Homework9

Dima Mikhaylov

11/5/2021

Part 1

```
library(MASS)
#data(package = 'birthwt')
head(birthwt)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

```
?birthwt
```

a. Which of these variables are categorical? Ensure that R is viewing the categorical variables correctly. If needed, use the factor() function to force R to treat the necessary variables as categorical.

First, check data types of all the columns:

```
sapply(birthwt, class)
```

```
##       low       age       lwt      race     smoke       ptl        ht        ui
## "integer" "integer" "integer" "integer" "integer" "integer" "integer" "integer"
##       ftv       bwt
## "integer" "integer"
```

It seems that `race` - mother's race (1 = white, 2 = black, 3 = other) should be viewed as categorical. Apply `factor` function:

```
birthwt$race <- as.factor(birthwt$race)
class(birthwt$race)
```

```
## [1] "factor"
```

b. A classmate of yours makes the following suggestion: "We should remove the variable low as a predictor for the birth weight of babies." Do you agree with your classmate? Briefly explain. Hint: you do not need to do any statistical analysis to answer this question.

Yes, I agree, variable `'low'` is not a good predictor for response variable `'bwt'`. It is actually produced from the response variable. In other words, it should be removed to avoid collinearity problem.

c. Based on your answer to part 1b, perform all possible regressions using the regsubsets() function from the leaps package. Write down the predictors that lead to a first-order model having the best

First, drop variable ' `low` ' as discussed above, create new dataset `data` :

```
data <- within(birthwt, rm('low'))
head(data)
```

```
##    age lwt race smoke ptl ht ui ftv  bwt
## 85  19 182    2     0   0  0  1   0 2523
## 86  33 155    3     0   0  0  0   3 2551
## 87  20 105    1     1   0  0  0   1 2557
## 88  21 108    1     1   0  0  1   2 2594
## 89  18 107    1     1   0  0  1   0 2600
## 91  21 124    3     0   0  0  0   0 2622
```

Next, run all possible subsets:

```
library(leaps)
allregs <- regsubsets(bwt ~ ., data=data, nbest=1)
summary(allregs)
```

```
## Subset selection object
## Call: regsubsets.formula(bwt ~ ., data = data, nbest = 1)
## 9 Variables  (and intercept)
##       Forced in Forced out
## age       FALSE      FALSE
## lwt       FALSE      FALSE
## race2     FALSE      FALSE
## race3     FALSE      FALSE
## smoke     FALSE      FALSE
## ptl       FALSE      FALSE
## ht        FALSE      FALSE
## ui        FALSE      FALSE
## ftv       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          age lwt race2 race3 smoke ptl ht  ui  ftv
## 1  ( 1 ) " " " " " " "   " " "   " " " " "*" " "
## 2  ( 1 ) " " " " " " "   " " "   " " "*" "*" " "
## 3  ( 1 ) " " "*" " " "   " " "   " " "*" "*" " "
## 4  ( 1 ) " " " " " " "*"   "*"   "*"   " " " " "*" " "
## 5  ( 1 ) " " " " " " "*"   "*"   "*"   " " "*" "*" " "
## 6  ( 1 ) " " "*" "*" "*"   "*"   "*"   " " "*" "*" " "
## 7  ( 1 ) " " "*" "*" "*"   "*"   "*"   "*" "*" "*" " "
## 8  ( 1 ) "*" "*" "*" "*"   "*"   "*"   "*" "*" "*" " "
```

i. adjusted R2

The best R2 is given by the following coef:

```
coef(allregs, which.max(summary(allregs)$adjr2))
```

```
## (Intercept)         lwt       race2       race3       smoke          ht
##  2837.26392     4.24155  -475.05760  -348.15038  -356.32095  -585.19312
##          ui
##  -525.52390
```

ii. Mallow's Cp,

The best Mallows's C is given by the following coef:

```
coef(allregs, which.min(summary(allregs)$cp))
```

```
## (Intercept)         lwt       race2       race3       smoke          ht
##  2837.26392     4.24155  -475.05760  -348.15038  -356.32095  -585.19312
##          ui
##  -525.52390
```

iii. BIC.

The best BIC is given by the following coef:

```
coef(allregs, which.min(summary(allregs)$bic))
```

```
## (Intercept)         lwt       race2       race3       smoke          ht
##  2837.26392     4.24155  -475.05760  -348.15038  -356.32095  -585.19312
##          ui
##  -525.52390
```

d. Based on your answer to part 1b, use backward selection to find the best model according to AIC. Start with the first-order model with all the predictors. What is the regression equation selected?

```
regnull <- lm(bwt ~ 1, data=data)
regfull <- lm(bwt ~ ., data=data)
step(regfull, scope=list(lower=regnull, upper=regfull), direction = 'backward')
```

```
## Start:  AIC=2458.21
## bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv
##
##         Df Sum of Sq      RSS    AIC
## - ftv    1     38708 75741025 2456.3
## - age    1     58238 75760555 2456.3
## - ptl    1     95285 75797602 2456.4
## <none>               75702317 2458.2
## - lwt    1   2661604 78363921 2462.7
## - ht     1   3631032 79333349 2465.1
## - smoke  1   4623219 80325536 2467.4
## - race   2   6578597 82280914 2470.0
## - ui     1   5839544 81541861 2470.2
##
## Step:  AIC=2456.3
## bwt ~ age + lwt + race + smoke + ptl + ht + ui
##
##         Df Sum of Sq      RSS    AIC
## - age    1     79115 75820139 2454.5
## - ptl    1     91560 75832585 2454.5
## <none>               75741025 2456.3
## - lwt    1   2623988 78365013 2460.7
## - ht     1   3592430 79333455 2463.1
## - smoke  1   4606425 80347449 2465.5
## - race   2   6552496 82293521 2468.0
## - ui     1   5817995 81559020 2468.3
##
## Step:  AIC=2454.5
## bwt ~ lwt + race + smoke + ptl + ht + ui
##
##         Df Sum of Sq      RSS    AIC
## - ptl    1    117366 75937505 2452.8
## <none>               75820139 2454.5
## - lwt    1   2545892 78366031 2458.7
## - ht     1   3546591 79366731 2461.1
## - smoke  1   4530009 80350149 2463.5
## - race   2   6571668 82391807 2466.2
## - ui     1   5751122 81571261 2466.3
##
## Step:  AIC=2452.79
```

```
## bwt ~ lwt + race + smoke + ht + ui
##
##         Df Sum of Sq      RSS    AIC
## <none>               75937505 2452.8
## - lwt    1   2674229 78611734 2457.3
## - ht     1   3584838 79522343 2459.5
## - smoke  1   4950633 80888138 2462.7
## - race   2   6630123 82567628 2464.6
## - ui     1   6353218 82290723 2466.0
```

```
##
## Call:
## lm(formula = bwt ~ lwt + race + smoke + ht + ui, data = data)
##
## Coefficients:
## (Intercept)          lwt        race2        race3        smoke           ht
##    2837.264        4.242     -475.058     -348.150     -356.321     -585.193
##          ui
##    -525.524
```

Selected model has AIC of AIC=2452.79: bwt ~ lwt + race + smoke + ht + ui

Part 2

a. The output below is obtained after using the step() function using forward selection, starting with a model with just the intercept term. What is the model selected based on forward selection?

The model with smallest AIC=-132.94 should be selected in this case: Share ~ discount + promo + price

b. Your client asks you to explain what each step in the output shown above means. Explain the forward selection procedure to your client, for this output.

In short, forward selection procedure starts with a simple intercept-only model and adds explanatory variables one by one trying to mimimize AIC SCORE.

Step 1: regress response variable ' `Share'` on constant 1 - this produces a benchmark intercept-only model with AIC=-94.8

Step 2: use 1 best predictor ' `discount` '- this produces AIC=-128.14

Step 3: use 2 best predicttors ' `discount` ' and ' `promo` ' - this produces AIC=-129.69

Step 4: additing ‘ `price` ’ lowers the score more down to AIC=-132.94

Step 5: stop here, becuase adding the remaining variables ‘ `time` ’ or ‘ `nielsen` ’ does not decrease AIC any further. Actually selection ‘ `none` ’ is the best choice because the resulting combination produces same lower score of -132.94, as in Step #3.

c. Your client asks if he should go ahead and use the models selected in part 2a. What advice do you have for your client?

It is the best model in terms of AIC. This result can be checked by running backward selection or stepwise regression - they all should produce similar results if minimal AIC is desirable. Maximazing expected entropy with BIC Schwarz criterion should also lead to similar results; this could be used to check the validity of the original findings.

My other advice would be to consider additional measures of penalized fit, such as AdjR2 or Mallows’s C, and compare and contrast all the findings. Important is that possibly a better model can be developed if we allow for higher order or interaction terms.

Finally, before deploying a model it is vital to check if regression assumptions are met, at least constant variance and zero means of the residuals. This should inform what model to select.

Part3.

Your client asks you to compare and contrast between R2 and the adjusted R2, specifically: name one advantage of R2 over the adjusted R2, and name one advantage of the adjusted R2 over R2.

Advantage of R2 over the adjusted R2: since R2 assumes that every single independent variable explains the variation in the response variable, it has intuitive interpretation as a proportion of variance accounted for by the model.

Advantage of the adjusted R2 over R2: since it focuses on the independent variables that actually affect variation in the response variable, it can be used for feature importances. For example, one can use adjusted R2 to choose between y=x1+x2 or y=x1+x2+x3. This measure does not have stand alone intuitive interpretation though and should be only used to compare two modeles.

Part 4

Include the function your group wrote to compute the PRESS statistic:

```
#Custom function to calculate PRESS statistics:

PRESS <- function(model) {
    i <- residuals(model)/(1 - lm.influence(model)$hat)
    sum(i^2)
}
```