This quarter, we explored classification methods, particularly logistic regression, as a key tool for predicting binary outcomes. Our learning included the following key concepts:

1. Overview of Classification

Classification problems occur frequently in various domains, such as medical diagnosis, fraud detection, and genetics. Unlike regression, which predicts continuous outcomes, classification assigns labels to observations based on predictor variables. For example, predicting credit default based on annual income and monthly balance.

2. Why Not Linear Regression

Linear regression is not appropriate for qualitative response variables because it assumes a continuous outcome. Assigning numerical values (e.g., 1 for stroke, 2 for drug overdose, 3 for epileptic seizure) creates an artificial ordering. Linear regression can produce predictions outside the valid range (e.g., negative values or values greater than 1), making probability interpretation problematic.

3. Logistic Regression as a Solution

Logistic regression models the probability of an event occurring, ensuring outputs remain within the range [0,1]. For example, the probability of default given a balance value is modeled as: Pr(default=Yes|balance). Unlike linear regression, logistic regression provides a probabilistic framework for classification.

4. The Logistic Model

   - The logistic function maps predictor values to probabilities using the logit transformation: $p(X) = \frac{e^{\beta 0 + \beta 1 X}}{1 + e^{\beta 0 + \beta 1 X}}$
   - Odds Representation: $\frac{p(X)}{1-p(X)} = e^{\beta 0 + \beta 1 X}$
   - Log-Odds Transformation: $log(\frac{p(X)}{1-p(X)}) = \beta 0 + \beta 1 X$

The logistic function ensures valid probability estimates and allows for interpretation through log-odds.

5. Maximum Likelihood Estimation (MLE)

Instead of minimizing errors (as in OLS regression), logistic regression maximizes the likelihood function to estimate parameters.

   - The likelihood function is defined as: $l(\beta 0, \beta 1) = \prod_{i:yi=1} p(xi) \prod_{i':yi'=0} (1 - p(xi))$

The estimated coefficients maximize the probability of observing the given dataset.

6. Alternative Link Functions

In addition to the logistic function, we have also read about other link functions, including the probit function, complementary log-log function, and log-log function. But the logistic function is widely used due to its symmetric properties and ease of interpretation.

For the final project, we applied logistic regression to a clinical dataset(Heart Failure Clinical Records dataset) to identify key predictors of patient mortality. The outcome Y is death and the predictors X are Serum Creatinine and Ejection Fraction(determined using forward and backward selection methods) Then we ran a univariate logistic regression for each predictor. And the final conclusion is that higher levels of serum creatinine increased mortality risk and higher levels of ejection fraction decreased mortality risk.