**Features Selection using Ridge Regularization, Recursive Feature Elimination (RFE) with LASSO, and Sequential Feature Selection (SFS) with LASSO. Statistical modeling of Current Population Survey (CPS) data to predict wages**

**Background:**

The Current Population Survey (CPS) is used to supplement census information between census years. These data consist of a random sample of 534 persons from the 1985 CPS, with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience, occupational status, region of residence and union membership.

**Data cleaning**: No duplicated records nor missing values, which is rare with survey data.

**Data splitting**: Dataset was split into a training (80%) and test (20%) set.

**Transformations:** The target variable, WAGE, was log transformed (using scikitlearn *TransformedTargetRegressor)* since it was not normally distributed. Categorical variables *['SOUTH','SEX','UNION','RACE','OCCUPATION', 'SECTOR', 'MARR']* were encoded using OneHotEncoder method.

**Outliers**: Outlying data points were retained in the dataset. Not enough time to assess the impact to the model by removing one by one while observing changes to the coefficients, R-squared and MSE.

**Multicollinearity:** Ridge and LASSO penalty (L2 and L1 regularizations respectively) handle Multicollinearity by shrinking the coefficients of correlated variables. Best alpha value (L2 penalty) used was 65.79.

**Dimension reduction/Feature importance:** Features importance, sorted by level of importance, by Permutation Importance method are shown below:

Permutation Importance:

feature importance\_mean importance\_std

0 EDUCATION 0.203168 0.037115

15 OCCUPATION\_Professional 0.053848 0.012644

13 OCCUPATION\_Management 0.042649 0.008747

17 OCCUPATION\_Service 0.036241 0.019059

5 SEX\_female 0.026668 0.015172

6 SEX\_male 0.026668 0.015172

2 AGE 0.021567 0.019482

14 OCCUPATION\_Other 0.013670 0.004482

16 OCCUPATION\_Sales 0.013595 0.006493

4 SOUTH\_yes 0.006162 0.004764

3 SOUTH\_no 0.006162 0.004764

11 RACE\_White 0.004118 0.002656

9 RACE\_Hispanic 0.003817 0.002804

10 RACE\_Other 0.001071 0.000885

12 OCCUPATION\_Clerical 0.000296 0.000118

18 SECTOR\_Construction -0.000064 0.000331

19 SECTOR\_Manufacturing -0.000290 0.003103

20 SECTOR\_Other -0.000698 0.003677

22 MARR\_Unmarried -0.000946 0.005574

21 MARR\_Married -0.000946 0.005574

1 EXPERIENCE -0.001227 0.012640

8 UNION\_not\_member -0.017452 0.015305

7 UNION\_member -0.017452 0.015305

Features selection/Best model by Ridge Regularization, Recursive Feature Elimination (RFE) with LASSO, and Sequential Feature Selection (SFS) with LASSO are shown below:

**Ridge Regularization** (Reduced model: Keeping features with p-value<0.05 only)

MSE: 21.892670999554635 R-squared: 0.31324677623391506

Significant Parameters with P-values <= 0.05:

EDUCATION: Coefficient=0.10096814998609079, P-value=3.324570362411413e-10

SOUTH\_no: Coefficient=0.014816429223958528, P-value=0.026043929449186577

SOUTH\_yes: Coefficient=-0.01460010776066053, P-value=0.026043929449186577

SEX\_female: Coefficient=-0.03943383741845604, P-value=0.014369215063869515

SEX\_male: Coefficient=0.041052703035547915, P-value=0.014369215063869515

OCCUPATION\_Management: Coefficient=0.05560326198328818, P-value=0.0003835817697092929

OCCUPATION\_Professional: Coefficient=0.05331775767678876, P-value=1.4623180376308908e-05

OCCUPATION\_Service: Coefficient=-0.054323879562551976, P-value=0.008964456303473556

F-test for the Model: F-test=5.581357467252827, p-value=0.30248602079490655

Model Equation for Significant Parameters:

WAGE = 8.031 + (0.101 \* EDUCATION) + (0.015 \* SOUTH\_no) - (0.015 \* SOUTH\_yes) - (0.039 \* SEX\_female) + (0.041 \* SEX\_male) + (0.056 \* OCCUPATION\_Management) + (0.053 \* OCCUPATION\_Professional) - (0.054 \* OCCUPATION\_Service)

**Recursive Feature Elimination (RFE) with LASSO**

RFE Parameter Estimates:

EDUCATION: Coef = 0.7849, p-value = 0.1024

AGE: Coef = 0.0795, p-value = 0.0000

UNION\_member: Coef = 1.6321, p-value = 0.0002

SECTOR\_Manufacturing: Coef = -0.1250, p-value = 0.0176

SECTOR\_Other: Coef = -1.8898, p-value = 0.9247

Intercept: -3.1247

R-squared: 0.20636366243242943 Adjusted R-squared: 0.19372614113358266

F-value: 16.329441316254222 F-test p-value: 2.5331856007938725e-14

Equation: WAGE = -3.1247 + (0.7849 \* EDUCATION) + (0.0795 \* AGE) + (1.6321 \* UNION\_member) + (-0.1250 \* SECTOR\_Manufacturing) + (-1.8898 \* SECTOR\_Other)

Mean Squared Error (MSE) on Validation Set: 19.3114

Mean Squared Error (MSE) on Test Set: 22.8655

**Sequential Feature Selection (SFS) with LASSO**

SFS Parameter Estimates:

EDUCATION: Coef = 0.7563, p-value = 0.0257

AGE: Coef = 0.0855, p-value = 0.0000

SEX\_male: Coef = 1.9997, p-value = 0.0001

UNION\_not\_member: Coef = -1.0331, p-value = 0.0001

MARR\_Unmarried: Coef = -0.5408, p-value = 0.1350

Intercept: -4.1466

R-squared: 0.22344203081253444 Adjusted R-squared: 0.21107645805477215

F-value: 18.069687121631635 F-test p-value: 9.347009181941542e-16

Equation: WAGE = -4.1466 + (0.7563 \* EDUCATION) + (0.0855 \* AGE) + (1.9997 \* SEX\_male) + (-1.0331 \* UNION\_not\_member) + (-0.5408 \* MARR\_Unmarried)

MSE on Validation Set: 18.2256 MSE on Test Set: 20.6308

**Comparison of Models:**

Method MSE R-squared F-Test Pvalue N-Features

0 Ridge (full model) 20.070148 0.370418 3.024860e-01 23

1 SFS 20.630757 0.223442 9.347009e-16 5

2 RFE 22.865513 0.206364 2.533186e-14 5

**Interpretation:**

**Best Model choice:** Sequential Feature Selection (SFS) with LASSO because of its being a parsimonious model and it is very easy to interpret. With 5 features it has an MSE value close to the MSE value of the Ridge full model.

Overall Model is significant (p-value for the F test <0.0001)

Adjusted R-squared= 0.21 meaning that 21% of the variance in wages in 1985 were explained by the variables in the model.

For every unit increase in education level, wages increased by 0.76 times while for every unit increase in age wages increased by 0.086. The wages were almost twice for males compared to females where as not belonging to a union decreased the wages by 1.03 times. The unmarried earned almost half the wages compared to the married.