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 data which I am working on, comes from different surveys. I of the data comes from Statistics Center of Iran. The data for 2006 and 2011 years are census data and almost all population is included in the census. The sponsor for all the survey and census is the Iranian Ministry of the interior. The name of data collection organization is SCI(Statistics Center of Iran). Data from 2001, 2008, 2009 also conducted by the same organization and same sponsor, the implementation method for these years is called sample household by the SCI. But the organization does give detailed information about the used sampling methods. The title of survey for these years is "Households Socio-economic Characteristics Survey". The surveys for 2010, 2012, 2013, and 2014 years are also sponsored by the Iranian Ministry of Interior and conducted by the SCI. The used method for these years is survey and the purpose of the surveys is collecting data for "General Census of Population & Housing". Unfortunately,

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 is no available information about sampling. The sections of SCI's website was unavailable to get the information. The website in Persian says that website will be unpdated, https://www.amar.org.ir/Install/UnderConstruction.htm?aspxerrorpath=/Default.aspx

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.

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`. Lastly, if your variable is a survey item/interview question you must provide the `wording` as well.

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)

Variable Index

```
## # A tibble: 18 x 3
##    gender  year value
##     <chr> <dbl> <dbl>
##  1    men  2001  13.2
##  2  women  2001  19.9
##  3    men  2006  10.0
##  4  women  2006  16.2
##  5    men  2008   9.1
##  6  women  2008  16.7
##  7    men  2009  10.8
##  8  women  2009  16.8
##  9    men  2010  11.9
## 10  women  2010  20.5
## 11    men  2011  10.5
## 12  women  2011  20.9
## 13    men  2012  10.4
## 14  women  2012  19.7
## 15    men  2013   8.6
## 16  women  2013  19.8
## 17    men  2014   8.8
## 18  women  2014  19.7
```

```
## 
## The downloaded binary packages are in
##  /var/folders/z9/b9hh4hpj6hl6x9r9dxjctv240000gn/T//RtmpQPxRuR/downloaded_packages
```

```
## 
##   There is a binary version available (and will be installed) but
##   the source version is later:
##           binary source
## tidyverse  1.1.1  1.2.1
## 
## 
## The downloaded binary packages are in
##  /var/folders/z9/b9hh4hpj6hl6x9r9dxjctv240000gn/T//RtmpQPxRuR/downloaded_packages
```

```
## # A tibble: 18 x 3
##    gender  year value
##     <chr> <dbl> <dbl>
##  1    men  2001  13.2
##  2  women  2001  19.9
##  3    men  2006  10.0
##  4  women  2006  16.2
##  5    men  2008   9.1
##  6  women  2008  16.7
##  7    men  2009  10.8
##  8  women  2009  16.8
##  9    men  2010  11.9
## 10  women  2010  20.5
## 11    men  2011  10.5
## 12  women  2011  20.9
## 13    men  2012  10.4
## 14  women  2012  19.7
## 15    men  2013   8.6
## 16  women  2013  19.8
## 17    men  2014   8.8
## 18  women  2014  19.7
```

```
# Create data.set object from "data" object (joint_unemployment_total)
joint_unemployment_total
```

```
## # A tibble: 18 x 3
##    gender  year value
##     <chr> <dbl> <dbl>
##  1    men  2001  13.2
##  2  women  2001  19.9
##  3    men  2006  10.0
##  4  women  2006  16.2
##  5    men  2008   9.1
##  6  women  2008  16.7
##  7    men  2009  10.8
##  8  women  2009  16.8
##  9    men  2010  11.9
## 10  women  2010  20.5
## 11    men  2011  10.5
## 12  women  2011  20.9
## 13    men  2012  10.4
## 14  women  2012  19.7
## 15    men  2013   8.6
## 16  women  2013  19.8
## 17    men  2014   8.8
## 18  women  2014  19.7
```

```
data_set <- as.data.set(joint_unemployment_total)

# Look at new data.set object
data_set
```

```
##
## Data set with 18 observations and 3 variables
##
##    gender year value
##  1    men 2001  13.2
##  2  women 2001  19.9
##  3    men 2006  10.0
##  4  women 2006  16.2
##  5    men 2008   9.1
##  6  women 2008  16.7
##  7    men 2009  10.8
##  8  women 2009  16.8
##  9    men 2010  11.9
## 10  women 2010  20.5
## 11    men 2011  10.5
## 12  women 2011  20.9
## 13    men 2012  10.4
## 14  women 2012  19.7
## 15    men 2013   8.6
## 16  women 2013  19.8
## 17    men 2014   8.8
## 18  women 2014  19.7
```

```r
codebook(data_set)
```

```
## ===========================================================================
##
##    gender 'gender is defined binary in the data and refers to men and
##    women in the Iran'
##
##    "The Statistics Center of Iran did not release the questionnaire of
##    surveys. However, in Iran mostly. I could not find the questionnaires
##    to write working for other variables"
##
## ---------------------------------------------------------------------------
##
##    Storage mode: character
##    Measurement: nominal
##
##    Values and labels     N     Percent
##
##         (unlab.vld.)    18   100.0 100.0
##
## ===========================================================================
##
##    year 'refers to the year which data is collected'
##
## ---------------------------------------------------------------------------
##
##    Storage mode: double
##    Measurement: interval
##
##             Min:  2001.000
##             Max:  2014.000
##            Mean:  2009.333
##        Std.Dev.:     3.771
##        Skewness:    -0.918
##        Kurtosis:     0.111
##
## ===========================================================================
##
##    value 'shows the percentage of unemployment for each gender in the
##    specified year'
##
## ---------------------------------------------------------------------------
##
##    Storage mode: double
##    Measurement: ratio
##
##             Min:   8.600
##             Max:  20.900
##            Mean:  14.639
##        Std.Dev.:   4.550
##        Skewness:   0.056
##        Kurtosis:  -1.659
```