

## Lab 12: Trends in Gender Occupational Segregation and Gender Equality among Young Cohorts, 1979-1999.

Note: cpsx.doc is the codebook for the data, and cpsapdx.doc gives the occupation codes.

In this lab we will be using data from the Current Population Survey (CPS) from 1979-1999 to study trends in wages, occupational segregation, and labor force participation for three cohorts of American workers.

For the lab, I have prepared a data set of synthetic cohorts (define) of individuals who were 21-35 in 1979 from the monthly CPS data from 1979-1999. Before we begin the lab, I want to:

- 1) Discuss the CPS
- 2) Show you the archived CPS data on the National Bureau of Economic Research (NBER) web page, and
- 3) Show you how a short Stata program made working with 20 years of monthly data very easy.

My setup for this lab can be taken as an example of data management procedures to extract relevant time-trend data from the CPS or other repeated-cross sectional data sets.

(1) The CPS is a monthly survey of about 60,000 households conducted by the Bureau of the Census for the Bureau of Labor Statistics. <http://www.bls.gov/cps/>

[More...]

(2) The NBER has an archive of CPS data. Let's inspect the NBER archive...

[http://www.nber.org/data/cps\\_index.html](http://www.nber.org/data/cps_index.html)

There are two archives: basic monthly data and supplements. (Look over supplement topics).

We are not going to import the raw data for this lab (because of time constraints).

However, the basic procedure for importing this data into Stata is similar to the process we used to bring data from IPUMS.

Basic steps for importing CPS data from the NBER web site:

- A) Download the zipped file.
- B) Unzip it (using pkunzip/winzip or gzip).
- C) Download the dictionary file or write your own using the codebook
- D) Bring the raw data into Stata using "infile using dictionary-name"

Note: you do not need to bring in *all* the data, only variables identified in the dictionary will be imported.

(3) The data that I used to construct today's data set consists of the "outgoing rotation groups" (ORG) from the CPS 1979-1999.

### **What are the Outgoing Rotation Groups?**

*Every household that enters the CPS is interviewed each month for 4 months, then ignored for 8 months, then interviewed again for 4 more months. Usual weekly hours/earning questions are asked only at households in their 4th and 8th interview. These outgoing interviews are the only ones included in the CD-*

ROM. New households enter each month, so one fourth the households are in an outgoing rotation each month.

I.e., the have data on hourly earnings. We will use this data to construct a time series of the male/female wage gap for these cohorts.

In my Unix space on Gromit (one of the CPC Unix machines), I have a directory with all of the data for the outgoing rotation groups.

Here is a listing of that directory:

```
gro:/home/tedmouw> cd /mouw/cpsorg/data
gro:/mouw/cpsorg/data> ls
morg79.dta.Z morg83.dta.Z morg87.dta.Z morg91.dta.Z morg95.dta.Z morg99.dta.Z
morg80.dta.Z morg84.dta.Z morg88.dta.Z morg92.dta.Z morg96.dta.Z pico.save
morg81.dta.Z morg85.dta.Z morg89.dta.Z morg93.dta.Z morg97.dta.Z
morg82.dta.Z morg86.dta.Z morg90.dta.Z morg94.dta.Z morg98.dta.Z
```

Here is a description of the data in these data sets:

(insert)

```
. use morg79
```

```
. des
```

```
Contains data from morg79.dta
```

```
obs:          328,406
vars:          52          2 Jun 2000 13:19
size:    25,615,668 (16.6% of memory free)
```

variable name	storage type	display format	value label	variable label
minsamp	byte	%8.0g		Month in sample (4 & 8 are depa
intmonth	byte	%8.0g		Interview month
hhid	str12	%12s		Household ID (12 digits)
state	byte	%8.0g		State
smsarank	byte	%8.0g		SMSA ranking
activlwr	byte	%8.0g		(r) Major Activity last week
hourslw	byte	%8.0g		How many hours last week?
reasonlw	byte	%8.0g		Reason <= 35 hours last week
absentlw	byte	%8.0g		Why absent from work last week?
classer	byte	%8.0g		(e&r) Class of worker
ind70	int	%8.0g		3-digit industry code (1970)
occ70	int	%8.0g		3-digit occupation code (1970)
lineno	byte	%8.0g		Persons line number in HH
relahh	byte	%8.0g		Relationship to household head
age	byte	%8.0g		Age
marital	byte	%8.0g		Marital status
race	byte	%8.0g		Race
sex	byte	%8.0g		Sex
veteran	byte	%8.0g		Veteran status
gradeat	byte	%8.0g		Highest grade attended
gradecp	byte	%8.0g		Whether completed highest grade
esr	byte	%8.0g		Employment status recode
weight	int	%12.0g		Final weight x100
smsastat	byte	%8.0g		SMSA status code
centcity	byte	%8.0g		Central city status
ethnic	byte	%8.0g		Ethnicity

ptstat	byte	%8.0g	Part-time status
ftpt79	byte	%8.0g	Full-time or part-time status
dind	byte	%8.0g	Detailed industry code
docc70	byte	%8.0g	Detailed occupation code
doinglw	byte	%8.0g	What was doing most of last wee
hours1wa	byte	%8.0g	How many hours last week all jo
uhours35	byte	%8.0g	Usually works >= 35 hrs at this
why35lw	byte	%8.0g	Why not at least 35 hours last
class	byte	%8.0g	Class of worker
uhours	byte	%8.0g	(dp) Usual hours
paidhr	byte	%8.0g	(dp) Paid by the hour
earnhr	int	%8.0g	(dp) Earnings per hour
uearnwk	int	%8.0g	(dp) Usual earnings per week
earnwt	long	%12.0g	(dp) Earnings weight for all ra
eligible	byte	%8.0g	Eligibility flag
uhourse	byte	%8.0g	(e&dp) Usual hours
paidhre	byte	%8.0g	(e&dp) Paid by the hour
earnhre	int	%8.0g	(e&dp) Earnings per hour
earnwke	int	%8.0g	(e&dp) Earnings per week
I25a	byte	%8.0g	(dp) Usual hours (I25a) allocat
I25b	byte	%8.0g	(dp) Paid by hour (I25b) alloca
I25c	byte	%8.0g	(dp) Earnings/hr (I25c) allocat
I25d	byte	%8.0g	(dp) Usl Earn/hr (I25d) allocat
uearnwke	int	%8.0g	(e&dp) Usual weekly earnings
year	byte	%8.0g	
smsa70	byte	%8.0g	

-----  
Sorted by:

In the Stata ado file getorg.ado I take a single one of these data sets, copy it to my lab12 directory, uncompress it, and take the relevant variables for the relevant cohorts.

Here is getorg.ado:

```
gro:/mouw/cpsorg/lab12> more getorg.ado
```

---

```
_____getorg.ado_____

program define getorg

local y "`1'"

! cp -f /mouw/cpsorg/data/morg`y'.dta.Z /mouw/cpsorg/lab12/morg`y'.dta.Z

! uncompress morg`y'.dta.Z

use morg`y'

gen age2=age-(year-79)

drop if age2>35 | age2<20
gen cohort=age2
recode cohort 20/25=1 26/30=2 31/35=3

replace earnhre=earnhre/100
replace earnhre=earnwke/uhourse if paidhre==2
replace earnhre=. if earnhre<1

keep age cohort sex marital ftpt* occ* year earnhre weight

compress

append using all

save all, replace

! rm -f morg`y'.dta

end
```

---

getorg.ado adds 1 year of ORG CPS data to all.dta. Now we need to run it for all the requested years. That is the job of getorg.do:

```
gro:/mouw/cpsorg/lab12> more getorg.do
```

```
-----getorg.do-----
```

```
program drop _all
set trace off
clear
set obs 1
gen age=1
save all, replace
clear
* set trace on
getorg 79
getorg 80
getorg 81
getorg 82
getorg 83
getorg 84
getorg 85
getorg 86
getorg 87
getorg 88
getorg 89
getorg 90
getorg 91
getorg 92
getorg 93
getorg 94
getorg 95
getorg 97
getorg 98
getorg 99
```

---

So, **all.dta** is the finished product of running getorg.do and getorg.ado. This type of example can be applied in other cases. An ado file such as getorg.ado can help you repeat a set of commands on different data sets. This is worth the time that goes into writing it because if I wanted to change my analysis—say to take all workers rather than just a few cohorts—all I have to do is change a few lines of getorg.ado and re-run the program.

occ.ado and occ.do operate similarly (inspect them to prove it to yourself), resulting in a dataset, occ.dta, of the average proportion female in 3-digit occupations *for all workers* for each year 1979-1999. What is interesting about this data set is you can track changes in the % female in specific occupations over time.

Key lesson: *let the computer do the work for you.* Learning a little bit of do-file and ado-file programming can greatly increase your effectiveness.

The resulting data sets for this lab are all.dta and occ.dta.

All.dta is a data set of all individuals from these three cohorts (20-25, 26-30, and 31-35 in 1979) from 1979-1999, with the following variables:

```
Contains data from all.dta
  obs:      2,146,927
  vars:      12
  size:     57,967,029 (30.9% of memory free)
  4 Apr 2002 11:26
```

---

variable name	storage type	display format	value label	variable label
sex	byte	%8.0g		Sex
age	float	%8.0g		Age
marital	byte	%8.0g		Marital status
occ80	int	%8.0g		3-digit occupation code (1980)
ftpt94	byte	%8.0g		Full-time or part-time status
earnhre	float	%8.0g		(e&dp) Earnings per hour
weight	float	%9.0g		Final weight x100
year	byte	%8.0g		
cohort	byte	%9.0g		
ftpt89	byte	%8.0g		Full-time or part-time status
ftpt79	byte	%8.0g		Full-time or part-time status
occ70	int	%8.0g		3-digit occupation code (1970)

The data set occ.do is a data set collapsed by occupation for each year 1979-1999, with the following variables:

```
Contains data from occ.dta
  obs:      2,220
  vars:      5
  size:     48,840 (99.8% of memory free)
  4 Apr 2002 11:58
```

---

variable name	storage type	display format	value label	variable label
occ80	int	%8.0g		3-digit occupation code (1980)
f	float	%9.0g	(mean) f	
year	float	%8.0g	(mean) year	
n	long	%8.0g	(count) sex	
zz	float	%9.0g		

[Note: in this data, occ80 is really occ70 if year<83]

The variable f is the proportion female in the 3-digit occupation for that year.

**If you sort this data by occ80 and year you can merge it to the all.dta.**

I would like you to work on answering these questions in lab today. You can work together to figure out the Stata code to answer each of these questions.

The youngest of these cohorts, age 20-25 in 1979 enters the labor force at a time when large cultural changes are presumably affecting American society, resulting in a greater opportunity for gender equality. How did this optimistic vision of equality play out over 20 years of their lives?

Lab assignment. Answer 1 of the following questions and bring your results to present in class next Tuesday. You can work together on the question. We will assign the questions evenly:

In this lab I would like you, collectively, to answer the following questions:

1) What are the overall trends in gender occupational segregation from 1979-1999? (use occ.dta)

For 3 cohorts of men and women aged 20-25, 26-30, and 31-35 in 1979 (use all.dta):

- 2) How did the average % female in their occupations change over their life course?
- 3) How did the % of workers (in all.dta) in predominately male or female occupations (i.e. >85% male/female in occ.dta) change over the life course?
- 4) How did labor force participation rates and marital status change as over time?
- 5) How did the male-female wage gap evolve over time?

Tips:

Some Stata commands that will be helpful:

Sort x y

Merge x y using z

Collapse x1 x2 [w=weight], by(year cohort)

Graph x1 x2 year, by(cohort) c(mm)

Gen b1=x>.9

Gen b2=x<.1

(1) Use definition of the index of dissimilarity (D) to write a Stata do file that will calculate D by year. Then make a table of D by year.

(2) % female in average female's or male's occupation →

Use occ.dta to put information about %female by occupation in the all.dta. Then the rest is up to you.

(3) merge with occ.dta. See bold sentence about occ.dta above.

(4-5) collapse the data. Example: how did X change over time for men and women?

collapse (mean) x [w=weight], by(year cohort sex)

sort sex

by sex: tab year cohort, sum(x)