README File for Data and Programs

"Violence While in Utero: The Impacts of Assaults During Pregnancy on Birth Outcomes"

Review of Economics and Statistics

RESTAT MS#23600

Janet Currie, Michael Mueller-Smith, Maya Rossin-Slater

A. New York City Births, Crimes, and Building Characteristics Data

i. Raw Data and Data Linkage

Here, we describe the sources of our three primary administrative data sets from New York City, and our process for linking them.

First, we used identifiable birth records from the New York City Department of Health and Mental Hygiene's Office of Vital Statistics. These records contain mothers' full maiden names, social security numbers, dates of birth, and residential addresses. Instructions for requesting similar access can be found here:

https://www1.nyc.gov/site/doh/data/data-sets/data-requests-application-process-for-identifiable-vital-statistics-data.page.

Second, we used confidential exact geocoded crime reports from the New York Police Department (NYPD). Our research project was added as an amendment to an existing agreement between the NYPD and the NYU Furman Center for Real Estate and Urban Policy that was proposed to and approved by NYPD Assistant Commissioner Phillip McGuire.

Third, publicly available building characteristics data was downloaded from the NYC Department of City Planning. We used the Primary Land Use Tax Lot Output (PLUTO) data, which can be found here: https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page.

One member of our research team (Michael Mueller-Smith) was allowed to access and merge these data sets on-site at the DOHMH. Then, these data were stripped from many of the identifiers (as described below), and released to us for analysis on the National Bureau of Economic Research (NBER) secure server. Our project has been approved by the NBER Institutional Review Board (IRB Ref. #14_006).

In order to link these data to each other, we used the standardized Bureau, Block, Lot (BBL) tax identifiers which are unique to New York City and uniquely identify buildings in the city. The PLUTO records already contains the BBL variable. In order to add the BBL to the birth records,

we used the New York City Department of City Planning's program known as "Geosupport" (specifically NYCgbat.exe), which can be downloaded here:

https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-gde-home.page.

We used Geosupport to perform fuzzy matching from maternal address on the birth certificate to the BBL. Finally, we used ARCGIS to link geocoded crime data to the nearest BBL using the NYC BBL shapefile available here:

https://data.cityofnewyork.us/Housing-Development/Building-Footprints/nqwf-w8eh.

We then created variables capturing the number of months between the criminal event and the child's birth month, as well as the number of months between the criminal event and the child's conception month: **Crime_RMB** and **Crime_RMC**.

Before releasing the data onto the NBER server, the data sets were stripped from exact maternal address information. Maternal social security numbers, full maiden names, and exact dates of birth were replaced by scrambled anonymized versions. Each birth record was assigned a unique identifier called **Study_ID.**

The data sets released to us for analysis on the NBER server are called:

birth_outcomes_chars.dta

maternal_address_chars.dta

direct_crime.dta

A codebook for all of the variables in these data sets is provided in the Excel spreadsheet titled **DataDictionary_NYC_births_crimes_bldgchars.xls**

We then used these data for most of the analysis, as detailed below.

ii. Analysis Code

The analysis is conducted using **Stata-16** software:

1) **clean_birthoutcomes.do** uses **birth_outcomes_chars.dta** and prepares it to be merged to the maternal address characteristics and crime data sets. It generates most of our outcome and control variables. The resulting saved data set is called **birth_outcomes_CLEAN.dta.**
2) **mergebirths_tocrime.do** first cleans **direct_crime.dta** to calculate the number of different types of crimes by month relative to conception. It then merges these data to the cleaned birth outcomes data generated in step (1), and then merges to **maternal_address_chars.dta** that provides information on building characteristics. These three data sets are merged to one other using **Study_ID**. The resulting saved data set is called **births_crimes.dta.**

3) **singlefamilyhome_analysis.do** uses **births_crimes.dta**, does some additional data cleaning, and then produces the following output in the manuscript: Figure 1, Tables 1 and 2 (columns 2-5), Table 3 (rows 1-2), Appendix Figure A2, and Appendix Tables A4, A5, A7, A8, A9, A10-Panel A, A11. It also produces output for Footnote 20 (number of assaults that occur after actual birth but before expected month of birth), the text in the Empirical Design section (F-stat for predicting assault during pregnancy), Footnote 28 (effects of burglary), and Appendix C (comparison with PRAMS data).
4) **singlefamilyhome_SIBS.do** uses **births_crimes.dta**, merges siblings, does some additional data cleaning, and then produces the following output in the manuscript: Table 3 (row 3), Appendix Tables A9 and A12.
5) **multifamilyhome_analysis.do** uses **births_crimes.dta**, does some additional data cleaning, and produces the following output in the manuscript: Appendix Table A10, Panel B. It also produces output for the text in Appendix E (re-weighting multi-family sample to be more representative of single-family sample).
6) **subsequent_fertility.do** uses **births_crimes.dta**, does some additional data cleaning, and produces the following output in the manuscript: Appendix Table A6.

B. Other Data Sets and Analyses

i. National Vital Statistics Data.

We used national vital statistics data from the National Center for Health Statistics (NCHS) for births in years 2006-2013 to create means of maternal and paternal characteristics in column (1) of Tables 1 and 2. We have access to the restricted version of these data (with geographic identifiers), and are therefore unable to post it publicly. Researchers can access the restricted version of the data by following instructions here: https://www.cdc.gov/nchs/nvss/nvss-restricted-data.htm.

The data are input into Stata using the code **input_nationalVS.do**, and then the statistics are generated in **nationalVS_stats.do** and **nationalVS_stats2.do**.

ii. NYPD Crime Data summary statistics.

nypd_summary_stats.do uses **Crime.dta**, the confidential geocoded crime data from the NYPD, to create the summary statistics for the NYPD crime reports presented in Figure A1 Panel b and Table A3.

iii. National Crime Victimization Survey Data.

ncvs_restat.do uses **NCVS_PERSONAL_VICTIMIZATION_1993-2018.csv**, which is publicly available at https://www.bjs.gov/developer/ncvs/index.cfm, to produce Panel A of Figure A1.

iv. Fragile Families and Child Wellbeing Study.

nyc_birth_ff_restat.do uses **ffmombspv3.dta** and **ffmom1ypv2.dta** (data publicly available to registered users of the Fragile Families and Child Wellbeing Study here: https://opr.princeton.edu/archive/restricted/Default.aspx) to produce the analysis presented in Appendix Tables D1 and D2.

v. PRAMS data.

Table C1 was constructed using a variety of sources. We downloaded and summed the annual count of the size of the estimated affected population of individuals self-reporting physical abuse during pregnancy in New York City from the PRAMStat System (https://www.cdc.gov/prams/prams-data/index.html). This data reporting system is no longer active, but the same information can currently be downloaded from this website: https://apps.health.ny.gov/public/tabvis/PHIG_Public/prams/reports/#annual. We took this estimate (28,593 pregnancies) and multiplied it by the average victim reporting rate to law enforcement for known violent offenders from the National Crime Victimization Survey (42%), to generate an estimate of the underlying true number of pregnancies with physical abuse reported to law enforcement from New York City between 2004 and 2012 (12,009 pregnancies). **singlefamilyhome_analysis.do** uses **births_crimes.dta** to produce the "Total Affected Pregnancies" row in Table C1, and the following two rows are calculated using the following:

Share mismeasured relative to PRAMS baseline: $1 - \frac{12{,}009}{"Total\ Affected\ Pregnancies"}$

Implied Scaling Factor for Estimates: $\frac{"Total\ Affected\ Pregnancies"}{12{,}009}$