PH245_HW3

```
library(ggplot2)
library(dummies)
```

```
## dummies-1.5.6 provided by Decision Patterns
```

```
setwd('/Users/xiaoyingliu/desktop')
getwd()
```

```
## [1] "/Users/xiaoyingliu/Desktop"
```

```
# Loading Data
data = read.table(file="hw3.txt", header=FALSE, quote="", sep=",")

length = nrow(data)
id= sort( c(seq(1, length)[data[,12]=='?'],
                seq(1, length)[data[,13]=='?']
               )
            )

data[,12] = as.numeric(data[,12]) - 2
data[,13] = as.numeric(data[,13]) - 1

predictors = data.matrix(data[,1:13])
response = data[,14]
response[response > 0] = 1

colnames(predictors) = c("age", "gender", "chestpain", "bldpressure", "chol",
      "bldsugar", "electrocardio", "heartrate", "angina", "STdepression",
      "STslope", "vessel", "thal")

# Removing patients without valid data
predictors = predictors[-id,]
response = response[-id]

stopifnot(nrow(predictors) == length(response) && nrow(predictors) == 297)

print("Predictors:")
```

```
## [1] "Predictors:"
```

```
head(predictors)
```

```
##      age gender chestpain bldpressure chol bldsugar electrocardio
## [1,]  63      1         1         145  233        1             2
## [2,]  67      1         4         160  286        0             2
## [3,]  67      1         4         120  229        0             2
## [4,]  37      1         3         130  250        0             0
## [5,]  41      0         2         130  204        0             2
## [6,]  56      1         2         120  236        0             0
##      heartrate angina STdepression STslope vessel thal
## [1,]       150      0          2.3       3     -1    1
## [2,]       108      1          1.5       2      2    0
## [3,]       129      1          2.6       2      1    2
## [4,]       187      0          3.5       3     -1    0
## [5,]       172      0          1.4       1     -1    0
## [6,]       178      0          0.8       1     -1    0
```

```
print("Response:")
```

```
## [1] "Response:"
```

```
head(response)
```

```
## [1] 0 1 1 0 0 0
```

#1(a)

```
# EDA

# 1.A.1: How many patients in the dataset had heart disease vs. no disease?
numHeartDisease = sum(response)
noHeartDisease = length(response) - numHeartDisease

print("Total number of patients:")
```

```
## [1] "Total number of patients:"
```

```
length(response)
```

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

```
stopifnot(length(response) == numHeartDisease + noHeartDisease)

print("Number of patients with heart disease:")
```

```
## [1] "Number of patients with heart disease:"
```

```
numHeartDisease
```

```
## [1] 137
```

```
print("Number of patients with no heart disease:")
```

```
## [1] "Number of patients with no heart disease:"
```

```
noHeartDisease
```

```
## [1] 160
```

```
"----------------------------------------------------------------"
```

```
## [1] "----------------------------------------------------------------"
```

```
# 1.A.2: Which predictors are numerical, which are categorical, and which are unclear?

print("Total number of predictors:")
```

```
## [1] "Total number of predictors:"
```

```
ncol(predictors)
```

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

```
print("Numeric predictor variables:")
```

```
## [1] "Numeric predictor variables:"
```

```
head(predictors[,c(1, 4, 5, 8, 10)])
```

```
##      age bldpressure chol heartrate STdepression
## [1,]  63         145  233       150          2.3
## [2,]  67         160  286       108          1.5
## [3,]  67         120  229       129          2.6
## [4,]  37         130  250       187          3.5
## [5,]  41         130  204       172          1.4
## [6,]  56         120  236       178          0.8
```

```
print("Categorical predictor variables:")
```

```
## [1] "Categorical predictor variables:"
```

```
head(predictors[,c(2, 3, 6, 9, 12, 13)])
```

```
##      gender chestpain bldsugar angina vessel thal
## [1,]      1         1        1      0     -1    1
## [2,]      1         4        0      1      2    0
## [3,]      1         4        0      1      1    2
## [4,]      1         3        0      0     -1    0
## [5,]      0         2        0      0     -1    0
## [6,]      1         2        0      0     -1    0
```

```
print("Unclear variables that could be treated as either numeric or categorical:")
```

```
## [1] "Unclear variables that could be treated as either numeric or categorical:"
```

```
head(predictors[, c(7, 11)])
```

```
##      electrocardio STslope
## [1,]             2       3
## [2,]             2       2
## [3,]             2       2
## [4,]             0       3
## [5,]             2       1
## [6,]             0       1
```

```
"________________________________________________________________"
```

```
## [1] "________________________________________________________________"
```

```
print("Gender breakdown % (0):")
```

```
## [1] "Gender breakdown % (0):"
```

```
(297-sum(predictors[,2]))/297 * 100
```

```
## [1] 32.32323
```

```
print("Gender breakdown % (1):")
```

```
## [1] "Gender breakdown % (1):"
```

```
sum(predictors[,2])/297 * 100
```

```
## [1] 67.67677
```

#1(b)

```
combinedDF = as.data.frame(cbind(predictors, response))
fit = glm(formula=response~., family="binomial", data=combinedDF)
summary(fit)
```

```
##
## Call:
## glm(formula = response ~ ., family = "binomial", data = combinedDF)
##
## 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)   -5.012690   2.893960  -1.732  0.08325 .
## 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 *
## chol           0.004816   0.003775   1.276  0.20202
## bldsugar      -0.991868   0.554947  -1.787  0.07389 .
## electrocardio  0.246117   0.185238   1.329  0.18396
## heartrate     -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
## STslope        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
```

#1(c)

```
combinedOneHotDF = dummy.data.frame(combinedDF, names=c("chestpain", "thal"))

# Dropping dependency variables
combinedOneHotDF$chestpain1 = NULL
combinedOneHotDF$thal1 = NULL



# Fit the logistic regression including new dummy variables
oneHotFit = glm(formula=response~., family="binomial", data=combinedOneHotDF)

summary(oneHotFit)
```

```
##
## Call:
```

```
## glm(formula = response ~ ., family = "binomial", data = combinedOneHotDF)
##
## 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)   -4.705373   3.027145  -1.554  0.12009
## age           -0.012296   0.024664  -0.499  0.61812
## gender         1.431422   0.513185   2.789  0.00528 **
## chestpain2     1.071153   0.753902   1.421  0.15537
## chestpain3     0.202175   0.648718   0.312  0.75530
## chestpain4     2.006802   0.652608   3.075  0.00210 **
## bldpressure    0.023981   0.011110   2.159  0.03089 *
## chol           0.004930   0.003944   1.250  0.21131
## bldsugar      -0.610758   0.599184  -1.019  0.30805
## electrocardio  0.255433   0.189565   1.347  0.17783
## heartrate     -0.021281   0.010821  -1.967  0.04922 *
## angina         0.739431   0.434687   1.701  0.08893 .
## STdepression   0.353095   0.230102   1.535  0.12490
## STslope        0.670508   0.371616   1.804  0.07118 .
## vessel         1.269290   0.271304   4.678 2.89e-06 ***
## thal0         -0.011430   0.795090  -0.014  0.98853
## thal2          1.429947   0.783279   1.826  0.06791 .
## ---
## 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
```

#1(d)

```
#The coefficient estimate for serum cholesterol is 0.00493.For every unit increase in blood serum
#cholesterol, a 0.00493 increase in the log odds of having heart disease compared to not having
#heart disease.The p-value for seeing a coefficient estimate is 0.2113. With an alpha of .05,
#we fail to reject the null hypothesis that blood serum cholesterol == 0.
```

#1(e)

```
#The coefficient estimate for chest pain type 4 was 2.006802. Compared to those having
#chest pain type 1,people having this chest pain see a 2.006802 increase in the log odds
#of having heart disease.
#The p-value of chestpain4 is 0.002105, thus we could reject our null hypothesis (alpha=.05)
#that chestpain4's coefficient == 0. Our results indicate that the coefficient estimate is
#indeed statistically significant to the model and predicting heart disease outcomes.
```

#1(f)

```
probabilityPredictions = as.numeric(predict(oneHotFit, combinedOneHotDF, type='response'))
```

```r
print("Reminder: 0=Heart Disease Absent; 1=Heart Disease Present")
```

```
## [1] "Reminder: 0=Heart Disease Absent; 1=Heart Disease Present"
```

```r
""
```

```
## [1] ""
```

```r
print("Head of Probability Predictions (%Chance that response was not 0)")
```

```
## [1] "Head of Probability Predictions (%Chance that response was not 0)"
```

```r
head(probabilityPredictions)
```

```
## [1] 0.23862882 0.99850760 0.99540963 0.23335447 0.03625226 0.04899566
```

```r
binaryResponsePredictions = as.numeric(probabilityPredictions >= .5)

print("Head of Response Predictions based on model")
```

```
## [1] "Head of Response Predictions based on model"
```

```r
head(binaryResponsePredictions)
```

```
## [1] 0 1 1 0 0 0
```

```r
print("Head of True responses for training set")
```

```
## [1] "Head of True responses for training set"
```

```r
head(combinedOneHotDF$response)
```

```
## [1] 0 1 1 0 0 0
```

```r
stopifnot(length(binaryResponsePredictions) == length(combinedOneHotDF$response))

accuracy = sum(binaryResponsePredictions == combinedOneHotDF$response)/length(binaryResponsePredictions
print("Model accuracy:")
```

```
## [1] "Model accuracy:"
```

```r
accuracy
```

```
## [1] 0.8619529
```

```r
print("Misclassification rate:")
```

```
## [1] "Misclassification rate:"
```

```r
1-accuracy
```

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