Public Economics (ECON 131)
Section #2: Empirical Tools and Connecting Theory to Data

Contents

Page

1 Empirical Tools **2**

1.1 A First View on OLS: Best fitting Line 2

1.2 A Second View on OLS: Conditional Expectation Function 4

1.3 A Final View on OLS: Correlation 7

1.4 Diff-in-Diff 8

2 Connecting Theory to Data: Immigration Example **12**

3 Extension Questions **14**

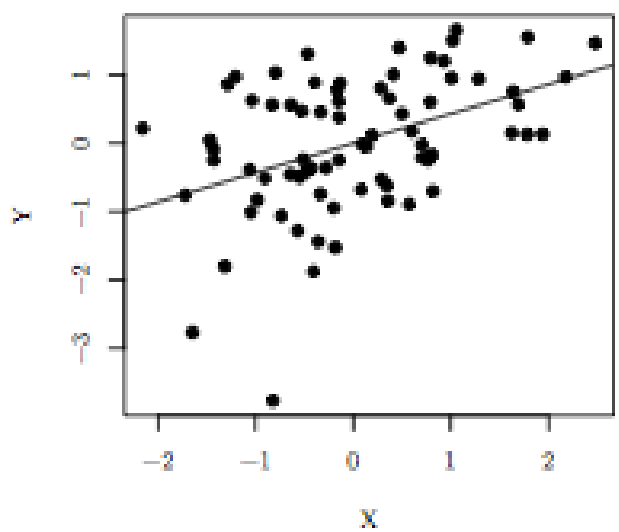
1 Empirical Tools

In this section we will begin to understand the tools available for us in studying public finance empirically.

1.1 A First View on OLS: Best fitting Line

Estimating by Ordinary Least Squares (OLS) is a particular way of fitting a line to a set of points. In particular, OLS fits the line that minimizes the sum of the *squares* of the distances to the points. [draw examples of scatterplots and lines which don't minimize the sum of squares]

Figure 1: Intro Idea: Best Fitting Line



Formally, assuming a regression line of the following form:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \tag{1}$$

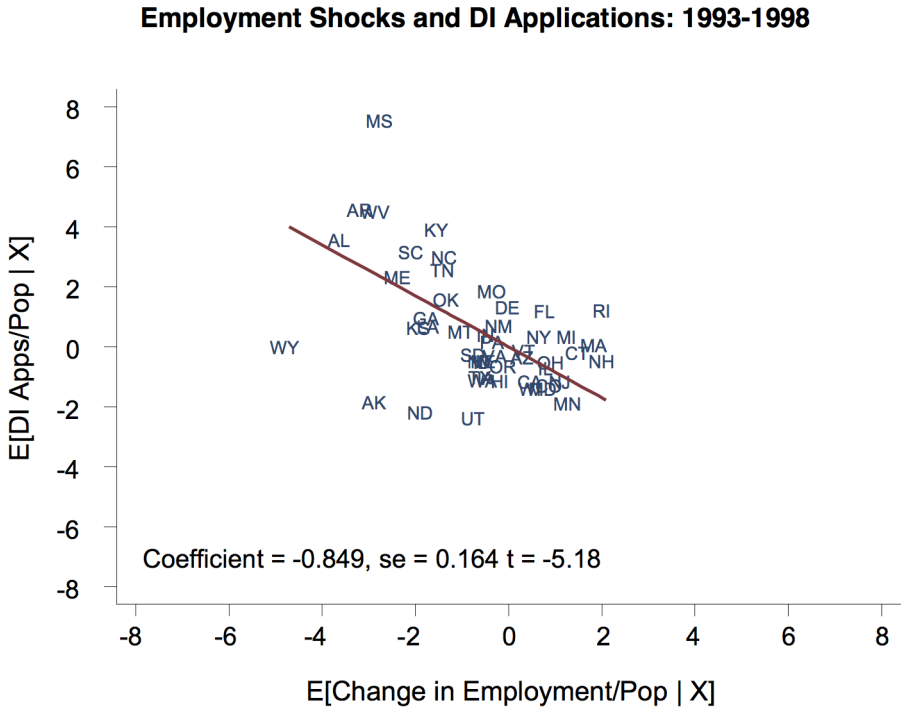
The line minimizes the sum of the squared vertical distances:

$$\min_{b_0, b_1} \sum_i (y_i - (b_0 + b_1 x_i))^2 \tag{2}$$

Disability Insurance Example

In the figure below OLS has been used to fit the best line through the scatterplot of states.

Figure 2: Disability Insurance and Labor Market Strength



The coefficient reveals that a 1 percentage point change in the employment rate is associated with 0.85 fewer DI applications per person. If this was a *causal* relation an explanation could be that during recessions people find it harder to find a job and are more inclined to apply for disability insurance to get an income.

1.2 A Second View on OLS: Conditional Expectation Function

A second way of viewing OLS is as the linear Conditional Expectation Function (CEF). In order to give meaning to this notion we need to first (re-)introduce a few concepts.

1. Expectations

$$\mathbb{E}(Y) = Y_1 \times Prob_1 + Y_2 \times Prob_2 + ... + Y_j \times Prob_j \tag{3}$$

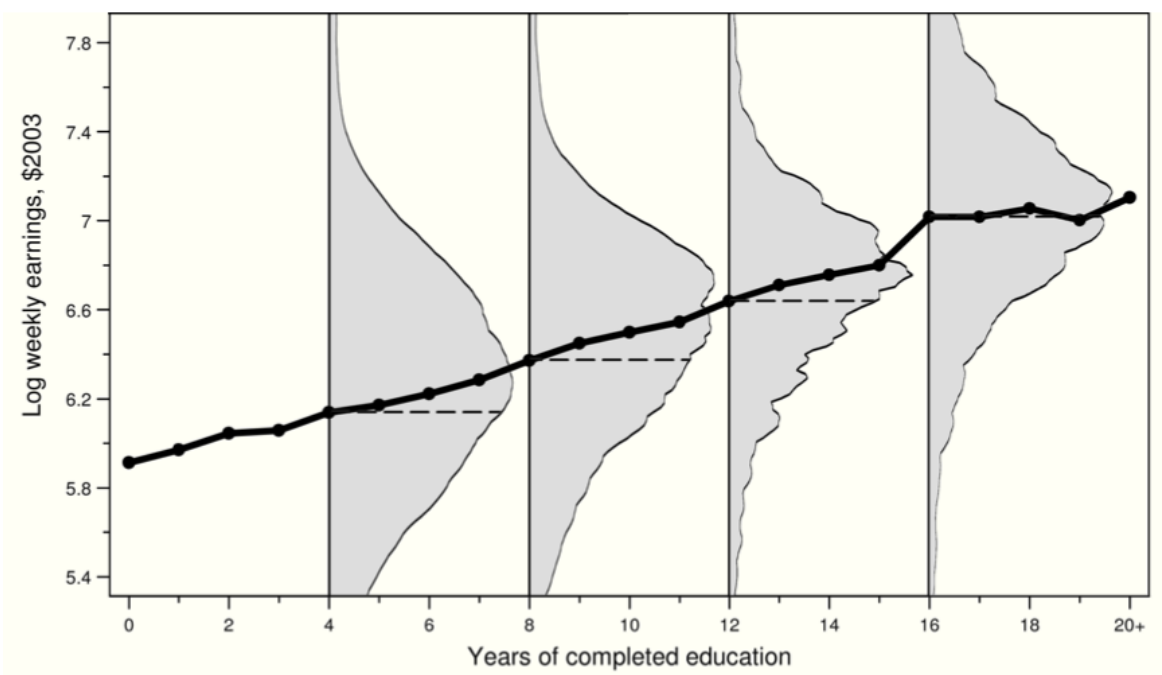
2. Variance

$$Var(Y) = \mathbb{E}[Y - \mu_y]^2 \tag{4}$$

3. Covariance

$$Cov(X, Y) = \mathbb{E}[X - \mu_x][Y - \mu_y] \tag{5}$$

The figure below depicts the conditional mean (/expectation) of the logarithm of weekly wages for each level of education. Note that about half of the mass is on either side of the dashed line in each of the distributions.



Since the line looks approximately linear, it does not seem far off to model the conditional expectation function as linear in this case, i.e.

$$\mathbb{E}[Y_i|X_i] = X_i\beta \tag{6}$$

To get to our OLS regression, note that by definition

$$y_i = \mathbb{E}[y_i|x_i] + \underbrace{\left(y_i - \mathbb{E}[y_i|x_i]\right)}_{\equiv \epsilon_i} \tag{7}$$

$$= \beta_0 + \beta_1 x_i + \epsilon_i \tag{8}$$

which in words states that Y_i can be decomposed into its Conditional Expectation Function plus a zero-conditional-mean error.

Regression Output

The table below shows the regression output for log earnings on schooling, equivalent to the previous figure. Like in the figure the intersect is at 5.84 (= \$344 per week), and the slope is 0.067, i.e. wages increase by 6.7% for each additional year of schooling. The 95% confidence interval for the slope is (0.0668, 0.0681), and is therefore significantly different from zero.

```
. regress earnings school, robust

      Source |       SS       df       MS              Number of obs =  409435
-------------+------------------------------           F(  1,409433) =49118.25
       Model |  22631.4793     1  22631.4793           Prob > F      =  0.0000
    Residual |   188648.31 409433  .460755019           R-squared     =  0.1071
-------------+------------------------------           Adj R-squared =  0.1071
       Total |  211279.789 409434   .51602893           Root MSE      =  .67879

------------------------------------------------------------------------------
             |               Robust                     Old Fashioned
    earnings |      Coef.   Std. Err.      t                Std. Err.       t
-------------+----------------------------------------------------------------
      school |   .0674387   .0003447   195.63               .0003043   221.63
      const. |   5.835761   .0045507  1282.39               .0040043  1457.38
------------------------------------------------------------------------------
```

Example: Height by gender

Let $Y = height, X = gender$, average male height = 6 ft, average female height = 5 ft 4 in. What is the CEF?

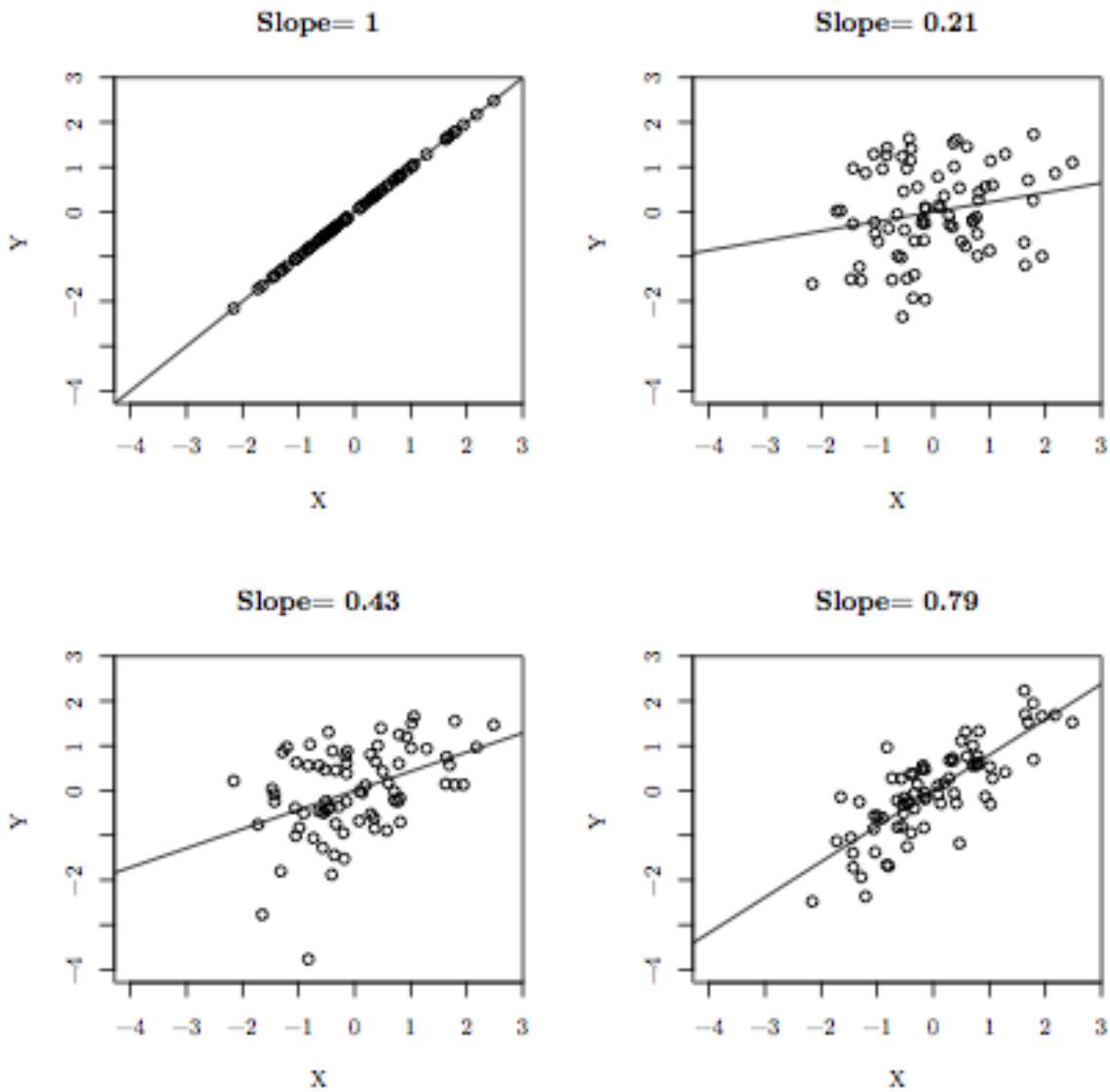
Rename X to D, our usual name for a dummy variable, and let $D = 1$ if the person is female, $D = 0$ if the person is male. Then $\mathbb{E}[Y|D] = 6 - \frac{2}{3} * D$ (measured in feet). This means that if we ran OLS we would get $\beta_0 = 6, \beta_1 = -\frac{2}{3}$.

1.3 A Final View on OLS: Correlation

The correlation between two variables, X and Y, is defined as $\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$.

What is β_1? $\beta_1 = \frac{Cov(x,y)}{Var(x)} = \rho_{xy}\frac{\sigma_y}{\sigma_x}$.

In the figure below, the variance of X and Y are roughly equal, so we may ignore the term $\frac{\sigma_y}{\sigma_x}$. The slopes therefore roughly reflect the correlation between X and Y. When high Y's and associated with high X's, and low Y's are associated with low X's, the correlation is strong and positive, and hence the slope coefficient is high. When all X's are associated with roughly the same Y's, the correlation and hence the slope is close to zero.



1.4 Diff-in-Diff

When working with OLS there are often issues with determining the direction of causality. In the example of regressing wages on schooling we cannot tell from the regression whether a one-year increase in schooling *causes* a 6.7% increase in wages. This would be true if the level of schooling was randomly assigned in the population, or if the mechanism assignment was independent of the level of wages. However, much economic research claims that this is not true; people with high ability tend to take more education *and* get higher wages. This is an example of a **missing variable problem** in determining causality. A classic example of **reverse causality** is a positive correlation between the rate of police men and crime. An OLS regression of crime rates on police men may give a positive correlation, but the causality may in fact be reverse - i.e. more police men does not cause more crime, but more crime may cause more police men.

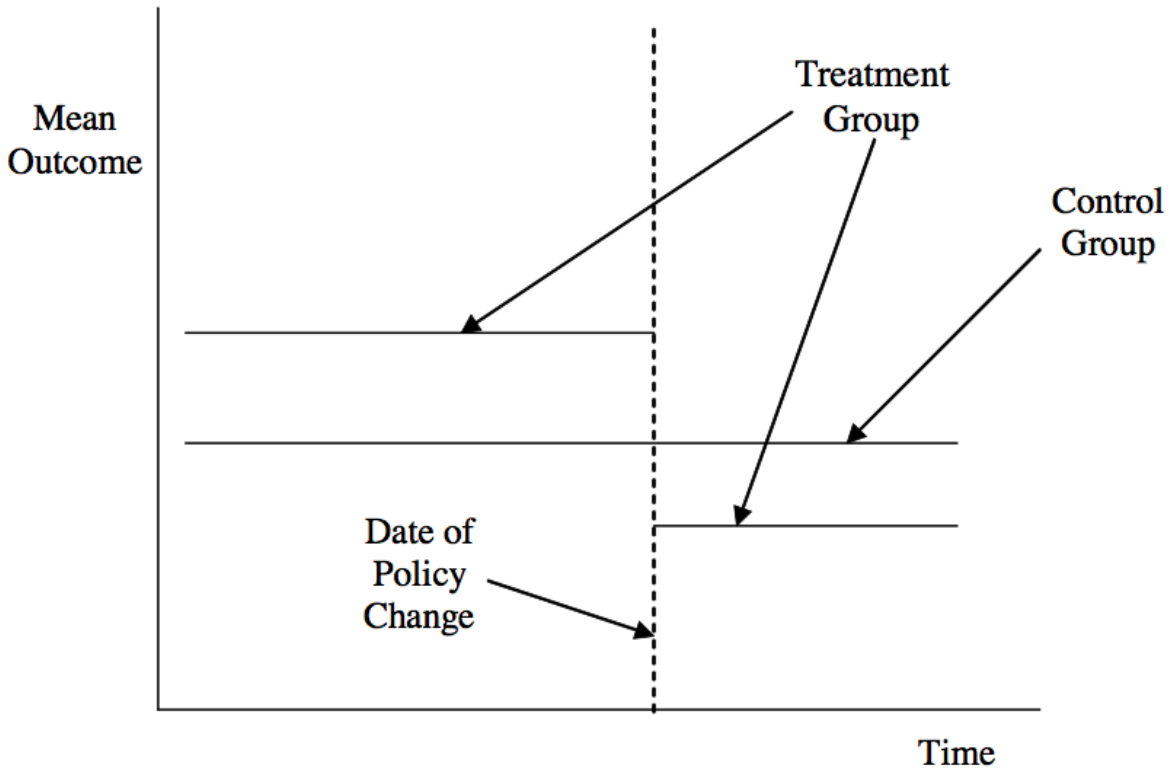
To alleviate these two problems with causality economists like to work with experiments where the treatment and control groups are (roughly) randomly assigned. If the groups are randomly assigned, then the difference between their outcomes is the causal effect of the experiment. With natural experiments the big issue is to determine whether the treatment and control groups are in fact randomly assigned.

One empirical tool often used to study natural experiments is difference-in-differences. This is used when a policy affects a group in the population and it is possible to find a comparable group not affected by the policy. Under diff-in-diff we do not need to assume that the two groups had identical outcomes prior to the implementation of the experiment, but we do need to assume that the outcomes would have grown in the same way for the groups if the experiment had not taken place (the **parallel trends/common trend** assumption).

The effect of the experiment, under this assumption, is found as

$$Effect = [After - Before]_{Treatment} - [After - Before]_{Control} \tag{9}$$

which is illustrated in the following graph


Mean
Outcome
Treatment
Group
Control
Group
Date of
Policy
Change
Time

Diff-in-Diff Example Gruber (2000): Canadian Disability Insurance

The figure below show the monthly flat rate (CAD) of the Quebec Pension Plan (QPP) and the Canada Pension Plan (CPP) over time. It shows that the CPP rate was lower than the QPP rate in 1973-1986 and then increased to reach the level of the QPP rate in 1987.

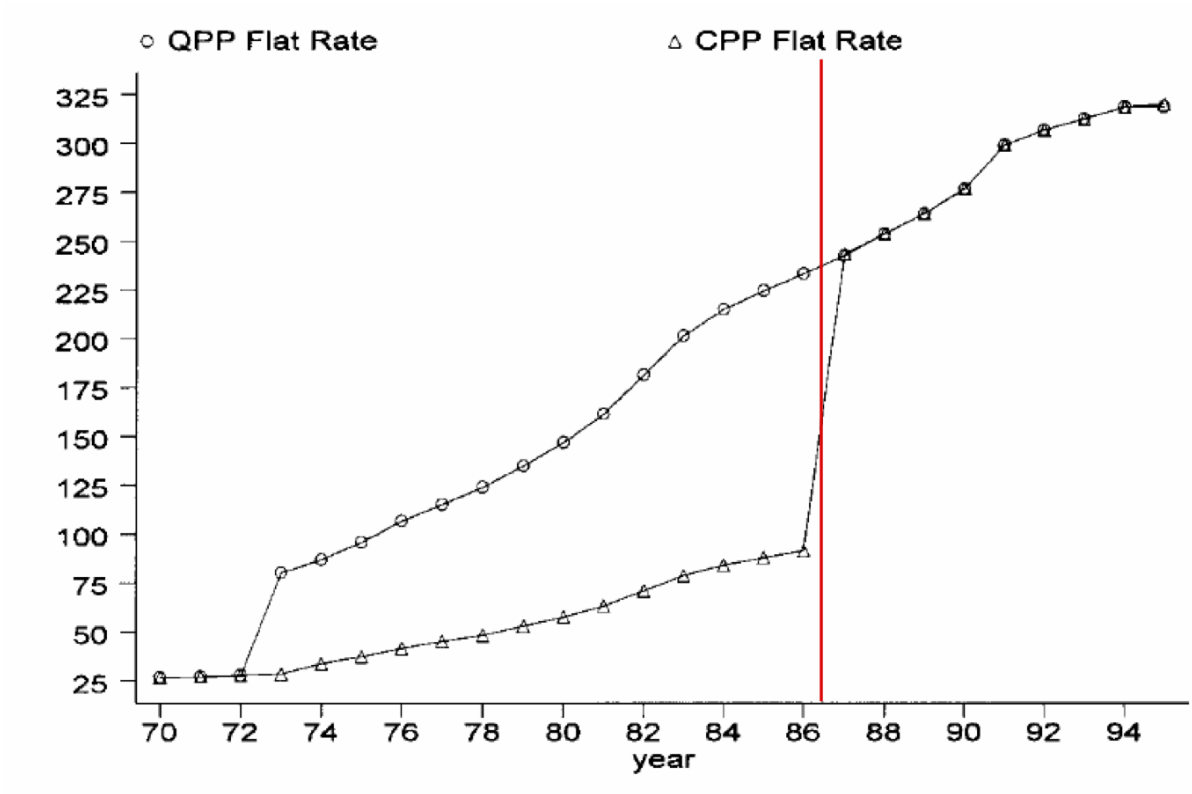


FIG. 1.—Flat-rate portion in Quebec and the rest of Canada

Under the assumption that the labor supply participation rate would grow similarly in Quebec and the rest of Canada had the QPP rate not changed in 1987 onwards, the effect of the policy on labor force participation can be found using equation 9. This is done in the table below. It shows that

$$Effect = [After - Before]_{CPP} - [After - Before]_{QPP} \tag{10}$$

$$= [.217 - .200] - [.246 - .256] \tag{11}$$

$$= 0.027 \tag{12}$$

i.e. under the parallel trend assumption the increase in the flat rate increased the non-employment rate by 2.7%.

Table of Diff-in-Diff from Gruber

TABLE 1
MEANS

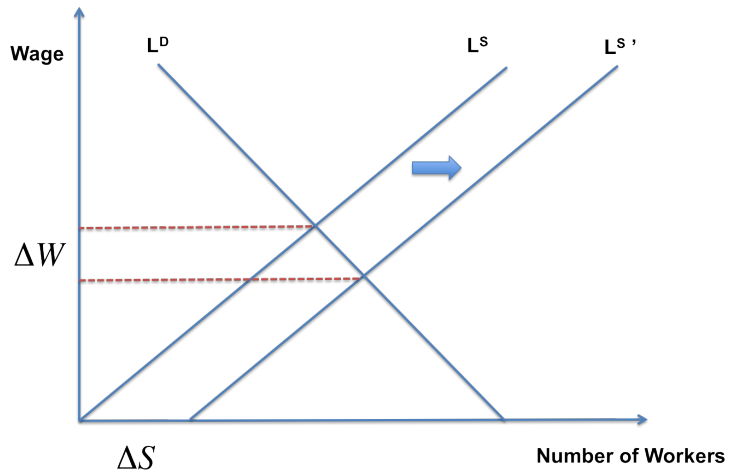
	CPP		QPP		DIFERENCE IN DIFFERENCE
	Before (1)	After (2)	Before (3)	After (4)	(5)
Benefits	5,134	7,776	6,878	7,852	1,668 (17)
Replacement rate	.245	.328	.336	.331	.088 (.003)
Not employed last week	.200	.217	.256	.246	.027 (.013)
Married?	.856	.856	.817	.841	−.024
Any kids < 17?	.367	.351	.354	.336	.002
Less than 9 years of education	.303	.274	.454	.421	.004

2 Connecting Theory to Data: Immigration Example

Say that there is a shock to the labor supply curve caused by sudden immigration, while the demand curve stays fixed.

What is the effect of immigrants on natives' wages?

Because the demand curve stays fixed, the sudden immigration represents a movement along the demand curve towards the right.



Regression

From the expression of the demand curve we can derive the following regression

$$\Delta W = \alpha + \beta \Delta S + \epsilon \tag{13}$$

where $\beta < 0$. But can we truly find that causal effect of the immigration shock on wages simply by running this regression?

Questions to consider

1. How do we define (/restrict) the labor market? How "big" is ΔS?
   - Skill Groups?
     - Does the influx of immigrants only affect certain skill groups? E.g. blue collar workers? Then we might want to look only at the blue collar labor markets.
   - Local labor markets?
     - Are only particular geographical areas affected?

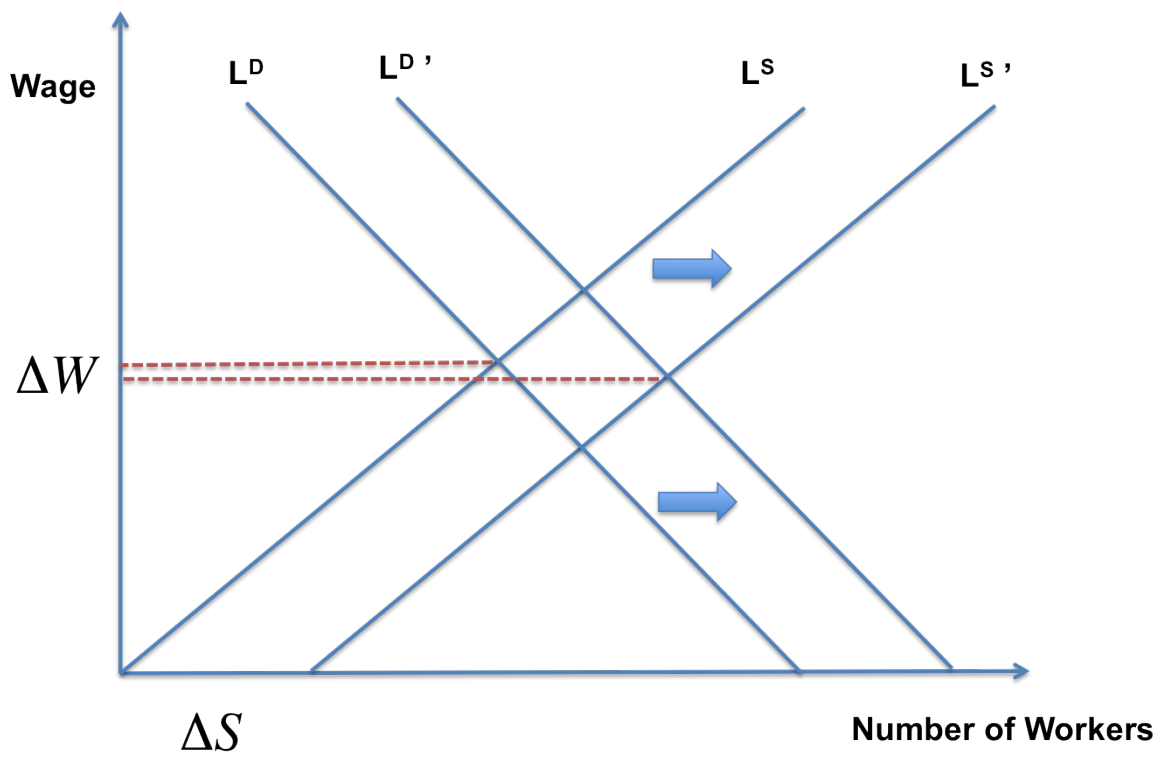
2. Does Demand shift right as well?

   - Larger population $\implies$ Higher local demand for goods and services $\implies$ Higher L^D
   - Constant $\frac{K}{L}$ Ratio $\implies$ Higher L^D
   - Trade theory $\implies$ Higher L^D

   If so, we won't be able to infer the demand curve from the OLS regression, and it is not clear whether wages fall.

3. What would have happened to wages otherwise?

   - Is there an underlying trend in wages due to technology/increases in educational level/... ?



What is the effect of immigrants on natives' wages?

If the

3 Extension Questions

1. Hypothetically, if we had before and after data on health insurance costs for a privately insuring households, and want to test if the rollout of Obamacare reduced the growth of healthcare costs. Write an OLS equation that could test for that using an indicator variable for the period before or after the rollout (takes values of 0 or 1 depending on if the data observation fulfills a condition)? What is an example of a problem that could make that coeficient not express a causal effect?

$$Costs_i = \beta_0 + \beta_1 After_i$$

   where $After_i$ is an indicator variable taking 1 if it's the post period. The issue is this may also capture other trends going on at the time such as if health insurance costs were already growing then this mixes the effect of time and the rollout into the effect β_1

2. Suppose different states rolled out the online health insurance exchange websites at different times. We decide to instead try to use Diff-in-Diffs to test if the rollout of the online health insurance exchange reduced the growth of healthcare costs. What is the treatment group? What is the control group? What do you need to check to see if this method is applicable?

   Treatment group is states that rolled out the online exchange, and had it working in an earlier period (e.g. first 3 months) vs the control group is those states that took longer (e.g. 3 months after original target). We need to check if parallel trends in the two groups health care cost growth existed before the obamacare rollout.

3. Suppose the test before for trends holds. You find the growth of healthcare costs in states with early online rollouts is 1.8% and the growth in healthcare costs in states with later rollouts is 2.3%. How would you interpret your results?

   The effect of an earlier online system rollout is reduced healthcare cost growth by 0.5% in the first year of Obamacare. This suggests at least this part of the obamacare

program prevents costs from growing as fast. Of course, you could also argue this is messier in real life since the online system that gets rolled out could be of different quality depending upon whether the government supports the program or not. So there's still a danger we've conflated this timing effect with some other effects such as differing quality being offered.