C528 Data Predictive Analytics and Business Strategy
Summer 2018
Homework 3, Team Purple

Benjamin Frazier
Sathya Thiruvengadam
Bassi Sidibe
Zach Toelle

Problem 1

1.a

The results of the difference-in-differences regression on total employment (full time employees, amangers, and ½ times part-time employees) between New Jersey and Pennsylvania is shown below.

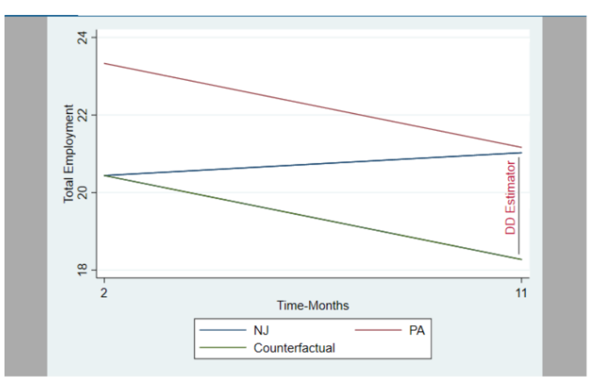
	(1) totalemploy
state	-2.892
	(1.194)
time	-2.166
	(1.516)
statetime	2.754
	(1.688)
_cons	23.33
	(1.072)

Standard errors in parentheses

Our result is difference-in-differences estimator of 2.75, which means that in this case a state that increased minimum wage saw relative increase in total employment. This contradicts the theoretical prediction of less employment given a rising minimum wage.

1.b

The graph below shows the changes in total employment for both states, as well as a line for the counterfactual. The difference between the New Jersey (NJ) line and the counterfactual represents the difference-in-differences estimator.



1.c

The results of the difference-in-differences regression on the price of a full meal (entrée, fries, and soda) between New Jersey and Pennsylvania is shown below.

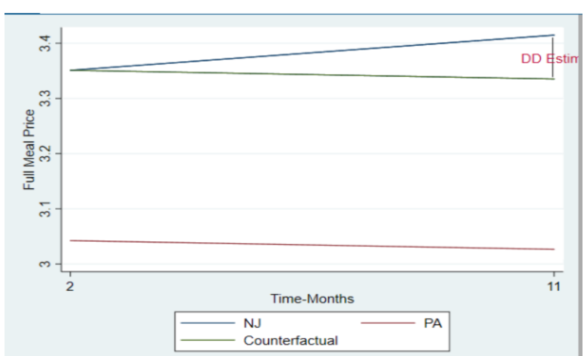
	(1)
	fmeal
state	0.309
	(0.0806)
time	-0.0157
	(0.104)
statetime	0.0794
	(0.116)
_cons	3.042
	(0.0723)

Standard errors in parentheses

Our result is a difference-in-differences estimator of .08, which means that in this case a state that increased minimum wage saw relative increase in the price of a full meal. This may imply that the increased cost imposed on the restaurants were passed on, at least in part, to the consumer.

1.d

The graph below shows the changes in the price of a full meal, as well as a line for the counterfactual. The difference between the New Jersey (NJ) line and the counterfactual represents the difference-in-differences estimator.



Problem 2

The following tables shows the results of models that will be discussed throughout Problem 2.

	(1) Pooled OLS	(2) Within Est	(3) First Diff
concen	-0.488	0.169	
	(0.0319)	(0.0294)	
y98	0.0287		
	(0.0177)		
y99	0.0313		
	(0.0177)		
y00	0.0906		
	(0.0177)		
diffconcen			0.171
			(0.0527)
_cons	5.355	4.953	0.0978
	(0.0232)	(0.0183)	(0.00564)
*.year	No	Yes	No

Standard errors in parentheses

2.a

The results of the pooled OLS model on on air routes between cities in 1998, 1999, and 2000 is shown in the first column of the table of the table above.

In this model, it is estimated that for every percent increase in market share of the largest carrier on the route, the fare will decrease by 48.8%. Since the pooled OLS model ignores unobserved heterogeneity, omitted factors may be biasing the estimate of the market concentration coefficient.

2.b

The results of the model using fixed effects with the within estimator on air routes between cities in 1998, 1999, and 2000 is shown in the second column of the table above.

In this model, it is estimated that for every percent increase in market share of the largest carrier on the route, the fare will increase by 16.9%. This result is more intuitive than the pooled OLS model, as you would expect fares to increase as market share of the largest carrier rises on a certain route.

2.c

The results of the first differences model on on air routes between cities in 1998, 1999, and 2000 is shown in the third column of the table above.

In this model, it is estimated that for every percent increase in market share of the largest carrier on the route, the fare will increase by 17.1%. This result is similar to the model using fixed effects with the within estimator.

2.d

In the pooled OLS model, a one unit increase in market share decreased the fare by 48.8%. This result is unintuitive, as you would expect fares to increase as market share of the largest carrier rises on a certain route. This may be due to bias caused by fixed effects that are not accounted for. The results of the fixed effects and first differences models were more realistic, with 16.9% and 17.1% increases in fares per once percent increase in market share.

2.e

A possible characteristic of a route that may be captured in a_i is holidays that may affect demand on that route specifically. A second is airport location - whether the airport is close to a major airline hub may also be captured in a_i. These factors could contribute to *concen* being endogenous especially the airport location, as there is a good chance these factors are correlated.

Problem 3

The following tables shows the results of models that will be discussed throughout Problem 3.

	(1) LPM	(2) LPM expendB	(3) Probit
main			
expendA	0.000466	0.000987	0.00618
	(0.000122)	(0.0000951)	(0.000949)
prtystrA	0.0152	0.00697	0.0295
	(0.00376)	(0.00273)	(0.0220)
democA	0.415	0.187	0.844
	(0.0731)	(0.0546)	(0.418)
expendB		-0.00114	-0.00808
		(0.0000877)	(0.00120)
_cons	-0.653	0.0720	-1.687
	(0.204)	(0.155)	(1.267)

Standard errors in parentheses

3.a

In order to estimate the probability of candidate A winning the election, we must first create a binary variable, *win,* which equals 1 when votes for party A exceeded 50%. The estimates of the following linear probability model are shown in column one of the table above.

$$win = \beta_0 + \beta_1 expendA + \beta_2 prtystrA + \beta_3 democA + \varepsilon$$

With this model, all independent variables are statistically significant ($p < .05$), and all independent variables have a positive relationship with the dependent variable *win*.

B_1, the coefficient for *expendA*, is .0004661. This is interpreted as for every $1000 increase in spending by party A, the probability of winning increases by .046 percentage points.

3.b

If *expendB* is excluded from the model, it may suffer from omitted variable bias. The expenditure of party B in the district is likely correlated with both the dependent variable (*win)* and an included independent variable. Its effects are being attributed to the independent variables currently in the model, giving them biased coefficients.

3.c

In the second column of the table above, a variable for expenditure by party B (*expendB)* is now included.

The coefficient for *expendA* (B_1) is .0009867. This is an increase from the same coefficient in the model that did not include *expendB*, .0004661. This is an indication that the effects of *expendB* were previously being attributed to *expendA* in the first model. This is interpreted as for every $1000 increase in spending by party A, the probability of winning increases by .099 percentage points.

3.d

The third column in the table above is a probit model estimating the probability of candidate A winning.

Of the included independent variables, all are considered statistically significant ($p < .05$) except for *partystrA (p = .181)*. According to the model, increasing *expendA, prtystrA,* or *democA* increase the probability of candidate A winning, while increasing *expendB* decreases the probability. To convert this estimate to an exact probability, a model output must be passed through a normal distribution.

3.e

Estimating the change in win chance due to spending an additional $50,000 (using sample means of the explanatory variables), we first calculate a starting point using the sample means.

Starting Point

(-1.686781 + (310.611*.006178) + (305.0885 * -.0080762) + (49.75723 * .0294829) + (.5549133 * .843552)) = -.297

Passed through a normal distribution: 38.3%

Extra $50,000 Spent

(-1.686781 + (360.611*.006178) + (305.0885 * -.0080762) + (49.75723 * .0294829) + (.5549133 * .843552)) = .012

Passed through a normal distribution: 50.5%

An increase of $50,000 in this scenario increases the win probability by 12.1%.