```
===========================================
EndMob_Test-replication_archive.zip
May 20, 2015
Contact: Ian M. Schmutte (schmutte@uga.edu)
===========================================

-------------
File Manifest
-------------

./code
        ./lib
                globaldefs.sas
                preamble.sas

        ./01_build
                01.10.annwage_stack.sas
                01.10.icf_licf.sas
                01.15_domjob_data.sas
                01.20.ref_dist_quantiles.sas


        ./02_test1
                01.mean_resids.sas
                02.mean_resid_trans.sas
                03.mean_resid_trans_analysis1.sas
                04.mean_resid_trans_analysis2.sas


        ./03_test2
                01.sort_job_strip.sas
                02.create_resid_deciles.sas
                03.firm_vars.sas
                04.select_years.sas
                05.test2.sas
                06.test2.jma.sas
                07.test2.jma.sas


./results
        ./test1
                ./endmob_dectrans_byresid_margin.ods
                ./endmob_resid_dist.csv
                ./mean_resid_trans_analysis1.lst

        ./test2
                ./chisquare.lst
                ./residuals_fdenom.csv
                ./residuals_wdenom.csv
                ./test2_results_fdenom.csv
                ./test2_results_wdenom.csv


Overview:
============================================================================
The files are the results of two tests for the presence of worker endogenous firm to firm mobility.  The
```

presence of endogenous mobility has been long presumed to exist, but these results actually represent the first formal tests. Endogenous mobility is present when uncontrolled factors (functions of residuals that are not orthogonal to the person and firm effect designs) in an Abowd Kramarz Margolis style wage equation are correlated with worker mobility patterns. Our two tests look at the issue from both the worker and the firm perspective.

The first test looks for a systematic relationship between the psi at the old and new firm as a worker changes jobs. Workers are classified into skill (fixed person effect) and match residual deciles. The match residual is the average residual for a (PIK, SEIN) pair over all years where the SEIN is the dominant job for the PIK. The test is designed to detect systematic variation in the relationship between the psi at the old and new firm across theta and match residual cells. The test results strongly suggest that endogenous mobility is present in LEHD data.

The second test looks at changes in the composition of workers over time. If endogenous mobility is present, the distribution of skill at the firm should vary significantly by the average residual for an SEIN at a point in time and the fixed firm effects at the SEIN. The test results strongly suggest that endogenous mobility is present in LEHD data.

Confidentiality of the underlying micro-data (both person and firm) is assured by construction. The samples used to estimate the statistics combine data for the UI-covered workforce for 28 states in the current snapshot. The reported statistics use information from every observation in their construction, so there should be no concern with small sample sizes.

Description of Analysis Samples:

===============================

The statistics are computed on a sample of LEHD data from 28 states in the s2004 snapshot. The sampling frame is the set of annualized employment histories reported in the Human Capital Estimates (vintage: vin_bls version 2.4 on NSF01).

The human capital estimates are derived from the following earnings model: y = Xb + theta + psi + epsilon. Where Xb represents a time varying component (experience), theta is a fixed person effect, psi is a fixed firm effect, and epsilon is a residual. These components are the primary variables used to construct the two tests.

This frame was drawn from the EHF for years between 1990-2003, annualized. The frame data were supplemented with additional individual characteristics (sex, age, race, imputed education) from the ICF, SEIN characteristics (NAICS 2002 Major Sector and estimated employment (recorded from the ES-202 in the ECF)) of each SEIN from the ECF. The Human Capital Estimates also provide data on imputed annual hours of work and the variance components of earnings due to unobservable person- and firm-specific heterogeneity.

From this frame, I selected dominant job records for all workers between the years 1999 and 2003, inclusive. Complete job history information for those workers was merged from the EHF. The data were processed to identify the quarter of transition between dominant employers.

The analysis sample for test 1 consists of all workers that change employers from one firm to another. A given worker can appear in the sample more than once, once for each firm to firm move. All records with a missing theta, psi, or residual are dropped. The final sample consists of over 104 million records (PIK, SEIN, year).

The analysis sample for test 2 is at the firm level. All firms with positive employment in 2001 are selected and their 2003 information is attached. Although 2003 firm information is used, only those firms alive in 2001 are retained. The distribution of the fixed person effects across the deciles is calculated along with the average residual at the firm in 2001. All records with missing psi or mean residual information are dropped. The final sample consists of over 4 million firms (SEINs).

****
DESCRIPTION OF RESULTS
****

Description of File: ./results/test1/endmob_resid_dist.csv
===
This table contains 10 rows and 1,000 columns. The 10 rows are the deciles of the match residual (average residual for the PIK-SEIN for all years that this was the dominant job) at the last job. The 1000 columns are derived from the interaction of three variables. The three variables are: the theta decile, the psi decile at the last job and the psi decile at the current job. The minimum cell size is 540 job transitions (PIK-SEIN pairs by match residual decile x theta decile x last psi decile x current psi decile).

With regard to confidentiality of the underlying data, all of the cells in the proposed table are sufficiently coarse to far exceed the required cell counts.

Description of File: ./results/test1/endmob_dectrans_byresid_margin.ods
===
This table contains the number of jobs in each current psi decile by the residual decile of the last job interacted with the psi decile of the last job. The table contains 100 rows and 10 columns. The 100 rows represent the interacted psi and residual deciles, while the 10 columns contain the psi counts across each current job psi decile. The minimum cell size is 6,658 job transitions.

With regard to confidentiality of the underlying data, all of the cells in the proposed table are sufficiently coarse to far exceed the required cell counts.

Description of File: ./results/test1/mean_resid_trans_analysis1.lst
===
These results are for test 1 and represent the summary statistics from a chi-square test of the following table (note that the table itself is NOT in the released data request). The decile of the residual at the last job by the interaction of three variables. The three variables are: the theta decile, the psi decile at the last job and the psi decile at the current job. The degrees of freedom in this test is 8,991. The summary statistics present no disclosure risk due to the high level of aggregation in the results.

Description of File: ./results/test2/chisquare.lst
===
This file contains the value of the test statistic for test 2, number of degrees of freedom, and the probability of observing the statistic (distributed chi-square) given the 900 degrees of freedom. The summary statistics present no disclosure risk due to the high level of aggregation in the results.

Description of File: ./results/test2/residuals_fdenom.csv and ./results/test2/residuals_wdenom.csv
===
These two files are identical and contain a 100 row table. Each row is one of the 100 possible psi decile and residual decile combinations. The table has the following columns: psi decile, mean residual decile, and the average change in the proportion of workers in a theta decile from 2001 to 2003 minus the average change in the proportion of workers in a theta decile from 2001 to 2003 for a given psi_decile. The minimum cell size in this table is 23,484 unique firms.

With regard to confidentiality of the underlying data, all of the cells in the proposed table are sufficiently coarse to far exceed the required cell counts.

Description of File: ./results/test2/test2_results_fdenom.csv and ./results/test2/test2_results_wdenom.csv

===

These two files are virtually identical and contain a 100 row table. Each table differs only by the denominator used to calculate the test statistic. For fdenom I use the number of firms and for wdenom I use the number of employees. Each row is one of the 100 possible psi decile and residual decile combinations. The table has the following columns: psi decile, mean residual decile, the number of firms, the chi-square contribution, the average change in the proportion of workers in a theta decile from 2001 to 2003 and the average change in the proportion of workers in a theta decile from 2001 to 2003 for a given psi_decile. The minimum cell size in this table is 23,484 unique firms.

With regard to confidentiality of the underlying data, all of the cells in the proposed table are sufficiently coarse to far exceed the required cell counts.

****

DETAILS OF PROGRAM SEQUENCE

****

Notes

============================

The code in this archive were prepared and run in the restricted-access computing environment of the U.S. Census Bureau. We have retained the original locations of the raw input files in the code. We grant any researcher with appropriate Census-approved projects permission to use the exact research files provided they are part of the approved project (a Census requirement). The file locations recorded in the code represent the last known location of the archive on the Census RDC network. We will cooperate with any researcher attempting to locate the input data for the purpose of replicating our results.

Those interested in applying our tests to different data can start with the code under ./02_test1 and ./03_test2. Use the documentation in the build sequence for guidance on how to prepare your data for the test.

==================================

The description that follows explains the sequence in which the code should be run to replicate our results.

Sequence to build analysis data

```
1.      Extract Input data from gzip archives
                1.      Use the get_hc_files.sh script to extract the hc_by_sein.sas7bdat
                        files to /temporary/saswork1/schmu005/hc
                2.      States Used are (28 total): al ca de fl ia id il in ks ky md me mn
                        mo nc nd nj nm ok or pa sc tx va vt wa wi wv
2.      01.10.annwage_stack.sas
                1.      Interleave the hc_by_sein files for each state by sein year
                2.      Interleave the annual_ecf_03 files for each state by sein year
                3.      Create sein_best_emp_avg by calculating the average of the 12
                        sein_best_emp variables (one observation per month).
                4.      Create shortnaics2002 using the first 2 digits of naics2002 to
                        create standard 2 digit NAICS sectors
                5.      Merge the stacked hc_by_sein files with the annual ECF variables
                        created above
                6.      Get the count of full (full_qtrs), continuous (cont_qtrs), and
                        discontinuous (dis_qtrs) quarters
                7.      Create an estimate of the weeks worked using the quarters worked
                        variables (weeks=full_qtrs*13 + cont_qtrs*6.5 + dis_qtrs*4.33)
                8.      Create weeklyhrs=hours_job/weeks
                9.      Create fulltime=weeklyhrs > 34
                10.     If a record is in both the hc_by_sein files and the ecf and is a
                        dominant job across all states used to create the HC estimates and
                        the year is in(1999-2004) (restriction imposed due to the
                        availability of  residence data) then in_univ=1
                11.     If in_univ=1 and fulltime and 18<=age<=70 and year =2002 and the
                        person is working at the end of Q2 then ref_dist=1
                12.     Finally, sort the data by pik sein
                13.     982,842,806 dominant jobs in the stacked dataset
                14.     Final Output File: netearn.domjob_sample.sas7bdat
                        (465,221,077 dominant jobs (PIK SEIN YEAR) with in_univ=1)
3.      01.10.icf_licf.sas
                1.      Load formats for standardized Kevin geography
                2.      Stack the icf and icft26 files
                3.      Bring in the pik_geo_2003 (residence location of PIK's in 2000 from
                        Decennial Census HCEF)
                4.      Bring in the PIK records from the hcef_pik_per_uniqpik file
                5.      Use the unduplicated hcef_pik_per_uniqpik file to select one PIK
                        record for each PIK on the pik_geo_2003 file, creating the
                        hcef_pik_geo_uniqpik file
                6.      Merge the stacked ICF, stacked ICFT26, LICF, and the
                        hcef_pik_geo_uniqpik file together by pik
                7.      The result of this merge gives us variables from the ICF along with
                        time varying residence information.
                8.      Final Output File: interwrk.ind_chars.sas7bdat (191,932,034 PIKs
                        with a record on the ICF)
4.      01.10.job_relations.sas (not used)
                1.      Stack the EHF
                2.      Create a PIK SEIN YEAR QUARTER file where each record represents an
                        active job (earnX>0)
                3.      Final Ouput File: netearn.job_relations.sas7bdat

5.      01.10.universe.sas (not used)
                1.      Sort and nodupkey the job_relations file to get the number of PIKs.
                        netearn.pik_universe.sas7bdat
                2.      Sort and nodupkey the job_relations file to get the number of SEINs
                        netearn.sein_universe.sas7bdat
6.      01.15.domjob_data.sas
                1.      Load formats that assigns FIPS county code to Core Based Statistical
                        Areas(CBSA)
                2.      Load a program that maps Census lat/long information to Census
                        Geography (blocks and tracts)
                3.      Merge the domjobs_sample dataset from 01.10.annwage_stack.sas with
                        the ind_chars dataset created in 01.10.icf_licf.sas
                4.      Create a consistent across time lat/long variable named pos
                5.      Create geocodefull
                6.      Create a consistent across time CBSA variable
                7.      Attach person specific characteristics
                8.      Final Output File: netearn.domjob_sample.sas7bdat
                        (465,221,077 PIK SEIN YEAR records)
7.      01.20.job_sample.sas (not used)
                1.      Stack the PHF files
                2.      Important Macro Variable Constants
```

3. eyear=2004 (last year of snapshot data)
4. endq=(eyear+1-1985)*4 (Number of quarters from 1985:1 to 2004:4)
5. rangeq=(endq-21)+1 (Number of quarters from 1990:1 to 2004:4)
6. Merge the domjob_sample file with the stacked PHF by PIK SEIN. Every year, this gives the complete earnings history at the job arrayed out for each job in domjob_sample.
7. Create dompattern by placing a "1111" pattern into a string of zeros of length rangeq. The "1111" pattern represents the location of the dominant job in time.
8. Create work by inserting the PHF work variable into a string that begins with 1990:1 instead of 1985:1
9. Sort the intermediate jobsample file by PIK YEAR SEIN
10. Determine the quarter worked patterns for each dominant job (including the transition point).
11. Final Output File: interwrk.jobsample2.sas7bdat

8. 01.20.ref_dist_quantiles.sas
   1. Create the reference distribution of psi, theta, and the residual. The reference sample is selected using the ref_dist variable (ref_dist=1) created in 01.10.annwage_stack.sas.
   2. Theta centiles are calculated for all of the persons where ref_dist=1
   3. Psi centiles are calculated for all of the firms where ref_dist=1
   4. Residuals are calculated over all jobs where ref_dist=1
   5. Output File: netearn.centiles_psi
   6. Output File: netearn.centiles_theta
   7. Output File: netearn.centiles_resid

Test 1

1. 01.mean_resids.sas
   1. Sort a subset of the domjob_sample dataset by PIK YEAR SEIN to create the job_strip dataset
   2. Calculate the mean residual at each job
   3. Output File: domjob_cats (PIK SEIN dataset of mean residuals at each job. 232,518,684 observations)
2. 02.mean_resid_trans.sas
   1. Load the centiles into arrays
   2. Read in domjob_cats and classify each theta, psi, and residual into the appropriate decile
   3. For every person that had more than one job, output a record for each job but the first one. This captures all of the job to job transitions.
   4. The values of sein, end_year, psi, and the residual for the last employer are captured on the current record. Theta is fixed and does not change across time and/or employers.
   5. Everything is now ready to do the test.
   6. Output File: domjob_cats2 (PIK SEIN dataset for job changers. 104,885,989 observations with 8,991 degrees of freedom)
3. 03.mean_resid_trans_analysis1.sas
   1. Run the Chi-Square test
4. 04.mean_resid_trans_analysis2.sas
   1. Create output datasets for table creation
   2. Create endmob_resid_dist.csv (100x10) table. The table is last_resid_dec*last_psi_dec*counts in each current psi decile cell. Minimum cell size is 6658

Test 2

```
1.      01.sort_job_strip.sas
                1.      Sort job_strip (structure is similar to domjob_sample
                        (jobs over time)) by SEIN YEAR PIK and remove any missing
                        observations.
                2.      Final Output File:interwrk.job_strip2.sas7bdat
                        (460,913,364 observations)
2.      02.create_resid_deciles.sas
                1.      Select the reference sample from domjob_sample
                        (year=2002 and fulltime=1 and employed at end of quarter 2
                        and resid~=.) Age was not available on the file and is not used.
                2.      Sort the data by SEIN (Only one year in the reference sample)
                3.      Calculate the mean residual for each firm
                4.      Calculate the weighted and unweighted centiles
                5.      Output File: centiles_resid_firm (1 observation with 100 vars,
                        one for each centile)
                6.      Output File: centiles_resid_firm_weight (1 observation with 100
                        vars, one for each centile)
3.      03.firm_vars.sas
                1.      Calculate the mean residuals for every firm at every point in time
                2.      Merge the mean residuals with the psi and theta information
                        (SEIN YEAR PIK). Classify each theta value into the appropriate
                        decile and then count up the number of workers in each decile at the
                        firm. Finally, classify psi and the mean_resid to the appropriate
                        decile and output the summary results for each firm at each point in
                        time.
                3.      Output File: interwrk.firm_data.sas7bdat (20,286,032 SEIN YEAR
                        observations)
4.      04.select_years.sas
                1.      Select two years of firm data and merge them together (currently
                        2001 and 2003). With the data for both time periods all on one
                        record, calculate the percentage of workers in each theta decile.
                        Take the difference between the two measures for each decile,
                        creating x_bar_j. The universe of firms are those active in 2001.
                        For firms that die, the percent in each theta category in the second
                        period is set to zero. This bounds the measure to between -1 and 1,
                        but since the percent in each category in time t does not always sum
                        to one, the sum of x_bar_j does not always equal zero. This allows
                        us to use all 10 categories in our test.
                2.      Sort the data by the psi (a), mean residual deciles (c) (100 values)
                        and then calculate the mean of x_bar_j, which is named x_bar_ac. Do
                        the same for a alone, giving us x_bar_a.
                3.      Merge and output the results
                4.      Output File: netearn.test2_ac.sas7bdat (100=10*10 observations,
                        sorted by PSI_DEC_S RESID_DEC_S)
                5.      Merge netearn.test2_ac.sas7bdat with the firm records to create a
                        firm level file
                6.      Create vx1-vx10=
                        sqrt(number of employees at the firm)*(x_bar_j-x_bar_ac)
                        This variable can be used to calculate the variance.
                7.      Output File: netearn.firms_s_t_sort_ac.sas7bdat( 4,145,918
                        observations, sorted by PSI_DEC_S RESID_DEC_S)
7.      06.test2.fdenom.sas
                1.      Calculate the test statistic using netearn.firm_s_t_sort_ac.sas7bdat
                        and netearn.test2_ac as inputs.
                2.      Use proc corr to output the variance matrix and mean vectors
                        (WEIGHTED variance denominator=number of firms)
```

3. Output File: netearn.test2_results.sas7bdat (100 observations with columns psi_decile, resid_decile, number of firms, chi_square contribution for that cell, x_bar_ac, x_bar_a) Across both tables, the minimum cell size is 23,484 unique firms.
4. Output File: test2_results.csv (same as above but in CSV format)
5. Output File: netearn.residuals_fdenom (100 observations with columns psi_decile, resid_decile, (x_bar_ac-x_bar_a)
6. Output File: residuals_fdenom (same as above but in CSV format)

8. 07.test2.wdenom.sas
    1. Calculate the test statistic using netearn.firm_s_t_sort_ac.sas7bdat and netearn.test2_ac as inputs.
    2. Use proc corr to output the variance matrix and mean vectors (WEIGHTED variance denominator=Sum of the weight (total employment))
    3. Output File: netearn.test2_results.sas7bdat (100 observations with columns psi_decile, resid_decile, number of firms, chi_square contribution for that cell, x_bar_ac, x_bar_a) Across both tables, the minimum cell size is 23,484 unique firms.
    4. Output File: test2_results.csv (same as above but in CSV format)
    5. Output File: netearn.residuals_wdenom (100 observations with columns psi_decile, resid_decile, (x_bar_ac-x_bar_a)
    6. Output File: residuals_wdenom (same as above but in CSV format)