Data Analytics for Business

Homework 1 – Part 2

Binary Response Models

In this homework exercise, we will apply both linear probability models and logit models to help a local bank to identify what kinds of customers are most likely to accept personal loan offers sent in mail.

Download the data file "**Loan.csv**," which contains the following fields. Complete the tasks outlined below.

Data Description

Column Name	Variable Description
Loan	Whether or not the customer accepted the personal loan offer: 1 – Yes; 0 – No
Income	Annual income of the customer (**in $1000s**)
Family	Family size of the customer
CCAvg	Average spending on credit cards per month (**in $1000s**)
Education	Education level: 1: Undergrad; 2: Graduate; 3: Advanced/Professional

Task 1: Linear Probability Model

1. Import the data into R. Convert "**Education**" into **factor** (you can use **as.factor()** function).
2. Run a linear probability model by regressing **Loan** on all the other variables (**Income**, **Family**, **CCAvg**, **Education**). Show the summary of the regression results. How do you interpret the coefficients in front of the two Education variables?
3. What does the fitted $\hat{y}$ ($\hat{y} = X\hat{\beta}$) from a linear probability model mean? Are there any customers with the fitted $\hat{y}$ being greater than 1 or less than 0 in this data set? Show some of those customers.

Task 2: Logit Model

4. Next, run a Logit model of **Loan** on the same set of independent variables (**Income**, **Family**, **CCAvg**, **Education**) using the **glm()** function. Show the summary of the estimation results.

5. Create the **confusion matrix** and calculate the **Percent Correctly Predicted (PCP)**, both at the overall level and for each possible outcome separately (i.e., PCP for $y = 1$ and $y = 0$, respectively). As we discussed in class, use the fraction of "success" in the original data as the threshold in predicting the binary outcomes.

6. Calculate the **predicted probability of success** according to the Logit model estimation results. Evaluate the probability at such values of the *X* variables: {**Income**, **Family**, **CCAvg**} equal their mean values in the original data and **Education="2".**

   - First calculate the probability **using the defining formula** as explained in class. You may do the calculation in R, but be explicit about the exact formula used and substitute in the exact numbers.
   - Then use the **predict(…, type="response")** function to obtain the predicted probability of success. The value should be the same as your own calculation result above.

7. Co-list and compare the coefficients from the linear probability model and the Logit model. Are the coefficients from the two models directly comparable? Why?

8. Calculate the partial effects of all independent variables based on the Logit model. Again, evaluate the probability at such values of the independent variables: {**Income**, **Family**, **CCAvg**} equal their mean values in the original data and **Education="2".** Also calculate the partial effects based on the linear probability model (**Hint**: simply the β coefficients). Co-list and compare the partial effects from the two models. Are they comparable?