Regression Practice

August 4, 2025

1 Regression Practice

Takyo software catalog firm that sells games and educational software. It started out as a software manufacturer and then added third-party titles to its offerings. It recently revised its collection of items in a new catalog, which it mailed out to its customers. This mailing yielded 2000 purchases. Based on these data, Takyo wants to devise a model for predicting the spending amount that a purchasing customer will yield.

The file `Tayko.csv` contains information on 2000 purchase. The next table describes the variables to be used in the problem (**the csv file contains additional variables**). For example, the columns source_a, source_c, source_b, … represent different sources or channels through which customers were acquired. Each column contains binary values (0 or 1) indicating whether a particular source was used to acquire the customer.

- **FREQ**: Number of transactions in the preceding year
- **Last_update_days_ago**: Number of dats since last update to customer record
- **Web order**: Whether customer purchased by Web order at least once
- **GENDER**: Male or female
- **Address_RES**: Whether it is a residential address
- **US**: Whether it is a US address
- **SPENDING (target)**: Amount spent by customer in test mailing (in dollars)

Questions:

1. **Data Exploration.** Begin by examining the dataset to understand the structure and types of data available.
    - Explore the relationship between spending and FREQ by a scatter plot (Spending against Freq).
    - (Optional) Generate a bar plot that compares the average spending between web orders and non-web orders. This visualization should help you assess whether placing orders through the web is associated with higher or lower spending.
2. **Machine Learning Model.** To fit a multiple linear regression for Spending:
    - Partition the 2000 records into 75% training vs. 25% test with the random seed set to 1.
    - Run a multiple linear regression for Spending vs. all six predictors. Write down the fitted predictive equation (the number of digits of precision should be set to 4).
    - Evaluate the predictive accuracy of the model by examining its performance on the **test** set using RMSE.