Monika Pawar

This assignment will analyze the data (HotelClickStream.xls) and interpret the results. This dataset includes clickstream data of online transactions for hotel booking in year 2011

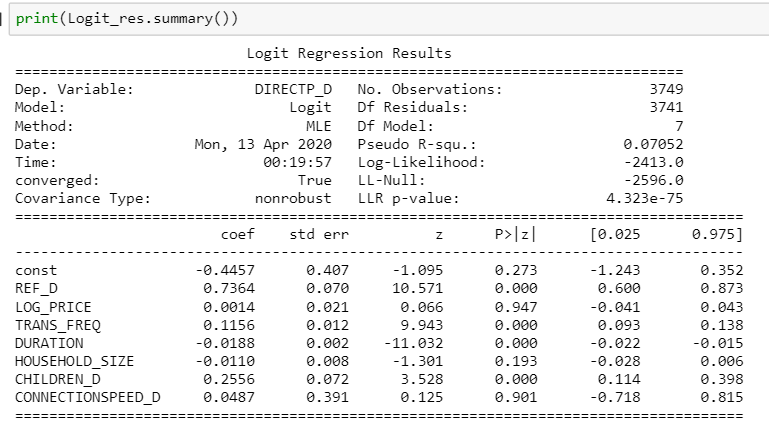
**SP20: Behavioral Analytics with Vsulztn**

***Click Stream Analytics Project Part 2***

**--------------------------------------------------------------------------------------------------------------------------------**

* **Ques 1 (b) Please report and interpret your (Logistic/Logit) regression results, which should include the interpretation of each of the regression coefficients.**

Logit regression results



* In the above result**, LOG\_PRICE, HOUSEHOLD\_SIZE, CONNECTIONSPEED\_D** these 3 variables have p-value greater than 0.05, hence it’s interpreted that they are *NOT* significantly correlated with DIRECTP\_D.
* Coef. – These are the values for the logistic regression equation for predicting the dependent

variable from the independent variable. They are in log-odds units. Similar to OLS regression, the prediction equation is

**Estimated Logit = log [ Π / (1 Π)] = β0 + β1\*REF\_D + β2 \*TRANS\_FREQ + β3\*DURATION + β4\*CHILDREN\_D**

* Estimated Logit = log [ Π / (1 Π)] = -0.446 + 0.736 \* REF\_D + 0.0+ 0.116 \* TRANS\_FREQ - 0.019 \* DURATION + 0.256 \* CHILDREN\_D
* Our Dependent Variable is DIRECTP\_D (which is the dummy variable indicating whether the transaction is incurred directly from a hotel website (1) or a third-party hotel website (0))
* The coefficient of each variable is understood as the change in log odds of the dependent variable DIRECTP\_D upon 1 unit increase in the predictor (IV), holding all other predictors constant.

**REF\_D**

Based on our result, when there is increase in 1-unit in REF\_D, we expect the log odds of DIRECTP\_D to increase by 0.736

The exponent of the co-efficient to get a odds ratio interpretation: exp (0.736) = 2.08 This means a user who goes through the referring website can have (2.08 -1) \* 100% = 108 % higher odds for booking the hotel through hotel website than if that user doesn’t go through the referring website.

Here the p-value of REF\_D is less than 0.05 it is significantly correlated with DIRECTP\_D.

**TRANS\_FREQ**

Based on our result, when there is increase in 1-unit TRANS\_FREQ, we expect the log odds of DIRECTP\_D to increase by 0.1156

The exponent of the co-efficient to get a odds ratio interpretation: exp (0.12) = 1.12 This means if a user who has 1 more transaction can have (1.12 -1) \* 100% = 12 % lesser odds for booking the hotel through hotel website, than if that user doesn’t not have 1 more transaction. Here the p-value of TRANS\_FREQ is less than 0.05 it is significantly correlated with DIRECTP\_D.

**DURATION**

Based on our result, when there is increase in 1- unit DURATION, we expect the log odds of DIRECTP\_D to decrease by 0.019

The exponent of the co-efficient to get a odds ratio interpretation: exp (-0.019) = 0.98 This means a user who has 1 more minute to shop can have (0.98 -1) \* 100% = -2 % lesser odds for booking the hotel through hotel website, than if that user doesn’t have 1 more minute. Here the p-value of

DURATION is less than 0.05 it is significantly correlated with DIRECTP\_D.

**CHILDREN\_D**

Based on our result, when there is increase in 1-unit CHILDREN\_D, we expect the log odds of DIRECTP\_D to increase by 0.256

The exponent of the co-efficient to get a odds ratio interpretation: exp (0.256) = 1.29 This means users who has child in household can have (1.29 -1) \* 100% = 29 % higher odds for booking the hotel through hotel website than if that user doesn’t have child.

Here the p-value of **CHILDREN\_D** is less than 0.05 it is significantly correlated with DIRECTP\_D.

* **1 (c) (Bonus: 10 points) Given the regression results, your interpretation, and your experience/research on internet shopping, what kind of improvement would you make on this model? e.g. IVs to be removed or new IVs to be added? Or even other regression methodologies. This question is designed to motivate your self-exploration beyond lecture/tutorial coverage**

### To improve the above model, I choose the following independent variables REF\_D, TRANS\_FREQ, PROD\_QTY, DURATION, PAGES\_VIEWED, CHILDREN\_D. Below are the observations/reasons for choosing these independent variables.

**•** We have created dummy variable as **REF\_D** for REF\_DOMAIN\_NAME. I infer that REF\_D which shows the referring website name through which the final purchase website was reached has high impact on DIRECTP\_D.

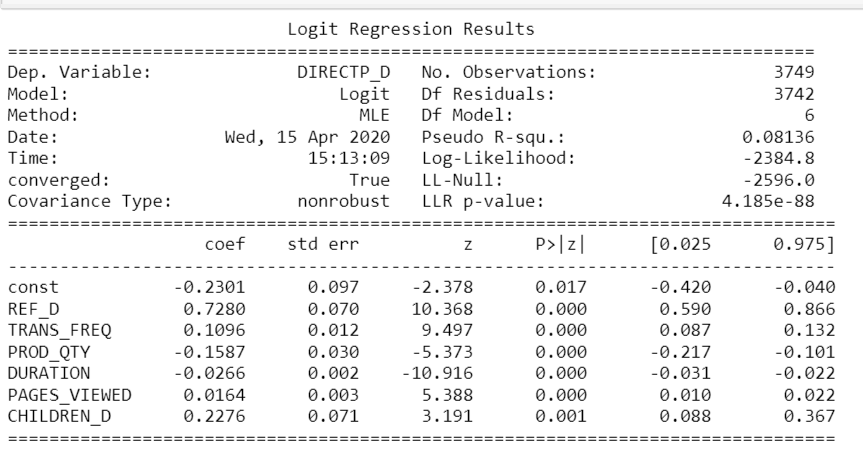
**•** I would like to add **TRANS\_FREQ** variable because if there are high number of TRANS\_FREQ we can say that customers are frequently booking the hotel. There is something called Hotel loyalty programs which is membership-only programs that almost every major hotel chain offers to all of their guests. These programs offer a number of benefits for their members and are completely free to join. Hotels offer these programs to encourage their guests to book direct instead of using third party booking sites. So many customers who stays at hotel frequently for work/ business purpose are member of these loyalty program to take advantages of cheaper rates, free cancellation, receive additional perks etc. and they end up using hotel website for bookings.

**•** I would like to remove **HOUSEHOLD\_SIZE, CONNECTIONSPEED\_D** from the above model because to my understanding and research total number of people in the household and whether the household has any good internet Connection speed or not are not significantly correlated with DIRECTP\_D.

**• PAGES\_VIEWED** should be one of the independent variables for DIRECTP\_D because on scientific level, there’s some web-browsing psychology at work here. Having a lot of relevant quality content on your website means visitors are likely to view a greater number of pages on your website, which mean they’re also likely to want to do business with you and book a room though hotel website.

**• DURATION** is one of the key metrics, which is correlated with DIRECTP\_D. People want to get the information they are looking in the website fast and easy so that they will find relevant results from which he/she will make a choice.

**• CHILDREN\_D** variable is also important, because user who have childrenalways prefer the hassle-free stay and child-friendly amenities and benefits, they trust more on hotel websites for direct booking.

****

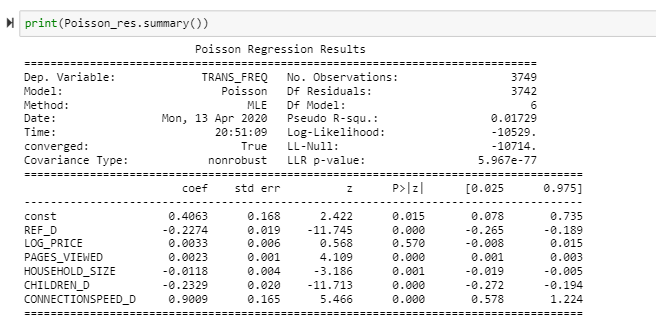
New Model Logit regression results

* In our new model we can see that p-value of all the variable is less than 0.05, which shows significant correlation with DIRECTP\_D.
* A pseudo R-squared when compared to another pseudo R-squared of the same type, on the same data, predicting the same outcome then the higher pseudo R-squared indicates which model better predicts the outcome.
* Based on above result summary of our new model shows that there is increase in Pseudo R-squared value, previously it was 0.07052 and to new model the value is 0.08136 which indicates our model is better now.

**------------------------------------------------------------------------------------------------------------------------------------------**

* **Ques 2 (b) Please report and interpret your Poisson regression results, which should include the interpretation of the regression coefficients.**

Poisson regression results

****

* In the above result**, LOG\_PRICE** have p-value greater than 0.05, hence it is *NOT* significantly correlated with TRANS\_FREQ.
* Coef. – It shows that a unit increase in the variables will either increase or decrease frequencies depending on whether the coefficient is positive or negative.

**Log(y) = β0 + β1 \* REF\_D +β2 \* PAGES\_VIEWED + β3 \* HOUSEHOLD\_SIZE + β4 \* CHILDREN\_D + β5 \***

**CONNECTIONSPEED\_D**

* Estimated Log(y) = 0.446 -0.227 \* REF\_D + 0.002 \* PAGES\_VIEWED - 0.012 \* HOUSEHOLD\_SIZE -0.233 \* CHILDREN\_D + 0.901 \* CONNECTIONSPEED\_D
* Our Dependent Variable is *TRANS\_FREQ* (Total number of transactions for the household) and a series of independent variables. This analysis shows the factors that are involved in determining frequency of booking the hotel. Form above Poisson regression result, we see that except LOG\_PRICE all other variables are statistically significant based on each p-value. This shows that these variables count towards people’s influence in making booking transactions.

**REF\_D**

Based on our result, if REF\_D increases by one unit, we can interpret as the change in the log of expected count (TRANS\_FREQ) will decrease by 0.23

Using exponent function on the coefficient to change the variable to actual number, exp (0.23) = 1.26 So, when a person who goes through reference website, could decrease the transaction frequency by (1.26-1) 100% = 26 %

**PAGES\_VIEWED**

Based on our result, when there is increase in 1-unit in PAGES\_VIEWED, we can interpret as the change in the log of expected count (TRANS\_FREQ) will increase by 0.002

Using exponent function on the coefficient to change the variable to actual number, exp (0.002) = 1.002 So, one more PAGES\_VIEWED, could increase the transaction frequency by (1.002-1) 100% = 0.2 %

**HOUSEHOLD\_SIZE**

Based on our result, when there is increase in 1-unit in HOUSEHOLD\_SIZE, we can interpret as the change in the log of expected count (TRANS\_FREQ) will decrease by 0.012

Using exponent function on the coefficient to change the variable to actual number, exp (0.012) = 1.012, which mean one more member in the household, could increase the transaction frequency by (1.012-1) 100% = 1.2 %

**CHILDREN\_D**

Based on our result, when there is increase in 1-unit in CHILDREN\_D, we can interpret as the change in the log of expected count (TRANS\_FREQ) will decrease by 0.233

Using exponent function on the coefficient to change the variable to actual number, exp (0.233) = 1.262 So, users who has child in household, could increase the transaction frequency by (1.262-1) 100% = 26.2 %

**CONNECTIONSPEED\_D**

Based on our result, when there is increase in 1-unit in CONNECTIONSPEED\_D, we can interpret as the change in the log of expected count (TRANS\_FREQ) will increase by 0.901

Using exponent function on the coefficient to change the variable to actual number, exp (0.901) = 2.462 So, users who has high speed internet connection, could increase the transaction frequency by (2.462-1) 100% = 146.2 %

### ------------------------------------------------------------------------------------------------------------------------------------------

### Ques 3. Please report and interpret your Negative Binomial regression mode which should include the interpretation of the regression coefficients.

### Negative Binomial results

### 

* In the above result**, LOG\_PRICE, HOUSEHOLD\_SIZE** have p-value greater than 0.05, hence they are *NOT* significantly correlated with TRANS\_FREQ.
* Coef. – Its shows same as Poisson regression, that a unit increase in the independent variables will change in the log of the TRANS\_FREQ.

**REF\_D**

Based on our result, REF\_D coefficient is negative, so if REF\_D is 1 it could lead to decrease in the TRANS\_FREQ by 0.23

Using exponent function on the coefficient to change the variable to actual number, exp (0.23) = 1.26, which mean when a person who goes through reference website, will decrease the transaction frequency by (1.26-1) 100% = 26 %

**PAGES\_VIEWED**

Based on our result, PAGES\_VIEWED coefficient is positive, so 1-unit increase could lead to a increase in the TRANS\_FREQ by 0.002

Using exponent function on the coefficient to change the variable to actual number, exp (0.002) = 1.002, which mean one more PAGES\_VIEWED, can increase the transaction frequency by (1.002-1) 100% = 0.2 %

**CHILDREN\_D**

Based on our result, CHILDREN\_D coefficient is negative, so if CHILDREN\_D is 1 it could lead to decrease in the TRANS\_FREQ by 0.238

Using exponent function on the coefficient to change the variable to actual number, exp (0.238) = 1.268, which mean users who has child in household, will decrease the transaction frequency by (1.268-1) 100% = 26.8 %

**CONNECTIONSPEED\_D**

Based on our result, CONNECTIONSPEED\_D, coefficient is positive, so if CONNECTIONSPEED\_D =1 it could lead to increase in the TRANS\_FREQ by 0.901

Using exponent function on the coefficient to change the variable to actual number, exp (0.901) = 2.462, which mean users who has high speed internet connection, will increase the transaction frequency by (2.462-1) 100% = 146.2 %

### ------------------------------------------------------------------------------------------------------------------------------------------

### Ques 4. (30 points) Please summarize your observations by comparing the results from 2 and 3

The main comparison between the two methods lays in the variance. Poisson regression assumes that the mean = variance, while Negative Binomial allows for mean < variance.

### In order to conclude, lets see some of the descriptive statistics of the dependent variable TRANS\_FREQ

### 

### Here, we can see that variance is much higher than mean so in this case we should choose the results of Negative Binomial regression

### The second comparison can be the p-values for each regression. In Poisson Regression we can see that only LOG\_PRICE p-value is not significant, rest all other variables p-value are significant. While for Negative Binomial Regression LOG\_PRICE and HOUSEHOLD\_SIZE both variable p-value is insignificant.