Huiyi Huang

ISYS 622

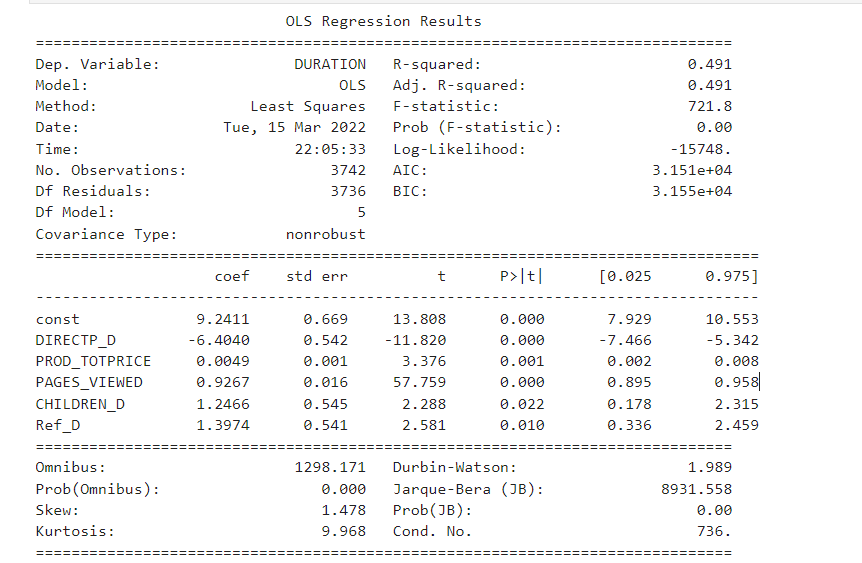
Mar 15th , 2022

Professor Wenqi Zhou

Project

PART A:

1. why do some people spend more time when booking hotels



1. Here is the final report for the linear regression of dependent variable: Duration. First of all, the equation needs to have a constant since the p-value is less than 0.05. And these five variables are statistically significant. Through the p-value of F-test, I can know that this linear regression is a better fit comparing to the intercept-only model. However, the R squared is not good. It means this model only captured almost 50% of variation.
2. The write up equation:

DURATION = 9.24 - 6.4\*DIRECTP\_D + 1.40\*REF\_D + 1.25\*CHILDREN\_D + 0.93\*PAGES\_VIEWED +0.005\*PROD\_TOTPRICE

The most important factor is DIRECTP\_D which means, if customer make transactions directly on website(1) or on third party website(0). The coefficient is -6.4. It means if a customer make purchase on website, will save 6.4 minutes on viewing the website. On the other side, a customer make transactions on third-party website, will spend 6.4 minuets on the web. As we all know third-party website might have complicated information and more choices; customers will completely spend longer time on that and support the linear regression model. The magnitude of the coefficient is larger than other factors and it means direct or not can make larger impact on the data.

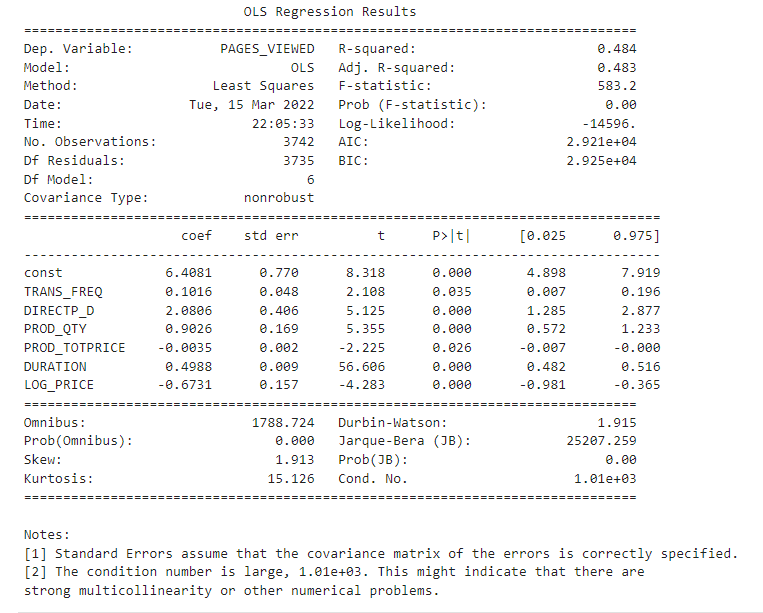
The second important factor is REF\_D, which means if a customer made purchases referred from other website(1) or not. The coefficient is 1.4. Customers made transactions being referred from other website will lead to spend 1.4 more minutes on viewing the page. This magnitude is not much but it caught part of the variations.

CHILDREN\_D, caught some variations too. It explained as: if a customer have any children, it can will make customers spend 1.25 minute more on the web. It is making sense, since people who have larger family should spend a little more time on choosing the hotel or booking a trip. The magnitude is not much.

PAGES\_VIEWED, explained as 1 customer view 1 more page, it will cost the customer 0.93 minutes more on the duration of viewing web. Although the logical is okay, but I think page viewed is not the cause of duration, and they should be in association.

PROD\_TOTPRICE, explained as 1 more dollar spend by customers will results in 0.005 minutes(0.3 seconds) of total viewing time for web. The impact is very minimal. Also, the predictor is in association with the duration not the cause of it.

**2.** why do some people visit more pages when booking hotels



1. The final set of independent variables I choose are: DIRECTP\_D , PROD\_QTY , PROD\_TOTPRICE , DURATION , LOG\_PRICE , TRANS\_FREQ, because they all have statistically significant p-value in the result table. I did not choose between these 6 variables, since the R-squared is only 0.484, and it will drop if I remove any of them. 0.484 of R-squared is not very good for capturing all the variations. However, the f-tests shows that this model is definitely fits better than the intercept-only model.
2. The write up equation:

PAGES\_VIEWED = 6.41 - 2.08\*DIRECTP\_D + 0.90\*PROD\_QTY -0.67\*LOG\_PRICE + 0.49\*DURATION+ 0.10\*TRANS\_FREQ -0.004\*PROD\_TOTPRICE

The most important factor is also DIRECTP\_D which means, if customer make transactions directly on website(1) or on third party website(0). The coefficient is -2.08. It means if a customer make purchase directly on website, he or she will have to see 2 pages less to make the final purchase. On the other side, a customer make transactions on third-party website, will have to see 2 more pages. As we all know third-party website might have complicated information and more choices; customers will have to view more pages. The magnitude of the coefficient is very small comparing to the mean of pages people needs to view(18 pages) , but larger than other factors in the model.

The second important factor is PROD\_QTY , which means the number of product purchased in this transaction. The coefficient is 0.90. Customers buy more products in one purchase will result in view almost 1 more page. This magnitude is not much but it caught part of the variations. I think when people trying to book both different hotels or flights, they will need to view more pages to make decisions. This predictor should be a causation.

LOG\_PRICE, explained as ln(1) more dollar spend by customers will results in view 0.67 page less. Similar to PROD\_TOTPRICE, means 1 more dollar paid, customer will have to view less 0.004 page. The magnitude of LOG\_PRICE is moderate, but minimal to PROD\_TOTPRICE. However, both of the predictors are homogeneous, and they are associated with the page viewed, not causes to page\_viewed.

DURATION, explained as 1 customer view 1 more minutes on the website, will result view half more page. The impact is moderate. But this is also associated with the number of page viewed, not the causation.

TRANS\_FREQ, explained as family who made 1 purchase in 2011, will have to see 0.10 more page. The impact is small but, I think it is reasonable. The more trip planned by a family, the more page the family needs to view. This predictor is causation.

1. In conclusion, whether people viewing on third party website or not influences both on total time spend at a site and how many pages people view. This is the reason for both. Differences:

Children and referred or not are influencing factors for the first model, but purchased product quantity and transaction frequency are influencing factors for the second model, except from association factors.

Reason for the difference could be:1) data’s characteristics. Duration is continuously variables, but page\_viewed is discrete. 2) Dummy variables. Based on the models I run I deduct that dummy variables capture variation more for continuous dependent variables. Quantitative variables capture variation more for discrete dependent variables.