Tarea 8: Introducción a Ciencia de Datos.

Rojas Gutiérrez Rodolfo Emmanuel

10 de noviembre de 2021

1. Ejercicios.

Ejercicio 1:

Solución. Para la resolución de este ejercicio, sea $X \in M_{N\times p}(\mathbb{R})$ una matriz de covariables, $y \in M_{N\times 1}(\mathbb{R})$ un vector columna de variables independientes y f_λ la función $f_\lambda : M_{p\times 1}(\mathbb{R}) \mapsto \mathbb{R}_+$ con regla de correspondencia:

$$f_\lambda(\beta) = \frac{1}{2N} \|y - X\beta\|^2 + \lambda \|\beta\|_1 , \tag{1}$$

donde, los símbolos $\|\|$ y $\|\|_1$ denotan de manera respectiva, la norma euclidiana en $\mathbb{R}^N$ y la norma l_1 en $\mathbb{R}^p$. Ahora, observe que gracias a la relación que guardan entre si, la norma euclidiana en $\mathbb{R}^N$ y el producto interno en $\mathbb{R}^N$, se tiene para cada $\beta \in M_{p\times 1}(\mathbb{R})$ que

$$\|y - X\beta\|^2 = \|y\|^2 + \|X\beta\|^2 - \langle y, X\beta\rangle . \tag{2}$$

Luego, recordando que para $\beta \in M_{p\times 1}(\mathbb{R})$, el producto $X\beta$ puede expresarse como una combinación lineal de la columnas de X, de la siguiente forma:

$$X\beta = \sum_{j=1}^{n} \beta_j X_j$$

donde, para $j \in \{1, \ldots, p\}$ se tiene que X_j denota a la j-ésima columna de X, mientras que, β_j representa a la j-esima entrada del vector columna β. Así, al sustituir la expresión anterior en la igualdad en (2), se obtiene que:

$$\begin{aligned} \|y - X\beta\|^2 &= \|y\|^2 + \|X\beta\|^2 - 2\left\langle y, \sum_{j=1}^{n} \beta_j X_j \right\rangle \\ &= \|y\|^2 + \|X\beta\|^2 - 2\sum_{j=1}^{n} \beta_j \langle y, X_j\rangle , \end{aligned} \tag{3}$$

donde, la segunda igualdad se debe a la bilinealidad del producto interno. Luego, dado que para cada $j \in \{1, \ldots, p\}$ se tiene que

$$\beta_j \langle y, X_j\rangle \leq |\beta_j \langle y, X_j\rangle| \leq |\beta_j| \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle| .$$

entonces, se sigue que

$$-\beta_j \langle y, X_j \rangle \geq -|\beta_j| \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle|, \text{ para cada } j \in \{1,\ldots,p\}.$$

Así, de la desigualdad anterior y de la igualdad en (3), se deduce que

$$\begin{aligned}\|y - X\beta\|^2 &\geq \|y\|^2 + \|X\beta\|^2 - 2\left[\sum_{j=1}^{p} \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle| \, |\beta_j|\right]. \\ &= \|y\|^2 + \|X\beta\|^2 - 2 \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle| \left[\sum_{j=1}^{p} |\beta_j|\right].\end{aligned}$$

Lo que implica que

$$\begin{aligned}\frac{1}{2N}\|y - X\beta\|^2 &\geq \frac{1}{2N}\left[\|y\|^2 + \|X\beta\|^2 - 2 \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle| \left[\sum_{j=1}^{p} |\beta_j|\right]\right] \\ &= \frac{1}{2N}\left[\|y\|^2 + \|X\beta\|^2 - 2 \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle| \, \|\beta\|_1\right].\end{aligned}$$

donde, la última desigualdad se debe a que $\beta = \begin{pmatrix}\beta_1 & \ldots & \beta_p\end{pmatrix}'$. De este modo, por la desigualdad anterior y la igualdad en (1), se obtiene que

$$\begin{aligned}f_\lambda(\beta) &\geq \frac{1}{2N}\left[\|y\|^2 + \|X\beta\|^2\right] - \frac{2}{2N} \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle| \, \|\beta\|_1 + \lambda \|\beta\|_1 \\ &= \frac{1}{2N}\|y\|^2 + \frac{1}{2N}\|X\beta\|^2 + \left(\lambda - \frac{1}{N} \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle|\right)\|\beta\|_1 \\ &\geq \frac{1}{2N}\|y\|^2 + \left(\lambda - \frac{1}{N} \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle|\right)\|\beta\|_1.\end{aligned}$$

donde, la segunda desigualdad se debe a que $(2N)^{-1}\|X\beta\|^2 \geq 0$, pues, es la norma euclidiana de $X\beta$. Finalmente, de la última desigualdad se deduce que si $\lambda \geq N^{-1} \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle|$, entonces:

$$f_\lambda(\beta) \geq \frac{1}{2N}\|y\|^2. \tag{4}$$

pues, en este caso el término $\left(\lambda - N^{-1} \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle|\right)\|\beta\|_1$ es mayor o igual a cero, ya que, $\|\beta\|_1$ es también mayor o igual a cero. Ahora, note que

$$\frac{1}{2N}\|y\|^2 = \frac{1}{2N}\|y - X0\|^2 + \lambda\|0\|_1 = f_\lambda(0).$$

De este modo, la desigualdad en (4) puede escribirse como

$$f_\lambda(\beta) \geq f_\lambda(0), \text{ siempre que } \lambda \geq \frac{1}{N} \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle|.$$

Así, de la arbitrariedad de $\beta \in M_{p\times 1}(\mathbb{R})$, se sigue que

$$f_\lambda(\beta) \geq f_\lambda(0), \text{ para cada } \beta \in M_{p\times 1}(\mathbb{R}) \text{ siempre que } \lambda \geq \frac{1}{N} \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle| .$$

Por lo que, de lo anterior es posible concluir que

$$\min_{\beta\in M_{p\times 1}(\mathbb{R})} f_\lambda(\beta) = f_\lambda(0), \text{ siempre que } \lambda \geq \frac{1}{N} \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle|$$

y, que

$$\operatorname{argmin}_{\beta\in M_{p\times 1}(\mathbb{R})} f_\lambda(\beta) = 0, \text{ siempre que } \lambda \geq \frac{1}{N} \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle| .$$

Por lo cual, de las dos igualdades anteriores, se concluye que es posible tomar el valor $\lambda_{\text{máx}}$ solicitado como:

$$\lambda_{\text{máx}} = \frac{1}{N} \max_{j\in\{1,\ldots,p\}} |\langle y, X_j\rangle| .$$

Finalmente, no se sabe si existe una forma analítica cerrada para $\lambda_{\text{mín}}$, no obstante, lo que si sabe es que al elegir $\lambda = 0$, las estimaciones Lasso para el vector de coeficientes, coinciden con las hechas por mínimo cuadrados. Más aún, bajo la normalidad de los términos de error, se tiene que el estimador de mínimos cuadrados para β, posee una distribución Gaussiana multivariada, por lo que, en este caso la probabilidad de que alguna de sus entradas sea igual a cero, es exactamente cero. Basados en esto, se piensa que el valor de $\lambda_{\text{mín}}$ debe elegirse muy cercano a cero, por ejemplo, $\lambda_{\text{mín}} = 10^{-3}$, con lo que. las estimaciones Lasso y de mínimos cuadrados deberían ser casi iguales y se esperaría que ningún termino estimado fuese nulo.

■

Ejercicio 2:

Solución. Primeramente, el conjunto de datos Prostate se presenta a continuación:[1]

lpsa	*pgg*45	*gleason*	*lcp*	*svi*	*lbph*	*age*	*lweight*	lcavol
-0.431	0	6	-1.386	0	-1.386	50	2.769	-0.580
-0.163	0	6	-1.386	0	-1.386	58	3.320	-0.994
-0.163	20	7	-1.386	0	-1.386	74	2.691	-0.511
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5.143	10	7	2.464	1	-1.386	52	3.396	2.907
5.478	80	7	1.558	1	1.558	68	3.774	2.883
5.583	20	7	2.904	1	0.438	68	3.975	3.472

Cuadro 1: Datos Prostate provenientes de la librería *lasso*2 de *R*.

Este conjunto de datos, forma parte de un estudio sobre la relación que existe entre el antígeno específico de prostata (lpsa) y otras variables clínicas. El objetivo de este ejercicio, será ajustar un modelo de regresión lineal para predecir lpsa en función del resto de variables, para luego comparar las variables seleccionadas con las que selecciona un método 'stepwise' hacia atrás. Para ello, sea

[1] Los datos al completo, pueden ser consultados en *R*.

$y \in M_{97\times 1}(\mathbb{R})$ el vector columna de datos, cuya i-esima entrada corresponde a la i-esima entrada de la primer columna en el Cuadro 1, menos la media muestral de esta primer columna, en notación:

$$y_i = lpsa_i - \overline{lpsa},\ \text{con } i \in \{1,\ldots,97\}.$$

Y, sea $X \in M_{97\times 8}(\mathbb{R})$ la matriz de datos, cuyas columnas coinciden con las columnas 2 a 9 de la tabla de datos en 1 y denote por $X^{(c)} \in M_{97\times 8}(\mathbb{R})$, a la versión de X cuyas columnas han sido centradas y escaladas para tener norma $N = 97$.[2] Entonces, el objetivo de este ejercicio sera ajustar un modelo de regresión lineal de la forma

$$y = X^{(c)}\beta + \varepsilon, \tag{5}$$

donde, ε es un vector de términos de error que se supondrán no correlacionados, de media cero y varianza homocedástica $\sigma^2 > 0$, mientras que, $\beta = \begin{pmatrix}\beta_1 & \ldots & \beta_8\end{pmatrix}' \in M_{8\times 1}(\mathbb{R})$ es un vector de coeficientes lineales. No obstante, puede que no todas las covariables en $X^{(c)}$ sean realmente significativas para el modelo,[3] por lo que, una forma de seleccionar que covariables serán incluidas en el mismo es mediante regresión Lasso. Por ende, el primer paso en este ejercicio, será el ajustar un modelo de regresión Lasso al minimizar, la función

$$f_\lambda(\beta) = \frac{1}{N}\left\|y - X^{(c)}\beta\right\|^2 + \lambda\left\|\beta\right\|_1,\ \text{con } \beta = \begin{pmatrix}\beta_1 & \ldots & \beta_8\end{pmatrix}' \in M_{8\times 1}(\mathbb{R}),$$

para algún $\lambda > 0$, el cual será elegido mediante validación cruzada de tal suerte que se minimice el error cuadrático promedio muestral, al seleccionar dicho λ como parámetro de penalización para el modelo a ajustar mediante regresión Lasso. Con esto en mente, se hizo uso de la librería *glmnet* de *R*, para explorar las estimaciones para el vector de parámetros β, obtenidas mediante regresión Lasso, al variar el valor del parámetro λ en un cierto rango. Ahora, para determinar el rango de posibles valores para el parámetro λ, se hizo uso de lo probado en el inciso **a)** de este ejercicio, por ende, se calculo el valor:

$$\lambda_{\text{máx}} = \frac{1}{N}\max_{1\le j\le 8}\left|\left\langle y, X_j^{(c)}\right\rangle\right| \approx 0.95065,$$

donde, para $j \in \{1,\ldots,8\}$, se tiene que $X_j^{(c)}$ representa a la j-ésima columna de $X^{(c)}$. De este modo, se eligió variar el parámetro λ en el intervalo $[\lambda_{\text{mín}}, \lambda_{\text{máx}}]$, donde, $\lambda_{min} = 10^{-2}$ se eligió muy cercano a cero, pues, con esta elección de valores para $\lambda_{\text{mín}}$ y $\lambda_{\text{máx}}$ se esperaría, por lo hecho en el inciso **a),** que existiesen valores de λ en $[\lambda_{\text{mín}}, \lambda_{\text{máx}}]$ para los que se tuvieran estimaciones Lasso para los coeficientes de β, desde el caso en el que todas estas estimaciones son cero, hasta el caso en el que ninguna de las mismas es cero. Habiendo determinado el rango de valores a tomar en cuenta para la elección del parámetro λ, se empleo la función *glmnet* para obtener la gráfica presentada en la Figura 1. Observe que, en la parte superior de dicha gráfica, se puede observar el número de covariables cuyo coeficiente de pendiente β_j asociado, tiene una estimación distinta de cero para el valor de λ indicado en el eje x. Note que, en esta gráfica se corrobora que al considerar valores para λ en el intervalo $[\lambda_{\text{mín}}, \lambda_{\text{máx}}]$, se esta considerando modelos que arrojan desde 0 hasta 8 de las estimaciones para los componentes de β, distintas de cero.

Ahora, para seleccionar el valor adecuado para λ dentro del intervalo $[\lambda_{\text{mín}}, \lambda_{\text{máx}}]$, se hizo uso de la función *cv.glmnet* la cual emplea la técnica de n-fold cross validation, para encontrar el

[2] Dado que el objetivo del modelo es únicamente predictivo, estos cambios de escala en las variables predictoras, como se mencionó en clase, no resultan relevantes.

[3] O que incluirlas a todas cause un sobre-ajuste, o traiga consigo problemas de multicolinealidad.

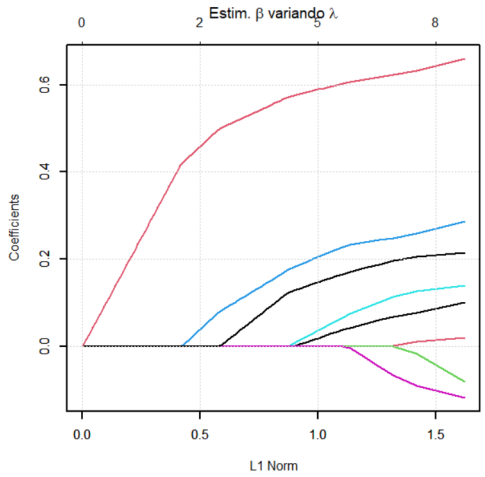

Figura 1: Estimaciones para el vector de coeficientes β obtenidas por Regresión Lasso, al variar el parámetro de penalización λ en el intervalo $[\lambda_{\text{mín}}, \lambda_{\text{máx}}]$.

valor de λ que produce el menor Error cuadrático promedio posible. De manera breve, el numero n en el nombre de esta metodología, denota el número de grupos en los que se parte de manera aleatoria el conjunto de datos originales, para llevar a cabo el ajuste del modelo Lasso con algún valor de λ especificado, usando para dicho ajuste todos los grupos de datos formados salvo uno,[4] para luego, probar que tan bueno es el ajuste del modelo en el grupo que quedo fuera, obteniendo una estimación del error cuadrático que comete el modelo, al hacer predicciones en este último grupo. Esto último, se realiza dejando fuera cada uno de los grupos en que se separo el conjunto, con lo que, se obtien un conjunto de estimaciones del error cuadrático promedio (ECP), que tiene el modelo ajustado usando el parámetro de penalización λ. Así, al promediar todos estas estimaciones del ECP, se obtiene una única estimación para el ECP que comete el modelo, cuando se elige por parámetro de penalización a λ. Así, para elegir el parámetro de penalización a usar, la idea fue llevar a cabo este algoritmo aplicándolo a los valores en el intervalo $[\lambda_{\text{mín}}, \lambda_{\text{máx}}]$ seleccionados y, elegir por parámetro de penalización para el modelo final, a aquel valor de λ asociado a la estimación más pequeña del ECP. Todo esto, como ya menciono con anterioridad, puede llevarse a acabo haciendo uso de la función *cv.glmnet*. Cabe destacar que, el paquete *glmnet* recomienda hacer uso del algoritmo 10-fold CV, no obstante, en Modelos Estadísticos *I* se vio que en el caso de Regresión Ridge, es preferible usar el algorito Leave-One-Out CV,[5] por ende, se llevaron ambos algoritmos a cabo. Ahora, entre los resultados producidos con *cv.glmnet*, se obtiene el valor de λ con el cual se consigue el menor error cuadrático promedio posible, por lo que, es importante mencionar que en este caso, se tiene que el valor de λ obtenido al usar el algoritmo 10-fold CV, es exactamente igual al valor de λ obtenido mediante el algoritmo Leave-One-Out CV y que, dicho valor de λ, el cual se denotará por λ^*, esta dado por:

$$\lambda^* = 0.029.$$

[4] A los datos usados para el ajuste, se les conoce como datos de entrenamiento. Mientras que, los datos que se excluyen al realizar el ajuste, se conocen como datos de prueba.

[5] Aquel que forma N grupos.

Lo anterior, puede corroborarse de manera visual al observar la gráfica en la Figura 2.[6] Pues, al lado izquierdo de dicho gráfico, puede ver una gráfica del $\log(\lambda)$ contra el ECP estimado mediante el uso del algoritmo 10-fold CV, cuando se usa el parámetro de penalización λ. Mientras que, de lado derecho se puede observar gráfica del $\log(\lambda)$ contra el ECP estimado mediante el uso del algoritmo Leave-One-Out CV, cuando se usa el parámetro de penalización λ. En ambas gráficas, la primer linea vertical punteada que aparece de izquierda a derecha, se encuentra remarcada a la altura $x = \log(\lambda^*)$. Así, se corrobora de manera visual, que el menor ECP estimado sea mediante 10-fold CV o mediante Leave-One-Out CV, se obtiene al tomar como parámetro de penalización a λ^*. Ahora, nombre de manera respectiva a las columnas de $X^{(c)}$, como las últimas 8 columnas de la

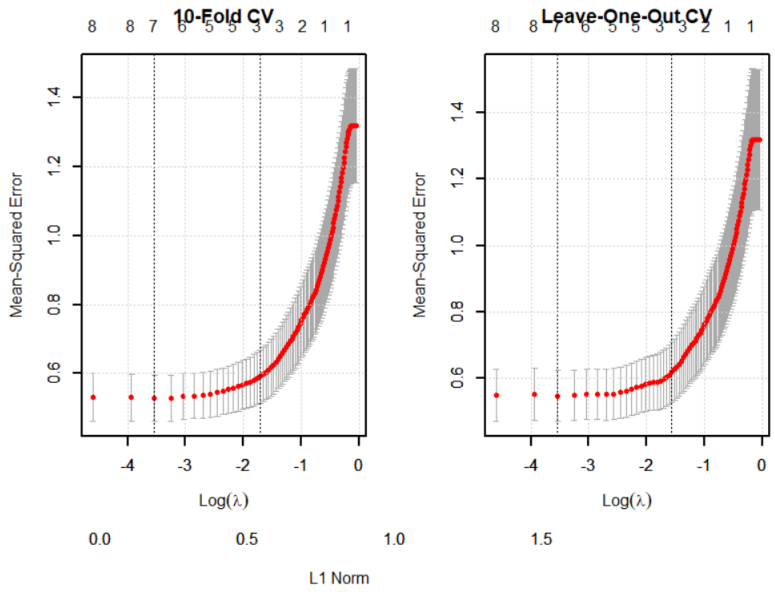

Figura 2: Gráfica de ECP estimado vs logaritmo del parámetro de penalización. Valores del parámetro de penalización en el intervalo $[\lambda_{\text{mín}}, \lambda_{\text{máx}}]$.

tabla de datos en el Cuadro 1, añadiendo el prefijo *sc* a los nombres de estas últimas 8 columnas, como una abreviación para hacer referencia al centrado y escalado hecho para obtener $X^{(c)}$. A modo de ejemplo de esto, la primer columna de $X^{(c)}$ se identificará con el nombre *scpgg*45. De este modo, para $i \in \{1, \ldots, 8\}$, note que β_i la i-ésima componente de β, es el parámetro de pendiente en el modelo en (1) asociado a la variable en la columna i, de la matriz $X^{(c)}$. Con esto en mente, se presenta en el Cuadro 2 los coeficientes estimados mediante el algoritmo Lasso, cuando se utiliza como parámetro de penalización a λ^*. Puede notar que, mediante la técnica de regresión Lasso se han incluido todas las variables salvo la variable en la columna 3, *sclcp*, de la matriz $X^{(c)}$. Por otro lado, otra técnica de selección de modelos es el algoritmo de selección automatizada 'stepwize' hacia atrás. Este algoritmo, consiste en ajustar el modelo en (5) mediante mínimos cuadrados,[7] para luego ver el efecto que tiene en el modelo el quitar una a una, cada una de las variables

[6] Las cual, también se ha producido al emplear la función *cv.glmnet*.
[7] Es decir, el modelo que contempla todas las variables.

Nombre columna	Coef	Estimación Lasso
scpgg45	β_1	0.0689
scgleason	β_2	0.0030
sclcp	β_3	0
scsvi	β_4	0.2495
sclbph	β_5	0.1136
scage	β_6	-0.0658
sclweight	β_7	0.1971
sclcavol	β_8	0.6223

Cuadro 2: Caption

en el mismo y ajustar con las covariables restantes, un modelo de regresión mediante mínimos cuadrados. Luego, si el quitar una de estas covariables, trae consigo el menor AIC en el modelo de regresión ajustado sin la misma, donde el término menor debe entenderse respecto a los AIC tanto del modelo completo, como de los modelos en los que no se considera a las otras variables, entonces dicha covariable es removida del modelo, lo que da paso a un modelo reducido al cual se le vuelve a aplicar esta misma técnica. El algoritmo concluye cuando el quitar covariables del modelo no trae consigo un menor AIC que el AIC del modelo más completo. Ahora, para aplicar este algoritmo puede usarse la función *step* de R, especificando el argumento $direction =' backward'$. Haciendo esto, se obtienen todos y cada uno de los pasos del algoritmo, necesarios para llegar del completo modelo a aquel en el cual ya no hay mejorías en el AIC, no obstante, en nuestro caso por cuestiones de espacio solo se pondrá la primer y última tablas producidas por *step*,[8] dichas tablas se presentan de manera respectiva en los Cuadros 3 y 4. Ahora, fije su atención en la columna AIC del Cuadro 3, en ella se muestra el AIC resultante al omitir del modelo completo la variable especificada en la columna Var. omitida, como puede ver el AIC del modelo completo es -60.322 y al remover la covariable *scgleason*, se obtiene el AIC más pequeño posible, por lo que, esta es la primer covariable removida del modelo por el algoritmo. Finalmente, en la columna AIC en el Cuadro 4, puede ver que el AIC más pequeño se obtiene cuando no se remueve ninguna covariable al modelo que considera únicamente a las covariables {scage,sclbph,sclweight,scsvi,sclcavol}, es decir, aquel en el que los coeficientes en β que no son cero, están dados por $\{\beta_4, \beta_5, \beta_6, \beta_7, \beta_8\}$. Finalmente, se ajusto con ayuda de la función lm de R, el modelo sugerido por el método 'stepwize' hacia atrás. Un resumen del ajuste anterior, puede verse en el Cuadro 5. Ahora, observando los Cuadros 2 y 5, se puede ver que ambos métodos de selección de variables, dejan fuera a la variable *sclcp*, más aún, las otras dos variables que deja fuera el método de selección 'stepwize' hacia atrás, tienen coeficientes estimados por Lasso muy cercanos a cero, mientras que, el resto de coeficientes se han estimado con el mismo signo por ambos métodos y las estimaciones resultan relativamente similares. Finalmente, por cuestiones de parsimonia del modelo y dado que como se comento previamente, las estimaciones por regresión Lasso para los coeficientes de scpgg45 y scgleason son cercanas a cero, elegiría el modelo seleccionado mediante el método 'stepwize' hacía atrás.

[8] Para ver este análisis en su totalidad puede ir al script adjunto.

Var. omitida	Df	Sum of Sq	*RSS*	*AIC*
- scgleason	1	0.0412	44.204	-62.231
- scpgg45	1	0.5258	44.689	-61.174
- sclcp	1	0.6740	44.837	-60.852
<ninguna>	-	-	44.163	-60.322
- scscage	1	1.5503	45.713	-58.975
- sclbph	1	1.6836	45.847	-58.693
- sclweight	1	3.5860	47.749	-54.749
- scsvi	1	4.9355	49.099	-52.045
- sclcavol	1	22.3722	66.535	-22.567

Cuadro 3: Primer paso del algoritmo 'stepwize' hacia atrás: El cual consiste en quitar variables al modelo original.

Var. omitida	Df	Sum of Sq	RSS	AIC
<ninguna>	-	-	45.526	-63.374
- scage	1	0.9593	46.485	-63.352
- sclbph	1	1.8568	47.382	-61.497
- sclweight	1	3.2250	48.751	-58.735
- scsvi	1	5.9517	51.477	-53.456
- sclcavol	1	28.7666	74.292	-17.87

Cuadro 4: Último paso del algoritmo 'stepwize' hacia atrás.

Nombre Col	Coeficiente	Estimación	t-valor	p-valor
scpgg45	β_1	0	-	-
scgleason	β_2	0	-	-
sclcp	β_3	0	-	-
scsvi	β_4	0.29693	3.468	0.000799
sclbph	β_5	0.16142	1.937	0.055803
scage	β_6	-0.11030	-1.392	0.167188
sclweight	β_7	0.20933	2.553	0.012329
sclcavol	β_8	0.66320	7.624	$2.16 \cdot 10^{-11}$

Cuadro 5: Estimaciones para el modelo ajustado haciendo uso de la función *step*.

■