Problem Set 2

This problem set is designed to perform predictive tasks using linear models.

You are allowed to work in groups of up to four students, but you must disclose the members of your group. Individual submissions are required. The code you submit may be identical to the one of the other group members, but we expect the comments and answers to the questions to be your own.

Your submission should consist of: (i) a R markdown document with code, figures and comments, and (ii) a zip folder containing all files needed for replication. You may upload these materials via the course's Canvas website (please do not email us with your homework submission).

The task of this problem set is to predict the number of airlines that operate in a certain market using market-level covariates.

1) Merge the existing market-level airline data you created in the previous problem set (for consistency, start from "*airline_data_market_level.R*") with additional data on market income, hubs, tourist destinations, and slot-controlled airports.

a) Download "*lookup_and_hub.R*". This contains data on which airlines have hubs at each domestic airport. Merge this data with your original dataset. The final variable should be an indicator variable that equals 1 if the route contains a hub at either endpoint.

b) Download "*vacations.R*". This contains data on which cities are classified as vacation destinations. Merge this data with your original dataset. Again, the final variable should be an indicator variable that equals 1 if either endpoint is a vacation destination.

c) Download "*data_income.R*". This contains data on the median income of each MSA that the airport is contained in. Merge this data with your original dataset. The final variable should be the geometric mean of the median income of the market's endpoints.

d) Download "*slot_controlled.R*". This contains data on whether or not an airport

is defined as slot-controlled by the FAA. Merge this data with your original dataset. The final variable should be an indicator variable that equals 1 if the market contains an airport that is slot-controlled.

e) Sort your data by *origin_airport_id* and then by *dest_airport_id*.

2) Divide the data into a training dataset and a test dataset.

a) Set the seed to 0 (set.seed(0)) so that your results are comparable to the solutions.

b) Begin by randomly allocating 50% of the data to the test set and the rest to the training set.

3) Estimate a linear model including six variables (the 4 covariates you created above as well as distance and market size). Compute the Adjusted R squared and calculate the MSE on the test sample.

4) Estimate a linear model with second order polynomials and all cross terms (again using the same six variables). Compute the Adjusted R squared and calculate the MSE on the test sample.

5) Perform a simple covariate selection procedure on the model in (4) using Algorithm 6.3 from James et al. (the textbook) using BIC and Adjusted R squared as criteria. Compute the out-of-sample MSE of the selected models. Compare with the fit and prediction performance of the OLS model

6) Fit to the data both a Ridge regression and a Lasso (again using the same six variables), each for three values of the tuning parameter: 0, 1, and 2. Produce a table that summarizes your estimates for both models and for each of the levels of the tuning parameters. What happens as the penalty parameter increases? Compute the out-of-sample MSE for all models and comment on the predictive performance.

7) Now, for both Lasso and Ridge, perform a 10-fold cross validation procedure to pick the tuning parameter, using the training dataset only. Compute the MSE in the test sample for both Lasso and Ridge. Comment on predictive performance for both models, and compare it with the models in (3)-(5).

8) Now step back, and consider the prediction task of determining the number of airlines that operate in a market. What linear supervised learning procedure seems most appropriate given the sample size and covariate structure? Why do you think is that?

9) Repeat 1 through 7 with the following variations:

a) Add in three additional predictors that are highly correlated with market distance, whether the market is a hub route, and market income. Here is some example code:

```
datam$noise1 = datam$average_distance_m + rnorm(nrow(datam),.01)
datam$noise2 = datam$hub_route + rnorm(nrow(datam),.01)
datam$noise3 = datam$market_income + rnorm(nrow(datam),.01)
```

b) Decrease the size of the training dataset by allocating 90 percent of the data to the test sample. Then allocate 98 percent of the data to the test sample.

Note: with three training sample sizes and two sets of covariates (one with noise variables and one without) there should be six sets of results. Tip: make this simple by writing your code in such a way that allows you to specify at the top which iteration you'd like to run with a change in sample size and/or a boolean that specifies whether or not you'd like to add noise.

c) Write a short note explaining why you think the MSE changes in the way that it does.

10) Now, go back to the simple linear model with 5 covariates and perform the covariate selection procedure (Algorithm 6.3) on this model. What seems to determine airline market structure?