```
===========================================
EndMob_Model-replication_archive.zip
April 19, 2016
Contact: Ian M. Schmutte (schmutte@uga.edu)
===========================================

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

./code
    ./01_data_construction
        config.sas
        01.02.subset_sample.sas
        01.03.freq_AKM_inputs.sas
        01.04.prepare_mcmc_vars.sas
        01.05.mcmc_var_stats.sas
        01.06.prepare_mcmc_samples.sas
        01.12.export_mcmc_inits.sas
        01.13.initial_mcmc_gamma_delta_inits.sas
        01.14.export_mcmc_transprobs.sas
        05.02.get_all_ehf_records.sas
        05.03.make_annual_earnings.sas
        05.04.balance_dominant_job_sample.sas
        05.05.get_individual_characteristics.sas
        05.06.get_hcep24_characteristics.sas
        05.07.create_real_ann_earn.sas
        05.08.cleanup_theta_psi.sas
        05.09.sample_reassembled.sas
        06.01.make_akm_mu.sas
        06.04.prepare_mcmc_vars_JBES.sas
        06.04b.prepare_pik_sein_strip.sas
        07.05c.make_mcmc_inits_JBES.sas


    ./02_preprocess
        prepInits.m
        prepHCData.m
        ./graph_coloring
            ./color_greedy_SEQ.m
            ./p01_extract_rmn.m
            ./p02_prepare_graph.m
            ./p03_color_graph.m



    ./03_runs
        runall_half_percent_AKMStart_10classes.m
        ./sampler
            AbilityUpdate.m
            MatchUpdate.m
            ProdUpdate.m
            RunSampler.m

    ./04_postprocess
        00_sample_w_AKM.sas
        export_mean_gibbs.m
        07.01.import_gibbs_model_out.sas
        07.02.br_gibbs_merge.sas
        07.03.br_gibbs_sein_level.sas
        07.04.ln_revenue_per_worker.sas
        ConvertOutput.m
        initSeq.m
        README-POST-20130510.txt
        README_REVENUE_ANALYSIS_2016-03-03.txt
        s01_reviewSamplerOutput.m
        s02_stack_samples.m
        s03_make_stats.m
        s04_corrMat_out.m
        s04_deltaMat_out.m
        s04_gammaMat_out.m
        s04_autocorr_out.m
        s04_latentProb_out.m
        s04_regParms_out.m
        s04_startingValues.m
        s04_wageParms_out.m
        script_run_postprocess.m

    ./05_results
        EmpTransProb_plots.m
        Gibbs_AvgMatchFx_Plot.m
        Gibbs_AvgMatchFX_TransCell.m
        Gibbs_SSAvgMatchFX_Plot.m
        LatentClassProbs.m
        wageParmsPlots.m


./released_data
    autocorr_Out.mat
    betaOut.mat
    corrMat.mat
    deltaOut.mat
    gammaOut.mat
    latentProbOut.mat
    README.txt
    wageParms_Out.mat
```

Overview:
==

The files in this folder are code and associated datafiles used to generate the results reported in `Endogenous Mobility' by John M. Abowd and Ian M. Schmutte. The raw data are restricted access files from the Longitudinal Employer Household Dynamics program of the U.S. Census Bureau. Most of the code in this folder is used to generate research samples from the restricted microdata and then estimate the Bayesian latent class model via the Gibbs sampler as described in the paper. Those parts of the estimation can only be replicated using the same confidential microdata.

The results in our manuscript are based on summaries of the posterior distribution of all model parameters that were approved for release by Census. Those output files can be found in the subfolder ./released_data, where they are stored in MATLAB format. The structure and contents of those files are described in great detail below.

Most of the results in the manuscript are based on manipulations of the released summaries. The code in ./05_results can be used to replicate the information in our output tables and figures when applied to the files in ./released_data.

****

DESCRIPTION OF RELEASED DATA

****

FILE NAME: autocorr_Out.mat
FILE DESCRIPTION: MATLAB output file storing a 2X3 matrix of estimated autocorrelation coefficients and associated MCSE
RESEARCH SAMPLE NUMBER: 1
RESEARCH OUTPUT PROGRAM: Same as above

The proposed release file records the mean and MCSE across all 9,968 samples of within-worker autocorrelation in residuals at lag 1, 2, and 3.

FILE NAME: betaOut.mat
FILE DESCRIPTION: MATLAB output file storing two 6x5 matrices `betaMean' and `betaMCSE'
These are the estimated coefficients of a linear regression of the predicted person, firm, match, and residual wage variation from the endogenous mobility model onto their least squares counterparts. For each MCMC draw, we regress the endogenous mobility values onto the least squares values. The file contains two output matrices: betaMean and betaMCSE. The matrix `betaMean' contains the mean of the regression coefficients across the 7,922 samples. The 5 columns correspond to the 5 endgenous mobility variables. The 6 rows correspond to the 6 regression coefficients (5 least squares values plus a constant). The matrix `betaMCSE' contains the MCSE of each regression coefficient computed using initSeq.m, arranged conformably.

FILE NAME: corrMat.mat
FILE DESCRIPTION: MATLAB output file storing two 10x10 matrices `corrMean' and `corrMCSE'
These are the correlations among log earnings and the parts of earnings attributed to person, firm, match, and residual variation under the endogenous mobility model and their least squares equivalents. For each MCMC draw, we compute the correlation among these 10 variables using the augmented data. The file contains two output matrices: output matrix `corrMean' contains the mean correlation matrix across the samples. The output matrix `corrMCSE' contains the MCSE of each correlation coefficient computed using initSeq.m

FILE NAME: deltaOut.mat
FILE DESCRIPTION: MATLAB output file storing two 10X11X10X11 matrices (deltaMean and deltaMCSE).
These programs compute the mean and MCSE across all samples of the probability of transitioning into a job with an employer of a given type conditional on separation and on the type of the worker, firm, and match in the origin job.

FILE NAME: gammaOut.mat
FILE DESCRIPTION: MATLAB output file storing two 10X11X10 matrices (gammaMean and gammaMCSE).
These programs compute the mean and MCSE across all samples of the separation probability for an employment relationship conditioned on the latent type of the worker, the employer, and the match. The file contains two (10x10x11) matrices: gammaMean and gammaMCSE.

FILE NAME: latentProbOut.mat
FILE DESCRIPTION: MATLAB output file storing two 10x2 stochastic vectors (piA_Out piB_Out) and two 10x10x10 arrays (piK_Mean and piK_MCSE).
The file contains estimates of latent class probabilities from our model. piA gives the probability of being in one of 10 worker heterogeneity classes, piB gives the probability of being in one of 10 employer heterogeneity classes, and piK gives the probability of a job being in one of 10 match heterogeneity classes, conditional on the latent type of the worker and the latent type of the firm in the match.

FILE NAME: startingValues.mat
FILE DESCRIPTION: MATLAB output file storing 10 MATLAB matrices ('Theta0','Psi0','Mu0','Alpha0','Sigma0','Gamma0,'Delta0','piA0','piB0','piK0', 'Beta0')
The file contains variables that summarize the values used to initialize the MCMC sampler. We construct deciles of estimated person, firm, and match effects from the AKM decomposition across the LEHD universe. To seed the sampler, we assign each worker, employer, and match to the decile of its corresponding (AKM-based) worker, employer, and match effect. The variables are described in detail below.

FILE NAME: wageParms_Out.mat,
FILE DESCRIPTION: MATLAB output file storing a 101x87 matrix `wageParms_Out'.
wageParms_Out.mat stores the output as a MATLAB matrix. The variables are the 87 parameters in the wage equation of our model (including non-free parameters determined by imposing linear restrictions after estimation). The data rows are configured as follows: row 1 contains the means from the samples, row 2 contains the MCSE computed using initSeq.m, rows 3--101 contain the 1st through 99th percentile for each parameter in the 7,922 draw sample.

****

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.

1. ./code/01_data_construction
    SAS code to build the analysis sample from raw LEHD Infrastructure files. The locations of raw data are in the library file ./lib/globaldefs.sas. The code is to be run in order according to the sequencing in the file names. Four things happen:
    1. Assemble the basic analysis data
    2. Draw a sample for estimating our model
    3. Generate starting values for the Gibbs sampler based on the AKM estimates of person and firm effects
    3. structure the data to pass to MATLAB

2. ./code/02_preprocess
    MATLAB code to read and reformat input data. We also apply the graph coloring alorithm to find disconnected groups of firms to facilitate parallelization. This code generates all of the MATLAB files needed to run the Gibbs sampler

3. ./code/03_runs
    runall_half_percent_AKMStart_10classes.m is a MATLAB script that sets configurable parameters, loads all needed files, and runs the Gibbs sampler. Our results are based on three parallel runs that were launched using this file.

4. ./code/03_runs/sampler
    The MATLAB functions and script files used by the Gibbs sampler.

5. ./code/04_postprocess
    MATLAB functions and scripts used to process the stored samples from the Gibbs sampler. This includes code to generate summaries of the AKM starting values, code to compute the Monte Carlo Standard Errors, and code that generates the posterior mean and MCSE for each estimated parameter. An associated README file describes how the code is to be run on output from the Gibbs sampler. This code sequence generates the complete set of data files in ./released_data that were released from the RDC, and upon which the results in the paper are based.

    Code in this folder also combines information from the AKM estimates and Gibbs output to perform the firm-level analysis of the relationship between earnings and revenue per worker. A README file (README_REVENUE_ANALYSIS_2016-03-03.txt) is included that describes the data files and associated code sequence.

6. ./code/05_results
    MATLAB scripts to generate summaries of the results stored in ./released_data. These generate images and .csv files on which the figures and tables in the paper are based.

VARIABLE DEFINITIONS
VARIABLE NAME: LnEarnReal
DEFINITION (include type, e.g., continuous, binary, etc.): Continuous. Dependent variable in

the wage equation.
SOURCE: gs_postProcess.m, initSeq.m, and SAS program 01.01.post_process.sas, and gs_postProcess_2_wageParms.m

VARIABLE NAME: Theta_1--10
DEFINITION: Continuous. Ability (individual) effect in the structural wage equation.
SOURCE: same as above

VARIABLE NAME: Psi_1--10
DEFINITION: Continuous. Productivity (employer) effect in the structural wage equation {.}
SOURCE: same as above

VARIABLE NAME: Mu_1--10
DEFINITION: Continuous. Quality (match) effect in the structural wage equation.
SOURCE: same as above

VARIABLE NAME: Alpha
DEFINITION: Continuous. Constant in the structural wage equation.
SOURCE: same as above

VARIABLE NAME: Sigma
DEFINITION: Continuous. Std. Dev of error term in the structural wage equation.
SOURCE: same as above

VARIABLE NAME: ThetaGibbs
DEFINITION: Continuous. For a given observation in the input data, the realized value of the ability (individual) effect in the structural wage equation.
SOURCE: same as above

VARIABLE NAME: PsiGibbs
DEFINITION: Continuous. For a given observation in the input data, the realized value of the productivity (employer) effect in the structural wage equation.
SOURCE: same as above

VARIABLE NAME: MuGibbs
DEFINITION: Continuous. For a given observation in the input data, the realized value of the quality (match) effect in the structural wage equation.
SOURCE: same as above

VARIABLE NAME: ResidGibbs
DEFINITION: Continuous. For a given observation in the input data, the realized value of the residual from the structural wage equation.
SOURCE: same as above

VARIABLE NAME: ThetaAKM
DEFINITION: Continuous. For a given observation in the input data, the realized value of the ability (individual) effect in the exogenous mobility wage equation.
SOURCE: same as above

VARIABLE NAME: PsiAKM
DEFINITION: Continuous. For a given observation in the input data, the realized value of the productivity (employer) effect in the exogenous mobility wage equation.
SOURCE: same as above

VARIABLE NAME: MuAKM
DEFINITION: Continuous. For a given observation in the input data, the realized value of the quality (match) effect in the exogenous mobility wage equation.
SOURCE: same as above

VARIABLE NAME: ResidAKM
DEFINITION: Continuous. For a given observation in the input data, the realized value of the residual from the exogenous mobility wage equation.
SOURCE: same as above

VARIABLE NAME: piA_1--10
DEFINITION: Continuous. Probability that a worker belongs to one of 10 latent ability classes.
SOURCE: same as above

VARIABLE NAME: piB_1--10
DEFINITION: Continuous. Probability that an employer belongs to one of 10 latent productivity classes.
SOURCE: same as above

VARIABLE NAME: piK_XXYYZZ
DEFINITION: Continuous. Probability that a match between a type XX worker and a type YY firm is in quality class ZZ where XX,YY, and ZZ run from 1--10.
SOURCE: same as above

VARIABLE NAME: gamma_XX_YY_ZZ
DEFINITION: Continuous. Separation probability for a match of quality class ZZ between a type XX worker and a type YY firm. XX and ZZ run from 1--10. YY runs from 1--11
SOURCE: same as above

VARIABLE NAME: delta_XX_YY_ZZ_MM
DEFINITION: Continuous. Probability of arriving in an employer of productivity type MM after separating from a type ZZ match in a type YY employer for worker of type XX. YY and MM run from 1--11. XX and ZZ run from 1--10.
SOURCE: same as above

VARIABLE NAME: Theta0
DEFINITION Estimated 10x1 vector of person class earnings effects associated with the initial classification of data according to the discretized (by population decile) AKM earnings components
SOURCE: StartingSummary.m

VARIABLE NAME: Psi0
DEFINITION Estimated 10x1 vector of employer class earnings effects associated with the initial classification of data according to the discretized (by population decile) AKM earnings components
SOURCE: StartingSummary.m

VARIABLE NAME: Mu0
DEFINITION Estimated 10x1 vector of match class earnings effects associated with the initial classification of data according to the discretized (by population decile) AKM earnings components
SOURCE: StartingSummary.m

VARIABLE NAME: Alpha0
DEFINITION Estimated constant in the earnings equation associated with the initial classification of data according to the discretized (by population decile) AKM earnings components
SOURCE: StartingSummary.m

VARIABLE NAME: Sigma0
DEFINITION Estimated standard deviation of the residual variance in the earnings equation associated with the initial classification of data according to the discretized (by population decile) AKM earnings components
SOURCE: StartingSummary.m

VARIABLE NAME: Gamma0
DEFINITION 10x11x10 matrix whose entries are the frequency of separation conditional on the worker employer and match class, using the initial classification of data according to the discretized (by population decile) AKM earnings components
SOURCE: StartingSummary.m

VARIABLE NAME: Delta0
DEFINITION 10x11x10x11 matrix whose entries are the observed frequency of transition to an employer of each latent type (including non-employment; class firm class 11) conditional on

separation and stratified by the worker employer and match class, using the initial classification of data according to the discretized (by population decile) AKM earnings components.
SOURCE: StartingSummary.m

VARIABLE NAME: piA0
DEFINITION 10x1 vector whose entries are the observed proportion of workers in each latent worker class, using the initial classification of data according to the discretized (by population decile) AKM earnings components.
SOURCE: StartingSummary.m

VARIABLE NAME: piB0
DEFINITION 10x1 vector whose entries are the observed proportion of employers in each latent employer class, using the initial classification of data according to the discretized (by population decile) AKM earnings components.
SOURCE: StartingSummary.m

VARIABLE NAME: piK0
DEFINITION 10x10x10 matrix whose entries are the observed proportion of matches in each latent match class, stratified by worker and employer class, using the initial classification of data according to the discretized (by population decile) AKM earnings components.
SOURCE: StartingSummary.m