OR/SYST-568 Applied Predictive Analytics (Spring 2019)

Assignment 2 – Linear Regression Models

Due Date: 02/28/2019 Thursday 11:59 PM

1. (30pts) Predicting Airfares on New Routes

Several new airports have opened in major cities, opening the market for new routes (a route refers to a pair of airports), and Southwest has not announced whether it will cover routes to/ from these cities. In order to price flights on these routes, a major airline collected information on 638 air routes in the United States. Some factors are known about these new routes: the distance traveled, demographics of the city where the new airport is located, and whether this city is a vacation destination. Other factors are yet unknown e. g., the number of passengers who will travel this route). A major unknown factor is whether Southwest or another discount airline will travel on these new routes. Southwest's strategy (point-to-point routes covering only major cities, use of secondary airports, standardized fleet, low fares) has been very different from the model followed by the older and bigger airlines (hub-and-spoke model extending to even smaller cities, presence in primary airports, variety in fleet, pursuit of high-end business travelers). The presence of discount airlines is therefore believed to reduce the fares greatly. The data file **Airfare** (provided in csv file) contains real data that were collected for the third quarter of 1996. They consist of the predictors and response listed in the following table. Note that some cities are served by more than one airport, and in those cases the airports are distinguished by their three-letter code.

S_CODE:	Starting airport's code
S_CITY:	Starting city
E_CODE:	Ending airport's code
E_CITY:	Ending city
COUPON:	Average number of coupons (a one-coupon flight is a non-stop flight, a two-coupon flight is a one stop flight, etc.) for that route
NEW:	Number of new carriers entering that route between Q3-96 and Q2-97
VACATION:	Whether a vacation route (Yes) or not (No); Florida and Las Vegas routes are generally considered vacation routes
SW:	Whether Southwest Airlines serves that route (Yes) or not (No)
HI:	Herfindel Index – measure of market concentration (refer to BMGT 681)
S_INCOME:	Starting city's average personal income
E_INCOME:	Ending city's average personal income
S_POP:	Starting city's population
E_POP:	Ending city's population
SLOT:	Whether either endpoint airport is slot controlled or not; this is a measure of airport congestion
GATE:	Whether either endpoint airport has gate constraints or not; this is another measure of airport congestion
DISTANCE	Distance between two endpoint airports in miles
PAX:	Number of passengers on that route during period of data collection
FARE:	Average fare on that route

a. Explore the numerical predictors and response (FARE) by creating a correlation table and examining some scatterplots between FARE and those predictors. What seems to be the best single predictor of FARE? (**Provide the correlation table and the scatter plots**).

b. Explore the categorical predictors excluding the first four (S_CODE, S_CITY, E_CODE, E_CITY) by computing the mean value of FARE according to each category. Which categorical predictor seems best for predicting FARE? (i.e. Find the categorical predictor that has the largest difference in **mean** FARE values between qualitative levels)

- R-code Hint: study and use "aggregate" function, for example
- summaryVacation <-aggregate(FARE~VACATION, data = mydata, FUN = mean)

c. Partition the data into training (60%) and validation (40%) sets (the random seed should be set at **value of 12345**). Build a simple model that involves only the two predictors identified in parts (a) and (b), and report the regression model results. After running the model, generate the plot of residual vs. predicted values and explain whether it violates any of the residual assumptions. (In reporting the regression results, copy and paste parameter estimates of the regression output and report the R-squared and BIC value in training dataset and MSE in validation set).

d. Assume that all variables are available to you now. Use backward selection method to build a regression model. You can ignore the first four predictors in generating the candidate models. (Copy and paste parameter estimates of the final model in the regression output and report the R-squared and BIC value in training dataset and MSE in validation set).

e. Compare the performance of models developed in parts (c) and (d) with in terms of the R-squared value, BIC value, and MSE measure. Which model is the better? Please explain briefly.

Submission:

1. Prepare a pdf file with answers to homework questions, with brief explanations for the analysis/implementation you perform. Please provide your R codes at the end of the submitted file as appendix or in a separate R file.
2. Name the file as "LastName, FirstName-HW2.pdf" (for example, "Ji,Ran-HW2.pdf") and submit it on blackboard.