Always Control for Year Effects in Panel Regressions!

Why is controlling for year effects important? Year effects (more simply known as "year dummies" or "dummies for each of the years in your dataset [excluding the first year]") capture the influence of aggregate (time-series) trends. As I have argued, time series data are not generally a meaningful source of causal inference because of the lack of compelling counterfactual. In fact, in many cases time series variables are spuriously related simply because of, say, the rising magnitude of aggregate variables (because of inflation, economic growth, population growth, etc.) For example, the U.S. time series relationship between two variables which almost certainly have no causal relationship -- the size of the trade deficit and health care spending -- produces an R-squared above 0.9!!! More generally, time series regressions – and therefore also panel regressions which fail to control for year effects -- pick up the influence of aggregate trends which have nothing to do with causal relationships.

Let us use the example of a city panel regression of crime rates on unemployment rates to see the potential issue of failing to control for year effects. I presented data on 46 U.S. cities in 1982 and 1987. Over this period, in the aggregate, average crime rates were rising and unemployment rates were falling. To see this, just look at the means of the city data by year:

	1982	1987
Crimes/1000 Population	97.71	103.87
	(23.91)	(34.78)
Unemployment Rate	10.05	5.89
	(3.45)	(1.51)
Sample Size (Cities)	46	46

Standard deviations in Parentheses.

Crime rates in the average city went from 97.7 to 103.8 per 1000 population between 1982 and 1987. Importantly, the rising crime rates in the aggregate are not thought to be causally tied to the falling unemployment rate (why would they be?) Instead, they were likely driven by demographic forces, rising inequality, or innovations in the drug trade. The massive fall unemployment rate (from 10% to 5.9%) is similarly not thought to be causally tied to rising crime rates, but due to business cycle forces (the end of a severe early-1980s recession).

Now let us use these data to run a panel regression of crime rates on unemployment rates, as we did in class, but including some specifications which fail to control for year effects (columns 1 and 3):

OLS Regressions of Crimes/1000 Popluation on Unemployment Rate

VARIABLES	Pooled Cross-Section No Year	Pooled Cross-Section w/Year	w/City Fixed Effects No Year	w/City Fixed Effects w/Year
	(1)	(2)	(3)	(4)
Unemployment Rate	-0.308	0.427	-0.0181	2.218
	(0.910)	(0.994)	(0.442)	(0.816)
Dummy for 1987		7.940		15.40
		(7.106)		(5.179)
Constant	103.2	93.42	100.9	75.41
	(8.642)	(10.46)	(3.882)	(8.916)
Observations	92	92	92	92
R-squared	0.001	0.012	0.864	0.891
City Fixed Effects?	N	N	Y	Y
Year Effects?*	N	Y	N	Y

Notes: Heteroskedasticity-Robust Standard errors in Parentheses. *"Year Effects" here really just means a dummy for 1987(!) since there are only two years of data, 1982 and 1987.

Notice that the regressions in columns 1 and 3, without year effects, have a negative coefficient on unemployment rate. This says higher unemployment rates are associated with lower crime rates![1] (Is this causal? Could we reduce crime rates just by firing people? You be the judge!) But this is an artifact of failing to control for year effects. Once we put in a control for year effects, the coefficient becomes (much more sensibly) positive. Put one way, the omitted variables bias from failing to control for year was negative, due to aggregate rising trends in crime and falling unemployment over the 1980s. This should be controlled for, because we do not want aggregate trends to influence our cross-city regression.

So….always control for year effects in panel regressions!

Another somewhat interesting thing is how much larger the R-squareds are in columns 3 and 4, which control for city fixed effects (city dummies). Fixed effects often capture a lot of the variation in the data. This often leads the standard errors to be larger, though that seems not to be true in this case.

[1] To be fair, neither coefficient is statistically significant. But the direction of bias should be clear.