HW3

Takao

10/26/2022

Takao Oba

Stats 101c

HW 3

Download Wine data from Bruinlearn week 4. Class is the response variable for this data

Q1

Split the data into 70% training and 30% testing using your birthday as a seed.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

```
## Warning: package 'purrr' was built under R version 4.1.2
## Warning: package ’dplyr’ was built under R version 4.1.2
```

```
## Warning: package ’stringr’ was built under R version 4.1.2
```

```
## Warning: package ’forcats’ was built under R version 4.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(MASS)
```

```
## Warning: package ’MASS’ was built under R version 4.1.2
```

```
##
## Attaching package: ’MASS’
##
## The following object is masked from ’package:dplyr’:
##
##     select
```

```
wine <- read_csv("/Users/takaooba/Downloads/Wine Fall 2021.csv")
```

```
## New names:
## Rows: 10000 Columns: 14
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): Wine.Color, Class dbl (12): ...1, fixed.acidity, volatile.acidity,
## citric.acid, residual.sugar...
## i Use ‘spec()‘ to retrieve the full column specification for this data. i
## Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
## * ‘‘ -> ‘...1‘
```

```
wine <- wine[, -1]
wine$Wine.Color <- as.factor(wine$Wine.Color)
head(wine)
```

```
## # A tibble: 6 x 13
##   Wine.C~1 fixed~2 volat~3 citri~4 resid~5 chlor~6 free.~7 total~8 density    pH
##   <fct>      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>
## 1 W            5.7    0.26    0.25    10.4   0.02        7      57   0.994  3.39
## 2 W            7.5    0.17    0.71    11.8   0.038      52     148   0.998  3.03
## 3 W            6.7    0.24    0.29    14.9   0.053      55     136   0.998  3.03
## 4 W            7.7    0.27    0.49     3.8   0.037      46     139   0.991  3.04
## 5 W            6.7    0.23    0.33     8.1   0.048      45     176   0.995  3.11
## 6 W            6.7    0.21    0.34     1.4   0.049      36     112   0.991  3.02
## # ... with 3 more variables: sulphates <dbl>, alcohol <dbl>, Class <chr>, and
## #   abbreviated variable names 1: Wine.Color, 2: fixed.acidity,
## #   3: volatile.acidity, 4: citric.acid, 5: residual.sugar, 6: chlorides,
## #   7: free.sulfur.dioxide, 8: total.sulfur.dioxide
```

(a)

Create a logistic regression using all predictors with a classification threshold of 0.5. Report your confusion matrices and error rates.

```r
set.seed(717)
dim(wine)
```

```
## [1] 10000    13
```

```r
# We will have that 7000 instances are training
# We will have that 3000 instances are testing

test.i <- sample(1:10000, 3000, replace = F)

# wine$Wine.Color <- ifelse(wine$Wine.Color == "W", 1, 0)
# wine$Wine.Color <- as.numeric(wine$Wine.Color)

wine.test <- wine[test.i,]
wine.train <- wine[-test.i,]


wine.model <- glm(as.factor(Class) ~., data = wine.train, family = "binomial")
wine.model
```

```
## 
## Call:  glm(formula = as.factor(Class) ~ ., family = "binomial", data = wine.train)
## 
## Coefficients:
##          (Intercept)           Wine.ColorW         fixed.acidity  
##            2.251e+02            -6.084e-01             2.144e-01  
##     volatile.acidity           citric.acid        residual.sugar  
##           -1.725e+00             2.901e-01             1.087e-01  
##            chlorides   free.sulfur.dioxide  total.sulfur.dioxide  
##           -3.500e+00             7.711e-03            -3.080e-03  
##              density                    pH             sulphates  
##           -2.332e+02             1.084e+00             1.293e+00  
##              alcohol  
##            1.762e-01  
## 
## Degrees of Freedom: 6999 Total (i.e. Null);  6987 Residual
## Null Deviance:       9704 
## Residual Deviance: 8988  AIC: 9014
```

```r
wine.pred <- predict(wine.model, wine.test[,1:12], type = "response")
wine.glm.pred <- rep("Bad", length(wine.pred))
wine.glm.pred[wine.pred >= 0.5] <- "Good"

wine.test.y <- wine.test$Class
```

```r
# testing
table(wine.glm.pred, wine.test.y)
```

```
##              wine.test.y
## wine.glm.pred  Bad Good
##          Bad  1021  563
##          Good  534  882
```

```r
mean(wine.glm.pred != wine.test.y)
```

```
## [1] 0.3656667
```

```r
# training
wine.pred <- predict(wine.model, wine.train[,1:12])
wine.glm.pred <- rep("Bad", length(wine.pred))
wine.glm.pred[wine.pred >= 0.5] <- "Good"
wine.train.y <- wine.train$Class

table(wine.glm.pred, wine.train.y)
```

```
##              wine.train.y
## wine.glm.pred  Bad Good
##          Bad  2993 2225
##          Good  518 1264
```

```r
mean(wine.glm.pred != wine.train.y)
```

```
## [1] 0.3918571
```

(b)

Create a LDA model using all predictors. Report your confusion matrices and error rates.

```r
set.seed(717)
model2 <- lda(Class ~ ., data = wine.train)

# Testing
testwine <- predict(model2, wine.test[1:12], type = "response")
table(testwine$class, wine.test$Class)
```

```
##
##         Bad Good
##   Bad  1028  576
##   Good  527  869
```

```
mean(testwine$class != wine.test$Class)
```

```
## [1] 0.3676667
```

```
# Training
trainwine <- predict(model2, wine.train[1:12], type = "response")
table(trainwine$class, wine.train$Class)
```

```
##
##         Bad Good
##   Bad  2331 1314
##   Good 1180 2175
```

```
mean(trainwine$class != wine.train$Class)
```

```
## [1] 0.3562857
```

(c)

Create a QDA model using all predictors. Report your confusion matrices and error rates.

```
set.seed(717)
wine.qda <- qda(Class ~ ., data = wine.train, method = "mle")

# Testing
testwine <- predict(wine.qda, wine.test[1:12], type = "response")
table(testwine$class, wine.test$Class)
```

```
##
##         Bad Good
##   Bad   807  375
##   Good  748 1070
```

```
mean(testwine$class != wine.test$Class)
```

```
## [1] 0.3743333
```

```
# Training
trainwine <- predict(wine.qda, wine.train[1:12], type = "response")
table(trainwine$class, wine.train$Class)
```

```
##
##         Bad Good
##   Bad  1798  866
##   Good 1713 2623
```

```r
mean(trainwine$class != wine.train$Class)
```

```
## [1] 0.3684286
```

(d)

Create a KNN model with k = 25 (Use numerical predictors only after scaling them)

```r
set.seed(717)
# We will want to extract the numerical predictors after scaling them
wine.num <- as.data.frame(scale(wine[,2:12]))

W.X.test <- wine.num[test.i,]

W.X.train <- wine.num[-test.i,]

W.Y.test <- wine$Class[test.i]
W.Y.train <- wine$Class[-test.i]

library(class)
```

```
## Warning: package 'class' was built under R version 4.1.2
```

```r
# Testing
wine.knn <- knn(W.X.train, W.X.test, W.Y.train, k = 25)
table(wine.knn, W.Y.test)
```

```
##         W.Y.test
## wine.knn  Bad Good
##     Bad  1020  495
##     Good  535  950
```

```r
mean(wine.knn != W.Y.test)
```

```
## [1] 0.3433333
```

```r
# Training
wine.knn <- knn(W.X.train, W.X.train, W.Y.train, k = 25)
table(wine.knn, W.Y.train)
```

```
##         W.Y.train
## wine.knn  Bad Good
##     Bad  2406 1085
##     Good 1105 2404
```

```
mean(wine.knn != W.Y.train)
```

```
## [1] 0.3128571
```

(e)

Compare and contrast between the models created parts A-D.

We will now compare and contrast the models generated from parts A-D. The error rates are as follows:

Logistic Regression Test: 0.3656667

Logistic Regression Train: 0.3918571

LDA Test: 0.3676667

LDA Train: 0.3562857

QDA Test: 0.3743333

QDA Train:0.3684286

KNN test: 0.3433333

KNN train: 0.3128571

Based on these error rates, the KNN is the best model, then LDA, then QDA, then logistic regression is the worst model. Overall, the error rates are relatively high in their 30s and 40s percent.

Q2

Use the full Wine data to: Use the LOOCV method and create the following:

(a)

Logistic regression. Report your confusion matrices and error rates.

```
set.seed(717)
# Logistic Regression

library(boot)
lr.model <- glm(factor(Class) ~ ., data = wine, family = binomial())
lr.model
```

```
##
## Call:  glm(formula = factor(Class) ~ ., family = binomial(), data = wine)
##
## Coefficients:
##          (Intercept)            Wine.ColorW          fixed.acidity
##            2.378e+02             -5.714e-01              2.462e-01
##     volatile.acidity            citric.acid         residual.sugar
##            -1.639e+00            1.816e-01            1.212e-01
##             chlorides   free.sulfur.dioxide  total.sulfur.dioxide
##            -2.571e+00            9.006e-03           -3.371e-03
##               density                   pH            sulphates
##            -2.475e+02            1.461e+00            1.120e+00
##               alcohol
##             1.715e-01
##
## Degrees of Freedom: 9999 Total (i.e. Null);  9987 Residual
## Null Deviance:       13860
## Residual Deviance: 12850     AIC: 12880
```

```
# cv.error <- cv.glm(wine, lr.model)$delta
# cv.error$K
# cv.error$delta
```

We attempted to run "cv.error <- cv.glm(wine, lr.model)$delta" but notice that the code takes too long to run. We will continue on with the confusion matrix and the error rates.

(b)

LDA. Report your confusion matrix and error rate.

```
set.seed(717)
library(MASS)
lda.LOOCV<- lda(Class~.,wine,CV = TRUE)
summary(lda.LOOCV)
```

```
##           Length Class  Mode
## class     10000  factor numeric
## posterior 20000  -none- numeric
## terms         3  terms  call
## call          4  -none- call
## xlevels       1  -none- list
```

```
# Confusion Matrix
table(lda.LOOCV$class,wine$Class)
```

```
##
##          Bad Good
##   Bad   3362 1892
##   Good  1704 3042
```

```
# Error Rate
mean(lda.LOOCV$class!=wine$Class)
```

```
## [1] 0.3596
```

(c)

QDA. Report your confusion matrix and error rate.

```
set.seed(717)
qda.LOOCV <- qda(Class ~ ., data = wine, CV = TRUE)

t = table(wine$Class, qda.LOOCV$class)
t
```

```
##
##         Bad Good
##   Bad  2576 2490
##   Good 1223 3711
```

```
mean(wine$Class != qda.LOOCV$class)
```

```
## [1] 0.3713
```

(d)

KNN with k = 25. Report your confusion matrix and error rate

```
set.seed(717)
# library(class)
#
# head(wine)
#
# length(wine[,-c(1,13)])
# length(wine[,13])
#
# # Testing
# wine.knn <- knn.cv(wine[,- c(1,13)], wine[,13], k = 25)
#
#
# table(wine.knn, W.Y.test)
# mean(wine.knn != W.Y.test)

X = as.matrix(wine[,-c(1,13)])
Y = as.factor(wine$Class)


knn.pred <- knn.cv(X,Y,k = 25)
length(knn.pred)
```

```
## [1] 10000
```

```r
length(wine$Class)
```

```
## [1] 10000
```

```r
knn.loocv.cm <- table(Predicted = knn.pred, wine$Class, dnn = c("Predicted", "Actual"))
knn.loocv.cm
```

```
##          Actual
## Predicted  Bad Good
##      Bad  3323 1840
##      Good 1743 3094
```

```r
knn.error <- mean(knn.pred != wine$Class)
knn.error
```

```
## [1] 0.3583
```

(e)

Compare and contrast the LOOCV error rates across the created models.

The error rate:

Linear Regression: ? lda: 0.3596 qda: 0.3713 KNN: 0.3583

Note that these are all error rates generated with the set seed of 717 (Takao Oba's birth date). Based on the generated models, we have that the best model is the KNN (a close tie to the lda model) and the worst model is the QDA. We are unsure about the confusion matrix and the error rate as the data takes too long to load, but we assume that the linear regression will be the worst model.

Q3

Use the full Wine data to: Use the CV 10-flod method and create the following:

(a)

Logistic regression. Report your confusion matrices and error rates.

```r
set.seed(717)
# Logistic Regression
lr.model <- glm(factor(Class) ~ ., data = wine, family = binomial())
lr.model
```

```
##
## Call:  glm(formula = factor(Class) ~ ., family = binomial(), data = wine)
##
```

```
## Coefficients:
##          (Intercept)           Wine.ColorW         fixed.acidity
##            2.378e+02            -5.714e-01             2.462e-01
##     volatile.acidity           citric.acid        residual.sugar
##           -1.639e+00             1.816e-01             1.212e-01
##            chlorides   free.sulfur.dioxide  total.sulfur.dioxide
##           -2.571e+00             9.006e-03            -3.371e-03
##              density                    pH             sulphates
##           -2.475e+02             1.461e+00             1.120e+00
##              alcohol
##            1.715e-01
##
## Degrees of Freedom: 9999 Total (i.e. Null);  9987 Residual
## Null Deviance:       13860
## Residual Deviance: 12850     AIC: 12880
```

```
cv.error10 <- cv.glm(wine, lr.model, K = 10)

# The K value
cv.error10$K
```

```
## [1] 10
```

```
# Error Rate
cv.error10$delta
```

```
## [1] 0.2260684 0.2260327
```

```
# install.packages("caret")
# library(caret)
```

(b)

LDA. Report your confusion matrix and error rate.

```
set.seed(717)
predfun.lda = function(train.x, train.y, test.x, test.y, negative){
  require("MASS") # for lda function
  lda.fit = lda(train.x, grouping=train.y)
  ynew = predict(lda.fit, test.x)$class
  # count TP, FP etc.
  out = confusionMatrix(test.y, ynew, negative=negative)
  return(out)
}

dim(wine)
```

```
## [1] 10000    13
```

```
names(wine)
```

```
##  [1] "Wine.Color"           "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"          "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"
## [10] "pH"                   "sulphates"            "alcohol"
## [13] "Class"
```

```
X = as.matrix(wine[,-c(1,13)])
Y = as.factor(wine$Class)
dim(X) # 10000 11
```

```
## [1] 10000    11
```

```
levels(Y) # "Bad", "Good"
```

```
## [1] "Bad"  "Good"
```

```
library(crossval)
l.cv.out <- crossval(predfun.lda, X, Y, K=10, B=1, negative="Bad")
```

```
## Number of folds: 10
## Total number of CV fits: 10
##
## Round # 1 of 1
## CV Fit # 1 of 10
## CV Fit # 2 of 10
## CV Fit # 3 of 10
## CV Fit # 4 of 10
## CV Fit # 5 of 10
## CV Fit # 6 of 10
## CV Fit # 7 of 10
## CV Fit # 8 of 10
## CV Fit # 9 of 10
## CV Fit # 10 of 10
```

```
l.cv.out
```

```
## $stat.cv
##         FP  TP  TN  FN
## B1.F1  176 301 330 192
## B1.F2  174 313 332 180
## B1.F3  181 315 326 179
## B1.F4  173 320 333 174
## B1.F5  175 315 332 179
## B1.F6  162 315 344 178
## B1.F7  186 288 321 206
## B1.F8  172 292 335 201
## B1.F9  153 304 354 189
## B1.F10 165 300 342 193
```

```
##
## $stat
##    FP    TP    TN    FN
## 171.7 306.3 334.9 187.1
##
## $stat.se
##       FP       TP       TN       FN
## 3.011275 3.451409 3.009060 3.413861
```

```
# Computing the various diagnostic errors
diagnosticErrors(l.cv.out$stat)
```

```
##       acc      sens      spec       ppv       npv       lor
## 0.6412000 0.6207945 0.6610738 0.6407950 0.6415709 1.1610050
```

```
# lda.LOOCV<- lda(Class~.,wine,CV = TRUE, k = 10)
# summary(lda.LOOCV)
#
# # Confusion Matrix
# table(lda.LOOCV$class,wine$Class)
#
# # Error Rate
# mean(lda.LOOCV$class!=wine$Class)
```

To find the error rate, we will utilize the accuracy and perform the operation 1 - accuracy. Thus, utilizing set.seed of 717, we have that the accuracy is 0.6412 thus the error rate will be 0.3589.

(c)

QDA. Report your confusion matrix and error rate.

```
set.seed(717)

predfun.qda = function(train.x, train.y, test.x, test.y, negative){
  require("MASS") # for lda function
  qda.fit = qda(train.x, grouping=train.y)
  ynew = predict(qda.fit, test.x)$class
  # count TP, FP etc.
  out = confusionMatrix(test.y, ynew, negative=negative)
  return(out)
}

l.cv.out <- crossval(predfun.qda, X, Y, K=10, B=1, negative="Bad")
```

```
## Number of folds: 10
## Total number of CV fits: 10
##
## Round # 1 of 1
## CV Fit # 1 of 10
## CV Fit # 2 of 10
## CV Fit # 3 of 10
## CV Fit # 4 of 10
## CV Fit # 5 of 10
## CV Fit # 6 of 10
## CV Fit # 7 of 10
## CV Fit # 8 of 10
## CV Fit # 9 of 10
## CV Fit # 10 of 10
```

```
l.cv.out
```

```
## $stat.cv
##          FP  TP  TN  FN
## B1.F1  259 367 247 126
## B1.F2  262 366 244 127
## B1.F3  230 374 277 120
## B1.F4  243 365 263 129
## B1.F5  241 377 266 117
## B1.F6  244 372 262 121
## B1.F7  264 358 243 136
## B1.F8  238 353 269 140
## B1.F9  241 363 266 130
## B1.F10 242 372 265 121
##
## $stat
##    FP    TP    TN    FN
## 246.4 366.7 260.2 126.7
##
## $stat.se
##       FP       TP       TN       FN
## 3.569002 2.347812 3.641733 2.319243
```

```
# Computing the various diagnostic errors
diagnosticErrors(l.cv.out$stat)
```

```
##       acc      sens      spec       ppv       npv       lor
## 0.6269000 0.7432104 0.5136202 0.5981080 0.6725252 1.1172163
```

```
# qda.LOOCV <- qda(Class ~ ., data = wine, CV = TRUE, k = 10)
#
# t = table(wine$Class, qda.LOOCV$class)
# t
# mean(wine$Class != qda.LOOCV$class)
```

Again, we will find the error rate by utilizing the accuracy. The accuracy is 0.6269 and thus the error rate is correspondingly 0.3731.

(d)

KNN with k = 25. Report your confusion matrix and error rate

```r
set.seed(717)
#install.packages("caret")
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.2

## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:boot':
##
##     melanoma

##
## Attaching package: 'caret'

## The following object is masked from 'package:crossval':
##
##     confusionMatrix

## The following object is masked from 'package:purrr':
##
##     lift
```

```r
control <- trainControl(method = "cv", number = 10)



fit <- train(Class ~ ., method = "knn", tuneGrid = expand.grid(k = 10), metric = "Accuracy", data = win

fit
```

```
## k-Nearest Neighbors
##
## 10000 samples
##    12 predictor
##     2 classes: 'Bad', 'Good'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 10000, 10000, 10000, 10000, 10000, 10000, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.6494865  0.2990126
##
## Tuning parameter 'k' was held constant at a value of 10
```

Thus, the error rate will be 1 minus the accuracy. We have that the error rate is 1-0.6494865 = 0.3505135.

(e)

Compare and contrast the 10-fold CV error rates across the created models.

The error rate:

Linear Regression: 0.2260684 lda: 0.3589 qda: 0.3731 KNN: 0.3505135

Note that these are all error rates generated with the set seed of 717 (Takao Oba's birth date). By looking at the error rates, we have that the best model is the linear regression and the worst model is the qda. The lda and KNN fairly have a close error rate.

Q4

Download the births data posted ccle week 4. (Use Regsubset function from Leap library) (STAT 101A material)

(a) Use the appropriate transformation to the response variable first (birth weight).

```
births <- read.csv("/Users/takaooba/Downloads/births 10000 Short F2021.csv")
births <- na.omit(births)
head(births)
```

```
##   Institution.type Plurality.of.birth Gender Race.of.child  Race Age.of.father
## 1                1                  1      2             1 White            50
## 2                1                  1      2             1 White            19
## 3                1                  1      2             1 White            37
## 5                1                  1      2             2 Black            39
## 6                1                  1      1             2 Black            20
## 7                1                  1      2             1 White            30
##   Age.of.mother Education.of.father..years. Education.of.mother..years.
## 1            24                          12                          15
## 2            18                           9                           9
## 3            35                          17                          17
## 5            31                          11                          16
## 6            19                          11                          12
## 7            27                          16                          16
##   Total.Preg BDead Terms Date.LBirth Month.LBirth Year.LBirth LOutcome Weeks
## 1          2     0     0       32004            3        2004        1    38
## 2          1     0     0           0            0           0        9    35
## 3          2     0     0      112003           11        2003        1    38
## 5          1     0     0           0            0           0        9    38
## 6          1     0     0           0            0           0        9    36
## 7          1     0     0           0            0           0        9    40
##   Prenatal Trimester.Prenatal Visits Birth.weight.group Marital Birth.Attendant
## 1        3                  1     10                  5       2               1
## 2        3                  1      9                  6       2               1
## 3        1                  1     20                  5       1               1
## 5        6                  2     12                  5       2               1
## 6        4                  2     10                  6       2               1
## 7        1                  1     20                  6       1               1
##   Numchild Month.Term Year.Term Low.Birth RaceMom RaceDad Mother.Minority
## 1        1          0         0      Norm       1       2           White
## 2        0          0         0      Norm       1       1           White
## 3        1          0         0      Norm       1       1           White
## 5        0          0         0      Norm       2       2        Nonwhite
## 6        0          0         0      Norm       2       1        Nonwhite
## 7        0          0         0      Norm       1       1           White
##   Father.Minority HispMom HispDad AveCigs Smoker AveDrink Wt.Gain
## 1        Nonwhite       N       N       0     No        0      50
## 2           White       N       N      23   Cigs        0      35
## 3           White       N       N       0     No        0      24
## 5        Nonwhite       N       N       0     No        0      30
## 6           White       N       M       0     No        0      10
## 7           White       N       N       0     No        0      37
##   Birth.Weight..g.
## 1         2865.875
## 2         3121.250
## 3         2667.250
## 5         2979.375
## 6         3036.125
## 7         3092.875
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.1.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.2
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:boot':
##
##     logit
```

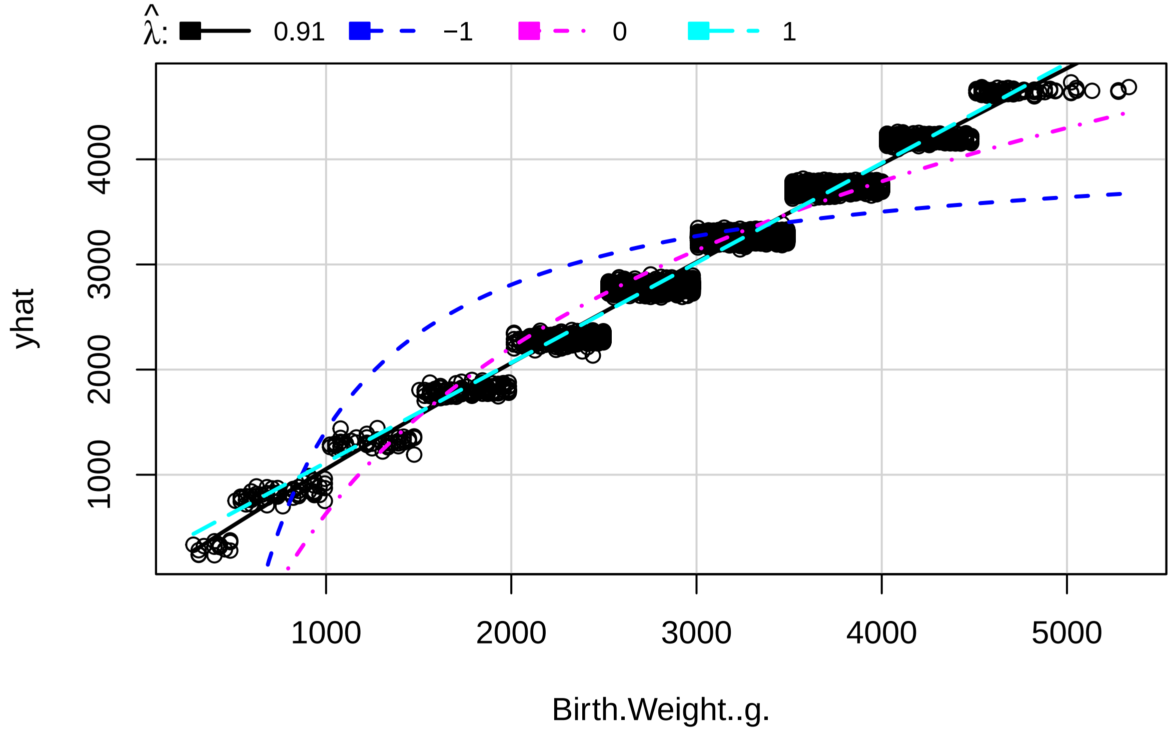
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
births.lm <- lm(Birth.Weight..g. ~., data = births)
summary(births.lm)
```

```
##
## Call:
## lm(formula = Birth.Weight..g. ~ ., data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -361.06 -111.83    1.04  110.19  647.04
##
## Coefficients: (4 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                113.630103  70.652080   1.608  0.10781
## Institution.type            -5.793001   4.112440  -1.409  0.15898
## Plurality.of.birth         -45.619579   9.005894  -5.066 4.17e-07 ***
## Gender                      -9.138465   3.147681  -2.903  0.00370 **
## Race.of.child               -0.670663   3.304553  -0.203  0.83918
## RaceOther                   16.159348  13.816763   1.170  0.24222
## RaceWhite                   20.212004   9.504359   2.127  0.03348 *
## Age.of.father               -0.505934   0.350712  -1.443  0.14918
## Age.of.mother                0.885874   0.441311   2.007  0.04474 *
## Education.of.father..years.  1.311397   0.816684   1.606  0.10837
## Education.of.mother..years. -0.293253   0.871284  -0.337  0.73645
## Total.Preg                   1.608333   2.168599   0.742  0.45832
## BDead                       12.209198  12.353121   0.988  0.32301
## Terms                       -1.312101   3.625692  -0.362  0.71744
## Date.LBirth                  0.005006   0.004744   1.055  0.29132
## Month.LBirth               -50.236160  47.676855  -1.054  0.29206
## Year.LBirth                        NA         NA      NA       NA
## LOutcome                     0.178140   1.101207   0.162  0.87149
## Weeks                       10.663338   0.773040  13.794  < 2e-16 ***
## Prenatal                     1.542984   2.316066   0.666  0.50530
## Trimester.Prenatal           5.204143   7.169221   0.726  0.46792
## Visits                       1.093058   0.462330   2.364  0.01809 *
## Birth.weight.group         454.118301   1.869896 242.857  < 2e-16 ***
## Marital                     -8.555404   4.165291  -2.054  0.04001 *
## Birth.Attendant              1.239004   2.480073   0.500  0.61738
## Numchild                           NA         NA      NA       NA
## Month.Term                   0.758097   0.837073   0.906  0.36515
## Year.Term                   -0.002471   0.003447  -0.717  0.47336
## Low.BirthNorm               20.857739   8.009332   2.604  0.00923 **
## RaceMom                            NA         NA      NA       NA
## RaceDad                     -3.924263   2.776333  -1.413  0.15756
## Mother.MinorityWhite               NA         NA      NA       NA
## Father.MinorityWhite        -6.396834   9.478744  -0.675  0.49978
## HispMomM                   -38.915071  51.140326  -0.761  0.44671
## HispMomN                   -40.024623  50.299538  -0.796  0.42622
## HispMomO                   -77.721620  61.040619  -1.273  0.20296
## HispMomP                   -35.992761  53.600230  -0.672  0.50192
## HispMomS                   -66.354014  51.505273  -1.288  0.19768
## HispMomU                    57.118474  89.186888   0.640  0.52191
## HispDadM                    -2.831209  48.063482  -0.059  0.95303
## HispDadN                    -9.641996  47.484435  -0.203  0.83910
## HispDadO                     2.828989  58.213494   0.049  0.96124
## HispDadP                   -13.091633  50.266620  -0.260  0.79453
## HispDadS                    24.339338  48.388548   0.503  0.61498
## HispDadU                        -79.559371  82.246593  -0.967  0.33341
## AveCigs                           1.695141   0.896925   1.890  0.05880 .
## SmokerNo                         39.508524  10.024921   3.941 8.18e-05 ***
## AveDrink                          4.645455  12.954389   0.359  0.71990
## Wt.Gain                           0.561398   0.119590   4.694 2.72e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 137.7 on 7816 degrees of freedom
## Multiple R-squared:  0.9491, Adjusted R-squared:  0.9488
## F-statistic:  3314 on 44 and 7816 DF,  p-value: < 2.2e-16
```

```
inverseResponsePlot(births.lm)
```



```
##       lambda        RSS
## 1  0.9093993  139157581
## 2 -1.0000000 1338373952
## 3  0.0000000  365771798
## 4  1.0000000  140559620
```

```
# The transformation with the lowest RSS is lambda = 0.9093993
```

Utilizing the inverseReversePlot, I determined that the transformation is with lambda of 0.9093993 because it has the lowest RSS value.

```r
# births.lm <- lm((Birth.Weight..g.)^0.9093993 ~., data = births)
weight <- (births$Birth.Weight..g.)^0.9093993

summary(births.lm)
```

```
##
## Call:
## lm(formula = Birth.Weight..g. ~ ., data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -361.06 -111.83    1.04  110.19  647.04
##
## Coefficients: (4 not defined because of singularities)
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 113.630103  70.652080   1.608  0.10781
## Institution.type             -5.793001   4.112440  -1.409  0.15898
## Plurality.of.birth          -45.619579   9.005894  -5.066 4.17e-07 ***
## Gender                       -9.138465   3.147681  -2.903  0.00370 **
## Race.of.child                -0.670663   3.304553  -0.203  0.83918
## RaceOther                    16.159348  13.816763   1.170  0.24222
## RaceWhite                    20.212004   9.504359   2.127  0.03348 *
## Age.of.father                -0.505934   0.350712  -1.443  0.14918
## Age.of.mother                 0.885874   0.441311   2.007  0.04474 *
## Education.of.father..years.   1.311397   0.816684   1.606  0.10837
## Education.of.mother..years.  -0.293253   0.871284  -0.337  0.73645
## Total.Preg                    1.608333   2.168599   0.742  0.45832
## BDead                        12.209198  12.353121   0.988  0.32301
## Terms                        -1.312101   3.625692  -0.362  0.71744
## Date.LBirth                   0.005006   0.004744   1.055  0.29132
## Month.LBirth                -50.236160  47.676855  -1.054  0.29206
## Year.LBirth                         NA         NA      NA       NA
## LOutcome                      0.178140   1.101207   0.162  0.87149
## Weeks                        10.663338   0.773040  13.794  < 2e-16 ***
## Prenatal                      1.542984   2.316066   0.666  0.50530
## Trimester.Prenatal            5.204143   7.169221   0.726  0.46792
## Visits                        1.093058   0.462330   2.364  0.01809 *
## Birth.weight.group          454.118301   1.869896 242.857  < 2e-16 ***
## Marital                      -8.555404   4.165291  -2.054  0.04001 *
## Birth.Attendant               1.239004   2.480073   0.500  0.61738
## Numchild                            NA         NA      NA       NA
## Month.Term                    0.758097   0.837073   0.906  0.36515
## Year.Term                    -0.002471   0.003447  -0.717  0.47336
## Low.BirthNorm                20.857739   8.009332   2.604  0.00923 **
## RaceMom                             NA         NA      NA       NA
## RaceDad                      -3.924263   2.776333  -1.413  0.15756
## Mother.MinorityWhite                NA         NA      NA       NA
## Father.MinorityWhite         -6.396834   9.478744  -0.675  0.49978
## HispMomM                    -38.915071  51.140326  -0.761  0.44671
## HispMomN                    -40.024623  50.299538  -0.796  0.42622
## HispMomO                    -77.721620  61.040619  -1.273  0.20296
## HispMomP                    -35.992761  53.600230  -0.672  0.50192
## HispMomS                    -66.354014  51.505273  -1.288  0.19768
## HispMomU                     57.118474  89.186888   0.640  0.52191
## HispDadM                     -2.831209  48.063482  -0.059  0.95303
## HispDadN                     -9.641996  47.484435  -0.203  0.83910
## HispDadO                      2.828989  58.213494   0.049  0.96124
## HispDadP                    -13.091633  50.266620  -0.260  0.79453
## HispDadS                     24.339338  48.388548   0.503  0.61498
## HispDadU                    -79.559371  82.246593  -0.967  0.33341
## AveCigs                       1.695141   0.896925   1.890  0.05880 .
## SmokerNo                     39.508524  10.024921   3.941 8.18e-05 ***
## AveDrink                      4.645455  12.954389   0.359  0.71990
## Wt.Gain                       0.561398   0.119590   4.694 2.72e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 137.7 on 7816 degrees of freedom
## Multiple R-squared:  0.9491, Adjusted R-squared:  0.9488
## F-statistic:  3314 on 44 and 7816 DF,  p-value: < 2.2e-16
```

```
# We can see that the adjusted R-squared increases by 0.007
```

Further, we will perform the transformation on the weight variable and generate the model correspondingly.

(b) Use Backwards Stepwise regression to determine a Least Squares model that predicts the birth weight based on best Mallows-Cp. Do this using set.seed(1128).

```
library(leaps)

set.seed(1128)

regfit.bck <- regsubsets(weight ~ ., data = births, method = "backward")
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 4 linear dependencies found

## Reordering variables and trying again:
```

```
summary(regfit.bck)
```

```
## Subset selection object
## Call: regsubsets.formula(weight ~ ., data = births, method = "backward")
## 49 Variables  (and intercept)
##                              Forced in Forced out
## Institution.type                 FALSE      FALSE
## Plurality.of.birth               FALSE      FALSE
## Gender                           FALSE      FALSE
## Race.of.child                    FALSE      FALSE
## RaceOther                        FALSE      FALSE
## RaceWhite                        FALSE      FALSE
## Age.of.father                    FALSE      FALSE
## Age.of.mother                   FALSE      FALSE
## Education.of.father..years.     FALSE      FALSE
## Education.of.mother..years.     FALSE      FALSE
## Total.Preg                      FALSE      FALSE
## BDead                           FALSE      FALSE
## Terms                           FALSE      FALSE
## Date.LBirth                     FALSE      FALSE
## Month.LBirth                    FALSE      FALSE
## LOutcome                        FALSE      FALSE
## Weeks                           FALSE      FALSE
## Prenatal                        FALSE      FALSE
## Trimester.Prenatal              FALSE      FALSE
## Visits                          FALSE      FALSE
## Birth.weight.group              FALSE      FALSE
## Marital                         FALSE      FALSE
## Birth.Attendant                 FALSE      FALSE
## Month.Term                      FALSE      FALSE
## Year.Term                       FALSE      FALSE
## Low.BirthNorm                   FALSE      FALSE
## RaceDad                         FALSE      FALSE
## Father.MinorityWhite            FALSE      FALSE
## HispMomM                        FALSE      FALSE
## HispMomN                        FALSE      FALSE
## HispMomO                        FALSE      FALSE
## HispMomP                        FALSE      FALSE
## HispMomS                        FALSE      FALSE
## HispMomU                        FALSE      FALSE
## HispDadM                        FALSE      FALSE
## HispDadN                        FALSE      FALSE
## HispDadO                        FALSE      FALSE
## HispDadP                        FALSE      FALSE
## HispDadS                        FALSE      FALSE
## HispDadU                        FALSE      FALSE
## AveCigs                         FALSE      FALSE
## SmokerNo                        FALSE      FALSE
## AveDrink                        FALSE      FALSE
## Wt.Gain                         FALSE      FALSE
## Birth.Weight..g.                FALSE      FALSE
## Year.LBirth                     FALSE      FALSE
## Numchild                        FALSE      FALSE
## RaceMom                         FALSE      FALSE
## Mother.MinorityWhite            FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: backward
##          Institution.type Plurality.of.birth Gender Race.of.child RaceOther
## 1  ( 1 ) " "              " "                " "    " "           " "
## 2  ( 1 ) " "              " "                " "    " "           " "
## 3  ( 1 ) " "              " "                " "    " "           " "
## 4  ( 1 ) " "              " "                " "    " "           " "
## 5  ( 1 ) " "              " "                " "    " "           " "
## 6  ( 1 ) " "              " "                " "    " "           " "
## 7  ( 1 ) " "              " "                " "    " "           " "
## 8  ( 1 ) " "              " "                " "    " "           " "
## 9  ( 1 ) " "              " "                " "    " "           " "
##          RaceWhite Age.of.father Age.of.mother Education.of.father..years.
## 1  ( 1 ) " "       " "           " "           " "                        
## 2  ( 1 ) " "       " "           " "           " "                        
## 3  ( 1 ) " "       " "           " "           " "                        
## 4  ( 1 ) " "       " "           " "           " "                        
## 5  ( 1 ) " "       " "           " "           " "                        
## 6  ( 1 ) " "       " "           " "           " "                        
## 7  ( 1 ) " "       " "           " "           " "                        
## 8  ( 1 ) " "       " "           " "           " "                        
## 9  ( 1 ) " "       " "           " "           " "                        
##          Education.of.mother..years. Total.Preg BDead Terms Date.LBirth
## 1  ( 1 ) " "                         " "        " "   " "   " "        
## 2  ( 1 ) " "                         " "        " "   " "   " "        
## 3  ( 1 ) " "                         " "        " "   " "   " "        
## 4  ( 1 ) " "                         " "        "*"   " "   " "        
## 5  ( 1 ) " "                         " "        "*"   " "   " "        
## 6  ( 1 ) " "                         " "        "*"   " "   " "        
## 7  ( 1 ) " "                         " "        "*"   " "   " "        
## 8  ( 1 ) " "                         "*"        "*"   " "   " "        
## 9  ( 1 ) " "                         "*"        "*"   " "   " "        
##          Month.LBirth Year.LBirth LOutcome Weeks Prenatal Trimester.Prenatal
## 1  ( 1 ) " "          " "         " "      " "   " "      " "               
## 2  ( 1 ) " "          " "         " "      "*"   " "      " "               
## 3  ( 1 ) " "          " "         " "      "*"   " "      " "               
## 4  ( 1 ) " "          " "         " "      "*"   " "      " "               
## 5  ( 1 ) " "          " "         " "      "*"   " "      " "               
## 6  ( 1 ) " "          " "         " "      "*"   "*"      " "               
## 7  ( 1 ) " "          " "         " "      "*"   "*"      " "               
## 8  ( 1 ) " "          " "         " "      "*"   "*"      " "               
## 9  ( 1 ) " "          " "         " "      "*"   "*"      " "               
##          Visits Birth.weight.group Marital Birth.Attendant Numchild Month.Term
## 1  ( 1 ) " "    " "                " "     " "             " "      " "       
## 2  ( 1 ) " "    " "                " "     " "             " "      " "       
## 3  ( 1 ) " "    " "                " "     " "             " "      " "       
## 4  ( 1 ) " "    " "                " "     " "             " "      " "       
## 5  ( 1 ) "*"    " "                " "     " "             " "      " "       
## 6  ( 1 ) "*"    " "                " "     " "             " "      " "       
## 7  ( 1 ) "*"    "*"                " "     " "             " "      " "       
## 8  ( 1 ) "*"    "*"                " "     " "             " "      " "       
## 9  ( 1 ) "*"    "*"                " "     " "             " "      "*"       
##          Year.Term Low.BirthNorm RaceMom RaceDad Mother.MinorityWhite
## 1  ( 1 ) " "       " "           " "     " "     " "                 
## 2  ( 1 ) " "       " "           " "     " "     " "                 
## 3  ( 1 ) " "       "*"           " "     " "     " "                 
## 4  ( 1 ) " "       "*"           " "     " "     " "                 
## 5  ( 1 ) " "       "*"           " "     " "     " "                 
## 6  ( 1 ) " "       "*"           " "     " "     " "                 
## 7  ( 1 ) " "       "*"           " "     " "     " "                 
## 8  ( 1 ) " "       "*"           " "     " "     " "                 
## 9  ( 1 ) " "       "*"           " "     " "     " "                 
##          Father.MinorityWhite HispMomM HispMomN HispMomO HispMomP HispMomS
## 1  ( 1 ) " "                  " "      " "      " "      " "      " "     
## 2  ( 1 ) " "                  " "      " "      " "      " "      " "     
## 3  ( 1 ) " "                  " "      " "      " "      " "      " "     
## 4  ( 1 ) " "                  " "      " "      " "      " "      " "
## 5  ( 1 ) " "                  " "      " "      " "      " "      " "
## 6  ( 1 ) " "                  " "      " "      " "      " "      " "
## 7  ( 1 ) " "                  " "      " "      " "      " "      " "
## 8  ( 1 ) " "                  " "      " "      " "      " "      " "
## 9  ( 1 ) " "                  " "      " "      " "      " "      " "
##          HispMomU HispDadM HispDadN HispDadO HispDadP HispDadS HispDadU AveCigs
## 1  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 2  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 3  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 4  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 5  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 6  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 7  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 8  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 9  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
##          SmokerNo AveDrink Wt.Gain Birth.Weight..g.
## 1  ( 1 ) " "      " "      " "     "*"
## 2  ( 1 ) " "      " "      " "     "*"
## 3  ( 1 ) " "      " "      " "     "*"
## 4  ( 1 ) " "      " "      " "     "*"
## 5  ( 1 ) " "      " "      " "     "*"
## 6  ( 1 ) " "      " "      " "     "*"
## 7  ( 1 ) " "      " "      " "     "*"
## 8  ( 1 ) " "      " "      " "     "*"
## 9  ( 1 ) " "      " "      " "     "*"
```

```
out <- summary(regsubsets(Birth.Weight..g. ~ ., data = births, method = "backward"))
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 4 linear dependencies found
```

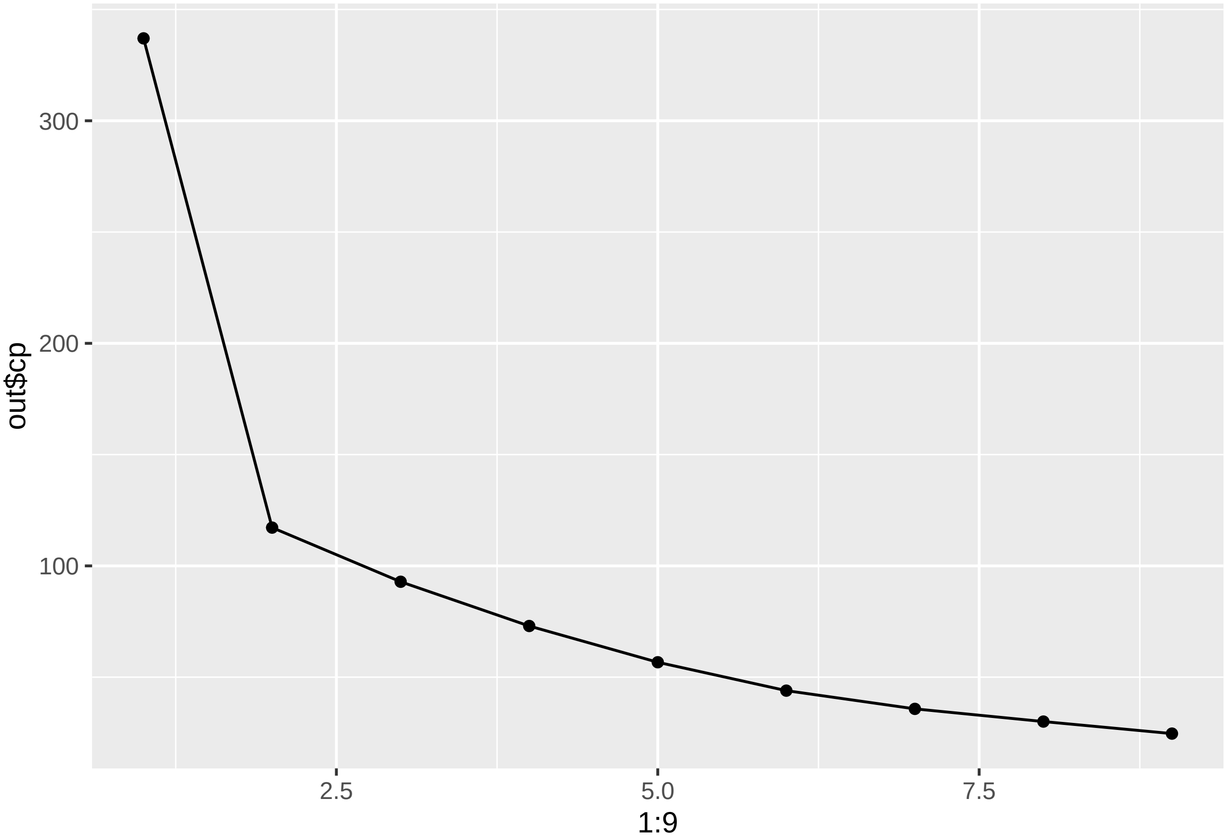
```
## Reordering variables and trying again:
```

```
# lr.model <- lm(Birth.Weight..g. ~ ., data = births.t)
# out <- summary(lr.model)

# regfit.bck <- regsubsets(x = births[,1:37], y = births[,38], method = "backward")
# summary(regfit.bck)
```

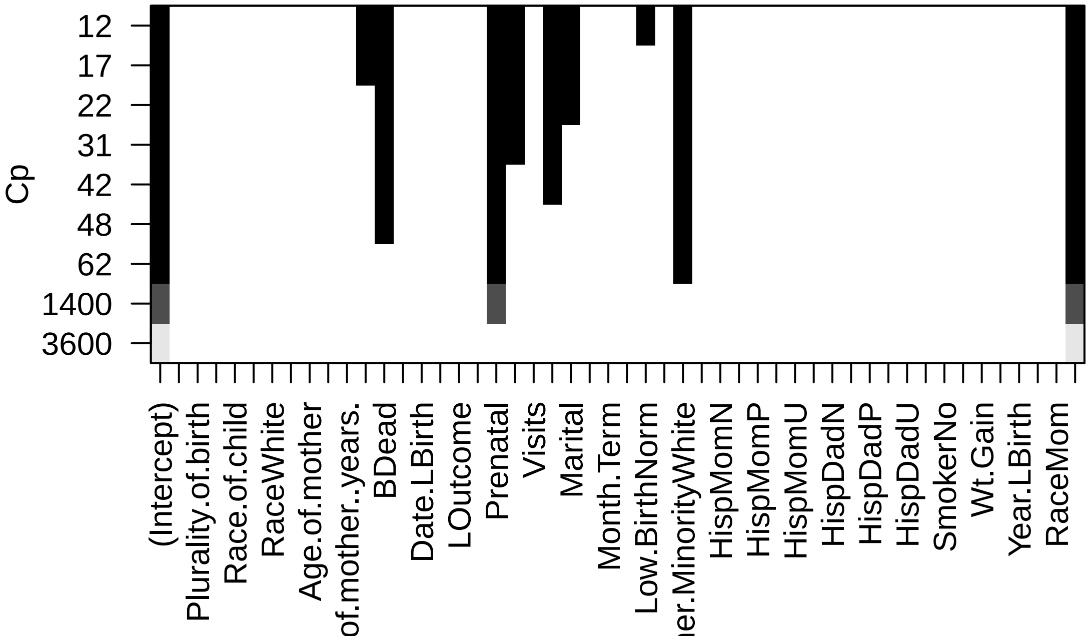
```
library(ggplot2)
qplot(1:9, out$cp) + geom_line()
```

```
## Warning: ‘qplot()‘ was deprecated in ggplot2 3.4.0.
```



Notice how the plot follows an elbow shape where the difference is initially very great at first and gets smaller and smaller as we move to the right of the x-axis. We will select up to 9 predictors

```
plot(regfit.bck, scale = "Cp")
```



Based on the plot, we can say that the significant predictors are

- Mother.Minority
- Father Minority
- Low Birth
- Marital
- Birth.weight.group
- Trimester.Prenatal
- Prenatal
- BDead
- Total.Preg

We will continue on to make a model utilizing these predictors

```
best.select <- lm(weight ~ Mother.Minority + Father.Minority + Low.Birth + Marital + Birth.weight.group
summary(best.select)
```

```
##
## Call:
## lm(formula = weight ~ Mother.Minority + Father.Minority + Low.Birth +
##     Marital + Birth.weight.group + Trimester.Prenatal + Prenatal +
##     BDead + Total.Preg, data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -204.506  -50.355    0.528   50.833  269.115
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          299.4916     4.9688  60.275  < 2e-16 ***
## Mother.MinorityWhite   5.7875     3.1530   1.836 0.066467 .
## Father.MinorityWhite   1.8456     3.1102   0.593 0.552938
## Low.BirthNorm         34.2335     3.4326   9.973  < 2e-16 ***
## Marital               -5.4896     1.6614  -3.304 0.000957 ***
## Birth.weight.group   205.0018     0.7530 272.232  < 2e-16 ***
## Trimester.Prenatal    -0.1456     3.2109  -0.045 0.963843
## Prenatal               1.2113     0.9916   1.222 0.221888
## BDead                  3.2117     5.4798   0.586 0.557822
## Total.Preg             0.1733     0.4807   0.361 0.718460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.91 on 7851 degrees of freedom
## Multiple R-squared:  0.9472, Adjusted R-squared:  0.9472
## F-statistic: 1.566e+04 on 9 and 7851 DF,  p-value: < 2.2e-16
```

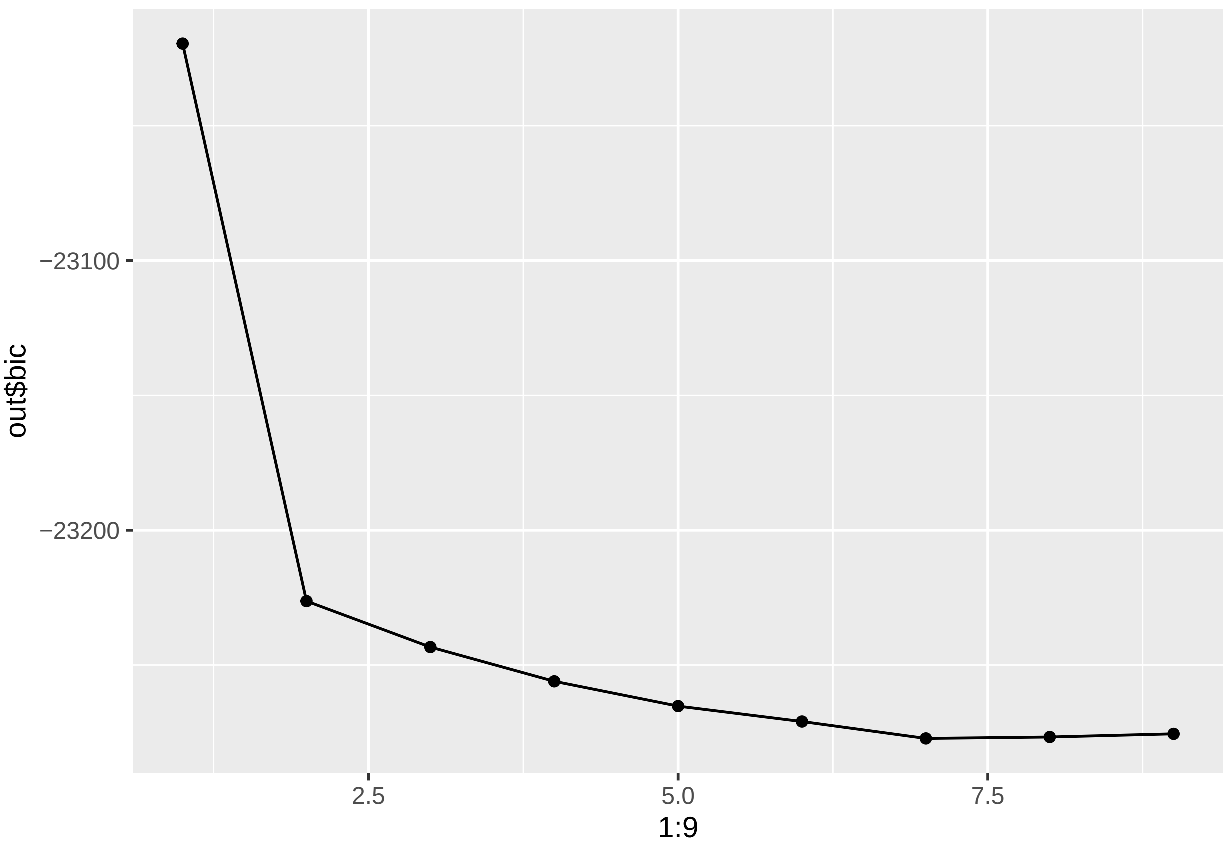
(c) Use Backwards Stepwise regression to determine a Least Squares model that predicts the birth weight based on best BIC. Do this using set.seed(1128).

```
library(ggplot2)
out <- summary(regsubsets(Birth.Weight..g. ~ ., data = births, method = "forward"))
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 4 linear dependencies found
```

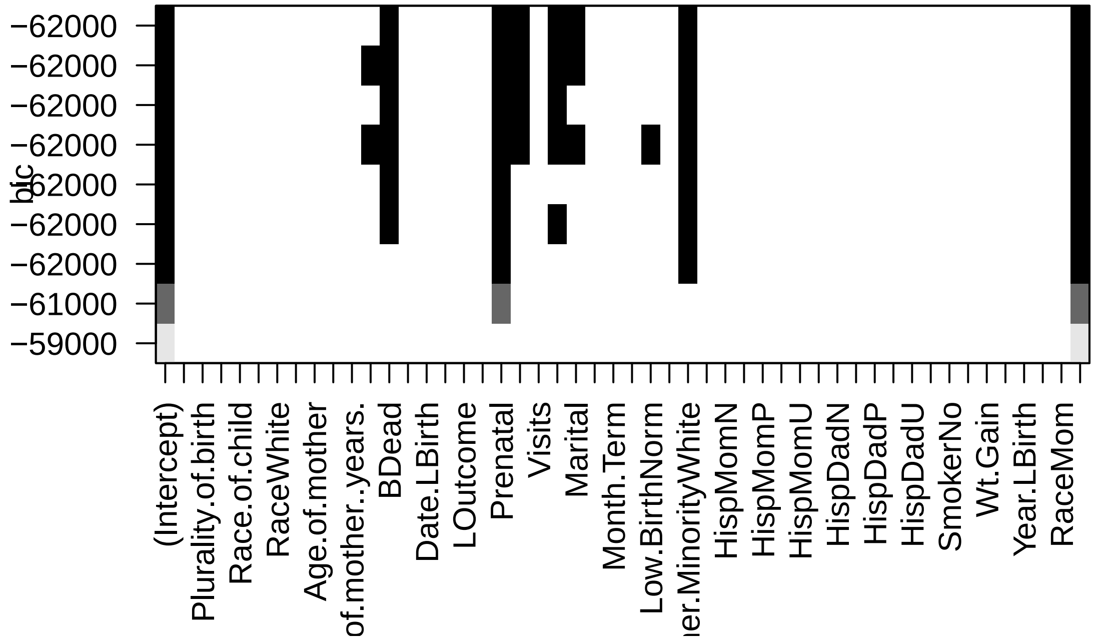
```
## Reordering variables and trying again:
```

```
qplot(1:9, out$bic) + geom_line()
```



Notice how the plot follows an elbow shape where the difference is initially very great at first and gets smaller and smaller as we move to the right of the x-axis. We will select up to 9 predictors.

```
plot(regfit.bck, scale = "bic")
```



Based on above graph, we have that the best predictors are - Mother.Minority - Father.Minority - Marital - Birth.weight.group - Trimester.Prenatal - Prenatal - BDead

Further we will make a model utilizing these predictors

```
best.select2 <- lm(weight ~ Mother.Minority + Father.Minority + Marital + Birth.weight.group + Trimeste
summary(best.select2)
```

```
##
## Call:
## lm(formula = weight ~ Mother.Minority + Father.Minority + Marital +
##     Birth.weight.group + Trimester.Prenatal + Prenatal + BDead,
##     data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -211.073  -50.891    1.093   51.890  257.762
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          301.3625     4.8609  61.998  < 2e-16 ***
## Mother.MinorityWhite   5.9774     3.1721   1.884 0.059553 .
## Father.MinorityWhite   1.2588     3.1274   0.403 0.687307
## Marital               -5.5302     1.6679  -3.316 0.000918 ***
## Birth.weight.group   209.9393     0.5708 367.792  < 2e-16 ***
## Trimester.Prenatal    -0.3210     3.2301  -0.099 0.920832
## Prenatal               1.4493     0.9965   1.454 0.145891
## BDead                  3.8847     5.4431   0.714 0.475440
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 62.3 on 7853 degrees of freedom
## Multiple R-squared:  0.9466, Adjusted R-squared:  0.9465
## F-statistic: 1.987e+04 on 7 and 7853 DF,  p-value: < 2.2e-16
```

(d) Use forward Stepwise regression to determine a Least Squares model that predicts the birth weight. based on best Mallows-Cp. Do this using set.seed(1128).

```
set.seed(1128)

regfit.fwd <- regsubsets(weight ~ ., data = births, method = "forward")
```

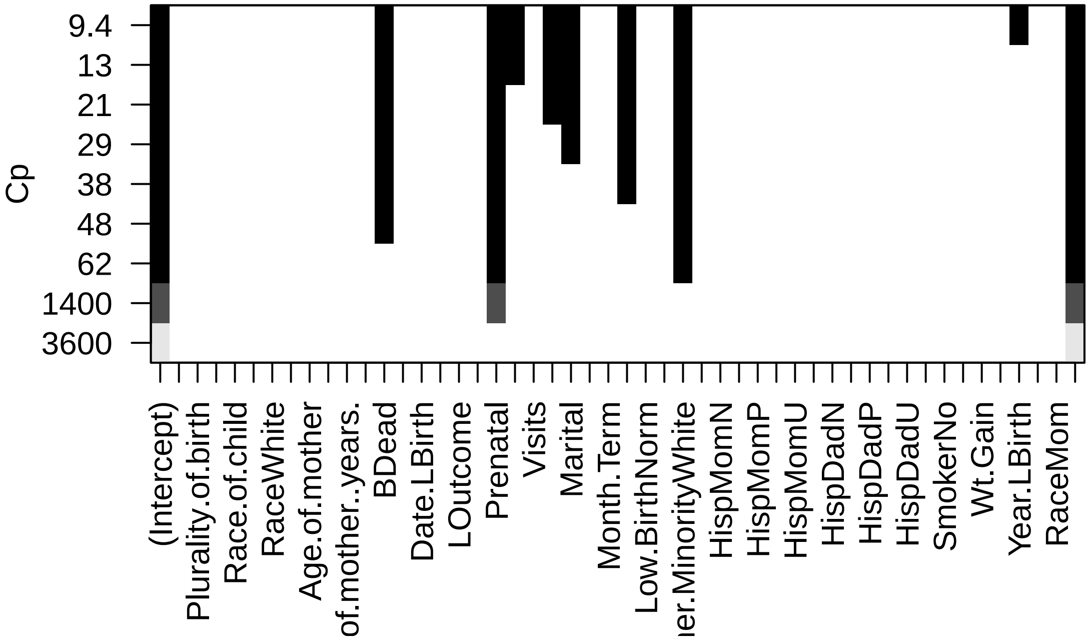
```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 4 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
summary(regfit.fwd)
```

```
## Subset selection object
## Call: regsubsets.formula(weight ~ ., data = births, method = "forward")
## 49 Variables  (and intercept)
##                             Forced in Forced out
## Institution.type                FALSE      FALSE
## Plurality.of.birth              FALSE      FALSE
## Gender                          FALSE      FALSE
## Race.of.child                   FALSE      FALSE
## RaceOther                       FALSE      FALSE
## RaceWhite                       FALSE      FALSE
## Age.of.father                   FALSE      FALSE
## Age.of.mother                   FALSE      FALSE
## Education.of.father..years.     FALSE      FALSE
## Education.of.mother..years.     FALSE      FALSE
## Total.Preg                      FALSE      FALSE
## BDead                           FALSE      FALSE
## Terms                           FALSE      FALSE
## Date.LBirth                     FALSE      FALSE
## Month.LBirth                    FALSE      FALSE
## LOutcome                        FALSE      FALSE
## Weeks                           FALSE      FALSE
## Prenatal                        FALSE      FALSE
## Trimester.Prenatal              FALSE      FALSE
## Visits                          FALSE      FALSE
## Birth.weight.group              FALSE      FALSE
## Marital                         FALSE      FALSE
## Birth.Attendant                 FALSE      FALSE
## Month.Term                      FALSE      FALSE
## Year.Term                       FALSE      FALSE
## Low.BirthNorm                   FALSE      FALSE
## RaceDad                         FALSE      FALSE
## Father.MinorityWhite            FALSE      FALSE
## HispMomM                        FALSE      FALSE
## HispMomN                        FALSE      FALSE
## HispMomO                        FALSE      FALSE
## HispMomP                        FALSE      FALSE
## HispMomS                        FALSE      FALSE
## HispMomU                        FALSE      FALSE
## HispDadM                        FALSE      FALSE
## HispDadN                        FALSE      FALSE
## HispDadO                        FALSE      FALSE
## HispDadP                        FALSE      FALSE
## HispDadS                        FALSE      FALSE
## HispDadU                        FALSE      FALSE
## AveCigs                         FALSE      FALSE
## SmokerNo                        FALSE      FALSE
## AveDrink                        FALSE      FALSE
## Wt.Gain                         FALSE      FALSE
## Birth.Weight..g.                FALSE      FALSE
## Year.LBirth                     FALSE      FALSE
## Numchild                        FALSE      FALSE
## RaceMom                         FALSE      FALSE
## Mother.MinorityWhite            FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: forward
##          Institution.type Plurality.of.birth Gender Race.of.child RaceOther
## 1  ( 1 ) " "              " "                " "    " "           " "
## 2  ( 1 ) " "              " "                " "    " "           " "
## 3  ( 1 ) " "              " "                " "    " "           " "
## 4  ( 1 ) " "              " "                " "    " "           " "
## 5  ( 1 ) " "              " "                " "    " "           " "
## 6  ( 1 ) " "              " "                " "    " "           " "
## 7  ( 1 ) " "              " "                " "    " "           " "
## 8  ( 1 ) " "              " "                " "    " "           " "
## 9  ( 1 ) " "              " "                " "    " "           " "
##          RaceWhite Age.of.father Age.of.mother Education.of.father..years.
## 1  ( 1 ) " "       " "           " "           " "
## 2  ( 1 ) " "       " "           " "           " "
## 3  ( 1 ) " "       " "           " "           " "
## 4  ( 1 ) " "       " "           " "           " "
## 5  ( 1 ) " "       " "           " "           " "
## 6  ( 1 ) " "       " "           " "           " "
## 7  ( 1 ) " "       " "           " "           " "
## 8  ( 1 ) " "       " "           " "           " "
## 9  ( 1 ) " "       " "           " "           " "
##          Education.of.mother..years. Total.Preg BDead Terms Date.LBirth
## 1  ( 1 ) " "                         " "        " "   " "   " "
## 2  ( 1 ) " "                         " "        " "   " "   " "
## 3  ( 1 ) " "                         " "        " "   " "   " "
## 4  ( 1 ) " "                         " "        "*"   " "   " "
## 5  ( 1 ) " "                         " "        "*"   " "   " "
## 6  ( 1 ) " "                         " "        "*"   " "   " "
## 7  ( 1 ) " "                         " "        "*"   " "   " "
## 8  ( 1 ) " "                         " "        "*"   " "   " "
## 9  ( 1 ) " "                         " "        "*"   " "   " "
##          Month.LBirth Year.LBirth LOutcome Weeks Prenatal Trimester.Prenatal
## 1  ( 1 ) " "          " "         " "      " "   " "      " "
## 2  ( 1 ) " "          " "         " "      "*"   " "      " "
## 3  ( 1 ) " "          " "         " "      "*"   " "      " "
## 4  ( 1 ) " "          " "         " "      "*"   " "      " "
## 5  ( 1 ) " "          " "         " "      "*"   " "      " "
## 6  ( 1 ) " "          " "         " "      "*"   " "      " "
## 7  ( 1 ) " "          " "         " "      "*"   " "      " "
## 8  ( 1 ) " "          " "         " "      "*"   "*"      " "
## 9  ( 1 ) " "          " "         " "      "*"   "*"      " "
##          Visits Birth.weight.group Marital Birth.Attendant Numchild Month.Term
## 1  ( 1 ) " "    " "                " "     " "             " "      " "
## 2  ( 1 ) " "    " "                " "     " "             " "      " "
## 3  ( 1 ) " "    " "                " "     " "             " "      " "
## 4  ( 1 ) " "    " "                " "     " "             " "      " "
## 5  ( 1 ) " "    " "                " "     " "             "*"      " "
## 6  ( 1 ) " "    "*"                " "     " "             "*"      " "
## 7  ( 1 ) "*"    "*"                " "     " "             "*"      " "
## 8  ( 1 ) "*"    "*"                " "     " "             "*"      " "
## 9  ( 1 ) "*"    "*"                " "     " "             "*"      " "
##          Year.Term Low.BirthNorm RaceMom RaceDad Mother.MinorityWhite
## 1  ( 1 ) " "       " "           " "     " "     " "
## 2  ( 1 ) " "       " "           " "     " "     " "
## 3  ( 1 ) " "       "*"           " "     " "     " "
## 4  ( 1 ) " "       "*"           " "     " "     " "
## 5  ( 1 ) " "       "*"           " "     " "     " "
## 6  ( 1 ) " "       "*"           " "     " "     " "
## 7  ( 1 ) " "       "*"           " "     " "     " "
## 8  ( 1 ) " "       "*"           " "     " "     " "
## 9  ( 1 ) " "       "*"           " "     " "     " "
##          Father.MinorityWhite HispMomM HispMomN HispMomO HispMomP HispMomS
## 1  ( 1 ) " "                  " "      " "      " "      " "      " "
## 2  ( 1 ) " "                  " "      " "      " "      " "      " "
## 3  ( 1 ) " "                  " "      " "      " "      " "      " "
## 4  ( 1 ) " "                  " "      " "      " "      " "      " "
## 5  ( 1 ) " "                  " "      " "      " "      " "      " "
## 6  ( 1 ) " "                  " "      " "      " "      " "      " "
## 7  ( 1 ) " "                  " "      " "      " "      " "      " "
## 8  ( 1 ) " "                  " "      " "      " "      " "      " "
## 9  ( 1 ) " "                  " "      " "      " "      " "      " "
##          HispMomU HispDadM HispDadN HispDadO HispDadP HispDadS HispDadU AveCigs
## 1  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 2  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 3  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 4  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 5  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 6  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 7  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 8  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
## 9  ( 1 ) " "      " "      " "      " "      " "      " "      " "      " "
##           SmokerNo AveDrink Wt.Gain Birth.Weight..g.
## 1  ( 1 ) " "      " "      " "     "*"
## 2  ( 1 ) " "      " "      " "     "*"
## 3  ( 1 ) " "      " "      " "     "*"
## 4  ( 1 ) " "      " "      " "     "*"
## 5  ( 1 ) " "      " "      " "     "*"
## 6  ( 1 ) " "      " "      " "     "*"
## 7  ( 1 ) " "      " "      " "     "*"
## 8  ( 1 ) " "      " "      " "     "*"
## 9  ( 1 ) "*"      " "      " "     "*"
```

```
plot(regfit.fwd, scale = "Cp")
```



Based on above graph, we have that the best predictors are - Mother.Minority - Father.Minority - Marital - Birth.weight.group - Trimester.Prenatal - Prenatal - BDead - Year.LBirth - Year.Term

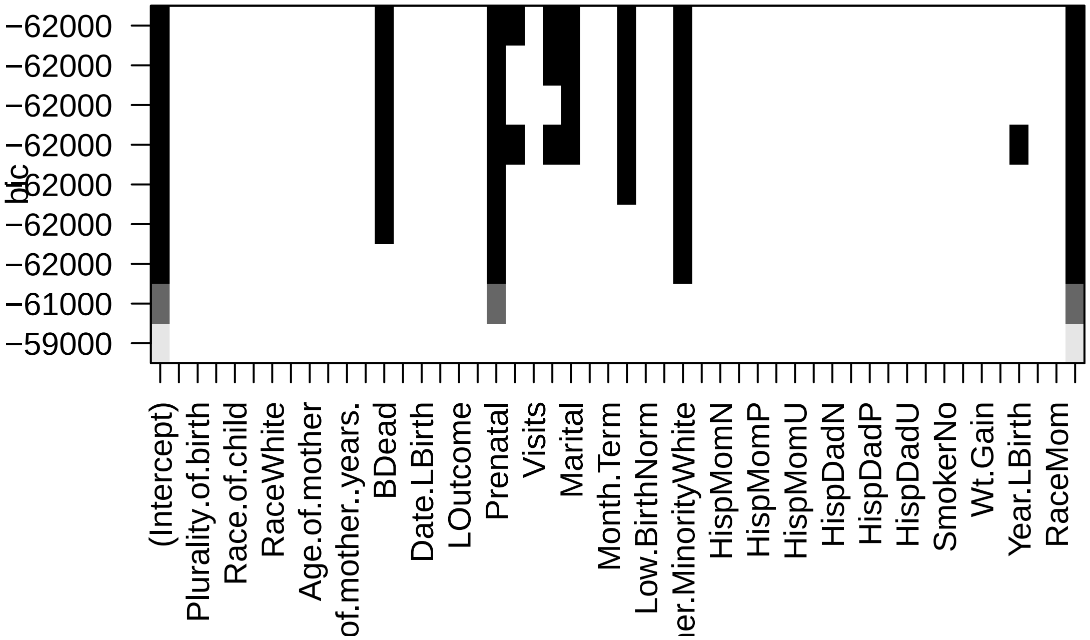
Further we will make a model utilizing these predictors

```
best.select3 <- lm(weight ~ Mother.Minority + Father.Minority + Marital + Birth.weight.group + Trimeste
summary(best.select3)
```

```
##
## Call:
## lm(formula = weight ~ Mother.Minority + Father.Minority + Marital +
##     Birth.weight.group + Trimester.Prenatal + Prenatal + BDead +
##     Year.LBirth + Year.Term, data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -210.954  -50.953    1.278   51.885  257.609
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.012e+02  4.944e+00  60.923  < 2e-16 ***
## Mother.MinorityWhite  5.981e+00  3.173e+00   1.885  0.05945 .
## Father.MinorityWhite  1.269e+00  3.129e+00   0.406  0.68501
## Marital              -5.495e+00  1.675e+00  -3.281  0.00104 **
## Birth.weight.group    2.099e+02  5.716e-01 367.260  < 2e-16 ***
## Trimester.Prenatal   -3.343e-01  3.231e+00  -0.103  0.91760
## Prenatal              1.442e+00  9.971e-01   1.446  0.14810
## BDead                 3.772e+00  5.467e+00   0.690  0.49023
## Year.LBirth           1.747e-04  7.302e-04   0.239  0.81095
## Year.Term            -6.819e-05  8.709e-04  -0.078  0.93759
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 62.3 on 7851 degrees of freedom
## Multiple R-squared:  0.9466, Adjusted R-squared:  0.9465
## F-statistic: 1.545e+04 on 9 and 7851 DF,  p-value: < 2.2e-16
```

(e) Use forward Stepwise regression to determine a Least Squares model that predicts the birth weight. based on best BIC. Do this using set.seed(1128).

```
plot(regfit.fwd, scale = "bic")
```



Based on above graph, we have that the best predictors are - BDead - Prenatal - Mother.Minority - Father.Minority - Marital - Birth.weight.group - Trimester.Prenatal - Year.Term

Further we will make a model utilizing these predictors

```
best.select4 <- lm(weight ~ BDead + Prenatal + Mother.Minority + Father.Minority + Marital + Birth.weig
summary(best.select4)
```

```
##
## Call:
## lm(formula = weight ~ BDead + Prenatal + Mother.Minority + Father.Minority +
##     Marital + Birth.weight.group + Trimester.Prenatal + Year.Term,
##     data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -211.094  -50.913    1.161   51.872  257.744
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.014e+02  4.883e+00  61.726  < 2e-16 ***
## BDead                 3.890e+00  5.444e+00   0.714 0.474941
## Prenatal              1.449e+00  9.966e-01   1.454 0.145966
## Mother.MinorityWhite  5.978e+00  3.172e+00   1.884 0.059549 .
## Father.MinorityWhite  1.255e+00  3.129e+00   0.401 0.688408
## Marital              -5.531e+00  1.668e+00  -3.316 0.000918 ***
## Birth.weight.group    2.099e+02  5.709e-01 367.763  < 2e-16 ***
## Trimester.Prenatal   -3.207e-01  3.230e+00  -0.099 0.920923
## Year.Term           -4.519e-05  8.655e-04  -0.052 0.958360
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 62.3 on 7852 degrees of freedom
## Multiple R-squared:  0.9466, Adjusted R-squared:  0.9465
## F-statistic: 1.739e+04 on 8 and 7852 DF,  p-value: < 2.2e-16
```

List the “best” Predictors. Write up a paragraph comparing results from parts b-d

From the built models, we have that the best predictors are the Mother.Minority, Father.Minority, Birth.weight.group, Trimester.Prenatal, Prenatal, BDead. We notice that all the models has the same amount of adjusted R-Squared. However, the backwards bic has the least amount of predictors, thus this will be the best model. Regarding the selection of the predictors, I assessed the various possibilities with the Teacher Assistant Mr. Yuantong Li has stated that we shall select all the predictors that are touching the very top line. In conclusion, we utilized the backwards and forward selection and further utilized Mallows Cp and BIC as well.