10

Selección de variables

Con el objetivo de considerar el "mejor" modelo de regresión posible, el experimentador debe seleccionar un conjunto de variables regresoras entre las observadas y, si es necesario, entre potencias y productos de las mismas. Una primera decisión fijará el tipo de relación funcional con la variable respuesta pero, en todo caso, la selección de un conjunto reducido de variables explicativas es un problema complicado. Si consideramos un número demasiado pequeño de variables es posible que la potencia del modelo se vea reducida y que las estimaciones obtenidas sean sesgadas, tanto de los coeficientes de regresión, como de las predicciones. Este sesgo se origina ya que los errores calculados con los datos observados pueden contener efectos no aleatorios de las variables desechadas. Por otra parte, un número muy grande de variables explicativas complica la utilidad práctica del modelo y, aunque mejora el ajuste aparente, aumenta la varianza de los estimadores de los parámetros.

Decidir el mejor conjunto de variables es prácticamente un arte, en el que algunas técnicas sirven de apoyo: test t de Student de los coeficientes de regresión, test F de significación de la regresión, estudio de la multicolinealidad, etc. Sin embargo, ya hemos alertado sobre la utilización ciega de los test t parciales para medir la importancia de las variables. Así pues, es preciso añadir algunas técnicas específicas para comparar modelos de regresión que pasamos a detallar.

10.1. Algunas ideas

10.1.1. La navaja de Occam

Durante la primera mitad del S. XIV, en el punto central del oscurantismo medieval, un monje franciscano, Guillermo de Occam (u Ockham) encendió una pequeña luz que marcó el camino a seguir, con una idea simple pero cortante como una navaja separó a la ciencia de la teología e inició el camino de una filosofía libre y abierta a la razón, esta idea es conocida como el **Principio de parsimonia**, "Principio de economía de pensamiento de Occam" o "La navaja de Occam" y desde entonces ha sido un pilar de la metodología científica demostrando su valor metodológico.

"Entia non sunt multiplicanda sine necesitate" (los entes no deben multiplicarse sin necesidad o no expliques con más lo que puedas explicar con menos) en una traducción un poco menos literal "es soberbia hacer con más lo que se puede hacer con menos". En términos más actuales podríamos decir que de las explicaciones posibles la más simple es la correcta y en lenguaje común: la explicación más simple es la mejor. Si el objetivo es la creación de modelos para explicar fenómenos, es decir de modelos científicos, el principio consiste en utilizar como regla de decisión entre varias hipótesis, la que resulte más simple. Aunque, como explica Leibniz, saber cuál es el punto de vista más simple requiere cierta sutileza.

Este principio, que ha sido tan eficaz en la ciencia, también se puede aplicar a los modelos lineales y la regresión: ante dos modelos prácticamente equivalentes en funcionalidades, es mejor elegir el más sencillo. Cuantos más elementos innecesarios existan en un diseño más se reduce su eficacia y aumenta la probabilidad de provocar confusión y errores. Cuantas más variables regresoras en

el diseño más esfuerzo se exige a los usuarios para percibir, comprender e interactuar, y también existen más posibilidades de que surjan problemas. Por eso es aconsejable, según este principio, reducir el número de variables a las mínimas necesarias para realizar correctamente las funciones propias del modelo lineal. Es preferible la sencillez a la complejidad.

10.1.2. Consejos

Se trata de explicar los datos de la forma más simple posible. Es preferible un modelo simple con un buen ajuste que un modelo muy complejo con un ajuste ligeramente mejor. Así, las variables predictoras con información redundante deben ser eliminadas. El "mejor" modelo de regresión es el modelo más sencillo que se ajusta a los datos razonablemente bien.

Como ya se ha dicho, la utilización de muchas variables predictoras, algunas innecesarias, complica el modelo pero además añade ruido a la estimación de parámetros y cantidades que nos interesan. En realidad, como dice Faraway[?], es un despilfarro de grados de libertad.

También con demasiadas variables nos podemos encontrar con el problema de la indeseable colinealidad que hemos tratado en la sección 8.5. Un modelo simple reduce este peligro.

En muchos casos, el coste efectivo en tiempo y dinero es también un buen argumento para reducir el número de variables explicativas al mínimo necesario.

Recordemos que la utilización de alguna transformación de las variables, tanto la respuesta como las predictoras, puede mejorar el ajuste sin necesidad de más variables. También la eliminación, tal vez temporalmente, de puntos influyentes y atípicos debe preceder a la selección, puesto que ello ayudará al mejor ajuste y simplificará el modelo.

En el caso de utilizar un modelo polinómico, se conviene en no eliminar los términos de orden inferior de una variable cuando el término de orden superior de la misma variable es necesario en el modelo. Si consideramos, por ejemplo, el modelo

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

y el parámetro β_2 resulta significativo, entonces el modelo contiene el término en x, aunque β_1 no sea significativo. La razón es que aunque en estas circunstancias eliminaramos el término en x, un cambio de escala $x \to x+a$ haría reaparecer dicho término. Sin embargo, un cambio de escala como éste no debería modificar el modelo.

Del mismo modo, en el caso de un modelo de superficie de respuesta como

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

no se puede eliminar el término de la interacción $x_1 x_2$ sin eliminar al mismo tiempo los términos x_1^2 y x_2^2. Debemos elegir entre un modelo de superficie cuadrática o un modelo lineal plano.

10.2. Selección paso a paso

El procedimiento se puede realizar hacia delante (*forward stepwise*) o hacia atrás (*backward stepwise*), seleccionando las variables una a una e incorporándolas desde el modelo inicial o eliminándolas desde el modelo completo en función de su contribución al modelo. Aunque es el método más utilizado por su facilidad de computación, este sistema tiene el inconveniente de que puede conducir a modelos distintos y no necesariamente óptimos.

10.2.1. Introducción progresiva

En la introducción progresiva o selección hacia adelante (*forward stepwise*) se incorpora como primera variable la de mayor F de significación de la regresión simple. La segunda variable se

selecciona por su mayor contribución al modelo que ya contiene la primera variable del paso anterior y así sucesivamente.

La selección de una variable se puede determinar con el valor F_{in} que separa la región crítica del contraste F a partir de un nivel de significación α fijado.

Si disponemos de los p-valores, el menor de ellos señala la variable que añadiremos al modelo. Dicho p-valor debe ser inferior al nivel α. En caso contrario no se selecciona ninguna variable y el procedimiento termina.

Ejemplo 10.2.1

Con los datos de la tabla 10.1 vamos a reducir el número de variables predictoras mediante la introducción progresiva de las variables desde el modelo más simple que únicamente contiene la constante.

y	x_1	x_2	x_3	x_4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

Tabla 10.1: Datos para la selección de variables

Partimos de las sumas de cuadrados residuales SCR *de los diversos modelos y calculamos la contribución de la variable añadida mediante un test F parcial. En la siguiente tabla observamos que la primera variable que entra es* x_4.

modelo	SCR	F	F_{in}
– – – –	2715.76		4.8443
1 – – –	1265.69	12.602	
– 2 – –	906.34	21.960	
– – 3 –	1939.40	4.403	
– – – 4	883.87	22.798	*entra x_4

Por ejemplo, el estadístico F para determinar la contribución de x_4 *se calcula así:*

$$F = \frac{2715.76 - 883.87}{883.87/(13-2)} = 22.798$$

El valor F_{in} *que indica si alguna variable debe entrar se calcula para* $\alpha = 0.05$ *con 1,11 grados de libertad.*

Para el siguiente paso partimos del modelo que incluye la variable x_4. *Deberemos tener en cuenta la modificación en los grados de libertad, ahora son 1 y 10. En la tabla observamos que la variable que entra es* x_1.

modelo	SCR	F	F_{in}
– – – 4	883.87		4.9646
1 – – 4	74.76	108.228	*entra x_1
– 2 – 4	868.88	0.173	
– – 3 4	175.74	40.294	

El siguiente paso es definitivo ya que el valor de F_{in} no permite la inclusión de ninguna otra variable.

modelo	SCR	F	F_{in}
1 – – 4	74.76		5.1174
1 2 – 4	47.97	5.026	
1 – 3 4	50.84	4.234	

El modelo obtenido por este método señala como variables regresoras a x_1 y x_4.

10.2.2. Eliminación progresiva

En el procedimiento de eliminación progresiva o selección hacia atrás (*backward stepwise*) se parte del modelo lineal completo y se elimina la variable con menor contribución. Como en el caso anterior debemos fijar un nivel de significación α, el mismo a lo largo de todo el procedimiento. Dicho nivel establece el valor F_{out} para descartar efectivamente una variable.

Ejemplo 10.2.2

Con los mismos datos del ejemplo anterior, que se pueden ver en la tabla 10.1, calculamos las F parciales de contribución de cada variable al modelo.

modelo	SCR	F	F_{out}
1 2 3 4	47.86		5.3176
– 2 3 4	73.81	4.338	
1 – 3 4	50.84	0.498	
1 2 – 4	47.97	0.018	*sale x_3
1 2 3 –	48.11	0.042	

En el siguiente paso, observaremos la contribución de cada variable al modelo establecido en el paso anterior.

modelo	SCR	F	F_{out}
1 2 – 4	47.97		5.1174
– 2 – 4	868.88	154.017	
1 – – 4	74.76	5.026	
1 2 3 –	57.90	1.863	*sale x_4

Finalmente, la siguiente tabla confirma la necesidad de conservar las variables no descartadas hasta el momento.

modelo	SCR	F	F_{out}
1 2 – –	57.90		4.9646
– 2 – –	906.34	146.535	
1 – – –	1265.69	208.599	

10.2.3. Regresión paso a paso

Se trata de combinar la selección hacia adelante (*forward selection*) con la eliminación de variables (*backward elimination*). Esto puede mejorar especialmente las situaciones en las que la incorporación o el rechazo de alguna variable al principio del procedimiento parece que no ha sido apropiada. Aunque hay varias formas de proceder, podemos empezar, por ejemplo, con el modelo más sencillo e ir incorporando variables. En cada paso revisaremos si alguna de las variables incorporadas no contribuye como se espera (dada la presencia de las otras).

Ejemplo 10.2.3

Con la información que tenemos en el ejemplo 10.2.1 podemos revisar la incorporación de las variables x_1 y x_4 observando la significación parcial de los parámetros asociados a estas variables. Compararemos el estadístico F (puede ser una t) con un valor $F_{in/out}$ fijado con el nivel de significación de la selección (el mismo para todo el procedimiento).

	g.l.	SC	CM	F	$F_{in/out}$
x_1	1	1450.08	1450.08	193.96	4.9646
x_4	1	1190.92	1190.92	159.30	
Residuo	10	74.76	7.48		

Tabla 10.2: Análisis de la varianza para contrastar la significación parcial de las variables x_1 y x_4 en la regresión paso a paso

Los dos valores de F de la tabla 10.2 son claramente significativos y las dos variables permanecen en la selección.

Inconvenientes

Aunque los procedimientos de selección paso a paso son de fácil computación, tienen algunos inconvenientes:

1) En primer lugar, el hecho de seleccionar o eliminar una variable en cada paso hace que no consideremos algunos modelos entre los que podría estar el "óptimo".

2) La elección del nivel de significación es arbitraria y no tiene que ser el 0.05 como siempre. De hecho, en problemas de predicción parece que un nivel del 15 %-20 % es más adecuado.

3) Se hacen muchos contrastes y no se tiene en cuenta el problema de las comparaciones múltiples. Además, la eliminación de las variables menos significativas tiende a incrementar la significación de las que quedan.

4) Estos procedimientos no están directamente relacionados con los objetivos de predicción o explicación y, tal vez, no nos ayuden a resolver el problema de investigación.

5) En general, la selección de variables paso a paso tiende a señalar modelos demasiado reducidos para los propósitos de la predicción.

Ejemplo 10.2.4

En los ejemplos anteriores 10.2.1, 10.2.2 y 10.2.3 hemos llegado a dos modelos distintos, ambos con dos variables. También es posible que el "mejor" modelo sea el que contempla como variables predictoras a x_1, x_2 y x_4.

10.3. Procedimientos basados en un criterio

Si tenemos k variables explicativas, podemos hacer 2^k modelos distintos. Se trata de seleccionar el mejor de ellos por algún criterio. En realidad, algunos algoritmos inteligentes permiten ajustar la búsqueda sin necesidad de ajustarlos todos.

Los criterios más utilizados son:

1) Coeficiente de determinación ajustado.
2) Error cuadrático de validación.
3) Estadístico C_P de Mallows.
4) Criterio de la Información de Akaike (AIC).

10.3.1. Coeficiente de determinación ajustado

Esta técnica consiste en calcular los coeficientes de determinación de todos los modelos posibles con la combinación de cualquier número de variables explicativas. Para evitar los problemas que justifican la definición 8.2.1 resulta obvio utilizar el coeficiente ajustado cuando hay muchas variables en juego. El objetivo es reconocer el modelo con mayor coeficiente, ello equivale a la menor varianza residual estimada o ECM. Sin embargo, si el número de variables es considerable esta técnica puede tener dificultades de cálculo.

Ejemplo 10.3.1

Con los datos de los ejemplos anteriores 10.2.1, 10.2.2 y 10.2.3, podemos estudiar qué modelo de dos variables es mejor.

modelo	R^2 ajustado
1 2 – –	0.9744
1 – – 4	0.9670

El modelo más adecuado según este criterio está definido por las variables x_1 y x_2. El valor de $\bar{R}^2$ es bastante alto de forma que no parece necesario añadir otra variable, salvo que la investigación lo requiera.

10.3.2. Error cuadrático de validación

Recordemos (ver definición 9.4) que la suma de cuadrados PRESS o *error cuadrático de validación* de un modelo es la suma de cuadrados de los residuos en función de las predicciones $\hat{y}_{i(i)}$ calculadas con el modelo de regresión sin la i-ésima observación:

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2 \tag{10.1}$$

El objetivo según este criterio es hallar el modelo con menor error cuadrático de validación.

Seguramente este criterio proporciona modelos con muchas variables, aunque eso se agradece en los problemas de predicción.

10.3.3. Criterio C_P de Mallows

Con este criterio se debe fijar en primera instancia un número P de parámetros, incluido el término independiente, aunque con posterioridad se podrá variar. Se trata de hallar el mejor modelo con P variables explicativas, incluida la constante, utilizando el estadístico de Mallows

$$C_P = \frac{\text{SCR}_P}{\hat{\sigma}^2} - (n - 2P)$$

donde SCR_P es la suma de cuadrados residual del modelo particular y $\hat{\sigma}^2$ un estimador de la varianza del modelo que acostumbra a ser el ECM del modelo completo.

Para el modelo completo $P = k+1$, el estadístico de Mallows es

$$C_{k+1} = \frac{\text{SCR}}{\text{ECM}} - (n - 2(k+1)) = n - (k+1) - (n - 2(k+1)) = k+1$$

También para todo modelo no completo se puede demostrar que, si el modelo es adecuado, $E(\text{SCR}_P) = (n - P)\sigma^2$ y aproximadamente $E(C_P) \approx P$. En consecuencia parece recomendable elegir los conjuntos para los que C_P sea aproximadamente P. Los modelos con un mal ajuste tendrán un C_P mayor que P.

Ejemplo 10.3.2

Seguimos con los mismos datos del ejemplo 10.2.1 y parece que un buen modelo es suficiente que tenga dos variables, por ello tomamos $P = 3$ para el criterio de Mallows.

Para calcular los coeficientes C_3 podemos utilizar la varianza residual o ECM de cada modelo. Por ejemplo, para el modelo con las variables x_1 y x_2 tenemos

$$C_3 = \frac{\text{ECM}_{12} \cdot 10}{\text{ECM}} - (13 - 2 \cdot 3) = \frac{5.790 \cdot 10}{5.983} - 7 = 2.6782$$

modelo	ECM	C_3
1 2 3 4	5.983	
1 2 – –	5.790	2.6782
1 – 3 –	122.707	198.0946
1 – – 4	7.476	5.4959
– 2 3 –	41.544	62.4377
– 2 – 4	86.888	138.2259
– – 3 4	17.574	22.3731

Es evidente que el menor C_3 corresponde al modelo con las variables x_1 y x_2.

Como hemos visto en el ejemplo y se deduce de la fórmula, el coeficiente C_P está íntimamente ligado a la varianza residual de cada modelo y, por ende, a $\overline{R}^2$ (también al AIC del apartado siguiente).

Un gráfico puede ayudar a decidir el modelo más adecuado. Los puntos del gráfico de dispersión que representan a cada modelo tienen como primera coordenada el número de parámetros P y como segunda el estadístico C_P. Nuestra elección debe ser el modelo con menor P que más se acerque por debajo a la bisectriz.

Ejemplo 10.3.3

Con los datos del ejemplo anterior y si consideramos todos los modelos posibles, se obtiene el gráfico 10.1. En este gráfico se ha evitado dibujar los puntos que representan modelos con un C_P muy elevado. La elección está entre el modelo 12 *y el* 134.

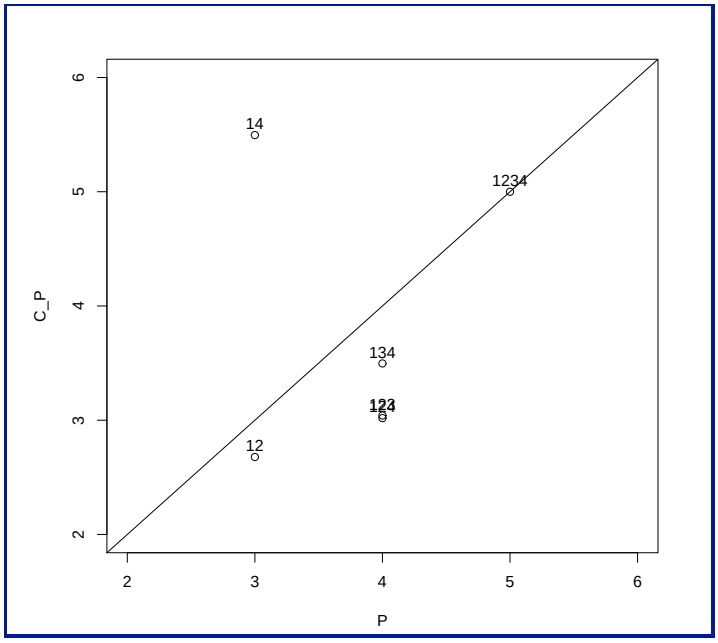

Figura 10.1: Gráfico de los estadísticos de Mallows para los datos del ejemplo 10.2.1

10.3.4. Criterio de la información de Akaike (AIC)

Akaike (1973, 1974) halló una relación formal (véase Burnham[?]) entre la información de Kullback-Leibler[1] y la teoría de la verosimilitud. El estimador propuesto por Akaike es

$$\ln\left(L(\hat{\theta}\,|\,\text{datos})\right) - K$$

donde L es la función de verosimilitud, $\hat{\theta}$ las estimaciones de máxima verosimilitud de los parámetros y K una constante para corregir el sesgo asintótico y no es arbitraria (como puede parecer en la literatura). Este estimador se multiplica por -2 (por razones históricas) y se tiene el criterio de información de Akaike general:

$$\text{AIC} = -2\ln\left(L(\hat{\theta}\,|\,\text{datos})\right) + 2K$$

En el caso particular de la regresión mínimo-cuadrática con errores normales (ver página 36), los estimadores de máxima verosimilitud de los parámetros β son los mínimo-cuadráticos, pero el estimador de σ^2 es SCR/n. Por todo ello (véase ejercicio 10.2), el estimador AIC correcto es

$$\text{AIC}_c = n + n\ln(2\pi) + n\ln(\text{SCR}/n) + 2(k+2) \tag{10.2}$$

Observemos que, en este caso, la constante K es el número total de parámetros, incluida la intercepción y la varianza del modelo.

Del mismo modo se define el estimador criterio de información de Bayes

$$\text{BIC}_c = n + n\ln(2\pi) + n\ln(\text{SCR}/n) + (\ln n)(k+2) \tag{10.3}$$

que aplica un castigo más duro a la falta de parsimonia pues añade $\ln n$ veces el número de parámetros.

El procedimiento que se basa en el criterio de información de Akaike (u opcionalmente Bayes) busca el modelo con menor AIC (o BIC). Ahora bien, como se trata de comparar modelos, es suficiente utilizar una fórmula simplificada del estadístico 10.2 que es

$$\text{AIC} = n\ln(\text{SCR}/n) + 2(k+1)$$

1. una medida de discrepancia entre un modelo aproximado y el modelo real o verdadero

donde $n \ln(\mathrm{SCR}/n)$ es la llamada *desviación* y $k+1$ es el número de parámetros β.
Del mismo modo, el estadístico BIC que se utiliza es

$$\mathrm{BIC} = n \ln(\mathrm{SCR}/n) + (\ln n)(k+1)$$

Ejemplo 10.3.4

Con los datos que hemos estudiado en todo este capítulo los modelos con menor AIC *son*

modelo	AIC	AIC_c
1 2 – 4	24.9739	63.8663
1 2 3 –	25.0112	63.9036
1 2 – –	25.4200	64.3124
1 3 – 4	25.7275	64.6200
1 2 3 4	26.9443	65.8367

Con este criterio el modelo elegido consta de las variables x_1, x_2 y x_4 que mejora ligeramente el modelo de dos variables x_1 y x_2.
En este caso no vale la pena utilizar el estadístico BIC *porque tenemos muy pocas variables.*

10.4. Ejemplos con R

Con los datos del ejemplo 10.2.1 que se muestran en la tabla 10.1 vamos a seleccionar las variables regresoras hacia adelante (forward). El sistema consiste en comparar los modelos, empezando por el más sencillo, con la función anova. Tomamos un nivel de significación $\alpha = 0.05$.
Primero introducimos los datos:

```
> y<-c(78.5,74.3,104.3,87.6,95.9,109.2,102.7,72.5,93.1,115.9,83.8,113.3,109.4)
> x1<-c(7,1,11,11,7,11,3,1,2,21,1,11,10)
> x2<-c(26,29,56,31,52,55,71,31,54,47,40,66,68)
> x3<-c(6,15,8,8,6,9,17,22,18,4,23,9,8)
> x4<-c(60,52,20,47,33,22,6,44,22,26,34,12,12)
```

Después definimos los modelos

```
> modelo0.lm <- lm(y~1)
> modelo1.lm <- lm(y~x1)
> modelo2.lm <- lm(y~x2)
> modelo3.lm <- lm(y~x3)
> modelo4.lm <- lm(y~x4)
```

y hacemos las comparaciones (sólo se muestra la última):

```
> anova(modelo0.lm,modelo1.lm)
...
> anova(modelo0.lm,modelo2.lm)
...
> anova(modelo0.lm,modelo3.lm)
...
> anova(modelo0.lm,modelo4.lm)
Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x4
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
```

```
1     12 2715.76
2     11  883.87  1   1831.90 22.799 0.0005762 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ahora disponemos de los p-valores de cada comparación. Así vemos que entra la variable x_4, ya que su p-valor es el más pequeño, pero superior a α.
El siguiente paso es añadir una variable al modelo

```
> modelo14.lm <- update(modelo4.lm,. ~ .+x1)
> modelo24.lm <- update(modelo4.lm,. ~ .+x2)
> modelo34.lm <- update(modelo4.lm,. ~ .+x3)
```

y volver a comparar con el modelo inicial ya establecido:

```
> anova(modelo4.lm,modelo14.lm)
...
> anova(modelo4.lm,modelo24.lm)
...
> anova(modelo4.lm,modelo34.lm)
...
```

El p-valor indica la incorporación de la variable x_1.
En el siguiente paso el procedimiento termina, ya que los p-valores son superiores al nivel α y no podemos incorporar ninguna variable más.

```
> modelo124.lm <- update(modelo14.lm,. ~ .+x2)
> modelo134.lm <- update(modelo14.lm,. ~ .+x3)
> anova(modelo14.lm,modelo124.lm)
...
> anova(modelo14.lm,modelo134.lm)
...
```

La selección de variables hacia atrás, partiendo del modelo completo, es mucho más sencilla de aplicar mediante la instrucción `summary` del modelo lineal.

```
> modelo.lm <- lm(y~x1+x2+x3+x4)
> summary(modelo.lm)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1750 -1.6709  0.2508  1.3783  3.9254

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
x1            1.5511     0.7448   2.083   0.0708 .
x2            0.5102     0.7238   0.705   0.5009
x3            0.1019     0.7547   0.135   0.8959
x4           -0.1441     0.7091  -0.203   0.8441
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-Squared: 0.9824,     Adjusted R-squared: 0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

Observemos que los contrastes t son equivalentes a los contrastes F, de forma que descartamos la variable x_3, ya que el contraste de significación de su parámetro tiene el mayor p-valor de todos. Reformulamos el modelo sin la variable excluida y miramos los p-valores.

```
> modelo.lm <- update(modelo.lm,. ~ .-x3)
> summary(modelo.lm)

Call:
lm(formula = y ~ x1 + x2 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0919 -1.8016  0.2562  1.2818  3.8982

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6483    14.1424   5.066 0.000675 ***
x1            1.4519     0.1170  12.410 5.78e-07 ***
x2            0.4161     0.1856   2.242 0.051687 .
x4           -0.2365     0.1733  -1.365 0.205395
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.309 on 9 degrees of freedom
Multiple R-Squared: 0.9823,     Adjusted R-squared: 0.9764
F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
```

Ahora la variable que excluiremos es x_4.
Finalmente el siguiente paso no excluye a ninguna otra variable, puesto que los p-valores son inferiores al nivel $\alpha = 0.05$.

```
> modelo.lm <- update(modelo.lm,. ~ .-x4)
> summary(modelo.lm)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
   Min     1Q Median     3Q    Max
-2.893 -1.574 -1.302  1.362  4.048

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.57735    2.28617   23.00 5.46e-10 ***
x1           1.46831    0.12130   12.11 2.69e-07 ***
x2           0.66225    0.04585   14.44 5.03e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.406 on 10 degrees of freedom
Multiple R-Squared: 0.9787,     Adjusted R-squared: 0.9744
F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-09
```

En S-PLUS hay una función `stepwise` que aplica estos métodos de selección paso a paso aunque como se ha indicado es mejor optar por un procedimiento con criterio. Para ello, en R se dispone de un paquete llamado `leaps` que no forma parte de la instalación básica y que debemos buscar entre las contribuciones. Una vez instalado en nuestro sistema podemos proceder:

```
> library(leaps)
```

```
> X<-matrix(c(x1,x2,x3,x4),ncol=4)
> g<-leaps(X,y,method="Cp")
```

El objeto `g` dispone de varios vectores entre los que destacaremos `g$which` con los modelos, `g$size` con el número de parámetros y `g$Cp` con los estadísticos. Atención al orden de los modelos, de menor a mayor C_P dentro de cada grupo con igual P.
Como en nuestro ejemplo nos interesa especialmente el caso $P = 3$ (dos variables) hacemos

```
> g$which[g$Cp==min(g$Cp[g$size==3])]
```

y nos indica el modelo 12.
Para dibujar el gráfico 10.1 las instrucciones son:

```
> etiquetas<-c("4","2","1","3","12","14","34","23","24","13",
+              "124","123","134","234","1234")
> plot(g$size,g$Cp,xlim=c(2,6),ylim=c(2,6),xlab="P",ylab="C_P")
> text(g$size,g$Cp+0.1,labels=etiquetas)
> abline(0,1)
```

La función `leaps` dispone también de los métodos `adjr2` y `r2`. También existe una función `regsubsets` que mejora la función `leaps` en algunos aspectos.

```
> datos<-data.frame(y,x1,x2,x3,x4)
> mejor_subconj<-regsubsets(y~x1+x2+x3+x4,data=datos)
> summary(mejor_subconj)
Subset selection object
Call: regsubsets.formula(y ~ x1 + x2 + x3 + x4, data = datos)
4 Variables  (and intercept)
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
1 subsets of each size up to 4
Selection Algorithm: exhaustive
         x1  x2  x3  x4
1  ( 1 ) " " " " " " "*"
2  ( 1 ) "*" "*" " " " "
3  ( 1 ) "*" "*" " " "*"
4  ( 1 ) "*" "*" "*" "*"
```

Para calcular el estadístico del criterio de información de Akaike en R se dispone de una función específica `AIC` para un objeto `lm` que justamente proporciona el AIC_c de la fórmula 10.2.

```
> modelo12.lm <- lm(y~x1+x2)
> AIC(modelo12.lm)
[1] 64.31239
```

Pero para utilizar el criterio y seleccionar el mejor modelo tenemos la función `step`. Realmente no hace falta evaluar todos los modelos y el algoritmo que se aplica compara modelos secuencialmente, de forma que se parece al método de selección paso a paso sin la desventaja de utilizar contrastes.

```
> step(modelo.lm)
```

El resultado por defecto es una selección desde el modelo completo hacia atrás, utilizando el estadístico AIC simplificado. Si queremos utilizar el estadístico BIC debemos añadir el parámetro `k=log(n)` a la función `step`.

10.5. Ejercicios

Ejercicio 10.1

Uno de los ejercicios más clásicos recoge los datos de 17 fábricas de Shanghai:

Output	SI	SP	I
12090	56	840	10.54
11360	133	2040	11.11
12930	256	2410	10.73
12590	382	2760	14.29
16680	408	2520	11.19
23090	572	2950	14.03
16390	646	2480	18.76
16180	772	2270	13.53
17940	805	4040	16.71
18800	919	2750	14.74
28340	1081	3870	29.19
30750	1181	4240	21.21
29660	1217	2840	12.45
20030	1388	3420	17.33
17420	1489	3200	24.40
11960	1508	3060	28.26
15700	1754	2910	19.52

donde Output es el rendimiento "per capita" en yuans chinos, SI el número de trabajadores en la fábrica, SP el terreno de la fábrica en metros cuadrados por trabajador y I la inversión en yuans por trabajador.

a) Ajustar un modelo de regresión que exprese la variable Output en función de las otras variables.

b) Añadir al modelo anterior las variables de segundo orden que se crean necesarias.

c) Utilizar el modelo del apartado anterior para maximizar el Output per capita.

Ejercicio 10.2

Probar que el estimador AIC en el caso de la regresión es

$$\mathrm{AIC}_c = n + n\ln(2\pi) + n\ln(\mathrm{SCR}/n) + 2(k+1+1)$$