Final Project: Analyzing Schooling Returns via a Regression Discontinuity Design

Jasmine Jia

I. Introduction

This report investigates the causal effects of introducing compulsory schooling requirements on educational attainment, earnings, marital status, and parental status. Utilizing a Regression Discontinuity Design (RDD) approach, this analysis leverages a natural policy change implemented on September 1, 1951, which mandated schooling until at least age 16 for individuals born on or after this date. By analyzing these data, our findings can provide essential insights for policymakers on the long-term benefits and social implications of compulsory education.

The research question addresses how introducing a minimum school leaving age policy influences years of schooling completed, individual's earnings, marital status, and parental status at age 45. The research question is highly policy-relevant because understanding how compulsory schooling affects earnings will help quantify the returns to education. This is crucial for justifying public expenditure on education and designing policies that optimize such investments for economic returns. Additionally, changes in educational requirements might affect personal decisions such as marriage and having children. For instance, increased education might delay these life events or alter demographic trends, which could have long-term implications for social policies such as housing, social welfare, and healthcare. Thirdly, compulsory schooling can level the playing field by ensuring that all children, regardless of background, receive a minimum level of education. This can promote social mobility and reduce inequality, which are significant considerations for any government.

Introducing compulsory schooling helps to solve the endogeneity issue that would be present if we solely used years of schooling as our treatment variable. Endogeneity - the case where there is an unobserved factor(s) in the error term that is correlated with X (here, years of schooling) and that also affects Y (e.g., earnings) - will give rise to biased estimates. There are several pathways through which endogeneity could be at play. For example, individuals may choose their level of education based on unobserved characteristics like motivation, ability, or family background, which are also correlated with earnings and other life outcomes. This is a self-

selection problem. Second, higher potential earnings could influence an individual's decision to pursue more education, resulting in concerns about reverse causality. Third, we have an omitted variable bias concern if other unmeasured factors such as regional economic conditions or parental education levels might influence educational attainment and the outcomes being studied.

A quasi-experimental research strategy is particularly useful in this context because it allows for estimating causal effects in this setting where a controlled experiment is not feasible, ethical, or practical. By exploiting a natural experiment where the treatment (compulsory schooling) is assigned based on an exogenous rule (birthdate relative to a policy cutoff), the quasi-experimental research strategy also helps to account for endogeneity issues.

II. Research Design

The research design employed to estimate the causal effect is a Regression Discontinuity Design (RDD), specifically utilizing a fuzzy RDD approach. This approach leverages the natural discontinuity created by the threshold date of September 1, 1951, to distinguish between individuals subjected to a new compulsory schooling policy and students who were not. This setup segments individuals into a control group (births between 1947 and pre-September 1, 1951) and a treatment group (births on or after September 1, 1951 through 1955) and compares the outcomes for individuals on either side of the cutoff. The assumption is that those just below and just above the cutoff are comparable in all significant respects, except for the treatment effect of the policy.

Unlike a sharp RDD, the assignment to treatment in a fuzzy RDD is not strictly deterministic at the cutoff date, reflecting some variability in compliance with the policy: in other words, some students not subject to the policy may still have stayed in school until at least age 16, and conversely, some students subject to the policy may not have adhered to it and dropped out before age 16. Because of this difference in treatment assignment and actual treatment status, the fuzzy RDD capitalizes on an instrumental variable (IV) approach by instrumenting treatment status with treatment assignment. In this study, we instrument whether a student attended school until at least age 16 with whether the student was subject to the minimum drop-out age policy.

The key variables in this analysis are defined as follows:

- **X (treatment variable):** The primary treatment variable indicates whether a student left school at or after age 16.
- **Y (outcome variable):** We study three long-term outcomes of interest, each measured at age 45: log earnings, marital status, and parental status.
- **Z (Instrument):** The instrument used is the minimum school age requirement, defined by whether an individual was born on or after September 1st, 1951, which is the cutoff date for the new compulsory schooling policy.

The use of this IV approach means that two critical conditions must be satisfied:

- **Inclusion condition**: This condition requires that the instrument (whether the student was subject to minimum school age requirement) function as a strong predictor of the treatment variable (completing schooling until age 16). A robust first-stage relationship between Z and X must support using Z as an instrument to isolate the exogenous variation in X (i.e., a t-statistic of absolute value 3 or greater). The results of our first-stage estimation (discussed below) demonstrate that this condition passes.
- **Exclusion condition:** This condition requires that the instrument must not directly affect the outcome variables except through its impact on the treatment variable. This means that any effect the policy has on earnings, marital status, and parental status must operate solely through its impact on extending required schooling years. This condition is crucial for maintaining the validity of the IV estimation, as it ensures that the instrument does not correlate with any other factors that might independently affect the outcome, potentially resulting in bias. In contrast to the inclusion condition, we cannot validate the exclusion condition via statistical tests.

III. Construction of the running variable

In applying RDD to our analysis, we constructed a running variable that measures the exact position of each observation relative to a predetermined cutoff. This study's cutoff date is the day the compulsory schooling policy was implemented: September 1st, 1951. Therefore, the running variable is calculated as the number of days between an individual's date of birth and the policy introduction date. To be more specific, for each participant, the running variable is computed by subtracting the policy date from the individual’s DOB. The result is an integer representing the number of days before or after the policy implementation. Individuals born before this date will

have negative running variables, while those born on or after will have positive running variables.

To address potential issues of granularity and overfitting in the RDD approach, the running variable is further segmented into binned categories based on 90-day intervals. Each 90-day period relative to the cutoff date forms a bin. Thus, individuals born within 90 days after September 1, 1951, fall into bin 0; those born 90 to 179 days after the cutoff fall into bin 1, and so on. Similarly, those born 1 to 90 days before the cutoff are in bin -1, 91 to 180 days before are in bin -2, etc.

Constructing binned running variables can help us ensure the precision and reliability of the RDD analysis. It allows for an examination of trends and discontinuities at and around the threshold, which is critical for validating the assumption of the RDD that units just on either side of the cutoff are comparable.

Furthermore, using bins facilitates a more manageable and less volatile analysis. It helps stabilize the estimates by pooling more data points, thereby mitigating the influence of outliers or anomalous values that could skew the results if each day was analyzed individually. Additionally, binning can enhance the interpretability of graphical representations of the data, making it easier to visualize and communicate the results.

IV. First Stage

a. Left school at age 16 or later

The first step in continuing our fuzzy regression discontinuity design analysis is to validate whether being subject to the age policy is a valid instrument for staying in school until at least age 16. Thus, we first examine whether the inclusion condition holds by running the following ordinary least squares (OLS) specification:

(1) $LeftSchAge16OrLater_i = \alpha_0 + \alpha_1 runvar_i + \beta_1 SubjectToPolicy_i + \alpha_2(runvar_i * SubjectToPolicy_i) + \epsilon_i$

Exhibit 1. First Stage OLS results (outcome = left school at age 16+).

```
t test of coefficients:

                                 Estimate   Std. Error t value            Pr(>|t|)
(Intercept)                   0.869285271  0.001840167  472.39 <0.0000000000000002 ***
runvar                        0.000000524  0.000002331    0.22                 0.82
subject_to_policy             0.130714729  0.001840167   71.03 <0.0000000000000002 ***
I(subject_to_policy * runvar) -0.000000524  0.000002331   -0.22                 0.82
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The above regression results clearly show a strong positive statistically significant relationship between being subject to the policy (born on or after September 1, 1951) and whether a student completed at least 16 years of schooling (i.e., t-statistic on the subject to policy indicator = 71.03). Thus, we find that the inclusion condition passes.

To visually assess the validity of the instrument, we plot both the sharp first stage and fuzzy first stage results. The sharp first stage plot examines assigned treatment against the running variable. All students with a birthdate before the cutoff have an assigned treatment of zero, and all other students have an assigned treatment status of one. As discussed, in actuality treatment was not deterministic in this manner: as seen in the fuzzy first stage plot, at least 85% of students in each bin to the left of the cutoff remained in school until at least age 16. As expected, the sharpest discontinuity occurs at the cutoff where the policy went into effect. There was almost 100% adherence to the policy for cohorts born on or after the cutoff date.

As an additional robustness check, we conducted a permutation test of 500 simulations: in each simulation, we randomly assigned the distance from the cutoff (thereby randomly assigning the instrumental variable) and re-estimated regression equation (1) using this randomly assigned running variable and instrument. As expected, we find that the distribution of the coefficients on the randomly assigned instrumental variable for these 500 simulations centers on zero (i.e., our instrumental variable is not just a good predictor of X by random chance). Our real first-stage estimate for the instrumental variable (0.131) is far to the right of the distribution from the permutation test (Figure 4). In this context, the coefficient means that students subject to the policy were 13.1% more likely to leave school at age 16 or higher compared to students not subject to the policy.

Figure 1.

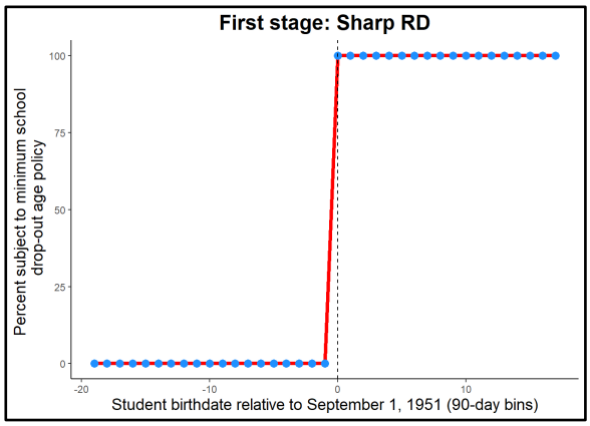


Figure 2.

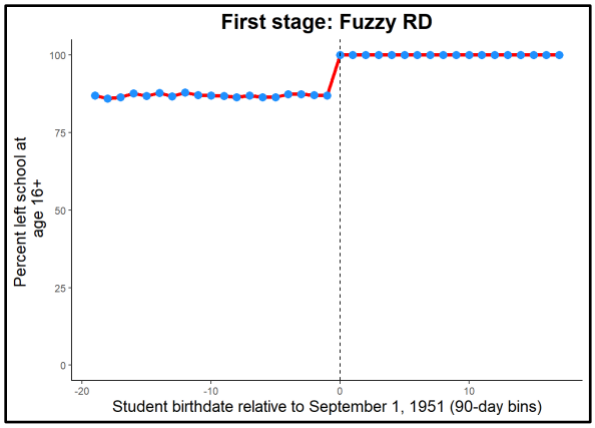


Figure 3.

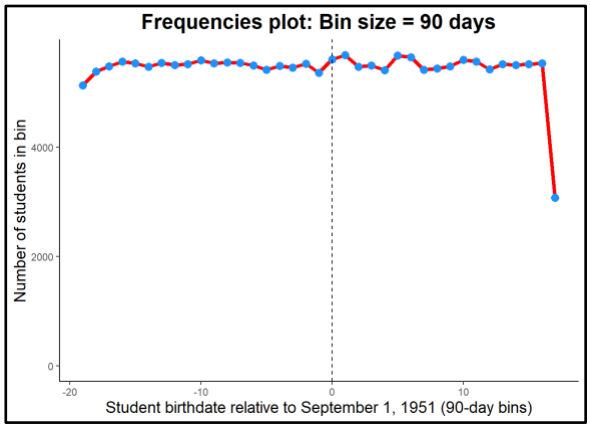
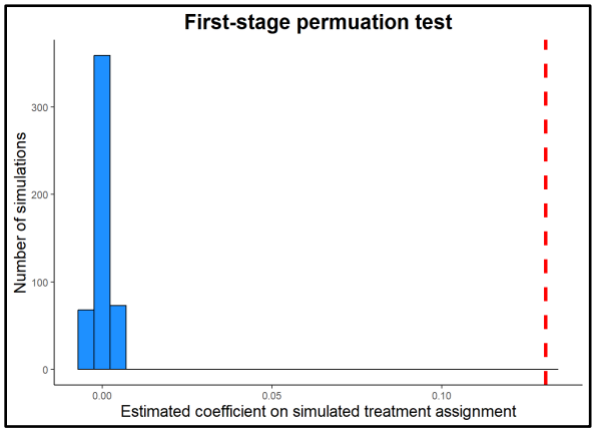


Figure 4.



b. Years of schooling

As part of the first stage, we also re-ran equation (1) but replaced the dependent variable as years of schooling:

(2) $YearsOfSchooling_i = \alpha_0 + \alpha_1 runvar_i + \beta_1 SubjectToPolicy_i + \alpha_2(runvar_i * SubjectToPolicy_i) + \epsilon_i$

Exhibit 2. First Stage OLS results (outcome = years of schooling).

```
Coefficients:
                          Estimate   Std. Error t value            Pr(>|t|)
(Intercept)           13.068417584  0.014442631  904.85 <0.0000000000000002 ***
runvar                 0.000000573  0.000014674    0.04                 0.97
subject_to_policy      0.190122089  0.020720728    9.18 <0.0000000000000002 ***
I(subject_to_policy * runvar)  0.000029191  0.000021939    1.33                 0.18
```

Similarly, we find that being subject to the policy is a positive and statistically significant ($p < 0.001$) predictor of years of total schooling attained. The coefficient indicates that students who were subject to the policy, on average, completed 0.19 more years of total schooling than did students not subject to the policy.

V. Assumptions

In the context of RDD, two other important assumptions must hold. First, students must not be able to choose their treatment assignment which translates to the choice of being born before September 1, 1951 vs. on or after this date. Students themselves have no choice in the birth date. However, depending on when the announcement was made about this policy, treatment manipulation could plausibly occur. For example, if the forthcoming policy was announced in the summer of 1950; this could presumably have factored into family planning. In contrast, if the policy was announced after September 1951 but before the students entered grade school (e.g., perhaps in September 1953), there would be no way to manipulate treatment status. We graphically validate whether we see evidence that families could choose a student's treatment status by examining the bin size of bins above and below the threshold (i.e., looking for bunching). While we do see a decrease in frequency between bin -2 and bin -1, as well as an increase between bin 0 and bin 1, it is by a marginal amount (Figure 3). In zooming out and looking at all bins, we generally see that the bin size stays relatively constant, consistent with the necessary assumption. Note that the bin size for the last bin is much smaller than the rest as a result of the binning methodology and is not a symptom of treatment manipulation issues (i.e., data stopped being collected during this 90-day bin).

The second assumption that must hold is that there are no other discontinuous changes at the cutoff. We would traditionally check this assumption by examining covariate index plots for all covariates in the analysis; however, since this dataset does not include any covariates, this step is skipped.

Finally, as noted above, in all graphs in this paper we use bin sizes of approximately 90 days (equating to roughly three months), as this provides roughly 20 bins both pre- and post-cutoff, respectively. Additionally, all analyses use the full available bandwidth (i.e., 8 years of data).

VI. Second Stage

Next, we evaluate our policy outcomes of schooling over 16 years of age using whether an individual was subject to the threshold as an instrument. The three outcomes of interest we focus on are earnings at age 45, whether the individual is married at age 45, and whether an individual has children at age 45.

For each of these three outcomes, we estimate a two-stage least squares (2SLS) regression to obtain the estimated relationship between compulsory schooling and the outcome of interest. That is, we use the predictions on X from the first stage (i.e., $\widehat{LeftSchoolAfter16_\iota}$) as the independent variable in the second-stage equation. We define each second-stage equation below.

a. Earnings

We use the following equation to evaluate this impact:

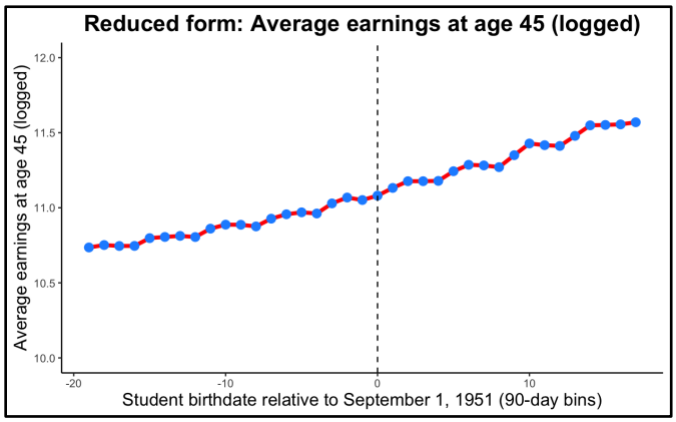
$$(3)\ Earnings45_i = \beta_0 + \beta_1 runvar_i + \beta_2 \widehat{LeftSchoolAfter16_\iota} + \beta_3(\widehat{LeftSchoolAfter16_\iota} * runvar_i) + \epsilon_i$$

Exhibit 3. 2SLS results (outcome = earnings).

```
Coefficients:
                                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)                     10.9394943  0.0371873  294.17 < 0.0000000000000002 ***
left_sch_age16_or_later          0.1335516  0.0397821    3.36              0.00079 ***
I(left_sch_age16_or_later * runvar)  0.0009451  0.0000417   22.68 < 0.0000000000000002 ***
runvar                          -0.0006127  0.0000387  -15.82 < 0.0000000000000002 ***
```

From the regression results, we see that the coefficient for leaving school after age 16 is 0.134, or a 13.4% increase in earnings at age 45 relative to those who left school before age 16 (recall that the earnings outcome is logged). This coefficient is statistically significant at $p < 0.001$. The coefficient on the interaction between runvar (distance from the threshold) and attending school until at least age 16 is 0.0095, meaning that the slope of earnings growth for those that left school at age 16 or later is 0.95 percentage points

Figure 5.



higher compared to the slope of earnings growth for those not subject to the policy. Adding the coefficient of the running variable and the instrumental variables, we can interpret that the rate of change in earnings for those who left school after age 16 is 0.033 percentage points.

b. Marriage

We also look at the relationship remaining in school until at least age 16 has on marital status at 45. We use the following equation to evaluate this impact:

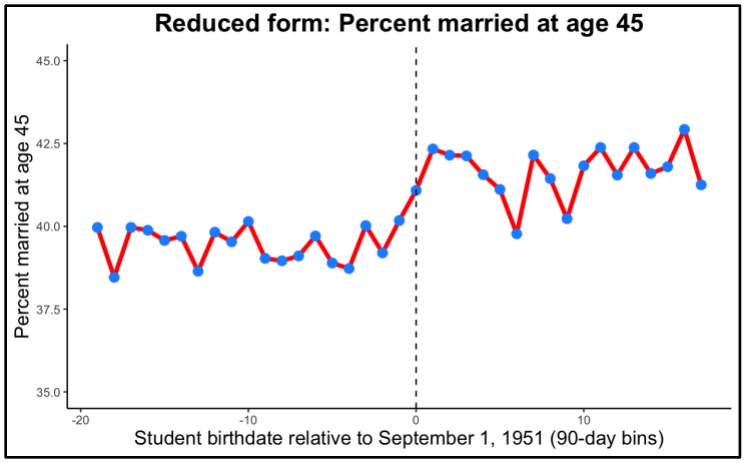
$$(4)\ Married45_i = \beta_0 + \beta_1 runvar_i + \beta_2 \widehat{LeftSchoolAfter16}_i + \beta_3(\widehat{LeftSchoolAfter16}_i * runvar_i) + \epsilon_i$$

Exhibit 4. 2SLS results (outcome = marriage).

```
Coefficients:
                                 Estimate Std. Error t value             Pr(>|t|)
(Intercept)                     0.2598050  0.0311468    8.34 < 0.0000000000000002 ***
left_sch_age16_or_later         0.1543565  0.0333201    4.63            0.0000036 ***
I(left_sch_age16_or_later * runvar) 0.0000281  0.0000349    0.81                 0.42
runvar                         -0.0000251  0.0000324   -0.77                 0.44
```

From the regression results, we see that the coefficient for leaving school after age 16 is 0.154 or a 15.4% probability increase in being married at age 45 than someone who left school before age 16. This coefficient is statistically significant at $p < 0.001$. The coefficient on the interaction between runvar and attending school until at least age 16 is 0.0000281 and not statistically significant, meaning that the slope in probability of being married is about the same on either side of the cutoff, apart from this discontinuity.

Figure 6.



c. Children

We also look at the relationship schooling over the age of 16 has on whether an individual has children at age 45. We use the distance to the threshold of the mandatory schooling policy as an instrument on schooling to isolate the effect of schooling directly on whether an individual has children.

We use the following equation to evaluate this impact:

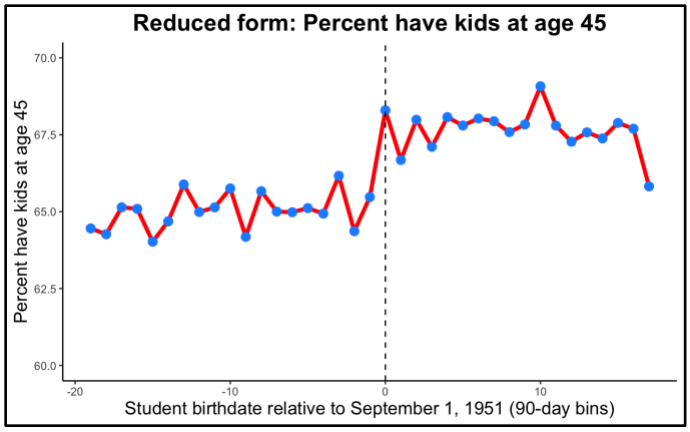
$$(5)\ HasKids45_i = \beta_0 + \beta_1 runvar_i + \beta_2 \widehat{LeftSchoolAfter16}_i + \beta_3(\widehat{LeftSchoolAfter16}_i * runvar_i) + \epsilon_i$$

Exhibit 5. 2SLS results (outcome = children).

```
Coefficients:
                                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)                      0.4848990  0.0297866   16.28 < 0.0000000000000002 ***
left_sch_age16_or_later          0.1937781  0.0318650    6.08        0.0000000012 ***
I(left_sch_age16_or_later * runvar) -0.0000432  0.0000334   -1.29                0.20
runvar                           0.0000411  0.0000310    1.32                0.19
```

From the regression results, we see that the coefficient for leaving school after age 16 is 0.1937781 or a 19.38% probability increase in having children at age 45 than someone who left school before age 16. This coefficient is statistically significant at $p < 0.001$. The coefficient on the interaction between runvar and attending school until at least age 16 is 0.0000411 and not statistically significant, meaning that the slope in probability of having children is about the same on either side of the cutoff, apart from this discontinuity.

Figure 7.



VII. Sensitivity Analysis: Permutation Testing

Permutation testing was employed to assess the robustness of estimated causal effects of schooling on various outcomes, including earnings, marital status, and number of children, based on randomized treatment assignments. For all three outcomes, a histogram was plotted to visualize the distribution of estimated coefficients on simulated treatment assignments. The histogram revealed that the absolute coefficient, indicated by the dashed red line, fell significantly to the right of the distribution. This finding suggests that the observed effect of schooling on earnings, marital status, or having children at age 45 is unlikely to be due to random chance alone. The statistical significance of this observation was underscored by the distance of the actual coefficient from the center of the distribution, indicating a level of significance beyond what would be expected by chance. This significant deviation from zero in

the real data provides strong support for the hypothesis that staying in school until at least age 16 has a substantial and statistically significant positive effect on earnings, being married, and the presence of children at age 45.

Figure 8.

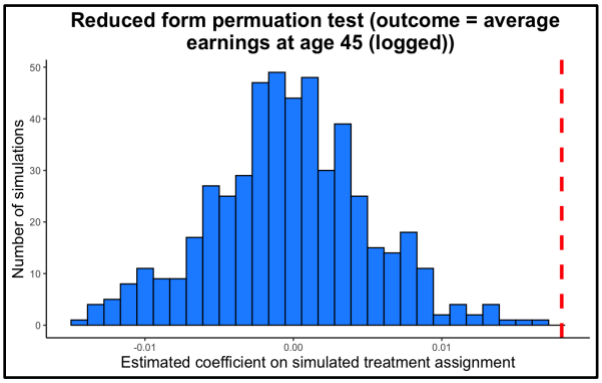


Figure 9.

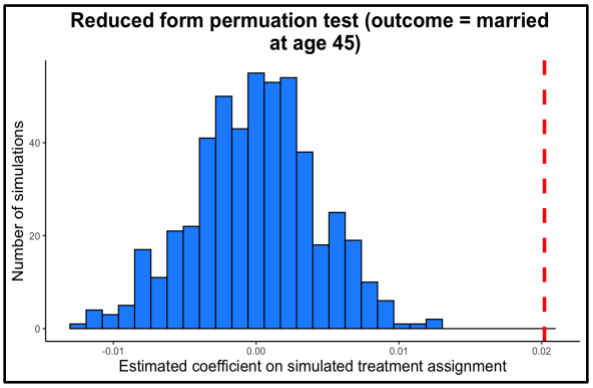
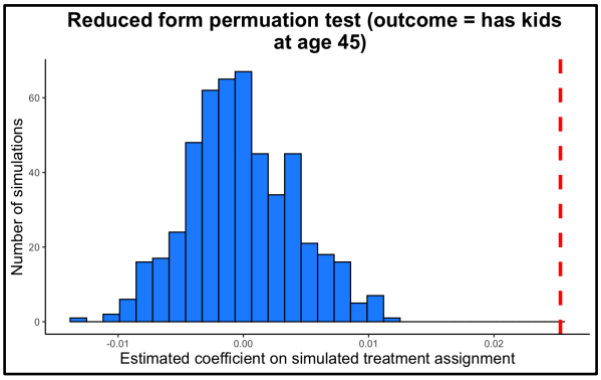


Figure 10.



VIII. Conclusion

The findings from our RDD analysis on the causal effects of compulsory schooling requirements demonstrate that increasing the minimum school leaving age not only enhances educational attainment but also significantly impacts long-term economic and social outcomes. Specifically, we observe that extending compulsory schooling correlates with higher earnings, a higher probability of marriage, and an increased likelihood of having children by age 45. These results indicate that the educational policy not only brings about economic returns but also societal impacts, such as potentially altered demographic trends. Additionally, the instrumental variable approach employed successfully addresses potential endogeneity issues, lending credibility to our conclusions that compulsory schooling directly influences these outcomes.