Problem Set 1

Tate Mason

2025-09-14

Problem 1: Conceptual Problem

Part 1: What is the effect of whether a child watches TV on their math skill?

A difference in differences approach could be used in this case. We could run an experiment where we randomly assign a group of children to watch a certain amount of TV each day, while another group of children do not watch any TV. We would then measure the math skills of both groups before and after the experiment. The difference in the change in math skills between the two groups would give us an estimate of the effect of watching TV on math skills. Problems could arise from children having differing levels of parental involvement, access to educational resources (tutors), or differing levels of innate ability.

Part 2: Is working around retirement age good for the person's health?

An RD approach could be used in this case. We set the discontinuity at the age of retirement, comparing health outcomes for individuals just below and just above retirement age. There is no reason to believe that individuals just below and just above retrirement age would be systematically different in terms of health, other than the fact that one group is working and the other is not. Problems arise from things like pre-existing conditions, potentially leading to early retirement, differing levels of exposure to virus/danger, as well as differences in access to healthcare.

Part 3: Is the racial wage gap in part due to discrimination against racial minorities?

- Note: Consider only two groups: racial minorities vs. racial majority.

For this question, implementation of an IV model would be appropriate. Using an instrument that affects the likelihood of being a racial minority but does not directly affect wages (e.g., historical segregation policies or geographic location) could help isolate the effect of being a minority on wages. Problems arise from the difficulty in finding a valid instrument that satisfies the relevance and exclusion restriction criteria, as well as potential confounding factors that may still influence wages. There is also the option to use a proxy variable. For instance, using something like housing outcomes or neighborhood characteristics as a proxy would allow estimation of effects across races without directly biasing the results.

Part 4: What is the effect of whether the mother receives welfare money support while the child is young on the child's future income (by age 40)?

Problem 2: Coding

Part 1: Creating dataset

1. Set random seed to 1
2. N = 10000
3. Draw ϵ_i^D $/perp$ $\epsilon_i^Y \sim N(0,1)$
4. Draw $U_i \sim N(0, 0.5)$ (Note: s.d. is 0.5)
5. Create $Z_i = \mathbb{1}(z_i > 0.5)$ where z_i is randomly drawn from a uniform distribution on $[0,1]$
6. Create $D_i = \mathbb{1}(\alpha_0 + \alpha_Z Z_i + \alpha_U U_i + \epsilon_i^D > 0)$ such that $\alpha_0 = -4, \alpha_Z = 5, \alpha_U = 4$
7. Create $Y_i = \beta_0 + \beta_D D_i + \beta_Z Z_i + \beta_U U_i + \epsilon_i^Y$ such that $\beta_0 = 3, \beta_D = 2, \beta_Z = 0, \beta_U = 6$

- $\beta_Z Z_i + \beta_U U_i + \epsilon_i^Y = \epsilon_i$

```
df <- data.frame(
  id = 1:10000,
  epsilon_D = rnorm(10000, mean = 0, sd = 1),
  epsilon_Y = rnorm(10000, mean = 0, sd = 1),
  U_i = rnorm(10000, mean = 0, sd = 0.5)
)

z_i <- runif(10000, min = 0, max = 1)
df <- df %>%
  mutate(
    Z_i = as.numeric(z_i > 0.5),
    D_i = as.numeric(-4 + 5 * Z_i + 4 * U_i + epsilon_D > 0),
    Y_i = 3 + 2 * D_i + 0 * Z_i + 6 * U_i + epsilon_Y
  )
```

Part 2: Estimating the effect of D on Y with OLS

```r
OLS <- lm(Y_i ~ D_i, data = df)
summary(OLS)
```

```
Call:
lm(formula = Y_i ~ D_i, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.2473  -2.0084  -0.1544   1.9659   9.2807 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.02336    0.03642   55.56   <2e-16 ***
D_i          4.72039    0.06021   78.40   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.9 on 9998 degrees of freedom
Multiple R-squared:  0.3807,    Adjusted R-squared:  0.3807 
F-statistic:  6146 on 1 and 9998 DF,  p-value: < 2.2e-16
```

Part 3: Estimating the effect of D on Y with IV

```r
IV <- ivreg(Y_i ~ D_i | Z_i, data = df)
summary(IV)
```

```
Call:
ivreg(formula = Y_i ~ D_i | Z_i, data = df)

Residuals:
      Min        1Q    Median        3Q       Max 
-11.14794  -2.10898   0.01567   2.11343  10.84223 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  2.92404    0.04738   61.71   <2e-16 ***
D_i          2.25817    0.09716   23.24   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.133 on 9998 degrees of freedom
Multiple R-Squared: 0.2771, Adjusted R-squared: 0.2771
Wald test: 540.1 on 1 and 9998 DF,  p-value: < 2.2e-16
```

Part 4: OLS of D on Z (with a constant)

- yields: $\hat{D} = (L(D|Z))$ and $\tilde{D} = D - \hat{D}$

  a) Regress Y on $\hat{D}$ (with a constant)

  b) Regress Y on D and $\tilde{D}$ (with a constant)

- Explain why the coefficient on $\hat{D}$ in a) is the same as the coefficient on D in b). Explain why both are also the same as the IV estimate from Part 3. What is the intuition behind the coefficient on $\tilde{D}$ in b)? Optional: explain the relationship between the standard errors of the estimates in a), b), and Part 3.

a)

```r
df <- df %>%
  mutate(
    D_hat = predict(lm(D_i ~ Z_i, data = df)),
    D_tilde = D_i - D_hat
  )
model_a <- lm(Y_i ~ D_hat, data = df)
summary(model_a)
```

```
Call:
lm(formula = Y_i ~ D_hat, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-11.614  -2.516   0.199   2.388  12.973

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.92404    0.05463   53.52   <2e-16 ***
D_hat        2.25817    0.11203   20.16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.612 on 9998 degrees of freedom
Multiple R-squared:  0.03905,   Adjusted R-squared:  0.03896 
F-statistic: 406.3 on 1 and 9998 DF,  p-value: < 2.2e-16
```

b)

```r
model_b <- lm(Y_i ~ D_i + D_tilde, data = df)
summary(model_b)
```

```
Call:
lm(formula = Y_i ~ D_i + D_tilde, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.9638  -1.7894   0.0297   1.8385   9.4420 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.92404    0.04077   71.72   <2e-16 ***
D_i          2.25817    0.08360   27.01   <2e-16 ***
D_tilde      4.46237    0.11255   39.65   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.696 on 9997 degrees of freedom
Multiple R-squared:  0.4649,    Adjusted R-squared:  0.4648 
F-statistic:  4342 on 2 and 9997 DF,  p-value: < 2.2e-16
```

$\beta_{\hat{D}}$ in model a) is the same as β_D in model b) because both models are essentially capturing the same variation in D that is explained by Z. This is the same as the IV estimate from **3** because the IV method captures the variation in D that is correlated with Z. The coefficient on $\tilde{D}$ in model b) captures the variation in D that is not explained by Z, which is unrelated to the instrument and thus does not contribute to the estimation of the causal effect of D on Y.

Part 5: Change DGP s.t. $\beta_Z = 1$, redo 3. Then change DGP s.t. $\beta_Z = -1$, redo 3.

- Explain why the IV estimates of β_D are biased in these two cases, and why the bias changes sign when β_Z changes sign. Optional: explain the intution for why the bias of the estimator of the coefficient on D changes sign when $\beta_Z = \{0, 1\}$.

Part 6: