Assignment 12 Solution

The correct answer is in bold font

Question 1: The following is a suitable form of cumulative probability distribution that can be employed in logit/probit class of models.

(a) $\frac{z}{1-z}$. Hint: When z -> -∞, the model approaches to a value of -1, which is not desirable.
(b) $\frac{1}{1+z}$. Hint: When z is in the following interval (-1,0), the model attains a value of more than 1, which is not desirable.
(c) $\boldsymbol{\frac{1}{1+e^{-z}}}$**. Hint: This is an appropriate 'S' form cumulative probability distribution function which ranges from 0 to 1 for different values of z.**
(d) e^z. Hint: This expression attains a value of more than 1 for z>0, which is not desirable.

Question 2: The following is a correct statement in the context of parameter interpretation for logit/probit models

(a) The interpretation is similar to linear probability models [Hint: Logit/probit class of models are non-linear in parameters and the relationship between the independent and dependent variable follows a cumulative probability function.]
(b) **Parameter interpretation requires computation of marginal effects [Hint: The impact of the independent variable on the dependent variable varies with the magnitude of the independent variable]**
(c) The coefficient β_2 in $z_i = \beta_1 + \beta_2 x_2 + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i$ measures the percentage increase in the probability of observing the event of interest, i.e., $P(y_i = 1)$. [Hint: This interpretation is correct for linear probability model but not logit/probit class of non-linear models.]
(d) Model fitting requires minimization of residual sum of squares [Hint: The OLS procedure entails minimizing the residual sum of squares. However, Logit/Probit models are not estimated through OLS procedure]

Question 3: The following is a correct statement in the context of training and testing the classification algorithm

(a) A model that fits well the insample data will also fit well with the out-of-sample data. [Hint: Often one can overfit the model within a given sample; such model may not offer a good out-of-sample fit]
(b) **Pseudo R-square is an appropriate measure of goodness-of-fit for logit/probit class of models. [Hint: Pseudo R-square measure compares a restricted model with the proposed model with the help of log-likelihood function.]**
(c) A model that classifies positives (1's) accurately will also necessarily classify negatives (0's) accurately. [Hint: There is a trade-off between classifying positives (1's) and negatives (0s) accurately]

(d) If the observations are symmetrically (50:50) distributed between 1's and 0's, then logit and probit approaches offer very different results. [Hint: If 1's and 0's is distributed symmetrically (50:50), then logit and probit approaches offer similar results.]

Question 4: The following is an incorrect statement in the context of logit model.

(a) The model is nonlinear in parameters. Hint: Cumulative logisitic function is non-linear in parameters.
(b) Model can be estimated through maximum likelihood estimation (MLE) method. Hint: Since the model is non-linear in parameters, it cannot be estimated using OLS procedure, hence MLE method is appropriate.
(c) The cumulative logit function appears like an S shaped curve. Hint: The cumulative logit function approaches asymptotically towards '0' on the left side and '1' on the right side.
(d) **The normal distribution function appears like an S shaped curve. Hint: The normal distribution appears like a bell-shaped curve. Cumulative normal distribution function appears like an 'S' curve.**

Question 5: The following is a correct statement in the context of logit model.

(a) Log of odds ratio is non-linearly related to the model variables. Hint: Log of odds is a linear function of zi, i.e., a linear combination of the independent variables.
(b) For a very large threshold ~1, sensitivity is close to one. Hint: Sensitivity is close to zero, since all the positives (1s) are classified as false negatives (0s).
(c) For a very small threshold ~0, specificity is close to one. Hint: Specificity is close to zero, since all the negatives (0s) are classified as false positives (1s).
(d) **A high value of threshold results in low sensitivity and high specificity. Hint: Increase in the thresholding value results in low proportion of 'true positives' and high proportion of 'true negatives'.**

Question 6: The following is an incorrect statement in the context of logit/probit models.

(a) These models are estimated using maximum likelihood estimation (MLE). Hint: Since these models are non-linear in parameters, the models are estimated with MLE.
(b) R-square is a poor goodness-of-fit (GoF) indicator for logit/probit class of models. Hint: Since the dependent variable is a discrete binary (1s and 0s) type and the model does not minimize the residual sum of squares, R-square measure is inappropriate as GoF indicator.
(c) Area under the ROC curve is an appropriate measure of model performance. Hint: ROC curve captures the trade-off between specificity and sensitivity at different threshold values.
(d) **Model coefficients measure the impact of the independent variables on the dependent variable. Hint: The impact of independent variables on the dependent variable is measured with marginal effects.**

Question 7: The following is a correct statement in the context of logit model.

(a) The model is linear in parameters. Hint: Model is non-linear in parameters.

(b) Model can be estimated through ordinary least square method (OLS). Hint: Since the model is non-linear in parameters, it cannot be estimated using OLS procedure.
(c) The cumulative logit function appears like a bell-shaped curve. Hint: The cumulative logit function appears like an S curve (from 0 to 1)
(d) **Thresholding is required because we do not observe probabilities. Hint: To convert probabilities estimated from the logit model into 1s and 0s (i.e., real life observed binary events), we perform thresholding.**

Question 8: Using historical data, a bank manager estimates the logit function for bank default applications. Here $x_1 = Income, x_2 = Wealth, and\ x_3 = Dept\ servicing.$ Here, Y=1 indicates the default event and Y=0 indicates no default. The following cumulative logit function is estimated from the given data. $F(z_i) = \hat{P}_i(Y = 1) = \frac{1}{1+e^{-(0.05+0.4x_1-0.7x_2+0.8x_3)}}$. Assuming the average values of $\bar{x}_1 = 1.5;\ \bar{x}_2 = 0.3, and\ \bar{x}_3 = 0.2$. On average, what is the correct interval in which the probability that the borrower will default, i.e., $\hat{P}_i(Y = 1)$ will lie.

(a) 0.00-0.20[Hint: $\hat{P}_i(Y = 1) = \frac{1}{1+e^{-(0.05+0.4*1.5-0.7*0.3+0.8*0.2)}}$]
(b) 0.20-0.40 [Hint: $\hat{P}_i(Y = 1) = \frac{1}{1+e^{-(0.05+0.4*1.5-0.7*0.3+0.8*0.2)}}$]
(c) 0.40-0.60 [Hint: $\hat{P}_i(Y = 1) = \frac{1}{1+e^{-(0.05+0.4*1.5-0.7*0.3+0.8*0.2)}}$]
(d) 0.60-0.80 [Hint: $\boldsymbol{\hat{P}_i(Y = 1) = \frac{1}{1+e^{-(0.05+0.4*1.5-0.7*0.3+0.8*0.2)}} = 0.65}$]

Question 9: Using historical data, a bank manager estimates the logit function for bank default applications. Here $x_1 = Income, x_2 = Wealth\ and\ x_3 = Dept\ servicing.$ Here, Y=1 indicates the default event and Y=0 indicates no default. The following logit function is estimated from the given data. $F(z_i) = \hat{P}_i(Y = 1) = \frac{1}{1+e^{-(0.05+0.4x_1-0.7x_2+0.8x_3)}}$. Assuming the average values of $\bar{x}_1 = 1.5;\ \bar{x}_2 = 0.3, and\ \bar{x}_3 = 0.2$. Using the value of $\hat{P}_i(Y = 1)$ computed in previous question no *, compute the correct interval for log of odds ratio, i.e., $Log\ (\frac{\hat{P}_i}{1-\hat{P}_i})$. Compute using natural logarithm.

(a) 0.05-0.25[Hint: $Log\ (Odds) = Log\ \left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = 0.05 + 0.4x_1 - 0.7x_2 + 0.8x_3$]
(b) 0.25-0.45 [Hint: $Log\ (Odds) = Log\ \left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = 0.05 + 0.4x_1 - 0.7x_2 + 0.8x_3$]
(c) 0.45-0.65 [[Hint: $\boldsymbol{Log\ (Odds) = Log\ (\frac{\hat{P}_i}{1-\hat{P}_i}) = 0.05 + 0.4 * 1.5 - 0.7 * 0.3 + 0.8 * 0.2 = 0.60}$]
(d) 0.65-0.85 [Hint: $Log\ (Odds) = Log\ \left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = 0.05 + 0.4x_1 - 0.7x_2 + 0.8x_3$]

Question 10: Using historical data, a bank manager estimates the logit function for bank default applications. Here $x_1 = Income, x_2 = Wealth\ and\ x_3 = Dept\ servicing.$ Here, Y=1 indicates the default event and Y=0 indicates no default. The following logit function is estimated from the given data. $F(z_i) = \hat{P}_i(Y = 1) = \frac{1}{1+e^{-(0.05+0.4x_1-0.7x_2+0.8x_3)}}$. Assuming the average values of $\bar{x}_1 = 1.5;\ \bar{x}_2 = 0.3, and\ \bar{x}_3 = 0.2$. Using the value of $\hat{P}_i(Y = 1)$ computed in the previous question no *, compute the correct interval for the increase in probability of default [$\hat{P}_i(Y = 1)$] for a one unit increase in x_1(also called as marginal effect of x_1).

(a) 0.00-0.10 [Hint: Marginal effects= $\boldsymbol{\beta_1 * F(z_i) * \left(1 - F(z_i)\right) = 0.4 * 0.65 * (1 - 0.65) = 0.091\ or\ 9.10\%}$

(b) 0.10-0.20 [Hint: Marginal effects= $\beta_1 * F(z_i) * \left(1 - F(z_i)\right)$]

(c) 0.20-0.30 [Hint: Marginal effects= $\beta_1 * F(z_i) * \left(1 - F(z_i)\right)$]

(d) 0.30-0.40 [Hint: Marginal effects= $\beta_1 * F(z_i) * \left(1 - F(z_i)\right)$]