Neha Maddali

Problem 1:

a) Misclassification rate: (25+32) / 218 = 0.26146
False negatives are worse in this case - saying someone is healthy when they're actually sick. Change the posterior probability threshold so that it is easier to classify someone as sick.

b) 1) $\delta_0(x_1) = x_1*3.4 / 4.5 + 3.4/2*4.5 + \log(0.32)$
$\delta_1(x_1) = x_1*5.1 / 4.5 + 5.1/*24.5 + \log(0.68)$
If $\delta_0(x_1) > \delta_1(x_1)$, then the test observation will be assigned to Y=0
If $\delta_0(x_1) \leq \delta_1(x_1)$, then the test observation will be assigned to Y=1
2) a threshold of 0.5 would give the smallest possible test misclassification rate. The average of the probabilities of Y in the test set being in either class is 0.35+0.65 / 2 = 0.5. So having this as the threshold will give the smallest overall test misclassification rate.

c) In a dataset, when p predictors are large relative to the sample size, calculating the distance between the points will be computationally expensive. Basically, the complexity of the KNN algorithm increases when there are too many predictors so the predictions can be affected.

d) i) LDA because logistic regression breaks down when the sample size is this small
ii) logistic regression because it works better when the sample size is larger
iii) K-NN because the decision boundary is complicated and highly non-linear. KNN is best for this setting.

e) I think LDA will perform better on the test set. LDA is less flexible than QDA. So it might have a higher bias but could have a smaller variance. This means that LDA is less likely to overfit and could do better on the test set.

f) I think QDA will perform better on the test set. QDA leads to models with smaller bias and a higher variance. This means we could have a more flexible model. This could result in a model that is overfit and performs well on the training set, but not the test set.

g) The probability that X is X is 1. We do not need to worry about estimating P(X) because this is estimating the probability on X. With 1 as a denominator, it can be disregarded.

Problem 2:

a) pr = exp(B0+B1*x1+B2*x2+B3*x3) / (1+exp(B0+B1*x1+B2*x2+B3*x3))

b)

glm.pred	0	1
0	108	12
1	35	345

Misclassification rate: 0.094

c)

	0	1
0	105	6
1	38	351

Misclassification rate: 0.088

d) K = 5

```
          test.Y
knn.pred   0   1
       0  97  12
       1  46 345
```

Misclassification rate: 0.116

e) LDA did the best based on the misclassification rate. Logistic regression did the next best and KNN did the worst.

Problem 3:

a)
```
Prior probabilities of groups:
     Down        Up
0.4477157 0.5522843

Group means:
            Lag2
Down -0.03568254
Up    0.26036581

Coefficients of linear discriminants:
            LD1
Lag2 0.4414162
```

b) Predicted probability of up: 0.5730006
Predicted probability of down: 0.4269994
I would classify this as Up. This doesn't match what we observe, as it is actually Down

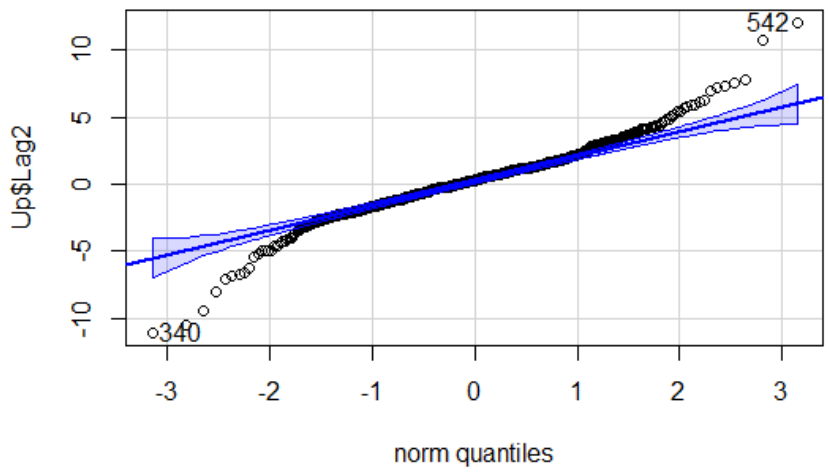
c) Predicted probability of up: 0.5263445
Predicted probability of down: 0.4736555
I would classify that as Up. This doesn't match what we observe, as it is actually Down.

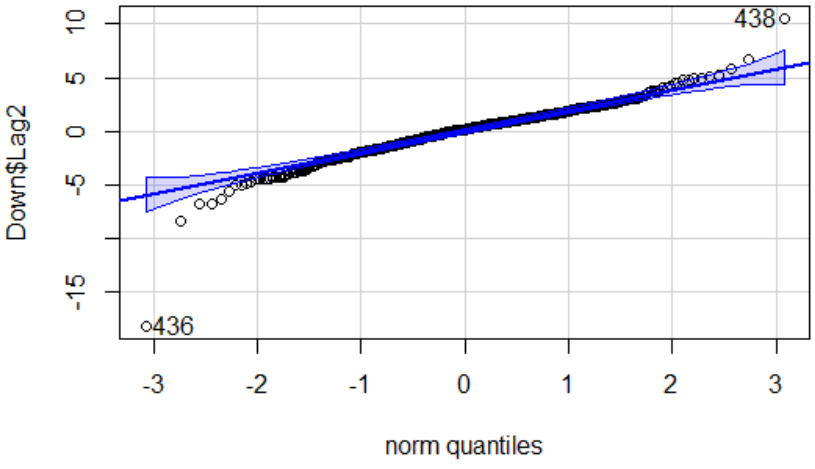
d)
```
      Down Up
Down     9  5
Up      34 56
```

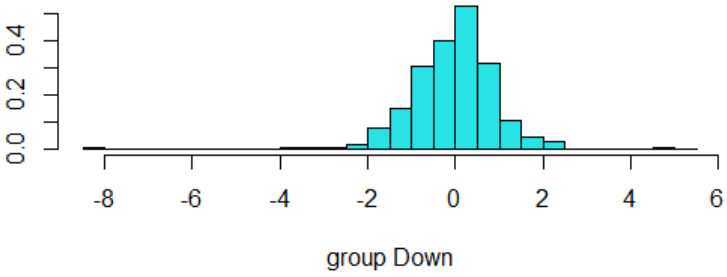
Correct predictions: 65/104
Accuracy rate is 62.5%

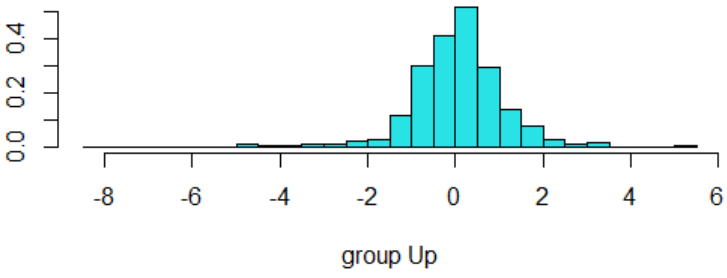
e) LDA makes these two assumptions:

1) $P(X_1 \mid Y = k)$ is a normal distribution. This assumption roughly holds as the points roughly follow the reference line in the quantile-quantile plots.
2) $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$. This assumption doesn't seem to hold because variance for Lag2 when Direction = Up is 5.756206 and variance for Lag2 when Direction = Down is 5.25.2452









f)

	Down	Up
Down	0	0
Up	43	61

Correct predictions: 61/104

Accuracy rate is 58.65%

g) I chose K=3 for KNN

```
set.seed(1)
train.X = cbind(training.data$Lag2)
test.X = cbind(test.data$Lag2)
train.Y = cbind(training.data$Direction)
knn.pred = knn(train.X, test.X, train.Y, k=3)
table(knn.pred, test.data$Direction)
```

```
knn.pred Down Up
       1   16 20
       2   27 41
```

Correct predictions: 57/104

Accuracy rate is 54.80%

h) It looks like LDA provides the best results on this data because of an accuracy rate of 62.5% (the highest fraction of correct predictions)