Supporting files information for Autor, Levy and Murane (2003, *QJE*)
07/11/08

1). Consistent occupation crosswalk files (occ8090.zip)

Contains:

occ80.dta
occ90.dta
1980-COC-notes.pdf
COC-Occupations-1980.pdf
COC-Occupations-1990.pdf
occ8090.txt

Description: these Stata datasets are the crosswalk tools from Autor-Katz-Kreuger (1998) for making comparable Census occupation codes for different time periods (1983 - 1999). occ8090.txt contains directions for using these datasets.

The procedure for using the datasets:

1. Open your data set (ex: the 1983 CPS)
2. Rename the 3-digit occupation variable to occ80
3. Merge occ80 using occ80.dta - This step appends our codes onto your data.
4. The consistent occupation code appended is occ8090

Note:

- The crosswalk is strictly an aggregation of the 80/90 changes. If two occupations overlapped due to recoding, we merged them. Otherwise, we just renumbered consistently.
- A few tiny occupations had to be dropped altogether in both years for consistency. These receive an occ8090 code of "." - i.e., missing. After merging in the codes, you may want to drop if occ8090==.

2). DOT means by occupation (DOT-occ-data-sets.zip)

Contains:

dot77-6070-gen.dta
dot77-6070.dta
dot77-70-gen.dta
dot77-70.dta
dot77-8090-gen.dta
dot77-8090.dta
dot91-6070-gen.dta
dot91-6070.dta
dot91-70-gen.dta
dot91-70.dta
dot91-8090-gen.dta
dot91-8090.dta

DOT-occ-data-sets.txt

How to read the file names:

The prefix (dot77, dot91) - the variables are from the 1977 4th edition or 1991 revised 4th edition of the DOT.

The subsequent number (6070, 70, 8090) - the set of occupation codes these are aggregated to.

6070 - the 1960-70 crosswalk aggregation
70 - the 1970 occupation codes
8090 - the 1980 and 1990 codes (which are almost identical)

The suffix (gen) - the variables are coded by gender. Otherwise, they are weighted averages of males and females.

DOT task measure variables (see ALM Appendix Table 1):

ehf - Nonroutine manual
finger - Routine manual
dcp - Non-routine cognitive/interactive
sts - Routine cognitive
math - Non-routine cognitive/analytical

Each DOT variable is coded on a continuous scale from 0 to 10, where higher numbers refer to more intensive use of the task.

3) Consistent industry crosswalk files (ind6090.zip)

Contains:

cpsind60.dta
ind60.dta
ind70.dta
ind80.dta
Ind8090docs.doc
ind90.dta
Indkey.xls
ind6090.txt

Description: These datasets contain the crosswalks from CICs of their respective decades and the consistent industry codes. The Microsoft Excel file indkey.xls explains what each of the ind6090 industries is, both at the ind6090 and ind7090 level.

The variables that contain these codes are called:

ind6090 - Consistent 60-90
ind8090 - Consistent 80-90
ind7080 - Consistent 70-80
ind7090 - Consistent 70-90

The longer in time the consistent series, the greater the level of industry aggregation.

These data sets also include variables called dindXXXX and mindXXXX where XXXX are years. The dind variables are consistent detailed industries (2-digit) and the mind variables are consistent major industries (1-digit).

The procedure for using these datasets (with the possible Stata commands):

1. Open your census (or CPS) dataset (ex: the 1960 CPS)
2. Rename the industry variable to indXX (ex: ind60)

```
rename ind ind60
```

1. Merge your dataset on the ind60 key using the indXX data set of the appropriate year (ex: ind60.dta). This step appends our codes onto your data, so the merged data set now contains the corresponding ind6090, ind7090, etc. codes.

```
sort ind60
merge ind60 using ind60.dta
tab _merge * Note: all industries should be matched *
summ ind6090
```

3. Aggregate your observations to the appropriate industry means using the industry codes of your choice.

```
collapse (mean) var1 var2 ... var2, by(ind6090)
```

Note:

- In the 1970 Census, there are a number of miscellaneous industries assignments which correspond to broad industry categories but do not have a specific classification. These industry aggregates do not have an ind6090 correspondence. We imputed (i.e., randomly assigned) workers in these industries to detailed industry categories within the broad aggregates according to the distribution of employment among the detailed industries within the aggregates. We subsequently assigned ind6090 codes according to the imputed industry assignments. You may need to follow a similar procedure.
- The CIC code changes in the CPS always lag the CIC code changes in the Census by two or more years. If you are trying to get consistent CIC assignments with the CPS, pay close attention to which coding regime is in use in which year.
- The 1960 CICS in the CPS are different again from the 1960 CICs used in the Census. Let me know if you need a crosswalk from the CPS 1960 CICs to the Ind6090 scheme.

4) DOT means by industry (ind-dot-means.zip)

Contains:
inddotmeans6098-77.dta
inddotmeans6098-91.dta
ind-dot-means.txt

Description: The data sets inddotmeans6098-77.dta and inddotmeans6098-91.dta provide the means of our DOT variables by industry-education-year cells for 140 consistent industries covering the entire economy for years 1960-1998. This crosswalking scheme can be used for CIC industries but does not cover the switchover to the NAICS (in the 2000 Census). Note that the crosswalking scheme is purely an aggregation technique. It does not reallocate employment among industries. For our industry level data, 140 industries is the only scheme we have. In our occupation-level data, things are considerably less aggregated.

How to read the file names:
The suffix (77, 91) - the task measures calculated from the 1977 DOT or the 1991 DOT (the '91 DOT provides a very limited update to the '77 DOT).

DOT task measure variables (see ALM Appendix Table 1):

ehf - Nonroutine manual
finger - Routine manual
dcp - Non-routine cognitive/interactive
sts - Routine cognitive
math - Non-routine cognitive/analytical

ind6090-consistent industry code
year - This is coded in a tricky way. Please see the variable yeartxt, which lists what data source corresponds to which years.

lswt - Gives you employment in the industry-year-education cell in FT equivalent employment

edcat - Each industry mean is calculated over 5 education groups
edcat == 0 All
edcat == 1 HS dropouts
edcat == 2 HS grads
edcat == 3 Some college
edcat == 4 College +

These variables tell you the education shares in each industry-year-education:
hsd
hsg
smc
clg - college exactly
gtc - greater than college