data_analysis_3.R

khagen

Sat Feb 22 11:30:50 2020

```r
# $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
#                              Data Analysis with R                              #
#                                Khagendra Adhikari                             #
#                                      2016                                      #
#                                                                               #
# $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$

#   Use packages
library(MASS)
library(ggplot2)
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##      as.Date, as.Date.numeric
```

```r
library(boot)
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'lattice'

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

```r
library(DAAG)
```

```
##
## Attaching package: 'DAAG'

## The following object is masked from 'package:MASS':
##
##      hills
```

```r
#                                  birthwt {MAAS}
#   Risk Factors Associated with Low Infant Birth Weight.
# ********************************************************************************
##   The birthwt data frame has 189 rows and 10 columns.
##   The data were collected at Baystate Medical Center, Springfield, Mass during 1986.
# ********************************************************************************

# structure of dataset birthwt.
str(birthwt)
```

```
## 'data.frame':    189 obs. of  10 variables:
##  $ low  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age  : int  19 33 20 21 18 21 22 17 29 26 ...
##  $ lwt  : int  182 155 105 108 107 124 118 103 123 113 ...
##  $ race : int  2 3 1 1 1 3 1 3 1 1 ...
##  $ smoke: int  0 0 1 1 1 0 0 0 1 1 ...
##  $ ptl  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ht   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ui   : int  1 0 0 1 1 0 0 0 0 0 ...
##  $ ftv  : int  0 3 1 2 0 0 1 1 1 0 ...
##  $ bwt  : int  2523 2551 2557 2594 2600 2622 2637 2637 2663 2665 ...
```

```
# low:    indicator of birth weight less than 2.5 kg.
#
# age:    mother's age in years.
#
# low:    mother's weight in pounds at last menstrual period.
#
# race:   mother's race (1 = white, 2 = black, 3 = other).
#
# smoke:  smoking status during pregnancy.
#
# ptl:    number of previous premature labours.
#
# ht:     history of hypertension.
#
# ui:     presence of uterine irritability.
#
# ftv:    number of physician visits during the first trimester.
#
# bwt:    birth weight in grams.



# ŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧ*********************************ŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧ
#                                      Ans.1                                         #
# ŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧ*********************************ŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧŧ
```

```r
attach(birthwt)
```

```
#   Attach dataset birthwt, so the objects in the database can be accessed
#   by simply giving their names.
```

```r
(race <- factor(race, labels = c("white", "black", "other")))
```

```
##   [1] black other white white white other white other white white other
##  [12] other other other white white black white other white other white
##  [23] white other other white white white black black black white black
##  [34] white black white white white white white black white black white
##  [45] white white white other white other white other white white other
##  [56] other other other other other other other other white other other
##  [67] other other white black white other other black white black white
##  [78] white black white white white other other other other other white
```
```

2

```
##   [89] white white white other white white white white white white white
##  [100] white white white other white other black white white white black
##  [111] white other white white white other white other white other white
##  [122] other white white white white white white white white other white
##  [133] black other other other other black other white white white other
##  [144] other white white black white other other other white white white
##  [155] white other black white black other white other other other black
##  [166] white other other white white black black black other other white
##  [177] white white white black other other white other white other other
##  [188] black white
## Levels: white black other
```

```
#   The factor function creates factor variable from numeric variable.
#   The levels in variable race has been modified as "white" for 1, "black"
#   for 2, and "other" for 3.

(ptd <- factor(ptl > 0))
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [12] FALSE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  TRUE
##  [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  TRUE
##  [67] FALSE FALSE TRUE  TRUE  TRUE  FALSE FALSE FALSE FALSE FALSE FALSE
##  [78] TRUE  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [89] FALSE FALSE FALSE FALSE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE
## [100] FALSE FALSE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [111] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [122] FALSE FALSE FALSE FALSE FALSE TRUE  FALSE FALSE FALSE TRUE  FALSE
## [133] FALSE TRUE  FALSE FALSE FALSE TRUE  FALSE FALSE FALSE TRUE  FALSE
## [144] FALSE FALSE FALSE FALSE TRUE  FALSE FALSE TRUE  FALSE FALSE TRUE
## [155] FALSE TRUE  FALSE TRUE  FALSE FALSE FALSE TRUE  FALSE FALSE TRUE
## [166] TRUE  TRUE  FALSE TRUE  FALSE FALSE FALSE FALSE FALSE FALSE TRUE
## [177] FALSE FALSE TRUE  FALSE TRUE  FALSE FALSE TRUE  FALSE FALSE FALSE
## [188] FALSE FALSE
## Levels: FALSE TRUE
```

```
#   Create new logic variable from  numeric variable  with
#   TRUE for  ptl > 0 and FALSE otherwise.

table(ftv)
```

```
## ftv
##   0   1   2   3   4   6
## 100  47  30   7   4   1
```

```
#   Display variable ftv in tabulated form. i.e. elements with corresponding frequency.

(ftv <- factor(ftv))
```

```
##   [1] 0 3 1 2 0 0 1 1 1 0 0 1 0 2 0 0 0 3 0 1 2 3 1 0 2 0 0 2 0 1 1 1 1 1 0
##  [36] 2 2 0 2 1 2 2 1 0 0 0 4 0 2 0 1 0 0 2 0 0 0 0 0 0 0 2 0 0 0 1 2 6 1 2
##  [71] 0 2 1 0 0 0 1 4 0 0 1 0 0 0 0 0 0 0 0 1 0 2 0 0 0 1 1 0 0 1 1 0 0 1 0
## [106] 0 1 0 2 4 2 1 2 1 0 1 0 0 2 1 1 0 1 0 2 2 1 0 1 1 0 2 0 0 0 0 1 1 0 1
## [141] 0 0 0 1 0 2 2 0 0 0 1 2 0 0 0 0 3 1 0 0 0 1 0 0 0 0 4 0 1 0 1 0 0 0 0
## [176] 0 1 3 0 2 1 3 0 0 2 2 0 0 3
```

```
## Levels: 0 1 2 3 4 6
#   Convert variable ftv of type numeric to factor.

levels(ftv) # Orignial levels.
```

```
## [1] "0" "1" "2" "3" "4" "6"
```

```
(levels(ftv)[-(1:2)] <- "2+")
```

```
## [1] "2+"
```

```
#   New level "2+" to all the levels of ftv except 1st ("0") and 2nd ("1") levels.
#   i.e. replace levels 2 or highter with new level "2+".

levels(ftv) # new levels after rename.
```

```
## [1] "0"  "1"  "2+"
```

```
table(ftv)
```

```
## ftv
##   0   1  2+
## 100  47  42
```

```
#   Display the factor variable ftv in tabulated form.

bw <- data.frame(low = factor(low), age, lwt, race, smoke = (smoke > 0),
                 ptd, ht = (ht > 0), ui = (ui > 0), ftv)

#   reate new data frame bw updaing variable "low" as a factor,
#   and variable "smoke", "ht", and "ui" as logic.

detach()

#   Detach the attachment birthwt such that dataset birthwt should be
#   referenced in order to access the variable. i.e. birthwt$race.

rm(race, ptd, ftv)

#   Remove objects in lists (race, ptd, ftv).


# ʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕ*********************************ʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕ
#                                          Ans.2                                      #
# ʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕ*********************************ʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕʕ

model <- glm(low ~ age * ftv, family = binomial(link = logit), data = bw)
summary(model)
```

```
##
## Call:
## glm(formula = low ~ age * ftv, family = binomial(link = logit),
##     data = bw)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5691  -0.8951  -0.7343   1.3331   2.2096
```

```
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.52863    1.03326  -1.479  0.13903
## age          0.04320    0.04563   0.947  0.34376
## ftv1         1.61096    1.89076   0.852  0.39420
## ftv2+        6.39702    2.29784   2.784  0.00537 **
## age:ftv1    -0.09550    0.07940  -1.203  0.22907
## age:ftv2+   -0.29211    0.10215  -2.860  0.00424 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 219.78  on 183  degrees of freedom
## AIC: 231.78
## 
## Number of Fisher Scoring iterations: 5
#   Yes, the data from this study confirm that ftv is an important
#   predicator of low birthweight. The interation effect is significant
#   between age of mother and "2+" number of physician visits during
#   the first trimester.


# ££££££££££££££££££££££££££££££££*****************************************££££££££££££££££££££££££££
#                                       Ans.3                                                     #
# ££££££££££££££££££££££££££££££££*****************************************££££££££££££££££££££££££££

ftv.level <- levels(bw$ftv)

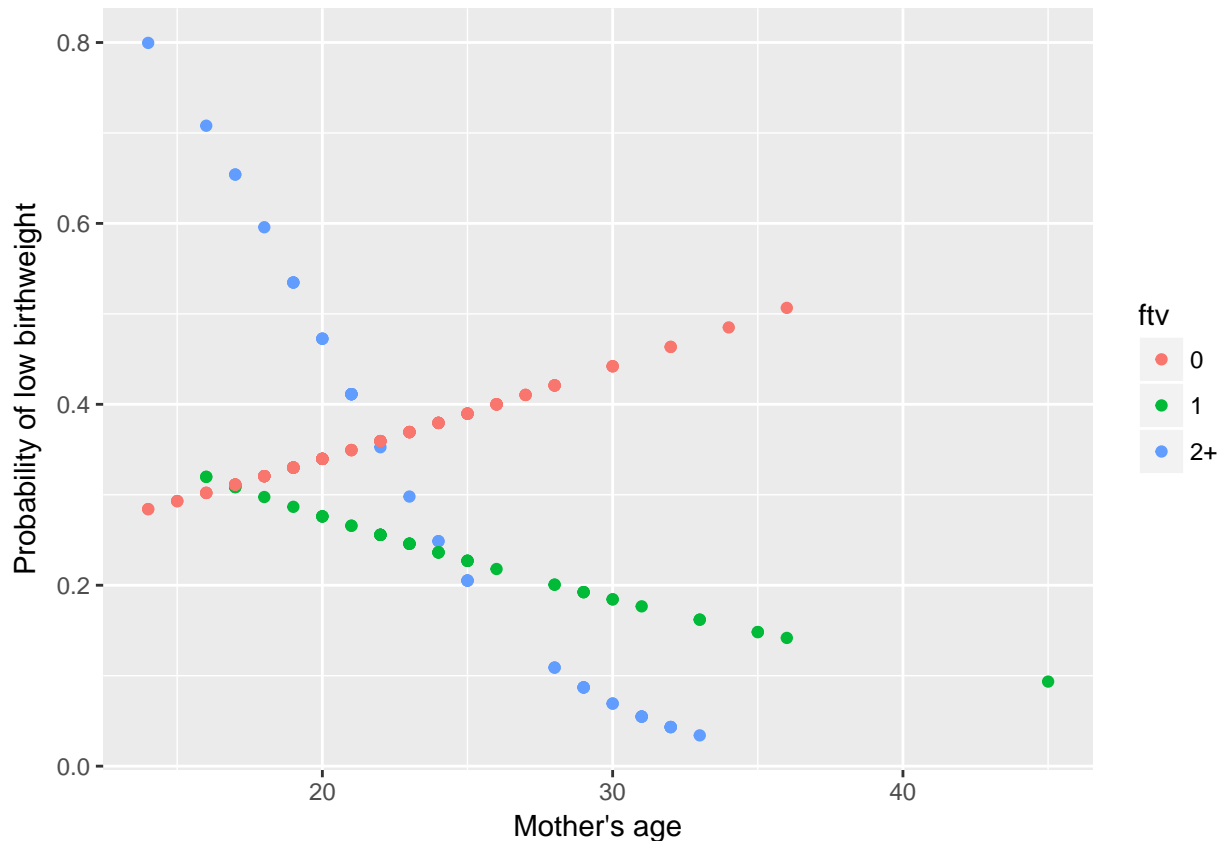est.prob.age20 <- predict.glm(model,
                    newdata = data.frame(age = 20, ftv = ftv.level),
                    type = "response", se.fit=TRUE)

est.prob.age30 <- predict.glm(model,
                    newdata = data.frame(age = 30, ftv = ftv.level),
                    type = "response", se.fit=TRUE)

#   Estimated probability of low birthweight for 20 and 30 years mothers
#   for each ftv-catagories
(est.prob.low.bw <- data.frame(ftv = ftv.level, est.prob.age20$fit, est.prob.age30$fit))

##   ftv est.prob.age20.fit est.prob.age30.fit
## 1   0          0.3396891         0.44208491
## 2   1          0.2761189         0.18439381
## 3  2+          0.4725540         0.06920055

#   Fitted values vs. age plot for each ftv-grups.
ggplot(bw, aes(bw$age, model$fitted.values, colour = ftv))+
    geom_point()+
    labs(x = "Mother's age", y = "Probability of low birthweight")
```

```
#   @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
#   Interpretation based on the plot: #
#   @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

##   The mothers, who had no physician visit (ftv = 0), have a chance of having
##   uderweight baby is low but increase with the mother's age.

##   The mothers, who had one physician visit (ftv = 1), have a chance of having
##   uderweight baby is low and also decrease with the mother's age.

##   The young mothers, who had 2+ physician visits (ftv = 2+), have a higher chance of
##   having uderweight baby and but decrease sharply with the mother's age.


# ϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾ*******************************************ϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾ
#                                             Ans.4                                       #
# ϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾ*******************************************ϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾϾ

p30 <- est.prob.age30
lower.limit <- p30$fit - 1.96*p30$se.fit
upper.limit <- p30$fit + 1.96*p30$se.fit

#   95% cofidence interval of PI for 30 years old women in each ftv group.
(conf.interval <- data.frame( ftv.level, p30$fit, lower.limit, upper.limit))

##   ftv.level    p30.fit lower.limit upper.limit
## 1         0 0.44208491  0.24204902   0.6421208
```

```
## 2         1 0.18439381  0.02727720   0.3415104
## 3        2+ 0.06920055 -0.03508703   0.1734881

#   The probabilty can't be zero.
#   Set the lower.limit  to zero if it is negative.

conf.interval$lower.limit[conf.interval$lower.limit < 0] <- 0.0
conf.interval

##   ftv.level    p30.fit lower.limit upper.limit
## 1         0 0.44208491   0.2420490   0.6421208
## 2         1 0.18439381   0.0272772   0.3415104
## 3        2+ 0.06920055   0.0000000   0.1734881
# %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%************************************%%%%%%%%%%%%%%%%%%%%%%%%%%
#                                            Ans.5                                                 #
# %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%************************************%%%%%%%%%%%%%%%%%%%%%%%%%%

#    *****************************?????????????????????????????*****************************
##   The likelihood ratio test (LR-test) can be performed either using lrtest()
##   from lmtest package or using anova() function from base R. The reduced
##   model has less log likelihood but it is necessary to test whether the observed
##   difference is significan or not?. Here, the null hypothesis is that the
##   reduced model is true.
#    ***************************************************************************************
# %%%%%%%%%
# Ans.5.a #***************************************************************************************
# %%%%%%%%%

# model.a includes all explanatory variables from datset bw.
model.a <- glm(low ~ ., family = binomial(link = logit), data = bw)

# %%%%%%%%%
# Ans.5.b #***************************************************************************************
# %%%%%%%%%

#   model.b include interaction effect to model.a.
model.b <- update(model.a,~.+age:ftv)

#   compare two models uisng anova. We use chi-squared test as our
#   models are binomial fits.
anova(model.a, model.b, test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: low ~ age + lwt + race + smoke + ptd + ht + ui + ftv
## Model 2: low ~ age + lwt + race + smoke + ptd + ht + ui + ftv + age:ftv
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       178     195.48
## 2       176     183.00  2   12.475 0.001955 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# Alternative method gives the same result.
lrtest(model.a, model.b)

## Likelihood ratio test
```

```
##
## Model 1: low ~ age + lwt + race + smoke + ptd + ht + ui + ftv
## Model 2: low ~ age + lwt + race + smoke + ptd + ht + ui + ftv + age:ftv
##    #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   11 -97.738
## 2   13 -91.500  2 12.475   0.001955 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#   Here, p value shows that the model.b does singnificantly improve the model fit.

# %%%%%%%%%
# Ans.5.c #***********************************************************************
# %%%%%%%%%

summary(model.a)
```

```
##
## Call:
## glm(formula = low ~ ., family = binomial(link = logit), data = bw)
##
## Deviance Residuals:
##     Min      1Q   Median      3Q      Max
## -1.7038  -0.8068  -0.5008   0.8835   2.2152
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.82302    1.24471   0.661  0.50848
## age         -0.03723    0.03870  -0.962  0.33602
## lwt         -0.01565    0.00708  -2.211  0.02705 *
## raceblack    1.19241    0.53597   2.225  0.02609 *
## raceother    0.74069    0.46174   1.604  0.10869
## smokeTRUE    0.75553    0.42502   1.778  0.07546 .
## ptdTRUE      1.34376    0.48062   2.796  0.00518 **
## htTRUE       1.91317    0.72074   2.654  0.00794 **
## uiTRUE       0.68019    0.46434   1.465  0.14296
## ftv1        -0.43638    0.47939  -0.910  0.36268
## ftv2+        0.17901    0.45638   0.392  0.69488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 195.48  on 178  degrees of freedom
## AIC: 217.48
##
## Number of Fisher Scoring iterations: 4
#   From summary of model.a, we can see that explanatory variables
#   "lwt", "race", "ptd", and "ht"  had a significant efffect
#   using singificance level 0.05.

model.c <- glm(low ~ lwt + race+ ptd + ht,
               family = binomial(link = logit), data = bw)
```

```
#   compare model.c with  model.a using anova.
anova(model.c, model.a, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: low ~ lwt + race + ptd + ht
## Model 2: low ~ age + lwt + race + smoke + ptd + ht + ui + ftv
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       183     205.40
## 2       178     195.48  5   9.9217  0.07749 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#   compare model.c with model.a using anova.
anova(model.c, model.b, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: low ~ lwt + race + ptd + ht
## Model 2: low ~ age + lwt + race + smoke + ptd + ht + ui + ftv + age:ftv
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       183     205.4
## 2       176     183.0  7   22.397  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#   According to LR-test, model.b is preferred.
AIC(model.a, model.b, model.c)
```

```
##         df      AIC
## model.a 11 217.4755
## model.b 13 209.0006
## model.c  6 217.3972
```

```
#   According to AIC, model.b with smalest AIC is the best.

#   The residuals also suggest model.b but they are not signigicant enough
#   to select one model over other. i.e ther are of same order.



# %%%%%%%%
# Ans.5.d #************************************************************************
# %%%%%%%%

model.d <- stepAIC(model.a, scope = ~. ^2)
```

```
## Start:  AIC=217.48
## low ~ age + lwt + race + smoke + ptd + ht + ui + ftv
##
##               Df Deviance    AIC
## + age:ftv      2   183.00 209.00
## - ftv          2   196.83 214.83
## - age          1   196.42 216.42
## <none>             195.48 217.48
## - ui           1   197.59 217.59
## + smoke:ui     1   193.76 217.76
```

```
## + lwt:smoke    1    194.04 218.04
## + ptd:ui       1    194.24 218.24
## + lwt:ui       1    194.28 218.28
## + ptd:ftv      2    192.38 218.38
## + ptd:ht       1    194.54 218.54
## + age:ptd      1    194.58 218.58
## + age:ht       1    194.59 218.59
## + age:smoke    1    194.61 218.61
## + race:ui      2    192.63 218.63
## - smoke        1    198.67 218.67
## + smoke:ht     1    195.03 219.03
## + smoke:ptd    1    195.16 219.16
## - race         2    201.23 219.23
## + race:smoke   2    193.24 219.24
## + lwt:ptd      1    195.35 219.35
## + lwt:ht       1    195.44 219.44
## + age:lwt      1    195.46 219.46
## + age:ui       1    195.47 219.47
## + ht:ftv       2    193.99 219.99
## + lwt:ftv      2    194.19 220.19
## + smoke:ftv    2    194.47 220.47
## + age:race     2    194.58 220.58
## + lwt:race     2    194.63 220.63
## + race:ptd     2    194.83 220.83
## - lwt          1    200.95 220.95
## + race:ht      2    195.19 221.19
## + ui:ftv       2    195.32 221.32
## - ht           1    202.93 222.93
## - ptd          1    203.58 223.58
## + race:ftv     4    193.81 223.81
##
## Step:  AIC=209
## low ~ age + lwt + race + smoke + ptd + ht + ui + ftv + age:ftv
##
##                Df Deviance    AIC
## + smoke:ui     1    179.94 207.94
## + lwt:smoke    1    180.89 208.89
## - race         2    186.99 208.99
## <none>              183.00 209.00
## + ptd:ui       1    181.42 209.42
## + lwt:ui       1    181.90 209.90
## + ptd:ht       1    182.05 210.05
## - smoke        1    186.11 210.11
## + age:smoke    1    182.16 210.16
## + race:ui      2    180.32 210.32
## + age:ptd      1    182.50 210.50
## - ui           1    186.61 210.61
## + smoke:ht     1    182.71 210.71
## + lwt:ptd      1    182.75 210.75
## + smoke:ptd    1    182.81 210.81
## + age:ht       1    182.90 210.90
## + age:ui       1    182.96 210.96
## + age:lwt      1    183.00 211.00
## + lwt:ht       1    183.00 211.00
```

```
## + race:smoke    2    181.23 211.23
## + lwt:ftv       2    181.44 211.44
## + ptd:ftv       2    181.57 211.57
## + age:race      2    181.62 211.62
## + smoke:ftv     2    181.65 211.65
## + ht:ftv        2    181.81 211.81
## + lwt:race      2    182.55 212.55
## + race:ht       2    182.78 212.78
## + race:ptd      2    182.85 212.85
## - lwt           1    188.88 212.88
## + ui:ftv        2    182.94 212.94
## - ht            1    190.13 214.13
## - ptd           1    191.05 215.05
## + race:ftv      4    181.69 215.69
## - age:ftv       2    195.48 217.48
##
## Step:  AIC=207.94
## low ~ age + lwt + race + smoke + ptd + ht + ui + ftv + age:ftv +
##     smoke:ui
##
##                 Df Deviance    AIC
## - race           2    183.07 207.07
## <none>                179.94 207.94
## + lwt:smoke      1    178.34 208.34
## + ptd:ht         1    178.88 208.88
## - smoke:ui       1    183.00 209.00
## + ptd:ui         1    179.07 209.07
## + age:ptd        1    179.35 209.35
## + age:smoke      1    179.37 209.37
## + smoke:ptd      1    179.58 209.58
## + lwt:ptd        1    179.61 209.61
## + lwt:ui         1    179.76 209.76
## + age:ht         1    179.78 209.78
## + smoke:ht       1    179.82 209.82
## + age:lwt        1    179.84 209.84
## + age:ui         1    179.86 209.86
## + lwt:ht         1    179.94 209.94
## + lwt:ftv        2    178.25 210.25
## + ptd:ftv        2    178.53 210.53
## + smoke:ftv      2    178.64 210.64
## + race:smoke     2    178.73 210.73
## + age:race       2    178.84 210.84
## + ht:ftv         2    178.88 210.88
## + race:ui        2    179.13 211.13
## + ui:ftv         2    179.50 211.50
## + race:ht        2    179.52 211.52
## + lwt:race       2    179.68 211.68
## + race:ptd       2    179.86 211.86
## - lwt            1    187.15 213.15
## - ht             1    187.66 213.66
## + race:ftv       4    178.51 214.51
## - ptd            1    188.83 214.83
## - age:ftv        2    193.76 217.76
##
```

```
## Step:  AIC=207.07
## low ~ age + lwt + smoke + ptd + ht + ui + ftv + age:ftv + smoke:ui
##
##               Df Deviance    AIC
## <none>              183.07 207.07
## + lwt:smoke  1     181.40 207.40
## + ptd:ui     1     181.88 207.88
## + ptd:ht     1     181.92 207.92
## + race       2     179.94 207.94
## + age:smoke  1     181.97 207.97
## + age:ht     1     182.64 208.64
## + age:ptd    1     182.69 208.69
## + lwt:ptd    1     182.73 208.73
## + lwt:ui     1     182.76 208.76
## + smoke:ptd  1     182.85 208.85
## + age:lwt    1     182.92 208.92
## - smoke:ui   1     186.99 208.99
## + age:ui     1     182.99 208.99
## + smoke:ht   1     183.02 209.02
## + lwt:ht     1     183.06 209.06
## + smoke:ftv  2     181.48 209.48
## + lwt:ftv    2     181.69 209.69
## + ptd:ftv    2     181.85 209.85
## + ui:ftv     2     182.28 210.28
## + ht:ftv     2     182.41 210.41
## - ht         1     191.21 213.21
## - lwt        1     191.56 213.56
## - ptd        1     193.59 215.59
## - age:ftv    2     199.00 219.00
```

```r
#   The final model with smallest AIC is
#   Step:  AIC=207.94
#   low ~ age + lwt + race + smoke + ptd + ht + ui + ftv + age:ftv + smoke:ui
#   Based on AIC value, we can conclude that this is the optimal model.

optimal.m <- glm(low ~ age + lwt + race + smoke + ptd + ht + ui + ftv + age:ftv + smoke:ui,
                 family = binomial(link = logit), data = bw)


# %%%%%%%%%
# Ans.5.e #***********************************************************************
# %%%%%%%%%


#   The prediction accuracy based on 10-fold cross validation is
CVbinary(optimal.m, print.details = FALSE)$acc.cv
```

```
## [1] 0.7195767
```

```r
# Average of 10 prediction accuracy rates based on 5-fold cross validation.
acc.cv.list <- NULL
for (j in 1:10)
  {
    acc.cv.list[j] <- CVbinary(optimal.m, nfolds = 5,
                               print.details = FALSE)$acc.cv
  }
ave.predict.acc <- mean(acc.cv.list)
```

```r
paste("Average predication accuracy is", round(mean(ave.predict.acc)*100, 1), "%.")
```

```
## [1] "Average predication accuracy is 71.7 %."
# @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
```