Econometric Methods:
Solutions to Empirical Exercise 10.1
Chapter 10: Regression with Panel Data
Stock & Watson, 3^{rd} Edition

Zaeen de Souza Deepti Goel *

Azim Premji University
09 February 2022

* Solution key prepared jointly by Zaeen and Deepti. R code and presentation in Rmarkdown by Zaeen.

Contents

Background: Empirical Exercise 10.1

These are the solutions to **E10.1** from **Chapter 10** of *Introduction to Econometrics (Updated Third edition)* by Stock & Watson. You should have the following on your computer in order to check answers/run the code and follow the questions in this assignment:

- An updated version of `R` and `Rstudio`.
- The following packages installed:
  - `ggplot2`
  - `readxl`
  - `stargazer`
  - `fixest`
- The dataset called `Guns`.
- The data description pdf to understand the variables being used.

Reading guide

All the code needed to complete the assignments is within this document. R code will be in a grey box and will look like this:

```
summary(iris)
```

And all R output i.e what R shows you once you run some code, will have `#` signs next to it, and will look like this:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.300   5.100   5.800   5.843   6.400   7.900
```

As far as possible these guides will show the **exact output** that comes from running code in `R`, and at times will use formatted tables made in latex. The results themselves, will be identical. Some things to note, that might make output look different accross different computers:

- R reports things like p-values using scientific notation, but some computers report the numbers with many trailing zeros.
- If you have an old version of `R` or `Rstudio` it is highly recommended that you update it using the following code:

```
# Use this to update R from within RStudio
install.packages("installr")
library(installr)
# This last command, will open up a download prompt; choose yes/no accordingly.
updateR()
```

For updating `Rstudio`, un-install your version of `RStudio`, and download a fresh version from the `RStudio` website.

Loading the data and libraries

The following code sets the working directory to the folder where you have downloaded the data, loads the libraries needed fo the assignment and loads the excel dataset.

```
# Loading excel files
install.packages("fixest")
library(fixest)
library(readxl)
library(ggplot2)
library(dplyr)

# Setting working directory - this is unique to your computer
setwd("~Zaeen de Souza/Chapter 10 Panel Data")

# Loading the data as 'pset_data'
pset_data <- read_excel("Guns.xlsx")
```

E10.1 Problem Context

Some U.S. states have enacted laws that allow citizens to carry concealed weapons. These laws are known as "shall-issue" laws because they instruct local authorities to issue a concealed weapons permit to all applicants who are citizens, are mentally competent, and have not been convicted of a felony. (Some states have some additional restrictions.) Proponents argue that if more people carry concealed weapons, crime will decline because criminals will be deterred from attacking other people. Opponents argue that crime will increase because of accidental or spontaneous use of the weapons. In this exercise, you will analyze the effect of concealed weapons laws on violent crimes. On the textbook website, www.pearsonglobaleditions .com/Stock_Watson, you will find the data file Guns, which contains a balanced panel of data from the 50 U.S. states plus the District of Columbia for the years 1977 through 1999. A detailed description is given in Guns_Description, available on the website.

Exercise E10.1

a. Estimate (1) a regression of ln(vio) against shall and (2) a regression of ln(vio) against shall, incarc_rate, density, avginc, pop, pb1064, pw1064, and pm1029.

We will estimate the following pooled models, using OLS:

$$ln(Y_{it}) = \alpha + \beta_1 Shall_{it} + \varepsilon_{it} \tag{1}$$

$$ln(Y_{it}) = \alpha + \beta_1 Shall_{it} + \beta' \mathbf{X}_{it} + \varepsilon_{it} \tag{2}$$

Where $ln(Y_{it})$ is the natural log of violent crimes per 100,000 in State i in year t. $\mathbf{X}$ in (2) is a vector of controls that includes: incarc_rate, density, avginc, pop, pb1064, pw1064, and pm1029. The results are in table 1.

```
# vcov manipulates the variance--covariance matrix for robust inference
model_1 <- feols(ln_vio ~ shall,
                         data = pset_data,
                         vcov = "HC1")
model_2 <- feols(ln_vio ~ shall + incarc_rate + density + avginc +
                         pop + pb1064 + pw1064 + pm1029,
                         data = pset_data, vcov = "HC1")
```

a i. Interpret the coefficient on shall in regression (2). Is this estimate large or small in a "real-world" sense?

The coefficient $\widehat{\beta}_1$ in regression (2) suggests that on average, across the time period (1977-1999), States that had a "Shall-Carry" law, experienced 36.8% fewer violent crimes per 100,000, than those States that did not have such laws. To get a "real-world" sense of what this means, we can do the following. First, we estimate mean violent crimes per 100,000 across the sample:

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Vio_{it} = 503$$

We then calculate $(\bar{Y} \cdot \widehat{\beta}_1)$ i.e 36.8% of $\bar{Y}$. This tells us that the presence of these laws was associated with an average reduction of 185 violent crimes per 100,000.

a ii. Does adding the control variables in regression (2) change the estimated effect of a shall-carry law in regression (1) as measured by statistical significance? As measured by the "real-world" significance of the estimated coefficient?

The control variables pull the coefficient down by around 0.075 log points. This seems to suggest that the pooled regression with no controls is overestimating the relationship between the laws and violent crime. Statistically, speaking, the significance is unchanged.

a iii. Suggest a variable that varies across states but plausibly varies little—or not at all—over time and that could cause omitted variable bias in regression (2).

Different States have different institutions such as the quality of their police forces, criminal rehabilitation programs, and as well, citizens preferences for guns themselves. These introduce an omitted variable bias in both (1) and (2).

Table 1: E.10.1 a

	ln(Violent Crime/100,000)	
	(1)	(2)
Shall	-0.443***	-0.368***
	(0.048)	(0.035)
Incaration Rate		0.002***
		(0.0002)
Pop. Density		0.027*
		(0.014)
Avg. Per Capita Income		0.001
		(0.007)
Total Population		0.043***
		(0.003)
% Black		0.081***
		(0.020)
% White		0.031***
		(0.010)
% Young Male		0.009
		(0.012)
(Intercept)	6.13***	2.98***
	(0.019)	(0.609)
R^2	0.087	0.564
Observations	1,173	1,173

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

b. Do the results change when you add fixed state effects? If so, which set of regression results is more credible, and why?

We will now estimate the following linear model using State fixed effects:

$$ln(Y_{it}) = \alpha_i + \beta_1 Shall_{it} + \beta' \mathbf{X}_{it} + \varepsilon_{it}$$

Where, $ln(Y_{it})$, *Shall* and $\mathbf{X}$ are as defined earlier. α_i are State fixed effects.

```
# Every variable after the | sign is used as a fixed effect
statefe <- feols(ln_vio ~ shall + incarc_rate + density + avginc +
                          pop + pb1064 + pw1064 + pm1029 | stateid,
                          data = pset_data, vcov = "HC1")
```

Once we control for state fixed effects, the coefficient changed from -0.368 to -0.046, a substantial drop in magnitude. The regression with State fixed effects is more credible as we have enlisted some potential time-invariant state specific variables that could bias our coefficient if we do not account for them.

c. Do the results change when you add fixed time effects? If so, which set of regression results is more credible, and why?

We will now estimate a twoway fixed effects models.

$$ln(Y_{it}) = \alpha_i + \delta_t + \beta_1 Shall_{it} + \beta' \mathbf{X}_{it} + \varepsilon_{it}$$

Where, $ln(Y_{it})$, *Shall* and $\mathbf{X}$ are as defined earlier. α_i are State fixed effects and δ_t are the Year fixed effects.

```
# Every variable after the | sign is used as a fixed effect
# add multiple FEs by using the + sign
twfe <- feols(ln_vio ~ shall + incarc_rate + density + avginc +
                       pop + pb1064 + pw1064 + pm1029 | stateid + year,
                       data = pset_data, vcov = "HC1")
```

$\widehat{\beta}_1$, which can be seen in table 2 column 2, is not statistically significant anymore, and is also far smaller in magnitude than the pooled OLS regression in table 1, column 2. This implies that the effect seen in the pooled model was being driven by unobserved omitted variables that are State-specific and time-invariant AND year-specific and State-invariant.

Table 2: E.10.1 b and c

	ln(Violent Crime/100,000)	
	(1)	(2)
Shall	-0.046**	-0.028
	(0.020)	(0.019)
Incaration Rate	-7.1×10^{-5}	7.6×10^{-5}
	(9.73×10^{-5})	(8.29×10^{-5})
Pop. Density	-0.172	-0.092
	(0.105)	(0.065)
Avg. Per Capita Income	-0.009	0.0010
	(0.007)	(0.007)
Total Population	0.011	-0.005
	(0.010)	(0.007)
% Black	0.104***	0.029
	(0.017)	(0.021)
% White	0.041***	0.009
	(0.005)	(0.009)
% Young Male	-0.050***	0.073***
	(0.008)	(0.019)
State Fixed Effects	✓	✓
Year Fixed Effects		✓
R^2	0.941	0.956
Within R^2	0.218	0.056
Observations	1,173	1,173

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

d. Repeat the analysis using ln(rob) and ln(mur) in place of ln(vio).

The code is below, and the results are in table 3.

```
twfe_vio <- feols(ln_vio ~ shall + incarc_rate + density + avginc +
                          pop + pb1064 + pw1064 + pm1029 | stateid + year,
                          data = pset_data, vcov = "HC1")

twfe_mur <- feols(ln_mur ~ shall + incarc_rate + density + avginc +
                          pop + pb1064 + pw1064 + pm1029 | stateid + year,
                          data = pset_data, vcov = "HC1")

twfe_rob <- feols(ln_rob ~ shall + incarc_rate + density + avginc +
                          pop + pb1064 + pw1064 + pm1029 | stateid + year,
                          data = pset_data, vcov = "HC1")
```

Table 3: E.10.1 d

	ln(Violent Crime/100,000) (1)	ln(Murders/100,000) (2)	ln(Robberies/100,000) (3)
Shall	-0.028	-0.015	0.027
	(0.019)	(0.027)	(0.025)
Incaration Rate	7.6×10^{-5}	-0.0001	3.14×10^{-5}
	(8.29×10^{-5})	(0.0001)	(0.0001)
Pop. Density	-0.092	-0.544***	-0.045
	(0.065)	(0.127)	(0.089)
Avg. Per Capita Income	0.0010	0.057***	0.014
	(0.007)	(0.010)	(0.010)
Total Population	-0.005	-0.032***	1.64×10^{-5}
	(0.007)	(0.009)	(0.011)
% Black	0.029	0.022	0.014
	(0.021)	(0.040)	(0.034)
% White	0.009	-0.0005	-0.013
	(0.009)	(0.013)	(0.012)
% Young Male	0.073***	0.069***	0.105***
	(0.019)	(0.025)	(0.026)
State Fixed Effects	✓	✓	✓
Year Fixed Effects	✓	✓	✓
R^2	0.956	0.923	0.962
Within R^2	0.056	0.116	0.049
Observations	1,173	1,173	1,173

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

e. In your view, what are the most important remaining threats to the internal validity of this regression analysis?

The twoway fixed effects regression accounts for time invariant, unit specific unobserved heterogeneity and unit invariant, year specific unobserved heterogeneity. While this identification strategy inoculates against omitted variables that fall into these two categories, there are *at least*, the following existing threats to internal validity.

1. The main threat comes from omitted variables that are both state AND time varying such as say, state and year specific in-migration rates. Time and State fixed effects do not account for OVB (Omitted variable bias) from such specifications.
2. Recent literature has discredited the use of TWFE, when multiple units get treated at different times and there is potential for differential effects over time and across units. It can be shown that under such circumstances, the strict exogeneity assumption is violated. The good news is that there *are* recent methods to tackle this problem (Callaway and Sant'Anna 2021; Roth et al. 2022).
3. Recent literature says that in order to get correct standard errors that account for correlated errors within states, one must cluster standard errors at state level. While clustering will not change the magnitude of the effect (i.e $\widehat{\beta}_1$), it might change whether or not the coefficient is statistically different from 0 or not.

f. Based on your analysis, what conclusions would you draw about the effects of concealed weapons laws on these crime rates?

Assuming that the additional assumption needed to use the TWFE models, hold good, this analysis suggests that the naive pooled regression overestimates the effect of Shall Issue gun laws on crime. After controlling for a series of known time varying correlates of Shall Issue gun laws and crime, time invariant State specific unobserved heterogeneity, State invariant year specific unobserved heterogeneity, we can conclude that Shall Issue gun laws have no significant effect on crime—where crime is defined as ln(murders per 100,000), ln(Robberies per 100,000) and ln(Violent crimes per 100,000).

Extra Material 1: Cluster Robust Inference in R

In order to cluster standard errors within the `fixest` package, we can do the following. Note the syntax is slightly different now—`cluster = ~ stateid` tells `fixest` that we want to cluster our errors at the State level. You can compare table 4 to table 3 and look at the differences in standard errors.

```
# cluster tells R to cluster at the level some variable - in this at the State level.
twfe_vio <- feols(ln_vio ~ shall + incarc_rate +
                           density + avginc +
                           pop + pb1064 +
                           pw1064 + pm1029 | stateid + year,
                           data = pset_data,
                           cluster = ~ stateid)

twfe_mur <- feols(ln_mur ~ shall + incarc_rate +
                           density + avginc +
                           pop + pb1064 +
                           pw1064 + pm1029 | stateid + year,
                           data = pset_data,
                           cluster = ~ stateid)

twfe_rob <- feols(ln_rob ~ shall + incarc_rate +
                           density + avginc +
                           pop + pb1064 +
                           pw1064 + pm1029 | stateid + year,
                           data = pset_data,
                           cluster = ~ stateid)
```

Table 4: Clustered Standard Errors

	ln(Violent Crime/100,000) (1)	ln(Murders/100,000) (2)	ln(Robberies/100,000) (3)
Shall	-0.028	-0.015	0.027
	(0.041)	(0.038)	(0.052)
Incaration Rate	7.6×10^{-5}	-0.0001	3.14×10^{-5}
	(0.0002)	(0.0004)	(0.0003)
Pop. Density	-0.092	-0.544*	-0.045
	(0.124)	(0.319)	(0.198)
Avg. Per Capita Income	0.0010	0.057***	0.014
	(0.016)	(0.017)	(0.025)
Total Population	-0.005	-0.032	1.64×10^{-5}
	(0.015)	(0.021)	(0.026)
% Black	0.029	0.022	0.014
	(0.050)	(0.076)	(0.084)
% White	0.009	-0.0005	-0.013
	(0.024)	(0.020)	(0.033)
% Young Male	0.073	0.069	0.105
	(0.052)	(0.042)	(0.073)
State Fixed Effects	✓	✓	✓
Year Fixed Effects	✓	✓	✓
R^2	0.95618	0.92254	0.96171
Within R^2	0.05635	0.11562	0.04910
Observations	1,173	1,173	1,173

Clustered (State) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Extra Material 2: Computation Speed in R

By now you will know enough `R` to know that base functions are often slow. At this point, it is recommended that you shift from using the `R` base function `lm()` or even user contributed ones like `plm()`, to the newer package called `fixest()`, which we used in this problem set. `Fixest` is substantially faster, and offers easier methods to use a wide range of robust standard errors. For those interested in how different `R` regression functions fare in terms of computation speed, figure 1[1], shows you how powerful (and fast) `fixest` is.

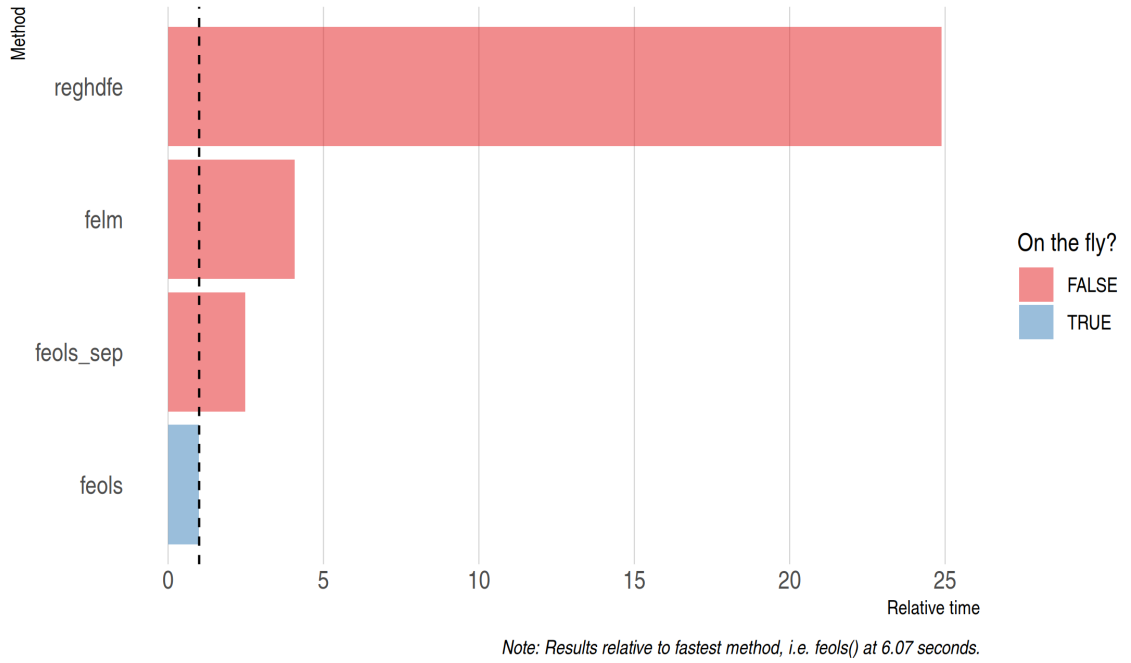

Figure 1: Source: Shared by Grant Mcdermott, Assistant Prof. at the University of Oregon.

References

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225 (2): 200–230.

Roth, Jonathan, Pedro Sant'Anna, Alyssa Bilinski, and John Poe. 2022. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature." *Working Paper*.

[1]On the fly means you can adjust SEs within the function.