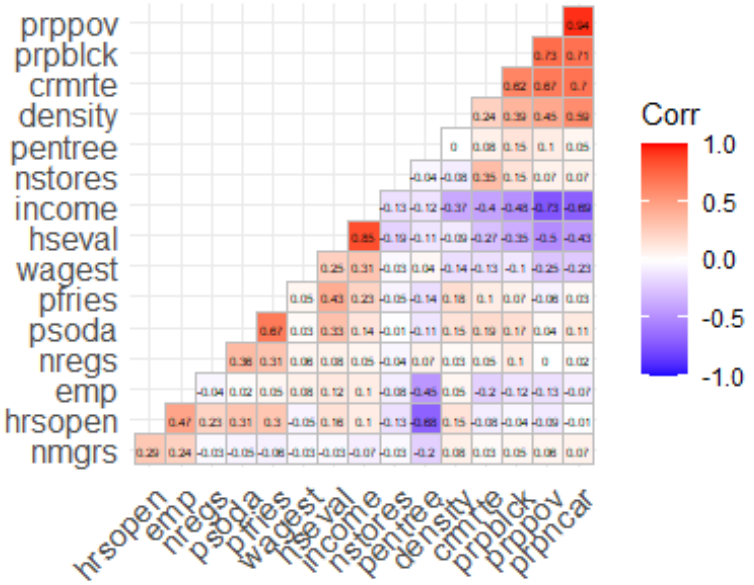
Proyecto Final Econometría

Arturo González Moya

3 de abril de 2021

a) El conjunto de datos que vamos a estudiar en este trabajo está contenido en la librería "wooldridge"de *R* y se llama "discrim". El conjunto de datos fue creado para ver si existe una discriminación en los precios de las cadenas de comida rápida en función de la raza y las características de ingresos de un área en Estados Unidos, más concretamente compara los estados de New Jersey y Pensilvania. Este conjunto de datos contiene 410 observaciones y 37 variables. En este trabajo la variable que intentaremos predecir (o variable dependiente) será la variable *psoda*. Una vez limpios los datos explicaremos que variables hemos seleccionado.

Para comenzar, los datos necesitan una limpieza ya que contienen valores perdidos y existe multicolinealidad perfecta entre diferentes variables independientes. Veamos detalladamente la multicolinealidad entre variables pero para ello, no mostraremos las variables como *lpfries, lincome, ldensity* ya que están en otra escala.



Como podemos observar, existen muchas variables que tienen, en valor absoluto, una alta correlación entre ellas. Esto hace que los supuestos de la regresión no se cumplan. Lo que se ha realizado ha sido la eliminación de variables que más correlación tenían con otras hasta que no se observa una correlación alta entre variables independientes.

También se han eliminado las variables *chain, state* ya que expresan lo mismo que las variables *NJ, BK, KFC, RR* que están en formato one hot encoding y las variables *hrsopen2, pentree2, wagest2, nregs2, psoda2, pfries2, nmgrs2, emp2* ya que están medidas en otro tiempo y generarían un sesgo.

Tras esta limpieza, el número de variables que seleccionamos son 21 y son las siguientes:

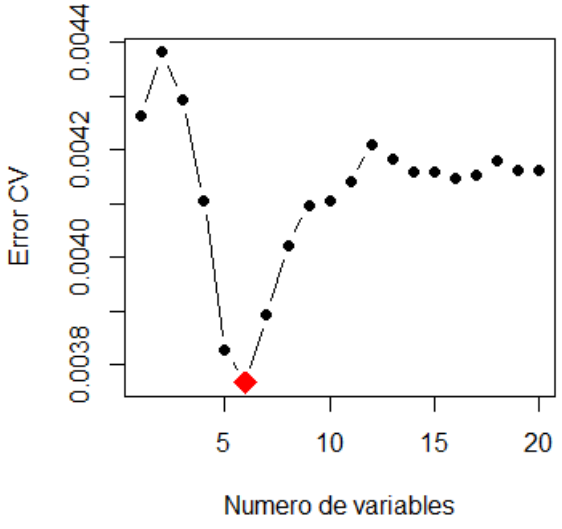
- **psoda**: precio de la soda mediana, primera vez que se mide. (Esta es la variable dependiente)
- **pfries**: precio de las patatas pequeñas, primera vez que se mide.
- **pentree**: precio del entrante (hamburguesa o pollo).
- **wagest**: salario inicial, primera vez que se mide.
- **nmgrs**: número de gerentes, primera vez que se mide.
- **nregs**: número de registros, primera vez que se mide.
- **emp**: número de empleados, primera vez que se mide.
- **hrsopen**: Horas que un establecimiento está abierto.
- **compown**: Variable binaria con valores 1 si es propiedad de la empresa, 0 si no.
- **density**: densidad de la población en una ciudad.
- **crmrte**: ratio de crimen en una ciudad.
- **prpblck**: proporcion de gente de color.
- **nstores**: número de tiendas.
- **income**: ingresos medios de una familia.
- **NJ**: Variable binaria. 1 si está en New Jersey, 0 si no
- **BK**: Variable binaria. 1 si es el establecimiento Burger King, 0 si no.
- **KFC**: Variable binaria. 1 si es el establecimiento Kentucky Fried Chicken, 0 si no.
- **RR**: Variable binaria. 1 si es el establecimiento Roy Rogers, 0 si no.
- **lpfries**: Logaritmo de *pfries*.
- **lincome**: Logaritmo de *income*.
- **ldensity**: Logaritmo de *density*.

Separamos en conjunto de entrenamiento y conjunto de test y realizamos la regresión por mínimos cuadrados de las variables que hemos seleccionado. En esta regresión, nuestra categoría de control son los establecimientos de Wendy's que se encuentran en Pensilvania.

Realizamos la predicción con el conjunto de test y calculamos la raíz del error cuadrático medio que es de 0,06750068.

b) En este apartado ajustamos una regresión por MCO utilizando la mejor selección de subconjuntos. Primero seleccionaremos el número de variables que minimicen el error de validación cruzada 10 veces y luego utilizaremos la regla del codo.

Para el primer caso, el menor erro mediante validación cruzada 10 veces se da con 6 variables. Veamoslo gráficamente:



La regresión sería la siguiente:

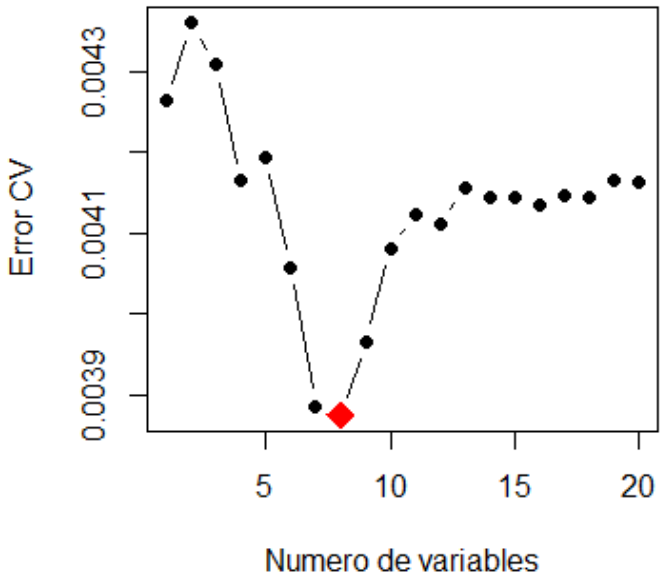
$$psoda = 0{,}54143894 + 0{,}40113678 pfries + 0{,}03083829 pentree + 0{,}05293740 prpblck + 0{,}03844590 NJ + 0{,}07960776 BK + 0{,}08529877 RR$$

Al realizar la predicción sobre el conjunto de prueba, obtenemos un RMSE de 0,06432277.

Utilizando la regla del codo, seleccionaríamos 5 variables. El RMSE de la predicción con el conjunto de test es de 0,06587767, que es mayor que el obtenido con las 6 variables anteriores.

c) Ahora repetiremos lo que hemos realizado en el apartado anterior pero utilizando el método de selección hacia adelante.

El número de variables que minimizan el error cuadrático medio de validación cruzada 10 veces es 8. Veamoslo gráficamente.



La regresión sería la siguiente:

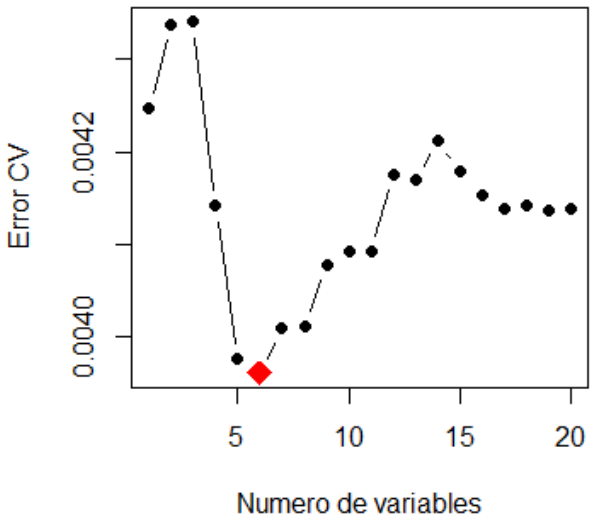
$$psoda = 0{,}557291666 + 0{,}399472877 pfries + 0{,}025266272 pentree + 0{,}005116091 nregs + 0{,}049776012 prpblck + 0{,}037582231 NJ + 0{,}075815690 BK + 0{,}071978166 RR - 0{,}005768185 nmgrs$$

El error de predicción de este modelo utilizando el conjunto de test es de 0,0649306.

Si ahora seleccionamos el número de variables utilizando la regla del codo, vemos que nos dice de coger 7 variables y el error de la predicción utilizando el conjunto de test es de 0,06487047, que es menor que el del modelo anterior con 8 variables.

d) En este apartado, aplicaremos los métodos de los dos apartados anteriores pero seleccionaremos el número de variables mediante validación cruzada 5 veces.

Empezamos con el método de mejor selección de subconjuntos. El menor error de validación cruzada 5 veces nos dice que tenemos que seleccionar 6 variables. Lo podemos ver en el siguiente gráfico.



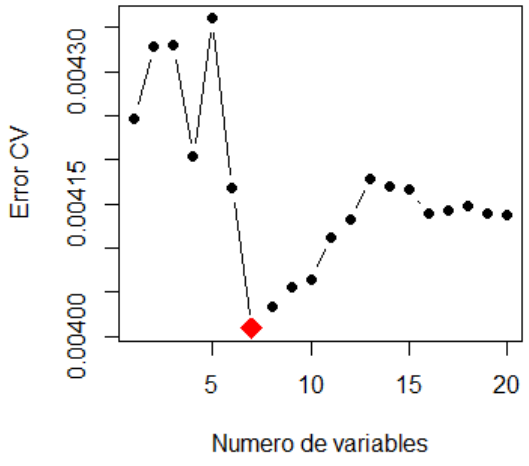
La regresión sería la misma que la obtenida en el apartado b), por lo tanto, su error de prueba es de 0,06432277. Utilizando el método del codo para seleccionar las variables también nos dice que seleccionemos 5 variables como en el apartado b).

Pasamos al método de selección por pasos hacia adelante. Si miramos el menor error de validación cruzada 5 veces nos indica que debemos seleccionar 7 variables. La regresión sería la siguiente:

$$psoda = 0{,}534582220 + 0{,}402342370 pfries + 0{,}026140860 pentree + 0{,}005035495 nregs + 0{,}049838893 prpblck \\ +0{,}038115294 NJ + 0{,}073818034 BK + 0{,}071774515 RR$$

El error de prueba cometido en la predicción utilizando el conjunto de test es de 0,06440577.

Si seleccionamos el número de variables mediante la regla del codo, deberíamos de seleccionar 7 variables. Veamoslo en el siguiente gráfico.



El error de prueba obtenido con este modelo de 7 variables sería 0,06487047 que es mayor que el del modelo anterior.

e) Creamos una tabla con los modelos que llevamos hasta ahora para ver cual es el mejor.

Selección de modelos		
Modelo	Error de prueba	Número variables
Mínimos cuadrados ordinarios	0.06750068	
Selección por subconjuntos CV 10	0.06432277	6
Selección por pasos hacia adelante CV 10	0.06493060	8
Selección por subconjuntos CV 5	0.06432277	6
Selección por pasos hacia adelante CV 5	0.06487047	7

Cuadro 1: Selección de modelos

Vemos que el error mínimo se da cuando seleccionamos 6 variables, por lo que nuestro modelo será este:

$$psoda = 0{,}54143894 + 0{,}40113678 pfries + 0{,}03083829 pentree + 0{,}05293740 prpblck + 0{,}03844590 NJ + 0{,}07960776 BK + 0{,}08529877 RR$$

Cabe recalcar que los errores de prueba entre los diferentes métodos no son muy diferentes.

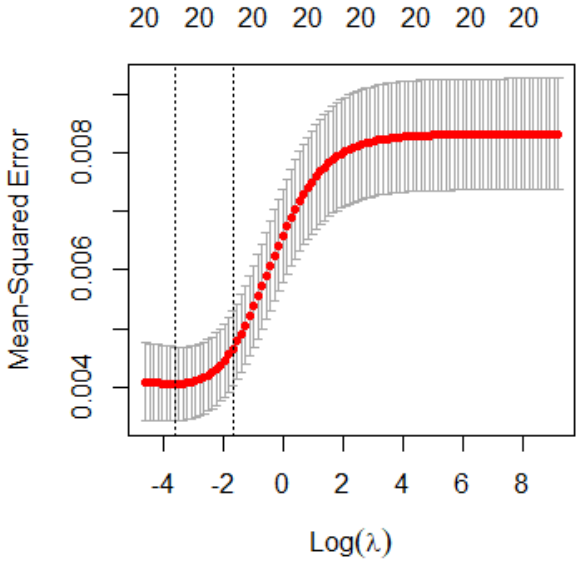
f) Como hemos dicho anteriormente, nos quedaremos con el modelo con 6 variables. Ya que el modelo que hemos elegido es el de mejor selección de subconjuntos, los p-valores estarán sesgados. Para arreglarlo, utilizaremos el método de **Bonferroni-Holm**.

Los p-valores obtenidos son: $3{,}75e-16$ para *pfries*, 0,000513 para *pentree*, 0,017370 para *prpblck*, 0,000329 para *NJ*, $1{,}11e-08$ para *BK* y $6{,}72e-08$ para *RR*. Lo que hemos de hacer es seleccionar los p-valores que sean menores que 0,05/6 y estas variables serán las que su coeficiente es significativo.

Se ha obtenido que todas las variables menos *prpblck* son significativas. Calculamos la regresión con las nuevas variables y vemos su RMSE con el conjunto de prueba que es de 0,06587767. El error obtenido con la regresión de 6 variables era de 0,06432277, por lo que la regresión con 5 variables es peor que con 6.

g) En este apartado ajustaremos una regresión Ridge en el conjunto de entrenamiento, obteniendo el parámetro λ mediante validación cruzada 10 veces.

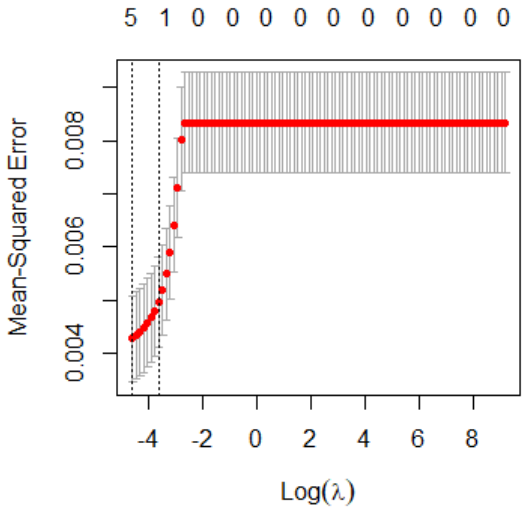
La validación cruzada nos dice que el mejor lambda es 0,02656088. Veamoslo en el siguiente gráfico.



Vemos que el error mínimo se da cuando $log(\lambda)$ se encuentra entre -3 y -4. Por lo tanto realizamos la predicción sobre los datos de test y vemos cuál es el RMSE, que tiene un valor de 0,06518127.

Si tomamos el valor de λ óptimo mediante la regla del codo (que es de 0,1873817), vemos que el error de prueba es de 0,06786258 que es mayor que el obtenido con el λ de validación cruzada.

h) Ahora realizaremos un modelo LASSO obteniendo el parámetro λ mediante validación cruzada 10 veces. El valor de λ que minimiza el error de validación cruzada 10 veces es 0,01, el mismo que en el apartado anterior. Esto lo podemos observar en el siguiente gráfico.



Vemos que el error mínimo se da cuando $log(\lambda)$ es un poco mas pequeño que -4. Si miramos el RMSE obtenido en la predicción utilizando el conjunto de test, vemos que es de 0,06841878.

Si seleccionamos λ utilizando la regla del codo (que tendría un valor de 0,02656088), obtendríamos un RMSE de prueba de 0,07257891 que es notablemente mayor que el obtenido el valor de λ de validación cruzada.

i) En este apartado, ajustaremos los modelos de Ridge y LASSO pero seleccionando λ mediante validación cruzada 5 veces.

Empezaremos con el modelo Ridge. El valor de λ seleccionado a partir de validación cruzada 5 veces es 0,0231013. El RMSE de prueba obtenido con el conjunto de test es 0,06524422.

Si ahora seleccionamos el valor de λ mediante la regla del codo, el valor que obtenemos es 0,1417474 y el RMSE obtenido con el conjunto de test es 0,06687203, que es mayor que el error obtenido mediante validación cruzada 5 veces.

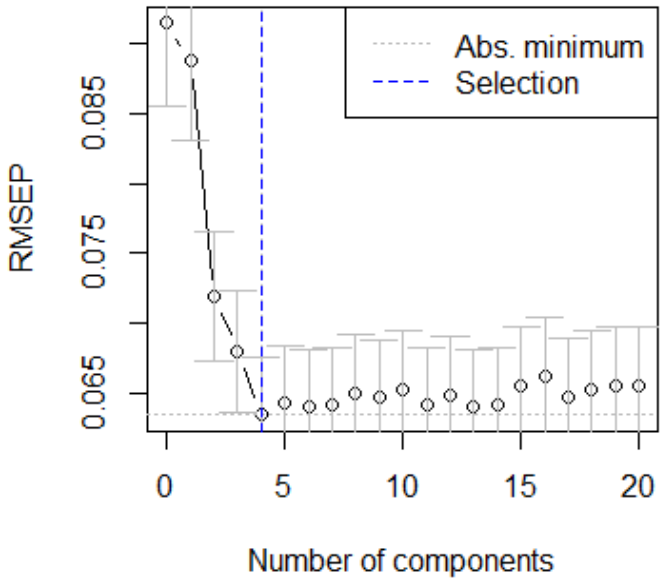
Pasamos ahora a ajustar el modelo LASSO utilizando validación cruzada 5 veces para determinar el valor de λ. Obtenemos que este valor es de 0,01. El RMSE del conjunto de test es 0,06841878, que es igual que el error obtenido ajustando el modelo LASSO mediante validación cruzada 10 veces.

Si tomamos ahora el valor de λ utilizando la regla del codo, obtenemos un valor de 0,03053856 y un RMSE, utilizando el conjunto de prueba, de 0,07313567.

j) En este apartado, ajustaremos un modelo de Componentes Principales y seleccionaremos el número de componentes mediante validación cruzada 10 y 5 veces. Para este modelo, escalaremos los datos ya que hay variables medidas en diferentes escalas.

Comenzamos seleccionando el número de componentes principales que nos indique la validación cruzada 10 veces. Esta nos dice que tomemos 13 componentes principales. El RMSE con el conjunto de test utilizando estas 13 componentes principales es 0,06706057.

Si ahora seleccionamos el número de componentes principales mediante la regla del codo, hemos de tomar 4. Podemos verlo en el siguiente gráfico.



El RMSE obtenido con este número de componentes principales es de 0,06408815, que es menor al error con 13 variables y además el modelo es más simple al tener menos variables.

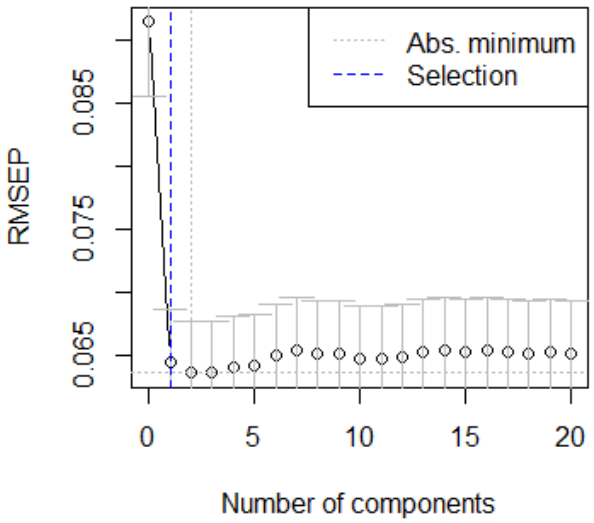
Pasamos a seleccionar el número de componentes principales mediante validación cruzada 5 veces. Según los resultados obtenidos, el número de componentes que minimiza el error de validación cruzada 5 veces es 13. El RMSE obtenido en la predicción con el conjunto de prueba es de 0,06706057.

Si seleccionamos el número de componentes principales mediante la regla del codo, nos dice de tomar 4 componentes, igual que con la regla del codo en validación cruzada 10 veces. El RMSE obtenido con el conjunto de prueba es 0,06408815, que es menor que el obtenido con el número de componentes principales seleccionadas con validación cruzada 5 veces y este modelo de 4 variables es más simple.

k) En este apartado ajustaremos un modelo PLS, seleccionando el número de componentes principales mediante validación cruzada 10 y 5 veces. Además, seleccionaremos también el número de componentes principales mediante la opción *randomization* en **R**. Para este modelo, escalaremos los datos ya que hay variables medidas en diferentes escalas.

Comenzamos utilizando validación cruzada 10 veces para seleccionar el número de componentes principales. Vemos que el error mínimo de validación cruzada 10 veces nos indica que seleccionemos 3 componentes principales. El RMSE obtenido con el conjunto de prueba es 0,06495082.

Si utilizamos la regla del codo, hemos de seleccionar una componente principal para nuestro modelo, como podemos ver en el siguiente gráfico.

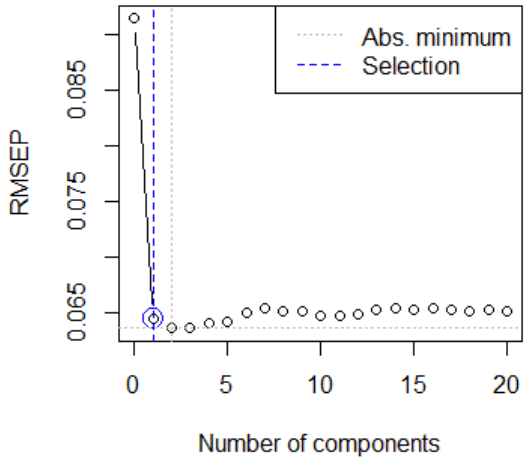


Este modelo tiene un RMSE con el conjunto de prueba de 0,06681841, que es un error mayor que utilizando 3 componentes principales.

Si ahora seleccionamos el número de componentes principales mediante validación cruzada 5 veces, vemos que hemos de tomar 2 componentes principales. El RMSE cometido con este modelo sobre el conjunto de test es de 0,06384467.

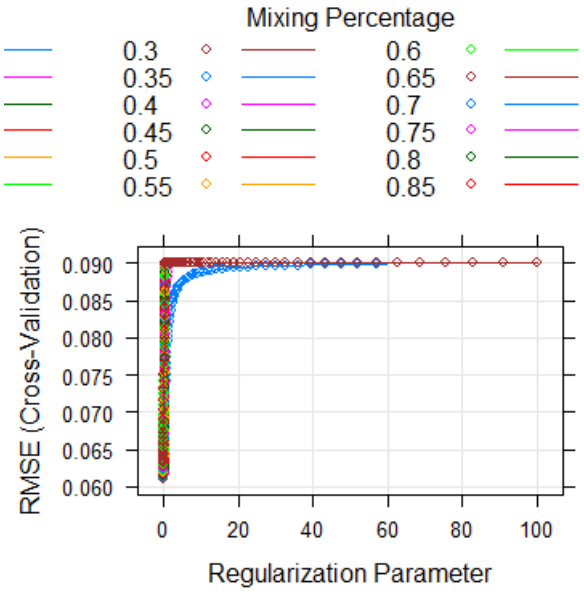
Utilizando la regla del codo con validación cruzada 5 veces, nos dice de seleccionar una única componente principal, obteniendo un RMSE con este modelo en el conjunto de test de 0,06681841.

Por último en este apartado, seleccionaremos el número de componentes principales mediante la opción *randomization* pero antes explicaremos que hace esta opción. Primero, el modelo PLS lo que hace es eliminar componentes principales que tengan alta correlación y realiza una regresión por MCO. Realizar *randomization* en **R** lo que hace es tomar el número de componentes principales que minimicen la curva de validación cruzada y va eliminando componentes principales hasta que no encuentre que hay una diferencia significativa (que de base es $\alpha = 0{,}01$). En nuestro conjunto de datos, esta opción nos dice que seleccionemos una componente principal como podemos ver el en el siguiente gráfico.



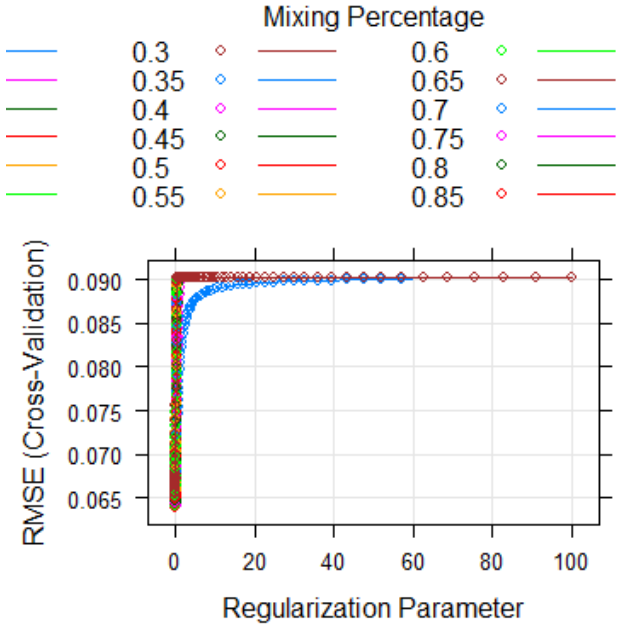
Este modelo tiene un RMSE utilizando el conjunto de test de 0,06681841, igual que en el obtenido con el modelo seleccionando las variables con la regla del codo en validación cruzada 5 veces.

l) En este apartado ajustaremos un modelo LASSO con la restricción de Red Elástica. Seleccionaremos los parámetros α y λ mediante validación cruzada 10 veces. Los parámetros que escogemos (por la validación cruzada 10 veces) son $\alpha = 0{,}15$ y $\lambda = 0{,}01$ como podemos ver en el siguiente gráfico.



Este método selecciona 13 variables y el modelo con estas 13 variables sobre el conjunto de test tiene un RMSE de 0,06553641.

m) Lo que haremos ahora será ajustar un modelo LASSO con la restricción de Red Elástica (igual que en el apartado anterior) pero utilizando validación cruzada 5 veces para seleccionar los valores de α y λ. Con nuestro conjunto de entrenamiento, obtenemos estos parámetros $\alpha = 0{,}2$ y $\lambda = 0{,}01204504$ como podemos ver en el siguiente gráfico.



El número de variables seleccionadas para el modelo son 11 y el RMSE obtenido con el conjunto de test es de 0,06559801.

n) Comenzamos contrastando la significación los coeficientes del modelo Ridge obtenido con validación cruzada 10 veces. Observamos que solo 3 variables son significativas, por lo que si realizamos un nuevo modelo con estas variables, obtenemos un error con el conjunto de prueba de 0,07188722 con un valor de λ de 0,01.

El siguiente que estudiaremos es el modelo LASSO cuyo coeficiente fue elegido con validación cruzada 10 veces. En este modelo tenemos que son significativas las mismas 3 variables que antes. Realizando el modelo LASSO con estas 3 variables tenemos que el error de prueba es de 0,071706 con un valor de λ de 0,01.

Seguimos con el modelo Ridge con el coeficiente elegido mediante validación cruzada 5 veces. Las variables significativas son las mismas 3 que antes, por lo que si creamos el nuevo modelo Ridge con estas variables, obtenemos que el error de prueba es de 0,07188722 con un valor de λ de 0,01.

El siguiente será el modelo LASSO con el coeficiente elegido mediante validación cruzada 5 veces. Las variables significativas que hemos cogido siempre en este apartado, por lo que si creamos el nuevo modelo LASSO con estas variables, obtenemos que el error de prueba es de 0,071706 con un valor de λ de 0,01.

Acabaremos con los modelos de LASSO con restricción de red elástica, seleccionando los parámetros con validación cruzada 5 y 10 veces. En ambos casos, las variables que son significativas son las 3 que hemos estado seleccionando en este apartado. Además, coincide que en ambos casos los nuevos parametros son $\alpha = 0$ y $\lambda = 0{,}1$, que son los del modelo Ridge que hemos obtenido antes, por lo que sus errores de prueba son 0,07188729.

o) En este apartado ajustaremos dos modelos RLASSO donde en uno el parámetro λ será elegido mediante la penalización dependiente de los datos y en el otro será elegido mediante la penalización independiente de los datos.

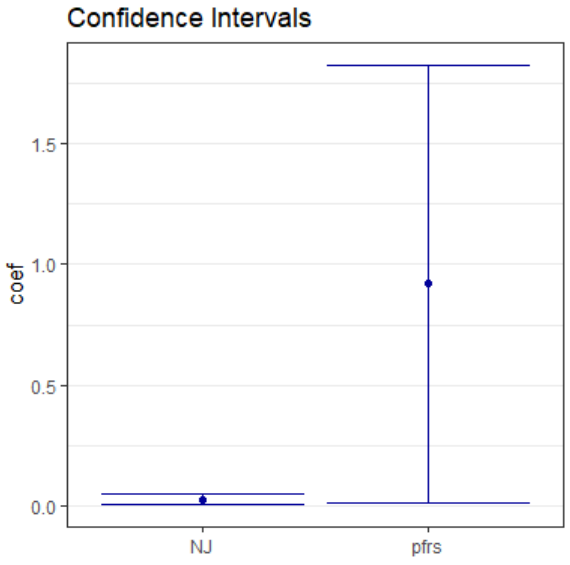
En ambos modelos hemos obtenido el mismo resultado, que es considerar la siguiente regresión:

$$psoda = 0{,}51919 + 0{,}54127 pfries + 0{,}031770 NJ$$

El error obtenido con el conjunto de prueba es de 0,06904211.

p) Como en el apartado anterior hemos obtenido los mismos modelos, entonces solo tenemos que plantear un solo modelo en este apartado. Podemos observar que el p-valor obtenido para *pfries* es 0,0469 y para *NJ* es de 0,0192, por lo que ambos son significativos a un 5 %.

El intervalo de confianza obtenido para *pfries* es $(-0{,}02492851, 1{,}86281711)$ y para *NJ* es $(0{,}00360638, 0{,}04917085)$. Veamoslo gráficamente.



q) En este apartado realizaremos una tabla con todos los modelos que hemos hecho en esto últimos apartados y elegiremos el mejor.

Selección de modelos	
Modelo	Error de prueba
Ridge CV 10	0.06518127
Ridge CV 5	0.06524422
LASSO CV 10	0.06841878
LASSO CV 5	0.06841878
Componentes Principales CV 10 (método codo)	0.0640881
Componentes Principales CV 5 (método codo)	0.0640881
PLS CV 10	0.06495082
PLS CV 5	0.06384467
PLS randomization	0.06681841
LASSO Red Elástica CV 10	0.06553641
LASSO Red Elástica CV 5	0.06750068
RLASSO penalización dependiente	0.06904211
RLASSO penalización independiente	0.06904211

Cuadro 2: Selección de modelos

Podemos observar que el modelo que menor error comete es el de Mínimos cuadrados parciales con validación cruzada 5 veces. Podemos observar que no hay mucha diferencia entre los diferentes modelos.

El modelo elegido en este apartado es mejor que el elegido en el apartado **f)**, ya que el mejor modelo del apartado **f)** tenía un error de 0,06432277.

r) En el modelo de Mínimos cuadrados parciales con validación cruzada 5 veces, seleccionábamos 3 componentes principales que son las siguientes:

Variables	Componentes PLS		
	Componente 1	Componente 2	Componente 3
pfries	0.0193228364	0.0226364274	0.0245113201
pentree	-0.0035258215	0.0016080442	0.0030130199
wagest	0.0004716088	-0.0004538095	-0.0001519289
nmgrs	-0.0019018551	-0.0048859627	-0.0056212057
nregs	0.0104423693	0.0115580218	0.0116664363
hrsopen	0.0084894794	0.0057469208	0.0064386459
emp	0.0008507683	-0.0030539476	-0.0029794142
compown	0.0014120345	0.0012149057	0.0009483646
density	0.0040551290	0.0014663455	-0.0015633886
crmrte	0.0052879543	0.0076061170	0.0062039954
prpblck	0.0049973165	0.0069620394	0.0056314565
nstores	0.0003501430	0.0020904303	0.0018545894
income	0.0029818374	0.0016801611	0.0013656471
lpfries	0.0192656105	0.0223697392	0.0240251798
lincome	0.0026253634	0.0016583980	0.0019836851
ldensity	0.0043180700	0.0016652264	-0.0013058669
NJ	0.0103938511	0.0144908323	0.0150178331
BK	0.0024505381	0.0062738508	0.0133279653
KFC	-0.0064888188	-0.0026205479	-0.0027563854
RR	0.0106868222	0.0081672045	0.0061578891

Cuadro 3: Componentes de PLS con CV 5

Podemos observar que todas las variables influyen poco en cada componente del PLS y casi todas lo hacen de forma positiva, menos por ejemplo las variables *KFC* y *nmgrs* que tienen influencia negativa en las componentes.