**EKT 720 Assignment 7 (a)**

*Grouped Data*

1. Linear Probability Model
   * Uses the typical regression model with outcomes being qualitative based on the conditional probability P(Y=1|X).
   * Since it is a probability, the restriction on Y is that
   * Calculate relative frequency, , and use it to model
   * Transform model and use Weighted Least Squares so as to remedy heteroscedasticity.
   * Estimate the parameters and transform back to original equation so that statistical inferences can be made.
   * Shortcomings of LPM without transformation:
     + Non-normality of disturbance terms
     + Heteroscedastic variance of the disturbance term
     + Possibility of the nonfulfillment of the restriction
   * Unreliable as goodness of fit measure
2. Logit Model
   * Preferred due to mathematical simplicity
   * Based on logistical distribution function   
      where
   * Model is nonlinear in parameters
     + Transformed using the natural log so that the parameters are linear by transforming the odds equation
   * To estimate, use the relative frequency so that
   * Steps for estimating the logit regression
     + Calculate probability of ‘success’ for each group using relative frequency
     + Obtain logits for each X where
     + Transform equation by using weights to resolve heteroscedasticity
     + Estimate by WLS (OLS on transformed model)
     + Calculate confidence intervals and test hypotheses if is reasonably large
3. Probit Model
   * Estimating model emerging from the normal distribution CDF
   * Based on utility theory, where the utility index () is given by
   * The probability of an event occurring is given by
   * Using relative frequency, can be calculated using   
      .

*Individual Data*

1. LPM
   * Similar to grouped data except with individual data, the probability used is for the respective observation and not with reference to a group.
2. Logit model
   * Estimated using maximum likelihood for each individual observation using the probability for each observation and follows similar steps to that of the GLOGIT.

*Measuring the goodness-of-fit*

1. Count
   * Given by:
   * Classify as 1 ; classify as 0
2. Hosmer-Lemeshow Test
   * Chi-squared goodness-of-fit for grouped data
   * Sample is divided into subgroups ranging from smallest to largest
   * The conventional method is to separate the groups in 10s.
   * The HL statistic is based on the Pearson’s statistic given by  
      where g=number of groups with degrees of freedom=g-2
   * Highly dependent on groupings chosen
3. Gini Index / ROC curves
   * Gini-measure of equality which ranges from 0 to 1
   * ROC curves
     1. Shows the performance of a binary variable
     2. Uses ‘true positive rate’ vs ‘false positive rate’