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 is applicable here. 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 factors like pre-existing conditions or 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.

Using a model which employs an instrumental variable would likely be the best course. The goal would be to find an instrument which is simultaneously affiliated with race, but uncorrelated with wage. The difficulty arises in this search. For instance, using something like neighborhood as an IV is surely correlated with race, given historical policies surrounding how cities were segregated, but there is also likely correlation with wage.

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)?

There are two approaches which would be equally effective, though both run into similar issues. First, a diff-in-diff approach could be used. This would allow for simple interpretation of the effect of income outcomes for those whose families were recipients of welfare benefits vs. those who were not. However, difficulty arises in data access/integrity issues. It is difficult to maintain good data over 40 years for a good sample, and is also hard to be granted access to such data if it exists. An RD setup would produce similar ease of interpretation. To implement this type of approach, set the cutoff line at the income threshold for receiving welfare benefits, then examine the difference in outcomes for those just above and below the line. Data access would be similarly difficult here. Across both setups, an issue also arises in that income is not a fixe state. For instance, there may be years where the mother is on welfare, but times where she is not. Thus, the children who spent, say, a couple of months on welfare vs. the kids who spent their entire adolescence in that state should have systematically different outcomes.

Problem 2: Coding

Part 1: Creating dataset

1. Set random seed to 1
2. N = 10000
3. Draw $\epsilon_i^D \perp\!\!\!\perp \epsilon_i^Y \sim N(0,1)$
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)
```

```r
)

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
  )
```

```r
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 which 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 $\beta_Z = \{0, 1\}$.

```r
df <- df %>%
  mutate(
    Y_i = 3 + 2 * D_i + 1 * Z_i + 6 * U_i + epsilon_Y
)
IV5a <- ivreg(Y_i ~ D_i | Z_i, data = df)
summary(IV5a)
```

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

Residuals:
      Min        1Q    Median        3Q       Max
-11.08396  -1.92037  -0.00676   1.95614  10.35569

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.86006    0.04371   65.44   <2e-16 ***
D_i          3.80869    0.08963   42.49   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.89 on 9998 degrees of freedom
Multiple R-Squared: 0.4263, Adjusted R-squared: 0.4262 
Wald test:  1806 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
df <- df %>%
  mutate(
    Y_i = 3 + 2 * D_i + -1 * Z_i + 6 * U_i + epsilon_Y
)
IV5b <- ivreg(Y_i ~ D_i | Z_i, data = df)
summary(IV5b)
```

```

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

Residuals:
     Min       1Q   Median       3Q      Max 
-11.2119  -2.3955   0.1247   2.2961  12.2946 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.98802    0.05216  57.287  < 2e-16 ***
D_i          0.70764    0.10696   6.616 3.88e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.449 on 9998 degrees of freedom
Multiple R-Squared: 0.09203,    Adjusted R-squared: 0.09194 
Wald test: 43.77 on 1 and 9998 DF,  p-value: 3.881e-11
```

Part 6:

```r
df <- df %>%
  mutate(
    D_i = as.numeric(-4 + 10 * Z_i + 4 * U_i + epsilon_D > 0),
    Y_i = 3 + 2 * D_i + 1 * Z_i + 6 * U_i + epsilon_Y
)
```

```r
IV6a <- ivreg(Y_i ~ D_i | Z_i, data = df)
summary(IV6a)
```

```

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

Residuals:
      Min        1Q    Median        3Q       Max 
-11.10825  -2.14575   0.05882   2.16869  10.92009 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.88435    0.04642   62.13   <2e-16 ***
D_i          3.22000    0.06556   49.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.134 on 9998 degrees of freedom
Multiple R-Squared: 0.2509, Adjusted R-squared: 0.2508 
Wald test:  2412 on 1 and 9998 DF,  p-value: < 2.2e-16 
```

```r
df <- df %>%
  mutate(
    D_i = as.numeric(-4 + 10 * Z_i + 4 * U_i + epsilon_D > 0),
    Y_i = 3 + 2 * D_i + -1 * Z_i + 6 * U_i + epsilon_Y
  )
IV6b <- ivreg(Y_i ~ D_i | Z_i, data = df)
summary(IV6b)
```

```

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

Residuals:
      Min        1Q    Median        3Q       Max 
-11.19456  -2.17591   0.02275   2.14348  11.89132 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.97066    0.04770   62.28   <2e-16 ***
```

```
D_i           1.12828    0.06736   16.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.22 on 9998 degrees of freedom
Multiple R-Squared: 0.06005,    Adjusted R-squared: 0.05996
Wald test: 280.5 on 1 and 9998 DF,  p-value: < 2.2e-16
```

Part 7:

D_i by formation cannot deviate from compliance. When $Z_i = 0$, The sum of $\alpha_0, \alpha_U, \epsilon_i^D$ is negative as $\alpha_0 = -4$ while $\alpha_U * U + \epsilon_i^D$ is not greater than 4, thus implying $D_i = 0$ when $Z_i = 0$.

Part 8:

```
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,
    pZ = length(Z_i == 1)/length(id),
    pD = length(D_i == 1)/length(id),
    compliers = (pZ*D_hat)/pD
  )
length(df$compliers)
```

```
[1] 10000
```

As shown in the last part, there can be no always-takers or never-takers as people will not defy treatment D_i

Part 9:

```
df <- df %>%
  mutate(
    Z_i = as.numeric(z_i > 0.5),
    D_i = as.numeric(-4 + 10 * Z_i + 4 * U_i + epsilon_D > 0),
    Y_i = 3 + 2 * D_i + 0 * Z_i + 6 * U_i + epsilon_Y,
    pZ = length(Z_i == 1)/length(id),
    pD = length(D_i == 1)/length(id),
    compliers = (pZ*D_hat)/pD
  )
length(df$compliers)
```

```
[1] 10000
```

The proportion remains the same as the coefficient on Z_i does not effect anything when $Z_i = 0$. Thus, all units will comply.