Udit Lal  
ECON 4835 Forecasting

HW#6

1. Run a linear probability model (LPM) regression that predicts whether a consumer defaulted based on their credit scores and on the loan amount. We will call this model reglp. The code runs this model for you, but you need to report and discuss the results. In particular
   1. Report the coefficient estimates and discuss which factors significantly affect the probability of default.
      1. credit\_scoreHIGH: -0.185  
         credit\_scoreLOW: 0.117  
         loan\_amountLOW: -0.077  
         loan\_amountMEDIUM: -0.097
   2. What is the benchmark category for credit\_score that is not reported in the regression table?
      1. credit\_scoreAVERAGE
   3. What is the benchmark category for loan\_amount that is not reported in the regression table?
      1. Loan\_amountHIGH
   4. In 1-2 sentences, informally discuss if the LPM model is suitable for predicting defaults in this application. If you think it is suitable, use a predicted versus actual values graph, a mathematical argument, or report some of the summary statistics for summary(mydata$predlp) to support your argument. If you think the model is not suitable, use a predicted versus actual values graph, a mathematical argument, or report some of the summary statistics for summary(mydata$predlp) to support your argument.
      1. The LPM has a weakness when trying to predict binary outcomes which in this case is loan accepted/rejected. This is because the LPM has continuous outcomes which need to be further analyzed before coming to an accurate prediction.
2. Run a logistical regression model that predicts whether a consumer defaults based on their credit score and on the loan amount. The code runs this model for you, but you need to report and discuss the results. The model is called small\_reg. In particular
   1. Report the coefficient estimates and discuss their significance.
      1. credit\_scoreHIGH: -0.774  
         credit\_scoreLOW: 0.480  
         loan\_amountLOW: -0.317  
         loan\_amountMEDIUM: -0.4
   2. Based on your results, how are credit scores linked to defaults?
      1. The results follow logic in that credit score has an inverse relationship with the rate of defaults.
3. Run a logistical regression that predicts purchases based on all available data. Call this regression big\_reg. NOTE: this model is also included in the code for you. However, if you are experimenting with different models and with the code and are writing this regression yourself, make sure that you’re not accidentally including predlp and predp\_small in your model. Those are predicted probabilities. You only want to include the actual data.
   1. Report all of the coefficients and their significance.
      1. A screenshot of a computer program

         Description automatically generated
   2. Does the purpose of the loan make a difference when it comes to increasing the probability of default? If yes, which loans types (by purpose) significantly increase the probability of default?
      1. Yes, loans for credit cards, debt consolidation, major purchases, vacation, and weddings all are high risk loans which increase rates of defaults. Loans for Education, home improvement, mortgages, medical, moving, energy, and small business all lower the risk of defaulting.
   3. In 1-2 sentences, informally discuss the predictive performance of this model. Does it do better than a naïve model? Does it seem to do better than the small model based on a gain curve? Is it hard to tell? You can use any of the approaches to assess the predictions (actual vs predicted for both the small and the large model or Gain Curves for both models).
      1. The gain curve for the “small” model has a much higher true positive rate which means that it is more accurate at determining positive results.
4. Split up the data into a test/train sample (70% train, 30% test) and re-estimate both the small and the large logistical regression model. Again, the code is provided to you, you will need to read the results.
   1. How may defaults are correctly predicted using the small model for the test sample? How many defaults are incorrectly classified as non-defaults using the small model for the test sample? What is the overall accuracy of the small model in the test sample? Pay attention to which model is reported first in the code.
      1. 1129 true positive results.   
         740 true negative results.  
         There are 1869 correct predictions out of 3394 meaning the model is 55% correct.
   2. How may defaults are correctly predicted using the large model for the test sample? How many non-defaults are incorrectly classified as defaults using the large model for the test sample? What is the overall accuracy of the large model in the test sample? Pay attention to which model is reported first in the code.
      1. 994 true positive  
         1079 true negative   
         2073 correct results out of 3394 meaning the model is 61% correct.
   3. Report the ROC curves for both models on a single graph. If you use ROC as visual supporting evidence, which model is favored by the data? The small or the large? Again, pay attention to which model is first in the code.
      1. A graph of a curve

         Description automatically generatedSmall is first and big is second