Name: Leung Ko Tsun SID:20516287

```
> #MATH3424HW5
> #Q1a
> q1data <- read.table("./Downloads/BreastCancer.txt", header = TRUE, sep = ",")
> q1data$ClassIndex <- rep(0, nrow(q1data))
> q1data$ClassIndex[which(q1data$Class == "benign")] <- 1
> q1model1 <- glm(ClassIndex ~ . - Class, data = q1data, family = "binomial")
> summary(q1model1)

Call:
glm(formula = ClassIndex ~ . - Class, family = "binomial", data = q1data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4698  -0.0222   0.0619   0.1153   3.4841

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     10.10394    1.17488   8.600  < 2e-16 ***
Cl.thickness    -0.53501    0.14202  -3.767 0.000165 ***
Cell.size        0.00628    0.20908   0.030 0.976039
Cell.shape      -0.32271    0.23060  -1.399 0.161688
Marg.adhesion   -0.33064    0.12345  -2.678 0.007400 **
Epith.c.size    -0.09663    0.15659  -0.617 0.537159
Bare.nuclei     -0.38303    0.09384  -4.082 4.47e-05 ***
Bl.cromatin     -0.44719    0.17138  -2.609 0.009073 **
Normal.nucleoli -0.21303    0.11287  -1.887 0.059115 .
Mitoses         -0.53484    0.32877  -1.627 0.103788
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 102.89  on 673  degrees of freedom
AIC: 122.89

Number of Fisher Scoring iterations: 8
```

The class benign is set to be 1 and malignant is set to be 0. Result for the logistic model is shown above.

```
> G <- q1model1$null.deviance - q1model1$deviance
> G
[1] 781.462
> qchisq(0.95, df = 9)
[1] 16.91898
```

The G value is 781.462, which is significantly greater than the 95% quantile of chi-square distribution, which is 16.91898. So, we can conclude that the model is significant.

b).

```
> #q1b
> z <- qnorm(0.975)
> z
[1] 1.959964
> beta_hat <- q1model1$coefficients[2]
> beta_hat
Cl.thickness
  -0.5350141
> se <- 0.14202
> Conf_low <- beta_hat - z * se
> Conf_low
Cl.thickness
  -0.8133682
> Conf_high <- beta_hat + z * se
> Conf_high
Cl.thickness
    -0.25666
```

The 95% confidence interval for beta_hat{CI.thickness} is [-0.8134, -0.2567]. For the hypothesis testing for beta_{cell.shape}, we can see that the p-value of beta_hat{cell.shape} is larger than 0.1, so we fail to reject the null hypothesis at the significant level of 0.1.

c).

```
> #Q1c
> q1model2 <- glm(ClassIndex ~ Cl.thickness + Cell.shape + Marg.adhesion + Bare.nuclei + Bl.cromatin, data =
 q1data, family = "binomial")
> summary(q1model2)

Call:
glm(formula = ClassIndex ~ Cl.thickness + Cell.shape + Marg.adhesion +
    Bare.nuclei + Bl.cromatin, family = "binomial", data = q1data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3713  -0.0234   0.0624   0.1242   3.2982

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   9.74114    1.04989   9.278  < 2e-16 ***
Cl.thickness -0.62576    0.13373  -4.679 2.88e-06 ***
Cell.shape   -0.48994    0.15379  -3.186 0.001444 **
Marg.adhesion -0.33918    0.11221  -3.023 0.002505 **
Bare.nuclei  -0.37330    0.09381  -3.979 6.91e-05 ***
Bl.cromatin  -0.55731    0.16341  -3.411 0.000648 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 112.57  on 677  degrees of freedom
AIC: 124.57

Number of Fisher Scoring iterations: 8
```

The summary of the model is shown above. We can see that the AIC values of the reduced model and the full model are similar. So, we can conclude that the reduced model and the full model are equally predictable regression model. For further comparison, we can make a hypothesis testing that H0 = reduced model is suitable versus H1 = full model is suitable, where p = 5, p + q = 9, and 2[L(p+q) - L(p)] = 9.68 < qchisq(0.01,4) = 13.7267. So, we fail to reject the null hypothesis, implying that we should use the reduced model instead of the full model.

d).

P(Class = "benign" | Cl.thickness=6, Cell.shape=3, Marg.adhesion=8, Bare.nuclei=2, Bl.cromatin=5) = 0.1506149

```
> #Q1d
> q1d_data <- c(1,6,3,8,2,5)
> e <- exp(sum(q1model2$coefficients * q1d_data))
> e / (1+e)
[1] 0.1506149
```

e).

```
> q1e_data <- cbind(q1data$Cl.thickness, q1data$Cell.shape, q1data$Marg.adhesion, q1data$Bare.nuclei, q1data
$Bl.cromatin)
> q1e_data <- cbind(as.data.frame(q1e_data), q1data$ClassIndex)
> colnames(q1e_data) <- c("Cl.thickness", "Cell.shape", "Marg.adhesion", "Bare.nuclei", "Bl.cromatin", "Clas
sIndex")
> bestglm(q1e_data, IC="AIC", family = binomial)$BestModel
Morgan-Tatar search since family is non-gaussian.

Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)

Coefficients:
  (Intercept)   Cl.thickness     Cell.shape  Marg.adhesion    Bare.nuclei    Bl.cromatin
       9.7411        -0.6258        -0.4899        -0.3392        -0.3733        -0.5573

Degrees of Freedom: 682 Total (i.e. Null);  677 Residual
Null Deviance:      884.4
Residual Deviance: 112.6        AIC: 124.6
```

Using bestglm() function in R, the model in part (c) has the smallest AIC (124.57). So the model is part (c) is the best model by the AIC method, and the following variables are used: CI.thickness, Cell.shape, Marg.adhesion, Bare.nuclei, and Bl.cromatin.