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

\* Homework4

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

\* 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 4 on Canvas

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

\* Topic 1: Forecasting Future Returns in a Time-Series

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

\*1. Open the excel file StateHousePriceData

\* The dataset includes three vari

\*State: State name

\*Stateid: Numerical Identifier for each state

\* year

\* us\_price: US median housing price

\* mortgage\_rate: US mortgate rates

\* adult\_population: Adult population in each state-year

\* unemployment: Unemployment rate for each state-year

\* price: state median housing price

clear

\*2. Keep only observations where the stateid is 0. This is the national housing data

keep if stateid==0

\*3. \* Use Tsset to create a yearly time series. (use help tsset to read about this function)

tsset year, yearly

\*4. Create a new variable to estimate annual US returns

gen return = 100\*(us\_price - l.us\_price)/l.us\_price

\*5. Run a regression with returns at the y-variable and lagged returns at the x-variable. Then create a scatter plot with returns as the y-variable and lagged returns as the x-variable. Place the linear fit on top of the scatter plot

\* What is the coefficient?

\* This is called 1-year momentum and it is very strong in housing data

reg return l.return

twoway scatter return l.return || lfit return l.return, title("Returns")

\*The coefficient is .8541

\*6. Split the sample into two sets: 2000-2012 and 2013-2022. Consider the first 12 years are your training set and the next 10 are the testing data. Test if returns over the prior year predict returns this year using the training set (years 2000-2012)

reg return l.return if year<2013

\*7. Predict returns using the lagged returns

predict predict\_return1

\*8. Estimate the residual of the prediction

predict residual\_return1, res

\*9. Repeat steps 6-8 but including both returns from last year and lagged mortgage rates as x-variables

reg return l.return l.mortgage\_rate if year<2013

predict predict\_return2

predict residual\_return2, res

\*10. Estimate the MSE of the two forecasts by analyzing the years 2002-2012 (the training years)

sum residual\* if year<2013, detail

\*11. Estimate the MSE of the two forecasts by analyzing the years 2013-2022 (the testing years)

\* Do mortgate rates predict returns beyond prior returns?

sum residual\* if year >= 2013, detail

\*12. Z-score mortgage rates

\* Don't worry about Z-scoring returns

egen mean = mean(mortgage\_rate)

egen sd = sd(mortgage\_rate)

gen zscore\_mortgage\_rate = (mortgage\_rate-mean)/sd

\* YOU ARE Z SCORING BECAUSE YOU ARE PENALIZING LARGE NUMBERS

\*13. Run LASSO on across all years (2000-2022) with return as the y-variable, and returns from the prior year and mortgage rates from the prior year as x-variables

\*To do this, you will need to first create new variables for lagged variables

\* We run our analysis on the full sample as years 2000-2013 have too few observations for LASSO to work well

gen l1return = l.return

gen l1mortgage\_rate = l.zscore\_mortgage\_rate

lasso linear return l1return l1mortgage

\*14. Estimate the coefficients from the lasso estimation using the command:

lassocoef, display(coef, postselection)

\* How many coefficients are there. Why?

/\* There are three coefficients. Both lagged mortgage rates and lag returns are valuable together.

Strong momentum, means there's a positive relationship between last year and this year.

In this case, we see that rising mortgage rates would give lower returns...makes sense.

\*/

\*15. Predict the salary estimate from the lasso estimation using predict

predict predict\_lasso

\*16. Graph the prediction from LASSO along with the scatter plot of the actual returns for 2000-2022

twoway (scatter return year) (connected predict\_lasso year, sort)

graph save LASSO,replace

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

\* Topic 2: Forecasting Future Returns in a Panel Series

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

\*17. Reopen the excel file StateHousePriceData

clear

import excel "/Users/henryvelasquez/Documents/MBA/MBA S3/2nd 7/Machine Learning in Finance K579/HW4/StateHousePriceData.xls", sheet("Sheet1") firstrow clear

\*18. Drop observations where the stateid is 0. This is the national housing data

keep if stateid != 0

\*19. \* Use xtset to create a state-year panel series. (use help xtset to read about this function)

xtset stateid year, yearly

\*20. Estimate the TOTAL housing market return for each state over 2000-2022. This should be a single number for each state-year

/\*gen hMarketReturn = 100\*(us\_price - l.us\_price)/l.us\_price \*/

/\* ^^^this gives you a lagged return \*/

gen return22 = 100\*(price - l22.price)/l22.price

\*21. Estimate the TOTAL percent change in adult population in each state between 2000-2022.

\* Then create a scatter plot with 2000-2022 returns as the y-variable and 2000-2022 population change as the x-variable. Fit a line on top of the scatter plot

gen adultPpChange = ((adult\_population/l22.adult\_population)-1)\*100

twoway (scatter return22 adultPpChange) (lfit return22 adultpPChange)

\*23. Estimate annual returns for each state. Then create a scatter plot with returns as the y-variable and lagged returns as the x-variable. Place the linear fit on top of the scatter plot

gen stateReturns = 100\*(us\_price - l.us\_price)/l.us\_price

twoway (scatter stateReturns l.stateReturns) (lfit stateReturns l.stateReturns)

\*24. Split the sample into two sets: 2000-2012 and 2013-2022. Consider the first 12 years are your training set and the next 10 are the testing data. Test if returns over the prior year predict returns this year using the training set (years 2000-2012)

reg stateReturns l.stateReturns if year < 2013

\*25. Predict returns using the lagged returns

predict predictedStateReturns

\*26. Estimate the residual of the prediction

predict residualStateReturns, residuals

\*27. Repeat steps 6-8 but including both returns from last year and lagged unemployment rates as x-variables

reg stateReturns l.stateReturns l.unemployment

predict stateReturns

predict stateReturns, residual

\*28. Estimate the MSE of the two forecasts by analyzing the years 2002-2012 (the training years)

summarize residual\* if year < 2013, detail

\*29. Estimate the MSE of the two forecasts by analyzing the years 2013-2022 (the testing years)

\* Do unemployment rates predict returns beyond prior returns?

summarize residual\* if year >= 2013, detail

\*30. Z-score unemployment rates

\* Don't worry about Z-scoring returns

egen mean = mean(unemployment)

egen sd = sd(unemployment)

gen zscore\_unemployment = (unemployment-mean)/sd

\*31. Run LASSO on across 2000-2022 with return as the y-variable, and returns from the prior year and mortgage rates from the prior year as x-variables

\*To do this, you will need to create new variables for lagged variables

gen laggedState = l.stateReturns

gen laggedMortgage = l.mortgage\_rate

lasso linear stateReturns laggedState laggedMortgage

\*32. Estimate the coefficients from the lasso estimation using the command:

\*lassocoef, display(coef, postselection)

\* How many coefficients are there. Why?

lassocoef, display(coef, postselection)

/\*--------------------------

| active

---------------+----------

laggedState | .8477519

laggedMortgage | -2.134039

\_cons | 11.57024

--------------------------

\*/

\*There are 3 coefficients. Lagged State Returns and Lagged Mortgage rates are valuable together...it's the best model.

/\*Class notes:

time series dont go along with machine learning...

time series you're looking at different time periods... vs a holistic view

\*/