Homework9

Hyun Suk (Max) Ryoo (hr2ee)

11/7/2021

Set Up

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.2
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyv
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v readr   1.4.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'stringr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

```
## -- Conflicts ------------------------------------------------------------------------ tidyverse_c
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x MASS::select()  masks dplyr::select()
```

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.0.2
```

```
data <- birthwt
head(data)
```

```
##    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
```

1-A

The following variables are categorical.

- low (indicator of birth weight less than 2.5 kg.)
- race (mother's race (1 = white, 2 = black, 3 = other))
- smoke (smoking status during pregnancy.)
- ht (history of hypertension.)
- ui (presence of uterine irritability.)

Making sure R forces the columns as categorical variables.

```
data$low <- factor(data$low)
data$race_tri_cat <- factor(data$race)
levels(data$race_tri_cat) <- c("White", "Black", "Other")
data$smoke <- factor(data$smoke)
data$ht <- factor(data$ht)
data$ui <- factor(data$ui)
head(data)
```

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

1-B

I agree. the low variable is an indicator of brithweight less than 2.5kg. The response variable we are trying to measure is the birth weight in grams. The low variable is directly related to the response so we should not use it for analysis.

1-C

```r
allreg <- regsubsets(bwt~age+lwt+race_tri_cat+smoke+ptl+ht+ui+ftv, data=data, nbest=8)
summary(allreg)
```

```
## Subset selection object
## Call: regsubsets.formula(bwt ~ age + lwt + race_tri_cat + smoke + ptl +
##     ht + ui + ftv, data = data, nbest = 8)
## 9 Variables  (and intercept)
##                   Forced in Forced out
## age                   FALSE      FALSE
## lwt                   FALSE      FALSE
## race_tri_catBlack     FALSE      FALSE
## race_tri_catOther     FALSE      FALSE
## smoke1                FALSE      FALSE
## ptl                   FALSE      FALSE
## ht1                   FALSE      FALSE
## ui1                   FALSE      FALSE
## ftv                   FALSE      FALSE
## 8 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          age lwt race_tri_catBlack race_tri_catOther smoke1 ptl ht1 ui1 ftv
## 1  ( 1 ) " " " " " "               " "               " "    " " " " "*" " "
## 1  ( 2 ) " " " " " "               " "               "*"    " " " " " " " "
## 1  ( 3 ) " " "*" " "               " "               " "    " " " " " " " "
## 1  ( 4 ) " " " " " "               " "               " "    "*" " " " " " "
## 1  ( 5 ) " " " " " "               " "               " "    " " "*" " " " "
## 1  ( 6 ) " " " " " "               "*"               " "    " " " " " " " "
## 1  ( 7 ) " " " " "*"               " "               " "    " " " " " " " "
## 1  ( 8 ) "*" " " " "               " "               " "    " " " " " " " "
## 2  ( 1 ) " " " " " "               " "               " "    " " "*" "*" " "
## 2  ( 2 ) " " " " " "               " "               "*"    " " " " "*" " "
## 2  ( 3 ) " " "*" " "               " "               " "    " " " " "*" " "
## 2  ( 4 ) " " " " "*"               " "               " "    " " " " "*" " "
## 2  ( 5 ) " " " " " "               "*"               " "    " " " " "*" " "
## 2  ( 6 ) " " " " " "               " "               " "    "*" " " "*" " "
## 2  ( 7 ) "*" " " " "               " "               " "    " " " " "*" " "
## 2  ( 8 ) " " " " " "               "*"               "*"    " " " " " " " "
## 3  ( 1 ) " " "*" " "               " "               " "    " " "*" "*" " "
## 3  ( 2 ) " " " " " "               "*"               "*"    " " " " "*" " "
## 3  ( 3 ) " " " " " "               " "               "*"    " " "*" "*" " "
## 3  ( 4 ) " " "*" "*"               " "               " "    " " " " "*" " "
## 3  ( 5 ) " " "*" " "               " "               "*"    " " " " "*" " "
## 3  ( 6 ) " " " " "*"               " "               "*"    " " " " "*" " "
## 3  ( 7 ) " " " " " "               "*"               " "    " " "*" "*" " "
## 3  ( 8 ) " " " " "*"               "*"               " "    " " " " "*" " "
## 4  ( 1 ) " " " " "*"               "*"               "*"    " " " " "*" " "
## 4  ( 2 ) " " " " " "               "*"               "*"    " " "*" "*" " "
## 4  ( 3 ) " " "*" "*"               " "               " "    " " "*" "*" " "
## 4  ( 4 ) " " "*" " "               " "               "*"    " " "*" "*" " "
## 4  ( 5 ) " " "*" "*"               " "               "*"    " " " " "*" " "
## 4  ( 6 ) " " "*" " "               "*"               "*"    " " " " "*" " "
## 4  ( 7 ) " " " " "*"               " "               "*"    " " "*" "*" " "
## 4  ( 8 ) " " " " "*"               "*"               " "    " " "*" "*" " "
## 5  ( 1 ) " " " " "*"               "*"               "*"    " " "*" "*" " "
## 5  ( 2 ) " " "*" "*"               "*"               "*"    " " " " "*" " "
```

```
## 5  ( 3 ) " " "*" "*"              " "               "*"    " " "*" "*" " "
## 5  ( 4 ) " " "*" " "              "*"               "*"    " " "*" "*" " "
## 5  ( 5 ) " " " " "*"              "*"               "*"    "*" " " "*" " "
## 5  ( 6 ) " " "*" "*"              "*"               " "    " " "*" "*" " "
## 5  ( 7 ) " " " " "*"              "*"               "*"    " " " " "*" "*"
## 5  ( 8 ) "*" " " "*"              "*"               "*"    " " " " "*" " "
## 6  ( 1 ) " " "*" "*"              "*"               "*"    " " "*" "*" " "
## 6  ( 2 ) " " " " "*"              "*"               "*"    "*" "*" "*" " "
## 6  ( 3 ) "*" " " "*"              "*"               "*"    " " "*" "*" " "
## 6  ( 4 ) " " " " "*"              "*"               "*"    " " "*" "*" "*"
## 6  ( 5 ) " " "*" "*"              "*"               "*"    "*" " " "*" " "
## 6  ( 6 ) "*" "*" "*"              "*"               "*"    " " " " "*" " "
## 6  ( 7 ) " " "*" "*"              "*"               "*"    " " " " "*" "*"
## 6  ( 8 ) " " "*" "*"              " "               "*"    "*" "*" "*" " "
## 7  ( 1 ) " " "*" "*"              "*"               "*"    "*" "*" "*" " "
## 7  ( 2 ) "*" "*" "*"              "*"               "*"    " " "*" "*" " "
## 7  ( 3 ) " " "*" "*"              "*"               "*"    " " "*" "*" "*"
## 7  ( 4 ) "*" " " "*"              "*"               "*"    "*" "*" "*" " "
## 7  ( 5 ) " " " " "*"              "*"               "*"    "*" "*" "*" "*"
## 7  ( 6 ) "*" " " "*"              "*"               "*"    " " "*" "*" "*"
## 7  ( 7 ) "*" "*" "*"              "*"               "*"    "*" " " "*" " "
## 7  ( 8 ) " " "*" "*"              "*"               "*"    "*" " " "*" "*"
## 8  ( 1 ) "*" "*" "*"              "*"               "*"    "*" "*" "*" " "
## 8  ( 2 ) " " "*" "*"              "*"               "*"    "*" "*" "*" "*"
## 8  ( 3 ) "*" "*" "*"              "*"               "*"    " " "*" "*" "*"
## 8  ( 4 ) "*" " " "*"              "*"               "*"    "*" "*" "*" "*"
## 8  ( 5 ) "*" "*" "*"              "*"               "*"    "*" " " "*" "*"
## 8  ( 6 ) "*" "*" "*"              " "               "*"    "*" "*" "*" "*"
## 8  ( 7 ) "*" "*" " "              "*"               "*"    "*" "*" "*" "*"
## 8  ( 8 ) "*" "*" "*"              "*"               " "    "*" "*" "*" "*"
```

1-C-I

The best model for Adjusted R^2.

```
max(summary(allreg)$adjr2)
```

```
## [1] 0.2153525
```

```
which.max(summary(allreg)$adjr2)
```

```
## [1] 41
```

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

```
##       (Intercept)                lwt race_tri_catBlack race_tri_catOther
##        2837.26392            4.24155        -475.05760        -348.15038
##            smoke1                ht1                ui1
##         -356.32095        -585.19312        -525.52390
```

The predictors for the best adjusted R^2 had predcitors of

- lwt
- race_black
- race_other
- smoke
- ht

- ui

1-C-II

The best model for Mallow's C_p

```
min(summary(allreg)$cp)
```

```
## [1] 4.556107
```

```
which.min(summary(allreg)$cp)
```

```
## [1] 41
```

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

```
##       (Intercept)               lwt race_tri_catBlack race_tri_catOther
##         2837.26392           4.24155        -475.05760        -348.15038
##             smoke1               ht1               ui1
##         -356.32095        -585.19312        -525.52390
```

The predictors for Mallow's C_p had predcitors of

- lwt
- race_black
- race_other
- smoke
- ht
- ui

1-C-III

The best model for BIC

```
min(summary(allreg)$bic)
```

```
## [1] -15.27446
```

```
which.min(summary(allreg)$bic)
```

```
## [1] 41
```

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

```
##       (Intercept)               lwt race_tri_catBlack race_tri_catOther
##         2837.26392           4.24155        -475.05760        -348.15038
##             smoke1               ht1               ui1
##         -356.32095        -585.19312        -525.52390
```

The predictors for Mallow's BIC had predcitors of

- lwt
- race_black
- race_other
- smoke
- ht
- ui

All three adjusted R^2, Mallow's C_p, and BIC led to the same model.

1-D

Starting with the first-order model with all the predictors, backward selection was done to find the best model according to the AIC.

```
##intercept only model
regnull <- lm(bwt~1, data=data)
##model with all predictors
regfull <- lm(bwt~age+lwt+race_tri_cat+smoke+ptl+ht+ui+ftv, data=data)
step(regfull, scope=list(lower=regnull, upper=regfull), direction="backward")
```

```
## Start:  AIC=2458.21
## bwt ~ age + lwt + race_tri_cat + 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_tri_cat  2   6578597 82280914 2470.0
## - ui            1   5839544 81541861 2470.2
##
## Step:  AIC=2456.3
## bwt ~ age + lwt + race_tri_cat + 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_tri_cat  2   6552496 82293521 2468.0
## - ui            1   5817995 81559020 2468.3
##
## Step:  AIC=2454.5
## bwt ~ lwt + race_tri_cat + 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_tri_cat  2   6571668 82391807 2466.2
## - ui            1   5751122 81571261 2466.3
##
## Step:  AIC=2452.79
## bwt ~ lwt + race_tri_cat + 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_tri_cat  2   6630123 82567628 2464.6
## - ui            1   6353218 82290723 2466.0
```

```
##
## Call:
## lm(formula = bwt ~ lwt + race_tri_cat + smoke + ht + ui, data = data)
##
## Coefficients:
##       (Intercept)                lwt  race_tri_catBlack  race_tri_catOther
##          2837.264              4.242           -475.058           -348.150
##            smoke1                ht1                ui1
##          -356.321           -585.193           -525.524
```

The selected Regression is stated as follows.

$bwt = 837.264 + 4.242lwt - 475.058I_1 - 348.150I_2 - 356.321smoke - 585.193ht - 525.524ui$

In the above equation, I_1 and I_2 represent the indicators for if the subject was of the race black or other respectivly.

2-A

The model selected based on forward selection utilized the following variables as predictors.

- discount
- promo
- price

2-B

The algorithm for forward selection can be broken down into many steps. At the very beginning (the base case), the model starts with 0 predictors, which means none of the variables are used in the prediction model. From this base case the algorithm kicks off.

Step 1: Select one predictor to utilize in the model. In this case it is adding one variable to the base case.

Step 2: Calculate the AIC for the model fit. If the AIC is smaller than the current AIC and the smallest AIC, the predictor is added.

Step 3: Continue this process (Step 1 and Step 2) until no smaller AIC is found or number of predictors run out.

2-C

Before defaulting to use the model outputted it is critical to check the assumptions of linear regerssion with tools such as Residual Plot, ACF Plot, and QQ plot. Also, a good sanity check is to see if the predictors selected actually makes sense. Does the study being conducted and the equation align well? Is this equation useful? These are the points / advice I would give to the client.

3

An advantage of adjusted R^2 over R^2 is that adjusted R^2 is being resistent to adding not useful parameters in the model. Adding a parameter to a model, even though that paramter is useless, will increase the R^2 value. If the added parameter is useless than the adjusted R^2 will catch this nature and decrease which would lead to a simplistic model. Adjust R^2 is good to find the regression with good fit and simplicity.

One advantage of R^2 is the interpretation being easy to understand for a given model. R^2 measures the proportion of variance caused by the model. The adjusted R^2 cannot output this information.

4

The function our group wrote to compute the PRESS Satistic form the guided question set.

```
press.computation <- function(model) {
  linear.model <- model
  influence <- lm.influence(linear.model)
  denom  = sapply(influence$hat, function(x) 1 - x)
  division = influence$wt.res / denom
  squared = division^2
  stat = sum(squared)
  return(stat)
}
```