Two Stage Least Squares Using Excel

With an Application to the Gun Debate

Background

During a career of economic research using sophisticated econometric and statistical software, I rarely had to know how to do the analysis manually: the software always did it for me. But, now retired, my access to that software is gone and I confronted the tedious task of doing some research using Microsoft Excel. That research involved estimation of a structural model using Two Stage Least Squares (TSLS). Excel worked well for estimation of the basic parameters of the model, but it provided no "canned" method for correctly estimating the variances of parameter estimates so that statistical inferences could be made.

This "note" shows just how to use Excel—or any spreadsheet software capable of matrix manipulations—to estimate TSLS. It ends with an application to the question, "Are homicides and suicides in the U. S. causally related to the number of guns?" In that portion we show the Excel spreadsheets used to make correct calculations of the variance-covariance matrix of a structural equation's parameters, and we compare the corrected t-statistics with those generated by an OLS estimation of the equation.

Structural Model Estimation by Two-Stage Least Squares

Notation: a "hat" over a vector or matrix indicates "predicted" value from an OLS regression
a "cup" (inverted hat) indicates a residual from an OLS regression

The Model

In the general structural equations model there are M structural equations in which J of the regressors are endogenous variables: variables that are correlated with the error term (a change in *u* with other variables constant changes *y* and that feeds back onto *z*). Thus, causation works both ways in a structural model.

(1) M structural equations, each of the form

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta}_i + \mathbf{Z}\boldsymbol{\varnothing}_i + \boldsymbol{u}_i \qquad i = 1, \ldots, M$$

where there are N observations and

$\mathbf{y}_i$ is an N x 1 vector of observations on the kth independent variable

$\mathbf{X}$ is an N x K matrix of *exogenous* variables

$\boldsymbol{\beta}_i$ is a K x 1 vector of coefficients to be estimated

$\mathbf{Z}$ is an N x J matrix of “included” *endogenous* variable

$\boldsymbol{\emptyset}_i$ is a J x 1 vector of coefficients to be estimated

$\boldsymbol{u}_i$ is a N x 1 matrix of random errors, $\mathbf{u} \sim N(0, \sigma^2 \mathrm{I})$

There are also J "instrumental equations," each describing the relationship of one of the J endogenous regressors to a set of exogenous regressors.

(2)

J instrumental equations, each of the form

$$\mathbf{z_j} = \mathrm{W}\boldsymbol{\pi_j} + \boldsymbol{v_j} \qquad \mathrm{j} = 1, \ldots, \mathrm{J}$$

$\mathbf{W}$ is a N x P matrix of observations on P “ instrumental variables

$\boldsymbol{\pi_j}$ is a P x 1 vector of coefficients to be estimated

$\boldsymbol{v_j}$ is a N x 1 vector of random errors, $\boldsymbol{v} \sim N(0, \sigma_v^2 \mathrm{I})$

Note: the coefficient vectors can be different for each instrumental or structural equation by setting some elements of $\boldsymbol{\beta_k}$ *and* $\boldsymbol{\pi_j}$ *to zero.*

The First-Stage Regression

The first stage in TSLS estimation uses Ordinary Least Squares (OLS) to create "instruments" for each endogenous regressor This requires choosing exogenous variables that are correlated with each of the J endogenous regressors but have no feedback to those endogenous regressors. Suppose you have P exogenous variables.

The exogenous variables are formed into an N x P matrix (W), where P is the number of exogenous variables and N is the number of observations. Then the following steps are applied to estimate each of the J equations explaining each "instrument variable" (z_j).

- Estimate, using OLS, each instrumental equation z_j = W$\pi_j + v_j$ j = 1. 2, . . . , J

  The estimators are the standard OLS estimators. Thus, for each instrumental equation we have

  $\hat{\pi}_j$ = (W'W)$^{-1}$W'z_j— the coefficient estimator

  $\widehat{\Omega}$ = $\hat{s}_v^2$(W'W)$^{-1}$ — the variance-covariance matrix of coefficients

  $\hat{s}_v^2 = \frac{1}{N-K}\Sigma\hat{v}_j^2$ — the sample variance of the estimated error term v_j

- Save the N x 1 vectors of fitted and residual values of z_j (denoted $\hat{z}_j$ and $\check{z}_j$ respectively) and form them into two N x J matrices $\hat{Z}$ and $\check{Z}$.

The Second Stage

- The i[th] structural equation in (1) can be written as:

$$y_i = X\beta_i + \hat{Z}\theta_i + €_i$$

with error term $€_i = u_i + \check{Z}\theta_i$

Note: X is K x N; Z, $\hat{Z}$ and $\check{Z}$ are M x N; θ_i is M x1; and β_i is K x 1

Each structural equation can be rewritten as

$$y_i = \hat{Q}\gamma_i + €_i \quad \text{where } \hat{Q} = \begin{vmatrix} X & \hat{Z} \end{vmatrix} \text{and } \gamma_i = \begin{vmatrix} \beta_i \\ \theta_i \end{vmatrix}$$

Note: $\hat{Q}$ is N x (K + M); γ_i is (K+M) x 1; β_i is a Kx1 vector of coefficients of exogenous regressors; and θ_i is an Mx1 vector of coefficients of endogenous regressors

- Estimate each structural equation using OLS to derive

  Estimated OLS coefficients: $\hat{\gamma} = (\hat{Q}'\hat{Q})^{-1}Q'y_i$

  OLS Standard Error of Estimate: $\hat{s}_€^2 = \frac{1}{N-(M+K)}\Sigma\widehat{€}^2$

  OLS variance-covariance matrix: $\widehat{\Omega} = \hat{s}_€^2(\hat{Q}'\hat{Q})^{-1}$
  where $\hat{s}_€^2 = \frac{1}{N-(M+K)}\Sigma\widehat{€}^2$

So far so good: the estimated OLS coefficients ($\hat{\gamma}$) need no further adjustments—they are asymptotically unbiased, meaning that as the sample size grows the estimates $\hat{\gamma}_i$ approaches the population parameters γ_i .

But the OLS estimate of the variance-covariance matrix is incorrect for statistical inference. Recall that the errors in the second-stage regressions (€) are calculated as

$$€_i = u_i + \check{Z}\theta_i$$

Thus, because the second-stage regression uses the *fitted* values for each endogenous regressor rather than the *actual* values, the effect of the residuals in Ž are compounded into the error term for each structural equation.

<u>Correcting the Variance-Covariance Matrix</u>

<u>The "Correct" Variance-Covariance Matrix</u>

The OLS estimation of the coefficient vector $\hat{\gamma}_i$ using the fitted values of instrument variables is

$$\hat{\gamma} = (\hat{Q}'\hat{Q})^{-1}\hat{Q}'y$$

The difference between the coefficients and their population values is

$$(\hat{\gamma} - \gamma) = (\hat{Q}'\hat{Q})^{-1}Q'(u + \check{Z}\hat{\theta})$$

The correct variance-covariance matrix is

$$\hat{\Omega} = \boldsymbol{E}((\hat{\gamma} - \gamma)(\hat{\gamma} - \gamma)'$$

Thus, $$\hat{\Omega} = \sigma_u^2(\hat{Q}'\hat{Q})^{-1} + (\hat{Q}'\hat{Q})^{-1}\hat{Q}'[\check{Z}(\hat{\theta}\hat{\theta}')\check{Z}']\hat{Q}(\hat{Q}'\hat{Q})^{-1}$$

Note: the first part of $\hat{\Omega}$ is the standard OLS estimator for the variance-covariance matrix. The second part is the adjustment necessary to derive the variance-covariance matrix for the second-stage regression.

There are two steps in making the adjustments:

- Compute $\hat{s}_u^2$, the correct estimator for σ_u^2.
- Compute the correct elements in the $\hat{\Omega}$ matrix

Find the Correct Estimator for σ_u^2, i.e. ($\hat{s}_u^2$)

The first step is to find the correct estimator for σ_u^2. Recall that the new structural equation error term is $\widehat{€}_i = \hat{u}_i + \check{Z}\hat{\theta}_i$, so $\hat{u}_i = \widehat{€}_i - \check{Z}\hat{\theta}_i$ is the implicit error term in the first-stage regression if the actual Z had been used instead of the fitted $\hat{Z}$. Thus, the vector $\hat{u}_i$ can be calculated using the estimates from the second stage OLS error vector ($€_i$), the estimated instrumental coefficients ($\hat{\theta}$), and the matrix of residuals in the instrumental variable equations ($\check{Z}$). In short

- Compute $\hat{u} = \widehat{€} - \check{Z}\hat{\theta}$
- Calculate $\hat{s}_u^2 = \frac{1}{N-(M+K)} \Sigma \hat{u}_i^2$

Compute the Elements in $\widehat{\Omega}$

Above we've seen that

$$\widehat{\Omega} = \hat{s}_u^2(\hat{Q}'\hat{Q})^{-1} + (\hat{Q}'\hat{Q})^{-1}\hat{Q}'[\check{Z}(\hat{\theta}\hat{\theta}'))\check{Z}']\widehat{Q'Q}\ (\hat{Q}'\hat{Q})^{-1}$$

To compute the elements you need $(\hat{Q}'\hat{Q})^{-1}$, $(\hat{Q}'\hat{Q})^{-1}\hat{Q}'$, the first-stage residual matrix $\check{Z}$, and the second stage coefficient estimates $\hat{\theta}$. These can be formed from the available data and estimates.

Though I found it a tedious task, the elements of $\widehat{\Omega}$ can be calculated using Excel (see Appendix)

A General Solution:
Multiple Exogenous *and* Endogenous Regressors

Note that the intercept ("constant term") is designated **1** (a an N x 1 vector of 1's) and is classed as an exogenous regressor. The simplest structural model is a single equation without an intercept and with only one endogenous regressor (i.e., K = 0, M = 1). Here we describe the most general case.

<u>The Data</u>

Let X be a N x K matrix of exogenous regressors
Z be a N x M matrix of endogenous regressors
$\hat{Z}$ be a N x M matrix of fitted values from first-stage regression
$\check{Z}$ be a N x M matrix of residual values from first stage regressions
and
$\hat{Q}$ = [X $\hat{Z}$] be the N x (K+M) matrix of all second-stage regressors

The variance-covariance matrix for the second-stage regression coefficients is

$$\Omega = \sigma_u^2(\hat{Q}'\hat{Q})^{-1} + (\hat{Q}'\hat{Q})^{-1}\hat{Q}'[(\hat{Z}'\check{Z})(\theta\theta')(\check{Z}\hat{Z}')]Q(\hat{Q}'\hat{Q})^{-1}$$

Let $A = (\hat{Q}'\hat{Q})^{-1}_{(K+M)x(K+M)}$ $B = [\hat{Z}'\check{Z}]_{MxM}$ and $\hat{\theta}' = (\hat{\theta}_1 \ \hat{\theta}_2 \ \ldots \ \hat{\theta}_M)_{1xM}$

so we can write $\Omega = \sigma_u^2 A^{-1} + A^{-1}B[\theta\theta']B'A'^{-1}$

where

$$A = \begin{vmatrix} \hat{Q}_1{}'\hat{Q}_1 & \hat{Q}_1{}'\hat{Q}_2 & .. & \hat{Q}_1{}'\hat{Q}_M \\ \hat{Q}_2{}'\hat{Q}_1 & \hat{Q}_2{}'\hat{Q}_2 & .. & \hat{Q}_2{}'\hat{Q}_M \\ . & . & . & . \\ \hat{Q}_M{}'\hat{Q}_1 & \hat{Q}_M{}'\hat{Q}_2 & .. & \hat{Q}_M{}'\hat{Q}_M \end{vmatrix}$$

$$B = \begin{vmatrix} \hat{Z}_1{}'\check{Z}_1 & \hat{Z}_1{}'\check{Z}_2 & .. & \hat{Z}_1{}'\check{Z}_{1M} \\ \hat{Z}_2{}'\check{Z}_1 & \hat{Z}_2{}'\check{Z}_2 & .. & \hat{Z}_2{}'\check{Z}_{1M} \\ . & . & . & . \\ \hat{Z}_M{}'\check{Z}_1 & \hat{Z}_M{}'\check{Z}_2 & .. & \hat{Z}_M{}'\check{Z}_{1M} \end{vmatrix} = \begin{vmatrix} 0 & \hat{Z}_1{}'\check{Z}_2 & \ldots & \hat{Z}_1{}'\check{Z}_{1M} \\ \hat{Z}_2{}'\check{Z}_1 & 0 & \ldots & \hat{Z}_2{}'\check{Z}_{1M} \\ . & . & . & . \\ \hat{Z}_M{}'\check{Z}_1 & \hat{Z}_M{}'\check{Z}_2 & \ldots & 0 \end{vmatrix}$$

$$\hat{\theta}\hat{\theta}' = \begin{vmatrix} \theta_1^2 & \theta_1\theta_2 & .. & \theta_1\theta_M \\ \theta_2\theta_1 & \theta_2^2 & .. & \theta_2 \\ \theta_2\theta_1 & \theta_2^2 & .. & \theta_2\theta_M \\ \theta_M\theta_1 & \theta_M\theta_2 & & \theta_M^2 \end{vmatrix}$$

Finding the variance-covariance Matrix

Recall that the variance-covariance matrix of the estimators $\hat{\theta}$ is

$$\hat{\Omega} = \ = \sigma_u^2 (Q'Q)^{-1} + (Q'Q)^{-1} Q' [\check{Z}(\hat{\theta}\hat{\theta}')\check{Z}'] Q (Q'Q)^{-1}$$

A central matrix in the TSLS estimation is

$\hat{Q}$ = $[\mathbf{X} \quad \hat{\mathbf{Z}}]$ from which

$$\hat{Q}'\hat{Q} = [\mathbf{X} \quad \hat{\mathbf{Z}}]\,[\mathbf{X} \quad \hat{\mathbf{Z}}]' = \begin{vmatrix} \mathbf{X'X} & \mathbf{X'\hat{Z}} \\ \mathbf{\hat{Z}'X} & \mathbf{Z'\hat{Z}} \end{vmatrix}$$

Block Inversion, a property of matrix algebra, says that

$$\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix}^{-1} = \begin{vmatrix} \mathbf{(A - BD^{-1}C)^{-1}} & \mathbf{-(A - BD^{-1}C)^{-1}BD^{-1}} \\ \mathbf{-D^{-1}C(A - BD^{-1}C)^{-1}} & \mathbf{D^{-1}C(A - BD^{-1}C)^{-1}\,BD^{-1}} \end{vmatrix}$$

so $(Q'Q)^{-1}$ can be written as

$$\underline{(Q'Q)^{-1}} = \begin{vmatrix} \mathbf{X'X} & \mathbf{X'\hat{Z}} \\ \mathbf{\hat{Z}'X} & \mathbf{Z'\hat{Z}} \end{vmatrix}^{-1} = \begin{vmatrix} \mathbf{A^*} & \mathbf{B^*} \\ \mathbf{C^*} & \mathbf{D^*} \end{vmatrix}$$

where

$\mathbf{A^* = \quad [\, X'X - (X'\hat{Z})(\hat{Z}'\hat{Z})^{-1}(\hat{Z}'X\,)]^{-1}}$

$\mathbf{B^* = \ - \ [\, X'X - (X'\hat{Z})(\hat{Z}'\hat{Z})^{-1}(\hat{Z}'X\,)]^{-1}}(\mathbf{X'\hat{Z}})(\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{-1}$

$\mathbf{C^* = \ - \{[\, X'X - (X'\hat{Z})(\hat{Z}'\hat{Z})^{-1}(\hat{Z}'X\,)]^{-1}\}}$

$\mathbf{D^* = \ \{}(\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{-1}\mathbf{(\hat{Z}'X)\ [\, X'X - (X'\hat{Z})(\hat{Z}'\hat{Z})^{-1}(\hat{Z}'X\,)]^{-1}}(\,\mathbf{X'\hat{Z}})(\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{-1}\mathbf{\}}$

Thus,

$$(\hat{Q}'\hat{Q})^{-1}\hat{Q}' = \begin{vmatrix} \mathbf{A^*} & \mathbf{B^*} \\ \mathbf{C^*} & \mathbf{D^*} \end{vmatrix} [\mathbf{X} \quad \hat{\mathbf{Z}}]' = \begin{vmatrix} \mathbf{A^*X' + B^*\hat{Z}'} \\ \mathbf{C^*X' \quad D^*\hat{Z}'} \end{vmatrix}$$

An additional important matrix is the N x N matrix $[(\hat{\mathbf{Z}}'\check{Z}\,)(\hat{\theta}\hat{\theta}')(\check{\mathbf{Z}}'\hat{\mathbf{Z}})]$. Both matrices can be formed from the second-stage regression output.

The Simplest Cases

Case 1: A Single Endogenous Regressor, No Intercept

Suppose the structural equation has no intercept and only one endogenous regressor. An equivalent way to describe it is that all variables are defines=d as deviations from the sample means. This is the easiest possible case.

In this case the complicated expression for $\widehat{\Omega}$ reduces to

$$\Omega = \hat{s}_u^2(\hat{z}'\hat{z})^{-1} + \hat{\theta}^2(\hat{z}'\hat{z})^{-1}ç(\hat{z}'\check{z})'(\hat{z}'\hat{z})^{-1}$$

A convenient property of OLS estimation is that the sum of the products of the residuals and the fitted regressors is zero, i.e. $\hat{z}'\check{z} = 0$. Thus, the complicated second term entirely vanishes in this case. In the one-regressor case we have

$$\Omega = \hat{s}_u^2(\hat{z}'\hat{z})^{-1} \text{ and } \mathrm{Var}(\hat{\theta}^2) = \frac{\hat{s}_u^2}{\Sigma\hat{z}^2}$$

Problem

You've estimated the first-stage regression to obtain vector $\hat{z}$ as well as the second-stage regression

$$y = \hat{z}\theta + € \quad \text{with error vector} \quad € = u + \check{z}\theta$$

What is the variance of the estimated coefficient $(\hat{\theta})$?

Answer

The estimator for the variance of $\hat{\theta}$θ* is

$$\mathrm{Var}\,(\hat{\theta}) = \frac{\hat{s}_u^2}{R^2\Sigma z^2}$$

where

R^2 is the R-Squared for the IV equation
s_u^{*2} is the adjusted error variance reported in the IV equation
$\sum z^2$ is the sum-of-squares of the endogenous variable

The t-statistic is $t = \frac{\theta^*}{\sqrt{var(\theta^2)}}$.

Note that because $0 < R^2 < 1$, Var $(\hat{\theta})$ *must* be greater than the variance $\hat{\theta}$ reported by the second-stage OLS output.

Case 2: One Exogenous and Two Endogenous Regressors

This is the format of the model used in the application that follows. The exogenous regressor in this example is a constant serving as the intercept.

Define the following matrices and vectors:

$\hat{Q}_{Nx3} = \begin{vmatrix} \mathbf{1} & \hat{\mathbf{z}}_1 & \hat{\mathbf{z}}_2 \end{vmatrix}$ $\hat{Q}'_{3xN} = \begin{vmatrix} \mathbf{1}' \\ \hat{\mathbf{z}}_1' \\ \hat{\mathbf{z}}_2' \end{vmatrix}$ $\check{Z}_{Nx2} = [\, \check{\mathbf{z}}_1 \;\; \check{\mathbf{z}}_2 \,]$ $\hat{\theta} = \begin{vmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{vmatrix}$ γ_0 = intercept

where $\mathbf{1}$ is an Nx1 vector of 1's for the constant term, $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2$ are Nx1 vectors of fitted values for the two instruments, $\check{\mathbf{z}}_1, \check{\mathbf{z}}_2$ are the associated Nx2 residual vectors, γ_0 is the estimated intercept coefficient, and $\hat{\theta}$ is a 2x1 vector of coefficients on the instrumental variables.

Note that only the endogenous regressor coefficients are

$$\text{Then} \quad \Omega = \sigma_u^2(\hat{Q}'\hat{Q})^{-1} + (\hat{Q}'\hat{Q})^{-1}Q'[\check{Z}(\theta\theta')\check{Z}']\hat{Q}(\hat{Q}'\hat{Q})^{-1}$$

which can be written as $\Omega = \sigma_u^2(\hat{Q}'\hat{Q})^{-1} + \boldsymbol{ßß'}$ where $\boldsymbol{ß} = (\hat{Q}'\hat{Q})^{-1}\check{Z}\theta$

where

$$\hat{Q} = \begin{vmatrix} 1 & \hat{z}_{11} & \hat{z}_{21} \\ 1 & \hat{z}_{12} & \hat{z}_{22} \\ .. & .. & .. \\ 1 & \hat{z}_{1N} & \hat{z}_{2N} \end{vmatrix}$$

$$\hat{Q}'\hat{Q} = \begin{vmatrix} 1'1 & 1'\hat{z}_1 & 1'\hat{z}_2 \\ \hat{z}_1'1 & \hat{z}_1'\hat{z}_1 & \hat{z}_1'\hat{z}_2 \\ \hat{z}_2'1 & \hat{z}_2'\hat{z}_1 & \hat{z}_2'\hat{z}_2 \end{vmatrix}$$

$$\check{Z}\theta = [\, \check{\mathbf{z}}_1 \;\; \check{\mathbf{z}}_2 \,]\hat{\theta} = \begin{vmatrix} \check{z}_{11}\hat{\theta}_1 + \check{z}_{12}\hat{\theta}_2 \\ \check{z}_{12}\hat{\theta}_1 + \check{z}_{22}\hat{\theta}_2 \\ .... \quad\quad ... \\ \check{z}_{1N}\hat{\theta}_1 + \check{z}_{2N}\hat{\theta}_2 \end{vmatrix}$$

An Application: Guns, Homicides and Suicides

This application is taken from a paper available at *www.fortunearchive.com* (scroll to the bottom of the index page and select "Guns in America.")

International opinion—and much American opinion—is clear: there is an obvious causal connection between the number of firearms and the number of homicides, so obvious that obtuse gunowners can't see it. This has been demonstrated by data across countries—countries with more guns per capita have more homicides by gun; the U.S. is a standout on both per capita gun numbers and percapita homicides (particularly if you exclude very violent nations from the data). It has also been demonstrated by data across states—states with more guns per capita appear to have more homicides-by-gun. This is what explains why America has a high murder rate—we Americans just have too many guns!

There are a variety of flaws in this logic. First,

- C*orrelation does not prove causation; the fact that more guns appear to be associated with more homicides is no proof that more guns cause more homicides.*

Second, homicides by gun have a clear socioeconomic and ethnic flavor:

- *Roughly 80 percent of homicides are done by non-whites to non-whites.*

This raises the question of whether violence among the poor and more crime-prone population is a major reason for homicides, not guns. It also raises the question of whether illegal guns—the most common guns in non-white areas—are the source of the association between guns and homicides.
In addition, there are a couple of factoids that raise questions about the guns-homicides association.
Here are two:

- *Three percent of adult gun owners hold fifty percent of America's guns.*

- *Between 1994 and 2015 the population adjusted rates of both violent crimes and homicides has declined steadily while the number of guns increased from 192 million in 1994 to over 265 million in 2015.*

The first factoid suggests that if more guns cause more homicides, there should be a plethora of murders by three percent of gun owners—the "supper-gunners." But there is no evidence that those with more guns murder more people. Furthermore, if these guns are in safe hands then only half of the gun stock is "in play" for homicidal purposes: the unsafe American gun supply is only half of the recorded number. This would take America out of the stratosphere of gun ownership.

The second factoid simply points out that in America there is no evidence that over the past 25 years more guns means more murders.

I set out to look into the guns-homicide connection using statistical analysis of data on relevant variables in the 51 states (including DC) in or about 2012. Some of the variables used were exogenous (per capita personal income in the state, median age, male-to-female ratio, degree of urbanization, non-white percentage of population). Two variables were treated as endogenous—the per capita number of guns owned in the state, and the per capita number of guns reported lost or stolen.

Table 1 below shows the results of regressing state homicide rate and suicide rate (per 100,000 population) the gun variables. This is done by OLS, a method that would be appropriate if all regressors were exogenous.

<u>OLS Estimation</u>

Table 1
Ordinary Least Squares Regressions

	Dependent Variables			
Independent Variable	**HOMICIDES per 100K**		**SUICIDES per 100K**	
	Coefficient	***t-Statistic***	***Coefficient***	***t-Statistic***
Constant	**- 75.90**	**- 2.13**	- 181.47	- 1.91
Pers. Income (per capita)	- 0.00015	- 0.99	-0.00062	- 1.61
Income Inequality (Gini)	+ 38.26	+ 1.69	+ 67.18	+ 1.11
Median Age (Years)	+ 0.37	+ 1.14	+ 1.11	+ 1.31
Gender (Male/Female)	**+0.61**	**+ 2.32**	+ 1.33	+ 1.95
Urbanization (percent)	- 5.97	- 1.27	+ 12.32	+0 .98
Race (% Black)	+ 3.66	+ 0.47	+ 3.85	+ 0.19
Stolen Guns (per 100 pop)	**+ .0163**	**+ 3.15**	+ 0.0133	+ 0.97
Guns Owned (per 100 pop)	**- 0.1323**	**- 2.06**	-0.0934	- 0.56
Adjusted R^2	*0.65*		*0.05*	

Bold face text shows statistically significant variables (5%)

The first thing to note is that *nothing* explains suicides. They follow an entirely different pattern—if there is a pattern—than homicides. However, homicides do have some statistically significant regressors. In particular, both stolen guns and guns owned play a statistically significant role in explaining homicides. As expected, stolen guns contribute directly to homicides. But guns-owned are inverse factors—the more guns owned in a state, the f*ewer* the homicides.

Clearly, this does not support the view that the volume of guns is the cause of the high homicides (and suicides) experienced in America. But perhaps there is endogeneity biasing the results, as when homicides induce purchase of fewer or more guns--more guns as people arm for self-defense, or fewer guns as people become more fearful of gun deaths. So let's try TSLS estimation to mitigate the effects of endogeneity.

TSLS Estimation

To do this we assume that stolen guns and guns owned are endogenous regressors and we regress each on all of the other exogenous variables. The results are reported in Table 2.

Table 2
Ordinary Least Squares Regressions

Independent Variable	**Dependent Variables** — **STOLEN GUNS per 100** ***Coefficient***	***t-Statistic***	**GUNS OWNED per 100K** ***Coefficient***	***t-Statistic***
Constant	+ 2.14	+ 1.47	- 104.23	- 0.89
Pers. Income (per capita)	**+0.00002**	**+ 3.78**	- 0.0005	- 1.58
Income Inequality (Gini)	+ 0.72	+ 0.69	+ 96.04	+ 1.16
Median Age (Years)	**- 0.05**	**- 4.47**	- 0.95	- 1.11
Gender (Male/Female)	-0.01	- 1.19	**+ 1.87**	**+ 2.53**
Urbanization (percent)	+ 2.00	+ 0.99	**- 45.69**	**- 3.35**
Race (% Black)	**+0.92**	**+ 3.13**	+ 8.80	+ 0.37
Adjusted R^2	*0.81*		*0.85*	

Bold face text shows statistically significant variables (5%)

Stolen guns are more common in higher income states , in younger states, and in states with higher proportions of blacks. Guns owned are driven by gender (more males buy guns than females) and by urbanization (guns are more common in *less* urbanized states).

Finally, Table 3 tells us the link between method of death and guns, purged of the endogeneity that might taint Table 1. The bottom line is unchanged—stolen guns matter, the number of guns doesn't—though the coefficient is now positive. Only to the extent that a larger stock of guns allows more stolen guns is there a link between guns and homicides. Suicides, on the other hand, are inexplicable using our data.

The good news for those who claim that guns and homicides are directly related is that the coefficient on guns-owned is now (slightly) positive, a sharp contrast with the OLS results in Table 1. The bad news is that it is not statistically significant.

Table 3
TSLS Second Stage Regressions

Independent Variable	**HOMICIDES per 100K** *Coefficient*	**HOMICIDES per 100K** *t-Statistic*	**SUICIDES per 100K** *Coefficient*	**SUICIDES per 100K** *t-Statistic*
Constant	+ 0.62	+ 0.52	**+ 7.9393**	**+ 2.32**
Stolen Guns (fitted)	**+ 10.09**	**+ 4.66**	- 2.0677	- 0.20
Guns Owned (fitted)	+ 0.04	+ 1.70	+ 0.9453	+ 0.41

t-statistics are corrected for errors introduced by TSLS.
Bold face text shows statistically significant variables (5%)

Estimates of Standard Errors and t-Statistics: Direct OLS vs. TSLS

In Table 4 we compare the standard errors and t-statistics generated directly by OLS estimation of the equations in Table 3 with those resulting from correct adjustment of TSLS estimation.

Table 4
TSLS Second Stage Regressions

Dependent Variable: HOMICIDES per 100K

Independent Variable	**Coefficient**	**OLS** *Std Error*	**OLS** *t-Statistic*	**TSLS** *Std Error*	**TSLS** *t-Statistic*
Constant	0.62	+ 0.65	+ 0.94	0.73	+ 0.52
Stolen Guns (fitted)	10.09	+ 1.93	+ 5.24	2.17	+ 4.66
Guns Owned (fitted)	0.04	+ 0.02	+ 1.91	0.02+	+ 1.70

As expected, the standard errors of the estimated coefficients are higher with TSLS than with OLS, and the t-statistics are correspondingly lower.

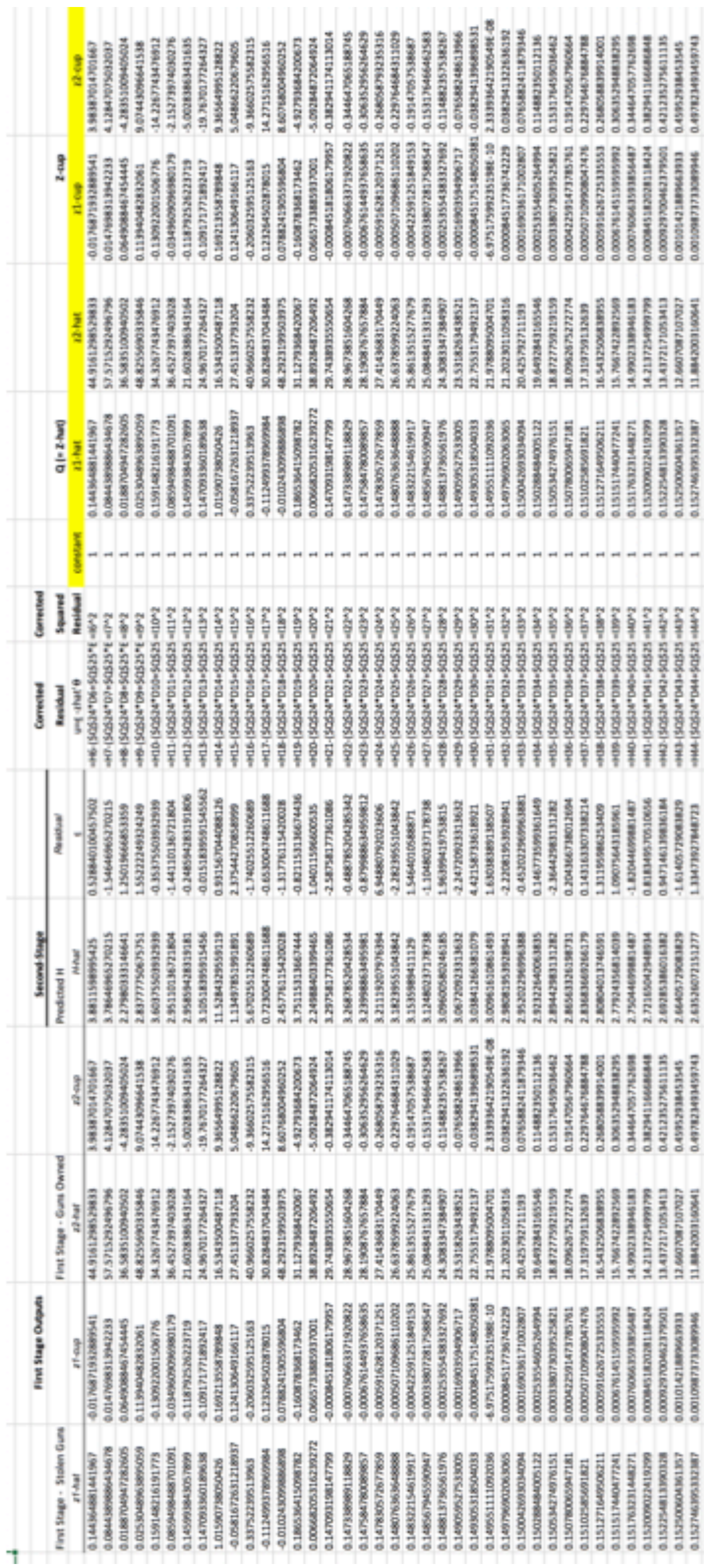
Appendix: Data Set

State	2011 Gun Deaths (per 100K) GUN DEATHS	2010 Gun Homicides (per 100K) HOMICIDES	2010 Gun Suicides (per 100K) SUICIDES	2010 Population (in 100Ks) POPULATION	2011 Median Income (per HH) MEDIAN Y	2012 Pers. Inc. per Capita PERS. INC.	2010 Income Ineq (Gini) INEQUALITY	2013 Gender Males per F GENDER	2013 Median Age (Years) MEDIAN AGE	2015 Race-Black (%) BLACK	2010 Urbanization (%) URBAN	Stolen/Lost Guns (per 100) STOLEN GUNS	2013 Guns Owned (per 100) GUNS OWNED
Alabama	17.6	4.41	13.19	48.027	$41,415	$23,606	0.4847	94.33	37.90	26.40%	59.00%	0.126677688	48.9
Alaska	19.8	2.24	**17.56**	7.227	$67,825	$33,062	0.4081	108.52	33.80	3.40%	66.00%	0.09920882	61.7
Arizona	14.1	3.53	10.57	64.825	$46,709	$25,715	0.4713	98.74	35.90	4.20%	89.80%	0.083779341	32.3
Arkansas	16.8	4.39	12.41	29.380	$38,758	$22,883	0.4719	96.45	37.40	15.50%	56.20%	0.139245379	57.9
California	7.7	3.25	4.45	376.919	$57,287	$30,441	0.4899	98.83	35.20	5.90%	95.20%	0.028226215	20.1
Colorado	11.5	1.51	9.99	51.168	$55,387	$32,357	0.4586	100.48	36.10	4.00%	86.20%	0.050988939	34.3
Connecticut	4.4	2.71	1.69	35.807	$65,753	$39,373	0.4945	94.83	40.00	10.30%	88.00%	0.027201317	16.6
Delaware	10.3	3.09	7.21	9.071	$58,814	$30,488	0.4522	93.94	38.80	21.60%	83.30%	0.037921588	5.2
D.C	18.44	12.46	5.98	6.180	$63,124	$45,877	0.5420	89.52	33.80	48.90%	100.00%	1.185120939	25.9
Florida	11.9	3.51	8.39	190.575	$44,299	$26,582	0.4852	95.60	40.70	16.10%	91.20%	0.065963386	32.5
Georgia	12.6	3.93	8.67	98.152	$46,007	$25,615	0.4813	95.38	35.30	30.90%	75.10%	0.1314898	31.6
Hawaii	2.6	0.07	2.53	13.748	$61,821	$29,736	0.4420	100.32	38.60	2.00%	91.90%	0.010765124	45.1
Idaho	14.1	1.14	12.96	15.850	$43,341	$23,938	0.4503	100.39	34.60	0.60%	70.60%	0.068581091	56.9
Illinois	8.6	2.93	5.67	128.693	$53,234	$30,417	0.4810	96.24	36.60	14.30%	88.50%	0.025658047	26.2
Indiana	13	3.29	9.71	65.169	$46,438	$25,140	0.4527	96.83	37.00	9.20%	72.40%	0.073255442	33.8
Iowa	8	0.71	7.29	30.623	$49,427	$29,507	0.4729	95.71	37.13	14.53%	87.67%	0.146248016	29.4
Kansas	11.4	2.78	8.62	28.712	$48,964	$29,485	0.4731	95.50	37.17	14.57%	88.52%	0.146578326	28.6
Kentucky	13.7	2.36	11.34	43.694	$41,141	$29,463	0.4733	95.29	37.21	14.61%	89.37%	0.146908635	27.9
Louisiana	19.3	10.16	9.14	45.748	$41,734	$29,441	0.4735	95.09	37.26	14.64%	90.22%	0.147238945	27.1
Maine	10.9	0.9	10.00	13.282	$46,033	$29,419	0.4738	94.88	37.30	14.68%	91.06%	0.147569254	26.4
Maryland	9.7	4.7	5.00	58.283	$70,004	$29,397	0.4740	94.67	37.34	14.72%	91.91%	0.147899563	25.7
Massachusetts	3.1	2.02	1.08	65.875	$62,859	$29,376	0.4742	94.47	37.39	14.76%	92.76%	0.148229873	24.9
Michigan	12	5.06	6.94	98.762	$45,981	$29,354	0.4745	94.26	37.43	14.80%	93.61%	0.148560182	24.2
Minnesota	7.6	0.82	6.78	53.449	$56,954	$29,332	0.4747	94.05	37.47	14.84%	94.45%	0.148890492	23.5
Mississippi	17.8	7.46	10.34	29.785	$36,919	$29,310	0.4749	93.85	37.51	14.88%	95.30%	0.149220801	22.7
Missouri	14.4	4.64	9.76	60.107	$45,247	$29,288	0.4751	93.64	37.56	14.91%	96.15%	0.14955111	22.0
Montana	16.7	0.76	15.94	9.982	$44,222	$29,266	0.4754	93.43	37.60	14.95%	97.00%	0.14988142	21.2
Nebraska	9	2.5	6.50	18.426	$50,296	$29,244	0.4756	93.23	37.64	14.99%	97.84%	0.150211729	20.5
Nevada	13.8	3.07	10.73	27.233	$48,927	$29,222	0.4758	93.02	37.69	15.03%	98.69%	0.150542039	19.8

New Hampshire	6.4	0.53	5.87	13.182	$62,647	$29,200	0.4761	92.81	37.73
New Jersey	5.7	3.07	2.63	88.212	$67,458	$29,179	0.4763	92.61	37.77
New Mexico	15.5	2.98	12.52	20.822	$41,963	$29,157	0.4765	92.40	37.82
New York	4.2	4.12	0.08	194.652	$55,246	$29,135	0.4767	92.19	37.86
North Carolina	12.1	3.87	8.23	96.564	$43,916	$29,113	0.4770	91.99	37.90
North Dakota	11.8	0.93	10.87	6.839	$51,704	$29,091	0.4772	91.78	37.95
Ohio	11	3.54	7.46	115.450	$45,749	$29,069	0.4774	91.57	37.99
Oklahoma	16.5	3.64	12.86	37.915	$43,225	$29,047	0.4776	91.37	38.03
Oregon	11	1.05	9.95	38.719	$46,816	$29,025	0.4779	91.16	38.08
Pennsylvania	11.2	3.97	7.23	127.429	$50,228	$29,003	0.4781	90.95	38.12
Rhode Island	5.3	0.57	4.73	10.513	$53,636	$28,982	0.4783	90.75	38.16
South Carolina	15.2	5.41	9.79	46.792	$42,367	$28,960	0.4786	90.54	38.21
South Dakota	10	0.68	9.32	8.241	$48,321	$28,938	0.4788	90.33	38.25
Tennessee	15.4	3.92	11.48	64.034	$41,693	$28,916	0.4790	90.13	38.29
Texas	10.6	2.91	7.69	256.747	$49,392	$28,894	0.4792	89.92	38.34
Utah	12.6	0.97	11.63	28.172	$55,869	$28,872	0.4795	89.71	38.38
Vermont	9.2	0.75	8.45	6.264	$52,776	$28,850	0.4797	89.51	38.42
Virginia	10.2	2.58	7.62	80.966	$61,882	$28,828	0.4799	89.30	38.47
Washington	8.7	1.25	7.45	68.300	$56,835	$28,806	0.4801	89.09	38.51
West Virginia	14.3	2.87	11.43	18.554	$38,482	$28,785	0.4804	88.89	38.55
Wisconsin	9.7	1.47	8.23	57.118	$50,395	$28,763	0.4806	88.68	38.60
Wyoming	16	2.01	13.99	5.636	$56,322	$28,741	0.4808	88.47	38.64

Appendix: Excel Spreadsheets

RGRESSION DATA



CALCULATION OF STANDARD ERRORS AND t-STATISTICS

Q'	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)
	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)
	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)	=TRANSPOSE(X6:M56)
Q'Q = Z'Z	=MMULT((R6:BP8),(X6:M56))	=MMULT((R6:BP8),(X6:M56))	=MMULT((R6:BP8),(X6:M56))						
	=MMULT((R6:BP8),(X6:M56))	=MMULT((R6:BP8),(X6:M56))	=MMULT((R6:BP8),(X6:M56))						
	=MMULT((R6:BP8),(X6:M56))	=MMULT((R6:BP8),(X6:M56))	=MMULT((R6:BP8),(X6:M56))						
(Q'Q)$^{-1}$	=MINVERSE(R10:T12)	=MINVERSE(R10:T12)	=MINVERSE(R10:T12)						
	=MINVERSE(R10:T12)	=MINVERSE(R10:T12)	=MINVERSE(R10:T12)						
	=MINVERSE(R10:T12)	=MINVERSE(R10:T12)	=MINVERSE(R10:T12)						
(Q'Q)$^{-1}$Q'	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))
	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))
	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))	=MMULT((R14:T16),(R6:BP8))
θ		(Q'Q)$^{-1}$Q'(Z-cq*θ)					Corrected s^{2}		
0.616	=MMULT((R18:BP20),(Q28:Q78))	=MMULT((R18:BP20),(Q28:Q78))	=MMULT((R18:BP20),(Q28:Q7				=SUM(J6:J56)/48		
10.0864	=MMULT((R18:BP20),(Q28:Q78))	=MMULT((R18:BP20),(Q28:Q78))	=MMULT((R18:BP20),(Q28:Q7			Part 1			
0.042	=MMULT((R18:BP20),(Q28:Q78))	=MMULT((R18:BP20),(Q28:Q78))	=MMULT((R18:BP20),(Q28:Q7			s^{2} (Q'Q)-1			
					= X23*R14:T16	= X23*R14:T16	= X23*R14:T16		=MMULT((R23:T25),TRANSPOSE(R23:T25))
Z-cq*θ	((Q'Q)$^{-1}$Q'(Z-cq(θ))*((Q'Q)-1Q')'				= X23*R14:T16	= X23*R14:T16	= X23*R14:T16	PLUS	=MMULT((R23:T25),TRANSPOSE(R23:T25))
=N6*Q24+O6*Q25	=MMULT((R23:T25),TRANSPOSE(R23:T	=MMULT((R23:T25),TRANSPOSE(R23:T	=MMULT((R23:T25),TRANSPOSE(R23:		= X23*R14:T16	= X23*R14:T16	= X23*R14:T16		=MMULT((R23:T25),TRANSPOSE(R23:T25))
=N7*Q24+O7*Q25	=MMULT((R23:T25),TRANSPOSE(R23:T	=MMULT((R23:T25),TRANSPOSE(R23:T	=MMULT((R23:T25),TRANSPOSE(R23:						
=N8*Q24+O8*Q25	=MMULT((R23:T25),TRANSPOSE(R23:T	=MMULT((R23:T25),TRANSPOSE(R23:T	=MMULT((R23:T25),TRANSPOSE(R23:					VARIANCES	
=N9*Q24+O9*Q25							=V26+Z26		
=N10*Q24+O10*Q25								=W27+AA27	
=N11*Q24+O11*Q25									=X28+AB28
=N12*Q24+O12*Q25							STANDARD ERRORS		
=N13*Q24+O13*Q25							=SQRT(X31)		
=N14*Q24+O14*Q25								=SQRT(Y32)	
=N15*Q24+O15*Q25									=SQRT(Z33)
=N16*Q24+O16*Q25									
=N17*Q24+O17*Q25								T-STATS	
=N18*Q24+O18*Q25							= Q23/X35		
=N19*Q24+O19*Q25								= Q23/Z37	
=N20*Q24+O20*Q25									= Q24/Y41
=N21*Q24+O21*Q25									
=N22*Q24+O22*Q25									