hw7

Mohammad Mottahedi

March 10, 2016

```
setwd("~/Dropbox/spring2016/discretDataAnalysis/hw_solutions/hw7")
set.seed(1234)
```

problem 1

part a

```
kyphosis <- array(data= c(rep(1,17), rep(0,22)))
age <- array(data=c(12, 15, 42, 52, 59, 73, 82, 91, 96, 105, 114,
                    120, 121, 128, 130, 139, 157,
                    1, 1, 2, 8, 11, 18, 22, 31, 37, 61, 72,
                    81, 97, 112, 118, 127, 131, 140, 151, 159, 177, 206))

df <- data.frame(age, kyphosis)

result <- glm(kyphosis ~ age, data=df, family = binomial(link= "logit"))
summary(result)
```

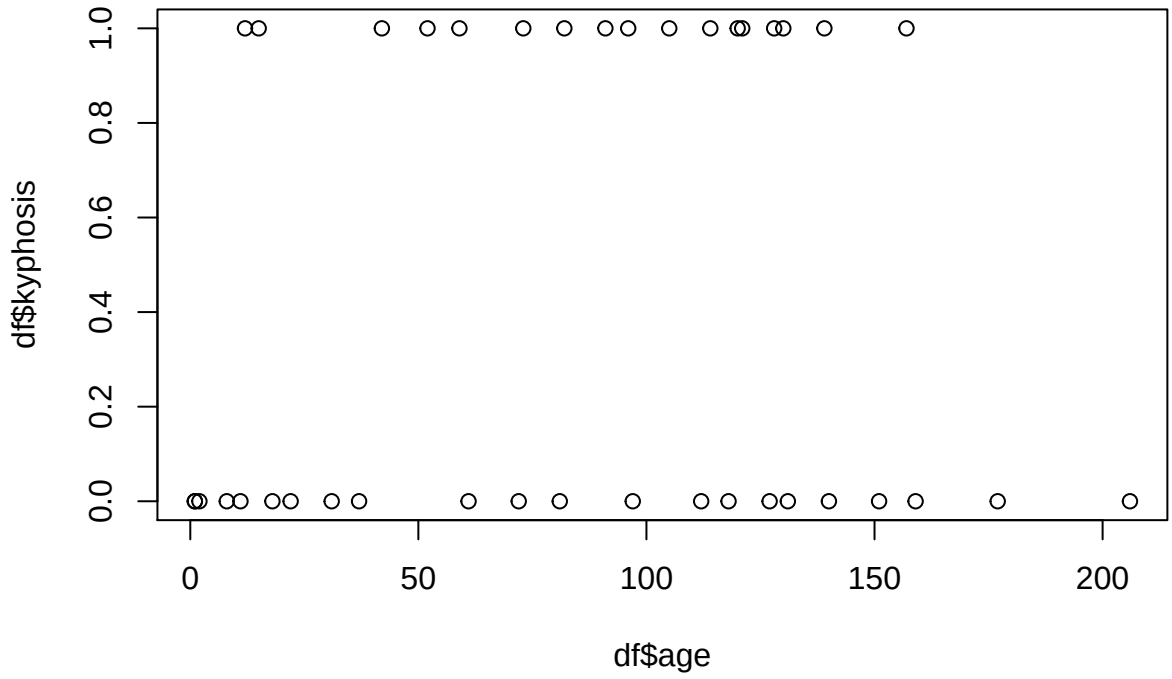
```
##
## Call:
## glm(formula = kyphosis ~ age, family = binomial(link = "logit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2418  -1.0753  -0.9586   1.2530   1.3981
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.545878   0.601992  -0.907    0.365
## age          0.003379   0.005906   0.572    0.567
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53.423  on 38  degrees of freedom
## Residual deviance: 53.093  on 37  degrees of freedom
## AIC: 57.093
##
## Number of Fisher Scoring iterations: 4
```

$logit(\pi) = -0.546 + 0.0059(age)$

Assuming the Type I error of 0.05, we can't reject the null hypothesis that $\beta_1 = 0$, so there is a strong evidence the relationship between the age and kyphosis in not significant.

part b

```
#plot(result$linear.predictors,residuals(result, type="pearson"))
plot(df$age, df$kyphosis)
```



age has higher dispersion in absence of kyphosis.

The age data is ungrouped and we can assume the scale factor is equal to 1.

```
overdispersion <- sum(residuals(result, type = 'pearson') ^ 2) / (dim(df)[1] - 1 )
overdispersion
```

```
## [1] 1.024818
```

```
summary(result, dispersion = overdispersion, correlation = TRUE, symbolic.cor = TRUE)
```

```
##
## Call:
## glm(formula = kyphosis ~ age, family = binomial(link = "logit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2418  -1.0753  -0.9586   1.2530   1.3981
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.545878   0.609417  -0.896    0.370
## age          0.003379   0.005978   0.565    0.572
##
## (Dispersion parameter for binomial family taken to be 1.024818)
```

```
##
##     Null deviance: 53.423  on 38  degrees of freedom
## Residual deviance: 53.093  on 37  degrees of freedom
## AIC: 57.093
##
## Number of Fisher Scoring iterations: 4
##
## Correlation of Coefficients:
##
## (Intercept) 1
## age         + 1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

```
#library(dispmod)
#glm.binomial.disp(result, maxit = 1000)
```

part c

```
result <- glm(kyphosis ~ age + I(age^2), data=df, family = binomial(link="logit"))
summary(result)
```

```
##
## Call:
## glm(formula = kyphosis ~ age + I(age^2), family = binomial(link = "logit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4585  -1.0233  -0.5091   1.0348   1.7826
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.0379796  0.9979085  -2.042   0.0411 *
## age          0.0605359  0.0273232   2.216   0.0267 *
## I(age^2)    -0.0003394  0.0001620  -2.096   0.0361 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53.423  on 38  degrees of freedom
## Residual deviance: 46.731  on 36  degrees of freedom
## AIC: 52.731
##
## Number of Fisher Scoring iterations: 4
```

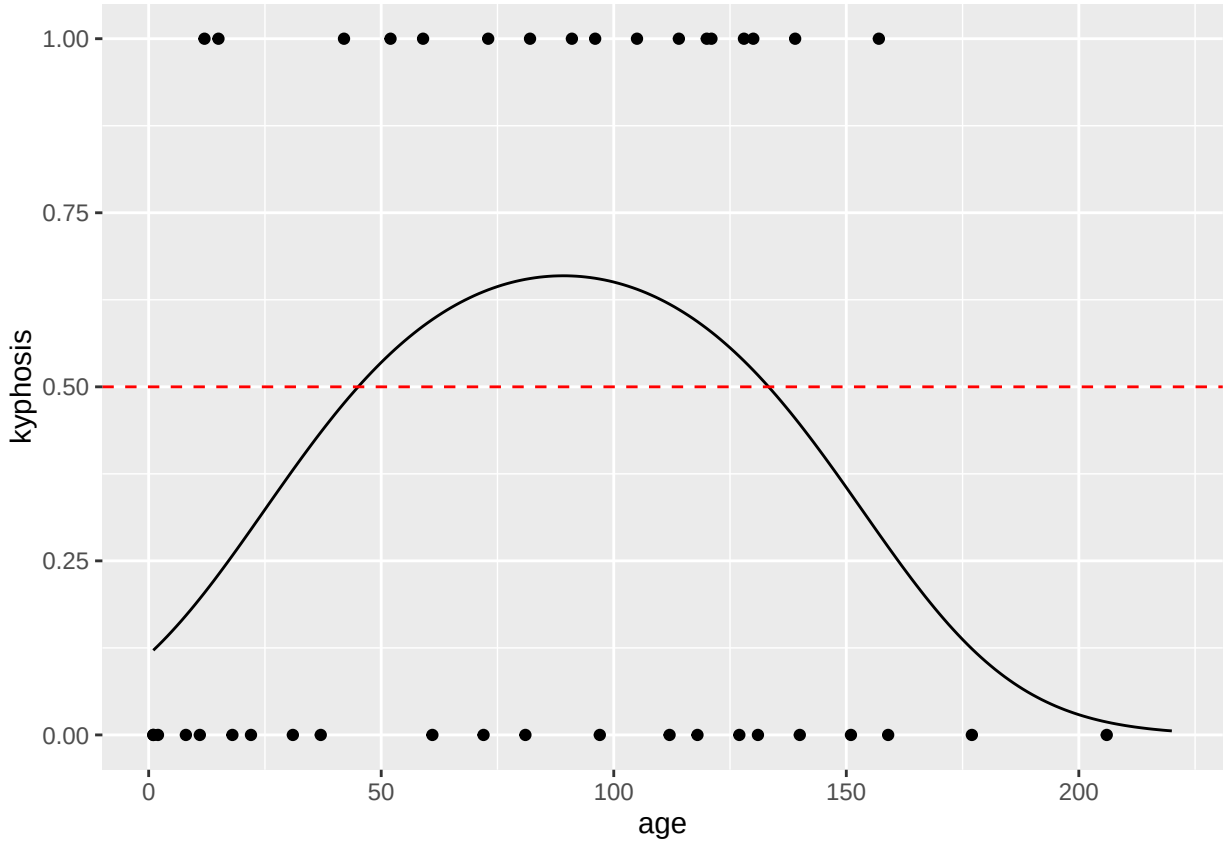
The deviance is smaller compared to prevous model and the null model. All estimated coefficient have have p-value less than 0.5 and are significant. The model deviance is still large which and the model is not a good fit (p-value= 0.108).

```
x <- data.frame(age=1:220, age_sq=(1:220)^2)
x["pred"] <- predict(result, x , type='response')
x['y_hat'] <- (exp(result$coefficients[1] + result$coefficients[2]*x$age + result$coefficients[3]*x$age_
    (1 + exp(result$coefficients[1] + result$coefficients[2]*x$age + result$coefficients[3]*x$age_sq))

library(ggplot2)

p <- ggplot(data=df, aes(x=age, y=kyphosis))+ geom_point() + geom_line(data=x,aes(x=x$age, y=x$y_hat))+

print(p)
```



problem 2

part a

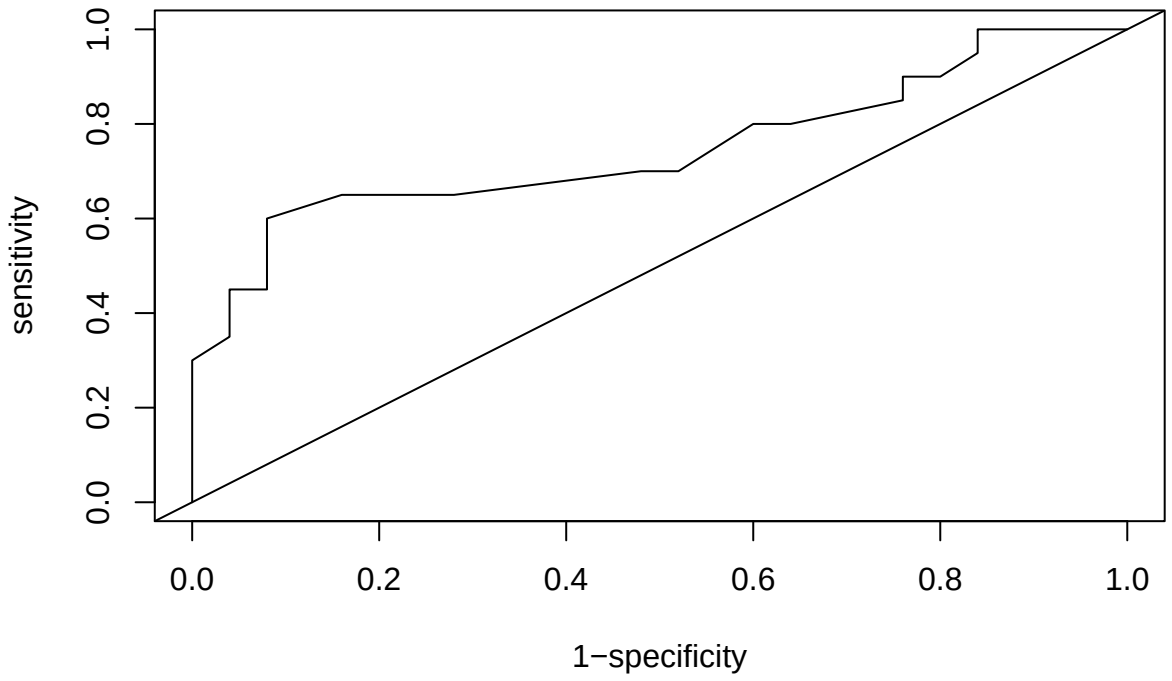
```
donner <- read.csv('donner.txt', header=F, sep=' ')
colnames(donner) <- c('age','sex','survive')

result <- glm(survive ~ age + sex, data=donner, family=binomial())
summary(result)
```

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

```
## glm(formula = survive ~ age + sex, family = binomial(), data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7445  -1.0441  -0.3029   0.8877   2.0472
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.23041    1.38686   2.329   0.0198 *
## age         -0.07820    0.03728  -2.097   0.0359 *
## sex         -1.59729    0.75547  -2.114   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

```
roc.plot(result$fitted.values[donner$survive == 1], result$fitted.values[donner$survive == 0] )
```




```
## $area
## [1] 0.746
##
## $cutoff
##        32
## 0.3363082
##
```

```
## $sensopt
## [1] 0.6
##
## $specopt
## [1] 0.92
```

AUC = 0.75 can be interpreted to mean that a randomly selected person in surviver group has a 75% higher survival probability than a person not in the survival group.

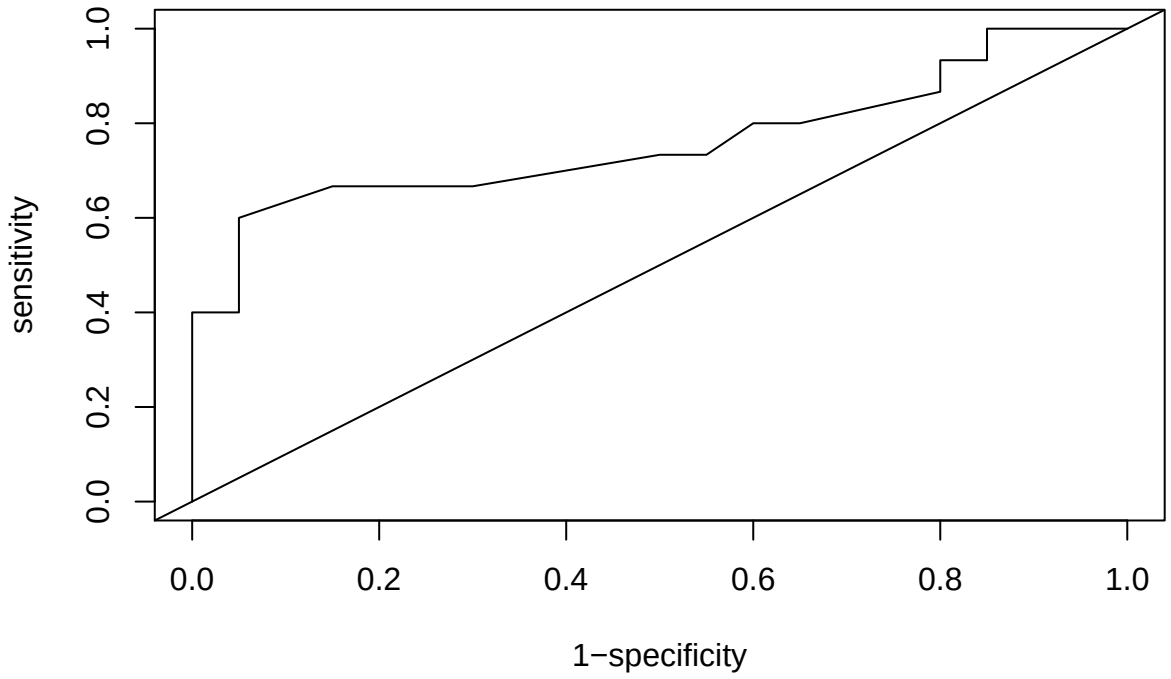
part b

```
train.indx <- sample(45, 35)
test <- donner[-train.indx,]
train <- donner[train.indx,]

result <- glm(survive ~ age + sex, data=train, family=binomial())
summary(result)
```

```
##
## Call:
## glm(formula = survive ~ age + sex, family = binomial(), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4489  -1.0247  -0.2396   0.9279   2.2499
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.10662    1.99905   2.054   0.0399 *
## age         -0.09893    0.04962  -1.994   0.0462 *
## sex         -2.00387    1.02249  -1.960   0.0500 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47.804  on 34  degrees of freedom
## Residual deviance: 38.662  on 32  degrees of freedom
## AIC: 44.662
##
## Number of Fisher Scoring iterations: 5
```

```
roc.plot(result$fitted.values[train$survive == 1], result$fitted.values[train$survive == 0] )
```

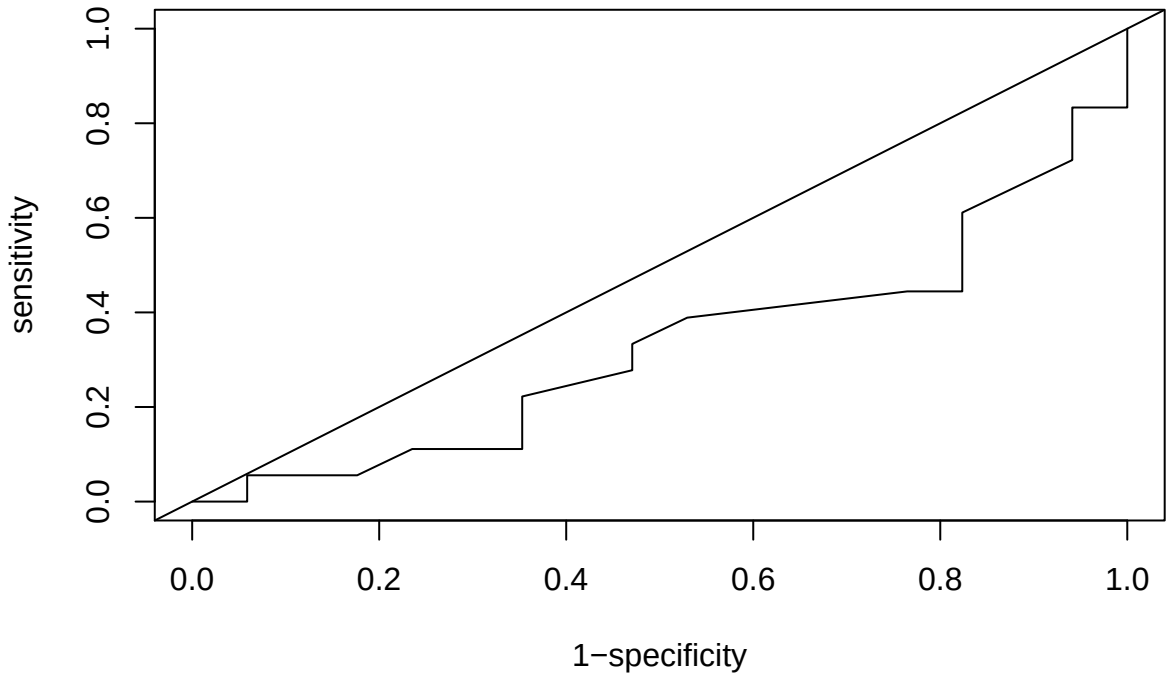



```
## $area
## [1] 0.76
##
## $cutoff
##         4
## 0.2962681
##
## $sensopt
## [1] 0.6
##
## $specopt
## [1] 0.95
```

The AUC value for the training data is equal to 0.76 which is close to the AUC predicted in part a.

part c

```
pred <- predict(result, test[,c('age','sex')])
roc.plot(result$fitted.values[test$survive == 1], result$fitted.values[test$survive == 0] )
```




```
## $area
## [1] 0.3267974
##
## $cutoff
##        18
## 0.9323069
##
## $sensopt
## [1] 1
##
## $specopt
## [1] 0
```

“‘ AUC for the test data is equal to 0.32 which is less than 0.5, which means the model is worst than randomly guessing the response values.