Trabajo Final Econometría

Sheila Fernández Martín

Ejercicio 1.

El conjunto de datos 'card' proviene del paquete 'wooldridge', que incluye datos para análisis econométricos. El paquete 'wooldridge' se basa en ejemplos del libro 'Introductory Econometrics: A Modern Approach de Jeffrey M. Wooldridge'. Específicamente, 'card' incluye datos utilizados en estudios sobre el impacto de la educación en los ingresos.

El conjunto de datos cuenta con 3010 observaciones de 34 variables pero dado que hay algunas muestras que no nos ofrece información para todas las variables, consideraremos el conjunto de datos con 1600 observaciones de 34 variables.

Observamos la relación entre las variables y eliminamos las variables redundantes 'educ', 'age', 'reg661', 'reg662', 'reg663', 'reg664', 'reg665', 'reg666', 'reg667', 'reg668'. También eliminamos la variable 'wage' para evitar problemas de multicolinealidad y permitir una estimación más precisa de los coeficientes. El modelo resultante es:

```
## lwage ~ id + nearc2 + nearc4 + fatheduc + motheduc + weight +
##     momdad14 + sinmom14 + step14 + reg669 + south66 + black +
##     smsa + south + smsa66 + enroll + KWW + IQ + married + libcrd14 +
##     exper + expersq
```

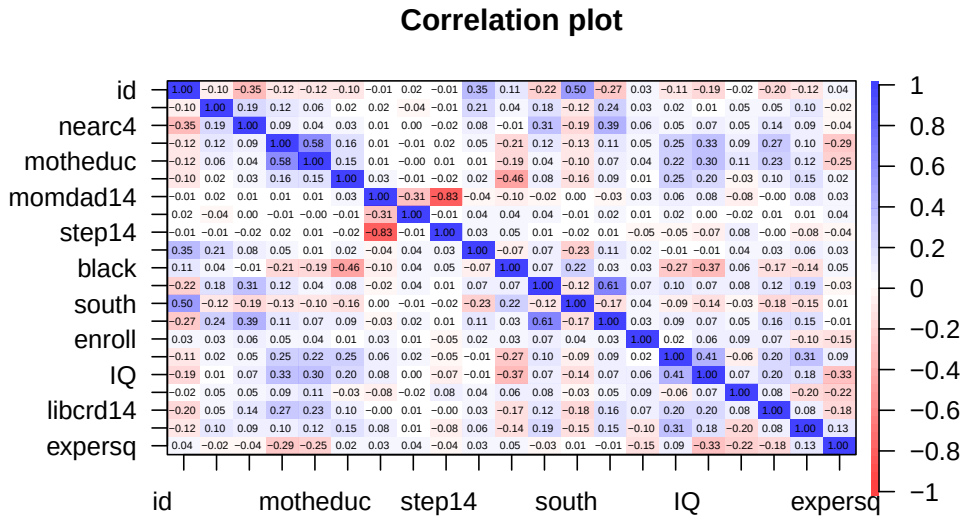
Utilizaremos dicho modelo para asegurar que sea matemáticamente sólido, estadísticamente interpretable y computacionalmente eficiente.

Eliminamos la variable que esté más correlacionada con las demás variables explicativas si el FIV es superior a 5 y volvemos a repetir el análisis hasta que no queden más variables explicativas con un FIV superior a 5.

De este modo, en primer lugar, eliminamos la variable 'exper', a continuación volvemos a analizar el FIV y finalmente, eliminamos la variable 'south66'.

Observamos que ya no hay ninguna variable con un VIF superior a 5 por lo que hacemos el diagrama de calor (Heat map) eliminando del modelo original las variables: 'educ', 'age', 'reg661', 'reg662', 'reg663', 'reg664', 'reg665', 'reg666', 'reg667', 'reg668', 'wage', 'exper' y 'south66'.

Por tanto, las variables que usaremos en adelante son: 'id', 'nearc2', 'nearc4', 'fatheduc', 'motheduc', 'weight', 'momdad14', 'sinmom14', 'step14', reg669', 'black', 'smsa', 'south', 'smsa66', 'enroll', Kww', 'IQ', 'married', 'libcrd14' y 'expersq'.



Ejercicio 2.

Ajustamos un modelo de mínimos cuadrados ordinarios con las variables seleccionadas en el apartado anterior:

```
##
## Call:
## lm(formula = lwage ~ ., data = Card)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48795 -0.21954  0.01948  0.24303  1.48125
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.494e+00  1.393e-01  39.450  < 2e-16 ***
## id          -1.309e-06  1.005e-05  -0.130 0.896349
## nearc2       4.805e-02  2.022e-02   2.376 0.017614 *
## nearc4       8.830e-03  2.398e-02   0.368 0.712806
## fatheduc    -3.120e-03  3.603e-03  -0.866 0.386746
## motheduc     1.226e-02  4.172e-03   2.938 0.003346 **
## weight       1.065e-07  6.940e-08   1.535 0.125057
## momdad14    -3.999e-02  1.016e-01  -0.394 0.693873
## sinmom14    -4.545e-02  1.750e-01  -0.260 0.795130
## step14      -1.815e-01  1.150e-01  -1.579 0.114603
## reg669       4.706e-02  4.031e-02   1.168 0.243184
## black       -1.234e-02  3.889e-02  -0.317 0.751113
## smsa         1.483e-01  2.800e-02   5.295 1.36e-07 ***
## south       -6.242e-02  2.728e-02  -2.288 0.022288 *
## smsa66       5.597e-03  2.735e-02   0.205 0.837881
```

```
## enroll      -1.298e-01  3.105e-02  -4.181 3.06e-05 ***
## KWW          1.126e-02  1.515e-03   7.429 1.79e-13 ***
## IQ           2.553e-03  7.967e-04   3.205 0.001379 **
## married     -3.820e-02  4.825e-03  -7.917 4.53e-15 ***
## libcrd14     9.021e-03  2.467e-02   0.366 0.714653
## expersq      6.556e-04  1.706e-04   3.844 0.000126 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3759 on 1579 degrees of freedom
## Multiple R-squared:  0.2055, Adjusted R-squared:  0.1954
## F-statistic: 20.42 on 20 and 1579 DF,  p-value: < 2.2e-16
```

Observamos que el intercepto es 5.494, que hace referencia al valor del logaritmo del salario esperado si todas las variables explicativas son cero.

Los valores de la columna 'Estimate' determinan el cambio esperado en la variable objetivo (lwage) asociado a un cambio unitario en la variable explicativa correspondiente, manteniendo las demás constantes. Por ejemplo, por cada unidad que aumente 'weight', el salario aumenta un 1.065e-05%.

Para evaluar si los coeficientes son significativos, analizamos la columna 'Pr(>|t|)'. Observamos así, que las variables 'motheduc', 'smsa', 'enroll', 'KWW', 'IQ', 'married' y 'expersq' son estadísticamente significativas a un nivel de signigicación del 1% y que todas las demás no lo son porque tienen un p-valor superior a 0.01.

Con el valor de 'Multiple R-squared' determinamos que el modelo explica el 20.55% de la variabilidad del logaritmo de los salarios, lo cual es una proporción muy baja indicando que el modelo no es bueno.

El p-valor global del modelo es 2.2e-16 < 0.01, luego es estadísicamente significativo globalmente a un nivel de significación de α=1%.

Ejercicio 3.

Para realizar el contraste de significatividad conjunta, primero analizamos el modelo:

```
##
## Call:
## lm(formula = lwage ~ ., data = Card1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44830 -0.21643  0.02075  0.23799  1.54958
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.293e+00  1.505e-01  35.171  < 2e-16 ***
## id          -7.877e-06  1.190e-05  -0.662 0.507927
## nearc2       4.753e-02  2.024e-02   2.348 0.019000 *
## nearc4       3.724e-03  2.411e-02   0.154 0.877266
## fatheduc    -2.359e-03  3.595e-03  -0.656 0.511799
## motheduc     1.250e-02  4.156e-03   3.008 0.002670 **
## weight       7.557e-08  6.966e-08   1.085 0.278183
## momdad14    -3.325e-02  1.012e-01  -0.328 0.742640
## sinmom14    -3.115e-02  1.744e-01  -0.179 0.858257
## step14      -1.710e-01  1.146e-01  -1.492 0.135824
```

```
## reg669       6.448e-02  4.447e-02   1.450 0.147246
## south66      4.605e-02  4.514e-02   1.020 0.307831
## black       -2.065e-02  3.905e-02  -0.529 0.597031
## smsa         1.492e-01  2.789e-02   5.350 1.01e-07 ***
## south       -8.510e-02  3.691e-02  -2.305 0.021276 *
## smsa66       6.204e-03  2.727e-02   0.228 0.820049
## enroll      -1.238e-01  3.097e-02  -3.999 6.67e-05 ***
## KWW          1.053e-02  1.522e-03   6.916 6.73e-12 ***
## IQ           3.130e-03  8.089e-04   3.870 0.000113 ***
## married     -3.562e-02  4.860e-03  -7.329 3.68e-13 ***
## libcrd14     1.138e-02  2.459e-02   0.463 0.643623
## exper        4.019e-02  1.098e-02   3.659 0.000261 ***
## expersq     -1.413e-03  5.901e-04  -2.395 0.016743 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3744 on 1577 degrees of freedom
## Multiple R-squared:  0.2126, Adjusted R-squared:  0.2017
## F-statistic: 19.36 on 22 and 1577 DF,  p-value: < 2.2e-16
```

Comparando este resultado con el anterior, observamos que los coeficientes estimados de las variables son muy similares en magnitud y significancia. Los dos modelos tienen un R cuadrado y Adjusted R cuadrado muy similares luego las diferencias no son estadísticamente relevantes. Finalmente, comentar que ambos tienen un valor de F muy bajo con un p-value cercano a 0, indicando que son estadísticamente significativos en general. Sin embargo, el modelo del ejercicio anterior cuenta con 20 predictores y F-statistic = 20.42 mientras que, este cuenta con 22 predictores y F-statistic = 19.36, lo que sugiere que añadir más variables (como 'south66' y 'exper') puede estar penalizando el modelo en términos de ajuste.

A continuación, miramos cuales son las variables que no son individualmente significativas al 1%.

```
##  [1] "id"       "nearc2"   "nearc4"   "fatheduc" "weight"   "momdad14"
##  [7] "sinmom14" "step14"   "reg669"   "south66"  "black"    "south"
## [13] "smsa66"   "libcrd14" "expersq"
```

Realizamos el contraste de significatividad conjunta de dichas variables considerando como hipótesis nula que sus coeficientes son 0.

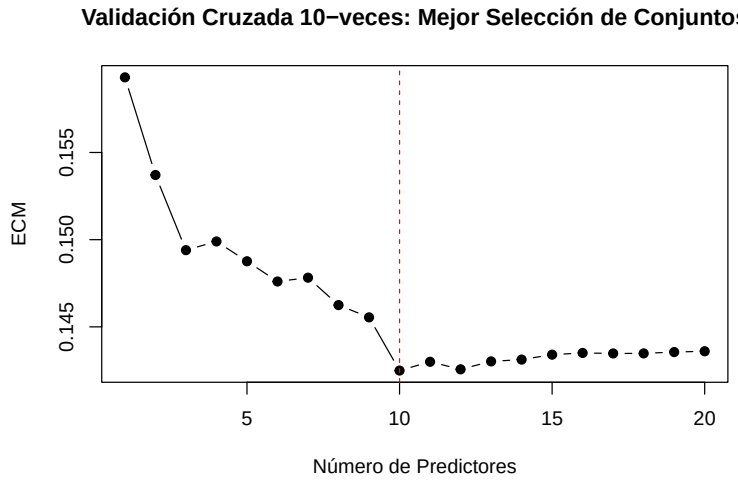
Como el p-valor es $0.0002497 < 0.01$, rechazamos la hipótesis nula indicando que las variables no significativas individualmente al 1% tienen un efecto conjunto estadísticamente significativo en el modelo.

Ejercicio 4.

Ajustamos un modelo de mínimos cuadrados ordinarios en el conjunto de entrenamiento y calculamos su error de prueba. Dicho error es una métrica que mide el desempeño de un modelo ajustado al conjunto de entrenamiento al realizar predicciones sobre el conjunto de prueba. En este caso, calculamos el Error Cuadrático Medio (ECM) como la métrica de error de prueba y su valor es: 0.1383376. Luego el promedio de los errores es 0.3719376, que es un valor bastante alto indicando que el modelo no tiene un buen ajuste en los datos de prueba.

Ejercicio 5.

Vamos a hacer la Validación Cruzada 10-Veces utilizando la Mejor Selección de Conjuntos.



Vemos que selecciona un modelo de 10 variables. Ahora realizamos mejor selección de subconjuntos en el conjunto de datos con el fin de obtener el modelo de 10 variables y observamos que es el siguiente:

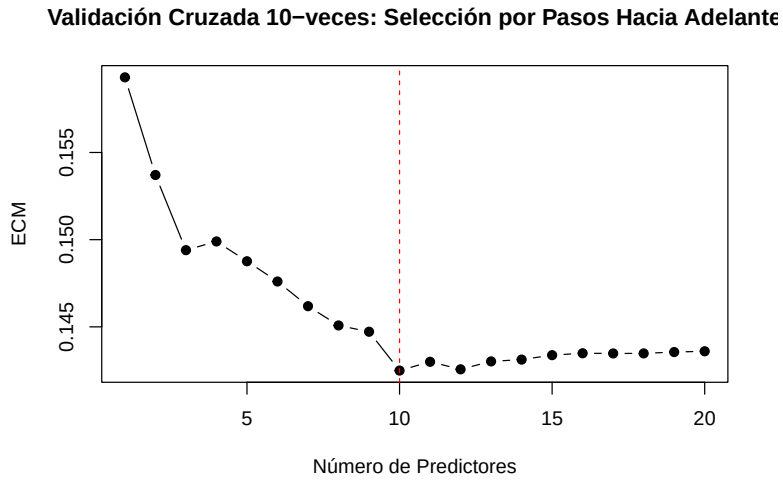
```
##   (Intercept)        nearc2      motheduc        step14          smsa
##  5.4647337490  0.0669184803  0.0115368063 -0.1961587288  0.1445142061
##         south        enroll           KWW            IQ       married
## -0.0733510399 -0.1231310148  0.0116630460  0.0025869320 -0.0374020467
##       expersq
##  0.0006105716
```

Finalmente, el ECM es 0.1385224, por lo que la raíz es: 0.372186, que es un valor bastante alto indicando que el modelo no tiene un buen ajuste en los datos de prueba.

El modelo seleccionado por la regla del codo incluye 8 variables con un ECM de prueba de 0.1422681.

Ejercicio 6.

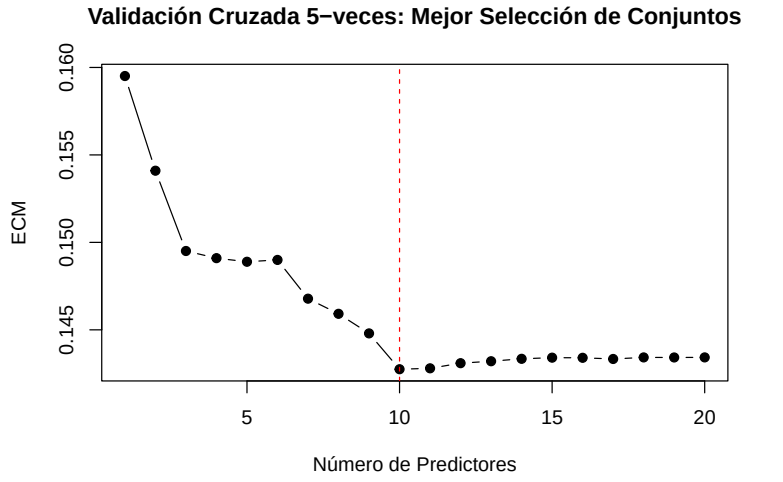
Vamos a hacer la Validación Cruzada 10-Veces utilizando la Selección por Pasos Hacia Adelante.



Vemos que de nuevo selecciona un modelo de 10 variables por lo que el modelo de 10 variables es el mismo que en el apartado anterior y el valor de sus errores también. Sin embargo, al aplicar la regla del codo, utiliza 7 variables en vez de 8.

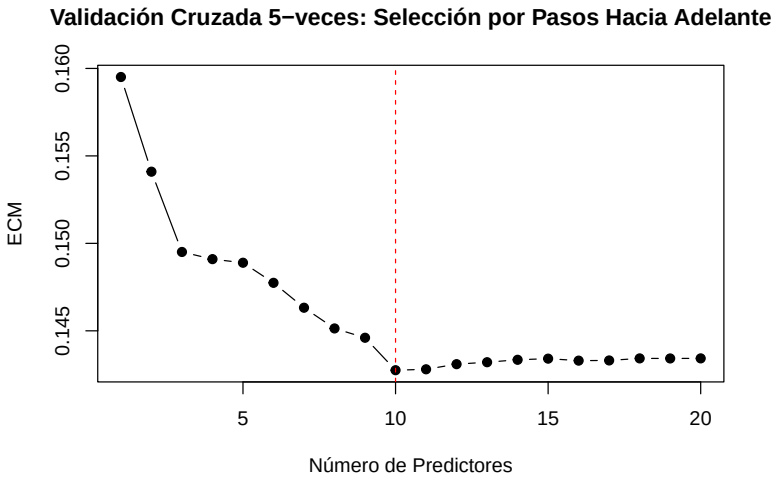
Ejercicio 7.

VC-5: Mejor Selección de Conjuntos



Vemos que de nuevo selecciona un modelo de 10 variables por lo que el modelo de 10 variables es el mismo que en el apartado anterior y el valor de sus errores también.

VC-5: Selección por Pasos Hacia Adelante



También selecciona un modelo de 10 variables por lo que el modelo de 10 variables es el mismo que en los apartados anteriores y el valor de sus errores también.

Ejercicio 8.

Modelo	ECM	Regla del Codo
VC 10-veces: Mejor Selección de Conjuntos	0.1385224	0.142286
VC 10-veces: Selección Hacia Adelante	0.1385224	0.142286
VC 5-veces: Mejor Selección de Conjuntos	0.1385224	0.142286
VC 5-veces: Selección Hacia Adelante	0.1385224	0.142286

Table 1: Tabla de los errores de prueba de cada modelo

No hay diferencias en los errores de prueba entre los distintos enfoques. Esto sugiere que, en términos prácticos, no hay una diferencia sustancial en el rendimiento de los modelos para los diferentes enfoques de selección de variables y números de pliegues en la validación cruzada.

Ejercicio 9.

Escogeremos el modelo de la Validación Cruzada 5-veces utilizando Selección Hacia Adelante para reducir el tiempo de cómputo.

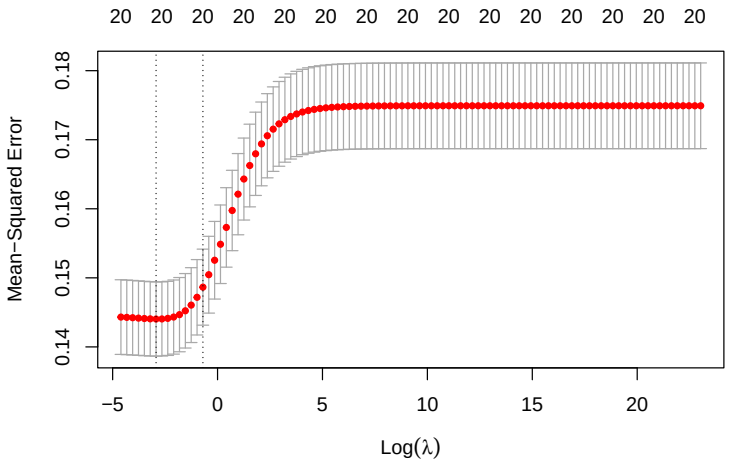
Ajustamos el modelo y obtenemos el p-valor de los coeficientes.

```
##  (Intercept)       nearc2     motheduc       step14         smsa        south
## 0.000000e+00 5.974848e-03 1.907021e-03 1.191386e-02 5.434919e-12 1.423412e-04
##       enroll          KWW           IQ      married      expersq
## 3.969315e-05 9.452461e-15 3.901697e-04 3.030302e-15 2.638403e-05
```

Comparamos la columna 'Pr(>|t|)' con el nivel de significación de $\alpha = 5\%$. Observamos así, que todas las variables son estadísticamente significativas a este nivel de significación pues tienen un p-valor inferior a 0.05.

Ejercicio 10.

Regresión Ridge: VC-10



Mejor lambda (mínimo error): 0.05336699.

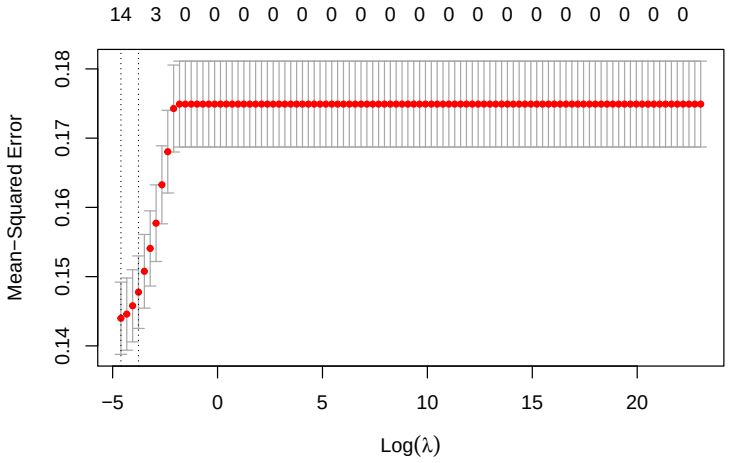
Error de prueba (MSE) con mejor lambda: 0.1387658.

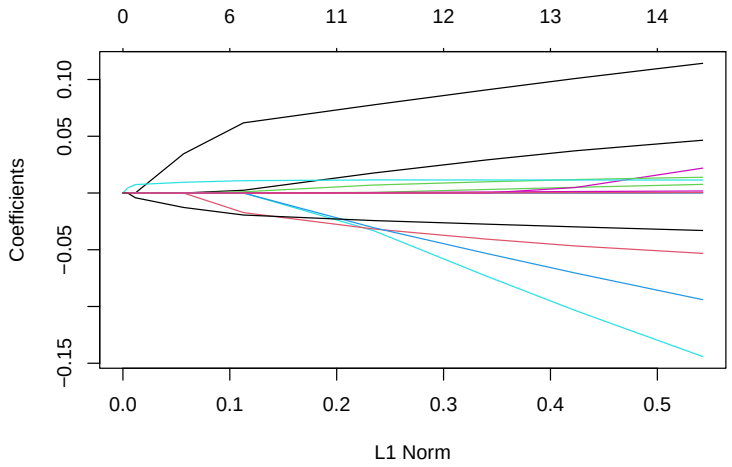
Lambda con la regla del codo: 0.4977024.

Error de prueba (MSE) con lambda con la regla del codo: 0.1469486.

Ejercicio 11.

Modelo LASSO: VC-10





Mejor lambda (mínimo error): 0.01.

Error de prueba (MSE) con mejor lambda: 0.1389268.

Lambda con la regla del codo: 0.0231013.

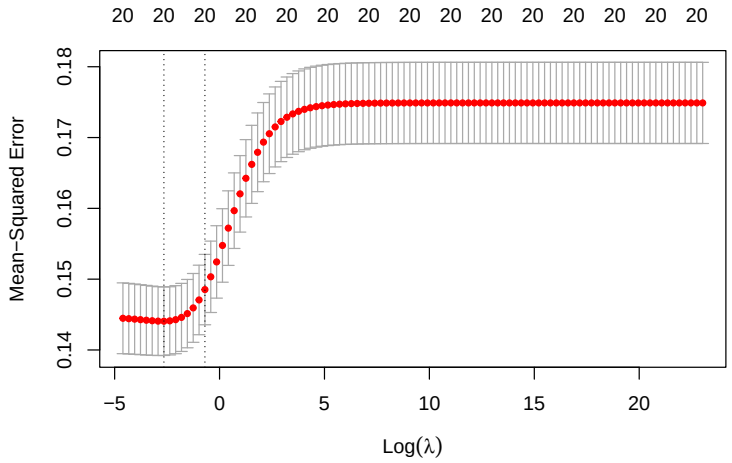
Error de prueba (MSE) con lambda del codo: 0.1433635.

Número de coeficientes diferentes de cero: 14.

Número de coeficientes diferentes de cero con lambda por la regla del codo: 12.

Ejercicio 12.

Regresión Ridge: VC-5



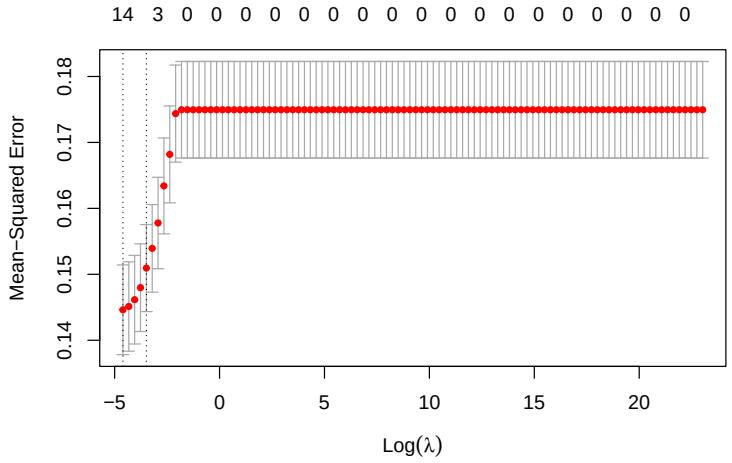
Mejor lambda (mínimo error), VC-5: 0.07054802.

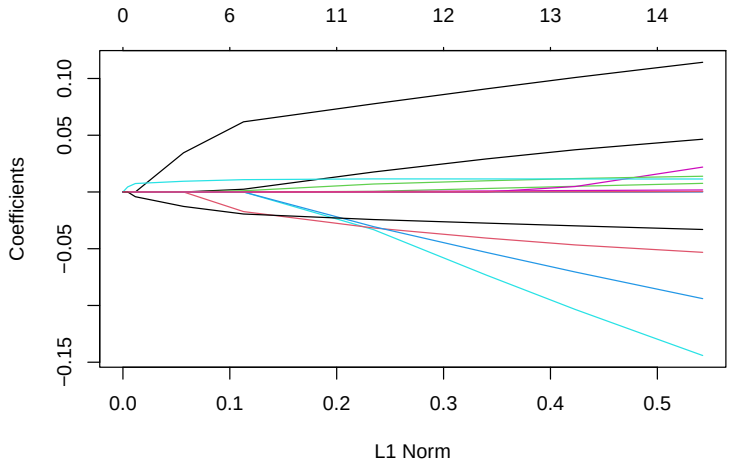
Error de prueba (MSE) con mejor lambda, VC-5: 0.1390268.

Lambda con la regla del codo, VC-5: 0.4977024.

Error de prueba (MSE) con lambda con la regla del codo, VC-5: 0.1469486.

Modelo LASSO: VC-5





Mejor lambda (mínimo error): 0.01.

Error de prueba (MSE) con mejor lambda: 0.1389268.

Lambda con la regla del codo: 0.03053856.

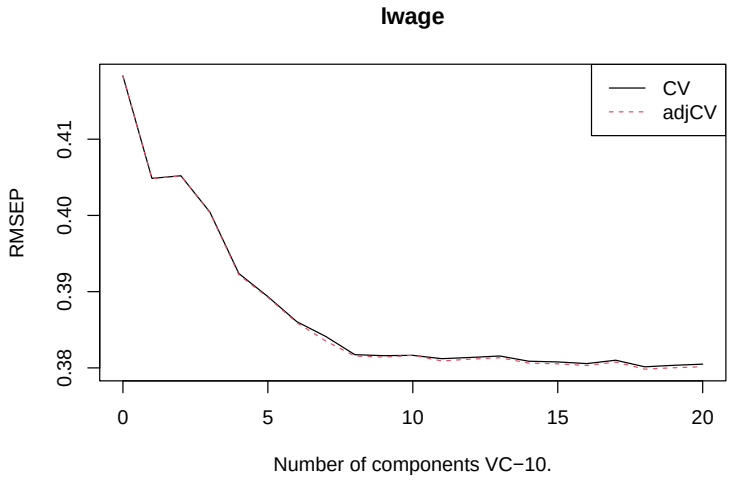
Error de prueba (MSE) con lambda del codo: 0.1474282.

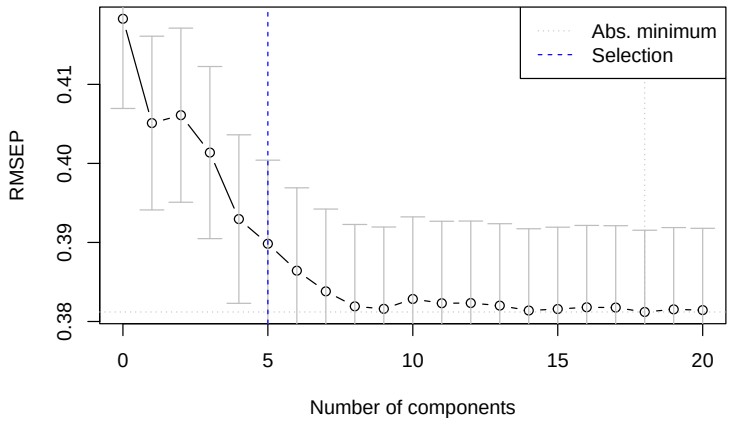
Número de coeficientes diferentes de cero: 14.

Número de coeficientes diferentes de cero con lambda por la regla del codo: 11.

Ejercicio 13.

Componentes principales: VC-10



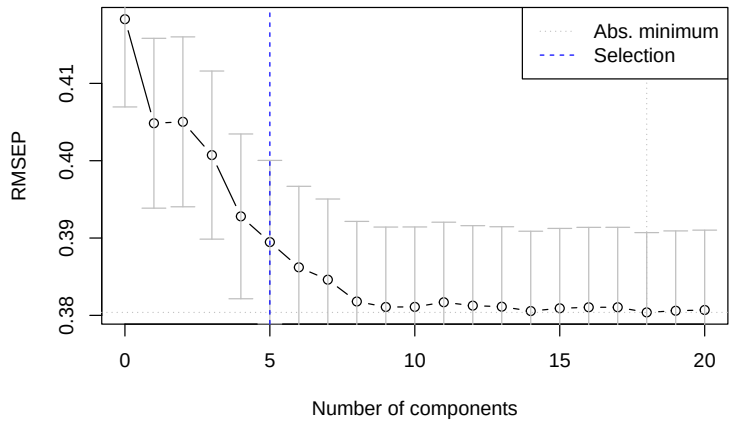


Número óptimo de componentes principales (VC-10): 5.

Error de prueba (VC-10): 0.1496024.

Componentes principales: VC-5

Obtenemos los mismos resultados que al usar 10 pliegues en vez de 5.

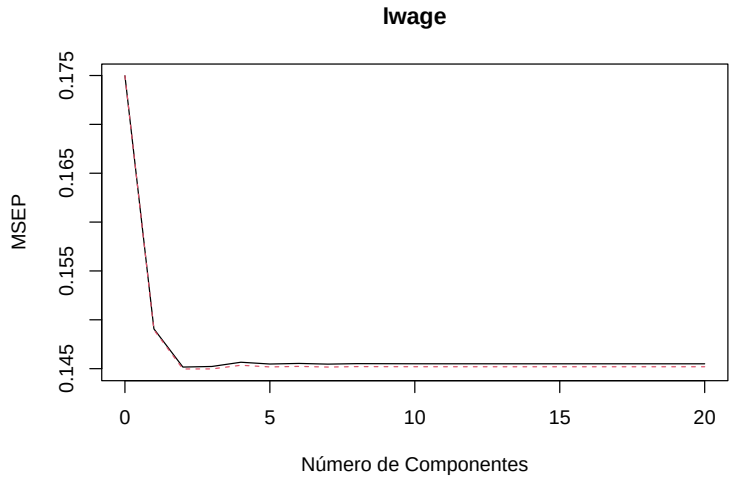


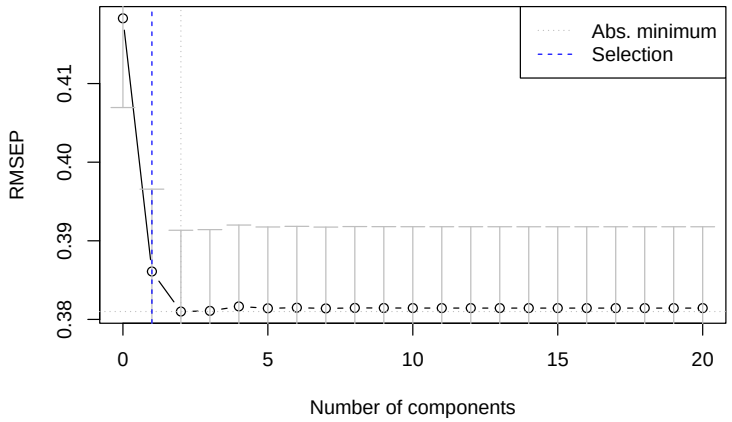
Número óptimo de componentes principales (VC-5): 5.

Error de prueba (VC-5): 0.1496024.

Ejercicio 14.

PLS: VC-10



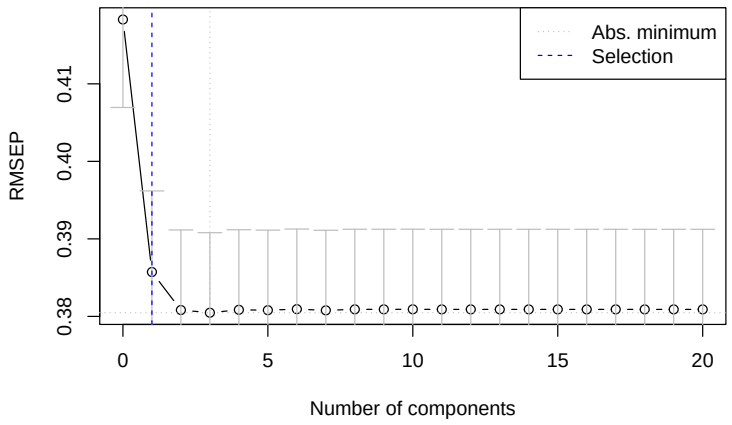


Número óptimo de componentes principales (VC-10): 1.

Error de prueba (VC-10): 0.1477611.

PLS: VC-5

De nuevo obtenemos los mismos resultados que al usar 10 pliegues en vez de 5.



Número óptimo de componentes principales (VC-5): 1.

Error de prueba (VC-5): 0.1477611.

Ejercicio 15.

Modelo	**ECM**
Ridge (VC-10)	0.1387658
Ridge (Regla del codo VC-10)	0.4977024
LASSO (VC-10)	0.1389268
LASSO (Regla del codo VC-10)	0.1433635
Ridge (VC-5)	0.1390268
Ridge (Regla del codo VC-5)	0.1469486
LASSO (VC-5)	0.1389268
LASSO (Regla del codo VC-5)	0.1474282
PCR (VC-10)	0.1496024
PCR (VC-5)	0.1496024
PLS (VC-10)	0.1477611
PLS (VC-5)	0.1477611

Table 2: Tabla de los errores de prueba de cada modelo

En general, los errores de prueba resultantes entre los enfoques son bastante similares, siendo el modelo el de Ridge mediante validación cruzada con 10 pliegues. Por otro lado, el modelo seleccionado en el apartado 9 cuenta con un error igual a 0.1385224 mientras que el error del modelo PCR tiene un error igual a 0.1387658 por lo que, se puede concluir que cuentan con una mayor precisión los modelos de Validación Cruzada 5-veces o 10-veces utilizando tanto Selección Hacia Adelante como Mejor Selección de Conjunto.