**Homework 3**

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

There are 20 points in this assignment. The minimum increment is 0.5 point. Solve them and fill the answers in the blank space.

1. **Continue question 2 of homework assignment 2.** HW3\_Airfares\_Selected.csv contains the data with the selected predictors (no missing values but the predictors have not been standardized or coded). Allocate 20% of the data to the test partition. Choose VACATION\_No, SW\_No, SLOT\_Free, GATE\_Constrained as the redundant dummies to drop.

a. Use the remaining 80% of the data as the training partition to build two LASSO regression models with the following pre-specified penalty levels respectively: alpha=0.01 and alpha=0.1. Organize the estimated coefficients and the intercept of both models into the following tables respectively. (2 points)

|  |  |  |  |
| --- | --- | --- | --- |
| Pre-specified penalty level: alpha=0.01 | | | |
| Predictor | **Estimated coefficient** | **Predictor** | **Estimated coefficient** |
| VACATION\_Yes |  | E\_POP |  |
| SW\_Yes |  | SLOT\_Controlled |  |
| S\_INCOME |  | GATE\_Free |  |
| E\_INCOME |  | DISTANCE |  |
| S\_POP |  | Intercept |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Pre-specified penalty level: alpha=0.1 | | | |
| Predictor | **Estimated coefficient** | **Predictor** | **Estimated coefficient** |
| VACATION\_Yes |  | E\_POP |  |
| SW\_Yes |  | SLOT\_Controlled |  |
| S\_INCOME |  | GATE\_Free |  |
| E\_INCOME |  | DISTANCE |  |
| S\_POP |  | Intercept |  |

b. Do you see any zero estimated coefficient in either of the models in (a)? For each model, if you do not, please explain why there is no zero estimated coefficient. If you do, please explain why those estimated coefficients are zero. (2 points)

|  |  |  |
| --- | --- | --- |
| **Model** | **Yes/No** | **Why** |
| Model with alpha=0.01 |  |  |
| Model with alpha=0.1 |  |  |

c. Conduct the necessary analysis to rank the predictors based on their importance. If one predictor is the *n*-th most important one, write number *n* in its column Rank of Importance. (1 point)

|  |  |  |  |
| --- | --- | --- | --- |
| Predictor | Rank of importance | Predictor | Rank of importance |
| VACATION\_Yes |  | E\_POP |  |
| SW\_Yes |  | SLOT\_Controlled |  |
| S\_INCOME |  | GATE\_Free |  |
| E\_INCOME |  | DISTANCE |  |
| S\_POP |  |  | |

d. Find the numerical predictors among all the predictors and rank them based on how strongly they are correlated with the dependent variable (note that the correlation can be either positive or negative). For example, 1means the most strongly correlated and 2 means the second most strongly correlated. Compare the ranking based on the importance in (c) and the ranking based on the degree of correlation in (d). Are two rankings consistent with each other within the numerical predictors? (2 points)

|  |  |  |  |
| --- | --- | --- | --- |
| Numerical predictor | Rank of correlation | Numerical predictor | Rank of correlation |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

e. Use cross-validation to select the model based on predictive performance. Set the number of folds for cross-validation to 3. Organize the estimated coefficients and the intercept as well as the corresponding penalty level alpha of the final selected model in the following table. (2 points)

|  |  |  |  |
| --- | --- | --- | --- |
| Predictor | Estimated coefficient | Predictor | Estimated coefficient |
| VACATION\_Yes |  | E\_POP |  |
| SW\_Yes |  | SLOT\_Controlled |  |
| S\_INCOME |  | GATE\_Free |  |
| E\_INCOME |  | DISTANCE |  |
| S\_POP |  | Intercept |  |
| Penalty level alpha: | | | |

f. How is the penalty level of the final selected model in (e) determined? (1 point)

g. Fill the answers in the following table. Among all three models, which one has the lowest ASE over the test partition? Why does that model have the lowest ASE over the test partition? (2 points)

|  |  |
| --- | --- |
| Model | ASE over the test partition |
| Model in (e) |  |
| Model with alpha=0.01 in (a) |  |
| Model with alpha=0.1 in (a) |  |

h. Use the final selected model in (e) to predict the average fare on a route with the following characteristics. The prediction outcome should be in its unstandardized value. Show the calculation and explain your answer. (4 points) **(Need the class content on Oct. 14th)**

VACATION = No, SW = No, S\_INCOME = $28,760, E\_INCOME = $27,664, S\_POP = 4,557,004, E\_POP = 3,195,503, SLOT = Free, GATE = Free, DISTANCE = 1976 miles.

i. Southwest Airline announces that it will serve the same route. Based on the final selected model in (e), do you expect the average ticket fare of the major airline for that route (not counting Southwest) to increase or decrease? How much do you expect the average fare to increase or decrease? Show your calculations. (2 points) **(Need the class content on Oct. 14th)**

j. Now we change our way to code the categorical variables by choosing a different redundant dummy to drop. Specifically, instead of dropping SLOT\_Free, we drop SLOT\_Controlled. Fill the answers in the following table. (2 points) **(Need the class content on Oct. 14th)**

|  |  |
| --- | --- |
| **Questions** | **Yes/No** |
| Will it affect the estimated coefficients or the intercept of the final selected model? |  |
| Will it change the penalty level of the final selected model? |  |
| Will it affect the average squared error over the test partition? |  |
| Will it affect the predicted value for the route in (h)? |  |

Submit your Python code with the filename [DM2020] HW3\_YOURFULLNAME.ipynb