Trabajo final

A. Conjunto de datos

Corresponde trabajar con la base de datos `wage1` de la librería `wooldridge` estableciendo la semilla `206` y con `lwage` como variable dependiente. El data set contiene datos de la encuesta de población de EE.UU. de 1976 respecto de las siguientes variables:

- `wage` y `lwage`: salario medio por hora y su transformación logarítmica. Se descarta `wage` porque `lwage` es la variable dependiente; no tendría sentido usar su transformación como explicativa.
- `educ`: años de educación.
- `exper` (y `expersq`): años de experiencia laboral (y la misma al cuadrado para determinar el efecto de valores elevados, trabajadores mayores)
- `tenure` (y `tenursq`): años con el empleo actual (y la misma al cuadrado para determinar el efecto de valores elevados, trabajadores con más antigüedad)
- `nonwhite`: variable categórica respecto de la raza, con valor 1 si no es blanco.
- `female`: variable categórica respecto del sexo, con valor 1 si es mujer.
- `married`: variable categórica respecto del estado civil, con valor 1 si es mujer.
- `numdep`: número de personas dependientes (hijos) a cargo.
- `smsa`: variable categórica respecto de la zona de residencia, con valor 1 si es urbana.
- `northcen`, `south` y `west`: variables categóricas respecto de la región de residencia. De estas se desprende que la categoría de referencia representa el Este de EE.UU.
- `construc`, `trcommpu`, `trade`, `services`, `profserv`, `profocc`, `clerocc` y `servocc` son variables categóricas respecto del sector de ocupación. Se desconoce la categoría de referencia, pero podrían ser las ocupaciones en el sector primario.

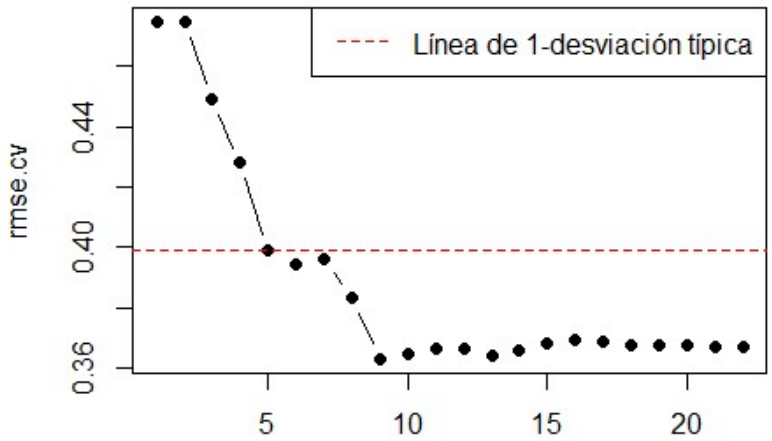
Se comprueba que no existe multicolinealidad perfecta entre las variables. Se asume, además, que, si bien las variables explicativas pueden influir sobre el nivel salarial, el salario no afecta a los años de educación, de experiencia laboral, a la raza, al sexo, la zona, la región de residencia ni al sector de ocupación. En algunas circunstancias podría considerarse que el salario sí que afecta a la antigüedad (trabajadores con salarios más elevados serían más reticentes a cambiar de trabajo; para el empleador, trabajadores con salarios más elevados supondrían un despido más caro). Sin embargo, los EE.UU. eran ya en 1976 una economía liberal con muy baja protección de los puestos de trabajo; por tanto, se asume que `tenure` es exógena. El número de hijos también podría ser un factor endógeno: un trabajador con salario más elevado a priori podría

permitirse tener una familia más extensa. Pero, en cualquier caso, se asume que esta decisión es independiente del nivel salarial, por lo menos en la mayoría de los casos.

Se cargan los datos sin la variable `wage` y se efectúa una regresión por MCO con `lwage` como variable dependiente y las demás como explicativas. El error (MSE) de prueba resulta en 0.1609.

B. Mejor Selección de Conjuntos (VC 10-veces)

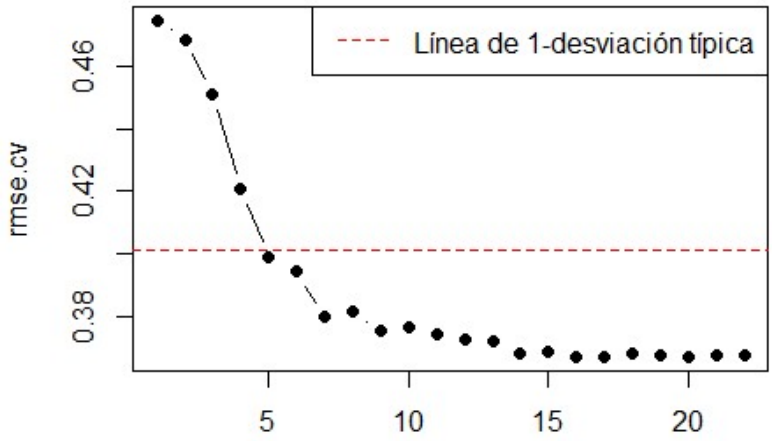
Por este método se calculan todos los modelos lineales con todas las combinaciones posibles de variables entre 0 (solo el intercepto) y 22 y se selecciona el que reporta el menor MSE mediante VC 10-veces. Posteriormente, mediante la regla del codo, se puede asumir un MSE mayor (a una distancia máxima de 1 desviación típica) a cambio de la simplicidad del uso de menos variables explicativas.



Aplicando la regla del codo se seleccionan **6 variables**, obteniéndose un MSE de prueba de **0.1696**.

C. Selección por Pasos hacia Adelante (VC 10-veces)

Por este proceso, a partir de un modelo solo con el intercepto, se añaden variables una a una escogiendo cada vez aquella que más disminuye el MSE. De todas las regresiones se escoge aquella con el número de regresores tal que minimiza el MSE de VC.

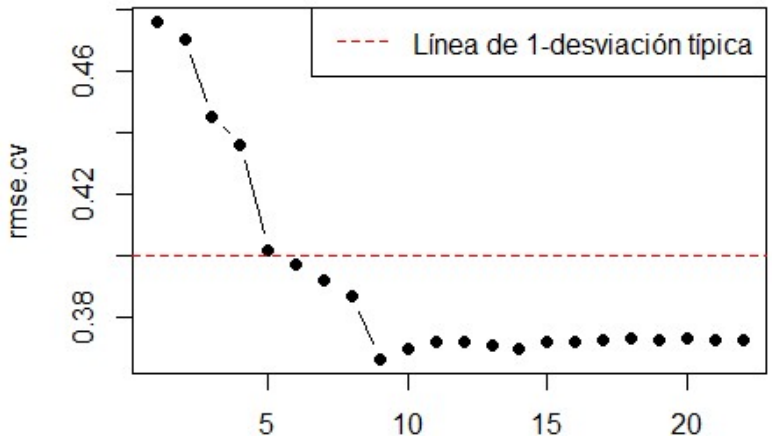


Aplicando la regla del codo se seleccionan **5 variables**, obteniéndose un MSE de prueba de **0.1651**.

D. Mejor Selección de Conjuntos y Selección por Pasos hacia Adelante (VC 5-veces)

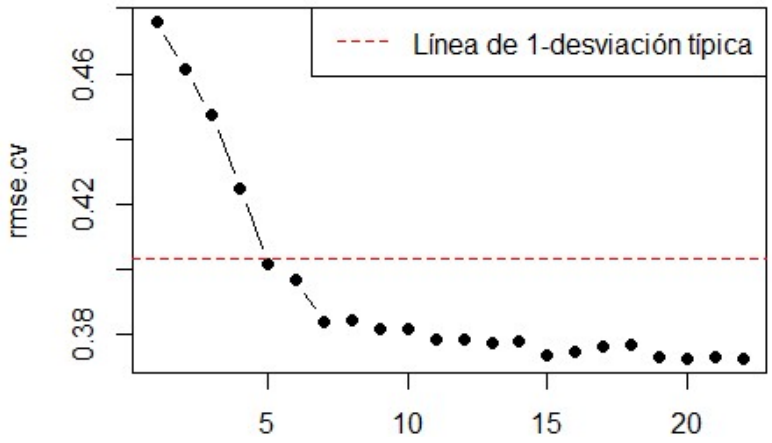
Aquí se repiten los procesos de los apartados B y C con VC 5 veces.

D.1. Mejor Selección de Conjuntos (VC 5-veces)



Aplicando la regla del codo se seleccionan **6 variables**, obteniéndose un MSE de prueba de **0.1696**.

D.2. Selección por Pasos hacia Adelante (VC 5-veces)



Aplicando la regla del codo se seleccionan **5 variables**, obteniéndose un MSE de prueba de **0.1651**.

E. Resumen de resultados

	MCO	MSS.VC10	MSS.VC5	MSHA.VC10	MSHA.VC5
RMSE	0.1609	0.1696	0.1696	0.1651	0.1651

F. Selección del mejor modelo

Del apartado E se sabe que el mejor modelo es MCO con todas las variables explicativas.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.8940168	0.1318315	6.7815103	0.0000000
educ	0.0441226	0.0088824	4.9674253	0.0000010
exper	0.0270546	0.0055968	4.8339367	0.0000019
tenure	0.0251399	0.0071276	3.5271047	0.0004699

nonwhite	0.0136412	0.0598942	0.2277554	0.8199551
female	-0.2699470	0.0406997	-6.6326461	0.0000000
married	0.0157473	0.0421204	0.3738639	0.7087077
numdep	-0.0242575	0.0157517	-1.5399962	0.1243681
smsa	0.1722506	0.0455846	3.7786983	0.0001821
northcen	-0.0432550	0.0492583	-0.8781258	0.3804134
south	-0.0320239	0.0479117	-0.6683952	0.5042747
west	0.0525875	0.0613380	0.8573385	0.3917817
construc	-0.0220009	0.1042152	-0.2111100	0.8329111
ndurman	-0.1233413	0.0696197	-1.7716444	0.0772304
trcommpu	-0.1087629	0.1055787	-1.0301595	0.3035703
trade	-0.3380434	0.0592953	-5.7010138	0.0000000
services	-0.3553896	0.0739820	-4.8037287	0.0000022
profserv	-0.0822468	0.0635773	-1.2936497	0.1965482
profocc	0.2401081	0.0551351	4.3549078	0.0000170
clerocc	0.0685419	0.0631951	1.0846072	0.2787622
servocc	-0.0723435	0.0637633	-1.1345633	0.2572517
expersq	-0.0005481	0.0001183	-4.6323102	0.0000049
tenursq	-0.0004357	0.0002341	-1.8613969	0.0634364

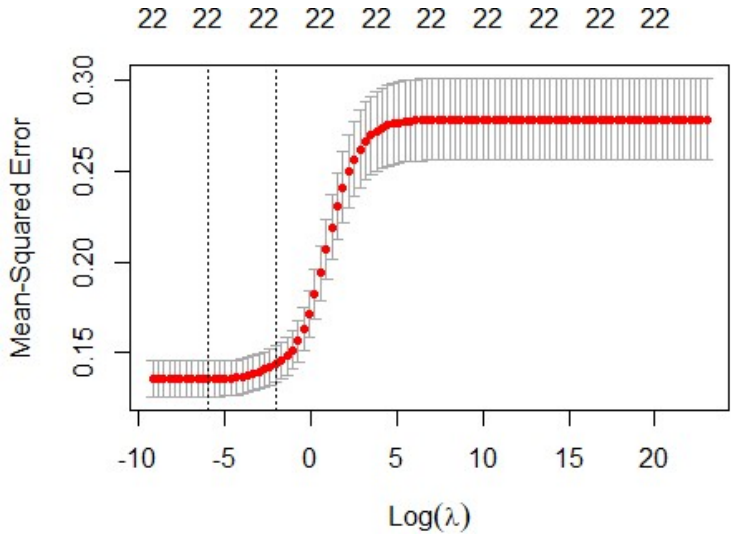
Como se considera que la estimación por MCO produce coeficientes eficientes e insesgados no es necesario aplicar ninguna corrección a los niveles de significación. Sin embargo, se hace necesario quitar muchas variables explicativas. Se va quitando una cada vez hasta que no queda ningún coeficiente con *p-valor* por encima de 0.05.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7778076	0.1183485	6.572180	0.0000000
educ	0.0459144	0.0084759	5.417049	0.0000001
exper	0.0290455	0.0049330	5.887954	0.0000000
tenure	0.0130736	0.0029617	4.414261	0.0000130
female	-0.2556802	0.0363653	-7.030885	0.0000000
smsa	0.2067706	0.0430491	4.803129	0.0000022
trade	-0.2805740	0.0411568	-6.817202	0.0000000
services	-0.2856238	0.0607697	-4.700100	0.0000036
profocc	0.2146007	0.0457834	4.687302	0.0000038
servocc	-0.1169125	0.0550800	-2.122595	0.0343955
expersq	-0.0005746	0.0001062	-5.410306	0.0000001

El MSE de prueba tras quitar todas las variables con *p-valor* por encima de 0.05 es 0.1633, que empeora las cifras del error reportado por el MCO con todas las variables.

G. Regresión Ridge (VC 10-veces)

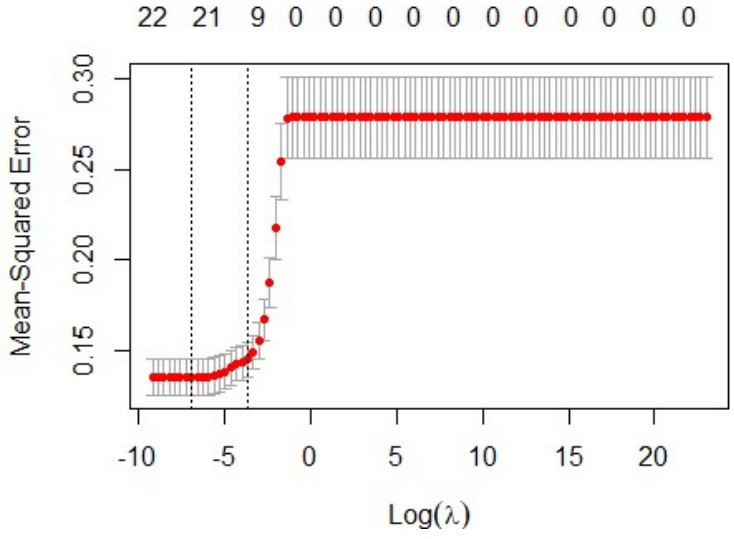
La regresión ridge es un método de contracción de los coeficientes estimados de las variables. El algoritmo busca minimizar el MSE penalizando el número de variables (la suma de los regresores al cuadrado), ponderada por el parámetro de ajuste λ. Se estiman tantas regresiones como valores de λ quieran probarse para encontrar el MSE por VC.



Aplicando la regla del codo se selecciona un valor de λ de **0.1292** y se obteniene un MSE de prueba de **0.1622**.

H. Regresión LASSO (VC 10-veces)

La regresión LASSO también es un método de contracción de los coeficientes. El algoritmo también busca minimizar el MSE penalizando el número de variables (esta vez como la suma de los valores absolutos de los regresores), ponderada por el parámetro de ajuste λ. Igual que con ridge, se estiman tantas regresiones como valores de λ quieran probarse para encontrar el MSE por VC. En esta aproximación es posible determinar un número óptimo de variables con coeficientes distintos de cero.

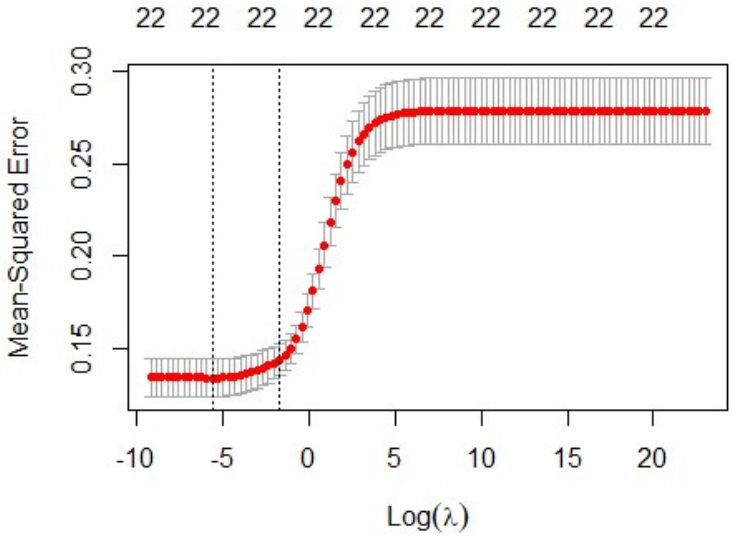


En este caso se seleccionan **11 variables** con coeficientes distintos de cero y, tras aplicar la regla del codo para determinar λ (**0.0254**), se obtiene un MSE de prueba de **0.1628**.

I. Regresión Ridge y LASSO (VC 5-veces)

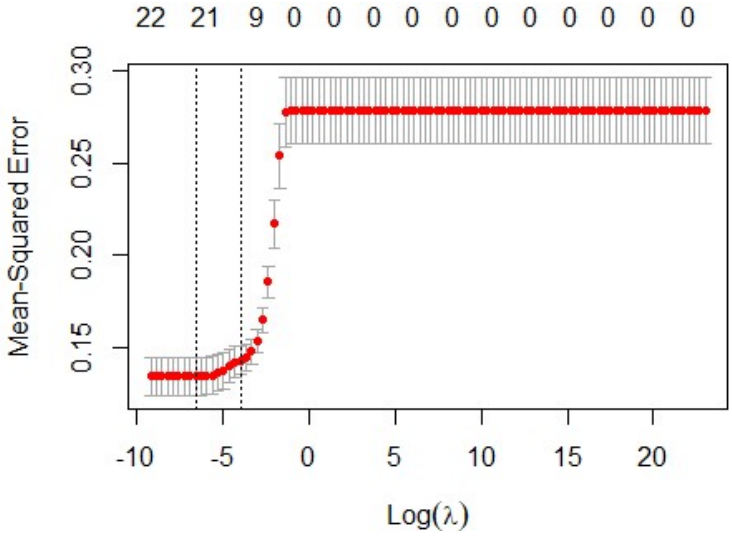
Se repite el proceso de los apartados G y H para VC 5 veces.

I.1. Regresión Ridge (VC 5-veces)



Aplicando la regla del codo se selecciona un valor de λ de **0.1789** y se obtiene un MSE de prueba de **0.1639**.

I.2. Regresión LASSO (VC 5-veces)

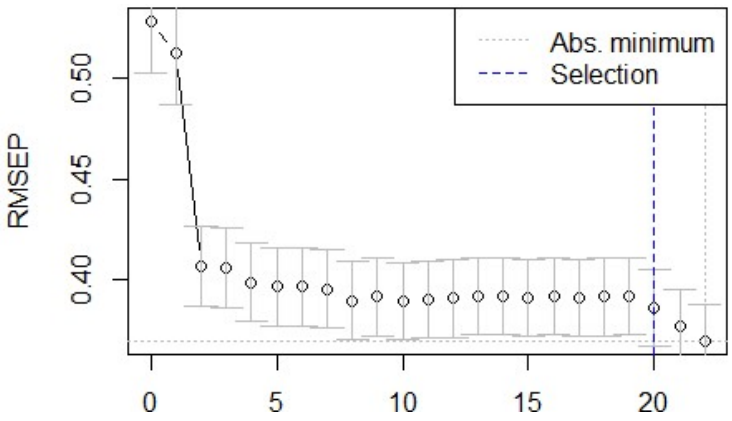


Se seleccionan **11 variables** con coeficientes distintos de cero y, tras aplicar la regla del codo para determinar λ (**0.0183**), se obtiene un MSE de prueba de **0.1626**.

J. Componentes principales (CP)

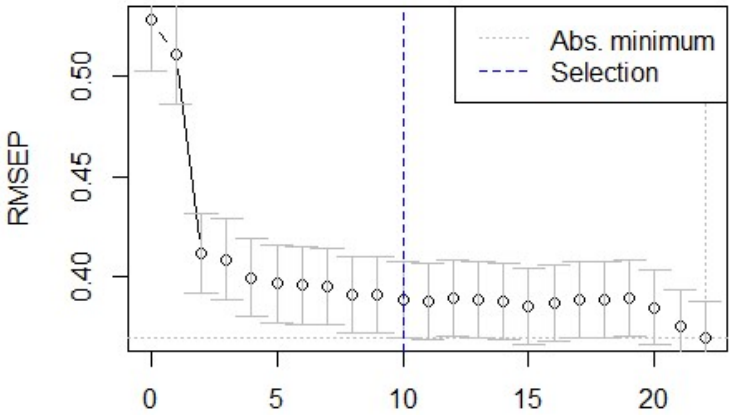
El método de CP efectúa combinaciones lineales del conjunto de variables de forma que se recoja la máxima varianza de las variables, reduciendo la dimensionalidad. Los sucesivos CP son ortogonales unos con otros. Se selecciona el menor número posible de CP que está como máximo a una desviación típica del mínimo del MSE de prueba de VC.

J.1. Componentes principales (VC 10-veces)



Se seleccionan **20 CP** y se obtiene un MSE de prueba de **0.1653**.

J.2. Componentes principales (VC 5-veces)



Se seleccionan **10 CP** y se obtiene un MSE de prueba de **0.1624**.

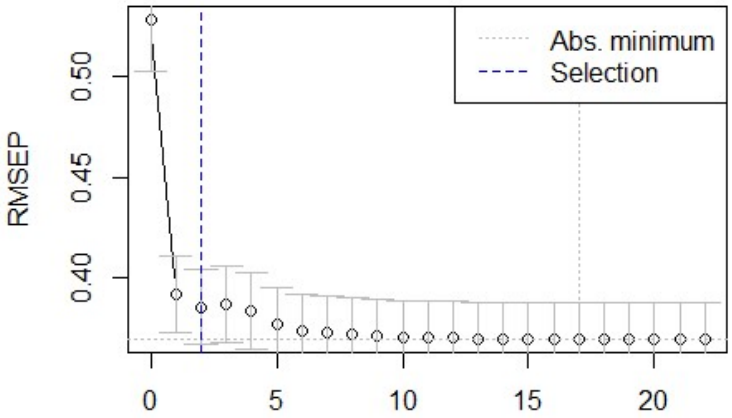
J.3. Resumen de resultados

Método	CP	RMSE
10-FOLDS-CV	20	0.1653
5-FOLDS-CV	10	0.1624

K. Mínimos Cuadrados Parciales (PLS)

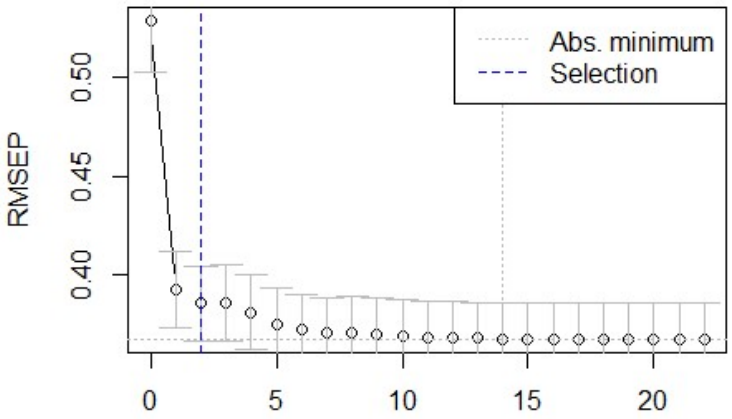
PLS es muy parecido a CP pero intenta maximizar la covarianza entre las variables dependiente e independientes. En esta aproximación se determina el número de CP de acuerdo con el del valor mínimo del MSE de VC y también según el método del codo. Por cada uno de los métodos se determina el MSE de prueba de VC.

K.1. Mínimos Cuadrados Parciales (VC 10-veces)



De acuerdo con el criterio del error mínimo de VC se escogen **16 CP** y se obtiene un MSE de prueba de **0.1608**. Por la regla del codo se utilizan **2 CP** y se reporta un MSE de prueba de **0.1623**.

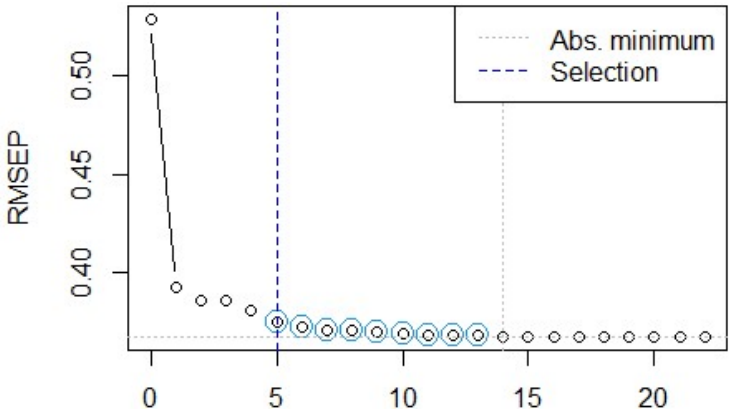
K.2. Mínimos Cuadrados Parciales (VC 5-veces)



De acuerdo con el criterio del error mínimo de VC se escogen **16 CP** y se obtiene un MSE de prueba de **0.1608**. Por la regla del codo se utilizan **2 CP** y se reporta un MSE de prueba de **0.1623**.

K.3. Mínimos Cuadrados Parciales (Regla de la Permutación)

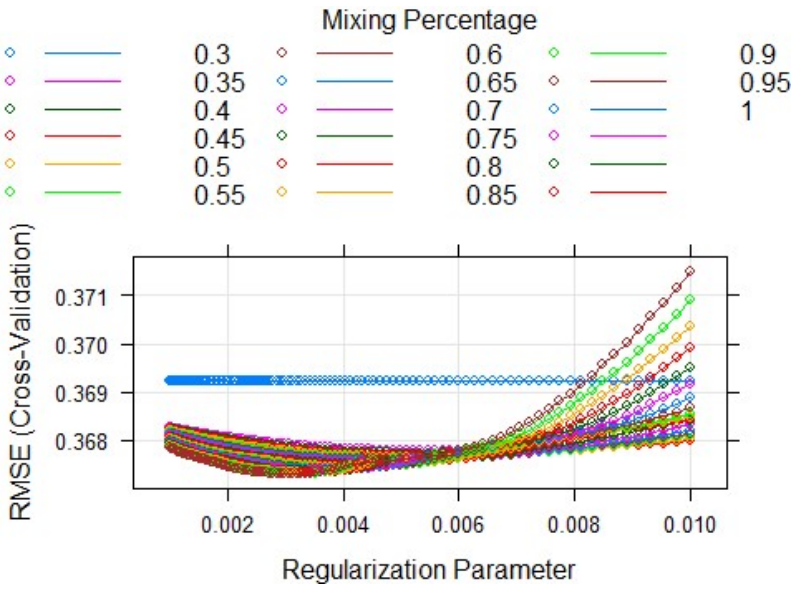
La aproximación "randomization" primero selecciona el número de CP con menor MSE de CV. Luego se van quitando CP y se comprueba si los MSE de los modelos sucesivos son significativamente superiores que en el modelo de referencia (con un 5% de significación). Esto lleva a considerar cada modelo respecto un valor; se selecciona aquel con menor número de CP que no es significativamente peor que el de referencia.



Se seleccionan **5 CP** y se obtiene un MSE de prueba de **0.1623**.

L. Red elástica (VC 10-veces)

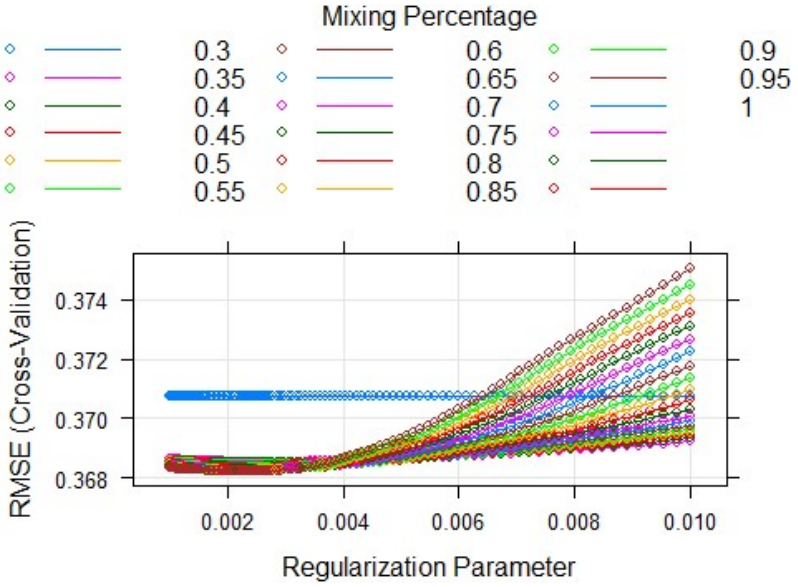
La red elástica "combina" los métodos de penalización de ridge y LASSO de acuerdo con un factor α que otorga más o menos peso a cada método. El método se sintoniza de acuerdo con λ tal como se ha hecho en los apartados G, H e I y con α entre 0 y 1.



La red elástica reporta un MSE de prueba por VC 10 veces de **0.1591** con $\boldsymbol{\alpha = 1}$ (LASSO) y $\boldsymbol{\lambda = 0.0031}$. **21 variables** tienen coeficientes estimados distintos de cero.

M. Red elástica (VC 5-veces)

Se repite el proceso del apartado L con VC 5 veces.



La red elástica reporta un MSE de prueba por VC 5 veces de **0.1596** con $\boldsymbol{\alpha = 1}$ (LASSO) y $\boldsymbol{\lambda = 0.002}$. **21 variables** tienen coeficientes estimados distintos de cero.

N. Selección de modelos Ridge, LASSO y red elástica con inferencia

Para contrastar la significación individual de los coeficientes de los modelos de contracción es necesario hacer uso de algún método alternativo que permita hacer

inferencia. Se opta por usar *bootstrap* para determinar intervalos de confianza al 5%. Si 0 está contenido en esos intervalos puede decirse que la variable no es significativa.

Bootstrap es una técnica de remuestreo a partir de la cual se generan una serie de muestras con reemplazamiento sobre la muestra original. Por cada una se hace una regresión que tendrá sus propios coeficientes, la media aritmética de los cuales representarán los verdaderos coeficientes del modelo.

N.1. Ridge por VC 10 veces

Se escoge un modelo con 10 variables que reporta un MSE de prueba de 0.1707.

N.2. Ridge por VC 5 veces

Acaba escogiendo un modelo idéntico al de VC 10 veces con el mismo MSE de prueba.

N.3. LASSO por VC 10 veces

Se escoge un modelo con 9 variables que reporta un MSE de prueba de 0.1745.

N.4. LASSO por VC 5 veces

Acaba escogiendo un modelo idéntico al de VC 10 veces con el mismo MSE de prueba.

N.5. Red elástica VC 10 veces

En este caso se escoge un modelo con 9 variables. Con este modelo la red elástica reporta un error de prueba por VC 10 veces de 0.1667 con α = 0.2 y λ = 0.1485.

N.6. Red elástica VC 5 veces

Acaba escogiendo un modelo idéntico al de VC 10 veces con los mismos valores de α y λ, y reportando el mismo MSE de prueba.

O. LASSO riguroso y ajustes Post-LASSO

Los métodos de VC producen resultados distintos según el número de partes que se hayan efectuado. Son modelos poco robustos y sin una justificación teórica. El LASSO robusto es un método de alta dimensión capaz de penalizar con dependencia de los datos o independientemente de estos. El análisis Post-LASSO efectúa un reajuste del modelo con las variables seleccionadas con el LASSO robusto.

Además, estos métodos son capaces de proponer intervalos de confianza y estadísticos t consistentes para los estimadores LASSO bajo heterocedasticidad y también permiten contrastar modelos conjuntamente. El proceso de inferencia se ve en el apartado P.

O.1. Penalización independiente de los datos

Cuando introducimos una penalización independiente de los datos el valor de λ se escoge de acuerdo con una fórmula teórica dada. El error de prueba del LASSO riguroso

independiente de los datos es 0.1648 y se obtienen 11 variables con coeficientes distintos de cero.

O.2. Penalización independiente de los datos (ajuste post-LASSO)

El MSE de prueba del LASSO riguroso independiente de los datos con el ajuste Post-LASSO es 0.1641 y se obtienen 8 variables con coeficientes distintos de cero.

O.3. Penalización dependiente de los datos

Con penalización dependiente de los datos el valor de λ se escoge teniendo en cuenta el diseño de la matriz. En este caso acaba escogiendo un modelo idéntico al de O.1. con el mismo MSE de prueba.

O.4. Penalización dependiente de los datos (ajuste post-LASSO)

Acaba escogiendo un modelo idéntico al de O.2. con el mismo MSE de prueba.

P. LASSO riguroso y ajustes Post-LASSO con inferencia

Se usan las estimaciones del anterior para seleccionar las variables significativas y efectuar las regresiones LASSO robustas directas y reestimadas (Post-LASSO).

P.1. Penalización independiente de los datos con inferencia

Seleccionadas 9 variables, el MSE de prueba del LASSO riguroso independiente de los datos del modelo con las variables significativas es 0.1686.

P.2. Penalización independiente de los datos con inferencia (ajuste post-LASSO)

El error de prueba del LASSO riguroso independiente de los datos del modelo con las variables significativas y ajustado a posteriori es 0.1747.

P.3. Penalización dependiente de los datos con inferencia

Acaba escogiendo un modelo idéntico al de P.1. con el mismo MSE de prueba.

P.4. Penalización dependiente de los datos con inferencia (ajuste post-LASSO)

Acaba escogiendo un modelo idéntico al de P.2. con el mismo MSE de prueba.

Q. Resultados

	Ridge	LASSO	R. elástica	Ridge (inf)	LASSO (inf)	R. elástica (inf)
MSE.VC10	0.1622	0.1628	0.1591	0.1707	0.1745	0.1667
MSE.VC5	0.1639	0.1626	0.1596	0.1707	0.1745	0.1667

	CP, codo	PLS, min	PLS, codo	PLS, perm
MSE.VC10	0.1653	0.1608	0.1623	-
MSE.VC5	0.1624	0.1608	0.1623	-
MSE.LOO	-	-	-	0.1623

	LASSO R	LASSO R aj	LASSO R (inf)	LASSO R aj (inf)
Indep.	0.1648	0.1641	0.1686	0.1747
Dep	0.1648	0.1641	0.1686	0.1747

R. Modelo Final (Coeficientes y significación)

El modelo con el menor MSE de prueba es Red Elástica por VC 10 veces. Por este método no es posible obtener directamente los coeficientes del modelo ni sus p-valores. Para estimarlos es necesario usar *bootstrap* sobre el modelo con las variables seleccionadas (todas excepto `construc`) y los valores fijos de α y λ hallados en L.

Con los valores medios de los coeficientes y sus desviaciones típicas es posible calcular los estadísticos t y, con estos y el número de observaciones, determinar los p-valores.

	Estimate	Std.Error	t_value	p_values
(Intercept)	0.9062116	0.1333028	6.7981	0.0000000
educ	0.0432734	0.0101388	4.2681	0.0000245
exper	0.0203763	0.0052696	3.8668	0.0001280
tenure	0.0216387	0.0071826	3.0127	0.0027486
nonwhite	0.0108409	0.0600622	0.1805	0.8568528
female	-0.2618945	0.0386505	-6.7760	0.0000000
married	0.0304832	0.0367522	0.8294	0.4073417
numdep	-0.0155931	0.0132195	-1.1796	0.2388535
smsa	0.1734195	0.0440759	3.9346	0.0000977
northcen	-0.0326893	0.0399377	-0.8185	0.4135386
south	-0.0240120	0.0361247	-0.6647	0.5066131
west	0.0645336	0.0592452	1.0893	0.2766717
ndurman	-0.0894865	0.0590792	-1.5147	0.1306147
trcommpu	-0.0695564	0.0821215	-0.8470	0.3974882
trade	-0.3040731	0.0519338	-5.8550	0.0000000
services	-0.3130504	0.0801539	-3.9056	0.0001097
profserv	-0.0425269	0.0530702	-0.8013	0.4233985
profocc	0.2322954	0.0466836	4.9760	0.0000010
clerocc	0.0439659	0.0514443	0.8546	0.3932500
servocc	-0.0949463	0.0569215	-1.6680	0.0960675
expersq	-0.0004031	0.0001130	-3.5681	0.0004017
tenursq	-0.0003002	0.0002621	-1.1453	0.2527449