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

\* Homework 3

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

\* 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 3 on Canvas

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

\* Topic 1: Over-Fitting

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

\*1. Import the SalaryData excel file into Stata

clear

import excel "/Users/henryvelasquez/Documents/MBA/MBA S3/2nd 7/Machine Learning in Finance K579/HW3/SalaryData.xlsx", sheet("Sheet1") firstrow case(lower)

\* The dataset includes three variables:

\* age: age/experience of the worker

\* salary: salary of the worker

\*TestData: Identifies the testing set as a 1, training set at a 0

\*2. Run a lin`ear regression with salary as the y-variable and age as the x-variable ONLY for the training set (meaning the TestData is 0)

\* Why do you exclude observations in the testing set?

reg salary age if testdata == 0

/\* You exclude observations because we want to look at training data only. We'll be comparing RMSE. \*/

\*3. Using the regression above, predict the salary for each individual in both the training and testing set.

predict predictedSalary

li age salary testdata predictedSalary

/\*

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

| age salary testdata predic~y |

|------------------------------------|

1. | 25 135000 0 146843 |

2. | 55 260000 0 261662.2 |

3. | 27 105000 0 154497.7 |

4. | 35 220000 0 185116.1 |

5. | 60 240000 0 280798.7 |

|------------------------------------|

6. | 65 265000 0 299935.3 |

7. | 45 270000 0 223389.2 |

8. | 40 300000 0 204252.6 |

9. | 50 265000 0 242525.7 |

10. | 30 105000 0 165979.6 |

|------------------------------------|

11. | 30 166000 1 165979.6 |

12. | 26 78000 1 150670.3 |

13. | 58 310000 1 273144.1 |

14. | 29 100000 1 162152.3 |

15. | 40 260000 1 204252.6 |

|------------------------------------|

16. | 27 150000 1 154497.7 |

17. | 33 140000 1 177461.5 |

18. | 61 220000 1 284626 |

19. | 27 86000 1 154497.7 |

20. | 48 276000 1 234871.1 |

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

\*/

\*4. Estimate the residuals for both the training and testing set

predict residualSalary, residuals

li age salary testdata predictedSalary residualSalary

/\*

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

| age salary testdata predic~y residua~y |

|------------------------------------------------|

1. | 25 135000 0 146843 -11843.04 |

2. | 55 260000 0 261662.2 -1662.201 |

3. | 27 105000 0 154497.7 -49497.66 |

4. | 35 220000 0 185116.1 34883.9 |

5. | 60 240000 0 280798.7 -40798.73 |

|------------------------------------------------|

6. | 65 265000 0 299935.3 -34935.25 |

7. | 45 270000 0 223389.2 46610.85 |

8. | 40 300000 0 204252.6 95747.38 |

9. | 50 265000 0 242525.7 22474.32 |

10. | 30 105000 0 165979.6 -60979.57 |

|------------------------------------------------|

11. | 30 166000 1 165979.6 20.42867 |

12. | 26 78000 1 150670.3 -72670.35 |

13. | 58 310000 1 273144.1 36855.88 |

14. | 29 100000 1 162152.3 -62152.27 |

15. | 40 260000 1 204252.6 55747.38 |

|------------------------------------------------|

16. | 27 150000 1 154497.7 -4497.656 |

17. | 33 140000 1 177461.5 -37461.49 |

18. | 61 220000 1 284626 -64626.03 |

19. | 27 86000 1 154497.7 -68497.66 |

20. | 48 276000 1 234871.1 41128.93 |

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

\*/

\*5. Estimate the MSE separately for the training set and testing set. Which is larger. Why?

summarize residualSalary if testdata==0

summarize residualSalary if testdata==1

/\*

TestingData is larger because we used the training data only to estimate the coefficients.

\*/

\*6. Create a scatter plot of the Training data and graph the linear prediction on top. Save the graph using: graph save

twoway (scatter salary age if testdata==0) (connected predictedSalary age if testdata==0, sort)

graph save polynomialTest, replace

\*7. Create a scatter plot of the Testing data and graph the linear prediction on top. Save the graph using: graph save

twoway (scatter salary age if testdata==1) (connected predictedSalary age if testdata==1, sort)

graph save polynomialTest2, replace

\*8. Create new variables:

gen age2 = age^2

gen age3 = age^3

gen age4 = age^4

gen age5 = age^5

\*9. Repeat steps 2-7 for each polynomial expansion. This means considering four new regressions:

\* second-order (or quadratic) polynomial: salary = a + b(age) + c(age^2)

reg salary age1 age2 if testdata == 0

predict predictedSalary2

predict residualSalary2, residuals

summarize residualSalary2 if testdata==0

summarize residualSalary2 if testdata==1

twoway (scatter salary age if testdata==0) (connected predictedSalary2 age if testdata==0, sort)

graph save polynomialTest2\_0, replace

twoway (scatter salary age if testdata==1) (connected predictedSalary2 age if testdata==1, sort)

graph save polynomialTest2\_1, replace

\* third-order polynomial: salary = a + b(age) + c(age^2)

reg salary age1 age2 age3 if testdata == 0

predict predictedSalary3

predict residualSalary3, residuals

summarize residualSalary3 if testdata==0

summarize residualSalary3 if testdata==1

twoway (scatter salary age if testdata==0) (connected predictedSalary3 age if testdata==0, sort)

graph save polynomialTest3\_0, replace

twoway (scatter salary age if testdata==1) (connected predictedSalary3 age if testdata==1, sort)

graph save polynomialTest2\_1, replace

\* fourth-order polynomial: salary = a + b(age) + c(age^3) + d(age^4)

reg salary age1 age2 age3 age4 if testdata == 0

predict predictedSalary4

predict residualSalary4, residuals

summarize residualSalary4 if testdata==0

summarize residualSalary4 if testdata==1

twoway (scatter salary age if testdata==0) (connected predictedSalary4 age if testdata==0, sort)

graph save polynomialTest4\_0, replace

twoway (scatter salary age if testdata==1) (connected predictedSalary4 age if testdata==1, sort)

graph save polynomialTest4\_1, replace

\* fifth-order polynomial: salary = a + b(age) + c(age^3) + d(age^4) + e(age^5)

reg salary age age2 age3 age4 age5 if testdata == 0

predict predictedSalary5

predict residualSalary5, residuals

summarize residualSalary5 if testdata==0

summarize residualSalary5 if testdata==1

twoway (scatter salary age if testdata==0) (connected predictedSalary5 age if testdata==0, sort)

graph save polynomialTest5\_0, replace

twoway (scatter salary age if testdata==1) (connected predictedSalary5 age if testdata==1, sort)

graph save polynomialTest5\_1, replace

\*What happens to the MSE as we increase the polynomial of the testing data?

\*What happens to the MSE as we increase the polynoial of the training data?

/\*testing data, MSE went down for the second order polynomial, and then increased after that.

training data, the MSE decreased after the second order polynomial, then increased for the third

order polynomial, and then decreased again during the fourth and fifth order polynomial.

\*/

\*10. Combine all graphs for the training data into a single graph using the code: graph combine

\* In a separate figure, Combine all graphs for the testing data into a single graph using the code

graph combine "polynomialTest\_0" "polynomialTest2\_0" "polynomialTest3\_0" "polynomialTest4\_0" "polynomialTest5\_0"

graph combine "polynomialTest\_1" "polynomialTest2\_1" "polynomialTest3\_1" "polynomialTest4\_1" "polynomialTest5\_1"

\*\*\*USE "graph combine" command

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

\* Topic 1: Regularization

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

\*laso linear salary zscore\_age\*

\*lassocoef, display(coef,postselection)

\*11. Z-score scale all five age variables:

\* New Value = (Old Value -Mean)/(Standard Deviation)

egen meanAge1 = mean(age)

egen meanAge2 = mean(age2)

egen meanAge3 = mean(age3)

egen meanAge4 = mean(age4)

egen meanAge5 = mean(age5)

egen sd1 = sd(age)

egen sd2 = sd(age2)

egen sd3 = sd(age3)

egen sd4 = sd(age4)

egen sd5 = sd(age5)

gen zscored1 = (age-meanAge1)/sd1

gen zscored2 = (age2-meanAge2)/sd2

gen zscored3 = (age3-meanAge3)/sd3

gen zscored4 = (age4-meanAge4)/sd4

gen zscored5 = (age5-meanAge5)/sd5

\*12. Estimate the fifth-order polynomial using the Z-scored variables using only the training data

reg salary zscored1 zscored2 zscored3 zscored4 zscored5 if testdata==0

\*13. Predict the Z-score scaled salary from the regression estimates

predict predict\_salary5\_zscore

\*14. Estimate coefficients under a Lasso Estimation for only the training data using the command:

\* lasso linear y-variable x-variables if Test==0

lasso linear salary zscored1 zscored2 zscored3 zscored4 zscored5 if testdata==0

/\*

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

| No. of Out-of- CV mean

| nonzero sample prediction

ID | Description lambda coef. R-squared error

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

1 | first lambda 51228.73 0 -0.2227 5.93e+09

30 | lambda before 3449.82 2 0.4829 2.51e+09

\* 31 | selected lambda 3143.348 2 0.4829 2.51e+09

32 | lambda after 2864.101 2 0.4803 2.52e+09

34 | last lambda 2377.827 2 0.4695 2.57e+09

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

\*/

\*15. 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

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

zscored1 | 127522.2

zscored5 | -75628.39

\_cons | 204382

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

There are three coefficients. All of the others increase the MSE and thus these are the coefficients that produce the bestfit line.

\*/

\*16. Predict the salary estimate from the lasso estimation using predict

predict lassoPrediction

\*17. Using only the training data, create one graph with (1) a scatter plot of the z-score scaled salary and age, (2) the 5th-order polynomial estimates, and (3) the lasso estimates

twoway (scatter salary zscored1 if test==0) (connected lassoPrediction zscored1 if test==0, sort) (connected predict\_salary5\_zscore zscored1 if test==0, sort)

graph save Lasso\_Test0, replace

\*18. Using only the testing data, create one graph with (1) a scatter plot of the z-score scaled salary and age, (2) the 5th-order polynomial estimates, and (3) the lasso estimates

twoway (scatter salary zscored1 if test==1) (connected lassoPrediction zscored1 if test==1, sort) (connected predict\_salary5\_zscore zscored1 if test==1, sort)

graph save Lasso\_Test1, replace

/\* 19. combine the graphs \*/

graph combine "Lasso\_Test0" "Lasso\_Test1"