Part 0 – Open Stata, and make your own do-file

- Start Stata through the start menu button
- In de white command window type `doedit` to start de do-file editor. Place the Stata screen on the left and the do file editor on the right such that you can easily switch between the two.
- Save the empty do-file as a new do-file under an applicable name such as ectrcs_did.do in a directory that you want to use for this course, for example H:\ectrcs
- In the first two lines of the do-file type

```
cls                                //this clears the screen
clear all                          //this clears the memory
cd "H:/ectrcs"                     //this is your path
```

- If you want to use your own computer type instead:

```
cd "~/ectrcs"                //this is your MAC path cd
"c:/ectrcs"                  //this is your PC path
```

Part 1 – Differences in Differences

In this computer exercise we will analyze the relation between wages and employment using quasi experimental data (Chapter 13 of S&W). More in particular, we will estimate the impact of an increase in the minimum wage on low-skilled workers. Economic theory predicts that in a competitive market, demand will fall when prices increase. A minimum wage increase should thus lead to unemployment if the labor market is competitive.

Until the 1980s empirical evidence of the impact of minimum wages on employment was provided by regression models of aggregated time series data. The dependent variable typically is country wide teenage employment,[1] which is explained by a host of factors including minimum wages. Empirical results indicate that a 10% increase in the minimum wage leads to 1%-3% decrease in teenage employment.

The main problems with the use of multiple regression models based on aggregated time series are (1) omitted variables; (2) measurement error; (3) simultaneous causality of employment (quantity) and wages (price). As a reaction on the aggregated time series approach Card and Krueger (1994), hereafter denoted as CK94, use quasi-experimental data. You can find the original article on Canvas.

We will use the data of CK94 to replicate their main results (see below). Download the file `did.dta` from the Canvas site. A description of all variables can be found in the file `did.pdf`. Available are data on employment in 410 fast food restaurants in New Jersey (NJ) and Pennsylvania (PA) before and after an increase in minimum wages in New Jersey on 1 April 1992.

[1] Given that it is mainly teenagers that earn wages close to the minimum wage, focusing on this group is more interesting than the total population that seeks employment.

1. The unit of observation is a particular restaurant. Total employment is defined as the number of full-time employees (inclusive management) plus 0.5 times the number of part-time employees. This is the dependent variable in the empirical analysis. Generate this variable for both periods (`generate emptot1 = emppt*0.5 + empft + nmgrs` and `generate emptot2 = emppt2*0.5 + empft2 + nmgrs2`). Regress emptot1 on a constant term and the dummy variable state, which has the value 1 for a restaurant in the treatment group (`state==1` if restaurant located in NJ) and zero for the control group (`state==0` for PA restaurants). Do not forget to use heteroskedasticity-robust standard errors (i.e. type `, robust`).
2. The estimated regression from question 1 is a simple regression model with a dummy variable regressor. The estimated coefficient of the constant term is equal to the sample average of the control group, hence in this case average employment in PA restaurants. The coefficient of the state dummy measures the difference in average employment between treatment group (NJ restaurants) and control group (PA restaurants) before the minimum wage increase (pretreatment). The estimation results are equal to those in row 1, columns (*i*) and (*iii*) of Table 3 in CK94. To reproduce row 1,[2] column (*ii*) `generate statealt = 1 - state` and type `regress emptot1 stateal t, robust`. Think carefully about what the coefficients of the constant and the variable statealt actually measure.
3. Proceed in the same way (but now use emptot2 as dependent variable) to generate the results in row 2, columns (*i*)-(*iii*) of Table 3. Basically what you get is the differences estimator: a comparison of post treatment (after the minimum wage increase) outcomes. The estimation results show a slightly negative employment effect of -0.14 employees, although not significantly different from zero (standard error equals 1.07). If the zero conditional mean assumption holds we should get an unbiased estimator of the treatment effect. However, it is likely that average employment in NJ and PA restaurants differs for other reasons too. If we do not control for these pretreatment differences the differences estimator suffers from an omitted variables problem.
4. The differences-in-differences estimator will automatically control for these pretreatment characteristics by looking only at changes in employment (instead of the level of employment). Generate now employment changes (`generate emptotd = emptot2 - emptot1`) and regress `emptotd` on a constant and the state dummy. Again this is a simple regression and the coefficient of the constant term measures the average change in employment in PA restaurants. This coefficient equals –2.28 and is significant, i.e. employment has on average decreased in PA restaurants. The coefficient of the dummy regressor state measures the difference in average employment changes between NJ and PA restaurants. This coefficient is called the differences-in-differences (DID) estimator and equals 2.75 (row 4, column (*iii*)). Also we can deduct from these estimation results that the average employment change in NJ restaurants equals –2.28+2.75=0.47 (row 4, column(*ii*)). Based on intuition and

[2] Small differences with CK94 table 4 are due to rounding errors.

economic theory we would expect that an increase in the minimum wage leads to lower employment holding constant other relevant factors. Because we analyze employment changes over almost a year these other relevant factors are probably not constant. The validity of the DID estimator depends on the assumption that these other relevant factors of employment have had a more or less similar impact on NJ en PA restaurants. If that is the case we would expect a negative DID estimate. The empirical results of CK94, however, show the opposite! This is the main empirical result of CK94.

5. The DID estimate in part 4. can also be obtained by regressing employment on a treatment indicator while including both time and restaurant fixed effects. In order to do this we have to "reshape" the data. First we keep only the restaurants in the data for which we have data on employment both before and after the change in the minimum wage, a so called balanced panel, by typing `keep if emptot1!=. & emptot2!=.` We keep only the relevant variables in the data by typing `keep restaurant_id state emptot1 emptot2`. Next we create a dataset in long form by typing `reshape long emptot, i( restaurant_id) j(time)`. By typing `browse` you can see that we now have two observations for each restaurant.
6. We are now going to obtain the DID estimator of the effect of the minimum wage on employment. We first let Stata know which variable indicates the entity (restaurant) by typing `xtset restaurant_id`. Next, we create a variable indicating that a restaurant has been treated by the minimum wage increase by typing `gen treated=0` and next `replace treated=1 if state==1 & time==2`. Obtain the DID estimator by typing `xi: xtreg emptot treated i.time, fe robust`. Compare the results to the results in part 4.
7. It is also possible to obtain the DID estimator by regressing employment on the treatment indicator, a state dummy and a time dummy. Type `xi: regress emptot treated i.state i.time, robust`. Compare the estimated coefficient and the standard error on the variable treated with the estimates obtained in part 6. Why is the standard error in part 6. smaller?
8. We go back to the data set in "wide form" by typing `use did.dta, clear.` We have to create the employment variables again by typing

```
generate emptot1 = empp t*0.5 + empft + nmgrs
generate emptot2 = emppt2*0.5+empft2+nmgrs2
generate emptotd = emptot2-emptot1
```

9. So far the control group has been PA restaurants. CK94 also provide DID estimates for an alternative control group, i.e. NJ restaurants with a relatively high starting wage in the first period (before the minimum wage increase). They divide the restaurants in three groups, i.e. low (starting wage equals $4.25), middle (wage between $4.26 and $4.99) and high (wage higher than $4.99). Generate with the following Stata commands four binary variables representing the low, middle, and high categories as well as a category for which no starting wage is observed:

```
generate low=0
generate mid=0
generate high=0
generate dna=0
```

```
replace low=1 if wage_st==4.25
replace mid=1 if wage_st>4.25 & wage_st<5.00
replace high=1 if wage_st>=5.00
replace dna=1 if wage_st==.
```

10. Consider now the NJ restaurants only. There are in total 331 NJ observations (`state==1`) of which 101 restaurants have a starting wage equal to \$4.25, 140 between \$4.26 en \$ 4.99 and 73 equal or higher than \$5.00. Also 17 restaurants do not have reported a starting wage at all (`dna==0`), hence we leave them out of the analysis. Using the remaining 331-17=314 observations regress emptot on low, mid and high. The precise Stata command is `regress emptot1 low mid high if state==1 & dna==0, noconstant robust`. In other words, we leave out the constant term by adding the option "`noconstant`" to the estimation command. Why?
11. From the Stata output table you see that only 305 observations have been used in estimation. The reason is that there are 9 missing observations on employment. If you have specified all commands properly your estimates should equal those reported in row 1, columns (*iv*)-(*vi*) of Table 3 (slight differences are due to rounding errors).
12. Perform the same regressions as in part 10., but now with `emptot2` and `emptotd` as dependent variables. You will get the estimation results of rows 2 and 4, columns (*iv*)(*vi*). It is seen that employment increased in restaurants with a low initial starting wage. However, the opposite is true for the control group, i.e. restaurants with an initial high starting wage. CK94 judge these results as additional evidence for the earlier surprising result that an increase in the minimum wage causes more employment. In other words, comparing restaurants influenced by the minimum wage increase with either PA restaurants or NJ restaurants with high initial starting wage as control group leads to similar outcomes.
13. The validity of especially the control group consisting of high starting wage restaurants can be questioned. The assumption is that the control group is not affected by the minimum wage increase. However, inspecting the wage changes in this control group we observe quite a few restaurants with a wage decrease. Type `tabulate wage_st wage_st2 if state==1 & dna==0 & high==1` to show the joint frequency distribution for the control group (`state==1` and `high==1`) of starting wages before and after the minimum wage increase. From the total of 71 restaurants in the control group with complete information, 38 restaurants went from \$5.00 to the new minimum wage of \$5.05. Only 6 went from \$5.00 to an even higher starting wage. These results assert the validity of the control group (NJ restaurants with an initial high starting wage). However, it is also the case that 18 of the 23 restaurants with an initial wage higher than the new minimum wage decreased their starting wage to the level of the new minimum wage (\$5.05). This is remarkable and at least questions the validity of this alternative control group used by CK94.

Finally, we invite you to download the paper of Card & Krueger (1994) and have a look yourself at their other empirical results. As said before their main contribution to the minimum wage literature is to use quasi-experimental firm level data instead of observational aggregated time series data. Their empirical analysis does not go beyond the use of linear regression. Hence, you will be able to reproduce all their estimation output with your current knowledge of the multiple regression model and its application to experimental data (see Chapter 13 of S&W).

Table 1: Card & Krueger (1994), Table 3

780 *THE AMERICAN ECONOMIC REVIEW* *SEPTEMBER 1994*

TABLE 3—AVERAGE EMPLOYMENT PER STORE BEFORE AND AFTER THE RISE IN NEW JERSEY MINIMUM WAGE

	Stores by state			Stores in New Jersey[a]			Differences within NJ[b]	
Variable	PA (i)	NJ (ii)	Difference, NJ – PA (iii)	Wage = \$4.25 (iv)	Wage = \$4.26–\$4.99 (v)	Wage ≥ \$5.00 (vi)	Low–high (vii)	Midrange–high (viii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	–2.89 (1.44)	19.56 (0.77)	20.08 (0.84)	22.25 (1.14)	–2.69 (1.37)	–2.17 (1.41)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	–0.14 (1.07)	20.88 (1.01)	20.96 (0.76)	20.21 (1.03)	0.67 (1.44)	0.75 (1.27)
3. Change in mean FTE employment	–2.16 (1.25)	0.59 (0.54)	2.76 (1.36)	1.32 (0.95)	0.87 (0.84)	–2.04 (1.14)	3.36 (1.48)	2.91 (1.41)
4. Change in mean FTE employment, balanced sample of stores[c]	–2.28 (1.25)	0.47 (0.48)	2.75 (1.34)	1.21 (0.82)	0.71 (0.69)	–2.16 (1.01)	3.36 (1.30)	2.87 (1.22)
5. Change in mean FTE employment, setting FTE at temporarily closed stores to 0[d]	–2.28 (1.25)	0.23 (0.49)	2.51 (1.35)	0.90 (0.87)	0.49 (0.69)	–2.39 (1.02)	3.29 (1.34)	2.88 (1.23)

Notes: Standard errors are shown in parentheses. The sample consists of all stores with available data on employment. FTE (full-time-equivalent) employment counts each part-time worker as half a full-time worker. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing.

[a]Stores in New Jersey were classified by whether starting wage in wave 1 equals \$4.25 per hour ($N = 101$), is between \$4.26 and \$4.99 per hour ($N = 140$), or is \$5.00 per hour or higher ($N = 73$).

[b]Difference in employment between low-wage (\$4.25 per hour) and high-wage (≥ \$5.00 per hour) stores; and difference in employment between midrange (\$4.26–\$4.99 per hour) and high-wage stores.

[c]Subset of stores with available employment data in wave 1 and wave 2.

[d]In this row only, wave-2 employment at four temporarily closed stores is set to 0. Employment changes are based on the subset of stores with available employment data in wave 1 and wave 2.