

# ASAP Assignment 2

Angelo Barisano; 508903

September 28th, 2022

# 1 Difference in Difference

## 1.1 Task 1: Derive Diff-In-Diff Coefficients

The canonical difference-in-difference equation is expressed in regression form (1-1):

$$(1-1) \ y_{it} = \beta_0 + \beta_1 D_i + \beta_2 T_t + \beta_3 D_i T_t + \epsilon_{it}$$

The terms in 2, 3, & 4 are relevant in describing the potential outcome assumption in difference in difference analysis. Difference in difference suggests that we compare the difference between treatment and control before and after the treatment introduction ( $t = 0$ ), shown as:

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

Subsequently, the four outcomes described in (1-2) yield the following outcomes:

- $E = (y_{T=1}|D=1)$ ; Because we are observing the outcome for the post-(treatment) test for treatment group, this yields  $\beta_0, \beta_1, \beta_2, \beta_3$
- $E = (y_{T=0}|D=1)$ ; Because we are observing the outcome of the pre-(intervention) test for the treatment, this will yield  $\beta_0, \beta_1$
- $E = (y_{T=1}|D=0)$ ; Because we are observing the outcome for the post-(treatment) test for the control group, this yields  $\beta_0, \beta_2$
- $E = (y_{T=0}|D=0)$ ; Because we are observing the outcome of the pre-(intervention) test for the control, this will yield  $\beta_0$

Note that the disturbance term ( $\beta_0$ ) is left out as it is assumed to be independent of the treatment allocation; as such we will ignore it here. Additionally, the constant is present in all four outcomes; thus, it also technically cancels out in the following equation. This originates from the assumption that the covariates measurements in pre- and post test yield the same results for both treatment and control subjects. Equivalently, by substitution the aforementioned outcomes into (1-2), the following results can be deduced (excluding the constant):

$$(1-3) \ E(y_{it}) = [E(y_{T=1}|D=1) - E(y_{T=0}|D=1)] - [E(y_{T=1}|D=0) - E(y_{T=0}|D=0)]$$

$$(=) \ E(y_{it}) = [(\beta_1 + \beta_2 + \beta_3) - (\beta_1)] - [(\beta_2)]$$

$$(=) \ E(y_{it}) = (\beta_2 + \beta_3) - \beta_2 = \beta_3$$

Table 1: Descriptive Statistics of Numeric Independent and Dependent Variable

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Unemp. Rate	6.762	1.462	2.600	5.700	6.800	7.700	11.400
Number Children	1.193	1.382	0	0	1	2	9
Family Income	15,255.320	19,444.250	0.000	5,123.418	9,636.664	18,659.180	575,616.800
Earnings	10,432.480	18,200.760	0.000	0.000	3,332.180	14,321.220	537,880.600
Age	35.210	10.157	20	26	34	44	54
Education (y)	8.806	2.636	0	7	10	11	11
Unearned Income	4.823	7.123	0.000	0.000	2.973	6.864	134.058

Notes: N = 13746

## 1.2 Task 2: Provide Graphical "evidence" for the presence of the DiD effect

Figure 1 displays the outcomes for both treatment and control group for earnings (earn), family income (finc), and work-participation indicator (work) over the period from 1991 to 1996. The treatment in this case is defined as a binary state of females with children (one or many) ( $D = 1$ ) and females without children ( $D = 0$ ). The tax credit (EITC) is assumed to be introduced on the 1st of January 1993.

Figure 1 a) displays the development of earnings and the introduction of the EITC is marked by the

Table 2: Diff-in-Diff Matrix

	dperiod	Earning		Family Income		Work Participation	
		Childless	Children	Childless	Children	Childless	Children
Before	1	14,203.900	7,290.380	19,159.190	12,140.900	0.580	0.450
After	2	13,507.900	8,277.200	18,218.950	13,111.690	0.570	0.480
Difference		-696.000	986.810	-940.240	970.800	-0.010	0.030

Notes: N = 5927 Childless; N = 7819 Has one or more Children.

vertical line (as in the other graphics). Females without children earned considerably more on average than females with children (pre-period  $mean = 7290.38$ ) during the pre-period, starting out at slightly less than \$14800 and dropping to \$13000 at its lowest, with an average pre-intervention earnings of \$14203.90. Overall, both groups displayed a downward trend before 1993. After 1993, the supposed introduction of the EITC, both groups experienced an increase in earnings; However, the childless group did not display a continuous increase in earnings (post-period  $mean = 13507.90$ ) when compared to females with children (post-period  $mean = 8277.20$ ). This might suggest that the introduction of the EITC is associated to the average earnings of women with children but not for women without children. Figure 1 b) shows the same features as the plot before but the outcome is now Family Income (numbers discussed in exercise 3 and 4). Overall, the general trend is similar for both groups as in Figure 1 a); it may be noted that the trend for females with child is less pronounced than in a). Nonetheless, the conclusion is the same as in a): During the pre-period, childless women earned more, but decreased stronger than women with children, while the post treatment movement is not continuous (dropping to pre treatment levels) for childless women when compared to women with children, which experience continuous growth on average.

Figure 1 c) displays workparticipation of females with and without children. Again, childless women start out higher at around 54% workparticipation vs 46% workparticipation for women with children. Moreover, the trends are somewhat comparable to the two preceding graphics, as females with children display an increase to 50.3% (and then a slight drop in the following year), while childless women tend to decrease: As such, women with children again experience an average net gain in work participation during the post period, while childless women drop below pre-treatment levels.

### 1.3 Task 3: Summary Statistics for data

The dataset is not balanced and we cannot assure that it is fixed. Nonetheless, for the purpose of this exercise, it is assumed that it is balanced and fixed. The data contains yearly records from 1991 to 1996; the reported values per variable are thus averages over all six years. A total of 13746 records were made, split by year into 2610, 2449, 2342, 2255, and 2085 observations respectively. Of the total 13746 observations, 7819 have one or more children, and 5927 have no children.<sup>1</sup> Of the relevant outcome variables (earnings, family income, work participation), the average family income was \$15255.32 ( $std = 19444.24$ ,  $median = 9636.66$ ). This large spread around the mean was partially addressed by Figure 1 a) displaying a large difference between women with and without children; this suggests a strong heterogeneity in the data wrt. to the outcomes. Additionally, as was to be expected by the financial nature of family income, there is a severe positive skewness of 7.06. Earnings displays a similar pattern as family income ( $mean = 10432.48$ ,  $std = 18200.76$ ,  $median = 3332.18$ ,  $skew = 6.766$ ), which is supported by Figure 1 b). As expected, earnings and family income correlate strongly with a Pearson correlation coefficient of .93 ( $p < 0.001$ ); thus, earnings and family income are assumed to behave similarly as outcome variables. Finally, work participation shows that 51.3% of records over all six years are in employment (std. and median have no meaning here as it is a binary variable). Considering possible covariates, the average years in education is 8.8 ( $std = 2.636$ ), with 11 years in the median. On average, per year, one woman had 1.19 children ( $std = 1.382$ ,  $median = 1$ ), with 56.9% of respondents having one child or more (N = 7819). Of the respondents 60% were people of colour. Unemployment rate by state was on average 6.76% ( $std = 1.462$ ,  $median = 7.7$ ) and unearned income at \$4823 ( $std = 7123$ ,  $median = 6864$ ).

### 1.4 Task 4: Diff in Diff Matrix

Table 2 reports the pre-post intervention averages for the two groups - Women with children and women without children - for earnings, family income, and work participation proportion. As shown in Figure 1 a) childless women have on average higher earnings in both pre- and post period than women with

<sup>1</sup> additional summary statistics can be found for women with and without children in Appendix 4.1

children. However, only women with children show a positive average gain over the post intervention period of \$986.82, when compared to the average over the pre treatment period, while childless women even drop in earnings in the post period. Using formula (1-2) this results in a *naïve* DiD effect of \$1682.81 for women with children. Again, a similar observation can be made for family income which displays a DiD of \$1911.04. Finally, work participation also followed the trend observed in Figure 1 c), suggesting a DiD effect of 0.04 (or 4% point gain ov women with children over childless women). Please note that these are just pre-post averages of the groups, complementing the descriptive statistics from part 2-3 **maybe describe a bit more**

## 1.5 Task 5: Analyze the DiD regression

### ADD POPULATION REGRESSION

#### Tutoial 1 plot for heteroscedasiticty fitted vs standardized residuals

**No model without theory** Considering possible covariates, earnings, workparticipation, and family income all tend to vary with age (inverted u). Moreover, higher unemployment implies more labour supply and suppression of wages, family income and (obviously) work participation. Furthermore, higher education positively correlates with earnings, family income, and workparticipation. Finally, the US is known for racial tensions; people of colour tend to show lower incomes and lower work participation in general.

#### CITE A COUPLE OF THINGS FOR THIS!!

**Results** For brevity reasons, Table 3 contains models for all three outcome variables, with the first model per outcome (1, 4, 7) containing a baseline model with only covariates. Models 2, 5, 8 contain only the DiD effects, and Models 3, 6, 9 contain the full model respectively.

Considerig earnings, the covariate model (1) shows only that nonwhite (indicator variable) ( $\hat{\beta} = -2037.990$ ,  $p < 0.001$ ) and age ( $\hat{\beta} = 96.038$ ,  $p < 0.01$ ) are significant. Overall, this model explains 0.6% of the variation in earnings. Proceeding to only the DiD effect (model 2), as expected, women with children earn on average \$8596.33 ( $p < 0.01$ ) less than childless women.<sup>2</sup> Subsequently, the value from Table 2 for the DiD effect is statistically significant at conventional levels, suggesting that women with children gain a \$1682.81 ( $p < 0.01$ ) in the post treatment/intervention period (compared to pre period) (1993 <) when compared to childless women; overall explaining 2.6% of the total variation in earnings. This effect is important to be examined more closely: As implied in task 4, while childless women on average earn \$ 8596.33 more than women with children, childless women (see Table 2) drop by \$696 when comparing the before and after period, while women with children increased. These two effects cancel each other out, most likely leading to the insignificant period coefficient. Moreover, this suggests that the comparison groups (women with vs women without children) are not comparable; as such we will not be able to discern causality from this exercise. Finally, model (3) considers the full model. Focusing on the DiD effect after including covariates, we see a slight increase in the coefficient ( $\hat{\beta} = 1722.360$ ,  $p < 0.01$ ) with stable standard errors; suggesting that women with children in the post period earn \$1722.36 more than childless women. However, the inclusion of the covariates does not suggest a considerable change in the direction of the outcome. Moving to family income as an outcome, a similar insights can be drawn. However, now all covariates are significant - model explaining 1.2% of the variation in family income. Again, the DiD effect alone in model (5) is significant ( $\hat{\beta} = 1911.04$ ,  $p < 0.01$ ); suggesting again that women with childrens' family income increases by \$1911.04 after the intervention (when compared to the pre-period) when compared to childless women. Additionally, women with children, again, have a lower family income of \$-8929.33 ( $p < 0.01$ ). This model explains 2.2% of the total variation in family income. The inclusion of covariates does not change this insight dramitically; womens' with children family income still improve significantly by \$2006.06 during the post period when compared to childless women and its coeificeint also increases, but the insight is the same as before ( $p < 0.01$ ). Finally, considering work participation; the covariates in model (7) are all significant at the conventional level in predicting work participation, explaining 1.9% of total variation in this outcome. As observed during task 1 through 4, work participation rises for women with children during the post intervention period (model 8) ( $\hat{\beta} = 0.031$ ,  $p < 0.1$ ); however, the effect is only marinally significant. As was to be expected, women with children show a significantly lower work participation ( $\hat{\beta} = -0.159$ ,  $p < 0.01$ ) when compared to childless women (15.9\$ less). When combining covariates and the base model in model (9), the overall

<sup>2</sup>Note: the period indicator is not really relevant in such an exercise as compares the total trend change for two converging groups  $y$

trend is the same as in the aforementioned models. The coefficient for the DiD effect rises to 0.033 ( $p < 0.1$ ), but is still only marginally significant. Thus, overall, the inclusion of the covariates does increase the coefficients of all three DiD effects, but only marginally. This means that the overall effect, or direction of the coefficients was not changed drastically. Moreover, the inclusion of the covariates shows that the standard errors are stable across all models, suggesting no problem with multicollinearity.



(a) Annual Earnings by Females with(out) Children



(b) Family Earnings Earnings by Females with(out) Children



(c) Work Participation by Females with(out) Children

Figure 1: Pre-Post Intervention of EICT Credit for Women with(out) Children

**Robustness Tests** Overall, due to the standard errors between the models for each outcome variable respectively staying stable, this suggests that there is generally less necessity to include robust standard errors. Running the Breusch-Pagan test on all models in Table 3 - only model (1) and model (4) (both covariate-only models) fail to reject the null hypothesis of heteroscedastic residuals. Finally, Table 11 in the appendix reports robust standard errors (white) and clustered by state. Overall, the standard errors are stable across models, which suggests, that there is little need to use robust standard errors.<sup>3</sup>

<sup>3</sup>Stability in coefficients also implies no multicollinearity; additionally, standardized coefficients are not of use here because we are interested in the DiD effect.

Table 3: NON-ROBUST REGRESSION RESULTS PART 3

	<i>Dependent variable:</i>								
	earn			finc			work		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Constant	7,977.343*** (1,159.632)	14,899.900*** (828.375)	12,958.640*** (1,550.012)	11,872.150*** (1,234.711)	20,099.430*** (886.522)	16,218.430*** (1,655.347)	0.414*** (0.032)	0.582*** (0.023)	0.532*** (0.043)
Age	96.038*** (15.489)		22.555 (15.922)	144.373*** (16.492)		78.717*** (17.004)	0.003*** (0.0004)		0.002*** (0.0004)
Unemp. Rate	33.433 (108.495)		133.948 (114.614)	256.102** (115.519)		372.861*** (122.403)	-0.018*** (0.003)		-0.018*** (0.003)
Education (y)	8.153 (60.120)		66.337 (59.579)	-177.633*** (64.013)		-125.305** (63.628)	0.016*** (0.002)		0.017*** (0.002)
Race	-2,037.990*** (323.611)		-1,255.622*** (326.237)	-3,109.127*** (344.563)		-2,438.387*** (348.408)	-0.058*** (0.009)		-0.043*** (0.009)
Has Child		-8,596.327*** (1,093.444)	-8,394.973*** (1,096.506)		-8,929.330*** (1,170.197)	-8,269.567*** (1,171.022)		-0.159*** (0.030)	-0.150*** (0.030)
dperiod		-695.997 (485.413)	-536.491 (500.046)		-940.239* (519.486)	-515.553 (534.028)		-0.005 (0.013)	-0.024* (0.014)
has_children1:dperiod		1,682.810*** (642.099)	1,722.360*** (641.893)		1,911.035*** (687.171)	2,006.060*** (685.515)		0.031* (0.018)	0.033* (0.018)
Observations	13,746	13,746	13,746	13,746	13,746	13,746	13,746	13,746	13,746
R <sup>2</sup>	0.006	0.026	0.027	0.012	0.022	0.028	0.019	0.012	0.027
Adjusted R <sup>2</sup>	0.006	0.026	0.027	0.012	0.022	0.027	0.018	0.012	0.026
Residual Std. Error	18,150.240	17,965.670	17,956.450	19,325.370	19,226.750	19,176.730	0.495	0.497	0.493
F Statistic	20.153***	121.691***	54.794***	43.406***	105.245***	56.166***	65.588***	54.906***	54.374***

*Note:* N = 13746. Non Robust Standard Errors applied. "White" is reference category for "non-White" categorical variable. First model of each outcome only covariates; second model only DiD effect with causal variables; third model complete. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

## 1.6 Task 6: Subset analysis

### 1.6.1 Women with Children compared based on high & low education levels

Please note: after subsetting the data per subtask, interaction effects were used in order to maintain statistical power of the model (the interpretation is the same). For brevity reasons, the covariate only models are left out (intuition does not change when compared to table 3). Considering now only women with children: the "difference in groups" as women with high vs low education, table 4 reports a baseline and a complete model per outcome. The comparison group is low education women with children. Considering earnings, (model 1 & 2), the DiD coefficient is interpreted as follows: during the post intervention period (compared to the pre-intervention period), high educated women gain 951.50\$ on average ( $p = 0.216$ ) compared to low educated women; which however is not significant. This is broadly maintained when only considering the base model ( $\hat{\beta} = 871.46$ ,  $p < 0.26$ ). A similar observation can be made for work participation, which is in both cases, the base model (5) and with covariates (6), displaying an insignificant decrease in work participation for highly educated women (compared to low educated women) in the post intervention period when compared to the pre-period ( $\hat{\beta} = -0.019$ ,  $p = 0.462$ ). Finally, family income observed a marginally significant increase of \$1345.12 at the 9.4% percent level for highly educated women during the post period compared to low educated women, while the base model is insignificant at the 15.2% level for the DiD effect of education and intervention period.

Table 4: SUBSECTION ANALYSIS SINGLE WOMEN WITH CHILDREN FOR ALTERNATING LOW/ HIGH EDUCATION LEVELS

	<i>Dependent variable:</i>					
	earn		finc		work	
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	8,743.660*** (1,108.213)	4,625.713*** (1,691.462)	13,755.130*** (1,166.890)	4,175.858** (1,768.019)	0.422*** (0.037)	0.527*** (0.057)
edu_lvl	-3,427.529*** (1,311.705)	-3,177.416** (1,306.105)	-3,629.042*** (1,381.156)	-3,079.124** (1,365.220)	0.002 (0.044)	0.006 (0.044)
dperiod	370.227 (653.474)	416.590 (658.753)	142.756 (688.073)	453.880 (688.568)	0.011 (0.022)	-0.009 (0.022)
age		174.206*** (20.082)		272.964*** (20.991)		0.004*** (0.001)
urate		-34.414 (127.034)		261.583** (132.784)		-0.025*** (0.004)
nonwhite1		-2,547.489*** (368.752)		-3,380.644*** (385.442)		-0.061*** (0.012)
edu_lvl:dperiod	871.461 (773.065)	951.502 (767.593)	1,166.212 (813.996)	1,345.136* (802.335)	0.021 (0.026)	0.019 (0.026)
Observations	7,819	7,819	7,819	7,819	7,819	7,819
R <sup>2</sup>	0.005	0.020	0.004	0.033	0.002	0.016
Adjusted R <sup>2</sup>	0.004	0.020	0.003	0.033	0.001	0.016
Residual Std. Error	14,923.420	14,810.030	15,713.570	15,480.340	0.499	0.495
F Statistic	12.717***	26.977***	9.460***	44.915***	4.884***	21.557***

*Note:* N = 7819 Single Women have Children. N = 5593 high education (years of education  $\geq 9$  years); N = 2226 low education (years of education  $< 9$  years); Non Robust Standard Errors applied. "White" is reference category for "non-White" categorical variable. \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .

### 1.6.2 Women with and without Children compared keeping education level (low) constant

Considering now only low educated women; alternating again on whether a women has a child or not, the insights are to be found in Table 5. The subsection analysis shows that the post intervention period earnings for low educated women with children decrease by \$413.45 when compared to women without children in the base model; however all insignificant ( $p = 0.730$ ). This does not change for the complete model ( $\hat{\beta} = -475.403$ ,  $p < 0.691$ ). Same can be observed for the outcome variables Family Income (main model:  $\hat{\beta} = -179.637$ ,  $p = 0.887$ ) and work participation (main model:  $\hat{\beta} = 0.012$ ,  $p = 0.702$ ). Thus, there is no statistical support that work participation rises by 1.2% for women with children compared

to childless women during the post intervention period and likewise for family income decreases.<sup>4</sup>

### 1.6.3 Conclusion subsections

Overall, the introduction of the tax credit does not seem to be significant across the different outcomes and subgroups.

Table 5: SUBSECTION ANALYSIS SINGLE WOMEN WITH/ WITHOUT CHILDREN FOR CONSTANT (LOW) EDUCATION LEVELS

	<i>Dependent variable:</i>					
	earn		finc		work	
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	11,066.700*** (1,457.343)	4,281.758 (2,602.848)	17,494.530*** (1,534.185)	8,507.214*** (2,735.084)	0.501*** (0.038)	0.441*** (0.068)
has_children1	-2,323.038 (2,026.514)	-2,402.722 (2,030.169)	-3,739.399* (2,133.367)	-3,267.950 (2,133.311)	-0.080 (0.053)	-0.087 (0.053)
dperiod	783.677 (858.662)	1,378.102 (882.127)	322.393 (903.937)	1,127.629 (926.943)	-0.004 (0.023)	-0.007 (0.023)
age		29.868 (29.856)		83.479*** (31.373)		0.001 (0.001)
urate		651.163*** (216.731)		845.832*** (227.742)		-0.002 (0.006)
nonwhite1		332.812 (664.392)		-2,403.220*** (698.146)		0.081*** (0.017)
has_children1:dperiod	-413.449 (1,194.473)	-475.403 (1,193.612)	-179.637 (1,257.455)	-152.761 (1,254.253)	0.015 (0.031)	0.012 (0.031)
Observations	4,311	4,311	4,311	4,311	4,311	4,311
R <sup>2</sup>	0.006	0.009	0.010	0.016	0.003	0.008
Adjusted R <sup>2</sup>	0.006	0.008	0.009	0.015	0.002	0.007
Residual Std. Error	18,962.540	18,944.100	19,962.390	19,906.540	0.498	0.497
F Statistic	9.304***	6.559***	14.690***	11.920***	4.494***	6.121***

*Note:* N = 4311 Single Women have Children (years of education < 9 years). N = 2085 has no children; N = 2226 has children; Non Robust Standard Errors applied. "White" is reference category for "non-White" categorical variable. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

## 2 Instrumental Variable approach

### 2.1 Task 1: Endogeneity problems and correlation with u - A3

When considering the problem of mean independence (A3), there can be multiple root causes to this issue; omitted variable bias, sampling bias, attrition bias, simultaneous causality bias. All of these issues introduce backdoor pathways **CITE ECI HERE**, which bias the (eg. OLS) estimator severely. A prominent example in what way education effect on earnings could be biased is through an omitted variable. A classical example for this is general ability, which commonly determines not only the degree of success in education, but also future earnings. Through this, education would correlate with the disturbance term, violating A3. Thus, neglecting to include this confounder, will lead to a biased estimate of the effect of years of education on earnings. A second example could be socioeconomic factors. Not only tend wealthier families to send their children to better schools, but also do such children participate more often in tertiary education; thereby impacting years of education. Contrarily, worse off families may require their children to drop out of school to contribute to the family income (considering that it is the 1950s this is not unlikely). Moreover, wealthy family tend to create more job opportunities for their children through connections and inheritance etc.. Thus, socioeconomic status of the family not only impacts the duration (years) of education of their child, but also their later income. This is not only describing omitted variable bias, but also (self) selection bias into another (higher ed) treatment group; another detrimental bias.

<sup>4</sup>this would have been very interesting; because if work would go up significantly while income drops, this would be crazy.



## 2.2 Task 2: Summary statistics for this task

Table 6 contains the descriptive statistics of the numeric variables. Overall 329,509 observations are in the data on people born between 1930 and 1939. The age range of study subectes ranges from 40 to 50, with a median age at 45 ( $mean = 44.645$ ,  $std = 2.940$ ). The average years of education was 12.77 ( $std = 3.281$ ,  $median = 12$ ). Finally, for interpretation purposes here,  $lnwage$  ( $mean = 5.90$ ,  $std = 0.679$ ,  $median = 6.257$ ) was rescaled to wage, which displayed a strong positive skew of 26.39, mean of \$439.47 ( $std = 364.941$ ) and a median of \$521.85. The log scaled wage will be used during the analysis. Moreover, 86.3% of respondents were married. SMSA indicates whether a study subject lives in rural or urban areas. Quarter of birth, the IV, distributes reasonably equally, with Q1 N = 81671, Q2 N = 80138, Q3 N = 86856, and Q4 N = 80844 - it is influencial to when a study subject will start schooling; eg. suggesting that subjects from the fourth quarter generally obtain more years of study.

Table 6: Instrumental Vairable Approach Descriptive Statistics of Numeric Indeependent and Dependent Variable

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Age	44.645	2.940	40	42	45	47	50
Education in Years	12.770	3.281	0	12	12	15	20
lnwage	5.900	0.679	-2.342	5.637	5.952	6.257	10.532
Wage	439.471	364.941	0.096	280.481	384.711	521.848	37,499.990
Quarter of Birth	2.506	1.112	1	2	3	3	4

Notes: N = 329,509; Wage is backwards transformed from  $lnWage$ ; Categorical variables described in text.

## 2.3 Task 3: Diagnositcs of Year Of Birth as instrument: Is it relevant?

A suitable instrument fulfills two conditions: 1) Cleanliness; meaning that the instrument only has an impact on the outcome thorough the causal variable (here years of education). This assumption cannot be tested, but is based in theory (and logic). 2) The instrument is relevant, meaning that the instrument significantly predicts (and preferably strongly correlates) with the causal ("to be instrumentalized") variable. This assumption can be tested i.a. thorough conducting an F test on the first stage of 2SLS. In order to assess the relevance of the instrument statisically, one regresses the casual variable on the instrument and exogenous covariates - i.e. the first stage of the 2SLS (this can also be retrieved from the 2SLS function in R). Table 7 contains both quarter of birth as a factor and quantitaitive variable.

Firstly, all models reported show a significant F statistic ( $F = 100.653$ ,  $p < 0.01$ ;  $F = 414.645$ ,  $p < 0.01$ ;  $F = 30.009$ ,  $p < 0.01$ ;  $F = 249.859$ ,  $p < 0.01$ ); which promises that qob might fulfill the relevane criterion.

Considering first the basis models (1 & 3) without the inclusion of the covariates, one can observe for quarter of birth (qob) as a numeric (model 1), that it significantly predicts years of education ( $\hat{\beta} = 0.052$ ,  $p < 0.01$ ) at conventional levels. Similarly, quarter two to four are all significantly different from quarter one (Q2:  $\hat{\beta} = 0.057$ ,  $p < 0.01$ ; Q3:  $\hat{\beta} = 0.117$ ,  $p < 0.01$ ; Q4:  $\hat{\beta} = 0.151$ ,  $p < 0.01$ ); Similarly, this can be also observed when alteranting the catgory of comparison for quarter. Considering the relevance of qob when combined with covariates (model 2 & 3), a similar picutre can be observed. It is assumed that the covaraites are exogenoeus. Interpreted as a numeric variable, qob is significant at conventional levels in predicting years of education ( $\hat{\beta} = 0.032$ ,  $p < 0.01$ ). In model (4), most categories are also significantly different from the first quarter in years of education (Q2:  $\hat{\beta} = -0.004$ ,  $p = 0.812$ ; Q3:  $\hat{\beta} = 0.052$ ,  $p < 0.01$ ; Q4:  $\hat{\beta} = 0.088$ ,  $p < 0.01$ ). Statistically, it can be asumed qob is a relevant instrument.<sup>5</sup>

Finally, when considering Figure 2, one can seethat on average, Q4 and Q3 tend to have significantly more years of education than Q1 and Q2 - this might also explain the aforementioend results regarding Q2 difference to Q1 in model (4) in table 7. Overall, the instrument appears to meet the relevance criterion.

## 2.4 Task 4: Conduct IVreg of the effect of effect of education on log wages, using quarter of birth as the instrument; are robust SE needed?

For the purpose of this analysis, it is assumed that the control variables are all exogenoeus. Table 8 reports the corresponding outputs; model (1) and (3) consider qob as a numeric variable and model (2) and (4)

<sup>5</sup>Note: the correlation is weak at 0.0175,  $p \geq 0.01$

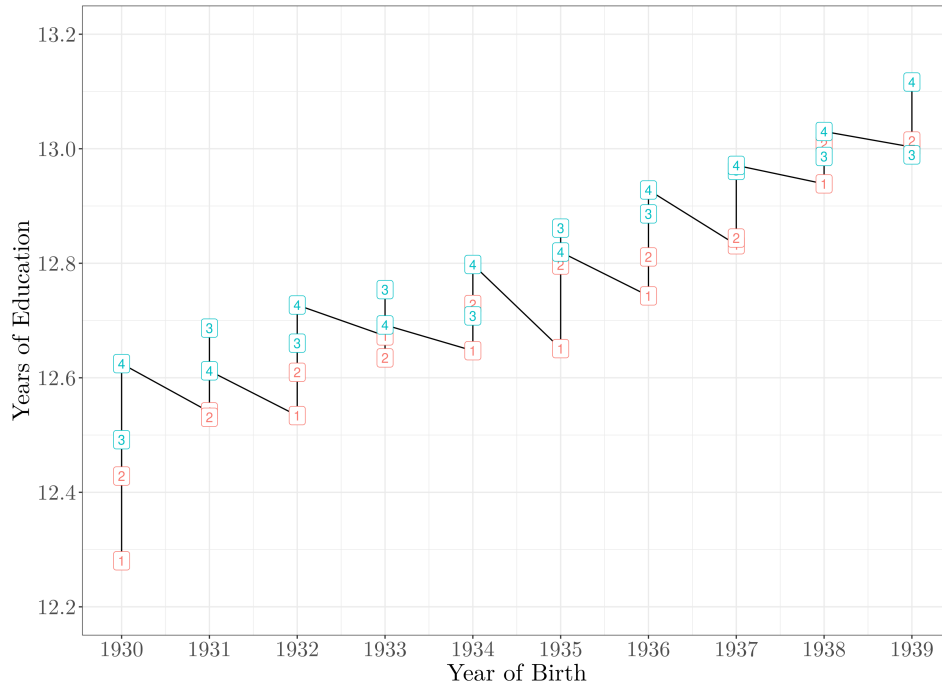


Figure 2: Displaying association between Quarter of Birth and Years of Education

Table 7: FIRST STAGE REGRESSION OUTPUT - RELEVANCY

	Dependent variable:			
	(1 - numeric)	(2 - numeric)	(3 - categorical)	(4 - categorical)
Constant	12.641*** (0.014)	15.150*** (0.091)	12.688*** (0.011)	15.213*** (0.091)
qob	0.052*** (0.005)	0.032*** (0.005)		
age		-0.060*** (0.002)		-0.060*** (0.002)
married		0.248*** (0.017)		0.248*** (0.017)
qob_fac2			0.057*** (0.016)	-0.004 (0.016)
qob_fac3			0.117*** (0.016)	0.052*** (0.016)
qob_fac4			0.151*** (0.016)	0.088*** (0.016)
Observations	329,509	329,509	329,509	329,509
R <sup>2</sup>	0.0003	0.004	0.0003	0.004
Adjusted R <sup>2</sup>	0.0003	0.004	0.0003	0.004
Residual Std. Error	3.281	3.275	3.281	3.275
F Statistic	100.653***	414.645***	34.009***	249.859***

Note: N = 329,509 - degrees of freedom are not reported for F tests; \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

consider it as a categorical variable.<sup>6</sup> Models (5) and (6) consider an additional instrument (discussed in part 5). The control variables chosen are SMSA and whether married or not. Married was included because the proportion of married men tends to be higher for higher educated men, suggesting more education but also more income. Moreover, SMSA indicates whether a person lives in a rural or urban surrounding; people in urban areas may have better access to education and better work opportunities. Considering the base models (1 - numeric) and (2 - categorical), we can observe that the choice between qob as a numeric or categorical variable instrument barely changes the coefficients and standard errors; reporting a significant positive coefficient for the years of education coefficient (model 1:  $\hat{\beta} = 0.099$ ,  $p < 0.01$ ; model 2:  $\hat{\beta} = 0.103$ ,  $p < 0.01$ ). For example, for model (2), this implies that one year more we can see that if education increases by one year wage increases by 10.3%.

Moving to model (4 - categorical), which contains SMSA and married as exogenous controls, the direction and magnitude, in addition to reported standard errors, for years of education do not change considerably when compared to model (2) (same for model 3 and 1), reporting a significant positive coefficient for education ( $\hat{\beta} = 0.100$ ,  $p < 0.01$ ). Meaning, that an increase by one year in years of education is associated with a 10% higher wage. Also the control variables are significant which are both categoricals (SMSA:  $\hat{\beta} = -0.148$ ,  $p < 0.01$ ; married:  $\hat{\beta} = -0.255$ ,  $p < 0.01$ ); However, due to them being controls, we refrain from interpreting them here as we did with education. Thus, the inclusion of (assumed) exogeneous controls does not impact change insights derived from the model.

Regarding robustness, the breusch pagan test was significant at conventional levels for all six models (not surprising at this sample size), suggesting no issues regarding heteroscedasticity. Moreover, when applying (White) standard errors, the non standardized standard errors maintain stability, suggesting that this research is not suffering from heteroscedasticity (see Appendix Table 14). Finally, an analysis of the (Residual Distribution) QQ Plot of model (4) (see Appendix Figure 3) also do not suggest extreme deviations in the residuals from a constant. Thus, statistical inference is not impacted by the inclusion of robust standard errors.

Table 8: IV REGRESSION OUTPUT

	<i>Dependent variable:</i>					
	lnwage					
	(1 numeric)	(2 categorical)	(3 numeric)	(4 categorical)	(5 numeric)	(6 categorical)
Constant	4.633*** (0.250)	4.590*** (0.249)	4.461*** (0.249)	4.425*** (0.248)	5.514*** (0.076)	5.511*** (0.076)
educ	0.099*** (0.020)	0.103*** (0.020)	0.098*** (0.020)	0.100*** (0.019)	0.015** (0.006)	0.015** (0.006)
SMSA			-0.151*** (0.022)	-0.148*** (0.022)	-0.243*** (0.007)	-0.243*** (0.007)
married			0.256*** (0.006)	0.255*** (0.006)	0.280*** (0.004)	0.280*** (0.004)
Observations	329,509	329,509	329,509	329,509	329,509	329,509
R <sup>2</sup>	0.098	0.094	0.124	0.120	0.082	0.083
Adjusted R <sup>2</sup>	0.098	0.094	0.124	0.120	0.082	0.083
Residual Std. Error	0.645	0.646	0.635	0.637	0.650	0.650

Note: N = 329,509 - degrees of freedom are not reported for F tests. SMSA comparison category is rural; married comparison category is not married. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

## 2.5 5 – Manual OLS

Note: this question was not particularly precise wrt. whether to run a manual 2SLS or just normal OLS.

Model (1) and (2) in Table 9 report the OLS regressions of lnwage on the endogeneous education with and without covariates. Models (3) and (4) report the manual 2SLS for quarter of birth as a numeric. Models (5) and (6) report quarter of birth as a factor.

Considering model (2), standard OLS with covariates, the overall effect of education on is significantly positive ( $\hat{\beta} = 0.067$ ,  $p < 0.01$ ), which does not deviate drastically from its (assumed) exogeneous counterpart in the other models in Table 8 (and 9). However, we can observe a considerably lower standard error for the years of education coefficient in model (2) when compared to Table 8 models 1 through 6.

<sup>6</sup>It may be noted that qob should be interpreted as a categorical variable in actuality.

Table 9: IV REGRESSION OUTPUT

	Dependent variable:					
	lnwage					
	OLS (1)	OLS (2)	HandIVREG(3)	HandIVREG(4)	HandIVREG(5)	HandIVREG(6)
Constant	4.995*** (0.004)	4.847*** (0.005)	4.633*** (0.263)	4.461*** (0.260)	4.590*** (0.262)	4.425*** (0.259)
educ	0.071*** (0.0003)	0.067*** (0.0003)				
SMSA		-0.185*** (0.003)		-0.151*** (0.023)		-0.148*** (0.023)
married		0.265*** (0.003)		0.256*** (0.007)		0.255*** (0.007)
educ.hat			0.099*** (0.021)	0.098*** (0.020)	0.103*** (0.020)	0.100*** (0.020)
Observations	329,509	329,509	329,509	329,509	329,509	329,509
R <sup>2</sup>	0.117	0.145	0.0001	0.041	0.0001	0.041
Adjusted R <sup>2</sup>	0.117	0.145	0.0001	0.041	0.0001	0.041
Residual Std. Error	0.638	0.628	0.679	0.665	0.679	0.665
F Statistic	43,782.560***	18,646.640***	23.145***	4,738.984***	25.088***	4,739.535***

Note: ALL ARE IV REGRESSION OUTPUTS!!! tests. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

Moving on now to the manual 2SLS, as expected, the reported coefficients in models 3 through 6 in Table 9 correspond exactly with those reported in Table 8 (models 1 through 4). However, the manual 2SLS does not correct for the fact that years of education (hat) originates as a predicted outcome of the first stage. Thus, the standard errors differ slightly. Considering how which model to prefer, OLS vs model of methods or IVREG, the Hausman's  $\chi^2$  can be used to assess which estimator is to be preferred. These are output from the `summary(IVREG_model, diagnostics = TRUE)` in R, together with a test for weak instruments and, in case of overidentification, the Sargan-Hansen  $\chi^2$  over-identification test. To keep it short, two models are considered for demonstrative purposes: 1) Table 8 model (3) where `qob` is a numeric variable (so no overidentification), and 2) Table 8 model (5) is considered where `yob` (year of birth) was added as an instrument to demonstrate over identification. Please note that technically, the inclusion of `qob` as a factor implies overidentification of the first stage; this will be ignored in this question.

Considering Table 10, the weak instruments test, which is just a partial F-test on the first stage of 2SLS, reports significant F statistics for both models (model 3:  $FStatistic = 100.159$ ,  $p < 0.01$ ; model 5:  $FStatistic = 557.21$ ,  $p < 0.01$ ), suggesting that the relevance criterion is fulfilled. Moving to the Wu-Hausman test however, the models differ. Model (3) (`qob` numeric model) reports an insignificant test statistic ( $\chi^2 = 2.477$ ,  $p = 0.116$ ), suggesting a preference for the OLS estimator. Contrarily, Model (5) reports a significant test statistic ( $\chi^2 = 82.16$ ), suggesting that the usage of model of methods (2SLS) is preferred over OLS in this case.<sup>7</sup> Finally, considering the question of overidentification, this is not applicable to model (3), when considering `qob` as a numeric; it would, however, be relevant if we considered `qob` as a factor. Model (5), on the other hand, contains a second instrument, year of birth, as a demonstrative example. As can be seen in Table 10, the Sargan-Hansen  $\chi^2$  test of overidentification reports a significant test statistic ( $\chi^2 = 18.84$ ,  $p < 0.01$ ). Rejecting the null hypothesis implies that independence of the instruments of the residuals cannot be assumed. Thus, the instruments themselves are assumed to be endogenous themselves and overidentification poses a problem for model (5). It may be interesting to note that if we considered `qob` as a factor, this test statistic would be insignificant, suggesting that the `qob` as a factor is exogenous and, thus, overidentification is no problem.

df1 df2 statistic p-value Weak instruments 1 329505 100.159 2e-16 \*\*\* Wu-Hausman 1 329504 2.477 0.116 Sargan 0 NA NA NA

Diagnostic tests: df1 df2 statistic p-value Weak instruments 2 329504 557.21 2e-16 \*\*\* Wu-Hausman 1 329504 82.16 2e-16 \*\*\* Sargan 1 NA 18.84 1.42e-05 \*\*\*

<sup>7</sup>Please note that I included this example on purpose

## 2.6 6 –

**Relevance:** Contrarily, this assumption can be tested: it states that the instrument used for the "to be instrumentalized" variable is strong, meaning that there is a relevant correlation between the instrument and the independent variables. Note: a correlation between the instrument and the independent variables beyond the biased variable is welcome. It only becomes a problem when the instrument and the dependent variable are related; but there will always be some correlation in that regard. To this end, an ANOVA is run on the two-stage model in order to conduct the F test, which helps with multiple outputs: Wu-Hausman, Sargan, and F test (the former two are only relevant if the model is overidentified by the instrument)

Elaborate on possible causes of concern that could render your instrumental variable identification strategy invalid. Base your argumentation on the required assumptions of instrumental variables as taught in the lecture. An arbitrarily made-up example would be: "Birthdays in the US have been recorded with a varying delay between 1 and 6 months. Hence, the measurement of school years for dropout students could be inaccurate."

There are multiple ways which could invalidate the instrumental variable identification strategy. 1) the instrument may not be relevant for the endogenous variable; it is notable that correlation does not necessarily guarantee relevance. Thus, it might actually be the case that qob is just not a relevant instrument and that the correlation is not strong enough. 2) The (supposed) variable years of education is actually not endogenous; remember that this is mostly based on theory and not on tests. However, the most interesting point to discuss is considering 3) the cleanliness of the instrument. This would mean, that qob does not only have an effect on lnwage through the endogenous variable alone, but rather is a confounder itself, thus, having an impact on lnwage. 4) the instrument is endogenous itself; such as measurement errors or it may be related to further backdoor paths such as socioeconomic status.

Cleanliness; Endogeneity of independent; Relevance – Cleanliness

1. not clean instruments! this would lead to violation of mean independence
2. weak instruments leads to mean independence violation as well because the instruments simply do not work as they should
3. google some stuff and see internet

reasons: not clear instrument (or not clean) (make a plot; or if the instrument is not relevant or weak)

## 3 Bibliography

## 4 Appendix

Table 10: Descriptive Statistics of ECIC; With Children

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Family Income	12,750.390	15,739.050	0.000	4,652.465	8,425.197	15,218.720	410,507.600
Earnings	7,909.934	14,956.930	0.000	0.000	1,110.727	11,107.270	366,095.500
Age	32.717	8.630	20	25	32	39	54
Education	9.001	2.408	0	7	10	11	11
Education Years	4.840	5.872	0.000	0.071	3.761	7.070	102.958
Unearned Income	2.097	1.209	1	1	2	3	9
Count Children	0.466	0.499	0	0	0	1	1

Notes: N = 7819

Table 11: Descriptive Statistics of ECIC; Without Children

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Family Income	18,559.860	23,041.780	0.000	5,793.092	11,912.950	24,391.010	575,616.800
Earnings	13,760.260	21,301.400	0.000	0.000	7,664.014	19,447.610	537,880.600
Age	38.498	11.046	20	28	40	49	54
Education	8.549	2.889	0	7	10	11	11
Education Years	4.800	8.496	0.000	0.000	1.248	6.528	134.058
Unearned Income	0.000	0.000	0	0	0	0	0
Count Children	0.574	0.494	0	0	1	1	1

Notes: N = 5927

Table 12: NON-ROBUST REGRESSION RESULTS PART 3

	<i>Dependent variable:</i>					
	earn		finc		work	
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	14,899.900*** (828.375)	12,958.640*** (1,550.012)	20,099.430*** (886.522)	16,218.430*** (1,655.347)	0.582*** (0.023)	0.532*** (0.043)
has_children1	-8,596.327*** (1,093.444)	-8,394.973*** (1,096.506)	-8,929.330*** (1,170.197)	-8,269.567*** (1,171.022)	-0.159*** (0.030)	-0.150*** (0.030)
dperiod	-695.997 (485.413)	-536.491 (500.046)	-940.239* (519.486)	-515.553 (534.028)	-0.005 (0.013)	-0.024* (0.014)
age		22.555 (15.922)		78.717*** (17.004)		0.002*** (0.0004)
urate		133.948 (114.614)		372.861*** (122.403)		-0.018*** (0.003)
ed		66.337 (59.579)		-125.305** (63.628)		0.017*** (0.002)
nonwhite1		-1,255.622*** (326.237)		-2,438.387*** (348.408)		-0.043*** (0.009)
has_children1:dperiod	1,682.810*** (642.099)	1,722.360*** (641.893)	1,911.035*** (687.171)	2,006.060*** (685.515)	0.031* (0.018)	0.033* (0.018)
R <sup>2</sup>	0.026	0.027	0.022	0.028	0.012	0.027
Adjusted R <sup>2</sup>	0.026	0.027	0.022	0.027	0.012	0.026
Residual Std. Error	17,965.670	17,956.450	19,226.750	19,176.730	0.497	0.493
F Statistic	121.691***	54.794***	105.245***	56.166***	54.906***	54.374***

Note: N = 13746. Non Robust Standard Errors applied. "White" is reference category for "non-White" categorical variable.

Table 13: NON-ROBUST REGRESSION RESULTS PART 3

	<i>Dependent variable:</i>								
	earn			finc			work		
	Basic(1)	White(2)	Cluster(3)	Basic(4)	White(5)	Cluster(6)	Basic(7)	White(8)	Cluster(9)
has_children1	−8,394.973*** (1,096.506)	−8,394.973*** (1,056.274)	−8,394.973*** (1,057.069)	−8,269.567*** (1,171.022)	−8,269.567*** (1,146.704)	−8,269.567*** (1,147.577)	−0.150*** (0.030)	−0.150*** (0.030)	−0.150*** (0.030)
dperiod	−536.491 (500.046)	−536.491 (564.674)	−536.491 (565.075)	−515.553 (534.028)	−515.553 (618.997)	−515.553 (619.444)	−0.024* (0.014)	−0.024* (0.014)	−0.024* (0.014)
age	22.555 (15.922)	22.555 (19.063)	22.555 (19.077)	78.717*** (17.004)	78.717*** (20.183)	78.717*** (20.198)	0.002*** (0.0004)	0.002*** (0.0004)	0.002*** (0.0004)
urate	133.948 (114.614)	133.948 (114.009)	133.948 (114.096)	372.861*** (122.403)	372.861*** (123.441)	372.861*** (123.537)	−0.018*** (0.003)	−0.018*** (0.003)	−0.018*** (0.003)
ed	66.337 (59.579)	66.337 (59.876)	66.337 (59.937)	−125.305** (63.628)	−125.305* (67.467)	−125.305* (67.538)	0.017*** (0.002)	0.017*** (0.002)	0.017*** (0.002)
nonwhite1	−1,255.622*** (326.237)	−1,255.622*** (341.701)	−1,255.622*** (341.929)	−2,438.387*** (348.408)	−2,438.387*** (368.776)	−2,438.387*** (369.027)	−0.043*** (0.009)	−0.043*** (0.009)	−0.043*** (0.009)
has_children1:dperiod	1,722.360*** (641.893)	1,722.360*** (647.317)	1,722.360*** (647.764)	2,006.060*** (685.515)	2,006.060*** (698.628)	2,006.060*** (699.118)	0.033* (0.018)	0.033* (0.018)	0.033* (0.018)
Constant	12,958.640*** (1,550.012)	12,958.640*** (1,622.488)	12,958.640*** (1,623.778)	16,218.430*** (1,655.347)	16,218.430*** (1,771.017)	16,218.430*** (1,772.450)	0.532*** (0.043)	0.532*** (0.042)	0.532*** (0.042)
Observations	13,746	13,746	13,746	13,746	13,746	13,746	13,746	13,746	13,746
R <sup>2</sup>	0.027	0.027	0.027	0.028	0.028	0.028	0.027	0.027	0.027
Adjusted R <sup>2</sup>	0.027	0.027	0.027	0.027	0.027	0.027	0.026	0.026	0.026
Residual Std. Error	17,956.450	17,956.450	17,956.450	19,176.730	19,176.730	19,176.730	0.493	0.493	0.493
F Statistic	54.794***	54.794***	54.794***	56.166***	56.166***	56.166***	54.374***	54.374***	54.374***

Note: N = 13746. Robust Standard Errors applied. "White" is reference category for "non-White" categorical variable. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

Table 14: WHITE-ROBUST REGRESSION RESULTS QOB NUMERIC ONLY

	<i>Dependent variable:</i>					
	lnwage					
	Basic(1)	White(2)	Basic(3)	White(4)	Basic(5)	White(6)
educ	0.099*** (0.020)	0.099*** (0.020)	0.098*** (0.020)	0.098*** (0.020)	0.015** (0.006)	0.015** (0.006)
SMSA			-0.151*** (0.022)	-0.151*** (0.022)	-0.243*** (0.007)	-0.243*** (0.007)
married			0.256*** (0.006)	0.256*** (0.007)	0.280*** (0.004)	0.280*** (0.004)
Constant	4.633*** (0.250)	4.633*** (0.249)	4.461*** (0.249)	4.461*** (0.248)	5.514*** (0.076)	5.514*** (0.076)
Observations	329,509	329,509	329,509	329,509	329,509	329,509
R <sup>2</sup>	0.098	0.098	0.124	0.124	0.082	0.082
Adjusted R <sup>2</sup>	0.098	0.098	0.124	0.124	0.082	0.082
Residual Std. Error	0.645	0.645	0.635	0.635	0.650	0.650

Note: \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

## 5 Code



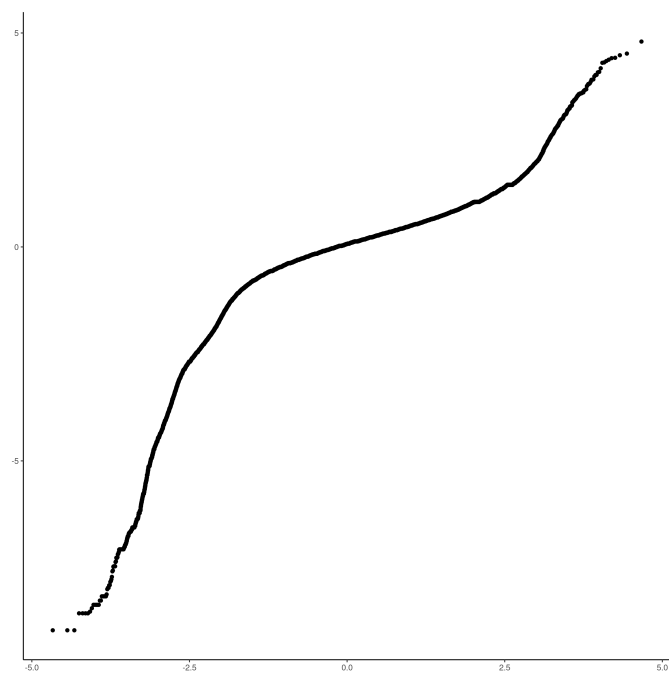


Figure 3: Residual Distribution Plot