correlacionado con la respuesta . Sea éste $x_{j1}$. En la dirección de este predictor se realiza el mayor *salto* posible hasta que otro predictor, $x_{j2}$, presente una correlación parecida a la anterior, con los residuos del modelo anteriormente resultante. En este punto, en lugar de seguir en la dirección del $x_{j1}$, LARS procede en una dirección *equian - gular* entre ambos predictores, hasta que una tercera variable $x_{j3}$ entre en el círculo de los regresores *más correlacionados*, en cuyo caso, LARS busca la dirección *equiangular* entre $x_{j1}$,$x_{j2}$,$x_{j3}$, es decir, la dirección *least angle* hasta que una cuarta variable $x_{j4}$ entra en el círculo o conjunto de los *más correlacionados*. Y así sucesivamente, hasta agotar el número de regresores disponible.

De esta manera, LARS va construyendo $\hat{y} = X\hat{\beta}$ en sucesivas etapas, añadiendo una variable más cada vez. Y, en consecuencia, tras *k* etapas, sólo hay *k* coeficientes de los $\hat{\beta}$ distintos de cero en el modelo.

## IX. Grados de libertad de Lars y el estadistico C,

Para determinar el número de coeficientes, y por tanto, de variables a incluir en un modelo, es necesario un criterio de selección. En nuestro caso va a ser del tipo $C_p$ de Mallows (1973).

Un modelo bueno debería predecir bien. Por tanto el error cuadrático medio de la predicción puede ser un criterio a tener presente a la hora de juzgar la idoneidad de un modelo.

Tenemos así el estadístico $C_p$. Este estadístico fue ideado por Colin Mallows y Cuthbert Daniel en 1973, de forma que tuviera pequeños valores enteros como resultado.

La definición original de $C_p$ es la siguiente:

$$C_p = \frac{RSS_p}{\hat{\sigma}_c^{\,2}} + 2p - n \qquad [10]$$

siendo $RSS_p$ la suma de residuos al cuadrado del modelo con *p* regresores; $\hat{\sigma}_c^{\,2}$ el estimador de la varianza del término de error del modelo completo (el modelo con todos los posibles regresores); *p* el número de regresores del modelo en consideración, y *n* el número de observaciones por predictor.

Si un modelo predice bien, entonces $C_p \simeq p$. Si un modelo predice mal, $C_p$ será mucho mayor que $p$. En definitiva, el modelo a tener en cuenta es el que tenga un $p$ pequeño, y un valor de $p$ en torno o por debajo de $p$.

Vengamos a nuestro caso.

Sea $\hat{y} - f(y)$, y $x_1, x_2, ..., x_m$, un conjunto de $m$ predictores, fijados en sus valores observados. Supondremos que la $y$ es homoscedástica y generada por el modelo $y \sim (E(y), \sigma^2 I)$, lo que significa que los valores de la están incorrelacionados, con media $E(y)$ y varianza $\sigma^2$.

Efron et al. (2004, p. 423), demuestran que los grados de libertad ($gl$) del estimador $\hat{y} - f(y)$, pueden escribirse como

$$gl - \sum \text{cov}(\hat{y}_i, y_i) / \sigma^2 \qquad [11]$$

y la fórmula del estadístico tipo riesgo Cp pasa a ser

$$C_p(\hat{y}) = \frac{\| y - \hat{y} \|^2}{\sigma^2} - n + 2(gl) \qquad [12]$$

Si $\sigma^2$ y los $gl$ fueran conocidos, entonces $C_p(\hat{y})$ sería un estimador insesgado del verdader riesgo.

Tratándose de un estimador lineal $\hat{y} - My$, los grados de libertad coinciden con el valor de la traza de la matrız $M$, por lo que los grados de libertad coinciden con los de la estimación por mínimos cuadrados ordinarios, y, a la vez, con la propuesta de Mallows (1973) sobre $C_p$.

La aplicación de $C_p$ requiere la estimación previa de $\hat{y}$, $\sigma^2$ y de los $gl$. La importante conclusión es que

$$gl(\hat{y}) \simeq k \qquad [13]$$

lo que permite escribir

$$C_p \simeq \frac{\| y - \hat{y} \|^2}{\hat{\sigma}^2} - n + 2k \qquad [14]$$

fórmula que coincide con el estimador $C_p$ en la estimación mínimo cuadrática, sobre la base de $k$ predictores, y con la gran ventaja de que LARS no necesita de cálculos adicionales.

Efron et al. (2004) recalcan que la fórmula del $C_p$ anterior sólo es aplicable a LARS, no a *Lasso* ni a S*tagewise*, salvo que se cumpla la condición de que la matriz $M$ de predictores sea ortogonal (condición del *cono positivo*).

## X. Grados de libertad en Lasso

En la ya mencionada investigación de Zuo, Hastie y Tibshrani (2004), no publicada pero disponible en la WEB, se profundiza en los grados de libertad de este algoritmo, en el sentido de extender y fundamentar lo que en Efron et al. (2004) se formulaba como una conjetura. Aquí se muestra que el número de coeficientes distintos de cero es un estimador insesgado de los grados de libertad del procedimiento *Lasso*.

Así mismo, se pasa revista a otros criterios de selección de variables como son: el $C_p$, ya mencionado, el AIC (Akaike Information Criterion) y el BIC (Bayesian Information Criterion). La utilización de estos criterios dentro del algoritmo LARS, proporcionan unos estadísticos eficientes para la selección de modelos. Un resultado importante en este trabajo, es la propuesta del *BIC-Lasso* como el principal criterio de selección de variables.

A la luz de esta investigación, las fórmulas para el AIC y el BIC son las siguientes:

$$AIC(\hat{y}) = \frac{\| y - \hat{y} \|^2}{n\sigma^2} + \frac{2}{n} gl(\hat{y}) \qquad [15]$$

y

$$BIC(\hat{y}) = \frac{\| y - \hat{y} \|^2}{n\sigma^2} + \frac{\log(n)}{n} gl(\hat{y}) \qquad [16]$$

y como el estimador de los grados de libertad: $gl(\hat{y})$, es el número de predictores en el modelo: $k$, basta con sustituir este valor en las fórmulas anteriores para obtener el valor de los respectivos estadísticos.

Como se recuerda en el trabajo, AIC y BIC poseen diferentes propiedades asintóticas. Si el verdadero modelo de regresión no está entre el conjunto de modelos candidatos, el criterio AIC, en el límite, alcanza el menor error cuadrático medio entre los modelos candidatos, y el estimador del AIC converge a la tasa óptima en el sentido del *minimax*, tanto si el el verdadero modelo está o no está entre los candidatos. Por otro lado, el criterio BIC selecciona el verdadero modelo de una manera consistente. Si el verdadero modelo está entre el conjunto de modelos candidatos, la probabilidad de seleccionar el verdadero modelo según BIC se acerca a la unidad al aumentar el tamaño de la muestra: $n \to \infty$.

Desde el punto de vista de la supresión de coeficientes, el *AIC-lasso* tiende a incluir más coeficientes no-cero de los verdaderos, mientras que el *BIC-lasso* is más preciso cuando se trata de establecer el número de variables a incluir en el modelo. Zuo et al., (2004), páginas 17 y 18.

Otro resultado al que se llega en Zuo et al. (2004) es el reconocimiento de que el establecimiento de un criterio que combinara las habilidades del AIC y BIC no es viable. Es decir, un criterio de selección que sea a la vez consistente y óptimo, cuando el estadístico de referencia es el error cuadrático medio, no es factible: el criterio de selección tiene que optar entre el óptimo en la predicción y la consistencia, al elegir el modelo.

## XI. LARS EN ACCIÓN

Vamos a presentar de un modo práctico el comportamiento de LARS. Lo haremos utilizando un conjunto de datos famoso por la alta correlación presente entre sus variables. Nos referimos a los datos de Longley.

Una de las razones para considerar estos datos se debe al hecho de que en una obra reciente de Heiberger y Holland: *Statistical Analysis and Data Display* (2004), los datos de Longley son utilizados para ilustrar el proceso de selección de variables en un modelo. Heiberger y Holland primero realizan una selección manual del modelo, y luego presentan el procedimiento de selección automática *forward*, con ayuda de los programas estadísticos S-PLUS y SAS. El modelo seleccionado automáticamente no les satisface, y ello les

sirve de motivo para destacar la superioridad del procedimiento manual sobre el automático.

Nosotros utilizamos el paquete LARS, implementado en el programa estadístico R, para la determinación del modelo más apropiado a estos datos, que completamos manualmente, para llegar a un modelo que no coincide con el de Heiberger y Holland, pero que tiene mayor contenido económico que el propuesto por ellos.

## XII. LOS DATOS DE LONGLEY

Longley publicó su muestra de datos en 1967. Las variables recogen datos de la economía norteamericana referentes al PIB, al deflator del PIB (1954=100), a la cifra de desempleados, a los trabajadores en las Fuerzas Armadas, a la población, a las cifras de personas empleadas, y al número de años: 1947-1962, un total de 16 años. El modelo a estimar relaciona la cifra de personas empleadas con el resto de las variables.

Para simplificar la presentación de resultados, renombramos las variables de la muestra original como sigue:

1. GNP deflator: $x_1$
2. GNP: $x_2$
3. Unemployment: $x_3$
4. Armed.Forces: $x_4$
5. Population: $x_5$
6. Year: $x_6$
7. Employed: $y$.

El modelo lineal con la variable $y$ como respuesta, e incluyendo todas las variables, el **modelo 1**, es ahora:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + \beta_5 x_{5t} + \beta_6 x_{6t} + u_t$$

siendo $u_t$ el término de error del modelo.

El modelo estimado da los resultados siguientes.

```
summary(modelo1, cor=T)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41011 -0.15767 -0.02816  0.10155  0.45539

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.482e+03  8.904e+02  -3.911 0.003560 **
x1           1.506e-02  8.492e-02   0.177 0.863141
x2          -3.582e-02  3.349e-02  -1.070 0.312681
x3          -2.020e-02  4.884e-03  -4.136 0.002535 **
x4           1.033e 02  2.143e 03   4.822 0.000944 ***
x5          -5.110e-02  2.261e-01  -0.226 0.826212
x6           1.829e+00  4.555e-01   4.016 0.003037 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3049 on 9 degrees of freedom
Multiple R-Squared: 0.9955,     Adjusted R-squared: 0.9925
F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10

Correlation of Coefficients:
   (Intercept) x1    x2    x3    x4    x5
x1 -0.20
x2  0.82       -0.65
x3  0.84       -0.55  0.95
x4  0.55        0.35  0.47  0.62
x5 -0.41        0.66 -0.83 -0.76 -0.19
x6 -1.00        0.19 -0.80 -0.82 -0.55  0.39

vif(modelo1)
        x1         x2         x3         x4         x5         x6
 135.53244 1788.51348   33.61889    3.58893  399.15102  758.98060
```

Las estimaciones muestran claramente las consecuencias de la multicolinealidad presente, enfatizadas por el alto valor del *vif: variance inflation factor*, de las variables.

Invocamos LARS, y como resultado obtenemos:

```
R-squared: 0.995
Sequence of LAR moves:
     x2 x3 x4 x6 x1 x5
Var   2  3  4  6  1  5
Step  1  2  3  4  5  6
```

Esta tabla nos muestra el orden en el que LARS ha incluido las variables en el modelo lineal. Es decir, el orden en que los coeficientes de las mismas han dejado de valer cero.

La evolución gráfica del proceso, viene recogida en la figura I.

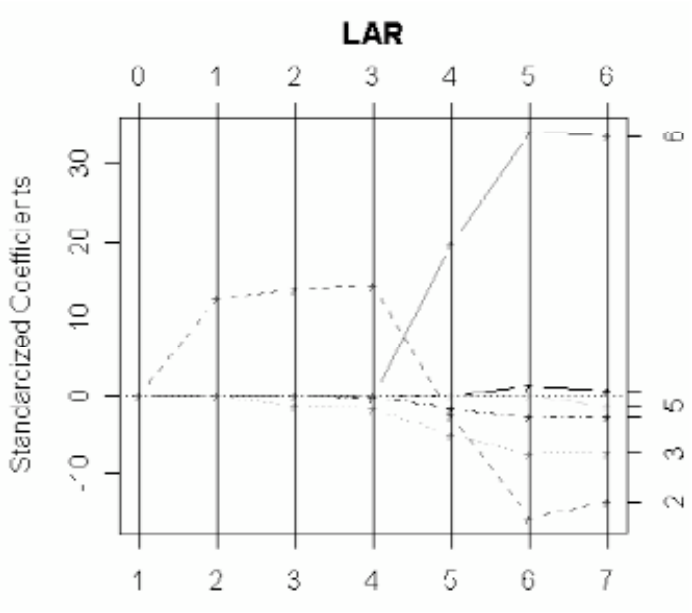


FIGURA I

Para mejor interpretar esta figura, conviene recordar que *Df: degrees of freedom*, coincide y representa el número de etapas.

Veamos los valores de los estadísticos , AIC y BIC:

```
Cp
     0           1          2           3          4          5
1976.712035  59.471238  31.783215  29.316469  8.546093  5.054627
6
7.000000

BIC
     0          1         2          3          4         5          6
124.419502  4.640239  2.958025  2.852140  1.602270  1.432348 1.602221

AIC
     0          1         2          3          4         5          6
124.419502  4.591952  2.861451  2.707279 1.409131  1.190914   1.312500
```

Dado que los tres estadísticos coinciden en alcanzar el valor mínimo en la quinta etapa, establezcamos el siguiente modelo de regresión, incluyendo en él las cinco primeras variables seleccionadas por LARS. Es decir:

$$y_t = \beta_0 + \beta_1 x_{2t} + \beta_2 x_{3t} + \beta_3 x_{4t} + \beta_4 x_{6t} + \beta_5 x_{1t} + u_t$$

Es el **modelo 2**. Al estimarlo por mínimos cuadrados ordinarios, obtenemos:

```
summary(modelo2,cor=T)

Residuals:
     Min       1Q   Median       3Q      Max
 0.39012  0.14342  0.03560  0.09728  0.46136

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.566e+03  7.724e+02  -4.615 0.000957 ***
x2           4.213e 02  1.762e 02   2.391 0.037891 *
x3          -2.104e-02  3.029e-03  -6.945 3.97e-05 ***
x4          -1.042e-02  2.002e-03  -5.207 0.000397 ***
x6           1.869e+00  3.994e-01   4.680 0.000867 ***
x1           2.772e-02  6.075e-02   0.456 0.657984

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.29 on 10 degrees of freedom
Multiple R-Squared: 0.9955,     Adjusted R-squared: 0.9932
F-statistic: 437.9 on 5 and 10 DF,  p-value: 2.266e 11

Correlation of Coefficients:
   (Intercept) x2    x3    x4    x6
x2  0.94
x3  0.88        0.87
x4  0.53        0.57  0.74
x6 -1.00       -0.94 -0.88 -0.53
x1  0.10       -0.24 -0.11 -0.30 -0.10

vif(modelo2)
        x2         x3         x4         x6         x1
546.870494  14.289620   3.460846 644.626426  76.641401
```

Estos resultados muestran que este segundo modelo tampoco es satisfactorio. Eliminemos la variable $x_6$ dado el alto valor de su *vif*, y la correlación negativa perfecta de su coeficiente con el término constante del modelo, además de la alta correlación con otros coeficientes.

El nuevo modelo, **modelo 3**, pasa a ser:

$$y_t = \beta_0 + \beta_1 x_{2t} + \beta_2 x_{3t} + \beta_3 x_{4t} + \beta_4 x_{1t} + u_t$$

Su estimación presenta el siguiente resultado:

```
summary(modelo3,cor=T)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70231 -0.22605 -0.01062  0.15518  1.08617

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.083570   5.942719   8.428 3.97e-06 ***
x2           0.035263   0.010361   3.404  0.00589 **
x3          -0.008538   0.002434  -3.508  0.00490 **
x4          -0.005495   0.002900  -1.895  0.08463 .
x1           0.056263   0.102940   0.547  0.59559
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4939 on 11 degrees of freedom
Multiple R-Squared: 0.9855,     Adjusted R-squared: 0.9802
F-statistic: 186.8 on 4 and 11 DF,  p-value: 4.968e-10

Correlation of Coefficients:
   (Intercept) x2    x3    x4
x2  0.97
x3  0.36        0.26
x4  0.34        0.27  0.69
x1 -0.99       -0.98 -0.43 -0.42

vif(modelo3)
       x2        x3        x4        x1
65.200243  3.180198  2.503275 75.868790
```

Aunque los resultados han mejorado, sigue siendo muy alto el *vif* de $x_1$. Al eliminarla, tenemos el **modelo 4**:

$$y_t = \beta_0 + \beta_1 x_{2t} + \beta_2 x_{3t} + \beta_3 x_{4t} + u_t$$

Al estimarlo, obtenemos:

```
summary(modelo4,cor=T)

Residuals:
     Min       1Q   Median       3Q      Max
 0.83085  0.22306  0.01735  0.10699  1.08090

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.306461   0.716342  74.415  < 2e-16 ***
x2           0.040788   0.002207  18.485 3.49e-10 ***
x3          -0.007968   0.002134  -3.734  0.00285 **
x4          -0.004828   0.002552  -1.892  0.08286 .
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.4793 on 12 degrees of freedom
Multiple R-Squared: 0.9851,     Adjusted R-squared: 0.9814
F-statistic: 264.4 on 3 and 12 DF,  p-value: 3.189e-11

Correlation of Coefficients:
   (Intercept) x2    x3
x2  0.20
x3 -0.61       -0.78
x4 -0.68       -0.71  0.63

vif(modelo4)
      x2       x3       x4
3.140867 2.596610 2.058847
```

Ahora, los resultados de esta estimación son aceptables. El último paso para la validación de este modelo, es el análisis de los residuos, que presentan un comportamiento estadísticamente aceptable: tienen

distribución normal son homoscedásticos y están incorrelacionados. Por razones de espacio, no se reproducen.

## XIII. CONCLUSIONES

El estudio de los datos de Longley ha venido motivado, en parte, por el reciente análisis de Heiberger y Holland (2004), antes aludido. Ellos llegan, manualmente, cf. página 247, a seleccionar el modelo:

$$\begin{array}{rcccc} \hat{y}_t = & 1797.221 & 0.015x_{3t} & 0.008x_{4t} & 0.956x_{6t}; R^2 = 0.993 \\ & (68.642) & (0.002) & (0.002) & (0.036) \end{array}$$

entre paréntesis, los errores estándar de los coeficientes estimados. Los estadísticos *vif* para las variables de este modelo son: 3.318, 2.223 y 3.89, respectivamente.

Utilizando el procedimiento automático *forward selection*, el modelo al que Heiberger y Holland llegan, relaciona el número de empleados con el PIB, con el número de desempleados, con el número de empleados en las Fuerzas Armadas y con el número de años, modelo que no agrada a dichos autores por varias razones de carácter estadístico, y, además, por la relación que este modelo establece entre el número de empleados y el PIB, que es negativa, en contra de toda lógica.

El modelo al que hemos llegado en nuestro estudio, relaciona el número de empleados con el PIB, de una manera positiva, y negativamente con el número de desempleados y con los empleados en las Fuerzas Armadas, algo lógicamente correcto.

Es decir, nuestro modelo final es:

$$\begin{array}{l} \hat{y}_t = 53.30646 + 0.04079x_{2t} - 0.00797x_{3t} - 0.00483x_{4t}; R^2 = 0.985 \\ \quad (0.7163) \quad (0.0022) \quad (0.0021) \quad (0.0026) \end{array}$$

entre paréntesis, los errores estándar de los coeficientes estimados. Este modelo final es consecuencia de un proceso automático, completado manualmente. Conviene poner de relieve que, de esta manera, la selección del mejor modelo no ha resultado ajena al objetivo de establecer un modelo, con sentido económico, a los datos de Longley.

Se corrobora así una de las conclusiones del debate que se reproduce en Efron et al. (2004), y que manifiestan los Autores en su contestación final a los argumentos de los participantes, a saber, que los procedimientos automáticos permiten una «honesta evaluación del error de la estimación», y son un primer acercamiento y un paso positivo en la búsqueda de un modelo apropiado, dada la masa ingente de información disponible.

## XIV. BIBLIOGRAFÍA


CAR, (Companion to Applied Regression), version: 1.0-14 , John Fox, CRAN (The Comprehensive R Archive Network), Viena, 2004.

CHATTERJEE, S.; HADI, A. S., y PRICE, B., *Regression Analysis by Example*, 3.ª ed., John Wiley & Sons, New York; 2000.

EFRON, B. y HASTIE, T., «Least Angle Regression» en *The Annals of Statistics*, 32 (2004) 407-499.

FARAWAY, J., *Linear Models with R*, Chapman & Hall/CRC, Boca Raton, Florida, 2005.

HASTIE, T.; TIBSHIRANI, R., y FRIEDMAN, J., *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, Springer-Verlag, New York 2001.

HEIBERGER, R. M., y HOLLAND, B., *Statistical Analysis and Data Display. An Intermediate Course with Examples in S-PLUS, R, and SAS*, Springer-Verlag, New York, 2004.

JOHNSON, R. A., y WICHERN, D. W., *Applied Multivariate Statistical Analysis*, 5a. ed., Prentice Hall, Inc. Upper Saddle River, NJ 2002.

LARS, (Least Angle Regression, Lasso and Forward Stagewise), version: 0.9-5, Trevor Hastie and Brad Efron, CRAN (The Comprehensive R Archive Network), Viena 2004.

LONGLEY, J. W., «An appraisal of least-squares programs from the point of view of the user», en *Journal of the American Statistical Association*, 62 (1967) 819-841.

MALLOWS, C., «Some comments on Cp», *Technometrics*, 15 (1973) 661-675.

MILLER, A., *Subset Selection in Regression*, 2.ª ed., Chapman & Hall/CRC, Boca Raton, Florida 2002.

OSBORNE, M.; PRESNELL, B., y TURLACH, B., «A new approach to variable selection in least squares problems», en *IMA Journal of Numerical Analysis*, 20 (2000) 389-403.

R, Version 2.0.1, *The R Foundation for Statistical Computing*, CRAN (The Comprehensive R Archive Network), Viena 2004.

TIBSHIRANI, R., «Regression shrinkage and selection via the Lasso», *Journal of the Royal Statistical Society, B*, 58 (1996) 267-288.

WEISBERG, S., *Applied Linear Regression*, 2.ª ed., John Wiley & Sons, New York 1985.
STEIN, C., «Estimation of the mean of a multivariate normal distribution», *Annals of Statistics*, 9 (1981)1135-1151.
ZOU, H.; HASTIE, T.; y TIBSHIRANI, R., *On the «Degrees of Freedom» of the Lasso,* Dept. of Statistics, Stanford University, 2004: www.stat.stanford.edu/~hastie/pub.htm