Universidad de los Andes

Mínería de Datos

Sección: Modelos Predictivos

Taller Tranformación y Selección de variables

Sergio A. Mora Pardo
Universidad de los Andes
s.morap@uniandes.edu.co
cod. 201920547
Bogotá D.C.

Cindy Zulima Alzate Román
Universidad de los Andes
c.alzate@uniandes.edu.co
cod. 201920019
Bogotá D.C.

Jahir Stevens Rodriguez Riveros
Universidad de los Andes
js.rodriguezr@uniandes.edu.co
cod. 201819361
Bogotá D.C.

supervised by
Dr. Carlos Valencia

March 9, 2020

Minería de Datos, Taller 2

Abstract

Solución al taller de selección y transformación de variables de Minería de Datos. Se utilizó LATEX y las siguientes librerías en R:

```
> library(pls)
> library(leaps)
> library(ISLR)
> library(MASS)
> library(tidyverse)
> library(xtable)
```


Contents

1 Datos

La base de datos "Boston" contenida en el paquete MASS contiene información para predecir el valor de las casas ("medv") dependiendo de ciertas caracter?ticas del vecindario. Para cargar los datos use:

```
> datos=Boston
> head(datos)
```

```
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
  medv
1 24.0
2 21.6
3 34.7
4 33.4
5 36.2
6 28.7
```

Para comparar los modelos, puede usar los últimos 100 datos como test y el resto como train.

```
> n = nrow(datos)
> train=datos[c(1:(n-100)),]
> test=datos[c((n-100):n),]
```

2 Primer punto

Usando la metodología de selección de variables exhaustiva y tipo "fordward", encuentre el mejor modelo lineal predictivo para predecir el valor medio de las casas ("medv") usando las demás variables como predictores. Explique los modelos resultantes.

2.1 Metodología Exahustiva

Se implementa la metodología exahustiva de selección de variables de la siguiente forma:

```
> reg_subset=regsubsets(medv~.,train,nvmax=13,method="exhaustive")
> reg_sub_summary=summary(reg_subset)
> reg_sub_summary
```

```
Subset selection object
Call: regsubsets.formula(medv ~ ., train, nvmax = 13, method = "exhaustive")
13 Variables  (and intercept)
        Forced in Forced out
crim        FALSE      FALSE
zn          FALSE      FALSE
indus       FALSE      FALSE
chas        FALSE      FALSE
nox         FALSE      FALSE
rm          FALSE      FALSE
age         FALSE      FALSE
dis         FALSE      FALSE
rad         FALSE      FALSE
tax         FALSE      FALSE
ptratio     FALSE      FALSE
```

```
black       FALSE      FALSE
lstat       FALSE      FALSE
1 subsets of each size up to 13
Selection Algorithm: exhaustive
          crim zn  indus chas nox rm  age dis rad tax ptratio black lstat
1  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " " "     " "   " "
2  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " " "     " "   "*"
3  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " " "*"   " "   "*"
4  ( 1 )  " "  " " " "   " "  " " "*" " " "*" " " " " "*"   " "   "*"
5  ( 1 )  " "  " " " "   " "  "*" "*" " " "*" " " " " "*"   " "   "*"
6  ( 1 )  " "  " " " "   "*"  "*" "*" " " "*" " " " " "*"   " "   "*"
7  ( 1 )  "*"  " " " "   " "  "*" "*" " " "*" "*" " " "*"   " "   "*"
8  ( 1 )  "*"  " " " "   " "  "*" "*" " " "*" "*" "*" "*"   " "   "*"
9  ( 1 )  "*"  "*" " "   " "  "*" "*" " " "*" "*" "*" "*"   " "   "*"
10  ( 1 ) "*"  "*" " "   "*"  "*" "*" " " "*" "*" "*" "*"   " "   "*"
11  ( 1 ) "*"  "*" "*"   "*"  "*" "*" " " "*" "*" "*" "*"   " "   "*"
12  ( 1 ) "*"  "*" "*"   "*"  "*" "*" " " "*" "*" "*" "*"   "*"   "*"
13  ( 1 ) "*"  "*" "*"   "*"  "*" "*" "*" "*" "*" "*" "*"   "*"   "*"
```

```
> reg_sub_summary$cp[which.min(reg_sub_summary$cp)]
```

```
[1] 8.772625
```

2.2 Metodología Fordward

Se implementa la selección de variables de la sigueinte forma bajo la metodología Fordward:

```
> reg_forw=regsubsets(medv~.,train,method="forward",nvmax=13)
> reg_forw_summary=summary(reg_forw)
> reg_forw_summary
```

```
Subset selection object
Call: regsubsets.formula(medv ~ ., train, method = "forward", nvmax = 13)
13 Variables  (and intercept)
        Forced in Forced out
crim        FALSE      FALSE
zn          FALSE      FALSE
indus       FALSE      FALSE
chas        FALSE      FALSE
nox         FALSE      FALSE
rm          FALSE      FALSE
age         FALSE      FALSE
dis         FALSE      FALSE
rad         FALSE      FALSE
tax         FALSE      FALSE
ptratio     FALSE      FALSE
black       FALSE      FALSE
lstat       FALSE      FALSE
1 subsets of each size up to 13
Selection Algorithm: forward
          crim zn  indus chas nox rm  age dis rad tax ptratio black lstat
1  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " " "     " "   " "
2  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " " "     " "   "*"
3  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " " "*"   " "   "*"
4  ( 1 )  " "  " " " "   " "  " " "*" " " "*" " " " " "*"   " "   "*"
5  ( 1 )  " "  " " " "   " "  "*" "*" " " "*" " " " " "*"   " "   "*"
6  ( 1 )  " "  " " " "   "*"  "*" "*" " " "*" " " " " "*"   " "   "*"
7  ( 1 )  " "  "*" " "   "*"  "*" "*" " " "*" " " " " "*"   " "   "*"
8  ( 1 )  "*"  "*" " "   "*"  "*" "*" " " "*" " " " " "*"   " "   "*"
9  ( 1 )  "*"  "*" " "   "*"  "*" "*" " " "*" "*" " " "*"   " "   "*"
```

```
10  ( 1 ) "*"  "*" " "   "*"  "*" "*" " " "*" "*" "*" "*"     " "   "*"
11  ( 1 ) "*"  "*" "*"   "*"  "*" "*" " " "*" "*" "*" "*"     " "   "*"
12  ( 1 ) "*"  "*" "*"   "*"  "*" "*" " " "*" "*" "*" "*"     "*"   "*"
13  ( 1 ) "*"  "*" "*"   "*"  "*" "*" "*" "*" "*" "*" "*"     "*"   "*"
```

```
> reg_forw_summary$cp[which.min(reg_forw_summary$cp)]
```

```
[1] 8.772625
```

```
> which.min(reg_forw_summary$cp)
```

```
[1] 10
```

2.3 Metodología Backward

Al ver que las dos implementaciones anteriores seleccionaron las mismas variables. Se implementa la metodología backward con el fin de comprobar si existe algúna diferencia en su resultado.

```
> reg_back=regsubsets(medv~.,train,method="backward",nvmax=13)
> reg_back_summary=summary(reg_back)
> reg_back_summary
```

```
Subset selection object
Call: regsubsets.formula(medv ~ ., train, method = "backward", nvmax = 13)
13 Variables  (and intercept)
        Forced in Forced out
crim        FALSE      FALSE
zn          FALSE      FALSE
indus       FALSE      FALSE
chas        FALSE      FALSE
nox         FALSE      FALSE
rm          FALSE      FALSE
age         FALSE      FALSE
dis         FALSE      FALSE
rad         FALSE      FALSE
tax         FALSE      FALSE
ptratio     FALSE      FALSE
black       FALSE      FALSE
lstat       FALSE      FALSE
1 subsets of each size up to 13
Selection Algorithm: backward
         crim zn  indus chas nox rm  age dis rad tax ptratio black lstat
1  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " "     " "   " "
2  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " "     " "   "*"
3  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " " "     " "   "*"
4  ( 1 )  " "  " " " "   " "  " " "*" " " "*" " " " "     " "   "*"
5  ( 1 )  "*"  " " " "   " "  " " "*" " " "*" " " " "     " "   "*"
6  ( 1 )  "*"  " " " "   " "  " " "*" " " "*" "*" " "     " "   "*"
7  ( 1 )  "*"  " " " "   " "  "*" "*" " " "*" "*" " "     " "   "*"
8  ( 1 )  "*"  " " " "   " "  "*" "*" " " "*" "*" "*"     " "   "*"
9  ( 1 )  "*"  "*" " "   " "  "*" "*" " " "*" "*" "*"     " "   "*"
10  ( 1 ) "*"  "*" " "   "*"  "*" "*" " " "*" "*" "*"     " "   "*"
11  ( 1 ) "*"  "*" "*"   "*"  "*" "*" " " "*" "*" "*"     " "   "*"
12  ( 1 ) "*"  "*" "*"   "*"  "*" "*" " " "*" "*" "*"     "*"   "*"
13  ( 1 ) "*"  "*" "*"   "*"  "*" "*" "*" "*" "*" "*"     "*"   "*"
```

```
> reg_back_summary$cp[which.min(reg_back_summary$cp)]
```

```
[1] 8.772625
```

```
> which.min(reg_back_summary$cp)
```

```
[1] 10
```

La base de datos Boston contiene información de 506 viviendas que describen datos propios de la vivienda y datos de la ciudad de ubicación de la vivienda, el objetivo para este ejercicio, es elegir las variables que mejor describan el comportamiento de la variable de respuesta (medv).

Para realizar esta elección se hace uso de los métodos secuenciales inteligentes exhaustivo, forward y backward. Con los tres modelos se llega a la misma conclusión con la metrica $C_p de Mallows$, donde selecciona el modelo 10 que tiene un Mallows's CP = 8.772625 y que contiene las siguiente variables:

- crim
- zn
- chas
- nox
- rm
- dis
- rad
- tax
- ptratio
- black
- lstat

Excluyendo de las siguiente variables:

- age
- indus

Es decir selecciona 11 de las 13 variables que son la que explican la varianza del comportamiento de la mediana de las viviendas expresadas en miles de dólares. Luego de aplicar los 3 métodos secuenciales, se excluyeron las variables "indus" y "age" que corresponden a:

indus La proporción de acres de negocios no minoristas por ciudad. **age** Proporción de unidades ocupadas por sus propietarios construidas antes de 1940.

3 Segundo punto

Con las variables predictoras (X) calcule e interprete los dos primeros componentes principales escalando las variables. Si tiene duda sobre lo que significan las variables, use "?Boston"

Cálculo de los componentes principales

```
> pp = princomp(datos,scores=TRUE)
> pp$loadings
```

```
Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
crim                                 0.272  0.930  0.157  0.151  0.106
zn                    -0.632  0.763
indus                                      -0.126  0.860         -0.465
chas
nox
rm
age                    0.752  0.641
dis                                               -0.110                 0.112
rad                                         0.231 -0.360 -0.400 -0.797 -0.134
```

```
tax      0.949 -0.293
ptratio                                                    -0.154  0.973
black   -0.291 -0.956
lstat                                0.460         0.169 -0.813  0.277
medv                                -0.829  0.242  0.223 -0.378  0.177  0.133
        Comp.11 Comp.12 Comp.13 Comp.14
crim
zn
indus
chas                     1.000
nox                              1.000
rm               -0.998
age
dis      0.984
rad
tax
ptratio -0.123
black
lstat
medv

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.071  0.071  0.071  0.071  0.071  0.071  0.071  0.071  0.071
Cumulative Var  0.071  0.143  0.214  0.286  0.357  0.429  0.500  0.571  0.643
               Comp.10 Comp.11 Comp.12 Comp.13 Comp.14
SS loadings      1.000   1.000   1.000   1.000   1.000
Proportion Var   0.071   0.071   0.071   0.071   0.071
Cumulative Var   0.714   0.786   0.857   0.929   1.000
```

> pp %>% summary()

```
Importance of components:
                            Comp.1     Comp.2      Comp.3       Comp.4
Standard deviation     175.6386232 78.9839281 28.65028069 16.333041612
Proportion of Variance   0.8045722  0.1627056  0.02140833  0.006957594
Cumulative Proportion    0.8045722  0.9672778  0.98868609  0.995643684
                             Comp.5      Comp.6       Comp.7       Comp.8
Standard deviation     8.768743607 6.825773380 4.1266429051 3.6785390152
Proportion of Variance 0.002005394 0.001215147 0.0004441388 0.0003529195
Cumulative Proportion  0.997649078 0.998864225 0.9993083634 0.9996612829
                              Comp.9      Comp.10      Comp.11      Comp.12
Standard deviation     2.9819469043 1.647979e+00 1.048736e+00 4.663730e-01
Proportion of Variance 0.0002319128 7.083176e-05 2.868514e-05 5.672726e-06
Cumulative Proportion  0.9998931957 9.999640e-01 9.999927e-01 9.999984e-01
                            Comp.13      Comp.14
Standard deviation     2.428758e-01 5.403254e-02
Proportion of Variance 1.538485e-06 7.614400e-08
Cumulative Proportion  9.999999e-01 1.000000e+00
```

Interpretación de componentes principales

Como vemos luego de realizar el modelo bajo los componentes principales se alcanza a explicar cerca del 96% de la varianza de los datos originales. El primer componente explica el 80.45% de la varianza y el segundo componente explica el 16.27%. Este primer componente se compone con los dos primeros componentes principales (Tax y Black).

4 Tercer punto

Corra los modelos de regresión por componentes principales y por "partial least squares" y compare el poder predictivo con el MSE en test de estos con el modelos de regresión lineal del punto anterior.

4.1 Modelo Lineal

```
> lm1=lm(medv~crim+zn+chas+nox+rm+dis+rad+tax+ptratio+black+lstat,data=train)
> summary(lm1)

Call:
lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
    tax + ptratio + black + lstat, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-19.8902  -2.6783  -0.6409   1.7064  25.5497

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.822677   6.127979   4.867 1.64e-06 ***
crim         -0.191247   0.044739  -4.275 2.40e-05 ***
zn            0.042950   0.014041   3.059 0.002373 **
chas          1.980786   0.888175   2.230 0.026298 *
nox         -13.632103   4.207164  -3.240 0.001296 **
rm            4.739935   0.469203  10.102  < 2e-16 ***
dis          -1.346653   0.199311  -6.757 5.10e-11 ***
rad           0.442107   0.084468   5.234 2.70e-07 ***
tax          -0.014484   0.004259  -3.400 0.000742 ***
ptratio      -0.791276   0.138608  -5.709 2.24e-08 ***
black        -0.002311   0.006554  -0.353 0.724547
lstat        -0.524320   0.054923  -9.546  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.839 on 394 degrees of freedom
Multiple R-squared:  0.7378,    Adjusted R-squared:  0.7304
F-statistic: 100.8 on 11 and 394 DF,  p-value: < 2.2e-16

> pred=predict(lm1,test)
> mse=mean((test$medv-pred)^2)
> mse

[1] 32.83844
```

4.2 Modelo bajo componetes principales

```
> lm2=pcr(medv~crim+zn+chas+nox+rm+dis+rad+tax+ptratio+black+lstat,
+         data=train,scale=T,validation="CV")
> summary(lm2)

Data:         X dimension: 406 11
        Y dimension: 406 1
Fit method: svdpc
Number of components considered: 11

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV           9.331    7.839    7.302    5.837    5.674    5.466    5.274
```

```
adjCV        9.331    7.835    7.455    5.821    5.672    5.454    5.261
       7 comps  8 comps  9 comps  10 comps  11 comps
CV       5.267    5.283    5.306     5.139     5.027
adjCV    5.255    5.271    5.292     5.124     5.013

TRAINING: % variance explained
      1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X       39.62    52.68    64.90    74.23    82.01    88.32     92.7    95.49
medv    29.88    41.87    62.39    64.63    67.76    70.24     70.7    70.70
      9 comps  10 comps  11 comps
X       97.54     99.06    100.00
medv    70.88     72.63     73.78
```

```
> predpp=predict(lm2,test)
> msepp=mean((test$medv-predpp)^2)
> msepp
```

```
[1] 25.81188
```

4.3 Modelo bajo mínimos cuadrados parciales

```
> lm3=plsr(medv~crim+zn+chas+nox+rm+dis+rad+tax+ptratio+black+lstat,
+          data=train,scale=T,validation="CV")
> summary(lm3)
```

```
Data:         X dimension: 406 11
        Y dimension: 406 1
Fit method: kernelpls
Number of components considered: 11

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV           9.331    6.507    5.197    5.115    5.049    5.017    5.008
adjCV        9.331    6.503    5.190    5.104    5.037    5.004    4.995
       7 comps  8 comps  9 comps  10 comps  11 comps
CV       4.998    4.993    4.994     4.996     4.996
adjCV    4.985    4.981    4.982     4.984     4.984

TRAINING: % variance explained
      1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X       36.69    51.60    59.21    65.47    73.05    80.43    85.97    90.52
medv    52.36    70.64    72.31    73.21    73.63    73.72    73.75    73.77
      9 comps  10 comps  11 comps
X       95.21     97.24    100.00
medv    73.78     73.78     73.78
```

```
> predpl=predict(lm3,test)
> msepl=mean((test$medv-predpl)^2)
> msepl
```

```
[1] 30.48446
```

4.4 Comparación y conclusión

```
> data.frame("Modelo Lineal"= mse,
+            "Modelo PCA" = msepp,
+            "Modelo PLS"  = msepl) %>% xtable()
```

	Modelo.Lineal	Modelo.PCA	Modelo.PLS
1	32.84	25.81	30.48

Según el MSE tenemos que el modelo lineal es de 32.84, seguido del Mínimos Cuadrados Parciales con 30.48 y finalmente, el modelo de mejor ajuste es el modelo bajo la metodología de Componentes Principales 25.81. Es decir, el modelo que mejor se ajusta es el modelo bajo componentes princiales.