Appendix C: Data Extraction and Wrangling

Christina Wang and Fei Wang

December 13, 2015

Download Data

This appendix describes the broader dataset and details about how to create the analytic dataset used here. The complete dataset has a size of 170.7 MB and is available as `CollegeScorecard_Raw_Data.zip` at https://collegescorecard.ed.gov/data/. The website also contains a number of `Featured Downloads` that provide quick access to some of the data in which users may be most interested.

The package, after unzipping, includes 18 comma-separated value (csv) files of annually compiled data, ranging from 1996 to 2013 (e.g. data from year 1996 is stored in file `MERGED1996_PP.csv`), a data dictionary, and other documentations. Please note that variabels and descriptions may change in different years.

For our analysis, we use dataset `MERGED2011_PP.csv`. Consult data dictionary `CollegeScorecardDataDictionary-09-12-2015.pdf` for labels and notes on each variable contained in the dataset.

Import data

Copy the unzipped file `MERGED2011_PP.csv` (size: 119.3MB) into the folder where you saved your R Markdown file. The size of the file is 365KB. Set the working directory to your current folder. Run the following code (or R markdown file) in the same folder where you have copied the file. You may need to install `readr` package to run the `read_csv` function.

```
df2011 <- read_csv("MERGED2011_PP.csv")
```

Select variables based on codebook

The full explanation for the names of the variables can be found in Codebook.

```
df_full <- df2011 %>%
  filter(HIGHDEG >= 3) %>% #choose institutions offering Graduate and Bachelor degree
  select(INSTNM, ZIP, st_fips, region, CONTROL, # UNITID for some reason cannot be selected
         DEBT_MDN_SUPP, GRAD_DEBT_MDN, GRAD_DEBT_N, WDRAW_DEBT_MDN, WDRAW_DEBT_N,
         mn_earn_wne_p10, md_earn_wne_p10, pct10_earn_wne_p10,
         pct25_earn_wne_p10, pct75_earn_wne_p10, pct90_earn_wne_p10,
         ADM_RATE, PCTPELL, C150_4)

df_full <- df_full %>%
  mutate(DEBT_MDN_SUPP = extract_numeric(DEBT_MDN_SUPP),
         GRAD_DEBT_MDN = extract_numeric(GRAD_DEBT_MDN),
         GRAD_DEBT_N = extract_numeric(GRAD_DEBT_N),
         WDRAW_DEBT_MDN = extract_numeric(WDRAW_DEBT_MDN),
         WDRAW_DEBT_N = extract_numeric(WDRAW_DEBT_N),
         mn_earn_wne_p10 = extract_numeric(mn_earn_wne_p10),
         md_earn_wne_p10 = extract_numeric(md_earn_wne_p10),
         pct10_earn_wne_p10 = extract_numeric(pct10_earn_wne_p10),
         pct25_earn_wne_p10 = extract_numeric(pct25_earn_wne_p10),
         pct75_earn_wne_p10 = extract_numeric(pct75_earn_wne_p10),
```

```
        pct90_earn_wne_p10 = extract_numeric(pct90_earn_wne_p10),
        ADM_RATE = extract_numeric(ADM_RATE),
        PCTPELL = extract_numeric(PCTPELL),
        C150_4 = extract_numeric(C150_4)
  )
```

Save the dataset

The filtered dataset based on the raw dataset of MERGED2011_PP.csv. is stored as `df2011.rda`.

```
save(df_full, file = "df2011.rda")
write.csv(df_full, file = "df2011.csv")
# The size of `df2011.rda` is 127.7KB
```