Trabajo Final

Reinier Mujica

13 de enero de 2019

### Inicialización

Limpiamos todas las variables del espacio de trabajo y el terminal de R.

rm(list=ls())  
cat("\014")

library(knitr)

Cargamos el paquete de wooldridge y los datos

library(wooldridge)  
attach(bwght2)

Fijamos la semilla

set.seed(100)

### Limpiar los datos

Vamos a remover los valores NA de los datos

datos = na.omit(bwght2)

### División del conjunto de datos

train.size = round(dim(datos)[1] \* 0.7)  
train = sample(1:dim(datos)[1], train.size)  
test = -train  
datos.train = datos[train, ]  
datos.test = datos[test, ]  
test.size = dim(datos.test)[1]

### Preguntas:

#### a)

El conjunto de datos es bwght2 y contiene información sobre nacimientos. Estos datos han sido obtenidos de certificados denacimiento y desfunción y también de información de natalidad y mortalidad del Centro Nacional de Estadística para la Salud. El dataset contiene varias variables con diversa información como: edad de los padres, educación de los padres, número de visitas prenatales, peso al nacer, e información acerca de la cantidad de cigarillos y bebida que consumia la madre durante el embarazo. También se incluye información del feto.

Los nombres de las variables que contiene el dataset son:

[1] "mage" "meduc" "monpre" "npvis" "fage" "feduc" "bwght"   
 [8] "omaps" "fmaps" "cigs" "drink" "lbw" "vlbw" "male"   
[15] "mwhte" "mblck" "moth" "fwhte" "fblck" "foth" "lbwght"   
[22] "magesq" "npvissq"

La variable dependiente que va ha ser explicada es **lbwght**, que es el logaritmo de la variable **bwght** que es el peso al nacer.

Vamos a ajustar un modelo de mínimos cuadrados ordinarios (con todas las variables explicativas) en el conjunto de entrenamiento e imprimir el error de prueba obtenido.

El error de prueba del MCO nos queda 5.187389710^{-4}.

Los coeficientes B los podemos ver a continuación.

(Intercept) mage meduc monpre npvis   
 7.032634e+00 -1.245338e-03 -4.244261e-04 5.920562e-04 1.056039e-04   
 fage feduc bwght omaps fmaps   
 2.315328e-04 -4.885967e-06 2.978545e-04 9.290155e-04 9.733477e-03   
 cigs drink lbw vlbw male   
 3.931007e-04 -6.183647e-04 -1.540387e-01 -3.316268e-01 -7.498069e-04   
 mwhte mblck moth fwhte fblck   
 3.538706e-03 2.381276e-03 NA -5.610199e-03 2.087637e-03   
 foth magesq npvissq   
 NA 1.585596e-05 -6.467952e-06

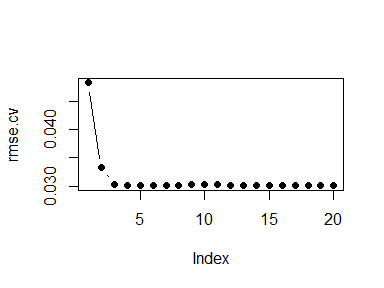
Podemos concluir que dos coeficientes fueron omitidos por estar las respectivas variables muy correlacionadas(**moth** y **foth**). También como se puede ver en el resultado de la instrucción **summary** solo las variables:

**bwght** **fmaps**  
**lbw**  
**vlbw**

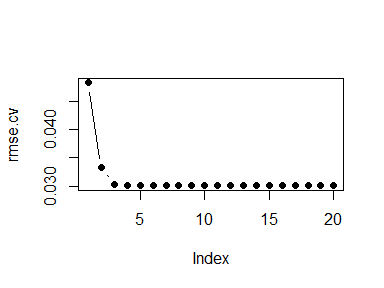
Son significativas para predecir **lbwght**.

#### b)

Procedemos a eliminar del Dataset las dos variables correlacionadas **moth** y **foth**

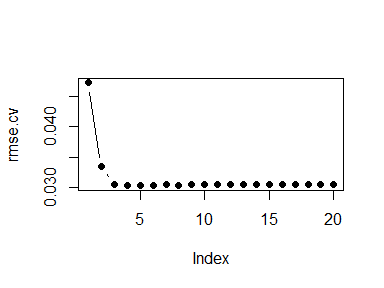
Luego ajustamos un MCO utilizando la Mejor Selección de Conjuntos  El error de prueba obtenido es mas pequeño que con el MCO ordinario, este es 5.119140210^{-4}.

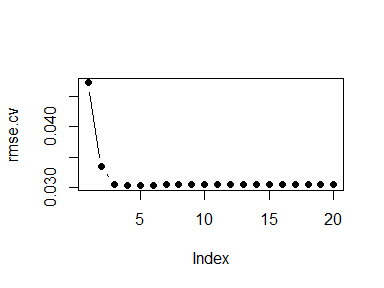
#### c)

Ahora usaremos la Selección por Pasos Hacia Adelante  El error del MCO con Forward step wise es 5.119140210^{-4}.

#### d)

Ahora repetiremos los apartados **b** y **c**, pero utilizando la Validación Cruzada 5-Veces.

El gráfico de 5-CV del **MCO with Subset Selection** nos queda asi: 

El gráfico de 5-CV del **MCO with Forward step wise** nos queda asi: 

[1] 0.000511914

#### e)

A continuación se muestran los modelos utilizados anteriormente con el tipo de validación cruzada y su error de prueba correspondiente.

|  |  |  |  |
| --- | --- | --- | --- |
|  | None | 5 Cross Validation | 10 Cross Validation |
| MCO | 0.0005187 | NA | NA |
| MCO with Subset Selection | NA | 0.0005119 | 0.0005119 |
| MCO with Forward step wise | NA | 0.0005119 | 0.0005119 |

Como se puede observar en la tabla no hay mucha diferencia entre los errores de prueba obtenidos de estos enfoques.

#### f)

Para seleccionar el modelo que minimiza el error, como no hay mucha diferencia entre los errores anteriores, escogimos el último modelo analizado **MCO with Forward step wise**. En el análisis de este modelo se determinaron las variables significativas al 5%, son las que se usan a continuación.

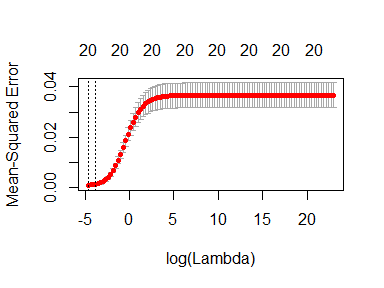
El modelo con sólo cuatro variables explicativas con un nivel de significación del 5% (**bwght, fmaps, lbw, vlbw**) y un error de 5.187389710^{-4}, es un poco mas preciso respecto al error del MCO: 5.187389710^{-4}.

Los **p-valores** se muestran a continuación:

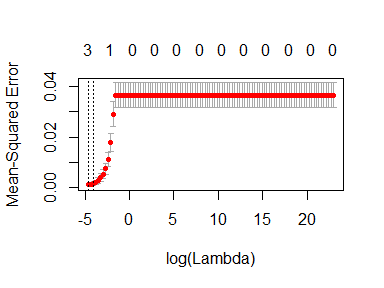
$coefficients  
 Estimate Std. Error t value Pr(>|t|)  
(Intercept) 7.0123631237 1.703342e-02 411.682667 0.000000e+00  
bwght 0.0002976461 1.585174e-06 187.768796 0.000000e+00  
fmaps 0.0105802459 1.834398e-03 5.767694 1.038118e-08  
lbw -0.1543394286 9.001872e-03 -17.145259 1.041566e-58  
vlbw -0.3309693951 1.372829e-02 -24.108564 7.864423e-104

#### g)

Regresión de Ridge

 El error de prueba obtenido en la regresión de Ridge es 5.855055810^{-4}.

#### h)

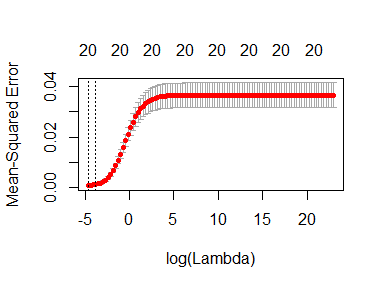
Modelo de LASSO  El error de prueba obtenido en el modelo de LASSO es 4.357229910^{-4}.

Para obtener los coeficientes del modelo para el mejor lambda usamos la función coef

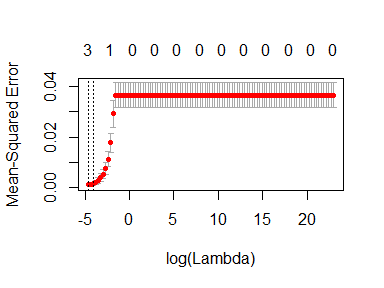
Los coeficientes diferentes de 0 son 3.

#### i)

Regresión de Ridge con CV 5 veces

 El error de Ridge con CV 5 veces nos queda 5.855055810^{-4}.

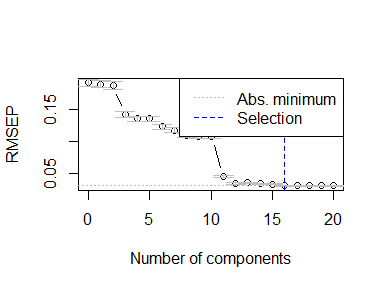
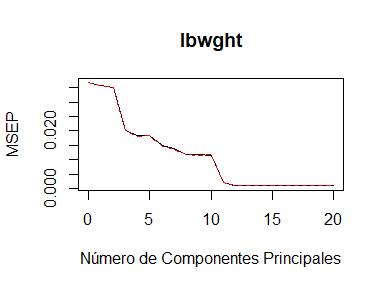
Modelo de LASSO con CV 5 veces

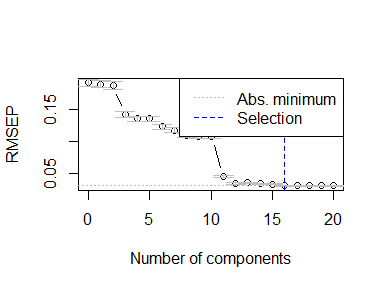
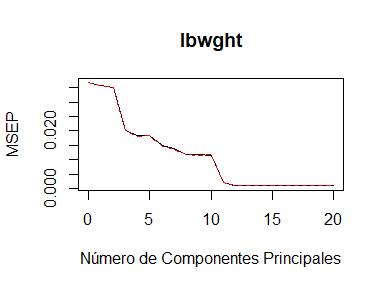


El error de LASSO con CV 5 veces nos queda 4.357229910^{-4}.

Para obtener los coeficientes del modelo para el mejor lambda usamos la función coef

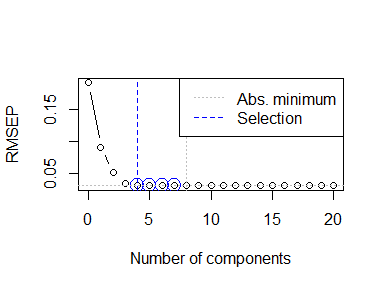
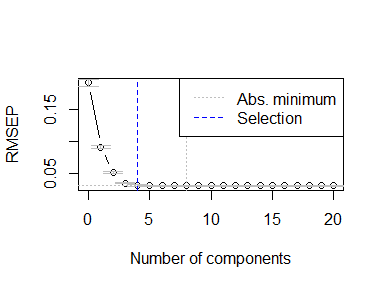
#### j)

Modelo de Componentes Principales en el conjunto de entrenamiento  Usando el Metodo de Componentes Principales con una Validación Cruzada 10 veces obtenemos la cantidad de componentes M = 16 y un error de prueba de 5.175971410^{-4}.

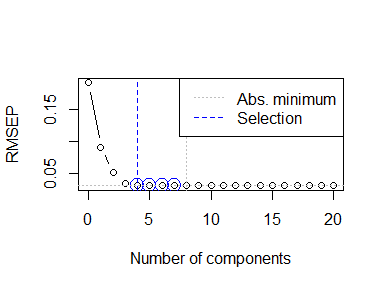
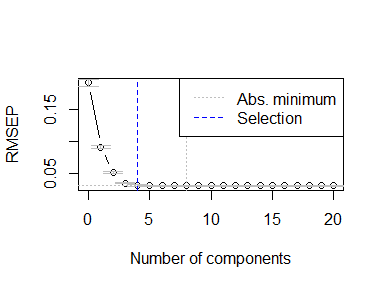
El mismo método pero con la validación cruzada 5 veces seria:  Usando el Metodo de Componentes Principales con una Validación Cruzada 5 veces obtenemos la cantidad de componentes M = 16 y un error de prueba de 5.175971410^{-4}.

#### k)

PLS con Validación cruzada 10 veces

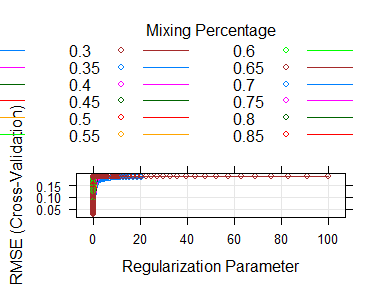


PLS con Validación cruzada 5 veces



#### l)

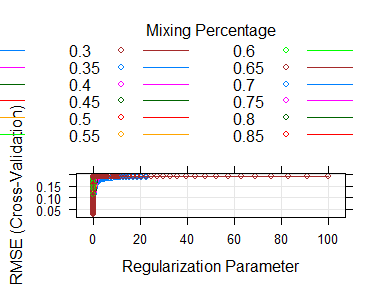
Modelo LASSO con la restricción de Red Elástica (LASSO with Elastic Net) en el conjunto de entrenamiento, con el α y el λ elegido mediante la Validación Cruzada 10-Veces.



El error de prueba obtenido es 4.796094410^{-4} y la cantidad de coeficientes distintos de 0 es 4.

#### m)

Modelo LASSO con la restricción de Red Elástica (LASSO with Elastic Net) en el conjunto de entrenamiento, con el α y el λ elegido mediante la Validación Cruzada 5-Veces.



21 x 1 sparse Matrix of class "dgCMatrix"  
 1  
(Intercept) 7.0950292590  
mage .   
meduc .   
monpre .   
npvis .   
fage .   
feduc .   
bwght 0.0002826996  
omaps .   
fmaps 0.0070610993  
cigs .   
drink .   
lbw -0.1609479918  
vlbw -0.3115200778  
male .   
mwhte .   
mblck .   
fwhte .   
fblck .   
magesq .   
npvissq .

[1] 0.0004959331

El error de prueba obtenido es 4.959330910^{-4} y la cantidad de coeficientes distintos de 0 es 4.

#### o)

Ajusta un modelo Rigorous LASSO (RLASSO) sobre el conjunto de entrenamiento con el lambda elegido mediante la penalización dependiente de los datos

Informa el error de prueba obtenido junto con el número de coeficientes estimados diferentes de cero.

Total de variables = **20** Variables seleccionadas = **1** p-value = **0**

[1] 0.0009358474

El error de prueba de Rigorous Lasso con penalización dependiente de los datos es 9.358473810^{-4}

Con el lambda elegido mediante la penalización independiente de los datos

Informa el error de prueba obtenido junto con el número de coeficientes estimados diferentes de cero.

Total de variables = **20** Variables seleccionadas = **1** p-value = **0**

[1] 0.0009358474

El error de prueba de Rigorous Lasso con penalización independiente de los datos es 9.358473810^{-4}

Calcula el error de prueba de los dos modelos mediante el ajuste de Post-LASSO.

Informa el error de prueba obtenido junto con el número de coeficientes estimados diferentes de cero.

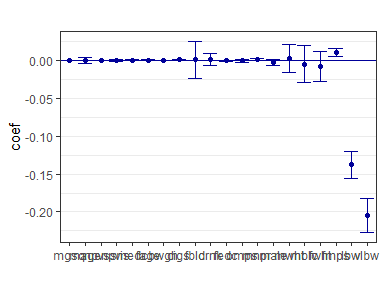
Total de variables = **20** Variables seleccionadas = **1** p-value = **0**

[1] 0.0005005301

El error de prueba Post-Lasso es 5.005300510^{-4}

#### p)

Contrastar la significación individual de los coeficientes estimados del modelo final. Utiliza el nivel de significación del 5%.



Sólo las variables \***bwght, fmaps, lbw y vlbw** son significativas sobre un nivel de confianza del 1%

Ahora eliminamos las variables no significativas

Las cuatro variables son muy significativas.

Informa el error de prueba obtenido

[1] 0.0007593627

El error de prueba obtenido es 7.593626810^{-4}

Basándose en contrastes de significación individual, selecciona modelos que contengan sólo variables significativas al 5% de significación y estima el error de prueba de estos modelos (mediante el ajuste de Post-LASSO).

[1] 0.000511914

El error de prueba de estos modelos es 5.119140210^{-4}

#### q)

A continuación se muestran los modelos utilizados anteriormente con el tipo de validación cruzada y su error de prueba correspondiente.

|  |  |  |  |
| --- | --- | --- | --- |
|  | None | 5 Cross Validation | 10 Cross Validation |
| RIDGE | NA | 0.0005855 | 0.0005855 |
| LASSO | NA | 0.0004357 | 0.0004357 |
| PCA | NA | 0.0005176 | 0.0005176 |
| PLS | NA | 0.0005496 | 0.0005496 |
| LASSO with Elastic Net | NA | 0.0004959 | 0.0004796 |
| Rigorous LASSO Independent L | 0.0009358 | NA | NA |
| Rigorous LASSO Dependent L | 0.0009358 | NA | NA |
| Rigorous LASSO Post-Lasso | 0.0005005 | NA | NA |

Como sucede en el inciso **f** no hay mucha diferencia en el error de prueba de los modelos analizados. Sin embargo el modelo **LASSO** presenta un menor error que los modelos analizados en el inciso **f**.