Case Study 3: Predicting Airfare on Routes

This is a Category B Assignment - Group Assignment:

Your group may not receive help from anyone outside your group. All questions concerning this assignment should be addressed to your professor. It is an honor code offense to give help to other groups and individuals or receive assistance from other groups and individuals.

DATA DESCRIPTION

The following problem takes place in the United States in the late 1990s, when many major US cities were facing issues with airport congestion, partly as a result of the 1978 deregulation of airlines. Both fares and routes were freed from regulation, and low-fare carriers such as Southwest began competing on existing routes and starting nonstop service on routes that previously lacked it. Building completely new airports is generally not feasible, but sometimes decommissioned military bases or smaller municipal airports can be reconfigured as regional or larger commercial airports. There are numerous players and interests involved in the issue (airlines, city, state and federal authorities, civic groups, the military, airport operators), and an aviation consulting firm is seeking advisory contracts with these players.

The firm needs predictive models to support its consulting service. One thing the firm might want to be able to predict is fares, in the event a new airport is brought into service. The firm starts with the file Case3.csv which contains real data that were collected between Q3-1996 and Q2-97.

Variable	Description
S CODE	Starting airport's code
S.CITY	Starting city
ECODE	Ending airport's code
E_CITY	Ending city
COUPON	Average number of coupons (a one-coupon flight is a nonstop flight, a two-coupon flight is a one-stop flight, etc.) for that route
NEW	Number of new carriers entering that route between 03-96 and 02-97
VACATION	Whether (Yes) or not (No) a vacation route
SW	Whether (Yes) or not (No) Southwest Airlines serves that route
HI	Herfindahl index: measure of market concentration
SINCOME	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 or not either endpoint airport is slot controlled (this is a measure of airport congestion)
GATE	Whether or not either endpoint airport has gate constraints (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

The dataset is provided for you in Blackboard and is called Case3.csv.

INSTRUCTIONS

THE MAIN GOAL OF THE ASSIGNMENT IS TO EXPLORE THE FACTORS AFFECTING FARE.

- In an R script file, load the Case3.csv dataset and using the appropriate functions to inspect the data frame.
- Using the dataset Case3.csv, create a linear model that explains at least 40% of the variance in FARE.
- Ignore the first four variables: S CODE, S.CITY, ECODE, E_CITY
- Test your model for the assumptions: linearity, heteroskedasticity, normality of errors, and multicollinearity.
- Transform x or y variables to combat skewness which leads to heteroskedasticity (if present in your model). Give a statement of why you chose what you did and how that affected the model.
- Delete any observations that are considered strong outliers help pass the assumptions. Describe how you found those outliers and why you chose them.
- Provide your final regression equation.
- Provide the final correlation (if a log Y model was used) or adjusted R square value (if no log Y model was used) that is applicable to the model.
- Provide a summary in comments interpreting how well the model did.
- In a team RScript file, include the following:
    - A R file that saves and opens without edits.
    - Code that runs procedurally from top to bottom without error.
    - Your team's name along with individual names listed of people that contributed in comments at the start of the script
    - Data loaded properly directly from your working directory (no subfolders) with summary functions for the dataset.
    - One final regression model with assessment with an interpretation of the beta coefficients.
    - A test of the assumptions and an assessment (in comments) as to whether you think they were violated or not in your final model.
    - An assessment about what your findings do to the integrity of the model.
    - Any other visualizations that describe the relationship or just the variables along with how those visualizations help you.
    - The results along with your implication (in comments) of why we care about the relationship describing what useful information someone can extract from what you found.
- One required R Script file per group should be submitted via blackboard by the due date listed in the system.

Note – splitting the dataset into training and testing groups is not required for this case study.

GROUP CONTRIBUTION AND ASSESSMENT

- Everyone in the group must contribute and write a portion of the code
- Groups will be peer evaluated through teammates.
- Participation and score through teammates will be incorporated into your final average.

Failure to complete a group evaluation on TEAMMATES will result in a 10%-point reduction

RUBRIC

- A (100) – This assignment is considered exemplary. The R script file was clean and free from errors and code ran from start to finish. All parts included. Assessments correct and complete.
- (90) – This assignment is considered well-done. The R script file was free from errors and code ran from start to finish. All parts included. Assessments mostly correct but missing minor detail.
- B (82) - This assignment is considered proficient. The R script file had minor errors preventing it from running without edits. All parts included. Assessments mostly correct but needed clarity and detail.

- C (75) - This assignment is considered acceptable. The R script file had errors preventing it from running without edits. Most parts requested were included. Some assessments incorrect or needed clarity and detail.
- F (60) – This assignment is considered underwhelming. The R script file was submitted, but major detail was missing and or inaccurate. R file Includes some of the necessary components. Not all assessments correct.
- F (0) – No R file submitted.