****

**BUDT758T  
  
DATA MINING AND PREDICTIVE ANALYTICS**

**Homework 1**

**NAME (in capitals): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

* Please submit on Canvas.
* Your submission should consist of this document (with answers filled in in the appropriate places).
* Please ensure that answers are appropriately numbered and clearly legible.
* In the space below please enter the following text and initial below: “I pledge on my honor that I have not given or received unauthorized assistance on this assignment.”

|  |
| --- |
| HONOR PLEDGE:    YOUR INITIALS: |

The goal of this homework is to review the fundamental concepts of regression modeling learned in earlier classes. You will also begin to develop facility with the statistical programming language *R*.

**The Assignment**

The data in the accompanying file “Airline data 2.csv” (posted on Canvas) was assembled by Professor Robert Windle of the Smith School with assistance from Oliver Yao. The file contains information on 627 air routes in the United States. A route refers to a pair of airports. Note that some cities are served by more than one airport. In such cases, the airports are distinguished by their 3-letter code. The data was collected for the third quarter of 1996 (3Q96). The variables in the data set are:

1. S\_CODE: starting airport’s code
2. S\_CITY: starting city
3. E\_CODE: ending airport’s code
4. E\_CITY: ending city
5. 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
6. NEW: number of new carriers entering that route between Q3-96 and Q2-97
7. VACATION: whether a vacation route (Yes) or not (No); Florida and Las Vegas routes are generally considered vacation routes
8. SW: whether Southwest Airlines serves that route (Yes) or not (No)
9. HI: Herfindahl Index – airlines use this as a measure of market concentration (a larger value indicates greater concentration)
10. S\_INCOME: starting city’s average personal income
11. E\_INCOME: ending city’s average personal income
12. S\_POP: starting city’s population
13. E\_POP: ending city’s population
14. SLOT: whether either endpoint airport is slot controlled or not; this is a measure of airport congestion
15. GATE: whether either endpoint airport has gate constraints or not; this is another measure of airport congestion
16. DISTANCE: distance between two endpoint airports in miles
17. PAX: number of passengers on that route during period of data collection
18. FARE: average fare on that route

There are two goals of the study. The first is to predict the FARE as a function of the other variables. The second is to determine how the presence of Southwest Airlines affects fares.

We will **not** use the first four attributes (S\_CODE, S\_CITY, E\_CODE, and E\_CITY) in our analysis.

**The Assignment**

Please answer all questions in the dedicated space and upload on Canvas. Please ensure that your numbering of questions matches those below.

1. **Working with data and regression in R**The VACATION, SW, GATE and SLOT variables will have type *Factor*. The dollar sign before FARE is likely to create a problem – you may choose to address this in *R*, or directly in *Excel* by changing the format from currency to number. Using the resulting new dataset run a multivariable regression for FARE, with all numerical variables (i.e. of type num or int) and the four factors above as independent variables.  
   1. What is the resulting R2?
   2. State precisely what effect the value of SLOT has on the predicted FARE:
   3. What is the predicted fare of a leg that has COUPON = 1, NEW = 3, VACATION = No, SW = No, HI =6000, S\_INCOME = $2000, E\_INCOME = $2000, S\_POP = 4,000,000, E\_POP=7,150,000, SLOT=Free and GATE = Constrained, DISTANCE = 1000, and PAX = 6000?
   4. Do you have any reservations about your predicted fare? If so, explain why.
2. **Exploratory Analysis** 
   1. Use the PAIRS function in *R* to depict pairwise scatterplots of all numerical variables in the data set.
   2. Present a scatterplot of FARE (Y-axis) and DISTANCE (X-axis). Use different colors or symbols to distinguish routes where *Southwest* is present. Briefly describe what you observe.
   3. High correlations between independent variables can be problematic. Present a table of linear correlations and identify any values greater than 0.7.
3. **More Regression Modeling**Run a simple regression to determine the effect of the presence of *Southwest* on FARE. Compare the coefficient with the corresponding value in (1) and explain discrepancies, if any.
4. **Further Analysis**
   1. A senior consultant in the airline industry has indicated that the presence of *Southwest* on Vacation routes has been driving prices down on these legs. Add this domain knowledge to your regression model and run a new multivariable linear regression. Describe how you added this feature to the model.
   2. What is the resulting R2?
   3. Now how would you quantify the effect of SW on the fare?
5. **Comparing two different models**

We will now build a model for FARES using the following explanatory variables: Factor variables for VACATION and SW, HI, S\_INCOME, E\_INCOME, S\_POP, E\_POP, DISTANCE, PAX. Then we will compare this with the model *without* the two INCOME variables.

* 1. Run the regression with and without the INCOME variables.
  2. On the basis of the output evaluate which model is better for predicting FARES. *Explain carefully how you made this determination*.