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

\* 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

cd"C:\Users\haniu\OneDrive\Desktop\Deepa\Deepa\Finance Core B\Business\Homework 4"

import excel "StateHousePriceData.xls", firstrow clear

\* (8 vars, 1,196 obs)

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

keep if stateid==0

\* (1,173 observations deleted)

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

tsset year, yearly

\* Time variable: year, 2000 to 2022

\* Delta: 1 year

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

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

\* (1 missing value generated)

\*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

regress return l.return

twoway (scatter return l.return) (lfit return l.return)

\*

\* Source | SS df MS Number of obs = 21

\*-------------+---------------------------------- F(1, 19) = 24.76

\* Model | 575.480964 1 575.480964 Prob > F = 0.0001

\* Residual | 441.545922 19 23.2392591 R-squared = 0.5658

\*-------------+---------------------------------- Adj R-squared = 0.5430

\* Total | 1017.02689 20 50.8513443 Root MSE = 4.8207

\*

\*------------------------------------------------------------------------------

\* return | Coefficient Std. err. t P>|t| [95% conf. interval]

\*-------------+----------------------------------------------------------------

\* return |

\* L1. | .8540549 .1716254 4.98 0.000 .4948389 1.213271

\* |

\* \_cons | 1.193151 1.267714 0.94 0.358 -1.460205 3.846508

\*------------------------------------------------------------------------------

\* The coefficient is 0.85 (when you have very high t value that means it has very strong co-relation)

\*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

\* Source | SS df MS Number of obs = 11

\*-------------+---------------------------------- F(1, 9) = 14.79

\* Model | 352.790112 1 352.790112 Prob > F = 0.0039

\* Residual | 214.717271 9 23.8574746 R-squared = 0.6216

\*-------------+---------------------------------- Adj R-squared = 0.5796

\* Total | 567.507383 10 56.7507383 Root MSE = 4.8844

\*

\*------------------------------------------------------------------------------

\* return | Coefficient Std. err. t P>|t| [95% conf. interval]

\*-------------+----------------------------------------------------------------

\* return |

\* L1. | .7775915 .2022114 3.85 0.004 .3201576 1.235025

\* |

\* \_cons | -.2195597 1.559642 -0.14 0.891 -3.747714 3.308595

\*------------------------------------------------------------------------------

\*you can make two sets as well training sets testing sets

\*7. Predict returns using the lagged returns

predict predict\_return1

\*(option xb assumed; fitted values)

\*(2 missing values generated)

\*8. Estimate the residual of the prediction

predict residual\_return1, res

\*(2 missing values generated)

\*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

\* Source | SS df MS Number of obs = 11

\*-------------+---------------------------------- F(2, 8) = 7.94

\* Model | 377.461337 2 188.730668 Prob > F = 0.0126

\* Residual | 190.046046 8 23.7557558 R-squared = 0.6651

\*-------------+---------------------------------- Adj R-squared = 0.5814

\* Total | 567.507383 10 56.7507383 Root MSE = 4.874

\*

\*-------------------------------------------------------------------------------

\* return | Coefficient Std. err. t P>|t| [95% conf. interval]

\*--------------+----------------------------------------------------------------

\* return |

\* L1. | .9428227 .2588501 3.64 0.007 .3459133 1.539732

\* |

\*mortgage\_rate |

\* L1. | -2.54494 2.497277 -1.02 0.338 -8.30367 3.21379

\* |

\* \_cons | 14.17013 14.2057 1.00 0.348 -18.58827 46.92854

\*-------------------------------------------------------------------------------

\*(option xb assumed; fitted values)

\*(2 missing values generated)

\*(2 missing values generated)

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

sum residual\* if year<2013, detail

\* Residuals

\*-------------------------------------------------------------

\* Percentiles Smallest

\* 1% -6.764799 -6.764799

\* 5% -6.764799 -6.373153

\*10% -6.373153 -4.66231 Obs 11

\*25% -4.66231 -4.190833 Sum of wgt. 11

\*

\*50% 1.201838 Mean 2.71e-08

\* Largest Std. dev. 4.633759

\*75% 3.871366 3.541477

\*90% 4.833732 3.871366 Variance 21.47173

\*95% 5.461523 4.833732 Skewness -.364975

\*99% 5.461523 5.461523 Kurtosis 1.524217

\*

\* Residuals

\*-------------------------------------------------------------

\* Percentiles Smallest

\* 1% -6.244652 -6.244652

\* 5% -6.244652 -6.011444

\*10% -6.011444 -5.144586 Obs 11

\*25% -5.144586 -2.64919 Sum of wgt. 11

\*

\*50% 1.846508 Mean -2.17e-08

\* Largest Std. dev. 4.359427

\*75% 3.714037 3.330096

\*90% 3.783609 3.714037 Variance 19.0046

\*95% 5.526742 3.783609 Skewness -.3522863

\*99% 5.526742 5.526742 Kurtosis 1.551986

\*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

\* Residuals

\*-------------------------------------------------------------

\* Percentiles Smallest

\* 1% -.6680025 -.6680025

\* 5% -.6680025 -.375581

\*10% -.5217918 .8608919 Obs 10

\*25% .8608919 2.022337 Sum of wgt. 10

\*

\*50% 2.066846 Mean 3.628574

\* Largest Std. dev. 4.248841

\*75% 7.693228 2.234127

\*90% 10.19252 7.693228 Variance 18.05265

\*95% 12.22031 8.164739 Skewness .9398687

\*99% 12.22031 12.22031 Kurtosis 2.585197

\* Residuals

\*-------------------------------------------------------------

\* Percentiles Smallest

\* 1% -5.582418 -5.582418

\* 5% -5.582418 -4.306479

\*10% -4.944448 -4.255045 Obs 10

\*25% -4.255045 -3.737934 Sum of wgt. 10

\*

\*50% -3.301047 Mean -2.097296

\* Largest Std. dev. 3.407782

\*75% -1.212717 -3.153123

\*90% 3.938425 -1.212717 Variance 11.61298

\*95% 5.053646 2.823203 Skewness 1.223032

\*99% 5.053646 5.053646 Kurtosis 3.119256

\* Mortgage rates doesn't predict returns beyond prior returns

\*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

\*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\_rate

\* (2 missing values generated)

\* (1 missing value generated)

\*Lasso linear model No. of obs = 21

\* No. of covariates = 2

\*Selection: Cross-validation No. of CV folds = 10

\*

\*--------------------------------------------------------------------------

\* | No. of Out-of- CV mean

\* | nonzero sample prediction

\* ID | Description lambda coef. R-squared error

\*---------+----------------------------------------------------------------

\* 1 | first lambda 5.234869 0 -0.0609 51.37697

\* 56 | lambda before .0313822 2 0.6158 18.60509

\* \* 57 | selected lambda .0285943 2 0.6158 18.60437

\*--------------------------------------------------------------------------

\* lambda selected by cross-validation.

\*Note: Minimum of CV function not found; lambda selected based on stop()

stopping criterion.

\*14. 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

\*----------------+----------

\* l1return | .8477519

\*l1mortgage\_rate | -2.898341

\* \_cons | .8202486

\*---------------------------

\*Legend:

\* b - base level

\* e - empty cell

\* o - omitted

\* 2 coefficient as we need both of them to predict the model

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

predict predict\_lasso

\*(options xb penalized assumed; linear prediction with penalized coefficients)

\*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

\*Blue is the actual returns and the red ones are our prediction and its been drwan with aline to connect the dots

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

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

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

\*17. Reopen the excel file StateHousePriceData

clear

cd"C:\Users\haniu\OneDrive\Desktop\Deepa\Deepa\Finance Core B\Business\Homework 4"

import excel "StateHousePriceData2.xls", firstrow clear

\* (8 vars, 1,196 obs)

\*other way to do it

\*clear

\*import excel "StateHousePriceData.xls", firstrow clear

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

drop if stateid==0

\* (23 observations deleted)

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

xtset stateid year, yearly

\* Panel variable: stateid (strongly balanced)

\* Time variable: year, 2000 to 2022

\* Delta: 1 year

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

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

\* (1,126 missing values generated)

\*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 populationchange\_return = 100\*(adult\_population -l22.adult\_population)/l22.adult\_population

twoway (scatter return22 populationchange\_return) (lfit return22 populationchange\_return)

\* (1,122 missing values generated)

\*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 return\_state = 100\*(price - l.price)/l.price

twoway (scatter return\_state l.return\_state) (lfit return\_state l.return\_state)

\* (69 missing values generated)

\*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 return\_state l.return\_state if year<2013

\* Source | SS df MS Number of obs = 543

\*-------------+---------------------------------- F(1, 541) = 708.06

\* Model | 19028.7929 1 19028.7929 Prob > F = 0.0000

\* Residual | 14539.1497 541 26.8745835 R-squared = 0.5669

\*-------------+---------------------------------- Adj R-squared = 0.5661

\* Total | 33567.9425 542 61.9334733 Root MSE = 5.1841

\*

\*------------------------------------------------------------------------------

\*return\_state | Coefficient Std. err. t P>|t| [95% conf. interval]

\*-------------+----------------------------------------------------------------

\*return\_state |

\* L1. | .743705 .027949 26.61 0.000 .6888031 .7986068

\* |

\* \_cons | -.0190228 .2355221 -0.08 0.936 -.4816727 .4436271

\*------------------------------------------------------------------------------

\*25. Predict returns using the lagged returns

predict predict\_return\_state

\*(option xb assumed; fitted values)

\*(120 missing values generated)

\*26. Estimate the residual of the prediction

predict residual\_return\_state, res

\*(120 missing values generated)

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

reg return\_state l.return\_state l.unemployment if year<2013

predict predictreturn\_state2

predict residual\_return\_state2, res

\* Source | SS df MS Number of obs = 543

\*-------------+---------------------------------- F(2, 540) = 370.19

\* Model | 19410.7535 2 9705.37674 Prob > F = 0.0000

\* Residual | 14157.1891 540 26.2170168 R-squared = 0.5783

\*-------------+---------------------------------- Adj R-squared = 0.5767

\* Total | 33567.9425 542 61.9334733 Root MSE = 5.1203

\*

\*------------------------------------------------------------------------------

\*return\_state | Coefficient Std. err. t P>|t| [95% conf. interval]

\*-------------+----------------------------------------------------------------

\*return\_state |

\* L1. | .8166221 .0335704 24.33 0.000 .7506775 .8825668

\* |

\*unemployment |

\* L1. | .4742257 .1242417 3.82 0.000 .2301694 .7182819

\* |

\* \_cons | -3.062249 .8305332 -3.69 0.000 -4.693721 -1.430777

\*------------------------------------------------------------------------------

\*(option xb assumed; fitted values)

\*(120 missing values generated)

\*(120 missing values generated)

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

sum residual\_return\_state\* if year <2013, detail

\* Residuals

\*-------------------------------------------------------------

\* Percentiles Smallest

\* 1% -14.7665 -25.44798

\* 5% -8.504498 -21.00803

\*10% -5.800522 -18.51707 Obs 543

\*25% -2.789996 -18.26004 Sum of wgt. 543

\*

\*50% .4800765 Mean 3.42e-09

\* Largest Std. dev. 5.179286

\*75% 2.677038 16.32603

\*90% 5.293305 17.6816 Variance 26.825

\*95% 7.960683 18.83642 Skewness -.2745454

\*99% 14.83299 22.33533 Kurtosis 5.970307

\*

\* Residuals

\*-------------------------------------------------------------

\* Percentiles Smallest

\* 1% -14.03712 -26.51

\* 5% -8.618551 -20.84751

\*10% -6.01502 -19.5795 Obs 543

\*25% -2.432163 -17.23622 Sum of wgt. 543

\*

\*50% .5470016 Mean 4.20e-09

\* Largest Std. dev. 5.1108

\*75% 2.595443 15.80663

\*90% 5.077143 18.18469 Variance 26.12028

\*95% 7.595941 18.74438 Skewness -.3064086

\*99% 13.99029 22.41683 Kurtosis 6.244941

\*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?

sum residual\_return\_state\* if year >=2013, detail

\* Residuals

\*-------------------------------------------------------------

\* Percentiles Smallest

\* 1% -4.788319 -10.66153

\* 5% -1.763824 -7.804771

\*10% -.6008727 -5.370143 Obs 510

\*25% .6202862 -4.840329 Sum of wgt. 510

\*

\*50% 1.882926 Mean 3.181752

\* Largest Std. dev. 4.718055

\*75% 4.16872 21.19082

\*90% 10.3367 23.10831 Variance 22.26004

\*95% 13.19605 26.76654 Skewness 1.653449

\*99% 19.38373 27.05459 Kurtosis 6.938402

\* Residuals

\*-------------------------------------------------------------

\* Percentiles Smallest

\* 1% -5.255073 -11.89072

\* 5% -2.018459 -9.255178

\*10% -.3842926 -7.492087 Obs 510

\*25% .9804087 -6.395482 Sum of wgt. 510

\*

\*50% 2.309397 Mean 3.193749

\* Largest Std. dev. 4.284937

\*75% 4.120831 19.3112

\*90% 9.144993 22.7787 Variance 18.36068

\*95% 11.33434 24.46033 Skewness 1.405425

\*99% 17.26897 25.90124 Kurtosis 7.553117

\* Unemployment rates doesn't predict returns beyond prior returns

\*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 unemployment from the prior year as x-variables

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

gen lreturn\_state = l.return\_state

gen lunemployment = l.unemployment

lasso linear return\_state lreturn\_state lunemployment

\* (120 missing values generated)

\* (51 missing values generated)

\*Lasso linear model No. of obs = 1,053

\* No. of covariates = 2

\*Selection: Cross-validation No. of CV folds = 10

\*

\*--------------------------------------------------------------------------

\* | No. of Out-of- CV mean

\* | nonzero sample prediction

\* ID | Description lambda coef. R-squared error

\*---------+----------------------------------------------------------------

\* 1 | first lambda 5.248342 0 0.0042 54.36391

\* 60 | lambda before .0216862 2 0.5465 24.75841

\* \* 61 | selected lambda .0197597 2 0.5465 24.75814

\*--------------------------------------------------------------------------

\* lambda selected by cross-validation.

\*Note: Minimum of CV function not found; lambda selected based on stop()

\* stopping criterion.

\*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

\*--------------+----------

\*lreturn\_state | .8825664

\*lunemployment | .7750828

\* \_cons | -3.519545

\*-------------------------

\*Legend:

\* b - base level

\* e - empty cell

\* o - omitted

\* 2 coefficient as we need both of them to predict the model