Stat 6021: Homework Set 9

Tom Lever

11/03/22

1. You will continue to use the `birthwt` data set from the `MASS` package for this question. The data were collected at Baystate Medical Center, Springfield, MA in 1986. The data contain information regarding weights of newborn babies as well as potential predictors. Before proceeding, be sure to read the documentation about the data set by typing `?birthwt`. The birthweight of newborns may be related to characteristics of their mothers during pregnancy.

   (a) Which of these variables is 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.

   The following predictors are discrete and categorical:

   - *low* (0 indicates newborn birthweight is less than 2.5 kg, 1 indicates newborn birthweight is greater than or equal to 2.5 kg),
   - *race* (1 indicates white, 2 indicates black, 3 indicates other),
   - *smoke* (0 indicates non-smoking, 1 indicates smoking),
   - *ptl* (value represents number of previous premature labors in $\{0, 1, 2, 3\}$),
   - *ht* (0 indicates no history of hypertension, 1 indicates history of hypertension),
   - *ui* (0 indicates no presence of uterine irritability, 1 indicates presence of uterine irritability), and
   - *ftv* (value represents number of physician visits during the first trimester in $\{0, 1, 2, 3, 4, 6\}$)

   On loading the `MASS` package and the `birthwt` data frame, R interprets the columns corresponding to these variables as vectors of integers.

```
library(MASS)
library(TomLeversRPackage)
birthwt$low <-
    convert_to_categorical_vector(birthwt$low, c("N", "Y"))
birthwt$race <-
    convert_to_categorical_vector(birthwt$race, c("white", "black", "other"))
birthwt$smoke <-
    convert_to_categorical_vector(birthwt$smoke, c("N", "Y"))
birthwt$ptl <-
    convert_to_categorical_vector(birthwt$ptl, unique(birthwt$ptl))
birthwt$ht <-
    convert_to_categorical_vector(birthwt$ht, c("N", "Y"))
birthwt$ui <-
    convert_to_categorical_vector(birthwt$ui, c("N", "Y"))
birthwt$ftv <-
    convert_to_categorical_vector(birthwt$ftv, unique(birthwt$ftv))
head(birthwt, n = 3)
```

```
##    low age lwt  race smoke ptl ht ui ftv  bwt
## 85   N  19 182 black     N   0  N  Y   0 2523
## 86   N  33 155 other     N   0  N  N   2 2551
```

```
## 87    N  20 105 white     Y   0  N  N   3 2557
```

(b) A classmate 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.

I agree. The predictor *low* is dependent on the response / birth weight *bwt*.

```
library(dplyr)
birthwt <- birthwt %>% select(-low)
head(birthwt, n = 3)
```

```
##    age lwt  race smoke ptl ht ui ftv  bwt
## 85  19 182 black     N   0  N  Y   0 2523
## 86  33 155 other     N   0  N  N   2 2551
## 87  20 105 white     Y   0  N  N   3 2557
```

(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

i. adjusted R^2,

```
library(leaps)
subset_selection_object <- regsubsets(
    bwt ~ .,
    data = birthwt,
    nbest = 2,
    really.big = TRUE
)
summary_for_subset_selection_object <- summary(subset_selection_object)
adjusted_R2 <- summary_for_subset_selection_object$adjr2
index_of_model_with_maximum_adjusted_R2 <- which.max(adjusted_R2)
coefficients <- coef(
    subset_selection_object, index_of_model_with_maximum_adjusted_R2
)
predictors <- names(coefficients[2:length(coefficients)])
predictors
```

```
## [1] "lwt"       "raceblack" "raceother" "smokeY"    "ptl1"      "ptl3"
## [7] "htY"       "uiY"
```

ii. Mallow's C_p, and

```
Cp <- summary_for_subset_selection_object$cp
index_of_model_with_minimum_Cp <- which.min(Cp)
coefficients <- coef(subset_selection_object, index_of_model_with_minimum_Cp)
predictors <- names(coefficients[2:length(coefficients)])
predictors
```

```
## [1] "lwt"       "raceblack" "raceother" "smokeY"    "ptl1"      "ptl3"
## [7] "htY"       "uiY"
```

iii. Schwartz Bayesian Information Criterion ($BIC_{Schwartz}$)

```
BICSchwartz <- summary_for_subset_selection_object$bic
index_of_model_with_minimum_BICSchwartz <- which.min(BICSchwartz)
coefficients <- coef(
    subset_selection_object, index_of_model_with_minimum_BICSchwartz
)
```

```r
predictors <- names(coefficients[2:length(coefficients)])
predictors
```

```
## [1] "lwt"       "raceblack" "raceother" "smokeY"    "htY"       "uiY"
```

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

```r
intercept_only_model <- lm(bwt ~ 1, data = birthwt)
full_model <- lm(bwt ~ ., data = birthwt)
step(
    full_model,
    scope = list(lower = intercept_only_model, upper = full_model),
    direction = "backward"
)
```

```
## Start:  AIC=2457.87
## bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv
## 
##         Df Sum of Sq      RSS    AIC
## - ftv    5   1560988 72480820 2452.0
## - age    1     38831 70958663 2456.0
## <none>               70919832 2457.9
## - lwt    1   2336000 73255832 2462.0
## - ptl    3   4012319 74932152 2462.3
## - smoke  1   2640100 73559932 2462.8
## - ht     1   3066897 73986729 2463.9
## - race   2   4574573 75494405 2465.7
## - ui     1   5867573 76787405 2470.9
## 
## Step:  AIC=2451.99
## bwt ~ age + lwt + race + smoke + ptl + ht + ui
## 
##         Df Sum of Sq      RSS    AIC
## - age    1     22593 72503413 2450.1
## <none>               72480820 2452.0
## - ptl    3   3351765 75832585 2454.5
## - lwt    1   2721251 75202072 2457.0
## - ht     1   3461344 75942164 2458.8
## - smoke  1   4107465 76588285 2460.4
## - race   2   5345011 77825831 2461.4
## - ui     1   6457308 78938128 2466.1
## 
## Step:  AIC=2450.05
## bwt ~ lwt + race + smoke + ptl + ht + ui
## 
##         Df Sum of Sq      RSS    AIC
## <none>               72503413 2450.1
## - ptl    3   3434092 75937505 2452.8
## - lwt    1   2728077 75231490 2455.0
## - ht     1   3440618 75944031 2456.8
## - smoke  1   4092457 76595870 2458.4
## - race   2   5499752 78003165 2459.9
## - ui     1   6435422 78938835 2464.1
```

```
##
## Call:
## lm(formula = bwt ~ lwt + race + smoke + ptl + ht + ui, data = birthwt)
##
## Coefficients:
## (Intercept)          lwt    raceblack    raceother       smokeY         ptl1
##    2834.324        4.306     -445.614     -310.801     -330.181     -294.426
##        ptl2         ptl3          htY          uiY
##     -15.485     1266.335     -573.974     -542.511
```

```
best_model <- lm(bwt ~ lwt + race + smoke + ptl + ht + ui, data = birthwt)
summarize_linear_model(best_model)
```

```
##
## Call:
## lm(formula = bwt ~ lwt + race + smoke + ptl + ht + ui, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1830.54  -441.19    44.76   482.39  1626.09
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2834.324    241.596  11.732  < 2e-16 ***
## lwt            4.306      1.659   2.595  0.01024 *
## raceblack   -445.614    144.003  -3.094  0.00229 **
## raceother   -310.801    111.480  -2.788  0.00588 **
## smokeY      -330.181    103.875  -3.179  0.00174 **
## ptl1        -294.426    143.421  -2.053  0.04154 *
## ptl2         -15.485    293.639  -0.053  0.95800
## ptl3        1266.335    654.582   1.935  0.05462 .
## htY         -573.974    196.937  -2.915  0.00402 **
## uiY         -542.511    136.105  -3.986 9.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 636.4 on 179 degrees of freedom
## Multiple R-squared:  0.2747, Adjusted R-squared:  0.2383
## F-statistic: 7.534 on 9 and 179 DF,  p-value: 2.505e-09
##
## E(y | x) =
##     B_0 +
##     B_lwt * lwt +
##     B_raceblack * raceblack +
##     B_raceother * raceother +
##     B_smokeY * smokeY +
##     B_ptl1 * ptl1 +
##     B_ptl2 * ptl2 +
##     B_ptl3 * ptl3 +
##     B_htY * htY +
##     B_uiY * uiY
## E(y | x) =
##     2834.32356097846 +
##     4.30561293695914 * lwt +
##     -445.614100326541 * raceblack +
```

```
##     -310.801223670066 * raceother +
##     -330.181304321706 * smokeY +
##     -294.425985626396 * ptl1 +
##     -15.4853803035702 * ptl2 +
##     1266.33524702617 * ptl3 +
##     -573.973791172762 * htY +
##     -542.510732694048 * uiY
## Number of observations: 189
## Estimated variance of errors: 405046.999482501
## Prediction R2: -Inf
## Multiple R:  0.524161996342691   Adjusted R:  0.488139839961811
## Critical value t(alpha/2 = 0.05/2, DFRes = 179): 1.97330543384147
## Critical value F(alpha = 0.05, DFR = 9, DFRes = 179): 1.93249997278723
```

Let $\boldsymbol{\beta}_{predictor}$ be a column vector of the coefficients of the non-reference indicator variables associated with predictor *predictor*. Let $\boldsymbol{predictor}$ be a column vector of the non-reference indicator variables associated with predictor *predictor*. The MLR equation selected is

$$bwt = \beta_0 + \beta_{lwt}\ lwt + \boldsymbol{\beta}_{race} \cdot \boldsymbol{race} + \boldsymbol{\beta}_{smoke} \cdot \boldsymbol{smoke} + \boldsymbol{\beta}_{ptl} \cdot \boldsymbol{ptl} + \boldsymbol{\beta}_{ht} \cdot \boldsymbol{ht} + \boldsymbol{\beta}_{ui} \cdot \boldsymbol{ui}$$