**PROJECT: RTC etc Paper**

**PURPOSE: Explain contents of Empirical\_Examples Folder**

**AUTHOR: Maddie McKelway**

**DATE: March 3, 2016**

This folder contains subfolders for each project we’re hoping to use as an empirical example in the paper. Each folder will contain materials provided by the original authors themselves as well as materials we’ve used to prep them. Prepping involves the following steps. I found it helpful to go in this order, but certainly not necessary to stick to this order as long as each step is completed!

1. Read the original paper and familiarize yourself with the replication materials the authors provided. Identify the main results of the paper (this will usually be results from a regression of selected outcomes on a treatment dummy and other control variables). Then compile the data/code needed to replicate these results exactly.
2. Find the questionnaire given at baseline and the datasets containing the entire set of baseline responses to all questions and for all subjects. This data represents the full set of potential controls for this study. Note much of this data is likely to be raw and untouched by the original authors. Datasets and questionnaires should be in the subfolders here but if they are not contact the authors.
3. Prep the baseline data for analysis. Ensure that everything you do is well documented and easily reversible/corrected. To do this well, you’ll need to be familiar with the survey. You’ll need to do the following:

* You’re only looking at baseline responses.
* Replace numerical responses like “6”, “9”, “99”, etc. that often represent missing values in survey data with “.”
* Ensure all variables are numeric. See note below about how to handle string variables.
* Ensure no variable names start with "\_".
* Ensure all variables are uniquely determined by the first 12 characters of their name.
* In general, each question on the survey should correspond to a control variable. But there are a few important exceptions I’ve found are important to look out for:
* String variables.
* Many surveys have multiple-choice questions or multiple-choice-multiple-answer questions that will be coded like “A,” “B,” “CD,” etc. You can turn these into sets of dummies for an observation containing a particular response.
* If there is some way to handle certain string variables made obvious in the paper itself or in the replication files (ex. certain keywords in a string response that are important and always turned into dummies) follow that strategy.
* Otherwise, drop string variables.
* Keep track of categorical variables like region that are numeric but not ordered numerically in a meaningful way. The cleaning code I’ve written will handle these variables but you’ll need to keep a list of which variables these are.
* Some survey questions are complete irrelevant (for example, the name or ID of the interviewer). Drop these.
* Be wary of multiple variables that represent a single quantity when combined. It may make sense to combine these into the single quantity, then to drop some of the original variables. For example a particular date may be coded in three variables: date, year, month. The date itself can probably be dropped. The year and month on their own might be relevant to keep. You may also want to combine the three into “days since date X” or “months since date X.”
* I’ve likely forgotten or not anticipated some necessary cleaning here. The general principle is that the dataset you make should be full of regression ready controls; think of prepping variables that would make sense as controls in a regression. If you have any questions, let me know!

1. Merge the baseline data to the data needed to replicate the main findings of the paper. The final dataset should contain: ID’s, outcome variables, treatment variables, controls used in the paper, and the full set of baseline controls. The dataset should be able to replicate the main findings but have tons of potential extra controls.

Once prepped, these raw datasets will be fed into a LASSO cleaning file and then into the LASSO. The cleaning file is in the Clean\_data\_program subfolder. Any cleaning that happens in there of course does not need to be done here.