Problem Set 1

PPPA 8022
Due in class, on paper, February 11, 2015

Some overall instructions:

- Please use a do-file (or its SAS or SPSS equivalent) for this work – do not program interactively!
- 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.
- Make formal tables to present your results – don't use statistical software output.
- This problem set uses some large data. For the Census data, I have put the full dataset up on Blackboard, and I've also put a smaller version. For the CPS, only the small one would fit.
- If the question is insufficiently clear, explain what assumptions you made to reach your final estimates.

1. Fixed Effects

For this problem, we'll use Decennial Census/American Community Survey data from IPUMS-USA for 1950 and 2010 (for 2010, the 1-year ACS). Data are available on Blackboard. The large versions, once for each year, have the years in the title (1950 and 2010); the small version is ipumscen.dta.zip. Note that analysis using the 1950 sample must use weights (perwt); for simplicity (if not correctness), please use Stata's aweights or the equivalent.

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 wages differ significantly across time, and present these results in a well-labeled table. Beware of missing values.

See answer in (c)

(b) Show that you can replicate these results by using grouped data, properly weighted. (For an example, see MHE Table 3.1.3.) You can use education as a grouping if you'd like; choose another variable if you'd prefer.

Please see the separate pdf on this question (I preferred to write that file in a different typesetting system, since it has a lot of math!).

(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). Update your table with these real wages.

My final table looks like

Wage Type	Statistic	1950	2010	t-test for difference of means
A. Small Sample				
Nominal	mean	2,100.4	41,365.9	184.8
	std error	18.8	193.7	
Real $2013	mean	19,723	44,137	63.8
	std error	176.3	206.6	
Observations		11,145	78,629	
B. Full Sample				
Nominal	mean	2,089	41,507	585.3
	std error	5.8	61.5	
Real $2013	mean	19,613	44,288	204.8
	std error	54.9	65.6	
Observations		111,680	787,469	

Both wages differ significantly. Accounting for inflation gets rid of somewhat less than half of the difference.

I use a t-test that does not assume equal variance for the two samples.

(c) Suppose we would like to know whether husbands earn higher real wages than wives. 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 main effect to vary between 1950 and 2010. Present these results in a table.

Interpret the main result as the specification changes across the table. Is the specification with family fixed effects superior? Why? or why not?

See table in (d).

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.

(d) The previous estimation included age linearly. Use two methods to relax this assumption. Interpret the results. Which method do 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; it is better to re-scale age^4 rather than not report the coefficient (like I do below).

Small Sample Results

	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	
	(13.50)	(12.70)	(13.80)	(13.80)	(4075.20)	
Male	26520.0***	26485.8***	26772.5***	17527.5***	17508.3***	17486.5***
	(293.80)	(273.30)	(292.70)	(761.00)	(754.50)	(754.60)
1{year is 2010}	29338.3***	13711.5***				
	(399.00)	(479.40)				
Male*1{year is 2010}				10842.5***	10761.3***	10785.7***
				(824.00)	(817.00)	(817.10)
Age^2					215.6	
					(141.90)	
Age^3					-2.6	
					(2.10)	
Age^4					0	
					0.00	
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	115,489	115,489	115,489	115,489	115,489	115,489

Large Sample Results

	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***	
	(4.30)	(4.00)	(4.10)	(4.10)	(1198.50)	
Male	26745.3***	26563.2***	26609.7***	17657.0***	17744.2***	17740.3***
	(93.30)	(86.80)	(86.70)	(215.70)	(213.70)	(213.70)
1{year is 2010}	29658.1***	13996.0***				
	(126.60)	(152.30)				
Male*1{year is 2010}				10669.2***	10519.7***	10524.7***
				(235.40)	(233.20)	(233.20)
Age^2					144.2***	
					(41.80)	
Age^3					-1.3*	
					(0.60)	
Age^4					0	
					0.00	
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	1,152,661	1,152,661	1,152,661	1,152,661	1,152,661	1,152,661

2. Difference-in-difference

Now let's use the IPUMS –CPS. I've put this on Blackboard, but only the small sample, called ipumscps.dta.zip. 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 will 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.*

See Figure 1 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. 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:
(1) incwage = b0 + b1 trend + b2 trend*treatment + e
(2) incwage = b0 + b1 time + b2 time*treatment + e
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 (aka fixed effects).

We test H0: b2=0. For equation (1), all we need is a t-test for whether b2=0. I find a t-value = 9.9/7.4 = 1.3, so we cannot reject b2=0. For equation (2), we want to know whether all the b2 are jointly 0. We do the second with an F test (H0: b2,1963 = b2,1964 = ... = b2,1999 = 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 table (with standard errors) that does the same analysis, and then do a regression. What are the results? Do the two methods yield similar findings?

The summary table is 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.

		Untreated	Treated	Difference	Difference-in-difference
Before	mean	13,617.1	14,968.0	1,350.9	
	se	44.5	69.1	52.9	
	obs	156,871	74,195		
After	mean	37,620.8	38,226.6	605.8	-745.1
	se	161.8	231.2	165.2	156.5
	obs	88,091	45,422		

To do the regression, I estimate the following equation:
incwage = b0 + b1 **time** + b2 **state** + b3 treatment*after + b4 **age** + b5 **race** + e

Bolded variables are vectors. Note that "after" is subsumed by time, which is less restrictive than "after" would be. Similarly, we don't need to include a separate "treatment" indicator, since the state fixed effects add up to the treatment indicator. Results are in column 1 in the 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) Now suppose that the policy targeted only men. This suggests a triple difference estimation strategy. Write the estimating equation. Make a simple table that does this triple difference, and then do a regression that does the same.

The summary table is below.

				Differences		
		Before	After	Single	Double	Triple
Treated						
Men	mean	17,824.8	46,577.6	28,752.8		
	sd	102.7	382.8	368.8		
	obs	42,434	23,813			
	variance	10,538.9	146,538.8			
Women		11,151.2	29,023.9	17,872.7	10,880.1	
		80.4	225.3	210.5	302.8	
		31,761	21,609			
		6,463.2	50,754.2			
Untreated						
Men		16,175.5	45,655.1	29,479.7		
		66.9	266.8	258.2		
		90,262	45,948			
		4,481.2	71,163.4			
Women		10,150.3	28,861.0	18,710.8	10,768.9	111.2
		49.5	162.0	154.2	207.1	220.9
		66,609	42,143			
		2,446.2	26,232.1			

To do the regression, I estimate the following equation:
incwage = b0 + b1 **time** + b2 **state** + b3 treatment*after*male + b4 **age** + b5 **race** + b6 male + b7 after*male + b8 male*treatment + b9 after*treatment + e

The second column in Table 1 presents results from this regression. The triple difference yields a significantly significant $6,500 dollar decrease in income for men, relative to women, in states with treatment, after the treatment.

(e) Explain and implement one method to correct the results from part (c) for serial correlation. For simplicity, ignore the covariates. Describe your method and present your results.

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 we'll 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 1. This method finds no significant difference in the treated states after the treatment.

Figure 1: Testing the Equal Pre-trend Assumption

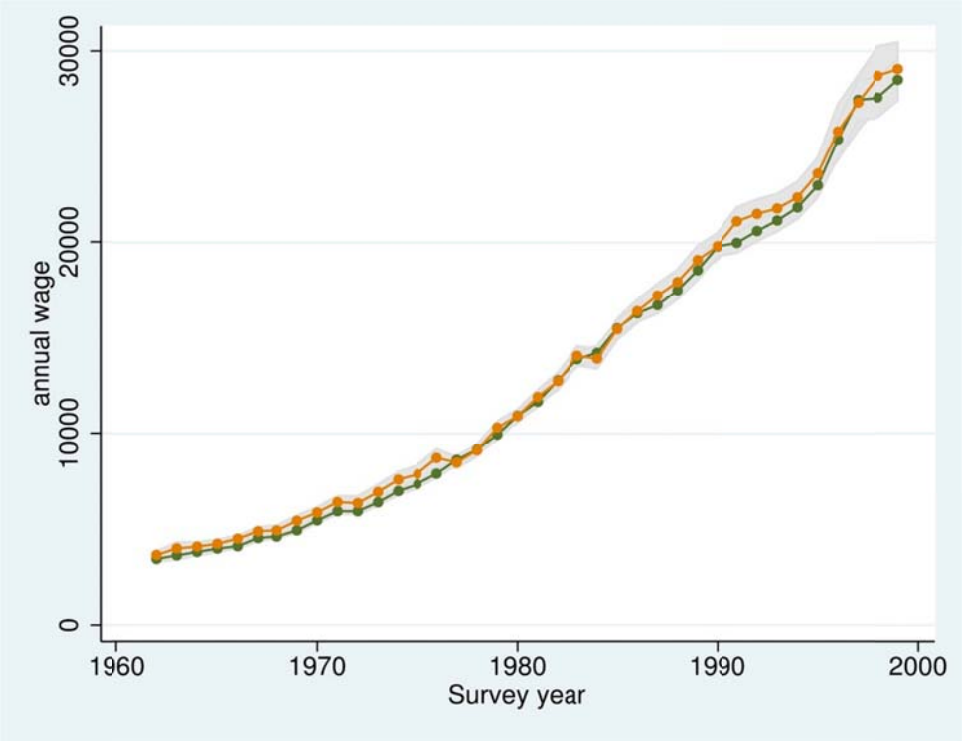

Table 1

	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	0.190	0.213	0.973
Observations	364,579	364,579	102