Recitation 5: Panel Data and Binary Dependent Variables

Matthew Alampay Davis

June 27, 2023

```
# Load all datasets used in this Notebook
guns <- read_dta('data/handguns.dta') # Practice Question 1
smoking <- read_dta('data/Smoking.dta') # Practice Question 2
```

Practice Question 1: Stock-Watson Empirical Exercise 10.1

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.

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 are using the feols function ("fixed effects estimation by OLS") in the fixest package. Note also that this is panel data where the unit is defined by the state variable and the time is defined by the year variable. In this context, we prefer to use standard errors clustered at the state level:

```
guns %<>% mutate(log.vio = log(vio))

mod.1a1 <- feols(log.vio ~ shall, data = guns)
mod.1a2 <- feols(log.vio ~ shall, data = guns)
```

```
etable(mod.1a1, mod.1a2, cluster = 'state', digits = 6, markdown = T)
```

Dependent Variable:	log.vio	
Model:	(1)	(2)
Variables		
Constant	6.13492***	6.13492***
	(0.079027)	(0.079027)
shall	-0.442965***	-0.442965***
	(0.157018)	(0.157018)
Fit statistics		
Observations	1,173	1,173
R^2	0.08664	0.08664
Adjusted R^2	0.08586	0.08586

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

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

The coefficient is -0.368, which suggests that shall-issue laws is associated with a violent crime rate 36% lower. The p-value associated with this estimate is less than 0.001, indicating it is a statistically significant effect. The magnitude of this effect is also clearly large in a real-world sense.

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

The coefficient in (1) is -0.443; in (2) it is -0.369. Both are highly statistically significant. Adding the control variables results in a small drop in the coefficient.

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).

There are several examples. Here are two: Attitudes towards guns and crime, and quality of police and other crime-prevention programs.

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

```
mod.1b <- feols(log.vio ~ shall + incarc_rate + density + avginc + pop + pb1064 + pw1064 + pm1029 | sta
etable(mod.1a2, mod.1b, cluster = 'state', markdown = T)
```

Dependent Variable:	log.vio	
Model:	(1)	(2)
Variables		
Constant	6.135***	
	(0.0790)	
shall	-0.4430***	-0.0461
	(0.1570)	(0.0418)
incarc_rate		-7.1×10^{-5}
		(0.0003)
density		-0.1723
		(0.1376)
avginc		-0.0092
		(0.0130)
pop		0.0115
		(0.0142)
pb1064		0.1043***
		(0.0327)
pw1064		0.0409***
		(0.0135)
pm1029		-0.0503**
		(0.0207)
Fixed-effects		
state		Yes
Fit statistics		
Observations	1,173	1,173
R^2	0.08664	0.94111
Within R^2		0.21779

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

Note that we here include state fixed effects just by referring to the "state" variable to the regression formula with a vertical bar "|" separating it from the regular part of the formula.

In this new regression with state fixed effects, the coefficient on shall falls to -0.046, a large reduction. Evidently there was important omitted variable bias in (2). Further, the estimate is not statistically significantly different from zero.

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

```
mod.1c <- feols(log.vio ~ shall + incarc_rate + density + avginc + pop + pb1064 + pw1064 + pm1029 | sta
etable(mod.1a2, mod.1b, mod.1c, cluster = 'state', markdown = T)
```

Dependent Variable:		log.vio	
Model:	(1)	(2)	(3)
Variables			
Constant	6.135***		
	(0.0790)		
shall	-0.4430***	-0.0461	-0.0280
	(0.1570)	(0.0418)	(0.0407)
incarc_rate		-7.1×10^{-5}	7.6×10^{-5}
		(0.0003)	(0.0002)
density		-0.1723	-0.0916
		(0.1376)	(0.1239)
avginc		-0.0092	0.0010
		(0.0130)	(0.0165)
pop		0.0115	-0.0048
		(0.0142)	(0.0152)
pb1064		0.1043***	0.0292
		(0.0327)	(0.0495)
pw1064		0.0409***	0.0092
		(0.0135)	(0.0238)
pm1029		-0.0503**	0.0733
		(0.0207)	(0.0525)
Fixed-effects			
state		Yes	Yes
year			Yes
Fit statistics			
Observations	1,173	1,173	1,173
R^2	0.08664	0.94111	0.95618
Within R^2		0.21779	0.05635

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

The coefficient falls further to -0.028. The coefficient is insignificantly different from zero.

Testing the joint significance of the time fixed effects. There isn't an in-built function to do this (since it's an unusual test) but here's how you could do it:

```
# Estimate the model but with time dummies explicitly included
mod.1c.dummies <- feols(log.vio ~ shall + incarc_rate + density + avginc + pop + pb1064 + pw1064 + pm10
# Test the linear hypothesis that all the dummies (variables beginning with factor) have a coefficient
wald(mod.1c.dummies, cluster = 'state', keep = 'year', )
```

```
## Wald test, H0: joint nullity of factor(year)78, factor(year)79, factor(year)80, factor(year)81, facto
##  stat = 21.6, p-value < 2.2e-16, on 22 and 1,092 DoF, VCOV: Clustered (state).
```

The year fixed effects are jointly statistically significant with an F-statistic of 21.6, so this regression seems better specified than (3).

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

Here's a very compact way of displaying all 12 regressions that questions 1a-1d is asking for:

```r
guns %<>% mutate(log.rob = log(rob),
                 log.mur = log(mur))
mods.1d <- feols(c(log.vio, log.rob, log.mur) ~ # vector of all outcome variables of interest
                   # Cumulative stepwise function (csw) to run models, cumulatively adding more regress
                   csw(shall,
                       incarc_rate + density + avginc + pop + pb1064 + pw1064 + pm1029,
                       # Include fixed effects here too as factor variables
                       factor(state),
                       factor(year)),
                 data = guns,
                 # Cluster SEs for all these regressions at the state level
                 cluster = ~ state)
etable(mods.1d,
       # display 95% confidence interval rather than standard errors
       coefstat = 'confint', ci = 0.95,
       # cluster all regressions' standard errors at the state level
       cluster = 'state',
       # treat all factor variables (state and/or year) as fixed effects,
       keepFactors = F,
       # include more statistics
       fitstat = ~ r2 + ar2 + rmse + wald + wf, markdown = T)
```

Dependent Variables:	log.vio					
Model:	(1)	(2)	(3)	(4)	(5)	
Variables						
Constant	6.135***	2.982	4.037***	3.972***	4.873***	
	[5.976; 6.294]	[-1.370; 7.333]	[2.539; 5.534]	[1.743; 6.201]	[4.641; 5.105]	[
shall	-0.4430***	-0.3684***	-0.0461	-0.0280	-0.7733***	
	[-0.7583; -0.1276]	[-0.5972; -0.1395]	[-0.1319; 0.0396]	[-0.1116; 0.0556]	[-1.226; -0.3209]	[-0
incarc_rate		0.0016***	-7.1×10^{-5}	7.6×10^{-5}		
		[0.0004; 0.0028]	[-0.0006; 0.0004]	[-0.0004; 0.0005]		[-0
density		0.0267	-0.1723	-0.0916		
		[-0.0566; 0.1100]	[-0.4548; 0.1102]	[-0.3460; 0.1629]		[-0
avginc		0.0012	-0.0092	0.0010		
		[-0.0472; 0.0496]	[-0.0358; 0.0174]	[-0.0329; 0.0348]		[-0
pop		0.0427***	0.0115	-0.0048		
		[0.0192; 0.0663]	[-0.0177; 0.0407]	[-0.0360; 0.0265]		[0
pb1064		0.0809	0.1043***	0.0292		
		[-0.0625; 0.2242]	[0.0372; 0.1714]	[-0.0726; 0.1309]		[-0
pw1064		0.0312	0.0409***	0.0092		
		[-0.0373; 0.0997]	[0.0132; 0.0685]	[-0.0396; 0.0580]		[-0
pm1029		0.0089	-0.0503**	0.0733		
		[-0.0596; 0.0774]	[-0.0928; -0.0078]	[-0.0345; 0.1811]		[-0
Fixed-effects						
state			Yes	Yes		
year				Yes		
Fit statistics						
R^2	0.08664	0.56426	0.94111	0.95618	0.12081	
Adjusted R^2	0.08586	0.56126	0.93804	0.95297	0.12006	
RMSE	0.61683	0.42605	0.15663	0.13511	0.89472	
Wald (joint nullity)	7.9586	62.125	32.631	54.370	11.789	

Clustered (state) co-variance matrix, 95% confidence intervals in brackets
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

The quantitative results are similar to the results using violent crimes: there is a large estimated effect of concealed weapons laws in specifications (1) and (2). This effect is spurious and is due to omitted variable bias as specification (3) and (4) show.

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

There is potential two-way causality between this year's incarceration rate and the number of crimes. Because this year's incarceration rate is much like last year's rate, there is a potential two-way causality problem. There are similar two-way causality issues relating crime and shall.

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

The most credible results are given by the two-way fixed effects model. The 95% confidence interval for shall contains the value of 0. Thus, there is no statistically significant evidence that concealed weapons laws have any effect on crime rates.

Practice Question 2: Stock-Watson Empirical Exercise 11.2

Believe it or not, workers used to be able to smoke inside office buildings. Smoking bans were introduced in several areas during the 1990s. Supporters of these bans argued that in addition to eliminating the externality of secondhand smoke, they would encourage smokers to quit by reducing their opportunities to smoke.

In this assignment, you will estimate the effect of workplace smoking bans on smoking, using data on a sample of 10,000 U.S. indoor workers from 1991 to 1993. The dataset contains information on whether individuals were or were not subject to a workplace smoking ban, whether the individuals smoked, and other individual characteristics.

a) Estimate the probability of smoking for (i) all workers, (ii) workers affected by workplace smoking bans, and (iii) workers not affected by workplace smoking bans.

Using a linear probability model:

```
mods.2a <- feols(smoker ~ csw0(smkban),
                 data = smoking,
                 se = 'HC1')
etable(mods.2a, fitstat = ~ r2 + ar2 + pr2 + f, markdown = T)
```

Dependent Variable:	smoker	
Model:	(1)	(2)
Variables		
Constant	0.2423***	0.2896***
	(0.0043)	(0.0073)
smkban		-0.0776***
		(0.0090)
Fit statistics		
R^2		0.00780
Adjusted R^2		0.00770
Pseudo R^2		0.00685
F-test	NaN	78.559

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

There is a 24% chance of smoking for all workers, 29.0% for workers affected by the ban and 0.290-0.08 = 21.2% for those affected by the ban

b) What is the difference in the probability of smoking between workers affected by a workplace smoking ban and workers not affected by a workplace smoking ban? Use a linear probability model to determine whether this difference is statistically significant.

We answered the first part above: 7.8 percentage points is the difference. This difference is significant as the associated p-value is less than 0.001.

c) Estimate a linear probability model with smoker as the dependent variable and the following regressors: smkban, female, age, age2, hsdrop, hsgrad, colsome, colgrad, black, and hispanic. Compare the estimated effect of a smoking ban from this regression with your answer from (b). Suggest an explanation, based on the substance of this regression, for the change in the estimated effect of a smoking ban between (b) and (c).

d) Test the hypothesis that the coefficient on smkban is 0 in the population version of the regression in (c) against the alternative that it is nonzero, at the 5% significance level.

e) Test the hypothesis that the probability of smoking does not depend on the level of education in the regression in (c). Does the probability of smoking increase or decrease with the level of education?

f) Repeat c-e using a probit model

g) Repeat c-e using a logit model

For the linear probability model:

```
smoking %<>% mutate(age2 = age^2)
mods2.lpm <- feols(smoker ~ csw0(smkban,
                                 female + age + age2 + hsdrop + hsgrad + colsome + colgrad + black + hisp
                   data = smoking,
                   se = 'HC1')
etable(mods2.lpm, fitstat = ~ r2 + ar2 + pr2 + f, digits = 6, markdown = T)
```

Dependent Variable:	smoker		
Model:	(1)	(2)	(3)
Variables			
Constant	0.242300***	0.289595***	-0.014110
	(0.004285)	(0.007262)	(0.041423)
smkban		-0.077558***	-0.047240***
		(0.008952)	(0.008966)
female			-0.033257***
			(0.008568)
age			0.009674***
			(0.001895)
age2			-0.000132***
			(2.19×10^{-5})
hsdrop			0.322714***
			(0.019488)
hsgrad			0.232701***
			(0.012590)
colsome			0.164297***
			(0.012625)
colgrad			0.044798***
			(0.012044)
black			-0.027566*
			(0.016078)
hispanic			-0.104816***
			(0.013975)
Fit statistics			
R^2		0.00780	0.05699
Adjusted R^2		0.00770	0.05605
Pseudo R^2		0.00685	0.05135
F-test	NaN	78.559	60.371

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

From model (3) the estimated difference is -0.047, smaller than the effect in model (2). Evidently (2) suffers from omitted variable bias. That is, smkban may be correlated with the education/race/gender indicators or with age. For example, workers with a college degree are more likely to work in an office with a smoking ban than high-school dropouts, and college graduates are less likely to smoke than high-school dropouts.

The p-value on the coefficient of smkban is less than 0.001 so the coefficient is statistically significant at the 1% level.

```
wald(mods2.lpm[[3]], keep = c('hsdrop', 'hsgrad', 'colsome', 'colgrad'))
```

```
## Wald test, H0: joint nullity of hsdrop, hsgrad, colsome and colgrad
##  stat = 140.1, p-value < 2.2e-16, on 4 and 9,989 DoF, VCOV: Heteroskedasticity-robust.
```

The p-value of the joint hypothesis test is less than 2.2e-16 so the coefficients are significant. The omitted education status is "Masters degree or higher." Thus the coefficients show the increase in probability relative to someone with a postgraduate degree. For example, the coefficient on Colgrad is 0.045, so the probability of smoking for a college graduate is 0.045 (4.5%) higher than for someone with a postgraduate degree. Similarly, the coefficient on HSdrop is 0.323, so the probability of smoking for a college graduate is 0.323

(32.3%) higher than for someone with a postgraduate degree. Because the coefficients are all positive and get smaller as educational attainment increases, the probability of smoking falls as educational attainment increases.

The coefficient on Age2 is statistically significant. This suggests a nonlinear relationship between age and the probability of smoking. In fact it is a negative quadratic with a probability-maximizing age of

```
-0.009674/(2*-0.000132)
```

```
## [1] 36.64394
```

For the probit and logit models:

```
mods2.probit<- feglm(smoker ~ csw0(smkban,
                                   female + age + age2 + hsdrop + hsgrad + colsome + colgrad + black + hisp
                     family = binomial(link = 'probit'),
                     data = smoking,
                     se = 'HC1')
etable(mods2.probit, fitstat = ~ r2 + ar2 + pr2 + f, digits = 6, markdown = T)
```

Dependent Variable:	smoker		
Model:	(1)	(2)	(3)
Variables			
Constant	-0.698923***	-0.554568***	-1.73493***
	(0.013712)	(0.021229)	(0.151675)
smkban		-0.244806***	-0.158630***
		(0.027872)	(0.029176)
female			-0.111732***
			(0.028866)
age			0.034511***
			(0.006855)
age2			-0.000468***
			(8.25×10^{-5})
hsdrop			1.14161***
			(0.073444)
hsgrad			0.882670***
			(0.060658)
colsome			0.677118***
			(0.061675)
colgrad			0.234683***
			(0.065572)
black			-0.084279
			(0.053828)
hispanic			-0.338274***
			(0.050155)
Fit statistics			
Pseudo R^2		0.00695	0.05441

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

```
mods2.logit<- feglm(smoker ~ csw0(smkban,
                                  female + age + age2 + hsdrop + hsgrad + colsome + colgrad + black + hi
                    family = binomial(link = 'logit'),
                    data = smoking,
                    se = 'HC1')
etable(mods2.logit, fitstat = ~ r2 + ar2 + pr2 + f, digits = 6, markdown = T)
```

Dependent Variable:	smoker		
Model:	(1)	(2)	(3)
Variables			
Constant	-1.14011***	-0.897351***	-2.99918***
	(0.023340)	(0.035298)	(0.265352)
smkban		-0.415340***	-0.262029***
		(0.047198)	(0.049501)
female			-0.190773***
			(0.049200)
age			0.059937***
			(0.011828)
age2			-0.000818***
			(0.000143)
hsdrop			2.01685***
			(0.134055)
hsgrad			1.57850***
			(0.115781)
colsome			1.22998***
			(0.117664)
colgrad			0.446583***
			(0.126394)
black			-0.156034*
			(0.091295)
hispanic			-0.597173***
			(0.086229)
Fit statistics			
Pseudo R^2		0.00695	0.05475

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

h) Predictions

i. Mr. A is white, non-Hispanic, 20 years old, and a high school dropout. Using the probit regression and assuming that Mr. A is not subject to a workplace smoking ban, calculate the probability that Mr. A smokes. Carry out the calculation again, assuming that he is subject to a workplace smoking ban. What is the effect of the smoking ban on the probability of smoking?

ii. Repeat for Ms. B, a female, black, 40-year-old college graduate.

iii. Repeat (i)–(ii) using the linear probability model.

iv. Repeat (i)–(ii) using the logit model.

```
preds <- data.frame(names = rep(c('Mr. A', 'Ms. B'), each = 2),
                    smkban = c(0, 1, 0, 1),
                    female = c(0, 0, 1, 1),
                    age = c(20, 20, 40, 40),
                    hsdrop = c(1, 1, 0, 0),
                    hsgrad = c(0, 0, 0, 0),
                    colsome = c(0, 0, 0, 0),
                    colgrad = c(0, 0, 1, 1),
                    black = c(0, 0, 1, 1),
                    hispanic = c(0, 0, 0, 0)) %>%
  mutate(age2 = age^2)

preds$probit.predictions <- predict(mods2.probit[[3]], newdata = preds)
preds$lpm.predictions <- predict(mods2.lpm[[3]], newdata = preds)
preds$logit.predictions <- predict(mods2.logit[[3]], newdata = preds)
preds.show <- select(preds, names, smkban, contains('predictions'))
preds.show
```

```
##   names smkban probit.predictions lpm.predictions logit.predictions
## 1 Mr. A      0          0.4641020      0.44937213         0.4723103
## 2 Mr. A      1          0.4017831      0.40213226         0.4078402
## 3 Ms. B      0          0.1436957      0.14596103         0.1405121
## 4 Ms. B      1          0.1107609      0.09872116         0.1117418
```

So differences for Mr. A and Ms. B respectively are

```
# Mr. A
preds$probit.predictions[2]-preds$probit.predictions[1]
```

```
## [1] -0.06231886
```

```
preds$lpm.predictions[2]-preds$lpm.predictions[1]
```

```
## [1] -0.04723987
```

```
preds$logit.predictions[2]-preds$logit.predictions[1]
```

```
## [1] -0.06447005
```

```
# Ms. B
preds$probit.predictions[4]-preds$probit.predictions[3]
```

```
## [1] -0.03293474
```

```
preds$lpm.predictions[4]-preds$lpm.predictions[3]
```

```
## [1] -0.04723987
```

```
preds$logit.predictions[4]-preds$logit.predictions[3]
```

```
## [1] -0.02877033
```

v. Based on your answers to (i)–(iv), do the logit, probit, and linear probability models differ? If they do, which results make most sense? Are the estimated effects large in a real-world sense?

They differ a bit but are pretty consistent with one another. The estimated effects are on the order of 6 percentage points for Mr. A and 3 percentage points for Ms. B, which is large.