1. While both of them assumes a **Gaussian distribution** of the data, the main difference between LDA and QDA is that in a LDA, we assume **the Simgas (covariance matrix)** are equal in all classes, while in a QDA we assume they have different Sigmas. Further, LDA is specific to a **linear boundary** between different classifiers while QDA applies to non-linear boundary.

In a LDA, we assume same covariance matrix, so when we take the natural log of the distribution, the **quadratic term** can be canceled out. However, by assuming different covariance matrices, these terms would not be treated as constants that do not depend on k, thus **won’t cancel out**. This will lead to a term of polynomial of degree 2(i.e., -1/2\*X’\*Sigma^-1\*X), and thus results in nonlinear boundaries.

QDA has a quadratic covariance matrix for each pariwise combinations of p features.

1. In a logit model, we find the parameters Beta0 and Beta 1 to maximize **the likelihood function**, which is the products of probabilities of the observed 0s (i.e., (1-P(x)) and 1s (i.e., P(x)=Pr(y=1|X=x)) in the data by setting the **first order conditions to zero** and find the optimized values of estimated parameters.

It is similar to linear regression except it is a binary classification method. Instead of fitting a line to the data a sigmoid function is fit to the data. The objective function is a sigmoidal curve. Logistic regression does not fit the objective function through least squares as is done for linear regression. Logistic regression does not have the same concept as residuals to calculate the least squares. Instead Logistic regression uses maximum likelihood. This works by selecting p(x). Thisis then used to calculate the likelihood to observe the classes depicted in the data. The likelihood for all the points in the data are then multiplied together to give the total likihood of the data given p(x). This process is repeated for multiple probabilities p(x) until the p(x) that gives the maximum likelihood is selected.

1. Since the p-value of X is less than 0.05, X is highly significant indicating X **is** **a useful** feature for the prediction. A negative coefficient of X suggests **larger value of X** will decrease the probability of getting 1 as the result (i.e., tend to be **classified as 0**).
2. X is **useless** in this case because performing a t-test (or F-test if 2 covariates) and getting the means are not statistically different will result in the p-value to be large if we run the regressions, indicating the covariates are not going to be significant.
3. True Positive rate=# of true positives predicted/# of true positives

False Positive rate=# of false positives predicted/# of true positives

We can show that no matter what value r is, the TP rate and FP rate will be **approximately 0.5**, given the Bernoulli and independent distributed of X. To be more specific, define the event Z to be the outcome of the classification results, where Z=1 if X=Y=1 (TP), and Z=0 if X=0 but Y=1(FP) :

The True Positive rate=Pr(Z=1 |Y=1)= Pr(Y=1|Z=1)\*Pr(Z=1)/Pr(Y=1) by Bayes rule.

Here by setting, the distribution of Y is independent from X and Z, thus Pr(Y=1|Z=1)=Pr(Y=1).

Thus by cancelling the terms, we can show that the True Positive rate eventually=Pr(Z=1)=0.5 by setting. The False positive rate=0.5 as well (approximately).