**Individual Assignment 1: Regression Review with R**

The goal is to predict the FARE as a function of the other variables. Please answer all questions. Supply supporting documentation and show calculations as needed (for example for the RMSE you may want to include a picture of the error measures from the R output). **Please submit a single well-formatted PDF or Word file**. The instructor should not need to go searching for your answers! You include your script as an Appendix, however it will not be graded, but is often useful when providing feedback.

**Note** that the detailed instructions refer to R – you are however free to use any other software.

1. **Data Exploration & Visualization**
2. Using the graphical capabilities of R (or the software of your choice) provide a single plot that captures some aspects of the data. Include the plot as a clearly marked Exhibit.
   1. See Exhibit 1 attached
3. What do you observe from the plot? How could your observation influence your regression model (or why would it not)?
   1. On observation, fare price has a positive relationship with distance. Fare price also seems to generally lower when SW is present compared to fare price when SW is not present.
   2. Presence of SW seems to be a variable that has influence on the fare price, therefore I would want to find out whether this variable has a negative or positive impact on the accuracy of the linear regression model, along with all other variable determined.

2. **Fitting a linear regression model**

1. Following the scripts available online (Data Mining with R) adjust them to:
   1. Randomly partition the data into 70% training and 30% validation, setting the seed to 1.
   2. Run a multivariable regression with all appropriate variables (note that starting end ending airport indicators are probably not very useful variables).

HINT: To remover variables from a data frame in R, one can use the -sign (please refer to the “R Tips and Tricks from Week 1” script).

Provide a summary of the model (that includes the values of the regression coefficients) or otherwise include it as a clearly marked and well formatted Exhibit (please refer to the “R Tips and Tricks from Week 1” script on how to export the needed information).

See Exhibit 2 attached

1. What is the resulting RMSE on the training data?
   1. 34.16087
2. On the validation data?
   1. 38.389
3. From your model, how would you quantify the effects of GATE on the predicted FARE? Please be precise in your interpretation, thinking back to your earlier data analysis class.
   1. I would create a dummy variable for the GATE (1 being free, 0 being constrained) and set one of them to be the constant and the other as an independent variable. That way I can see the changes in fare price on average, holding everything else constant.
4. What is the predicted fare of a leg that has COUPON = 1, NEW = 3, VACATION = No, SW = No, HI = 6000, S\_income = $25000, E\_income = $30000, S\_POP = 4,000,000, E\_POP=7,150,000, SLOT = Free and GATE = constrained, DISTANCE = 1000, and PAX = 6000?
   1. = -8.955E+00

+(-7.412E+00)

+(9.422E-03)\*6000

+(1.391E-03)\*25000

+0.0019\*30000

+(4.719E-06)\*4000000

+(4.430E-06)\*7150000

+(-1.305E+01)

+(7.643E-02)\*1000

+(-9.947E-04)\*6000

= $239.10

1. **Variable Selection**

Experiment with variable selection methods (please refer to Step 8 in the Data Mining with R). Minimally implement forward, backwards and stepwise regression models.

1. From your experiments – pick a model as your final regression model. Provide a summary of the model or otherwise include it as a clearly marked Exhibit.
   1. See Exhibit 3 attached
2. Why did you select this particular model? Please provide quantitative reasoning.
   1. This model removed COUPON variable. which subsequently increased R-Squared (how closely the data are to the fitted line) from 0.8092 in exhibit 2 to 0.8101 in exhibit 3. The higher R-Squared suggested that the model is more accurate. P-value remain minimal (<0.01), suggesting that the model default hypothesis is true.
3. What is the resulting RMSE on the training data?
   1. 34.07567
4. On the validation data?
   1. 38.41205
5. **The Impact of SW**

A senior consultant in the airline industry has indicated that the presence of Southwest on vacation routes has significantly been driving prices down on these legs, beyond other routes.

Add this domain knowledge to your regression model from Section 3 by creating an interaction variable

**HINT**: To create an interaction term in R refer to the “Good to know Module 1”. As an example, here is how I created an interaction term indicating a male applicant asking for credit for Type 1 in the Credit data (you need to implement the same but in your case you want an interaction term for SW =1 and Vacation =1):



Either add this term to both your training and validation data or add it to the airline data and then repartition.

See exhibit 4 attached

1. What is the resulting RMSE on the training and validation data?
   * + 1. Training = 32.5269
       2. Validation = 36.04463
2. How would you quantify the effect of SW on the fare on vacation routes vs. non-vacation routes (using your model)? Does the data support the consultant’s claim? HINT: Carefully think about which variables determine the fare on each type of route (your have both vacation and non-vacation routes, and for each type, some routes are served by SW and some are not, this will require you thinking back to your Data Analysis course).
   * + 1. On average, holding everything else constant, the fare price will increase 5.293e+01 or $52.93 in extra for vacation route when SW is present. The data does not support the consultant’s claim.

Exhibit 1

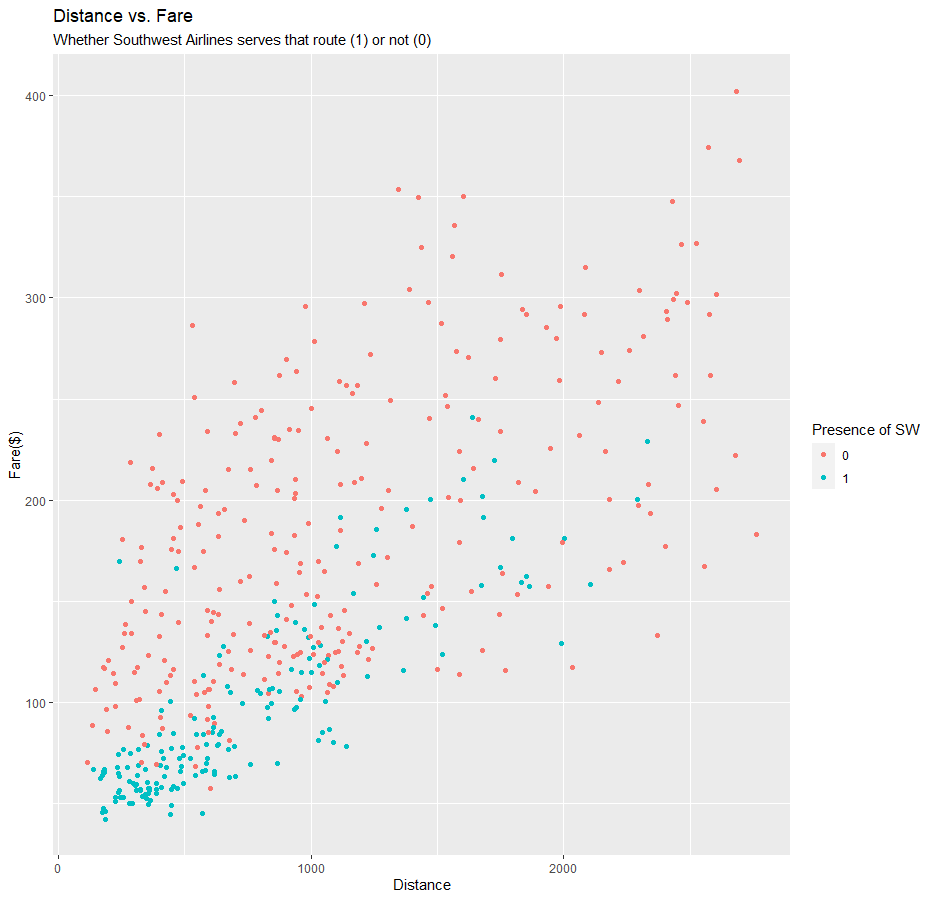


Exhibit 2

Text

Description automatically generated

Exhibit 3

Text, table

Description automatically generated

Exhibit 4

Text, table

Description automatically generated