Steps to process data

Updated May 26, 2013 by Qui Nguyen

Step 1: Get data

ssh into 50.116.35.140 (our linode with the database on it)

username: student, pwd: usual

run **sudo mysqldump -u root -p -T /tmp pmma --skip-tz-utc**

* This copies the contents of the database into /tmp
* You can then copy the files onto your computer using scp or other file transfer method
* Most of the files are unnecessary; you only need the txt files corresponding to the tables in the database (stimuli, users, users\_stimuli\_startstop, users\_stimuli\_enforcement, users\_stimuli\_actions, users\_stimuli\_browsinghistory, users\_stimuli\_order, errors)

Get data from Qualtrics, in csv format

Step 2: Process MIT data

1. Remove extra startstop entries as follows

* Order startstop file by startstopid, reverse order (can be done with Excel). Save as users\_stimuli\_startstop\_sorted.txt
* In the same directory, run startstop\_clean.py to get users\_stimuli\_startstop\_clean.txt. Later steps will use this file.

2. Delete all test ids from data files, if they are present (\*\*Optional)

3. Open Stata and do the following .do file: mit\_data\_clean1.do

Read the comments. Some things might have to be changed a little bit.

4. Run count\_actions.py and count\_enforce.py. These scripts will use the actions and enforcement files to count how many times people took actions and were enforced, and put the information into actions\_count.txt and enforcement\_count.txt.

5. Use Stata to run mit\_data\_merge.do.

Should delete parts referring to s4, since there were only 3 stimuli in the US.

6. Use Stata to run mit\_data\_clean2.do. Again, small modifications might have to be made. Now we have the final Stata file with all the MIT data, and a csv with the data as well.

7. Clean the browsing data by taking out test ids (optional). Save as browsing\_clean.txt. Next, run browsing\_format.py to format the browsing file into browsing\_format.txt.

Step 3: Merge with Qualtrics data

1. Done in Stata with something like mit\_gongos\_merge.do. You’ll need to read in all the columns from the Qualtrics csv with the variable names you want.

* If you need to clean the data or remove unnecessary columns from the Qualtrics file, you should do that before this step.
* Then, merge the data with the final Stata file you got above in Step 3.7.
* Now you have a final data file, both in Stata (mit\_gongos\_merge.dta) and in txt format (mit\_gongos\_merge.txt). Feel free to change the file names since it’s not Gongos anymore....

Step 4: Add involvement measures

1. Basically, write some python scripts to process the browsing/action files to create whatever involvement variables you want. You can see what I did in the python files (google\_actions, edmunds\_actions, etc). For the China pretest, I just did some really basic measures for both Weibo and Baidu - if they clicked on any part of the story.

2. Then read those new variables in and merge them into the final data file, using Stata. Example commands in data\_involvement.do, though they will have to be adapted.

Step 5: Put data in SAS for analysis

1. Export the data from Stata to SAS. First, use saveold in Stata to get a Stata 9 file. Then you can use StatTransfer on equity to convert it into SAS.

2. Once you have a SAS file, add premc and postmc (pre Chevy model consideration and post). See sample code in \_\_\_\_.

* For the Chinese data, since model names were in Chinese, I replaced the Chinese with English for the models of interest (Malibu, Cruze, Aveo) to make this step easier.