Entrega final

David Prokes

Introducción

En el presente trabajo se han asignado el dataframe de *injury* de la librería *wooldridge* y la semilla aleatoria 641. Siendo solicitada a emplear a lo largo del documento como variable dependiente a predecir *ldurat*. Concretamente, en el documento *Entrega final.Rmd* puede verse con mayor detalle el procedimiento completo mediante código de los resultados expuestos en este documento. Mostrándose aquí, por tanto, aquellos resultados que se consideran relevantes.

Preparación de los datos

En cuanto a la preparación inicial de los datos del dataframe *injury*, primeramente, se han cargado todas las librerías necesarias, así como se han transformado a variables de tipo factor todas aquellas variables cualitativas con más de 2 niveles (no binarias). Por otra parte, se dividen los datos en dos conjuntos: 85% de los datos para entrenamiento y el 15% de prueba.

(1)

El conjunto de datos de *injury* pertenece a la colección de datos empleada en el libro de *Wooldridge, Introducción a la Econometría: Un Enfoque Moderno*, accesible mediante la librería *wooldridge*. Concretamente, *injury* procede del artículo: *B.D. Meyer, W.K. Viscusi, and D.L. Durbin (1995), "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," American Economic Review 85, 322-340*; donde se recopilan datos acerca de la compensación y duración de las lesiones laborales en un experimento natural en Michigan y Kentucky.

En el dataframe se encuentran un total de 7150 observaciones y 30 variables. Así, la variable dependiente solicitada (***ldurat***) se trata de la transformación logarítmica de la variable ***durat***, que captura la duración (probablemente en días) de la prestación. Mientras que, otras variables relevantes serían: el coste médico total durante la prestación (***totmed***), el beneficio percibido de la prestación (***benefit***), la retribución semanal percibida en la semana anterior a la baja (***prewage***), o la variable dummy sobre la aplicación del incremento en el beneficio de la prestación (***afchnge***).

A continuación, se muestran una serie de estadísticos principales para aquellas variables cuantitativas. De esta manera, se puede observar que en el conjunto de datos se presenta una duración mínima de 0.25 de la prestación (probablemente sean 2 horas de jornada laboral de 8 horas) y un máximo de 182 días. La edad mínima registrada es de 12 años y una máxima de 98.

En cuanto a los salarios previos semanales, existe una alta variabilidad desde 81.78 dólares a 1583.10 dólares, viéndose una distribución sesgada con cola hacia la derecha similar al caso de ***totmed*** y ***benefit***. Mientras que, destaca el máximo coste

médico total de alrededor de 2 millones de dólares respecto al beneficio semanal máximo de la prestación de 742.22 dólares.

Tabla 1: Principales estadísticos de las variables explicativas cuantitativas

	durat	*age*	*prewage*	*totmed*	*benefit*
Min.	0.25	12.00	81.78	0.00	14.87
Median	4.00	32.00	263.85	328.60	149.93
Mean	9.92	34.71	329.73	1714.40	162.92
Max.	182.00	98.00	1583.10	2323376.50	742.22

Nota: En determinados estados y empleos concretos se permite el empleo juvenil.

A continuación, se muestra un mapa de calor de las correlaciones entre las variables cuantitativas en el conjunto de datos de entrenamiento. En él destaca una correlación alta entre ***benefit*** y ***lprewage*** de 0.73, así como entre las correspondientes transformaciones logarítmicas como era de esperar.

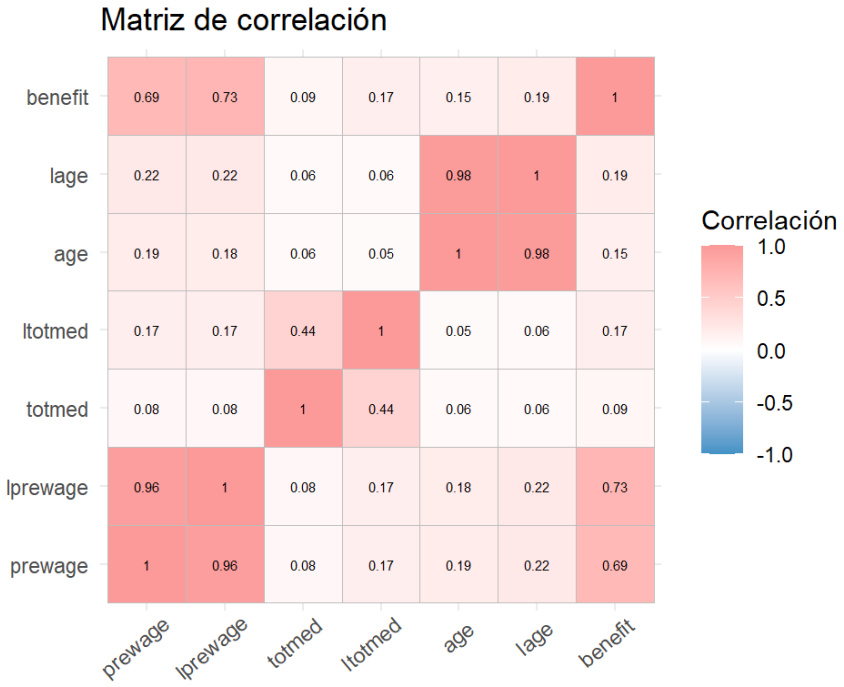

Ilustración 1: Mapa de correlación de variables explicativas cuantitativas (heatmap)

Por otra parte, aplicamos el criterio del factor inflacionario de la varianza. Para ello, primero excluimos una variable dummy para las categorías mutuamente excluyentes para evitar la multicolinealidad perfecta: ***ky*** con ***mi***, ***injtype*** incluye todas las lesiones *[head, neck, upextr, trunk, lowback, lowextr, occdis]*, ***indust*** incluye *[manuf, construc]*; y con combinaciones lineales o transformaciones logarítmicas: ***afhigh***, ***highlpre***, ***prewage***, ***lage***, ***totmed*** (se han optado por las transformaciones logarítmicas o no en función de su VIF; aunque en el caso de ***totmed*** presentaba un VIF inferior a 5 pero por preferencia se ha mantenido únicamente ***ltotmed*** por tener el menor VIF).

De modo que, con el criterio aplicado, se eliminarían por presentar una elevada multicolinealidad las siguientes variables (con VIF superior a 5): ***injdes*** (menos manejable e interpretable que ***injtype***) y ***lprewage*** (muy correlacionada probablemente

con ***highearn***). Por lo tanto, para escoger las variables explicativas y poder predecir ***ldurat*** podría ser razonable mantener las variables, cuyo coeficiente es individualmente significativo, derivadas del modelo que hemos empleado para el criterio del factor inflacionario de la varianza:

- ***ltotmed***: el coste médico total determina la gravedad de la baja y por tanto su duración.
- ***benefit***: un beneficio semanal más elevado puede incentivar una mayor duración de la prestación.
- ***age***: la edad podría determinar la resiliencia o gravedad y por tanto la duración de la baja.
- ***male***: el género puede ser una variable interesante a medir.
- ***highearn***: se podría pensar que individuos con altos ingresos tienen un comportamiento distintivo en cuanto al beneficio de la prestación.
- ***hosp***: la hospitalización determina en gran medida la duración de la baja.
- ***indust***: el tipo de industria puede determinar la frecuencia y la duración de la baja, dado el entorno en el que se expone con mayor o menor riesgo.
- ***injtype***: el tipo de baja determina en gran medida la duración de la misma.

(2)

En este apartado, se muestran las interpretaciones (ceteris paribus) del modelo estimado, con la muestra de datos completa de *injury*. Cabe destacar que todos los coeficientes son individualmente significativos al 1% dado que todos los p-valor son menores al 0.01. Por otro lado, alrededor de un 32% de la varianza de ***ldurat*** es explicada por el modelo (R^2 ajustado), así como los coeficientes son conjuntamente significativos con el contraste F de significancia global. La siguiente fórmula es la empleada para interpretar las variables de tipo dummy a partir de sus coeficientes:

$$\Delta\% = \left(e^{\beta} - 1\right) \times 100$$

- ***ltotmed***: Un incremento del 1% en ***totmed*** conlleva, en promedio, a un incremento del 0.35% en la duración de la prestación (elasticidad).
- ***benefit***: Un incremento en un dólar de ***benefit*** conlleva, en promedio, a un incremento del 0.28% en la duración de la prestación.
- ***age***: Un año adicional de edad conlleva, en promedio, a un incremento del 0.77% en la duración de la prestación.
- ***male***: La fórmula para interpretar variables dummy en un modelo log-lineal es la indicada anteriormente. Por lo que, ser hombre respecto a ser mujer implica, en promedio, una reducción del 13.22% en la duración de la prestación.
- ***highearn***: Aplicando la fórmula anterior, los trabajadores con ingresos altos, en promedio, presentan una duración de la prestación menor en un 10.86% respecto a los ingresos bajos.
- ***hosp***: Aplicando la fórmula anterior, las personas hospitalizadas, en promedio, presentan una duración de la prestación mayor en un 27.91% respecto a los no hospitalizados.

- ***indust***: Cada coeficiente representa el efecto de vincularse a cada tipo de industria sobre la duración de la prestación respecto a la categoría base (industria manufacturera o sector primario). Viéndose que, la industria 2 (sector de transformación y, en especial, construcción) es la que incide en un mayor efecto en promedio sobre la duración respecto a la categoría base.
- ***injtype***: Cada coeficiente representa el efecto de cada tipo de lesión sobre la duración de la prestación respecto a la categoría base (lesión de cabeza). Destacando la lesión 7 (enfermedad ocupacional) por tener un mayor efecto en cuanto a la duración, en promedio, de la prestación.

(3)

En este apartado, se estima el modelo con los mismos datos de *injury* completos como en el apartado (2), no obstante, teniendo en cuenta todas las variables explicativas (menos ***durat*** por coherencia; además el modelo automáticamente excluye aquellas variables que presentan multicolinealidad perfecta). De modo que, con los resultados que se extraen, el modelo 2 obtuvo un mayor R^2 ajustado de 0.3211 frente al 0.3193 del modelo 3. En este caso, el mayor número de variables perjudica ligeramente al modelo 3 por la presencia probablemente de elevada multicolinealidad entre algunas variables (antes identificadas mediante VIF). En cuanto al RSE (Error estándar residual) en el modelo 3 es de 1.074 inferior al 1.078 del modelo 2, posiblemente por un sobreajuste a la muestra dado el mayor número de variables explicativas.

Calculando el contraste F de significancia conjunta sobre las 13 variables no significativas individualmente al 1%, obtenemos un valor del estadístico de 8.64. Mientras que, el valor crítico con 6796 grados de libertad y del 1% de significancia es de 2.13. Por lo que, se rechaza la hipótesis nula, implicando que al menos alguna de las 13 variables es significativa al 1%. De modo que, es posible que el correspondiente coeficiente se encuentre distorsionado en el modelo 3 completo, por la presencia de elevada multicolinealidad, pasando a ser no significativo individualmente.

$$F = \frac{(0.3218 - 0.3106)/13}{(1 - 0.3218)/(6796)} = 8.63655$$

(4)

Ajustando un modelo con todas las variables explicativas mediante el conjunto de entrenamiento, y posteriormente evaluando la predicción con el conjunto de prueba, da como resultado un ECM (Error Cuadrático Medio) de 11.92 y una raíz del ECM de 3.45. Siendo estos valores muy elevados en comparación con el promedio de la variable ***ldurat*** de 1.33 en el conjunto de datos total de entrenamiento. Dichos resultados dan un indicio de efectos de elevada multicolinealidad y sobreajuste a los datos de entrenamiento (R^2 ajustado de 0.333) por el elevado número de explicativas.

(5)

La Mejor Selección de Conjuntos consiste en un procedimiento mediante el cual se obtienen los mejores modelos o subconjuntos para cada número de variables posible,

a través de la iteración de todas las combinaciones de variables. Por otro lado, la validación cruzada de 10-veces consiste en crear 10 particiones de la muestra, donde todas menos una se emplean para estimar el modelo, mientras que con la partición restante se calcula el ECM. Este proceso se aplica de manera iterativa para cada una de las 10 particiones, y sobre cada uno de los modelos resultantes de la Mejor Selección de Conjuntos.

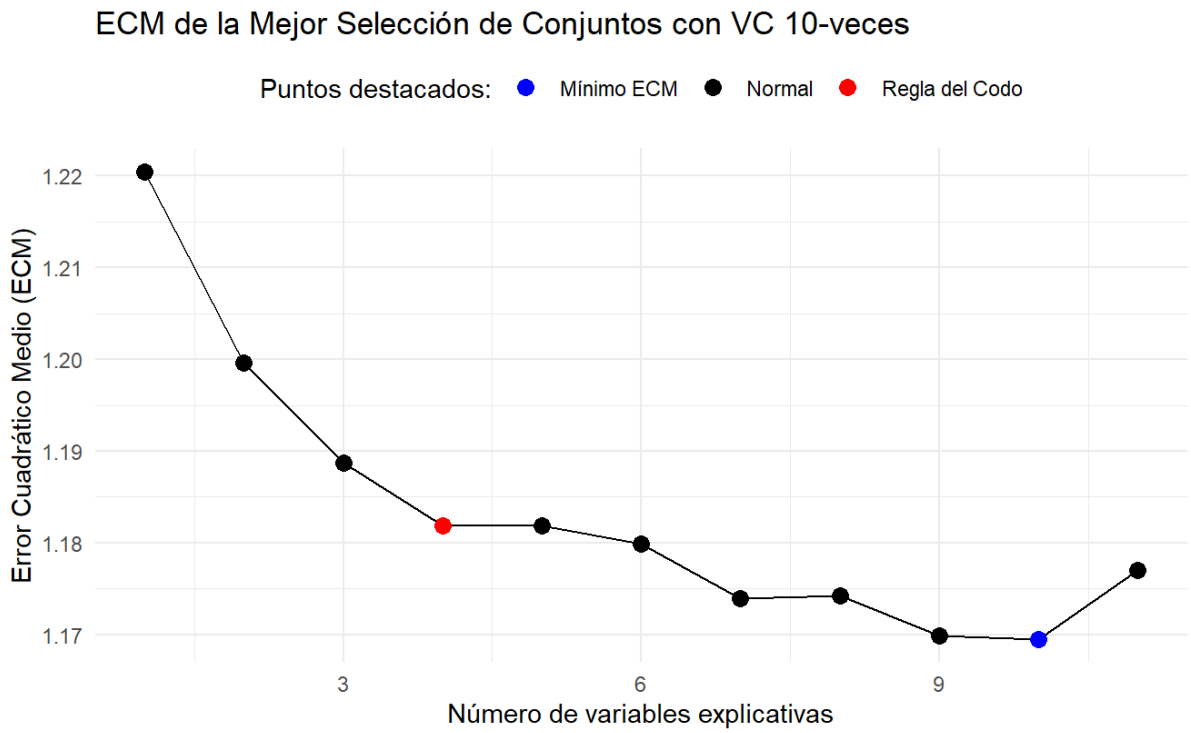

Ilustración 2: Regla del codo aplicada a los ECM de la Mejor Selección de Conjuntos

Así, en la gráfica anterior vemos que, pese a que el ECM mínimo se obtiene con el modelo que tiene 10 de variables; mediante la regla del codo basado en el criterio de parsimonia, se obtiene como modelo óptimo aquel que tiene 4 variables explicativas. Concretamente, con el modelo óptimo resultante se estima un error de prueba de 1.18. Por otro lado, estimando el modelo óptimo en el conjunto de entrenamiento hemos obtenido un R^2 ajustado de 0.30.

(6)

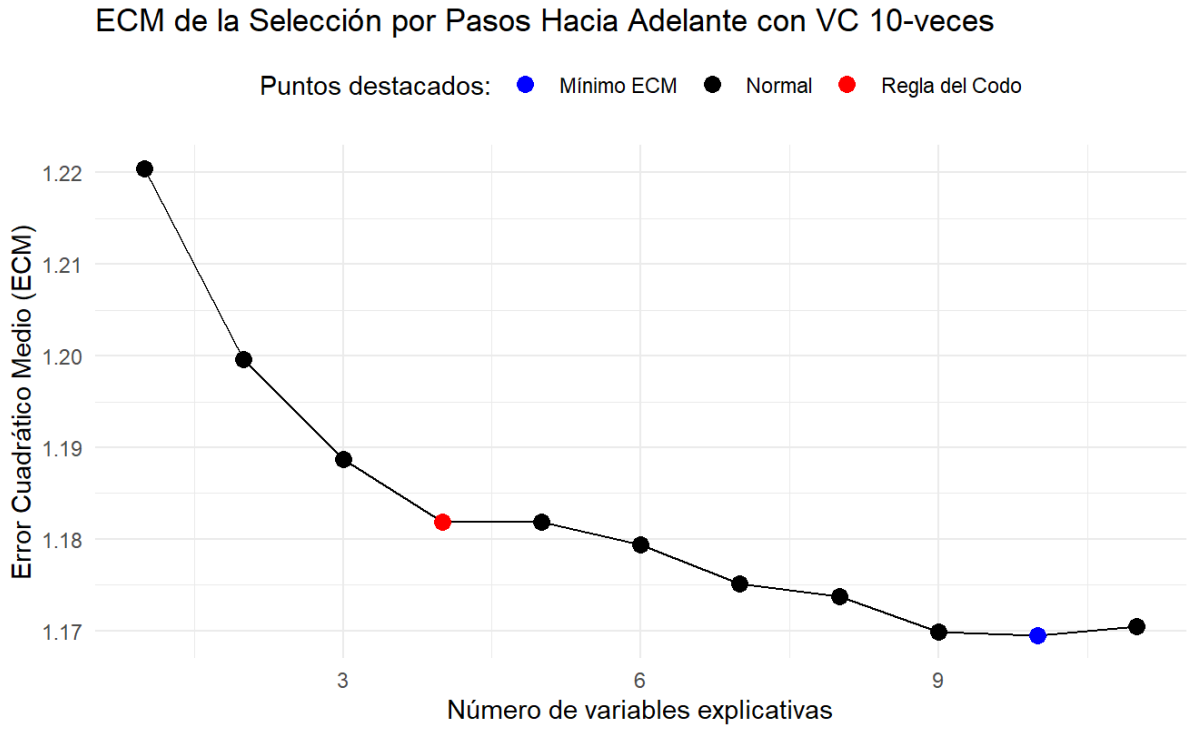

Ilustración 3: Regla del codo aplicada a los ECM de la Selección por Pasos Hacia Adelante

En la gráfica anterior, podemos ver que el modelo resultante del método de la Selección por Pasos Hacia Adelante con el ECM mínimo, de nuevo es el modelo con 10 variables. Asimismo, aplicando la regla del codo, el modelo con 4 variables sería el óptimo manteniendo un buen compromiso entre ECM mínimo y parsimonia. Por lo que su R^2 ajustado es de 0.30 en el conjunto de entrenamiento y el error de prueba estimado de 1.18, siendo el mismo modelo con los mismos coeficientes al del apartado anterior.

(7)

Aplicando la validación cruzada 5-veces no presenta diferencias en cuanto a los modelos óptimos extraídos respecto a la validación cruzada 10-veces, siendo incluso computacionalmente más barato. De modo que en ambos métodos se obtiene, de nuevo, el mismo modelo con idénticos coeficientes, aplicando el criterio del codo. No obstante, vemos que el ECM mínimo para ambos métodos con VC 5-veces, se presenta en el modelo con 9 variables, mientras que, en la VC 10-veces era el correspondiente al modelo con 10 variables.

(8)

Tabla 2: Error de prueba por modelo seleccionado mediante la Mejor Selección de Conjuntos y la Selección por Pasos Hacia Adelante

	Mejor Selección	Hacia Adelante
VC 5-veces	1.180105	1.180105
VC 10-veces	1.180105	1.180105

Nota: Los modelos óptimos derivados de cada método coinciden, empleando las mismas variables explicativas.

Así, en la tabla anterior podemos ver que ambos métodos, independientemente de la validación cruzada empleada de 5-veces o 10-veces, aplicando la regla del codo se seleccionan el mismo modelo, cuyo resultado se expone a continuación. Este hecho ocurre dado que, para una cantidad moderada-baja de variables explicativas posibles (11 en este caso), ambos métodos convergen a una misma selección óptima. Cabe destacar que ambos métodos se ven afectados de manera negativa por la introducción de variables con elevada multicolinealidad. Por ello, respecto a las 28 variables que incluye la base de datos original (sin contar ***durat***), tras aplicar el criterio del factor inflacionario de la varianza, las variables disponibles se reducen a 11.

Tabla 3: Modelo óptimo derivado de los métodos de la Mejor Selección de Conjuntos y la Selección por Pasos Hacia Adelante

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*	
(Intercept)	-1.4099039	0.0670136	-21.039	< 2e-16	***
hosp	0.2455078	0.0377702	6.500	8.60e-11	***
age	0.0083852	0.0010563	7.939	2.37e-15	***
benefit	0.0021726	0.0002201	9.870	< 2e-16	***
ltotmed	0.3426369	0.0096500	35.506	< 2e-16	***

Residual standard error:	1.087	*Degrees of freedom*	6817
Multiple R-squared:	0.304	*Adjusted R-squared:*	0.3036
F-statistic:	744.3	*p-value:*	< 2.2e-16

*Nota: Los niveles de significación son indicados mediante '***' para el 0.1%, '**' para el 1%, '*' para el 5% y '.' para el 10%.*

(9)

Teniendo en cuenta las tablas del apartado anterior, ambos métodos coinciden con el modelo empleado, independientemente de la validación cruzada de 5-veces o 10-veces. De modo que, en los resultados expuestos anteriormente, se observa que todos los coeficientes son significativos al 1%, dado que los correspondientes p-valor son claramente inferiores al 0.01.

(10)

La regresión Ridge consiste en una regresión lineal con la característica de añadir una penalización al tamaño de los coeficientes, ideal para afrontar el problema de multicolinealidad elevada. Por tanto, a la hora de escoger los datos de entrenamiento, no será necesario eliminar aquellas variables que presentaban un VIF superior a 5 en el apartado (1), pero sí se deberán excluir aquellas que pueden ocasionar

multicolinealidad perfecta. Además, se han omitido las variables de ***durat*** y ***injdes***, esta última por ser un código categórico de 4 dígitos correspondiente a la descripción de la lesión, difícilmente manejable en modelos de regresión.

Por otro lado, tanto el modelo de regresión Ridge como LASSO pueden verse afectados negativamente cuando se introducen variables con valores de rangos diversos. Por lo que, será conveniente escalar los valores a media 0 y desviación típica 1. Así, habiendo escalado aquellas variables cuantitativas del conjunto de datos, se estima una regresión Ridge en el conjunto de entrenamiento, resultando en un error de prueba sobre el conjunto de prueba de 1.21. En la gráfica concretamente se puede encontrar la línea discontinua sobre el valor de lambda óptimo de 0.30, cuyo logaritmo es de -1.20 representado en la gráfica.

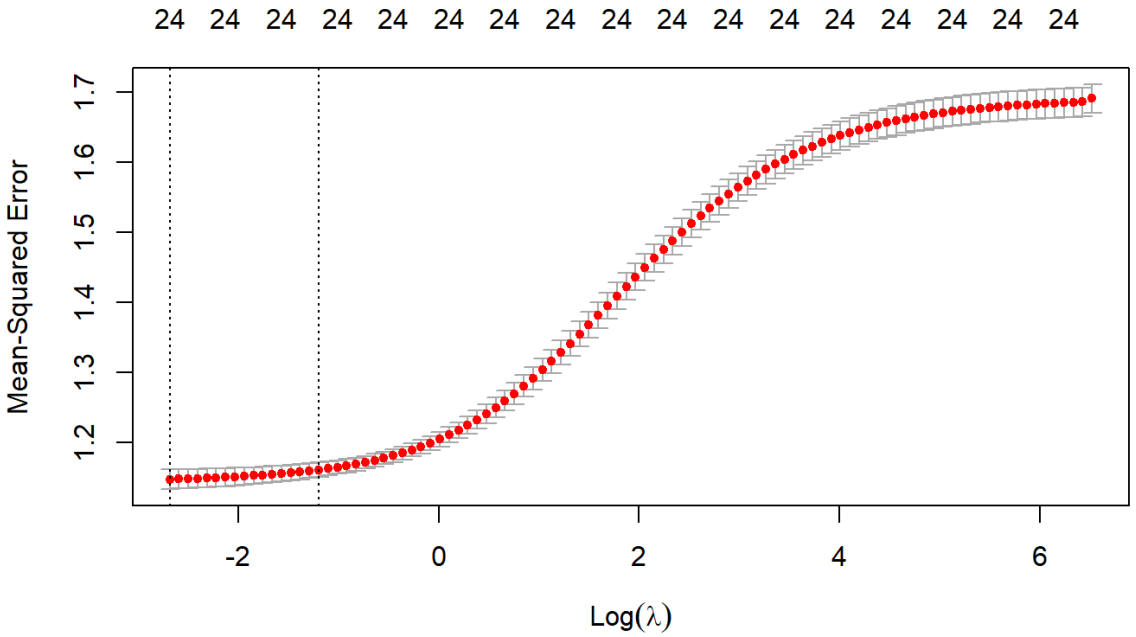

Ilustración 4: ECM para cada lambda en la regresión Ridge

(11)

La regresión LASSO es similar a la regresión Ridge, ambas caracterizadas por añadir una penalización al tamaño de los coeficientes, pero en el caso de LASSO se aplica una selección de variables, es decir, algunos coeficientes pueden obtener un valor nulo, ideal para situaciones con muchas variables explicativas de las que se esperan que haya alguna irrelevante.

Por lo que, en la gráfica presentada a continuación, vemos que mediante la aplicación de la validación cruzada en la regresión LASSO, se obtiene un lambda óptimo de 0.02 inferior al obtenido con la regresión Ridge. Viéndose representado con una línea discontinua su logaritmo con un valor de -3.92. Obteniendo un error de prueba de 1.187.

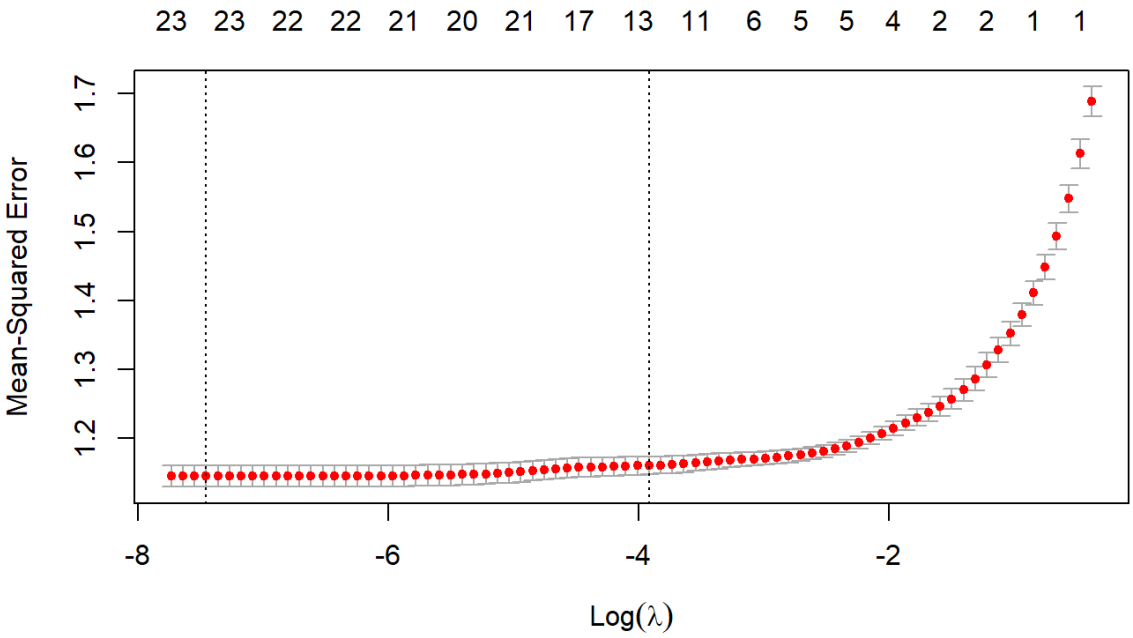

Ilustración 5: ECM para cada lambda en la regresión LASSO

En cuanto a los coeficientes, en la siguiente tabla se pueden ver las que se han omitido mediante un coeficiente nulo. Así, el número de coeficientes distintos de 0 es de 13, contando el intercepto. En otras palabras, se considera el efecto de 12 variables explicativas en el modelo LASSO estimado.

Tabla 4: Coeficientes estimados por regresión LASSO con VC 10-veces

Variable	*Estimate*	*Variable*	*Estimate*
(Intercept)	1.2677	*injtype7*	0.1543
afchnge	0.0000	*injtype8*	0.0000
highearn	0.0000	*age*	0.0202
male	-0.0705	*prewage*	0.0000
married	0.0000	*totmed*	0.1987
hosp	0.1801	*benefit*	0.1109
indust2	0.1462	*ky*	0.0000
indust3	0.0758	*afhigh*	0.0000
injtype2	0.0673	*lprewage*	0.0000
injtype3	0.0000	*lage*	0.0649
injtype4	0.0000	*ltotmed*	0.5024
injtype5	0.0000	*highlpre*	0.0000
injtype6	0.0032		

Nota: El método LASSO permite implementar una selección de variables estableciendo como nulos los coeficientes.

(12)

En este apartado, se estiman de nuevo los modelos de Ridge y LASSO con la diferencia de emplear validación cruzada 5-veces para hallar el lambda óptimo, siempre aplicando la regla del codo. Las diferencias producidas en cuanto a la regresión Ridge

se centran en un aumento ligero del lambda escogido de 0.36 y del error de prueba calculado de 1.212, respecto al lambda de 0.30 seleccionado con validación cruzada 10-veces y el correspondiente error de prueba de 1.206.

Mientras que, respecto a la regresión LASSO, se obtiene un lambda óptimo de 0.03 ligeramente superior al 0.02 obtenido con validación cruzada 10-veces. Así como, el error de prueba para validación cruzada 5-veces es de 1.192 muy similar al 1.187 de 10-veces. En cuanto al número de coeficientes distintos de 0 es de 11, contando el intercepto. En otras palabras, se considera el efecto de 10 variables explicativas en el modelo LASSO estimado.

Tabla 5: Coeficientes estimados por regresión LASSO con VC 5-veces

Variable	*Estimate*	*Variable*	*Estimate*
(Intercept)	1.2776	*injtype7*	0.0383
afchnge	0.0000	*injtype8*	0.0000
highearn	0.0000	*age*	0.0063
male	-0.0251	*prewage*	0.0000
married	0.0000	*totmed*	0.1927
hosp	0.1643	*benefit*	0.0974
indust2	0.0771	*ky*	0.0000
indust3	0.0261	*afhigh*	0.0000
injtype2	0.0000	*lprewage*	0.0000
injtype3	0.0000	*lage*	0.0729
injtype4	0.0000	*ltotmed*	0.4997
injtype5	0.0000	*highlpre*	0.0000
injtype6	0.0000		

Nota: El método LASSO permite implementar una selección de variables estableciendo como nulos los coeficientes.

(13)

El modelo de Componentes Principales consiste en un método que integra el Análisis de Componentes Principales en la regresión lineal clásica. Así, el ACP resulta ideal para reducir la dimensionalidad de variables cuantitativas altamente correlacionadas, eliminando el problema de incluirlas por separado y tener que tratar con los efectos sobre los coeficientes de la multicolinealidad elevada. De modo que, el modelo de Componentes Principales también se puede integrar con la validación cruzada y el método correspondiente del codo, a la hora de seleccionar el número óptimo *M* de componentes principales a incluir en la regresión lineal.

Por lo que, en este apartado se aplica tanto la validación cruzada 5-veces como 10-veces para elegir el *M* óptimo. De esta manera, ambos casos resultan en la elección del mismo *M* óptimo con la regla del codo, siendo 18 componentes principales incluidos en la regresión lineal derivada del modelo. En otras palabras, se incluyen aquellos que explican una proporción significativa de la varianza entre las variables explicativas. Respecto al error de prueba que se obtiene es de 1.18, tratándose ambos del mismo modelo.

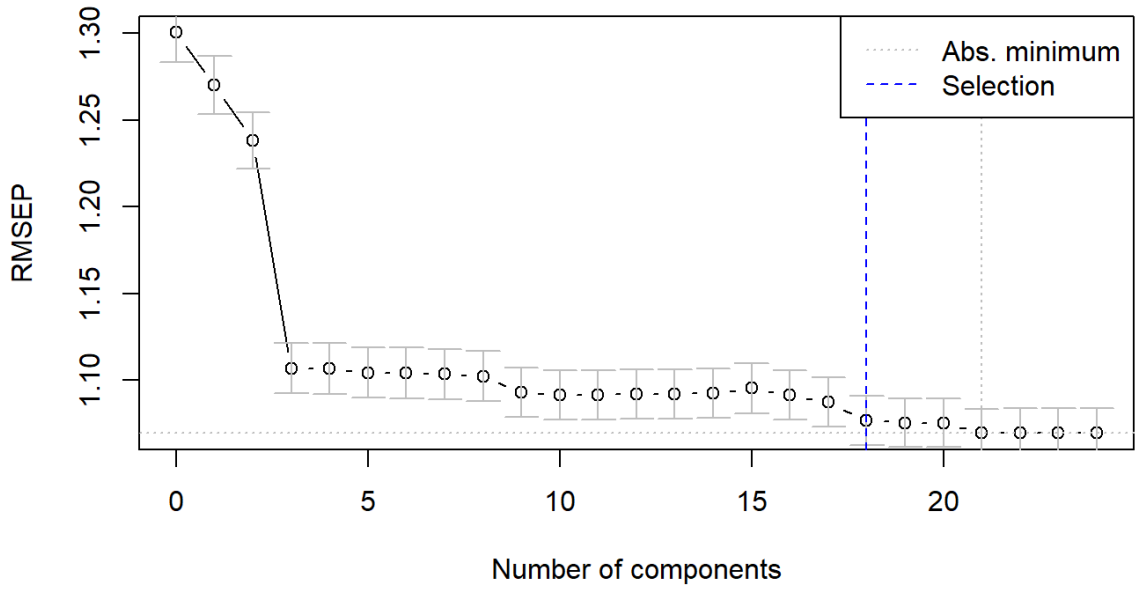

Ilustración 6: ECM por número de componentes principales incluidos con VC 10-veces (Componentes Principales)

En la gráfica anterior se puede ver el *M* que minimiza el ECM representado a través de una línea discontinua gris, y el *M* seleccionado con el método del codo con una línea discontinua azul. Aunque, se podría pensar que incluir 18 componentes principales a partir de 25 variables puede ser poco parsimonioso, viéndose gráficamente como realmente el ECM se estabiliza con 4 componentes principales, y con una segunda estabilización con 9 componentes principales. Esta última idea veremos que se aplicará en el siguiente apartado con PLS.

(14)

El modelo de Mínimos Cuadrados Parciales (PLS) consiste en una técnica similar al modelo de Componentes Principales, con la diferencia de escoger el número óptimo de componentes principales únicamente en base a la varianza explicada de los predictores. En otras palabras, M no se selecciona en base al error de prueba estimado en la regresión lineal, como sí lo hacía el modelo de Componentes Principales visto anteriormente.

Por lo que, estimando el modelo PLS en el conjunto de entrenamiento, se obtiene un *M* óptimo con la regla del codo de 3 componentes principales a incluir en la regresión. De nuevo, aquí, el *M* óptimo resultante es independiente de que validación cruzada empleemos de 5-veces o 10-veces. El error de prueba del modelo seleccionado es de 1.20. Como podría esperarse el error de prueba es superior al obtenido en el modelo de Componentes Principales, ya que en este último se captura una mayor parte de la varianza de las características.

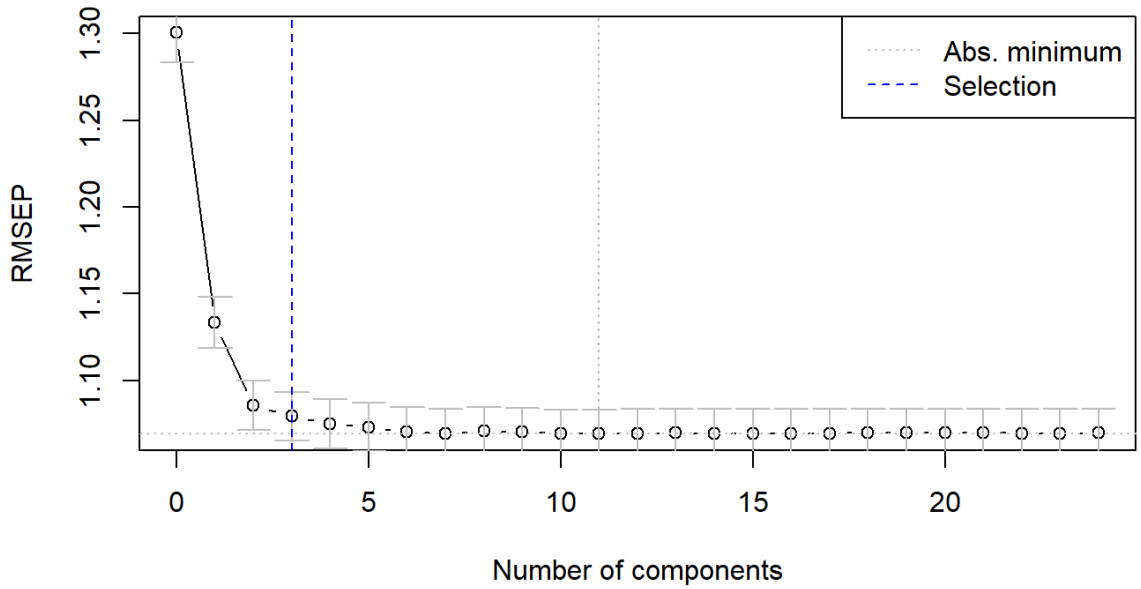

Ilustración 7: ECM por número de componentes principales incluidos con VC 10-veces (RLS)

(15)

Tabla 6: Error de prueba y parámetro optimizado por modelo

	Ridge	LASSO	CP	PLS
CV5 Error	1.212114	1.191957	1.1879	1.2007
CV10 Error	1.205938	1.186844	1.1879	1.2007
Lambda CV5*	0.363542	0.031619	*NA*	*NA*
Lambda CV10*	0.301819	0.019858	*NA*	*NA*
M CV5*	*NA*	*NA*	18	3
M CV10*	*NA*	*NA*	18	3

Nota: En las regresiones Ridge y LASSO se estima un lambda óptimo, mientras que en el modelo de Componentes Principales (CP) y PLS se estima un número M óptimo de componentes principales a incluir en la regresión lineal.

En la tabla anterior podemos ver los errores de prueba resultantes de estimar cada modelo sobre el conjunto de entrenamiento (85% de los datos) y evaluar la capacidad predictiva sobre el conjunto de prueba (15% de los datos). Todo ello, empleando para cada enfoque la validación cruzada 5-veces y 10-veces para hallar el parámetro óptimo respectivo.

Por tanto, el modelo CP es que obtiene un menor error de prueba, siendo un enfoque óptimo si únicamente se desea predecir la variable ***ldurat***. Como hemos comentado previamente, este enfoque tiene la particularidad de reducir la dimensionalidad de los predictores, hallando componentes principales ortogonales que capturan gran parte de la varianza total en el espacio de las características. Siendo 18 componentes principales el número *M* óptimo para este modelo.

En cuanto a la comparativa con el apartado (9), de este se extrajo que todos los modelos convergían hacia el mismo número de variables óptimo de 4, con los mismos coeficientes, y por ende, el mismo error de prueba estimado de 1.180105. Siendo esta opción en este caso más atractiva respecto a los modelos estimados en los apartados (10)-(14). Esto se debe al relativo bajo número de variables explicativas disponibles en el dataset que no permite exprimir el potencial de la reducción de la dimensionalidad de los modelos (13) y (14).

Sin embargo, en conjuntos de datos con un gran número de características, podría resultar más atractivo aplicar un modelo de reducción de dimensionalidad como el CP, ya que en estos casos es más probable que el impacto de la multicolinealidad sea mayor y no permita introducir tantas variables como se desearían en los modelos de selección (Mejor Selección de Conjuntos y Selección por Pasos Hacia Adelante). Dando mejores resultados en las estimaciones con los errores de prueba. Además de que, sería costoso computacionalmente trabajar con los modelos de selección, ya que iteran sobre un gran número de variables para estimar una cantidad significativa de modelos.