Advanced Statistics and Programming - Individual Assignment 2

Thanh Dat Nguyen - 532618tn

29 September 2022

Difference-in-Difference analysis: female labor force participation

The following section studies the effects of the 1993 'Earned Income Tax Credit' intervention (from here on the EITC) using the difference-in-difference (from here on DiD) analysis, where the introduction of the policy serves as a treatment and divides the timeline into a pre- and post-treatment period; the presence of children divides the pool into the treatment and control group.

The basic difference-in-difference equation has the following form:

$$y_{it} = \beta_0 + \beta_1 D_i + \beta_2 T_i + \beta_3 D_i T_t + \epsilon_{it} \tag{1}$$

The coefficients in the above equations can also be used to express the following conditional expected values of y (through substitution), which are key to the difference-in-difference analysis:

- $E = (y_{T=1}|D = 1) = \beta_1 + \beta_2 + \beta_3$ - expected value of y in the post-treatment period for the treatment group
- $E = (y_{T=0}|D = 1) = \beta_1$ - expected value of y in the pre-treatment period for the treatment group
- $E = (y_{T=1}|D = 0) = \beta_2$ - expected value of y in the post-treatment period for the control group
- $E = (y_{T=0}|D = 0) = 0$ - expected value of y in the pre-treatment period for the control group

The difference-in-difference effect is the result of the subtraction of the difference in the post-treatment and pre-treatment period between the two groups, expressed as:

$$[E(y_{T=1}|D = 1) - E(y_{T=0}|D = 1)] - [E(y_{T=1}|D = 0) - E(y_{T=0}|D = 0)] \tag{2}$$

Substituting the above yields:

$$[(\beta_1 + \beta_2 + \beta_3) - \beta_1] - [\beta_2 - 0] = \beta_3 \tag{3}$$

So β_3 is the difference-in-difference effect.

Visualization of difference-in-difference effect

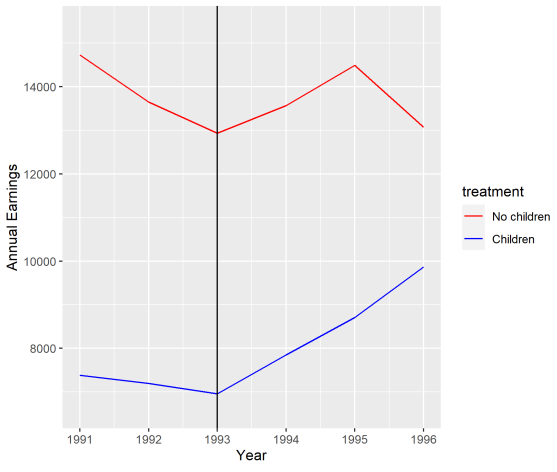

Figure 1: DiD Effect in the Annual Earnings

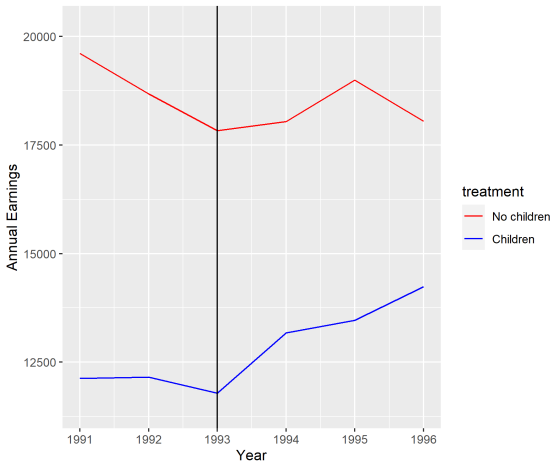

Figure 2: DiD Effect in the Family Income

The basis for this difference-in-difference analysis is the assumption that the treatment (the women with children) and control group (the women without children) react differently to the treatment at hand (the introduction of the EITC). As the EITC applied to women with children and was supposed to serve as a

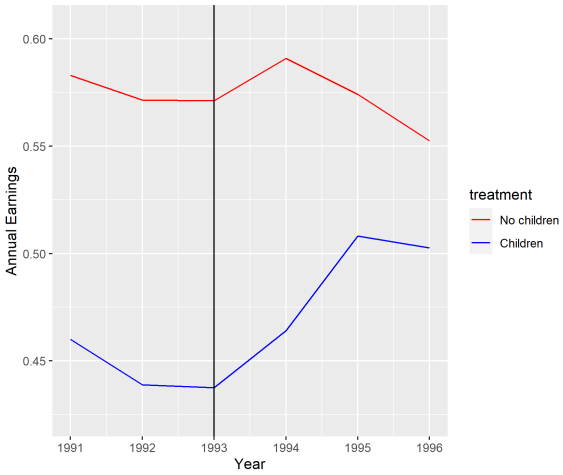

Figure 3: DiD Effect in the Working Status of the Female Labor Force

financial incentive to stimulate more female labor participation, the dependent variable y should experience a greater reaction for the treatment group compared to the control group. In the analysis, multiple dependent variables can be chosen to measure the effect, such as annual earnings, annual family income and the working/non-working indicator. To demonstrate the DiD effect, a line graph is plotted for each of the variables, connecting the average values across the different years. The graphs also include a vertical line at the year 1993 to distinguish the different treatment periods.

As can be seen in each of the graphs, the year 1993 serves as a clear breaking point and greatly alters how each line behaves from that point onward. It is interesting to see that the introduction of EITC also positively affected the control group, even though the incentive did not apply to them. As expected, however, the treatment group benefited much more from the incentive, as their graphs have much more positive slope.

Summary Statistics of the Labor Data

Table 1: Summary Statistics of the Sample

Statistic	N	Mean	St. Dev.	Min	Max
state	13,746	54.525	27.135	11	95
year	13,746	1,993.347	1.703	1,991	1,996
urate	13,746	6.762	1.462	2.600	11.400
children	13,746	1.193	1.382	0	9
nonwhite	13,746	0.601	0.490	0	1
finc	13,746	15,255.320	19,444.250	0.000	575,616.800
earn	13,746	10,432.480	18,200.760	0.000	537,880.600
age	13,746	35.210	10.157	20	54
ed	13,746	8.806	2.636	0	11
work	13,746	0.513	0.500	0	1
unearn	13,746	4.823	7.123	0.000	134.058

The provided data set covers the period between the years 1991 and 1996 and includes women aged 20-54 with less than high-school education (in the US this means less than 12 years of education). As can be seen, the unemployment rate varies greatly across the different US states, with there being states with a rate as high as 11.4% and as low as 2.6%. When inspecting the number of children for women, while there

are extreme cases of up to 9 children present in the data, most women on average have 1 child. What is interesting is that the data set is mainly made up of the Hispanic/Black ethnicity (around 60%) rather than the white ethnicity. When observing the variables regarding income (annual earnings and annual family income), there is great disparity in the observations as well, since there are values as high as 500000 (dollars) in the data set. This is a sizeable sum even for today's standards and there is a chance these observations could skew the above figures, especially when considering that most of the extreme observations occur after the year 1993. Finally, the data is split relatively evenly between working (51%) and non-working women.

Calculating the Difference in Difference effect

To be able to calculate the DiD effect, we fist define the period pre- and post-treatment. As the policy took effect from that tax year 1993 onwards, the period betweeen 1991-1992 is considered the pre-treatment period (period = 1) and 1993-1996 period is considered the post-treatment period (period = 2).

Table 2: DiD Effect in the Annual Earnings

	period	No Children	Children
Before	1	$14,203.900$	$7,290.383$
After	2	$13,507.900$	$8,277.196$
Difference		-695.997	986.813

Table 3: DiD Effect in the Family Income

	period	No Children	Children
Before	1	$19,159.190$	$12,140.900$
After	2	$18,218.950$	$13,111.690$
Difference		-940.239	970.796

Table 4: DiD Effect in Working Status

	period	No Children	Children
Before	1	0.577	0.450
After	2	0.573	0.476
Difference		-0.005	0.026

The control group (no children) experienced a decrease in all of the dependent variables following the EITC intervention, while the treatment group (with children) experienced an increase - so the annual earnings, family income and the working participation all increased. The DiD effect would be calculated by subtracting the change in means of the control group from the change in means of the treatment group. So for the annual earnings the effect would be 1682.81 [= 986.813 -(- 695.997)], for Family Income 1911.035 (both in dollars) and for Working Status 0.031 (can be thought of as percentage of people declaring themselves as working).

Running the DiD Regression

Now that we have described the DiD effect and demonstrated its meaning and calculation, we can finally run the regression model as was described in the model equation at the beginning. The equation is applied three times, once for each dependent variable that is studied. When interpreting the results, please keep in mind that the base case for the regressions are period 1 (pre-treatment - 1991 and 1992) and no treatment (no children). Therefore, the value of the coefficients describe the effect of having children and moving to the post-treatment period on the dependent variable.

As can be seen from the table, in general, having a child has a negative effect on all three dependent variables, as the treatment variable has a negative coefficient in all three columns - it reduces both earnings and income by almost 7000 dollars ($p < 0.01$) and reduce workforce participation by almost 13% ($p < 0.01$) (when all other things are held constant). The basis for introducing the policy was thus sensible. The rest of the results exactly reflect tables 2, 3 and 4 from above. The coefficients for the period variable are negative (for the base case of women with no children), so introducing the initiative generally led to a decrease in the dependent variables for the control group (although this effect is not statistically significant and there is great

Table 5: Simple DiD Regression for Annual Earnings, Family Income and Working Status

	Dependent variable:		
	earn	finc	work
	(1)	(2)	(3)
Constant	$14,203.900^{***}$	$19,159.190^{***}$	0.577^{***}
	(387.548)	(414.751)	(0.011)
Children	$-6,913.517^{***}$	$-7,018.295^{***}$	-0.128^{***}
	(510.988)	(546.857)	(0.014)
Post-treatment	−695.997	-940.239^{*}	−0.005
	(485.413)	(519.486)	(0.013)
Children:post-treatment	$1,682.810^{***}$	$1,911.035^{***}$	0.031^{*}
	(642.099)	(687.171)	(0.018)
Observations	13,746	13,746	13,746
R^2	0.026	0.022	0.012
Adjusted R^2	0.026	0.022	0.012
Residual Std. Error (df = 13742)	17,965.670	19,226.750	0.497
F Statistic (df = 3; 13742)	121.691^{***}	105.245^{***}	54.906^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01

amount of variance in observations). These are exactly equal to the difference in means of the dependent variables after the introduction of the intervention for the control group. The interaction coefficient also equals to the DiD effect that was calculated before (which is in line with the transformation of the equation done above - β_3), meaning that women with children benefited from the new policy.

There are other independent variables in the data set, however. Do the results change as we introduce more control variables? Table 6 does not suggest that to be the case as the coefficients corresponding to the DiD effect still hover around the same level. The main difference is that the interaction coefficient is no longer significant in the regression for the working participation dependent variable ($\beta = 0.028, p = 0.103$). As such, it cannot be confidently stated that the intervention resulted in an increase in the working status of women with children. Inspecting further, it is found that this particular model (with the working status as dependent variable) would benefit from robust standard errors, as heteroskedasticity has been detected in the model with the Breusch-Pagan test. This is seemingly unnecessary, however, as the significance levels and the standard errors remain roughly the same.

Table 6: DiD Regressions with Control Variables

	Dependent variable:			
	earn	finc	work	
	OLS	*OLS*	*OLS*	*Robust errors*
	(1)	(2)	(3)	(4)
Constant	$10,514.550^{***}$	$15,340.540^{***}$	0.278^{***}	0.278^{***}
	(2,736.148)	(2,929.111)	(0.075)	(0.075)
Children	$-6,783.937^{***}$	$-6,518.676^{***}$	-0.118^{***}	-0.118^{***}
	(516.131)	(552.531)	(0.014)	(0.014)
Post-treatment period	$-1,132.252^{**}$	$-1,236.923^{**}$	−0.011	−0.011
	(530.881)	(568.320)	(0.015)	(0.015)
age	28.003^{*}	79.570^{***}	0.002^{***}	0.002^{***}
	(15.918)	(17.041)	(0.0004)	(0.0004)
ed	155.143^{***}	−37.083	0.020^{***}	0.020^{***}
	(60.151)	(64.394)	(0.002)	(0.002)
urate	-585.239^{***}	-471.910^{**}	−0.008	−0.008
	(203.931)	(218.313)	(0.006)	(0.006)
Children:post-treatment period	$1,519.981^{**}$	$1,766.566^{***}$	0.028	0.028
	(638.741)	(683.787)	(0.017)	(0.017)
Observations	13,746	13,746	13,746	13,746
R^2	0.044	0.040	0.053	0.053
Adjusted R^2	0.040	0.036	0.049	0.049
Residual Std. Error (df = 13689)	17,835.510	19,093.330	0.488	0.488
F Statistic (df = 56; 13689)	11.156^{***}	10.105^{***}	13.578^{***}	
Note:				*p<0.1; **p<0.05; ***p<0.01

Effect Heterogeneity

Finally, effect heterogeneity within the data set is examined. To be specific, we compare whether high-education mothers reacted differently to the introduction of the EITC policy. To do this, the effect of the EITC is compared between mothers with high education and mothers with low education. Subsequently, we compare the effect between low education mothers (less than 9 years of education) and low education women without children. A similar approach as above is taken, with the results presented below.

Table 7: DiD Effect Comparing High and Low Education Mothers

	Dependent variable:		
	earn	finc	work
	(1)	(2)	(3)
Constant	$9,113.888^{***}$	$13,897.890^{***}$	0.433^{***}
	(516.756)	(544.116)	(0.017)
High education	$-2,556.068^{***}$	$-2,462.830^{***}$	0.024
	(611.812)	(644.206)	(0.020)
Post-treatment period	370.227	142.756	0.011
	(653.474)	(688.073)	(0.022)
High education:post-treatment	871.461	1,166.212	0.021
	(773.065)	(813.996)	(0.026)
Observations	7,819	7,819	7,819
R^2	0.005	0.004	0.002
Adjusted R^2	0.004	0.003	0.001
Residual Std. Error (df = 7815)	14,923.420	15,713.570	0.499
F Statistic (df = 3; 7815)	12.717^{***}	9.460^{***}	4.884^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01

For the interpretation of this table, the treatment group is the high education group, while low education group is the control group. Interestingly, according to the results, receiving higher education actually had a significant negative effect on the earnings ($\beta = -2556.068, p < 0.01$) and family income ($\beta = -2462.830, p < 0.01$) for a mother, while having a miniscule effect on the working status/participation. The introduction of the EITC had positive, although insignificant effect on all three dependent variables. Inspecting the interaction coefficient, however, it is clear that the high-education mothers benefited more from the policy, as their increase in the averages of the dependent variables were markedly higher (the average annual earnings and family income difference was 871 and 1166 dollars respectively, and the working status rose by 2.1% more compared to low education mothers).

Secondly, we compare the effect of the intervention in the low education women group, but between those with children and those without. Here, the interpretation of the table below stays the same as in the original DiD regression (assignment to treatment is in relation to children again).

Similar results are echoed as to those in the original DiD regression - having children negatively affects all three dependent variables. This time, however, the introduction of the policy led to an increase in the annual earnings and the family income for women without children, while only slightly negatively affecting their working participation. Most surprisingly, though, the women with children in this group actually benefit less from the introduction of the policy, with the coefficient for the interaction term being negative. This effect, however, is highly insignificant and the standard error for the coefficient is extremely high.

Table 8: DiD Effect Comparing Low Education Women with and without Children

	Dependent variable:		
	earn	finc	work
	(1)	(2)	(3)
Constant	$11,850.380^{***}$	$17,816.920^{***}$	0.497^{***}
	(679.840)	(715.686)	(0.018)
Children	$-2,736.488^{***}$	$-3,919.035^{***}$	-0.065^{***}
	(945.162)	(994.998)	(0.025)
Post-treatment period	783.677	322.393	−0.004
	(858.662)	(903.937)	(0.023)
Children:post-treatment period	−413.449	−179.637	0.015
	(1,194.473)	(1,257.455)	(0.031)
Observations	4,311	4,311	4,311
R^2	0.006	0.010	0.003
Adjusted R^2	0.006	0.009	0.002
Residual Std. Error (df = 4307)	18,962.540	19,962.390	0.498
F Statistic (df = 3; 4307)	9.304^{***}	14.690^{***}	4.494^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01

Instrumental Variable analysis: effect of compulsory schooling on wages

In the following section, an analysis will be conducted to determine the effect of education on the future earnings of students. The analysis deals with data on people born between the years 1930 and 1939 in the US and contains data such as age, years of education, weekly earnings, marital status and more.

The Need for Instrumental Analysis

It is the standard societal assumption that education should help you make a better living in the future. Measuring the effect, however, is difficult and standard OLS regression can easily be biased because of the multitude of unobservable variables present. Given that the standard OLS analysis is run, there are a number of things that could bias the results.

An obvious condition that could skew the results are the parents of a given student. If the student comes form a middle to upper class household, where both parents went through the standard schooling system, it is highly likely that their child will follow in their footsteps, when seeing their success. These parents can also have an impact on weekly earnings, as they can have the connections to set their child up for a good career compared to other students. The opposite would also be true - coming from a low-income household, where the parents dropped out of the schooling system, it can be likely for the student to drop out as well.

Another condition that can also bias these results would be the IQ of the students. Students with high IQ, who probably have no problem following the content in class are more likely to stay in education longer. Since they also assumedly like learning, they will probably pursue higher education as well, increasing their years of education. Highly intelligent people also arguably make good employees since they can be thought to be highly versatile and capable. On the other hand, students with a lower IQ may struggle with school

content, and therefore enjoy learning less. Lower IQ can also have an adverse effect on the wages since they may be unable to perform the higher paying and more demanding jobs.

Summary Statistics of the Education Data

Table 9: Summary Statistics of the Education Sample

Statistic	N	Mean	St. Dev.	Min	Max
age	329,509	44.645	2.940	40	50
educ	329,509	12.770	3.281	0	20
lnwage	329,509	5.900	0.679	−2.342	10.532
married	329,509	0.863	0.344	0	1
qob	329,509	2.506	1.112	1	4
SMSA	329,509	0.186	0.389	0	1
yob	329,509	1,934.603	2.905	1,930	1,939

There are almost 330000 observations in the set. The data is relatively even split in terms of age, as the mean is close to the average of the minimum and maximum values - the same statement applies to the quarter of birth, and obviously year of birth. Inspecting the education years variable, it can be said that on average, most people at least finish high school education, as the standard number of years required to reach that state is 12. It is also interesting to see that the grand majority (more than 86% of the sample is married, which may reflect the societal norm of that era. Another interesting insight relates to the SMSA variable (indicator of the person's living situation), which shows that only a fraction of the people live in an urban area (roughly 18%).

Relevance of the Instrumental Variable

Before we move on with the instrumental variable analysis, it is important to determine whether the selected instrument (quarter of birth) is actually a good instrument for years of education. To do this, the relevance criterion is tested - i.e. whether a change in the instrumental variable causes a (substantial) change in the endogenous variable, which in this case are the years of education. To accomplish this, we can run a linear regression between the years of education and quarter of birth and inspect the coefficient and the p-value. Before running the regression, however, let us visualize the relationship with a bar chart comparing the averages across different quarters of birth:

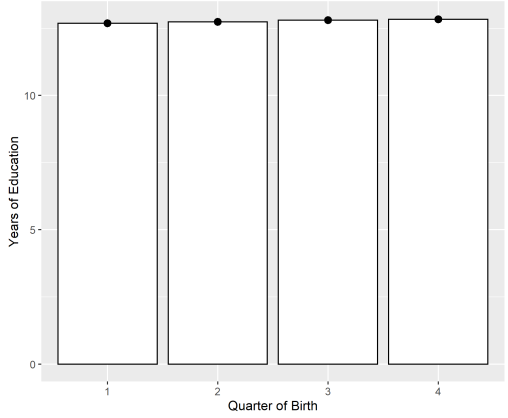

Figure 4: Quarter of Birth Against Years of Education

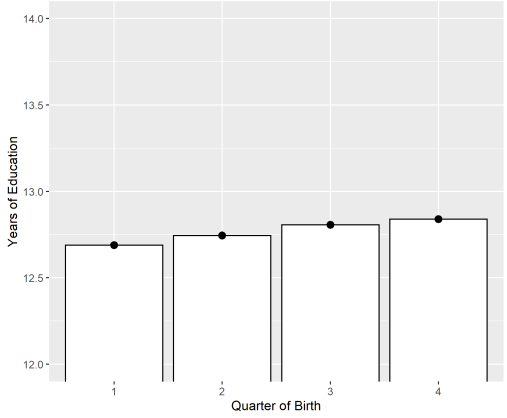

Figure 5: Quarter of Birth Against Years of Education - Zoomed In

Looking at Figure 4, we can barely see any differences between the groups of people born in a different quarter of birth. All groups seem to hover around the average value of 12.5 years of education that was identified in the summary table. Zooming in on Figure 5, however, it can be seen that there are differences in averages present, albeit small. The question, then is whether this positive effect of the quarter of birth on years of education is significant, which can be answered with a simple linear regression:

Table 10: Endogenous Regression

	Dependent variable:
	educ
Constant	12.688***
	(0.011)
qob2	0.057***
	(0.016)
qob3	0.117***
	(0.016)
qob4	0.151***
	(0.016)
Observations	329,509
R^2	0.0003
Adjusted R^2	0.0003
Residual Std. Error	3.281 (df = 329505)
F Statistic	34.009*** (df = 3; 329505)
Note:	*p<0.1; **p<0.05; ***p<0.01

The regression confirms our findings from the plots, wherein a higher quarter of birth has a positive influence on the number of years in education. All the effects are also highly significant ($p < 0.01$) and they are also jointly significant ($F = 34.009, p < 0.01$). Interpreting the results, specifically we can see that compared to people born in the first quarter of the year, people in the second quarter stay in education 0.057 years longer, people in the third quarter 0.117 years longer and people in the fourth quarter 0.151 years longer. People born in the second half have a notably larger coefficient, which may be due to the fact that they reach the legal dropout age earlier (as was established in the assignment description) and can, therefore, leave school earlier. Nonetheless, since the coefficients are all significant, the variable can be said to be relevant, although it does not affect the variable education years to a large extent.

Running the Instrumental Variable Analysis

Now that we have determined that quarter of birth is a suitable instrument, we can run the actual regression. Below are the results of three regressions: a simple instrumental variable estimator model, an instrumental variable estimator model with additional control variables, and finally the model with control variables, but with robust standard errors.

Table 11: Instrumental Variable Analysis Results

	Dependent variable:		
	lnwage		
	(Simple IV)	(IV with control)	(IV with control (robust))
educ	0.103^{***}	0.137^{***}	0.137^{***}
	(0.020)	(0.031)	(0.031)
age		0.008^{***}	0.008^{***}
		(0.002)	(0.002)
Married		0.243^{***}	0.243^{***}
		(0.010)	(0.010)
Urban		-0.106^{***}	-0.106^{***}
		(0.035)	(0.035)
Constant	4.590^{***}	3.603^{***}	3.603^{***}
	(0.249)	(0.488)	(0.489)
Observations	329,509	329,509	329,509
R^2	0.094	0.035	0.035
Adjusted R^2	0.094	0.035	0.035
Residual Std. Error	0.646 (df = 329507)	0.667 (df = 329504)	0.667 (df = 329504)
Note:			*p<0.1; **p<0.05; ***p<0.01

Running the simple instrumental variable estimator analysis yields significant results for the single coefficient that is included ($p < 0.01$). Interpreting the result, it implies that a 1 year increase in the years of education is associated with a 10.3 % increase in wages. When we include additional control variables (specifically the individual's age, their marriage status and whether they live in an urban are), this effect of education years on wages increases - now an additional year is associated with a 13.7 % increase in the wages of an individual ($p < 0.01$). All the other variables are also highly significant ($p < 0.01$) and have a positive effect on wages, except the variable indicating the living situation, which actually suggests that individuals living in an urban area earn 10.6% less. Expanding the model for robust standard errors, it can be seen that the results remain the same and retain their level of significance. It looks like the standard errors did not change between the two models, but that is not the case - the change was simply so small that it is not visible with the three decimal places that are utilized in the table. Surprisingly, when adding control variables to the model, our R^2 and adjusted R^2 actually decreased.

OLS and IV Model Comparison

Finally, let us compare the original OLS model (assuming no endogenous variables) with the IV estimator model that was just created. The OLS model uses the same variables as the last IV model that included the control variables and both the models in the table are using robust standard errors. The results of each of the model are presented in Table 12 below.

Table 12: OLS and IV Model Results

	Dependent variable:	
	lnwage	
	OLS	*instrumental variable*
	(1)	(2)
educ	0.067***	0.137***
	(0.0004)	(0.031)
qob2	0.004	
	(0.003)	
qob3	0.010***	
	(0.003)	
qob4	0.007**	
	(0.003)	
age	0.004***	0.008***
	(0.0004)	(0.002)
Married	0.264***	0.243***
	(0.004)	(0.010)
Urban	−0.184***	−0.106***
	(0.003)	(0.035)
Constant	4.671***	3.603***
	(0.018)	(0.489)
Observations	329,509	329,509
R^2	0.145	0.035
Adjusted R^2	0.145	0.035
Residual Std. Error	0.628 (df = 329501)	0.667 (df = 329504)
F Statistic	8,009.031*** (df = 7; 329501)	
Note:		*p<0.1; **p<0.05; ***p<0.01

Comparing the two models, we can see there are vast differences in the coefficient values, although all of them remain significant in both models ($p < 0.01$) despite the usage of (White's) robust standard errors. Under the OLS model, the effect of the education years is much smaller, with an additional year of education resulting in only 6.7% increase in wages compared to the 13.7% estimated in the IV model. Age also has a smaller effect and the marriage status is associated with a 2% higher increase compared to the IV model. Living in an urban area has even more negative consequences under the OLS model compared to the IV model coefficient.

We can run several tests to evaluate which of the two models to choose and whether our choice to conduct an instrumental variable analysis was a redundant one. Firstly, the "weak instruments" (a partial F-test) is run to compare the model with control variables and the instrument with the model with control variables and without the instrument, and determine whether there is a significant increase in explanatory power. The test results are significant ($F = 14.390, p < 0.01$), so we assume the instrument qob is strong and therefore relevant. This reflects the linear regression results, that were presented in Table 10, as the instrumental variable had small but significant coefficients. The next test is the Wu-Hausman test, which tests for the assumed exogeneity of the independent variables (H0 is that they are indeed exogenous) - if this test is significant, we opt for the IV estimators, which is the case for our model ($p = 0.0184$). Finally, the Sagan-Hansen over-identification test is performed, which checks whether the over-identification of the model violated the cleanness assumption. Note that the Sagan test is only possible in this case because our instrumental variable qob is defined as a factor (and is therefore made up of four dummy variables, $L > K$) - if qob was treated as a numerical variable, it would be necessary to add another instrumental variable to the model to check for over-identification. Since the result is equal to 6.014 and significant ($p = 0.0495$), it seems the exogeneity assumption is violated by the over-identification. From the Table 10 regression results, it can be seen that significant variation occurs between halves of year rather than quarter, so splitting them into quarters may have caused this result. Other than that, the IV model is preferred.

Possible Concerns over the Instrumental Variable Identification Strategy

While the model has favorable results and should be chosen over the basic OLS model, there are still some concerns present over whether the instrument was identified correctly. There are particular ways that the instrumental variable (quarter of birth) can violate the two assumptions necessary to establish an instrument - cleanness and relevance.

Regarding cleanness, in more affluent families, it may be the case that the families attempt to have children at a specific time of year for either superstitious reasons or because of established practice. For example, certain months of the year are associated with career success or better health (this is more of a case of superstition but it has also been proven scientifically to a certain extent). A more realistic reason would be established practice - many middle class and wealthier families believe that if they enter their children into the schooling system later (so if they have have children after the school in-take period), they will naturally have an advantage over the other children as they will be oldest in the class, which will presumably lead to greater academic performance or athletic performance. Since these children also come from these wealthy families, they will most likely have a higher wage later in life - therefore, the quarter of birth can indirectly have an effect on the individuals' wages in this way.

When it comes to relevance, there was already some evidence within the data that quarter of birth is not a very strong instrument as the linear regression in table 10 showed the coefficients for each quarter were quite low. The F-statistic value for the "weak instrument" test was also low despite it being significant. This can be the result of the families' assumption of the advantages of being an older child in the class mentioned above cancelling out the effect on the dropout age (since more developed children should have an easier time staying in school). Children born during this period would also have lived through a number of wars, and conscription into the army happens during certain periods of the year, which means that this could be the underlying cause of the effect on the education years (and wages).

Appendix - R code

```
####Session set-up####
####Session set-up####
library(tidyverse)
library(stargazer)
library(plyr)
library(reshape2)
library(lmtest)
library(lfe)
library(Hmisc)
library(AER)

dir <- "C:/Users/tnguyen10/Private/School/Block 1/ASAP/Assignment 2/"

dirProg <- paste0(dir, "Programs/")
dirData <- paste0(dir, "Data/")
dirResults <- paste0(dir, "Results/")

##Importing the datasets##
labor <- read.csv(file = paste0(dirData, "DiD_dataset-1.csv"), header = TRUE)
dfschooling <- read.csv(file = paste0(dirData, "IV_dataset.csv"), header = TRUE)

####Preparing the data####
dflabor <- labor
str(dflabor)
dflabor$state <- as.factor(dflabor$state)
dflabor$year <- as.integer(dflabor$year)
dflabor$urate <- as.numeric(dflabor$urate)
dflabor$nonwhite <- as.integer(dflabor$nonwhite)
dflabor$work <- as.integer(dflabor$work)
##labor and non-white are factor variables but before running summary statistic,
##I leave them as numeric so I can make conclusions about the general sample (stargazer ignores factor)

#Besides making sure all variables are in correct format, I introduce new ones
#treatment
dflabor$treatment <- with(dflabor, ifelse(children > 0, 1, 0))
dflabor$period <- with(dflabor, ifelse(year > 1992, 2, 1))
dflabor$treatment <- as.factor(dflabor$treatment)
dflabor$period <- as.factor(dflabor$period)

#-------------------------------------------------------------------------------
# Part 1
#-------------------------------------------------------------------------------

####Part 1 - Task 2####
#earn, finc, work#
line <- ggplot(dflabor, aes(year, earn, colour = treatment))

##Creating the plot for annual earnings##
line + stat_summary(geom = 'line') + geom_vline(xintercept = 1993) +
  labs(x = "Year", y = "Annual Earnings") +
  scale_color_manual(labels = c("No children", "Children"),
                     values = c("red", "blue"))
```

```
?ggsave
ggsave(filename = "DiDearning.png", path = dirResults, width = 6, height = 5)

##Creating the plot for family income##
linefinc <- ggplot(dflabor, aes(year, finc, colour = treatment))
linefinc + stat_summary(geom = 'line') + geom_vline(xintercept = 1993) +
  labs(x = "Year", y = "Family Income") + labs(x = "Year", y = "Annual Earnings") +
  scale_color_manual(labels = c("No children", "Children"),
                     values = c("red", "blue"))
ggsave(filename = "DiDfincome.png", path = dirResults, width = 6, height = 5)

##Creating the plot for work##
linework <- ggplot(dflabor, aes(year, work, colour = treatment))
linework + stat_summary(geom = 'line') + geom_vline(xintercept = 1993) +
  labs(x = "Year", y = "Working population ratio") + labs(x = "Year", y = "Annual Earnings") +
  scale_color_manual(labels = c("No children", "Children"),
                     values = c("red", "blue"))
ggsave(filename = "DiDwork.png", path = dirResults, width = 6, height = 5)

##alternative: https://nateapathy.com/2019/08/06/dd-viz/

####Part 1 - Task 3####
stargazer(dflabor, type = "text")
view(filter(dflabor, earn > 300000))


####Part 1 - Task 4####
##Summary table##
dflabor[1:5, ]
n_distinct(dflabor$state)

##Find average annual earnings per period and per women with children/no children##
avgEarn <- ddply(dflabor, .(period, treatment), summarise,
                 avgEarnings = mean(earn, na.rm = TRUE))

##Make table of the outcomes for earnings##
tmp <- dcast(avgEarn, period ~ treatment, value.var = "avgEarnings")
tmp <- rbind(tmp, tmp[2,]-tmp[1,])
rownames(tmp) <- c("Before", "After", "Difference")
tmp[3, "period"] <- NA

stargazer(tmp, summary = FALSE, align = TRUE, type = "text")

##Find average family income per period and per women with children/no children##
avgFinc <- ddply(dflabor, .(period, treatment), summarise,
                 avgFincome = mean(finc, na.rm = TRUE))

##Make table of the outcomes for finc##
tmp2 <- dcast(avgFinc, period ~ treatment, value.var = "avgFincome")
tmp2 <- rbind(tmp2, tmp2[2,]-tmp2[1,])
rownames(tmp2) <- c("Before", "After", "Difference")
tmp2[3, "period"] <- NA

stargazer(tmp2, summary = FALSE, align = TRUE)
```

```r
##Find average working status per period and per women with children/no children##
avgWork <- ddply(dflabor, .(period, treatment), summarise,
                 avgWorkstatus = mean(work, na.rm = TRUE))

##Make table of the outcomes for work##
tmp3 <- dcast(avgWork, period ~ treatment, value.var = "avgWorkstatus")
tmp3 <- rbind(tmp3, tmp3[2,]-tmp3[1,])
rownames(tmp3) <- c("Before", "After", "Difference")
tmp3[3, "period"] <- NA

stargazer(tmp3, summary = FALSE, align = TRUE, type = "text")

####Part 1- Task 5####
##Running the actual regression##
#THe simple models#
str(dflabor)
mdlearn <- earn ~ treatment + period + treatment:period
mdlfinc <- finc ~ treatment + period + treatment:period
mdlwork <- work ~ treatment + period + treatment:period

regearn <- lm(mdlearn, data = dflabor)
regfinc <- lm(mdlfinc, data = dflabor)
regwork <- lm(mdlwork, data = dflabor)

stargazer(regearn, regfinc, regwork, intercept.bottom = FALSE, align = TRUE)

#Expanding the models with control variables - age, ed, urate, state
regearnlong <- lm(earn ~ treatment + period + treatment:period + age + ed
                  + urate + as.factor(state), data = dflabor)
regfinclong <- lm(finc ~ treatment + period + treatment:period + age + ed
                  + urate + as.factor(state), data = dflabor)
regworklong <- lm(work ~ treatment + period + treatment:period + age + ed
                  + urate + as.factor(state), data = dflabor)

stargazer(regearnlong, regfinclong, regworklong, intercept.bottom = FALSE,
          align = TRUE, no.space = TRUE,
          omit = "state")

#Are robust standard errors necessary?
bptest(regworklong)
regworkrob <- felm(work ~ treatment + period + treatment:period + age + ed +
                     urate + as.factor(state), data = dflabor)

stargazer(regearnlong, regfinclong, regworklong, regworkrob,
          intercept.bottom = FALSE, align = TRUE, no.space = TRUE,
          omit = "state")

summary(regworkrob, robust = FALSE)

####Part 1 - Task 6####
dfwchildren <- subset(dflabor, children > 0)
dfwchildren$treatment <- with(dfwchildren, ifelse(ed >= 9, 1, 0))
dfwchildren$treatment <- as.factor(dfwchildren$treatment)
```

```r
str(dfwchildren)

regearn2 <- lm(mdlearn, data = dfwchildren)
regfinc2 <- lm(mdlfinc, data = dfwchildren)
regwork2 <- lm(mdlwork, data = dfwchildren)

stargazer(regearn2, regfinc2, regwork2,intercept.bottom = FALSE, align = TRUE)

#for the sake of clarity, treatment is rewritten as education in the paper, but it serves the same purpo

##Subsetting on low education##
dflow <- subset(dflabor, ed < 9)
regearn3 <- lm(mdlearn, data = dflow)
regfinc3 <- lm(mdlfinc, data = dflow)
regwork3 <- lm(mdlwork, data = dflow)

stargazer(regearn3, regfinc3, regwork3,intercept.bottom = FALSE, align = TRUE,
          type = "text", report=('vc*p'))
stargazer(regearn3, regfinc3, regwork3,intercept.bottom = FALSE, align = TRUE)


#------------------------------------------------------------------------------
# Part 2
#------------------------------------------------------------------------------
####Preparing the data - Part 2####
dfedu <- subset(dfschooling, select = c(age, educ, lnwage, married, qob, SMSA, yob))
str(dfedu)
stargazer(dfedu)
dfedu$married <- as.factor(dfedu$married)
#also choosing to define qob as categorical variable since it's a nominal variable
dfedu$qob <- as.factor(dfedu$qob)
dfedu$SMSA <- as.factor(dfedu$SMSA)

####Part 2 - Task 3####
#So one way we can check for relevance is to run a linear regression of
#the endogenous x instrumental
bar <- ggplot(dfedu, aes(qob, educ))

##Creating the plot for education years against the quarter of birth##
bar + stat_summary(fun.y = mean, geom = "bar", fill = "white", colour = "Black") +
  stat_summary(fun.data = mean_cl_normal, geom = "pointrange") +
  labs(x = "Quarter of Birth", y = "Years of Education") +
  coord_cartesian(ylim = c(12,14))
?ggsave
ggsave(filename = "Barzoomed.png", path = dirResults, width = 6, height = 5)

##running linear regression models##
lmqob <- lm(educ ~ qob, data = dfedu)
stargazer(lmqob, intercept.bottom = FALSE, align = TRUE)

####Part 2 - Task 4 - Running the IV regression####

##IVREG##
rslt2SLS.A <- ivreg(lnwage ~ educ|qob, data = dfedu)
```

```r
summary(rslt2SLS.A)
stargazer(rslt2SLS.A, type = "text")

##Adding other control variables - age, married, SMSA(but still keeping qob as the sole instrument)##
rslt2SLS.B <- ivreg(lnwage ~ educ + age + married + SMSA|qob + age + married +
                      SMSA, data = dfedu)
stargazer(rslt2SLS.B, type = "text")

#Adding control variables actually improved the results - now one additional year
#year of education is associated with a 13.7% increase in wages

##Creating robust standard errors (White)##
seWhite <- sqrt(diag(vcovHC(rslt2SLS.B, type = "HC0")))

stargazer(rslt2SLS.A, rslt2SLS.B, rslt2SLS.B, se = list(NULL, NULL, seWhite))
#seemingly no change but when we check in more detail
seWhite
summary(rslt2SLS.B)
#there is a change but it is in very small decimal

####Part 2 - Task 5####
#Create the exact same model but now with OLS instead of IV - we create the one
#with the control variables and compare (also includes robust errors)
modelOLS <- lm(lnwage ~ educ + qob + age + married + SMSA, data = dfedu)
seWhiteOLS <- sqrt(diag(vcovHC(modelOLS, type = "HC0")))
stargazer(modelOLS, rslt2SLS.B, se = list(seWhiteOLS, seWhite))
stargazer(modelOLS, rslt2SLS.B, type = "text")

summary(rslt2SLS.B, diagnostics = TRUE)
```