methodology

Grigor Dimitrov

September 23, 2016

Introduction

The data utilized for this research has been collected from panelists who participate in a large consumer panel in the Netherlands. The data has two main components, behavioral data and survey data. The methodological section of this paper explains the sampling, data collection and analysis procedures that have been conducted, in the following way.

- Panel and sampling. Sampling based on behavioral data. A sample of respondents has been pulled out of the behavioral dataset based on their online behavior on popular tourism related websites in the Netherlands.
- Survey. An online survey has been administered among the sampled panelists to reveal the "unobservables" from the perspective of the available behavioral data. Namely, these include data points related to trip characteristics associated with respondents' latest tourist related purchases as well as question regarding their personality.
- Data processing.The full behavioral data of the respondents who successfully completed the survey has been sampled out of the full behavioral panel dataset.
- Categorization. All of the unique websites of the sampled behavioral dataset have been classified into categories. In such a way it was possible to assess whether a certain website was travel related or not.
- Analysis. The aggregated information from the behavioral data has been regressed over the survey data to reveal the impact the unobservable personality attitudes and traits on the tourism related online behavior while controlling for the trip characteristics.

In the following sections I will first focus on explaining the behavioral data and the technology behind it. Afterwards, the 5 major steps of the methodology are explained in details. Then I proceed with the descriptive results of each of the datasets (survey and behavioral). Finally, I report the analysis results.

Behavioral data & Technology

The behavioral data, also referred to as observational data, reflects the online behavior of the consumers. It consists of records of the interaction made via consumers' digital devices and the Internet.

The behavioral data has been collected via a technology developed and provided for this research by Wakoopa. The company is a provider of a tracking technology. The technology is utilized primarily for market research purposes. Similarly to the market research consumer panels, where panelists enroll to participate in online surveys for incentives, Wakoopa provides its technology to market research consumer panel companies that are interested in tracking the online behavior of their panelists. After enrolling into the panel and giving their consent to be tracked, panelists install an application on their devices i.e. desktop, mobile and tablet. The tracking software collects every interaction of the panelists' devices on the Internet which consists of path or the address the panelists are reaching and the duration of the visit. The software works in different manner over the different operation systems platforms and devices, but the final result consists of recording raw data containing events. Each event has the address the participant accessed, the duration of the interaction and the client requesting the information i.e. browser, app etc.

Steps

Panel and sampling

Using Wakoopa's panel it was possible to 'pre-screen' relevant respondents for the purposes of the analysis, that could be invited to participate in the survey I have conducted. The respondents that I was looking for, should have been active on tourism related websites and should have conducted a purchase on such websites in the period of January 2015 to June 2016. Herein, I describe the process that was used to reach such subjects. Upon starting of the project the panel used for the research had 6682 active panelists and 7103 active devices. The majority of the panelists were being active only on desktop. See *Appendix #1. Panelists* and *Appendix #2. Devices* for reference.

First, I started looking into tourism related websites in the Netherlands. Initially I used 300 domains to account for the majority of the tourism related internet traffic in the Netherlands according to the internet analytics company SimilarWeb. Afterwards, based on this data I exported the activity of the whole panel over those websites ranging from 01/2015 to 06/2016. The data of this activity was manually analysed looking for the end pages of the payments, also referred to as "confirmation" pages or "thank you" pages. Confirmation pages are the pages where a customer has been redirected after conducting a purchase at a company website. In general, the analysis consists of looking up keywords within the travel domains URLs and marking down the common patterns with the aid of regular expressions (See @mitkov2005oxford for discussion). Identified were the patterns of "confirmation" pages and data for the participants visited from the period of 01/03/2016 to 16/06/2016 was exported. Using this information, it was possible to identify a sample composed of 949 respondents which were to be invited into the online survey as described in the second step below.

Given the estimation of the incidence rate provided by the panel supplier, 20%, and the initially desired sample of 500 respondents more respondents were needed. Therefore, an additional random sample of 123 respondents was selected based on whether they were active on the 300 initial domains but without evidence for their purchases from the data.

Survey

The selected 1039 panelists were invited to participate in an online survey, that aimed to reveal their attitudinal and personal characteristics as well as the trip characteristics of their last travel. The fieldwork was conducted during the last week of June 2016. Out of the 1039 invitations sent 872 started the survey, which resulted in 495 completed interviews. (Note to self, move this to a separate cleaning section). The data was further cleaned by accounting for speeders and flatliners. Speeders were panelists that have finished the survey within less than half of the average length of the interview. Flatliners refer to respondents who have answered all of the grid questions in a straight line. Moreover, a few respondents attempted to participate in the survey multiple times and thus, they were excluded from further analysis. Thus, the final dataset consists of 426 observations.

The survey consisted of several parts. First, a screening criteria was used. It accounted for the number of travel related purchases since March 2016. Respondents with no travel related purchases were not allowed to further proceed. Then, subjects who have not purchased neither flight nor accommodation were also excluded from the questionnaire. Business only travellers were not relevant for the analysis, thus they were also screened out, leaving only panelists who have gone on a leisure trip or both on a leisure and business trip. The second part was a demographics section, which consisted of question regarding age, gender and income. The next part of the questionnaire, was in regards to the trip characteristics, which was adapted from @Roehl1992 and contained information about moment of booking, destination, planning horizon, information sources used i.e. Internet, advice from friend and relatives, tourist information office, travel agent etc., products purchased online, duration of the trip or number of nights spent away from home, whether the destination was visited before and how many times, number of travel companions, indication whether there was children on the trip and whether subjects visited friends or relatives during their trip.

The next two sections aimed to reveal more about the subjects personality and risk and uncertainty attitude. Items assessing risk and uncertainty originate from @Quintal2009. Respondents reported their risk attitude

on three item scale which the authors adapt from @donthu1996infomercial. Uncertainty attitude has been assessed on four items scale which authors adapt from @yoo2002testing and the scale is based on @hofstede1980motivation UA items. I have chosen these scales as they are the shortest reliable scales for self-assessing the risk and uncertainty attitudes. Using such scales I can make sure to not tire respondents and keep them engaged. Personality traits related to openness, consciousness, extraversion, agreeableness and neuroticism have been assessed using the short scale from the big five inventory proposed by @Gosling2003. In addition to the ten item scale an eleventh item was added as XXXX doesn't perform well in the ten item scale. The eleven items were derived from a validated Dutch translation of the BIG5 inventory from @denissen2008development.

Data processing

After having collected and cleaned all the data, I further proceed with data processing and analysis. All of the data processing and the analyzation tasks of this paper were done in R Studio. All of the code, including libraries can be seen in the appendix.

The data processing includes merging down the two main behavioral streams of data that originate from desktop and mobile. Additional to that it is needed to classify and derive all the tourism related domains that will be further used for analysis. This is an important step as it will allow to analysis what the relationship between the online travel behavior and the traveller's attitudinal characteristics are. Categorization of the domains was done using machine learning classification algorithms provided by uClassify.com. The domains from the desktop and mobile behavioral data and all of the mobile apps are further classified using keywords.

Next step is, converting the raw data set onto micromoments. They are also further used in the analysis as I examine the relationship between one's micromoment frequency and subject's risk and uncertainty attitude. Micromoments are essentially user sessions where users were active within a given moment of time. For example, if a panelist access a certain domain and spend five minutes on it and then become inactive for more than 3 minutes, this will result in a factor variable grouping all of the observation within those five minutes of activity. If a travel related website has been visited during the micromoment, a dummy variable is assigned to this moment to indicate this. Furthermore, purchases of travel related products were included in the data based on the initial dataset used for pooling the sample out the panel for the survey. The final dataset includes aggregated information about panelists' activity over two main levels. 1. top-level i.e. total activity 2. low-level micromoment level.

This section includes the data processing tasks done on the behavioral data including descriptive statistics and variable derivation, classification and aggregation on the different levels intended for further analysis.

Starting point of the analysis of the behavioral data include procession the data and rendering it in a format suitable for running the analysis. In its initial form the data has been exported in a format containing the following variables: for the desktop data (sample data, see: *Appendix #3 Desktop data*) and for the mobile data (sample data, see: *Appendix #4 Mobile data*). The next step is adding *"host"* variable to both datasets. The host variable contains domains and the subdomains that are going to be used to for the categorization whether the domains were travel related or not. The next query on the data included identifying all of the micromoments in the data. First, the data has been subsetted on *"panelist_id"* level, sorted by the the timestamp *"used_at"* and assigned into list where each element is the full data for each individual respondent. The sum of *"used_at"* variables and *"active_seconds"*, the duration in seconds respondent spend on the page, were compared to *"used_at"* variable of the next observation. If differences larger than five minutes were found all of the variables prior to this difference were grouped together under a common factor variable.

Table 1: Desktop dataset

used_at	host	panelist_id	url	active_seconds	browser_name	mmid	Class_Travel	purchase

Table 2: Mobile dataset

panelist_id	device_id	scheme	url	domain	app_id	app_name	used_at	connection	duration	h

Categorization

The categorization procedure includes using uClassify machine learning algorithms and also using keywords. The full activity coming from desktop and mobile devices resulted in 194,534 unique domains. The categorization algorithm has been responsible for classifying all domains that have more than ten visits or 47,818 domains in total works as follows. A web-scraper designed for these projects accesses a the collection of domains and collects all of the information on the page, then removes the HTML elements along with the punctuation and renders down the information only to a part that is visible to the website visitors. Then, it passes this information to an application programmable interface (API) that returns the probability of this text being into sixteen different categories including travel.

The full list of categories includes:

NULL

Websites with content with the highest probability to be travel were assigned value of a dummy variable 1 or 0 otherwise.

Due to technical and time contains the API service wasn't able to classify all of the domains. Therefore, the full dataset of unique domains names has also been scanned for keywords. Keywords include travel, tourism, accommodation, hotels, flight etc. (full list of the keywords can be found in the code in the appendix). If any of the keywords appear within the domain name, the respective domain has also been assigned to the list of travel domains. The same keyword approach has been used for classification of the application on mobile devices, where the apps have been classified using their names.

Analysis

The following section contains an explanation of the main techniques performed during the analysis along with their assumptions, followed by the results of the ordinary least squares diagnostic tests of the restricted model. Once the functional form of the restricted model has been selected I proceed with variable selection in order to come up with the final model. Finally, I ran the diagnostics of tests over the final model again.

Factor Analysis

Factor analysis is a widely used technique used for for explaining the variance in several variables by smaller set of latent variables. As in the current case it is often used to consolidate several survey variables onto their "underlying" factors in order to reduce the dimensionality of the data. Factor analysis groups variables together, that is, using a large amount of variables one can potentially reduce them to certain factors representing the latent underlying factors representing them by accounting the similar patterns in the variables. The intuition behind the anaysis is as follows. The analysis groups together observed, correlated variables into smaller groups of unobserved (latent) variables [@yong2013beginner].

In this case I use factor analysis to to reduce the seven survey items rearding the risk and uncertainty attitude down to two constructs namely risk and unertainty. Also to reduce the eleven item scale of BIG5 to 5 factors representing each of on the of the five personlity traits.

Regression Analysis

For testing the hypotheses of this paper, regression analysis will be utilized. The regression model or ordinary least squares (hereafter OLS) is the "cornerstone of econometrics" [@verbeek2008guide]. It aims at explaining a variable, y, in terms of another variable, x. In other words, using OLS researchers are able to find how will y vary as x changes, the ultimate goal being to infer the causal effect x has on y. Using such models allows to find relationships between various variables, present the effect the independent variables, xi have on the dependent variable, y in order to be able to make predictions.

The general linear regression models is represented as follows:

$$y = \beta_0 + X_1\beta_1 + X_2\beta_2 + ... + X_k\beta_k + \varepsilon$$

Where: y is the dependent variable

X_1 to X_k are the independent variables, which explain y

β_0 is the intercept, indicating the expected value of y when all the independent variables are equal to 0

β_1 to β_k are the coefficients which determine the effect x has on y

ε is the error term

Goodness of fit and model selection

The standard measures of fit include the R-squared and the adjusted R-squared, which measures the variance that is explained in the model for the independent variable by the dependent variables. The measure can be interpreted directly. For example if the R-squared is equal to 0.45, it means that the variables included in the model explain 45% of the variation of the independent variable, y. The higher the value, the higher its predictive power. However, it should be noted that adjusted R-squared penalizes for the additional number of parameters. Thus, applying additional variables to the model, I should test if they are jointly significant in order to assess whether they are relevant or not in the model.This is typically applied by using the F-test (Wald test). Using both the R-squared, the overall F-test and applying the F-test to certain variables I can compare best which model fits the data best.

Akaike information criterion (AIC)

Model selection has been done over Akaike information criterion (AIC) introduced by @akaike1974new. AIC is a metric traditionally used for model selection. It compares the goodness of fit for a number of explanatory variables and penalizes for each additional explanatory variable.

BLUE Assumptions

There are several assumptions that need to be met when applying OLS explained in the section below. Namely Gauss-Markov assumptions for full ideal conditions for OLS. The model needs to be best linear unbiased estimator" ("BLUE") [@verbeek2008guide]. It is crucial for the assumptions to be met as to compute unbiased and consistent estimates that explain the variation in the dependent variable. Now, I will go through each assumption: Linear in parameters This implies that the model should have linear parameters, b, however, there can be nonlinearities in the variables, x. This assumption is met as my specified model does not include non-linearities in the parameters.

Normality

The error term's should follow a normal distribution. In large datasets, however, even if the error term does not follow a normal distribution the regression estimators are 'asymptotically normally distributed', meaning that following non-normal distribution is not crucial as the estimates will still be consistent and unbiased. The Shapiro-Wilk test can is adopted here and results presented below. The test works under null hypothesis: "the sample comes from normally distributed population" @shapiro1965analysis

Random sample

The data collection should be done randomly, meaning that the each subject should have the same probability of being selected. In this research, both in the behavioral and survey data collection parts, I can say that subjects were randomly selected for further analysis.

Multicollinearity

Multicollinearity implies that there is no perfect linear relationship between the independent (explanatory) variables as this can lead to 'unreliable regression estimated' [@verbeek2008guide]. For example, adding both male and female in the analysis would lead to perfect collinearity (as male + female =1) and the estimations would not work. In this example, removing one of the variables would solve the problem, however there can be other variables that are highly correlated. Having multicollinearity would not lead to biased estimates, but to inaccurate estimates. In such a case, excluding variables from the model should be considered. There are no tests that specifically look for multicollinearity, however there are certain indications. For instance, having two variables that are jointly significant (have big F-statistics), but independently are not significant can be a sign of multicollinearity

Homoscedasticity

Homoscedasticity implies that the variance of the error term should be the same for all values of the independent variables. If this does not hold, there is problem with heteroscedasticity meaning that the estimates of the regression are inconsistent due to inaccuracy of their standard errors, meaning that the t-statistics and thus the significance level of the estimates is not valid anymore. To test for homoscedasticity, I perform the Breusch-Pagan test, which hypothesize that there is constant variance of the error terms. Endogeneity The last assumption is crucial to be met as otherwise the regression estimates are biased and inconsistent. Endogeneity implies that there is correlation between an independent variable and the error term. There are several reasons why this assumption does not fold: 1 The model is misspecified. That is, nonlinearities are missing from the model or interaction effects are not accounted for. To account for that I perform the Ramsey-Reset test. The tests adds fitted values on power and re-estimates the model. The intuition behind it is that if non linear combination of independent variables can explain the dependent variable there are evidence the model is misspecified. The Ramsey-Reset test work under null hypothesis that the model has no important omitted non-linearities [@ramsey1974classical] 2 Endogeneity, meaning that we are either missing important variables that explain the variance in the independent variables or we have reverse causality, that means that there can be a loop of causality between the independent and dependent variable.

Stepwise regression

The idea of stepwise regression has been introduced by @hastie1992statistical and further improved by @ripley2002modern. It is an iterative function ran over a restricted model and a set of candidate models. Each candidate model consists of a different set of explanatory variables. The function computes iteratively Akaike information criterion (AIC) values for the models comparing them to the best performing models from the previous iteration and based on the performance chooses whether to continue the loop with the new model or remain with the old one. The final output is the best performing model.

Model selection

Restricted Model

There are two distinct sources of data resulting in three data sets to be investigated.

1. Dataset from desktop
2. Dataset from mobile

3. Combined dataset from desktop and mobile

Furthermore, there are five dependent variables that are intended to be examinated during this research. Namely,

1. Number of travel micro-moments
2. Number of unique travel domains visited
3. Number of travel pageviews
4. Total time in seconds spent on travel domains
5. Total length in seconds of travel micro-moments

The difference between the last two is as follows. Total time in seconds has been measured once a panelist arrives at a certain website that has been classified as travel, whereas length is the total length (i.e. the difference between the start and end) of micro-moments amongst which a participant visited a travel related domain.

The restricted models of each one of the three datasets take the following form:

Table 3: Restricted model

Model	Dependent variables	Independent variables	Control *	Control2 **
# 1	Number of travel micro-moments	Risk + Uncertainty + Interaction	Total micro-moments	Total purchases + Device + Days active in the panel
# 2	Number of unique travel domains visited		Total domains	
# 3	Number of travel pageviews		Total pageviews	
# 4	Total time in seconds spent on travel domains		Total time	
# 5	Total length in seconds of travel micro-moments		Total micro-moments length	

* As there was only one purchase detected on mobile devices this term wasn't present in mobile model
** The variable device is present only in the combined model. It has value 1 is a respondent has been active only on desktop 2. Only on mobile and 3 if the respondent has been active on mobile and desktop. Obviously, in the models containing only desktop and only mobile data the variable has no variance therefore it is excluded. Since the respondents are selected only based on their desktop behavior there are no respondents who has been active only on mobile. Therefore the device variable in the combined dataset has only values 1 and 3.

Running Ramsey-Reset test per each on of the basic models showed evidence that the functional form of the models is not well specified, thus the models were rejected. Consequently, all of the dependent variables along with their corresponding controls were transformed into logs accounting for additional five models per each dataset. The results from consequent running of Ramsey-Reset test still were not satisfying, as the test still showed significant results indicating that the functional form of the model is not well specified. Consequently, the corresponding control variables accounting for total activity and purchases were transformed into binary variables with values 1 indicating a participant belongs to a highly active groups of participants (having values above the mean of the sample) and 0 vise versa. The latter transformation accounted for investigating ten more models per each dataset reaching the total number of thirty models to be reviewed. The functional

form of the model has been selected based on the dataset containing desktop data due to the following reasons: it is the most complete in terms of number of observation, the participants were selected based on their desktop behavior.

Based on Ramsey-Reset test, I select the following functional form of the model for Desktop, Mobile and Combined datasets. Shapiro-Wilk test for normality also performs best for this functional form, yet the sample is large enough so we can relax the normality assumption.

Model	Dependent variables	Independent variables	Control *	Control2 ** ***
# 1	log(#) of travel micro-moments	Risk + Uncertainty + Interaction	(D) Total micro-moments	(D) Total purchases + Device + Days active in the panel
# 2	log(#) of unique travel domains visited		(D) Total domains	
# 3	log(#) of travel pageviews		(D) Total pageviews	
# 4	log(#) time in seconds spent on travel domains		(D) Total time	
# 5	log(#) length in seconds of travel micro-moments		(D) Total micro-moments length	

* Note: as there was only one purchase detected on mobile devices this term wasn't present in mobile model.
** Note: The variable device is present only in the combined model. It has value *1* if the respondent has been active only on desktop *2* only on mobile and *3* if the respondent has been active on mobile and desktop. Obviously, in the models containing only desktop and only mobile data the variable has no variance therefore it is excluded. Since the respondents are selected only based on their desktop behavior there are no respondents who has been active only on mobile. Therefore the device variable in the combined dataset has only values *1* and *3*.
*** Dummy variables take value 1 is the respondent is activity is above the average activity of the sample and 0 otherwise.

A table with the results from the restricted model along with the model performance metrics can be found in the appendix:

```
##
## ================================================================================
##                                        Dependent variable:
##                                -------------------------------------------------
##                                log(MM)  log(Domains) log(PV)  log(Time) log(Length)
##                                  (1)        (2)        (3)       (4)        (5)
## --------------------------------------------------------------------------------
## Risk:seek                       -0.16     -0.25**    -0.42*** -0.47***    -0.28**
##                                 (0.12)     (0.12)     (0.13)   (0.13)     (0.12)
##
## Uncertainty:seek                -0.15     -0.26**    -0.31*** -0.32***    -0.21*
##                                 (0.11)     (0.11)     (0.12)   (0.12)     (0.11)
##
## Days                           0.01***    0.01***    0.004*** 0.004***    0.01***
##                                (0.001)    (0.001)    (0.002)   (0.002)    (0.001)
```

```
##
## D MM                              1.17***
##                                    (0.10)
##
## D Domains                                    1.08***
##                                               (0.10)
##
## D PV                                                    0.88***
##                                                          (0.11)
##
## D Time                                                                  0.86***
##                                                                          (0.11)
##
## D Length                                                                                  1.34***
##                                                                                           (0.10)
##
## D Purchase                        0.50***    0.62***    0.86***   0.87***    0.48***
##                                    (0.11)     (0.11)     (0.11)   (0.11)     (0.10)
##
## Risk x Uncertainty                  0.27       0.32      0.66**   0.83**      0.48
##                                    (0.30)     (0.31)     (0.32)   (0.32)     (0.29)
##
## Constant                          3.12***    3.59***    5.66***   8.94***   10.99***
##                                    (0.27)     (0.27)     (0.29)   (0.29)     (0.26)
##
## ------------------------------------------------------------------------------------
## Observations                        426        426        426       426        426
## R2                                  0.36       0.35       0.30      0.31       0.41
## Adjusted R2                         0.35       0.34       0.29      0.30       0.41
## Residual Std. Error (df = 419)      0.97       0.98       1.02      1.03       0.93
## F Statistic (df = 6; 419)        39.89***   37.21***   29.69*** 31.07***   49.37***
## ====================================================================================
## Note:                                                    *p<0.1; **p<0.05; ***p<0.01
```

Table: Desktop data, restricted model

```
##
## ================================================================================================================
## Residual_Standard_Err F.Stat NumDF FDenDF R.Sq Adj.R.Sq Shapiro_Wilk_Stat Shapiro_Wilk_P.val Reset_St
## ----------------------------------------------------------------------------------------------------------------
## 0.97                  39.89    6    419   0.36   0.35         1.00               0.94           2.03
## 0.98                  37.21    6    419   0.35   0.34         1.00               0.24           1.20
## 1.02                  29.69    6    419   0.30   0.29         0.99               0.16           0.79
## 1.03                  31.07    6    419   0.31   0.30         0.98              0.0001          1.31
## 0.93                  49.37    6    419   0.41   0.41         0.99               0.01           1.97
## ----------------------------------------------------------------------------------------------------------------
```

Table: Desktopr data, restricted model tests

```
##
## ================================================================================
##                                        Dependent variable:
##                                    ---------------------------------------------
```

```
##                                log(MM)  log(Domains) log(PV)  log(Time) log(Length)
##                                  (1)        (2)        (3)       (4)        (5)
## ------------------------------------------------------------------------------------
## Risk:seek                       0.36       0.29       0.42     0.96       1.26
##                                (0.39)     (0.31)     (0.42)   (0.91)     (0.92)
##
## Uncertainty:seek               0.84**     0.66**      0.54     1.49*     2.26**
##                                (0.38)     (0.30)     (0.41)   (0.89)     (0.89)
##
## Days                           0.01**    0.01***     0.004    0.02***    0.02***
##                               (0.003)    (0.002)    (0.003)   (0.01)     (0.01)
##
## D MM                          1.50***
##                                (0.36)
##
## D Domains                               2.28***
##                                          (0.29)
##
## D PV                                               2.55***
##                                                     (0.40)
##
## D Time                                                        3.09***
##                                                               (0.82)
##
## D Length                                                                 3.13***
##                                                                          (0.82)
##
## Risk x Uncertainty             -2.50      -2.07      -2.07    -7.86**    -8.74**
##                                (1.58)     (1.25)     (1.72)   (3.71)     (3.75)
##
## Constant                        0.10       0.03       0.37     1.60       1.58
##                                (0.46)     (0.36)     (0.50)   (1.08)     (1.08)
##
## ------------------------------------------------------------------------------------
## Observations                     101        101        101      101        101
## R2                               0.36       0.55       0.42     0.37       0.36
## Adjusted R2                      0.33       0.52       0.38     0.33       0.33
## Residual Std. Error (df = 95)    1.50       1.19       1.64     3.52       3.56
## F Statistic (df = 5; 95)       10.84***   22.94***   13.49*** 10.99***   10.72***
## ====================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

Table: Mobile data, restricted model

```
##
## ==========================================================================================================
## Residual_Standard_Err F.Stat NumDF FDenDF R.Sq Adj.R.Sq Shapiro_Wilk_Stat Shapiro_Wilk_P.val Reset_St
## ----------------------------------------------------------------------------------------------------------
## 75.41                  1.89    5     95   0.09   0.04         0.46                0              0.41
## 211.98                42.72    5     95   0.69   0.68         0.65                0            205.85
## 59.27                 10.91    5     95   0.36   0.33         0.56                0             25.23
## 106,753.30             5.67    5     95   0.23   0.19         0.38                0             15.69
## 157,353.90             3.87    5     95   0.17   0.13         0.40                0              3.63
## ----------------------------------------------------------------------------------------------------------
```

Table: Mobile data, restricted model tests

	Dependent variable:				
	log(MM)	log(Domains)	log(PV)	log(Time)	log(Length)
	(1)	(2)	(3)	(4)	(5)
Risk:seek	-0.20	-0.20	-0.35***	-0.38***	-0.19
	(0.13)	(0.13)	(0.13)	(0.14)	(0.15)
Uncertainty:seek	-0.13	-0.15	-0.23*	-0.19	-0.17
	(0.12)	(0.12)	(0.13)	(0.13)	(0.14)
Days	0.01***	0.01***	0.01***	0.01***	0.01***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
D MM	0.91***				
	(0.11)				
D Domains		1.01***			
		(0.10)			
D PV			0.82***		
			(0.11)		
D Time				0.77***	
				(0.12)	
D Length					0.98***
					(0.12)
D Purchase	0.59***	0.62***	0.90***	0.85***	0.51***
	(0.11)	(0.11)	(0.12)	(0.12)	(0.13)
Risk x Uncertainty	0.25	0.27	0.58*	0.58	0.48
	(0.32)	(0.31)	(0.34)	(0.35)	(0.37)
Constant	2.67***	3.16***	5.08***	8.10***	10.01***
	(0.29)	(0.28)	(0.30)	(0.31)	(0.32)
Observations	429	429	429	429	429
R2	0.30	0.34	0.30	0.29	0.29
Adjusted R2	0.29	0.33	0.29	0.28	0.28
Residual Std. Error (df = 422)	1.03	1.00	1.07	1.13	1.16
F Statistic (df = 6; 422)	30.57***	36.53***	29.52***	29.22***	29.27***
Note:				*p<0.1; **p<0.05; ***p<0.01	

Table: Combined data, restricted model

```
## ==========================================================================================
## Residual_Standard_Err F.Stat NumDF FDenDF R.Sq Adj.R.Sq Shapiro_Wilk_Stat Shapiro_Wilk_P.val Reset_S
## ------------------------------------------------------------------------------------------
## 0.87                  70.65    6    422   0.50   0.49         1.00               0.36          10.42
## 0.81                  90.89    6    422   0.56   0.56         0.98             0.0001           5.80
## 0.91                  68.73    6    422   0.49   0.49         1.00               0.28           2.72
## 0.97                  62.99    6    422   0.47   0.46         0.97             0.0000          12.86
## 1.06                  49.46    6    422   0.41   0.40         0.75                  0           1.30
## 267.80                12.88    6    422   0.15   0.14         0.73                  0           2.48
## ------------------------------------------------------------------------------------------
```

Table: Combined data, restricted model tests

Final Model

Based on the selected functional form of the model(s) I proceed with stepwise selection in order to select the final model across desktop, mobile and combined dataset. First, the models are presented then results of the OLS performance tests are shown and discussed.

Results: https://www.dropbox.com/s/x6y4xl3sh4elr3u/Screenshot%202016-09-15%2018.58.17.png?dl=0 https://www.dropbox.com/s/fas9fw1y5jgo1pz/Screenshot%202016-09-15%2018.58.34.png?dl=0 https://www.dropbox.com/s/axaanyn7mfpctbt/Screenshot%202016-09-15%2018.58.45.png?dl=0

Tests: https://www.dropbox.com/s/dsl9e1zhton9w02/Screenshot%202016-09-15%2018.57.21.png?dl=0