Question 4a : Missing Data

i. (2 Marks) What is Missing Data? Discuss the implications of Missing Data in the context of a statistical analysis.

ii. (3 Marks) Compare and contrast the following types of missing data: Missing At Random, Missing Not At Random, Missing Completely at Random.

iii. (3 Marks) Discuss some of the traditional techniques for dealing with Missing Data, making reference to the limitations of each.

iv. (2 Marks) Briefly describe the technique of Multiple Imputation.

1 Logistic Regression

What is the Logit Function The logit function that you use in logistic regression is also known as the link function because it connects, or links, the values of the independent variables to the probability of occurrence of the event defined by the dependent variable.

$$\text{logit}[E(Y)] = log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

Give a brief description of the purpose of the Cox-Snell R-square statistics

In standard regression, R (or R squared in particular) gives you an idea of how powerful your equation is at predicting the variable of interest. An R close to 1 is a very strong prediction, whereas a small R, closer to zero, indicates a weak relationship.

There is no direct equivalent of R for logistic regression.

However, to keep people happy who insist on an R value, statisticians have come up with several R-like measures for logistic regression. They are not R itself, R has no meaning in logistic regression.

Some of the better known ones are:

- Cox and Snell's R-Square
- Pseudo-R-Square
- Hagle and Mitchell's Pseudo-R-Square

2 Supervised Learning

What is the difference between supervised and unsupervised learning?

The difference is that in supervised learning the 'categories' are known.

In unsupervised learning, they are not, and the learning process attempts to find appropriate 'categories'. In both kinds of learning all parameters are considered to determine which are most appropriate to perform the classification. Whether you chose supervised or unsupervised should be based on whether or not you know what the 'categories' of your data are. If you know, use supervised learning. If you do not know, then use unsupervised.

Give an example of a supervised learning methodology and an unsupervised learning methodology