Codebook Assignment

November 14, 2017

This assignment is meant to direct the focus of your poster project while also exposing you to the importance of a well-documented codebook. To that end you should select only the variables you are planning to include in your potential analysis, be that a linear regression model or a data visualization.

Your final codebook document should be clearly labelled with the header titles mentioned in the following questions. For parts 2 and 3 you should write in complete sentences as if you were writing an essay. **Do not make lists.** There should not be any code in your knitted PDF, just the output of the `codebook` function. To turn in your final assignment commit your PDF and Rmd files to Git and push them to GitHub.

1. Create a new R Markdown file and name it after your data, followed by the word codebook (e.g.. my in-class example might be *PSID_CDI_Codebook*).

2. In the first section of your codebook, titled **Study Design**, describe the study design including the purpose of the study, the sponsor of the study, the name of the data collection organization, and the specific methodology used including the mode of data collection, the method of participant recruitment (if any), and the length of the field period. (12 points)

Study Design

The purpose of this study is to unveil the differences among the various Asian subgroups in King County, to address problems in understanding, narratives and policies that are being generated from the general categorization of the Asian American community. The data has been captured from the Census 2010, as most recent surveys from the American Community Service do not depict all the subgroups the study will portray. The datasets have been generated from American Fact Finder, filtered by specific geographical location: by each tract in King County as the unit of analysis. Further in the study, I will be generating simple tables utilizing each subgroup as unit of analysis and displaying socioeconomic parameters of income/household and education attainment/B.A or higher with total population. On the next two quarters, the study will narrow down a more speficic research question to depict the experience of a particular segment of the Asian American community in King County via qualitative research. "The experiece of Asian American artists entering the art world in Seattle area". I will purposefully seek self-identified Asian American, first generation(immigrant), female, LGBT respondents. I am interest in how intersections of identity categorizations have informed, challenged or supported these individuals in their work (individual) and entrance in this specific sphere of the society (external).I will pay particular attention in how policies have played, or not in the experience. As mentioned, it will be accomplished within the period of two quarters, potentially 10 interviews, divided by two groups, A(Cambodian/Laotian/Hmong/Vietnamese, Filipino) and B(Japanese, Chinese, Taiwanese, Korean). The selection and comparison of these two groups have been informed by the quantitative research, as they are part of two distinct socioeconomic groups. The interviewee will also have to have displayed at least one art work/enhibit in the public sphere - either a public space sponsored by an organization or the government, gallery or museum. The responses will be framed in levels of integration or marginalization, as well as language and experience. I will utilize Bonilla-Silva's Colorblind Racism framework of Abstract Liberalism, Naturalization, Cultural Racism and Minimization of Racism. I am assuming from my personal experience that individuals with these characteristics have a unique perspective and awareness of the society and sensititve to levels of marginalizations. Accessiblity to respondents will start through personal connections, and I will count on these connections to lead to other respondents. The general time line will be: interviews to be conducted within 8 weeks, mid february until mid April. Transcripts and analysis developed in the following 6 weeks, until the deadline by the end of May.

3. The second section, **Sampling** should clearly document all available sampling information. This includes a description of the population, the methods used to draw the sample, and any special conditions associated with the sample (i.e groups that were oversampled). (12 points)

Sampling

The sampling population in King County Asian American community. The sampling frame are individuals who self-identified as Asian American, first generation (immmigrant), female, LGBT, artists. I will contrast against self-identified White, male, heterosexual, non-immmigrant, artitst. The method is initially non randomized by personal connections, and later by introduction. As the questions can be fairly personal and I hope to understand in-depth possible experiences of marginalization, discrimination and vulnerability; I hope the introductions and personal connections might be an advantage for the respondents to feel a level of trust to share personal experiences. The sample size that I am aiming is 10 interviews. 6 from the Asian American group, and 4 from the White group. The interviews will not utilize a survey or questionnaire format.

4. Section three should be titled **Variable Index**. Here you should utilize the sample code shown in lecture (and reproduced below) to build a `data.set` version of your dataset. You will need to install and load the `memisc` package.

Variable Index

The two distinct visual representations will be a historgram and facet maps. For the Histogram, the variables are the each Asian subgroup in the x-axis and the population proportional number in relationship with the total population in King County. For the maps, I will be displaying maps for each Asian subgroup and the geographical location it is represented in King County by tract. The purpose of this study is a quantitative descriptive explanatory analysis of the Asian subroups in King County, I will not be utilizing a survey or questionnaire, therefore the steps below are not applicable.

Each variable in your dataset should be given a `description` and a unit of `measurement`. If there are `labels` associated with the underlying numeric values of the data those should be specified as should any `missing.value` codes including `NA`.

Once you've added the information above to your `data.set` object, make a call to `codebook()` to have your variable index printed out. (26 points)

```
##
## The downloaded binary packages are in
##  /var/folders/dm/pz9f8jgs199g21t077m_9nrh0000gn/T//RtmpvvN5s8/downloaded_packages
```

```
## Warning: package 'memisc' was built under R version 3.4.3
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

“‘{r}

Create data.set object from “data” object (tbl)

data_set_AsianAlone <- as.data.set(AsianAlone)

Look at new data.set object

View(data_set_AsianAlone)

Creating variable index for data_set

data_set_AsianAlone$POPGROUP.display.label < -data_set_AsianAlone$**POPGRO**

data_set_AsianAlone <- within(data_set_AsianAlone,{

Description of the variables #description(POPGROUP.id) <- “reference number of the subgroup” #description(POPGROUP.display.label) <- “description of the subgroup” #description(GEO.id) <- “larger number reference of the tract” #description(GEO.id2) <- “more specific reference of the tract number” #description(GEO.display.label) <- “descrition of tract, county and state” #description(D001) <- “population count” #})

Wording of survey item/interview questions (if applicable)

Wording is not applicable for my dataset #wording(sad) <- “Select the number next to the a sentence from the group that #best describes your feelings during the last two weeks.” #wording(okay) <- “Select the number next to the a sentence from the group that #best describes your feelings during the last two weeks.” #wording(selfhate) <- “Select the number next to the a sentence from the group #that best describes your feelings during the last two weeks.” #wording(cry) <- “Select the number next to the a sentence from the group that #best describes your feelings during the last two weeks.” #wording(irritability) <- “Select the number next to the a sentence from the #group that best describes your feelings during the last two weeks.” #wording(appearance) <- “Select the number next to the a sentence from the #group that best describes your feelings during the last two weeks.” #wording(isolation) <- “Select the number next to the a sentence from the #group that best describes your feelings during the last two weeks.” #wording(friends) <- “Select the number next to the a sentence from the group #that best describes your feelings during the last two weeks.” #wording(loved) <- “Select the number next to the a sentence from the group #that best describes your feelings during the last two weeks.” #wording(work_out) <- “Select the number next to the a sentence from the group #that best describes your feelings during the last two weeks.”

type of measurement

measurement(POPGROUP.id) <- “nominal” measurement(POPGROUP.display.label) <- “nominal” measurement(GEO.id) <- “nominal” measurement(GEO.id2) <- “nominal” measurement(GEO.display.label) <- “nominal” measurement(D001) <- “ratio”

labels associated with underlying numeric #It doesn’t apply to my dataset

any annotation/notes you wish to add

#This dataset has already been transformed to generate a more meaningful statistical analysis. Both by the proportion of the population of each Asian subgroup with the total population, and by representation of each subgroup by tract

How are missing values coded? # which(is.na(AsianAlone)) resulted in 0, telling me that there are no missing values.

})

\pagebreak

#```{r}
codebook(data_set_AsianALone)

Shorter, more efficient version of code above

```
# Creating variable index for data_set (truncated, more efficient code)
data_set <- within(data_set,{
  # Description of the variables
  description(id) <- "Unique person identification number (1968 Interview
                      Number(ER30001) + the Person Number(ER30002))"
  description(age) <- "Preloaded age of child from coverscreen (in years)"
  description(cdi) <- "Child Depression Index"
  description(sad) <- "Sad"
  description(okay) <- "Do Things Okay"
  description(selfhate) <- "Hate Myself"
  description(cry) <- "Crying"
  description(irritability) <- "Bother Me"
  description(appearance) <- "Look Okay"
  description(isolation) <- "Feel Alone"
  description(friends) <- "Plenty of Friends"
  description(loved) <- "Loves Me"
  description(work_out) <- "Things Will Work Out"

  # Wording of survey item/interview questions (if applicable)
  foreach(x=c(sad, okay, selfhate, cry, irritability, appearance,
              isolation, friends, loved, work_out),{
    wording(x) <- "Select the number next to the a sentence from the group that
    best describes your feelings during the last two weeks."
  })

  # type of measurement
  measurement(id) <- "interval"
  measurement(age) <- "ratio"
  foreach(x=c(cdi, sad, okay, selfhate, cry, irritability, appearance,
              isolation, friends, loved, work_out),{
    measurement(x) <- "ordinal"
  })

  # labels associated with underlying numeric
  labels(sad) <- c(
    "I am sad once in a while"                    =  1,
    "I am sad many times"                         =  2,
    "I am sad all the time"                       =  3)
  labels(okay) <- c(
    "I do most things O.K."                       =  1,
    "I do many things O.K."                       =  2,
    "I do everything wrong"                       =  3)
  labels(selfhate) <- c(
    "I hate myself"                               =  1,
    "I do not like myself"                        =  2,
    "I like myself"                               =  3)
  labels(cry) <- c(
    "I feel like crying every day"                =  1,
    "I feel like crying many days"                =  2,
    "I feel like crying once in awhile"           =  3)
  labels(irritability) <- c(
```

```
    "Things bother me all the time"                 =  1,
    "Things bother me many times"                   =  2,
    "Things bother me once in a while"              =  3)
  labels(appearance) <- c(
    "I look O.K."                                   =  1,
    "There are some bad things about my looks"      =  2,
    "I look ugly          "                         =  3)
  labels(isolation) <- c(
    "I do not feel alone"                           =  1,
    "I feel alone many times"                       =  2,
    "I feel alone all the time"                     =  3)
  labels(friends) <- c(
    "I have plenty of friends"                      =  1,
    "I have some friends, but I wish I had more"    =  2,
    "I do not have any friends"                     =  3)
  labels(loved) <- c(
    "Nobody really loves me"                        =  1,
    "I am not sure if anybody loves me"             =  2,
    "I am sure that somebody loves me"              =  3)
  labels(work_out) <- c(
    "Nothing will ever work out for me"             =  1,
    "I am not sure if things will work out for me"  =  2,
    "Things will work out for me O.K."              =  3)
  foreach(x=c(sad, okay, selfhate, cry, irritability, appearance,
              isolation, friends, loved, work_out),{
    labels(x) <- c(
    "Don't know"                                    =  8,
    "Not applicable; Answer refused"                =  9,
    "Inappropriate; child age 8-11"                 =  0)
  })

  # any annotation/notes you need to add to further explicate certain variables
  annotation(cdi)["Note"] <- "This variable is the raw score of the CDI. Items
  `sad`, `okay`, `selfhate`, `irritability`, `friends` were recoded to: 1=0; 2=1;
  3=2 and items `cry`, appearance`, `isolation`, `loved`, and `work_out` were
  recoded to: 1=2; 2=1; 3=0. The 10 items were then summed if valid data points
  were available on all 10 items."

  # How are missing values coded?
  foreach(x=c(cdi, sad, okay, selfhate, cry, irritability, appearance,
              isolation, friends, loved, work_out),{
    missing.values(x) <- c(8, 9, 0)
  })

})
```