-Following the lecture and textbooks, please answer the following questions An Introduction to Statistical Learning (with python application) Chapter 4 Exercise: (5), (6)

4.5) We now examine the differences between LDA and QDA.

(a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

Answer: As given Bayes decision boundary is linear and if data has no noise or complexity then LDA will perform better on both training and testing set than QDA.

Regardless of Bayes decision boundary, QDA will perform better than LDA on the training set because it is more flexible and leads to a closer fit. However, LDA perform better than QDA on the testing set.

(b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

Answer: Because of additional flexibility QDA will perform better than LDA on both testing and training set when the Bayes decision boundary is non-linear. QDA with additional flexibility may best fit in some of the non-linearity.

(c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

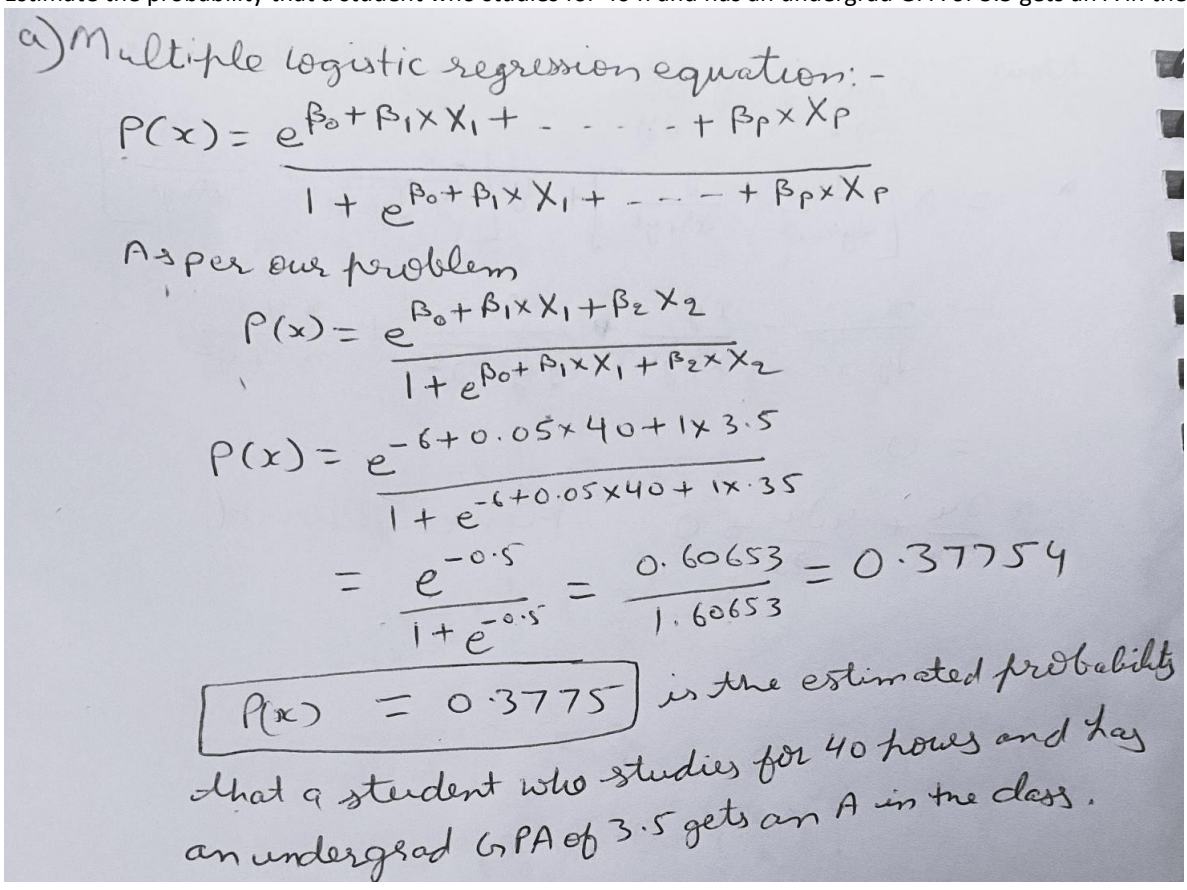
Answer: As the sample size n increases the test prediction accuracy of QDA relative to LDA will improve because with the increase of sample size variance is reduced which further decrease the overfitting of QDA due to its excessive flexibility. So, with sample size increment QDA perform better in prediction.

(d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

Answer: False, QDA is excessive flexible in comparison of LDA which will cause it to overfit while testing when Bayes decision boundary is linear. LDA always offers better performance since it is unbiased and less prone to noise. With less parameters and capability to generalize better to unseen data LDA is more likely to provide a better test error rate.

6. Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6, \hat{\beta}_1 = 0.05, \hat{\beta}_2 = 1.$

a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.



b) How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?

b) for 50·1 chance $P(x) = 0.5$, putting same in above equation

$$p(x) = \frac{e^{-6+0.05\times X_1 + 1\times 3.5}}{1+e^{-6+0.05\times X_1+1\times 3.5}} \Rightarrow$$

$$0.5 = \frac{1}{2} = \frac{e^{-6+0.05\times X_1+1\times 3.5}}{1+e^{-6+0.05\times X_1+1\times 3.5}}$$

from above we can say that $e^{-6+0.05\times X_1+1\times 3.5} = 1$ to match the given condition.

$$e^{-6+0.05\times X_1+1\times 3.5} = 1$$

$$-6+0.05\times X_1+1\times 3.5 = \log(1) = 0$$

$$-6+0.05X_1+3.5 = 0$$

$$0.05X_1 - 2.5 = 0$$

$$X_1 = \frac{2.5}{0.05} = 50$$

So to get A with 50% probability a student needs to study 50 hours.