Jin Kweon (3032235207) HW3

Jin Kweon

10/23/2017

```
data <- read.table(file="Data-HW3-CHeartDisease.dat", header=FALSE, quote="", sep=",")
dim(data)
```

```
## [1] 303  14
```

Professor allowed us to use the R codes he gave for hints!!!

Ref: http://archive.ics.uci.edu/ml/datasets/heart+Disease

Remove missing values

```
#remove missing values with "?" in 12th and 13th columns
id.ms <- sort(c(seq(1, nrow(data))[data[, 12] == "?"], seq(1, nrow(data))[data[, 13] == "?"]))
data2 <- data[-id.ms, ]

#Change the data into numeric.
data2[, 12] <- as.numeric(data2[, 12]) - 2 #factor messed up numbers
data2[, 13] <- as.numeric(data2[, 13]) - 1 #factor messed up numbers

#Set X and Y
design <- data.matrix(data2[, 1:13])
response <- data2[, 14]

#The goal is to distinguish presence (values 1,2,3,4) from absence (value 0) of
#heart disease given the 13 attributes
response[response > 0] <- 1


colnames(design) <- c("age", "gender", "chestpain", "bldpressure", "cholestroal",
                      "bloodsugar", "electrocardio", "maxheartrate", "angina", "STdepression",
                      "STslopepeakexercise", "vessel", "thal")
response <- as.matrix(response)

colnames(response) <- "heartdisease"

dim(design)
```

```
## [1] 297  13
```

```
dim(response)
```

```
## [1] 297   1
```

Comment:

We removed 6 “?” (missing values) on the 12th and 13th columns. Since the repsponse should be categorical, I transformed the values 1 - 4 into 1 and 0 into 0. That way, I could do logisitc regression. (one of classification)

So the values on the 14th column were all dummy numbers.

Part a

```
response <- as.data.frame(response)

count(response, "heartdisease")
```

```
##   heartdisease freq
## 1            0  160
## 2            1  137
```

Comment:

Q.Among the 297 remaining patients, how many had heart disease and how many had none?

So, from the dplyr outputs, I can say that 137 people had heart disease, and 160 people had none.

Q.Which predictors are numerical, which are categorical, and which are unclear?

1. The first column, age, is numerical.
2. The second column, gender, is categorical. (0 and 1 do not have numerical meaning)
3. The third column, chest pain type, is little bit unclear but closed to categorical (the number represents the characteristic, but also can be measured) - part c) says we are going to regard this as categofical variables.
4. The fourth column, resting blood pressure, is numerical.
5. The fifth column, serum cholestoral, is numerical.
6. The sixth column, fasting blood sugar, is categorical.
7. The seventh column, resting electrocardiographic result, is unclear. (0, 1, and 2) - part c) says we are going to regard this as numerical variables. (closed to categorical) - don't know whether 0, 1, and 2 measure something or random numbers.
8. The eigth column, maximum heart rate achieved, is numerical.
9. The ninth column, exercise induced angina, is categorical.

10. The tenth column, ST depression induced by exercise, is numerical. (has decimal points)
11. The eleventh column, the slope of the peak exercise ST segment, is unclear. (1, 2, and 3) - part c) says we are going to regard this as numerical variables. (little bit closed to categorical) - don't know whether 1, 2, and 3 measure something or random numbers.
12. The twelvth column, number of major vessels colored by flourosopy, is numerical. -> R: The number of vessels actually means the number.
13. The thirteenth column, thal, is categorical. - part c) says we are going to regard this as categorical variables.
14. Definitely, response (presence/absence of heart disease) is categorical, as 0/1 does not numerically mean anything.

Most of them are fairly clear.

Part b

```r
logfit <- glm(heartdisease ~., family = binomial, data = cbind(design, response))

summary(logfit)
```

```
##
## Call:
## glm(formula = heartdisease ~ ., family = binomial, data = cbind(design,
##     response))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8042  -0.5263  -0.1860   0.4161   2.3676
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -6.993701   2.893938  -2.417  0.01566 *
## age                  -0.014057   0.024036  -0.585  0.55866
## gender                1.319688   0.486718   2.711  0.00670 **
## chestpain             0.578582   0.191335   3.024  0.00250 **
## bldpressure           0.024182   0.010727   2.254  0.02418 *
## cholestroal           0.004816   0.003775   1.276  0.20202
## bloodsugar           -0.991868   0.554947  -1.787  0.07389 .
## electrocardio         0.246117   0.185238   1.329  0.18396
## maxheartrate        -0.021183   0.010275  -2.062  0.03923 *
## angina               0.915651   0.414003   2.212  0.02699 *
## STdepression         0.249909   0.212418   1.176  0.23940
## STslopepeakexercise  0.582699   0.362317   1.608  0.10778
## vessel               1.267008   0.265723   4.768 1.86e-06 ***
## thal                 0.714003   0.202068   3.533  0.00041 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 409.95  on 296  degrees of freedom
## Residual deviance: 203.86  on 283  degrees of freedom
## AIC: 231.86
##
## Number of Fisher Scoring iterations: 6
```

Comment:

It shows that vessel and thal are significant.

Part c

Reference: https://cran.r-project.org/web/packages/dummies/dummies.pdf

```
#See how many groups are in 3rd and 13th columns
levels(as.factor(design[,3]))
```

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

```
levels(as.factor(design[,13]))
```

```
## [1] "1" "2" "3"
```

```
#Change to dummy matrix for these two columns
col3chestdummy <- dummy(design[,3])
col13thaldummy <- dummy(design[,13])

colnames(col3chestdummy) <- c("chest1", "chest2", "chest3", "chest4")
colnames(col13thaldummy) <- c("thal1", "thal2", "thal3")

head(col3chestdummy)
```

```
##      chest1 chest2 chest3 chest4
## [1,]      1      0      0      0
## [2,]      0      0      0      1
## [3,]      0      0      0      1
## [4,]      0      0      1      0
## [5,]      0      1      0      0
## [6,]      0      1      0      0
```

```r
head(col13thaldummy)
```

```
##      thal1 thal2 thal3
## [1,]     0     1     0
## [2,]     1     0     0
## [3,]     0     0     1
## [4,]     1     0     0
## [5,]     1     0     0
## [6,]     1     0     0
```

```r
newdesign <- cbind(design[,c(1,2)], col3chestdummy, design[,4:12], col13thaldummy)
newdesignonecolout <- newdesign[,-c(6, 18)]

#Fitting
logfit2 <- glm(heartdisease ~., family = binomial, data = cbind(newdesign, response))
summary(logfit2)
```

```
## 
## Call:
## glm(formula = heartdisease ~ ., family = binomial, data = cbind(newdesign, 
##     response))
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.7145  -0.5436  -0.1444   0.3264   2.7316  
## 
## Coefficients: (2 not defined because of singularities)
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         -2.537914   2.807710  -0.904 0.366044    
## age                 -0.012296   0.024664  -0.499 0.618120    
## gender               1.431422   0.513185   2.789 0.005282 ** 
## chest1              -2.006802   0.652608  -3.075 0.002105 ** 
## chest2              -0.935649   0.556725  -1.681 0.092835 .  
## chest3              -1.804627   0.492607  -3.663 0.000249 ***
## chest4                     NA         NA      NA       NA    
## bldpressure          0.023981   0.011110   2.159 0.030889 *  
## cholestroal          0.004930   0.003944   1.250 0.211306    
## bloodsugar          -0.610758   0.599184  -1.019 0.308052    
## electrocardio        0.255433   0.189565   1.347 0.177829    
## maxheartrate        -0.021281   0.010821  -1.967 0.049224 *  
## angina               0.739431   0.434687   1.701 0.088931 .  
## STdepression         0.353095   0.230102   1.535 0.124903    
## STslopepeakexercise  0.670508   0.371616   1.804 0.071184 .  
## vessel               1.269290   0.271304   4.678 2.89e-06 ***
## thal1               -1.441377   0.418558  -3.444 0.000574 ***
## thal2               -1.429947   0.783279  -1.826 0.067912 .  
## thal3                      NA         NA      NA       NA    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 409.95  on 296  degrees of freedom
## Residual deviance: 194.83  on 280  degrees of freedom
## AIC: 228.83
##
## Number of Fisher Scoring iterations: 6
```

```
logfit3 <- glm(heartdisease ~., family = binomial, data = cbind(newdesignonecolout, response))
summary(logfit3)
```

```
##
## Call:
## glm(formula = heartdisease ~ ., family = binomial, data = cbind(newdesignonecolout,
##     response))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7145  -0.5436  -0.1444   0.3264   2.7316
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.537914   2.807710  -0.904 0.366044
## age                  -0.012296   0.024664  -0.499 0.618120
## gender                1.431422   0.513185   2.789 0.005282 **
## chest1               -2.006802   0.652608  -3.075 0.002105 **
## chest2               -0.935649   0.556725  -1.681 0.092835 .
## chest3               -1.804627   0.492607  -3.663 0.000249 ***
## bldpressure           0.023981   0.011110   2.159 0.030889 *
## cholestroal           0.004930   0.003944   1.250 0.211306
## bloodsugar           -0.610758   0.599184  -1.019 0.308052
## electrocardio         0.255433   0.189565   1.347 0.177829
## maxheartrate         -0.021281   0.010821  -1.967 0.049224 *
## angina                0.739431   0.434687   1.701 0.088931 .
## STdepression          0.353095   0.230102   1.535 0.124903
## STslopepeakexercise   0.670508   0.371616   1.804 0.071184 .
## vessel                1.269290   0.271304   4.678 2.89e-06 ***
## thal1                -1.441377   0.418558  -3.444 0.000574 ***
## thal2                -1.429947   0.783279  -1.826 0.067912 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 409.95  on 296  degrees of freedom
## Residual deviance: 194.83  on 280  degrees of freedom
## AIC: 228.83
##
## Number of Fisher Scoring iterations: 6
```

```
newdesignonecolout2 <- newdesign[,-c(3, 16)]

logfit4 <- glm(heartdisease ~., family = binomial, data = cbind(newdesignonecolout2, response))
summary(logfit4)
```

```
##
## Call:
## glm(formula = heartdisease ~ ., family = binomial, data = cbind(newdesignonecolout2,
##     response))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7145  -0.5436  -0.1444   0.3264   2.7316
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -5.986093   2.938058  -2.037 0.041607 *
## age                 -0.012296   0.024664  -0.499 0.618120
## gender               1.431422   0.513185   2.789 0.005282 **
## chest2               1.071153   0.753902   1.421 0.155371
## chest3               0.202175   0.648718   0.312 0.755304
## chest4               2.006802   0.652608   3.075 0.002105 **
## bldpressure          0.023981   0.011110   2.159 0.030889 *
## cholestroal          0.004930   0.003944   1.250 0.211306
## bloodsugar          -0.610758   0.599184  -1.019 0.308052
## electrocardio        0.255433   0.189565   1.347 0.177829
## maxheartrate        -0.021281   0.010821  -1.967 0.049224 *
## angina               0.739431   0.434687   1.701 0.088931 .
## STdepression         0.353095   0.230102   1.535 0.124903
## STslopepeakexercise  0.670508   0.371616   1.804 0.071184 .
## vessel               1.269290   0.271304   4.678 2.89e-06 ***
## thal2                0.011430   0.795090   0.014 0.988530
## thal3                1.441377   0.418558   3.444 0.000574 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 409.95  on 296  degrees of freedom
## Residual deviance: 194.83  on 280  degrees of freedom
## AIC: 228.83
##
## Number of Fisher Scoring iterations: 6
```

```r
newdesignonecolout3 <- newdesign[,-c(4, 16)]


logfit5 <- glm(heartdisease ~., family = binomial, data = cbind(newdesignonecolout3, response))
summary(logfit5)
```

```
##
## Call:
## glm(formula = heartdisease ~ ., family = binomial, data = cbind(newdesignonecolout3,
##     response))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7145  -0.5436  -0.1444   0.3264   2.7316
##
## Coefficients:
```

```
##                       Estimate Std. Error z value Pr(>|z|)    
## (Intercept)          -4.914940   2.846093  -1.727 0.084184 .  
## age                  -0.012296   0.024664  -0.499 0.618120    
## gender                1.431422   0.513185   2.789 0.005282 ** 
## chest1               -1.071153   0.753902  -1.421 0.155371    
## chest3               -0.868978   0.616290  -1.410 0.158535    
## chest4                0.935649   0.556725   1.681 0.092835 .  
## bldpressure           0.023981   0.011110   2.159 0.030889 *  
## cholestroal           0.004930   0.003944   1.250 0.211306    
## bloodsugar           -0.610758   0.599184  -1.019 0.308052    
## electrocardio         0.255433   0.189565   1.347 0.177829    
## maxheartrate         -0.021281   0.010821  -1.967 0.049224 *  
## angina                0.739431   0.434687   1.701 0.088931 .  
## STdepression          0.353095   0.230102   1.535 0.124903    
## STslopepeakexercise  0.670508   0.371616   1.804 0.071184 .  
## vessel                1.269290   0.271304   4.678 2.89e-06 ***
## thal2                 0.011430   0.795090   0.014 0.988530    
## thal3                 1.441377   0.418558   3.444 0.000574 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 409.95  on 296  degrees of freedom
## Residual deviance: 194.83  on 280  degrees of freedom
## AIC: 228.83
## 
## Number of Fisher Scoring iterations: 6
```

Comment:

Since glm() function automatically creates an intercept, one of the columns from two dummy matrices will give NA (singularity errors). To fix them, I can manually, take out one column each for two dummy matrices manually. When I manually take each of them out, I can take out any one column in each dummy matrix; however, just to follow how R works, I will take out the last dummy column in each dummy variable. (Otherwise, the coefficients are different, but means the same thing though)

Edited:

Since part e) asked us to interpret the coefficient estimate in front of the predictor [3] chest pain type 4, I decide to take out the first column for each dummy variable (part d), e), and f) literally said "Based on the logistic regression fit from (c)." So, I will stick to this regression fit for the following questions)

Part d

```r
summary(logfit4)$coefficients
```

```
##                         Estimate  Std. Error     z value     Pr(>|z|)
## (Intercept)         -5.986093033 2.938057804 -2.03743202 4.160677e-02
## age                 -0.012295651 0.024664460 -0.49851694 6.181197e-01
## gender               1.431422404 0.513184959  2.78929142 5.282351e-03
## chest2               1.071152889 0.753901685  1.42081244 1.553713e-01
## chest3               0.202174825 0.648717977  0.31165288 7.553043e-01
## chest4               2.006801827 0.652608055  3.07504912 2.104679e-03
## bldpressure          0.023981075 0.011110050  2.15850292 3.088875e-02
## cholestroal          0.004930394 0.003944373  1.24998181 2.113062e-01
## bloodsugar          -0.610758340 0.599183766 -1.01931724 3.080524e-01
## electrocardio        0.255433049 0.189564760  1.34747117 1.778285e-01
## maxheartrate        -0.021280893 0.010820888 -1.96664943 4.922364e-02
## angina               0.739430871 0.434686711  1.70106620 8.893056e-02
## STdepression         0.353095180 0.230102261  1.53451417 1.249032e-01
## STslopepeakexercise  0.670508183 0.371615790  1.80430488 7.118352e-02
## vessel               1.269289907 0.271304027  4.67847795 2.890122e-06
## thal2                0.011429945 0.795089870  0.01437566 9.885303e-01
## thal3                1.441377014 0.418557883  3.44367428 5.738666e-04
```

```r
summary(logfit4)$coefficients[8, ]
```

```
##    Estimate  Std. Error     z value    Pr(>|z|)
## 0.004930394 0.003944373 1.249981811 0.211306192
```

Comment:

Q. Interpret the coefficient estimate in front of the predictor [5] serum cholestoral.

The estimated coefficient is approximately 0.004930.

1. The odds of getting heartdisease with every one unit (mg/dl) increase in serum cholestoral is around $e^{0.004930} \approx$ 1.004942 times higher, keeping everything is fixed.
2. A one unit (mg/dl) increase in serum cholestoral is associated with an increase in the log odds of getting heartdisease by around $e^{0.004930} \approx$ 1.004942 units.

Q. Test the null hypothesis that this coefficient equals zero, what is the p-value of this test?

The null hypothesis is $H_0 : \beta_{cholestoral} = 0$, so from the t-test: $\frac{\hat{\beta}_{cholestoral} - \beta_{cholestoral}}{s.e.(\beta_{cholestoral})} \sim t_{n-p-1}$, the p-value of the test is around 0.211306192.

Q. If the significance level is set at 0.05, what is your conclusion of this hypothesis test?

Since signficance level of 0.05, I would *not* reject the null, as p-value is larger than 0.05. So, it means that coefficient of serum cholestoral should be zero. So, coefficient of serum cholestoral is not significant in the logistic regression fit.

Part e

```
summary(logfit4)$coefficients
```

```
##                        Estimate  Std. Error     z value     Pr(>|z|)
## (Intercept)        -5.986093033 2.938057804 -2.03743202 4.160677e-02
## age                -0.012295651 0.024664460 -0.49851694 6.181197e-01
## gender              1.431422404 0.513184959  2.78929142 5.282351e-03
## chest2              1.071152889 0.753901685  1.42081244 1.553713e-01
## chest3              0.202174825 0.648717977  0.31165288 7.553043e-01
## chest4              2.006801827 0.652608055  3.07504912 2.104679e-03
## bldpressure         0.023981075 0.011110050  2.15850292 3.088875e-02
## cholestroal         0.004930394 0.003944373  1.24998181 2.113062e-01
## bloodsugar         -0.610758340 0.599183766 -1.01931724 3.080524e-01
## electrocardio       0.255433049 0.189564760  1.34747117 1.778285e-01
## maxheartrate       -0.021280893 0.010820888 -1.96664943 4.922364e-02
## angina              0.739430871 0.434686711  1.70106620 8.893056e-02
## STdepression        0.353095180 0.230102261  1.53451417 1.249032e-01
## STslopepeakexercise 0.670508183 0.371615790  1.80430488 7.118352e-02
## vessel              1.269289907 0.271304027  4.67847795 2.890122e-06
## thal2               0.011429945 0.795089870  0.01437566 9.885303e-01
## thal3               1.441377014 0.418557883  3.44367428 5.738666e-04
```

```
summary(logfit4)$coefficients[6, ]
```

```
##    Estimate  Std. Error     z value    Pr(>|z|)
## 2.006801827 0.652608055 3.075049123 0.002104679
```

```
summary(logfit5)$coefficients
```

```
##                        Estimate  Std. Error     z value     Pr(>|z|)
## (Intercept)        -4.914940144 2.846093169 -1.72690768 8.418425e-02
## age                -0.012295651 0.024664460 -0.49851694 6.181197e-01
## gender              1.431422404 0.513184959  2.78929142 5.282351e-03
## chest1             -1.071152889 0.753901685 -1.42081244 1.553713e-01
## chest3             -0.868978064 0.616289884 -1.41001513 1.585352e-01
## chest4              0.935648939 0.556725256  1.68062959 9.283488e-02
## bldpressure         0.023981075 0.011110050  2.15850292 3.088875e-02
## cholestroal         0.004930394 0.003944373  1.24998181 2.113062e-01
## bloodsugar         -0.610758340 0.599183766 -1.01931724 3.080524e-01
## electrocardio       0.255433049 0.189564760  1.34747117 1.778285e-01
## maxheartrate       -0.021280893 0.010820888 -1.96664943 4.922364e-02
## angina              0.739430871 0.434686711  1.70106620 8.893056e-02
## STdepression        0.353095180 0.230102261  1.53451417 1.249032e-01
## STslopepeakexercise 0.670508183 0.371615790  1.80430488 7.118352e-02
## vessel              1.269289907 0.271304027  4.67847795 2.890122e-06
## thal2                    0.011429945 0.795089870  0.01437566 9.885303e-01
## thal3                    1.441377014 0.418557883  3.44367428 5.738666e-04
```

Comment:

It may be helpful to write down the "sub-model" for every combination of the categorical variables. The coefficient of chest4 in each regression will be involved with some intercept in sub-models involving (x,y) pairs where chest 4, but the interpretation differs slightly. In the first it represents some mean shift form the baseline chest 1, while in the second the baseline is chest 2 instead. (Another way is to take out the intercept and keep all dummy columns)

So, as I keep taking out the first column of the two logistic fits, intercept has the same effect from thal dummy variable (related to thal 1). So, the intercept from the first model is the outcome as chest 1 and thal 1, and the intercept from the second model is the outcome as chest 2 and thal 1. As you can see, the estimate coefficients from the chest 2 from the first model and chest 1 from the second model are the same (up to the sign difference), as the first one means difference between chest 2 and chest 1 (with effect of thal 1), and the second one means the difference between chest 1 and chest 2 (with effect of thal 1). So, the real coefficient for chest 4 (with effect of thal 1) is *around -3.97929121* (as 0.935648939 - 4.914940144 from the second model or 2.006801827 - 5.986093033 from the first model). However, this still cannot be said the real estimator, since there is another dummy variable, thal affecting the intercept.

Since the question specifically asked me to interpret the coefficient estimates for predictor [3] chest pain type 4, I will twick the model a little bit. Instead of taking out the last dummy column in part c), I will take out the first column for two dummy matrices.

Q. Interpret the coefficient estimate in front of the predictor [3] chest pain type 4.

The estimated coefficient is approximately 2.006802. (type 4 stands for asymptomatic)

1. The odds of getting heartdisease with every one unit increase in **difference between chest pain type 4 and chest pain type 1** is approximately $e^{2.006802} \approx$ 7.439488 times higher, keeping everything is fixed.
2. A one unit increase in **difference between chest pain type 4 and chest pain type 1** is associated with an increase in the log odds of getting heartdisease by around $e^{2.006802} \approx$ 7.439488 units.

Q. Test the null hypothesis that this coefficient equals zero, what is the p-value of this test?

The null hypothesis is $H_0 : \beta_{chest4} = 0$, so from the t-test: $\frac{\hat{\beta_{chest4}} - \beta_{chest4}}{s.e.(\beta_{chest4})} \sim t_{n-p-1}$, the p-value of the test is approximately 0.002104679.

Q. If the significance level is set at 0.05, what is your conclusion of this hypothesis test?

Since signficance level of 0.05, I would *reject* the null, as p-value is smaller than 0.05. So, it means that coefficient of chest pain type 4 should not be zero. So, coefficient of chest pain type 4 is significant in the logistic regression fit.

Part f

Reference: https://stats.stackexchange.com/questions/146294/what-is-misclassification-rate-how-do-we-calculate-it

```r
summary(logfit4)
```

```
##
## Call:
## glm(formula = heartdisease ~ ., family = binomial, data = cbind(newdesignonecolout2,
##     response))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7145  -0.5436  -0.1444   0.3264   2.7316
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -5.986093   2.938058  -2.037 0.041607 *
## age                 -0.012296   0.024664  -0.499 0.618120
## gender               1.431422   0.513185   2.789 0.005282 **
## chest2               1.071153   0.753902   1.421 0.155371
## chest3               0.202175   0.648718   0.312 0.755304
## chest4               2.006802   0.652608   3.075 0.002105 **
## bldpressure          0.023981   0.011110   2.159 0.030889 *
## cholestroal          0.004930   0.003944   1.250 0.211306
## bloodsugar          -0.610758   0.599184  -1.019 0.308052
## electrocardio        0.255433   0.189565   1.347 0.177829
## maxheartrate        -0.021281   0.010821  -1.967 0.049224 *
## angina               0.739431   0.434687   1.701 0.088931 .
## STdepression         0.353095   0.230102   1.535 0.124903
## STslopepeakexercise  0.670508   0.371616   1.804 0.071184 .
## vessel               1.269290   0.271304   4.678 2.89e-06 ***
## thal2                0.011430   0.795090   0.014 0.988530
## thal3                1.441377   0.418558   3.444 0.000574 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 409.95  on 296  degrees of freedom
## Residual deviance: 194.83  on 280  degrees of freedom
## AIC: 228.83
##
## Number of Fisher Scoring iterations: 6
```

```
#Classfication
count(predict(logfit4, type = "response") > 0.5)
```

```
##       x freq
## 1 FALSE  173
## 2  TRUE  124
```

```
one <- rep(1, nrow(response))
yhat <- rep(0,nrow(response))
yhat[fitted(logfit4) > 0.5] <- 1 #fitted outputs yhat as cutoff is 0.5
mean(one != yhat) #not match only if yhat is classfied into 0 -> 0 is predicted to be 58.24%.
```

```
## [1] 0.5824916
```

```
#Misclassification rate
sum(response == yhat)
```

```
## [1] 256
```

```
sum(response != yhat)
```

```
## [1] 41
```

```
mean(response != yhat)
```

```
## [1] 0.1380471
```

Comment:

Q. Classify the 297 patients into the presence class or absence class, using the cutoff probability 0.5.

I found that if the cut off is 0.5, then, 124 patients are classfied into the presence class, and 173 are classfied into the absence class. (0.5824 for absence class probability when cutoff probability is 0.5)

Q. What is the misclassification rate for this logistic regression fit?

Misclasfficiation rate can be calculated as $\frac{1}{n}\sum_i I(y_i - \hat{y}_i)$. So, misclassfication rate for this logistic regression fit is approximately 0.13805.