CSCI E-106: W 14: Variable Selection (preliminary)

5/2/2019

Contents

```
## package 'Biobase' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\P\AppData\Local\Temp\RtmpWSAs6E\downloaded_packages
```

Packages used:

1. leaps - for computing stepwise regression
2. MASS - datasets and functions dealing with Modern and APpl
3. caret - for easy machine learning workflow
4. tidyverse - for easy data manipulation and visualization

Learning about Leaps()

leaps() performs an exhaustive search for the best subsets of the variables in x for predicting y in linear regression, using an efficient branch-and-bound algorithm. It is a compatibility wrapper for regsubsetsdoes the same thing better. Since the algorithm returns a best model of each size, the results do not depend on a penalty model for model size: it doesn't make any difference whether you want to use AIC, BIC, CIC, DIC, . . .

```
# Fit the full model

full.model <- lm(Fertility ~ ., data = swiss)

# just to see the first 10 rows of the (given) swiss data

head(swiss, n = 10)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
## Broye             83.8        70.2          16         7    92.85
## Glane             92.4        67.8          14         8    97.16
## Gruyere           82.4        53.3          12         7    97.67
## Sarine            82.9        45.2          16        13    91.38
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
## Broye                    23.6
## Glane                    24.9
## Gruyere                  21.0
## Sarine                   24.4
```

```r
# just to see the last 10 rows of the (given) swiss data
tail(swiss, n = 10)
```

```
##              Fertility Agriculture Examination Education Catholic
## Sion              79.3        63.1          13        13    96.83
## Boudry            70.4        38.4          26        12     5.62
## La Chauxdfnd      65.7         7.7          29        11    13.79
## Le Locle          72.7        16.7          22        13    11.22
## Neuchatel         64.4        17.6          35        32    16.92
## Val de Ruz        77.6        37.6          15         7     4.97
## ValdeTravers      67.6        18.7          25         7     8.65
## V. De Geneve      35.0         1.2          37        53    42.34
## Rive Droite       44.7        46.6          16        29    50.43
## Rive Gauche       42.8        27.7          22        29    58.33
##              Infant.Mortality
## Sion                     18.1
## Boudry                   20.3
## La Chauxdfnd             20.5
## Le Locle                 18.9
## Neuchatel                23.0
## Val de Ruz               20.0
## ValdeTravers             19.5
## V. De Geneve             18.0
## Rive Droite              18.2
## Rive Gauche              19.3
```

```r
# Stepwise regression model Step: used to choose a model by AIC in a stepwise
# algorithm first we can use it to go backward

step(full.model, scope = formula(full.model), direction = "backward")
```

```
## Start:  AIC=190.69
## Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## - Examination       1     53.03 2158.1 189.86
## <none>                          2105.0 190.69
## - Agriculture       1    307.72 2412.8 195.10
## - Infant.Mortality  1    408.75 2513.8 197.03
## - Catholic          1    447.71 2552.8 197.75
## - Education         1   1162.56 3267.6 209.36
```

```
## 
## Step:  AIC=189.86
## Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
## 
##                     Df Sum of Sq    RSS    AIC
## <none>                           2158.1 189.86
## - Agriculture        1    264.18 2422.2 193.29
## - Infant.Mortality   1    409.81 2567.9 196.03
## - Catholic           1    956.57 3114.6 205.10
## - Education          1   2249.97 4408.0 221.43
```

```
## 
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic + 
##     Infant.Mortality, data = swiss)
## 
## Coefficients:
##      (Intercept)       Agriculture         Education          Catholic  
##          62.1013           -0.1546           -0.9803            0.1247  
## Infant.Mortality  
##           1.0784
```

```
# we can use it to go forward too
step(full.model, scope = formula(full.model), direction = "forward")
```

```
## Start:  AIC=190.69
## Fertility ~ Agriculture + Examination + Education + Catholic + 
##     Infant.Mortality
```

```
## 
## Call:
## lm(formula = Fertility ~ Agriculture + Examination + Education + 
##     Catholic + Infant.Mortality, data = swiss)
## 
## Coefficients:
##      (Intercept)       Agriculture       Examination         Education  
##          66.9152           -0.1721           -0.2580           -0.8709  
##         Catholic  Infant.Mortality  
##           0.1041            1.0770
```

###With the backwards model the start at AIC = 190.69 has it dropped from the model. At 189.86 nothing is dropped from the model.

```
leaps(x = swiss[, 2:6], y = swiss[, 1], names = names(swiss)[2:6], method = "Cp")
```

```
## $which
##   Agriculture Examination Education Catholic Infant.Mortality
## 1       FALSE       FALSE      TRUE    FALSE            FALSE
## 1       FALSE        TRUE     FALSE    FALSE            FALSE
## 1       FALSE       FALSE     FALSE     TRUE            FALSE
## 1       FALSE       FALSE     FALSE    FALSE             TRUE
## 1        TRUE       FALSE     FALSE    FALSE            FALSE
## 2       FALSE       FALSE      TRUE     TRUE            FALSE
## 2       FALSE       FALSE      TRUE    FALSE             TRUE
## 2       FALSE        TRUE     FALSE    FALSE             TRUE
## 2       FALSE        TRUE      TRUE    FALSE            FALSE
## 2        TRUE       FALSE      TRUE    FALSE            FALSE
## 2        TRUE        TRUE     FALSE    FALSE            FALSE
## 2       FALSE        TRUE     FALSE     TRUE            FALSE
## 2       FALSE       FALSE     FALSE     TRUE             TRUE
## 2        TRUE       FALSE     FALSE    FALSE             TRUE
## 2        TRUE       FALSE     FALSE     TRUE            FALSE
## 3       FALSE       FALSE      TRUE     TRUE             TRUE
## 3        TRUE       FALSE      TRUE     TRUE            FALSE
## 3       FALSE        TRUE      TRUE    FALSE             TRUE
## 3       FALSE        TRUE      TRUE     TRUE            FALSE
## 3        TRUE       FALSE      TRUE    FALSE             TRUE
## 3        TRUE        TRUE      TRUE    FALSE            FALSE
## 3       FALSE        TRUE     FALSE     TRUE             TRUE
## 3        TRUE        TRUE     FALSE    FALSE             TRUE
## 3        TRUE        TRUE     FALSE     TRUE            FALSE
## 3        TRUE       FALSE     FALSE     TRUE             TRUE
## 4        TRUE       FALSE      TRUE     TRUE             TRUE
## 4       FALSE        TRUE      TRUE     TRUE             TRUE
## 4        TRUE        TRUE      TRUE     TRUE            FALSE
## 4        TRUE        TRUE      TRUE    FALSE             TRUE
## 4        TRUE        TRUE     FALSE     TRUE             TRUE
## 5        TRUE        TRUE      TRUE     TRUE             TRUE
##
## $label
## [1] "(Intercept)"      "Agriculture"      "Examination"
## [4] "Education"        "Catholic"         "Infant.Mortality"
##
## $size
##  [1] 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 6
##
## $Cp
##  [1] 35.204895 38.483494 66.746668 72.546431 79.376482 18.486158 19.846060
##  [8] 23.827488 28.135883 35.997988 38.324902 38.654479 52.540917 54.450565
## [15] 64.094684  8.178162 11.014774 14.252399 20.438243 21.663853 22.954993
## [22] 25.175215 25.342889 38.442616 46.855566  5.032800  9.993398 11.961249
## [29] 12.720023 26.643242  6.000000
```

```
leaps(x = swiss[, 2:6], y = swiss[, 1], names = names(swiss)[2:6], method = "r2")
```

```
## $which
##   Agriculture Examination Education Catholic Infant.Mortality
## 1       FALSE       FALSE      TRUE    FALSE            FALSE
## 1       FALSE        TRUE     FALSE    FALSE            FALSE
## 1       FALSE       FALSE     FALSE     TRUE            FALSE
## 1       FALSE       FALSE     FALSE    FALSE             TRUE
## 1        TRUE       FALSE     FALSE    FALSE            FALSE
## 2       FALSE       FALSE      TRUE     TRUE            FALSE
## 2       FALSE       FALSE      TRUE    FALSE             TRUE
## 2       FALSE        TRUE     FALSE    FALSE             TRUE
## 2       FALSE        TRUE      TRUE    FALSE            FALSE
## 2        TRUE       FALSE      TRUE    FALSE            FALSE
## 2        TRUE        TRUE     FALSE    FALSE            FALSE
## 2       FALSE        TRUE     FALSE     TRUE            FALSE
## 2       FALSE       FALSE     FALSE     TRUE             TRUE
## 2        TRUE       FALSE     FALSE    FALSE             TRUE
## 2        TRUE       FALSE     FALSE     TRUE            FALSE
## 3       FALSE       FALSE      TRUE     TRUE             TRUE
## 3        TRUE       FALSE      TRUE     TRUE            FALSE
## 3       FALSE        TRUE      TRUE    FALSE             TRUE
## 3       FALSE        TRUE      TRUE     TRUE            FALSE
## 3        TRUE       FALSE      TRUE    FALSE             TRUE
## 3        TRUE        TRUE      TRUE    FALSE            FALSE
## 3       FALSE        TRUE     FALSE     TRUE             TRUE
## 3        TRUE        TRUE     FALSE    FALSE             TRUE
## 3        TRUE        TRUE     FALSE     TRUE            FALSE
## 3        TRUE       FALSE     FALSE     TRUE             TRUE
## 4        TRUE       FALSE      TRUE     TRUE             TRUE
## 4       FALSE        TRUE      TRUE     TRUE             TRUE
## 4        TRUE        TRUE      TRUE     TRUE            FALSE
## 4        TRUE        TRUE      TRUE    FALSE             TRUE
## 4        TRUE        TRUE     FALSE     TRUE             TRUE
## 5        TRUE        TRUE      TRUE     TRUE             TRUE
##
## $label
## [1] "(Intercept)"      "Agriculture"      "Examination"
## [4] "Education"        "Catholic"         "Infant.Mortality"
##
## $size
##  [1] 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 6
##
## $r2
##  [1] 0.4406156 0.4171645 0.2150035 0.1735189 0.1246649 0.5745071 0.5647800
##  [8] 0.5363016 0.5054845 0.4492484 0.4326045 0.4302471 0.3309201 0.3172607
## [15] 0.2482782 0.6625438 0.6422541 0.6190960 0.5748498 0.5660833 0.5568480
## [22] 0.5409672 0.5397679 0.4460681 0.3858919 0.6993476 0.6638654 0.6497897
## [29] 0.6443624 0.5447723 0.7067350
```

```
leaps(x = swiss[, 2:6], y = swiss[, 1], names = names(swiss)[2:6], method = "adjr2")
```

```
## $which
##   Agriculture Examination Education Catholic Infant.Mortality
## 1       FALSE       FALSE      TRUE    FALSE            FALSE
## 1       FALSE        TRUE     FALSE    FALSE            FALSE
## 1       FALSE       FALSE     FALSE     TRUE            FALSE
## 1       FALSE       FALSE     FALSE    FALSE             TRUE
## 1        TRUE       FALSE     FALSE    FALSE            FALSE
## 2       FALSE       FALSE      TRUE     TRUE            FALSE
## 2       FALSE       FALSE      TRUE    FALSE             TRUE
## 2       FALSE        TRUE     FALSE    FALSE             TRUE
## 2       FALSE        TRUE      TRUE    FALSE            FALSE
## 2        TRUE       FALSE      TRUE    FALSE            FALSE
## 2        TRUE        TRUE     FALSE    FALSE            FALSE
## 2       FALSE        TRUE     FALSE     TRUE            FALSE
## 2       FALSE       FALSE     FALSE     TRUE             TRUE
## 2        TRUE       FALSE     FALSE    FALSE            TRUE
## 2        TRUE       FALSE     FALSE     TRUE           FALSE
## 3       FALSE       FALSE      TRUE     TRUE            TRUE
## 3        TRUE       FALSE      TRUE     TRUE           FALSE
## 3       FALSE        TRUE      TRUE    FALSE            TRUE
## 3       FALSE        TRUE      TRUE     TRUE           FALSE
## 3        TRUE       FALSE      TRUE    FALSE            TRUE
## 3        TRUE        TRUE      TRUE    FALSE           FALSE
## 3       FALSE        TRUE     FALSE     TRUE            TRUE
## 3        TRUE        TRUE     FALSE    FALSE            TRUE
## 3        TRUE        TRUE     FALSE     TRUE           FALSE
## 3        TRUE       FALSE     FALSE     TRUE            TRUE
## 4        TRUE       FALSE      TRUE     TRUE            TRUE
## 4       FALSE        TRUE      TRUE     TRUE            TRUE
## 4        TRUE        TRUE      TRUE     TRUE           FALSE
## 4        TRUE        TRUE      TRUE    FALSE            TRUE
## 4        TRUE        TRUE     FALSE     TRUE            TRUE
## 5        TRUE        TRUE      TRUE     TRUE            TRUE
##
## $label
## [1] "(Intercept)"      "Agriculture"      "Examination"
## [4] "Education"        "Catholic"         "Infant.Mortality"
##
## $size
##  [1] 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 6
##
## $adjr2
##  [1] 0.4281849 0.4042126 0.1975591 0.1551527 0.1052130 0.5551665 0.5449973
##  [8] 0.5152244 0.4830065 0.4242143 0.4068138 0.4043492 0.3005074 0.2862271
## [15] 0.2141091 0.6390004 0.6172951 0.5925213 0.5451882 0.5358100 0.5259304
## [22] 0.5089417 0.5076587 0.4074217 0.3430471 0.6707140 0.6318526 0.6164364
## [29] 0.6104921 0.5014173 0.6709710
```

```r
models <- regsubsets(Fertility ~ ., data = swiss, nvmax = 5, method = "seqrep")
summary(models)
```

```
## Subset selection object
## Call: regsubsets.formula(Fertility ~ ., data = swiss, nvmax = 5, method = "seqrep")
## 5 Variables  (and intercept)
##                  Forced in Forced out
## Agriculture          FALSE      FALSE
## Examination          FALSE      FALSE
## Education            FALSE      FALSE
## Catholic             FALSE      FALSE
## Infant.Mortality     FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: 'sequential replacement'
##          Agriculture Examination Education Catholic Infant.Mortality
## 1  ( 1 ) " "         " "         "*"       " "      " "
## 2  ( 1 ) " "         " "         "*"       "*"      " "
## 3  ( 1 ) " "         " "         "*"       "*"      "*"
## 4  ( 1 ) "*"         "*"         "*"       "*"      " "
## 5  ( 1 ) "*"         "*"         "*"       "*"      "*"
```

```r
models1 <- regsubsets(Fertility ~ ., data = swiss, nvmax = 5, method = "backward")
summary(models1)
```

```
## Subset selection object
## Call: regsubsets.formula(Fertility ~ ., data = swiss, nvmax = 5, method = "backward")
## 5 Variables  (and intercept)
##                  Forced in Forced out
## Agriculture          FALSE      FALSE
## Examination          FALSE      FALSE
## Education            FALSE      FALSE
## Catholic             FALSE      FALSE
## Infant.Mortality     FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: backward
##          Agriculture Examination Education Catholic Infant.Mortality
## 1  ( 1 ) " "         " "         "*"       " "      " "
## 2  ( 1 ) " "         " "         "*"       "*"      " "
## 3  ( 1 ) " "         " "         "*"       "*"      "*"
## 4  ( 1 ) "*"         " "         "*"       "*"      "*"
## 5  ( 1 ) "*"         "*"         "*"       "*"      "*"
```

```r
models2 <- regsubsets(Fertility ~ ., data = swiss, nvmax = 5, method = "forward")
summary(models2)
```

```
## Subset selection object
## Call: regsubsets.formula(Fertility ~ ., data = swiss, nvmax = 5, method = "forward")
## 5 Variables  (and intercept)
##                  Forced in Forced out
## Agriculture          FALSE      FALSE
## Examination          FALSE      FALSE
## Education            FALSE      FALSE
## Catholic             FALSE      FALSE
## Infant.Mortality     FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: forward
##          Agriculture Examination Education Catholic Infant.Mortality
## 1  ( 1 ) " "         " "         "*"       " "      " "
## 2  ( 1 ) " "         " "         "*"       "*"      " "
## 3  ( 1 ) " "         " "         "*"       "*"      "*"
## 4  ( 1 ) "*"         " "         "*"       "*"      "*"
## 5  ( 1 ) "*"         "*"         "*"       "*"      "*"
```

##Using Polynomial Regression

```r
Y <- swiss$Fertility
x1 <- swiss$Agriculture - mean(swiss$Agriculture)
x2 <- swiss$Examination - mean(swiss$Examination)
x3 <- swiss$Education - mean(swiss$Education)
x4 <- swiss$Catholic - mean(swiss$Catholic)
x5 <- swiss$Infant.Mortality - mean(swiss$Infant.Mortality)

# with interaction
```

```r
f1 <- lm(Y ~ (x1 + x2 + x3 + x4 + x5)^2)
print(f1)
```

```
##
## Call:
## lm(formula = Y ~ (x1 + x2 + x3 + x4 + x5)^2)
##
## Coefficients:
## (Intercept)           x1           x2           x3           x4
##   70.293960    -0.168622    -0.325271    -0.889362     0.050353
##          x5        x1:x2        x1:x3        x1:x4        x1:x5
##    0.866190     0.021373     0.019060     0.002626     0.063698
##       x2:x3        x2:x4        x2:x5        x3:x4        x3:x5
##    0.075174    -0.001533     0.171015    -0.007132     0.033586
##       x4:x5
##    0.009919
```

```r
# Note that we have named the centered variables x1 and x2.  We also will need
# the second order terms for the model:

x1sq <- x1^2
x2sq <- x2^2
x1x2 <- x1 * x2

x3sq <- x3^2
x4sq <- x4^2
x3x4 <- x3 * x4

f2 <- lm(Y ~ x1 + x2 + x2sq + x1x2)
print(f2)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x2sq + x1x2)
##
## Coefficients:
## (Intercept)           x1           x2         x2sq         x1x2
##    70.48735     -0.08108     -1.09809      0.02212      0.01415
```

```r
# we can try different alternatives

f3 <- lm(Y ~ x3 + x4 + x4sq + x3x4)
print(f3)
```

```
##
## Call:
## lm(formula = Y ~ x3 + x4 + x4sq + x3x4)
##
## Coefficients:
## (Intercept)           x3           x4         x4sq         x3x4
##  70.6176587   -0.8439322    0.0910897   -0.0006319   -0.0099459
```