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

\* Topic 1: Over-Fitting

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

\*1. Import the SalaryData excel file into Stata

\* 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

clear

cd"C:\Users\haniu\OneDrive\Desktop\Deepa\Deepa\Finance Core B\Business\Homework 3"

import excel "SalaryData.xlsx", firstrow clear

\* (3 vars, 20 obs)

\*2. Run a linear 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?

regress salary age if TestData==0

\* Source | SS df MS Number of obs = 10

\*-------------+---------------------------------- F(1, 8) = 9.43

\* Model | 2.6244e+10 1 2.6244e+10 Prob > F = 0.0153

\* Residual | 2.2259e+10 8 2.7823e+09 R-squared = 0.5411

\*-------------+---------------------------------- Adj R-squared = 0.4837

\* Total | 4.8503e+10 9 5.3892e+09 Root MSE = 52748

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

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

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

\* age | 3827.305 1246.189 3.07 0.015 953.5872 6701.023

\* \_cons | 51160.42 56360.29 0.91 0.391 -78806.64 181127.5

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

\*As for regression we need a training set (i.e., to run the data) and the validation data set (i.e., testing data), we exclude the testing data so as to train and run the regression model.

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

predict predict\_salary1

\*(option xb assumed; fitted values)

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

predict residual\_salary1, res

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

summarize residual\_salary1 if TestData==0, detail

summarize residual\_salary1 if TestData==1, detail

\*

\* Residuals

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

\* Percentiles Smallest

\* 1% -60979.57 -60979.57

\* 5% -60979.57 -49497.66

\*10% -55238.61 -40798.73 Obs 10

\*25% -40798.73 -34935.25 Sum of wgt. 10

\*

\*50% -6752.623 Mean -.0000244

\* Largest Std. dev. 49731.13

\*75% 34883.9 22474.32

\*90% 71179.11 34883.9 Variance 2.47e+09

\*95% 95747.38 46610.85 Skewness .5452581

\*99% 95747.38 95747.38 Kurtosis 2.313258

\* Residuals

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

\* Percentiles Smallest

\* 1% -72670.35 -72670.35

\* 5% -72670.35 -68497.66

\*10% -70584 -64626.03 Obs 10

\*25% -64626.03 -62152.27 Sum of wgt. 10

\*50% -20979.57 Mean -17615.28

\* Largest Std. dev. 49989.62

\*75% 36855.88 20.42867

\*90% 48438.15 36855.88 Variance 2.50e+09

\*95% 55747.38 41128.93 Skewness .2515126

\*99% 55747.38 55747.38 Kurtosis 1.476178

\* The MSE for the training data set will always be higher as they match with the original data set

\*6. 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==0) (connected predict\_salary1 age if TestData==0, sort)

graph save Polynomial1\_TestData0, replace

\*7. 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==1) (connected predict\_salary1 age if TestData==1, sort)

graph save Polynomial1\_TestData1, replace

\*8. Create new variables:

\* age2 = age^2

\*age3 = age^3

\*age4 = age^4

\*age5 = age^5

generate age1=age

generate age2=age^2

generate age3=age^3

generate age4=age^4

generate 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)

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

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

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

\*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?

\*second-order (or quadratic) polynomial

regress salary age age2 if TestData==0

predict predict\_salary2

predict residual\_salary2, res

summarize residual\_salary2 if TestData==0, detail

summarize residual\_salary2 if TestData==1, detail

twoway (scatter salary age if TestData==0) (connected predict\_salary2 age if TestData==0, sort)

graph save Polynomial2\_TestData0, replace

twoway (scatter salary age if TestData==1) (connected predict\_salary2 age if TestData==1, sort)

graph save Polynomial2\_TestData1, replace

\*third-order polynomial

regress salary age age2 age3 if TestData==0

predict predict\_salary3

predict residual\_salary3, res

summarize residual\_salary3 if TestData==0, detail

summarize residual\_salary3 if TestData==1, detail

twoway (scatter salary age if TestData==0) (connected predict\_salary3 age if TestData==0, sort)

graph save Polynomial3\_TestData0, replace

twoway (scatter salary age if TestData==1) (connected predict\_salary3 age if TestData==1, sort)

graph save Polynomial3\_TestData1, replace

\*fourth-order polynomial

regress salary age age2 age3 age4 if TestData==0

predict predict\_salary4

predict residual\_salary4, res

summarize residual\_salary4 if TestData==0, detail

summarize residual\_salary4 if TestData==1, detail

twoway (scatter salary age if TestData==0) (connected predict\_salary4 age if TestData==0, sort)

graph save Polynomial4\_TestData0, replace

twoway (scatter salary age if TestData==1) (connected predict\_salary4 age if TestData==1, sort)

graph save Polynomial4\_TestData1, replace

\*fifth-order polynomial

regress salary age age2 age3 age4 age5 if TestData==0

predict predict\_salary5

predict residual\_salary5, res

summarize residual\_salary5 if TestData==0, detail

summarize residual\_salary5 if TestData==1, detail

twoway (scatter salary age if TestData==0) (connected predict\_salary5 age if TestData==0, sort)

graph save Polynomial5\_TestData0, replace

twoway (scatter salary age if TestData==1) (connected predict\_salary5 age if TestData==1, sort)

graph save Polynomial5\_TestData1, replace

\* TestData0 TestData1

\* 2.47E+09 2.50E+09

\* 1.08E+09 1.13E+09

\* 1.02E+09 1.27E+09

\* 4.76E+08 1.40E+09

\* 1.66E+08 1.50E+09

\*As we see for the training set that the SME is decreasing as we increase the polynomial order.

\*But for testing data set the SME decreases as we increase the polynomial order untill order three but post the dip SME increases as the polynomial order increases.

\*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 "Polynomial1\_TestData0" "Polynomial2\_TestData0" "Polynomial3\_TestData0" "Polynomial4\_TestData0" "Polynomial5\_TestData0"

graph combine "Polynomial1\_TestData1" "Polynomial2\_TestData1" "Polynomial3\_TestData1" "Polynomial4\_TestData1" "Polynomial5\_TestData1"

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

\* Topic 1: Regularization

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

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

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

egen Mean\_age1 = mean(age1)

egen Mean\_age2 = mean(age2)

egen Mean\_age3 = mean(age3)

egen Mean\_age4 = mean(age4)

egen Mean\_age5 = mean(age5)

egen sd1 = sd(age1)

egen sd2 = sd(age2)

egen sd3 = sd(age3)

egen sd4 = sd(age4)

egen sd5 = sd(age5)

gen zscore\_age1 = (age1 - Mean\_age1)/sd1

gen zscore\_age2 = (age2 - Mean\_age2)/sd2

gen zscore\_age3 = (age3 - Mean\_age3)/sd3

gen zscore\_age4 = (age4 - Mean\_age4)/sd4

gen zscore\_age5 = (age5 - Mean\_age5)/sd5

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

reg salary zscore\_age1 zscore\_age2 zscore\_age3 zscore\_age4 zscore\_age5 if TestData==0

\*

\* Source | SS df MS Number of obs = 10

\*-------------+---------------------------------- F(5, 4) = 25.10

\* Model | 4.7004e+10 5 9.4008e+09 Prob > F = 0.0040

\* Residual | 1.4983e+09 4 374568956 R-squared = 0.9691

\*-------------+---------------------------------- Adj R-squared = 0.9305

\* Total | 4.8503e+10 9 5.3892e+09 Root MSE = 19354

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

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

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

\* zscore\_age1 | -3.15e+07 9604181 -3.28 0.031 -5.81e+07 -4810946

\* zscore\_age2 | 1.28e+08 4.04e+07 3.18 0.034 1.62e+07 2.41e+08

\* zscore\_age3 | -2.00e+08 6.58e+07 -3.04 0.038 -3.83e+08 -1.74e+07

\* zscore\_age4 | 1.41e+08 4.88e+07 2.89 0.045 5368907 2.76e+08

\* zscore\_age5 | -3.77e+07 1.38e+07 -2.73 0.053 -7.61e+07 663525.9

\* \_cons | 196823.7 6477.354 30.39 0.000 178839.7 214807.8

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

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

predict predict\_salary5\_zscore

\*(option xb assumed; fitted values)

\*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 zscore\_age1 zscore\_age2 zscore\_age3 zscore\_age4 zscore\_age5 if Test==0

\*Lasso linear model No. of obs = 10

\* No. of covariates = 5

\*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 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

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

\* lambda selected by cross-validation.

\*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

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

\* zscore\_age1 | 127522.2

\* zscore\_age5 | -75628.39

\* \_cons | 204382

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

\*Legend:

\* b - base level

\* e - empty cell

\* o - omitted

\*There are only 2 coefficients, as Lasso tends to omit covariates that have small coefficients in favor of irrelevant ones (variables not in the true model) that are correlated with the error term.

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

predict predict\_lasso

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

\*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 zscore\_age1 if TestData==0) (connected predict\_lasso zscore\_age1 if TestData==0, sort) (connected predict\_salary5\_zscore zscore\_age1 if TestData==0, sort)

graph save LASSO\_Training, 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 zscore\_age1 if TestData==1) (connected predict\_lasso zscore\_age1 if TestData==1, sort) (connected predict\_salary5\_zscore zscore\_age1 if TestData==1, sort)

graph save LASSO\_Testing, replace

\*19. Combine both graphs above using the code: graph combine

graph combine "LASSO\_Testing" "LASSO\_Training"