Readme File for "One-Child Policy and the Rise of Man-Made Twins"

Wei Huang, Xiaoyan Lei and Yaohui Zhao

This readme file will help researchers to replicate the results in our paper.

1. **Public available data**

We here provide the public available data:

Census data: **Census 1982 (1982dta.dta) and Census 1990 (census1990.dta)**;

These two data can be downloaded at China Survey Data Network. To do so, you need to register first. The contact is on the website. Tel: (86)10-62759399 and Email: csdn@ccer.edu.cn.

CHNS relevant files: **m10pexam.dta (raw data), m10relatives.dta (raw data) and twins_chns_use.dta** (Intermediate data produced to run regressions). The CHNS data are available at the website below, and we downloaded the data in 2014. (http://www.cpc.unc.edu/projects/china.)

The main variable in this paper, One-Child Policy fine, **fines.dta** are provided, too. This file are from Avi Ebenstein (ebenstein@mscc.huji.ac.il).

2. **Proprietary data and How to get access**

The proprietary data used in this study are **Census 2000 and 2005 Population Study sample.** These data are not provided due to our data use agreement with Census Bureau. The ways for other researchers to get access to the proprietary census data are as follows.

There are three requirements for the potential readers to get the data. Formal access: 1. cooperate with Census Bureau and National Bureau of Statistics of the People's Republic of China（http://www.stats.gov.cn/ ）on a relevant research project which will use the data; 2, submit the data application and proposal in person to the census bureau in China and get their formal approval; 3, get the census bureau's approval and they will share the official sample of the data with the (approved) scholars under certain conditions. In each year, there are some collaborative projects with the bureau and that's the reason some scholars in China have the access to the data. Our data access is based on a correlation project and the Data Use Agreement.

Because the strict selection among the projects with census bureau in China, as far as we know, some scholars have another version of the confidential data. We have replicated our results using this sample and find the results are close and consistent. Before doing so, please make sure the variables names are used consistently across different datasets.

To help the scholars to better understand our results and to replicate them, we

provide detailed do files on line to show how we get our results in details and also provide an automatically generated log file to check. If the reader wants to replicate part of our results with a different model or other details about our sample, please contact Wei Huang (weihuang@fas.harvard.edu), corresponding author of this paper, and send your Stata 13 codes to him. Wei could help you to get the results done and send you relevant log files and results. We are not allowed to give data access to anyone else based on the data use agreement.

3. Documentation of the proprietary data

To help other researchers to replicate the results in our paper, we use Stata 13 to document the census data 1982, 1990, 2000 and 2005. Because the original data are labeled in Chinese, this do file include all the English interpretation of the variable contained in the raw data. In addition, the original questionnaires for the four waves data are also provided, including **82census.jpg, 1990 Census 问卷.jpg, 2000 Census.pdf, 2005 年全国 1%人口抽样调查表.xls.** These are original data we get from the census bureau.

To make the results more replicable, we provide the do file **"Twins_Documentation.do"** to show summary statistics *(Observataions, Means, Standard Deviations, Minimum value, Maximum value)*, description *(Variable definitions, storage type, format)* as well as the correlation matrix between key variables in the log file named **Twins-Document.log.**

4. Results in xls and txt

All the xls and txt files show the main outcomes of this project. We present the version xls and txt file used for the results in the paper. Please allow very very small differences in the number of observations (<20 observations relative to over 5 million) in the regressions in the log file. This difference is caused due to drop in the duplicated observations in Stata.

For the CHNS data, we provide the do file (also in the Twins_Data.do), raw data (said above), the log file (also in Twins-Data.log) and the regression results (table4.xls and table4.txt).

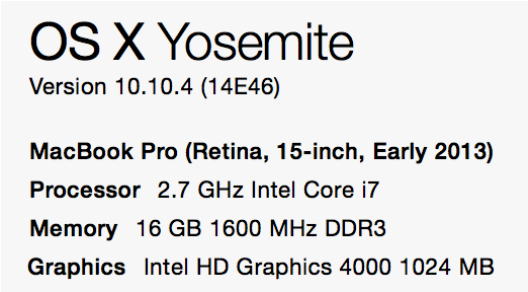
5. Analysis codes and computer details

Twins_Data.do provides the analysis codes for our project and Twins-Data.log provide corresponding log file **Twins-Data.log** for the readers to check. This do file can be smoothly conducted on the condition that the **proprietary data – Census 2000 and Census 2005, are located in the "raw data" path, and the current path is the current folder.**

The software the authors use is Stata 13.1 MP. Current update level is 17 Apr 2015. Note that we use programs named outreg2 and outsum in the Stata do file. Please install them before running your codes. The software info:

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   13.1   Copyright 1985-2013 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     MP - Parallel Edition            College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

14-user 8-core Stata network perpetual license:
       Serial number:  501306203335
         Licensed to:  National Bureau of Economic Research
                       Cambridge, MA 02138
```

Operation System:

The operation system of the authors are mac OS X Yosemite. The OS info: