**Homework 3**

Name: Tyler DiNapoli-Chiappelli\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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).

a. Standardize data and code predictors

a. 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 | -0.592885 | E\_POP | 0.054358 |
| SW\_Yes | -0.632651 | SLOT\_Controlled | 0.106474 |
| S\_INCOME | 0.022521 | GATE\_Free | -0.279766 |
| E\_INCOME | 0.062799 | DISTANCE | 0.600097 |
| S\_POP | 0.033497 | Intercept | 0.539930 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Pre-specified penalty level: alpha=0.1** | | | |
| **Predictor** | **Estimated coefficient** | **Predictor** | **Estimated coefficient** |
| VACATION\_Yes | -0.142356 | E\_POP | 0.081628 |
| SW\_Yes | -0.334210 | SLOT\_Controlled | 0.000000 |
| S\_INCOME | 0.053697 | GATE\_Free | -0.000000 |
| E\_INCOME | 0.051722 | DISTANCE | 0.514107 |
| S\_POP | 0.017274 | Intercept | 0.137609 |

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 | No | A penalty weight of alpha = .01 is not enough to drive any coefficients to zero in the OLS+LASSO alpha calculation. All coefficients hold enough explanatory power, as measured on aggregate by the model’s MSE, to survive the imposed penalty of .01 |
| Model with alpha=0.1 | Yes | With a penalty level of alpha = .1, SLOT\_Controlled and GATE\_Free have been driven to zero (or at least so near zero that Python’s level of precision calls it zero). The tenfold increase in the penalty level forced the algorithm to shrink the coefficients so much their explanatory power fell to zero. |

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)

For clarity, please note that these rankings are derived from comparing the **absolute values** of the estimated coefficients after finding the optimal alpha using the K-Folds = 5 method.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | **Rank of importance** | **Predictor** | **Rank of importance** |
| VACATION\_Yes | 2 | E\_POP | 7 |
| SW\_Yes | 1 | SLOT\_Controlled | 5 |
| S\_INCOME | 9 | GATE\_Free | 4 |
| E\_INCOME | 6 | DISTANCE | 3 |
| S\_POP | 8 |  | |

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** |
| DISTANCE | 1 |
| E\_INCOME | 2 |
| E\_POP | 3 |
| S\_INCOME | 4 |
| S\_POP | 5 |

The rankings of the numerical predictors in the correlation matrix are decently consistent with the rankings of predictor importance as per the output of the K-Folds LASSO regression with optimal alpha. DISTANCE claims the top spot in both cases among the numerical variables, followed by E\_INCOME and E\_POP. However, the relationships shown in the correlation matrix and the estimated coefficients from the regression switch the importance of S\_POP and S\_INCOME.

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 | -0.614716 | E\_POP | 0.049143 |
| SW\_Yes | -0.641363 | SLOT\_Controlled | 0.132911 |
| S\_INCOME | 0.019475 | GATE\_Free | -0.309793 |
| E\_INCOME | 0.062388 | DISTANCE | 0.605753 |
| S\_POP | 0.030709 | Intercept | 0.564667 |
| Penalty level alpha: .004 | | | |

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

By calling the LassoCV function, we are asking Python to run the regression (according to the Python user documentation) 100 times with equally spaced alphas along the regularization path. The model that corresponds with the lowest mean-squared error will be selected for the user. Therefore, the penalty level associated with this model is the alpha present in the final selected model.

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) | 0.3266243850921795 |
| Model with alpha=0.01 in (a) | 0.32721956500281546 |
| Model with alpha=0.1 in (a) | 0.4311941761059571 |

The model with the lowest ASE over the test partition is the model in part (e), the 3 k-folds cross-validation model. There are two likely reasons why this model performed the best out of the three. First, for this model we let Python find the optimal alpha for us algorithmically, rather than manually specifying it ourselves. Second, by implementing the K-folds method, we asked Python to run several models on our training data and then average their performances together. This iterative method dilutes the negative effects of splitting our data into the 20% test partition and the 80% training partition, to the extent we got unlucky and included a concentration of outliers in one of the partitions.

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.

*Please see the Jupyter notebook submitted for this assignment for the calculations wanted for this question.*

Per the calculations displayed in the submitted Jupyter notebook, the predicted fare for this route will be $252.16. At face value, this result is a bit surprising as $252.16 is in approximately the 85th percentile of the historical Fares data. However, here it is important to note two things. First, looking at the coefficients in the optimal model, it can be observed that only the coefficients on the categorical variables carry a negative sign. Given the model inputs above, we can see that all of the categorical variables dummies will be “off” except for GATE. Thus, the only component of the model pulling “down” the price is GATE. The second important observation is that the above numerical inputs into our model are in the high percentiles for their respective variables in the historical data. Therefore, each predictor’s contribution to the final price is large, relative to the numbers in the historical data.

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)**

***Please see the Jupyter notebook submitted for this assignment for the calculations wanted for this question.***

Based on the final selected model, the ticket price will decrease. The SouthWest dummy carries a negative coefficient which, when on, will decrease the final price. After running the model with the dummy on, we can see that the unstandardized price has fallen from $252.16 to $203.34. This means that the SouthWest dummy has an unstandardized coefficient of approximately 48.82 (252.16 – 203.34).

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? | Yes |
| Will it change the penalty level of the final selected model? | No |
| Will it affect the average squared error over the test partition? | Yes, but only at the 16th decimal place. |
| Will it affect the predicted value for the route in (h)? | No |

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