

## Statistical Analysis of “Pampers develops a rash- a rash of market share”

# Estimate the regression of brand preference on the ratings of nine attributes. Interpret the results.

**Solution:**

As given in the case, P&G’s marketing department has collected the survey data from 300 mothers of infants to understand the competitive diaper market. The first variable represents brand preference (Y). The dependent and independent variables are as follows:

**Dependent Variable:**

Y = Brand Preference

**Independent Variable:**

|  |  |
| --- | --- |
| x1 | COUNT PER BOX |
| x2 | PRICE |
| x3 | VALUE |
| x4 | UNISEX |
| x5 | STYLE |
| x6 | ABSORBENCY |
| x7 | LEAKAGE |
| x8 | COMFORT/SIZE |
| x9 | TAPING |

After running the regression in SPSS, we got the following output. Let’s try to interpret the results:

Table 1 gives the model summary for the regression with the given dependent and independent variables. The adjusted R square of 0.722 indicates that the 72.2% of the variation in dependent variable is explained by the independent variable. Table 2 i.e., the ANOVA table shows the variance explained by the regression versus the variance that the regression model cannot explain. The p-value of F-statistics (0.000 < 0.05) clearly indicates that the model is significant at 5% level of significance and is a good fit on given data.

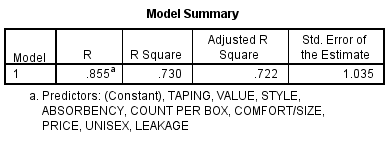


Table : Model Summary of Regression

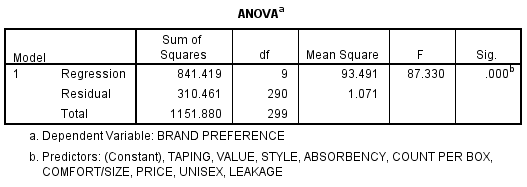


Table : ANOVA Table

As the model is significant, we can proceed with interpreting the significance and sign of the coefficient of the independent variables which is given in Table 3. The results indicate that the variables count per box, unisex, absorbency are significant at 5% level of significance and are contributing positively to brand preference which is our dependent variable. Rest all other independent variables are coming out to be insignificant.

One important problem to notice is also that the model suffers from the problem of multicollinearity. The VIF value for the variable leakage is more than 10 and for absorbency it is 9.4 which signifies the presence of multicollinearity in the model. Due to this multicollinearity in the independent variables, the following problems could be present in the model:

1. Multiple independent variables might come insignificant
2. The estimated sign of the coefficient variable might not be correct

The regression equation for the model is given as:

Y = -3.458 x1 + 0.503 x2 + 0.108 x3 + 0.016 x4 + 0.472 x5 - 0.065 x6 + 0.341 x7 + 0.252 x8 + 0.106 x9

Which clearly indicates that one unit increase in independent variable count per box will increase the dependent variable brand preference score by 0.504 units. The similar interpretation can be done for the other independent variables.

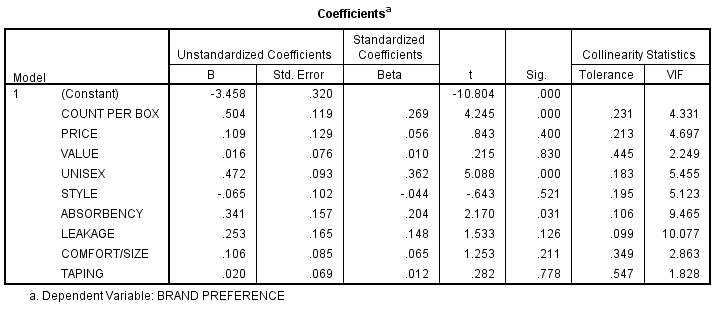


Table :Regression Results

Hence, to avoid these problems with so many independent variables, the managers of P&G’s marketing department must proceed with the ‘Factor Analysis.’

# Carry out a chapter analysis of the nine attributes. Label the factors and interpret the results. Save the factors scores and use them as independent variables in running the regression of brand preference on attributes. Interpret all the results and also compare the regression results so obtained with the regression as obtained in question 1.

**Solution:**

**Part I: Factor Analysis:**

The results of the factor analysis are as follows:

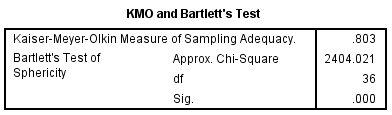


Table : KMO and Bartlett's Test

As given in table 4, the measure of sampling adequacy from the KMO test is 0.803 and the Bartlett’s test of Sphericity is significant indicates that factor analysis can be applied to the given set of data.

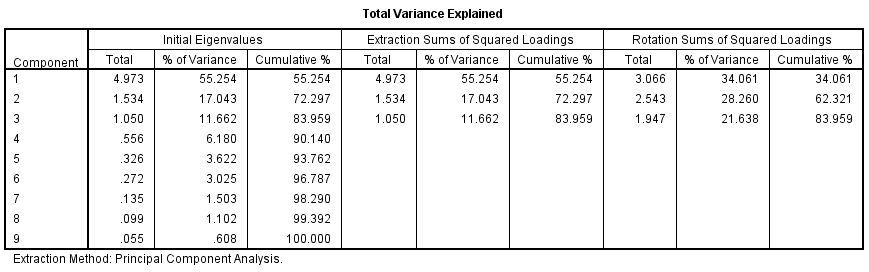


Table : Factor Analysis: Total Variance Explained

From the analysis, we got three factors for our data as given in table 5 which are explaining a total of 83.95% of the variation in the independent variables i.e., 83.95% of the variation in all the nine independent variables is captured in the three factors generated. The percentage of variation explained by the first, second and third factor after rotation is 34.06%, 28.26% and 21.64% respectively.

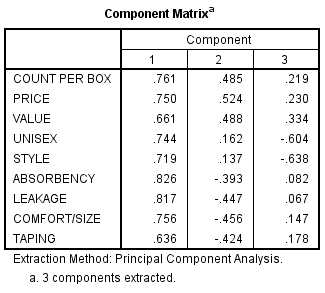


Table : Factor Analysis - Component Matrix

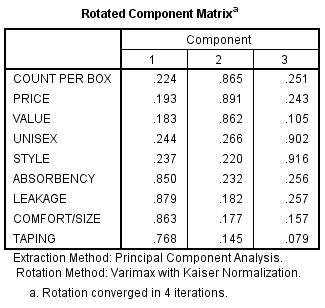


Table : Factor Analysis – Rotated Component Matrix

Let us take 0.75 as our cut of point to name the factors from the rotated component matrix:

Factor 1 consist of the independent variables: 1. Absorbency 2. Leakage 3. Comfort/Size and 4. Taping. These independent variables typically indicate the functionality or the features of the Pampers product of each of the brand. So, let’s name this factor as “Functionality/Features rating of the Pampers of respective brand.”

**Factor 2 consist of the independent variables:** Count per box, price and value. This factor can be clearly defined as “**value for money**.”

**Factor 3 consist of the independent variables:** Unisex and Style. The third factor can be labelled as “**Fashion Attributes of the product.**”

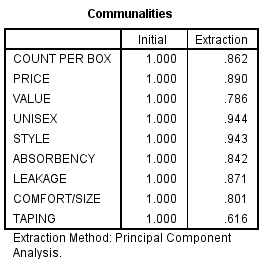
****

Table : Factor Analysis - Communalities

The table 8 gives the communalities of the factor analysis for the given nine independent variables. The extraction value for 0.862 for the count per box indicates that the total of 86.2% of the variation in this independent variable is explained by all the three factors combined. Similarly, 89% of the variation in price as an independent variable is explained by all the three factors combined.

Finally, the factor score equations for the given factors for each respondent can be written with the help of Component score coefficient score matrix as given in table 9 below:

**Factor score for 1st factor =** -0.069 x1 - 0.085 x2 - 0.057 x3 - 0.109 x4 - 0.109 x5 + 0.309 x6 + 0.329 x7 + 0.343 x8 + 0.343 x8 + 0.319 x9

**Factor score for 2st factor =** 0.398 x1 + 0.419 x2 + 0.437 x3 - 0.096 x4 - 0.125 x5 - 0.048 x6 - 0.08 x7 - 0.058 x8 - 0.044 x9

**Factor score for 3rd factor =** -0.055 x1 - 0.062 x2 - 0.16 x3 + 0.585 x4 + 0.609 x5 - 0.03 x6 - 0.024 x7 - 0.097 x8 - 0.131 x9

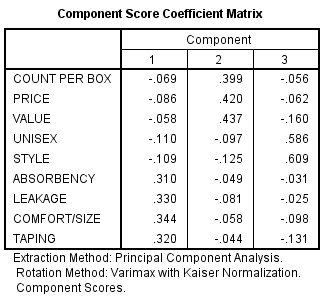
****

Table : Factor Analysis - Component Score coefficient Matrix

**Part II: Regression Analysis using the factor scores:**

Now, after running the regression using factor score, we got the following results:

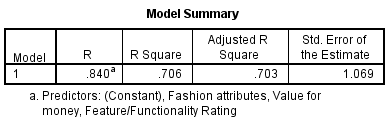


Table : Model Summary of Regression

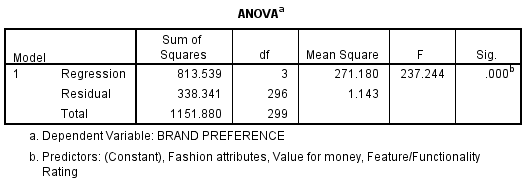


Table : ANOVA Table from Regression Analysis

Model summary from Table 3 indicates that, around 70.6% variation in the dependent variable i.e., brand preference is explained by the three factors. Whereas, with all the 9 independent variables, the model was about to explain around 72.2% of the variation, so the information loss is very less with factor analysis while on the benefit side, we are using just the 3 factors instead of all the 9 variables which is amazing.

The ANOVA table indicates that the model is significant at 1% level of significance which means that the model with three factors can explain a significant amount of variation in dependent variable as compared to the error variance in the model.

Table 12 shows the coefficient analysis of the regression model using factors. Note that all the three factors are statistically significant. The VIF value of one for all the three factor indicates that the factor regression model has solved the problem of multicollinearity which was originally there with all 9 independent variables model.

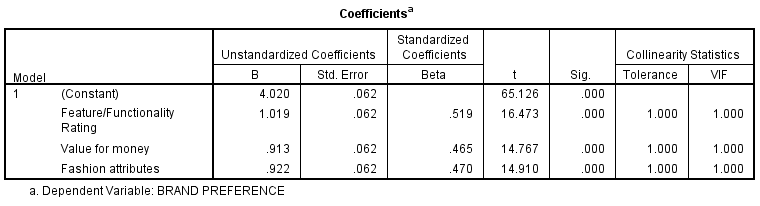


Table : Regression Coefficient Analysis

The regression equation with the help of factors as independent variables can be given as:

Y = 4.020 + 1.019 \* F1 + 0.913 \* F2 + 0.922 F3

Where,

F1 = Feature/Functionality Rating

F2 = Value for money

F3 = Fashion attributes

The coefficient value of 1.109 for Feature/Functionality Rating score indicates that one unit increase in the score of this factor will bring 1.109 unit change in the value of the dependent variable brand preference. Similarly, one unit change in the value for money factor will bring the 0.913 unit change in dependent variable.

So, in comparison of the results of two regression models, we can conclude that:

1. Almost all the variance explained by the original model is also explained by the factor regression model with only three independent variables.
2. In factor model, all the three factors came out to be significant but, in original 9 independent variable model, most of the variables were insignificant.
3. And most importantly, the factor model has solved the problem of multicollinearity in the independent variables. The three-factor model was not suffering from the problem of multicollinearity hence it is more reliable and satisfies the condition of linear regression.

# Carry out a discriminant analysis by grouping the brand preference into two groups using a score of one to four as low preference and five to seven as high preference. Use the ratings of attributes as independent variables. Interpret the results.

**Solution:**

After grouping the brand preference into two groups as per the condition in question, out of 300 respondents, 170 respondents fall in low brand preference while 130 respondents fall in high brand preference as shown in table 13.

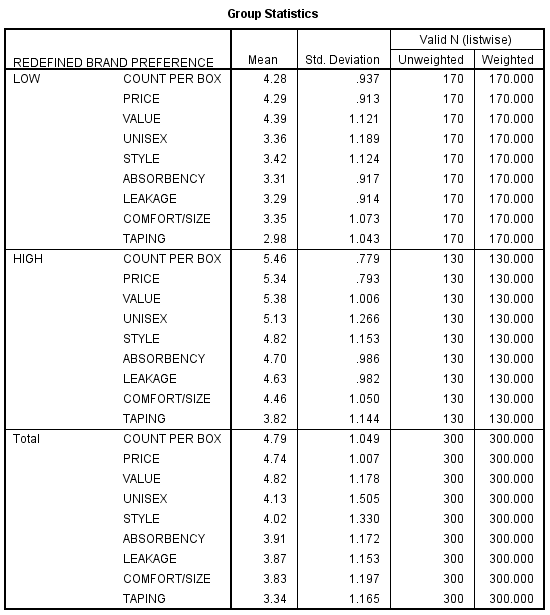


Table : Discriminant Analysis - Group Statistics

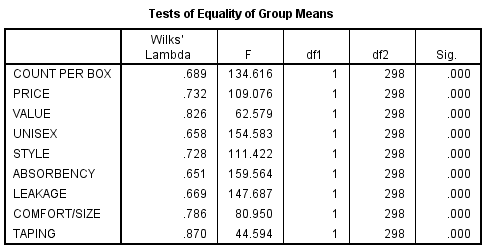


Table : Test of equality of group mean using F-statistics

The significance of F-statistics in the table 14 for each independent variable indicates that on the basis of initial analysis, all the variables seem to have discriminating power into high and low preference.

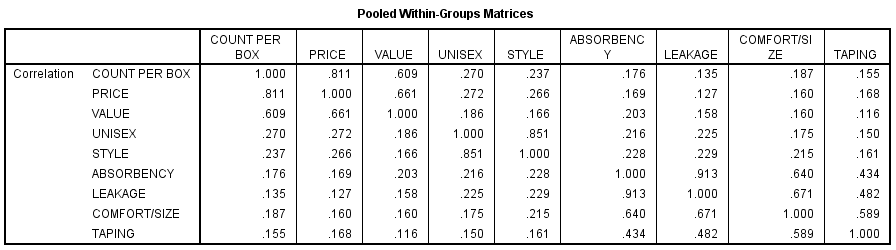


Table : Pooled within-group Matrices

The independent variables count per box and price have a high correlation of 0.811. Also, the correlation between Leakage and absorbency is 0.913 which is again very high. So, the model might be suffering from the problem of multicollinearity. Hence, the reliability of the model might be less and researcher should be cautious about it.

The dependent variable in Discriminant analysis is always categorical and, in our case, it is high or low preference of the customer towards a particular brand of pampers. The estimated unstandardized discriminant function coefficients can be given using the table 16 below as:

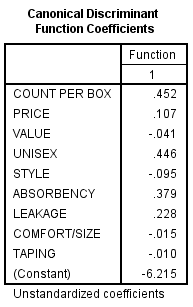


Table : Canonical Discriminant Function Coefficients

Y = - 6.215 + 0.451 x1 + 0.107 x2 - 0.04 x3 + 0.446 x4 - 0.095 x5 + 0.379 x6 + 0.228 x7 - 0.015 x8 - 0.009 x9

So, this discriminant function will give the discriminant score for each of the respondent which will help to classify the respondent into prospective high or low preference customer.

The eigenvalues for the above estimated discriminant function is 1.061. as show in table 17 below.

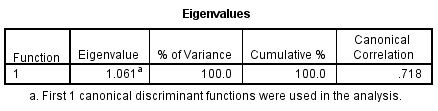


Table : Eigenvalues and Canonical Correlation table

The value of the canonical correlation for our discriminant function is 0.718, the square of which come out to be 0.5155, which means that 51.55% of the variance in discriminating model between a prospective high/low preference customer is due to the changes in nine predictor variables.

**Classification of the cases using the Discriminant Function:**

Group centroids which represent the mean discriminant scores of the high and low preference customers is calculated using the function values in table 18 below:

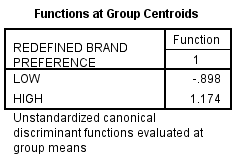


Table :Functions at Group Centroids

= - 0.00013

The above cut off score is equivalent to zero. So, the customer having a discriminant score of less than zero will be categorized as low preference consumer and a customer having a discriminant score of more than zero will be categorized as high preference consumer.

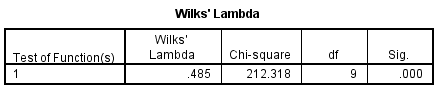


Table : Wilks Lambda

The value of Wilk’s Lambda is 0.485 which is significant at 1%. Hence, our model is significant and can be used for further discriminant analysis. The lower the value of Wilk’s Lambda, the better it is and whose significance is tested using the chi-square test.

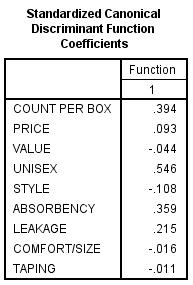


Table : Standardized Canonical Discriminant Function Coefficients

The absolute value of Standardized Canonical Discriminant Function Coefficients indicates the relative importance of the independent variables in discriminating between the high and low preference for the given brand of pampers. As the absolute value of Unisex variable is coming out to be highest, we can say that the weather the pampers is unisex or not has the highest discriminating power amongst all the independent variables followed by count per box with the second highest absolute value of Standardized Canonical Discriminant Function Coefficients of 0.394. Similar hierarchy of importance of the variables in discriminating can be formed as per the order of absolute descending values of Standardized Canonical Discriminant Function Coefficients.

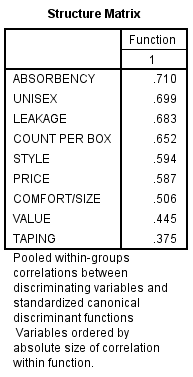


Table : Structure Matrix

Structure matrix is another way of finding the relative contribution of predictor variables in discriminating between the high/low brand preference. The highest value of absorbency followed by the unisex indicates the top two discriminating variables respectively. Similarly, higher the value in the structure matrix for a particular variable, the higher will be the importance in discrimination.

Note that the order of importance is different according to standardized coefficient and the structure matrix. So, as per the rule, we will finally follow the order as per the Standardized Canonical Discriminant Function Coefficients.

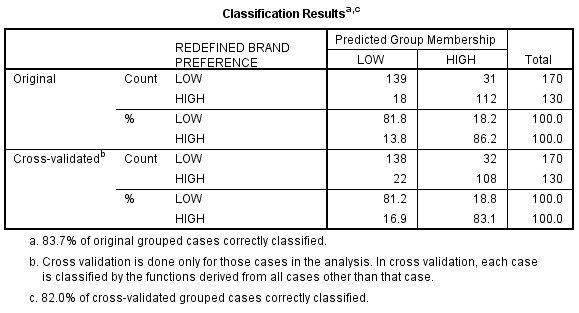


Table :Classification Results

The overall accuracy of the model for the given data is 83.67% as calculated from table 22 which is a good accuracy than the null model.

# Carry out the above exercise (as given in question 3) by using a logit regression. Compare the results in two cases.

**Solution:**

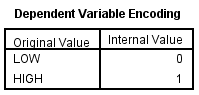
****

Table ‑:Dependent Variable Encoding

The dependent variable i.e., brand preference is encoded as 0 for low brand preference and 1 for high brand preference.

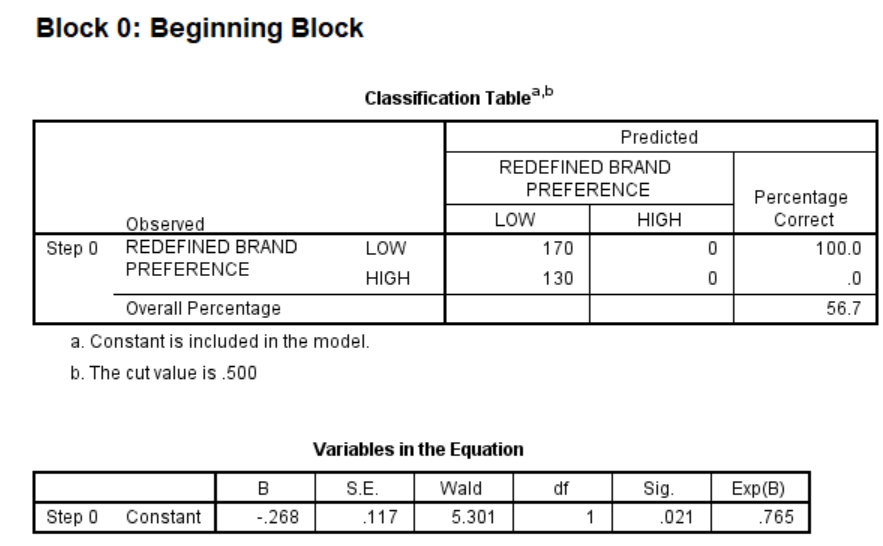
****

Table ‑: Null Model

Table 4-2 indicates that the null model only has the accuracy of 56.5% in discriminating between the low brand preference and high brand preference respondents. So, me must proceed by adding the independent variables in the model.

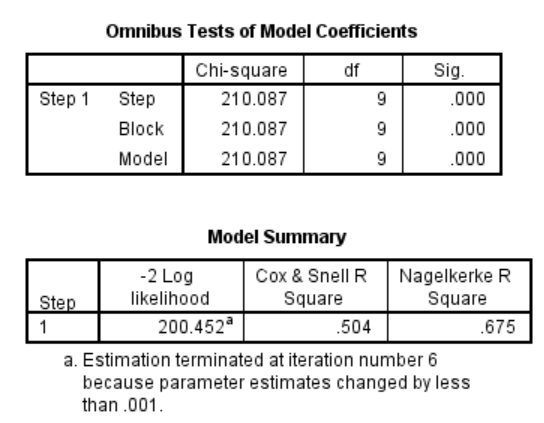
****

Table ‑: Omnibus Tests of Model Coefficients and Model Summary

The step, block and Model chi-square values for the Omnibus Tests of Model Coefficients are coming to be significant at 1% level of significance. This means that the model is good fit and we can proceed with the further interpretation of the model.

Table 4-3 also gives the model summary. Cox and Snell R-square shows that 50.4% of the variance is explained by the model. While, Nagelkerke R-square shows that 67.5% variance is explained by the model.

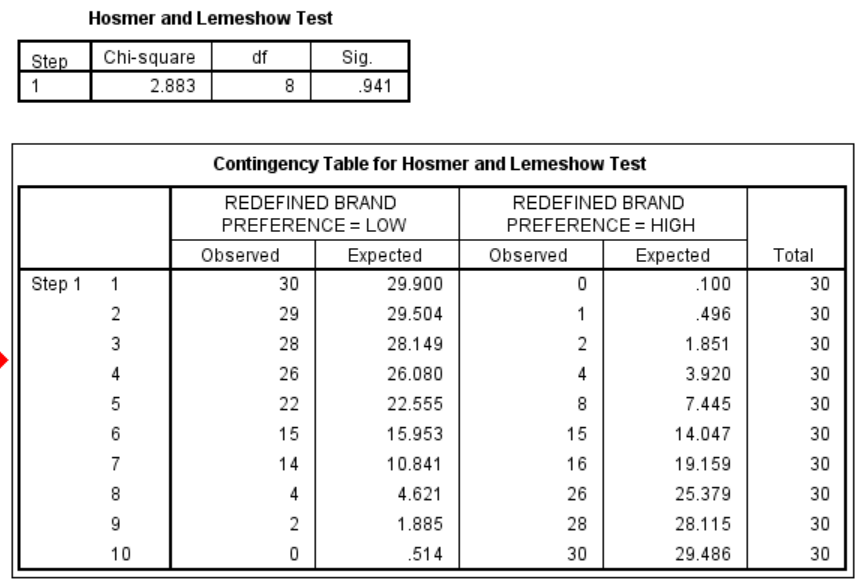


Figure ‑:Hosmer and Lemeshow Test

Figure 4-4 shows the results of Hosmer and Lemeshow Test of model fit. The insignificant p-value of 0.941 indicates that there are no differences between the fitted values of the model and the actual values.

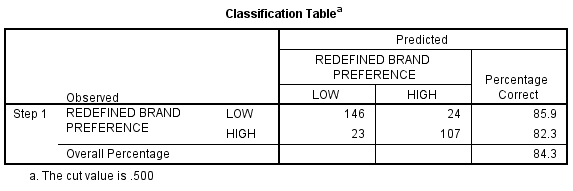


Figure ‑: Classification Table

Table 4-5 gives the Classification Table for the estimated and actual brand preference. The model gives an accuracy of 84.3% which is really good than the null model (having accuracy of 56.7%.)

The logistic equation can be written as:

Ln(p/1-p) = Log (odds in favor of high brand preference) = -14.280 + 1.171 x1 + 0.056 x2 + 0.047 x3 + 0.786 x4 -0.125 x5 + 0.813 x6 + 0.424 x7 + -0.014 x8 -0.003 x9

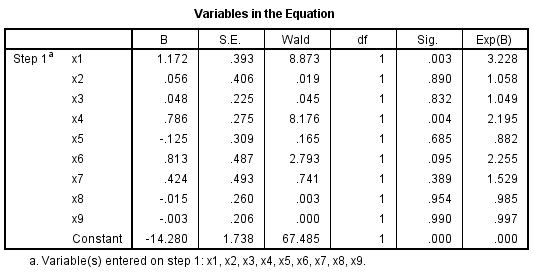


Table ‑: Variables in the Equation

Variables x1 (count per box) and x4 (unisex) are significant at 5% level of significance, while variable x6 (absorbency) is significant at 10% level of significance. Rest all other variables came out to be insignificant.

Importance of significant variables in discriminating between high and low preference is given by:

1. x1 i.e., count per box with an exp(B) value of 3.228 indicating that if the count per box goes up by one unit, then the probability of the respondent having higher preference for the brand is 3.228-times the probability of the same respondent having lower preference for that brand.
2. Similarly, x6 i.e., absorbency with an exp(B) value of 2.255 is second important in discriminating, indicating that if the score for absorbency goes up by one unit, then the probability of the respondent having higher preference for the brand is 2.255-times the probability of the same respondent having lower preference for that brand.
3. Similar interpretation can be made for the other variables as well.

**Comparison of the results of discriminant analysis and logistic regression:**

The accuracy of discriminant analysis was 83.67% whereas the accuracy of logistic regression model is 84.3% which is slightly higher. Also, there is a difference in the relative importance given to an independent variable in both the models. With the advancement of research logistic models are more preferred over the discriminant models due to the sigmoid function used for the discrimination and having lesser restriction related to normality as compared with linear regression model

# Carry out a three groups discriminant analysis by redefining the preference variable as follows:

1. The score of one to two may be regarded as low preference.
2. The score of three to five may be regarded as medium preference.
3. The score of six to seven may be regarded as high preference.

**Solution:**

After doing the classification of the brand preference into low, medium and high preference group the number of respondents in each category came out to be 79, 140 and 81 respectively as shown in the table 5-1 below.

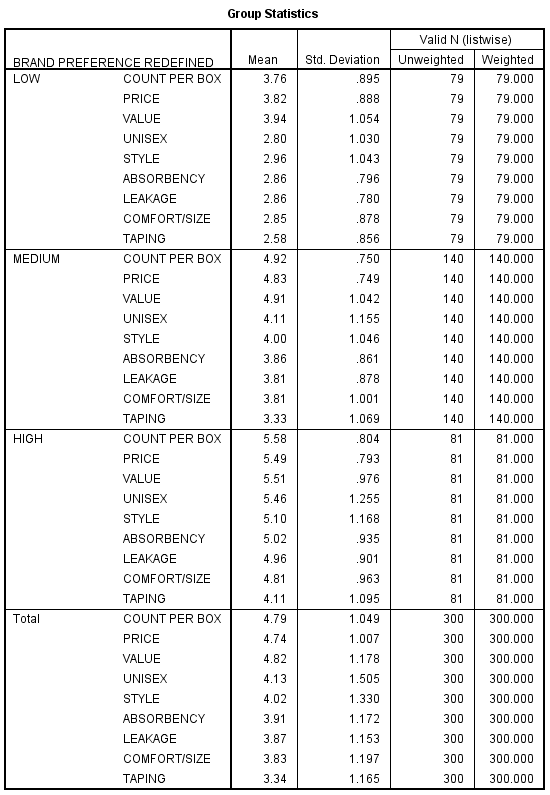


Figure ‑:Group statistics

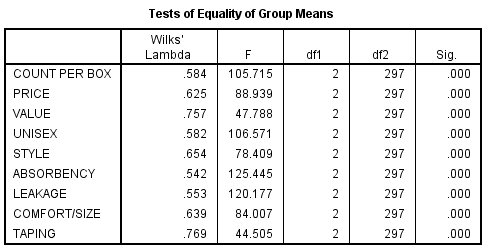


Figure ‑: Test of equality of group mean using F-statistics

The significance of F-statistics in the table 5-2 for each independent variable indicates that on the basis of initial analysis, all the variables seem to have discriminating power into high, medium and low preference.

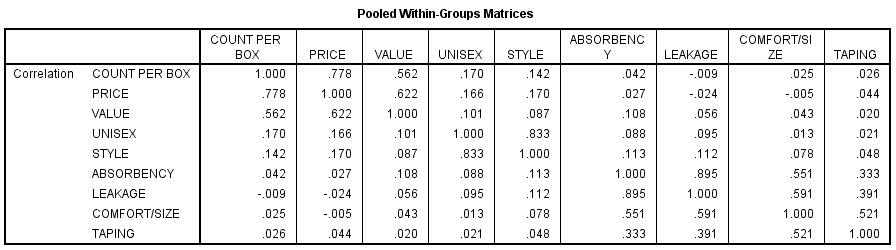


Figure ‑:Pooled within-group Matrices

The independent variables count per box and price have a high correlation of 0.778. Also, the correlation between Leakage and absorbency is 0.895 which is again very high. So, the model might be suffering from the problem of multicollinearity. Hence, the reliability of the model might be less and researcher should be cautious about it.

The dependent variable in Discriminant analysis is always categorical and, in our case, it is high, medium or low preference of the customer towards a particular brand of pampers. The estimated unstandardized discriminant function coefficients for the given set of variables can be given using the table 5-4 below as:

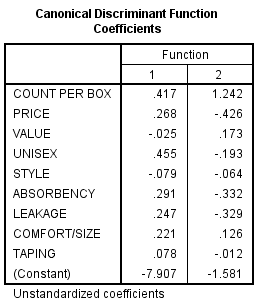


Figure ‑:Canonical Discriminant Function Coefficients

As we have three categories for the dependent variable, we have got two Canonical Discriminant Functions. But, for the purpose of the discrimination of new case, we will consider the discriminant score that we got from the first function only.

Y = -7.907 + 0.416 x1 + 0.268 x2 - 0.024 x3 + 0.455 x4 - 0.079 x5 + 0.291 x6 + 0.246 x7 + 0.22 x8 + 0.077 x9

So, this discriminant function will give the discriminant score for each of the respondent which will help to classify the respondent into prospective high or low preference customer.

The eigenvalues for the above estimated discriminant function are 2.059 and 0.029 as show in table 5-5 below.

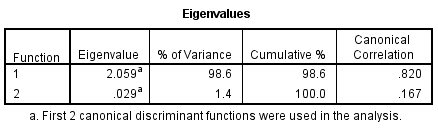


Figure ‑: Eigenvalues and Canonical Correlation table

The Eigenvalue of 2.059 for the 1st function is high and indicates that the out of the two functions, the 1st function is explaining 98.6% of the variance. While the eigen value of the second function is 0.029 which is comparatively very low indicating only 1.4% of the total variance is explained by the second function.

The value of the canonical correlation for our discriminant function 1 which is our main discriminating function is 0.82, the square of which come out to be 0.6724, which means that 67.24% of the variance in discriminating model between a prospective high/medium/low preference customer is due to the changes in nine predictor variables.

**Classification of the cases using the Discriminant Function:**

Group centroids which represent the mean discriminant scores of the high, medium and low preference customers is calculated using the function values in table 18 below:

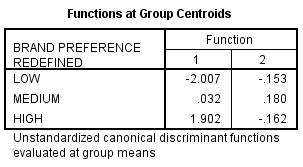


Figure ‑: Functions at Group Centroids

The weighted mean of the low and medium brand preference is calculated as:

The weighted mean of the medium and high brand preference is calculated as:

Cut off score 1 and 2 can be used for the interpretation as follows:

1. The participants for which the estimated cut-off score will be below -0.7035 will be categorized as low preference consumers.
2. The participants for which the estimated cut-off score will be above -0.7035 but below 0.7174 will be categorized as medium preference consumers.
3. The participants for which the estimated cut-off score will be above 0.7174 will be categorized as high preference consumers.

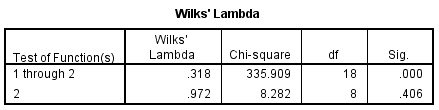


Figure ‑: Wilks Lambda

The value of Wilk’s Lambda for function 1 through 2 is 0.318 which is significant at 1%. Hence, our model using function one is significant and can be used for further discriminant analysis. The lower the value of Wilk’s Lambda, the better it is and whose significance is tested using the chi-square test.

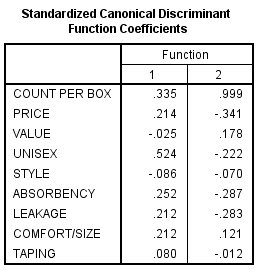


Figure ‑: Standardized Canonical Discriminant Function Coefficients

The absolute value of Standardized Canonical Discriminant Function Coefficients for function one indicates the relative importance of the independent variables in discriminating between the high, medium and low preference for the given brand of pampers. As the absolute value of Unisex variable is coming out to be highest with the value of 0.524, we can say that the weather the pampers is unisex or not has the highest discriminating power amongst all the independent variables followed by count per box with the second highest absolute value of Standardized Canonical Discriminant Function Coefficients of 0.335. Similar hierarchy of importance of the variables in discriminating can be formed as per the order of absolute descending values of Standardized Canonical Discriminant Function Coefficients.

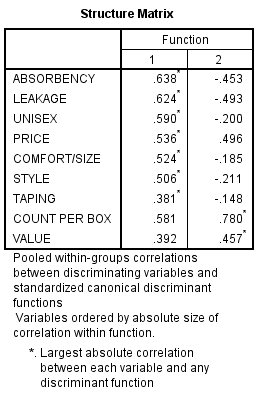


Figure ‑: Structure Matrix

Structure matrix is another way of finding the relative contribution of predictor variables in discriminating between the high/medium/low brand preference. The highest value of absorbency followed by the leakage indicates the top two discriminating variables respectively. Similarly, higher the value in the structure matrix for a particular variable, the higher will be the importance in discrimination.

Note that the order of importance is different according to standardized coefficient and the structure matrix. So, as per the rule, we will finally follow the order as per the Standardized Canonical Discriminant Function Coefficients for determining the importance of each variable in discrimination.