Huiyi Huang

ISYS 622

April 5, 2022

Professor Wenqi Zhou

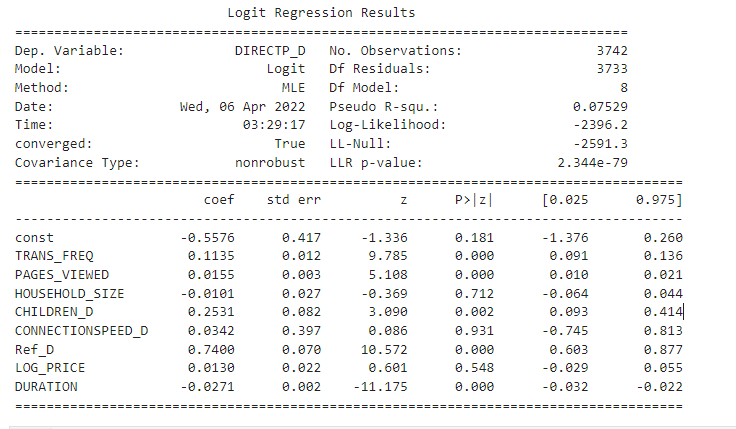
Project Phase 2

1. *It can be very obvious that more frequent purchases are desirable. What about direct purchasing from the hotel website, instead of third-party site/aggregator sites? Why do we care about the mechanism underlying the direct purchase? Love to hear your thoughts.*

Why customers purchase more directly on hotel webpage, there are some reasons, for example, loyalty program, the more loyal customers be, the more benefit they gain; standards, some customers with child want to live in a specific hotel, so they do not want to book it on the third-party website; being referred to, when they search specific hotel on Google, they can be referred to the hotel webpage to book, etc. There could be other reasons, when we understand better about it, we can understand the regression in a large picture. We have the assumption of causes and relationships, so we could throw them into regression to see if we could have better model on predicting the results, or analysis principal variables.

2. (a) *Please use DIRECTP\_D as your Dependent Variable (DV). For your Independent Variables (IV), you may consider, but not limited to, REF\_D, LOG\_PRICE, TRANS\_FREQ, DURATION, HOUSEHOLD\_SIZE, CHILDREN\_D, and CONNECTIONSPEED\_D. Feel free to add any descriptive statistics (e.g. over new vars that you didn’t look into in project #1, correlation matrix, etc.) into this analysis that may help you understand the context and the data as well as interpret the results.*

The first logistic regression result:



2. (b) *Please report and interpret your regression results, which should include the interpretation of each of the regression coefficients. what is this impact? Or is it simply an association, instead of causality? Why possibly do we observe this impact? What is this magnitude of this impact?*

* As the p-value shows, we can see that HOUSEHOLD\_SIZE, LOG\_PRICE, and CONNECTIONSPEED\_D are not statistically significant, because the p-value is larger than 0.05.
* The equation can be written as:

Log(p/1-p) = - 0.56 + 0.11\* TRANS\_FREQ + 0.02 \* PAGE\_VIEWED + 0.25 \* CHILDREN\_D + 0.74 \* REF\_D – 0.03 \* DURATION

The dependent variable is DIRECTP\_D, which means that whether a customer book directly on a hotel website or from other third party website.

Graphical user interface

Description automatically generatedThese are the exponential of the coefficients.

The calculation from odd ratio to percentages: (odd ratio-1)\*100%

TRANS\_FREQ : explained as 1 more purchase in a year will results in log odds of (DIRECTP\_D) increasing by 0.11. e^0.11 to get the odds ratio(1.12), meaning that if customer have 1 more transaction in a year, it will have 12% more odds to have them buy tickets on hotel website.

PAGE\_VIEWED : explained as 1 more page viewed results in log odds of (DIRECTP\_D) increasing by 0.02. e^0.02 to get the odds ratio(1.02), meaning that if customer have 1 more page viewed, it will have 2% more odds to have them buy tickets on hotel website.

CHILDREN\_D : explained as if a family have children it will results in log odds of (DIRECTP\_D) increasing by 0.25. e^0.25 to get the odds ratio(1.29), meaning that if customer have children, it will have 29% more odds to have them buy tickets on hotel website.

Ref\_D : explained as if a customer make orders on reference website it will results in log odds of (DIRECTP\_D) increasing by 0.74. e^0.74 to get the odds ratio(2.10), meaning that if customer make orders on reference website, it will have 110% more odds to have them buy tickets on hotel website.

DURATION: explained as if a customer makes orders on reference website it will results in log odds of (DIRECTP\_D) decreasing by 0.03. e^0.03 to get the odds ratio(0.97), meaning that if customer make orders on reference website, it will have 0.03% less odds to have them buy tickets on hotel website.

2(c) *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 trials and errors.*

IV I choose or remove:

Ref\_D: It is important that reference website is used, since customers can be referred to the hotel website and to make transactions. It is causality to the DV.

TRANS\_FREQ: I believe this could be a causality why customers want to book hotel on the hotel website, because there might be loyalty program to the specific hotel, and customers could gain further benefits for keeping loyal to the hotel.

PAGE\_VIEWED: This IV should be included, since it’s understandable that more pages customers looked, the more likely they will make orders on the hotel website page. I think it is causality not association.

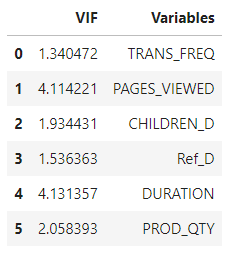
DURATION: Similar to PAGE\_VIEWED, the more time customers spend on the hotel web page the more likely they are going to book hotel on it. Causality too.

CHILDREN: I not sure the relationship, but I believed that a family has children or not have impact on making orders on hotel website. For example, if there is no children, a customer could choose either one of the hotel or third-party website. If they have children, they will consider more when booking hotel. This could turn them to hotel website and it could be the cause.

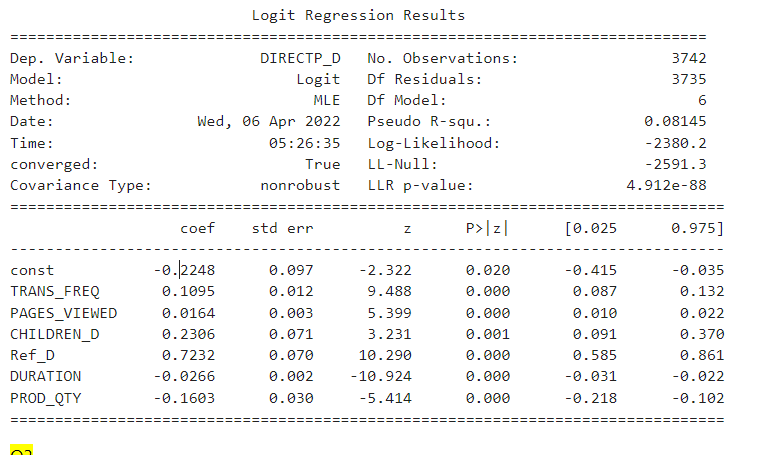
PROD\_QTY: this IV I want to add in because the more product customer want to buy and they still buying on the hotel website, for a reason is there is loyalty programs, so they buy more there could be more benefit. Same reason with keeping TRANS\_FREQ. It could be causality.

HOUSEHOLD\_SIZE, and CONNECTIONSPEED\_D: I removed these two, since I think them are only associative to the DV, and they are not statistically significant in regressing DIRECTP\_D.

LOG\_PRICE: I removed it, because I think it is only associative to the DV and it cannot influence people to book hotel on the website. Also, it is not statistically significant, so it’s better to drop it.

This is the VIF result for the 6 IVs I choose.

As the VIF result shows, all five variables are not highly multicollinearity to the rest of them, so there is no need to remove variable. And the final regression result is shown below.



First of all, IVs are all statistically significant. And there is no problem of high multicollinearity. Secondly, the R square is improving by 6%, from 75% to 81%, which means the model explain more variations comparing to the first regression.

Different from what I expected, PROD\_QTY’s coefficient is negative. Explained as, meaning that if customer buy 1 more thing in one order, it will have 14% less odds to have them buy tickets on hotel website.

3.(a) *Please use TRANS\_FREQ as your DV; and REF\_D, LOG\_PRICE, PAGES\_VIEWED, HOUSEHOLD\_SIZE, CHILDREN\_D, and CONNECTIONSPEED\_D as your IVs. Feel free to add any descriptive statistics into this analysis that may help you understand the context and the data as well as interpret the results.*

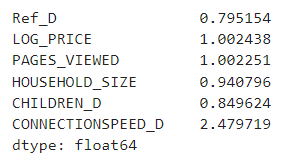
Table

Description automatically generated

* As the p-value shows, we can see that LOG\_PRICE is not statistically significant, because the p-value is larger than 0.05.
* The equation can be written as:

Log(lamda) = 0.51-0.23\* REF\_D +0\* PAGES\_VIEWED – 0.06 \*HOUSEHOLD\_SIZE – 0.16\*CHILDREN\_D +0.9\* CONNECTIONSPEED\_D

* The dependent variable is TRANS\_FREQ, which means that how many transactions customers made during 2011.

These are the exponential of the coefficients.

Ref\_D : explained as if a customer make orders on reference website it will results in log odds of (TRANS\_FREQ) decreasing by 0.22. e^-0.22 to get the odds ratio(0.79), meaning that if customer make orders on reference website, it will have decrease the transaction frequency by 21%. Maybe people do not like being referred to other web page to make transaction, it discourages the chance people could spend in 2011. It should be a causality.

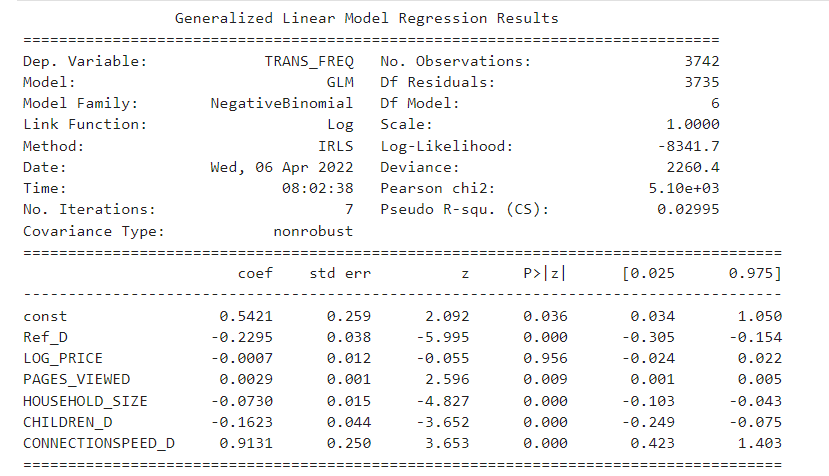
PAGE\_VIEWED : explained as 1 more page viewed results in log odds of (TRANS\_FREQ) increasing by 0.002. e^0.002 to get the odds ratio(1.002), meaning that if customer have 1 more page viewed, it will increase the transaction frequency by 0.2%. As you can see, the magnitude of this IV is very minimal. However, it could be a causality, since the more page customers viewed, the more likely customers will make the order.

HOUSEHOLD\_SIZE: explained as 1 more people live in the house, it will result in log odds of (TRANS\_FREQ) decrease by 0.06. e^0.06 to get the odds ratio(0.94) It means that one more people live in the house will decrease the transaction frequency by 6%. I think this is only associated with DV, there should no impact for how many people live in together on making travel plans.

CHILDREN\_D : explained as if a family have children it will results in log odds of (TRANS\_FREQ) decreasing by 0.16. e^0.16 to get the odds ratio(0.84), meaning that if customer have children, it will decrease 16% of the transaction frequency. This could be a causality, one of the reasons could be, when a family has a newborn baby, it is not easy for them to decide to go out, hence, the transaction rate can decrease.

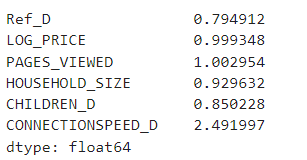
CONNECTIONSPEED\_D : explained as if a customer has high internet connection speed it will results in log odds of (TRANS\_FREQ) increasing by 0.9. e^0.9to get the odds ratio(2.48), meaning that if customer has high internet connection speed, it could increase the transaction frequency by 148%. This is a causality, since the fast internet could bring more transactions, all spending activity could be quick.

4. *Please repeat the analysis in question 3 using the Negative Binomial Regression model. Feel free to add any descriptive statistics into this analysis that may help you understand the context and the data as well as interpret the results.*



* As the p-value shows, we can see that LOG\_PRICE is not statistically significant, because the p-value is larger than 0.05.
* The equation can be written as:

Log(lamda) = 0.54-0.23\* REF\_D + 0.003\* PAGES\_VIEWED – 0.07 \*HOUSEHOLD\_SIZE – 0.16\*CHILDREN\_D +0.9\* CONNECTIONSPEED\_D

These are the exponential of the coefficients.

Ref\_D : explained as if a customer make orders on reference website it will results in log odds of (TRANS\_FREQ) decreasing by 0.23. e^-0.23 to get the odds ratio(0.79), meaning that if customer make orders on reference website, it will have decrease the transaction frequency by 21%. Maybe people do not like being referred to other web page to make transaction, it discourages the chance people could spend in 2011. It should be a causality.

PAGE\_VIEWED : explained as 1 more page viewed results in log odds of (TRANS\_FREQ) increasing by 0.003. e^0.003 to get the odds ratio(1.003), meaning that if customer have 1 more page viewed, it will increase the transaction frequency by 0.3%. As you can see, the magnitude of this IV is very minimal. However, it could be a causality, since the more page customers viewed, the more likely customers will make the order.

HOUSEHOLD\_SIZE: explained as 1 more people live in the house, it will result in log odds of (TRANS\_FREQ) decrease by 0.07. e^0.07 to get the odds ratio(0.93) It means that one more people live in the house will decrease the transaction frequency by 7%. I think this is only associated with DV, there should no impact for how many people live in together on making travel plans.

CHILDREN\_D : explained as if a family have children it will results in log odds of (TRANS\_FREQ) decreasing by 0.16. e^0.16 to get the odds ratio(0.84), meaning that if customer have children, it will decrease 16% of the transaction frequency. This could be a causality, one of the reasons could be, when a family has a newborn baby, it is not easy for them to decide to go out, hence, the transaction rate can decrease.

CONNECTIONSPEED\_D : explained as if a customer has high internet connection speed it will results in log odds of (TRANS\_FREQ) increasing by 0.91. e^0.91 to get the odds ratio(2.49), meaning that if customer has high internet connection speed, it could increase the transaction frequency by 149%. This is a causality, since the fast internet could bring more transactions, all spending activity could be quick.

5. *Please summarize your observations by comparing the results from 2 and 3 and provide the underlying explanations/justifications to common factors and unique influencing factors.*

* The NB regression’s coefficient is slightly larger than the Poisson one.
* The NB regression has higher R square value as 0.029, and Poisson has 0.019. The NB regression captured more variance than the Poisson.
* A screenshot of a computer

  Description automatically generated with medium confidence
* For Poisson model it asks for the mean = the variance, but in this dataset, they are not equal. Could it be the reason why NB model performs better?