Individual Assignment 2

Eran Kan Kohen

29th September 2022

1 Difference in Difference

1.1 Equation

There are two expected outcomes for the model:

$$E = (y|D = 1) = \beta 0 + \beta 1 + (\beta_2 + \beta_3)T_t$$

$$E = (y|D = 0) = \beta 0 + \beta_2 T_t$$

Hence:

$$E = (y_T = 1|D = 1) = \beta 0 + \beta 1 + \beta_2 + \beta_3$$

$$E = (y_T = 0|D = 1) = \beta 0 + \beta 1$$

$$E = (y_T = 1|D = 0) = \beta 0 + \beta_2$$

$$E = (y_T = 0|D = 0) = \beta 0$$

The values above are plugged into the the difference in difference equation below:

$$[E(y_T = 1|D = 1) - E(y_T = 0|D = 1)] - [E(y_T = 1|D = 0) - E(y_T = 0|D = 0)]$$

$$[(\beta 0 + \beta 1 + \beta_2 + \beta_3) - (\beta 0 + \beta 1)] - (\beta 0 + \beta_2 - \beta 0)$$

$$(\beta_2 + \beta_3) - \beta_2$$

$$\mathrm{E} = \beta_3$$

1.2 Plots

In order to investigate the data set, multiple plots regarding the highlighted variables were generated.

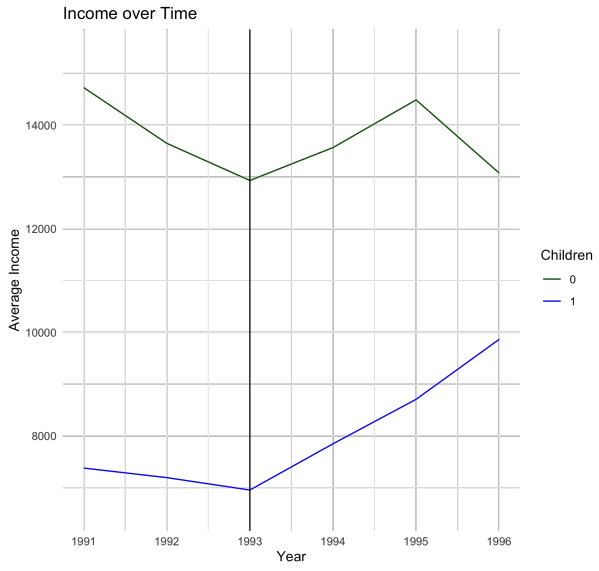

Figure 1: Average Income DiD

The first plot highlights the difference-in-difference effect for average income. As seen above, before the intervention, the average income was decreasing for all women and the intervention had a positive effect with women starting to earn more. However, there was a difference between women with

children and women without children. In 1996, women with children kept the positive slope whereas women without children had a negative slope and saw the average income decrease.

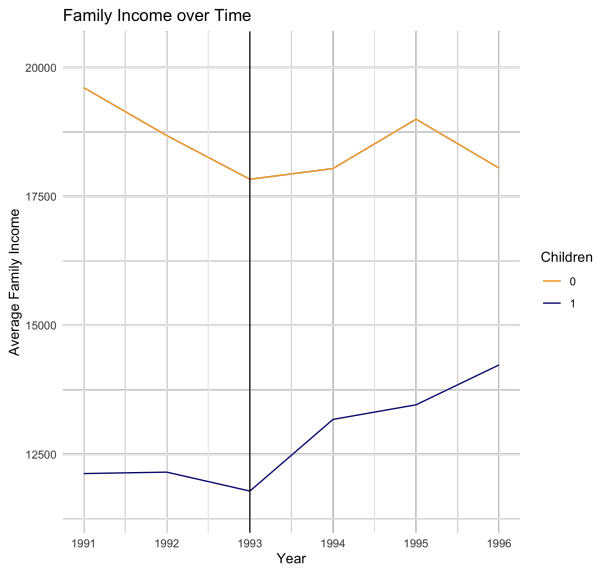

Figure 2: Average Family Income DiD

The second plot is about the family income and it mirrors the first graph to a degree. Similar to the first plot, the intervention benefited all women until 1996. Afterwards, women with children returned to their average wage before the intervention.

Figure 3: Working Percentage DiD

The final plot focuses on the working percentage. As seen above, around 57% of women without children were working in 1991, whereas the rate was around 46% for women without children. The intervention initially boosted the employment rate for both demographics, however, the rate started to drop for women without children after the year 1994. The rate for women without children was 55% in 1996 and 50% for women with children. The intervention closed the gap.

1.3 Summary

After the plot analysis, a summary table of the variables were generated.

Table 1:

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
inter_binary	13,746	0.632	0.482	0	1

As seen above, the data set has 13746 separate data points. For all of the variables, the standard deviation is almost as high or even higher than the mean which indicates a great deal of spread in the data. Moreover, for both the Annual Family Income and Annual Income, the mean is very low compared to the maximum value. Overall, it can be said that the data is very spread over but most of the population is on the low side.

The table also highlights information on other variables such as the amount of children and the race of the woman. One of the interesting findings from this data is that a woman has 9 children.

The variable inter_binary represents the data points before and after the intervention. 1 stands for After and 2 stand for before. The mean for the variable is 1.368, which indicates that there are more values after the intervention than before the intervention.

1.4 Difference-in-Difference Effects

In order to illustrate the difference-in-difference effects better than the plots, tables for each of the dependent variables were generated.

Table 2: DiD Effect on Average Income

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

The first table highlights the difference-in-difference effect of the intervention for the average income. The table shows the change of income for women with and without children. As seen above, the intervention helped women with children and fulfilled its intention. However, it seems like it backfired for women without children as the income has decreased.

Table 3: DiD Effect on Average Family Income

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

The table above shows the intervention effects on the average family income. As seen above, similar to the average income, the intervention seems to have benefited women with children while deteriorating women without children.

The final graph reflects the intervention effect on the working percentage for women. Similar to the previous graphs, the intervention seems to have helped women with children and not women without children. Moreover, it

Table 4: DiD Effect on Working Percentage

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

seems that overall, women without children earn more and work more than women with children.

1.5 Difference-in-Difference Regression

The previous section investigated the difference-in-difference effect between the variables of children and intervention. This illustrated the effect of the EITC policy. The regression models below further investigates the difference-in-difference effect with different control variables according to the dependent variable.

1.5.1 Average Income

Table 5:

	Dependent variable:			
	earn			
	(1)	(2)	(3)	(4)
inter_binary	−1,569.338***	−695.997	−664.001	−667.264
	(587.648)	(485.413)	(485.255)	(485.277)
children_binary	−8,981.633***	−6,913.517***	−6,759.319***	−6,783.592***
	(607.270)	(510.988)	(512.384)	(513.247)
nonwhite			−1,192.168***	−1,163.538***
			(316.693)	(318.621)
ed				48.054
				(58.686)
inter_binary:children_binary	2,811.027***	1,682.810***	1,720.579***	1,717.765***
	(760.648)	(642.099)	(641.870)	(641.887)
Constant	15,539.450***	14,203.900***	14,798.600***	14,375.110***
	(471.276)	(387.548)	(418.338)	(665.200)
Observations	9,435	13,746	13,746	13,746
R^2	0.041	0.026	0.027	0.027
Adjusted R^2	0.040	0.026	0.027	0.027
Residual Std. Error	17,449.930 (df = 9431)	17,965.670 (df = 13742)	17,957.060 (df = 13741)	17,957.280 (df = 13740)
F Statistic	133.522*** (df = 3; 9431)	121.691*** (df = 3; 13742)	94.899*** (df = 4; 13741)	76.051*** (df = 5; 13740)
Note:				*p<0.1; **p<0.05; ***p<0.01

The first table looks at the regression model for variable the variable earn. The first two columns has the initial model, the third model includes the race of the women as a binary variable indicating whether the women is white or not. The final model adds the education level of the women as well.

The difference between the first two columns is due to the first model on the left only focusing on a subset of women with high education. This model also happens to have the highest R^2 and all of its component are statistically significant. It indicates that for women with high education, there is a negative correlation between their income and both the intervention and having children.

The three remaining models have very similar R^2 values and not much separating them. Moreover, the standard errors for the models are quite similar with not much distinguishing them. It seems that education is not a statistically significant factor in this relationship whereas the race of the

women is.

1.5.2 Average Family Income

Table 6:

	Dependent variable:		
	finc		
	(1)	(2)	(3)
inter_binary	−940.239*	−974.915*	−714.400
	(519.486)	(519.148)	(485.452)
children_binary	−7,018.295***	−6,584.122***	−6,809.680***
	(546.857)	(554.320)	(518.328)
age		78.423***	19.513
		(16.817)	(15.781)
unearn			959.977***
			(21.590)
inter_binary:children_binary	1,911.035***	1,941.117***	1,699.430***
	(687.171)	(686.683)	(642.090)
Constant	19,159.190***	16,162.160***	13,656.510***
	(414.751)	(764.740)	(717.269)
Observations	13,746	13,746	13,746
R^2	0.022	0.024	0.147
Adjusted R^2	0.022	0.024	0.146
Residual Std. Error	19,226.750 (df = 13742)	19,212.250 (df = 13741)	17,963.960 (df = 13740)
F Statistic	105.245*** (df = 3; 13742)	84.489*** (df = 4; 13741)	472.720*** (df = 5; 13740)
Note:			*p<0.1; **p<0.05; ***p<0.01

The second table sheds a light on the difference-in-difference effect for the family income. The table includes models which add the age and the unearned income of the women into the equation. The latter model has the highest R^2 among all of the models with 15%. It seems that the unearned income explains a great deal of the variation in family income.

Furthermore, the models all seem to have similar standard error values, with the notable exception being the standard error value for the constant of the initial model being lower compared to the other models.

1.5.3 Working Proportion

Table 7:

	Dependent variable:			
	work			
	(1)	(2)	(3)	(4)
inter_binary	−0.019	−0.005	−0.030**	−0.025*
	(0.016)	(0.013)	(0.014)	(0.014)
children_binary	−0.131***	−0.128***	−0.124***	−0.118***
	(0.017)	(0.014)	(0.014)	(0.014)
urate			−0.025***	−0.021***
			(0.003)	(0.003)
nonwhite				−0.050***
				(0.009)
inter_binary:children_binary	0.047**	0.031*	0.032*	0.034*
	(0.021)	(0.018)	(0.018)	(0.018)
Constant	0.590***	0.577***	0.757***	0.754***
	(0.013)	(0.011)	(0.025)	(0.025)
Observations	10,502	13,746	13,746	13,746
R^2	0.010	0.012	0.016	0.019
Adjusted R^2	0.010	0.012	0.016	0.018
Residual Std. Error	0.497 (df = 10498)	0.497 (df = 13742)	0.496 (df = 13741)	0.495 (df = 13740)
F Statistic	36.620*** (df = 3; 10498)	54.906*** (df = 3; 13742)	57.324*** (df = 4; 13741)	52.169*** (df = 5; 13740)
Note:				*p<0.1; **p<0.05; ***p<0.01

The final regression table focuses on the working population of women. Similar to the first table, the first two columns deal with the initial model. The other two columns gradually add the unemployment rate and the race of the women into the model.

The first column deals with a very specific subset. An analysis of the unemployment revealed that 75% of the women lives in a state with less than 7.7% unemployment. In order to remove the remaining 25% who live in high unemployment, the first model only focuses on the women aforementioned women who live in states less than 7.7% unemployment. This seems to have been a bad decision since both the standard errors and the R^2 have deteriorated compared to the initial model (column 2).

The remaining models have almost identical standard error values and very similar R^2 values which all round up to 1% and 2%. It is saddening

that once again the race of the women is a statistically significant factor in the working rate of women.

1.6 Education Subsections

1.6.1 High vs Low Education for Parents

The difference-in-difference analysis was expanded further by investigating the education level of women. In order to analyze the the effect of education, the women having children was held constant. Three separate tables for the previously investigated variables were generated.

Table 8: Average Income of Women with Children

	Low Edu	High Edu
Before	$9,113.888$	$6,557.819$
After	$9,484.115$	$7,799.508$
Difference	370.227	$1,241.689$

The first table highlights the difference in average income of women before and after the intervention. The table draws a very clear picture with highly educated women having a much greater difference in their income. While less educated women saw a 370 dollar increase in their average income, high educated women saw a 1200 dollar increase. This is a huge difference, with underlines the importance of education.

However, another interesting fact is that less educated women were and still are making more money then high educated women. Maybe more education does not correlate with high income?

Table 9: Average Family Income of Women with Children

	Low Edu	High Edu
Before	13,897.890	11,435.060
After	14,040.640	12,744.020
Difference	142.756	1,308.968

The second table focuses on family income. Similar to the previous table, the intervention seems to have boosted women with higher education. The families of women with higher education has benefited a great deal from the intervention whereas the families of less educated women have seen an increase of 1% through the intervention. It can even be argued that the intervention has not helped the families, but the wages were increased over time through inflation.

Table 10: Working Percentage of Women with Children

	Low Edu	High Edu
Before	0.433	0.457
After	0.444	0.489
Difference	0.011	0.032

The final table looks at the working percentage. Again, the table mimics the previous ones and shows that the intervention has helped women with higher education. An expected result from the table shows that a greater proportion of highly educated women are employed compared to women with less education. This is not surprising since higher education is likely to have helped the women find jobs.

However, this finding is very interesting when combined with the insights gathered from the previous tables. The previous tables have shown that women with less education make more money and their families have an

higher income as well. Apparently this is in spite of a lower proportion of less educated women working. This indicates that women with less education receive much better wages compared to women with more education.

The findings shown in the tables above were summarized below through their regression analyses.

Table 11: High vs Low Education Regression Analyses

	Dependent variable:		
	earn	finc	work
	(1)	(2)	(3)
inter_binary	370.227	142.756	0.011
	(653.474)	(688.073)	(0.022)
edu_factor1	$-2{,}556.068^{***}$	$-2{,}462.830^{***}$	0.024
	(611.812)	(644.206)	(0.020)
inter_binary:edu_factor1	871.461	1,166.212	0.021
	(773.065)	(813.996)	(0.026)
Constant	$9{,}113.888^{***}$	$13{,}897.890^{***}$	0.433^{***}
	(516.756)	(544.116)	(0.017)
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

The DiD effects are clearly visible in the third row of the table. It should be highlighted that the R^2 values are quite low which indicates that the models are not very explanatory of the relationship.

1.6.2 Children vs No Children for Low Educated

The final analysis of the section focuses on less educated women. The following tables look at the three dependent variables and compares women with and without children.

Table 12: Average Income of Less Educated Women

	No Children	Children
Before	11,850.380	9,113.888
After	12,634.050	9,484.115
Difference	783.677	370.227

The initial table looks at average income of women with low education. It seems that women without children have higher income and were more positively affected by the intervention. This is probably due to women without children being able to focus more on their careers.

Table 13: Average Family Income of Less Educated Women

	No Children	Children
Before	17,816.920	13,897.890
After	18,139.310	14,040.640
Difference	322.393	142.756

The second table looks at the average family income. One again, this table is very similar to the average income table as it shows families without children making more money compared to families without children. The intervention has had a greater positive influence on families without children as well.

Table 14: Working Percentage of Less Educated Women

	No Children	Children
Before	0.497	0.433
After	0.493	0.444
Difference	-0.004	0.011

Finally, last table focuses on the working percentage. It seems that a greater proportion of women without children are working compared to women with children. However, after the intervention the rates did not change much. The difference for women without children is -0.4% and for women with children is 1.1%. This is a very small difference and hence is not substantial enough to make valid claims.

Similar to section 1.6.1, the findings were summarized below by their regression analyses.

Table 15: Children vs No Children Regression Analyses

	Dependent variable:		
	earn	finc	work
	(1)	(2)	(3)
inter_binary	783.677	322.393	−0.004
	(858.662)	(903.937)	(0.023)
children_binary	−2,736.488***	−3,919.035***	−0.065***
	(945.162)	(994.998)	(0.025)
inter_binary:children_binary	−413.449	−179.637	0.015
	(1,194.473)	(1,257.455)	(0.031)
Constant	11,850.380***	17,816.920***	0.497***
	(679.840)	(715.686)	(0.018)
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

The table above clearly shows the DiD effects in the third row.

2 Instrumental Variable Analysis

2.1 Possible Biases

The following section will analyse the relationship between education and wage for men people born in the US between 1930 and 1939. Before beginning in the analysis, it is important to underline two possible biases which could have affected the data in unobserved ways.

The first bias could occur due to the Korean and Vietnamese wars. The sample size would most probably be drafted for both of the wars and their participation in the army could have halted their education or could have paused their working careers. Both of these possible interruptions could have contaminated the data.

The second bias could derive from the lack of information surrounding the degree of the person. The data lacks any indicator of which degree the person acquired or the highest level of education one completed. It is possible that the years of education is disrupted by people who had to re-take a year of education, or people who had to study longer due to their degree. It is very possible that someone who studied a specific degree is earning more than another person with a different degree.

2.2 Summary

Before the analysis, the data was investigated.

Table 16: IV Data Summary

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

The table above presents a summary of the relevant variables. As seen clearly, the data set has 329509 data points and gives insight on factors such as education, wage, and marriage status. It should be highlighted that wage is recorded as in the log of the weekly wage. So, the minimum wage of -2.3

is not someone who loses money every week.

Moreover, the table shows that the age of the data ranges from 40 to 50, indicating that the data was pulled in 1980. The average years to education spent being 12.7 years. It is saddening that there is a data point with 0 years of education. Finally, SMSA is a binary variable indicating the data points' living condition. The mean value of 0.18 indicates that most of the data points are not living in urban areas.

2.3 Relevance Criterion

The relevance criterion indicates that the instrumental variables in the model should highly correlate with the independent variable. In order to test this before adding qob as an instrumental variable, the correlation between the variables were tested.

Table 17: OLS Regression between Education and Qob

	Dependent variable:
	educ
qob	0.052^{***}
	(0.005)
Constant	12.641^{***}
	(0.014)
Observations	329,509
R^2	0.0003
Adjusted R^2	0.0003
Residual Std. Error	3.281 (df = 329507)
F Statistic	100.653^{***} (df = 1; 329507)
Note:	*p<0.1; **p<0.05; ***p<0.01

The table above shows the results of the OLS regression between qob

and education. As seen clearly, the relationship is statistically significant, however both the coefficient and the R^2 is quite low.

Moreover, a Pearson Correlation test revealed that there is a significant correlation between the variables. However, the correlation coefficient is 0.017, which is very low. In order to better illustrate the relationship, a bar chart was generated.

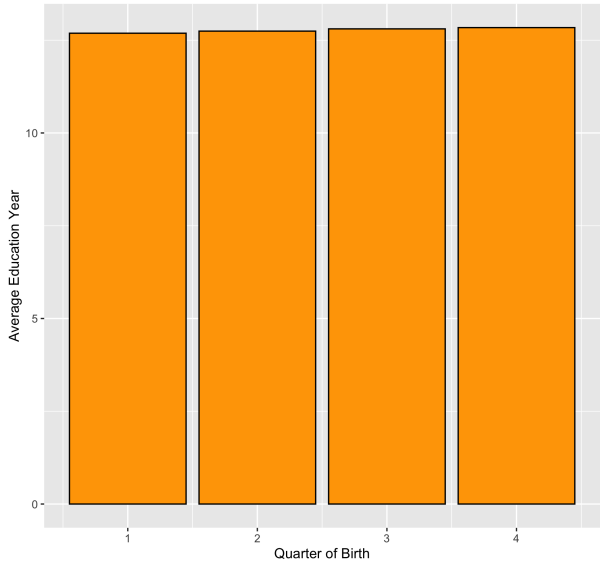

Figure 4: Average Education Years in Qobs

As seen from the figure above, even though there is a statistically significant correlation, the average education per qob is very similar. If looked very closely, one can see that from left to right, there is a very small gradually increase, which indicates that people born on the last quarter tend to have a longer education journey compared to people who were born on the first quarter. However, as stated, this is a very small difference.

2.4 IV Regression

2.4.1 Regression Comparison

Initially, the instrumental variable regression was ran with qob as an instrumental variable, added to highlight the relationship between education and lnwage. In order to further investigate the relationship, two separate models with control variables were also generated. The first addition included age as a control variable whereas the second had marriage as a control variable. The results are presented below.

Table 18: IV Regression Models

	Dependent variable:		
	lnwage		
	(1)	(2)	(3)
educ	0.099***	0.165***	0.099***
	(0.020)	(0.035)	(0.019)
age		0.010***	
		(0.002)	
married			0.249***
			(0.006)
Constant	4.633***	3.327***	4.419***
	(0.250)	(0.536)	(0.244)
Observations	329,509	329,509	329,509
R^2	0.098	−0.088	0.114
Adjusted R^2	0.098	−0.088	0.114
Residual Std. Error	0.645 (df = 329507)	0.708 (df = 329506)	0.639 (df = 329506)
Note:			*p<0.1; **p<0.05; ***p<0.01

The table above illustrates three distinct regression models. The initial model shows the relationship between education and lnwage, with qob acting as an instrumental variable for education. As seen from the coefficient, an

extra year of education leads to 0.1 units more lnwage. Considering that the standard deviation is 0.68 units of lnwage in the data, a 0.1 unit increase can be considered substantial.

The other two models use the initial groundwork but try to improve it by adding control variables. The second model which adds age as a control variable has a very interesting result. Even though both education and age are statistically significant the R^2 is negative. This indicates that the model fits worse than a horizontal line. The final model includes marriage as a control variable and, once again, has all of the variables as statistically significant. Moreover, the R^2 is higher compared to the initial model. The standard errors are slightly less as well.

Overall, it can be stated that marriage is a good control variable as it explains a decent part of the variance and improves the model. However, the model without marriage itself is also decent as it explains 10% of the variance in lnwage.

2.4.2 Heteroskedasticity

Before concluding the section, the standard errors of the model was checked to meet the heteroskedasticity assumption. Initially a Breusch-Pagan test was conducted which revealed that the model is heteroskedastic. Hence, robust standard errors were checked in the model. The check revealed that the standard errors remain constant even with robust standard errors, thus it can be stated that robust standard errors do not affect the inferences.

2.5 OLS Regression and Comparison

2.5.1 Regression Comparison

After investigating the IV regression models, an OLS model was also generated. In order to find the optimal model for the data set and goal, the models were compared in the table below.

Table 19: OLS vs IV Regression

	Dependent variable:	
	lnwage	
	OLS	*instrumental variable*
	(1)	(2)
educ	0.071***	0.099***
	(0.0003)	(0.020)
Constant	4.995***	4.633***
	(0.004)	(0.250)
Observations	329,509	329,509
R^2	0.117	0.098
Adjusted R^2	0.117	0.098
Residual Std. Error (df = 329507)	0.638	0.645
F Statistic	43,782.560*** (df = 1; 329507)	
Note:		*p<0.1; **p<0.05; ***p<0.01

The table above compares a simple OLS model which investigates the relationship between education and lnwage with the instrumental regression model presented in section 2.4. As seen above, the OLS model has lower standard error values and has a higher R^2 score. This could indicate that the OLS is a better fit for the data and the relationship in question. However, one can't be sure of this claim until conducting further tests.

2.5.2 Hausman Test

In order to substantiate the claim in the previous section, a Wu-Hausman test was carried out. The test had a p value of 0.143. This score is not able to satisfy either of the expected scores of 0.1 or 0.05. Hence, the results indicates that the model with the instrumental variable is not better than the simple OLS model.

On the other hand, the diagnostics test showed that the weak instruments test is significant. So, it can be stated that even though a simple OLS model should be preferred for the analysis, qob is a good instrumental variable for education.

2.6 Assessing the Strategy

The analysis has been completed and insight has been gathered on the relationship between education, qob, and wage. However, as a final note, the possible biases surrounding the instrumental regression should be analysed. There is a chance that due to an unobserved reason, the data could have been contaminated or the logic was hurt.

For example, in section 2.1, it was mentioned that between the years of 1930 and 1980, the US was heavily involved in the Korea and Vietnam wars. There is a good chance that the wars affected the data, either in the form of people drafted in the army causing unobserved variance, or through attrition bias. Attrition bias concern types of errors which occur due to a non-random sample selection. There is a chance that army veterans were not included in the analysis or people who would have been included did not survive due to the war.

Moreover, it is possible that, due to an omitted variable, a certain aspect of the relationship may have been left unexplored. For example, it is likely that different parent profiles value education more than others. Any information regarding the parents of the sample could have shown a completely different aspect of the explored relationship. For instance, it could have been found that wealthy parents' children tend to study longer whereas children from less wealthy families study much less. This variable could have acted as a useful instrumental variable.