Mod 09 HW

Matt Scheffel

2022-11-07

```
library(MASS)
data(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
```

1A

The categorical variables in this data set are "low", "race", "smoke", "ht," & "ui".

```
birthwt$low = factor(birthwt$low)
birthwt$race = factor(ifelse(birthwt$race == 1, "white",
                             ifelse(birthwt$race == 2, "black", "other")))
birthwt$smoke = factor(birthwt$smoke)
birthwt$ht = factor(birthwt$ht)
birthwt$ui = factor(birthwt$ui)
```

1B

Yes, "low" should be dropped as a predictor. Since we are trying to predict the birth weight of the babies, having a variable which already tells us that the baby's birth weight is low does not make sense.

1C

```
#install.packages("leaps")
library(leaps)
allreg <- regsubsets(bwt ~., data=birthwt, nbest=2)
summary(allreg)
```

```
## Subset selection object
## Call: regsubsets.formula(bwt ~ ., data = birthwt, nbest = 2)
## 10 Variables  (and intercept)
##           Forced in Forced out
## low1          FALSE      FALSE
```

```
## age           FALSE      FALSE
## lwt           FALSE      FALSE
## raceother     FALSE      FALSE
## racewhite     FALSE      FALSE
## smoke1        FALSE      FALSE
## ptl           FALSE      FALSE
## ht1           FALSE      FALSE
## ui1           FALSE      FALSE
## ftv           FALSE      FALSE
## 2 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          low1 age lwt raceother racewhite smoke1 ptl ht1 ui1 ftv
## 1  ( 1 ) "*"  " " " " " " "       " "       " "    " " " " " " " "
## 1  ( 2 ) " "  " " " " " " "       " "       " "    " " " " "*" " "
## 2  ( 1 ) "*"  " " " " " " "       " "       " "    " " " " "*" " "
## 2  ( 2 ) "*"  " " " " " " "       "*"       " "    " " " " " " " "
## 3  ( 1 ) "*"  " " " " " " "       "*"       " "    " " " " "*" " "
## 3  ( 2 ) "*"  " " " " " " "       " "       "*"    " " " " "*" " "
## 4  ( 1 ) "*"  " " " " " " "       "*"       "*"    " " " " "*" " "
## 4  ( 2 ) "*"  " " " " " " "       "*"       " "    " " "*" "*" " "
## 5  ( 1 ) "*"  " " " " " " "       "*"       "*"    " " "*" "*" " "
## 5  ( 2 ) "*"  " " " " " " "       "*"       "*"    "*" " " "*" " "
## 6  ( 1 ) "*"  "*" " " " " "       "*"       "*"    "*" " " "*" " "
## 6  ( 2 ) "*"  " " " " " " "       "*"       "*"    "*" "*" "*" " "
## 7  ( 1 ) "*"  "*" " " " " "       "*"       "*"    "*" "*" "*" " "
## 7  ( 2 ) "*"  "*" "*" " " "       "*"       "*"    " " "*" "*" " "
## 8  ( 1 ) "*"  "*" "*" " " "       "*"       "*"    "*" "*" "*" " "
## 8  ( 2 ) "*"  "*" "*" "*" "       "*"       "*"    " " "*" "*" " "
```

i.

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

```
##  (Intercept)         low1          age          lwt    racewhite       smoke1
##  3311.658408 -1120.502431    -7.726118     1.323854   207.499857  -177.562773
##          ptl          ht1          ui1
##    82.182693  -178.819660  -338.917971
```

Predictors: low1, age, lwt, racewhite

ii.

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

```
## (Intercept)        low1   racewhite      smoke1         ui1
##   3306.5934  -1131.3271    194.9801   -158.8879   -308.2584
```

Predictors: low1, racewhite, smoke1, ui1

iii.

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

```
## (Intercept)        low1         ui1
##   3363.4380  -1190.4560   -318.7814
```

Predictors: low1, ui1

For all of these, we could have dropped the "low" predictor as discussed earlier and it would not be counted for each model. The following code could be used:

birthwt = subset(birthwt, select = -c(low))

1D

```
regnull <- lm(bwt~1, data=birthwt)

regfull <- lm(bwt~., data=birthwt)
step(regnull, scope=list(lower=regnull, upper=regfull), direction="forward")
```

```
## Start:  AIC=2492.76
## bwt ~ 1
##
##         Df Sum of Sq      RSS    AIC
## + low    1  61573224 38396432 2313.9
## + ui     1   8059031 91910625 2478.9
## + race   2   5015725 94953931 2487.0
## + smoke  1   3625946 96343710 2487.8
## + lwt    1   3448639 96521017 2488.1
## + ptl    1   2391041 97578614 2490.2
## + ht     1   2130425 97839231 2490.7
## <none>               99969656 2492.8
## + age    1    815483 99154173 2493.2
## + ftv    1    339993 99629663 2494.1
##
## Step:  AIC=2313.91
## bwt ~ low
##
##         Df Sum of Sq      RSS    AIC
## + ui     1   2354601 36041831 2303.9
## + race   2    957345 37439086 2313.1
## + smoke  1    417505 37978926 2313.8
## <none>               38396432 2313.9
## + lwt    1    284886 38111546 2314.5
## + ht     1     71336 38325096 2315.6
## + ftv    1      7961 38388470 2315.9
## + age    1       929 38395503 2315.9
## + ptl    1        61 38396371 2315.9
##
## Step:  AIC=2303.95
## bwt ~ low + ui
##
##         Df Sum of Sq      RSS    AIC
## + race   2    959016 35082815 2302.8
## <none>               36041831 2303.9
## + smoke  1    349884 35691947 2304.1
## + ht     1    233610 35808222 2304.7
## + lwt    1    115944 35925887 2305.3
```

```
## + ptl    1     94378 35947453 2305.4
## + age    1     13684 36028147 2305.9
## + ftv    1       169 36041662 2305.9
##
## Step:  AIC=2302.85
## bwt ~ low + ui + race
##
##         Df Sum of Sq      RSS    AIC
## + smoke  1    934437 34148378 2299.8
## <none>               35082815 2302.8
## + ht     1    200349 34882466 2303.8
## + lwt    1    135291 34947524 2304.1
## + age    1     97756 34985059 2304.3
## + ptl    1     69302 35013513 2304.5
## + ftv    1      2897 35079918 2304.8
##
## Step:  AIC=2299.75
## bwt ~ low + ui + race + smoke
##
##        Df Sum of Sq      RSS    AIC
## <none>              34148378 2299.8
## + ht    1    205506 33942872 2300.6
## + ptl   1    178354 33970024 2300.8
## + age   1    161097 33987281 2300.8
## + lwt   1     88189 34060188 2301.3
## + ftv   1     11185 34137193 2301.7
```

```
##
## Call:
## lm(formula = bwt ~ low + ui + race + smoke, data = birthwt)
##
## Coefficients:
## (Intercept)         low1          ui1    raceother    racewhite       smoke1
##     3294.45     -1131.17      -309.28        16.48       206.36      -157.30
```

```
step(regfull, scope=list(lower=regnull, upper=regfull), direction="backward")
```

```
## Start:  AIC=2304.73
## bwt ~ low + age + lwt + race + smoke + ptl + ht + ui + ftv
##
##         Df Sum of Sq      RSS    AIC
## - ftv    1      9552 33263661 2302.8
## - ptl    1    263368 33517477 2304.2
## - age    1    270851 33524960 2304.3
## - lwt    1    308633 33562742 2304.5
## - ht     1    332675 33586785 2304.6
## <none>               33254109 2304.7
## - smoke  1   1088898 34343008 2308.8
## - race   2   1706722 34960832 2310.2
## - ui     1   2509469 35763578 2316.5
## - low    1  42448208 75702317 2458.2
##
## Step:  AIC=2302.79
## bwt ~ low + age + lwt + race + smoke + ptl + ht + ui
##
```

```
##         Df Sum of Sq      RSS    AIC
## - ptl    1    267027 33530689 2302.3
## - age    1    299213 33562875 2302.5
## - lwt    1    300805 33564466 2302.5
## - ht     1    324794 33588455 2302.6
## <none>               33263661 2302.8
## - smoke  1   1084502 34348163 2306.8
## - race   2   1699490 34963151 2308.2
## - ui     1   2502617 35766278 2314.5
## - low    1  42477363 75741025 2456.3
##
## Step:  AIC=2302.3
## bwt ~ low + age + lwt + race + smoke + ht + ui
##
##         Df Sum of Sq      RSS    AIC
## - age    1    231737 33762425 2301.6
## - lwt    1    251768 33782456 2301.7
## - ht     1    319904 33850593 2302.1
## <none>               33530689 2302.3
## - smoke  1    939840 34470529 2305.5
## - race   2   1653607 35184295 2307.4
## - ui     1   2282970 35813659 2312.8
## - low    1  42301896 75832585 2454.5
##
## Step:  AIC=2301.6
## bwt ~ low + lwt + race + smoke + ht + ui
##
##         Df Sum of Sq      RSS    AIC
## - lwt    1    180447 33942872 2300.6
## - ht     1    297763 34060188 2301.3
## <none>               33762425 2301.6
## - smoke  1    873670 34636095 2304.4
## - race   2   1467913 35230338 2305.6
## - ui     1   2228338 35990763 2311.7
## - low    1  42175080 75937505 2452.8
##
## Step:  AIC=2300.61
## bwt ~ low + race + smoke + ht + ui
##
##         Df Sum of Sq      RSS    AIC
## - ht     1    205506 34148378 2299.8
## <none>               33942872 2300.6
## - smoke  1    939594 34882466 2303.8
## - race   2   1511671 35454543 2304.8
## - ui     1   2344367 36287239 2311.2
## - low    1  44668862 78611734 2457.3
##
## Step:  AIC=2299.75
## bwt ~ low + race + smoke + ui
##
##         Df Sum of Sq      RSS    AIC
## <none>               34148378 2299.8
## - smoke  1    934437 35082815 2302.8
## - race   2   1543569 35691947 2304.1
```

```
## - ui     1   2201541 36349919 2309.6
## - low    1  46921303 81069681 2461.2
```

```
##
## Call:
## lm(formula = bwt ~ low + race + smoke + ui, data = birthwt)
##
## Coefficients:
## (Intercept)         low1    raceother    racewhite       smoke1          ui1
##     3294.45     -1131.17        16.48       206.36      -157.30      -309.28
```

The regression equation selected using backwards selection is the last one, with the lowest AIC of 2299.75:

Step: AIC=2299.75 bwt ~ low + race + smoke + ui

4

PRESS Statistic:

```
press <- function(regmodel) {
sum((regmodel$residuals) / (1-lm.influence(regmodel)$hat))**2
}
```

2) a) Following the AIC method, the model selected based on forward selection is the model with the lowest AIC score. Using this criteria, the model selected is the 4th model with an AIC score of -132.94:

$$Share \sim discount + promo + price$$

b) Output explanation:

The stepwise model begins with the "start"; the intercept and no variables. After the first model (with just the intercept), we continue adding variables to the model in an effort to reduce the AIC score.

1) $Share = \beta_0$ w/ AIC = -94.8

Next, we add discount as a variable:

2) $Share = \beta_0 + \beta_1(discount)$ w/ AIC = -126.14

Next, we add promo as a variable!

3) $Share = \beta_0 + \beta_1(discount) + \beta_2(promo)$ w/ AIC = -129.69

Finally, we add price as a variable!

4) $Share = \beta_0 + \beta_1(discount) + \beta_2(promo) + \beta_3(price)$ w/ AIC = -132.94

We do not add other variables (such as "time", "nielsen") to the model as they result in higher AIC scores.

c) I would say the models in part 2a are a reasonable choice, but that the client should clarify more details about the model before making a decision. It would be a good idea to run a regression and check the model's p-values (significance and that the coefficients make sense (are they positive for a discount, etc.)

3) R^2 vs. adjusted R^2:

R^2 is a measure of fit that indicates how much variation of a dependent variable is explained by the independent variables in a regression model. The adjusted R^2 adjusts the statistics based on how many independent variables are added to the model.

R^2 over adjusted R^2:

R^2 gives a direct interpretation of the proportion of variance in dependent variables in reference to the independent variables. the adjusted R^2 does not give this interpretation.

Adjusted R^2 over R^2

Adjusted R^2 gives a more precise look at the correlation in the model since it takes into account precisely how many independent variables are added into the model, whereas R^2 does not.

4) ~~*~~ See Q4 on Rmd file ~~*~~