Prof. James MacKinnon QUEEN'S UNIVERSITY Department of Economics

Economics 882 Winter, 2025

Assignment 2

Due: February 25, 2025

This assignment uses a simulated dataset called `loan-defs.csv`. The variables are `default` (whether or not someone defaulted on their credit card debt), `age`, `balance` (the most recent balance), `educ` (their education level, from 1 to 5), `gender` (male = 1, female = 2), `income` (in dollars per year), `own` (1 if they own a house or condo, 0 if they rent), and `miss` (the number of missed payments over the past year).

1. The dataset contains 20,000 observations. Start by randomly splitting the sample into a training sample with 15,000 observations and a test sample with 5000. [10]

2. Using the training sample, estimate one or more linear probability models to explain `default`. Find one that you think makes sense and fits well. [15]

3. Using the training sample, estimate one or more logistic regression models to explain `default`. Find one that you think makes sense and fits well. [15]

4. Using the training sample, and starting with the model you chose for question 3, estimate at least one logistic lasso model and at least one logistic ridge model to explain `default`. Use x-fold cross-validation (for an appropriate choice of x) to determine the tuning parameters. Does using either form of regularization seem to be a good idea in this case? [15]

5. Using the test sample, compute two confusion matrices for each of the best model of question 2 and the best model of either question 3 or question 4, depending on your answer to the final part of question 4. One should be based on a threshold of 0.5, and the other on a threshold of 0.2. What do you conclude? [15]

6. Using the test sample, plot ROC curves for the best linear probability model of question 2, the best logistic regression model of question 3, and the best logistic lasso or ridge model of question 4. What do you conclude? [15]

7. Using the entire sample, plot four kernel density estimates for the logarithm of `income`. Two of these should use the default bandwidh and the Gaussian or Epanechnikov kernels. The other two should use the Epanechnikov kernel with either 2/3 of the default bandwith or 1.5 times the default bandwidth. Which density or densities do you like best? Which ones do you dislike? Does this variable seem to be symmetrically distributed? Explain. [15]