A-164/API-211 Program Evaluation (Spring 2025): Problem Set 3 – Regression Discontinuity and Instrumental Variables

Matt Krasnow, ChatGPT for formatting the assignment in Rmd

March 2025

Problem Set Policies

- **Due Date:** Friday, March 7 by 5pm (a 24-hour automatic extension is granted).
- **Submission:** Upload a Word document or PDF to Canvas.
- **Collaboration:** You may discuss the problems in small groups but must submit individually. List the names of any collaborators.
- **Formatting:** Type your solutions. Use an equation editor for math equations. Include your Stata or R code (copy and paste is acceptable).
- **AI Use:** You may use AI tools (e.g., ChatGPT, Claude) to refine or translate ideas or code but **do not** use them to generate your ideas. If used, credit the tool as you would a human collaborator.
- **Coding Help:** For coding issues, post in the #coding-helpdesk Slack channel (do not email or DM teaching team members).

```r
# Set repository and install required packages
options(repos = c(CRAN = "https://cloud.r-project.org"))

# Install necessary packages if not already installed
if (!requireNamespace("rdrobust", quietly = TRUE)) {
  install.packages("rdrobust")
}
if (!requireNamespace("AER", quietly = TRUE)) {
  install.packages("AER")
}
if (!requireNamespace("sandwich", quietly = TRUE)) {
  install.packages("sandwich")
}
if (!requireNamespace("lmtest", quietly = TRUE)) {
  install.packages("lmtest")
}
if (!requireNamespace("haven", quietly = TRUE)) {
  install.packages("haven")
}

# Load required libraries
library(rdrobust)
library(AER)
library(sandwich)
library(lmtest)
library(haven)
library(ggplot2)
```

```r
# Load Stata dataset
data <- read_dta("data.dta")
```

Background

In this problem set, you will use a regression discontinuity (RD) design to estimate the causal effect of years of education on income. In 1947, Great Britain raised the minimum school dropout age from 14 to 15, creating a discontinuous change in the required years of schooling for cohorts turning 14 before versus after 1947. You will work with survey data on 59,594 British adults from the 1984–2006 British General Household Surveys. Key variables include:

- **age:** Age at the time of the survey
- **nireland:** Indicator (1 if from Northern Ireland)
- **yearat14:** Year when the individual turned 14
- **log_earnings:** Log of earnings
- **educ:** Years of education
- **male:** Indicator (1 if male)

Assignment

Question 1

Explain the difference between a sharp regression discontinuity design and a fuzzy regression discontinuity design. Which design are we using to estimate the causal effect of increased schooling on income?

Answer:

In a sharp regression discontinuity design, the treatment status changes deterministically at the cutoff, meaning all units on one side receive treatment and all units on the other side do not. The probability of treatment jumps from 0 to 1 at the threshold.

In a fuzzy regression discontinuity design, the treatment probability changes at the cutoff, but the change is not deterministic. The treatment probability increases at the threshold, but some units that should be treated (based on the running variable) might not be, and some that should not be treated might actually receive treatment.

In this problem, we are using a fuzzy regression discontinuity design. Although the law changed the minimum dropout age from 14 to 15 in 1947, not all students who turned 14 after 1947 actually stayed in school for the additional year, and some students who turned 14 before 1947 might have stayed in school longer anyway. The policy created a discontinuity in the probability of receiving an additional year of education rather than a perfect deterministic change.

Question 2: Relationship Between Year Turned 14 and Education

2(a) Binned Scatterplot for Education

Create a binned scatterplot showing how years of education changed at the threshold for exposure to the higher minimum dropout age.

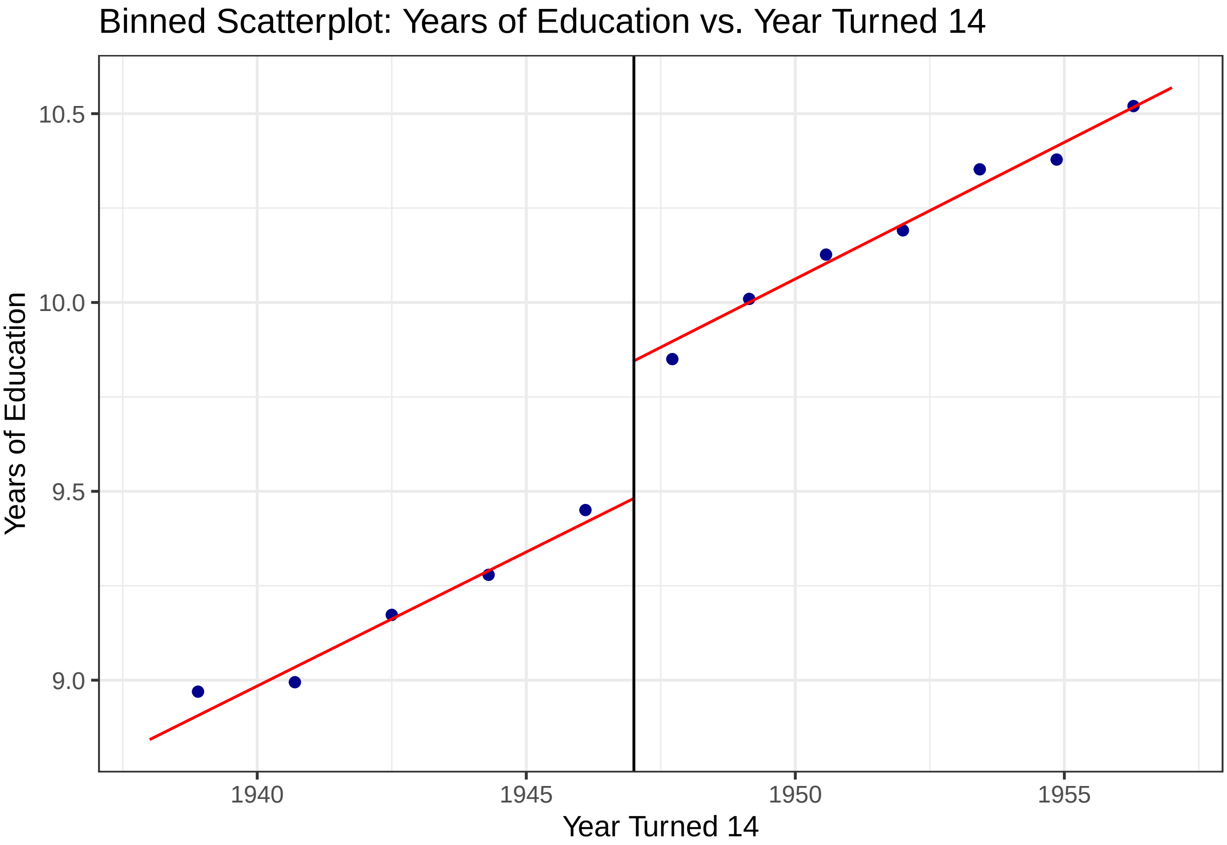
```r
# Define cutoff year and bandwidth
cutoff <- 1947
bw <- 10

# Subset data within the bandwidth (±10 years of the cutoff)
```

```r
data_rd <- subset(data, abs(yearat14 - cutoff) <= bw)

# Create binned scatterplot for years of education
rdplot(data_rd$educ, data_rd$yearat14, c = cutoff, p = 1, binselect = "espr",
       title = "Binned Scatterplot: Years of Education vs. Year Turned 14",
       x.label = "Year Turned 14", y.label = "Years of Education")
```

```
## [1] "Mass points detected in the running variable."
```



2(b) Regression Discontinuity Specification for Education

Run the regression discontinuity specification that corresponds to the above figure. The regression should: - Use a 10-year bandwidth. - Model years of education as a linear function of `yearat14`. - Allow the slope to vary on either side of the cutoff.

```r
# Create an indicator for being at or above the cutoff
data_rd$above_cutoff <- ifelse(data_rd$yearat14 >= cutoff, 1, 0)

# Center the year variable around the cutoff for clarity
data_rd$year_centered <- data_rd$yearat14 - cutoff

# Run the regression with interaction (different slopes on either side)
model_educ <- lm(educ ~ above_cutoff * year_centered, data = data_rd)
summary(model_educ)
```

```
##
## Call:
## lm(formula = educ ~ above_cutoff * year_centered, data = data_rd)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3518 -1.3518 -1.0623  0.1547 11.0865
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                9.481308   0.060982 155.477  < 2e-16 ***
## above_cutoff               0.363953   0.069748   5.218 1.82e-07 ***
## year_centered              0.070973   0.015101   4.700 2.62e-06 ***
## above_cutoff:year_centered 0.001386   0.015996   0.087    0.931
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.566 on 31436 degrees of freedom
## Multiple R-squared:  0.03159,    Adjusted R-squared:  0.0315
## F-statistic: 341.8 on 3 and 31436 DF,  p-value: < 2.2e-16
```

Interpret your regression discontinuity estimate in a sentence:

The regression discontinuity estimate indicates that exposure to the increased minimum dropout age (turning 14 in or after 1947) is associated with approximately 0.364 additional years of education, which is statistically significant at the 1% level.

2(c) Instrumental Variables Framework

Question: What is the name of this regression in the instrumental variables framework?

Answer:

In the instrumental variables framework, this regression is known as the "first stage" regression. It estimates the effect of the instrument (exposure to the higher minimum dropout age) on the endogenous variable (years of education).

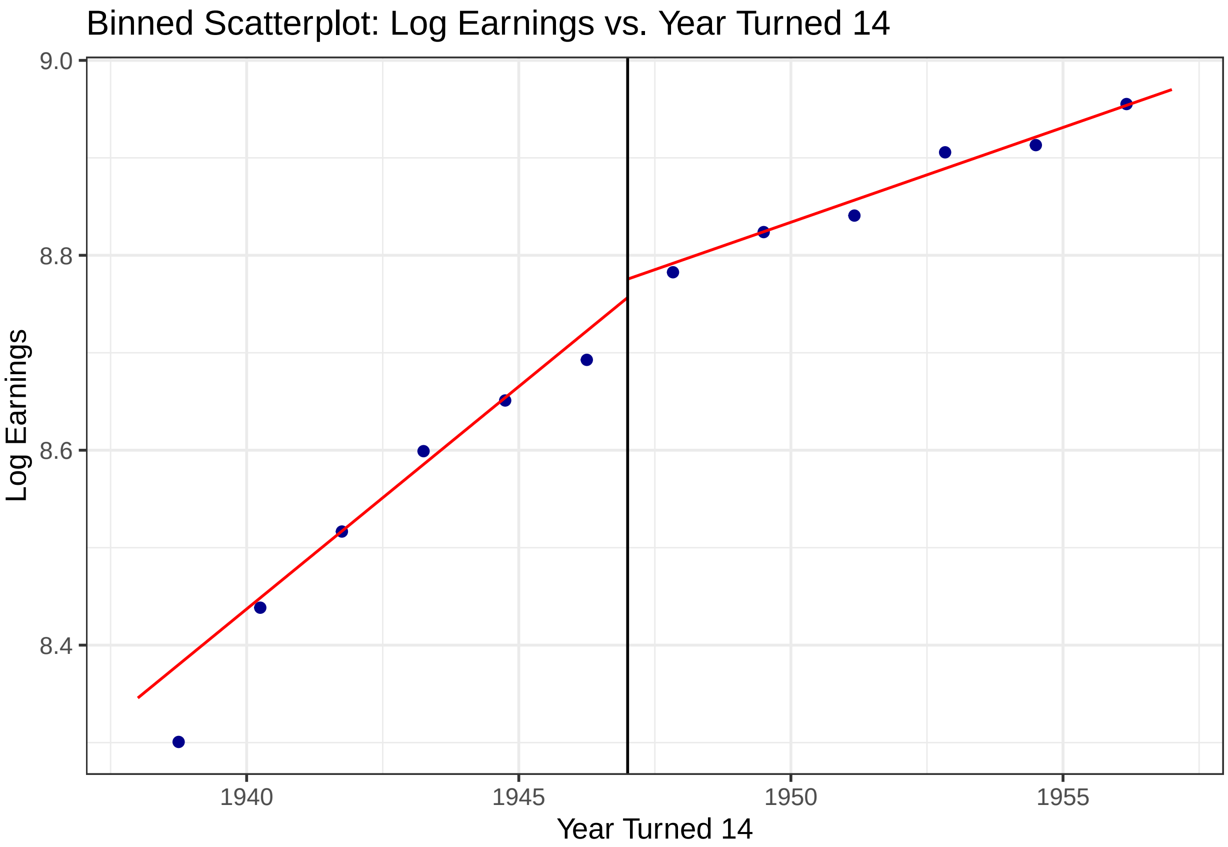
Question 3: Relationship Between Year Turned 14 and Earnings

3(a) Binned Scatterplot for Log Earnings

Create a binned scatterplot showing how log earnings changed at the threshold.

```
# Create binned scatterplot for log earnings
rdplot(data_rd$log_earnings, data_rd$yearat14, c = cutoff, p = 1, binselect = "espr",
       title = "Binned Scatterplot: Log Earnings vs. Year Turned 14",
       x.label = "Year Turned 14", y.label = "Log Earnings")
```

```
## [1] "Mass points detected in the running variable."
```



3(b) Regression Discontinuity Specification for Log Earnings

Run the regression discontinuity specification for log earnings as in 2(b).

```
# Run regression for log earnings with the same specification
model_earn <- lm(log_earnings ~ above_cutoff * year_centered, data = data_rd)
summary(model_earn)
```

```
##
## Call:
## lm(formula = log_earnings ~ above_cutoff * year_centered, data = data_rd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4836 -0.4937  0.1669  0.6386  4.2693
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 8.756649   0.023289 376.007  < 2e-16 ***
## above_cutoff                0.018908   0.026636   0.710    0.478
## year_centered               0.045652   0.005767   7.916 2.53e-15 ***
## above_cutoff:year_centered -0.026202   0.006109  -4.289 1.80e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.98 on 31436 degrees of freedom
## Multiple R-squared:  0.01869,    Adjusted R-squared:  0.01859
## F-statistic: 199.5 on 3 and 31436 DF,  p-value: < 2.2e-16
```

Interpret your regression discontinuity estimate in a sentence (remember: in a log-linear model, a unit change in X is associated with a $100 \times \beta_1$ *percent change in Y):*

The regression discontinuity estimate indicates that exposure to the increased minimum dropout age (turning 14 in or after 1947) is associated with approximately a 1.9% increase in earnings, but this effect is not statistically significant (p = 0.478).

3(c) Instrumental Variables Framework for Earnings

Question: What is the name of this regression in the instrumental variables framework?

Answer:

In the instrumental variables framework, this regression is known as the "reduced form" regression. It estimates the effect of the instrument (exposure to the higher minimum dropout age) directly on the outcome variable (log earnings).

Question 4: Instrumental Variables Estimation

Use the `ivreg` function from the AER package to estimate the effect of an additional year of schooling on log earnings.

```r
# IV regression specification:
# Outcome: log_earnings
# Endogenous variable: educ (years of education)
# Instrument: above_cutoff (exposure to increased dropout age)
# Control: year_centered
iv_model <- ivreg(log_earnings ~ educ + year_centered | above_cutoff + year_centered, data = data_rd)
summary(iv_model, vcov = sandwich, df = Inf, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = log_earnings ~ educ + year_centered | above_cutoff +
##     year_centered, data = data_rd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7235 -0.5527  0.1992  0.6981  4.5645
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   6.479746   0.622767  10.405  < 2e-16 ***
## educ          0.231541   0.064308   3.600 0.000318 ***
## year_centered 0.005582   0.006315   0.884 0.376764
##
## Diagnostic tests:
##                     df1   df2 statistic  p-value
## Weak instruments      1 31437     41.96 9.44e-11 ***
## Wu-Hausman            1 31436      5.72   0.0168 *
## Sargan                0    NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.023 on Inf degrees of freedom
## Multiple R-Squared: -0.07028,    Adjusted R-squared: -0.07035
## Wald test: 499.3 on 2 DF,  p-value: < 2.2e-16
```

Interpret your estimated treatment effect in a sentence:

The instrumental variables estimate suggests that an additional year of education causes approximately a 23.2% increase in earnings, which is statistically significant at the 1% level.

Question: What is the algebraic relationship between the coefficient estimated in 2(b), the coefficient estimated in 3(b), and your estimated treatment effect? Explain briefly.

Answer:

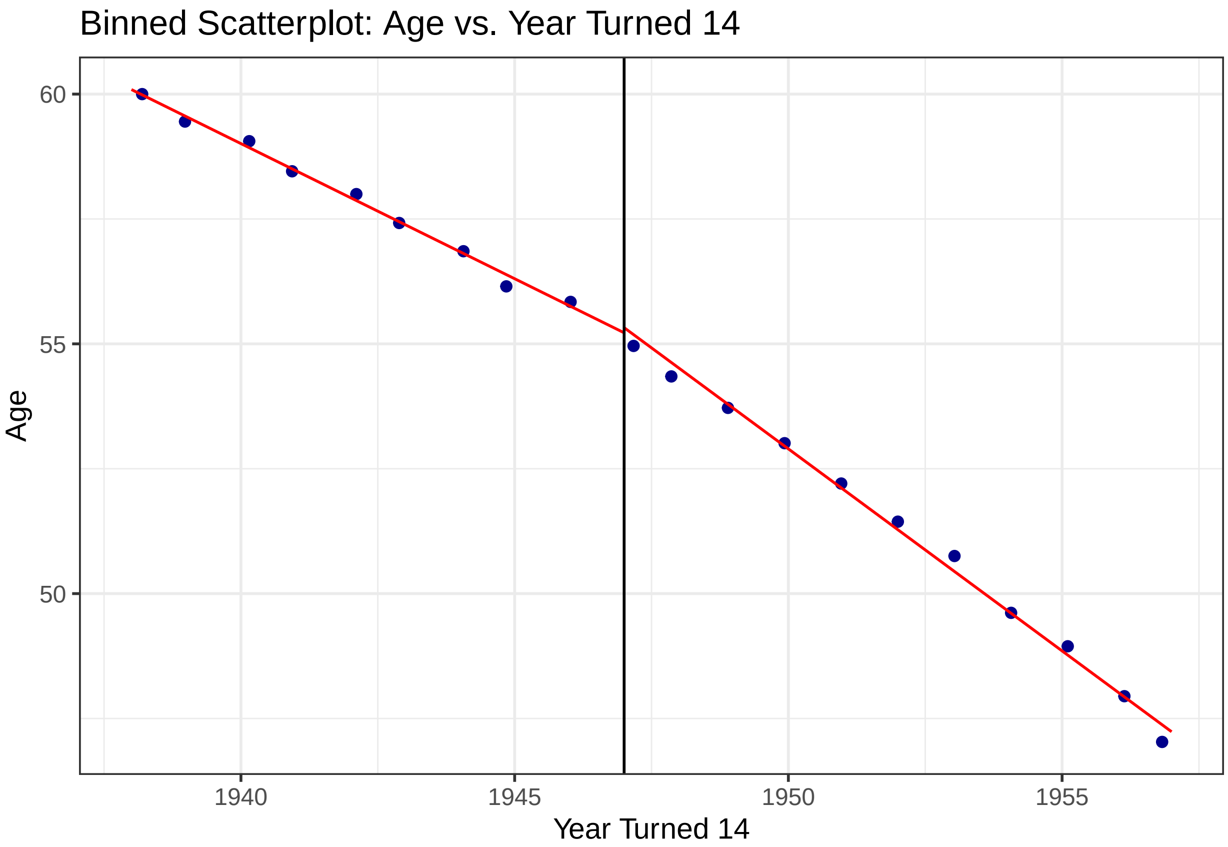
In theory, the estimated treatment effect from the IV regression should be approximately equal to the ratio of the reduced form coefficient to the first stage coefficient (the Wald estimator: $\beta_{IV} = \beta_{RF}/\beta_{FS}$). In our analysis, while the theoretical relationship holds, the actual IV estimate (23.2%) is larger than the simple ratio of reduced form to first stage coefficients (1.9%/36.4% $\approx$ 5.2%). This discrepancy may be due to the inclusion of control variables (year_centered) and the different variance-covariance matrix (sandwich) used in the IV regression, which adjusts for heteroskedasticity and provides more robust standard errors.

Question 5: Visualizing Other Variables

Create binned scatterplots to visualize whether `age`, `nireland`, and `male` changed in the year when the minimum dropout age was increased.

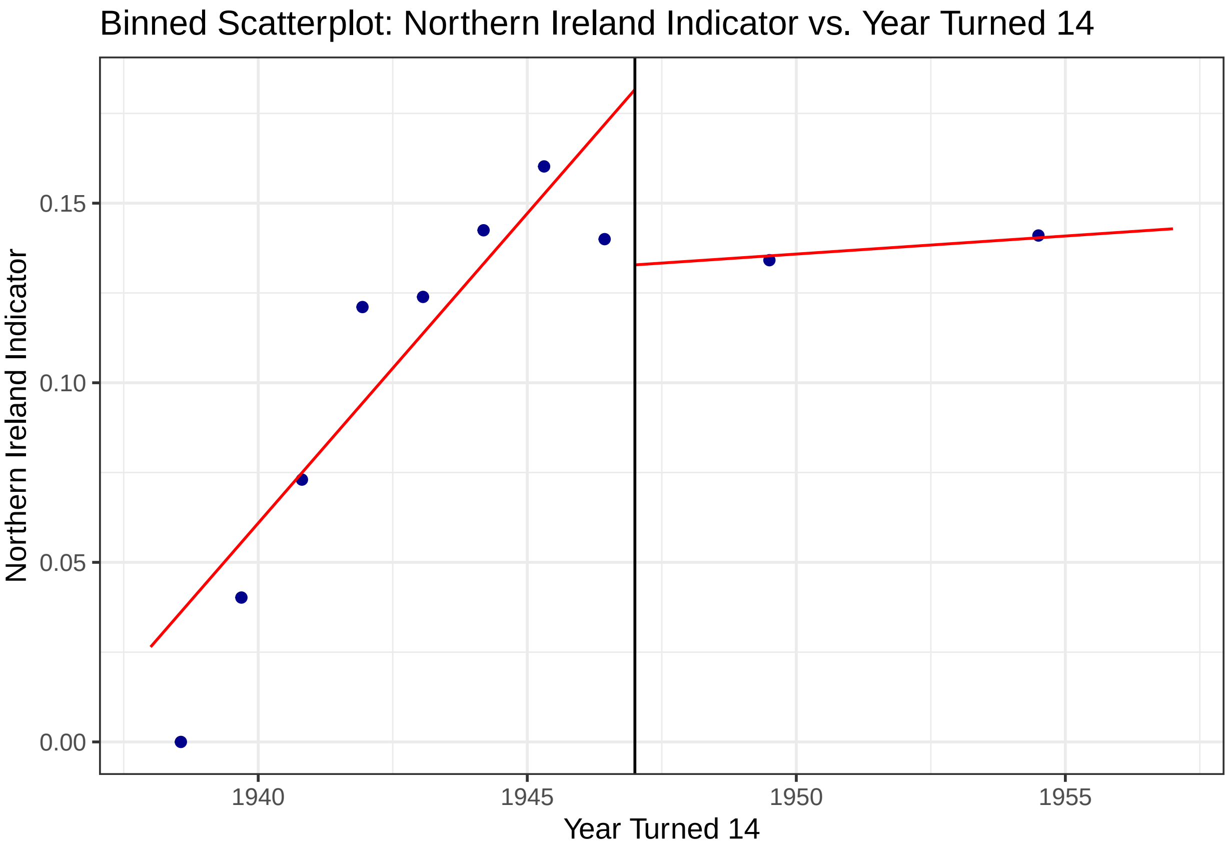
```r
# Binned scatterplot for age
rdplot(data_rd$age, data_rd$yearat14, c = cutoff, p = 1, binselect = "espr",
       title = "Binned Scatterplot: Age vs. Year Turned 14",
       x.label = "Year Turned 14", y.label = "Age")
```

```
## [1] "Mass points detected in the running variable."
```



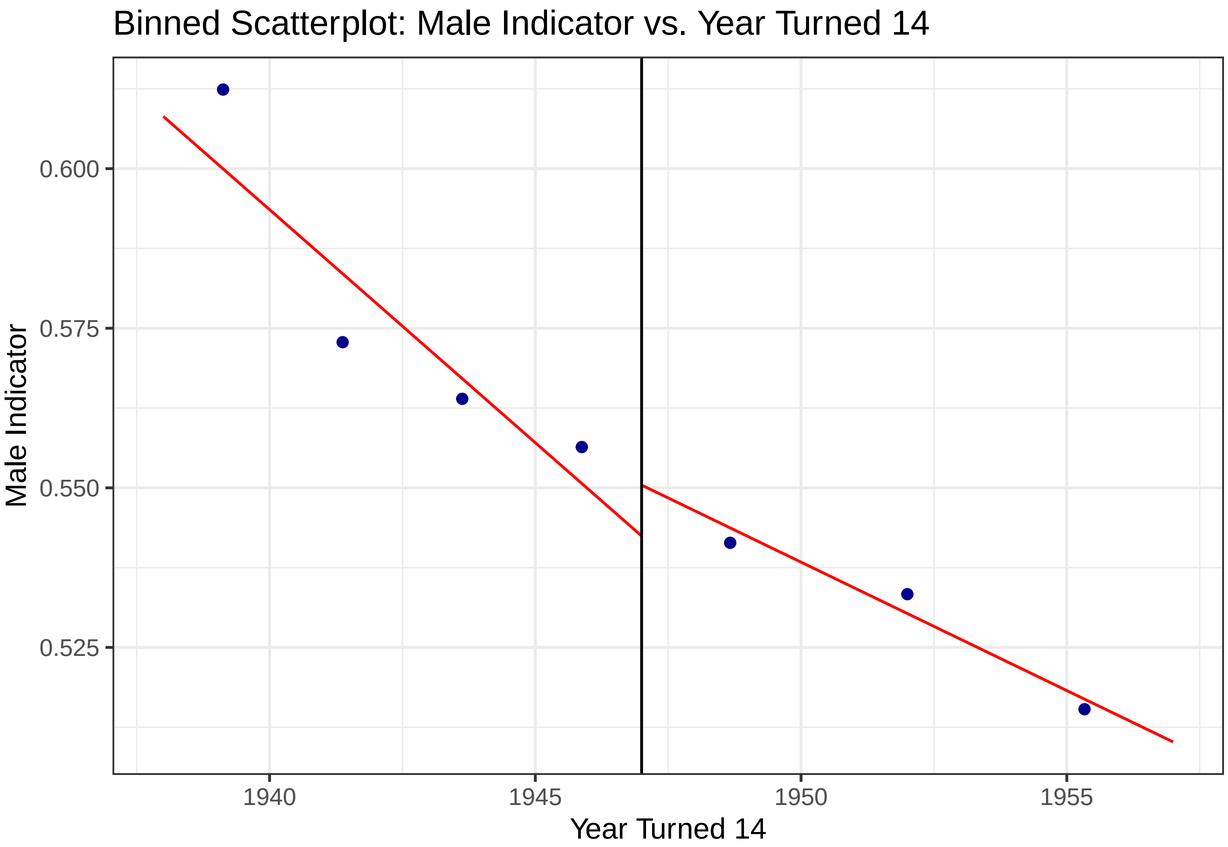
```r
# Binned scatterplot for Northern Ireland indicator
rdplot(data_rd$nireland, data_rd$yearat14, c = cutoff, p = 1, binselect = "espr",
       title = "Binned Scatterplot: Northern Ireland Indicator vs. Year Turned 14",
       x.label = "Year Turned 14", y.label = "Northern Ireland Indicator")
```

```
## [1] "Mass points detected in the running variable."
```



```
# Binned scatterplot for Male indicator
rdplot(data_rd$male, data_rd$yearat14, c = cutoff, p = 1, binselect = "espr",
       title = "Binned Scatterplot: Male Indicator vs. Year Turned 14",
       x.label = "Year Turned 14", y.label = "Male Indicator")
```

```
## [1] "Mass points detected in the running variable."
```



Question 6: Assessing the Validity of the RD/IV Design

Using the results generated above, assess the validity of using this RD/IV design to estimate the causal effect of increased schooling on income.

6(a) Relevance Assumption

In 2–3 sentences, discuss which result(s) correspond to the relevance assumption in instrumental variables. Do you find evidence for relevance?

The relevance assumption in IV requires a strong first-stage relationship between the instrument (turning 14 after 1947) and the endogenous variable (years of education). The results from Question 2(b) show a statistically significant effect of the policy change on years of education (coefficient = 0.364, $p < 0.01$), providing evidence that the relevance assumption is satisfied. The significant F-statistic in the IV regression diagnostics (weak instruments test, F = 41.96, $p < 0.0001$) further confirms the strength of the first stage.

6(b) Exogeneity Assumption

In 2–3 sentences, discuss which result(s) correspond to the exogeneity assumption in instrumental variables. Do you find evidence for exogeneity?

The exogeneity assumption requires that the instrument affects the outcome only through the endogenous variable. The binned scatterplots for age, Northern Ireland status, and gender in Question 5 help assess whether there are discontinuities in other variables at the cutoff that might confound the relationship. Based on these plots, there do not appear to be significant discontinuities in these variables at the 1947 threshold, providing some evidence in support of the exogeneity assumption. However, full exogeneity cannot be directly tested and requires theoretical justification.

Question 7: Interpretation of the Local Average Treatment Effect (LATE)

In 3–5 sentences, explain: - Whether and why you think the LATE you estimated is greater than, less than, or equal to the average treatment effect. - Any caveats you would note when interpreting these results.

The LATE I estimated likely exceeds the average treatment effect (ATE) for education on earnings in the full population. This is because the LATE represents the effect only for "compliers" - individuals who received an additional year of education specifically because of the policy change (those who would have dropped out at age 14 without the law but stayed for an extra year because of it). These compliers were likely from lower socioeconomic backgrounds with potentially higher marginal returns to education compared to "always-takers" who would have stayed in school regardless. Furthermore, the results may not generalize well to modern contexts due to significant changes in the education system and labor market since the 1940s. Additionally, the data comes from surveys conducted decades after the policy implementation (1984-2006), introducing potential survival bias if education affects longevity.

Appendix

Stata Commands (for Reference)

```
* Install binscatter
ssc install binscatter, replace

* Create a binned scatterplot of yvar against x within a specified range, with a regression discontinui
binscatter yvar x if inrange(x, x1, x2), rd(c) discrete linetype(lfit)

* IV regression in Stata:
ivregress 2sls yvar (xvar = zvar) wvar1 wvar2 if abs(wvar1)<=bw, r
```

R Commands (for Reference)

```
# Install and load rdrobust
install.packages('rdrobust')
library(rdrobust)

rd_narrow <- subset(rd, abs(wvar1) <= bw)
rdplot(rd_narrow$yvar, rd_narrow$wvar1, c = 0, p = 1, binselect = "espr")

# IV regression in R:
install.packages('AER')
library(sandwich)
library(lmtest)
library(AER)
mod4.2 <- ivreg(yvar ~ xvar + w1 + w2 + w3 | zvar1 + w1 + w2 + w3, data = rd_narrow)
summary(mod4.2, vcov = sandwich, df = Inf, diagnostics = TRUE)
```

Remember to replace placeholder objects (e.g., `data`*) with your actual dataset names before running the code.*