Solo 2: Discrete Choice Experiment

Name: Young, Brent

MSDS 450 Section #: 55

Quarter: Winter 2019

Introduction

Problem

The purpose of Solo 2 Assignment is to analyze and interpret Star Technologies Company (STC) choice-based conjoint (CBC) task data using Hierarchical Bayes (HB) Multinomial Logit (MNL) models that allow for the price sensitivity of respondent choices to be brand specific. These models will be estimated using Markov Chain Monte Carlo simulation, will be interpreted in regard to attributes' effects on stated preferences, and will be used to predict choices for Obee Juan's (Star's product development manager) two additional scenarios (along with calculation descriptions of attribute level partworths and attribute importance). The analysis will also take into account the possible effects of attributes on preferences of prior STC product ownership. The goal of this analysis is to help Star Technologies Company (STC) better understand what tablet configuration(s) they should go to the market with based on consumer preference. As a result, in order to accomplish this, we will explain how the data was prepared for modeling, the modeling procedures that were used, provide recommendations, and address any factors or limitations that Obee and STC should consider in using the results when deciding on what tablet to produce moving forward.

Exploratory Data Analysis

Structure and Description of Datasets & Variables

The structure of the respondent dataset includes 424 rows (respondents) and 55 columns (includes uuid variable, specific question coding, etc.) and was collected from a sample of tablet owners and likely buyers by Neverending Marketing Insights. The online questionnaire, that included a choice-based (CBC) task, consisted of questions pertaining to prior ownership of STC products, choice sets, buyer interest, and demographics. The 36 choice sets were built with the following five attributes (4 attributes with 3 discrete levels and 2 dummy variables; 1 attribute with 4 discrete levels and 3 dummy variables; blue text below: denotes reference = -1 for effects coding):

- Screen- 3 levels: 5 inch, 7 inch, 10 inch (levels: 0,1,2; *2 dummies using effects coding*)
- RAM- 3 levels: 8 Gb, 16 Gb, 32 Gb (Gb = "gigabytes") (levels: 0,1,2; *2 dummies using effects coding*)
- Processor- 3 levels: 1.5 GHz, 2 GHz, 2.5 GHz (GHz = "gigahertz") (levels: 0,1,2; *2 dummies using effects coding*)
- Price- 3 levels: $199, $299, $399 (levels: 0,1,2; *2 dummies using effects coding*)
- Brand- 4 levels: STC, Somesong, Pear, Gaggle (level codes: 0,1,2,3; *3 dummies using effects coding*)

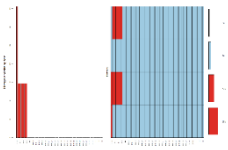
Respondents chose their most preferred alternative out of each choice set that included 3 choices (e.g., 1, 2, or 3), each described as specific combinations of the attribute levels above *(108 possible choices using Fractional Factorial Design)*. The task was also designed to allow estimation of the two-way interaction between brand and price *(3 variables that linearly multiply brand by price)* that Obee was interested in learning more about. Therefore, overall there are 14 predictor variables with 11 main-effect factors and 3 interaction variables (2 Screen dummies + 2 RAM dummies + 2 Processor dummies + 2 Pricing dummies + 3 Brand dummies + 3 Brand*Price interaction variables). The respondent data is stored in an R data file called stc-cbc-respondents-v3.RData and was accompanied with a stc-v3-datamap.txt that describes the variables in the respondent data file. The respondent data was stored in data frame called resp.data.v3, with 424 rows (respondents) and columns 4 to 39 representing the 36 choice sets with their recorded response. The data also included a task/design file called stb-dc-task-cbc-v3.csv, which provides the choice task plan that mimics the levels above. Additionally, an R file called efCode.RData (requires: dummies package) was also provided which contains R functions for effects coding that will be used to code the attributes and levels as the predictor variables for the MNL model. Lastly, in order to fulfill Obee's request of estimating preference shares for two different choice scenarios that were not part of the original choice tasks, additional files called extra-scenarios.csv and extra-scenarios-v3.csv were also provided. The files include description of alternatives in these two scenarios.

Raw Data – Preferred Choices

Appendix A shows a summary of the actual frequencies of the 36 choice sets (top), with options 1, 2, and 3 on the left. For instance, for DCM1_2, it shows how many people like options 1, 2, and 3 in choice set 1 (e.g., 316 out of 424 respondents liked option 1). Highlighted in yellow also shows the top 4 picked options. For instance, here were the top 4 picked options (A) Scr 7", RAM 16 gB, Proc 2.5 GHZ, Price $199, STC; (B) Scr 7", RAM 16 gB, Proc 2.5 GHZ, Price $199, Gaggle (C) Scr 10", RAM 8 gB, Proc 2 GHZ, Price $199, Somesong and (D) Scr 10", RAM 8 gB, Proc 2 GHZ, Price $199, STC. The results show that the options A and B are the same, but the brand is different. Additionally, options C and D show that people were willing to upgrade from Scr 7 to Scr 10, even though it meant a smaller RAM (8 gB) and processor (Proc 2 GHZ). We will obtain additional insights using post processing.

Data Preparation/Data Pre-Processing

Data Cleaning: Addressing Missing Values & Imputation



After conducting summary statistics on the dataset, the results showed that there are 1,238 NA's. The histogram on the right shows the total number of missing values for all the variables in the dataset using the VIM package. For instance, the following variables had missing values: vList3 (356, 84%) and D4r1 to D4r6 (147 each, 35%). To address the missing values, I decided to fill in the NA's for vList3 using 3 and D4r1 to D4r6 using 0. For instance, for vList3, I noticed that values 1 or 2 were only provided for "Owns or has owned STC remote or audio accessory". However, for the "Other" category (not an STC customer and has not ever been), there was no value assigned to this category. As a result, I decided to fill in the NA's for vList3 using 3 so that we could identify this group. Additionally, for D4r1 to D4r6, I noticed that respondents who were not parents (D3), either chose to skip or were authorized to skip question D4, which pertains to the age (s) of the children living in their household. Therefore, I decided to fill in the NA's for D4r1 to D4r6 using 0. After conducting the imputation, my summary statistics showed that there were no more NA's. This is important since we can now analyze the respondent population prior to modeling and get a better idea of what the population looks like.

Data Preparation for Modeling

In order to estimate Hierarchical Bayes Multinomial Logit (MNL) regression models using the rhierMnlDP() in the R package "bayesm, the data needs to be in the required form. As a result, there are 2 main steps that I conducted *(excludes loading the data, see page 2)*. First, I created an X matrix or design matrix called X.matrix, which is a matrix of categorical predictor variables representing STC's choice task attributes and levels (108 rows and 14 columns). This entailed using the function efcode.att.f (), which contains effects coding and then loading in a matrix called task.mat that describes our choice design (e.g., "screen", "RAM", "processor", "price", "brand) using the function efcode.attmat.f(), resulting in X.mat. I then created a separate matrix called X.BrandByPrice that included the interaction between brand and price. This entailed getting the vector of prices and centering it on its mean (pricevec), extracting the last 3 columns from X.mat that represent brand (X.brands), and then multiplying each column in X.brands by pricevec. I then combined X.mat and X.BrandsByPrice to get X.matrix, which we'll use for choice modeling. *Note: an indicator variable for STC product ownership was also created called zowner (1 if a respondent has ever owned an STC product, otherwise, make it equal to zero). This will be used later on for modeling purposes.* Second, I created the necessary list data structure for input to the bayesm package rhierMnlDP(). This entailed extracting the responses that STC's survey participants provided to the 36 choice questions into a data frame called ydata and converting it into a matrix. Next, a list of data lists was created called lgtdata, which contains data for each respondent and is a list with two elements (e.g., X.matrix and ydata).

Buyer Interest

Observations: Appendix 1 shows histograms and averages (denoted by a red dashed line) of buyer interest or potential activity in the next 12 months in regards to purchasing a new tablet (top left), purchasing a new smart phone (bottom left), using cloud storage for storing personal digital content (top right), and taking an online course to improve skills (bottom right). The y-axis of the histogram represent the counts, while the x-axis represents interest values of 1 to 10 (1 = Not Interested and 10 = Extremely Interested). The histograms illustrate bimodal distributions in all of the histograms with averages for new tablet (6.3), cloud storage (6.1), new smart phone (6.1), and taking an online course (4.9). Overall, the results show that there is medium to high interest for new tablet, cloud storage, and new smart phone, but low interest for taking an online course.

Respondent Demographics

Observations: Appendix 2 shows barplots of respondent demographics so that we can obtain a better idea of our survey respondent population and check for distributions prior to modeling. The barplots represent proporations of the respondent population (y-axis = percentage and x-axis = demographic/category). The top left graph shows that 84% of respondents are not an STC customer or has never been a STC customer, while 16% own/owned STC products. Furtheremore, there is a slight right skew for age range (bottom left) with 24.5% of respondents falling in-between the 25-34 age group. Additionally, the data shows that 65% of respondents are parents (top right) *(summary statistics showed somewhat equal representation of children at all age groups, see appendix 2 – bottom chart)*, 50% are equally male/female (middle left), and majority of respondents are from

California (9%), Florida (7%), New York (5%), Pennsylvania (7%), and Texas (6%) (bottom right). Furthermore, in regards to household income, majority of respondents fell in-between $50,000 to less than $75,000 (normal distribution) (middle right).

Modeling Procedure

Modeling Strategy

Multinomial logistic regression is a classification technique that is similar to logistic regression, but instead of classifying a response variable that has 2 outcomes, we are classifying a response variable that has more than two classes (e.g., 3 alternatives) (James, et al., 2010). As a result, given that the goal of this assignment is to model the response as a function of the product design attributes (e.g., utility = E(Y) = E(log odds ratio) = log(p/(1-p)) = b1X1 + b2X2 +… + b14X14, where X's are dummy variables)) and the fact that the population isn't homogenous, I decided to fit a Hierarchical Bayes Multinomial Logit Model using rhierMnlDP. This entails using the Bayesian approach, which assumes a prior distribution *(represents uncertainty about the parameter before obtaining the data; beta is not constant, but assumes a random variable)* of the parameters of the model, given the collected data, and models the posterior distribution *(revised probability distribution of beta that contains less uncertainty, after collecting the data, which is obtained using the prior distribution and the data that we conducted; likelihood function * prior distribution)* using Monte Carlo Markov Chain simulation with Generic Metropolis-Hastings Algorithm (Lynch, 2007). Generic Metropolis Hastings is an algorithm that draws samples from a probability distribution using the proportional height (Lynch, 2007). This is accomplished by 'wandering around' (aka: random walk) in such a way that the amount of time spent at each location is proportional to the height of the distribution (Lynch, 2007). After performing MCMC, its performance will then be evaluated by convergence and mixing (Lynch, 2007). The purpose of convergence is to ensure that the simulation results accurately capture the posterior density we are measuring, while the purpose of mixing is to ensure that the simulation sampled from all areas of the density. This will be accomplished using trace plots for convergence and taking every kth value or number of lags to compensate for auto correlation *(destroying the dependent samples)* (Lynch, 2007). In the end, given that it's a Bayesian approach and the fact that parameters have distributions, allows us to summarize each beta or combine all the betas using the average of betas as estimates, which allows us to derive impacts and insights of every feature.

Estimation of HB MNL Models

In order to address assignment items 2 to 5, we will estimate two HB MNL models using MCMC and the function rhierMnlDP from the bayesm package. The first model will not include the covariate indicating previous ownership of a STC product, while the second model will include this covariate. These models will be interpreted using the respondent level estimates in regard to attributes' effects on stated preferences.

Fitting HB MNL Model 1

In order to use the function rhierMnlDP (), we created inputs that consisted of data and specifications on how it should run. This entailed specifying 100,000 iterations, keeping every 5^{th} sample to adjust for autocorrelation, and creating the data list that rhierMnlDP () expects using p = 3 (p= the choice set size) and lgtdata=lgtdata called Data1. *Note: I also tried 5k, 10k, 20k, 30k, 50k, etc. iterations and with different people, but the burn in period was still difficult to determine. As a result, I ultimately decided upon 100,000 given that it contained more iterations and allowed for more data points. I also noticed that the overall beta means were similar across all the various iterations.* We then obtained the output from testrun1 and extracted betadraw out of testrun 1 called betadraw1. *Note: betadraw" is an array that has the draws (i.e. samples from marginal posterior distributions) for the regression coefficients (e.g., 424 rows or respondents, 14 columns, and 20,000 blocks).* We then utilized trace plots to help us determine the burn-in period so that we know how many points to throw away, while the remaining points converge to the truth and have stabilized. The plots in appendix B(i) show the trace plots for betadraw1 (x-axis = iteration and y-axis = sampled value or parameter/data) for respondents 1 and 2 for the various betas. The plots show that "possible" convergence has occurred at iteration 10,000 for majority of the plots, although we went with a conservative estimate (e.g., 'burn in' period to the 'right' as much as possible) given that there isn't a clear burn-in period. This is denoted by the red vertical dashed line. As a result we kept points 10001 to 20,000 and threw away the rest.

Interpretation: The table in appendix B(ii) shows the overall beta means by pooling the average value of the betas from all the respondents (e.g., betameansoverall) and the odds ratio which was computed by taking the exponent of betameansoverall. The baseline was computed using the following formula: X1*(-1) +X2*(-1), while the comparison odds ratio between attributes was computed by taking the difference between the log odds ratios of X1 & X2 and then taking the exponent, where X is the coefficient. Appendix B(ii) also includes a graph of the partworths. The following insights below have been interpreted using the following interpretation rules: log (odds ratio): beta = 0 (neutral), beta >1 (higher preference), and beta <1 (lower preference);

odds ratio = 1 (neutral), odds ratio > 1 (higher preference), and odds ratio <1 (lower preference). Therefore the model is: Y = log(p/(1-p)) = -0.19090*X1 + 0.44756*X2 + 0.09217*X3 + 0.64282*X4 + 1.02724*X5+ 1.32718*X6 + 0.29413*X7 – 2.94293*X8 – 0.20091*X9 + 0.07543*X10 – 0.32018*X11 + 0.05585*X12 + 0.03968*X13 + 0.01683*X14. The results show the following insights: The odds of preference of a screen size of 5 inches, 7 inches, and 10 inches is 0.77, 0.83 and 1.56 times greater than that of not preferring. The probability of preference for a screen size of 10 inches is 56% higher than that of non-preference. Additionally, if everything else stays the same, people prefer a screen size of 10 inches, 2 times more than a screen size of 7 and 5 inches. The odds of preference of a RAM size of 8 gB, 16 gB, and 32 gB is 0.48, 1.10 and 1.90 times greater than that of not preferring. The probability of preference for a RAM size of 16 gB and 32 gB is 10% and 90% higher than that of non-preference. Furthermore, if everything else stays the same, people prefer RAM 32 gB, 1.7 and 4 times more than RAM 16 gB and RAM 8 gB. The odds of preference of a processor speed of 1.5 GHz, 2 GHz, and 2.5 GHz is 0.09, 2.79 and 3.77 times greater than that of not preferring. Additionally, if everything else stays the same, people prefer 2.5 GHz, 1.4 and 40 times more than 2.0 GHz and 1.5 GHz. The odds of preference of a price of $199, $299, and $399 is 14.14, 1.34 and 0.05 times greater than that of not preferring. The probability of preference for a price of $299 is 34% higher than that of non-preference. Furthermore, if everything else stays the same, people prefer a price of $199, 11 and 268 times more than a price of $299 and $399. Additionally, out of all odds ratios, the results show that people prefer the $199 price the most given that it received the highest odds ratio and prefer the $399 price the least given the lowest odds ratio. Finally, the odds of preference of a brand of STC, Somesong, Pear, and Gaggle is 1.56, 0.82, 1.08, and 0.73 times greater than that of not preferring. The probability of preference for STC and Pear is 56% and 8% higher than that of non-preference. Additionally, if everything else stays the same, people prefer STC, 1.5 times or 50% more than Pear. Overall, the results show that respondents seem to have a high preference for tablets with a 10 inch screen, RAM of 32 gB, processor of 2.5 GHz, price of $199, and the STC brand. The results also show a medium preference for RAM of 16 gB, processor of 2.0 GHz, a $299 price point, and the Pear brand. Although the former is more preferred than the latter.

Price Sensitivity: In regards to the interaction interpretation between STC and Pear, if everything else stays the same, at price = -1 ($199), price = 0 ($299), and price = 1 ($399), then people prefer STC over Pear 1.69 times, 1.45 times, and 1.24 times, respectively. Therefore, STC is always preferred than Pear when price isn't mentioned, but STC is 'more' preferred to Pear when price =$199 as opposed to price = $299 or $399. Therefore, when answering the question whether which brand people prefer, it comes down to price. In this case, there would be a significant interaction between brand and price. However, in the scenario above, no matter what the price is, we always prefer STC over Pear, which means there's no significant interaction between brand and price. Additionally, when looking at the mean beta coefficients for the brand and price interactions, all the betas are close to 0 and have little variation across the different brands. Overall, this provides evidence that price sensitivity does not vary over the brands.

Fitting HB MNL Model 2 with Prior STC Ownership as a Covariate

Fitting HB MNL Model 2 involved a similar process to HB MNL Model 1, but this time the covariate called zowner was included in the model. This entailed creating a new version of Data1 called Data2 and centering the z covariate so that we could "demean" zowner and make the result a 1 column matrix. We then obtained the output from testrun2, which now includes a new component called Deltadraw. Deltadraw is a matrix with rows = saved iterations (20,000) and number of columns = number of regression predictors in the X.matrix (14). We then extracted betadraw out of testrun 2 called betadraw2. As a result we kept points 10001 to 20,000 and threw away the rest to be consistent with what was done in Model 1.

Interpretation, Model Comparison, & Covariate Impact: The table in appendix C shows the overall beta means and delta means by pooling the average value of the betas from all the respondents and the odds ratio which was computed by taking the exponent of betameansoverall and deltameansoverall. Appendix C also includes a graph of the partworths. Therefore, the model is: Y = log(p/(1-p)) = -0.18212*X1 + 0.47072*X2 + 0.09324*X3 + 0.65682*X4 + 1.06174*X5+ 1.35701*X6 + 0.32255*X7 – 3.10643*X8 – 0.21239*X9 + 0.07641*X10 – 0.34895*X11 + 0.07158*X12 + 0.03546*X13 + 0.00945*X14. The results show that when comparing the betameansoverall Log(Odds Ratio) for Model 2 with Model 1, there is virtually no change in the beta coefficients. Therefore, the beta coefficient interpretations mentioned in Model 1 are the same as Model 2. However, when analyzing the deltameansoverall for Model 2, which was used to assess the impact of ownership of a STC product has on the effects of attribute on preferences, the results show the correlation of ownership on different attributes. For instance, the following insights illustrate that ownership of a STC product impacts attribute preference. For example, the results "directionally" show that ownership of STC products negatively impacts 7 and 10 inch screen size, positively impacts RAM size of 16 gB, negatively impacts RAM size of 32 gB, positively impacts processor speed of 2 and 2.5 GHz, negatively impacts price of $299 and $399, and negatively impacts Somesong and Gaggle, but positively impacts the Pear brand. This is vastly different than what we saw in Model 1 where respondents prefer tablets with a 10 inch screen, RAM of 32 gB, processor of 2.5 GHz, price of $199, and the STC brand and a medium preference for RAM of 16 gB, processor of 2.0 GHz, a $299 price point, and the Pear brand. Overall, this illustrates that the ownership of a STC product impacts respondent's preferences. However, more research and data will

need to be collected since only 68 out of the 424 respondents have prior product ownership (16%), which was seen in our EDA. Overall, this shows that the covariate is a significant predictor of attribute preferences and possibly illustrates two things: (1) tablet attributes that STC is possibly lacking in the market and (2) Pear is a fierce rival.

Customer Choice Predictions, Goodness of Fit and Validation

Model	Accuracy	Multi Class AUC	log likelihood
HB MNL Model 1	0.8839099	0.9078	-5789.52
HB MNL Model 2 (w/ Covariate)	0.8844995	0.9081	-5789.52

Making choice predictions for our two models involved computing the means of each beta, using the means as estimates of the betas (e.g., Model is: $Y_{ij} = X_{ij}b + e_{ij}$), substituting the betas with our means and then converting the log(odds) to probabilities of preferences (e.g., $P(Yi = j) = \exp(X_{ij}b) / S_j \exp(X_{ij}b)$). To accomplish this, we calculated the posterior means for each beta and for each respondent (e.g., betameansoverall; 424 by 14). We then obtained the product of our X.matrix and each subject's vector of mean betas called xbeta (108 by 424), reorganized (putting respondents in rows with choice sets "stacked" within subjects) and exponentiated xbeta (15264 by 3), and then divided each row by its sum. This was then used to obtain the predicted choice probabilities and predicted choices. These results were then used to calculate GOF metrics such as accuracy, AUC, and log likelihood. The results show that both models performed similarly and admirably (see table above for HB MNL Model 1 and 2). See appendix D for the results, confusion matrix explanations, and the AUC curve for both models.

Predictions for 36 Choice Sets

Making predictions for the 36 choice sets involve 2 options: (A) use individual respondent's model or (B) use overall beta means. Appendix E shows a summary of predicted frequencies of the 36 choice sets (top), with options (1, 2, 3) on the left using the individual respondent's model and a summary of the predicted probabilities and frequencies of the 36 choice sets (left), with options (1, 2, 3) on the top using the overall beta means for HB MNL Model 1 and HB MNL Model 2. Highlighted in yellow shows the top 4 picked options for A and B for both HB MNL Model 1 and HB MNL Model 2. The results showed that there were some overlap and differences between using the individual respondent's model and overall beta means. Additionally, the results showed that the top 4 options for HB MNL Model 1 and HB MNL Model 2 were the same. As a result, here were the top options that appeared using options A and B (italicized text indicates that it appeared in both A & B, blue text indicates it appeared in top 4 raw data preferred choices that were identified in our EDA). *(1A) Scr 7", RAM 16 gB, Proc 2.5 GHZ, Price $199, STC*; (2A) Scr 7", RAM 16 gB, Proc 2.5 GHZ, Price $199, Gaggle; *(3A) Scr 7", RAM 16 gB, Proc 2.5 GHZ, Price $199*, Pear (4A) Scr 10", RAM 8 gB, Proc 2 GHZ, Price $199, Somesong; (5A) Scr 10", RAM 8 gB, Proc 2 GHZ, Price $199, STC; and (6A) Scr 10", RAM 16 gB, Proc 2 GHZ, Price $299, Somesong. The results show that the options 1A, 2A and 3A are the same, but the brand is different. Additionally, options 4A and 5A show that people were willing to upgrade from Scr 7 to Scr 10, even though it meant a smaller RAM (8 gB) and processor (Proc 2 GHZ). However, option 6A shows that people are willing to pay more if it meant a bigger screen, more RAM, while sacrificing a smaller processor.

Predicting Obee's Two Additional Scenarios

In order to predict Obee's two additional scenarios, we used a similar approach as the section above by using the two options: (1) the overall beta means by pooling the average value of the betas from all the respondents (e.g., betameansoverall) and (2) individual respondent's model and then using voting kind of techniques. To accomplish this, we first created an X matrix called Xextra.matrix that included the appropriate effects coding for the two additional scenarios. We then took the means of the 14 coefficients and the means across all respondents, obtained the product of our Xextra.matrix and each subjects vector of mean betas called xbetaextra2, reorganized) and exponentiated xbetaextra2, and then divided each row by its sum, which was used to obtain the predicted choice probabilities. This was then used to obtain the predicted customer choices (e.g., preference share). The matrix on the right shows the predicted preference share for Obee's two additional scenarios, with scenarios 1 and 2 in the rows and the 3 choice set alternatives in the columns. For example, for scenario 1, the results show that a 10 inch screen, 32 Gb RAM, 2 GHz processor, $199 price, and STC brand is preferred (69%) over other potential tablets with similar attributes, but with less RAM and a different brand name (Gaggle (7%): 8 GB RAM and Pear (24%): 16 GB RAM, holding all other attributes constant). When using the individual respondent's model approach, the results show similar results as the overall approach. For scenario 2, the results show that a 5 inch screen, 8 Gb RAM, 1.5 GHz processor, $199 price, and STC brand is slightly preferred (52%) over a Gaggle brand (48%) with similar attributes, but with a larger RAM size of 16 Gb. However, when using the individual respondent's model approach, the results show conflicting results and show that the Gaggle brand with similar attributes, but with a larger RAM size of 16 Gb is preferred. Overall, this provides possible evidence that customers seem to value larger RAM size, when holding all other attributes constant. The 7 inch screen, 16 Gb RAM, 1.5 GHz processor, $399 price, and Pear brand seems to hold nearly no preference at all in this scenario.

```
          [,1]      [,2]        [,3]
[1,] 0.07129054 0.6920298 0.236679700
[2,] 0.51540350 0.4815826 0.003013914
```

```
> table(extra1)
extra1
  1   2   3
 94 227 103
> table(extra2)
extra2
  1   2   3
 99 243  82
```

Partworth & Attribute Importance Calculations

Attribute level partworths for Obee's respondents can be calculated using the model results by taking the mean or median of the beta coefficients of each attribute (14 betas) across all the respondents (see Appendix B & C for graphs for Models 1 & 2). To obtain another view of the data, we could also count the number of times an attribute level was chosen relative to the number of times it was available for selection since count proportions are closely related to conjoint (Orme, 2010). Relative importance for each attribute can then be calculated by taking the difference between the highest and lowest part-worth utility values (attribute utility range) for each attribute and then dividing it by the sum of the part-worth utility values (utility range total) multiplied by 100% (Orme, 2010). As a result, when summarizing attribute importance for all the respondents combined, we can then take the individual attribute importances for each respondent and then average them (Orme, 2010). For example, X1 importance = abs(beta1)/ {abs(beta1) + abs(beta2) +...+ abs(beta14)} *(Note: It is the percentage importance of each feature).* We would then do the same for the other features (e.g., X2, X3, etc.).

STC Tablet Recommendation

According to the partworths (see Appendix B(ii) & C)), our analysis above, and extra choice sets, respondents seem to have a high preference for tablets with a 10 inch screen, RAM size of 32 gB, processor speed of 2.5 GHz, price of $199, and the STC brand. The results also show a medium preference for RAM size of 16 gB, processor speed of 2.0 GHz, a $299 price point, and the Pear brand. Although the former is more preferred than the latter. Additionally, the predictions for the 36 choice sets and Obee's two additional scenarios revealed that respondents are willing to pay more if it meant a bigger screen and more RAM. Furthermore, for one of the scenarios, the results show that a 10 inch screen, 32 Gb RAM, 2 GHz processor, $199 price, and STC brand is preferred. As a result, I would recommend that STC go to market with two possible tablet designs using a tiered strategy approach – high end and low end model with large screens: (1) High-end Model: 10 inch screen, RAM size of 32 gB, processor speed of 2.5 GHz, price of $299, and the STC brand. (2) Low-end Model: 10 inch screen, RAM size of 16 gB, processor speed of 2.0 GHz, price of $199, and the STC brand. Additionally, the results show that Pear seems to be their greatest competitor. As a result, all marketing advertisements and commercials should be made in a way that it makes it look like STC's tablet is better than Pear.

Limitations

In regards to the limitations, there are seven primary limitations that I've identified that Obee and STC should consider in using the results to decide what tablet to produce. First, we have to remember that conjoint analysis predicts consumer preference, not market share (Orme, 2010). As a result, Obee and STC need to keep in mind that a limitation of conjoint analysis is that it assumes that the consumer is fully aware/educated about available brands and that all products are equally available, which may not always be the case (Orme, 2010). Additionally, conjoint experiments hold market forces constant, doesn't take into account potential market acceptance of the product, and some consumers may not have the financial or interest to buy the product (Orme, 2010). Therefore, conjoint analysis can provide good "directional" indicators that can help increase market share, but it won't be able to tell us how much actual market share will increase, the exact price sensitivity of a product, or the exact number of products that will be purchased by consumers (Orme, 2010). Second, in regards to survey design and data collection, some of the questions were not clearly labeled or assigned a value. For instance, for vList3, values of 1 or 2 were used for "customers who owns or has owned STC remote or audio accessory". Although it is implied that 1 represents the former and 2 represents the latter, it is not entirely clear. Also, as noted in the data cleaning section of the report, the "other" category (not an STC customer and has not ever been) was not assigned a value, so the data contained NA's. Third, the sample size for customers who owns or has owned STC products was extremely small (16%). Overall, this adds uncertainty, confusion, and therefore could skew the final results. Fourth, it's important to note that some individuals did not choose choices responsibly and selected the same option for all 36 choice sets. This could skew our results (e.g., difficult to determine burn-in period in trace plots). Fifth, when making predictions using the two different options (1) overall beta means vs. (2) individual respondent's model, the results seemed to differ depending on which approach was used. STC and Obee need to keep this in mind. Sixth, it was noted in the problem description that Obee collected "qualitative research" to better understand the attributes that impact tablet preferences in which he "believed" are the most important (e.g., retail unit price, screen size, processor speed, and RAM). However, using terms like "believed" seems like he was uncertain that these were the attributes that they should focus on. This is a limitation because Obee and STC may have focused on the wrong attributes from the beginning, which could have repercussions later on when they decide which tablet to produce. Lastly, it was not entirely clear what sampling approach they used (e.g., population of interest for their tablet product). As a result, a limitation that Obee and STC should consider when deciding what tablet to produce is that

the results may not be representative of their target population (e.g., U.S. and/or international market) and helps determine how validly you can generalize the results to the market of interest.

Recommendations for Additional Research

In regards to future work, there are four primary areas that I would suggest in regards to conducting additional research due to the limitations mentioned above. First, it would be beneficial to collect additional data such as more respondents who own or currently own STC products and information such as marital status, ethnicity/race, minority status, Hispanic/Latino, and education since these variables were not present. This would provide Obee and STC with additional information to conduct further research and ensure that the sample is representative of their target population (e.g., U.S. and/or international market). Second, Obee and STC need to make corrections to the survey. For example, in vList3, customers who owns or has owned STC remote or audio accessories, should be split up into separate categories and there should be a value that is assigned to the "other" category. Furthermore, for the question pertaining to children in the household (D4), there should have been a seventh option that said, "NA: I am not a parent". This would have prevented NA's from occurring in the data. Third, Obee and STC need to incorporate warm-up tasks to educate and familiarize the respondent and identify/remove the "worst fit" people, which could help increase the validity of the results. Lastly, it would have been beneficial for Obee to supplement his initial qualitative research with "quantitative research" (e.g., cluster analysis/market segmentation/profiling and survey) when deciding which attributes to focus on prior to conducting the choice-based conjoint (CBC). This would have helped him "confidently" narrow down what is the most important attributes in a tablet beforehand and ensure that the products are designed appropriately for particular market segments. For instance, there may be other tablet attributes besides brand, price, screen, RAM, and processor that consumers may have placed a higher importance on (e.g., battery life, operating system, etc.). Additionally, let's assume cluster analysis revealed 2 different segments, by examining the partworths and importances for each group, he could gain additional insight into the product features that might appeal to each segment.

Conclusion

In conclusion, we began this analysis by first conducting EDA to get a better feel of the dataset, buyer interest, preferred choices, and respondent demographics. We then performed missing value imputation and prepared the data for modeling. We then estimated two HB MNL models (one with the STC ownership covariate and one without) using MCMC. The results showed that respondents seem to have a high preference for tablets with a 10 inch screen, RAM size of 32 gB, processor speed of 2.5 GHz, price of $199, and the STC brand. The results also show a medium preference for RAM size of 16 gB, processor speed of 2.0 GHz, a $299 price point, and the Pear brand. Additionally, from a model validation standpoint, both models performed similarly and admirably. We then made predictions for the 36 choice sets and for Obee's two additional scenarios. Our final recommendation was to create two tablets using a tiered strategy approach (high end and low end model with large screens): (1) High-end Model: 10 inch screen, RAM size of 32 gB, processor speed of 2.5 GHz, price of $299, and the STC brand. (2) Low-end Model: 10 inch screen, RAM size of 16 gB, processor speed of 2.0 GHz, price of $199, and the STC brand. Additionally, the results show that Pear seems to be their greatest competitor. As a result, all marketing advertisements and commercials should be made in a way that it makes it look like STC's tablet is better than Pear. However with that said, we also saw that that ownership of STC products negatively impacts 7 and 10 inch screen size, positively impacts RAM size of 16 gB, negatively impacts RAM size of 32 gB, positively impacts processor speed of 2 and 2.5 GHz, negatively impacts price of $299 and $399, and negatively impacts Somesong and Gaggle, but positively impacts the Pear brand. However, more research and data will need to be collected since only 68 out of the 424 respondents have prior product ownership (16%), which was seen in our EDA. We then concluded the analysis by suggesting limitations of the study and recommendations for additional research.

References

1. Lynch, Scott (2007) Introduction to Applied Bayesian Statistics and Estimation for Social Scientists. New York: Springer.

2. Orme, B. K. (2010). *Getting started with conjoint analysis: Strategies for product design and pricing research* (2nd ed.). Madison, Wisc.: Research Publishers LLC.

3. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. New York: Springer Science + Business Media.

Appendix

Appendix A

Appendix A shows a summary of the actual frequencies of the 36 choice sets (top), with options (1,2,3) on the left. Highlighted in yellow also shows the top 4 picked options.

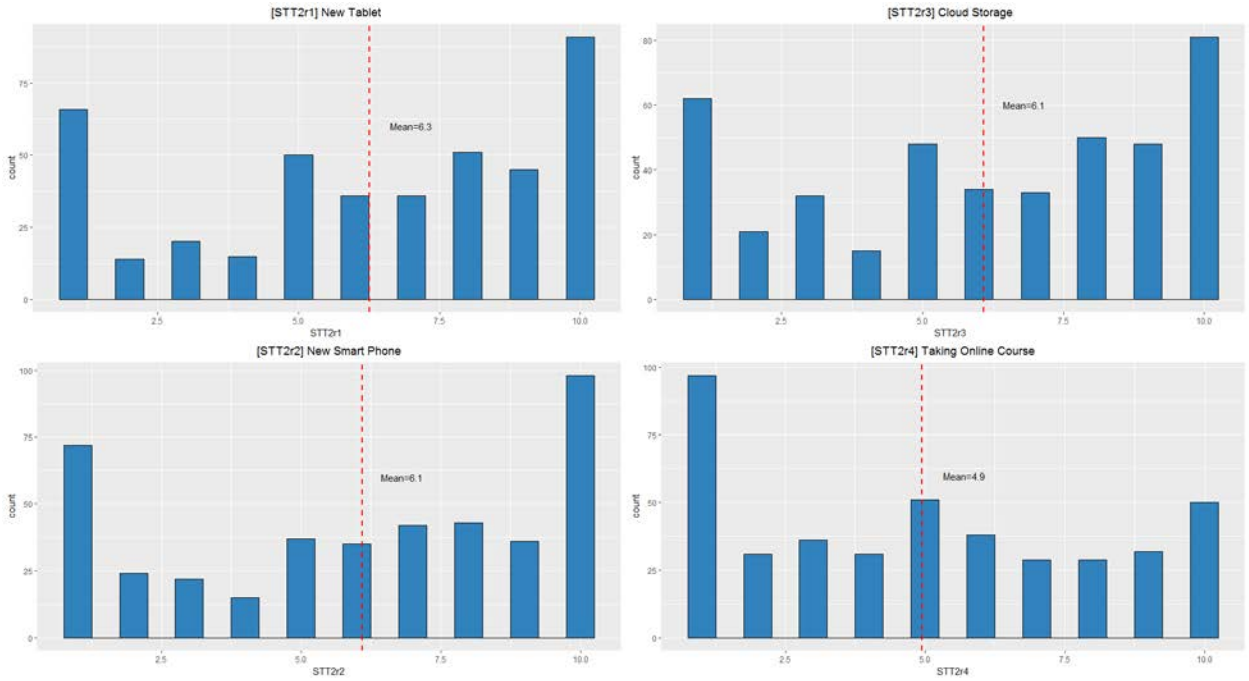
	DCM1_1	DCM1_2	DCM1_3	DCM1_4	DCM1_5	DCM1_6	DCM1_7	DCM1_8	DCM1_9	DCM1_10	DCM1_11	DCM1_12	DCM1_13	DCM1_14
[1,]	93	316	218	125	157	70	81	65	100	25	271	192	77	151
[2,]	242	45	180	28	132	83	182	85	21	307	113	217	75	148
[3,]	89	63	26	271	135	271	161	274	303	92	40	15	272	125

	DCM1_15	DCM1_16	DCM1_17	DCM1_18	DCM1_19	DCM1_20	DCM1_21	DCM1_22	DCM1_23	DCM1_24	DCM1_25	DCM1_26	DCM1_27
[1,]	18	65	22	64	41	303	225	114	183	37	71	32	109
[2,]	140	207	130	68	276	42	147	25	88	88	177	103	19
[3,]	266	152	272	292	107	79	52	285	153	299	176	289	296

	DCM1_28	DCM1_29	DCM1_30	DCM1_31	DCM1_32	DCM1_33	DCM1_34	DCM1_35	DCM1_36
[1,]	34	312	189	88	164	27	44	35	86
[2,]	308	67	209	43	121	114	207	116	36
[3,]	82	45	26	293	139	283	173	273	302

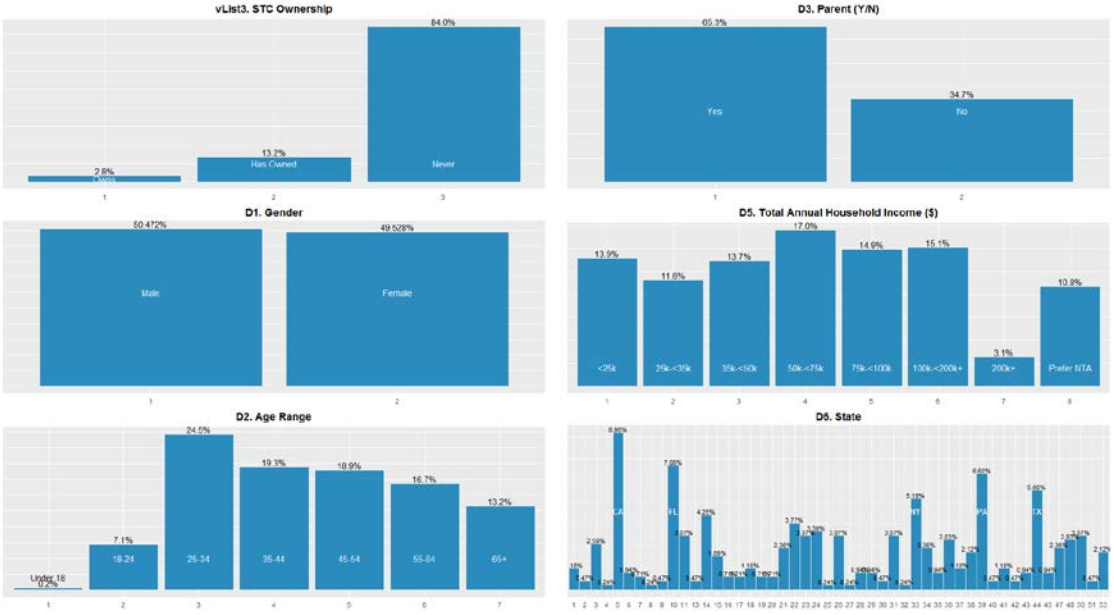
Appendix 1

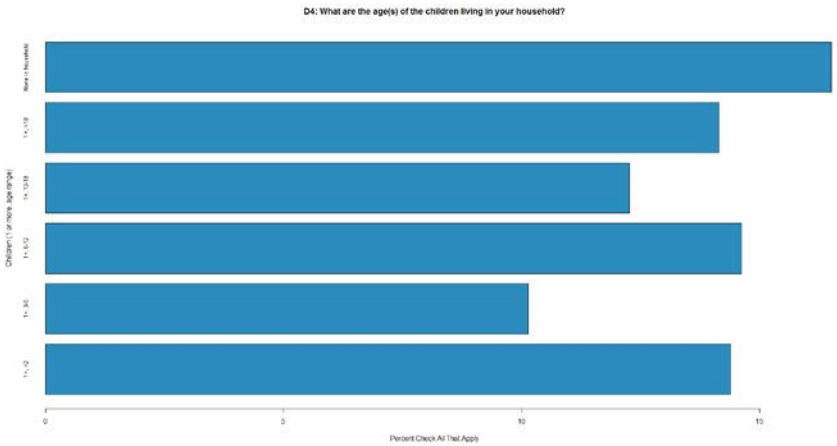
Appendix 1 shows histograms and averages (denoted by a red dashed line) of buyer interest or potential activity in the next 12 months in regards to purchasing a new tablet (top left), purchasing a new smart phone (bottom left), using cloud storage for storing personal digital content (top right), and taking an online course to improve relevant skills (bottom right). The y-axis of the histogram represent the counts, while the x-axis represents interest values of 1 to 10 (1 = Not At All Interested and 10 = Extremely Interested). The histograms illustrate bimodal distributions in all of the histograms with averages for new tablet (6.3), cloud storage (6.1), new smart phone (6.1), and taking an online course (4.9). Overall, the results show that there is medium to high interest for new tablet, cloud storage, and new smart phone, but low interest for taking an online course.



Appendix 2

Appendix 2 shows barplots of respondent demographics so that we can obtain a better idea of our survey respondent population and check for distributions prior to modeling. The barplots represent proporations of the respondent population (y-axis = percentage and x-axis = demographic/category). The top left graph shows that 84% of respondents are not an STC customer or has never been a STC customer. Furtheremore, there is a slight right skew for age range (bottom left) with 24.5% of respondents falling in-between the 25-34 age group. Additionally, the data shows that 65% of respondents are parents (top right) *(summary statistics showed somewhat equal representation of children at all age groups, see appendix 2 – bottom chart)*, 50% are equally male/female (middle left), and majority of respondents are from California (9%), Florida (7%), New York (5%), Pennsylvania (7%), and Texas (6%) (bottom right). Furthermore, in regards to household income, majority of respondents fell in-between $50,000 to less than $75,000 (normal distribution) (middle right).





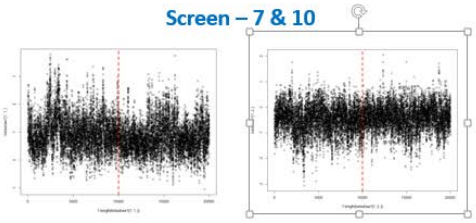
Appendix B(i)

Model 1

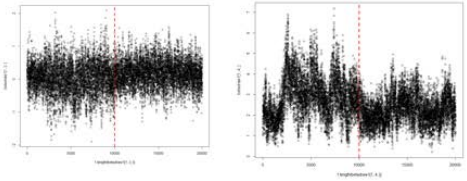
The plots below in appendix B(i) show the trace plots for betadraw1 (x-axis = iteration and y-axis = sampled value or parameter/data) for respondents 1 and 2 with the various betas. The plots show that "possible" convergence has occurred at iteration 10,000 for majority of the plots, although we went with a conservative estimate given that there isn't a clear burn-in period. This is denoted by the red vertical dashed line. As a result we kept points 10001 to 20,000 and threw away the rest.

Respondent # 1

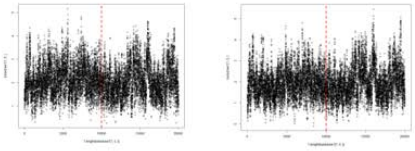
Screen – 7 & 10



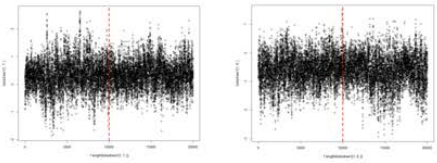
RAM – 16 & 32



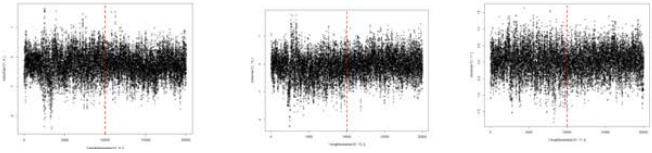
Processor 2 & 2.5



Price $299 & $399

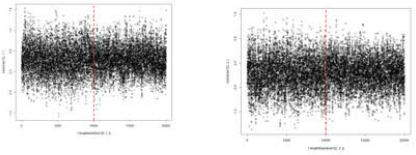


Brand – Somesong, Pear, Gaggle

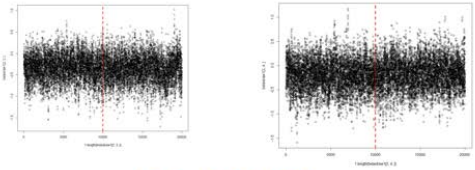


Respondent # 2

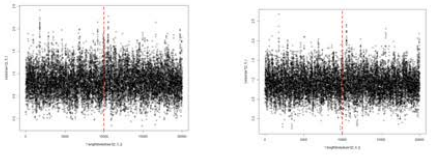
Screen – 7 & 10



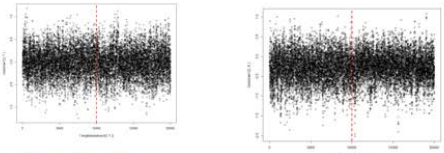
RAM – 16 & 32



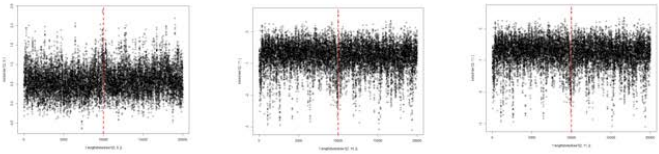
Processor 2 & 2.5



Price $299 & $399



Brand – Somesong, Pear, Gaggle



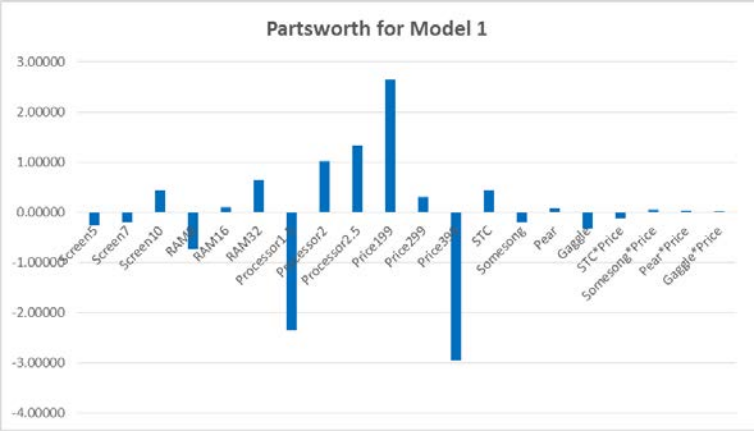
Appendix B(ii)

Model 1

The table in appendix B(ii) shows the overall beta means by pooling the average value of the betas from all the respondents (e.g., betameansoverall) and the odds ratio which was computed by taking the exponent of betameansoverall. It also shows odd ratio comparisons as well. The table in appendix B shows the overall beta means by pooling the average value of the betas from all the respondents (e.g., betameansoverall) and the odds ratio which was computed by taking the exponent of betameansoverall. The baseline was computed using the following formula: X1*(-1) +X2*(-1), while the comparison odds ratio between attributes was computed by taking the difference between X1 & X2 and then taking the exponent, where X is the coefficient. Appendix B also includes a graph of the partworths. Interpretation rules: Log(Odds Ratio): beta = 0 (neutral), beta >1 (higher preference), and beta <1 (lower preference); odds ratio = 1 (neutral), odds ratio > 1 (higher preference), and odds ratio <1 (lower preference). Therefore the model is: Y = log(p/(1-p)) = -0.19090*X1 + 0.44756*X2 + 0.09217*X3 + 0.64282*X4 + 1.02724*X5+ 1.32718*X6 + 0.29413*X7 – 2.94293*X8 – 0.20091*X9 + 0.07543*X10 – 0.32018*X11 + 0.05585*X12 + 0.03968*X13 + 0.01683*X14.

Dummy	Attribute	betameansoverall Log(OR)	exp(betameansoverall) Odds Ratio	Comparison (Difference)	Comparison Odds Ratio	Comparison Decription
Base	Screen5	-0.25666	0.77	0.70	2.02	10 inch vs. 5 inch
X1	**Screen7**	**-0.19090**	**0.83**	0.64	1.89	10 inch vs. 7 inch
X2	**Screen10**	**0.44756**	**1.56**			
Base	RAM8	-0.73500	0.48	1.38	3.97	32 RAM vs. 8 RAM
X3	**RAM16**	**0.09217**	**1.10**	0.55	1.73	32 RAM vs. 16 RAM
X4	**RAM32**	**0.64282**	**1.90**			
Base	Processor1.5	-2.35442	0.09	3.68	39.71	2.5 GHz vs. 1.5 GHz
X5	**Processor2**	**1.02724**	**2.79**	0.30	1.35	2.5 GHz vs. 2 GHz
X6	**Processor2.5**	**1.32718**	**3.77**			
Base	Price199	2.64880	14.14			
X7	**Price299**	**0.29413**	**1.34**	2.35	10.53	$199 vs. $299
X8	**Price399**	**-2.94293**	**0.05**	5.59	268.20	$199 vs $399
Base	STC	0.44567	1.56			
X9	**Somesong**	**-0.20091**	**0.82**			
X10	**Pear**	**0.07543**	**1.08**	0.37	1.45	STC vs. Pear
X11	**Gaggle**	**-0.32018**	**0.73**			
Base	STC*Price	-0.11236	0.89	0.52	1.69	STC vs. Price (-1)
X12	**Somesong*Price**	**0.05585**	**1.06**	0.37	1.45	STC vs. Price (0)
X13	**Pear*Price**	**0.03968**	**1.04**	0.22	1.24	STC vs. Price (1)
X14	**Gaggle*Price**	**0.01683**	**1.02**			

The graph below shows the overall beta means by pooling the average value of the betas from all the respondents for Model 1.



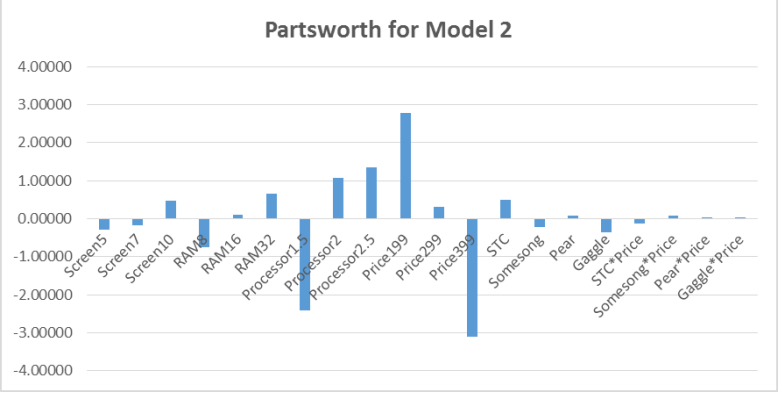
Appendix C

Model 2

The table in appendix C shows the overall beta means and delta means by pooling the average value of the betas from all the respondents and the odds ratio for betameansoverall which was computed by taking the exponent of betameansoverall. Appendix C also includes a graph of the partworths. Interpretation rules: Log(Odds Ratio): beta = 0 (neutral), beta >1 (higher preference), and beta <1 (lower preference); odds ratio = 1 (neutral), odds ratio > 1 (higher preference), and odds ratio <1 (lower preference). Therefore the model is: Y = log(p/(1-p)) = -0.18212*X1 + 0.47072*X2 + 0.09324*X3 + 0.65682*X4 + 1.06174*X5+ 1.35701*X6 + 0.32255*X7 – 3.10643*X8 – 0.21239*X9 + 0.07641*X10 – 0.34895*X11 + 0.07158*X12 + 0.03546*X13 + 0.00945*X14.

Dummy	Attribute	betameansoverall Log(OR)	exp(betameansoverall) Odds Ratio	Deltameansoverall Log(OR)
Base	Screen5	-0.28861	0.749307	
X1	**Screen7**	**-0.18212**	**0.833503**	**-0.08248**
X2	**Screen10**	**0.47072**	**1.601155**	**-0.03703**
Base	RAM8	-0.75006	0.472339	
X3	**RAM16**	**0.09324**	**1.097728**	**0.08370**
X4	**RAM32**	**0.65682**	**1.928640**	**-0.00557**
Base	Processor1.5	-2.41875	0.089033	
X5	**Processor2**	**1.06174**	**2.891394**	**0.20524**
X6	**Processor2.5**	**1.35701**	**3.884570**	**0.41529**
Base	Price199	2.78387	16.181539	
X7	**Price299**	**0.32255**	**1.380650**	**-0.02279**
X8	**Price399**	**-3.10643**	**0.044761**	**-0.60765**
Base	STC	0.48493	1.624064	
X9	**Somesong**	**-0.21239**	**0.808649**	**-0.25978**
X10	**Pear**	**0.07641**	**1.079402**	**1.09791**
X11	**Gaggle**	**-0.34895**	**0.705430**	**-0.08367**
Base	STC*Price	-0.11650	0.890030	
X12	**Somesong*Price**	**0.07158**	**1.074208**	**-0.19889**
X13	**Pear*Price**	**0.03546**	**1.036100**	**0.06122**
X14	**Gaggle*Price**	**0.00945**	**1.009498**	**0.16063**

The graph below shows the overall beta means by pooling the average value of the betas from all the respondents for Model 2.



Appendix D

HB MNL Model 1

Predicted Choice Probabilities (First 6)

The top denotes options 1 to 3, while the left side denotes the 424 respondents * 36 choice sets = 15,264.

```
             [,1]         [,2]         [,3]
[1,] 0.0000123441 0.7838459399 0.2161417160
[2,] 0.9900629802 0.0002556915 0.0096813283
[3,] 0.8585337375 0.1411377818 0.0003284806
[4,] 0.0033665514 0.0007242215 0.9959092271
[5,] 0.9452710607 0.0545941938 0.0001347455
[6,] 0.0169238430 0.7722132153 0.2108629418
```

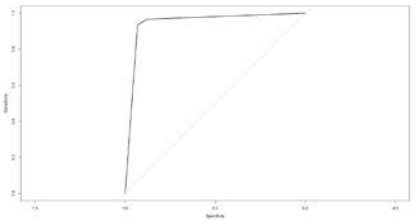
Confusion Matrix

The confusion matrix shows that the accuracy % = 13,410/15,264 = 88.4%. The diagonal is the 'correct' classification: 3686 +3753 +6053= 13,492, while the total number= 424*36 choice sets = 15,264. Predicted choice is denoted by custchoice and actual choice is denoted by ydatavec.

```
          ydatavec
custchoice    1    2    3
         1 3686  419  212
         2  227 3753  206
         3  291  417 6053
```

AUC Curve

Multi-class area under the curve: 0.9078



HB MNL Model 2 w/ Covariate

Predicted Choice Probabilities (First 6)

The top denotes options 1 to 3, while the left side denotes the 424 respondents * 36 choice sets = 15,264.

```
             [,1]         [,2]         [,3]
[1,] 2.111530e-06 0.8614611013 1.385368e-01
[2,] 9.972729e-01 0.0001100377 2.617054e-03
[3,] 9.370152e-01 0.0628287275 1.560652e-04
[4,] 4.544988e-04 0.0003061034 9.992394e-01
[5,] 9.790569e-01 0.0209054108 3.768907e-05
[6,] 3.410513e-02 0.8479026933 1.179922e-01
```

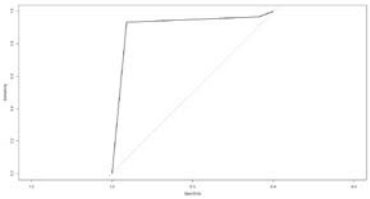
Confusion Matrix

The confusion matrix shows that the accuracy % = 13,501/15,264 = 88.4%. The diagonal is the 'correct' classification: 3691 +3766 +6044= 13,492, while the total number= 424*36 choice sets = 15,264. Predicted choice is denoted by custchoice and actual choice is denoted by ydatavec.

```
          ydatavec
custchoice    1    2    3
         1 3691  408  219
         2  226 3766  208
         3  287  415 6044
```

AUC Curve

Multi-class area under the curve: 0.9081



Appendix E

HB MNL Model 1

Prediction for 36 choice sets Using Individual Respondent's model

Shows a summary of the predicted frequencies of the 36 choice sets (top), with options (1,2,3) on the left. Highlighted in yellow also shows the top 4 picked options.

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19]
[1,]   87  344  235  110  184   68   73   59  106    32   301   196    63   178    19    49    18    57    34
[2,]  249   30  169   14  111   62  191   71   10   302    97   216    66   126   121   216   116    62   282
[3,]   88   50   20  300  129  294  160  294  308    90    26    12   295   120   284   159   290   305   108
     [,20] [,21] [,22] [,23] [,24] [,25] [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36]
[1,]   328   259   106   210    29    57    34    95    35   328   216    78   177    21    33    25    73
[2,]    28   117    12    56    79   180    81     8   308    53   186    30   102    92   218    99    26
[3,]    68    48   306   158   316   187   309   321    81    43    22   316   145   311   173   300   325
```

Predictions based on overall beta means, for 36 choice sets

Shows a summary of the predicted probabilities and frequencies of the 36 choice sets (left), with options (1,2,3) on the top. Highlighted in yellow also shows the top 4 picked options (frequencies).

```
          [,1]         [,2]         [,3]
 [1,] 0.020226846 0.9643478090 1.542534e-02
 [2,] 0.994801500 0.0044759618 7.225383e-04
 [3,] 0.758720300 0.2412608984 1.880204e-05
 [4,] 0.084058186 0.0039035024 9.120383e-01
 [5,] 0.673587581 0.1848477990 1.415646e-01
 [6,] 0.006297929 0.0496693487 9.440327e-01
 [7,] 0.158337973 0.4484662840 3.931957e-01
 [8,] 0.005313403 0.0361641630 9.585224e-01
 [9,] 0.055552028 0.0004267363 9.440212e-01
[10,] 0.003763769 0.9919699122 4.266319e-03
[11,] 0.974704382 0.0242433633 1.052255e-03
[12,] 0.362599951 0.6373866934 1.335595e-05
[13,] 0.064691026 0.0123346962 9.229743e-01
[14,] 0.416126481 0.4688733442 1.150002e-01
[15,] 0.004338608 0.1404918491 8.551695e-01
[16,] 0.081069157 0.7002454772 2.186854e-01
[17,] 0.004593104 0.0953370974 9.000698e-01
[18,] 0.051326739 0.0012024118 9.474708e-01
[19,] 0.011149273 0.9710213098 1.782942e-02
[20,] 0.990351820 0.0081398448 1.508335e-03
[21,] 0.632540700 0.3674264302 3.286969e-05
[22,] 0.036279809 0.0027487241 9.609715e-01
[23,] 0.509998146 0.2283392319 2.616626e-01
[24,] 0.002632963 0.0338787325 9.634883e-01
```

```
[25,] 0.104167129 0.4299145890 4.659183e-01
[26,] 0.002977573 0.0295307723 9.674917e-01
[27,] 0.031626203 0.0003540083 9.680198e-01
[28,] 0.005819433 0.9858069985 8.373569e-03
[29,] 0.982939141 0.0157138380 1.347021e-03
[30,] 0.469511972 0.5304660746 2.195296e-05
[31,] 0.053860501 0.0079902919 9.381492e-01
[32,] 0.451658656 0.3959573141 1.523840e-01
[33,] 0.003747717 0.0944224713 9.018298e-01
[34,] 0.078766272 0.6407888425 2.804449e-01
[35,] 0.003581664 0.0700197335 9.263986e-01
[36,] 0.039390232 0.0008691167 9.597407e-01
```

```
     [,1] [,2] [,3]
 [1,]    9  409    7
 [2,]  422    2    0
 [3,]  322  102    0
 [4,]   36    2  387
 [5,]  286   78   60
 [6,]    3   21  400
 [7,]   67  190  167
 [8,]    2   15  406
 [9,]   24    0  400
[10,]    2  421    2
[11,]  413   10    0
[12,]  154  270    0
[13,]   27    5  391
[14,]  176  199   49
[15,]    2   60  363
[16,]   34  297   93
[17,]    2   40  382
[18,]   22    1  402
[19,]    5  412    8
[20,]  420    3    1
[21,]  268  156    0
[22,]   15    1  407
[23,]  216   97  111
[24,]    1   14  409
[25,]   44  182  198
[26,]    1   13  410
[27,]   13    0  410
[28,]    2  418    4
[29,]  417    7    1
[30,]  199  225    0
[31,]   23    3  398
[32,]  192  168   65
[33,]    2   40  382
[34,]   33  272  119
[35,]    2   30  393
[36,]   17    0  407
```

HB MNL Model 2 w/ Covariate

Prediction for 36 choice sets Using Individual Respondent's model

Shows a summary of the predicted frequencies of the 36 choice sets (top), with options (1,2,3) on the left. Highlighted in yellow also shows the top 4 picked options.

```
    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19]
[1,]   88  341  235  111  183   66   75   59  105    33   301   194    63   174    17    47    19    57    37
[2,]  247   30  170   14  111   64  189   69   10   301    97   219    70   129   120   218   116    62   278
[3,]   89   53   19  299  130  294  160  296  309    90    26    11   291   121   287   159   289   305   109
     [,20] [,21] [,22] [,23] [,24] [,25] [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36]
[1,]   327   263   109   207    30    54    35    96    37   327   215    81   177    23    32    25    75
[2,]    30   113    12    60    78   185    80     8   306    54   187    31   102    91   221   101    27
[3,]    67    48   303   157   316   185   309   320    81    43    22   312   145   310   171   298   322
```

Predictions based on overall beta means, for 36 choice sets

Shows a summary of the predicted probabilities and frequencies of the 36 choice sets (left), with options (1,2,3) on the top. Highlighted in yellow also shows the top 4 picked options (frequencies).

```
         [,1]         [,2]         [,3]
 [1,] 0.020226846 0.9643478090 1.542534e-02
 [2,] 0.994801500 0.0044759618 7.225383e-04
 [3,] 0.758720300 0.2412608984 1.880204e-05
 [4,] 0.084058186 0.0039035024 9.120383e-01
 [5,] 0.673587581 0.1848477990 1.415646e-01
 [6,] 0.006297929 0.0496693487 9.440327e-01
 [7,] 0.158337973 0.4484662840 3.931957e-01
 [8,] 0.005313403 0.0361641630 9.585224e-01
 [9,] 0.055552028 0.0004267363 9.440212e-01
[10,] 0.003763769 0.9919699122 4.266319e-03
[11,] 0.974704382 0.0242433633 1.052255e-03
[12,] 0.362599951 0.6373866934 1.335595e-05
[13,] 0.064691026 0.0123346962 9.229743e-01
[14,] 0.416126481 0.4688733442 1.150002e-01
[15,] 0.004338608 0.1404918491 8.551695e-01
[16,] 0.081069157 0.7002454772 2.186854e-01
[17,] 0.004593104 0.0953370974 9.000698e-01
[18,] 0.051326739 0.0012024118 9.474708e-01
[19,] 0.011149273 0.9710213098 1.782942e-02
[20,] 0.990351820 0.0081398448 1.508335e-03
[21,] 0.632540700 0.3674264302 3.286969e-05
[22,] 0.036279809 0.0027487241 9.609715e-01
[23,] 0.509998146 0.2283392319 2.616626e-01
[24,] 0.002632963 0.0338787325 9.634883e-01
[25,] 0.104167129 0.4299145890 4.659183e-01
[26,] 0.002977573 0.0295307723 9.674917e-01
[27,] 0.031626203 0.0003540083 9.680198e-01
[28,] 0.005819433 0.9858069985 8.373569e-03
[29,] 0.982939141 0.0157138380 1.347021e-03
[30,] 0.469511972 0.5304660746 2.195296e-05
[31,] 0.053860501 0.0079902919 9.381492e-01
[32,] 0.451658656 0.3959573141 1.523840e-01
[33,] 0.003747717 0.0944224713 9.018298e-01
[34,] 0.078766272 0.6407888425 2.804449e-01
[35,] 0.003581664 0.0700197335 9.263986e-01
[36,] 0.039390232 0.0008691167 9.597407e-01
```

```
      [,1] [,2] [,3]
 [1,]    9  409    7
 [2,]  422    2    0
 [3,]  322  102    0
 [4,]   36    2  387
 [5,]  286   78   60
 [6,]    3   21  400
 [7,]   67  190  167
 [8,]    2   15  406
 [9,]   24    0  400
[10,]    2  421    2
[11,]  413   10    0
[12,]  154  270    0
[13,]   27    5  391
[14,]  176  199   49
[15,]    2   60  363
[16,]   34  297   93
[17,]    2   40  382
[18,]   22    1  402
[19,]    5  412    8
[20,]  420    3    1
[21,]  268  156    0
[22,]   15    1  407
[23,]  216   97  111
[24,]    1   14  409
[25,]   44  182  198
[26,]    1   13  410
[27,]   13    0  410
[28,]    2  418    4
[29,]  417    7    1
[30,]  199  225    0
[31,]   23    3  398
[32,]  192  168   65
[33,]    2   40  382
[34,]   33  272  119
[35,]    2   30  393
[36,]   17    0  407
```

R Code

```
#Solo 2 Assignment Code
#MSDS 450, Winter 2019

#Load Libraries
require(useful)
require(Hmisc)
library(HSAUR)
library(MVA)
library(HSAUR2)
library(fpc)
library(mclust)
library(lattice)
library(car)
library(proxy)
library(VIM) #Missingness Map
library(mice)
library(plyr)
library(likert) #Visualize Likert Scale Data
require(ggplot2)
library(reshape)
library(dummies) #For functions
library(bayesm) #Hierarchical Bayes Multinomial Logit Regression
library(knitr)
library(scales)

# Multiple plot function

multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
 library(grid)

 # Make a list from the ... arguments and plotlist
 plots <- c(list(...), plotlist)

 numPlots = length(plots)

 # If layout is NULL, then use 'cols' to determine layout
 if (is.null(layout)) {
  # Make the panel
  # ncol: Number of columns of plots
  # nrow: Number of rows needed, calculated from # of cols
  layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                   ncol = cols, nrow = ceiling(numPlots/cols))
 }

 if (numPlots==1) {
  print(plots[[1]])

 } else {
  # Set up the page
  grid.newpage()
  pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

  # Make each plot, in the correct location
  for (i in 1:numPlots) {
```

```
    # Get the i,j matrix positions of the regions that contain this subplot
    matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

    print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                    layout.pos.col = matchidx$col))
  }
 }
}

###### Load Data #####
setwd("~/R/MSDS 450/Solo 2")

# Load the dataset
load("stc-cbc-respondents-v3.RData") #Load respondent data
ls() #Load dataframe
str(resp.data.v3)
head(resp.data.v3)
#resp.data.v3

#Load dataframes and csv provided
resp.data.v3 <- resp.data.v3
resp.data.eda <- resp.data.v3 #For EDA
taskV3 <- taskV3 <- read.csv("stc-dc-task-cbc -v3.csv", sep="\t")
str(taskV3) #Levels indicated by 0,1,2,3 according to prob description; 108 rows, 7 columns
head(taskV3)
scenarios.data <- read.csv("extra-scenarios-v3.csv") #Descriptions of the two additional
ex_scen <- read.csv("extra-scenarios.csv")

#choice scenarios that you'll analyze after estimating your MNL model

# Load functions
load("efCode.RData")
ls() #Load dataframe
#Includes a couple of R functions that you'll use to code your attributes and levels
#as the predictor variables for MNL model.

str(efcode.att.f)
str(efcode.attmat.f)
str(resp.data.v3)
apply(resp.data.v3[4:39], 2, function(x){tabulate(na.omit(x))}) #summarizes info

######################################## Exploratory Data Analysis ########################################

#convert to Factors
resp.data.eda$D1 <- as.factor(resp.data.eda$D1)
resp.data.eda$D2 <- as.factor(resp.data.eda$D2)
resp.data.eda$D3 <- as.factor(resp.data.eda$D3)

resp.data.eda$D5 <- as.factor(resp.data.eda$D5)
resp.data.eda$D6 <- as.factor(resp.data.eda$D6)

#resp.data.eda$STT2r1 <- as.factor(resp.data.eda$STT2r1)
#resp.data.eda$STT2r2 <- as.factor(resp.data.eda$STT2r4)
#resp.data.eda$STT2r3 <- as.factor(resp.data.eda$STT2r4)
#resp.data.eda$STT2r4 <- as.factor(resp.data.eda$STT2r4)
```

```
#resp.data.eda$D4r1 <- as.factor(resp.data.eda$D4r1)
#resp.data.eda$D4r2 <- as.factor(resp.data.eda$D4r2)
#resp.data.eda$D4r3 <- as.factor(resp.data.eda$D4r3)
#resp.data.eda$D4r4 <- as.factor(resp.data.eda$D4r4)
#resp.data.eda$D4r5 <- as.factor(resp.data.eda$D4r5)
#resp.data.eda$D4r6 <- as.factor(resp.data.eda$D4r6)

#Descriptive Statistics
str(resp.data.eda)
head(resp.data.eda)
tail(resp.data.eda)
summary(resp.data.eda)
dim(resp.data.eda) #424  55
describe(resp.data.eda)

#Check for Missingness
sum(is.na(resp.data.eda)) #1238
sapply(resp.data.eda, function(x) sum(is.na(x)))
aggr_plot <- aggr(resp.data.eda, col=c('#9ecae1','#de2d26'), numbers=TRUE,prop=FALSE, sortVars=TRUE, labels=names(resp.dat
a.eda), cex.axis=.5, gap=2, ylab=c("Histogram of missing data","Pattern"))
#D4r1 to D4r6; vList3

#Check missing data percentage
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(resp.data.eda,2,pMiss)
#D4r1 to D4r6 (147 Each); vList3 (356)

#Conduct Imputation
#Impute D4r1 to D4r6 with 0 and vList3 with 3 which equates to "other"
resp.data.eda$D4r1[is.na(resp.data.eda$D4r1)] <- 0
resp.data.eda$D4r2[is.na(resp.data.eda$D4r2)] <- 0
resp.data.eda$D4r3[is.na(resp.data.eda$D4r3)] <- 0
resp.data.eda$D4r4[is.na(resp.data.eda$D4r4)] <- 0
resp.data.eda$D4r5[is.na(resp.data.eda$D4r5)] <- 0
resp.data.eda$D4r6[is.na(resp.data.eda$D4r6)] <- 0 #No children in household

#Impute vList3 with newly created level of 3 which equates to "other"
resp.data.eda$vList3[is.na(resp.data.eda$vList3)] <- 3

#Check N/A values have been removed
summary(resp.data.eda)
sapply(resp.data.eda, function(x) sum(is.na(x)))
sum(is.na(resp.data.eda))
aggr_plot <- aggr(resp.data.eda, col=c('#9ecae1','#de2d26'), numbers=TRUE,prop=FALSE, sortVars=TRUE, labels=names(resp.dat
a.eda), cex.axis=.5, gap=2, ylab=c("Histogram of missing data","Pattern"))

#Buyer Interest
#STT2: How interested are you in purchasing or doing each of the following in the next 12 months?
#Values: 1-10

#Histogram & Density of [STT2r1] Purchasing a new tablet
a1<-ggplot(resp.data.eda, aes(x=STT2r1)) +
  geom_histogram(aes(y=..count..),
                 binwidth=.5,
```

```
            colour="black", fill="#3182bd") +
  geom_vline(aes(xintercept=mean(STT2r1, na.rm=T)),
          color="red", linetype="dashed", size=1) +
  annotate("text", x = 7, y = 60, label = "Mean=6.3",color="black")+
  ggtitle("[STT2r1] New Tablet") +
  theme(plot.title = element_text(hjust = 0.5))
a1

#Histogram of [STT2r2] Purchasing a new smart phone
a2<-ggplot(resp.data.eda, aes(x=STT2r2)) +
  geom_histogram(aes(y=..count..),
            binwidth=.5,
            colour="black", fill="#3182bd") +
  geom_vline(aes(xintercept=mean(STT2r2, na.rm=T)),
          color="red", linetype="dashed", size=1) +
  annotate("text", x = 6.8, y = 60, label = "Mean=6.1",color="black")+
  ggtitle("[STT2r2] New Smart Phone") +
  theme(plot.title = element_text(hjust = 0.5))
a2

#Histogram of [STT2r3] Using cloud storage for storing your personal digital content
a3<-ggplot(resp.data.eda, aes(x=STT2r3)) +
  geom_histogram(aes(y=..count..),
            binwidth=.5,
            colour="black", fill="#3182bd") +
  geom_vline(aes(xintercept=mean(STT2r3, na.rm=T)),
          color="red", linetype="dashed", size=1) +
  annotate("text", x = 6.8, y = 60, label = "Mean=6.1",color="black")+
  ggtitle("[STT2r3] Cloud Storage") +
  theme(plot.title = element_text(hjust = 0.5))
a3

#Histogram of [STT2r4] Taking an online course to improve relevant skills
a4<-ggplot(resp.data.eda, aes(x=STT2r4)) +
  geom_histogram(aes(y=..count..),
            binwidth=.5,
            colour="black", fill="#3182bd") +
  geom_vline(aes(xintercept=mean(STT2r4, na.rm=T)),
          color="red", linetype="dashed", size=1) +
  annotate("text", x = 5.7, y = 60, label = "Mean=4.9",color="black")+
  ggtitle("[STT2r4] Taking Online Course") +
  theme(plot.title = element_text(hjust = 0.5))
a4

multiplot(a1, a2, a3, a4, cols=2)

#Respondent Demographics

#vList3. STC Ownership?
vList3c<-ggplot(resp.data.eda) +
  geom_bar( aes(vList3),colour="#2b8cbe",fill="#2b8cbe") +
  ggtitle("vList3. STC Ownership" ) +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5),axis.title.x=element_blank())+
  annotate("text", x = 1, y = 20, label = "Owns",color="black") +
  annotate("text", x = 2, y = 40, label = "Has Owned",color="white")+
```

```
  annotate("text", x = 3, y = 40, label = "Never",color="white")
vList3c #Count

vList3<-ggplot(resp.data.eda, aes(x = as.factor(vList3))) +
  geom_bar(aes(y = (..count..)/sum(..count..)),colour="#2b8cbe",fill="#2b8cbe") +
  geom_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +
  scale_y_continuous(labels = percent) +
  labs( title="vList3. STC Ownership", y = "", x = "")+
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5),axis.title.x=element_blank(),
        axis.text.y=element_blank(),axis.ticks=element_blank())+
  annotate("text", x = 1, y = 0.015, label = "Owns",color="white") +
  annotate("text", x = 2, y = 0.1, label = "Has Owned",color="white")+
  annotate("text", x = 3, y = 0.1, label = "Never",color="white")
vList3 #Percentage

#D1. What is your gender?
D1c<-ggplot(resp.data.eda) +
  geom_bar( aes(D1),colour="#2b8cbe",fill="#2b8cbe") +
  ggtitle("D1. Gender" ) +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5),axis.title.x=element_blank())+
  annotate("text", x = 1, y = 100, label = "Male",color="white") +
  annotate("text", x = 2, y = 100, label = "Female",color="white")
D1c #Count


D1<-ggplot(resp.data.eda, aes(x = as.factor(D1))) +
  geom_bar(aes(y = (..count..)/sum(..count..)),colour="#2b8cbe",fill="#2b8cbe") +
  geom_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +
  scale_y_continuous(labels = percent) +
  labs( title="D1. Gender", y = "", x = "")+
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5),axis.title.x=element_blank(),
        axis.text.y=element_blank(),axis.ticks=element_blank())+
  annotate("text", x = 1, y = 0.3, label = "Male",color="white") +
  annotate("text", x = 2, y = 0.3, label = "Female",color="white")
D1 #Percentage

#D2. What is your age range?
D2c<-ggplot(resp.data.eda) +
  geom_bar(aes(D2),colour="#2b8cbe",fill="#2b8cbe") +
  ggtitle("D2. Age Range" ) +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5),axis.title.x=element_blank())+
  annotate("text", x = 1, y = 5, label = "Under 18",color="black") +
  annotate("text", x = 2, y = 20, label = "18-24",color="white")+
  annotate("text", x = 3, y = 20, label = "25-34",color="white")+
  annotate("text", x = 4, y = 20, label = "35-44",color="white")+
  annotate("text", x = 5, y = 20, label = "45-54",color="white")+
  annotate("text", x = 6, y = 20, label = "55-64",color="white")+
  annotate("text", x = 7, y = 20, label = "65+",color="white")
D2c #Count

D2<-ggplot(resp.data.eda, aes(x = as.factor(D2))) +
  geom_bar(aes(y = (..count..)/sum(..count..)),colour="#2b8cbe",fill="#2b8cbe") +
  geom_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +
  scale_y_continuous(labels = percent) +
  labs( title="D2. Age Range", y = "", x = "")+
```

```
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5),axis.title.x=element_blank(),
        axis.text.y=element_blank(),axis.ticks=element_blank())+
  annotate("text", x = 1, y = 0.02, label = "Under 18",color="black") +
  annotate("text", x = 2, y = 0.05, label = "18-24",color="white")+
  annotate("text", x = 3, y = 0.05, label = "25-34",color="white")+
  annotate("text", x = 4, y = 0.05, label = "35-44",color="white")+
  annotate("text", x = 5, y = 0.05, label = "45-54",color="white")+
  annotate("text", x = 6, y = 0.05, label = "55-64",color="white")+
  annotate("text", x = 7, y = 0.05, label = "65+",color="white")
D2 #Percentage


#D3. Are you a parent?
D3c<-ggplot(resp.data.eda) +
  geom_bar(aes(D3),colour="#2b8cbe",fill="#2b8cbe") +
  ggtitle("D3. Parent (Y/N)" ) +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5),axis.title.x=element_blank())+
  annotate("text", x = 1, y = 100, label = "Yes",color="white") +
  annotate("text", x = 2, y = 100, label = "No",color="white")
D3c #Count

D3<-ggplot(resp.data.eda, aes(x = as.factor(D3))) +
  geom_bar(aes(y = (..count..)/sum(..count..)),colour="#2b8cbe",fill="#2b8cbe") +
  geom_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +
  scale_y_continuous(labels = percent) +
  labs( title="D3. Parent (Y/N)", y = "", x = "")+
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5),axis.title.x=element_blank(),
        axis.text.y=element_blank(),axis.ticks=element_blank())+
  annotate("text", x = 1, y = 0.3, label = "Yes",color="white") +
  annotate("text", x = 2, y = 0.3, label = "No",color="white")
D3 #Percentage

#D5. Total annual household income from all sources and before taxes for 2010?
D5c<-ggplot(resp.data.eda) +
  geom_bar(aes(D5),colour="#2b8cbe",fill="#2b8cbe") +
  ggtitle("D5. Total Annual Household Income ($)" ) +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5),axis.title.x=element_blank())+
  annotate("text", x = 1, y = 10, label = "<25k",color="white") +
  annotate("text", x = 2, y = 10, label = "25k-<35k",color="white")+
  annotate("text", x = 3, y = 10, label = "35k-<50k",color="white")+
  annotate("text", x = 4, y = 10, label = "50k-<75k",color="white")+
  annotate("text", x = 5, y = 10, label = "75k-<100k",color="white")+
  annotate("text", x = 6, y = 10, label = "100k-<200k+",color="white")+
  annotate("text", x = 7, y = 10, label = "200k+",color="white")+
  annotate("text", x = 8, y = 10, label = "Prefer NTA",color="white")
D5c #Count

D5<-ggplot(resp.data.eda, aes(x = as.factor(D5))) +
  geom_bar(aes(y = (..count..)/sum(..count..)),colour="#2b8cbe",fill="#2b8cbe") +
  geom_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25) +
  scale_y_continuous(labels = percent) +
  labs( title="D5. Total Annual Household Income ($)", y = "", x = "")+
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5),axis.title.x=element_blank(),
        axis.text.y=element_blank(),axis.ticks=element_blank())+
  annotate("text", x = 1, y = 0.02, label = "<25k",color="white") +
```

```
  annotate("text", x = 2, y = 0.02, label = "25k-<35k",color="white")+
  annotate("text", x = 3, y = 0.02, label = "35k-<50k",color="white")+
  annotate("text", x = 4, y = 0.02, label = "50k-<75k",color="white")+
  annotate("text", x = 5, y = 0.02, label = "75k-<100k",color="white")+
  annotate("text", x = 6, y = 0.02, label = "100k-<200k+",color="white")+
  annotate("text", x = 7, y = 0.02, label = "200k+",color="white")+
  annotate("text", x = 8, y = 0.02, label = "Prefer NTA",color="white")
D5 #Percentage

#D6. In which state do you live?
D6c<-ggplot(resp.data.eda) +
  geom_bar(aes(D6),colour="#2b8cbe",fill="#2b8cbe") +
  ggtitle("D6. State" ) +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5),axis.title.x=element_blank())+
  annotate("text", x = 5, y = 39, label = "CA",color="black") +
  annotate("text", x = 10, y = 31, label = "FL",color="black")+
  annotate("text", x = 32, y = 24, label = "NY",color="black")+
  annotate("text", x = 38, y = 29, label = "PA",color="black")+
  annotate("text", x = 43, y = 25, label = "TX",color="black")
D6c

D6<-ggplot(resp.data.eda, aes(x = as.factor(D6))) +
  geom_bar(aes(y = (..count..)/sum(..count..)),colour="#2b8cbe",fill="#2b8cbe") +
  geom_text(aes(y = ((..count..)/sum(..count..)), label = scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -0.25, siz
e=3) +
  scale_y_continuous(labels = percent) +
  labs( title="D6. State", y = "", x = "")+
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5),axis.title.x=element_blank(),
        axis.text.y=element_blank(),axis.ticks=element_blank())+
  annotate("text", x = 5, y = 0.045, label = "CA",color="white",fontface=2, size=3.5) +
  annotate("text", x = 10, y = 0.045, label = "FL",color="white",fontface=2, size=3.5)+
  annotate("text", x = 32, y = 0.045, label = "NY",color="white",fontface=2, size=3.5)+
  annotate("text", x = 38, y = 0.045, label = "PA",color="white",fontface=2, size=3.5)+
  annotate("text", x = 43, y = 0.045, label = "TX",color="white",fontface=2, size=3.5)
D6 #Percentage

#Mutiplots
multiplot(vList3, D1, D2, D3,D5, D6,cols=2)

#D4: What are the age(s) of the children living in your household?
#Values: 0-1
#0=Unchecked
#1=Checked

D4_MyMultResp_labels<-data.frame(resp.data.eda[48:53])

names(D4_MyMultResp_labels) <- make.names(names(D4_MyMultResp_labels))
colnames(D4_MyMultResp_labels)[1] <- "1+, <2"
colnames(D4_MyMultResp_labels)[2] <- "1+, 3-5"
colnames(D4_MyMultResp_labels)[3] <- "1+, 6-12"
colnames(D4_MyMultResp_labels)[4] <- "1+, 13-18"
colnames(D4_MyMultResp_labels)[5] <- "1+, >18"
colnames(D4_MyMultResp_labels)[6] <- "None in household"

str(D4_MyMultResp_labels)
```

```
D4_MyMultResp<-data.frame(Freq = colSums(D4_MyMultResp_labels[1:6]),
                          Pct.of.all.Resp = (colSums(D4_MyMultResp_labels[1:6])/sum(D4_MyMultResp_labels[1:6]))*100,
                          Pct.Check.All.That.Apply = (colSums(D4_MyMultResp_labels[1:6])/nrow(D4_MyMultResp_labels[1:6]))*100)
D4_MyMultResp


barplot(D4_MyMultResp[[3]]
      ,names.arg=row.names(D4_MyMultResp)
      ,main = "D4: What are the age(s) of the children living in your household?"
      ,xlab = "Percent Check All That Apply"
      ,ylab = "Children (1 or more, age range)"
      ,col = "#2b8cbe"
      ,horiz = TRUE
      ,cex.names = 0.8)

######################################## Ch. 1 Data Preparation ######################################

##### Create X Matrix or Design Matrix #####

#Effects Coding
xvec=c(0,1,2,3)
efcode.att.f(xvec) #Check

#Provide Matrix for Choice Design
task.mat <- as.matrix(taskV3[, c("screen", "RAM", "processor", "price", "brand")])
task.mat
dim(task.mat) #108 rows: 3 rows for each of 36 choice sets by 5 columns
head(task.mat)

#Produces Effects Coded Version of task.mat
X.mat=efcode.attmat.f(task.mat)
X.mat
dim(X.mat) #108 by 11
head(X.mat)

#Add interaction between brand and price to X.mat
#scale(taskV3$price,scale=FALSE)
pricevec=taskV3$price-mean(taskV3$price) #Get the vector of prices from taskV3 and center it on its mean by taking the elemen
t-wise difference of two numeric vectors.
head(pricevec)
str(pricevec)

#Extract the last 3 columns from X.mat that represent brand.
X.brands=X.mat[,9:11]
X.brands
dim(X.brands)
str(X.brands)

#Multiply each column in X.brands by pricevec.
X.BrandByPrice = X.brands*pricevec
X.BrandByPrice
dim(X.BrandByPrice)
str(X.BrandByPrice)
```

```
#Combine X.mat and X.BrandsByPrice to get the X matrix we'll use for choice modeling
X.matrix=cbind(X.mat,X.BrandByPrice)
X.matrix
dim(X.matrix) #108 by 14
str(X.matrix)
head(X.matrix)

#Check to ensure it's a postive #
det(t(X.matrix)%*%X.matrix)

#Create data frame for Y Vector or Y Response Data
ydata=resp.data.v3[,4:39]
str(ydata)
head(ydata)
summary(ydata) #No missing data within the responses
names(ydata) #Check to see if you have all 36 response variables
ydata=na.omit(ydata) #Make sure you have no missing data

ydatadf <- ydata
head(ydatadf)

ydata=as.matrix(ydata) #Convert ydata to matrix
dim(ydata) #424 by 36

#Create indicator variable for STC product ownership (vList3)
zowner <- 1 * ( ! is.na(resp.data.v3$vList3) ) #Equal to 1 if a respondent has ever owned an STC product, other 0.

#Create the list of data lists for HB MNL models - rhierMnlDP()
#Contains data for each respondent and is a list with two elements (X.matrix and ydata)
lgtdata = NULL # a starter placeholder for your list
for (i in 1:424) {
   lgtdata[[i]]=list(y=ydata[i,],X=X.matrix)
 }

length(lgtdata) #424
str(lgtdata)

################################### Ch. 2 Modeling - Fitting HB MNL Model ###################################

require(bayesm)

lgtdata=lgtdata #x and y data

#Specify 100k and kth = 5
mcmctest=list(R=100000,keep=5) #R = 100k iterations; kth = 5 to adjust for autocorrelation

#Create the "Data" list rhierMnlDP() expects
Data1=list(p=3,lgtdata=lgtdata)   # p is choice set size

#Output for testrun1
set.seed(123)
testrun1=rhierMnlDP(Data=Data1,Mcmc=mcmctest)
names(testrun1) #"betadraw" is an array that has the draws (i.e. samples from marginal posterior distributions) for the regressio
n coefficients
```

```
dim(testrun1$betadraw) #424 rows, 14 columns in X.matrix, 20000 blocks or samples produced by the (thinned) iterations of the algorithm

#Extract betadraw out of testrun1
betadraw1=testrun1$betadraw #betadraw1
betadraw1
dim(betadraw1) #424 people, 14 betas, x simulated values

#Respondent 1
#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[1,1,]),betadraw1[1,1,]) #1st: Person 1, 2nd: Beta 1 (e.g., dummy X1 or screen 7), 3rd: Iteration # after th
inning
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[1,2,]),betadraw1[1,2,]) #Screen10
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[1,3,]),betadraw1[1,3,]) #RAM16
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[1,4,]),betadraw1[1,4,]) #RAM32
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[1,5,]),betadraw1[1,5,]) #Proc2
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[1,6,]),betadraw1[1,6,]) #Proc2.5
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[1,7,]),betadraw1[1,7,]) #Price299
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[1,8,]),betadraw1[1,8,]) #Price399
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[1,9,]),betadraw1[1,9,]) #Somesong
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[1,10,]),betadraw1[1,10,]) #Pear
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[1,11,]),betadraw1[1,11,]) #Gaggle
abline(v=10000, col="red", lwd=3, lty=2)

#Respondent 2
```

```
#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[2,1,]),betadraw1[2,1,]) #1st: Person 1, 2nd: Beta 1 (e.g., dummy X1 or screen 7), 3rd: Iteration # after th
inning
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[2,2,]),betadraw1[2,2,]) #Screen10
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[2,3,]),betadraw1[2,3,]) #RAM16
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[2,4,]),betadraw1[2,4,]) #RAM32
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[2,5,]),betadraw1[2,5,]) #Proc2
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[2,6,]),betadraw1[2,6,]) #Proc2.5
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[2,7,]),betadraw1[2,7,]) #Price299
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[2,8,]),betadraw1[2,8,]) #Price399
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[2,9,]),betadraw1[2,9,]) #Somesong
abline(v=10000, col="red", lwd=3, lty=2)

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[2,10,]),betadraw1[2,11,]) #Pear

#Plot betadraw1 & Decide Burning Period (Convergance/Stablization)
plot(1:length(betadraw1[2,11,]),betadraw1[2,11,]) #Gaggle
abline(v=10000, col="red", lwd=3, lty=2)

#Respondent 1 Continued
#Density smooth histogram of converged points, post burn-in period
plot(density(betadraw1[1,1,10001:20000],width=2))
abline(v=0) ## vertical line
abline(v=mn) ## vertical line

#Compute the probability that person 1 beta 1 > 0
# Assumes Normal distribution of beta 1
mn <- mean(betadraw1[1,1,10001:20000])
sd <- sd(betadraw1[1,1,10001:20000])
mn
```

```
sd

prob <- pnorm(0,mean=mn, sd=sd, lower.tail = FALSE)
prob

#Empirical probability based on sample values
p1b1 <- betadraw1[1,1,10001:20000]
quantile(p1b1, probs = seq(0, 1, by= 0.1))

####Summary####
#Mean for Person 1 and Beta 1 (7 inch)
summary(betadraw1[1,1,10001:20000]) #Positive mean denotes like
exp(0.7929) #log(odds ratio)=7929; odds ratio = exp(0.7929)=2.209796

#Mean for Person 1 and Beta 2 (10 inch)
summary(betadraw1[1,2,10001:20000]) #Negative mean denotes dislike
exp(-0.38501) #log(odds ratio)=-0.38501; odds ratio = exp(-0.38501)=0.6804438

#Means of the 14 coefficients and the means across all respondents
betameansoverall<-apply(betadraw1[,,10001:20000],c(2),mean)
betameansoverall

#Percentiles of the 14 coefficients or betas
perc <- apply(betadraw1[,,10001:20000],2,quantile,probs=c(0.05,0.10,0.25,0.5 ,0.75,0.90,0.95))
perc #50% = median; 90% confidence obtained by 5% and 95%

#Extras
#Matrix of coefficient means by each respondent
apply(betadraw1[,,10001:20000],c(1,2),mean) #424 respondents, 14 mean coefficients.

#Compare Beta 1 vs. Beta 2 for Person 1
summary((betadraw1[1,1,10001:20000]-betadraw1[1,2,10001:20000]))
plot(density(betadraw1[1,1,10001:20000]-betadraw1[1,2,10001:20000],width=2))

########## Ch. 3 Modeling - Fitting a HB MNL model with Prior STC Ownership as a Covariate ############

#Center z covariate, "demean" zoner and make result a 1 column matrix
zownertest=matrix(scale(zowner,scale=FALSE),ncol=1)

#Create Data2 list from Data1 to Include zowner as a covariate
Data2=list(p=3,lgtdata=lgtdata,Z=zownertest)

#Output for testrun2
set.seed(123)
testrun2=rhierMnlDP(Data=Data2,Mcmc=mcmctest)
names(testrun2) #Check to see that Deltadraw is included
dim(testrun2$Deltadraw) #Deltadraw is a matrix with # rows = saved iterations, #number of columns = number of regression pre
dictors in the X.matrix (14):

#Extract betadraw out of testrun2
betadraw2=testrun2$betadraw #betadraw1
betadraw2
dim(betadraw2) #424 people, 14 betas, x simulated values

#Plot betadraw2 & Decide Burn-in Period (Convergance/Stablization)
```

```
plot(1:length(betadraw2[1,1,]),betadraw2[1,1,]) #1st: Person 1, 2nd: Beta 1 (e.g., dummy X1 or screen 7), 3rd: Iteration # after th
inning
abline(v=10000, col="red", lwd=3, lty=2)

#Density smooth histogram of converged points, post burn-in period
plot(density(betadraw2[1,1,10001:20000],width=2))
abline(v=0) ## vertical line
abline(v=mn) ## vertical line

# Assumes Normal distribution of beta 1
mn <- mean(betadraw2[1,1,10001:20000])
sd <- sd(betadraw2[1,1,10001:20000])
mn
sd

prob <- pnorm(0,mean=mn, sd=sd, lower.tail = FALSE)
prob

#Empirical probability based on sample values
p1b1 <- betadraw2[1,1,10001:20000]
quantile(p1b1, probs = seq(0, 1, by= 0.1))

### Summary ###
#Means of the 14 coefficients and the means across all respondents for Deltadraw
#Samples from the posterior distributions of the regression coefficients of the 14 betas on  (mean-centered) zownertest.
apply(testrun2$Deltadraw[10001:20000,],2,mean) #Postive = Like; #Negative = Dislike

#Percentiles of the 14 coefficients or betas
#Summary of the posterior distributions of correlations b/w choice models regression coefficients and zownertest variable
apply(testrun2$Deltadraw[10001:20000,],2,quantile,probs=c(0.05,0.10,0.25,0.5 ,0.75,0.90,0.95))

#Means of the 14 coefficients and the means across all respondents for Betadraw
betameansoverall2<-apply(betadraw2[,,10001:20000],c(2),mean)
betameansoverall2

#Percentiles of the 14 coefficients or betas
perc2 <- apply(betadraw2[,,10001:20000],2,quantile,probs=c(0.05,0.10,0.25,0.5 ,0.75,0.90,0.95))
perc2 #50% = median; 90% confidence obtained by 5% and 95%

betadraw2=testrun2$betadraw
dim(betadraw2)

########## Ch. 4 Make customer choice prediction using the individual respondent's model  #############
######################### & goodness of fit & validation ########################################

########Beta Draw 1########

#Means of the 14 coefficients and the means across all respondents from Betadraw 1
betameans <- apply(betadraw1[,,10001:20000],c(1,2),mean)
str(betameans)
dim(betameans) #424 by 14
dim(t(betameans))
dim(X.matrix)

#Get the product of our X.matrix and each subjects vector of mean betas
```

```
xbeta=X.matrix%*%t(betameans)
dim(xbeta) #108 by 424

#Reorganize xbeta data
xbeta2=matrix(xbeta,ncol=3,byrow=TRUE) #subjects in rows with choice set alternatives across columns
dim(xbeta2) #15264 by 3

#Exponentiate xbeta2
expxbeta2=exp(xbeta2)
dim(expxbeta2) #15264 by 3

#Obtain Predicted Choice Probabilities
rsumvec=rowSums(expxbeta2) #divide each row by its sum
pchoicemat=expxbeta2/rsumvec
pchoicemat
head(pchoicemat)
dim(pchoicemat)

#Provide prediction of the customer choice
custchoice <- max.col(pchoicemat)
custchoice
head(custchoice)
str(custchoice)

#Assess Model Fit

#Confusion Matrix
ydatavec <- as.vector(t(ydata))
str(ydatavec)
head(ydatavec)
cm<-table(custchoice,ydatavec)
cm
accuracy<-sum(diag(cm))/sum(cm)
accuracy #0.8839099

#ROC Curve and AUC
require("pROC")
roctest <- roc(ydatavec, custchoice, plot=TRUE)
auc(roctest)
roctestMC <- multiclass.roc(ydatavec, custchoice, plot=TRUE)
auc(roctestMC) #0.9078

#-2log(likelihood) test applies to nested models only
logliketest <- testrun2$loglike
mean(logliketest) #Lower the better
hist(logliketest) #-5789.52

#Prediction for 36 choice sets Using Individual Respondent's model
m <- matrix(custchoice, nrow =36,  ncol = 424)
m2 <- t(m)
apply(m2, 2, function(x){tabulate(na.omit(x))})

#Predictions based on overall beta means, for 36 choice sets
#We can predict the original 36 choice sets using the pooled model.
#The code below, provides the probabilities as well as the frequencies for the 36 choice sets #####
```

```
#Means of the 14 coefficients and the means across all respondents
betavec=matrix(betameansoverall,ncol=1,byrow=TRUE)

#Get the product of our X.matrix and each subjects vector of mean betas
xbeta=X.matrix%*%(betavec)
dim(xbeta)

#Reorganize xbeta2 data
xbeta2=matrix(xbeta,ncol=3,byrow=TRUE)
dim(xbeta2)

#Exponentiate expxbeta2
expxbeta2=exp(xbeta2)

#Obtain Predicted Choice Probabilities
rsumvec=rowSums(expxbeta2)

#Provide prediction of the customer choice probabilites
pchoicemat=expxbeta2/rsumvec
pchoicemat

#Provide prediction of the customer choice frequencies
pchoicemat2 <- round(pchoicemat*424,digits=0)
pchoicemat2

########Beta Draw 2########

#Means of the 14 coefficients and the means across all respondents from Betadraw 1
betameans <- apply(betadraw2[,,10001:20000],c(1,2),mean)
str(betameans)
dim(betameans) #424 by 14

#Get the product of our X.matrix and each subjects vector of mean betas
xbeta=X.matrix%*%t(betameans)
dim(xbeta) #108 by 424

#Reorganize xbeta data
xbeta2=matrix(xbeta,ncol=3,byrow=TRUE) #subjects in rows with choice set alternatives across columns
dim(xbeta2) #15264 by 3

#Exponentiate xbeta2
expxbeta2=exp(xbeta2)
dim(expxbeta2) #15264 by 3

#Obtain Predicted Choice Probabilities
rsumvec=rowSums(expxbeta2) #divide each row by its sum
pchoicemat=expxbeta2/rsumvec
pchoicemat
head(pchoicemat)
dim(pchoicemat)

#Provide prediction of the customer choice
custchoice <- max.col(pchoicemat)
custchoice
```

```
head(custchoice)
str(custchoice)

#Assess Model Fit

#Confusion Matrix
ydatavec <- as.vector(t(ydata))
str(ydatavec)
cm<-table(custchoice,ydatavec)
cm
accuracy<-sum(diag(cm))/sum(cm)
accuracy #0.8844995

#ROC Curve and AUC
require("pROC")
roctest <- roc(ydatavec, custchoice, plot=TRUE)
auc(roctest)
roctestMC <- multiclass.roc(ydatavec, custchoice, plot=TRUE)
auc(roctestMC) #0.9081

#-2log(likelihood) test applies to nested models only
logliketest <- testrun2$loglike
mean(logliketest) #Lower the better
hist(logliketest)

#Prediction for 36 choice sets Using Individual Respondent's model
m <- matrix(custchoice, nrow =36,  ncol = 424)
m2 <- t(m)
apply(m2, 2, function(x){tabulate(na.omit(x))})

#Predictions based on overall beta means, for 36 choice sets
#We can predict the original 36 choice sets using the pooled model.
#The code below, provides the probabilities as well as the frequencies for the 36 choice sets #####

#Means of the 14 coefficients and the means across all respondents
betavec=matrix(betameansoverall,ncol=1,byrow=TRUE)

#Get the product of our X.matrix and each subjects vector of mean betas
xbeta=X.matrix%*%(betavec)
dim(xbeta)

#Reorganize xbeta2 data
xbeta2=matrix(xbeta,ncol=3,byrow=TRUE)
dim(xbeta2)

#Exponentiate expxbeta2
expxbeta2=exp(xbeta2)

#Obtain Predicted Choice Probabilities
rsumvec=rowSums(expxbeta2)

#Provide prediction of the customer choice probabilites
pchoicemat=expxbeta2/rsumvec
pchoicemat
```

```
#Provide prediction of the customer choice frequencies
pchoicemat2 <- round(pchoicemat*424,digits=0)
pchoicemat2

########## Ch. 5 Predicting extra scenarios, as well as the 36 choice sets  ############
################### using betas from all the pooled respondents ########################

###### Predicting Extra Scenarios #####
ex_scen <- read.csv("extra-scenarios.csv")
Xextra.matrix <- as.matrix(ex_scen[,c("V1","V2","V3","V4","V5","V6","V7","V8","V9",
                           "V10","V11","V12","V13","V14")])
dim(Xextra.matrix)

###### Predict extra scenarios using the overall model######
#Means of the 14 coefficients and the means across all respondents
betavec=matrix(betameansoverall,ncol=1,byrow=TRUE)
dim(betavec)
betavec

#Get the product of our X.matrix and each subjects vector of mean betas
xextrabeta=Xextra.matrix%*%(betavec)
dim(xextrabeta)

#Reorganize xbetaextra2 data
xbetaextra2=matrix(xextrabeta,ncol=3,byrow=TRUE)
dim(xbetaextra2)

#Exponentiate expxbetaextra2
expxbetaextra2=exp(xbetaextra2)

#Obtain Predicted Choice Probabilities
rsumvec=rowSums(expxbetaextra2)

#Provide prediction of the customer choice
pchoicemat=expxbetaextra2/rsumvec
pchoicemat
dim(pchoicemat)

###### Predict extra scenarios based on individual models ######
xextrabetaind=Xextra.matrix%*%(t(betameans))
xbetaextra2ind=matrix(xextrabetaind,ncol=3,byrow=TRUE)

#xextrabetaind=Xextra.matrix%*%(t(betameansindividual))
#dim(xextrabetaind)
#xbetaextra2ind=rbind(matrix(xextrabetaind[1:3,],ncol=3,byrow=TRUE),
#                matrix(xextrabetaind[4:6,],ncol=3,byrow=TRUE))
dim(xbetaextra2ind)

##Exponentiate expxbetaextra2ind
expxbetaextra2ind=exp(xbetaextra2ind)

#Obtain Predicted Choice Probabilities
rsumvecind=rowSums(expxbetaextra2ind)

#Provide prediction of the customer choice
```

```
pchoicematind=expxbetaextra2ind/rsumvecind
dim(pchoicematind)
head(pchoicematind)

#Highest probability identifier
custchoiceind <- max.col(pchoicematind)
head(custchoiceind)
str(custchoiceind)

#Use individual respondents models to predict the extra scenario & then using 'voting' kind of techniques
#make the final prediction
#Each individual model is separately predicting each additional scenario.
extra1 <- custchoiceind[1:424]
extra2 <- custchoiceind[425:848]
table(extra1)
table(extra2)

##### Accuracy based on confusion matrix for each of the 424 respondents using individual models #####
############ for all the 36 choice sets - actual response vs predicted response ################

#Compute accuracy based on confusion matrix for each of the 424 respondents using individual models
resp_accuracy <- NULL
for (i in 1:424) {
  start <- i*36-35
  end <- i*36
  d <- table(factor(custchoice[start:end],levels = 1:3),
             factor(ydatavec[start:end], levels = 1:3))
  resp_accuracy[i] <- sum(diag(d))/sum(d)
}

#Plot Accuracy
plot(resp_accuracy, main = "Model Accuracy by Respondent")

#Create dataframe of accuracy for each of 424 respondents using individual models
respdf <- data.frame(resp_accuracy)
head(respdf)
str(respdf)

#Add respondent number to dataframe
head(ydatadf)
rn <- rownames(ydatadf)
rndf <- as.data.frame(rn)
resp_all <- cbind(rndf,respdf)
head(resp_all)
str(resp_all)

#Plot histogram of accuracy with frequency
hist(resp_all$resp_accuracy)

#Identify accuracy lower than 0.6
outlier <- subset(resp_all, resp_accuracy < 0.6)
outlier[order(outlier$resp_accuracy),]
```