Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts

Benjamin R. Handel

March, 2013

README: Instructions for Data Files / Code

This README file contains instructions for how to use the simulated data and code provided with this paper. While this document attempts to be comprehensive, the user can email the author handel@berkeley.edu with any additional questions on how to use the materials provided.

The data used in the main analysis are proprietary and protected by HIPAA rules governing Limited Data Sets. Proprietary here means that there is an agreement with the firm whose data are used that those data will not be placed in any public domain. Additionally, the agreement stipulates that the firm will not be identified or referred to by name in any public domain. HIPAA regulations stipulate that the type of detailed data we use for this study fall under the confines of a 'Limited Data Set.' For access, these regulations stipulate that only a direct and small group of researchers have access to the full data.

These restrictions mean that the author cannot share the actual data associated with the analysis in a public domain. The author is happy to discuss obtaining access to these kinds of claims datasets *in general* (e.g. providing instruction on what should and should not be done to acquire them from similar firms) but because of the identification restrictions the author cannot put other researchers in touch with the same firm these data are from to obtain access to those data.

As a result of these restrictions, the author has created simulated data that are similar in spirit to the actual data, so that readers / users can run the code / models studied in this paper (provided with the submission). While these simulated data have no actual link to the true data, they do allow users to work in an environment with similar variables and understand how to perform the analysis from the paper.

The remainder of this README files describes how to use the simulated data as well as the analytical code provided with this submission.

Cost Model

The first part of the structural analysis in the paper, described in section 4 in the main text, is to estimate a cost model that predictions ex ante distributions of out-of-pocket spending for each family and each potential plan that they could choose from the available set. The first step in this process is taking detailed medical claims and cost data and estimating distributions of total predicted health expenditures at the individual level. The second step is taking those individual estimates of total costs, and generating plan and family specific distributions. Note: these processes are described in detail in Online Appendix A.

The first code file supplied is:

'Cost_Model_Estimation-STATA_2011_1284.txt'

This file provides STATA code that describes how to estimate the individual total cost distributions by category of medical expenditure starting from individual level data on costs and mean predicted medical spending, output by the Johns Hopkins ACG program referenced in the paper. There is no corresponding simulated data to run this part of the code because (i) it requires very detailed claims / expense data and (ii) it is quite straightforward to run with another given dataset (both in practice and conceptually).

The first simulated data provide pick up from the end of the previous code file. These data are provided in the file:

'ASIN-Handel-SimulatedData_2011_1284.mat'

These data are provided in MATLAB format, which all subsequent code is for. These simulated data contain simulated data that mimic the output of the 'Cost_Model_Estimation-STATA_2011_1284.txt' code file. These data (each specific variable included) are described in detail in the second-half of the document:

'Data_Description_ASIN_2011_1284_Handel.pdf'

These data contain cost model estimates of total individual expenditures by category of medical expenditures. The next step is to take these estimates and create distributions of family-plan-time specific out-of-pocket expenditure risk which is used as an input into the choice model. This code for this process is supplied in the code file:

'Cost_Model_Implementation_2011_1284.m'

The output from this file, the distributions of out-of-pocket risk plus detailed individual demographics and attributes, would, in the main analysis, be a direct input into choice model estimation. Here, with simulated data, there is an additional interim step where the author has to simulate choices to add to these data based on the underlying fundamentals and parameters

similar to those estimated in the actual data. So, in the *actual* the data at this stage would already contain observed choices, while in the simulated data these have to be simulated in an interim step. As a result, one cannot take the output of 'Cost_Model_Implementation_2011_1284.m' run on 'ASIN-Handel-SimulatedData_2011_1284.mat' and use it as a direct input into the choice model estimation code.

Choice Model

The author has simulated choices based on the output of the 'Cost_Model_Implementation_2011_1284.m' code and processed the output into a data file that should be used as the direct input into the choice estimation code. In these data, the main unit of analysis / observation is the *family* while in the cost model implementation described above the data were at the individual level and being translated to the family level. These data are provided in the file:

'ASIN-ChoiceModelData-FINAL_2011_1284.mat'

These data (each specific variable included) are described in detail in the first-half of the document:

'Data_Description_ASIN_2011_1284_Handel.pdf'

These data can be used as a direct input into running the choice model estimation code. The choice model main estimation code is supplied in the file:

'EstimationCode_2011_1284.m'

This file shows how to run the non-linear optimization routine for the simulated maximum likelihood estimation, described in detail in Online Appendix B. The maximum likelihood routine uses a likelihood function which evaluates the parameters for each iteration of the optimization. The likelihood function used is supplied in the file:

'Likelihood_2011_1284.m'

Both the estimation code main file and the likelihood function have substantial commenting to help the user understand what is going on and to allow them to adapt the code in various ways to fit their own specific settings.

The estimates from the choice model, run on the simulated data, are provided at the end of the 'EstimationCode_2011_1284.m' file right after the non-linear optimization routine. This part of the code supplies *both* the estimated parameter values *and* the 'true' parameter values that the authors simulated the choices from in the interim step between the cost model and choice model estimation described above (again, this is only necessary in simulated data, since in the actual data choices are observed). The estimation results are close to the 'true' underlying parameters, and, as the number of families simulated becomes larger these numbers should converge to become the same. The estimation code file discusses this point in some more depth.

Sample code for one method for calculating standard errors is also supplied. In the analysis, we computed these parameter by parameter by inverting a standard likelihood ratio test and finding the parameter value that is equivalent to the 95 pctile (or 5^{th} pctile) of the parameter estimate distribution, which, assuming asymptotic normality is 2 standard deviations away from the true parameter estimate. This value supplied by the code can then be divided by 2 to find the parameter standard error. Note: This does not calculate the full covariance matrix of the parameters, just the diagonal standard element of this matrix (the variance / standard error) for each parameter. The code supplied is to compute the standard error for one parameter, the mean of the risk preference estimate, and similar code can be applied to compute these for all other parameters. This code, for expediency, has two parts, an optimization routine and a function that that optimization routine calls. These files are:

'Standard_Error_OneParam_2011_1284.m'

'Standard_Error_Function_2011_1284.m'

The final code file supplied is the code to compute the counterfactual simulations where inertia is reduced from the estimated value. Here, the code takes in the same simulated data file as the choice model estimation 'ASIN-ChoiceModelData-FINAL_2011_1284.mat' and performs the counterfactual analysis based on the assumed reduction in inertia and the estimated parameters from the simulated data. While the paper performs this analysis for a range of welfare treatments of inertia and a range of assumed inertia reduction, the code provides computes the impact when inertia is reduced by 75% and when inertia is not welfare relevant in and of itself. Also, while the main analysis studies the impact of reduced inertia over a 7 year time horizon, the code supplied does so for 5 because this code is repetitive and users can easily extend the code to add these extra years. The code comments easily how these assumptions can be changed, see tables 7 and 8 in the main paper for the full analysis of these issues with the actual data. The code file for all components of the counterfactual analysis, choices and welfare, is supplied as:

'Counterfactuals_2011_1284.m'

This summarizes all the files provided with the paper for readers to use to understand both how the code works and to actually use the code with the simulated data provided. Again, readers with questions should feel free to contact author Ben Handel at handel@berkeley.edu with any questions not answered in this document or in comments in the code.