ECON 8000/9000 Empirical Energy Economics
PS1: Replication Exercise: Cross-sectional and Panel-data Regression
Christy Zhou

Instruction and Notes:

- Deadline: Tuesday 1/20 @ 5pm
- Submission: Please compile all results into a single PDF.
- Notation: **[OptAll]** means optional for everyone. **[OptM]** means optional for master's students.
- Group: Everyone can work together, but only **up to two students** can submit the assignment together. If so, make sure you include both names on top when you submit.
- Here I write instructions using Stata. You can use other software if you like.
- This study has a replication package online. I recommend you follow my instructions below, which are slightly less complicated than the online package.

Q1. Replication of Severen and van Benthem (AEJ:AE 2022) Formative Experience and the Price of Gasoline
Q1.A Visualize the key temporal variation in fuel price:
Here we will reproduce Figure 1, which highlights the key variation used in this study.

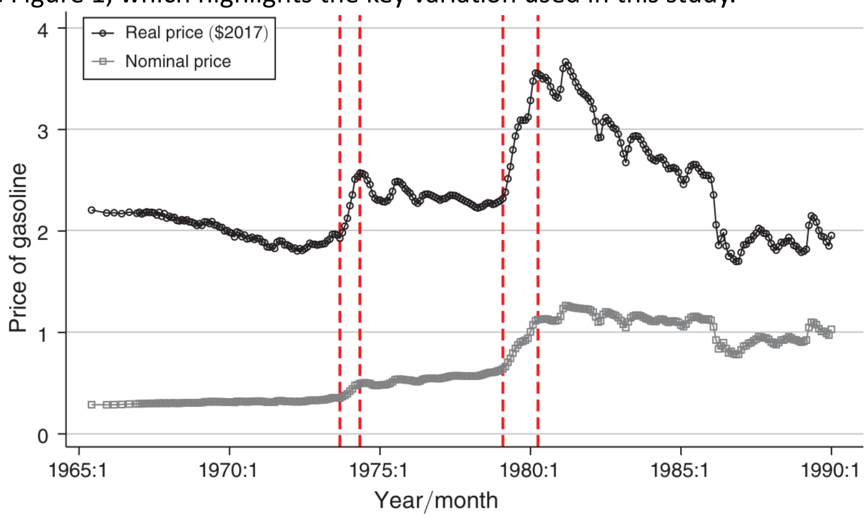

FIGURE 1. GASOLINE PRICES SPIKES IN THE UNITED STATES (1965–1990)

Info:

- Dataset: "PSdata_severen_venbenthem_data_gasprice_fig1.dta"
- Key vars: Nominal gas price: gas_price_nom and CPI: cpi

Task

- 1. Normalize gas price to 2017 USD as of January. Call it gas_price_2017.
- 2. Plot the above figure using twoway graph. You can use either scatter, connected, or line.

Q1.B Long-run Analysis Using Cross-Sectional Data: Effect of 1979 Oil Crisis on Future Driving Behavior
The initial test of whether the hypothesized effect exists is to examine the driving pattern in a (relatively) recent survey year and test whether the driving probability is affected by whether a person observed in the 2000 census was young enough to experience the oil price in 1979. Here, we will replicate Equation (1) with T_{cs} defined as a person who is 15 years or younger by 1980, and reproduce Appendix Table A.3 and A.4 Panel A. These regressions are essentially a linear-probability model (LPM), aka an OLS with the left-hand-side variable being a dummy.

$$(1) \qquad Y_i = \alpha + g(S_i) + \tau D_i + X_i' \lambda + \varepsilon_i,$$

Table A.3: Event study in turning 15 after 1979 on commuting behavior in 2000.

Model	Poly. order	Bandwidth (years) 2	3	4	5	6	7	8	9	10
Panel A: Effect on driving, no controls										
	1	-0.0050*	-0.0029+	-0.0026+	-0.0032**	-0.0026*	-0.0027**	-0.0032**	-0.0032**	-0.0029**
		(0.0022)	(0.0016)	(0.0014)	(0.0012)	(0.0011)	(0.0010)	(0.0009)	(0.0009)	(0.0008)

Table A.4: Event study in turning 15 after 1979 on transit usage and vehicle access in 2000.

Poly. order	Bandwidth (years) 2	3	4	5	6	7	8	9	10
Panel A: Transit usage									
1	0.0036*	0.0027*	0.0027**	0.0023**	0.0017*	0.0016*	0.0016**	0.0015**	0.0018**
	(0.0015)	(0.0011)	(0.0009)	(0.0008)	(0.0007)	(0.0007)	(0.0006)	(0.0006)	(0.0005)

Info:

- Dataset: Depending on the storage and computing capabilities of your machine, here I provide
  - Full sample "PSdata_severen_venbenthem_data_tableA3_ACS.dta", 1.3GB
  - 50% sample "PSdata_severen_venbenthem_data_tableA3_ACS_sample_50pct.dta", 480 MB
  - Qualitative result will be the same, though numbers will be a bit different
- Outcome Y_i:
  - For Appendix Table A.3, it is t_drive. It equals 1 if a person drives
  - For Appendix Table A.4 Panel A, t_transit. It equals 1 if a person uses pub transit
- Some vars:
  - Birth year: birthyr
  - Birthplace: blp
  - Demographics: sex, race, educd (educational attainment), hhincom (household income)
  - Individual person weight: perwt

Tasks:

- 1. Gen key right-hand-side vars:
  - Define treatment D as gen D = inrange(birthyr, 1965, .) // aka at most 15 in 1980
  - Define var S_i as the age in 1964, gen T = 1964 - birthyr
  - Then generate interaction gen DT = D*T
- 2. Redo Table A.3 and A.4 first row
  - First, run naïve version:
    - For column 1, run reg t_drive D if inrange(birthyr, 1964, 1966) [aw=perwt], robust
    - You can adjust 1966 in a loop to produce the remainder columns
  - Second, run a simple version that almost does Appendix Table A.3-A.4: Use birthyr as control for trend
    - For column 1, run reg t_drive D birthyear if inrange(birthyr, 1964, 1966) [aw=perwt], robust
    - Loop the rest
  - Third, run the actual spec authors did:
    - For column 1, run reg t_drive D T DT if inrange(birthyr, 1964, 1966) [aw=perwt], robust
    - Loop the rest
  - Explain why the pattern is different in the naïve version vs the versions when we control for the trend
  - Explain where the identification comes from for the regression produced in Table A.3-A.4
- 3. [OptAll] Redo Table A.3 and A.4 with demographic FE (race, sex, and educational attainment), log household income, and birthplace FE. I recommend using reghdfe with the FE absorbed.
  - In the paper, the authors use a non-white (as race>=2) dummy, but you can use race FE as well.
  - In the paper, the authors use a high school dummy (educd >=62) and a college dummy (educd >= 101), but you can use educational attainment FEs as well.
- 4. The authors call these "event studies". In a way, these are event studies, but only in a very specific way. Explain in what way these estimates are event studies.

- 5. [OptAll] When you do the replication exercise for Tasks in 1-2, if you can, produce the outputs in a single table.

Q1.C Panel Data Analysis: Effect of 1979 Oil Crisis on Future Driving Behavior

We will reproduce the main panel data regression results in Tables 1-2. Here, we will estimate Equation (2) for the same two outcomes that you did in part Q1.A.

$$(2) \qquad Y_{icst} = \theta T_{cs} + \kappa_s + \delta_t + \eta_{t-c} + X_{it}'\lambda + \varepsilon_{icst},$$

TABLE 1—THE EFFECT OF FORMATIVE GASOLINE PRICES ON DRIVING TO WORK

	1[drive] (1)	**1**[drive] (2)	**1**[drive] (3)	**1**[drive] (4)	**1**[drive] (5)	**1**[drive] (6)	**1**[drive] (7)
Exposure defined by age							
$P_{cs}^{\Delta 17,15}$	−0.0038	−0.0028	−0.0031	−0.0037	−0.0039	−0.0039	−0.0043
	(0.0010)	(0.0008)	(0.0009)	(0.0010)	(0.0010)	(0.0010)	(0.0009)

TABLE 2—THE EFFECT OF FORMATIVE GASOLINE PRICES ON OTHER CENSUS OUTCOMES

	Transit usage		Vehicle available			
	1[transit] (1)	**1**[transit] (2)	**1**[vehicle] (3)	**1**[vehicle] (4)	**1**[vehicle] (5)	**1**[vehicle] (6)
Exposure defined by age						
$P_{cs}^{\Delta 17,15}$	0.0029	0.0024	−0.0014	−0.0009	−0.0019	−0.0018
	(0.0007)	(0.0009)	(0.0008)	(0.0006)	(0.0009)	(0.0006)

Information

- Datasets: Depending on the storage and computing capabilities of your machine, here I provide
  - Original dataset: "PSdata_severen_venbenthem_data_table1_2_ACS.dta", 1.7GB
    - 50% sample: "PSdata_severen_venbenthem_data_table1_2_ACS_sample_50pct.dta" 850MB
    - 40% sample: "PSdata_severen_venbenthem_data_table1_2_ACS_sample_40pct.dta" 680MB
    - Note: Qualitative result will be the same if you use a subsample, but the numbers will be different.
  - The gas price data: "PSdata_severen_venbenthem_data_gasprice_tab1.dta"
- Outcome: We will focus on two outcomes: t_drive and t_transit.
- Some vars:
  - Stay in the same state (or employed in the same state): m_samestate == 1
  - All demographics can be found in d_*, including d_fem d_marr d_hs d_col d_black d_hisp.
  - For income, use hhincome or w_hhi.
  - Vars used in exercise in part B: blp, perwt
  - Census year: censusyear_all and state: stfips
  - State-by-CensusYear FE, generate using egen group from stfip and censusyear_all

Tasks

- 1. Create the key right-hand side variable $\Delta P^{15,17}$, denoted as d2gp_bp_age17
  - First, use the "PSdata_severen_venbenthem_data_gasprice_tab1.dta".
    - Do a tsset. Then create a 2-year difference relative to a base 2-year lag,
    - gen d2gp_bp = (gas_price_99-L2.gas_price_99)/L2.gas_price_99
    - Save this dataset. We will later merge into the year when a person turns 17
  - Second, use the census dataset, create year = birthyr + 17. Then merge the dataset you created above so that your main dataset has the main variable of interest d2gp_bp. You can rename it as d2gp_bp_age17.
  - [OptAll] If you are also interested in reproducing row 2 of Tables 1-2, then merge in the real gas price (gas_price_99) when a person turns 16, aka, create P^{16}. Hint: You would want to re-create year = birthyr + 16, merge into the same dataset, and create gas_price_99_age16.

- 2. Reproduce Table 1, first row, columns 1, 4, 5, and 6
- 3. Reproduce Table 2, first row, column 1
- 4. [OptAll] Reproduce Table 1, second row, columns 1, 4, 5, and 6
- 5. [OptAll] Reproduce Table 2, second row, column 1
- 6. Explain where the variation comes from

Q1.D [OptM] Your turn: What could you do with a similar dataset or a similar design?

This research builds on a publicly available dataset of individuals and leverages the variation in birth cohorts. Typically, repeated cross-sectional data, such as ACS-PUMS, have their limitations; however, a research design like this doesn't require a longitudinal data structure by utilizing variation in birth year.

- Could you think of some other questions that you can answer with a design in a similar style? It does not need to be limited to an energy-related topic, although that will be encouraged.
- You can brainstorm using ACS or other publicly available datasets.
- If you can, list 1-3 ideas/designs just as practice and explain what your design will allow you to estimate.