Development Microeconometrics Final

Question 1

I ran two tests to check for measurement errors of the suicide variables (total suicides e95, and suicides by firearm e955). First, I checked to make sure we didn't have any missing data and that the number of suicides by firearm was less than or equal to the total number of suicides. The latter was not a problem. The former, missing values, was only a problem for 1 instance for Sonoma County. As there were only 2 total suicides reported it is a small data point which should not affect the overall analysis. This entry is excluded from the regression results (using indicator code for inclusion).

Second, I checked to make sure that the total number of suicides was reasonable in the context of the county population (i.e. that suicides made up less than one half of one percent of the population).

As for the rest of the data, there are some records with missing values in the number of homicides and the UCR variable categories. The FBI UCR data is not included for the year 1993. It is not clear if the authors had this data when their study was done. I ran the regressions twice, both without this data and then with interpolated values for this data (the latter is used for the do file and log and the results in this pdf). Normally I would not include interpolated data for missing values, but as the paper already uses significant interpolation for census data (on a 10 year basis), and the task was to best match the paper's results, using the interpolated data (on a 1 year basis) seemed the best course to follow. Excluding the 1993 data doesn't significantly change any of the conclusions.

Also, we only have 196 of the largest counties versus the 200 listed in the paper. The paper indicates that the 5 counties in New York City are combined into one.

Some counties have a negative rural population in some years. I reset these to zero. All 22 negative values also appear in years outside of the 1980 to 1999 time period analyzed in the regression analysis, but would present issues in the full 1980-2004 dataset. Also, I checked for regression variable values of zero, as ln(0) would cause problems or exclusions. The regression equation uses urban, and not rural, so there aren't any problems with this field. None of the other fields had zero value.

Finally, I evaluated population by county to make sure there aren't any drastic levels of variation or swings in the range.

Question 2

I tabulate annual changes in FSS by county, and tabulate statistics including maximum annual change. 177 counties had a maximum annual change in FSS of over 10 percentage points, indicating that very large annual swings can occur in this proxy variable. The worst case is Rockland County in New York between 1993 and 1995 when the FSS fell from 69% to 14% and rose back to 67%.

I also looked at how often FSS changes by more than 10% in a year for each county. For the 1980-1999 time period, there are 534 occurrences of a 10% change, and many counties have multiple 10% swings. The paper states the goal of the FSS variable is to "estimate variation over time" in gun ownership rates "which require consistent estimates of gun prevalence over time, preferably at a sub national level."

High annual variability in the FSS on the county level would argue against FSS being a consistent enough estimator for their purposes.

Question 3

Table 1.1: Descriptive Statistics for County Data

	CL (2006)	Unweighted	Weighted
Full period (1980-1999)			
FSS in selected years	49.9	53.0	50.0
Homicide rate	11.0	9.1	11.3
Gun homicide rate	7.3	5.9	7.5
%Urban	92.6	89.7	93.7
%Black	14.0	12.7	14.3
%Female household head	18.0	16.7	17.2
# Suicides	195.8	83.4	199.0
FSS in selected years			
1980	48.0	52.5	48.0
1990	52.8	55.5	52.6
1999	48.0	50.9	48.2

Generally, the weighted calculation matches up very well to the statistics quoted in the paper. The homicide rate is slightly higher at 11.3 vs 11.0 per 100,000 people, as is the gun homicide rate at 7.5 vs 7.3 per 100,000. It is unclear what sample size and records are included in the paper. This table was calculated before the determination of the records to include in the regressions that follow but also reran using the record identifier codes (include, include04) without any significant changes.

The unweighted data, if used in the subsequent regression analysis, would probably lead to insignificant results for the proxy variable, as the mean proxy for gun ownership is higher, but both the homicide and the gun homicide rates are lower (later reviewed in question #10). So the unweighted results are likely to cause differences while the weighted results should be qualitatively if not quantitatively similar.

The rationale for weighting probably should be considered in two separate ways. For a descriptive statistics table, weighting seems irrelevant for determining averages, and gives incorrect values for FSS and the percentage of female headed household which have different measures than population. In terms of regression, the weighting rationale does seem reasonable, as the proxy relies on a very rare societal characteristic or choice that occurs annually for significantly less than one half of one percent of the population in order to reach a conclusion about another societal characteristic, gun ownership, which on average, makes up 50 percent of the population. In counties with lower populations and lower number of suicides, the number of suicides per year is likely to have a much larger variance, and these data points with higher variance get a smaller weight as they have less explanatory power. This version of the table, showing both unweighted and weighted, seems like a better compromise than the table in the paper.

Question 4

Presented below are condensed versions of table 2, where all variations include year and county fixed effects. The paper is compared to regressions run using the xtreg command which has two implications. First, xtreg requires a constant weighting factor, so an annually varying population cannot be used as the weight. Given that the data set of the 200 largest counties was chosen based on 1990 populations, I use the 1990 population as the weight (and I also tested other choices without significant impact, see log file). Second, as of 2008, xtreg fe overrides any robust error command with the clustered command due to a paper in Econometrica detailing how robust errors are not consistent in certain fixed effects model (see technical note). I also ran areg, which allows for robust errors, and also for a varying population weight. A comparison of those results with xtreg are in the log file and also are included in the appendix of this pdf. Areg allows for the comparison of robust and clustered errors. Clustered errors assume independence between each cluster, but correlation within the cluster. As we have aggregate data within each cluster, clustering of errors should be used rather than robust errors. As correlation is now present between observations of each cluster, the standard error increases as compared to that of robust standard errors and can be seen in the table in the appendix between columns [b] and [c] (using areg).

	Paper [1]	Replication [1a]	Replication [1d]	Paper [2]	Replication [2a]	Replication [2d]
Ln FSS (t-1)	0.100 (0.044)**	0.0735 (0.0436)*	0.109 (0.040)***	0.107 (0.037)***	0.087 (0.040)**	0.124 (0.038)***
Ln Rob (t)				0.139 (0.043)***	0.033 (0.029)	0.041 (0.037)
Ln Burg (t)				0.258 (0.068)***	0.161 (0.073)**	0.220 (0.099)**
Ln % black (t)						
Ln %Urb (t)						
Ln % same h						
Ln %fem hh						
R-squared	0.915	0.241	0.328	0.921	0.265	0.366
(rho)		0.905	0.890		0.890	0.867

Notes: Columns [1], [2], [1a], 2[a] use data from 1980-1999 and contain 3,822 entries. Columns [1d], [2d] include data through 2004 and contain 4,783 entries. Year and County fixed effects are included in all columns.

The 1980-1999 sample size replication in columns [1a] and [2a] are qualitatively similar for the proxy variable, but quantitatively less significant in both effect and statistical significance. The R^2 measure quoted in the paper corresponds to the rho measurement from the replication, measuring the amount of variance due to differences across panels and includes all the time and area dummy variable controls. I was unable to match the robbery coefficients or statistical significance for any of the 4 model variations. It may be that the 1993 data was significantly different than what I interpolated. Given that the 1993 data is still unavailable, it may be that the prior data used in the paper was incorrect and hence was pulled. The 1980-2004 sample does a better job at matching the effect and statistical significance for the proxy variable, as well as for burglaries. The appendix compares the xtreg results to areg.

	Paper [3]	Replication [3a]	Replication [3d]	Paper [4]	Replication [4a]	Replication [4d]
Ln FSS (t-1)	0.085	0.054	0.085	0.086	0.067	0.101
	(0.044)*	(0.040)	(0.041)**	(0.038)**	(0.037)*	(0.038)***
Ln Rob (t)				0.149	0.042	0.041
				(0.042)***	(0.028)	(0.032)
Ln Burg (t)				0.226	0.139	0.200
				(0.072)***	(0.069)**	(0.085)**
Ln % black (t)	0.233	0.157	0.172	0.278	0.151	0.142
	(0.166)	(0.125)	(0.113)	(0.164)*	(0.125)	(0.110)
Ln %Urb (t)	-0.389	-0.441	-0.294	-0.537	-0.588	-0.434
	(0.161)**	(0.468)	(0.400)	(0.157)***	(0.433)	(0.355)
Ln % same h	-10.209	-0.955	-0.679	-0.690	-0.859	-0.594
	(0.430)***	(0.382)**	(0.392)*	(0.419)	(0.378)**	(0.353)*
Ln %fem hh	0.790	0.628	0.976	-0.303	0.613	0.902
	(0.460)*	(0.372)*	(0.418)**	(0.413)	(0.367)*	(0.366)**
R-squared	0.918	0.263	0.357	0.923	0.284	0.389
(rho)		0.869	0.845		0.851	0.823

Notes: Columns [3], [4], [3a], 4[a] use data from 1980-1999 and contain 3,822 entries. Columns [3d], [4d] include data through 2004 and contain 4,783 entries. Year and County fixed effects are included in all columns.

The 1980-99 sample in columns [3a] and [4a] has a lower effect and statistical significance for the proxy variable than the paper, with no statistical significance for [3a]. The 1980-2004 sample in columns [3d] and [4d] does a much better job at matching the quantitative effect of the paper and increases the statistical significance.

Again, the coefficient and significance of robberies could not be matched, and is qualitatively very different, but of less concern for the main conclusions of the paper. The coefficient for burglaries is somewhat similar, but less statistically significant.

The coefficient for same house of -10.2 in [3] is obviously a paper typo. Most of the demographic controls are qualitatively similar and of same sign, but most of the statistical significance results do not match up. The percent of female head of households control is the only one with a different sign for the effect in [4], but is insignificant for the paper anyhow.

Overall, the most important component, the FSS proxy, is matched up fairly well, except in [3a]. The use of clustered errors under large sample size requires the assumption of stationarity and weak dependence which are not tested here or mentioned in the paper. The replication uses an unbalanced sample, and the do and log files have more comments on different tests and assumptions used to try to replicate the paper's results including: not using 1993 interpolated data, not lagging FSS, changing the weighting, using areg instead of xtreg and changes to what data is included. Overall, the results are fairly similar qualitatively, other than in [3a] as mentioned. The question and comments regarding robust versus clustered standard errors can be seen in the appendix (increasing standard errors between columns [c] and [b] for each run).

Question 5

	Paper [4]	Replication [4a]	Replication [5]	Replication [5b]
Ln FSS (t-1)	0.086	0.067	0.047	0.079
	(0.038)**	(0.037)*	(0.038)	(0.038)**
Ln Rob (t)	0.149	0.042	0.042	0.043
	(0.042)***	(0.028)	(0.031)	(0.036)
Ln Burg (t)	0.226	0.139	0.140	0.200
	(0.072)***	(0.069)**	(0.074)*	(0.092)**
Ln % black (t)	0.278	0.151	0.172	0.170
	(0.164)*	(0.125)	(0.120)	(0.107)
Ln %Urb (t)	-0.537	-0.588	-0.164	-0.109
	(0.157)***	(0.433)	(0.449)	(0.370)
Ln % same h	-0.690	-0.859	-0.667	-0.386
	(0.419)	(0.378)**	(0.369)*	(0.352)
Ln %fem hh	-0.303	0.613	0.444	0.714
	(0.413)	(0.367)*	(0.384)	(0.382)*
Ln pop			-0.374	-0.374
			(0.165)**	(0.166)**
R-squared	0.923	0.284	0.291	0.395
(rho)		0.851	0.880	0.858
Observations	3,822	3,822	3,822	4,783
Time Frame	1980-1999	1980-1999	1980-1999	1980-2004

Here, the model in equation 1.1 is revised by adding the natural log of population as a regressor on the right hand side, without adjusting the quotient of the dependent variable on the left hand side. The 1980-1999 data [5] with this revised model is not robust to adding the natural log of population as a regressor as the proxy variable is no longer statistically significant and the coefficient has decreased substantially. Burglaries become less statistically significant, although the coefficient remains unchanged. The regressor for those remaining in the same house also declines in statistical significance to the 10% level. The percent of urban population now becomes significant at the 10% level. Most importantly, the new population regressor is statistically significant at the 5% level and has a negative coefficient indicating a negative elasticity with the homicide rate. For every 1% increase in population, at the average population level, this model estimates a 0.374% decrease in the homicide rate, controlling for other regressors which are also made up of population. This seems counterintuitive, as I would expect a higher population to lead to more homicides. But the dependent variable is the homicide rate, not the number, so if population grows faster than the number of homicides, then the rate would decrease.

There is an issue of simultaneity here, as population is in the denominator of the homicide rate, and also in the equation as a regressor (and in the denominator of other regressors), so changes in population affect both the dependent and independent variables. It actually is more than simply a problem of simultaneity as the dependent variable with population in the denominator is being explained by population as an independent variable. Question 9 has more discussion of the problems with using population in the denominator, and how this is similar to an interaction term without the individual regressors included separately (especially the numerators).

The 1980-2004 data [5b] still maintains a sufficient effect and level of statistical significance but the magnitude of the coefficient has been reduced. So the larger data set is more robust.

Question 6

	Paper [4]	Replication [4a]	Replication [4d]	Replication [6a]	Replication [6d]
Ln FSS (t-1)	0.086	0.067	0.101	0.047	0.079
	(0.038)**	(0.037)*	(0.038)***	(0.038)	(0.038)**
Ln Rob (t)	0.149	0.042	0.041	0.042	0.043
	(0.042)***	(0.028)	(0.032)	(0.031)	(0.036)
Ln Burg (t)	0.226	0.139	0.200	0.140	0.200
	(0.072)***	(0.069)**	(0.085)**	(0.074)*	(0.092)**
Ln % black (t)	0.278	0.151	0.142	0.172	0.170
	(0.164)*	(0.125)	(0.110)	(0.120)	(0.107)
Ln %Urb (t)	-0.537	-0.588	-0.434	-0.164	-0.109
	(0.157)***	(0.433)	(0.355)	(0.449)	(0.370)
Ln % same h	-0.690	-0.859	-0.594	-0.667	-0.386
	(0.419)	(0.378)**	(0.353)*	(0.369)*	(0.352)
Ln %fem hh	-0.303	0.613	0.902	0.444	0.714
	(0.413)	(0.367)*	(0.366)**	(0.384)	(0.382)*
Ln pop				1.103	0.708
				(0.740)	(0.715)
R-squared	0.923	0.284	0.389	0.231	0.285
(rho)		0.851	0.823	0.880	0.858

The table from question 4 (equation 1.1) has been appended with the results of the new model in question 6 (equation 1.2), so the variable definitions in the leftmost column remove population from the denominator for the rightmost column (i.e. % urban is now # of urban residents for [6a] and [6d]). The 1980-1999 dataset is shown in [4a] and [6a]. The 1980-2004 dataset is shown in [4d] and [6d].

With the addition of population as a regressor, the proxy variable loses its statistical significance for the 1980-1999 dataset and is reduced in effect [6a]. The statistical significance of burglaries and female head of households also decreases. The longer dataset 1980-2004 remains statistically significant, but with reduced effect and significance level [6d]. The addition of population as a regressor and the removal of population from the denominator of the dependent variable reduces the robustness of the proxy variable. Compared to question 5, the coefficient of population is now positive as would be expected, with a higher population leading to a higher number of homicides. But it is also no longer statistically significant. Interpretation of the coefficients and the dependent variable are easier now as they are no longer similar to interaction terms.

Question 7

The following math develops the linear restriction on delta that transforms equation 1.2 into equation 1.1 when the restriction is applied (note, variable names truncated here to save space):

$$\log(\mathrm{e96}_{i,t}) = \alpha + \beta log\left(\frac{e955_{i,t-1}}{e95_{i,t-1}}\right) + \tilde{\gamma}_1 \log(rob_{i,t}) + \tilde{\gamma}_2 \log(burg_{i,t}) + \tilde{\gamma}_3 \log(blacks_{i,t}) + \tilde{\gamma}_4 \log(urban_{i,t}) + \tilde{\gamma}_5 \log(same_{i,t}) + \sigma log\left(\frac{f\mathrm{hh}_{i,t}}{house_{i,t}}\right) + \delta \log(pop_{i,t}) + d_i + d_t + \varepsilon_{i,t} \quad (1.2)$$

Impose the following restriction on δ, where δ = 1- $\tilde{\gamma}_1 - \tilde{\gamma}_2 - \tilde{\gamma}_3 - \tilde{\gamma}_4 - \tilde{\gamma}_5$ and substitute into equation (1.2).

$$\log(\mathrm{e96}_{i,t}) = \alpha + \beta log\left(\frac{e955_{i,t-1}}{e95_{i,t-1}}\right) + \tilde{\gamma}_1 \log(rob_{i,t}) + \tilde{\gamma}_2 \log(burg_{i,t}) + \tilde{\gamma}_3 \log(blacks_{i,t}) + \tilde{\gamma}_4 \log(urban_{i,t}) + \tilde{\gamma}_5 \log(same_{i,t}) + \sigma log\left(\frac{f\mathrm{hh}_{i,t}}{house_{i,t}}\right) + (1 - \tilde{\gamma}_1 - \tilde{\gamma}_2 - \tilde{\gamma}_3 - \tilde{\gamma}_4 - \tilde{\gamma}_5) \log(pop_{i,t}) + d_i + d_t + \varepsilon_{i,t}$$

Now redistribute the log(pop) throughout to get to equation (1.1) which includes subtracting 1*log(pop) from each side of the equation to get the quotient on the left-hand side.

$$\log(\mathrm{e96}_{i,t}) - \log(pop_{i,t}) = \alpha + \beta log\left(\frac{e955_{i,t-1}}{e95_{i,t-1}}\right) + \tilde{\gamma}_1 \log(rob_{i,t}) - \tilde{\gamma}_1 \log(pop_{i,t}) + \tilde{\gamma}_2 \log(burg_{i,t}) - \tilde{\gamma}_2 \log(pop_{i,t}) + \tilde{\gamma}_3 \log(blacks_{i,t}) - \tilde{\gamma}_3 \log(pop_{i,t}) + \tilde{\gamma}_4 \log(urban_{i,t}) - \tilde{\gamma}_4 \log(pop_{i,t}) + \tilde{\gamma}_5 \log(same_{i,t}) - \tilde{\gamma}_5 \log(pop_{i,t}) + \sigma log\left(\frac{f\mathrm{hh}_{i,t}}{\mathrm{house}_{i,t}}\right) + d_i + d_t + \varepsilon_{i,t}$$

Now combine the gamma coefficient terms.

$$\log\left(\frac{\mathrm{e96}_{i,t}}{pop_{i,t}}\right) = \alpha + \beta log\left(\frac{e955_{i,t-1}}{e95_{i,t-1}}\right) + \tilde{\gamma}_1 \log\left(\frac{rob_{i,t}}{pop_{i,t}}\right) + \tilde{\gamma}_2 \log\left(\frac{burg_{i,t}}{pop_{i,t}}\right) + \tilde{\gamma}_3 \log\left(\frac{blacks_{i,t}}{pop_{i,t}}\right) + \tilde{\gamma}_4 \log\left(\frac{urban_{i,t}}{pop_{i,t}}\right) + \tilde{\gamma}_5 \log\left(\frac{same_{i,t}}{pop_{i,t}}\right) + \sigma log\left(\frac{f\mathrm{hh}_{i,t}}{house_{i,t}}\right) + d_i + d_t + \varepsilon_{i,t} \quad (1.1)$$

Equation 1.1 can be derived from equation 1.2 (as shown above) if the appropriate linear restriction is placed on delta. However, given the dataset, this linear restriction may not be consistent, as is the case here. The linear restriction test utilizes the F-statistic as it requires the standard errors of the linear combination of the estimated coefficients, which includes covariance terms. The null hypothesis is that the restriction is true, or consistent with the data set. A large F-statistic (and small p-value) will reject the null hypothesis, indicating that the linear restriction is not consistent with the dataset. For both the 1980-1999 and 1980-2004 datasets, the restriction on delta has a large F-statistic and p-value less than the standard 5% or 10% level, so the null hypothesis of the restriction being consistent with equation 1.2 is rejected. Equation 1.1 is not a consistent model derived from 1.2, and is misspecified (assuming 1.2 is correctly specified). Thus 1.1 may have biased coefficients and errors if 1.2 is correctly specified.

Testing restriction **δ = 1- $\tilde{\gamma}_1 - \tilde{\gamma}_2 - \tilde{\gamma}_3 - \tilde{\gamma}_4 - \tilde{\gamma}_5$**		
	1980-1999	1980-2004
F (1,195)	5.12	5.11
Prob > F (p-value)	0.0247	0.0249

<u>More General Model - Derivation</u>

Start with the more general model, defined to be equation (1.3) with no fractions:

$$\boxed{\begin{aligned}\log(\mathrm{e96}_{i,t}) = \alpha + \beta_1 \log(e955_{i,t-1}) + \beta_2 \log(e95_{i,t-1}) + \tilde{\gamma}_1 \log(rob_{i,t}) + \tilde{\gamma}_2 \log(burg_{i,t}) + \tilde{\gamma}_3 \log(blacks_{i,t}) + \\ \tilde{\gamma}_4 \log(urban_{i,t}) + \tilde{\gamma}_5 \log(same_{i,t}) + \sigma_1 \log(f\mathrm{hh}_{i,t}) + \sigma_2 \log(house_{i,t}) + \delta \log(pop_{i,t}) + d_i + d_t + \varepsilon_{i,t} \quad (1.3)\end{aligned}}$$

Impose linear restrictions on equation (1.3) to get to equation (1.1).

The restrictions are as follows: (1) $\boldsymbol{\delta = 1 - \tilde{\gamma}_1 - \tilde{\gamma}_2 - \tilde{\gamma}_3 - \tilde{\gamma}_4 - \tilde{\gamma}_5}$; **(2)** $\boldsymbol{\beta_1 = -\beta_2}$; **(3)** $\boldsymbol{\sigma_1 = -\sigma_2}$

$$\begin{aligned}\log(\mathrm{e96}_{i,t}) = \alpha + \overbrace{\beta_1 \log(e955_{i,t-1}) + \beta_2 \log(e95_{i,t-1})}^{apply\ restriction\ 2:\ \beta_1=-\beta_2} + \\ \overbrace{\tilde{\gamma}_1 \log(rob_{i,t}) + \tilde{\gamma}_2 \log(burg_{i,t}) + \tilde{\gamma}_3 \log(blacks_{i,t}) + \tilde{\gamma}_4 \log(urban_{i,t}) + \tilde{\gamma}_5 \log(same_{i,t})}^{apply\ restriction\ 1} + \\ \overbrace{\sigma_1 \log(f\mathrm{hh}_{i,t}) + \sigma_2 \log(house_{i,t})}^{apply\ restriction\ 3\ \sigma_1=-\sigma_2} + \delta \log(pop_{i,t}) + d_i + d_t + \varepsilon_{i,t}\end{aligned}$$

Again, redistribute log(pop) by subtracting log(pop) from both sides and applying constraint 1. The result will be the restricted model in equation (1.1).

$$\boxed{\begin{aligned}\log\left(\frac{\mathrm{e96}_{i,t}}{pop_{i,t}}\right) = \alpha + \beta log\left(\frac{e955_{i,t-1}}{\mathrm{e95}_{i,t-1}}\right) + \tilde{\gamma}_1 \log\left(\frac{rob_{i,t}}{pop_{i,t}}\right) + \tilde{\gamma}_2 \log\left(\frac{burg_{i,t}}{pop_{i,t}}\right) + \tilde{\gamma}_3 \log\left(\frac{blac_{i,t}}{pop_{i,t}}\right) + \tilde{\gamma}_4 \log\left(\frac{urban_{i,t}}{pop_{i,t}}\right) + \tilde{\gamma}_5 \log\left(\frac{same_{i,t}}{pop_{i,t}}\right) + \\ \sigma log\left(\frac{f\mathrm{hh}_{i,t}}{\mathrm{house}_{i,t}}\right) + d_i + d_t + \varepsilon_{i,t} \quad (1.1)\end{aligned}}$$

Starting with a more general model (equation 1.3) that does not include any denominators, equation 1.1 can be derived by applying linear restrictions to certain parameters. The same restriction on delta is imposed, and the coefficients associated with suicide measures and for household data are set equal to each other (with opposite signs), resulting in equation 1.1, but also requiring that the coefficients for suicides and suicides by firearm have the same magnitude on both events (same for households and female head of households). This is a very strong assumption, and in this case, does not generally hold at the 5% or 10% critical levels.

Here, there are three linear restrictions which require a joint test, by comparing the explanatory power (residual sum of squares, "SSR") of the unrestricted model (equation 1.3) with the explanatory power of the restricted model (equation 1.1). If the "SSR" of the restricted model is significantly higher than the unrestricted model, then the restrictions reduce the explanatory power of the model by too large of an extent, indicating that the restricted model is misspecified. That is generally the case here with the restrictions placed on equation 1.3, given this dataset. The F-statistic is too high (p-value too small, although the 1980-2004 data is rejected at the 10% critical p-value rather than at 5%). Equation 1.3 is estimated in the do and log files. The F-test of the three linear restrictions is shown below for each dataset. Generally, the null hypothesis that the restrictions are consistent with the data is rejected at the 10% level, with the 1980-2004 dataset not being rejected at the 5% level.

Testing restrictions (1) $\boldsymbol{\delta = 1 - \tilde{\gamma}_1 - \tilde{\gamma}_2 - \tilde{\gamma}_3 - \tilde{\gamma}_4 - \tilde{\gamma}_5}$;
(2) $\boldsymbol{\beta_1 = -\beta_2}$; **(3)** $\boldsymbol{\sigma_1 = -\sigma_2}$

	<u>1980-1999</u>	<u>1980-2004</u>
F (3,195)	3.62	2.33
Prob > F (p-value)	0.014	0.076

Question 8

1980-1999 dataset

	Q4	Q5	Q6	Q8
$\tilde{\gamma}_1$ (robberies)	0.042	0.042	0.042	0.042
	(0.028)	(0.031)	(0.031)	(0.029)
$\tilde{\gamma}_2$ (burglaries)	0.139	0.140	0.140	0.139
	(0.069)**	(0.074)*	(0.074)*	(0.071)**
$\tilde{\gamma}_3$ (blacks)	0.151	0.172	0.172	0.151
	(0.125)	(0.120)	(0.120)	(0.128)
$\tilde{\gamma}_4$ (urban)	-0.588	-0.164	-0.164	-0.588
	(0.433)	(0.449)	(0.449)	(0.444)
$\tilde{\gamma}_5$ (same res)	-0.859	-0.667	-0.667	-0.859
	(0.378)**	(0.369)*	(0.369)*	(0.388)**
δ (population)		-0.374	1.103	2.115
		(0.165)**	(0.740)	(0.711)***

The restriction in column [Q8] on delta does match (delta = 1-sum(gammas)). [Q8] uses equation 1.2 from [Q6] but applies the restriction discussed in question 7, resulting in equation 1.1. Hence, the gamma coefficients in [Q8] match equation 1.1 in [Q4] exactly, verifying the constraint on equation 1.2, with only some slight differences in the standard error terms.

Assuming equation 1.2 in [Q6] is correctly specified, applying the constraint results in changing both the coefficients and standard errors of the gammas, including increasing the statistical significance of both burglaries and those that did not move residence. The magnitudes of the coefficient are also different, which can lead to inaccurate interpretations and conclusions.

Similar results are obtained for the 1980-2004 dataset and are included in the do and log files. As a programming note, cnsreg does not accommodate fixed effects directly, so a dummy variable regression is run instead, which returns a fixed effects regression (add i.fips to the regression command, i.year already present).

Question 9

	1980-1999	1980-2004
Natural Log of Proxy Gun Ownership % 1 yr lag = D1	-0.008	-0.015
	(0.027)	(0.025)
Natural log of # robberies = D1	0.012	-0.003
	(0.008)	(0.012)
Natural log of # burglaries = D1	0.053	0.036
	(0.027)*	(0.028)
Natural log of # Afr Amer = D1	0.181	0.246
	(0.142)	(0.165)
Natural log of # urbanites = D1	0.771	0.886
	(0.695)	(0.617)
Natural log of # residents who did not move = D1	-0.937	-0.561
	(0.366)**	(0.336)*
Natural Log of % Female Head House in pop = D1	1.318	1.046
	(0.350)***	(0.368)***
Natural Log of Population = D1	0.736	0.167
	(0.944)	(0.835)
Constant	-0.033	-0.024
	(0.005)***	(0.005)***
Observations	3,567	4,560
R-squared	0.054	0.043

After applying first differences, the proxy variable no longer has any explanatory power. The coefficient is very close to zero and is not statistically significant at all. The coefficients should not have changed very much in a first difference regression as compared to a fixed effects regression if the original coefficients were correct. So, there is not a robust relation between homicides and the proxy for gun ownership.

There are unobserved factors in the error term which can be constant or variable over time. The goal of first differencing is to eliminate any time constant omitted variable bias, so that constant unobserved effects in the error term are no longer correlated with the independent variables. This is done by controlling for them with the county dummy variable, and this is eliminated in the first differencing. (Side note: first differencing can make an integrated time series weakly dependent). When first differencing is applied here, the proxy no longer has any explanatory power on homicides.

First differencing with panel data can present some new issues. The coefficients may be biased if the explanatory variables do not vary much over time, or if explanatory variables aren't exogenous or if the explanatory variables are subject to measurement error (all of which is present here to some degree).

Nevertheless, the problem with equation 1.1, with including population in the denominators, is that it requires the coefficients of the independent variable to have the same impact on both the numerator and the denominator when it is very likely that the numerator and denominator act in different ways. Including population in the denominator can change the standard errors and the distribution of those errors, which then affects or biases the coefficients. Both the dependent and independent variables with population in the denominator are essentially similar to interaction terms, but the rest of the regression equation doesn't have the corresponding individual terms of the interaction as individual regressors (except for population in Q6, but the right hand side still lacks the numerators as stand alone regressors). This can also change the coefficients and standard errors. It also seems incorrect to use an interaction term as the dependent variable.

One other problem that might be worth exploring is if there is spurious correlation in equation 1.1 from having two independent I(1) variables, x and y, that are correlated to a third variable, z, and the relationship between x and y disappears when adding another independent variable z, leading to incorrectly high R^2 and t-statistics. This is briefly explored in question 10.

In determining which model is best to use, I would choose the first difference model or the general model 1.3 developed in the bonus part of question 7. The more general equation 1.3 found that the gun ownership proxy was statistically insignificant in explaining homicides and gets rid of the quasi interaction term "problem." The fixed effects and first differencing results should be similar, and without further exploration and explanation of the differences, simply accepting the fixed effects model showing statistically significant effects for the proxy would be irresponsible. I would also conclude that the proxy isn't strong enough to find a causal effect on the homicide rate. There were suspicions early on about the validity of this proxy given its annual variability and its very rare occurrence rate. I would instead search for a better proxy for gun ownership and use a better specified model.

<u>Question 10</u>

Table3: Sensitivity Analysis (1980-1999 dataset)

	Paper Ln(Hom)	Replication Ln(Hom)	Paper Ln(Gun)	Replication Ln(Gun)	Paper Ln(nongun)	Replication Ln(nongun)
Alternative specifications						
Baseline model, [4] table2	0.086 (0.038)**	0.067 (0.037)*	0.173 (0.049)***	0.148 (0.051)***	-0.033 (0.040)	-0.041 (0.037)
First Differences		-0.008 (0.027)		-0.036 (0.043)		
3 Baseline, unweighted	0.051 (0.043)	0.040 (0.033)	0.167 (0.043)***	0.132 (0.048)***	-0.061 (0.042)	-0.064 (0.037)*
First Differences				-0.012 (0.047)		-0.020 (0.044)

Specific portions of the top panel of table 3 are replicated, using the gun homicide rate and the nongun homicide rate as the dependent variable, and testing the unweighted data as well. The replications match the paper's results fairly well, both in quantitative effect and qualitatively. The proxy for gun ownership is statistically significant for predicting the gun homicide rate and is strongly positive in effect, for both the weighted and unweighted data. The proxy is not significant for nongun homicides for the weighted data, but is negative in value, which may correspond to the paper's hypothesis about the prevalence of guns causing a substitution effect for murder weapon choice. But it's not statistically significant, so the results are inconclusive. The unweighted sample is significant at the 10% level, and does have a negative coefficient. Here, it would seem, more guns reduce the number of nongun homicides, pointing towards a possible substitution effect in weapon choice.

I’ve also rerun the statistically significant results through a first differencing test similar to question 9. Removing the fixed effects has the same effect, eliminating all statistical significance and effect of the proxy variable for gun ownership on both the gun and non gun homicide rates. Thus the results and not robust and are questionable.

I also briefly explored the idea having two I(1) variables discussed earlier in question 9. I examined the dependent variable forms as well as the proxy independent variable forms only. I applied the Harris-Tzavalis test, which requires strongly balanced data, so I used the original dataset and removed only those counties with missing values in any of the homicide, population or firearm suicide categories. I apply the xtunitroot ht test in the do and log files. When applied to the number of homicides and the number of firearm suicides, both variables have unit root. After incorporating their respective denominators, the dependent variable no longer has unit root but the independent variable still does. Further tests for the other independent variables could be run to see if they also have unit root, and cointegration tests (such as xtcointtest pedroni) could further explore the issue of spurious regressions, but this would require significantly more data manipulation to get a balanced panel and significantly more time.