The Causes of Crime and Violence: 
A Guide for Empirical Researchers

Pablo Fajnzylber, Daniel Lederman and Norman Loayza

February 1999

1. Crime Data Sources

To facilitate comparisons across the case studies being commissioned for this project, some preliminary analyses should be common across all the studies in the project. As a first step, each study should provide a description of the available sources of crime data. All possible sources should be considered at this point, even in the cases where, because of possible limitations of the data, some sources will not be used in the statistical analysis of the causes of crime. This analysis should include a detailed international comparative analysis of the types of data sources available in several countries.

There are clearly two functions of this study. The first is to measure violent crime in each country or city and its trend over time. The second is to address the causes of violent crime. The same tools, i.e. basic data sets, will be used for each task. In addition, all the studies should shed some light on the particular causes of violent crime in their countries by comparing them to other countries. Hence, an international focus is essential in these studies.

The first type of data set is collected from victimization surveys. This is unquestionably the primary workhorse for measuring crime. To be really useful, this sort of data must have geographic identifiers which enable the research to link the individual to the community in which he lives and was robbed. Then the researcher also needs to know attributes of that community. These attributes may be drawn from broader national surveys or national census.

The second type of data set relies on official crime statistics. This data set is useful but relies on the quality of reported crime (and arrests) data. With this sort of data at the locality level, one might run regressions connecting city level characteristics with the level of reported crime. However, reporting (and recording) problems in some of the countries included in this project are so problematic that this type of data set has many problems.

The third type of data set involves homicide (or intentional injury) data from hospitals and morticians. This type of data is generally more reliable than official crime statistics. It is also less sensitive to changing definitions of crimes across cultures. In

---

1 Fajnzylber is Professor of Economics at the Federal University of Minas Gerais, Belo Horizonte, Brazil; Lederman is Economist with the Office of the Chief Economist for Latin America and the Caribbean (LAC) of the World Bank; Loayza is Senior Economist with the Central Bank of Chile and is on leave for the Development Economics Research Group of the World Bank. The document was written to serve as the terms of reference for the case studies to be written by Latin American research groups in five countries for the project, “Crime in Latin American Cities,” supported by the Regional Studies Program of the LAC Region of the World Bank. We are grateful for the suggestions and advice provided by professors Edwards Glaeser (Harvard University) and Jeffrey Grogger (UCLA).
principle, data on the victim can also provide a rich set of stylized facts about the nature of homicides in the country in question.

A final data source is the offender survey. This type of survey can both be taken through traditional survey methods where respondents are asked if they have been arrested (or less reliably if they have committed a crime). Alternatively, these surveys can be done at the point of arrest, or through surveys of prison population. For that type of survey to be effective it must be assumed that the police arrest a relatively random sample of the population of criminals. When this type of data set is merged with data on the population at large, then it is possible to identify how criminals differ from average citizens.

In addition, anecdotes and case studies may present valuable information on questions where data is hard to come by. All of these data sets together help to form a picture both of the level of crime and to identify potential causes of crime.

The presentation of the available sources of crime, victimization and offenders data should be summarized by means of a table including four columns: Data Source, Coverage, Description, and Observations. The second column should provide information on both the spatial and temporal coverage of each data source. In the case of the data from health and police reports, this column should also indicate whether crime and population data is available at the sub-city level—e.g. police precincts, neighborhoods or census units. The column labeled “Description” should include the list of variables available from each source, such as the types of crimes covered and, in the case of victimization and offender surveys, also the information provided by the surveys regarding demographic and/or socioeconomic characteristics of victims and/or offenders. The column labeled “Observations” should comment on possible limitations of the data and/or the methodologies used for its construction.

2. Descriptive Statistics on Crime and Victimization Rates

The second preliminary analysis to be undertaken by all case studies is the presentation of descriptive statistics on crime and victimization rates. Each study should produce a reliable estimate of the amount of violent crime in each country based on a variety of sources. The study should show both the level of crime, in a way that will be comparable across studies, and the trend of crime for as long a period as possible. If possible, crime rates should be presented at the national, the city and the sub-city level. This preliminary analysis should be applied in a comparative framework, where each country's statistics are compared and contrasted with those in other countries.

In order to provide as complete a picture as possible of the crime problem in each city, all the available sources should be used, and an analysis of their consistency should be undertaken. It would be particularly useful to see comparisons of homicide rates (or even homicide counts) from police and public health (or coroner's) reports. The comparison of homicide data from different sources will help provide a gauge of the adequacy of each country's aggregate data collection systems. Likewise, each study should include a simple analysis of age-offending (and age-victimization) profiles for major categories of crimes. Time and age profiles have proven invaluable to the development of theory and evaluation
of policy in the United States, and it seems likely that they would prove similarly valuable in the countries represented in this project.

Below, a suggestion of the steps to be taken by each study is presented:

i) **Aggregate data on homicides from health authorities**
   - Trend data should be presented for as long a period as possible, at both the national and the city level.
   - If possible, homicide rates (per 100,000 population) should be calculated for different demographic groups, formed on the basis of gender and age (i.e. males, females, population aged less than 15, between 15 and 34, and older than 34).
   - If data is available at the sub-city level, summary statistics of the distribution of homicide rates within the city should be presented.

ii) **Aggregate data on homicides and other violent crimes using police data.**
   - Trend data on crime rates (per 100,000 population) for as long a period as possible, at both the national and the city level.
   - If data is available at the sub-city level, summary statistics of the distribution of homicide and other crime rates within the city should be presented. The bivariate correlation between the various crime rates should be calculated.
   - An analysis of the consistency of the homicide data from health and police sources should be made. First, by comparing the levels and the trends of the crime rates calculated with data from different sources. Second, by calculating the correlation between the homicide rates calculated at the sub-city level using health and police data.

iii) **Victimization Surveys**
   - For each city and survey, a summary table should be presented containing the means (averages) of victimization (the rate of victims to the total number of individuals sampled) by type of crime and by type of individual (i.e. what is the victimization rate of young men, etc.).

3. **The Causes of Crime**

   As emphasized by professor Glaeser, we can think of the determinants of crime as coming in six separate categories: (1) opportunity cost of time, (2) social disorder, (3) returns to crime, (4) drugs and gangs, (5) weapons and (6) enforcement. Below we review the basic ways that can be used to test these different theories, and then present a suggested set of analysis to be undertaken by each study to investigate the possible determinants of crime rates and the probability of victimization.

   The opportunity cost of time is customarily measured by wages or unemployment, or other measures of poverty. In the case of victimization studies, this theory is tested either with wage variables at community level (so the subject’s probability of victimization is regressed on the average wage level in his community). This theory is also tested with the
characteristics of the victim. The assumption for this type of test is that victims often resemble criminals. This theory can also be tested with cross-sectional regressions where the poverty, inequality, wages or unemployment of an area are used to explain the level of crime. Ideally, it would be tested with offender level data where it can be examined whether offenders with lower opportunities in the legal sector are more likely to engage in crime.

Another recommendation that is applicable to most of the proposals is that, if possible, wages AND unemployment rates should be used as explanatory variables to estimate the effects of the labor market on crime. For years, researchers in the United States have noted that unemployment and crime are largely uncorrelated, and have concluded that labor market factors play little role in determining crime. Grogger (1998) analyzed this issue with individual-level data from the U.S., however, and found that wages were a strong determinant of crime even though crime was essentially independent of employment. The lesson is that the link between crime and the labor market is best measured using wages. The project would benefit by examining this particular question, since these findings for the U.S. may not hold in certain contexts that are common in developing countries (e.g., lack of effective unemployment insurance).

Social disorder (and social capital) are generally measured with variables like the level of migration into the area, the number of single-parent families, survey type questions on trust, and other non-economic variables. All studies should be able to test this hypothesis by using this type of variable at the community level and using victimization data. Cross-sectional or offender level data can also be used to test this hypothesis.

The returns to crime are generally measured by the proximity of rich people or business (or tourists) to poor people. This can be tested by using community level variables in cross-sectional studies. It can be tested in the victimization studies by examining whether controlling individual income raises the probability of victimization of the individual holding the level of poverty in the community constant.

Drugs and weapons are much harder to measure. In principle, disruptions to the drug trade or exogenous inflows of guns might be measured in particular cases. Individual studies should certainly try to do this when possible. However, measurement difficulties make it hard to be certain that either of these theories can be tested reliably.

Enforcement has a major identification problem because countries tend to focus enforcement activities where there are higher levels of crime. Thus empirically we tend to identify a false positive correlation between the level of enforcement and the level of crime. However, it still important to try and measure enforcement and examine how it differs across space and what the correlation between enforcement and crime may be. When reliable data on reported crimes and arrests is available, then the arrest rate may serve as a plausible proxy for enforcement. Alternatively, measures of the number of police personnel or expenditures on police may be used.

A possibility to deal with the spurious correlation between crime and enforcement variables is to use an instrumental variables approach to estimate the effect of law enforcement expenditures on violent crime rates. As mentioned, the problem with just regressing crime rates on expenditures is endogeneity: governments raise law enforcement
spending when crime is high, so one typically finds a positive relationship between crime and spending on police. What one needs is an instrumental variable that isolates spending increases that occur for reasons other than an increase in crime; Levitt (1997) proposes using electoral cycles, on the basis that the party in power may raise police spending in order to elicit votes. The approach works well on U.S. data, where there is political competition. Yet we understand that finding "good" instruments may be troublesome, and we leave it up to each research team to assess whether a particular instrumental variables approach is feasible. Nevertheless, if proper instruments are not found, and the regressions are run without instruments, then the interpretation of the resulting coefficients must consider the problem of endogeneity explicitly.

In many countries, the police are themselves a major part of the crime problem. Instead of serving as the servants of the public, police often support criminal activities and engage in them themselves. To measure this, questions regarding crimes by the police must be included in victimization studies. Anecdotal evidence (hopefully with some data) can also help to explore this issue.

Below we present the minimum steps that each study should take in order to implement the analysis of the causes of crime. It is worth emphasizing that these steps constitute a lower, rather than an upper bound for each study. Certainly each study would do well to present as much data and qualitative analysis as possible, using a wide range of different data sources and highlighting the issues that are the most relevant for the city or country. For instance, in the case of Brazil, the dissimilar incentives faced by the police in Rio and Sao Paulo deserve special attention; and in the case of El Salvador, the remains of the civil war (a demobilized army, a large number of weapons) must be taken into account. Similarly, in the case of Colombia, the drug trade is without doubt the most important issue. Thus, we urge each research team to deal with the issues that make crime and crime-prevention unique in each city. Hopefully, they can be dealt with in a statistically rigorous way in the context of analysis of victimization and crime rates. However, we understand that in some cases qualitative rather than econometric analysis is called for. In any case, what we are asking from every research team is to go well beyond the minimum requirements for the study and address in an original way the issues that are most relevant to their cities.

i) Individual Risk of Victimization

- Victimization surveys constitute the main data source for the study of the determinants of individual risks of victimization. However, as already mentioned, if geographic identifiers are available that allow the researcher to link individuals to the communities in which they live and/or were victimized, those surveys should be complemented with data on the social and economic characteristics of the corresponding communities. This data should be extracted from broader national household surveys and/or national census.

- Before proceeding to the estimation of a model of the individual risk of victimization, each study should present a table with a summary description of the explanatory variables that will be used in the analysis. This table should have four columns: Name of the Variable, Unit of Analysis (individual, household, neighborhood, police
precinct, etc.), Data Source, Year of Reference (e.g. the census year when this is the data source).

Next, a table with summary statistics for the explanatory variables (presented in the preceding table) should be constructed. This table should have six columns: Name of the Variable, Mean (and Standard Deviation in parenthesis) for the whole sample, Mean (and Standard Deviation in parenthesis) for non-victims, Mean (and Standard Deviation in parenthesis) for victims of non-economically motivated crimes, Mean (and Standard Deviation in parenthesis) for victims of economically motivated violent crimes, and Mean (and Standard Deviation in parenthesis) for victims of non-violent economically motivated crime.

The explanatory variables to be considered can be classified under three categories. Below, we present a tentative list of variables (and types of variables) that should be considered in each of the three categories. Of course, the specific variables to be used in each study will depend on the type and amount of data available for each city, as well as on the possibility of linking each surveyed individual/household to some particular aggregate – or community – for which socioeconomic data is available.

a) individual demographic characteristics of the victims:
   - gender;
   - age;
   - marital status;
   - racial/ethnic group;
   - household/family characteristics (gender of the head of the household, etc.).

b) individual socioeconomic (and other) characteristics of the victims:
   - family income;
   - family income as a percentage of mean income in the sample;
   - educational achievement;
   - employment status;
   - home ownership status;
   - home characteristics;
   - social capital indicators (church attendance, participation in community organizations, etc.);
   - gun availability in the household;
   - drug and alcohol consumption;
   - value of property taken;
   - victim knew offender;
   - victim reported crime to the police;
- an arrest was made in connection with the crime;

c) characteristics of the communities where the victims live:
- income variables (average level, rates of change and distribution);
- poverty indicators;
- unemployment rate;
- urban development indicators;
- average educational achievement;
- primary and secondary enrollment rates;
- variables representing the strength of the police and justice system (arrest rates, police personnel and expenditures);
- variables representative of the existence of profitable criminal activities (drug trafficking, gambling, etc.);
- average demographic characteristics of the community;
- average family characteristics (percent female head of household, etc.);
- urban density;
- gun availability in the community;
- incidence of drug and alcohol consumption;
- social capital variables at the community level.

Finally, a regression analysis should be performed in order to determine the effect of the above mentioned variables on the probability of being victimized. Probit or Logit techniques should be used. First, victimization should be regressed only on individual demographic characteristics. Second, victimization should be regressed on demographic, socioeconomic and other individual characteristics. Finally victimization should be regressed on individual and community characteristics.

It is likely that victimization surveys contain information on the type of crime suffered by the victim. Having various types of crime opens up the range of potential dependent variables. However, some grouping of crime types is necessary to, first, keep the econometric exercise of a manageable size, and, second, facilitate the interpretation of the estimated coefficients. We suggest the following categories of common crime:

Economically motivated crime and non-economically motivated crime. The first category can be subdivided in crime involving the use or threat of violence (for example, robbery, homicide (sometimes), and house burglary (sometimes)) and crime without violence (for example, theft, and house burglary (sometimes)). We presume that all non-economically motivated crime involves violence (for example, homicide (sometimes), assault, and rape).
Having these categories in mind, our suggestion is to work with two types of discrete dependent-variable models, namely, binomial and multinomial models. In the case of binomial models, we suggest the analysis of the following "events" (where the alternative "event" for each of them is given implicitly):

1. Event: being a victim of any type of crime
2. Event: being a victim of an economically motivated crime (with or without violence)
3. Event: being a victim of a violent crime (whether or not economically motivated)
4. Event: being a victim of a non-economically motivated crime

In the multinomial case, we suggest the analysis of the following models:

1. Events: being a victim of a non-economically motivated crime, being a victim of an economically motivated crime, and being a non-victim.
2. Events: being a victim of a violent crime, being a victim of an economically-motivated crime without violence, being a non-victim.

We want to stress that the groupings and econometric analysis using victimization-survey data depend largely on the characteristics of available surveys. The suggestions given here attempt to homogenize all case studies, but they should be considered with flexibility by each research team. Thus, each team is given freedom to study additional groupings of types of crime and their corresponding "events," in accordance with both the focus and data availability of each case study.

Finally, a technical point. It is important that all regressions with community level characteristics appropriately cluster the data to correct for correlation in the error terms within communities. The error term corresponding to each individual has two components, one corresponding to the individual himself (which is uncorrelated with other error terms) and the other corresponding to the community in which the crime was committed. The second component is, thus, common to all individuals victimized in a given community, which brings about a correlation between their error terms. This correlation must be controlled for in the process of estimation; otherwise, the estimated standard errors would be inconsistent. In addition, if the "community" component of the error term is correlated to any of the community-level variables, then the coefficient estimates will also be inconsistent. The way to control for the presence of the community component in the error term is to cluster observations by community and to allow community-specific intercepts in the estimation.

ii) Social and Economic Determinants of Crime Rates Over Time and Across City Subdivisions

Although the data from victimization surveys can be thought to be more reliable than the data from official health and police sources, the latter have the advantage of covering homicides and other violent crimes for which data is not available from
victimization surveys. Furthermore, official health and police statistics are produced more frequently than victimization surveys. For these reasons, it is important to perform, when possible, an analysis of the social and economic determinants of crime rates over time and across neighborhoods, precincts, census units or other city subdivisions for which data is available.

A regression analysis should be performed using the homicide rate (and/or other crime rates) calculated at the sub-city level as dependent variable. The explanatory variables should be the community level variables constructed for the analysis of the determinants of victimization risks (third group of variables listed above).

The econometric techniques to be used will depend on the characteristics of the data set to be constructed. Ideally, if data is available both across units and over time, panel data techniques should be used. However, whenever possible, even in the context of a cross-sectional analysis, an effort should be made to construct appropriate instrumental variables that correct for the possible endogeneity of some of the explanatory variables – especially those related to the strength of the police and judicial system.

At the city level, a complementary analysis to be performed consists in the calculation of the correlation between time-series of crime rates (particularly homicide rates), and time series of variables representative of the socioeconomic characteristics of the cities (again from the third group of variables listed above).

5. **Summary of the Tables (and Graphs) Related to the Measurement and Explanation of Crime and Victimization Rates**

1) Table describing the available sources of crime data;

2) Tables and/or graphs with the levels and trends of homicide rates using data from public health reports: by demographic group and at the national, city, and sub-city levels;

3) Tables and/or graphs with the levels and trends of homicide and other crimes rates using police data: at the national, city, and sub-city levels, and including correlations between the various crime rates;

4) Tables with mean victimization rates: by type of crime and by demographic group;

5) Table with a description of the variables to be used in the explanation of victimization and crime rates;

6) Table with summary statistics of the variables to be used in the explanation of victimization and crime rates: mean and standard deviation for the total sample, the victims only and the non-victimized individuals only;

7) Tables with the results of regression of victimization rates (for the various types of crimes) on the explanatory variables described in the previous tables;

8) Tables with the results of regressions performed using crime rates at the sub-city level (calculated from police and/or public health data) as dependent variables.
9) Tables and/or graphs with the time-series correlation between crime rates at the city level and the explanatory variables used in the regressions above, also calculated at the city level.

References
