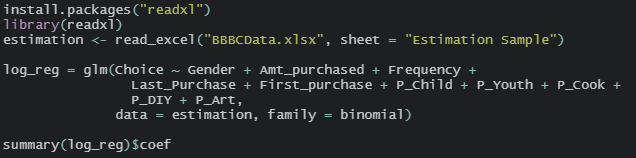
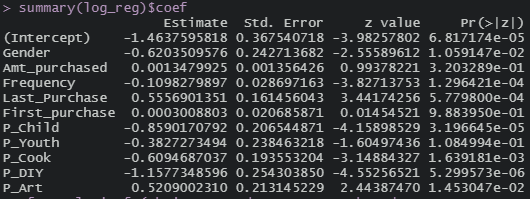
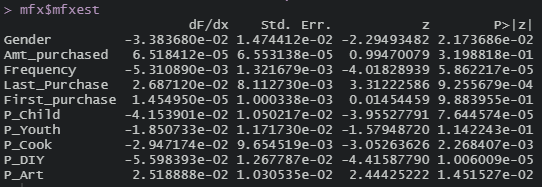
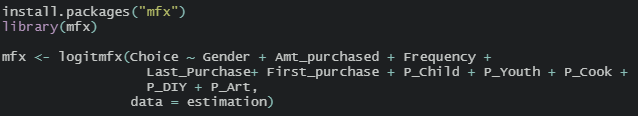
# Assignment 4

**Q1. Estimate a logistic regression model using “Choice” as the dependent variable and the following as explanatory variables: Gender, Amt\_purchased, Frequency, Last\_Purchase, First\_purchase, P\_Child, P\_Youth, P\_Cook, P\_DIY, and P\_Art. Report the regression results.**





**Q2. Summarize and interpret the results (so that a marketing manager can understand them). Which variables are statistically significant? Which seem to be economically “important”? Interpret the economic importance for some of the explanatory variables.**



From the logistic regression model, it can be concluded that only the following factors are proven to have statistically significant relationships with the customer’s decision on whether to buy the book - *The Art History of Florence*.

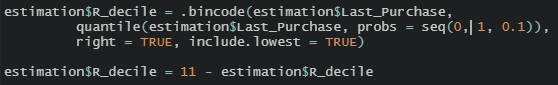
* Gender
  + While holding other variables constant, female customers have a higher probability of buying the book than male customers.
* Frequency
  + While holding other variables constant, the more purchases a customer made, the less likely the customer is going to buy the book.
* Last Purchase
  + While holding other variables constant, the longer it is since a customer made his/her last purchase, the more likely the customer buys the book.
* Number of children’s books purchased
  + While holding other variables constant, the more children’s books a customer purchased in the past, the less likely the customer buys the book.
* Number of cookbooks purchased
  + While holding other variables constant, the more cookbooks a customer purchased in the past, the less likely the customer buys the book.
* Number of do-it-yourself books purchased
  + While holding other variables constant, the more do-it-yourself books a customer purchased in the past, the less likely the customer buys the book.
* Number of art books purchased
  + While holding other variables constant, the more art books a customer purchased in the past, the more likely the customer buys the book.

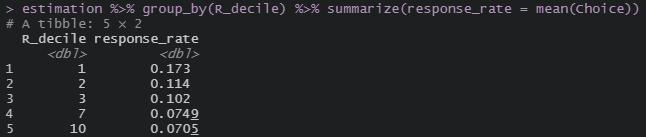
Other variables such as the total money spent on BBBC books, months since first purchase, and number of youth books purchased do not have statistically significant relationships with whether a customer buys the book.

By calculating the marginal probability change at mean of each variable, it can be noted that a few variables can have significantly more impact than others. The number of DIY books purchased in the past has the most impact on the probability of the customer buying the book. One DIY book purchased decreases the probability of buying The Art History of Florence by 5.60%. Similarly, one children’s book purchased decreases the probability of buying The Art History of Florence by 4.15%.

On the contrary, one more month since the customer’s last purchase increases the probability of buying The Art History of Florence by 2.69%. One more art book purchased increases the probability of buying The Art History of Florence by 2.52%.

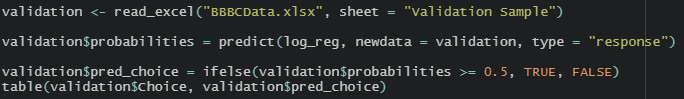
**Q3. Run a “pure R” model (i.e., a RFM model without F and M) with deciles and predict response rate for each decile group.**





The predicted response rate for each decile is listed in the table. Decile 1 has the largest number of months since last purchase and the highest response rate as well.

**Q4. Check the overall classification performance of the logistic regression model above on the validation sample. Create a table showing the fraction of observations which are correctly predicted by the model.**

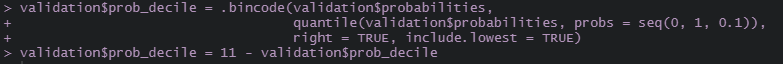


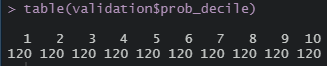




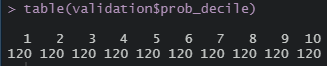
The confusion matrix is shown above. The accuracy of the logistic regression model tested on the validation dataset is 91.3%

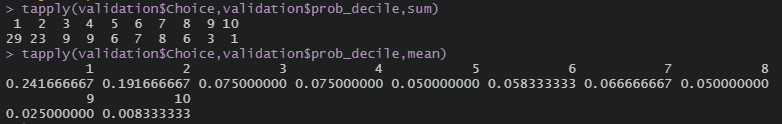
**Q5. Using your logistic regression result, assign each customer to a decile based on his/her predicted probability of purchase. (Hint: use .bincode function as we did in RFM.)**





**Q6. Using your logistic regression result, report the number of customers, the number of buyers of “The Art History of Florence,” and the average response rate to the offer by decile for the 1200 customers in the validation dataset, respectively.**





**Q7. Using your pure-R model, report the number of customers and the number of buyers of “The Art History of Florence,” by decile for the 1200 customers in the validation dataset.**

The number of buyers in each R decile is shown below.



The number of customers in each R decile is shown below.



**Q8. Use the computations above to create a table showing the lift and cumulative lift for each decile, for both logistic regression results and R(FM) results. You may want to use Excel for these calculations.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **R Decile** | | | | | | | | |
| **Decile** | **Num Customer** | **Cum Customer** | **Num Buyer** | **Cum Buyers** | **Response Rate** | **Lift** | **Cum Response Rate** | **Cum Lift** |
| 1 | 148 | 148 | 13 | 13 | 8.78% | 104.36 | 8.78% | 104.36 |
| 2 | 75 | 223 | 10 | 23 | 13.33% | 158.42 | 10.31% | 122.54 |
| 3 | 152 | 375 | 18 | 41 | 11.84% | 140.70 | 10.93% | 129.90 |
| 7 | 386 | 761 | 28 | 69 | 7.25% | 86.18 | 9.07% | 107.73 |
| 10 | 439 | 1200 | 32 | 101 | 7.29% | 86.61 | 8.42% | 100.00 |
| Total | 1200 |  | 101 |  | 8.42% |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Logistic Regression Decile** | | | | | | | | |
| **Decile** | **Num Customer** | **Cum Customer** | **Num Buyer** | **Cum Buyers** | **Response Rate** | **Lift** | **Cum Response Rate** | **Cum Lift** |
| 1 | 120 | 120 | 29 | 29 | 24.17% | 287.13 | 24.17% | 287.13 |
| 2 | 120 | 240 | 23 | 52 | 19.17% | 227.72 | 21.67% | 257.43 |
| 3 | 120 | 360 | 9 | 61 | 7.50% | 89.11 | 16.94% | 201.32 |
| 4 | 120 | 480 | 9 | 70 | 7.50% | 89.11 | 14.58% | 173.27 |
| 5 | 120 | 600 | 6 | 76 | 5.00% | 59.41 | 12.67% | 150.50 |
| 6 | 120 | 720 | 7 | 83 | 5.83% | 69.31 | 11.53% | 136.96 |
| 7 | 120 | 840 | 8 | 91 | 6.67% | 79.21 | 10.83% | 128.71 |
| 8 | 120 | 960 | 6 | 97 | 5.00% | 59.41 | 10.10% | 120.05 |
| 9 | 120 | 1080 | 3 | 100 | 2.50% | 29.70 | 9.26% | 110.01 |
| 10 | 120 | 1200 | 1 | 101 | 0.83% | 9.90 | 8.42% | 100.00 |
| Total | 1200 |  | 101 |  | 8.42% |  |  |  |

**Q9. Use the computations above to create a table showing the gains and cumulative gains for each decile, for both logistic regression results and R(FM) results. You may want to use Excel for these calculations.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **R Decile** | | | | | | |
| **Decile** | **Num Customer** | **Cum Customer** | **Num Buyer** | **Cum Buyers** | **Gains** | **Cum Gains** |
| 1 | 148 | 148 | 13 | 13 | 12.87% | 12.87% |
| 2 | 75 | 223 | 10 | 23 | 9.90% | 22.77% |
| 3 | 152 | 375 | 18 | 41 | 17.82% | 40.59% |
| 7 | 386 | 761 | 28 | 69 | 27.72% | 68.32% |
| 10 | 439 | 1200 | 32 | 101 | 31.68% | 100.00% |
| Total | 1200 |  | 101 |  | 8.42% |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Logistic Regression Decile** | | | | | | |
| **Decile** | **Num Customer** | **Cum Customer** | **Num Buyer** | **Cum Buyers** | **Gains** | **Cum Gains** |
| 1 | 120 | 120 | 29 | 29 | 28.71% | 28.71% |
| 2 | 120 | 240 | 23 | 52 | 22.77% | 51.49% |
| 3 | 120 | 360 | 9 | 61 | 8.91% | 60.40% |
| 4 | 120 | 480 | 9 | 70 | 8.91% | 69.31% |
| 5 | 120 | 600 | 6 | 76 | 5.94% | 75.25% |
| 6 | 120 | 720 | 7 | 83 | 6.93% | 82.18% |
| 7 | 120 | 840 | 8 | 91 | 7.92% | 90.10% |
| 8 | 120 | 960 | 6 | 97 | 5.94% | 96.04% |
| 9 | 120 | 1080 | 3 | 100 | 2.97% | 99.01% |
| 10 | 120 | 1200 | 1 | 101 | 0.99% | 100.00% |
| Total | 1200 |  | 101 |  | 8.42% |  |

**Q10. Create a chart showing the cumulative gains by decile along with a reference line corresponding to “no model,” for the logistic regression and R(FM).**

**Q11. What is the breakeven response rate? (Hint: No R is required.)**