Problem Set 3: Answer Key (answers in **bold)**

This assignment applies the RD ideas discussed in lectures to Clark and Royer (2013), henceforth CR.

Background and data

In April 1947, the British government raised the minimum school-leaving age from 14 to 15. People born before April 1933 could leave at 14, whereas those born in April 1933 or later had to remain until 15. Not everyone would otherwise have left at 14, so the reform did not increase everyone's schooling by one year. Education is measured by age left full-time education, a multivalued treatment. Increasing this age by one year is interpreted as an additional year of education.

The data extract on the course website (`ps3_rd.dta`) combines Health Survey of England (HSE) and General Household Survey (GHS) data. It contains 81,415 UK-born respondents from birth cohorts within ten years of the threshold. The variables are:

- `r`: birth month relative to April 1933 ($r = -1$ is March 1933; $r = 0$ is April 1933).
- `educend`: age left full-time education; `smoke_ever`: ever smoked (0/1).
- `sex`: male (1) or female (0); `ghs`: GHS (1) or HSE (0).
- `yintb`, `mintb`: interview year and month; `mob`: calendar month of birth.
- `age`: age in months at interview.

Let $D_i \equiv$ `educend`$_i$, $Y_i \equiv$ `smoke_ever`$_i$, and $Z_i \equiv 1(r_i \geq 0)$. Here Z denotes exposure to the new minimum leaving age, not educational attainment. Create the following variables:

```
use ps3_rd.dta, clear
assert _N == 81415
gen byte z = (r>=0) if !missing(r)
gen double zr = z*r
```

```
gen double r2 = r*r
gen double zr2 = z*r2
gen double r3 = r2*r
gen double zr3 = z*r3
gen double age2 = age*age/100
gen double age3 = age2*age/100
```

We rescale age2 and age3 to make estimates easier to interpret.

Important!

- Use heteroskedasticity-robust standard errors throughout.
- Report probability effects as percentage points, not percent changes.
- Note that an h-month observation window is defined as $-h \leq r \leq h$.

Q1: OLS estimates

(a) Follow CR and restrict the sample to $-79 \leq r < 0$ and $D \in \{14, 15\}$. Regress ever smoking on education and an intercept. Interpret the schooling coefficient. Why is this a useful baseline estimate for thinking about the causal effect of education induced by the reform?

The coefficient is -0.04259 (robust SE 0.01069; $N = 13{,}811$). Leaving school at 15 rather than 14 is associated with a 4.26 percentage-point lower probability of ever smoking. The OLS comparison examines the same schooling margin as the reform, which induced some individuals to leave school at 15 rather than 14. In that sense it serves as a useful benchmark. Of course, individuals who leave school at 14 may differ systematically from those who leave at 15, so the OLS estimate need not identify the causal effect.

(b) Re-estimate the OLS relationship controlling for sex, a third-order polynomial in age, interview-year and interview-month indicators, the GHS indicator, and calendar-month-of-birth indicators. Compare the estimate with Q1(a) and CR's OLS result in panel A of Table 5a (OLS column for "ever smoke"). By how much does adding these controls change the estimated relationship between education and smoking?

The coefficient is -0.03023 (robust SE 0.01045; $N = 13,811$), approxi-

mately the −0.030 benchmark in CR. The coefficient falls in magnitude from 4.26 to 3.02 percentage points, about a 29% reduction.

(c) Describe two unobserved characteristics that could bias the OLS estimate in Q1(b). For each characteristic, explain how it might be related to education and smoking, and how you would expect the estimate to change if you could control for it.

Patience may raise education and reduce smoking independently of education. Family resources could raise education and reduce smoking independently of education. In both cases, controlling for those variables could make the negative estimate smaller.

Q2: Assessing RD validity

(a) Can you think of any reasons why the key continuity assumption underlying RD might fail in this context? That is, why might individuals born in March 1933 differ systematically from those born in April 1933 in ways that affect their smoking behavior independently of their education? Describe at least one scenario that could violate this assumption, even if you do not find it to be plausible. [Hint: Here's an example scenario (that you are not allowed to use!). Suppose that a sudden change in medical practices in April 1933 affected the survival chances of babies born from that month onward. This could change the composition of surviving individuals born on either side of the threshold, potentially creating systematic differences in their later smoking behavior.]

Suppose another government policy applied only to individuals born in April 1933 or later and independently affected their propensity to smoke. This would generate a discontinuity in determinants of smoking unrelated to education, violating the continuity assumption.

(b) One way to assess RD validity is to examine whether the density of the running variable changes discontinuously at the threshold. Another way is to examine whether characteristics determined before the reform change discontinuously at the threshold. Explain how CR do these checks and what they find. [Hint: See Appendix Figure A1 of CR, available at `https://www.aeaweb.org/articles/materials/2394`.]

CR show there is no statistically significant discontinuity in the number of births at the threshold. They also examine the percentage of births

that are stillborn. The stillbirth percentage does not exhibit a statistically significant discontinuity at the April 1933 threshold. These checks provide evidence against discontinuous changes in the number and composition of birth cohorts.

(c) The data used in this exercise come from surveys conducted many decades after the individuals were born. Explain why this could generate a sample-selection (i.e., attrition) problem, even if the continuity assumption held in the original population.

We observe only individuals who survived and participated in the surveys. If the additional education induced by the reform affected survival or survey participation, the composition of the observed sample could differ on either side of the threshold. This could generate sample-selection bias even if the original birth cohorts were comparable. The evidence from (b) is based on population data collected in the 1930s, and hence does not rule out this possibility. To investigate this possibility, we can examine whether there are discontinuities in the number and composition of individuals observed in the survey data.

(d) Plot a one-month-bin histogram of respondents' birth cohorts for $-60 \leq r \leq 60$, marking $r = 0$. Does this provide any evidence of differential attrition? Next plot the fraction male by birth cohort over the same range. On the microdata, estimate

$$\texttt{sex}_i = a + \delta Z_i + br_i + cZ_i r_i + v_i$$

using $-46 \leq r \leq 46$. Report the jump and its robust SE. Do the graph or regression estimate provide any evidence of differential attrition?

The histogram exhibits substantial month-to-month variation but no clear break at the cutoff. This is evidence against an abrupt change in the number of observed respondents driven by differential attrition.

The estimated jump in fraction male is 0.00770 (SE 0.01133; $N = 31,385$): 0.77 percentage points, with a 95% CI of approximately $[-1.45, 2.99]$ percentage points. The two-sided p-value is 0.497. Neither the graph nor this regression gives clear evidence of a sex-composition break. This provides additional evidence against the possibility of differential attrition.

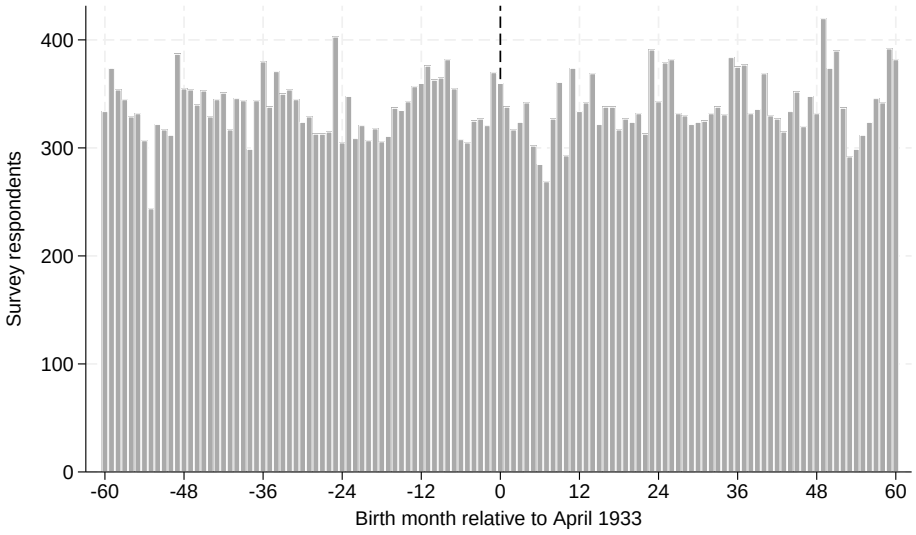

Figure 1: Survey respondents by birth month.

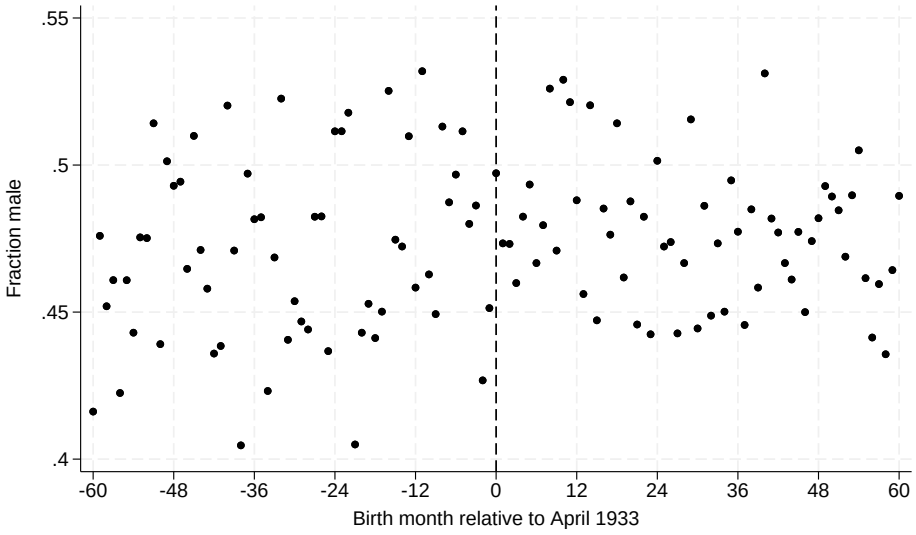

Figure 2: Fraction male by birth month; each marker is a cohort mean.

Q3: First-stage estimates

(a) Plot average age left education by birth month using the full sample, marking the cutoff (i.e., similar to the graph in the top-right of Figure 2 of CR). Describe the pattern you observe. Is there evidence of a discontinuity at the cutoff?

The cohort means show a marked upward shift at April 1933. The jump is smaller than one year because many people would already have stayed in school beyond age 14.

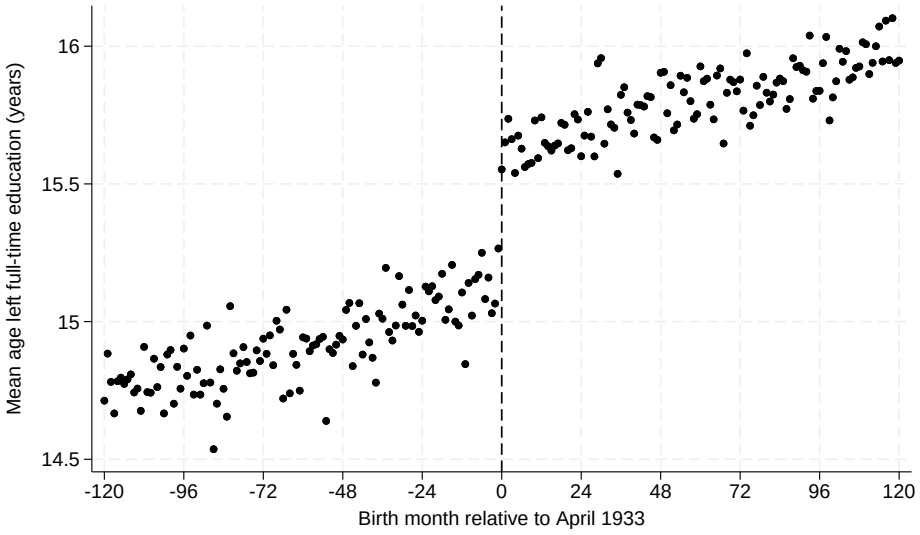

Figure 3: Mean age left education by birth month, in years.

(b) Using the individual data (46-month window and individuals with nonmissing education), regress D on $1, Z, r, Zr$. Interpret the estimated discontinuity in terms of years of education. Repeat this regression after adding sex, interview-year and interview-month indicators, the GHS indicator, calendar-month-of-birth indicators, and a third-order age polynomial. Report the result and check it is close to the estimate reported by CR in Table 2 column 1 (top row).

The unadjusted jump is 0.46366 years (robust SE 0.03422; $N = 31,345$). Exposure raised mean school-leaving age by about 0.46 years. With the listed covariates it is 0.44941 (SE 0.03474), close to the 0.450 estimate reported by CR.

(c) Using the same 46-month sample and covariates as in (b), compare the covariate-adjusted interacted-linear estimate with the following alternatives:

- A common quadratic, D on $1, Z, r, r^2$ and the same covariates
- An interacted quadratic, adding Zr and Zr^2 and retaining the same covariates

Explain the restrictions imposed by each functional form and assess the robustness of the estimated discontinuity.

The common quadratic forces both sides to have the same slope and curvature, allowing only an intercept shift. The interacted quadratic permits different slopes and curvature. The interacted-linear model permits different slopes but sets curvature to zero. The estimated jumps are similar across these different specifications. As expected, the more flexible interacted-quadratic is less precise.

Specification	Discontinuity	Robust SE
Interacted linear	**0.44941**	**0.03474**
Common quadratic	**0.44959**	**0.03468**
Interacted quadratic	**0.44460**	**0.05307**

(d) Summarize the evidence on whether exposure to the reform affected education.

The graph shows a clear upward shift at the cutoff, and both baseline estimates indicate increased education. The adjusted estimates remain similar across the functional forms. Together, this is strong evidence of a positive schooling first stage.

Q4: Reduced-form estimates

In CR, the preferred windows differ by outcome: 79 months for smoking and 46 months for schooling. For this question, use an observation window of 79 months and continue to exclude observations with missing education information.

(a) Graph average ever-smoking rates by birth month using the full sample. Describe the pattern you observe. Is there evidence of a discontinuity at the cutoff?

The cohort means show considerable variation, with no obvious discontinuity at the cutoff.

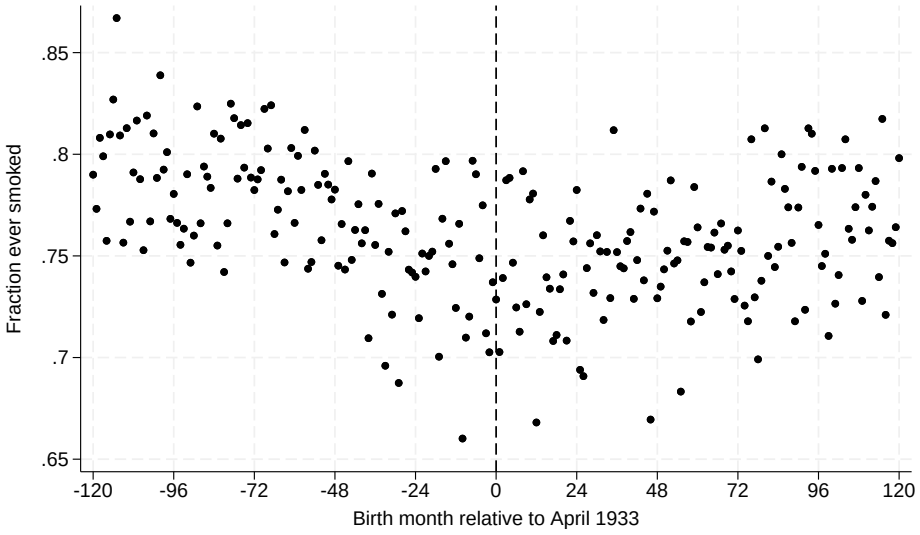

Figure 4: Mean ever smoked by birth month.

(b) Use individual data to estimate the linear-interacted reduced form of Y on $1, Z, r, Zr$. Repeat using the same 79-month sample after adding sex, interview-year and interview-month indicators, GHS, calendar-month-of-birth indicators, and a third-order age polynomial. Check that your covariate-adjusted estimate and standard error are close to those reported by CR (Table 5a, RF column). Interpret both estimated discontinuities and robust SEs. Compare their statistical significance at the 5% significance level.

The unadjusted discontinuity estimate is 0.01408 (SE 0.00889): an increase of 1.41 percentage points, not significant at 5% ($p = 0.113$). With the listed covariates the estimate is 0.01846 (SE 0.00869), close to the CR estimate.

The estimated effect corresponds to an increase of 1.85 percentage points, significant at 5% ($p = 0.034$). Both use $N = 38,677$.

(c) Using the same 79-month sample and covariates as in (b), compare the covariate-adjusted interacted-linear estimate with the following alternatives:

- A common quadratic, Y on $1, Z, r, r^2$ and the same covariates
- An interacted quadratic, adding Zr and Zr^2 and retaining the same covariates

Compare the estimated discontinuities and robust SEs across the three specifications.

I obtain the following estimates:

Specification	Discontinuity	Robust SE
Interacted linear (79 months)	**0.01846**	**0.00869**
Common quadratic (79 months)	**0.01828**	**0.00869**
Interacted quadratic (79 months)	**0.00607**	**0.01315**

The common quadratic changes little. Allowing different curvature yields a smaller estimate of 0.61 percentage points that is not statistically significant. All three specifications use $N = 38,677$.

(d) Summarize the evidence on whether exposure to the reform affected smoking.

The adjusted estimate in (b) suggests that exposure to the reform increased smoking by 1.85 percentage points. The common-quadratic estimate is similar and also statistically significant at 5%. However, the estimate is sensitive to functional form: when we use an interacted quadratic specification, the estimate is smaller and not statistically significant. Thus neither the estimated magnitude nor statistical significance is stable across all three specifications.

Q5: Fuzzy RD and instrumental variables

Throughout this question, use a 46-month window and restrict to observations with nonmissing education and smoking. Use the interacted-linear specification, including r and Zr, and the same covariates as the adjusted specifications in Q3 and Q4: sex, interview-year and interview-month indicators, the GHS indicator, calendar-month-of-birth indicators, and a third-order polynomial in age. Restrict to observations with nonmissing values of all required variables, and use exactly this common sample and the same covariates in the first-stage, reduced-form, and IV regressions.

(a) Estimate the education first stage and the smoking reduced form on this sample. Report the two discontinuities and calculate their ratio.

Both regressions use $N = 22,666$. The first-stage coefficient is 0.46535556 (SE 0.04080132) years and the reduced-form coefficient is 0.01297259 (SE 0.01155880) in probability units. Their ratio is:

$$\frac{0.01297259}{0.46535556} \simeq 0.02787672.$$

These estimates differ from those in Q3 and Q4 because Q5 uses a common estimation sample and a 46-month window for both regressions.

(b) Estimate the model by instrumental variables, instrumenting education with the cutoff indicator and including r, Zr, and the same covariates as in part (a), on the same common sample. Report the estimate, a robust SE and 95% confidence interval. Verify the estimate equals the ratio reported in part (a).

IV gives 0.02787672. As expected, this is equal to the ratio calculated in part (a). Its robust SE is 0.02501315, with a 95% CI of $[-0.02115, 0.07690]$. Thus one additional year of education is estimated to increase ever smoking by 2.79 percentage points, with a CI of roughly $[-2.11, 7.69]$ percentage points. The estimate is not statistically distinguishable from zero.

(c) Compare your IV estimate with the OLS estimates from Q1. What do you conclude about the effect of education on smoking?

The OLS estimates from Q1 suggest that additional education is associated with a lower probability of ever smoking, whereas the IV estimate is positive but statistically insignificant. Under the maintained common

linear causal model and RD identification assumptions, IV estimates the causal effect of an additional year of education. However, it is much less precise than OLS, and its confidence interval includes both negative and positive effects. The estimates also use different samples: Q1 restricts attention to individuals born before the reform who left school at ages 14 or 15, whereas the IV estimate uses individuals on both sides of the cutoff with a broader range of educational attainment. OLS also uses individual schooling variation, whereas IV uses variation predicted by cutoff exposure after adjusting for the included controls. Consequently, the estimates are not directly comparable.

References

Clark, D. and Royer, H. (2013). The effect of education on adult mortality and health: Evidence from Britain. *American Economic Review*, 103(6):2087–2120.