Capítulo 12

Selección de modelos.

12.1. Criterios para la comparación.

En ocasiones, ajustamos un modelo de regresión teniendo una idea clara de las variables que debemos incluir como regresores. Es más frecuente, sin embargo, el caso en que sólo tenemos una idea aproximada de la forma adecuada para nuestro modelo, y debemos decidir con criterio estadístico qué regresores deben ser incluidos.

Para enfrentar este tipo de situaciones necesitamos, por una parte, criterios de bondad de ajuste, capaces de permitirnos comparar distintos modelos ajustados a una misma muestra. Por otra, necesitamos estrategias de selección de variables que construyan de manera automática o semi-automática subconjuntos de todos los modelos posibles susceptibles de incluir el "mejor". Examinaremos en esta Sección el primer punto.

Es claro que no podemos preferir un modelo a otro simplemente porque su SSE es menor, dado que toda[1] variable que incluyamos en la regresión, tenga mucha o poca relación con la variable respuesta, reducirá SSE. Tenemos, pues, que buscar criterios más elaborados.

[1]Las únicas excepciones son aquellas variables correspondientes a columnas de la matriz de diseño X ortogonales a $\vec{y}$, o que son combinación lineal exacta de columnas correspondientes a variables ya presentes entre los regresores.

12.1.1. Maximización de $\overline{R}_p^2$.

Se define el *coeficiente de determinación corregido* así:

$$\overline{R}_p^2 = 1 - [1 - R_p^2] \times \frac{N-1}{N-p} \tag{12.1}$$

haciendo referencia el subíndice p al número de regresores presentes en el modelo. Si reescribimos la ecuación (13.1) en la forma:

$$1 - \overline{R}_p^2 = [1 - R_p^2] \times \frac{N-1}{N-p} \tag{12.2}$$

$$= \frac{SSE_p}{SST} \times \frac{N-1}{N-p} \tag{12.3}$$

vemos que mientras que el primer término de la derecha de (13.3) es monótono no creciente con p, el segundo es monótono creciente. Por consiguiente, el producto de ambos[2] puede crecer o decrecer al crecer p.

Es frecuente por ello utilizar $\overline{R}_p^2$ como criterio de ajuste. Aunque útil, veremos sin embargo que debe complementarse con otros criterios. Su exclusiva aplicación da lugar con gran probabilidad a modelos sobreparametrizados, como pone de manifiesto el siguiente teorema.

Teorema 12.1 *El estadístico $\overline{R}_p^2$ crece con la introducción de un parámetro en la ecuación de regresión si el estadístico Q_h asociado al contraste de significación de dicho parámetro verifica $Q_h > 1$.*

Demostración:[3]

Para contrastar la significación del $(p+1)$-ésimo parámetro, empleamos (Sección 7.2, pág. 77):

$$Q_h = \frac{SSE_p - SSE_{p+1}}{SSE_{p+1}} \times \frac{N-p-1}{1} \tag{12.4}$$

$$= \frac{(R_{p+1}^2 - R_p^2)}{1 - R_{p+1}^2} \times \frac{N-p-1}{1} \tag{12.5}$$

[2]Expresiones como la anterior con un término función de la suma de cuadrados de los residuos y otro interpretable como "penalización" por la introducción de parámetros adicionales, son ubicuas en la literatura estadística. La C_p de Mallows que se examina más abajo tiene la misma forma, como muchos criterios de ajuste utilizados sobre todo en el análisis de series temporales: Criterio de Información de Akaike (AIC), FPE, BIC, etc.

[3]Sigue a Haitovsky (1969).

de donde:

$$\begin{aligned}(1-R^2_{p+1})Q_h &= (R^2_{p+1}-R^2_p)(N-p-1) && (12.6)\\ Q_h - Q_h R^2_{p+1} &= (N-p-1)R^2_{p+1} - (N-p-1)R^2_p && (12.7)\\ Q_h + (N-p-1)R^2_p &= R^2_{p+1}\left[(N-p-1)+Q_h\right] && (12.8)\end{aligned}$$

Despejando R^2_{p+1} tenemos:

$$\begin{aligned}R^2_{p+1} &= \frac{Q_h + (N-p-1)R^2_p}{(N-p-1)+Q_h} && (12.9)\\ &= \frac{\frac{1}{N-p-1}Q_h + R^2_p}{1+\frac{1}{N-p-1}Q_h} && (12.10)\end{aligned}$$

De (13.10) y de la definición de $\overline{R}^2_{p+1}$ se deduce que:

$$\overline{R}^2_{p+1} = 1-[1-R^2_{p+1}]\times\frac{N-1}{(N-p-1)} \qquad (12.11)$$

Sustituyendo en esta expresión (13.10) llegamos a:

$$\begin{aligned}\overline{R}^2_{p+1} &= 1-\frac{[1-R^2_p]}{\frac{N-p-1+Q_h}{N-p-1}}\times\frac{N-1}{N-p-1} && (12.12)\\ &= 1-[1-R^2_p]\frac{N-1}{N-p-1+Q_h} && (12.13)\\ &= \underbrace{1-[1-R^2_p]\frac{N-1}{N-p}}_{\overline{R}^2_p}\underbrace{\frac{N-p}{N-p-1+Q_h}}_{t} && (12.14)\end{aligned}$$

Es evidente de (13.14) que $\overline{R}^2_{p+1} \geq \overline{R}^2_p$ si $Q_h > 1$, y viceversa[4]. Maximizar $\overline{R}^2_p$ implica introducir en la ecuación de regresión todos aquellos regresores cuyo estadístico Q_h sea superior a la unidad; pero esto ocurre con probabilidad $\approx 0{,}50$ incluso cuando $h\colon \beta_i = 0$ es cierta. Consecuentemente, el emplear este criterio en exclusiva conduciría con gran probabilidad al ajuste de modelos sobreparametrizados.

[4]Obsérvese que si el término t en (13.14) fuera la unidad —lo que acontece cuando $Q_h = 1$—, el lado derecho sería precisamente $\overline{R}^2_p$. Si $Q_h > 1$, t es menor que 1 y, como sólo multiplica al sustraendo en (13.14), el resultado es *mayor* que $\overline{R}^2_p$.

12.1.2. Criterio C_p de Mallows.

Supongamos que la variable aleatoria Y se genera realmente como prescribe el modelo $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$, no obstante lo cual ajustamos el modelo equivocado $Y = \tilde{X}\tilde{\beta} + \vec{\epsilon}$ con p parámetros. Una vez estimado, dicho modelo suministra las predicciones $\hat{Y}^{(p)}$. Un criterio para evaluar la adecuación del modelo estimado al real, sería el error cuadrático medio

$$ECM = E(\hat{Y}^{(p)} - X\vec{\beta}\,)'(\hat{Y}^{(p)} - X\vec{\beta}\,) \tag{12.15}$$

que sumando y restando $E(\hat{Y}^{(p)})$ dentro de cada paréntesis podemos descomponer así:

$$\begin{aligned} ECM &= E\left[(\hat{Y}^{(p)} - E(\hat{Y}^{(p)}))'(\hat{Y}^{(p)} - E(\hat{Y}^{(p)}))\right] \\ &\quad + E\left[(E(\hat{Y}^{(p)}) - X\vec{\beta}\,)'(E(\hat{Y}^{(p)}) - X\vec{\beta}\,)\right] \qquad (12.16) \\ &= \mathrm{Var}(\hat{Y}^{(p)}) + (\mathrm{Sesgo})^2. \qquad (12.17) \end{aligned}$$

El primer término no ofrece dificultad. Como

$$\hat{Y}^{(p)} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\vec{Y} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'(X\vec{\beta} + \vec{\epsilon}), \tag{12.18}$$

tenemos que

$$E[\hat{Y}^{(p)}] = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'X\vec{\beta}$$

y

$$\begin{aligned} ((\hat{Y}^{(p)} - E(\hat{Y}^{(p)}))'((\hat{Y}^{(p)} - E(\hat{Y}^{(p)})) &= \vec{\epsilon}\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\vec{\epsilon} \\ &= \vec{\epsilon}\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\vec{\epsilon} \\ &\sim \sigma^2\chi^2_p. \qquad (12.19) \end{aligned}$$

Falta el término de sesgo. Observemos que

$$\underbrace{E[(\vec{Y} - \hat{Y}^{(p)})'(\vec{Y} - \hat{Y}^{(p)})]}_{SSE} = \underbrace{E\left[(X\vec{\beta} - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'X\vec{\beta}\,)'(X\vec{\beta} - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'X\vec{\beta}\,)\right]}_{(\mathrm{Sesgo})^2}$$
$$+ \quad E\left[\vec{\epsilon}\,'(I - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}')\vec{\epsilon}\right].$$

Por consiguiente,

$$(\mathrm{Sesgo})^2 = E[SSE] - E[\sigma^2\chi^2_{N-p}]. \tag{12.20}$$

Sustituyendo en (13.17) tenemos entonces que

$$\begin{aligned} ECM &= E\left[SSE - \sigma^2\chi^2_{N-p}\right] + E\left[\sigma^2\chi^2_p\right] && (12.21)\\ &= E[SSE] - \sigma^2(N-p) + \sigma^2 p, && (12.22)\end{aligned}$$

y por consiguiente:

$$\frac{ECM}{\sigma^2} = E\left[\frac{SSE}{\sigma^2}\right] - N + 2p. \tag{12.23}$$

Minimizar esta última expresión es lo mismo que minimizar

$$E\left[\frac{SSE}{\sigma^2}\right] + 2p, \tag{12.24}$$

ya que N es constante. Como quiera que el valor medio en la expresión anterior no puede ser calculado y σ es desconocida, todo lo que podemos hacer es reemplazar (13.24) por la expresión análoga,

$$C_p = \frac{SSE}{\hat{\sigma}^2} + 2p. \tag{12.25}$$

A esta última expresión se la conoce como C_p de Mallows.

Para que se verifique la aproximación en (13.25) es preciso que $\hat{\sigma}^2 \approx \sigma^2$, lo que se consigue si la muestra es lo suficientemente grande y $\hat{\sigma}^2 = SSE^{(N-p-k)}/(N-p-k)$, estando entre los $(p+k)$ regresores incluídos los p necesarios. Incluso aunque entre dichos $(p+k)$ regresores haya algunos innecesarios, $\hat{\sigma}^2$ es insesgado; el precio que se paga por emplear más parámetros de los debidos en la estimación de σ^2 es una reducción en el número de grados de libertad (véase Sección 6.2).

De acuerdo con el criterio de Mallows, seleccionaremos el modelo que minimice C_p. La expresión (13.25) es otro ejemplo de criterio de ajuste con penalización. Cada nuevo parámetro que introducimos, reduce quizá SSE, pero esta reducción tiene un precio: el incremento del segundo sumando de (13.25) en 2. El efecto neto indica si el nuevo regresor es o no deseable.

> **Observación 12.1** De acuerdo con el criterio C_p de Mallows, dada una ecuación de regresión con unos ciertos regresores presentes, introduciremos un nuevo regresor si éste puede "pagar" su inclusión reduciendo SSE en, al menos, dos veces $\hat{\sigma}^2$. La maximización de $\overline{R}^2_p$, en cambio, requeriría en análoga situación introducir el mismo regresor si disminuye SSE en al menos una vez $\hat{\sigma}^2$. El criterio C_p de Mallows es más restrictivo[5].

[5]La comparación es aproximada tan sólo. El valor de $\hat{\sigma}^2$ que se emplea en el criterio C_p se obtiene, típicamente, ajustando el modelo más parametrizado (esto minimiza el riesgo de

Observación 12.2 Un estadístico se enfrenta con frecuencia a este dilema en su trabajo. ¿Hasta dónde procede llevar la complejidad del modelo a emplear? ¿Qué mejora en el ajuste de un modelo a la muestra justifica la adición de un nuevo parámetro?. O, si se prefiere, ¿Cuán afilada debe ser la navaja de Ockham? En el caso del modelo de regresión lineal, el criterio C_p suministra seguramente una navaja con el filo adecuado; argumentos alternativos llevan a criterios equivalentes o similares al C_p. Es un hecho notable y llamativo que por diversas vías se llegue siempre a análogos resultados, que tienen en común el medir la complejidad del modelo empleado como una función lineal o aproximadamente lineal del número de sus parámetros; más sobre esto en la Sección 13.1.5. En la Sección 13.1.4 se introduce la idea de la *validación cruzada*, que proporciona una forma alternativa de evaluar la bondad de ajuste de un modelo soslayando el empleo de una penalización basada en el número de parámetros.

12.1.3. Criterio AIC

Relacionado con el criterio C_p de Mallows, aunque válido de modo mucho más general y motivado de modo muy diferente, está el criterio AIC (**A**kaike's **I**nformation **C**riterion, o **A**n **I**nformation **C**riterion). Consiste en seleccionar el modelo minimizando

$$AIC(p) = -2\log_e\left[\max_{\vec{\theta}} \text{verosimilitud}(\vec{x}\,,\vec{\theta}\,)\right] + 2p$$

El primer término en la expresión anterior es, como en la C_p de Mallows, una medida de bondad de ajuste (disminuye al crecer el máximo de la verosimilitud); el segundo penaliza el número de parámetros en $\vec{\theta}$. Puede verse una justificación en Akaike (1972) (y en Akaike (1974), Akaike (1991)). Una explicación simplificada que sigue esencialmente a de Leeuw (2000) puede encontrarse en Tusell (2003), Sección **??**.

Cuando consideremos modelos de regresión lineal con normalidad, el uso de los criterios AIC y C_p daría resultados exactamente equivalentes si conociéramos σ^2 (ambos criterios difieren en tal caso en una constante; ver Venables and Ripley (1999a), pág. 185). Cuando σ^2 es desconocida y ha de ser estimada a partir de los datos, ambos criterios pueden diferir, pero son a efectos prácticos intercambiables. El criterio AIC no obstante es de ámbito mucho más

introducir sesgos en la estimación de σ^2, aunque seguramente nos hace despilfarrar algunos grados de libertad). Por el contrario, al utilizar el criterio basado en $\overline{R}_p^2$ introducimos el nuevo regresor si $Q_h > 1$ en (13.4), es decir, si la disminución $SSE_p - SSE_{p+1}$ en la suma de cuadrados de los residuos es mayor que $\hat{\sigma}^2 = SSE_{p+1}/(N-p-1)$, varianza estimada *en el modelo con* $p+1$ *regresores.*

general, y puede ser utilizado dondequiera que tengamos una verosimilitud, sea o no normal la distribución generadora de la muestra.

12.1.4. Residuos borrados y validación cruzada

Hemos visto que el problema de emplear como criterio para la selección de modelos alguno de los estadísticos de ajuste obvios (suma de cuadrados residual, R^2, o similar) estriba en que hay que tomar en consideración el diferente número de parámetros en cada modelo.

El problema consiste en que, al incrementar el número de parámetros, el modelo puede "seguir" más a la muestra, ajustando no sólo el comportamiento predecible sino incluso el puramente aleatorio Se adapta muy bien a *una* muestra —la que hemos empleado para estimarlo—, pero quizá no a otras.

Una solución consistiría en estimar los modelos con una muestra (muestra de entrenamiento o aprendizaje) y evaluarlos examinando su comportamiento en la predicción de *otra* diferente (muestra de validación). Actuando así, estaríamos a salvo de impresiones excesivamente optimistas: la suma de cuadrados de los residuos o R^2 que calculáramos para cada modelo reflejaría su capacidad de generalización: su comportamiento con otras observaciones distintas de las que han servido para estimarlo.

Lamentablemente, esto requiere dividir nuestra disponibilidad de observaciones en dos grupos: uno para estimar y otro para validar. El obtener un diagnóstico realista por este procedimiento requiere sacrificar en aras de la validación una preciosa fracción de muestra que habría permitido, quizá, estimar mejor.

¿Realmente es esto así? No; una vez que hemos decidido por el procedimiento anterior de fraccionar la muestra en dos para seleccionar el modelo mejor, podemos emplear *todas* las observaciones en reestimarlo.

La idea de la *validación cruzada* incorpora una mejora adicional al planteamiento anterior. No tenemos necesariamente que usar sólo una fracción de la muestra para validar. Podemos dividir la muestra en dos (o más) partes y emplear todas ellas en la validación. El ejemplo que sigue detalla los pasos a seguir haciendo validación cruzada por mitades.

Ejemplo 12.1 Consideremos una muestra de tamaño $N = 100$. Tenemos una colección de K modelos $\mathcal{M}_i$, $i = 1, \ldots, K$, posiblemente con diferente número de parámetros, de entre los que queremos seleccionar uno. Podemos dividir la muestra en dos trozos, A y B, de tamaños respectivos $N_A = N_B = 50$, y proceder así:

1. Con la muestra A estimaremos cada uno de los modelos $\mathcal{M}_i$.

2. Examinaremos el ajuste de los modelos así estimados a la muestra B, computando sumas de cuadrados residuales para cada uno de los modelos, $SSE_i^{(A)}$.
3. Con la muestra B estimaremos cada uno de los modelos $\mathcal{M}_i$.
4. Examinaremos el ajuste de los modelos así estimados a la muestra A, computando sumas de cuadrados residuales para cada uno de los modelos, $SSE_i^{(B)}$
5. Tanto $SSE_i^{(A)}$ como $SSE_i^{(B)}$ son estimaciones de las sumas de cuadrados de los residuos del modelo $\mathcal{M}_i$, cuando se utiliza en predicción sobre una muestra diferente de la que se ha empleado en su estimación. Podemos promediar ambas para obtener un único estadístico, $SSE_i = \frac{1}{2}(SSE_i^{(A)} + SSE_i^{(B)})$.
6. Seleccionaremos el modelo $\mathcal{M}_i$ tal que SSE_i es mínimo.

Observemos que nada nos constriñe a dividir la muestra en dos partes; podríamos dividirla en s partes, y proceder exactamente del mismo modo: utilizaríamos sucesivamente $s-1$ partes para estimar y la restante para evaluar $SSE_i^{(\ell)}$, $\ell = 1, \ldots, s$, (suma de cuadrados de los residuos al predecir en la muestra ℓ mediante el modelo $\mathcal{M}_i$ estimado con las restantes observaciones). Promediando los s valores $SSE_i^{(\ell)}$ obtendríamos el SSE_i del modelo $\mathcal{M}_i$.

El caso extremo consistiría en tomar $s = N$, y realizar el proceso dejando cada vez fuera una única observación (validación cruzada de tipo *leave one out*).

En muchas situaciones esta estrategia puede requerir un esfuerzo de cálculo formidable: ¡cada modelo ha de ser reestimado $(N-1)$ veces, dejando cada vez fuera de la muestra de estimacion una observación diferente! En regresión lineal, sin embargo, la diferencia entre la predicción de la observación i-ésima haciendo uso de todas las restantes y el valor observado de la misma es, simplemente, el residuo borrado, de cómoda y rápida obtención (véase Sección 12.1.4). Por tanto, utilizando la notación de dicha Sección,

$$\begin{aligned} SSE_i^{\ell} &= d_{\ell}^2 \qquad (\ell = 1, \ldots, N) \\ SSE_i &= N^{-1} \sum_{\ell=1}^{N} SSE_i^{\ell}. \end{aligned}$$

El modelo seleccionado es aquél al que corresponde un SSE_i más pequeño[6].

[6]Nótese que SSE_i es lo que se conoce también como suma de cuadrados de los residuos predictiva o PRESS; véase nota a pie de página de la Sección 12.1.4.

Fin del ejemplo ■

12.1.5. Complejidad estocástica y longitud de descripción mínima*

En esencia, seleccionar un modelo entraña adoptar un compromiso entre la bondad de ajuste y la complejidad, medida por el número de sus parámetros. Sabemos que un modelo lineal suficientemente parametrizado podría ajustar perfectamente la muestra, pero que ello no significa que sea idóneo: puede tener muy poca capacidad de generalización. Por el contrario, un modelo que no incluya los parámetros suficientes dara un ajuste susceptible de mejora. Se trata de alcanzar un equilibrio entre los dos objetivos en contradicción: un modelo dando buen ajuste y con los mínimos parámetros precisos.

Una aproximación intuitivamente atrayente al problema es la siguiente: tratemos de dar una descripción tan corta como sea posible de la evidencia (la muestra). Esto puede de nuevo verse como una apelación al principio de Ockham: construir "explicaciones" de la realidad que hacen uso del mínimo número de entidades.

La aproximación propuesta exige medir la longitud de la descripción que hagamos, y podemos para ello hacer uso de la Teoría de la Información. No podemos elaborar esta cuestión con detalle aquí (véase una buena introducción en Rissanen (1989), y detalles en Legg (1996)). En esencia, dado un modelo probabilistico podemos describir o codificar unos datos de modo compacto asignando a los más "raros" (menos probables) los códigos más largos.

> **Observación 12.3** Esta estrategia, de sentido común, es la que hace que al codificar en el alfabeto telegráfico de Morse la letra "e" (muy frecuente en inglés) se adoptara el código `.`, reservando los códigos más largos para caracteres menos frecuentes (ej: `-..-` para la "x").

Además de codificar los datos tenemos que codificar los parámetros del modelo probabilistico. La longitud total de descripción de la muestra $\vec{y}$ cuando hacemos uso del modelo probabilístico $\mathcal{M}_k$ haciendo uso del vector de parámetros $\vec{\theta}_k$ es entonces

$$\begin{aligned} MDL(\mathcal{M}_k; \vec{y}) &= (\text{Código necesario para } \vec{y}) && (12.26)\\ &+ \quad (\text{Código necesario para } \vec{\theta}_k). && (12.27) \end{aligned}$$

Un mal ajuste hará que el primer sumando sea grande; los datos muestrales se desvían mucho de lo que el modelo predice. Un modelo con un perfecto ajuste tendría un primer sumando nulo (porque las $\vec{y}$ se deducirían exactamente del modelo, y no requerirían ser codificadas), pero requeriría quizá muchos parámetros incrementando el segundo sumando.

El criterio MDL propone seleccionar el modelo $\mathcal{M}_k$ que minimiza (13.27). En el caso de modelos de regresión, el criterio MDL da resultados íntimamente emparentados asintóticamente con los precedentes (suma de cuadrados PRESS y C_p); véanse detalles en Rissanen (1989), Cap. 5.

12.2. Selección de variables.

Una aproximación ingenua al problema consistiría en estudiar la reducción en un cierto criterio (SSE, $\overline{R}_p^2$, C_p, ...) originada por la introducción de cada variable, y retener como regresores todas aquellas variables que dieran lugar a una reducción significativa. Desgraciadamente, esta estrategia no tiene en cuenta el hecho de que, a menos que las columnas de la matriz de diseño X sean ortogonales, la reducción en SSE originada por la inclusión de una variable depende de qué otras variables estén ya presentes en la ecuación ajustada.

Se impone, pues, emplear procedimientos más sofisticados. Relacionamos algunos de los más utilizados.

12.2.1. Regresión sobre todos los subconjuntos de variables.

De acuerdo con el párrafo anterior, la adopción de una estrategia ingenua podría dificultar el hallazgo de un modelo adecuado. Por ejemplo, puede bien suceder que una variable X_i, que debiera ser incluída en el modelo, no origine una reducción significativa de SSE cuando la introducimos después de X_j. Si esto ocurre, es claro que X_i no mostrará sus buenas condiciones como regresor mas que si es introducida con X_j ausente.

Una posible solución sería, dados p regresores, formar todos los posibles subconjuntos de regresores y efectuar todas las posibles regresiones, reteniendo aquélla que, de acuerdo con el criterio de bondad de ajuste que hayamos adoptado, parezca mejor.

El inconveniente es el gran volumen de cálculo que es preciso realizar. Piénsese que con p regresores pueden estimarse $2^p - 1$ diferentes regresiones. Si $p = 5$, $2^p - 1 = 31$; pero si $p = 10$, $2^p - 1 = 1023$, y para $p > 20$ habría que

realizar por encima de un millón de regresiones. Hay procedimientos para reducir y agilizar el cálculo[7], pero aún así éste puede resultar excesivo.

12.2.2. Regresión escalonada *(stepwise regression)*.

Se trata de un procedimiento muy utilizado que, aunque no garantiza obtener la mejor ecuación de regresión, suministra modelos que habitualmente son óptimos o muy próximos al óptimo, con muy poco trabajo por parte del analista. Describiremos el procedimiento de regresión escalonada "hacia adelante" *(forward selection procedure)*; la regresión escalonada "hacia atrás" *(backward elimination)* o mixta son variantes fáciles de entender.

En cada momento, tendremos una ecuación de regresión provisional, que incluye algunas variables (regresores incluidos) y no otras (regresores ausentes). Al comienzo del procedimiento, la ecuación de regresión no incluye ningún regresor. El modo de operar es entonces el siguiente:

1. Calcular los estadísticos Q_h para todos los regresores ausentes ($h\colon \beta_i = 0$).

2. Sea Q_h^* el máximo estadístico de los calculados en 1). Si $Q_h^* < \mathcal{F}$, siendo $\mathcal{F}$ un umbral prefijado, finalizar; la ecuación provisional es la definitiva. Si, por el contrario, $Q_h^* \geq \mathcal{F}$, se introduce la variable correspondiente en la ecuación de regresión.

3. Si no quedan regresores ausentes, finalizar el procedimiento. En caso contrario, reiniciar los cálculos en 1).

En suma, se trata de introducir las variables de una en una, por orden de mayor contribución a disminuir SSE, y mientras la disminución sea apreciable.

El procedimiento de regresion "hacia atrás" procede de manera análoga, pero se comienza con una ecuación que incluye todos los regresores, y se van excluyendo de uno en uno, mientras el incremento en SSE que dicha exclusión origine no sea excesivo. En el procedimiento míxto, por fin, se alterna la inclusión y exclusión de variables en la recta de regresión; ello permite que una variable incluida sea posteriormente desechada cuando la presencia de otra u otras hacen su contribución a la reducción de SSE insignificante.

Los criterios de entrada y salida de variables se fijan especificando sendos valores $\mathcal{F}_{\text{entrada}}$ y $\mathcal{F}_{\text{salida}}$ que deben ser superados (no alcanzados) por el Q_h^* correspondiente para que una variable pueda ser incluida (excluida)

[7]Véase Seber (1977), pag. 349 y ss.

en la regresión. Ambos umbrales pueden ser el mismo. Mediante su selección adecuada, puede lograrse un algoritmo "hacia adelante" puro (fijando $\mathcal{F}_{\text{salida}} = 0$, con lo que se impide el abandono de cualquier variable introducida), "hacia atrás" puro (fijando $\mathcal{F}_{\text{entrada}}$ muy grande, y comenzando con una ecuación de regresión que incluye todas las variables), o un procedimiento mixto arbitrariamente próximo a cualquiera de los dos extremos[8].

R: Ejemplo 12.1 *(selección automática de modelos)* El ejemplo siguiente muestra el uso de las funciones `leaps` (en el paquete del mismo nombre) para hacer regresión sobre todos los subconjuntos con criterios R^2, $\overline{R}^2$ ó C_p, `stepAIC` (en el paquete `MASS`) para hacer regresión escalonada con criterio AIC y algunas otras funciones ancilares.

Orimero generamos datos sintéticos del modo habitual. Como puede verse, hay muchos betas no significativos.

```
> set.seed(123457)
> X <- matrix(rnorm(1000),
+     ncol = 20)
> betas <- rep(0, 20)
> betas[c(3, 5, 7, 12)] <- 1:4
> y <- X %*% betas + rnorm(50)
> datos <- as.data.frame(cbind(X,
+     y))
> dimnames(datos)[[2]][21] <- "y"
> completo <- lm(y ~ ., datos)
```

Como puede verse, hay muchos betas no significativos:

```
> summary(completo)

Call:
lm(formula = y ~ ., data = datos)

Residuals:
```

[8]Podría pensarse en fijar niveles de significación para la entrada y salida de variables. Esto no se hace porque serían considerablemente arduos de computar; obsérvese que en un procedimiento *stepwise* se selecciona para entrar o salir de la ecuación de regresión la variable con un Q_h mayor (menor). Bajo la hipótesis de nulidad del correspondiente parámetro, un Q_h cualquiera se distribuye como una $\mathcal{F}$ de Snedecor con grados de libertad apropiados. *El mayor* (o menor) de los estadísticos Q_h en cada etapa, sigue una distribución diferente (véase Capítulo 9). El nivel de significación asociado al contraste implícito en la inclusión o exclusión de un regresor *no es* la probabilidad a la derecha (o izquierda) de $\mathcal{F}_{\text{entrada}}$ (o $\mathcal{F}_{\text{salida}}$) en una distribución $\mathcal{F}$ con grados de libertad apropiados.

```
   Min     1Q Median     3Q
-1.916 -0.550 -0.107  0.829
   Max
 2.204

Coefficients:
            Estimate Std. Error
(Intercept)  -0.0706     0.2227
V1            0.0408     0.2422
V2            0.1720     0.2603
V3            1.1884     0.2397
V4           -0.0238     0.2067
V5            2.0035     0.2022
V6            0.2633     0.2217
V7            2.9970     0.1875
V8           -0.1074     0.2804
V9            0.0514     0.2105
V10          -0.2367     0.2148
V11          -0.2053     0.2042
V12           4.0374     0.2212
V13           0.1137     0.2161
V14          -0.2115     0.2163
V15           0.0191     0.3076
V16           0.1206     0.2328
V17           0.0318     0.1972
V18          -0.0786     0.2108
V19           0.0879     0.2569
V20           0.0162     0.1949
            t value Pr(>|t|)
(Intercept)   -0.32     0.75
V1             0.17     0.87
V2             0.66     0.51
V3             4.96  2.9e-05 ***
V4            -0.11     0.91
V5             9.91  8.1e-11 ***
V6             1.19     0.24
V7            15.98  6.5e-16 ***
V8            -0.38     0.70
V9             0.24     0.81
V10           -1.10     0.28
V11           -1.01     0.32
V12           18.25  < 2e-16 ***
V13            0.53     0.60
```

```
V14          -0.98     0.34
V15           0.06     0.95
V16           0.52     0.61
V17           0.16     0.87
V18          -0.37     0.71
V19           0.34     0.73
V20           0.08     0.93
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.2 on 29 degrees of freedom
Multiple R-squared: 0.977,        Adjusted R-squared: 0.961
F-statistic:   61 on 20 and 29 DF,  p-value: <2e-16
```

Utilizamos ahora la función `leaps` para hacer regresión sobre todos los subconjuntos. Con 15 regresores, es un problema de talla modesta.

```
> library(leaps)
> mods <- leaps(x = X, y = y,
+     method = "Cp")
```

El objeto `mods` contiene información sobre todos los modelos estimados. Podemos ver como varía C_p y $\overline{R}^2$ con el número de regresores:

```
> postscript(file = "demo10.eps",
+     horizontal = FALSE, width = 5,
+     height = 9)
> opar <- par()
> par(mfrow = c(2, 1))
> plot(mods$size, mods$Cp,
+     main = "Cp versus talla modelos",
+     xlab = expression(p),
+     ylab = expression(C[p]))
> mods.r <- leaps(x = X, y = y,
+     method = "adjr2")
> plot(mods.r$size, mods.r$adjr2,
+     main = "R2 versus talla modelos",
+     xlab = expression(p),
+     ylab = expression(bar(R)^2))
> par(opar)
> dev.off()
```

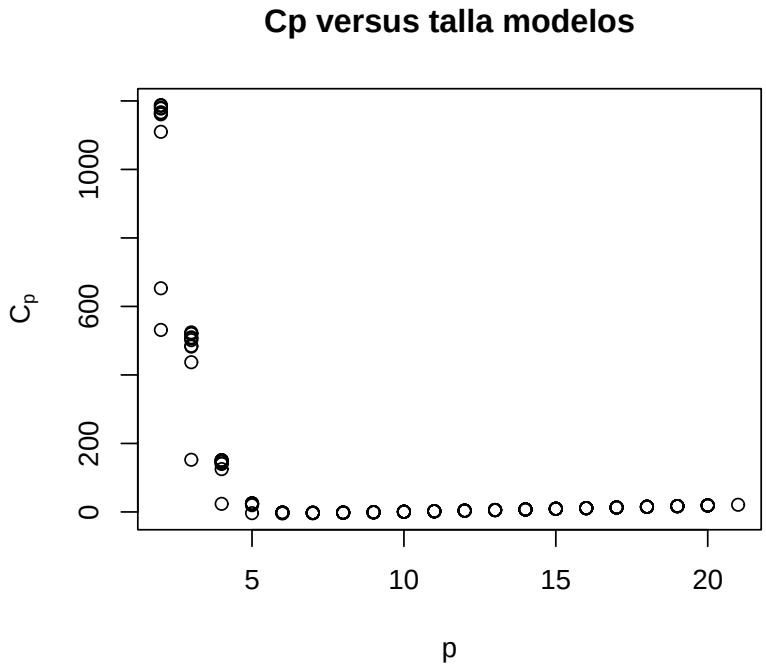
```
X11cairo
       2
```

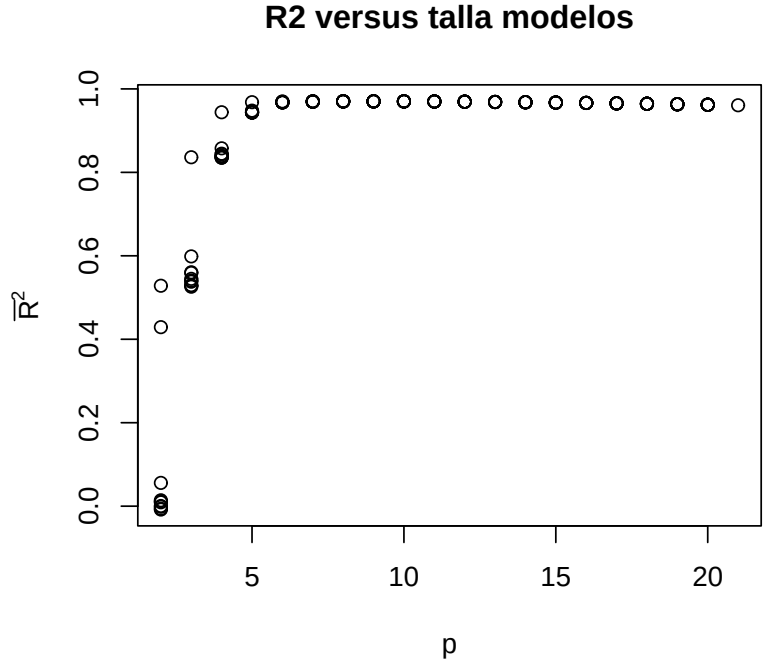
La Figura 13.1 muestra el comportamiento típico de los criterios C_p y $\overline{R}^2$. Se aprecia que, aunque de forma no muy notoria en este caso, el criterio $\overline{R}^2$ tiende a seleccionar modelos más parametrizados.

```
> mejores <- order(mods$Cp)[1:15]
> regres <- mods$which[mejores,
+     ]
> dimnames(regres)[[2]] <- dimnames(datos)[[2]][1:20]
> Cp <- mods$Cp[mejores]
> cbind(regres, Cp)
```

```
  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10
5  0  0  1  0  1  1  1  0  0   0
6  0  0  1  0  1  1  1  0  0   0
6  0  0  1  0  1  1  1  0  0   1
4  0  0  1  0  1  0  1  0  0   0
6  0  0  1  0  1  1  1  0  0   0
5  0  0  1  0  1  0  1  0  0   1
6  0  0  1  0  1  1  1  0  0   0
5  0  0  1  0  1  0  1  0  0   0
7  0  0  1  0  1  1  1  0  0   1
6  0  0  1  0  1  1  1  0  0   0
6  1  0  1  0  1  1  1  0  0   0
5  1  0  1  0  1  0  1  0  0   0
6  0  0  1  0  1  1  1  0  0   0
7  0  0  1  0  1  1  1  0  0   0
6  0  0  1  0  1  1  1  0  0   0
  V11 V12 V13 V14 V15 V16 V17
5   0   1   0   0   0   0   0
6   0   1   0   1   0   0   0
6   0   1   0   0   0   0   0
4   0   1   0   0   0   0   0
6   1   1   0   0   0   0   0
5   0   1   0   0   0   0   0
6   0   1   0   0   0   0   0
5   1   1   0   0   0   0   0
7   0   1   0   1   0   0   0
6   0   1   0   0   1   0   0
6   0   1   0   0   0   0   0
5   0   1   0   0   0   0   0
```

Figura 12.1: Valores de C_p y $\overline{R}^2$ para 141 modelos ajustados a los datos `UScrime`





```
6   0   1   0   0   0   0   1
7   1   1   0   1   0   0   0
6   0   1   1   0   0   0   0
  V18 V19 V20     Cp
5   0   0   0 -4.225
6   0   0   0 -3.491
6   0   0   0 -3.455
4   0   0   0 -3.453
6   0   0   0 -3.213
5   0   0   0 -3.150
6   0   1   0 -2.654
5   0   0   0 -2.550
7   0   0   0 -2.548
6   0   0   0 -2.518
6   0   0   0 -2.476
5   0   0   0 -2.405
6   0   0   0 -2.368
7   0   0   0 -2.365
6   0   0   0 -2.335
```

```
> mod1 <- lm(y ~ V3 + V4 +
+     V5 + V7 + V10 + V12 +
+     V16 + V17, data = datos)
> mod2 <- update(mod1, . ~
+     . + V1 + V2)
> summary(mod2)

Call:
lm(formula = y ~ V3 + V4 + V5 + V7 + V10 + V12 + V16 + V17 +
    V1 + V2, data = datos)

Residuals:
   Min     1Q Median     3Q
-1.611 -0.762  0.122  0.627
   Max
 2.237

Coefficients:
            Estimate Std. Error
(Intercept) -0.03573    0.18316
V3           1.08674    0.19721
V4          -0.00741    0.16766
V5           2.03931    0.16976
V7           3.05622    0.14772
```

```
V10         -0.27977    0.19088
V12          4.10685    0.18483
V16          0.08436    0.15101
V17          0.05185    0.14567
V1           0.16370    0.18257
V2          -0.00659    0.20666
            t value Pr(>|t|)
(Intercept)   -0.20     0.85
V3             5.51  2.5e-06 ***
V4            -0.04     0.96
V5            12.01  1.1e-14 ***
V7            20.69  < 2e-16 ***
V10           -1.47     0.15
V12           22.22  < 2e-16 ***
V16            0.56     0.58
V17            0.36     0.72
V1             0.90     0.38
V2            -0.03     0.97
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.11 on 39 degrees of freedom
Multiple R-squared: 0.973,        Adjusted R-squared: 0.966
F-statistic:  141 on 10 and 39 DF,  p-value: <2e-16
```

```
> mod3 <- update(mod1, . ~
+      . - V10 - V16 - V17)
> summary(mod3)

Call:
lm(formula = y ~ V3 + V4 + V5 + V7 + V12, data = datos)

Residuals:
    Min      1Q  Median      3Q
-2.0289 -0.6955  0.0539  0.7177
    Max
 2.5956

Coefficients:
            Estimate Std. Error
(Intercept)   0.0738     0.1596
V3            1.0693     0.1819
V4           -0.0410     0.1567
V5            1.9898     0.1603
```

```
V7            3.0484     0.1400
V12           4.1357     0.1642
            t value Pr(>|t|)
(Intercept)    0.46     0.65
V3             5.88  5.1e-07 ***
V4            -0.26     0.79
V5            12.41  5.7e-16 ***
V7            21.77  < 2e-16 ***
V12           25.19  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.09 on 44 degrees of freedom
Multiple R-squared: 0.971,        Adjusted R-squared: 0.967
F-statistic:  293 on 5 and 44 DF,  p-value: <2e-16
```

```
> m <- regsubsets(y ~ ., datos,
+     method = "forward")
> summary(m)

Subset selection object
Call: regsubsets.formula(y ~ ., datos, method = "forward")
20 Variables  (and intercept)
    Forced in Forced out
V1      FALSE      FALSE
V2      FALSE      FALSE
V3      FALSE      FALSE
V4      FALSE      FALSE
V5      FALSE      FALSE
V6      FALSE      FALSE
V7      FALSE      FALSE
V8      FALSE      FALSE
V9      FALSE      FALSE
V10     FALSE      FALSE
V11     FALSE      FALSE
V12     FALSE      FALSE
V13     FALSE      FALSE
V14     FALSE      FALSE
V15     FALSE      FALSE
V16     FALSE      FALSE
V17     FALSE      FALSE
V18     FALSE      FALSE
V19     FALSE      FALSE
V20     FALSE      FALSE
```

```
1 subsets of each size up to 8
Selection Algorithm: forward
         V1  V2  V3  V4  V5  V6
1  ( 1 ) " " " " " " " " " " " "
2  ( 1 ) " " " " " " " " " " " "
3  ( 1 ) " " " " " " " " "*" " "
4  ( 1 ) " " " " "*" " " "*" " "
5  ( 1 ) " " " " "*" " " "*" "*"
6  ( 1 ) " " " " "*" " " "*" "*"
7  ( 1 ) " " " " "*" " " "*" "*"
8  ( 1 ) " " " " "*" " " "*" "*"
         V7  V8  V9  V10 V11 V12
1  ( 1 ) " " " " " " " " " " "*"
2  ( 1 ) "*" " " " " " " " " "*"
3  ( 1 ) "*" " " " " " " " " "*"
4  ( 1 ) "*" " " " " " " " " "*"
5  ( 1 ) "*" " " " " " " " " "*"
6  ( 1 ) "*" " " " " " " " " "*"
7  ( 1 ) "*" " " " " "*" " " "*"
8  ( 1 ) "*" " " " " "*" " " "*"
         V13 V14 V15 V16 V17 V18
1  ( 1 ) " " " " " " " " " " " "
2  ( 1 ) " " " " " " " " " " " "
3  ( 1 ) " " " " " " " " " " " "
4  ( 1 ) " " " " " " " " " " " "
5  ( 1 ) " " " " " " " " " " " "
6  ( 1 ) " " "*" " " " " " " " "
7  ( 1 ) " " "*" " " " " " " " "
8  ( 1 ) " " "*" " " " " " " " "
         V19 V20
1  ( 1 ) " " " "
2  ( 1 ) " " " "
3  ( 1 ) " " " "
4  ( 1 ) " " " "
5  ( 1 ) " " " "
6  ( 1 ) " " " "
7  ( 1 ) " " " "
8  ( 1 ) "*" " "
```

```
> library(MASS)
> step <- stepAIC(completo,
+     scope = y ~ ., direction = "both",
+     trace = FALSE)
> summary(step)
```

```
Call:
lm(formula = y ~ V3 + V5 + V6 + V7 + V12, data = datos)

Residuals:
    Min      1Q  Median      3Q
-1.9495 -0.6503 -0.0349  0.5244
    Max
 2.6196

Coefficients:
            Estimate Std. Error
(Intercept)   0.0514     0.1518
V3            1.0256     0.1761
V5            2.0499     0.1557
V6            0.3046     0.1603
V7            3.0499     0.1346
V12           4.1077     0.1585
            t value Pr(>|t|)
(Intercept)    0.34    0.736
V3             5.82  6.1e-07 ***
V5            13.17  < 2e-16 ***
V6             1.90    0.064 .
V7            22.65  < 2e-16 ***
V12           25.91  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.05 on 44 degrees of freedom
Multiple R-squared: 0.973,        Adjusted R-squared: 0.97
F-statistic:  317 on 5 and 44 DF,  p-value: <2e-16
```

FIN DEL EJEMPLO ■

12.3. Modelos bien estructurados jerárquicamente

La facilidad con que los algoritmos presentados en este Capítulo producen modelos candidatos no debe hacer que el analista delegue demasiado en ellos. Un modelo ha de ser consistente con los conocimientos fiables que se tengan

acerca del fenómeno bajo estudio. Debe ser también interpretable. Prestemos algo de atención a este último requerimiento.

Imaginemos un modelo como el siguiente:

$$y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon. \tag{12.28}$$

En un caso así, frecuentemente el interés se centrará en dilucidar si la relación de X con Y es lineal o cuadrática —es decir, en contrastar la hipótesis $h : \beta_2 = 0$—.

Es frecuentemente el caso que X se mide en unidades en que tanto la escala como el origen son arbitrarios (como ocurría, por ejemplo, en el Ejercicio 3.10, pág. 41); y sería inconveniente que el contraste de h dependiera del origen y de la escala empleadas. Lo menos que debemos esperar de nuestra inferencia es que sea invariante frente a cambios en las unidades de medida.

Si en (13.28) reemplazamos X por $Z = aX + b$, obtenemos

$$\begin{aligned} y &= \beta_0 + \beta_1(aX + b) + \beta_2(aX + b)^2 + \epsilon \\ &= (\beta_0 + \beta_1 b + \beta_2 b^2) + (\beta_1 a + 2ab\beta_2)X + a^2\beta_2 X^2 + \epsilon \\ &= \beta_0^* + \beta_1^* X + \beta_2^* X^2 + \epsilon. \end{aligned} \tag{12.29}$$

En este nuevo modelo, $\beta_2^* = a^2\beta_2$ absorbiendo el cambio de escala en la X. Es fácil ver que es equivalente contrastar $h : \beta_2 = 0$ en (13.28) o $h : \beta_2^* = 0$ en (13.29); el contraste de la hipótesis "efecto cuadrático de X sobre Y", al menos, no se altera por el cambio de unidades. Sin embargo, sean cuales fueren β_1 y β_2, habrá coeficientes a, b anulando $\beta_1^* = (\beta_1 a + 2ab\beta_2)$ en (13.29). Ello hace ver que:

- No tiene sentido contrastar efecto lineal en un modelo que incluye término cuadrático, porque el contraste tendría un resultado diferente dependiendo de las unidades de medida.

- La inclusión de un término en X^2 *debe* ir acompañada de un término lineal y constante, si queremos que el modelo sea invariante frente a cambios en el origen y la escala.

La conclusión que extraemos es que los términos de orden superior deben estar acompañados de todos los términos de orden inferior —es decir, si incluimos un término cúbico, deben también existir términos cuadráticos y lineales, etc.—. Un modelo que cumpla con dicho requisito se dice que está jerárquicamente estructurado y en él podemos contrastar no nulidad del coeficiente del término jerárquico de orden superior, pero no de los inferiores. La misma conclusión es de aplicación a términos recogiendo interacciones:

si introducimos una variable compuesta como X_iX_j en el modelo, X_i y X_j deben también ser incluidas. Se suele decir que un modelo jerárquicamente bien estructurado verifica *restricciones de marginalidad* y que, por ejemplo, X_i y X_j son ambas marginales a X_iX_j.

Si regresamos al Ejercicio 3.10 en que se argüía la necesidad de utilizar un término β_0 veremos que se trata del mismo problema: necesitamos el término jerárquico inferior (la constante) cuando incluimos X dado que las unidades y el origen son arbitrarios. No es imposible que un modelo sin β_0 sea adecuado, pero lo normal es lo contrario.

Dependiendo de los programas que se utilicen, un algoritmo puede eliminar del modelo de regresión un término jerárquico inferior manteniendo otro de orden superior. Es responsabilidad del analista garantizar que ello no ocurra, manteniendo la interpretabilidad de los parámetros en toda circunstancia.

Complementos y ejercicios

12.1 Supongamos que hacemos regresión escalonada "hacia adelante". ¿Qué valor de $\mathcal{F}_{\text{entrada}}$ equivaldría a introducir regresores en el modelo en tanto en cuanto incrementen $\overline{R}_p^2$?

12.2 Las estrategias de regresión escalonada descritas (hacia adelante, hacia atrás, o mixta) exploran un subconjunto de los modelos posibles, añadiendo (omitiendo) en cada momento el regresor que parece con mayor (menor) capacidad explicativa de la variable respuesta. Puede perfectamente alcanzarse un óptimo local, al llegarse a un modelo en el que no es posible mejorar el criterio elegido (C_p, o cualquier otro) añadiendo u omitiendo regresores, pese a existir otro modelo mejor en términos de dicho criterio. ¿Mejoran nuestras expectativas de encontrar el óptimo global mediante regresión escalonada cuando las columnas de la matriz X de regresores son ortogonales? Justifíquese la respuesta.

12.3 En la Observación 13.1 se comparan los criterios de selección de modelos consistentes en maximizar $\overline{R}_p^2$ y C_p, viendo que el segundo es en general más restrictivo.

Consideremos ahora dos posibles modelos A y B de regresión con sumas de cuadrados de los residuos respectivamente SSE_A y SSE_B. El primer modelo utiliza sólo un subconjunto de los regresores presentes en el segundo (por tanto, $SSE_A \geq SSE_B$).

Para escoger entre los modelos A y B podríamos adoptar uno de los siguientes criterios:

1. Seleccionar el modelo B si la disminución en la suma de cuadrados respecto al modelo A es estadísticamente significativa, es decir, si:
$$Q_h = \frac{(SSE_A - SSE_B)}{q\hat{\sigma}^2} > \mathcal{F}^{\alpha}_{q,N-(p+q)}$$
siendo p el número de parámetros presentes en A y q el de los adicionales presentes en B.
2. Seleccionar el modelo B si su estadístico C_p es menor.

Supongamos además que el modelo B es el más parametrizado de los posibles (incluye todas las variables de que disponemos). ¿Qué relación existe entre ambos criterios?