[STAT W4702] Statistical Inference & Modelling Group Project

Babies

12 December 2015

Abstract

Data Set

This project was conducted on the Low Birth Weight dataset collected in 1986 at Baystate Medical Center, Springfield, Massachusetts as a part of a bigger study on the factors influencing newborn infants' health and risk of serious health problems potentially leading to death. This dataset is distributed as a part of `MASS` library and contains **189 observations** and **10 variables**, among which `bwt` represents the exact amount of newborn infant's weight in grams and is used as the variable of interest we are trying to predict. The other 9 variables stand for different factors related to mothers' physiological parameters, such as age, weight and race, their health-related habits and behavior during pregnancy (smoking habits, presence of uterine irritability and number of physician visits). Also there is a low birth weight indicator `low`, which is defined as a binary variable showing whether the weight of an infant is below 2500 grams or not. Brief description of each variable is provided in the table below.

The goal of our research is to identify relationship between these variables and infant weight and understand the influence of each of them on the explained variable. The project pursue both inferential and predictive goals as it is equally important to be able to obtain inference about factors affecting newborn's health and to be able to react on the potential health risks in a timely manner, when the model predicts the low birth weight outcome for a certain observation. In order to accomplish this goal we tried to fit multiple linear and non-linear models exploring the rationale that could provide the evidence for certain types of models and finding balance between interpretability and predictive power of the model.

Cleaning Dataset

For the purposes of the research the dataset was cleaned in the following way:

- factor variable `race` was assigned with proper labels `white`, `black` and `other`;
- physisian visits were converted to a factor variable `ftv` with 3 labels `0`, `1` and `2+`;
- response is defined as an exact amount of infant's weight from `bwt`;
- all the columns are assigned with meaningful names.

Variable description table and summary statistics of the tidy dataset are provided below.

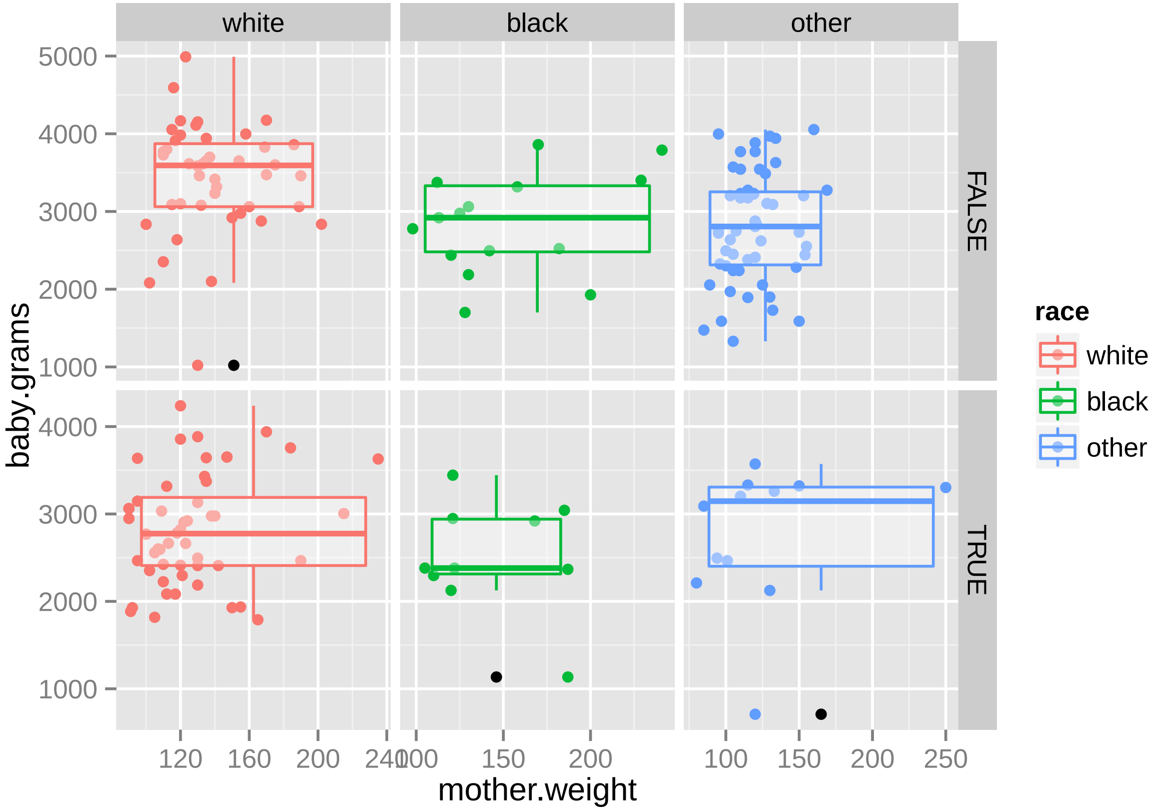
Variable	Description
`baby.grams`	weight of newborn infant in grams
`mother.age`	mother's age in years
`mother.weight`	mother's weight in pounds at last menstrual period
`race`	mother's race, factor variable with following labels: *white*, *black* or *other*
`smoke`	smoking status during pregnancy, binary variable
`prem.labor`	binary variable showing whether mother had premature labors before or not
`hypertension`	binary variable showing whether mother had hypertension or not

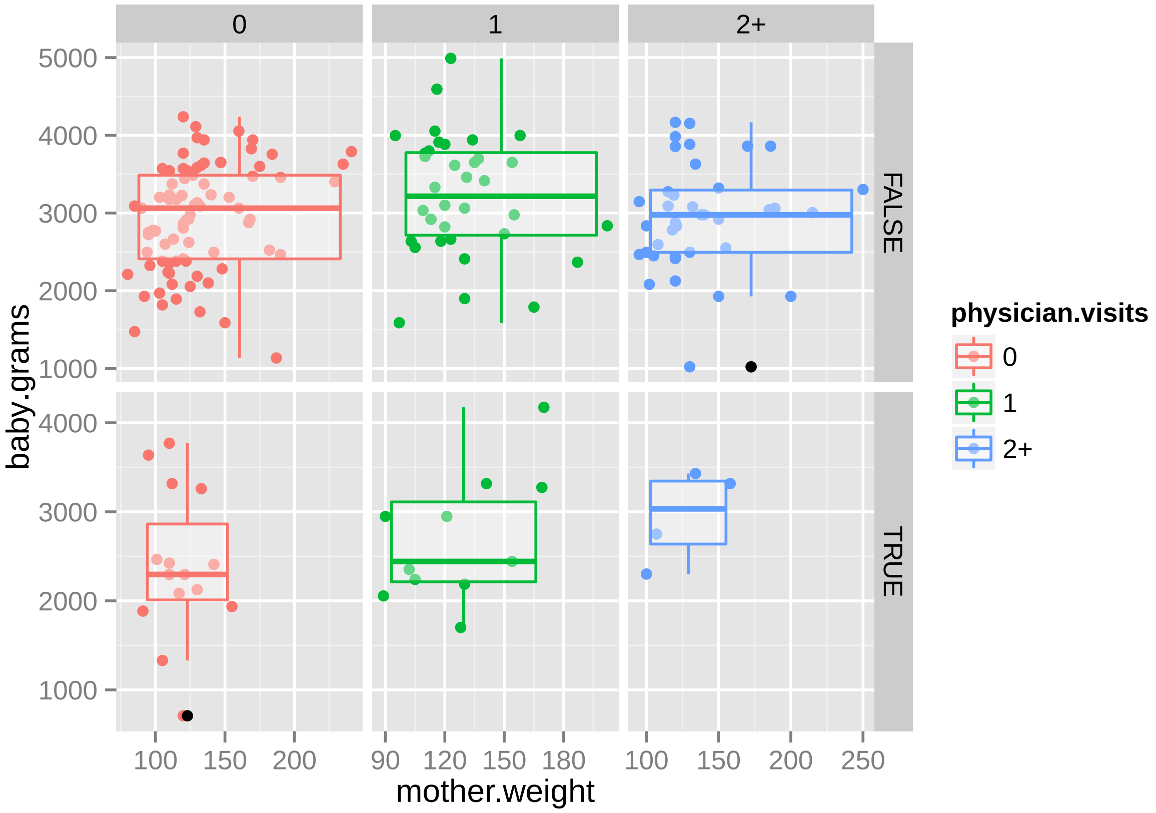
Variable	Description
`uterine`	binary variable showing presence of uterine irritability
`physician.visits`	number of physician visits during the first trimester: *0*, *1* or *2+*

```
##    baby.grams     mother.age     mother.weight      race       smoke
##  Min.   : 709   Min.   :14.00   Min.   : 80.0   white:96   Mode :logical
##  1st Qu.:2414   1st Qu.:19.00   1st Qu.:110.0   black:26   FALSE:115
##  Median :2977   Median :23.00   Median :121.0   other:67   TRUE :74
##  Mean   :2945   Mean   :23.24   Mean   :129.8              NA's :0
##  3rd Qu.:3487   3rd Qu.:26.00   3rd Qu.:140.0
##  Max.   :4990   Max.   :45.00   Max.   :250.0
##  prem.labor  hypertension     uterine        physician.visits
##  FALSE:159   Mode :logical   Mode :logical   0 :100
##  TRUE : 30   FALSE:177       FALSE:161       1 : 47
##              TRUE :12        TRUE :28        2+: 42
##              NA's :0         NA's :0
##
##
```

```
cor(bwt.grams[,1:3])
```

```
##               baby.grams mother.age mother.weight
## baby.grams    1.00000000 0.09031781     0.1857333
## mother.age    0.09031781 1.00000000     0.1800732
## mother.weight 0.18573328 0.18007315     1.0000000
```





```
set.seed(1)
train <- sample(1:nrow(bwt.grams), floor(0.75*nrow(bwt.grams)))
```

```
library(MASS)
data(birthwt)
bwt <- with(birthwt, {
  race <- factor(race, labels = c("white", "black", "other"))
  ptd <- factor(ptl > 0)
  ftv <- factor(ftv)
  levels(ftv)[-(1:2)] <- "2+"
  data.frame(low, age, lwt, race, smoke = (smoke > 0),
             ptd, ht = (ht > 0), ui = (ui > 0), ftv)
})
colnames(bwt) <- c("below.2500", "mother.age",
                   "mother.weight", "race",
                   "smoke", "prem.labor",
                   "hypertension", "uterine",
                   "physician.visits")

bwt.grams <- with(birthwt, {
  race <- factor(race, labels = c("white", "black", "other"))
  ptd <- factor(ptl > 0)
  ftv <- factor(ftv)
  levels(ftv)[-(1:2)] <- "2+"
  data.frame(bwt, age, lwt, race, smoke = (smoke > 0),
             ptd, ht = (ht > 0), ui = (ui > 0), ftv)
})
```

```r
colnames(bwt.grams) <- c("baby.grams", "mother.age",
                       "mother.weight", "race",
                       "smoke", "prem.labor",
                       "hypertension", "uterine",
                       "physician.visits")
summary(bwt)
```

```
##    below.2500       mother.age    mother.weight      race
##  Min.   :0.0000   Min.   :14.00   Min.   : 80.0   white:96
##  1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   black:26
##  Median :0.0000   Median :23.00   Median :121.0   other:67
##  Mean   :0.3122   Mean   :23.24   Mean   :129.8
##  3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0
##  Max.   :1.0000   Max.   :45.00   Max.   :250.0
##    smoke          prem.labor  hypertension     uterine
##  Mode :logical   FALSE:159   Mode :logical   Mode :logical
##  FALSE:115       TRUE : 30   FALSE:177       FALSE:161
##  TRUE :74                    TRUE :12        TRUE :28
##  NA's :0                     NA's :0         NA's :0
##
##
##  physician.visits
##  0 :100
##  1 : 47
##  2+: 42
##
##
##
```

```r
summary(bwt.grams)
```

```
##    baby.grams     mother.age    mother.weight      race      smoke
##  Min.   : 709   Min.   :14.00   Min.   : 80.0   white:96   Mode :logical
##  1st Qu.:2414   1st Qu.:19.00   1st Qu.:110.0   black:26   FALSE:115
##  Median :2977   Median :23.00   Median :121.0   other:67   TRUE :74
##  Mean   :2945   Mean   :23.24   Mean   :129.8              NA's :0
##  3rd Qu.:3487   3rd Qu.:26.00   3rd Qu.:140.0
##  Max.   :4990   Max.   :45.00   Max.   :250.0
##  prem.labor  hypertension     uterine        physician.visits
##  FALSE:159   Mode :logical   Mode :logical   0 :100
##  TRUE : 30   FALSE:177       FALSE:161       1 : 47
##              TRUE :12        TRUE :28        2+: 42
##              NA's :0         NA's :0
##
##
```

```r
bwt[0:10,]
```

```
##    below.2500 mother.age mother.weight  race smoke prem.labor hypertension
## 1           0         19           182 black FALSE      FALSE        FALSE
## 2           0         33           155 other FALSE      FALSE        FALSE
## 3           0         20           105 white  TRUE      FALSE        FALSE
## 4           0         21           108 white  TRUE      FALSE        FALSE
## 5           0         18           107 white  TRUE      FALSE        FALSE
## 6           0         21           124 other FALSE      FALSE        FALSE
## 7           0         22           118 white FALSE      FALSE        FALSE
## 8           0         17           103 other FALSE      FALSE        FALSE
## 9           0         29           123 white  TRUE      FALSE        FALSE
## 10          0         26           113 white  TRUE      FALSE        FALSE
##    uterine physician.visits
## 1     TRUE                0
## 2    FALSE               2+
## 3    FALSE                1
## 4     TRUE               2+
## 5     TRUE                0
## 6    FALSE                0
## 7    FALSE                1
## 8    FALSE                1
## 9    FALSE                1
## 10   FALSE                0
```

```
bwt.grams[0:10,]
```

```
##    baby.grams mother.age mother.weight  race smoke prem.labor hypertension
## 1        2523         19           182 black FALSE      FALSE        FALSE
## 2        2551         33           155 other FALSE      FALSE        FALSE
## 3        2557         20           105 white  TRUE      FALSE        FALSE
## 4        2594         21           108 white  TRUE      FALSE        FALSE
## 5        2600         18           107 white  TRUE      FALSE        FALSE
## 6        2622         21           124 other FALSE      FALSE        FALSE
## 7        2637         22           118 white FALSE      FALSE        FALSE
## 8        2637         17           103 other FALSE      FALSE        FALSE
## 9        2663         29           123 white  TRUE      FALSE        FALSE
## 10       2665         26           113 white  TRUE      FALSE        FALSE
##    uterine physician.visits
## 1     TRUE                0
## 2    FALSE               2+
## 3    FALSE                1
## 4     TRUE               2+
## 5     TRUE                0
## 6    FALSE                0
## 7    FALSE                1
## 8    FALSE                1
## 9    FALSE                1
## 10   FALSE                0
```

```
attach(bwt.grams)

library (leaps)
regfit.full=regsubsets(baby.grams~., bwt.grams, nvmax =19)
reg.summary = summary(regfit.full)
reg.summary$rsq
```

```
##  [1] 0.08061477 0.11225032 0.14782772 0.18905712 0.21364404 0.24039446
##  [7] 0.25042689 0.25537670 0.25647316 0.25682243
```

```r
par(mfrow =c(2,2))
plot(reg.summary$rss ,xlab=" Number of Variables ",ylab=" RSS", type="l")
plot(reg.summary$adjr2 ,xlab =" Number of Variables ", ylab=" Adjusted RSq",type="l")
max.adjr2=which.max (reg.summary$adjr2)
max.adjr2
```

```
## [1] 8
```

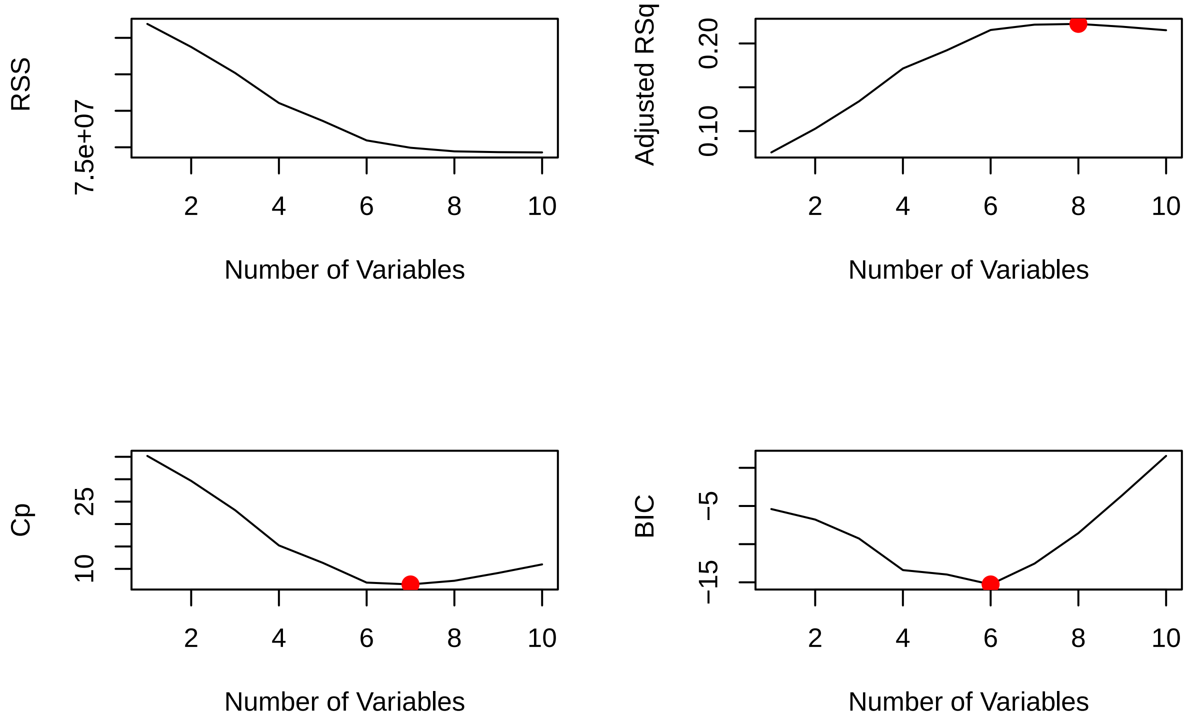
```r
points (max.adjr2, reg.summary$adjr2[max.adjr2], col ="red",cex =2, pch =20)


plot(reg.summary$cp ,xlab =" Number of Variables ", ylab="Cp", type='l')
min.cp= which.min (reg.summary$cp )
min.cp
```

```
## [1] 7
```

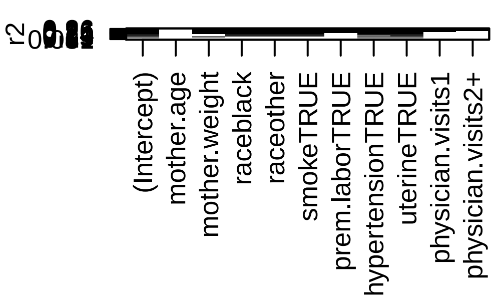
```r
points (min.cp, reg.summary$cp[min.cp], col ="red",cex =2, pch =20)


min.bic = which.min(reg.summary$bic)
min.bic
```

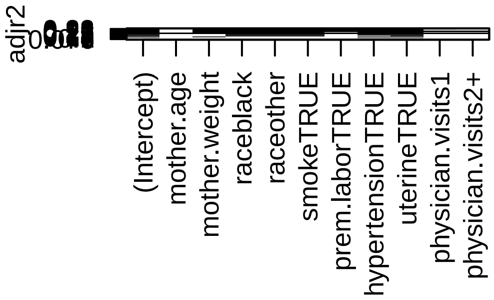
```
## [1] 6
```

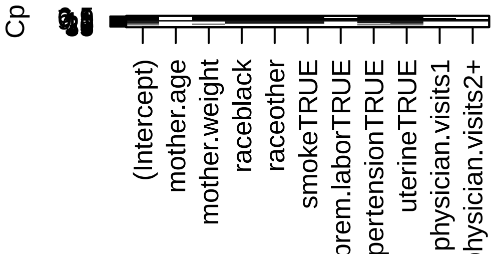
```r
plot(reg.summary$bic ,xlab=" Number of Variables ",ylab=" BIC", type='l')
points (min.bic, reg.summary$bic [min.bic], col =" red",cex =2, pch =20)
```

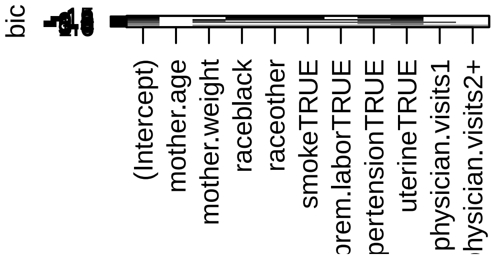


```r
plot(regfit.full ,scale ="r2")
plot(regfit.full ,scale ="adjr2")
plot(regfit.full ,scale ="Cp")
plot(regfit.full ,scale ="bic")
```









```r
coef(regfit.full, max.adjr2)
```

```
##       (Intercept)     mother.weight         raceblack         raceother 
##       2799.714010          4.194539       -453.359173       -305.169792 
##         smokeTRUE    prem.laborTRUE  hypertensionTRUE       uterineTRUE 
##       -294.468372       -235.263456       -577.857003       -478.599299 
## physician.visits1 
##        125.220667
```

```r
coef(regfit.full, min.cp)
```

```
##      (Intercept)    mother.weight        raceblack        raceother 
##      2871.512227         4.043831      -465.601219      -333.878191 
##        smokeTRUE   prem.laborTRUE hypertensionTRUE      uterineTRUE 
##      -325.081991      -207.834528      -573.799253      -491.143889
```

```r
coef(regfit.full, min.bic)
```

```
##      (Intercept)    mother.weight        raceblack        raceother 
##       2837.26392          4.24155       -475.05760       -348.15038 
##        smokeTRUE hypertensionTRUE      uterineTRUE 
##       -356.32095       -585.19312       -525.52390
```

```
classfit.full=regsubsets(below.2500~., bwt, nvmax =19)
class.summary = summary(classfit.full)
class.summary$rsq
```

```
##  [1] 0.07279919 0.09555397 0.12812223 0.14603025 0.16130952 0.17290333
##  [7] 0.18432185 0.19036572 0.19240164 0.19259390
```

```
par(mfrow =c(2,2))
plot(class.summary$rss ,xlab=" Number of Variables ",ylab=" RSS", type="l")
plot(class.summary$adjr2 ,xlab =" Number of Variables ", ylab=" Adjusted RSq",type="l")
max.adjr2=which.max (class.summary$adjr2)
max.adjr2
```

```
## [1] 8
```

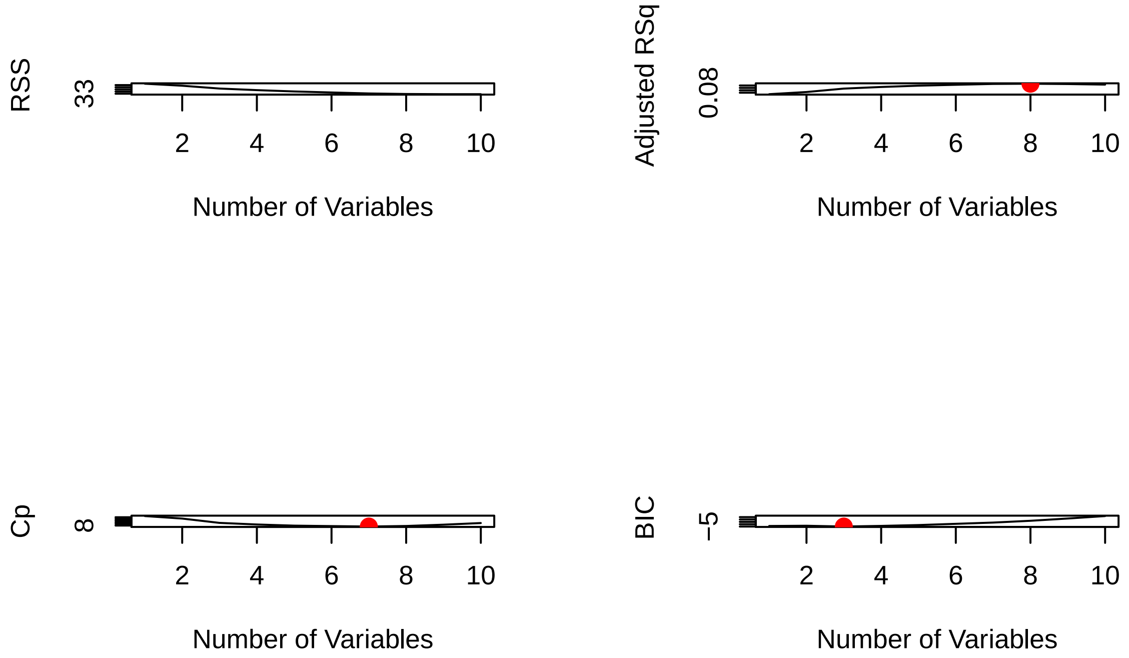
```
points (max.adjr2, class.summary$adjr2[max.adjr2], col ="red",cex =2, pch =20)


plot(class.summary$cp ,xlab =" Number of Variables ", ylab="Cp", type='l')
min.cp= which.min (class.summary$cp )
min.cp
```

```
## [1] 7
```

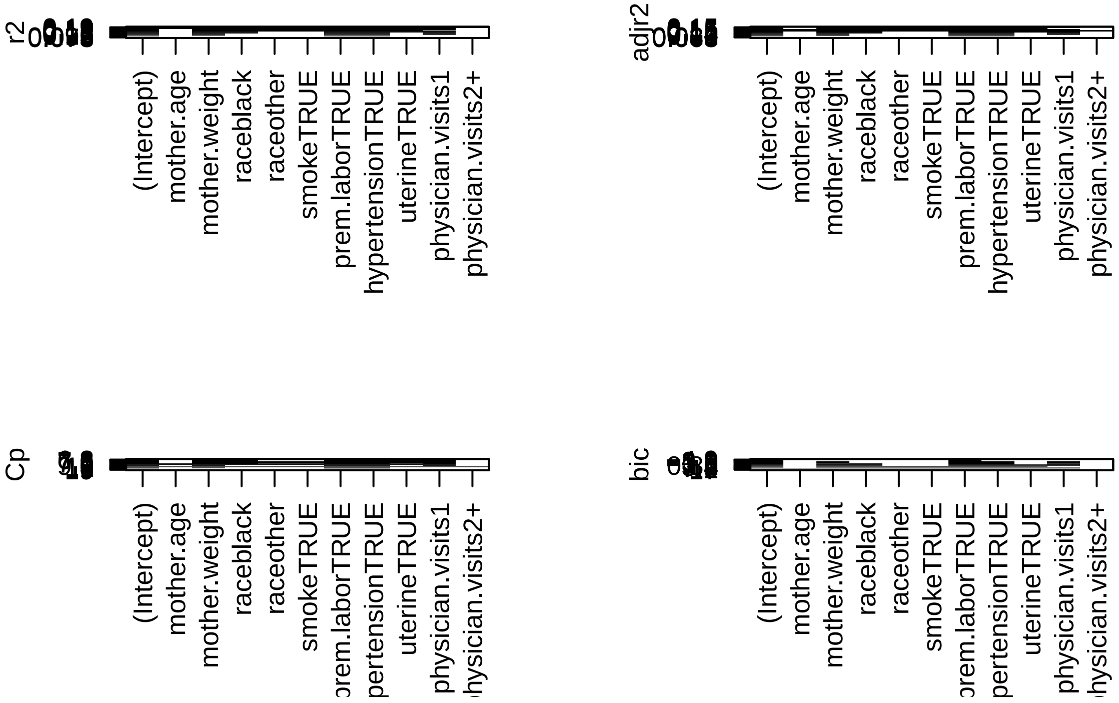
```
points (min.cp, class.summary$cp[min.cp], col ="red",cex =2, pch =20)


min.bic = which.min(class.summary$bic)
min.bic
```

```
## [1] 3
```

```
plot(class.summary$bic ,xlab=" Number of Variables ",ylab=" BIC", type='l')
points (min.bic, class.summary$bic [min.bic], col =" red",cex =2, pch =20)
```



```
plot(classfit.full ,scale ="r2")
plot(classfit.full ,scale ="adjr2")
plot(classfit.full ,scale ="Cp")
plot(classfit.full ,scale ="bic")
```



```
coef(classfit.full, max.adjr2)
```

```
##       (Intercept)     mother.weight         raceblack         raceother
##        0.47731870       -0.00269524        0.21446267        0.11814439
##         smokeTRUE    prem.laborTRUE  hypertensionTRUE       uterineTRUE
##        0.12582999        0.26509425        0.36294635        0.14095381
## physician.visits1
##       -0.08816014
```

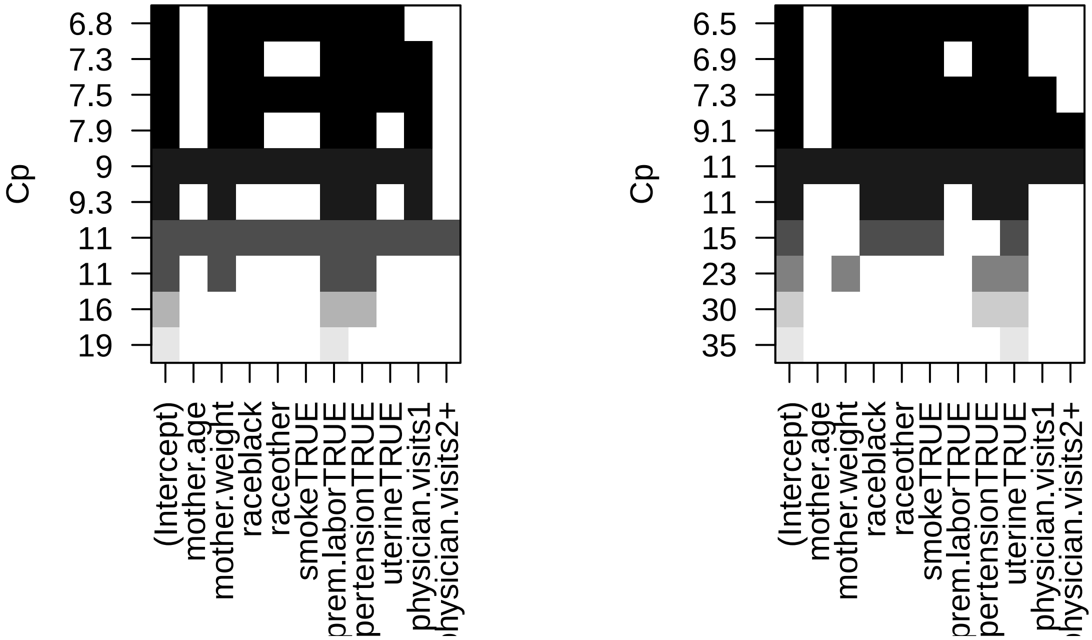
```
coef(classfit.full, min.cp)
```

```
##      (Intercept)    mother.weight        raceblack        raceother
##      0.426770003     -0.002589136      0.223081542      0.138356202
##        smokeTRUE   prem.laborTRUE hypertensionTRUE      uterineTRUE
##      0.147383151      0.245783235      0.360089535      0.149785677
```

```
coef(classfit.full, min.bic)
```

```
##      (Intercept)    mother.weight   prem.laborTRUE hypertensionTRUE
##      0.607928770     -0.002842709      0.313205471      0.370930320
```

```
#Let's compare classification and regression
par(mfrow =c(1,2))
plot(classfit.full ,scale ="Cp")
plot(regfit.full ,scale ="Cp")
```



```
#Logistic regression with the predictors selected by best subset
log.fit = glm( below.2500~ mother.weight+race+smoke+hypertension+uterine, family = binomial, data=bwt[t
summary(log.fit)
```

```
##
## Call:
## glm(formula = below.2500 ~ mother.weight + race + smoke + hypertension +
##     uterine, family = binomial, data = bwt[train, ])
```

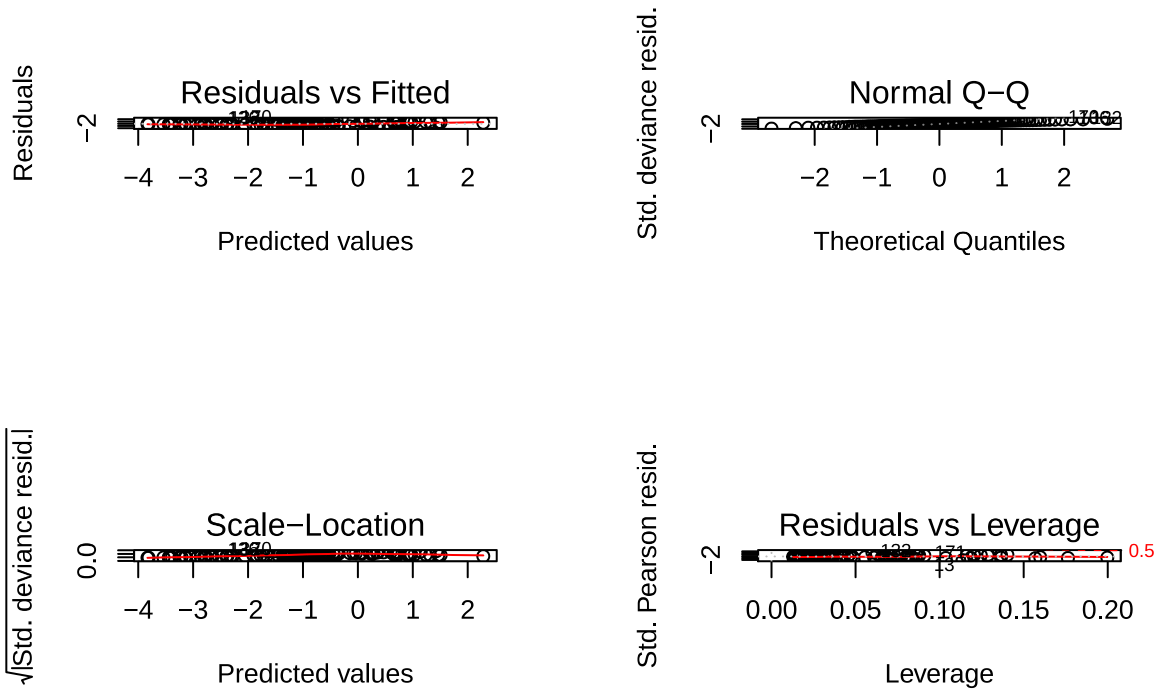
```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8144  -0.7984  -0.4335   0.8262   2.1800
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.143646   1.153607  -0.125  0.90090
## mother.weight    -0.019410   0.008422  -2.305  0.02119 *
## raceblack         1.671922   0.653759   2.557  0.01055 *
## raceother         1.395900   0.561766   2.485  0.01296 *
## smokeTRUE         1.543006   0.511265   3.018  0.00254 **
## hypertensionTRUE  2.023435   0.777850   2.601  0.00929 **
## uterineTRUE       1.041207   0.545658   1.908  0.05637 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 175.05  on 140  degrees of freedom
## Residual deviance: 142.72  on 134  degrees of freedom
## AIC: 156.72
##
## Number of Fisher Scoring iterations: 5
```

```r
confint(log.fit)
```

```
## Waiting for profiling to be done...
```

```
##                        2.5 %       97.5 %
## (Intercept)      -2.32618747  2.220864159
## mother.weight    -0.03731747 -0.004067736
## raceblack         0.41072752  3.005448309
## raceother         0.33601380  2.561460286
## smokeTRUE         0.58304495  2.610272893
## hypertensionTRUE  0.54469293  3.659423231
## uterineTRUE      -0.02878761  2.132040177
```

```r
par(mfrow = c(2, 2))
plot(log.fit)
```



```
pred.train <- predict(log.fit, type = "response")
low.train <- sapply(pred.train, function(x) {ifelse(x > 0.5, 1, 0)})
table(low.train, bwt$below.2500[train])
```

```
##
## low.train  0  1
##          0 87 25
##          1 10 19
```

```
mean(low.train == bwt$below.2500[train])
```

```
## [1] 0.751773
```

```
pred.test <- predict(log.fit, newdata = bwt[-train, -1], type = "response")
low.test <- sapply(pred.test, function(x) {ifelse(x > 0.5, 1, 0)})
table(low.test, bwt$below.2500[-train])
```

```
##
## low.test  0  1
##         0 29 11
##         1  4  4
```

```
mean(low.test == bwt$below.2500[-train])
```

```
## [1] 0.6875
```

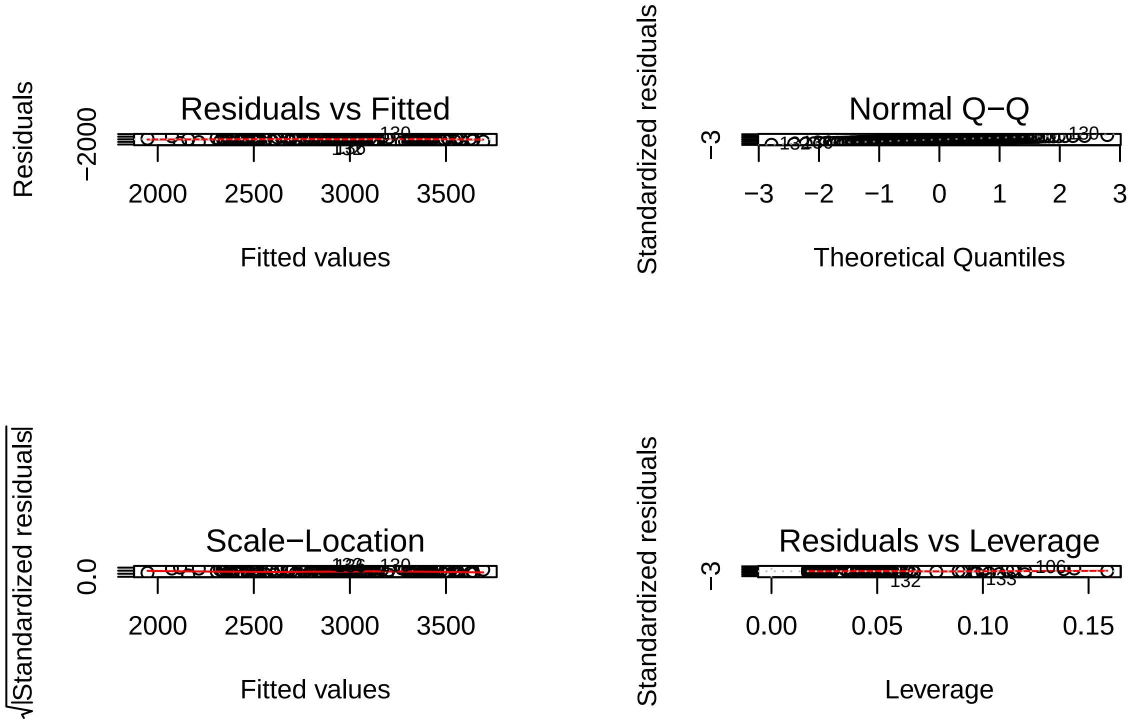
```
#Linear regression with the predictors selected by best subset
lm.fit = lm( baby.grams~ mother.weight+race+smoke+hypertension+uterine, data=bwt.grams)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = baby.grams ~ mother.weight + race + smoke + hypertension +
##     uterine, data = bwt.grams)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1842.14  -433.19    67.09   459.21  1631.03
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2837.264    243.676  11.644  < 2e-16 ***
## mother.weight       4.242      1.675   2.532 0.012198 *
## raceblack        -475.058    145.603  -3.263 0.001318 **
## raceother        -348.150    112.361  -3.099 0.002254 **
## smokeTRUE        -356.321    103.444  -3.445 0.000710 ***
## hypertensionTRUE -585.193    199.644  -2.931 0.003810 **
## uterineTRUE      -525.524    134.675  -3.902 0.000134 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 645.9 on 182 degrees of freedom
## Multiple R-squared:  0.2404, Adjusted R-squared:  0.2154
## F-statistic:   9.6 on 6 and 182 DF,  p-value: 3.601e-09
```

```
confint(lm.fit)
```

```
##                          2.5 %      97.5 %
## (Intercept)      2356.4706569 3318.057183
## mother.weight       0.9358509    7.547249
## raceblack        -762.3440159 -187.771193
## raceother        -569.8476393 -126.453123
## smokeTRUE        -560.4237850 -152.218115
## hypertensionTRUE -979.1080814 -191.278160
## uterineTRUE      -791.2496587 -259.798136
```

```
par(mfrow = c(2, 2))
plot(lm.fit)
```


Residuals vs Fitted
Residuals
−2000
130
2000
2500
3000
3500
Fitted values
Normal Q−Q
Standardized residuals
−3
130
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
0.0
130
2000
2500
3000
3500
Fitted values
Residuals vs Leverage
Standardized residuals
−3
106
132
133
0.00
0.05
0.10
0.15
Leverage