**Project Name** : Computer Choice Study **Author** : Suvro Banerjee

**Functional Description** :

In 1998 Microsoft introduced a new operating system. Computer manufacturers were interested in making predictions about the personal computer marketplace. To help manufacturers understand the market for personal computers, we (a bunch of Economics practitioners) conducted a computer choice study. It was a nation wide study, we identified people who expressed an interest in buying a new personal computer within the next year.

**Technical Specifications:**

The survey consisted of 16 pages, we call it as Choice Sets, with each Choice Set (i.e. each page) comprising 4 Product Profiles.

For each Choice Set, the survey respondents were asked to select the computer they most prefered (i.e of 4 product profiles in 1 Choice Set, the respondent was asked to choose 1 product profile).

Each product profile is defined in terms of attributes and corresponding levels. They are as follows. Also each product profile has the consumer’s binary choice i.e.

Choice = 1 if chosen

Choice = 0 if not chosen

|  |  |  |  |
| --- | --- | --- | --- |
| **Sn** | **Attribute** | **Level Code** | **Level Description** |
| 1 | Brand | Apple | Manufacturer : Apple |
| Compaq | Manufacturer : Compaq |
| Dell | Manufacturer : Dell |
| Gateway | Manufacturer : Gateway |
| HP | Manufacturer : HP |
| IBM | Manufacturer : IBM |
| Sony | Manufacturer : Sony |
| Sun | Manufacturer : Sun Microsystems |
| 2 | Compatibility | 1 | 65 % Compatible |
| 2 | 70 % Compatible |
| 3 | 75 % Compatible |
| 4 | 80 % Compatible |
| 5 | 85 % Compatible |
| 6 | 90 % Compatible |
| 7 | 95 % Compatible |
| 8 | 100 % Compatible |
| 3 | Performance | 1 | Just as fast |
| 2 | Twice as fast |
| 3 | Three times as fast |
| 4 | Four times as fast |
| 4 | Reliability | 1 | As likely to fail |
| 2 | Less likely to fail |
| 5 | Learn | 1 | 4 hours to learn |
| 2 | 8 hours to learn |
| 3 | 12 hours to learn |
| 4 | 16 hours to learn |
| 5 | 20 hours to learn |
| 6 | 24 hours to learn |
| 7 | 28 hours to learn |
| 8 | 32 hours to learn |
| 6 | Price | 1 | $1,000 |
| 2 | $1,250 |
| 3 | $1,500 |
| 4 | $1,750 |
| 5 | $2,000 |
| 6 | $2,250 |
| 7 | $2,500 |
| 8 | $2,750 |

**Predictive Modeling Tasks**

* To predict individual consumer choice.
* To estimate Brand Loyalty, Price Sensitivity, Feature Focus and Brand Switching
* Market Simulations ("What-if analyses") using the Consumer Preference and Choice Models.

1. **To Predict Individual Consumer Choice**
2. **Model Train & Test Regimen** :

In this context we build predictive models on 12 Choice Sets and test on 4 Choice Sets which we choose randomly.

Train Choice Sets : 1, 2, 4, 5, 6, 8, 9, 10, 12, 13, 14, and 16

Test Choice Sets : 3, 7, 11, and 15

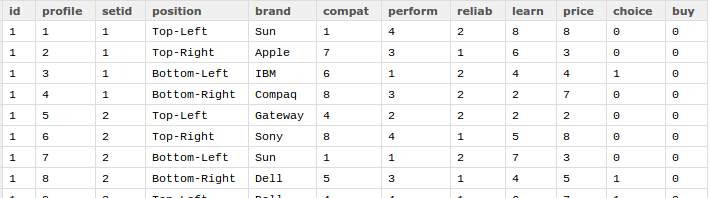
In total Training data set includes 12 \* 4 = 48 product profiles per individual.

In total Test data set includes 4 \* 4 = 16 product profiles per individual.

1. **Read in the Respondent Survey Data** :

The following table is the survey output from the respondents.

The csv file used is “*computer\_choice\_study.csv*”



* id : Consumer's ID (i.e. distinct consumers who responded)

In total 224 consumers responded.

* profile : value from 1 to 64, representing 64 different product profiles given to each

consumer for the survey.

* setid : value from 1 to 16, i.e 16 different Choice sets given to each consumer for the

survey, Note: each choice set has 4 product profiles.

* position : how the product profiles are aligned in each Choice Set, page position.
* <attributes> : brand, compat, perform, reliab, learn and price

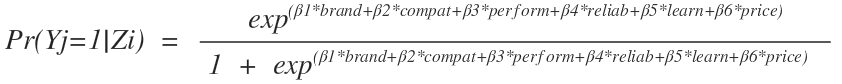
(6 attributes and their corresponding levels, explained earlier)

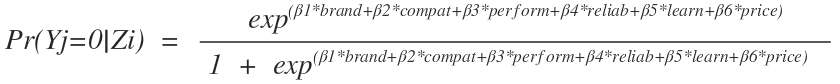
* choice : 0 or 1 for each Choice set (mandatory field)
* buy : 0 or 1 for each Choice set (optional field). Whether consumers going to buy it.

Note : There are in total 14,336 rows and 12 columns. (1 \* 4 \* 16 \* 224 = 14,336)

1. **Model Specifications :**

* Model used : Hierarchical Bayes (HB) Multinomial Logit Model (MNL)





Here,

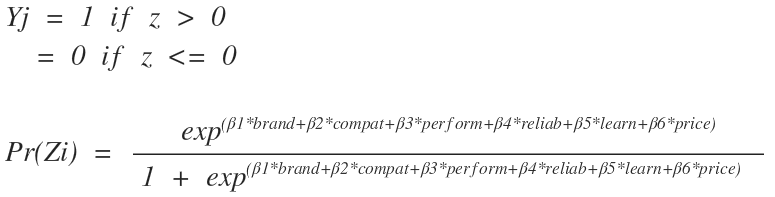
*Yj* = Consumer choice (discrete choice) for a brand, where 'j' is each consumer in the study, note there are 224 unique consumers in the study.

*Zi* = Utility of a brand, where i denotes all attributes for the specific brand chosen by the consumer “j”.

*i = ['brand', 'compat', 'perform', 'reliab', 'learn', 'price']*

The above model can also be seen a Hierarchical model with “Z” acting as a latent variable.

The Hierarchical model will look like as :



Our job here is to estimate the posterior probability distribution of the all the coefficients i.e

estimate .

As marketing data, at the individual-level is inherently discrete and lumpy. So, we will use the Bayesian analysis to this hierarchical model as the Bayes theorem takes a large sample concept like Sensitivity and Specificity i.e likelihood function distribution and transforms into a statistics so that inference can be made about a single record, i.e. posterior probability distribution.

But to compute the posterior probability distribution analytically we have to solve the complex product of likelihood function and prior probability.

The alternative and a great working tool is to used Markov Chain Monte Carlo Estimation (MCMC).

* Estimation Technique : MCMC

Motivation : If we keep drawing randomly the coefficient estimates, after a significant number of iterations the probability distribution of the coefficient estimates would come to an equilibrium. It will be independent of the state when we had started the draws. This is called Markov Chain.

If we can construct such a chain then we arbitrarily start from some point and iterate the Markov Chain many times. Eventually the draws we generate would appear as if they are coming from our target distribution. We then approximate the quantities of interest (e.g coef mean etc.) by taking the sample average of the draws (Monte Carlo components) after discarding a few initial draws.

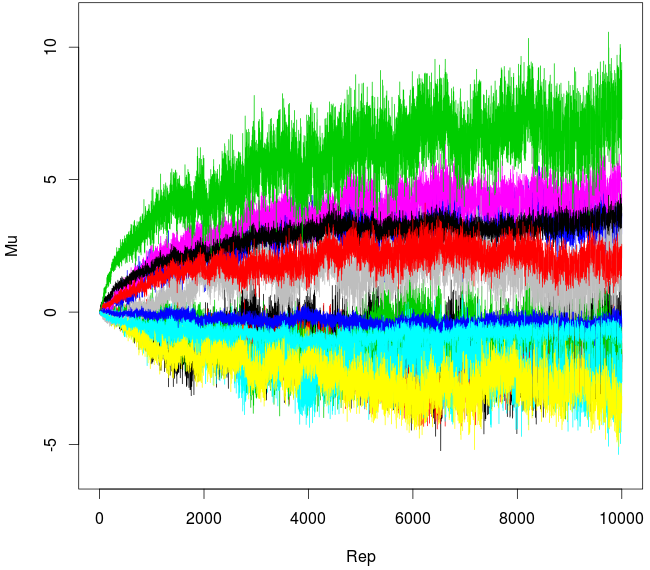
For our problem it works as below-

* Draw Z (Utility) given Y (Choice) and All Attributes (Brand, Compat, Perform, Reliab, Learn & Price)
* Draw all the Coefficients (Attribute Coefs) given the value of Z (Utility)
* Repeat 10,000 iterations so that the posterior distribution of the mean of all the coefficients go into an equilibrium.
* Take the estimate based on the last 2,000 iterations.

1. **Set-up and run MCMC estimation**

Install the R package : ChoiceModelR

* Set up the data frame as per the package specification.
* Set up different xcoding for categorical (brand : 0) and continuous (rest, 1) variables.
* Set up the constraints for all the attributes.
* Set up the run options.
* Set up the run parameter (10,000 iterations) and use parameter (last 2,000 iterations).



1. **Gather estimates from the equilibrium distributions**

We, now gather the data from HB posterior parameter distributions (betadraw.c) and create posterior.mean and posterior.sd

i.e for each consumer we take the mean of the last 2,000 coef estimates of all attributes.

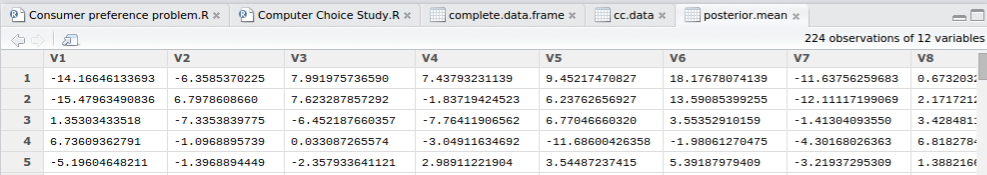
Estimate of Consumer’s preference towards all the brands and then based on the brand he/she chooses what is the relation b/w his/her Utility and the other product attributes (product heterogeneity.)

colnames(cc.priors) = c("A1B1", "A1B2", "A1B3", "A1B4", "A1B5", "A1B6", "A1B7", "A1B8",

"A2B1", "A3B1", "A4B1", "A5B1", "A6B1")

*# here A : Attributes (total 6 attributes, so A1:A6)*

*# here B : Brand (total 8 brands, so B1:B8)*



Note: The probability values which can be seen above is the **Probability of Logit**, so some of the values are negative, i.e



We’ll work with one row of respondent training data frame at a time and create choice predictions using the individual part-worths i.e

Matrix multiplication of the posterior.mean matrix and the survey result matrix to get terms like

etc.

As logit is linear in all the coefs, if we add up all the terms (**part-worths**) we get the logit probability, which is the measure of the individual consumer choice for each product profile.

e.g. for the 1st consumer and 1st product profile.

This is from the posterior mean matrix (obtained from MCMC prob estimates)-

Apple Compaq Dell Gateway HP IBM

-14.16646134 -6.3585370 7.991975737 7.43793231 9.45217471 18.17678074

Sony Sun

-11.63756260 -10.8963

Compat Perform Reliab Learn Price

6.732032e-01 9.703769e+00 2.769425e-01 -0.0003804059 -1.695590660

This is from the survey response

Brand compat perform reliab learn price

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sun | 1 | 4 | 2 | 8 | 8 |

The above values are standardized, so the matrix multiplication would result in-

-10.8963\*1 + 6.732032e-01\*(-3.5) + 9.703769e+00\*1.5 + 2.769425e-01\*0.5 + (-0.0003804059)\*3.5 + (-1.695590660)\*3.5

= Utility of the 1st product profile is -4.494285

Note: This prob value is the value of the logit and hence negative.

In the same way we calculate all the individual consumer choice for the training data, i.e.

4 product profile \* 12 choice sets \* 224 unique consumers = 10,752.

So we get 10,752 distinct training choice utility, using which we predict the individual consumer choice and compared with the true value.

Result : Training choice set sensitivity = 93.7 %

Using the training posterior distribution of the coefs estimates we predicted test set choice utility and the test set sensitivity was 52.6 %.

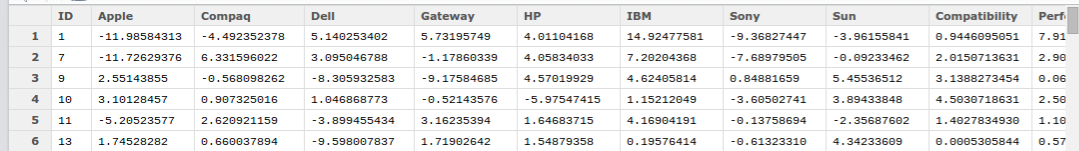
1. **To estimate Brand Loyalty, Price Sensitivity, Feature Focus and Brand Switching**

For this task we will be using the complete data from the computer choice study.

We’ll perform similar task as performed in the previous exercise with the objective to have the posterior distribution of the coefficients estimates.

For the full data set, the choice set sensitivity is 89.1 %.

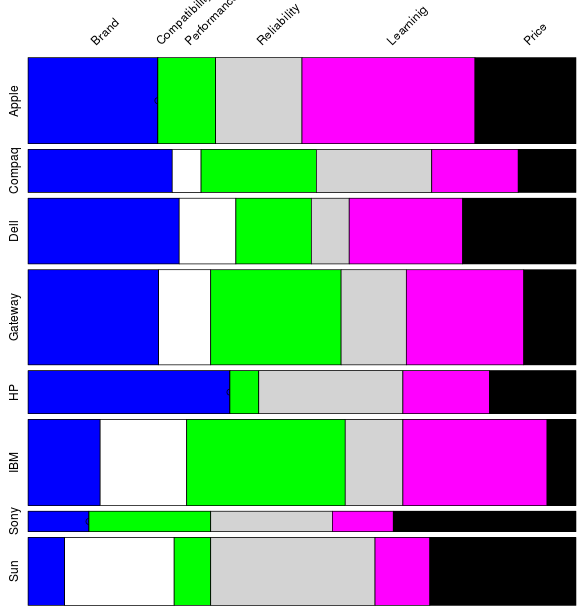
* Compute the individual-level partworths from the posterior mean distribution of the coef estimates obtained from the MCMC, i.e for each consumer we compute the worths of levels and their attributes.



* Compute Attribute Importance values for each attribute, which is the range of the levels within an attribute. Also compute the Relative Importance of each attribute.
* Figure out the top brand and the most valued attribute for each consumer.

In this part we have mined the original survey with 16 choice sets, we estimated conjoint measures (attributes and part-worths) at the individual level with an HB model and place consumers into groups based upon their revealed preferences for computer products.

Mosaic plot of joint frequencies of Top-Ranked Brands and Most Valued Attributes



The relative heights of rows correspond to the relative row frequencies of the Contingency table.

Here "Gateway", "IBM" and "Apple" have high relative row frequencies, i.e. Top-Ranked Brand

The relative width of columns corresponds to the cell frequencies within rows.

Here for "HP" brand "Brand" attribute is the most important.

For "Apple" brand, "Learning" attribute is relatively more important.

* Brand Loyalty, Price Sensitivity and Feature Focus - are the key inputs to models of Consumer Preference and Market Response.

We’ll use the **triplot / ternary plot** to understand these three relative measures (just between these three measures)

Note : Here feature importance is the most important feature after excluding Price and Brand.

So, for each consumer (id) we calculate the

Brand Loyalty

id.data$brand.loyalty[id] = id.data$brand.importance[id] / sum.importances

Price Sensitivity

id.data$price.sensitivity[id] = id.data$price.importance[id] / sum.importances

Feature Focus

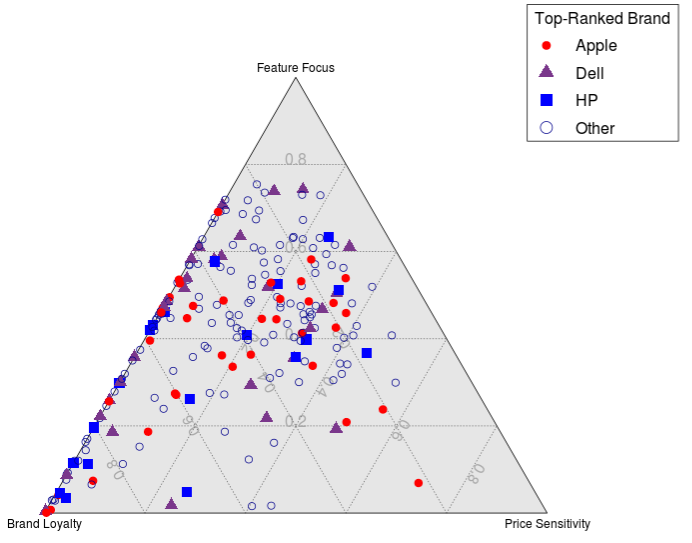
id.data$feature.focus[id] = id.data$feature.importance[id] / sum.importances

where,

sum.importances = id.data$brand.importance[id] +

id.data$price.importance[id] +

id.data$feature.importance[id]

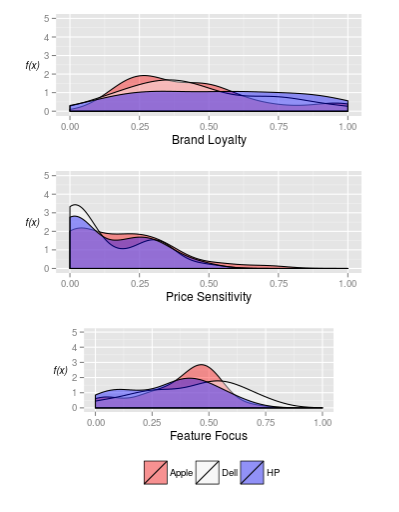


We can see a wide variability or heterogeneity in consumer preferences.

Let’s pick these three brands Apple, Dell and HP and select the subset of consumers for whom one of these brands is the top-ranked brand.

We use **density plots** to examine the "distributions of values" for Brand Loyalty, Price Sensitivity and Feature Focus across this subset of consumers.

It also shows the degree to which there is overlap in these distributions.



Note the x-axis has the attribute importances and y-axis is the function of that to the consumer i.e Consumer choice.

Let’s see the distribution first amongst this subset of consumers

Apple Compaq Dell Gateway HP IBM Sony Sun

38 0 29 0 19 0 0 0

* Consumers (29 respondents) who rated "Dell" as the highest tend to be less price-sensitive and more feature-focused than consumers who rate Apple and HP highest.
* Consumers (19 respondents) who rated "HP" as the highest tend to have higher brand-loyalty and less feature-focused than consumers rating Dell or Apple. Note: We had also seen this pattern in the Mosaic plot.
* Consumers (38 respondents) who rated "Apple" as the highest tend to be more price-sensitive (however this evidence is not very evident)

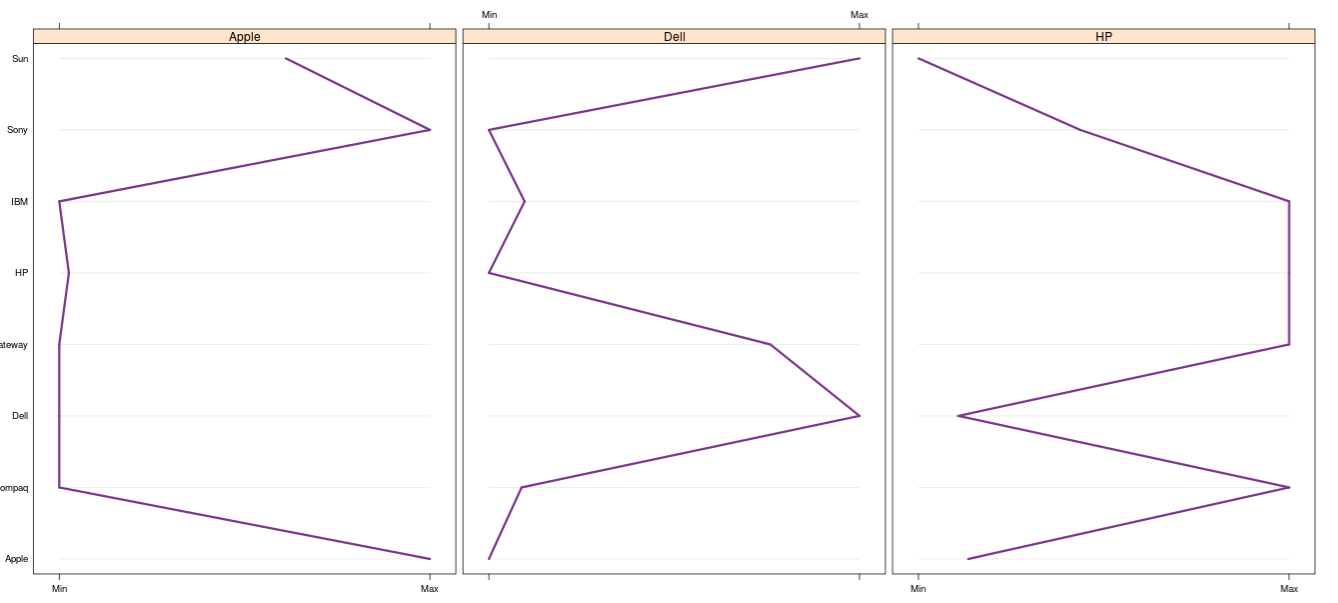
In general : In terms of the three ternary model measures, there is a considerable overlap across consumers rating Apple, Dell and HP highest.

* Brand Switching : extent to which consumers are open to switching from one brand to another.

**Parallel coordinate plots** may be used to explore the potential for brand switching.

Here we are grouping the part-worths of all the brands, grouped by the top brand either Apple, Dell or HP. Then we compute mean to have a proper pictorial view.

So, this parallel-coordinate plots show the mean part-worths for brands.



Lines farther to the right show stronger preference for a brand and stronger likelihood of switching to that brand.

* "Apple" consumers are most likely to switch to Sony or may be Sun.
* “Dell” consumers are most likely to switch to Gateway or Sun.
* “HP” consumers are most likely to switch to IBM, Gateway, or Compaq

1. **Market Simulations ("What-if analyses") using the Consumer Preference and Choice Models**.

Market simulations constructed from individual-level conjoint measures as we obtain from the Bayesian methods.

Hypothetical situation :

Suppose “Apple” computer wants to know the price to charge for their computer, given three other competitors in the market: Dell, Gateway, and HP.

In addition, suppose that they have an objective of commanding a 25% share in the market.

Constructing the Market Simulations (showing just 2 choice sets)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **setid** | **brand** | **compat** | **perform** | **reliab** | **learn** | **price** |
| **1** | 1 | Dell | 8 | 4 | 2 | 4 | 4 |
| **2** | 1 | Gateway | 6 | 2 | 1 | 2 | 2 |
| **3** | 1 | HP | 6 | 3 | 2 | 2 | 3 |
| **4** | 1 | Apple | 5 | 4 | 2 | 1 | 1 |
| **5** | 2 | Dell | 8 | 4 | 2 | 4 | 4 |
| **6** | 2 | Gateway | 6 | 2 | 1 | 2 | 2 |
| **7** | 2 | HP | 6 | 3 | 2 | 2 | 3 |
| **8** | 2 | Apple | 5 | 4 | 2 | 1 | 2 |

Note: In these two Choice Sets the only difference is the price of the Apple computer.

For the 1st one it is $1,000, for the 2nd one it is $1,250 and so on… till it is $2,750.

We do it for all the 224 unique consumers, i.e. 1\* 4 \* 8 \* 224 = 7,168 rows (product profiles) for the market simulations using these four brands.

Now, we use the posterior.mean which we computed using the MCMC method from the **Revealed Consumer Preference** in the earlier exercise to this Market Simulation problem.

So, we evaluate the Utility of each product profile in each choice set for each individual in the study. Note this is Simulated data and NOT on the respondent's survey data.

Below is a snapshot of the result-

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **id** | **setid** | **brand** | **compat** | **perform** | **reliab** | **learn** | **price** | **simulation.predicted.choice** |
| **1** | 1 | 1 | Dell | 8 | 4 | 2 | 4 | 4 | YES |
| **2** | 1 | 1 | Gateway | 6 | 2 | 1 | 2 | 2 | NO |
| **3** | 1 | 1 | HP | 6 | 3 | 2 | 2 | 3 | NO |
| **4** | 1 | 1 | Apple | 5 | 4 | 2 | 1 | 1 | NO |
| **5** | 1 | 2 | Dell | 8 | 4 | 2 | 4 | 4 | YES |
| **6** | 1 | 2 | Gateway | 6 | 2 | 1 | 2 | 2 | NO |
| **7** | 1 | 2 | HP | 6 | 3 | 2 | 2 | 3 | NO |
| **8** | 1 | 2 | Apple | 5 | 4 | 2 | 1 | 2 | NO |
| **9** | 1 | 3 | Dell | 8 | 4 | 2 | 4 | 4 | YES |
| **10** | 1 | 3 | Gateway | 6 | 2 | 1 | 2 | 2 | NO |
| **11** | 1 | 3 | HP | 6 | 3 | 2 | 2 | 3 | NO |

Below is taken from the Contingency table for the simulated.predicted.choice = “Yes”.

Note, there are 8 choice sets to given to each of the 224 consumers.

Here, the distribution of “Yes” is shown.

brand  
setid Dell Gateway HP Apple  
 1 91 20 31 82  
 2 98 22 33 71  
 3 101 22 37 64  
 4 105 25 41 53  
 5 105 28 43 48  
 6 111 28 44 41  
 7 115 29 46 34  
 8 115 30 46 33

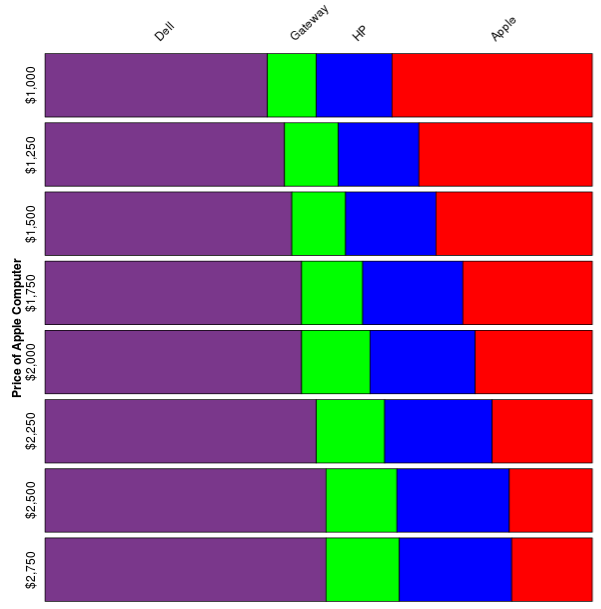
We, know that the price of Apple was set to minimum in choice set-1 and to maximum in the choice set-2. We could see that the distribution holds the law of demand, higher prices translate into lower market shares for Apple.

The below table will make it clear further-

Apple.Price Apple Dell Gateway HP  
1 1000 36.6 40.6 8.9 13.8  
2 1250 31.7 43.8 9.8 14.7  
3 1500 28.6 45.1 9.8 16.5  
4 1750 23.7 46.9 11.2 18.3  
5 2000 21.4 46.9 12.5 19.2  
6 2250 18.3 49.6 12.5 19.6  
7 2500 15.2 51.3 12.9 20.5  
8 2750 14.7 51.3 13.4 20.5

Note the objective for this Market Simulation was also to set the price of Apple computers so that it has at least 25 % of market share.

From the above table, Apple would need to set its price below $1,750 to capture a 25 % share.



**Appendix:**

This project was done on R-studio (R 3.1.3 on Ubuntu 14.04 LTS.)

The below Scripts, Utilities and data files were used and can be produced upon request.

1. Survey Data File (from the respondents) : computer\_choice\_study.csv
2. Predict Individual Consumer Choice : Computer\_Choice\_Study.R
3. Product Heterogeneity and Market Simulation : Consumer\_preference\_problem.R
4. Utility program to compute the choice utility : R\_utility\_program\_2.R
5. Utility program for Grid Graphics : R\_utility\_program\_3.R

**Reference:**

1. Hierarchical Bayes Models: A Practitioners Guide by Greg M. Allenby, Peter E. Rossi, and Robert E. McCulloch (January, 2005)
2. Modeling Techniques in Predictive Analytics with Python and R: A Guide to Data Science by Thomas W. Miller (October, 2014)