The Effect of Compulsory Education Requirement on Future Earnings in the U.S.

Arman Jahangiri[*], Yeting Wu[†], Frances Yuan[†]
[*]Department of Mathematics
[†]Department of Statistics
University of Washington

June 11, 2025

Contents

Contents

1 Introduction and Explaratory Analysis

This project investigates whether additional education causally affects labor market earnings, building on Angrist and Krueger (1991). U.S. compulsory schooling laws require children to remain in school until age 16-17, while school districts typically require children to turn six by January 1 to enter first grade. Consequently, children born earlier in the calendar year start school older, reach the legal dropout age sooner, and accumulate fewer years of education than those born later.

Assuming birth timing is random, quarter of birth serves as an **Instrumental Variable (IV)** for years of education. This strategy addresses endogeneity from unobserved ability, motivation, or family background. Our causal question is: *"Does an exogenous increase in education, induced by compulsory school attendance laws and entry cutoffs, lead to higher earnings?"*

We use the **1980 U.S. Census 5% Public Use Microdata Sample (PUMS)** from Angrist [1991]. The original dataset contains 1,063,634 observations on 27 variables for individuals born 1930-1949, with 167,139 having missing features. We focus on the cleaned subset (asciiQOB.zip) used by Angrist and Krueger [1991], containing 5 variables for 329,509 U.S.-born men, listed in Table 1. Variables were selected based on their IV framework roles and causal relevance; detailed justification appears in Appendix Table 3.

Table 1: Dataset Description with Summary Statistics

Variable	**Acronym**	**Range**	**Mean**	**Median**
Log Weekly Wage	`LWW`	$[-2.342,\ 10.532]$	5.900	5.952
Years of Education	`YOE`	$[0,\ 20]$	12.77	12.00
Quarter of Birth	`QOB`	$\{1,\ 2,\ 3,\ 4\}$	2.506	3.000
Year of Birth	`YOB`	$[1930,\ 1939]$	1934.6	1935
State of Birth	`SOB`	$\{1,\ldots,56\} \setminus \{3,7,14,43,52\}$	30.69	34.00

The correlation heatmap of the variables in Figure 1 indicates a highly positive linear association between YOE and LWW, but a very weak direct (linear) association between QOB and LWW.

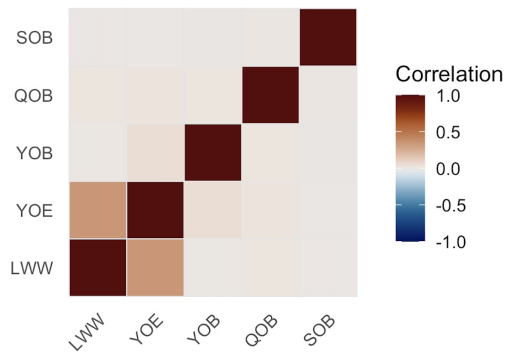


Figure 1: Correlation heatmap of the variables.

Some summary distributions are provided in the plot of Appendix Figure 7, which shows LWW by YOE, grouped by QOB. Each boxplot represents the distribution of LWW for individuals with the same YOE and QOB.

Table 2 in the Appendix illustrates the 51 geographic units in the dataset, which include the 50 U.S. states and the District of Columbia.

2 OLS and TSLS

Ordinary Least Squares (OLS): As our baseline approach, we estimate the relationship between education and wages using Ordinary Least Squares (OLS) with $Y = \texttt{LWW}$ and $X = \texttt{YOE}$: $\texttt{LWW}_i = \alpha_0 + \alpha_1 \texttt{YOE}_i + \alpha_2' \mathbf{X_i'} + \epsilon_{0i}$, where $\mathbf{X_i'}$ includes optional controls (race, age, place of birth). Regressing LWW on YOE yields $\widehat{\texttt{LWW}} = 4.995 + 0.071 \times \texttt{YOE}$ with $p < 2 \times 10^{-16}$, suggesting 7% wage increases per additional education year. This aligns with the statement *"The estimated monetary return to an additional year of schooling for those who are compelled to attend school by compulsory schooling laws is about 7.5 percent, which is hardly different from the ordinary-least-squares (OLS) estimate of the return to education for all male workers."*

- Angrist and Krueger [1991].

However, the coefficient α_1 is likely biased due to omitted variables that simultaneously influence both YOE and LWW, measurement error, and endogeneity of YOE. This endogeneity problem motivates our instrumental variable approach, which exploits exogenous variation in educational attainment to identify causal effects.

Two-Stage Least Squares (TSLS): We use TSLS with $Z = \texttt{QOB}$ as an instrument for $X = \texttt{YOE}$. The first stage regresses YOE on QOB: $\widehat{\texttt{YOE}_i} = \alpha_0 + \alpha_{\texttt{YOE,QOB}} \cdot \texttt{QOB}_i + u_i$, isolating exogenous variation in YOE to yield fitted values $\widehat{\texttt{YOE}}_i$. The second stage regresses $Y = \texttt{LWW}$ on fitted values: $\texttt{LWW}_i = \beta_0 + \beta^{\text{TSLS}}_{\texttt{LWW,YOE}} \cdot \widehat{\texttt{YOE}}_i + \varepsilon_i$.

The coefficient $\beta^{\text{TSLS}}_{\texttt{LWW,YOE}}$ identifies the Local Average Treatment Effect (LATE) for individuals whose education was influenced by QOB. Our results (Table 4) yield $\widehat{\texttt{LWW}} = 4.590 + 0.103 \cdot \widehat{\texttt{YOE}}$, with $\hat{\beta}^{\text{TSLS}} = 0.103$ ($p < 0.001$), indicating an additional year of schooling increases LWW by 10.3%. Figure 3 shows OLS and TSLS estimates are remarkably similar, confirming Angrist and Krueger [1991]'s finding that IV estimates closely match OLS results.

When YOE takes discrete values $\{0, 1, 2, \ldots, k\}$, each unit has corresponding potential outcomes $\texttt{LWW}(0), \texttt{LWW}(1), \ldots, \texttt{LWW}(k)$. The **Average Causal Effect (ACE)** comparing treatment levels x to x' is defined as:

$$\text{ACE}(x, x') = \mathbb{E}[\texttt{LWW}(x) - \texttt{LWW}(x')]$$

Figure 2 shows the estimated ACE of changing YOE from level x' to x, assuming random assignment (naive but illustrative). Brighter colors indicate larger earnings gains.

To verify QOB is related to LWW, we regress $\widehat{\texttt{LWW}} = \alpha + \gamma \cdot \texttt{factor(QOB)} + \varepsilon$. Table 5 (Appendix) shows small but significant coefficients (around 0.014).

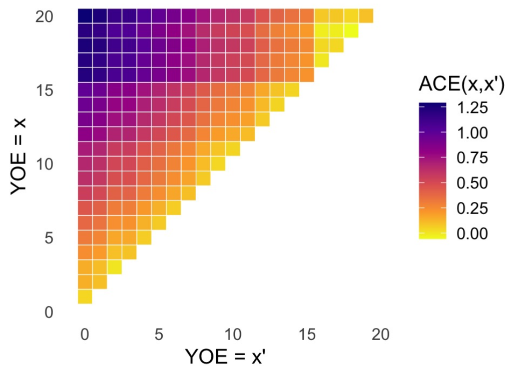


Figure 2: $\text{ACE}(x, x') = \mathbb{E}[\texttt{LWW} \mid \texttt{YOE} = x] - \mathbb{E}[\texttt{LWW} \mid \texttt{YOE} = x']$ for $0 \leq x' \leq x \leq 20$.

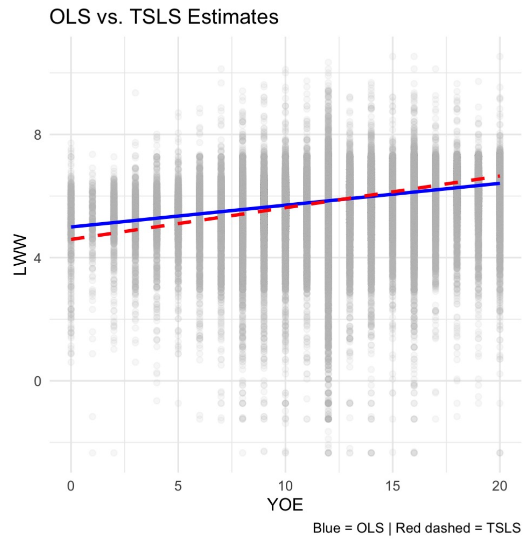


Figure 3: Comparison of OLS estimates (blue line) and TSLS estimates (red dashed line).

3 PC and FCI Algorithms

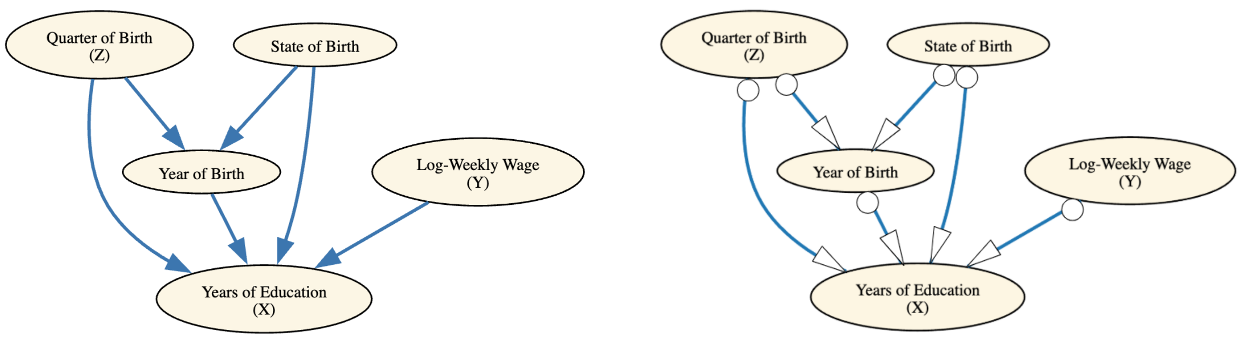
PC Algorithm: To provide a general overview on the causal structure between the variables, we employed the PC algorithm to explore the conditional independence structure among our five observed variables. By estimating this graph, we can also evaluate whether the assumptions needed for a valid IV analysis, particularly the Exclusion and Relevance are satisfied, and determine if any confounders are present.

Validating the IV Conditions: From the CPDAG estimated in Figure 4a, we observe the edge $\texttt{QOB} \to \texttt{YOE}$, satisfying the *Relevance* condition for an instrumental variable. We have further tested the effect of QOB on LWW for each YOB individually using the permutation test and achieved positive results - see Appendix 6.3. The path $\texttt{QOB} \to \texttt{YOB} \to \texttt{YOE}$ reflects a plausible temporal ordering, where QOB affects YOB, which in turn influences YOE. Additionally, the **absence** of a directed edge $\texttt{QOB} \to \texttt{LWW}$ ($\texttt{QOB} \not\to \texttt{LWW}$) supports the *Exclusion* restriction.

Moreover, the edges $\texttt{SOB} \to \texttt{YOE}$ and $\texttt{YOB} \to \texttt{YOE}$ suggest that SOB and YOB may act as confounders and should therefore be included as control variables. Finally, the presence of the edge LWW $\rightarrow$ YOE may indicate reverse causality or a statistical artifact; however, this does not invalidate the IV strategy as long as the conditional independence QOB $\perp$ LWW | YOE holds, a conditional independence that can be tested against marginal independence (found by PC Algorithm below) using the techniques provided by Guo and Richardson [2020].

Causal structure resulted from the PC Algorithm: The algorithm identifies independencies LWW $\perp$ YOB, LWW $\perp$ QOB, LWW $\perp$ SOB, and QOB $\perp$ SOB.

Due to the limitations of PC Algorithm (elaborated in Appendix), we will switch to the more conservative FCI Algorithm.



(a) CPDAG (PC Algorithm) (b) PAG (FCI Algorithm)

Figure 4: Comparison of causal structures inferred by PC and FCI algorithms

FCI Algorithm: The Fast Causal Inference (FCI) algorithm addresses PC's limitation by accounting for latent confounders, providing a more appropriate framework for our IV analysis where hidden common causes between YOE and LWW likely exist.

Under faithfulness, both PC and FCI identify identical conditional independence structures, which we observe in our analysis. However, FCI's Partial Ancestral Graph (PAG) in Figure 4b differs in causal interpretation: every directed edge ($\rightarrow$) becomes a circle-arrow edge ($\circ\rightarrow$). This transformation occurs because while PC assumes all causally relevant variables are observed, FCI acknowledges that latent confounders may exist, making definitive causal direction determination impossible from observational data alone.

The circle-arrow edges suggest hidden confounders explain observed associations. The relationship YOE $\circ \rightarrow$ LWW can be explained by unobserved variables like ability or family background acting as common causes (YOE $\leftarrow$ U $\rightarrow$ LWW). This aligns with economic theory suggesting multiple unmeasured factors simultaneously influence educational choices and labor market outcomes.

Our analysis reveals potential faithfulness violations. We observe both marginal independence (QOB $\perp$ LWW) and conditional independence (QOB $\perp$ LWW | YOE) holding approximately. This demonstrates a fundamental limitation: even sophisticated causal discovery algorithms cannot fully resolve endogeneity when faithfulness may not hold, explaining why IV approaches remain necessary despite their limitations.

4 Polytope Approach and ACE Bounds

To assess causal effects without functional form assumptions, we implement a non-parametric polytope approach following Richardson and Robins [2014]. This method transforms variables into binary indicators: Y (whether LWW exceeds median), X (whether YOE exceeds threshold k), and Z (whether born in first half-year). The method constructs conditional probabilities $\mathbf{p} = [p(y, x|z)]$ for all combinations, yielding an 8-element vector, verifies IV inequalities, and computes Average Causal Effect bounds: $\text{ACE} = P(Y(1) = 1) - P(Y(0) = 1)$.

For education thresholds $k \in \{0, \ldots, 20\}$, we examine bounds on "having education of at least k years" affecting "having above-median wages." Figure 5 shows the positive part of the interval increases with k up to $k = 12$, suggesting that having at least 12 YOE is more likely to be associated with the positive causal effect on LWW. This pattern highlights that the causal impact of YOE intensifies until 12 years, likely corresponding to high school completion.

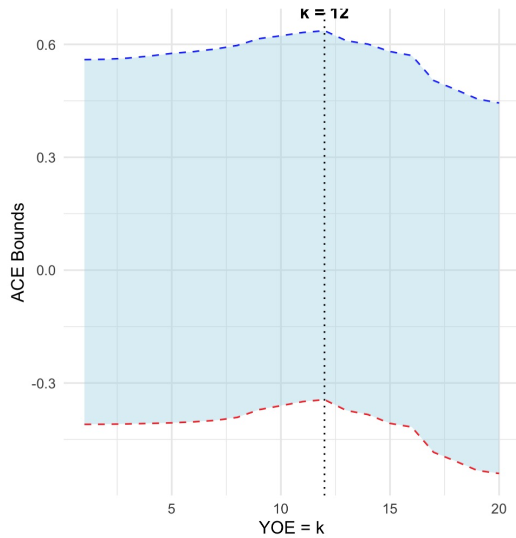


Figure 5: ACE bounds based on *T. S. Richardson and J. M. Robins (2014)*

IV inequalities were violated at extreme thresholds (0, 1, 20 years) but satisfied for thresholds 2-19 years. Table 7 presents bounds for select thresholds. **Crucially, all valid bounds are very wide (width** ≈ 0.99**)**, indicating very loose identification, **and all valid bounds contain zero**, meaning without parametric assumptions, we cannot rule out that education has no causal effect on above-median wages. Figure 6 visualizes the polytope geometry, showing feasible regions defined by IV inequalities and how ACE bounds derive from extreme points.

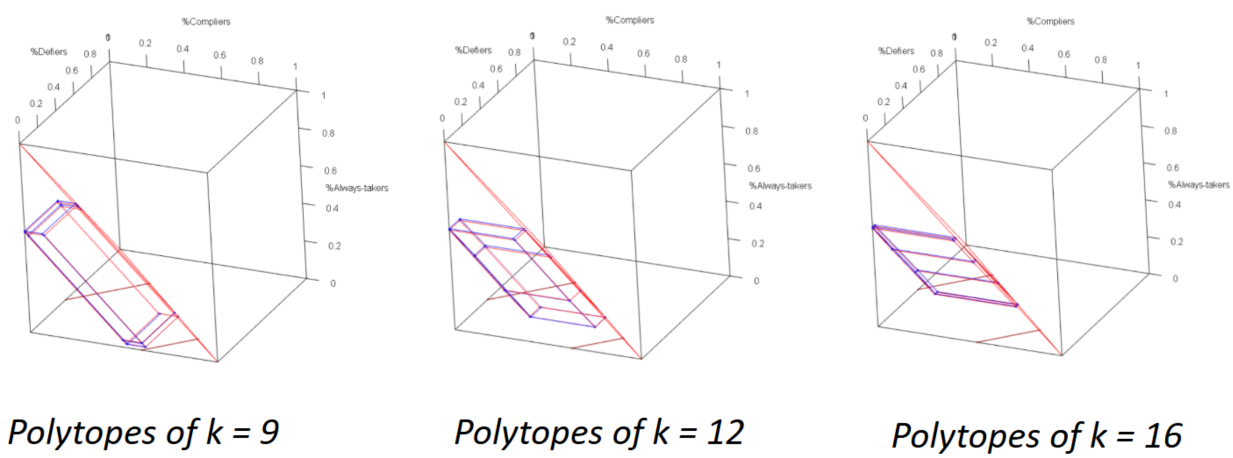


Figure 6: Polytopes of Dichotomized `YOE`, `LWW`, and `QOB`

5 Conclusion

Our analysis reveals a fundamental tension between methodological approaches. Parametric 2SLS estimates suggest significant 10.3% returns to education, while non-parametric bounds universally contain zero, meaning we cannot reject the null hypothesis of no causal effect without functional form assumptions. This contrast highlights the risks of relying solely on parametric precision in causal analysis.

Causal discovery provides crucial context for these conflicting results. While PC algorithm identifies potential causal effect structures, FCI transforms all directed edges into circle-arrow edges, acknowledging that causal directions cannot be definitively established from observational data when latent confounders exist. This confirms that unmeasured confounding remains a challenge even with instrumental variables.

The weak instrument problem emerges consistently across all approaches. Despite a first-stage F-statistic of 36.4, `QOB` explains less than 0.01% of education variation and correlates minimally (0.01) with wages. This weakness produces precise but potentially unreliable parametric estimates, while generating wide non-parametric bounds ($\approx$ 0.99 width) that transparently reflect the instrument's limited information content.

Our findings indicates the tradeoff between parametric precision versus non-parametric honesty. The 2SLS validity depends on linearity assumptions: the data alone cannot identify even the causal effect's sign. Non-parametric bounds honestly reflect the uncertainty inherent in weak instruments, showing what can be learned without restrictive assumptions.

The wide, zero-containing bounds demonstrate that while `QOB` is statistically significant, it induces insufficient variation for confident causal identification without parametric assumptions. This provides evidence of the `QOB` instrument's limitations and aligns with critiques in the literature.

References

J. D. Angrist. Angrist data archive. https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive, 1991. Datasets used: asciiqob.zip, QOB.rar, and NEW7080.rar.

J. D. Angrist and A. B. Krueger. Does compulsory school attendance affect schooling and earnings?*. *The Quarterly Journal of Economics*, 106(4):979–1014, 11 1991. doi: 10.2307/2937954. URL https://doi.org/10.2307/2937954.

F. R. Guo and T. S. Richardson. On testing marginal versus conditional independence. *arXiv preprint arXiv:1906.01850*, 2020. URL https://doi.org/10.48550/arXiv.1906.01850.

T. S. Richardson and J. M. Robins. ACE Bounds; SEMs with Equilibrium Conditions. *Statistical Science*, 29(3):363 – 366, 2014. doi: 10.1214/14-STS485. URL https://doi.org/10.1214/14-STS485.

6 Appendix

6.1 Visualization

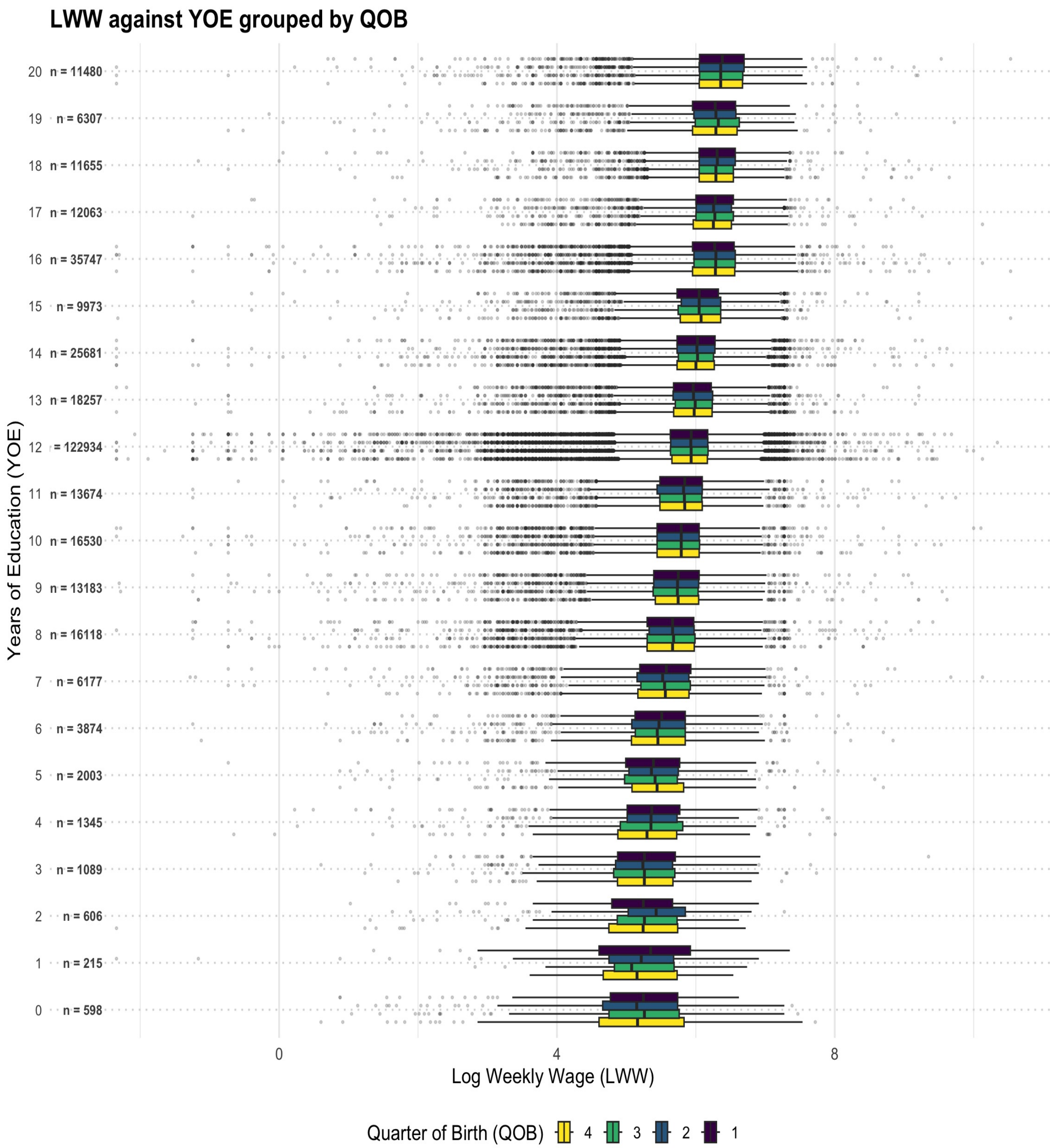


Figure 7: `LWW` against `YOE`, grouped by `QOB`

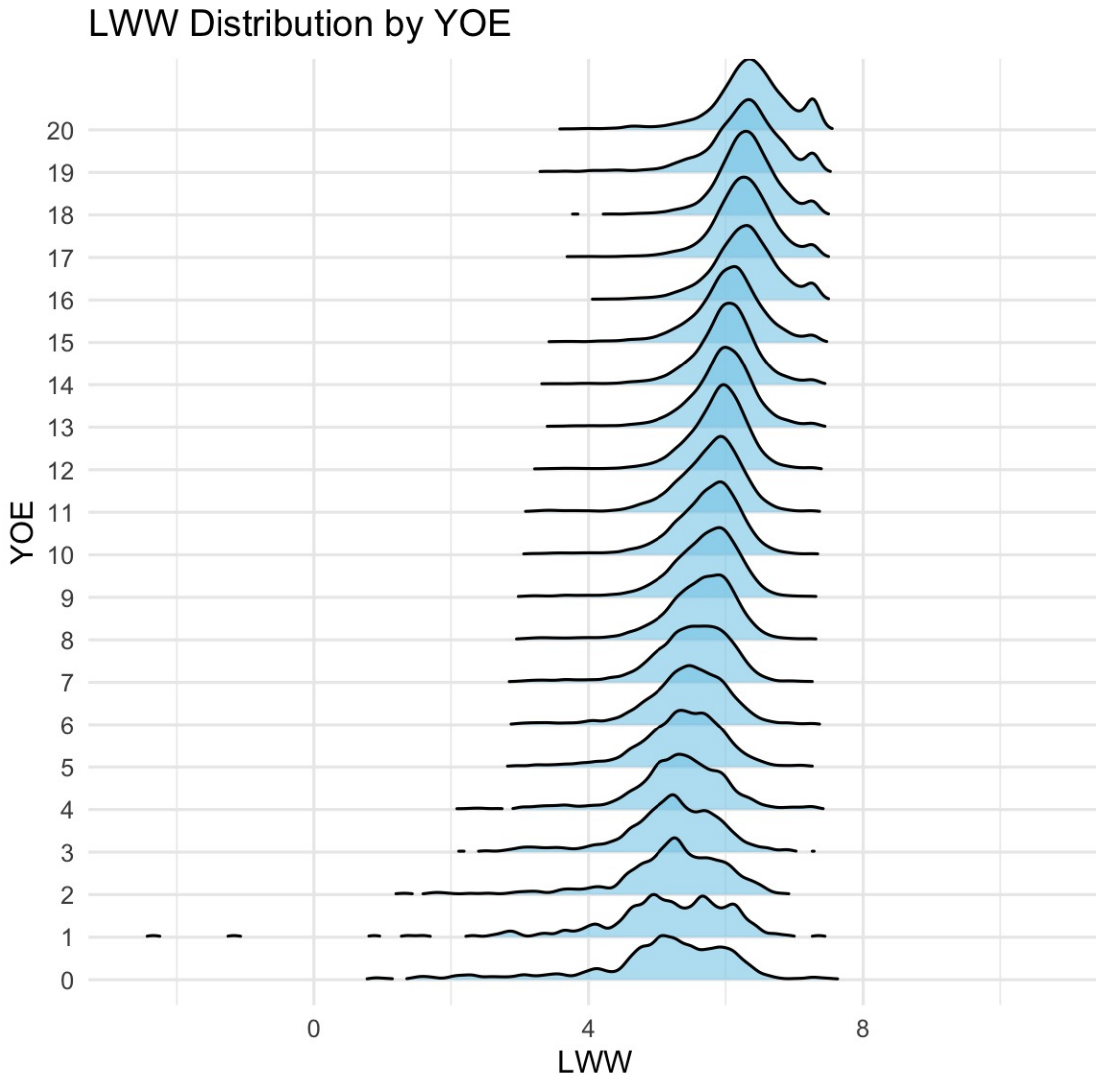


Figure 8: Distribution of `LWW` among different `YOE` values

6.2 Tables

Table 2: State Names and Codes

Code	State	Code	State
1	Alabama	30	Montana
2	Alaska	31	Nebraska
4	Arizona	32	Nevada
5	Arkansas	33	New Hampshire
6	California	34	New Jersey
8	Colorado	35	New Mexico
9	Connecticut	36	New York
10	Delaware	37	North Carolina
11	D.C.	38	North Dakota
12	Florida	39	Ohio
13	Georgia	40	Oklahoma
15	Hawaii	41	Oregon
16	Idaho	42	Pennsylvania
17	Illinois	44	Rhode Island
18	Indiana	45	South Carolina
19	Iowa	46	South Dakota
20	Kansas	47	Tennessee
21	Kentucky	48	Texas
22	Louisiana	49	Utah
23	Maine	50	Vermont
24	Maryland	51	Virginia
25	Massachusetts	53	Washington
26	Michigan	54	West Virginia
27	Minnesota	55	Wisconsin
28	Mississippi	56	Wyoming
29	Missouri		

Regression of `LWW` on `QOB` - Table 5: The estimated coefficients suggest that individuals with `QOB` = 3 and `QOB` = 4 earn, on average, significantly higher `LWW` than those with `QOB` =1. The effect for `QOB` = 2 is positive but not statistically significant, which makes sense as being closer to `QOB` = 1. While the explained variance is minimal, as expected from an instrument, the statistically significant coefficients on `factor(QOB)3` and `factor(QOB)4` confirm that `QOB` has a meaningful reduced-form effect on earnings. This supports the validity of `QOB` as a relevant instrument in the IV framework for estimating the causal effect of education on earnings.

Table 3: Justification for variable selection (5 out of 27) and subsetting the sample based on sex and age

Choice	**Reasoning**
QOB	Used as an instrument for YOE due to exogenous variation from school entry and compulsory laws.
YOE	Treatment variable whose causal effect is being estimated.
LWW	Main outcome variable measuring the returns.
YOB	Used to control for cohort trends; avoids redundancy with AGE.
SOB	Included to capture heterogeneity in compulsory schooling laws and school entry age policies across states.
Subsetting	"We mainly focus on 40–49 year-old men, whose wages are hardly related to age." (p. 995 of Angrist and Krueger [1991]). This avoids confounding due to life-cycle earnings growth. The focus on men avoids gender-specific patterns during this time period.
Features Removed	AGE and AGEQ (fully collinear with YOB and QOB), MARRIED, RACE, ENOCENT, SMSA, and 6 regional dummy variables such as MIDATL—excluded to avoid post-treatment bias. These were only included in robustness checks. Additionally, 7 unnamed variables were removed (v3, v7, v14, v15, v22, v23, v26), including a full-zero column (v26) and likely redundant/unclear encodings.

Table 4: Second-stage IV regression of LWW on predicted YOE_hat

	Estimate	Std. Error	t value	Pr($>|t|$)
(Intercept)	4.58978	0.26158	17.547	$< 2 \times 10^{-16}$ ***
YOE_hat	0.10260	0.02048	5.009	5.48×10^{-7} ***

Table 6: ACE Bounds for Key Education Thresholds (Binary Z)

Education Threshold	Lower Bound	Upper Bound	Contains Zero?
8 years (Middle School)	-0.4291	0.5633	Yes
12 years (High School)	-0.3668	0.6165	Yes
13 years (Some College)	-0.3606	0.6277	Yes
16 years (College Degree)	-0.3924	0.6002	Yes

Table 5: Regression of LWW on QOB

	Estimate	Std. Error	t value	Pr($>|t|$)
(Intercept)	5.891596	0.002375	2480.425	$< 2 \times 10^{-16}$ ***
factor(QOB)2	0.004540	0.003375	1.345	0.179
factor(QOB)3	0.014925	0.003309	4.511	6.45×10^{-6} ***
factor(QOB)4	0.013490	0.003368	4.006	6.19×10^{-5} ***

6.3 Permutation Test for QOB–YOE Pattern

To evaluate whether QOB influences YOE for each YOB, we analyzed average YOE across QOB for individuals born between 1930 and 1940. For each of these ten YOB cohorts, we ranked the four QOB groups (1 to 4) based on their average YOE. We defined a year as **favorable** if both QOB 3 and QOB 4 had higher average YOE than both QOB 1 and QOB 2. In other words, QOBs 3 and 4 should appear *after* QOBs 1 and 2 in the YOE ranking. This captures the hypothesis that individuals born later in the year (Q3/Q4) tend to receive more schooling—consistent with policy-driven schooling age cutoffs. **We observed 7 favorable years out of the 10.** To determine whether this is more than would be expected by chance, we conducted a permutation test. We simulated 10,000 datasets by randomly assigning QOB ranks and counted how often at least 7 favorable years occurred. The result yielded a **permutation p-value of 0.0005**, providing strong evidence against the null hypothesis of random QOB effects.

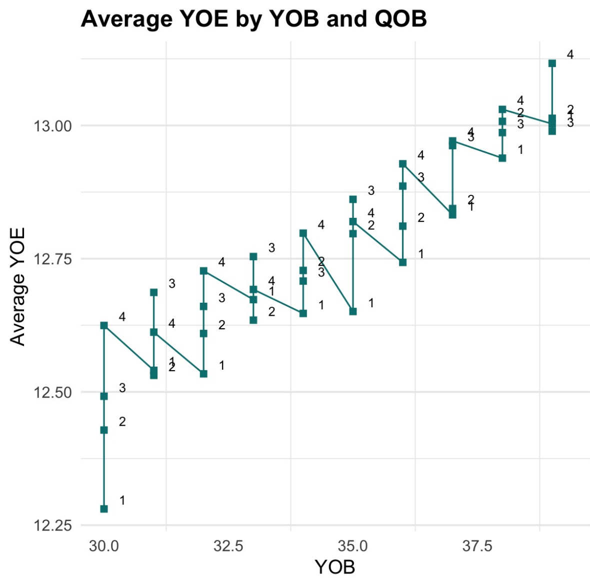


Figure A1: Average YOE by YOE and QOB. Each year displays four labeled points (QOB 1–4) connected by lines.

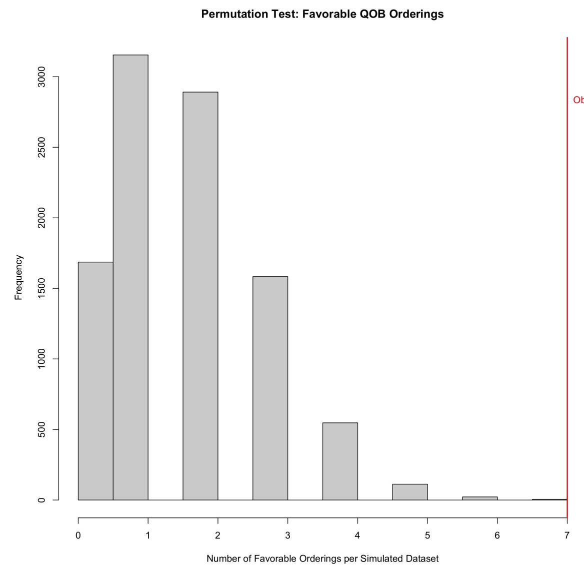


Figure A2: Permutation test distribution for the number of favorable QOB orderings in 10 simulated years. The red line indicates the observed value (7 favorable years).

These results suggest that QOB can be adopted as a valid instrument. This provides further justification for its use in our causal analysis of education and earnings.

6.4 Limitations of the PC Algorithm

While the PC algorithm provides valuable insights into the causal structure among our observed variables, it operates under the unrealistic assumption that all relevant confounders are observed and measured (no hidden variables). In our instrumental variable context, this assumption is likely violated since we are not considering everything, including but not limited to the omitted variables, which may be unobserved confounders affecting both YOE and LWW. The presence of such latent confounders means that the causal structure estimated by PC algorithm may be inaccurate or incomplete.

6.5 Non-Parametric IV Constraints for a Multi-Level Instrument

The non-parametric polytope approach, as described by Richardson and Robins [2014], relies on a set of linear inequalities that must be satisfied if a variable is a valid instrument. In the context of our analysis, the instrument -`QOB` - is not binary but has four levels. Therefore, if want to keep it to be 4 levels, the standard IV inequalities must be generalized. This generalization is a direct consequence of the core instrumental variable assumptions of Independence and the Exclusion Restriction. Because these assumptions must hold universally for all four quarters of birth, any pair of distinct quarters can be isolated and treated as its own valid binary instrument experiment. This necessitates the systematic, pairwise application of the original four inequalities across all possible pairings of the `QOB` levels. For the four levels of the instrument `QOB`, denoted by the set $\{z_1, z_2, z_3, z_4\}$, the inequalities must be checked for all $\binom{4}{2} = 6$ unique pairs of quarters. For any distinct pair (z_i, z_j) where $i, j \in \{1, 2, 3, 4\}$ and $i \neq j$, the following four inequalities must be satisfied by the conditional probabilities in the vector $\mathbf{p}$. This results in a total of $4 \times 6 = 24$ linear constraints to verify before computing bounds on the Average Causal Effect (ACE).

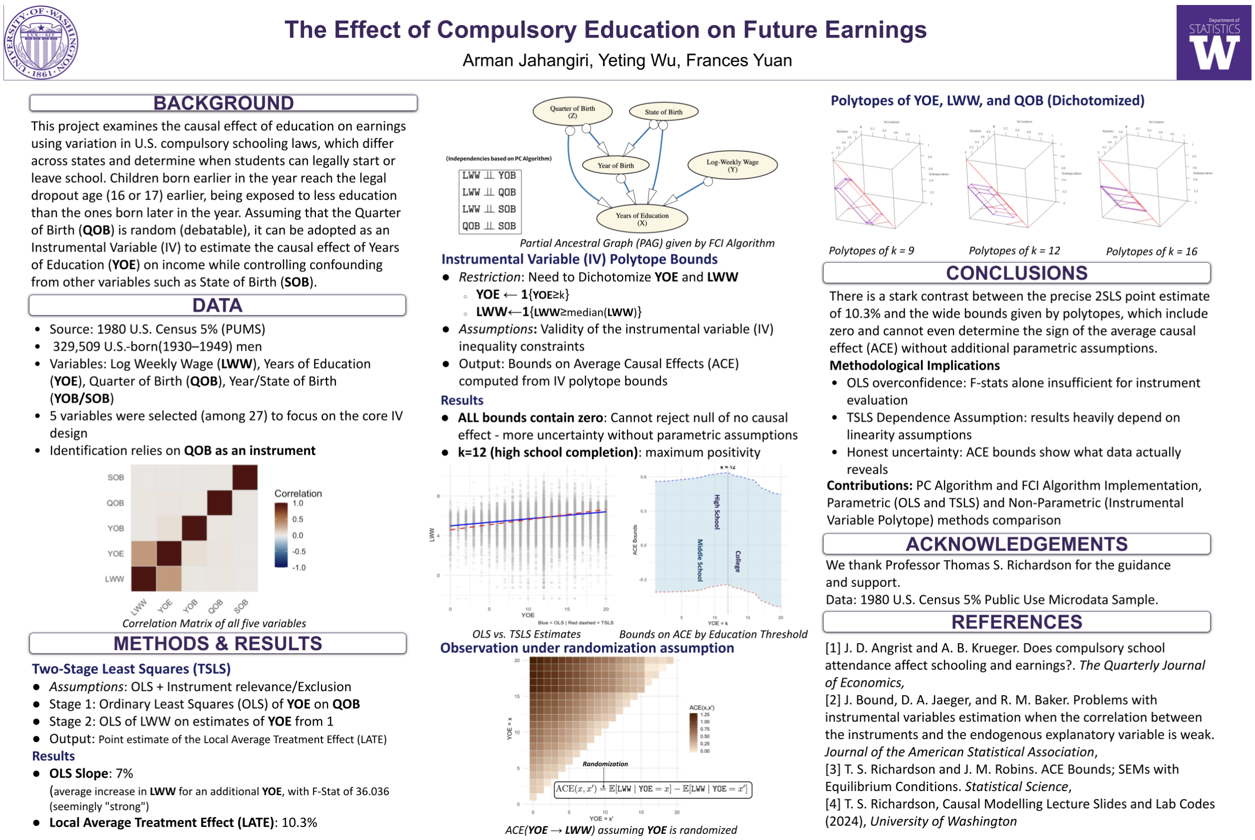
$$P(Y = 0, X = 0|Z = z_i) + P(Y = 1, X = 0|Z = z_j) \leq 1$$
$$P(Y = 1, X = 0|Z = z_i) + P(Y = 0, X = 0|Z = z_j) \leq 1$$
$$P(Y = 0, X = 1|Z = z_i) + P(Y = 1, X = 1|Z = z_j) \leq 1$$
$$P(Y = 1, X = 1|Z = z_i) + P(Y = 0, X = 1|Z = z_j) \leq 1$$

The resulting bounds on the ACE are very similar to those obtained using a binary Z. However, while the instrumental variable (IV) inequalities for the binary Z were only violated at thresholds $k = 0, 1, 20$, the IV conditions for the four-level Z, being stricter, were violated at $Z = 0, 1, 2, 3, 20$. The table below summarizes the ACE bounds.

Table 7: ACE Bounds for Key Education Thresholds (4-level Z)

Education Threshold	Lower Bound	Upper Bound	Contains Zero?
8 years (Middle School)	-0.4342	0.5611	Yes
12 years (High School)	-0.3753	0.6205	Yes
13 years (Some College)	-0.3632	0.6274	Yes
16 years (College Degree)	-0.3896	0.6016	Yes

7 Contributions, Photos, and Poster



Poster

Arman Jahangiri

TSLS, PC Algorithm, ACE Bounds, Consultations, Exploratory Analysis, Visualizations

Yeting Wu

Introduction, Exploratory Analysis, Data Preparation, Poster Design

Frances Yuan

Data Recognition, FCI Algorithm, Polytopes and interpretations, Interpretation, Poster Administration, Consultation