0.000    0.000    0.003    0.001

muneca                           -1.24    -1.51    -1.25    -1.59    -1.53
T-Value                          -2.85    -3.40    -2.66    -3.09    -3.00
P-Value                          0.005    0.001    0.008    0.002    0.003

antebraz                                   0.47     0.53     0.57     0.55
T-Value                                    2.60     2.86     3.08     2.99
P-Value                                   0.010    0.005    0.002    0.003

cuello                                             -0.37    -0.40    -0.40
```

```
T-Value                                          -1.65    -1.81    -1.83
P-Value                                          0.100    0.072    0.068

edad                                                      0.046    0.068
T-Value                                                    1.60     2.21
P-Value                                                   0.110    0.028

muslo                                                              0.22
T-Value                                                            1.91
P-Value                                                           0.057

S             4.88     4.46     4.39     4.34     4.33     4.31     4.29
R-Sq         66.17    71.88    72.77    73.50    73.79    74.06    74.45
R-Sq(adj)    66.03    71.65    72.44    73.07    73.26    73.43    73.71
C-p           72.9     20.7     14.2      9.3      8.6      8.0      6.3
```

Notar que el método "stepwise" produjo exactamente los mismos resultados que la selección hacia adelante. Usar un nivel de signficación menor del 15% para que una variable entre al modelo trae como consecuencia elegir un modelo final mas pequeño.

6.2 Método de los mejores subconjuntos

Para problemas con un número pequeño de variables predictoras k (con k menor que 8), se podrían calcular uno o dos criterios de selección para las 2^k regresiones posibles, luego se escogerían unos cuantos de estos modelos para un análisis más detallado y decidir sobre el mejor modelo. Lamentablemente hoy en día existen modelos con un gran número de variables predictoras, fácilmente se pueden encontrar problemas con más de 200 variables predictoras y ajustar 2^{200} modelos sería un trabajo computational bien pesado. Basándose en el algoritmo "Branch and Bound" (Ramificación y acotamiento) Hocking and Leslie (1967) propusieron un método para acelerar la búsqueda de los mejores subconjuntos de acuerdo a cierto criterio. Más tarde en 1974, Furnival and Wilson, propusieron un algoritmo llamado "Leaps and Bound" (Brincando y acotando) que permite elegir los mejores subconjuntos más eficientemente y este es el algoritmo adoptado por la mayoría de los programas estadísticos de computadoras.

6.3 Criterios para elegir el mejor modelo:

6.3.1 El coeficiente de Determinación R^2

La manera más básica de determinar el mejor modelo es eligiendo aquél que da un R^2 bastante alto con el menor número de variables predictoras posibles. Aparte del efecto de datos anormales que pueden afectar este criterio, hay otro problema pues un modelo con pocas variables siempre tendrá un R^2 menor o igual que un modelo que incluye un mayor número de variables, en consecuencia este criterio tendería a sugerirnos un modelo que contiene una buena cantidad de variables. Como una regla práctica se debería elegir un modelo con k variables si al incluir una variable adicional el R^2 no se incrementa sustancialmente, algo como un 5%, en términos relativos.

6.3.2 El R^2 ajustado

Para subsanar la tendencia del R^2 de elegir como mejor modelo aquel que tiene un gran número de variables predictoras, se ha definido un R^2 ajustado de la siguiente manera:

$$R^2_{ajus} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{n-1}{n-p}(1-R^2) \qquad (1)$$

donde p es el número de parámetros en el modelo. El R^2 ajustado podría disminuir al incluirse una variable adicional en el modelo. Nuevamente, el modelo que se busca es quel que tiene un R^2-ajustado alto con pocas variables.

6.3.3 La varianza estimada del error (s^2).

El mejor modelo será aquel que tenga la varianza estimada (o desviación estándar) del error más pequeña.

6.3.4 El C_p de Mallows.

La idea de este criterio, introducido por Mallows en 1973, es que el mejor modelo es áquel que no tiene ni mucha falta de ajuste ("underfitting") ni mucho sobreajuste ("overfitting") al ajustar los datos. Cuando hay falta de ajuste el estimado del valor predicho de la variable de respuesta tiene mucho sesgo y poca varianza, mientras que cuando hay "overfitting" la varianza del estimado del valor predicho es bastante alta, pero el sesgo es bajo. El cuadrado medio del error para un valor predicho sumando sobre todas las observaciones y considerando que hay p predictoras en el modelo, está dado por

$$\sum_{i=1}^{n} \frac{MSE(\hat{y}(x_i))}{\sigma^2} = \sum_{i=1}^{n} \frac{E[\hat{y}(x_i) - y(x_i)]^2}{\sigma^2} = \sum_{i=1}^{n} \frac{Var(\hat{y}(x_i)) + Sesgo^2(\hat{y}(x_i))}{\sigma^2} \qquad (2)$$

Puede ser demostrado que

$$\sum_{i=1}^{n} \frac{Var(\hat{y}(x_i))}{\sigma^2} = p \qquad (3)$$

y que

$$\sum_{i=1}^{n} \frac{Sesgo^2(\hat{y}(x_i))}{\sigma^2} = (n-p)\left(\frac{E(s_p^2) - \sigma^2}{\sigma^2}\right) \qquad (4)$$

El criterio de Mallows trata de encontrar un modelo donde tanto el sesgo como la varianza sean moderados. El estimado del lado derecho de la ecuación (2) es llamado el estadístico de Mallows y está dado por

$$C_p = p + (n-p)\frac{s_p^2}{s^2} - (n-p) = \frac{SSE_p}{s^2} - (n-2p) \qquad (5)$$

donde SSE_p es la suma de cuadrados del error del modelo que contiene p parámetros, incluyendo el intercepto, y s^2 es la varianza estimada con el modelo completo. Si un modelo con p parámetros es adecuado entonces $E(SSE_p)=(n-p)\sigma^2$. Luego, $E[SSE_p/s^2]$ es aproximadamente$(n-p)$ $\sigma^2/\sigma^2=(n-p)$. En consecuencia si el modelo fuera adecuado adecuado $E(C_p)=p$. Para decidir

acerca del valor de p se acostumbra a plotear C_p versus p. Los valores p más adecuados serán aquellos cercanos a la intersección de la gráfica con la línea C_p=p

La library **leaps** de R selecciona los mejores subconjuntos usando los criterios de R^2, R^2 ajustado y el C_p de Mallows. Aquí solo mostramos los resultados para el criterio Cp.

Ejemplo 2. Elegir los mejores subconjuntos de variables predictoras para el conjunto de datos **grasa** usando los criterios anteriores.

```
>#  El numero maximo de variables a entrar sera igual al numero de
> # predictoras del conjunto original
> maxvar<-dim(grasa)[2]
> #Mejor modelo usando Cp de mallows
> bcp<-leaps(grasa.x,grasa.y,method="Cp",nbest=1,names=nombres)
> bcp
$which
   edad  peso altura cuello pecho abdomen cadera muslo rodilla tobillo biceps
1  FALSE FALSE  FALSE  FALSE FALSE    TRUE  FALSE FALSE   FALSE   FALSE  FALSE
2  FALSE  TRUE  FALSE  FALSE  FALSE    TRUE  FALSE  FALSE   FALSE   FALSE  FALSE
3  FALSE  TRUE  FALSE  FALSE  FALSE    TRUE  FALSE  FALSE   FALSE   FALSE  FALSE
4  FALSE  TRUE  FALSE  FALSE  FALSE    TRUE  FALSE  FALSE   FALSE   FALSE  FALSE
5  FALSE  TRUE  FALSE   TRUE  FALSE    TRUE   FALSE FALSE   FALSE   FALSE  FALSE
6   TRUE  TRUE  FALSE   FALSE FALSE    TRUE   FALSE  TRUE    FALSE   FALSE  FALSE
7   TRUE  TRUE  FALSE   TRUE  FALSE    TRUE   FALSE  TRUE   FALSE   FALSE  FALSE
8   TRUE  TRUE  FALSE   TRUE  FALSE    TRUE   TRUE   TRUE   FALSE   FALSE  FALSE
9   TRUE  TRUE  FALSE   TRUE  FALSE    TRUE   TRUE   TRUE   FALSE   FALSE   TRUE
10  TRUE  TRUE  FALSE   TRUE  FALSE   TRUE   TRUE   TRUE   FALSE   TRUE   TRUE
11  TRUE  TRUE   TRUE   TRUE  FALSE   TRUE   TRUE   TRUE   FALSE   TRUE   TRUE
12  TRUE  TRUE   TRUE   TRUE  TRUE    TRUE   TRUE   TRUE   FALSE   TRUE   TRUE
13  TRUE  TRUE   TRUE   TRUE  TRUE    TRUE   TRUE   TRUE   TRUE    TRUE   TRUE
   antebrazo muneca
1     FALSE  FALSE
2     FALSE  FALSE
3     FALSE   TRUE
4      TRUE   TRUE
5      TRUE   TRUE
6      TRUE   TRUE
7      TRUE   TRUE
8      TRUE   TRUE
9      TRUE   TRUE
10     TRUE   TRUE
11     TRUE   TRUE
12     TRUE   TRUE
13     TRUE   TRUE

$label
 [1] "(Intercept)" "edad"        "peso"        "altura"      "cuello"
 [6] "pecho"       "abdomen"     "cadera"      "muslo"       "rodilla"
[11] "tobillo"     "biceps"      "antebrazo"   "muneca"

$size
 [1]  2  3  4  5  6  7  8  9 10 11 12 13 14
```

```
$Cp
 [1] 72.868837 20.690746 14.210205  9.314331  8.559272  7.664855  6.337654
 [8]  6.367146  7.249744  8.533156 10.065111 12.003988 14.000000

> p<-2:maxvar
> plot(p,bcp$Cp,type="l")
> title("Grafica del Cp de Mallows segun el tamano del modelo")
> lines(2:maxvar,2:maxvar)
```

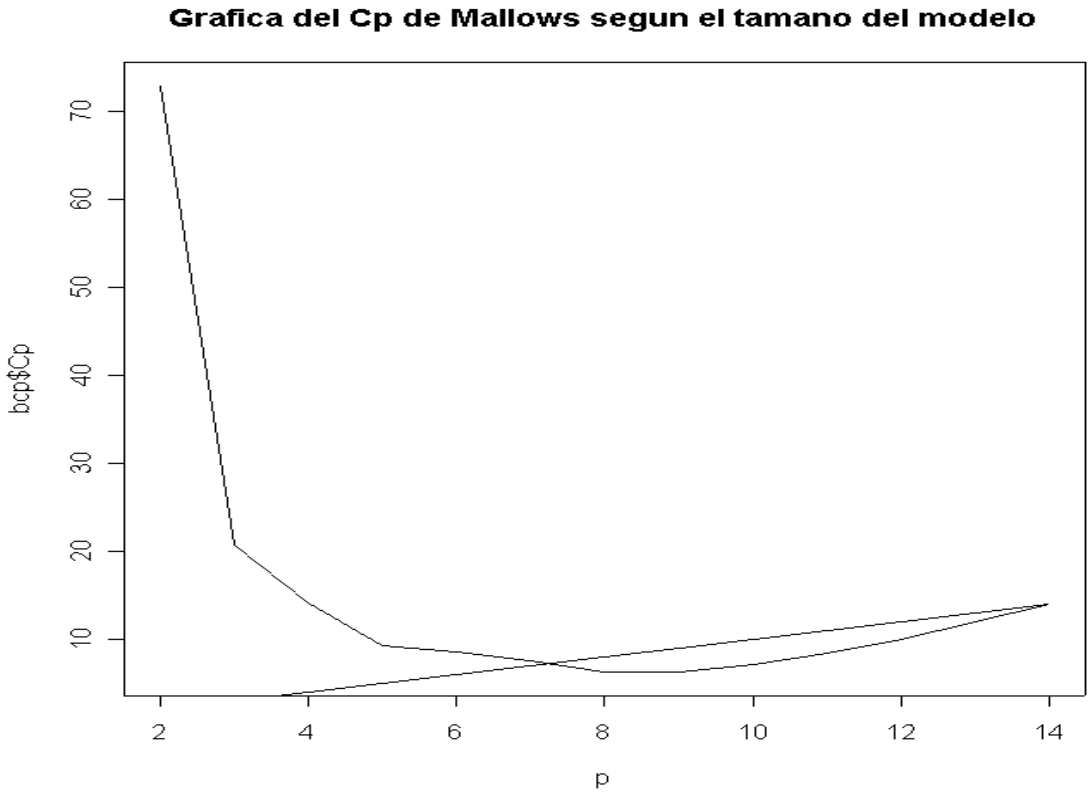

Figura 6.1 Plot del Cp de Mallows para determinar el número óptimo de variables predictoras.

Notar que la curva y la línea se intersectan alrededor de p=7.

MINITAB permite seleccionar los mejores subconjuntos basados en los criterios mencionados anteriormente. Se debe usar la secuencia **STAT ▸ Regression ▸ Best Subsets**.

Best Subsets Regression: grasa versus edad, peso, ...

```
Response is grasa

                                                  a
                                              r t   n
                                 a c   a c    o o b t m
                                 l u p d a m  d b i e u
                             e p t e e o d u  i i c b n
```

Vars	R-Sq	R-Sq(adj)	C-p	S	d a	e s	u r	l l	c h	m e	e r	s l	l l	l o	e s	r z	e a
1	66.2	66.0	72.9	4.8775						X							
1	49.4	49.2	232.2	5.9668					X								
2	71.9	71.7	20.7	4.4556		X				X							
2	70.2	70.0	36.6	4.5866						X							X
3	72.8	72.4	14.2	4.3930		X				X							X
3	72.4	72.0	18.0	4.4251		X		X		X							
4	73.5	73.1	9.3	4.3427		X				X						X	X
4	73.3	72.8	11.4	4.3609		X				X					X		X
5	73.8	73.3	8.6	4.3276		X		X		X						X	X
5	73.7	73.2	9.2	4.3336	X	X				X						X	X
6	74.1	73.5	7.7	4.3111	X	X				X		X				X	X
6	74.1	73.4	8.0	4.3138	X	X		X		X						X	X
7	74.4	73.7	6.3	4.2906	X	X		X		X		X				X	X
7	74.3	73.6	7.4	4.2998	X	X		X		X					X	X	X
8	74.7	73.8	6.4	4.2819	X	X		X		X	X	X				X	X
8	74.6	73.8	7.0	4.2872	X	X		X		X		X			X	X	X
9	74.8	73.8	7.2	4.2808	X	X		X		X	X	X			X	X	X
9	74.7	73.8	7.7	4.2851	X	X	X	X		X	X	X				X	X
10	74.8	73.8	8.5	4.2832	X	X		X		X	X	X		X	X	X	X
10	74.8	73.8	8.7	4.2850	X	X	X	X		X	X	X			X	X	X
11	74.9	73.7	10.1	4.2879	X	X	X	X		X	X	X		X	X	X	X
11	74.8	73.7	10.5	4.2920	X	X		X	X	X	X	X		X	X	X	X
12	74.9	73.6	12.0	4.2963	X	X	X	X	X	X	X	X		X	X	X	X
12	74.9	73.6	12.1	4.2968	X	X	X	X		X	X	X	X	X	X	X	X
13	74.9	73.5	14.0	4.3053	X	X	X	X	X	X	X	X	X	X	X	X	X

Lo que se muestra aquí son los dos mejores subcojuntos de variables para cada número de variables predictoras, excepto cuando se tiene el modelo que incluye todas las variables. De acuerdo al R^2 y R^2 ajustado el mejor modelo sería aquel que incluye solo dos variables predictoras: peso y abdomen. De acuerdo al C_p de Mallows se escogería el modelo que incluye 6 variables predictoras: Edad, peso, abdomen, muslo, antebrazo y muñeca. El C_p es de 7.7.

De acuerdo a la varianza estimada del error se escogería el modelo que incluye 4 variables predictoras: peso, abdomen, antebrazo y muñeca.

6.3.5 PRESS (Suma de cuadrados de Predicción)

El criterio suma de cuadrados de Predicción [PRESS], introducido por Allen en 1974, es una combinación de todas las regresiones posibles, análisis de residuales y "leave-one-out" (ver más adelante validación cruzada).
Supongamos que hay p parámetros en el modelo y que tenemos n observaciones disponibles para estimar los parámetros. En cada paso se deja de lado la i-ésima observacion del conjunto de datos y se calculan todas las regresiones posibles (más eficientemente se podrían calcular solamente los mejores subconjuntos de regresión que resultan de aplicar algún criterio, tal como el C_p de Mallows). Luego se calcula la predicción $\hat{y}_{(i)}$ para la observación que no fue incluida y se calcula el residual correpondiente $e_{(i)} = y_i - \hat{y}_{(i)}$, el cual es llamado el residual PRESS. Ya se vió en la sección 3.1.4 que la relación entre el residual PRESS y el residual usual $\hat{e}_i$ es

$$e_{(i)} = \frac{\hat{e}_i}{1 - h_{ii}} \qquad (6)$$

donde los h_{ii} representan los elementos de la diagonal de la matriz **H=X(X'X)$^{-1}$X.**
Si la diferencia entre el residual PRESS y el residual usual de una observación es bien grande entonces se considera que dicha observación es influencial.
La medida PRESS para el modelo de regresión que contiene p parámetros se define por:

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 \qquad (7)$$

Otra forma, equivalente de cálculo sería

$$PRESS = \sum_{i=1}^{n} \left(\frac{\hat{e}_i}{1-h_{ii}}\right)^2 \qquad (8)$$

Según el criterio PRESS el mejor modelo será aquel que tenga el valor de PRESS más bajo. Algunas veces se acostumbra usar un PRESS promedio, que es simplememente el PRESS dividido por el número de observaciones del conjunto de datos. No existe una function en R que calcule el PRESS pero ésta puede ser programada fácilmente. MINITAB calcula el PRESS cuando se hace regresión y cuando se hace el "stepwise".

Ejemplo 3. Calcular el PRESS para los mejores modelos según los criterios discutidos anteriormente del conjunto de datos grasa

```
> grasa=read.table("http://math.uprm.edu/~edgar/grasa.txt",header=T)
> PRESS=function (x)
{#x es un objecto que sale de aplicar lm
    sum(resid(x)^2/(1 - lm.influence(x)$hat)^2)
}
> lm1=lm(grasa~peso+abdomen, data=grasa)
> PRESS(lm1)
[1] 5109.1
> lm2=lm(grasa~peso+abdomen+antebrazo+muneca, data=grasa)
> PRESS(lm2)
[1] 4908.053
> lm3=lm(grasa~peso+abdomen+antebrazo+muneca+edad+muslo, data=grasa)
> PRESS(lm3)
[1] 4877.671
> lm4=lm(grasa~peso+abdomen+antebrazo+muneca+edad+muslo+cuello, data=grasa)
[1] 4840.64
```

Si buscamos un modelo parsimonioso, sería mejor elegir aquel que incluye 4 variables ya que si bien tiene un PRESS mayor que el de 6 variables, la diferencia no es mucha. Esta selección coincide con la que da el método de eliminación hacia atrás al nivel de significación del 5%.

6.3.6 Validación Cruzada (CV)

Fue introducido por Stone en 1974. La idea aquí es estimar el error de predicción dividiendo al azar el conjunto de datos en varias partes. En cada paso una de las partes se convierte en una muestra de prueba que sirve para validar el modelo y las restantes partes constituyen lo que es llamado una muestra de entrenamiento que sirve para construir el modelo. Por lo general se usan 10 partes y eso es llamado una "10 fold cross-validation" , ó n partes y en ese caso es llamado el método "leave-one-out"(dejar uno afiera). Este ultimo se relaciona bastante con el PRESS. El cálculo del error por validación cruzada usando K partes estará dado por:

$$CV = \frac{\sum_{i=1}^{K}\sum_{j=1}^{N_i}(y_j - \hat{y}_j^{(-i)})^2}{n}$$

donde $\hat{y}_j^{(-i)}$ representa el valor predicho para la j-ésima observación de la parte N_i usando una línea de regresión que ha sido estimada sin haber usado las observaciones de dicha parte.
La idea es escoger el mejor modelo como aquel que tiene el más error de validación cruzada promedio más pequeño. En el caso de ":leave-one-out" el error de predicción promedio es PRESS/n.
El cálculo de validación cruzada para regresión no está disponible en ninguno de los programas estadísticos usados en este texto. Se debe tiene que escribir un programa para obtenerlo, o usar solamente el método "leave-one-out". Nosotros hemos escrito la función **CV10reg** que estima el error promedio de predicción usando validación cruzada 10.

Ejemplo 4. Aplicar la función CV10reg al conjunto grasa.

```
> #Leyendo el conjunto de datos pero excluyendo los nombres de las columnas
> grasa<-read.table(file="c:/grasa.txt",header=F,skip=1)
> dim(grasa)
[1] 252  14
> #Estimando el error promedio de prediccion con todas las predictoras
> CV10reg(grasa,10)
Los estimados del error promedio de prediccion en cada repeticion son
 [1] 20.51001 19.91260 20.20943 19.94021 20.48995 20.54480 20.29945 20.40238
 [9] 19.98185 20.60561
El estimado del error promedio de prediccion por VC con el numero de repeticiones dado es
[1] 20.28963
> # Estimando el error promedio de prediccion usando: peso, abdomen
> CV10reg(grasa[,c(1,3,7)],10)
Los estimados del error promedio de prediccion en cada repeticion son
 [1] 20.21294 20.15848 20.43202 20.43359 20.31295 20.27680 20.15327 20.26616
 [9] 20.30443 20.24162
El estimado del error promedio de prediccion por VC con el numero de repeticiones dado es
[1] 20.27923
> # Estimando el error promedio de prediccion usando: peso, abdomen, antebrazo y muneca.
> CV10reg(grasa[,c(1,3,7,13,14)],10)
Los estimados del error promedio de prediccion en cada repeticion son
 [1] 19.26377 19.48586 19.63418 19.66251 19.63294 19.35995 19.40967 19.58761
 [9] 20.51080 19.54337
El estimado del error promedio de prediccion por VC con el numero de repeticiones dado es
[1] 19.60907
> # Estimando el error promedio de prediccion usando: edad, peso, abdomen, muslo, antebrazo y
muneca.
>CV10reg(grasa[,c(1:3,7,9,13,14)],10)
Los estimados del error promedio de prediccion en cada repeticion son
 [1] 19.61473 19.29554 19.23239 19.69176 19.52733 19.36840 19.66784 19.21728
 [9] 19.24110 19.15240
El estimado del error promedio de prediccion por VC con el numero de repeticiones dado es
[1] 19.40088
```

```
>
```

De los resultados previos el criterio CV nos sugiere que el mejor modelo es aquel que incluye las siguientes seis variables; edad, peso, abdomen, muslo, antebrazo y muñeca.

6.3.7 AIC

El criterio de información de Akaike (Akaike, 1973), tiene su origen en conceptos de teoría de información y está basado en la minimización de la distancia Kullback-Leibler entre la distribución de la variable de respuesta Y bajo el modelo reducido y bajo el modelo completo. Se define como,

$$AIC= -2*\text{máximo de la log likelihhod} +2p \qquad (9)$$

Donde p es el número de parámetros dle modelo. En particular para el caso de regresión, asumiendo que la varianza de las y's es estimada por SSE/n, la fórmula anterior se reduce a:

$$AIC=n\log[SSE_p/n]+2p \qquad (10)$$

Existen otras variantes a la formula (10). Un buen modelo es aquel con bajo AIC.

MINITAB no da el AIC, pero si aparece en SAS y S-Plus (aunque la versión que calculan es $AIC=[SSE_p/s^2] +2p$. Tanto en R como en S-Plus están disponibles la funciones step y stepAIC (de la librería MASS) que calcula el mejor modelo por el método "stepwise" basado en el criterio AIC.

Ejemplo 5. Seleccionar el mejor modelo de regresión para el conjunto grasa usando el criterio AIC usando los métodos "forward" y "backward".

```
># Metodo "backward"
># Primero hay que hallar la regresion con todas las variables predictoras
> #Hallando el mejor subconjunto usando stepwise y el criterio AIC
> l1<-lm(grasa~.,data=grasa)
> step(l1,scope=~.,direction="backward")
Start:  AIC= 749.36
 grasa ~ edad + peso + altura + cuello + pecho + abdomen + cadera +
    muslo + rodilla + tobillo + biceps + antebrazo + muneca

            Df Sum of Sq    RSS    AIC
- rodilla    1       0.1 4411.5  747.4
- pecho      1       1.1 4412.5  747.4
- altura     1       9.7 4421.2  747.9
- tobillo    1      11.4 4422.9  748.0
- biceps     1      20.9 4432.3  748.5
<none>                   4411.4  749.4
- cadera     1      37.5 4448.9  749.5
- muslo      1      49.6 4461.0  750.2
- peso       1      50.6 4462.1  750.2
- edad       1      68.3 4479.7  751.2
- cuello     1      76.0 4487.4  751.7
- antebrazo  1      95.5 4507.0  752.8
```

```
- muneca     1     170.1 4581.6  756.9
- abdomen    1    2261.0 6672.4  851.6

Step:  AIC= 747.36
 grasa ~ edad + peso + altura + cuello + pecho + abdomen + cadera +
    muslo + tobillo + biceps + antebrazo + muneca

          Df Sum of Sq    RSS    AIC
- pecho      1       1.1 4412.7  745.4
- altura     1       9.7 4421.2  745.9
- tobillo    1      12.1 4423.6  746.1
- biceps     1      20.8 4432.3  746.5
<none>                   4411.5  747.4
- cadera     1      37.4 4448.9  747.5
- peso       1      53.1 4464.6  748.4
- muslo      1      54.9 4466.4  748.5
- edad       1      74.1 4485.6  749.6
- cuello     1      78.4 4490.0  749.8
- antebrazo  1      96.8 4508.3  750.8
- muneca     1     170.5 4582.1  754.9
- abdomen    1    2269.9 6681.4  850.0

Step:  AIC= 745.43
 grasa ~ edad + peso + altura + cuello + abdomen + cadera + muslo +
    tobillo + biceps + antebrazo + muneca

          Df Sum of Sq    RSS    AIC
- altura     1       8.7 4421.3  743.9
- tobillo    1      12.4 4425.1  744.1
- biceps     1      20.1 4432.8  744.6
<none>                   4412.7  745.4
- cadera     1      36.3 4449.0  745.5
- muslo      1      60.1 4472.7  746.8
- peso       1      70.8 4483.5  747.4
- edad       1      73.8 4486.5  747.6
- cuello     1      79.5 4492.1  747.9
- antebrazo  1      95.6 4508.3  748.8
- muneca     1     170.0 4582.6  753.0
- abdomen    1    2879.4 7292.1  870.0

Step:  AIC= 743.92
 grasa ~ edad + peso + cuello + abdomen + cadera + muslo + tobillo +
    biceps + antebrazo + muneca

          Df Sum of Sq    RSS    AIC
- tobillo    1      13.3 4434.6  742.7
- biceps     1      22.4 4443.7  743.2
- cadera     1      30.4 4451.8  743.6
<none>                   4421.3  743.9
- muslo      1      68.8 4490.1  745.8
- cuello     1      77.1 4498.4  746.3
```

```
- edad       1      81.3 4502.6  746.5
- antebrazo  1      98.1 4519.4  747.5
- peso       1     119.6 4540.9  748.6
- muneca     1     181.3 4602.6  752.0
- abdomen    1    3178.5 7599.9  878.4

Step:  AIC= 742.68
 grasa ~ edad + peso + cuello + abdomen + cadera + muslo + biceps +
    antebrazo + muneca

          Df Sum of Sq    RSS    AIC
- biceps     1      20.7 4455.3  741.9
- cadera     1      31.7 4466.4  742.5
<none>                   4434.6  742.7
- muslo      1      72.3 4506.9  744.8
- edad       1      77.6 4512.2  745.1
- cuello     1      87.3 4521.9  745.6
- antebrazo  1      97.4 4532.0  746.2
- peso       1     107.2 4541.8  746.7
- muneca     1     168.0 4602.6  750.0
- abdomen    1    3182.0 7616.7  877.0

Step:  AIC= 741.85
 grasa ~ edad + peso + cuello + abdomen + cadera + muslo + antebrazo +
    muneca

          Df Sum of Sq    RSS    AIC
<none>                   4455.3  741.9
- cadera     1      36.5 4491.8  741.9
- cuello     1      79.1 4534.4  744.3
- edad       1      83.8 4539.1  744.5
- peso       1      93.0 4548.3  745.1
- muslo      1     100.7 4556.0  745.5
- antebrazo  1     140.5 4595.8  747.7
- muneca     1     166.8 4622.2  749.1
- abdomen    1    3163.0 7618.3  875.0

Call:
lm(formula = grasa ~ edad + peso + cuello + abdomen + cadera +     muslo + antebrazo +
muneca, data = grasa)

Coefficients:
(Intercept)         edad         peso       cuello      abdomen       cadera
  -22.65637      0.06578     -0.08985     -0.46656      0.94482     -0.19543
     muslo     antebrazo       muneca
    0.30239      0.51572     -1.53665
```

Las variables eliminadadas son: rodilla, pecho, altura, tobillo, y bíceps en ese orden. El mejor modelo según el método “backward” y usando el criterio AIC es el considera las siguientes 8 variables predictoras: edad, peso, cuello, abdomen , cadera, muslo, antebrazo, y muñeca.

```
>#Metodo “Forward”
> #Hallando primero la regresion con la variable predictora mas correlacionada V7
> l2=lm(grasa~abdomen,data=grasa)
>
step(l2,scope=~.+edad+peso+altura+cuello+pecho+cadera+muslo+rodilla+tobillo+biceps+antebr
azo+muneca,direction="forward")
Start:  AIC= 800.65
 grasa ~ abdomen

           Df Sum of Sq    RSS    AIC
+ peso      1    1004.2 4943.2  756.0
+ muneca    1     709.2 5238.3  770.6
+ cuello    1     614.5 5332.9  775.2
+ cadera    1     548.2 5399.2  778.3
+ altura    1     458.8 5488.7  782.4
+ rodilla   1     318.7 5628.8  788.8
+ tobillo   1     233.3 5714.1  792.6
+ edad      1     200.9 5746.5  794.0
+ pecho     1     195.5 5752.0  794.2
+ muslo     1     174.6 5772.9  795.1
+ biceps    1     135.3 5812.2  796.8
+ antebrazo 1      54.3 5893.2  800.3
<none>                 5947.5  800.6

Step:  AIC= 756.04
 grasa ~ abdomen + peso

           Df Sum of Sq    RSS    AIC
+ muneca    1     157.2 4786.1  749.9
+ cuello    1      86.9 4856.3  753.6
+ muslo     1      81.4 4861.9  753.9
+ antebrazo 1      66.9 4876.4  754.6
+ biceps    1      63.8 4879.4  754.8
+ altura    1      40.3 4903.0  756.0
<none>                 4943.2  756.0
+ rodilla   1       9.7 4933.5  757.5
+ edad      1       1.9 4941.3  757.9
+ tobillo   1       1.5 4941.7  758.0
+ pecho     1   0.01017 4943.2  758.0
+ cadera    1   0.00529 4943.2  758.0

Step:  AIC= 749.9
 grasa ~ abdomen + peso + muneca

           Df Sum of Sq    RSS    AIC
+ antebrazo 1     127.8 4658.2  745.1
+ biceps    1      88.7 4697.3  747.2
+ muslo     1      40.5 4745.6  749.8
<none>                 4786.1  749.9
+ cuello    1      25.2 4760.9  750.6
+ altura    1      23.4 4762.6  750.7
```

```
+ edad       1     21.2 4764.9  750.8
+ rodilla    1     20.5 4765.5  750.8
+ tobillo    1     15.0 4771.1  751.1
+ cadera     1      9.2 4776.8  751.4
+ pecho      1      1.3 4784.8  751.8

Step:  AIC= 745.07
 grasa ~ abdomen + peso + muneca + antebrazo

         Df Sum of Sq    RSS    AIC
+ cuello  1     51.1 4607.2  744.3
+ edad    1     38.4 4619.9  745.0
<none>               4658.2  745.1
+ biceps  1     33.9 4624.4  745.2
+ muslo   1     27.2 4631.0  745.6
+ rodilla 1     19.8 4638.4  746.0
+ tobillo 1     18.2 4640.1  746.1
+ altura  1     18.0 4640.2  746.1
+ cadera  1      3.5 4654.7  746.9
+ pecho   1      0.5 4657.7  747.0

Step:  AIC= 744.3
 grasa ~ abdomen + peso + muneca + antebrazo + cuello

         Df Sum of Sq    RSS    AIC
+ edad    1     47.9 4559.2  743.7
+ biceps  1     45.9 4561.2  743.8
<none>               4607.2  744.3
+ muslo   1     25.1 4582.1  744.9
+ altura  1     18.9 4588.3  745.3
+ cadera  1     11.0 4596.2  745.7
+ tobillo 1     10.7 4596.5  745.7
+ rodilla 1     10.4 4596.8  745.7
+ pecho   1  0.009572 4607.2  746.3

Step:  AIC= 743.66
 grasa ~ abdomen + peso + muneca + antebrazo + cuello + edad

         Df Sum of Sq    RSS    AIC
+ muslo   1     67.4 4491.8  741.9
+ biceps  1     48.1 4511.1  743.0
<none>               4559.2  743.7
+ altura  1     19.0 4540.3  744.6
+ tobillo 1     14.8 4544.5  744.8
+ rodilla 1      6.6 4552.7  745.3
+ cadera  1      3.2 4556.0  745.5
+ pecho   1      0.8 4558.4  745.6

Step:  AIC= 741.91
 grasa ~ abdomen + peso + muneca + antebrazo + cuello + edad +
    muslo
```

```
         Df Sum of Sq    RSS    AIC
+ cadera   1      36.5 4455.3  741.9
<none>                 4491.8  741.9
+ biceps   1      25.5 4466.4  742.5
+ tobillo  1      12.8 4479.1  743.2
+ altura   1       4.3 4487.5  743.7
+ pecho    1       0.8 4491.1  743.9
+ rodilla  1  0.002584 4491.8  743.9

Step:  AIC= 741.85
 grasa ~ abdomen + peso + muneca + antebrazo + cuello + edad +
    muslo + cadera

         Df Sum of Sq    RSS    AIC
<none>                 4455.3  741.9
+ biceps   1      20.7 4434.6  742.7
+ altura   1      11.7 4443.6  743.2
+ tobillo  1      11.6 4443.7  743.2
+ rodilla  1 3.651e-02 4455.3  743.8
+ pecho    1 9.904e-05 4455.3  743.9

Call:
lm(formula = grasa ~ abdomen + peso + muneca + antebrazo + cuello +      edad + muslo +
cadera, data = grasa)

Coefficients:
(Intercept)      abdomen         peso       muneca    antebrazo       cuello         edad
  -22.65637      0.94482     -0.08985     -1.53665      0.51572     -0.46656      0.06578
      muslo       cadera
    0.30239     -0.19543
```

Las variables que son seleccionadas en cada paso son: abdomen, peso, muñeca, antebrazo, cuello, edad, muslo y cadera en ese orden. El mejor modelo obtenido con el método “forward” es el mismo que se obtuvo con el método “backward”.

6.3.8 BIC

Este criterio fue introducido por Schwarz en 1978 y está basado en argumentos bayesianos. Se define por

$$BIC=nlog[SSE_p/n]+plog(n)$$

Notar que el BIC se diferencia del AIC solo en el ultimo término. Los criterios AIC y C_p de Mallows tienden a dar modelos óptimos más grandes que el criterio BIC. R y S-Plus dan una versión modificada del BIC. MINITAB no da el BIC, pero si aparece en SAS (ldonde es llamado SBC) .

Ejemplo 6: La siguiente salida en R muestra los valores de AIC y BIC, según las formulas dadas y siguiendo un método “forward”.

```
> forwabic(grasa[,2:14],grasa[,1],.15)
Seleccion Forward

p=numero de coeficientes en el modelo, p=1 es por el  intercepto
nvar=p-1=numero de variables  predictoras
add.var=la variable que ha sido anadida al modelo actual
pvmax=p-value de F-parcial correspondiente a la variable mas importante en cada paso

  p nvar   add.var  pvmax     aic     bic
2 2    1   abdomen 0.0000  800.645  807.704
3 3    2      peso 0.0000  756.040  766.628
4 4    3    muneca 0.0047  749.896  764.014
5 5    4 antebrazo 0.0098  745.075  762.722
6 6    5    cuello 0.1000  744.297  765.473
7 7    6     edad  0.1098  743.661  768.367
8 8    7     muslo 0.1098  741.909  770.144
>
```

El modelo elegido con el criterio AIC sería el mismo que fue elegido con el clásico método "forward". El modelo elegido con el criterio BIC sería aquel que consider a las predictoras: abdomen, peso, muñeca y antebrazo.

Ejemplo 7. A continuación se muestran los 15 mejores modelos para el conjunto de datos **grasa**, ordenados de acuerdo al Cp de Mallows, mostrando además los valores del AIC, BIC y SBC. Los resultados fueron obtenidos usando SAS version 8.

Number in Model	C(p)	R-Square	AIC	BIC	SBC	Variables in Model
**7	6.3377	0.7445	741.9088	744.5436	770.14425	edad peso cuello abdomen muslo antebraz muneca
8	6.3671	0.7466	741.8514	744.7178	773.61622	edad peso cuello abdomen cadera muslo antebraz muneca
8	6.9626	0.7459	742.4748	745.2951	774.23968	edad peso cuello abdomen muslo biceps antebraz muneca
9	7.2497	0.7477	742.6772	745.7373	777.97144	edad peso cuello abdomen cadera muslo biceps antebraz muneca
7	7.3761	0.7434	742.9864	745.5508	771.22181	edad peso cuello abdomen biceps antebraz muneca
8	7.6488	0.7452	743.1915	745.9589	774.95637	edad peso cuello abdomen muslo tobillo antebraz muneca
*6	7.6649	0.7410	743.3451	745.7048	768.05114	edad peso abdomen muslo antebraz muneca
9	7.7333	0.7472	743.1859	746.2041	778.48021	edad peso altura cuello abdomen cadera muslo antebraz muneca
9	7.7402	0.7472	743.1933	746.2108	778.48755	edad peso cuello abdomen cadera muslo tobillo antebraz muneca
6	7.9732	0.7406	743.6612	746.0031	768.36724	edad peso cuello abdomen antebraz muneca
6	8.0815	0.7405	743.7721	746.1077	768.47815	peso cuello abdomen biceps

						antebraz muneca
8	8.1043	0.7447	743.6660	746.3984	775.43091	edad peso altura cuello abdomen muslo antebraz muneca
9	8.1748	0.7468	743.6497	746.6295	778.94397	edad peso cuello abdomen muslo tobillo biceps antebraz muneca
8	8.2969	0.7445	743.8664	746.5841	775.63130	edad peso cuello pecho abdomen muslo antebraz muneca
8	8.3375	0.7445	743.9087	746.6232	775.67353	edad peso cuello abdomen muslo rodilla antebraz muneca

(*) sería el mejor modelo con el C_p de Mallows, porque aún cuando no es el modelo con el menor C_p , es un error bastante común elegir como el mejor modelo aquel con el menor C_p, su valor 7.66 está cerca a 6+1=7 (número de parámteros del modelo).

(**) sería el mejor modelo de acuerdo al AIC, porque aunque es el modelo con el segundo AIC más pequeño, se está eligiendo solo 7 variables.

6.3.9 Validación Cruzada Generalizada (CGV)

Este criterio fue introducido en 1979, por Golub, Heath and Whaba El cálculo de validación cruzada "leave-one out" es computacionalmente pesado y el GCV es una aproximación al "leave-one-out", que puede ser calculado más rápidamente.

Se define por

$$GCV = \frac{nSSE_p}{[n - tr(H_p)]^2} = \frac{nSSEp}{(n-p)^2}$$

donde H_p es la matriz HAT para el modelo que incluye a p variables. Luego, $tr(H_p)=p$. El modelo óptimo será aquel que incluye las p variables predictoras que hacen que GCV sea mínimo. El cálculo del GCV puede ser fácilmente programable en R o S-Plus. SAS da esta medida pero en su procedimiento para Regresión Noparamétrica.

Los resultados de GCV para el ejemplo grasa son como siguen:

```
> forwabic(grasa[,2:14],grasa[,1],.15)
Loading required package: leaps
Seleccion Forward

p=numero de coeficientes en el modelo, p=1 es por el intercepto
nvar=p-1=numero de variables predictoras
add.var=la variable que ha sido anadida al modelo actual
pvmax=p-value de F-parcial correspondiente a la variable mas importante en cada paso

  p nvar   add.var   pvmax     aic     bic         gcv
2 2    1   abdomen 0.0000 800.645  807.704  23.9802
```

```
3 3    2      peso     0.0000 756.040  766.628  20.0916
4 4    3    muneca     0.0047 749.896  764.014  19.6099
5 5    4 antebrazo     0.0098 745.075  762.722  19.2410
6 6    5    cuello     0.1000 744.297  765.473  19.1851
7 7    6      edad     0.1098 743.661  768.367  19.1408
8 8    7     muslo     0.1098 741.909  770.144  19.0128
>
```

6.3.10 Otros Criterios

Recientemente se han introducido muchos otros criterios para selección de variables en regresión entre los más conocidos están:

MDL: Longitud de descripción Mínima (Rissanen, 1978).
RIC: Criterio de Inflación del Riesgo (Foster y George, 1994)
CIC: Criterio de Inflación del Covarianza (Tibshirani and Knight, 1999)

Para más detalles acerca estos métodos veáse el texto "Subset selection in regression" por Alan Miller (2002).

6.3.11 Recomendaciones para elegir el mejor modelo

En cualquier problema las variables predictoras pueden ser clasificadas en 3 grupos:
a) Las que son importantes.
b) Las que uno no está seguro de su importancia. Es decir, podrian ser redundantes.
c) Las que no son relevantes para explicar el comportamiento de la variable de respuesta.

Lo que se recomienda es eliminar las variables tipo c) eligiendo un buen subconjunto de variables predictoras usando para ello los criterios C_p, AIC o BIC y luego aplicar "stepwise" para descartar las variables tipo b) y quedarnos con las variables tipo a) que son las que nos interesantes. Aplicando esta metodología a nuestro conjunto de datos grasa, vamos a considerar que las relevantes variables que se eligieron con el Cp son: edad, peso, abdomen, muslo, antebrazo y muñeca. Luego, aplicaremos el clásico "stepwise" con estas variables predictoras solamente. Usando MINITAB, obtenemos

Stepwise Regression: grasa versus edad, peso, ...

```
 Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15

Response is grasa on 6 predictors, with N = 252

Step            1       2       3       4
Constant    -39.28  -45.95  -27.93  -34.85

abdomen      0.631   0.990   0.975   0.996
T-Value      22.11   17.45   17.37   17.76
P-Value      0.000   0.000   0.000   0.000

peso                -0.148  -0.114  -0.136
```

```
T-Value              -7.11   -4.84   -5.48
P-Value              0.000   0.000   0.000

muneca                       -1.24   -1.51
T-Value                      -2.85   -3.40
P-Value                      0.005   0.001

antebrazo                             0.47
T-Value                               2.60
P-Value                              0.010

S            4.88    4.46    4.39    4.34
R-Sq        66.17   71.88   72.77   73.50
R-Sq(adj)   66.03   71.65   72.44   73.07
Mallows Cp   72.0    20.0    13.5     8.6
```

Solo quedan: abdomen, peso, muñeca y antebrazo como las variables predictoras más importantes. Así que las variables edad y muslo parecen ser redundantes, esto se explica porque estas variables están correlacionadas con otras predictoras. Este problema sera discutido en el próximo capítulo.

6.4 Otros métodos de Selección de variables

Existen muchos otros métodos de selección de variables en regresión, solo mencionaremos cuatro de ellos.

6.4.1. Método basados en remuestreo.

Bootstrapping (Efron, 1983)
El pequeño Boostrapping (Breiman, 1992)

6.4.2. Métodos basados en Regresión Penalizada

La idea de estos métodos es estimar la importancia de cada uno de las variables predictoras y luego considerar solamente en el modelo aquellas que no tienen una relevancia despreciable. Se relaciona con la metodología de regression ridge que se verá en el próximo capítulo. Dos de los métodos más usados son:

La Garrote (Breiman, 1995)
El Lasso (Tibshirani, 1996)

6.4.3 Métodos Bayesianos

Es considerado en gran detalle por Mitchel y Beauchamp (JASA, 1988). Supongamos que ya se tiene un conjunto de buenos modelos. La idea se basa en asignar probabilidades a priori a los coeficientes de cada uno de estos modelos que incluyen solo un subconjunto de predictoras e igualmente se asignan probabilidades a priori a cada uno de los modelos. Finalmente, se elige como mejor modelo aquel que tiene la probabilidad posterior más alta con respecto a la variable de respuesta.

6.4.4. Algoritmo Genéticos: En este caso el problema de selección de variables es considerado como un problema de optimización con respecto al número de variables predictoras que deben incluirse en el modelo. Luego, el problema de optimización es resuelto usando algoritmos Genéticos.

Ejercicios

1. Supongamos que se desea omitir la variable predictora Xj de un modelo de regresión con p parámetros (incluyendo el intercepto) y n observaciones. Si F_j es el estadístico para probar la hipótesis H: βj=0 demostrar que:

$$C_{p-1} = \frac{F_j SSE_p}{s^2(n-p)} + C_p - 2$$

2. Hacer selección de variables predictoras usando el conjunto de datos
Berkeley: La variable de respuesta es SOMA y las predictoras son WT2, HT2, WT9, HT9, LG9, ST9. Disponible en la página de internet del texto.

a) Los metodos "stepwise". Explicar los pasos de los procesos y justificar la terminación del mismo.
b) Los mejores usando los mejores subcojuntos con por lo menos 6 criterios. Explicar los resultados
c) Comparar los resultados obtenidos en a y b. Dar su seleccion final.

3. Hacer selección de variables predictoras usando el conjunto de datos

Highway: La variable de respuesta es TASA y todas las otras son predictoras

a) Los metodos "stepwise". Explicar los pasos de los procesos y justificar la terminación del mismo.
b) Los mejores usando los mejores subcojuntos con por lo menos 6 criterios. Explicar los resultados
c) Comparar los resultados obtenidos en a) y b). Dar su selección final.

4. Verificar las ecuaciones (3) y (4) de las sección 6.3.4 del texto.
5. Investigar la relación entre los criterios AIC, BIC, Cp de Mallows y R^2.