**OPIM 5604 B15 – Predictive Modeling Assignment** **Meghana Kasula (Net ID=mek15120)**

*"The work contained and presented here is my work and my work alone."*

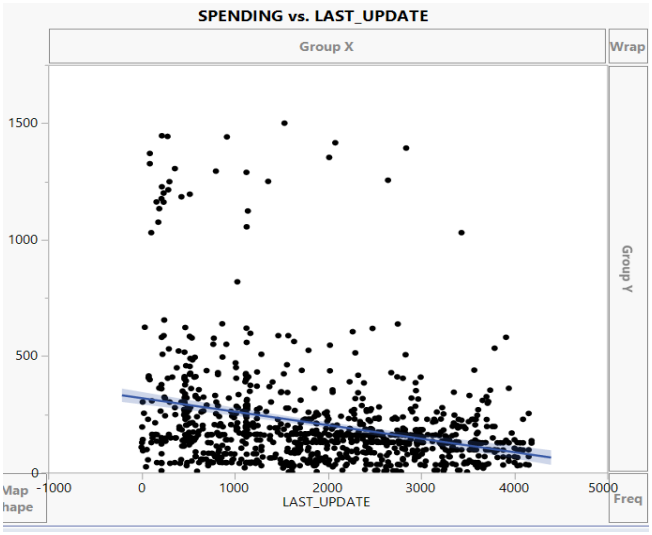
## 6.2 Predicting Software Reselling Profits.

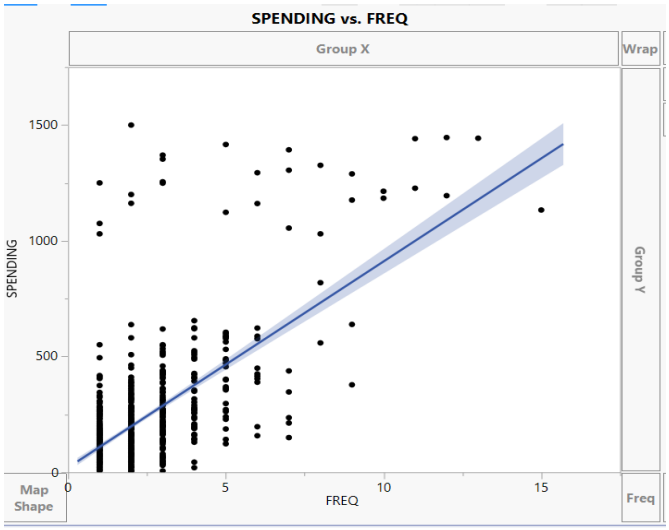
### a. Explore the spending amount by creating a tabular summary of the categorical variables and computing the average and standard deviation of spending in each category.

A-Following is the tabular summary

| | WEB | GENDER | ADDRESS_RES | ADDRESS_US | N Rows | Mean(SPENDING) | Std Dev(SPENDING) |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 39 | 177.58974359 | 131.51740288 |
| 2 | 0 | 0 | 0 | 1 | 131 | 212.83206107 | 237.0980697 |
| 3 | 0 | 0 | 1 | 0 | 11 | 206.18181818 | 104.05269644 |
| 4 | 0 | 0 | 1 | 1 | 48 | 195.125 | 144.16267776 |
| 5 | 0 | 1 | 0 | 0 | 26 | 299.61538462 | 361.74997741 |
| 6 | 0 | 1 | 0 | 1 | 152 | 209.31578947 | 246.97913742 |
| 7 | 0 | 1 | 1 | 0 | 5 | 137.4 | 53.125323528 |
| 8 | 0 | 1 | 1 | 1 | 44 | 191.77272727 | 125.28736525 |
| 9 | 1 | 0 | 0 | 0 | 34 | 255.94117647 | 219.29818377 |
| 10 | 1 | 0 | 0 | 1 | 157 | 219.77070064 | 277.2668083 |
| 11 | 1 | 0 | 1 | 0 | 9 | 230 | 92.54458385 |
| 12 | 1 | 0 | 1 | 1 | 57 | 181.78947368 | 147.93673079 |
| 13 | 1 | 1 | 0 | 0 | 35 | 160.51428571 | 106.64437092 |
| 14 | 1 | 1 | 0 | 1 | 203 | 201.12807882 | 219.85454992 |
| 15 | 1 | 1 | 1 | 0 | 8 | 183.25 | 171.23897253 |
| 16 | 1 | 1 | 1 | 1 | 41 | 159.17073171 | 122.41096814 |

### b. Explore the relationship between spending and each of the two continuous predictors by creating two scatterplots (SPENDING vs. FREQ, and SPENDING vs. LAST_UPDATE). Does there seem to be a linear relationship?




There seems to be a linear relationship. The linear equation has been partially fit. Not all the values are under the linear relation. First one is a negative equation. Second one is a equation with a positive slope.

c. To fit a predictive model for SPENDING:

i. Partition the 1000 records into training and validation sets. (Note that the dataset has a validation column—create your own validation column for this exercise.)

A-Validation set has been created.

| | source_a | source_c | source_b | source_d | source_e | source_m | source_o | source_h | source_r | source_s | source_t | source_u | source_p | source_x | source_w | Validation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Training |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | Training |
| 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Training |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | Training |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | Validation |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | Training |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | Training |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | Training |
| 9 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Training |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | Training |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | Training |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Training |
| 3 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Validation |
| 4 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Training |
| 5 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Training |

ii. Run a multiple linear regression model for SPENDING vs. all six predictors. Give the estimated predictive equation.

A-The formula for prediction is stated beside.

iii. Based on this model, what type of purchaser is most likely to spend a large amount of money?

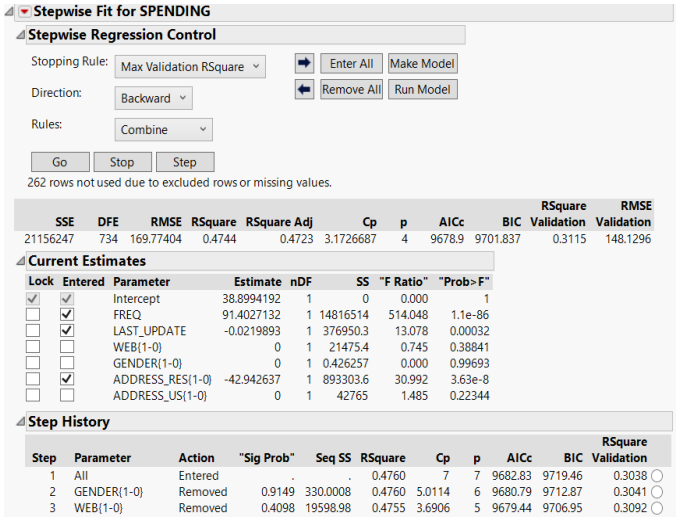
A-Based on the formula, we can see that frequency has the highest impact and hence, purchaser is most likely to spend a large amount if the frequency is high.

**Prediction Expression**

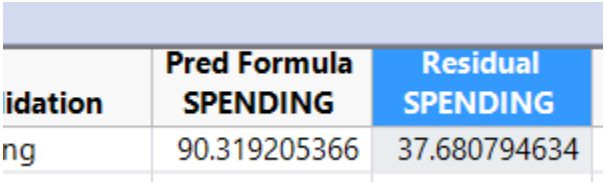
$$
\begin{aligned}
&43.5552808034149 \\
&+ 91.747540897913 * \text{FREQ} \\
&+ -0.0212438274742 * \text{LAST\_UPDATE} \\
&+ \text{Match}\left(\text{WEB}, \begin{cases} 0 \Rightarrow 5.18759802407864 \\ 1 \Rightarrow -5.1875980240786 \\ \text{else} \Rightarrow . \end{cases}\right) \\
&+ \text{Match}\left(\text{GENDER}, \begin{cases} 0 \Rightarrow -0.6729549583002 \\ 1 \Rightarrow 0.6729549583002 \\ \text{else} \Rightarrow . \end{cases}\right) \\
&+ \text{Match}\left(\text{ADDRESS\_RES}, \begin{cases} 0 \Rightarrow 43.0881461279384 \\ 1 \Rightarrow -43.088146127938 \\ \text{else} \Rightarrow . \end{cases}\right) \\
&+ \text{Match}\left(\text{ADDRESS\_US}, \begin{cases} 0 \Rightarrow 9.98756191229101 \\ 1 \Rightarrow -9.987561912291 \\ \text{else} \Rightarrow . \end{cases}\right)
\end{aligned}
$$

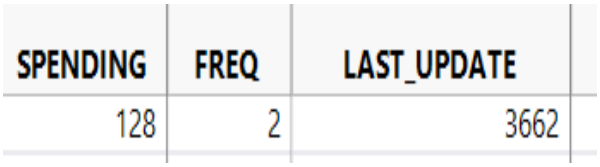
iv. If we used backward elimination to reduce the number of predictors, which predictor would be dropped first from the model?

A-For my model, Gender was the first variable that was removed. As shown in picture beside.



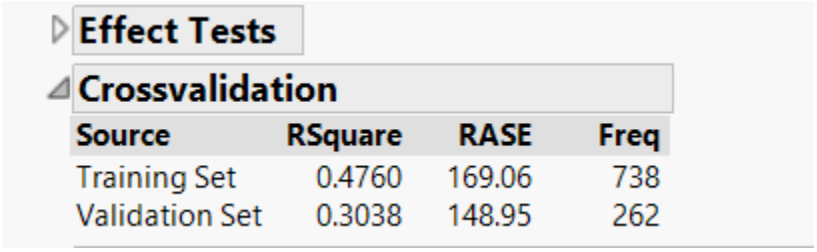

v. Show how the prediction and the prediction error are computed for the first purchase in the validation set.




After the regression model, I have saved the Pred formula Spending, Residual Spending. It can be seen in save columns, save Predicting Formula and Residuals.

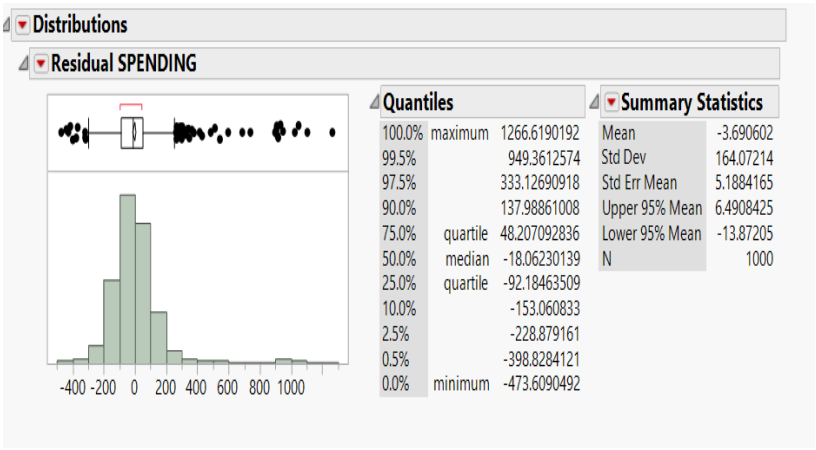
vi. Evaluate the predictive accuracy of the model by examining its performance on the validation set.




We can see that the Rsquare value is more for Training set rather than Validation set. So, that model is slightly better,

vii. Create a histogram of the model residuals. Do they appear to follow a normal distribution? How does this affect the predictive performance of the model?

The histogram is skewed on right. With a lot of outliers. There is no difference in the predictive performance since the residuals do not affect the predictive capabilities. They only mention the variance from the predicted value from the actual value.



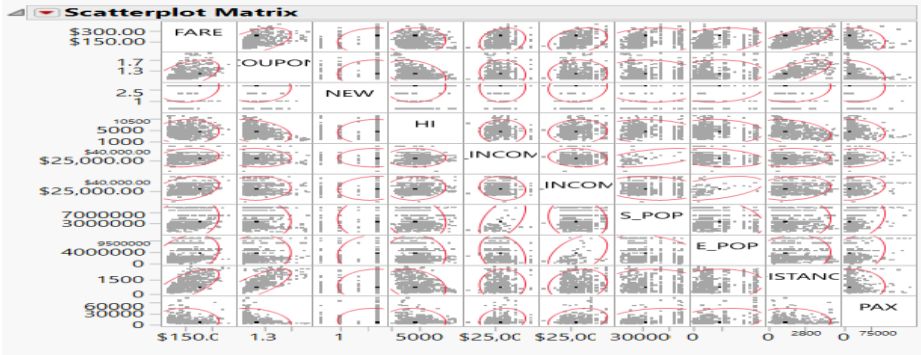

6.3 Predicting Airfare on New Routes6.3 a. Explore the numerical predictors and response (FARE) by creating histograms, a correlation table and a scatterplot matrix. Examining potential relationships between FARE and those predictors. What seems to be the best single predictor of FARE?

Multivariate

Correlations

| | FARE | COUPON | NEW | HI | S_INCOME | E_INCOME | S_POP | E_POP | DISTANCE | PAX |
|---|---|---|---|---|---|---|---|---|---|---|
| FARE | 1.0000 | 0.4965 | 0.0917 | 0.0252 | 0.2091 | 0.3261 | 0.1451 | 0.2850 | 0.6700 | -0.0907 |
| COUPON | 0.4965 | 1.0000 | 0.0202 | -0.3473 | -0.0884 | 0.0469 | -0.1078 | 0.0950 | 0.7468 | -0.3370 |
| NEW | 0.0917 | 0.0202 | 1.0000 | 0.0541 | 0.0266 | 0.1134 | -0.0167 | 0.0586 | 0.0810 | 0.0105 |
| HI | 0.0252 | -0.3473 | 0.0541 | 1.0000 | -0.0274 | 0.0824 | -0.1725 | -0.0625 | -0.3124 | -0.1690 |
| S_INCOME | 0.2091 | -0.0884 | 0.0266 | -0.0274 | 1.0000 | -0.1389 | 0.5172 | -0.2723 | 0.0282 | 0.1382 |
| E_INCOME | 0.3261 | 0.0469 | 0.1134 | 0.0824 | -0.1389 | 1.0000 | -0.1441 | 0.4584 | 0.1765 | 0.2600 |
| S_POP | 0.1451 | -0.1078 | -0.0167 | -0.1725 | 0.5172 | -0.1441 | 1.0000 | -0.2801 | 0.0184 | 0.2846 |
| E_POP | 0.2850 | 0.0950 | 0.0586 | -0.0625 | -0.2723 | 0.4584 | -0.2801 | 1.0000 | 0.1156 | 0.3147 |
| DISTANCE | 0.6700 | 0.7468 | 0.0810 | -0.3124 | 0.0282 | 0.1765 | 0.0184 | 0.1156 | 1.0000 | -0.1025 |
| PAX | -0.0907 | -0.3370 | 0.0105 | -0.1690 | 0.1382 | 0.2600 | 0.2846 | 0.3147 | -0.1025 | 1.0000 |

Scatterplot Matrix



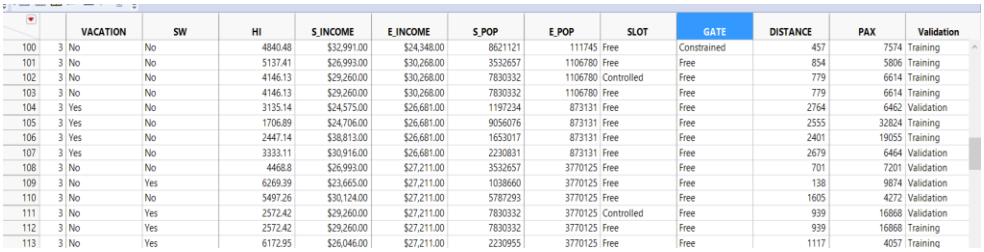

After examining the relationships between all the predictors and fare, I can conclude that Distance has the highest correlation with FARE. Hence, it is the best single predictor.

b. Explore the categorical predictors (excluding the first four) by computing the percentage of flights in each category. Create a tabular summary with the average fare in each category. Which categorical predictor seems best for predicting FARE?

| | VACATION | SW | SLOT | GATE | N Rows | Mean(FARE) |
|---|---|---|---|---|---|---|
| 1 | No | No | Controlled | Constrained | 18 | $206.798 |
| 2 | No | No | Controlled | Free | 114 | $208.330 |
| 3 | No | No | Free | Constrained | 78 | $210.582 |
| 4 | No | No | Free | Free | 119 | $196.184 |
| 5 | Yes | No | Controlled | Free | 28 | $139.902 |
| 6 | Yes | No | Free | Constrained | 20 | $137.314 |
| 7 | Yes | No | Free | Free | 67 | $143.977 |
| 8 | No | Yes | Controlled | Constrained | 1 | $74.280 |
| 9 | No | Yes | Controlled | Free | 17 | $110.171 |
| 10 | No | Yes | Free | Constrained | 7 | $139.960 |
| 11 | No | Yes | Free | Free | 114 | $96.951 |
| 12 | Yes | Yes | Controlled | Free | 4 | $131.605 |
| 13 | Yes | Yes | Free | Free | 51 | $89.811 |

South west is the best since for all the fares, when south west is “NO”, the fares are higher and they follow a trend.

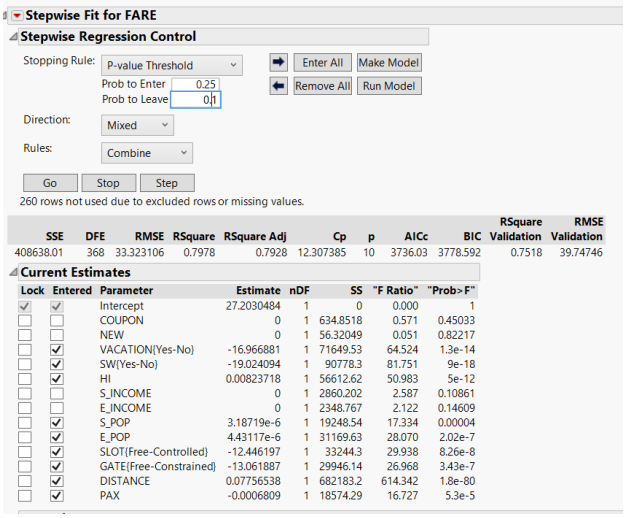
c. Find a model for predicting the average fare on a new route:
i. Partition the data into training and validation sets. The model will be fit to the training data and evaluated on the validation set.



| | | VACATION | SW | HI | S_INCOME | E_INCOME | S_POP | E_POP | SLOT | GATE | DISTANCE | PAX | Validation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 100 | 3 | No | No | 4840.48 | $32,991.00 | $24,348.00 | 8621121 | 111745 | Free | Constrained | 457 | 7574 | Training |
| 101 | 3 | No | No | 5137.41 | $26,993.00 | $30,268.00 | 3532657 | 1106780 | Free | Free | 854 | 5806 | Training |
| 102 | 3 | No | No | 4146.13 | $29,260.00 | $30,268.00 | 7830332 | 1106780 | Controlled | Free | 779 | 6614 | Training |
| 103 | 3 | No | No | 4146.13 | $29,260.00 | $30,268.00 | 7830332 | 1106780 | Free | Free | 779 | 6614 | Training |
| 104 | 3 | Yes | No | 3135.14 | $24,575.00 | $26,681.00 | 1197234 | 873131 | Free | Free | 2764 | 6462 | Validation |
| 105 | 3 | Yes | No | 1706.89 | $24,706.00 | $26,681.00 | 9056076 | 873131 | Free | Free | 2555 | 32824 | Training |
| 106 | 3 | Yes | No | 2447.14 | $38,813.00 | $26,681.00 | 1653017 | 873131 | Free | Free | 2401 | 19055 | Training |
| 107 | 3 | Yes | No | 3333.11 | $30,916.00 | $26,681.00 | 2230831 | 873131 | Free | Free | 2679 | 6464 | Validation |
| 108 | 3 | No | No | 4468.8 | $26,993.00 | $27,211.00 | 3532657 | 3770125 | Free | Free | 701 | 7201 | Validation |
| 109 | 3 | No | Yes | 6269.39 | $23,665.00 | $27,211.00 | 1038660 | 3770125 | Free | Free | 138 | 9874 | Validation |
| 110 | 3 | No | No | 5497.26 | $30,124.00 | $27,211.00 | 5787293 | 3770125 | Free | Free | 1605 | 4272 | Validation |
| 111 | 3 | No | Yes | 2572.42 | $29,260.00 | $27,211.00 | 7830332 | 3770125 | Controlled | Free | 939 | 16868 | Validation |
| 112 | 3 | No | Yes | 2572.42 | $29,260.00 | $27,211.00 | 7830332 | 3770125 | Free | Free | 939 | 16868 | Training |
| 113 | 3 | No | Yes | 6172.95 | $26,046.00 | $27,211.00 | 2230955 | 3770125 | Free | Free | 1117 | 4057 | Training |

Done using Validation with 60-40 division.

ii. Use mixed stepwise regression to reduce the number of predictors. You can ignore the first four predictors (S_CODE, S_CITY, E_CODE, E_CITY). Report the estimated model selected.



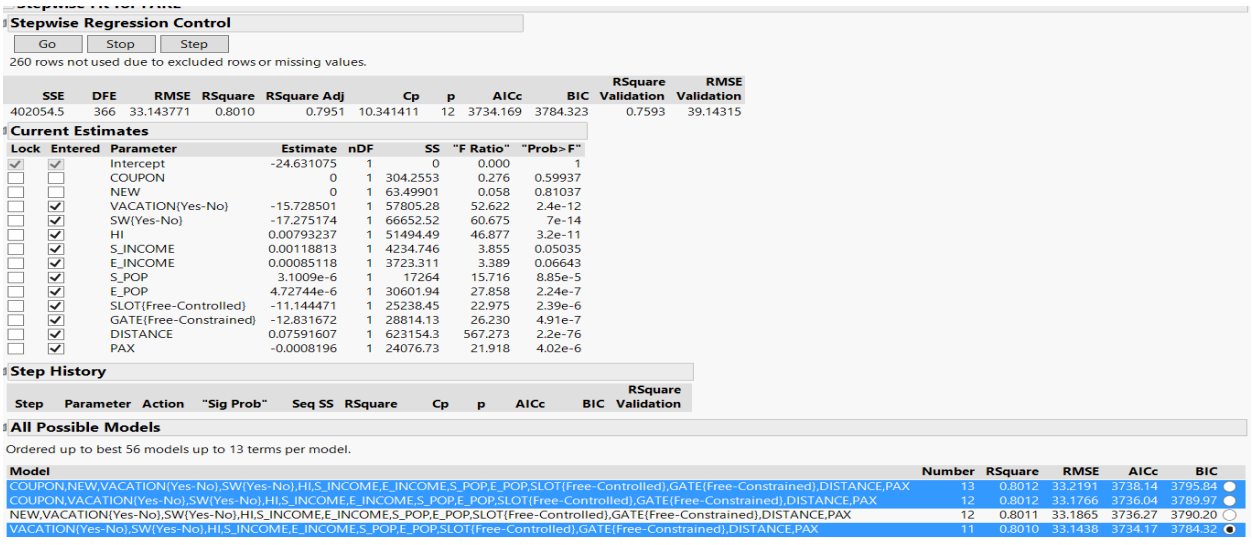

| SSE | DFE | RMSE | RSquare | RSquare Adj | Cp | p | AICc | BIC | RSquare Validation | RMSE Validation |
|---|---|---|---|---|---|---|---|---|---|---|
| 408638.01 | 368 | 33.323106 | 0.7978 | 0.7928 | 12.307385 | 10 | 3736.03 | 3778.592 | 0.7518 | 39.74746 |

Current Estimates

| Lock | Entered | Parameter | Estimate | nDF | SS | "F Ratio" | "Prob>F" |
|---|---|---|---|---|---|---|---|
| ✓ | ✓ | Intercept | 27.2030484 | 1 | 0 | 0.000 | 1 |
| | | COUPON | 0 | 1 | 634.8518 | 0.571 | 0.45033 |
| | | NEW | 0 | 1 | 56.32049 | 0.051 | 0.82217 |
| | ✓ | VACATION{Yes-No} | -16.966881 | 1 | 71649.53 | 64.524 | 1.3e-14 |
| | ✓ | SW{Yes-No} | -19.024094 | 1 | 90778.3 | 81.751 | 9e-18 |
| | ✓ | HI | 0.00823718 | 1 | 56612.62 | 50.983 | 5e-12 |
| | | S_INCOME | 0 | 1 | 2860.202 | 2.587 | 0.10861 |
| | | E_INCOME | 0 | 1 | 2348.767 | 2.122 | 0.14609 |
| | ✓ | S_POP | 3.18719e-6 | 1 | 19248.54 | 17.334 | 0.00004 |
| | ✓ | E_POP | 4.43117e-6 | 1 | 31169.63 | 28.070 | 2.02e-7 |
| | ✓ | SLOT{Free-Controlled} | -12.446197 | 1 | 33244.3 | 29.938 | 8.26e-8 |
| | ✓ | GATE{Free-Constrained} | -13.061887 | 1 | 29946.14 | 26.968 | 3.43e-7 |
| | ✓ | DISTANCE | 0.07756538 | 1 | 682183.2 | 614.342 | 1.8e-80 |
| | ✓ | PAX | -0.0006809 | 1 | 18574.29 | 16.727 | 5.3e-5 |

The screenshot beside is the model I got after generating mixed stepwise regression. It involves 9 variables.

iii. Repeat (ii) using All Possible Models instead of *Mixed stepwise* regression. Compare the resulting best model to the one you obtained in (ii) in terms of the predictors that are in the model.




| SSE | DFE | RMSE | RSquare | RSquare Adj | Cp | p | AICc | BIC | RSquare Validation | RMSE Validation |
|---|---|---|---|---|---|---|---|---|---|---|
| 402054.5 | 366 | 33.143771 | 0.8010 | 0.7951 | 10.341411 | 12 | 3734.169 | 3784.323 | 0.7593 | 39.14315 |

Current Estimates

| Lock | Entered | Parameter | Estimate | nDF | SS | "F Ratio" | "Prob>F" |
|---|---|---|---|---|---|---|---|
| ✓ | ✓ | Intercept | -24.631075 | 1 | 0 | 0.000 | 1 |
| | | COUPON | 0 | 1 | 304.2553 | 0.276 | 0.59937 |
| | | NEW | 0 | 1 | 63.49901 | 0.058 | 0.81037 |
| | ✓ | VACATION{Yes-No} | -15.728501 | 1 | 57805.28 | 52.622 | 2.4e-12 |
| | ✓ | SW{Yes-No} | -17.275174 | 1 | 66652.52 | 60.675 | 7e-14 |
| | ✓ | HI | 0.00793237 | 1 | 51494.49 | 46.877 | 3.2e-11 |
| | ✓ | S_INCOME | 0.00118813 | 1 | 4234.746 | 3.855 | 0.05035 |
| | ✓ | E_INCOME | 0.00085118 | 1 | 3723.311 | 3.389 | 0.06643 |
| | ✓ | S_POP | 3.1009e-6 | 1 | 17264 | 15.716 | 8.85e-5 |
| | ✓ | E_POP | 4.72744e-6 | 1 | 30601.94 | 27.858 | 2.24e-7 |
| | ✓ | SLOT{Free-Controlled} | -11.144471 | 1 | 25238.45 | 22.975 | 2.39e-6 |
| | ✓ | GATE{Free-Constrained} | -12.831672 | 1 | 28814.13 | 26.230 | 4.91e-7 |
| | ✓ | DISTANCE | 0.07591607 | 1 | 623154.3 | 567.273 | 2.2e-76 |
| | ✓ | PAX | -0.0008196 | 1 | 24076.73 | 21.918 | 4.02e-6 |

Step History

| Step | Parameter | Action | "Sig Prob" | Seq SS | RSquare | Cp | p | AICc | BIC | RSquare Validation |
|---|---|---|---|---|---|---|---|---|---|---|

All Possible Models

Ordered up to best 56 models up to 13 terms per model.

| Model | Number | RSquare | RMSE | AICc | BIC |
|---|---|---|---|---|---|
| COUPON,NEW,VACATION{Yes-No},SW{Yes-No},HI,S_INCOME,E_INCOME,S_POP,E_POP,SLOT{Free-Controlled},GATE{Free-Constrained},DISTANCE,PAX | 13 | 0.8012 | 33.2191 | 3738.14 | 3795.84 |
| COUPON,VACATION{Yes-No},SW{Yes-No},HI,S_INCOME,E_INCOME,S_POP,E_POP,SLOT{Free-Controlled},GATE{Free-Constrained},DISTANCE,PAX | 12 | 0.8012 | 33.1766 | 3736.04 | 3789.97 |
| NEW,VACATION{Yes-No},SW{Yes-No},HI,S_INCOME,E_INCOME,S_POP,E_POP,SLOT{Free-Controlled},GATE{Free-Constrained},DISTANCE,PAX | 12 | 0.8011 | 33.1865 | 3736.27 | 3790.20 |
| VACATION{Yes-No},SW{Yes-No},HI,S_INCOME,E_INCOME,S_POP,E_POP,SLOT{Free-Controlled},GATE{Free-Constrained},DISTANCE,PAX | 11 | 0.8010 | 33.1438 | 3734.17 | 3784.32 |

As we can see, the one in blue are the optimal regression models with best Rsquare value. But I would choose the third one, with 11 variables. Since the Rsquare values are similar, it is a little less expensive as it has less number of variables. The one with 13 values are is also a good model with high RSquare value.
So, the result in model ii has only 9 variables using mixed stepwise. Model iii has 11 variables which is more optimum and with high RSquare value.

## iv. Compare the predictive accuracy of both models (ii) and (iii) using measures such as RMSE, *Cp*, *AICc*, and Validation RSquare.

MODEL ii (With 9 variables)

| SSE | DFE | RMSE | RSquare | RSquare Adj | Cp | p | AICc | BIC | RSquare Validation | RMSE Validation |
|---|---|---|---|---|---|---|---|---|---|---|
| 408638.01 | 368 | 33.323106 | 0.7978 | 0.7928 | 12.307385 | 10 | 3736.03 | 3778.592 | 0.7518 | 39.74746 |

MODEL iii (with 11 variables)

| SSE | DFE | RMSE | RSquare | RSquare Adj | Cp | p | AICc | BIC | RSquare Validation | RMSE Validation |
|---|---|---|---|---|---|---|---|---|---|---|
| 402054.5 | 366 | 33.143771 | 0.8010 | 0.7951 | 10.341411 | 12 | 3734.169 | 3784.323 | 0.7593 | 39.14315 |

There is a very slight difference in Rsquare and Rsquare ADJ. There is a larger difference in BIC, SIC p, Cp value.

## v. Using model (iii), predict the average fare on a route with the following characteristics: COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S_INCOME = $28,760, E_INCOME = $27,664, S_POP = 4,557,004, E_POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782, DISTANCE = 1976 miles.

| Parameter | |
|---|---|
| Intercept | 0 |
| COUPON | 1.202 |
| NEW | 3 |
| VACATION[No] | 1 |
| SW[No] | 1 |
| HI | 4442.141 |
| S_INCOME | 28760 |
| E_INCOME | 27664 |
| S_POP | 4557004 |
| E_POP | 3195503 |
| SLOT[Controlled] | 0 |
| GATE[Constrained] | 0 |
| DISTANCE | 1976 |
| PAX | 12782 |
| = | 0 |

| | |
|---|---|
| Value | 290.5966461 |
| Std Error | 29.44861105 |
| t Ratio | 9.8679236724 |
| Prob>\|t\| | 1.326407e-20 |
| SS | 113468.88682 |

| | |
|---|---|
| Sum of Squares | 113468.88682 |
| Numerator DF | 1 |
| F Ratio | 97.375917603 |
| Prob > F | 1.326407e-20 |

The average Fare is - $290.59

vi. Using model (iii), predict the reduction in average fare on the route in (v) if Southwest decides to cover this route.

If southwest decides to cover this, the reduction is approximately $21.
The average fare now is $269.55

Custom Test

| Parameter | |
|---|---|
| Intercept | 0 |
| COUPON | 1.202 |
| NEW | 3 |
| VACATION[No] | 1 |
| SW[No] | 0 |
| HI | 4442.141 |
| S_INCOME | 28760 |
| E_INCOME | 27664 |
| S_POP | 4557004 |
| E_POP | 3195503 |
| SLOT[Controlled] | 0 |
| GATE[Constrained] | 0 |
| DISTANCE | 1976 |
| PAX | 12782 |
| = | 0 |

| | |
|---|---|
| Value | 269.55818949 |
| Std Err... | 30.155143932 |
| t Ratio | 8.9390450299 |
| Prob>\|t\| | 1.715606e-17 |
| SS | 93112.391472 |

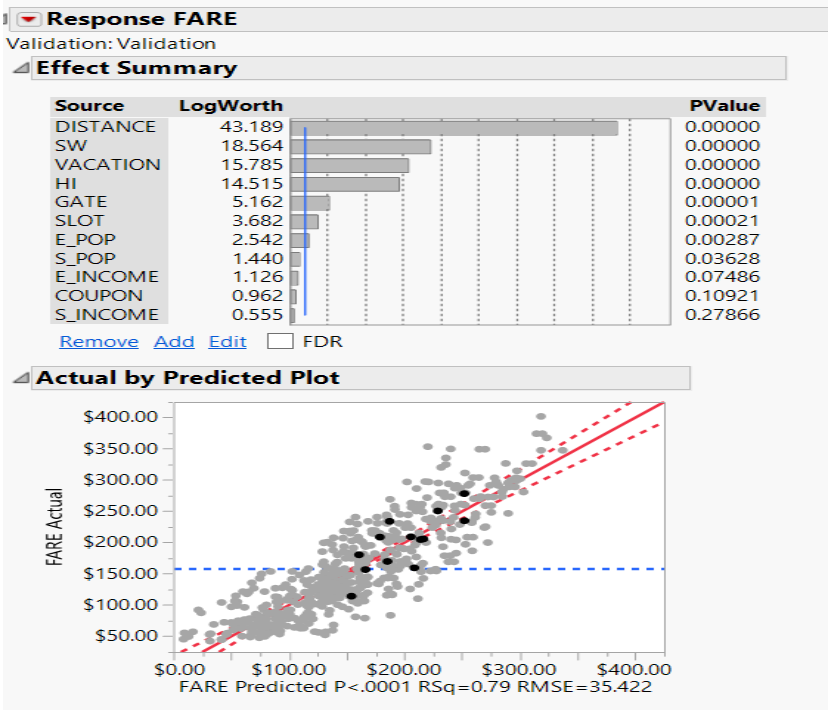
| | |
|---|---|
| Sum of Squares | 93112.391472 |
| Numerator DF | 1 |
| F Ratio | 79.906526046 |
| Prob > F | 1.715606e-17 |

vii. In reality, which of the factors will not be available for predicting the average fare from a new airport (i.e., before flights start operating on those routes)? Which ones can be estimated? How?

COUPON, NEW- not available

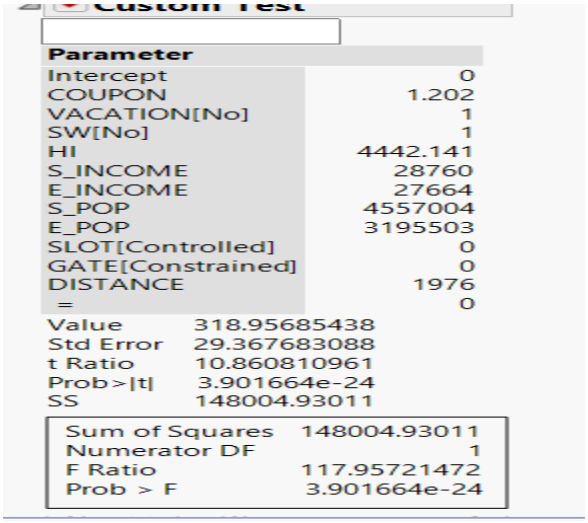
E_INCOME, S_INCOME, GATE,SLOT,PAX- estimated.

viii. Select a model that includes only factors that are available before flights begin to operate on the new route. Use an exhaustive search (All Possible Models) to find such a model.



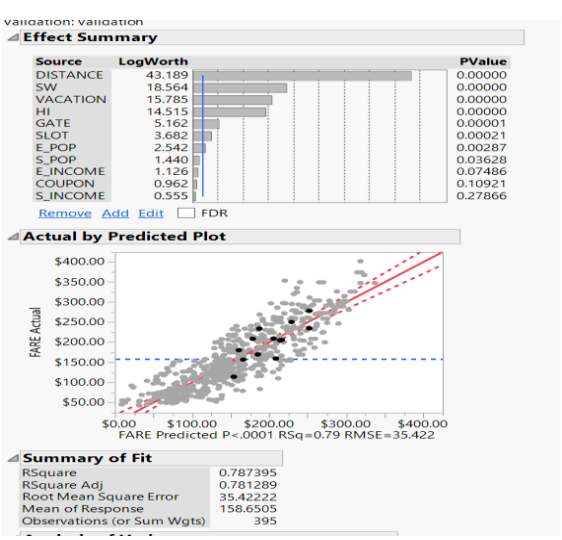

Following is the model with only available available and predictors which can be estimated values.

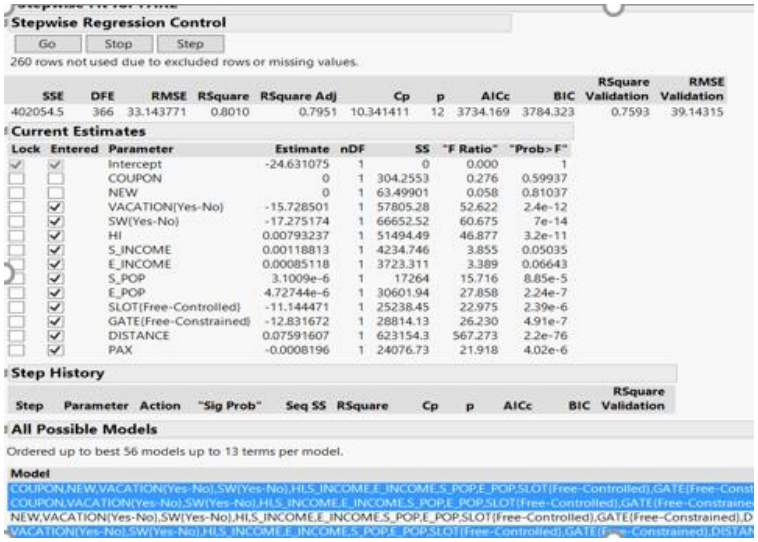
ix. Use model (viii) to predict the average fare on a route with characteristics COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S_INCOME = $28,760, E_INCOME = $27,664, S_ POP = 4,557,004, E_POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782, DISTANCE = 1976 miles.




The picture beside is the one with the values and the value of fare is roughly $318.95.

x. Compare the predictive accuracy of this model with model (iii). Is this model good enough, or is it worthwhile reevaluating the model once flights begin on the new route?




Model 3 is better due to more r square value. Yes, it is worthwhile.

d. In competitive industries, a new entrant with a novel business plan can have a disruptive effect on existing firms. If a new entrant's business model is sustainable, other players are forced to respond by changing their business practices. If the goal of the analysis was to evaluate the effect of Southwest Airlines' presence on the airline industry rather than predicting fares on new routes, how would the analysis be different? Describe technical and conceptual aspects.

Here, we can notice, the only variable that we can consider where can predict is the coupons. We can use the business acumen and predict and modulate the coupon prices. Hence, it’s the only other thing which can predicted using the other predictors.