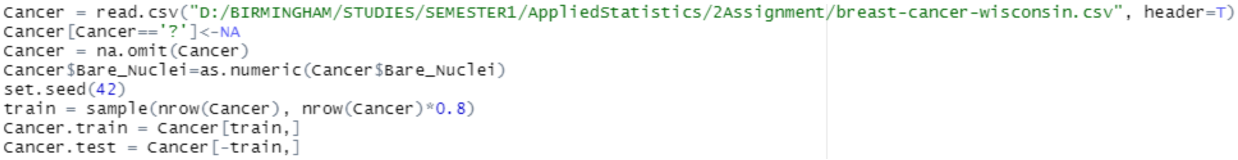
Applied Statistics

Assignment 2

1. This breast cancer dataset was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. (Wolberg and Mangasarian, 1990. *Proceedings of the National Academy of Sciences*). The objective is to predict the whether a tumor is benign or malignant based on several characteristics of the tumor. The variable information are as follows:

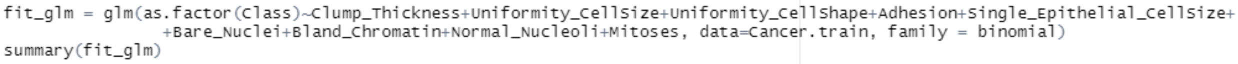
	Variable	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 - 10
3.	Uniformity of Cell Size	1 - 10
4.	Uniformity of Cell Shape	1 - 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class:	(2 for benign, 4 for malignant)

(a) Construct a training set randomly using 80% of the data and use the remaining 20% as the test set. Use `set.seed(42)` to set the seed in random sampling in R.

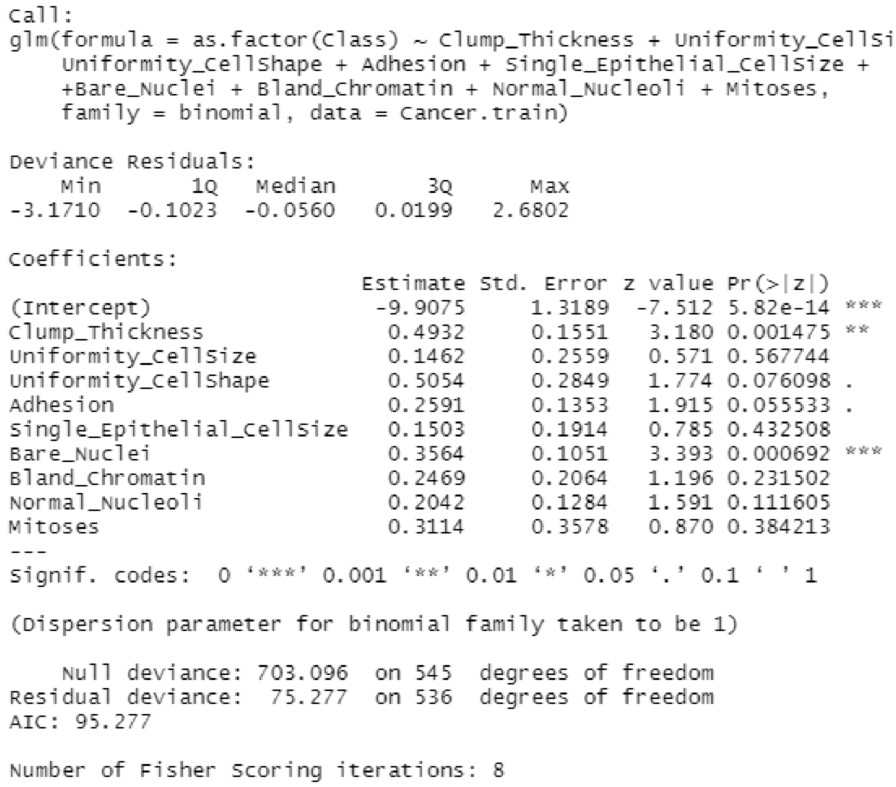


```
Cancer = read.csv("D:/BIRMINGHAM/STUDIES/SEMESTER1/AppliedStatistics/2Assignment/breast-cancer-wisconsin.csv", header=T)
Cancer[Cancer=='?']<-NA
Cancer = na.omit(Cancer)
Cancer$Bare_Nuclei=as.numeric(Cancer$Bare_Nuclei)
set.seed(42)
train = sample(nrow(Cancer), nrow(Cancer)*0.8)
Cancer.train = Cancer[train,]
Cancer.test = Cancer[-train,]
```

(b) Using variables 2-10, fit a logistic regression model to predict the class of the tumor using the training set. Comment on the significance of the individual variables.



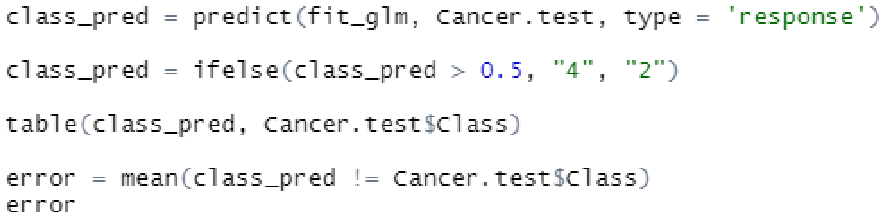
```
fit_glm = glm(as.factor(Class)~Clump_Thickness+Uniformity_CellSize+Uniformity_CellShape+Adhesion+Single_Epithelial_CellSize+
                 +Bare_Nuclei+Bland_Chromatin+Normal_Nucleoli+Mitoses, data=Cancer.train, family = binomial)
summary(fit_glm)
```



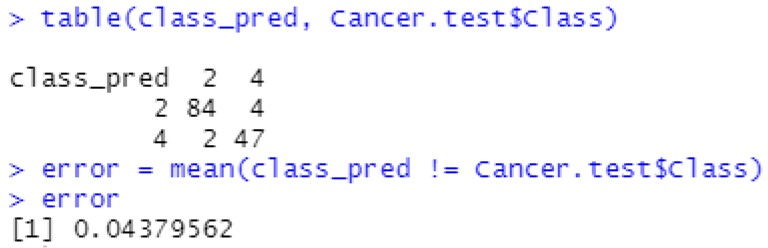
```
Call:
glm(formula = as.factor(Class) ~ Clump_Thickness + Uniformity_CellSize +
    Uniformity_CellShape + Adhesion + Single_Epithelial_CellSize +
    +Bare_Nuclei + Bland_Chromatin + Normal_Nucleoli + Mitoses,
    family = binomial, data = Cancer.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1710  -0.1023  -0.0560   0.0199   2.6802

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -9.9075     1.3189  -7.512 5.82e-14 ***
Clump_Thickness              0.4932     0.1551   3.180 0.001475 **
Uniformity_CellSize          0.1462     0.2559   0.571 0.567744
Uniformity_CellShape         0.5054     0.2849   1.774 0.076098 .
Adhesion                     0.2591     0.1353   1.915 0.055533 .
Single_Epithelial_CellSize   0.1503     0.1914   0.785 0.432508
Bare_Nuclei                  0.3564     0.1051   3.393 0.000692 ***
Bland_Chromatin              0.2469     0.2064   1.196 0.231502
Normal_Nucleoli              0.2042     0.1284   1.591 0.111605
Mitoses                      0.3114     0.3578   0.870 0.384213
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 703.096  on 545  degrees of freedom
Residual deviance:  75.277  on 536  degrees of freedom
AIC: 95.277

Number of Fisher Scoring iterations: 8
```

After calculating each variable’s P-value, we get Clump_Thickness (p = 0.0014), and Bare_Nuclei (p=0.0006) are significant. Other variables are insignificant.

(c) Compute the misclassification error for the test data from the above model.

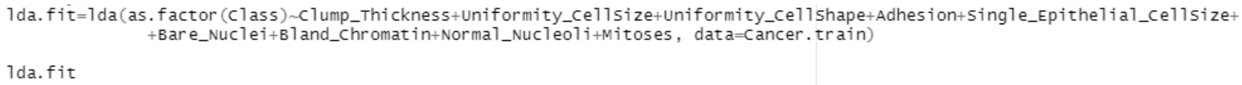


```
class_pred = predict(fit_glm, Cancer.test, type = 'response')

class_pred = ifelse(class_pred > 0.5, "4", "2")

table(class_pred, Cancer.test$Class)

error = mean(class_pred != Cancer.test$Class)
error
```

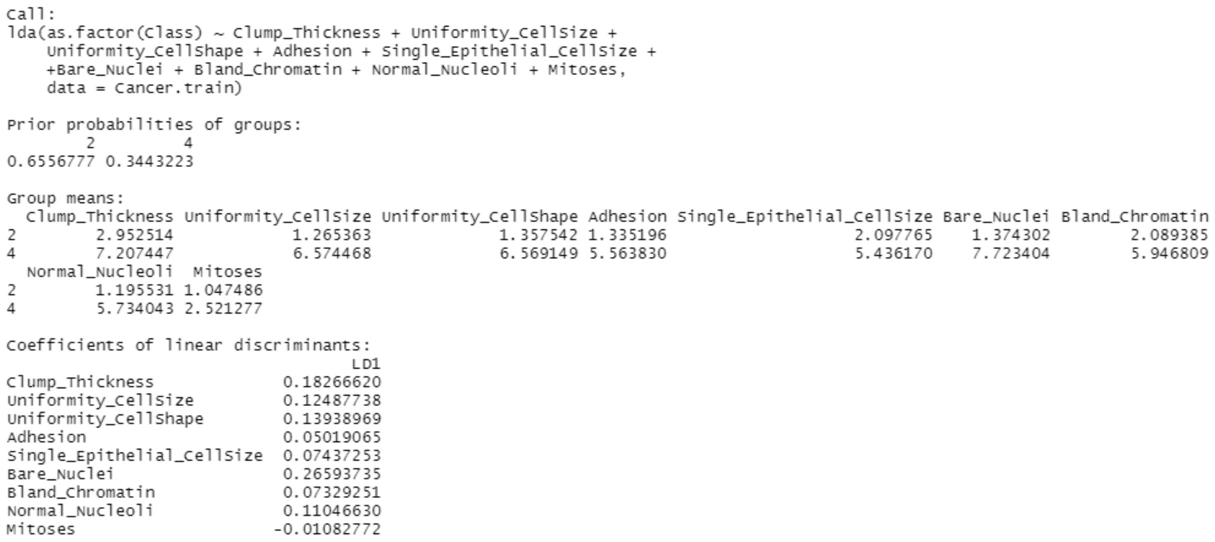


```
> table(class_pred, Cancer.test$Class)

class_pred  2  4
         2 84  4
         4  2 47
> error = mean(class_pred != Cancer.test$Class)
> error
[1] 0.04379562
```

(d) Fit a linear discriminant analysis model to the training set to predict the class of tumor. Report your R code and output.

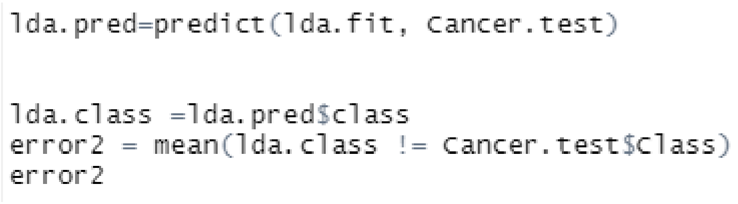


```
lda.fit=lda(as.factor(Class)~Clump_Thickness+Uniformity_CellSize+Uniformity_CellShape+Adhesion+Single_Epithelial_CellSize+
            +Bare_Nuclei+Bland_Chromatin+Normal_Nucleoli+Mitoses, data=Cancer.train)

lda.fit
```

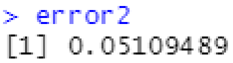


```
Call:
lda(as.factor(Class) ~ Clump_Thickness + Uniformity_CellSize + 
    Uniformity_CellShape + Adhesion + Single_Epithelial_CellSize + 
    +Bare_Nuclei + Bland_Chromatin + Normal_Nucleoli + Mitoses, 
    data = Cancer.train)

Prior probabilities of groups:
        2         4 
0.6556777 0.3443223 

Group means:
  Clump_Thickness Uniformity_CellSize Uniformity_CellShape Adhesion Single_Epithelial_CellSize Bare_Nuclei Bland_Chromatin
2        2.952514            1.265363             1.357542 1.335196                   2.097765    1.374302        2.089385
4        7.207447            6.574468             6.569149 5.563830                   5.436170    7.723404        5.946809
  Normal_Nucleoli  Mitoses
2        1.195531 1.047486
4        5.734043 2.521277

Coefficients of linear discriminants:
                                   LD1
Clump_Thickness             0.18266620
Uniformity_CellSize         0.12487738
Uniformity_CellShape        0.13938969
Adhesion                    0.05019065
Single_Epithelial_CellSize  0.07437253
Bare_Nuclei                 0.26593735
Bland_Chromatin             0.07329251
Normal_Nucleoli             0.11046630
Mitoses                    -0.01082772
```

(e) Compute the misclassification error for the test data from the above model.



```
lda.pred=predict(lda.fit, Cancer.test)


lda.class =lda.pred$class
error2 = mean(lda.class != Cancer.test$Class)
error2
```



```
> error2
[1] 0.05109489
```

(f) Comment on the relative performance of logistic regression and linear discriminant analysis.

Relying on errors from each model, logistic regression performs better in classifying this data set.

$\boxed{2}$ $Y \sim \text{Bernoulli}(0{,}5)$, $P(Y=1) = 1/2 = P(Y \neq 1)$

If $Y=1$, then $X \sim \text{Bernoulli}(p)$

$Y \neq 1$, $X \sim \text{Bernoulli}(q)$, $p > q$

(a) Derive $P(Y=1|X)$

By Bayes' theorem:

$$P(Y=1|X) = \frac{f_1(x) \cdot P(Y=1)}{f_0(x)P(Y\neq 1) + f_1(x)P(Y=1)}$$

$$f_1(x) = p^x(1-p)^{1-x}, \quad x=0,1$$

$$f_0(x) = q^x(1-q)^{1-x}, \quad x=0,1$$

$$P(Y=1|X) = \frac{\frac{1}{2}p^x(1-p)^{1-x}}{\frac{1}{2}q^x(1-q)^{1-x} + \frac{1}{2}p^x(1-p)^{1-x}} =$$

$$= \frac{p^x(1-p)^{1-x}}{q^x(1-q)^{1-x} + p^x(1-p)^{1-x}}$$

(b) What is the Bayes optimal classifier?

Bayes classifier is optimal, when there is the lowest probability of error

The Bayes Classification Rule:

$C(x) = j$ if $\pi_j f_j = \max\{\pi_0 f_0(x), \dots \pi_k f_k(x)\}$

In our case $k=1$, $\pi_0 = \pi_1 = 1/2$:

$$\begin{cases} c(x) = 0 & \text{if } \delta_0 > \delta_1 \ ① \\ c(x) = 1 & \text{if } \delta_0 < \delta_1 \ ② \end{cases}$$

① $\frac{1}{2}q^x(1-q)^{1-x} > \frac{1}{2}p^x(1-p)^{1-x}$

TRUE only when $x=0$

② $\frac{1}{2}q^x(1-q)^{1-x} < \frac{1}{2}p^x(1-p)^{1-x}$

TRUE only when $x=1$

$$\Rightarrow \begin{cases} C(x=0)=0 \\ C(x=1)=0 \end{cases} \rightarrow \boxed{C(x) = X, \ x=0,1}$$

(c) Compute the total probability of misclassification:

$$P(c(x) \neq x) = P(Y=0, X=1) + P(Y=1, X=0) =$$

$$= P(Y=0)\, P(X=1|Y=0) + P(Y=1)\, P(X=0|Y=1) \textcircled{=}$$

we know that

$$P(X=1|Y=1) = p$$

$$P(X=1|Y=0) = q$$

$$P(X=0|Y=1) = 1-p$$

$$P(X=0|Y=0) = 1-q$$

$$\textcircled{=} \frac{1}{2}(1-p) + \frac{1}{2}q$$