Problem Set 2

PPPA 8022
Due in class, on paper, February 13

- Please use a do-file (or its SAS or SPSS equivalent) for this work. Do not program interactively. It may seem faster at first, but it is inevitably slower.
- Turn in a typed up set of answers that answers the questions below. Also turn in a Stata .do file and its associated .log file.
- Make formal tables to present your results. Do not present statistical software output.
- I have provided Stata datasets, but you should feel free to do the analysis in whatever software you prefer. If you need to transfer to another format, use StatTransfer.
- This problem set uses some large data. For the Census data, I have posted full dataset as well a smaller version; use whichever you prefer. For the CPS, we are using a random sample.
- If the question is insufficiently clear, explain the assumptions you made to reach your final estimates.

1. Fixed Effects

For this problem, well use Decennial Census/American Community Survey data from IPUMS-USA for 1950 and 2010 (for 2010, the 1-year ACS). Data are linked on the handouts page. The large versions, which are one file per year, have the years in the title (1950 and 2010); the small version is ipumscen.dta.zip.

For purposes of **this problem set only** we will not use any survey-defined weights. This is totally wrong and you should never do it when you really analyze a dataset. We are doing it here so that 1(b) does not become extremely difficult.

The IPUMS website is `https://usa.ipums.org/usa/`, and it provides detailed information on the datasets and variables.

Let's examine the effect of education on wages.

(a) Start by finding the average wage (incwage) of prime age men (25 to 64) in 1950 and 2010. Test whether these average wages are significantly different in 1950 and 2010, and present these results in a well-labeled table. Beware of missing values and use a regression for the test. Write a few sentences to interpret your table.

See Tables **??** and **??**. The first table presents the result of a regression of `incwage` on $1\{year = 2010\}$. We find that in both the large and small samples that wages are higher in 2010: \$40,790 higher in the small sample and \$41,061 higher in the large sample. Not a shock!

Table **??** reports the equivalent difference in means. Note that the difference in means **is equivalent** to the difference you find in the regression. If it's not, you've done something wrong.

(b) Re-create the means in (a) by creating grouped data, and then properly weighting those grouped data to return to the mean in part (a) via regression (I am asking you to do something like *MHE* Table 3.1.3, but present in a table, rather than stata output). The grouping variable is your choice. I used education (so, average wages by years of education), but you could use age or any other variable for which you can create discrete categories. Present in a table, and write a few sentences explaining the logic of your answer.

First, a little explanation on the general issue of weighted averages. Suppose we have a dataset with N people, each denoted i. We observe wages, w_i. People further belong to group j, there are J groups in total, and each group has N_j people.

We can write the average wage as

$$\frac{\sum_{i=1}^{N} w_i}{N}$$

We can write the average for any group j as

$$\frac{\sum_{k=1}^{N_j} w_k}{N_j}$$

Any group j is a share of the total population, and we can write that share as N_j/N. Given this information, we can write a weighted total of the group averages as

$$\left(\frac{N_1}{N}\frac{\sum_{k=1}^{N_1} w_k}{N_1}\right) + \left(\frac{N_2}{N}\frac{\sum_{k=1}^{N_2} w_k}{N_2}\right) + \ldots + \left(\frac{N_J}{N}\frac{\sum_{k=1}^{N_J} w_k}{N_J}\right)$$

Note that this equation simplifies to the first equation for average wage.

There is nothing to show if you've done this correctly – the answer is the same as in (a).

(c) Make the wages in both surveys into constant 2013 dollars. Use the all urban consumers series from the Bureau of Labor Statistics (`http://www.bls.gov/cpi/data.htm`, and use the "all urban consumers" row, and use the "all items" series; using the December inflation number for each year is sufficient). There is no output to report for this step, but you may wish to do some checks for yourself to make sure your results are reasonable.

Nothing to report here, but your Stata code should reflect this work.

(d) Suppose we would like to know whether the average husband earns higher real wages than the average wife. Use a regression to estimate wages as a function of age, year, and being the husband (think about what sample you should use to do this, and explain what sample you chose. Make sure you only keep working age people.). Then re-estimate with a variety of sensible covariates. Then re-estimate with the covariates and family fixed effects (in Stata, I highly recommend areg). Then re-estimate to allow the effect of being a husband to vary between 1950 and 2010. Present all results in one table and interpret the coefficients in each regression, explaining why they change.

See Table **??** (small sample) and Table **??** (big sample).

I keep only observations where the person is married and the spouse is present. I limit the sample to ages 25 to 65, to be sure that people could be in the labor force.

In the first column, husbands earn substantially more than wives, and people earn more in 2010. Earnings seem to decline with age (but that is because we didn't also include age squared). Adding a variety of sensible covariates in the second column barely budges the husband result, though it affects the age and year coefficients.

Adding family fixed effects, as in the third column, soaks up a fair amount of the variation (look at the R-squared). It shrinks the husband coefficient, but it remains quite sizeable just a little less than the average wage.

The fourth column tells us that the effect has declined substantially over time – the interaction of being the husband and being 2010 is negative, and about half the average husband premium.

(e) The previous estimation included age linearly. Use two methods to relax this assumption and report the results in a table. Write a few sentences that interpret the results. Explain which method you prefer and why.

I relaxed the (crazy) linear assumption for age by including age, age^2, age^3, and age^4 in the fifth column, and then including age dummies in the sixth column. The difference between these two is enough to make a small difference in the coefficients of interest. The coefficient for age^4 was too small to be reported. This is not good practice! It is better to re-scale age^4 (to something like age$^4 * 1000$) rather than not report the coefficient.

See Table **??** for the small sample and Table **??** for the large sample.

2. Difference-in-difference

Now let's use the IPUMS CPS. The small sample (a random sample of the full dataset) is called ipumscps.dta.zip and is linked on the handouts page. Documentation for this dataset is available at `https://cps.ipums.org/cps/`. For the purposes of this problem set, treat each observation with equal weight. This is entirely wrong, and you should absolutely never do such a thing if you are doing a real project. Finally, beware of top-coded data!

(a) Pretend that MI, CA, AZ, NM, MN, OH, VA, KY, WV, MO, MS, GA, IA, NH, MA and ME all adopt a policy aimed at increasing wages that takes effect in 2000. For simplicity, we focus only on employed people. We hypothesize that treatment is random conditional on age and race. Use a figure to examine the parallel pre-trend assumption (the unconditional outcome, not conditional on covariates), and show this figure (note that making a legible picture may require some summary of the data; think about the best way to summarize the data). Use the variable incwage for annual wages. Write a few sentences that interpret the figure.

See Figure **??** at the end. I don't see any compelling difference between the two groups pre-treatment (in the pre-2000 era). I've added gray bands that show the 95% confidence intervals for the means. This is Stata command `rarea`, and it can be a very helpful way to show a lot of information.

(b) Use a regression to test whether the treated and untreated states have similar trends before the treatment is adopted, conditional on covariates. Report the results in a table, and write a few sentences that interpret the results of your test.

To do this, you should use only the pre-treatment data. I tested the equality of trends in two different ways:

$$\text{incwage}_{i,t} = \beta_0 + \beta_1 \text{trend}_t + \beta_2 \text{trend}_t * \text{treatment}_{i,t} + \epsilon \tag{1}$$

$$\text{incwage}_{i,t} = \beta_0 + \beta_1 \text{time}_t + \beta_2 \text{time}_t * \text{treatment}_{i,t} + \epsilon \tag{2}$$

The variable treatment is 1 if the state is ever treated, trend is a linear trend variable (1960=1. 1961=2, etc; though the exact number for each year is not consequential for the slope, only the intercept), and time is a full set of year dummy variables (that is, fixed effects).

We test $H_0 : \beta_2 = 0$. For Equation **??**, all we need is a t-test for whether $\beta_2 = 0$. I find a t-value = 9.9/7.4 = 1.3, so we cannot reject $\beta_2 = 0$. For Equation **??**, we want to know whether all the β_2 are jointly 0. We do the second with an F test ($H_0 : \beta_{2,1963} = \beta_{2,1964} = \ldots = \beta_{2,1999} = 0$). We cannot reject the hypothesis that the year*treatment coefficients are jointly zero.

```
    .  local testvals _IyeaXtre_1963;

    .  forvalues y=1964/1999
> ;
2.  local testvals ‘testvals’ = _IyeaXtre_‘y’;
```

```
3.  ;

   .  test ‘testvals’ = 0;

   ( 1) _IyeaXtre_1963 - _IyeaXtre_1964 = 0
( 2) _IyeaXtre_1963 - _IyeaXtre_1965 = 0
( 3) _IyeaXtre_1963 - _IyeaXtre_1966 = 0
( 4) _IyeaXtre_1963 - _IyeaXtre_1967 = 0
( 5) _IyeaXtre_1963 - _IyeaXtre_1968 = 0
( 6) _IyeaXtre_1963 - _IyeaXtre_1969 = 0
( 7) _IyeaXtre_1963 - _IyeaXtre_1970 = 0
( 8) _IyeaXtre_1963 - _IyeaXtre_1971 = 0
( 9) _IyeaXtre_1963 - _IyeaXtre_1972 = 0
(10) _IyeaXtre_1963 - _IyeaXtre_1973 = 0
(11) _IyeaXtre_1963 - _IyeaXtre_1974 = 0
(12) _IyeaXtre_1963 - _IyeaXtre_1975 = 0
(13) _IyeaXtre_1963 - _IyeaXtre_1976 = 0
(14) _IyeaXtre_1963 - _IyeaXtre_1977 = 0
(15) _IyeaXtre_1963 - _IyeaXtre_1978 = 0
(16) _IyeaXtre_1963 - _IyeaXtre_1979 = 0
(17) _IyeaXtre_1963 - _IyeaXtre_1980 = 0
(18) _IyeaXtre_1963 - _IyeaXtre_1981 = 0
(19) _IyeaXtre_1963 - _IyeaXtre_1982 = 0
(20) _IyeaXtre_1963 - _IyeaXtre_1983 = 0
(21) _IyeaXtre_1963 - _IyeaXtre_1984 = 0
(22) _IyeaXtre_1963 - _IyeaXtre_1985 = 0
(23) _IyeaXtre_1963 - _IyeaXtre_1986 = 0
(24) _IyeaXtre_1963 - _IyeaXtre_1987 = 0
(25) _IyeaXtre_1963 - _IyeaXtre_1988 = 0
(26) _IyeaXtre_1963 - _IyeaXtre_1989 = 0
(27) _IyeaXtre_1963 - _IyeaXtre_1990 = 0
(28) _IyeaXtre_1963 - _IyeaXtre_1991 = 0
(29) _IyeaXtre_1963 - _IyeaXtre_1992 = 0
(30) _IyeaXtre_1963 - _IyeaXtre_1993 = 0
(31) _IyeaXtre_1963 - _IyeaXtre_1994 = 0
(32) _IyeaXtre_1963 - _IyeaXtre_1995 = 0
(33) _IyeaXtre_1963 - _IyeaXtre_1996 = 0
(34) _IyeaXtre_1963 - _IyeaXtre_1997 = 0
(35) _IyeaXtre_1963 - _IyeaXtre_1998 = 0
(36) _IyeaXtre_1963 - _IyeaXtre_1999 = 0
(37) _IyeaXtre_1963 = 0
```

```
   F( 37,230832) = 0.71
Prob > F = 0.9093
```

These regressions each use 231,066 observations – less than the full dataset, since they omit data before 2000.

(c) Do a difference-in-difference regression to examine the effects of this policy on wages. Write the estimating equation you use. Start with a simple summary statistics table (with standard errors) that shows both differences and then the difference in difference (as in the Milligan paper). Now run a regression that is parallel to this estimate. Present both results in one or two tables and interpret the magnitudes in a few sentences.

See summary statistics in Table **??** below. We find significant single differences between the treated and untreated, before and after. The double difference is also significant (t=4.77). Wages declined by about $750 in the treated states, relative to the untreated ones, after the treatment.

To do the regression, I estimate the following equation:

$$\text{incwage}_{ist} = \beta_0 + \beta_1\text{time}_t + \beta_2\text{state}_s + \beta_3\text{treatment*after}_{ist} + \beta_4\text{age}_t + \beta_5(race)_i + \epsilon_{ist}$$

In general, a difference-in-difference estimating equation should include both parts of the interaction that yields the effect of interest (the rule is actually broader: you should always include both parts of an interaction separately.) Note that “after” can be created by adding time fixed effects, and that these time fixed effects are less restrictive than the specification with simply an “after” indicator. Similarly, we don’t need to include a separate “treatment” indicator, since a linear combination of state fixed effects yields the treatment indicator. Results are in column 1 of Table **??** below.

The regression, controlling for age, race, state and year, finds an insignificant $61 dollar decrease in earnings due to this fake policy.

(d) Explain and implement one method to correct the results from part (c) for serial correlation. For simplicity, do not use covariates. Describe your method, present your results in a table (or add to the previous table) and write a few sentences that explain what you find.

The simplest method for assessing the importance of serial correlation in (c) is to average the values of the pre- and post-treatment years and re-do the regressions (aka, shrink T to 2). Because we don’t observe each person for all years, we also need to collapse to the state level, so well have 2*51 observations. (You might think about doing something at a lower geographic level, say the county, but given the sample size I went ahead and averaged to the state level.)

In the early years of the CPS in the 1970s, it seems that they have strange state categories - state combinations, instead of states by themselves. So I use data after 1976 only, and have 102 observations (states plus DC); see Column 3 of Table **??**. This method finds no significant difference in the treated states after the treatment.

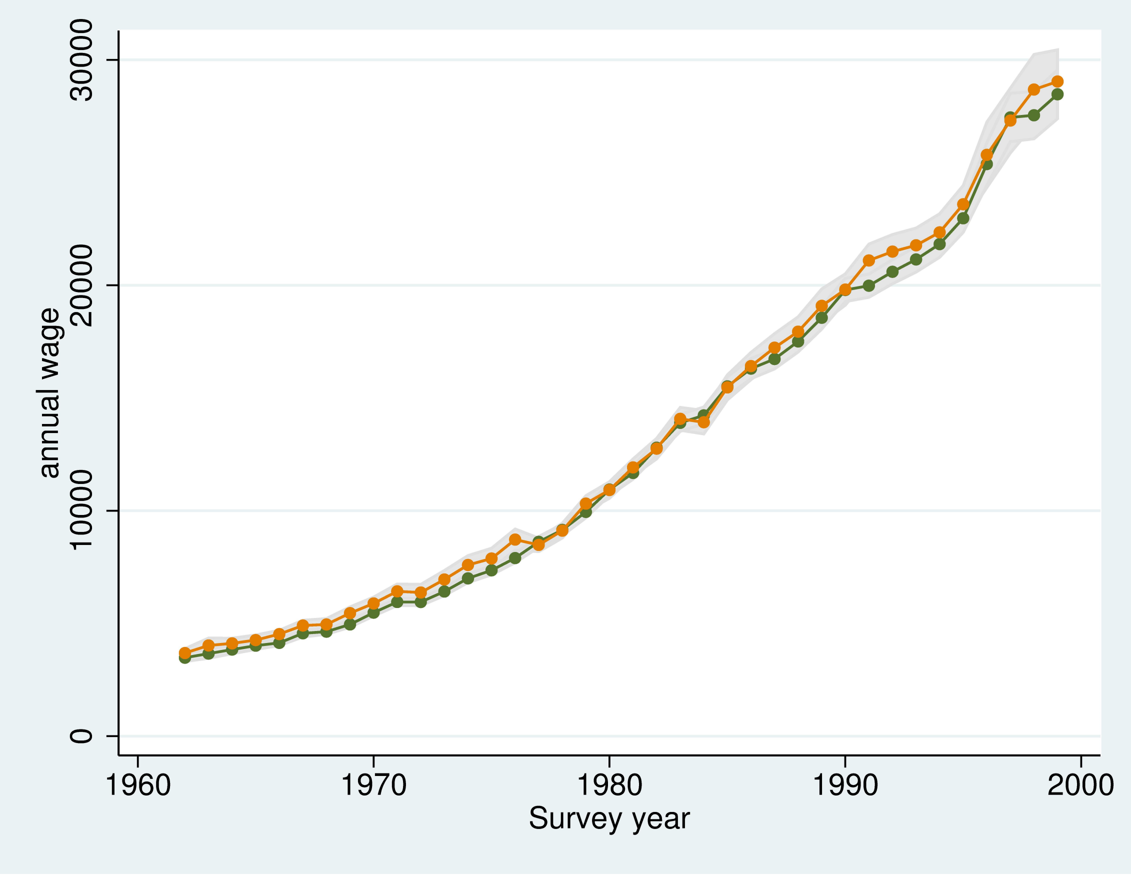

Figure 1: Parallel Pre-Trend Assumption

Table 1: Answer for 1(b)

Wage Type	Statistic	1950	2010	t-test for difference of means
A. Small Sample				
Nominal	mean	2181	42,970	186.0
	std error	18.7	201	
Real $2013	mean	20,477	45,849	65.2
	std error	175.3	214.1	
Observations		11,145	78,629	
B. Full Sample				
Nominal	mean	2,182	43,243	587.5
	std error	5.8	64.1	
Real $2013	mean	20,490	46,141	208.3
	std error	54.8	68.3	
Observations		111,680	787,469	

Table 2: Regression to Test Difference in Wages

	Small Sample	Big Sample
Male	40,790	41,061
	(533.0)	(170.2)
Observations	89,774	899,149

Table 3: Answer for 1 (d) and (e), Small Sample

	Only age, year, husband	With sensible covariates	With family fixed effects	With family FE, allowing main effect to vary	Parametric non-linear age	Non-parametric non-linear age
Age	-135.7***	-73.0***	-88.8***	-89.3***	-5871.3	
	(13.5)	(12.7)	(13.8)	(13.8)	(4075.2)	
Male	26520.0***	26485.8***	26772.5***	17527.5***	17508.3***	17486.5***
	(293.8)	(273.3)	(292.7)	(761)	(754.5)	(754.6)
1{year is 2010}	29338.3***	13711.5***				
	(399)	(479.4)				
Male*1{year is 2010}				10842.5***	10761.3***	10785.7***
				(824)	(817)	(817.1)
Age^2					215.6	
					(141.9)	
Age^3					-2.6	
					(2.1)	
Age^4					0	
					(0)	
Education FE		x	x	x	x	x
Race FE		x	x	x	x	x
Metro type FE		x	x	x	x	x
Age FE						x
R-squared	0.105	0.227	0.342	0.344	0.355	0.355
Observations	115489	115489	115489	115489	115489	115489

Table 4: Answer for 1 (d) and (e), Large Sample

	Only age, year, husband	With sensible covariates	With family fixed effects	With family FE, allowing main effect to vary	Parametric non-linear age	Non-parametric non-linear age
Age	-131.1***	-75.9***	-80.7***	-80.7***	-4186.1***	
	(4.3)	(4)	(4.1)	(4.1)	(1198.5)	
Male	26745.3***	26563.2***	26609.7***	17657.0***	17744.2***	17740.3***
	(93.3)	(86.8)	(86.7)	(215.7)	(213.7)	(213.7)
1{year is 2010}	29658.1***	13996.0***				
	(126.6)	(152.3)				
Male*1{year is 2010}				10669.2***	10519.7***	10524.7***
				(235.4)	(233.2)	(233.2)
Age^2					144.2***	
					(41.8)	
Age^3					-1.3*	
					(0.6)	
Age^4					0	
					(0)	
Education FE		x	x	x	x	x
Race FE		x	x	x	x	x
Metro type FE		x	x	x	x	x
Age FE						x
R-squared	0.106	0.228	0.244	0.245	0.259	0.259
Observations	1152661	1152661	1152661	1152661	1152661	1152661

Table 5: Answer for 2(c)

		Untreated	Treated	Difference	Difference-in-difference
Before	mean	13617.1	14968	1350.9	
	se	44.5	69.1	52.9	
	obs	156871	74195		
After	mean	37620.8	38226.6	605.8	-745.1
	se	161.8	231.2	165.2	156.5
	obs	88091	45422		

Table 6: Regressions, Problem 2

	Full Sample		State-year obs, 2 time periods
	DD	DDD	DD
	(1)	(2)	(3)
1{treatment}*1{after}	-61.3	3735.8***	436.2
	(233.5)	(323.1)	(1098.7)
1{treatment}*1{after}*1{male}		-6684.4***	
		(422.1)	
1{male}*1{after}		16604.9***	
		(207.2)	
1{male}*1{treatment}		7313.7***	
		(228.2)	
1{after}			19780.2***
			(615.4)
Age fixed effects	x	x	
Race fixed effects	x	x	
State fixed effects	x	x	x
Year fixed effects	x	x	
R-squared	0.19	0.213	0.973
Observations	364579	364579	102

Table 7: Answer for 2(d)

				Differences		
		Before	After	Single	Double	Triple
Treated						
Men	mean	17824.8	46577.6	28752.8		
	sd	102.7	382.8	368.8		
	obs	42434	23813			
	variance	10538.85	146538.8			
Women		11151.2	29023.9	17872.7	10880.1	
		80.4	225.3	210.5	302.8	
		31761	21609			
		6463.2	50754.2			
Untreated						
Men		16175.5	45655.1	29479.7		
		66.9	266.8	258.2		
		90262	45948			
		4481.2	71163.4			
Women		10150.3	28861	18710.8	10768.9	111.2
		49.5	162	154.2	207.1	220.9
		66609	42143			
		2446.2	26232.1			