\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Homework 5

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Instructions:

\* To create this document, first copy and paste the full text here into a .Do document (a STATA Do-File).

\* Below each question, write the code you used to answer the question

\* Next, write your actual answer to the question by commenting out your writing (by starting the line with a \*)

\* Next, copy and paste the entire document (my writing and yours) into a Word document. This will allow me to see your code on Canvas without downloading every homework.

\* The goal is that I should be able to copy and paste your entire text into a .Do File and run the code without any errors.

\* Finally, submit file as Homework 5 on Canvas

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Topic 1: Logistic Regressions

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*1. Import the LendingData excel file into Stata

\* The dataset includes three variables:

\* home\_ownership: Binary variable where a 1 denotes home ownership, 0 is a renter

\* income: income of borrower

\* dti: Debt to income Ratio

\* fico: Credit score

\* loan\_status: Binary variable where a 1 denotes repayment, 0 is default

\* TestData: Binary variable where a 1 denotes Test Data, 0 denotes Training Data

clear

cd "C:\Users\haniu\OneDrive\Desktop\Deepa\Deepa\Finance Core B\Business\Homework 5"

import excel "LendingClubData.xlsx", firstrow clear

\* (6 vars, 9,290 obs)

\*2. run a logistic regression on the training data with Loan Status as the y-variable and credit score as the x-variable for the Training DAta

logit loan\_status fico if TestData==0

\* 3. Predict Loan Status from the regression above

predict predict\_Loan\_Status

\* (option pr assumed; Pr(loan\_status))

\*4. Create a scatter plot with loan status as the y-variable and credit score as the x-variable. Plot the predicted loan status on top

twoway (scatter loan\_status fico) (connected predict\_Loan\_Status fico, sort)

\*5. Estimate four alternative logistic models using the training data:

\* Model 1: Loan Status is predicted by credit score

\* Model 2: Loan Status is predicted by credit score and Debt-to-Income

\* Model 3: Loan Status is predicted by credit score, Debt-to-Income, and Income

\* Model 4: Loan Status is predicted by credit score, Debt-to-Income, Income, and Home Ownership

\* For each model estimate the predicted loan\_status and the residual

\*Model 1

logit loan\_status fico if TestData==0

predict predict\_Loan\_Status

gen residual\_logit1 = loan\_status-predict\_Loan\_Status

\*Model 2

logit loan\_status fico dti if TestData==0

predict predict\_Loan\_Status2

gen residual\_logit2 = loan\_status-predict\_Loan\_Status2

\*Model 3

logit loan\_status fico dti income if TestData==0

predict predict\_Loan\_Status3

gen residual\_logit3 = loan\_status-predict\_Loan\_Status3

\*Model 4

logit loan\_status fico dti income home\_ownership if TestData==0

predict predict\_Loan\_Status4

gen residual\_logit4 = loan\_status-predict\_Loan\_Status4

\*6. Estimate model 4 again using the training data, but with a linear regression

\*Estmate the prediction and residual

regress loan\_status fico dti income home\_ownership if TestData==0

predict predict\_Loan\_Status5

gen residual\_logit5 = loan\_status-predict\_Loan\_Status5

\*7. Compare the MSE of the five models above using the test data

sum residual\* if TestData==1, detail

\*8. z-score all four of the x-variables

foreach var in home\_ownership income dti fico {

egen mean=mean(`var')

egen sd=sd(`var')

gen zscore\_`var' = (`var'-mean)/sd

drop mean sd

}

\*9. Use LASSO to test for the best model using the training data

\* Hint: the new command is lasso logit

lasso logit loan\_status zscore\* if TestData==0

\*10. Estimate the coefficients from the lasso estimation using the command:

\*lassocoef, display(coef, postselection)

\* How many coefficients are there. Why?

lassocoef, display(coef, postselection)

\*There are 4 coefficients since we need all of the data to predict it correctly

\*The values between the both will be different but lets us know that we require all the data to actually predict data properly.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Topic 2: Decision Criterion and ROC Curves

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*11. Keep only the Test Data

keep if TestData==1

\*12. You will need to estimate the True Positive Rate and False Positive Rate not just once, but at many z thresholds.

\* For practice, we will estimate these rates 100 times for every 0.01 increase in the Z-treshold from 0 to 1

\*The first step is to create two new variables

\* TruePositiveRate =.

\*FalsePositiveRate = .

gen TruePositiveRate=0

gen FalsePositiveRate=0

\*13. Next, we need to create a forvalues loop for each of the 100 estimates

\*Use the code

\*forvalues i = 0(.01)1 {

\* ADD YOUR CODE HERE

\* }

forvalues i = 0.01(0.01)1 {

gen dummy = 100\*`i'

gen predict\_repaid=0

replace predict\_repaid=1 if predict\_Loan\_Status4>`i'

gen TP = 0

replace TP=1 if predict\_repaid==1 & loan\_status==1

gen TN = 0

replace TN=1 if predict\_repaid==0 & loan\_status==0

gen FP = 0

replace FP=1 if predict\_repaid==1 & loan\_status==0

gen FN = 0

replace FN=1 if predict\_repaid==0 & loan\_status==1

sum TP TN FP FN

egen meanTP = mean(TP)

egen meanTN = mean(TN)

egen meanFP = mean(FP)

egen meanFN = mean(FN)

replace TruePositiveRate = meanTP/(meanTP+meanFN) if \_n==dummy

replace FalsePositiveRate = meanFP/(meanTN+meanFP) if \_n==dummy

drop dummy

drop TP predict\_repaid TN FP FN mean\*

}

\*\_n is basically which row you are on in the excel (or similar datasheet)

\* you can write capture, if these variables doesnt exists then drop them but if it does have those variables but capture it but keep going

\*14. Step 1 of the forvalue loop: Create a new variable

\* dummy = 100\*`i'

\*14. Step 2 of the forvalues loop: estimate which loans were paid based on the Z-thresholds for each observation

\*15. Step 3 of the loop: Estimate TP, TN, FP, and FN for each observations

\*16. Step 4 of the loop: Estimate the mean TP, TN, FP, and FN

\* Create a variable with the mean of each measure

\*17. Step 5 of the loop: replace the True Positive Rate and False Positive Rate variables if \_n== dummy

\*18. Step 6 of the loop: drop all new variable except the True Positive Rate and False Positive Rate

\*19. Keep only the first 100 observations

\* Create a scatter plot with the True Positive Rate as the y-variable and False Positive Rate as the x-variable

gen ID =\_n

keep if ID<101

\*keep if <\_101 (you can do it this way as well)

twoway (scatter TruePositiveRate FalsePositiveRate)

\*20. Add to the plot above a new line across the 45 degree line

gen x = 0

gen y = 0

forvalues i =0(0.01)1 {

gen dummy = 100\*`i'

replace x = 1-`i' if ID==dummy

replace y = 1-`i' if ID==dummy

drop dummy

}

twoway (scatter TruePositiveRate FalsePositiveRate) (lfit y x)