uwa6xv_M09_HW

Alanna Hazlett

2024-03-30

Problem 1

Use birthwt from MASS. The goal of the data set is to relate the birthweight of newborns with the characteristics of their mothers during pregnancy.

(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.

```
Data<-birthwt
?birthwt
#Checked all by using contrasts()
Data$race<-factor(Data$race)
#Don't need to factor these as they are already dummy coded
#Data$low<-factor(Data$low)
#Data$smoke<-factor(Data$smoke)
#Data$ht<-factor(Data$ht)
#Data$ui<-factor(Data$ui)
```

These are categorical: low, race, smoke, ht, ui.
(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.
We can remove this variable as it is the "indicator of birth weight less than 2.5 kg" which it just a bucketed/indicator version of our response variable, birthweight. Birthweight of the baby can not predict the birthweight of baby.
(c)
Based on your answer to part 1b, perform all possible regressions using the regsubsets() function from the leaps package (use nbest=1). Write down the predictors that lead to a first-order model having the best
i. adjusted R2
ii. Mallow's Cp
iii. BIC

```
#removing low from dataset, so our regression models don't include it
Data2<-Data[,2:10]
allreg <- leaps::regsubsets(bwt ~., data=Data2, nbest=1)
summary(allreg)
```

```
## Subset selection object
## Call: regsubsets.formula(bwt ~ ., data = Data2, 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 ) "*" "*" "*"   "*"   "*"   "*" "*" "*" " "
```

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

```
## [1] 6
```

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

```
## [1] 6
```

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

```
## [1] 6
```

All three criteria show the same model as the best option: Model 6, which is comprised of lwt, race2, race3, smoke, ht, and ui.

(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=Data2)
regfull<-lm(bwt~.,data=Data2)
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 = Data2)
##
## Coefficients:
## (Intercept)          lwt        race2        race3        smoke           ht
##    2837.264        4.242     -475.058     -348.150     -356.321     -585.193
##          ui
##    -525.524
```

The estimated candidate model regression equation is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 I_2 + \beta_4 x_3 + \beta_5 x_4 + \beta_6 x_5$$

$$y = 2837.264 + 4.242 x_{lwt} + -457.058 I_{race2} + -348.150 I_{race3} + -356.321 x_{smoke} + -585.193 x_{ht} + -525.524 x_{ui}$$

Problem 2

(a)
What is the model selected based on forward selection?
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.
We utilize the value of a term called AIC to inform our decision on choosing a model that best represents the relationship of the items that influence market share sales. We start by considering no items influencing market share sales, we then want to add one item at a time. Each time we do this the value of AIC changes, it decreases up until a certain point. At this point adding another item to our model would increase the value of AIC, indicating that it would perform worse than the model shown in the last step. When this occurs we have found a potential efficient model to represent the relationship of the items that influence market share sales. Based on our output from forward selection our model would be Share ~ discount + promo + price.
(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?
This is one potential model that we could utilize. We do have other criteria that also help us inform our decision on model selection. It would be wise to see what we conclude from those criteria as well. We need to assess if this model has the potential to solve our research question. If we did decide that this would be the model we would like to look at, we would need to check the regression assumptions to make sure that they are met prior to using the model.

Problem 3

- The advantage of R2 adjusted is that it takes into account the addition of unnecessary variables in the model. It penalizes for it, whereas R2 will only increase.

- One advantage to R2 is that it is easier to calculate using the sum of squares values. Another advantage is that R2 can be used to compare models with the same number of parameters.

Problem 4

```
press<-function(model) {
  pr<-resid(model)/(1-lm.influence(model)$hat)
  press<-sum(pr^2)
  return (press)
}
```