**Predictive with SAS**

**Dataset:**

The dataset has 3291 rows with 17 variables. The dataset shows the store data of four cracker brands - private, sunshine, Keebler, niabisco, their prices and whether display and features were present for these.



1. **OBS** : = Observation number

2. **Private, Keebler, Sunshine, Nabisco**: Indicator variables for which brand was chosen. Value of 1 indicates the brand that was chosen. Other 3 brands will be 0 in that observation.

3. **PricePrivate, PriceNabisco, PriceKeebler and PriceSunshine**: Prices that were offered by each brand for that purchase occasion.

4. **DisplPrivate**: = 1 if Private had a store display, =0 if Private did not have a store display

5. **DisplKeebler**: = 1 if Keebler had a store display, =0 if Keebler did not have a store display

6. **DisplSunshine**: = 1 if Sunshine had a store display, =0 if Sunshine did not have a store display

7. **DisplNabisco**: = 1 if Nabisco had a store display, =0 if Nabisco did not have a store display

8. **FeatPrivate**: = 1 if Private had a store feature, =0 if Private did not have a store feature

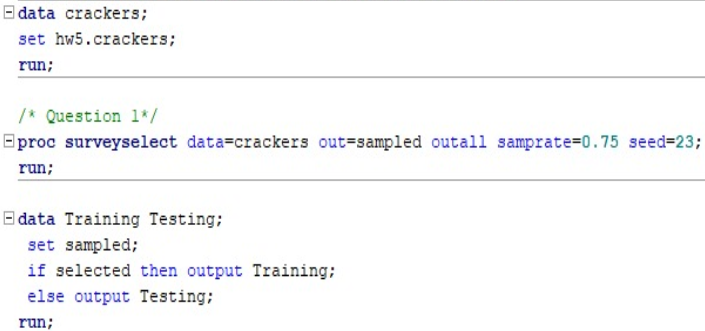
9. **FeatKeebler**: = 1 if Keebler had a store feature, =0 if Keebler did not have a store feature

10. **FeatSunshine**: = 1 if Sunshine had a store feature, =0 if Sunshine did not have a store feature

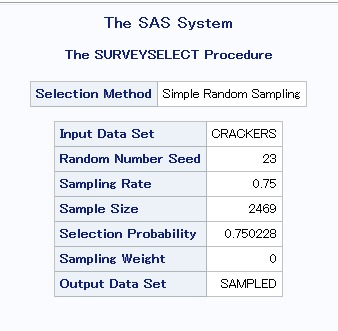
11. **FeatNabisco**: = 1 if Nabisco had a store feature, =0 if Nabisco did not have a store feature

**1) Use PROC SURVEYSELECT to sample the original data into training and testing data sets. Use 75% for training and 25% for testing. Use the seed= option to set random seed to a value of 23**

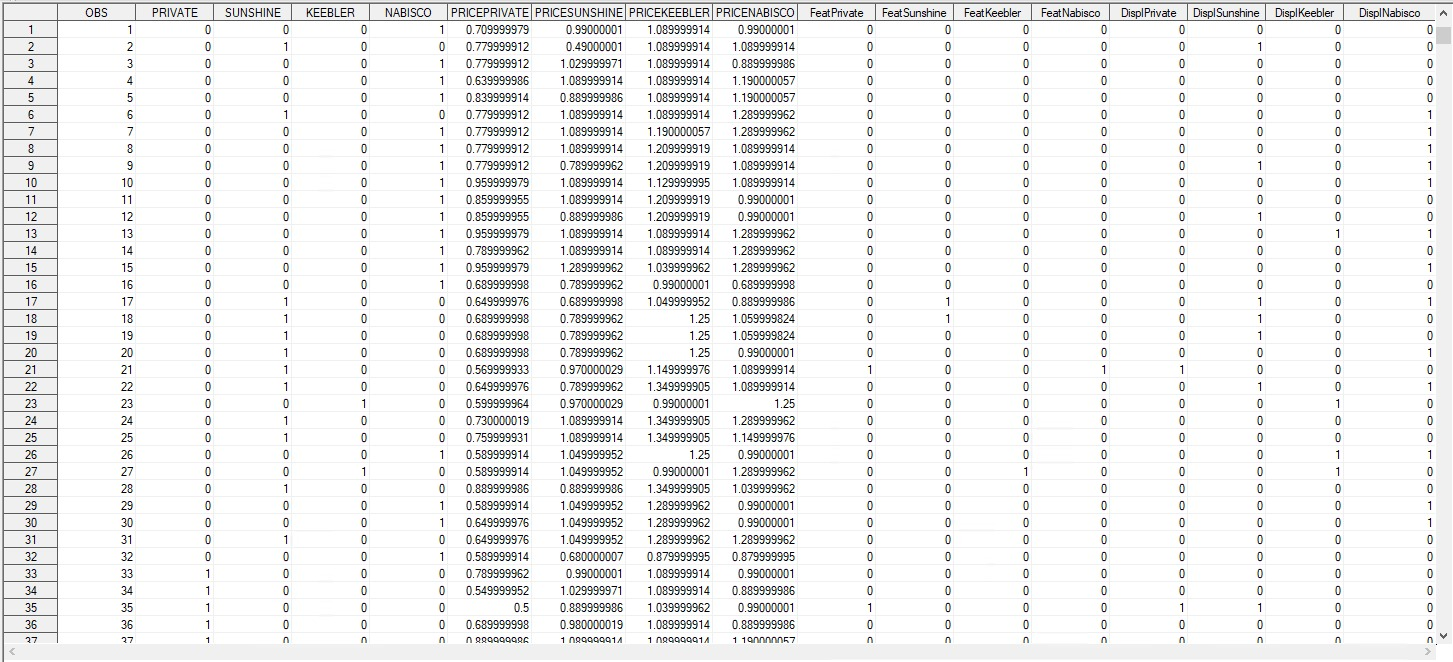
The dataset is split into train and test using 75 – 25 split. Surveyselect command is used for this with samprate =0.75 defining the train test split and seed =23 sets it to provide same output each time. It sets a selection identifier with which it selects the selected observations to train and test.



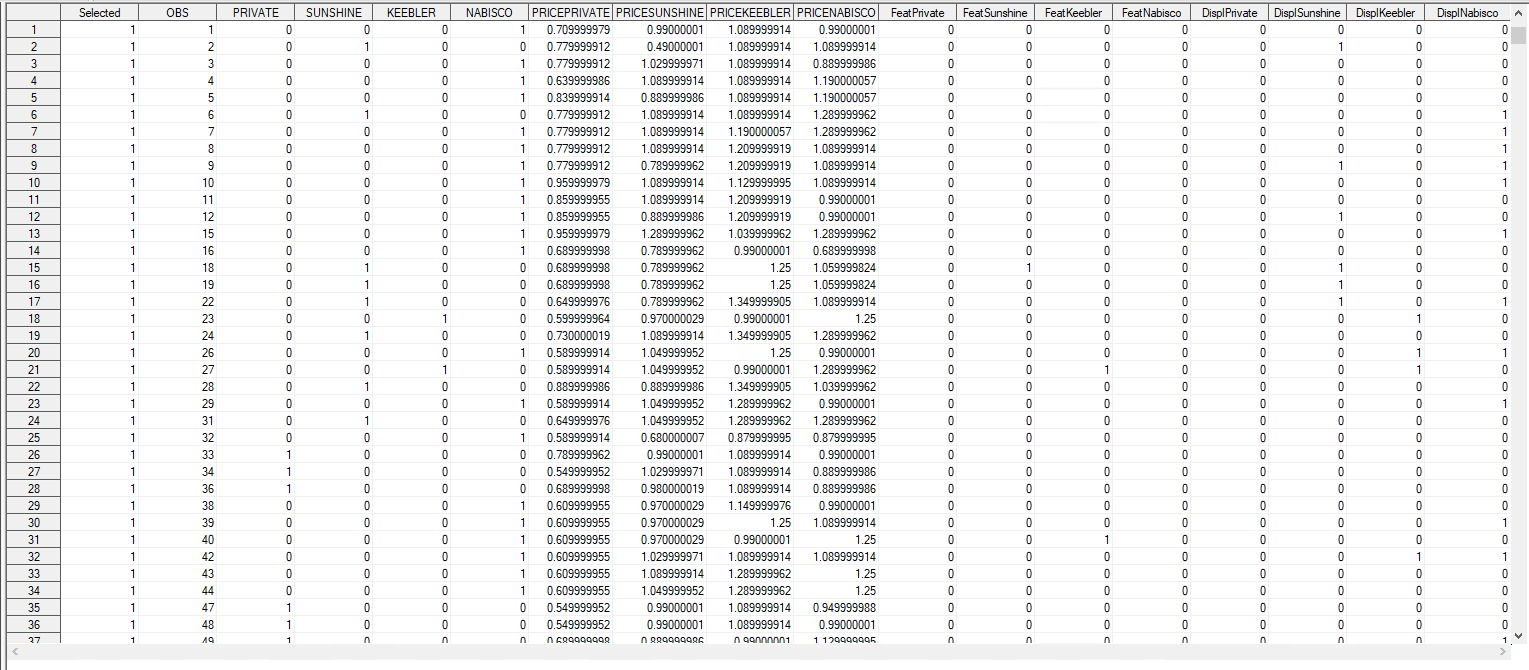
The output shows the sampling rate of 0.75 being used and the sample that is then obtained has a size of 2469.



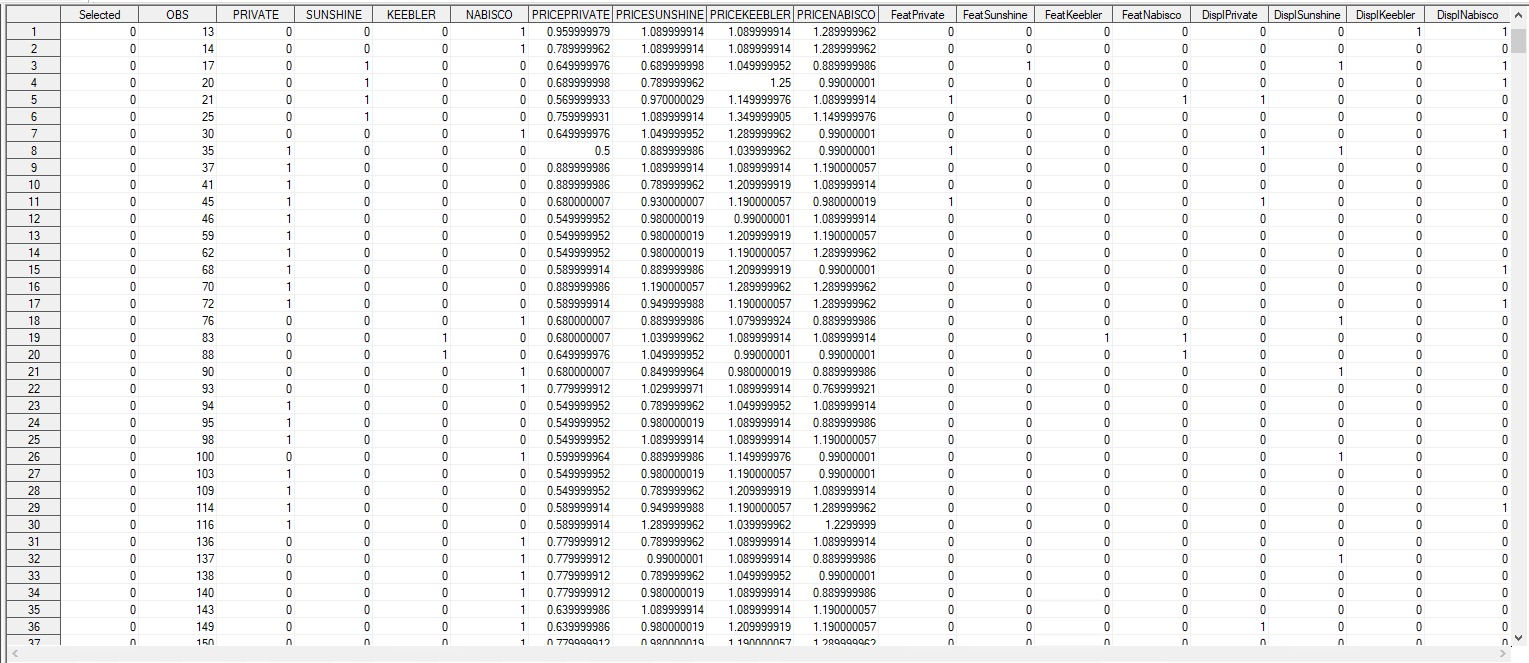
The dataset is shown in the below snip.



Train dataset – this has the selected column set to 1 as it is filtered in proc after selection.



Test dataset – this has the selected column set to 0 as it is filtered in proc after selection.



**2)** **The store manager would like to predict the choice probabilities for each brand of crackers depending on the price, display and promotion for all brands. What type of multinomial logit model would you estimate – a model with alternative-specific characteristics or with individual-specific characteristics? Write the general utility model to estimate this logit model.**

In the dataset provided there are no individual specific characteristics that pertain to a brand or customer alone. Thus, we will go with alternative specific characteristics.

General utility for this model is as defined:

V=β0+ β x price +α1 displ\*private + α2 feat\*private + α3 Sunshine + α4 displ\*Sunshine + α5 feat\*Sunshine + α6 Nabisco + α7 displ\*Nabisco + α8 feat\*Nabisco + α9 Keebler + α10 displ\*Keebler + α11 feat\*Keebler

Intercept for brand private α0, private = 0 as it is the reference baseline

**3) Is the data formatted as needed to estimate the above multinomial logit model using PROC LOGISTIC? If not, how should the data be formatted? Reformat the data as necessary.**

Multinomial logit model deals with prediction of more than one category in the target variable.

In the dataset given, it is seen that the observations don’t have separate rows for each brand or category. Rather it has one row specifying all details for one observation.

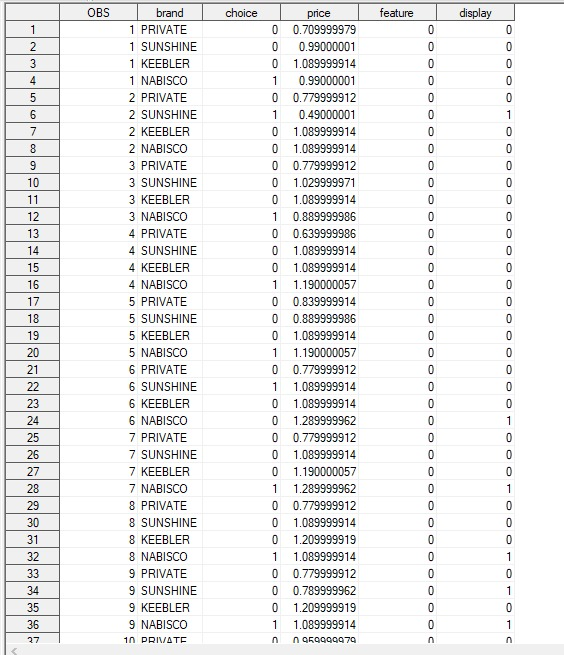
The alternative variables – display, feature, price should be mentioned separately for each brand for each observation and not as one observation having all details.

It should have a separate column identifying if a purchase is done or not for each brand with 0/1 as the choice variable.



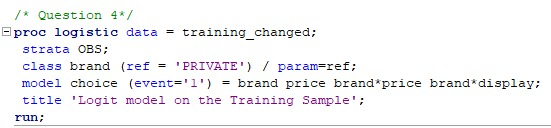
Thus, the dataset is read and re-arranged to reflect one observation per brand of crackers with a choice variable. It now has 9876 rows.

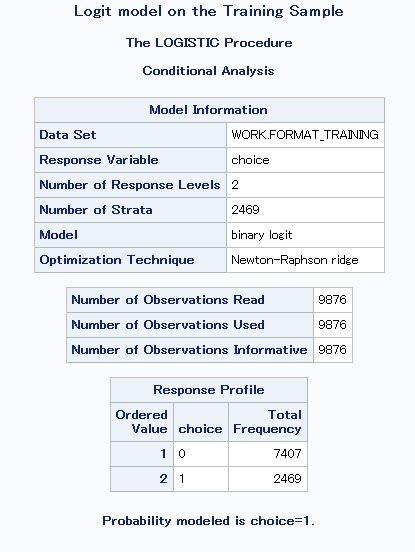
The headers are obs, brand, choice, price, display and feature.



**4.Estimate the logit model on the training sample using PROC LOGISTIC and report the estimation results (model parameters, significance)**

The logistic model is run on the training data. Strata is set as obs as this helps identify the customer transaction uniquely with the brand. The brand private is set as reference or baseline with 1 is the event of a purchase being made. The other brands are measured from the reference baseline of private brand





The response variable is choice as it shows whether a purchase is done or not. It is binary 0/1 which is shown in the response level. Of the 9876 observations we have 2469 values 1 showing the purchase being made and remaining 7407 rows are brands that weren’t purchased by that customer.

The class level information shows selection of brand setting private as baseline, the 4 brands are represented as

PRIVATE 0 0 0

KEEBLER 1 0 0

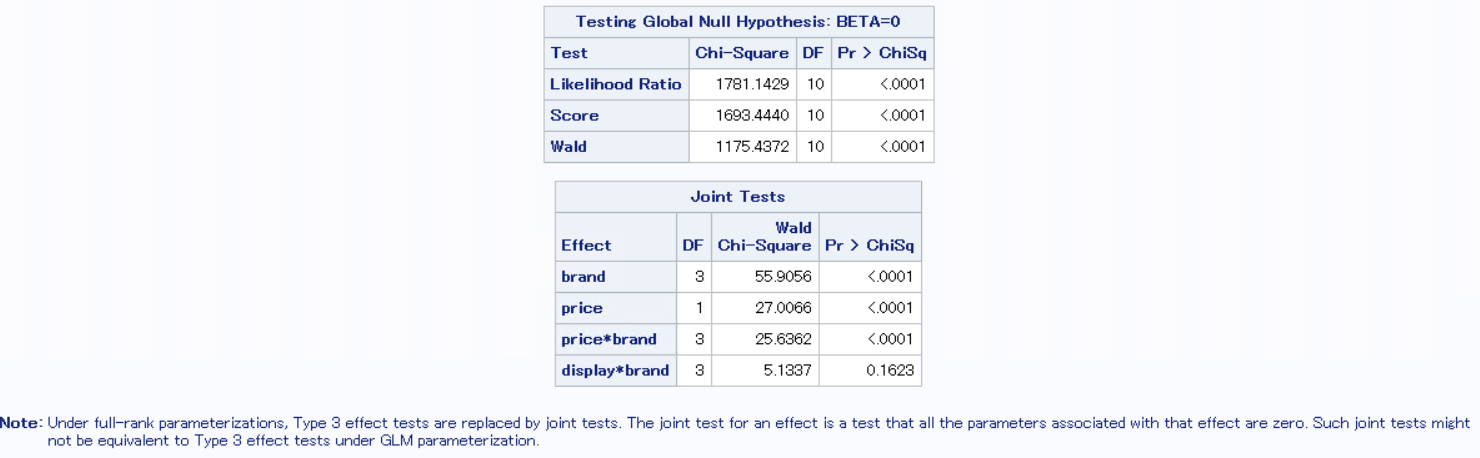
NABISCO 0 1 0

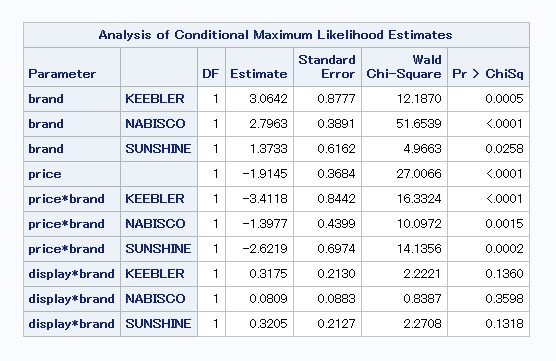
SUNSHINE 0 0 1



The strata summary shows customer with 3 brands obtaining 0 while 1 brand obtains 1 showing that it has been purchased.

The model fit statistics shows the fit of the model and how relevant it is.





The estimate shows which variables or interaction terms have a positive or negative effect. Estimates with +ve values have a positive relationship while those with -ve values have negative relationship. P value provides the significance of that relationship.

Price is the most affecting variable as an increase in price will always decrease of sale of the product across all brands. The display for sunshine, Keebler and Nabisco have a positive value but they are insignificant as shown by their pvalue.

Setting brand Private as reference, the coefficients are interpreted as:

• **Brand Keebler:** Brand Keebler is 3.0642 times higher preference with reference to brand Private and is statistically significant

• **Brand Nabisco**: Brand Nabisco is 2.7963 times higher preference with reference to brand Private, everything else constant, and is statistically significant.

• **Brand Sunshine**: Brand Sunshine is at 1.3733 times higher preference with reference to brand Private

• **Price**: A one-unit increase price of crackers will decrease by 1.9145 the choice of buying the crackers of any brand and is statically significant

• **Price\* Keebler**: With a unit increase in price, customers will prefer to buy private by -3.4118 times over keebler and it is statistically significant

• **Price\* Nabisco**: With a unit increase in price, customers will prefer to buy private by -1.3977 times over nabisco and it is statistically significant

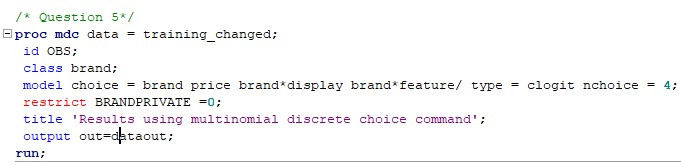
**• Price\* Sunshine**: With a unit increase in price, customers will prefer to buy private by -2.6219 times over sunshine and it is statistically significant

**• Display\* Keebler**: Customers prefer buying Keebler over private by 0.3175 times when properly displayed but it is not statistically significant

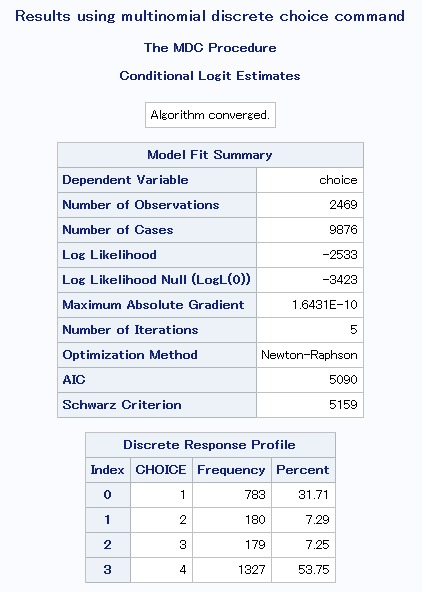
**• Display\* Nabisco**: Customers prefer buying Nabisco over private by 0.0809 times when properly displayed but it is not statistically significant

**• Display\* Sunshine**: Customers prefer buying Sunshine over private by 0.3205 times when properly displayed but it is not statistically significant

**5) Reproduce your results using multinomial discrete choice command PROC MDC**

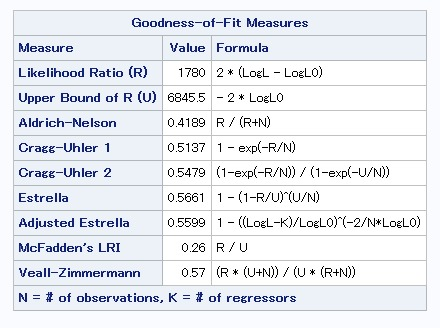


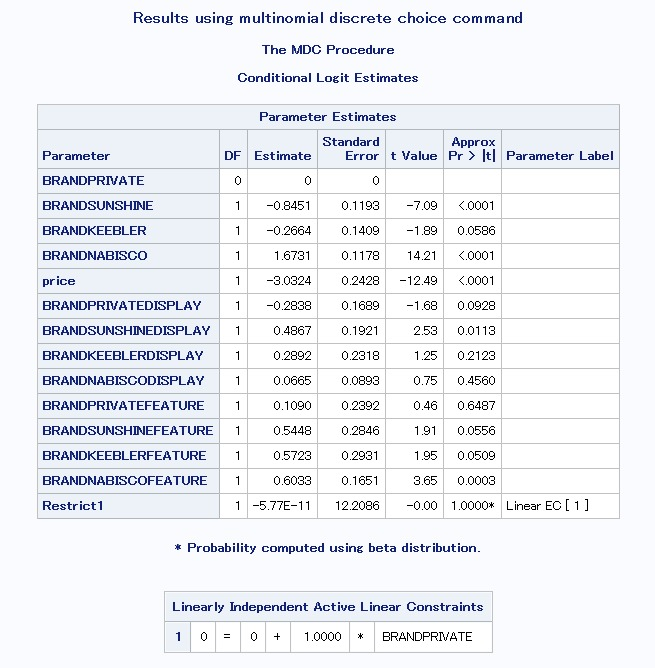
* The MDC procedure is used for multinomial logistic where the choice set contains many alternatives
* Class variable is used for Brand to create dummy variables for the categoricals
* Type=clogit is used for specifying the use of conditional logistic model
* The choice variable is provided as the Y term and the interaction and other alternatives are provided as X terms
* Nchoice =4 specifies the number of brands/categories the dataset holds.
* The restrict command is used to set a variable to baseline reference in our case, it is the brand private



The discrete response profile shows the different choice we have and the frequency with which it is chosen and the percent it is chosen to the other 3. Choice 4 has the highest frequency.

The MDC calculates 9 goodness of fit measures which evaluates the model. Of which most used R term as the measure in calculation.

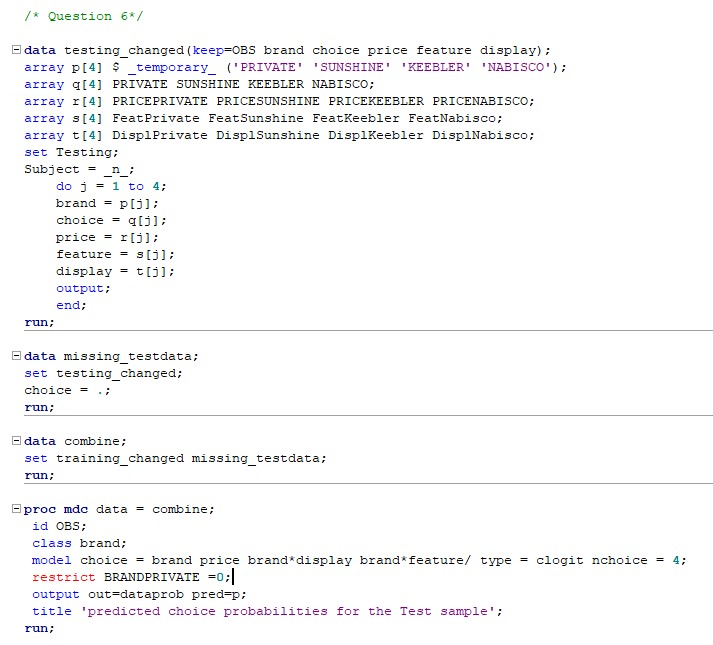




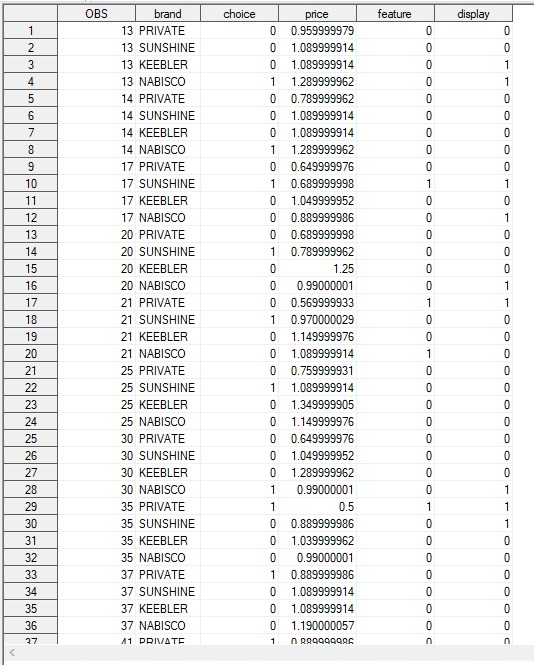
The parameter estimate shows the estimates of each variable with its pvalue showing the significance of that estimate. The estimate values take +ve and -ve values with +ve values showing a +ve relationship and -ve values showing a negative relationship.

6. Use PROC MDC to predict the choice probabilities for the test sample using the estimated model.

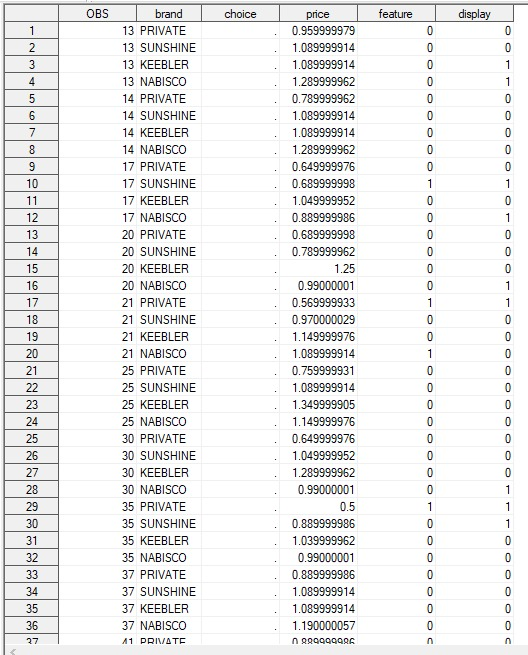
The dataset is with test data is chosen and the data is formatted to represent each brand for a customer as separate rows. The alternatives are shown separate for each brand per customer.



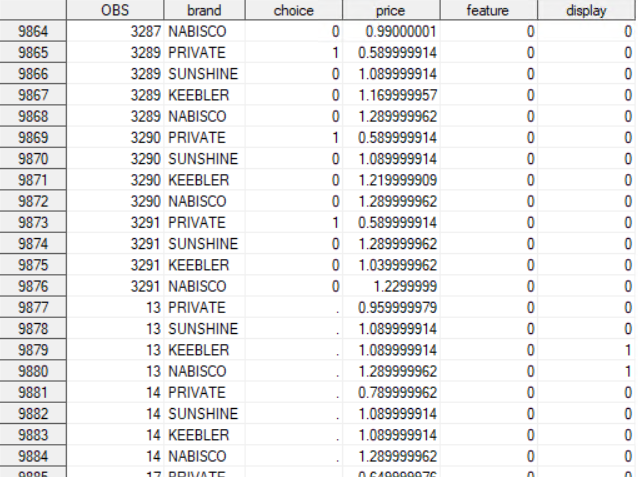
This shows the formatted dataset using separate row for each brand per customer instance.



The choice variable for the test data is made blank as we use the model to predict the values.



We can see the combined dataset has blanks for test data and choice variables are populated for train dataset



Thus, the logistic model is run on the dataset and the output is as follows.

