Constructing the Intermediate Data for *Cupid*

Alfred Galichon Bernard Salanié

July 27, 2021

Abstract

We set up the data for the empirical application in our "Cupid's Invisible Hand" paper. This data is very similar to that used in Choo and Siow's 2006 *Journal of Political Economy* article.

This document details the first step of the construction, which turns the Stata files in the `Input` subfolder of `Data` into the CSV files in the `Intermediate` subfolder.

The Python program `make_files_for_estimation.py` manages the second step and produces the text files in the `Output` subfolder of `Data`.

1 Working with the Data

For each year Y=71, 72, 81, 82, we downloaded three Stata datafiles from the Vital Statistics marriage data center:

- the main file, `marrY.dta`, contains the variables `year, statemarr, samplingweight` and `husbreform, husbforeign, husbresidstate, husbagemarr, husbrace2, husbeduc` along with the wife variables. It discards all marriages between two foreign partners, where "foreign" is everyone who is not in the list of states we use (see online appendix G.3 of our paper, or the variable `listStates` below).

- the files `husbforeignY.dta` and `wifeforeignY.dta` have the same variables; they have marriages with only one "foreign partner" (which are also in the corresponding `marrY.dta` file.) We add these foreign partners to the availables in their residence state. We will deal with the corresponding marriages in `DataStep2.py`.

In addition, we have one file from the ACS 1970 and 1980: `IpumsAvailables.dta`. This has variables `year, statefip, age, sex, agemarr, marrno, race, serial, myperwt, relate, hispan, educd, school, prevmarr, reform` for each person aged 16-75 who is in a Choo-Siow state and available for marriage (that is, not married). The weight `myperwt` is 20 or 50.

First we read them in, convert them to `R` datafiles, and add the foreigners to the availables. We drop the race and education variables. We also recode all state numbers so they refer to the Census codes; the translation stuff is stored in `listStates.RData`. And we drop marriages in New York City (we do not use New York State anyway.)

In the end we get a file `ChooSiowAvailables.csv` that has year of ACS, state, age, sex, sampling weight, and reform state indicator for every available man or woman in both waves; and a file `ChooSiowMarriages.csv` that has year of survey, state of marriage, and age/reform for both partners.

```
inputs_dir <- "./Input/"
outputs_dir <- "./Intermediate/"

library(foreign)
## first read ACS availables file
CSipums <- read.dta(file=paste(inputs_dir,
```

```
                                  "IpumsAvailables.dta", sep=''))
## drop race and education
CSipums <- data.frame(year=CSipums$year,
                      state=as.numeric(CSipums$statefip),
                       age=CSipums$age,sex=CSipums$sex,
                       weight=CSipums$myperwt,
                      reform=CSipums$reform)

## we need to recode the states
## format: Nonreform or Reform, statefip code in Census,
##  state name, state code in Vital Stats
##   Colorado only shows up in 1980, as Reform
listStates <-  rbind(
  c("N",1,"Alabama",1),
  c("R",2,"Alaska",2),
    c("R",6,"California",5),
    c("R80",8,"Colorado",6),
    c("N",9,"Connecticut",7),
    c("R",10,"Delaware",8),
  c("N",11,"District of Columbia",9),
  c("R",12,"Florida",10),
  c("R",13,"Georgia",11),
    c("R",15,"Hawaii",12),
 c("N",16,"Idaho",13),
c("N",17,"Illinois",14),
  c("N",18,"Indiana",15),
   c("N",19,"Iowa",16),
  c("R",20,"Kansas",17),
  c("N",21,"Kentucky",18),
  c("N",22,"Louisiana",19),
  c("N",23,"Maine",20),
 c("R",24,"Maryland",21),
  c("N",25,"Massachusetts",22),
  c("N",26,"Michigan",23),
 c("N",27,"Minnesota",24),
c("N",28,"Mississippi",25),
c("N",29,"Missouri",26),
```

```
 c("N",30,"Montana",27),
 c("N",31,"Nebraska",28),
 c("N",33,"New Hampshire",30),
 c("N",34,"New Jersey",31),
   c("R",37,"North Carolina",34),
   c("N",39,"Ohio",36),
  c("R",41,"Oregon",38),
 c("N",42,"Pennsylvania",39),
 c("N",44,"Rhode Island",40),
 c("R",45,"South Carolina",41),
 c("N",46,"South Dakota",42),
 c("N",47,"Tennessee",43),
 c("N",49,"Utah",45),
 c("N",50,"Vermont",46),
 c("R",51,"Virginia",47),
 c("N",54,"West Virginia",49),
 c("N",55,"Wisconsin",50),
 c("N",56,"Wyoming",51))

## take a vector of state codes from Vital Statistics
##  and translate it to state codes from Census
recodeState <- function(vsState) {
  nACS <- NROW(listStates)
  nobs <- NROW(vsState)
  numState <- numeric(nobs)
  for (i in 1:nACS) {
    numState[vsState==listStates[i,4]] <- as.numeric(listStates[i,2])
  }
  ## we return
  numState
}

## now read files from Vital Stats
## and add in the foreign partners
for (iyear in c(71,72,81,82)) {
  marrFile <- read.dta(file=paste(inputs_dir, "marr",iyear,".dta",sep=''))
  # we only need those variables
```

```
marrFile <- subset(marrFile, select=c(husbresidstate, wiferesidstate,
                                      statemarr, year,
                                      husbreform, wifereform,
                                      samplingweight,
                                      husbagemarr, wifeagemarr))
## we drop New York City
marrFile <- marrFile[(marrFile$statemarr != 33),]
marrFile$husbstate <- recodeState(marrFile$husbresidstate)
marrFile$wifestate <- recodeState(marrFile$wiferesidstate)
marrFile$statemarr <- recodeState(marrFile$statemarr)
marrFile$husbresidstate <- NULL
marrFile$wiferesidstate <- NULL
marrFile$year <- 1900+iyear
## drop race and education

if (iyear==71) {
  marrData <- marrFile
}
if (iyear > 71) {
  marrData <- rbind(marrData,marrFile)
}
mywt <- ifelse((iyear < 75),50,20)
myyear <- ifelse((iyear < 75),1970,1980)
husbforeignFile <- read.dta(file=paste(inputs_dir,
                                       "husbforeign",
                                       iyear,".dta",sep=''))
nhusbForeign <- NROW(husbforeignFile)
recodedHusbState <- recodeState(husbforeignFile$husbresidstate)
husbAvail <- data.frame(year=myyear+numeric(nhusbForeign),
                        state=recodedHusbState,
                        age=husbforeignFile$husbagemarr,
                        sex=1+numeric(nhusbForeign),
                        weight=mywt+numeric(nhusbForeign),
                        reform=husbforeignFile$husbreform)
wifeforeignFile <- read.dta(file=paste(inputs_dir,
                                       "wifeforeign",
                                       iyear,".dta",sep=''))
```

```
  nwifeForeign <- NROW(wifeforeignFile)
  recodedWifeState <- recodeState(wifeforeignFile$wiferesidstate)
  wifeAvail <- data.frame(year=myyear+numeric(nwifeForeign),
                          state=recodedWifeState,
                          age=wifeforeignFile$wifeagemarr,
                          sex=2+numeric(nwifeForeign),
                          weight=mywt+numeric(nwifeForeign),
                          reform=wifeforeignFile$wifereform)
  CSipums <- rbind(CSipums,husbAvail,wifeAvail)
}
```

2 Checking the Data

Now we replicate the numbers in Choo and Siow's Table 2.

```
availMenR70 <-
  sum((CSipums$sex==1 & CSipums$year==1970 & CSipums$reform==1)*
        CSipums$weight)
availMenR80 <-
  sum((CSipums$sex==1 & CSipums$year==1980 & CSipums$reform==1)*
        CSipums$weight)
availMenN70 <-
  sum((CSipums$sex==1 & CSipums$year==1970 & CSipums$reform==0)*
        CSipums$weight)
availMenN80 <-
  sum((CSipums$sex==1 & CSipums$year==1980 & CSipums$reform==0)*
        CSipums$weight)
availWomenR70 <-
  sum((CSipums$sex==2 & CSipums$year==1970 & CSipums$reform==1)*
        CSipums$weight)
availWomenR80 <-
  sum((CSipums$sex==2 & CSipums$year==1980 & CSipums$reform==1)*
        CSipums$weight)
availWomenN70 <-
  sum((CSipums$sex==2 & CSipums$year==1970 & CSipums$reform==0)*
        CSipums$weight)
```

```
availWomenN80 <-
  sum((CSipums$sex==2 & CSipums$year==1980 & CSipums$reform==0)*
        CSipums$weight)

marrRR70 <-
  sum((marrData$husbreform==1 & marrData$wifereform==1 & marrData$year < 1975)*
      marrData$samplingweight)
marrRN70 <-
  sum((marrData$husbreform==1 & marrData$wifereform==0 & marrData$year < 1975)*
      marrData$samplingweight)
marrNR70 <-
  sum((marrData$husbreform==0 & marrData$wifereform==1 & marrData$year < 1975)*
      marrData$samplingweight)
marrNN70 <-
  sum((marrData$husbreform==0 & marrData$wifereform==0 & marrData$year < 1975)*
      marrData$samplingweight)
marrRR80 <-
  sum((marrData$husbreform==1 & marrData$wifereform==1 & marrData$year > 1975)*
      marrData$samplingweight)
marrRN80 <-
  sum((marrData$husbreform==1 & marrData$wifereform==0 & marrData$year > 1975)*
      marrData$samplingweight)
marrNR80 <-
  sum((marrData$husbreform==0 & marrData$wifereform==1 & marrData$year > 1975)*
      marrData$samplingweight)
marrNN80 <-
  sum((marrData$husbreform==0 & marrData$wifereform==0 & marrData$year > 1975)*
      marrData$samplingweight)
```

We have 5.6772×10^6 men available in 1970 in reform states, and 1.048475×10^7 in non-reform states. In 1980 there are 9.34698×10^6 and 1.433696×10^7. For women the numbers are 6.536×10^6 and 1.301815×10^7 in 1970, and 1.046622×10^7 and 1.705538×10^7 in 1980.

Choo and Siow impute one marriage in each cell where there is none, which we did not do here. Our numbers of marriages RR were 1066122 in 1971–72, and 1270824 in 1981–82; for NN they were 2114747 in 1971–72, and 2117293 in 1981–82.

These numbers are very close to Choo and Siow's.

3 Creating the Intermediate Files

Finally, we convert the data to CSV format.

```
write.csv(CSipums, paste(outputs_dir, "ChooSiowAvailables.csv", sep=''),
          row.names=F)
write.csv(marrData, paste(outputs_dir, "ChooSiowMarriages.csv", sep=''),
          row.names=F)
```