MI Masetla
12333710

EKT 720 Assignment 7 (a)

Grouped Data

i. 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
$$0 \leq E(Y|X) \leq 1$$
- Calculate relative frequency, $P_i = \frac{n_i}{N_i}$, and use it to model
$\hat{P}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
- 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 R^2 as goodness of fit measure

ii. Logit Model
- Preferred due to mathematical simplicity
- Based on logistical distribution function
$P_i = \frac{1}{1+e^{-Z_i}} = \frac{e^Z}{1+e^Z}$ where $Z_i = \beta_1 + \beta_2 X_i$
- Model is nonlinear in parameters
  - Transformed using the natural log so that the parameters are linear by transforming the odds equation
$$\frac{P_i}{1-P_i} = e^{Z_i}$$
- To estimate, use the relative frequency so that
$$\hat{L}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$
- Steps for estimating the logit regression
  - Calculate probability of 'success' for each group using relative frequency
  - Obtain logits for each X where $\hat{L}_i = \ln(\frac{\hat{P}_i}{1-\hat{P}_i})$
  - Transform equation by using weights to resolve heteroscedasticity
  - Estimate by WLS (OLS on transformed model)

- Calculate confidence intervals and test hypotheses if N_i is reasonably large

iii. Probit Model

- Estimating model emerging from the normal distribution CDF
- Based on utility theory, where the utility index (I_i) is given by

$$I_i = \beta_1 + \beta_2 X_i$$

- The probability of an event occurring is given by

$$P_i = F(Z)$$

- Using relative frequency, $I_i$can be calculated using $I_i = F^{-1}(P_i)$.

Individual Data

i. LPM
   - Similar to grouped data except with individual data, the probability used is for the respective observation and not with reference to a group.

ii. 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

i. Count R^2
   - Given by:

$$Count\ R^2 = \frac{number\ of\ correct\ prediction}{total\ number\ of\ observations}$$

   - Classify $y \geq 0.5$ as 1 ; classify $y < 0.5$ as 0

ii. 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 $HL = \sum_{i=1}^{g} \sum_{j=0}^{1} \frac{(obs_{ij} - exp_{ij})^2}{exp_{ij}}$ where g=number of groups with degrees of freedom=g-2
   - Highly dependent on groupings chosen

iii. Gini Index / ROC curves
   - Gini-measure of equality which ranges from 0 to 1

MI Masetla
12333710

- ROC curves
    i. Shows the performance of a binary variable
    ii. Uses ‘true positive rate’ vs ‘false positive rate’