Problem 1: Investigation of Life Expectancy

(1) Predicting variables that affect the Life Expectancy are those with small p-values, which means that the corresponding coefficient has a high probability not equal to 0. The following variables affect the Life Expectancy:

Status	Measles	HIV/AIDS
Adult Mortality	BMI	GDP
Infant deaths	under five deaths	Income composition of resources
Alcohol	Polio	Schooling
Hepatitis B	Diphtheria	

R code for summary report of the linear model:

```
library(mice)
q1data = read.csv("Life Expectancy Data.csv", header=TRUE,sep=",")
# remove rows with missing Life expectancy
q1data = q1data[complete.cases(q1data[, "Life.expectancy"]),]
# remove Country
q1data = q1data[, ! names(q1data) %in% c('Country'), drop = F]
predictors = q1data[, ! names(q1data) %in% c('Life.expectancy'), drop = F]
# perform mice imputation, based on mean
tempData = mice(q1data,maxit=1,meth='mean')
X = complete(tempData,1)
q1model = lm(Life.expectancy~.,X)
summary(q1model)
```

Summary Report:

```
> summary(q1model)

Call:
lm(formula = Life.expectancy ~ ., data = X)

Residuals:
     Min       1Q   Median       3Q      Max
-22.2690  -2.2422  -0.0893   2.3846  16.5668

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     7.445e+01  3.476e+01   2.142 0.032299 *
Year                           -9.012e-03  1.739e-02  -0.518 0.604257
StatusDeveloping               -1.583e+00  2.702e-01  -5.860 5.14e-09 ***
Adult.Mortality                -1.979e-02  7.962e-04 -24.856  < 2e-16 ***
infant.deaths                   9.942e-02  8.428e-03  11.797  < 2e-16 ***
Alcohol                         6.088e-02  2.611e-02   2.331 0.019806 *
percentage.expenditure          8.755e-05  8.471e-05   1.033 0.301460
Hepatitis.B                    -1.401e-02  3.927e-03  -3.569 0.000365 ***
Measles                        -1.978e-05  7.662e-06  -2.581 0.009892 **
BMI                             4.412e-02  4.969e-03   8.878  < 2e-16 ***
under.five.deaths              -7.440e-02  6.177e-03 -12.045  < 2e-16 ***
Polio                           2.877e-02  4.469e-03   6.437 1.42e-10 ***
Total.expenditure               5.602e-02  3.458e-02   1.620 0.105372
Diphtheria                      4.037e-02  4.710e-03   8.571  < 2e-16 ***
HIV.AIDS                       -4.707e-01  1.765e-02 -26.663  < 2e-16 ***
GDP                             3.291e-05  1.302e-05   2.527 0.011545 *
Population                      2.712e-10  1.692e-09   0.160 0.872700
thinness..1.19.years           -8.170e-02  5.039e-02  -1.622 0.105001
thinness.5.9.years              6.291e-03  4.964e-02   0.127 0.899168
Income.composition.of.resources 5.609e+00  6.455e-01   8.689  < 2e-16 ***
Schooling                       6.741e-01  4.263e-02  15.814  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.05 on 2907 degrees of freedom
Multiple R-squared:  0.8204,    Adjusted R-squared:  0.8192
F-statistic:   664 on 20 and 2907 DF,  p-value: < 2.2e-16
```

(2) The 95% confidence interval of Adult Mortality is [-0.02135053, -0.01822836] and that of HIV/AIDS is [-0.50532955, -0.43609669]. As the intervals have probability of 95% to contain the true βs and the upper bound of the both intervals are negative, together with the small p-values obtained, I am confident that both Adult Mortality and HIV/AIDS have negative impact to Life Expectancy.

```
> confint(q1model,c('Adult.Mortality','HIV.AIDS'),level=0.95)
                      2.5 %      97.5 %
Adult.Mortality -0.02135053 -0.01822836
HIV.AIDS        -0.50532955 -0.43609669
```

(3) The 97% confidence interval of Schooling is [0.581555139, 0.7666618] and that of Alcohol is [0.004181476, 0.1175687]. Both have positive impact to Life Expectancy, but the impact by Schooling is much higher than Alcohol.

```
> confint(q1model,c('Schooling','Alcohol'),level=0.97)
                  1.5 %    98.5 %
Schooling 0.581555139 0.7666618
Alcohol   0.004181476 0.1175687
```

(4) There are 8 predictors with p-value less than 2e-16, including:

Adult Mortality	Diphtheria
Infant deaths	HIV/AIDS
BMI	Income composition of resources
under five deaths	Schooling

R code for summary report of the smaller model:

```
selected =
c('Adult.Mortality','infant.deaths','BMI','under.five.deaths','Diphtheria','HIV.AIDS','Income.composition.of.resources','Schooling','Life.expectancy')
q1smallmodel = lm(Life.expectancy~.,X[selected])
summary(q1smallmodel)
```

Summary Report:

```
> summary(q1smallmodel)

Call:
lm(formula = Life.expectancy ~ ., data = X[selected])

Residuals:
     Min       1Q   Median       3Q      Max
-21.9706  -2.2326  -0.0837   2.3125  19.4747

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     52.8180480  0.4674763  112.98   <2e-16 ***
Adult.Mortality                 -0.0210086  0.0008158  -25.75   <2e-16 ***
infant.deaths                    0.0892630  0.0084316   10.59   <2e-16 ***
BMI                              0.0559976  0.0047951   11.68   <2e-16 ***
under.five.deaths               -0.0683718  0.0062268  -10.98   <2e-16 ***
Diphtheria                       0.0517006  0.0037200   13.90   <2e-16 ***
HIV.AIDS                        -0.4626238  0.0180975  -25.56   <2e-16 ***
Income.composition.of.resources  6.9469276  0.6510239   10.67   <2e-16 ***
Schooling                        0.8409741  0.0420677   19.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.218 on 2919 degrees of freedom
Multiple R-squared:  0.8044,    Adjusted R-squared:  0.8039
F-statistic:  1501 on 8 and 2919 DF,  p-value: < 2.2e-16
```

(5) Using the new observation with the smaller model, the prediction of Life Expectancy is 83.30468 and the 99% confidence interval for the prediction is [72.19249, 94.41688].

```
> newdata = data.frame(Year=2008,Status='Developed',Adult.Mortality=125,infant.deaths=94,Alcohol=4.1,percentag
e.expenditure=100,Hepatitis.B=20,Measles=13,BMI=55,under.five.deaths=2,Polio=12,Total.expenditure=5.9,Diphther
ia=12,HIV.AIDS=0.5,GDP=5892,Population=1.34e6,Income.composition.of.resources=0.9,Schooling=18)
> predict(q1smallmodel, newdata,interval = "prediction",level=0.99)
       fit      lwr      upr
1 83.30468 72.19249 94.41688
```

Problem 2: Predicting Breast Cancer

(1) R code and summary report of the logistic regression model using all predictors:

```
q2train = read.csv("BreastCancer_train.csv", header=TRUE,sep=",")
q2test = read.csv("BreastCancer_test.csv", header=TRUE,sep=",")
library(mice)
# perform mice imputation, based on predictive mean matching
temptrainData = mice(q2train,maxit=10,meth='pmm',)
Xtrain = complete(temptrainData,1)
Xtrain$Class = as.factor(Xtrain$Class)
temptestData = mice(q2test,maxit=10,meth='pmm',)
Xtest = complete(temptestData,1)

modelq21 = glm(Class~.,data=Xtrain,family=binomial)
summary(modelq21)
```

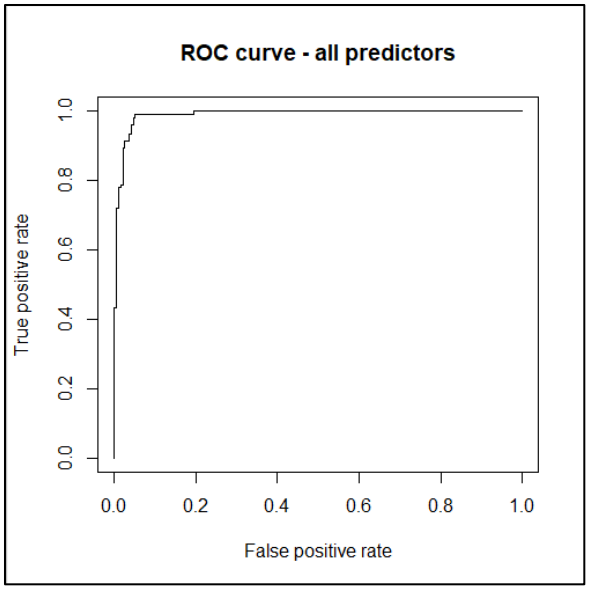
```
> summary(modelq21)

Call:
glm(formula = Class ~ ., family = binomial, data = Xtrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2732  -0.0821  -0.0508   0.0187   1.9232

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -9.949e+00  1.989e+00  -5.002 5.68e-07 ***
Id              -8.086e-07  1.246e-06  -0.649   0.5164
Cl.thickness     3.735e-01  2.161e-01   1.729   0.0839 .
Cell.size       -2.764e-01  2.782e-01  -0.994   0.3204
Cell.shape       5.609e-01  3.124e-01   1.796   0.0725 .
Marg.adhesion    3.656e-01  1.459e-01   2.505   0.0122 *
Epith.c.size     4.175e-01  2.676e-01   1.560   0.1187
Bare.nuclei      2.723e-01  1.219e-01   2.233   0.0256 *
Bl.cromatin      6.338e-01  2.671e-01   2.373   0.0176 *
Normal.nucleoli  2.042e-01  1.603e-01   1.274   0.2026
Mitoses          3.408e-01  4.152e-01   0.821   0.4118
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 514.144  on 399  degrees of freedom
Residual deviance:  53.605  on 389  degrees of freedom
AIC: 75.605

Number of Fisher Scoring iterations: 8
```

R code for predictions on test data, prediction table and ROC curve:

```
modelq21.probs = predict(modelq21,Xtest,type='response')
modelq21.pred = rep('benign',nrow(Xtest))
modelq21.pred[modelq21.probs>.5]='malignant'
table(modelq21.pred,Xtest$Class)
library(ROCR)
pred = prediction(modelq21.probs,Xtest$Class)
perf = performance(pred,"tpr","fpr")
plot(perf,main='ROC curve - all predictors')
```

Prediction table:

```
modelq21.pred benign malignant
    benign       189         9
    malignant      6        95
```



(2) R code and summary report of the logistic regression model using predictors: Cl.thickness, Cell.shape, Marg.adhesion, Bare.nuclei, Bl.cromatin:

```
modelq22 =
glm(Class~Cl.thickness+Cell.shape+Marg.adhesion+Bare.nuclei+Bl.cromatin,data=Xtrain,family=
binomial)
summary(modelq22)
```

```
> summary(modelq22)

Call:
glm(formula = Class ~ Cl.thickness + Cell.shape + Marg.adhesion +
    Bare.nuclei + Bl.cromatin, family = binomial, data = Xtrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2436  -0.1036  -0.0564   0.0229   1.9117

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.8199     1.4982  -6.555 5.58e-11 ***
Cl.thickness   0.4892     0.1942   2.519 0.011765 *
Cell.shape     0.4999     0.1995   2.506 0.012209 *
Marg.adhesion  0.3530     0.1238   2.850 0.004367 **
Bare.nuclei    0.2899     0.1073   2.703 0.006880 **
Bl.cromatin    0.7668     0.2296   3.340 0.000838 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 514.14  on 399  degrees of freedom
Residual deviance:  61.55  on 394  degrees of freedom
AIC: 73.55

Number of Fisher Scoring iterations: 8
```

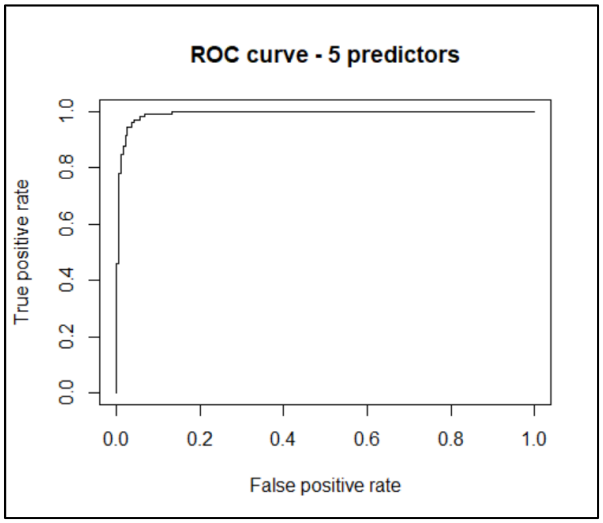
R code for predictions on test data, prediction table and ROC curve:

```
modelq22.probs = predict(modelq22,Xtest,type='response')
modelq22.pred = rep('benign',nrow(Xtest))
modelq22.pred[modelq22.probs>.5]='malignant'
table(modelq22.pred,Xtest$Class)

pred2 = prediction(modelq22.probs,Xtest$Class)
```

```
perf2 = performance(pred2,"tpr","fpr")
plot(perf2,main='ROC curve - 5 predictors')
```

Prediction table:

```
modelq22.pred benign malignant
    benign       190         8
    malignant      5        96
```



(3) R code and summary report of the LDA model using all predictors:

```
library(MASS)
modelq23 = lda(Class~.,data=Xtrain)
modelq23
```

```
> modelq23
Call:
lda(Class ~ ., data = Xtrain)

Prior probabilities of groups:
   benign malignant
   0.6575    0.3425

Group means:
               Id Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
benign    1130343     2.859316  1.273764   1.384030      1.296578     2.091255
malignant 1013844     7.124088  6.678832   6.627737      5.751825     5.321168
          Bare.nuclei Bl.cromatin Normal.nucleoli  Mitoses
benign       1.342205    2.030418        1.231939 1.064639
malignant    7.708029    6.262774        6.204380 2.773723

Coefficients of linear discriminants:
                         LD1
Id              -4.284140e-08
Cl.thickness     1.557594e-01
Cell.size        8.463676e-02
Cell.shape       1.195888e-01
Marg.adhesion    8.279502e-02
Epith.c.size     1.215521e-01
Bare.nuclei      2.430924e-01
Bl.cromatin      1.204167e-01
Normal.nucleoli  1.101310e-01
Mitoses         -3.098874e-02
```

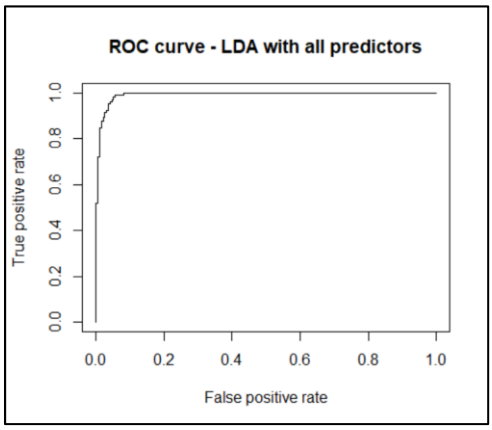
R code for predictions on test data, prediction table and ROC curve:

```
modelq23.pred = predict(modelq23,Xtest)
table(modelq23.pred$class,Xtest$Class)
```

```
pred3 = prediction(modelq23.pred$posterior[,2],Xtest$Class)
perf3 = performance(pred3,"tpr","fpr")
plot(perf3,main='ROC curve - LDA with all predictors')
```

Prediction table:

```
           benign malignant
benign        191        11
malignant       4        93
```



(4) R code and summary report of the LDA model using predictors: Cl.thickness, Cell.shape, Marg.adhesion, Bare.nuclei, Bl.cromatin:

```
modelq24 =
lda(Class~Cl.thickness+Cell.shape+Marg.adhesion+Bare.nuclei+Bl.cromatin,data=Xtrain)
modelq24
```

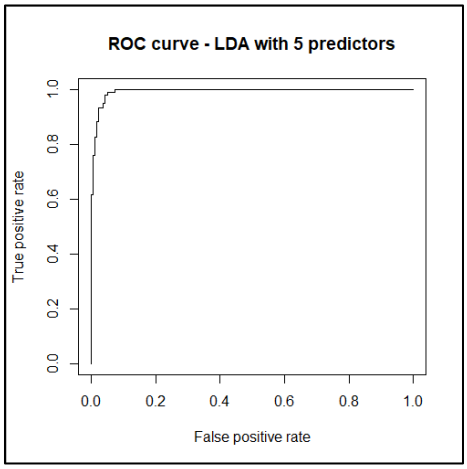
```
> modelq24
Call:
lda(Class ~ Cl.thickness + Cell.shape + Marg.adhesion + Bare.nuclei +
    Bl.cromatin, data = Xtrain)

Prior probabilities of groups:
   benign malignant
   0.6575    0.3425

Group means:
          Cl.thickness Cell.shape Marg.adhesion Bare.nuclei Bl.cromatin
benign        2.859316   1.384030      1.296578    1.342205    2.030418
malignant     7.124088   6.627737      5.751825    7.708029    6.262774

Coefficients of linear discriminants:
                    LD1
Cl.thickness  0.1637813
Cell.shape    0.2457094
Marg.adhesion 0.1075643
Bare.nuclei   0.2354190
Bl.cromatin   0.1932278
```

R code for predictions on test data, prediction table and ROC curve:

```
modelq24.pred = predict(modelq24,Xtest)
table(modelq24.pred$class,Xtest$Class)

pred4 = prediction(modelq24.pred$posterior[,2],Xtest$Class)
perf4 = performance(pred4,"tpr","fpr")
plot(perf4,main='ROC curve - LDA with 5 predictors')
```

Prediction table:

```
           benign malignant
benign        192        12
malignant       3        92
```



(5) R code and summary report of the QDA model using all predictors:

```
modelq25 = qda(Class~.,data=Xtrain)
modelq25
```

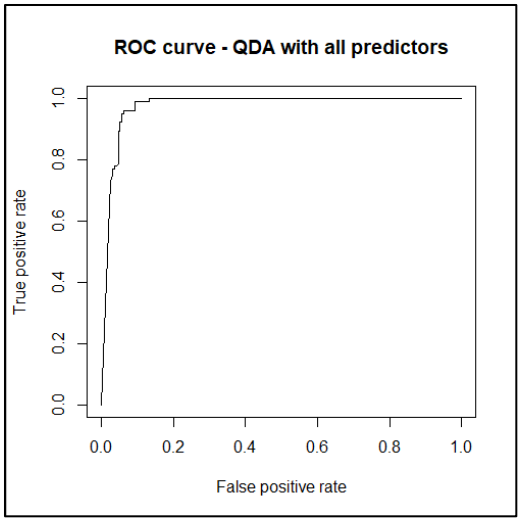
```
> modelq25
Call:
qda(Class ~ ., data = Xtrain)

Prior probabilities of groups:
   benign malignant
   0.6575    0.3425

Group means:
                Id Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
benign    1130343     2.859316  1.273764   1.384030      1.296578     2.091255
malignant 1013844     7.124088  6.678832   6.627737      5.751825     5.321168
          Bare.nuclei Bl.cromatin Normal.nucleoli  Mitoses
benign       1.342205    2.030418        1.231939 1.064639
malignant    7.708029    6.262774        6.204380 2.773723
```

R code for predictions on test data, prediction table and ROC curve:

```
modelq25.pred = predict(modelq25,Xtest)
table(modelq25.pred$class,Xtest$Class)

pred5 = prediction(modelq25.pred$posterior[,2],Xtest$Class)
perf5 = performance(pred5,"tpr","fpr")
plot(perf5,main='ROC curve - QDA with all predictors')
```

Prediction table:

```
           benign malignant
benign        178         4
malignant      17       100
```



(6) R Code for generating AUC for comparison:

```
a1 = as.numeric(performance(pred,"auc")@y.values)
a2 = as.numeric(performance(pred2,"auc")@y.values)
a3 = as.numeric(performance(pred3,"auc")@y.values)
a4 = as.numeric(performance(pred4,"auc")@y.values)
a5 = as.numeric(performance(pred5,"auc")@y.values)
AUC.df = data.frame(model=c('Logistic Regression with all predictors','Logistic Regression with
5 predictors','LDA with all predictors','LDA with 5 predictors','QDA with all
predictors'),AUC=c(a1,a2,a3,a4,a5))
AUC.df
```

Output:

```
                                    model       AUC
1 Logistic Regression with all predictors 0.9899901
2   Logistic Regression with 5 predictors 0.9921598
3                 LDA with all predictors 0.9923570
4                   LDA with 5 predictors 0.9936391
5                 QDA with all predictors 0.9762081
```

Among the 5 models, the AUC of using LDA with 5 specified predictors is the largest and that of using QDA with all predictors is the smallest.