Problem Set 1

Tate Mason

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)
)

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

```
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

```
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 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
)
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 have defiers. 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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 1),
    pC = 1 - pAT - pNT
  )

summarize(df, pC = mean(pC), pNT = mean(pNT), pAT = mean(pAT))
```

```
      pC    pNT    pAT
1 0.8216 0.1579 0.0205
```

```
cor(df$D_i, df$Z_i)
```

```
[1] 0.6694964
```

Majority are compliers, with about 82% compliers and a correlation of about 0.67 between D and Z.

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),
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 1),
    pC = 1 - pAT - pNT
  )

summarize(df, pC = mean(pC), pNT = mean(pNT), pAT = mean(pAT))
```

```
      pC    pNT    pAT
1 0.9782 0.0013 0.0205
```

```
cor(df$D_i, df$Z_i)
```

```
[1] 0.9570938
```

Proportion of compliers increases quite a bit, from about 0.82 to about 0.98. Correlation is now 0.96, a significant increase from before.

Part 10: How does correlation between $D$and$Z$ vary with the proportion of compliers when we change α_Z?

The correlation also increases from about 0.67 to about 0.96, implying the coefficient on Z has a strong effect on the correlation between D and Z.

Part 11: Irrespective of whether $\alpha_Z = 5$ or $\alpha_Z = 10$, there is no external validity problem (e.g. LATE $\neq$ ATE) in this case. Show that you can justify this assertion directly from item 1 alone.

Part 12: Repeat with seed values of 2, 3, and 4

```
set.seed(2)

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
  )
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
-10.1670  -2.0674  -0.1108   1.9913  10.4347

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.05792    0.03628   56.72   <2e-16 ***
D_i          4.64392    0.06083   76.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.912 on 9998 degrees of freedom
Multiple R-squared:  0.3683,    Adjusted R-squared:  0.3682
F-statistic:  5828 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-11.07161  -2.13429  -0.02852   2.17754  12.07252
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.96252    0.04739   62.52   <2e-16 ***
D_i          2.10146    0.09933   21.16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.156 on 9998 degrees of freedom
Multiple R-Squared: 0.2579, Adjusted R-squared: 0.2578
Wald test: 447.6 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-12.4818  -2.4946   0.1184   2.4921  13.0604

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.96252    0.05408   54.77   <2e-16 ***
D_hat        2.10146    0.11338   18.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.603 on 9998 degrees of freedom
Multiple R-squared:  0.03322,   Adjusted R-squared:  0.03312
F-statistic: 343.6 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-10.2498  -1.8127   0.0082   1.8597  10.5775

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.96252    0.04060   72.96   <2e-16 ***
D_i          2.10146    0.08511   24.69   <2e-16 ***
D_tilde      4.54446    0.11379   39.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.705 on 9997 degrees of freedom
Multiple R-squared:  0.4552,    Adjusted R-squared:  0.4551
F-statistic:  4176 on 2 and 9997 DF,  p-value: < 2.2e-16
```

```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
-10.35525  -1.95761  -0.02863   1.97623  11.55490

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.90669    0.04363   66.62   <2e-16 ***
D_i          3.67491    0.09146   40.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.906 on 9998 degrees of freedom
```

```
Multiple R-Squared: 0.4111, Adjusted R-squared: 0.4111
Wald test:  1614 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
-12.12745  -2.41072   0.07622   2.40888  12.59013

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.01835    0.05225  57.764  < 2e-16 ***
D_i          0.52802    0.10954   4.821 1.45e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.481 on 9998 degrees of freedom
Multiple R-Squared: 0.06862,    Adjusted R-squared: 0.06853
Wald test: 23.24 on 1 and 9998 DF,  p-value: 1.453e-06
```

```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
)
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
-11.142994  -2.137132  -0.006763   2.200596  12.102601

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92684    0.04624   63.29   <2e-16 ***
D_i          3.10706    0.06529   47.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.139 on 9998 degrees of freedom
Multiple R-Squared: 0.233,  Adjusted R-squared: 0.2329
Wald test:  2265 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
-11.13680  -2.16387  -0.03453   2.17185  12.10879

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.00065    0.04733   63.40   <2e-16 ***
D_i          1.02707    0.06683   15.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.213 on 9998 degrees of freedom
Multiple R-Squared: 0.04927,    Adjusted R-squared: 0.04918
Wald test: 236.2 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 1),
    pC = 1 - pAT - pNT
  )

summarize(df, pC = mean(pC), pNT = mean(pNT), pAT = mean(pAT))
```

```
      pC    pNT    pAT
1 0.8166 0.1658 0.0176
```

```r
cor(df$Z_i, df$D_i)
```

```
[1] 0.6637294
```

```r
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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 1),
    pC = 1 - pAT - pNT
  )

summarize(df, pC = mean(pC), pNT = mean(pNT), pAT = mean(pAT))
```

```
      pC    pNT    pAT
1 0.9809 0.0015 0.0176
```

```r
cor(df$Z_i, df$D_i)
```

```
[1] 0.962287
```

```r
set.seed(3)
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
  )
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
-10.3604  -2.0804  -0.1357   2.0174  10.5062

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.09125    0.03665   57.06   <2e-16 ***
D_i          4.54605    0.06148   73.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.942 on 9998 degrees of freedom
Multiple R-squared:  0.3535,    Adjusted R-squared:  0.3535
F-statistic:  5467 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-11.32054  -2.15633   0.01342   2.15176  12.24840

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.05137    0.04859   62.79   <2e-16 ***
D_i          1.84377    0.10258   17.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.214 on 9998 degrees of freedom
Multiple R-Squared: 0.2286, Adjusted R-squared: 0.2285
Wald test: 323.1 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-12.4803  -2.5186   0.1553   2.4288  14.0154

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.05137    0.05464   55.85   <2e-16 ***
D_hat        1.84377    0.11533   15.99   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.614 on 9998 degrees of freedom
```

```
Multiple R-squared:  0.02493,   Adjusted R-squared:  0.02483
F-statistic: 255.6 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-11.1236  -1.7829   0.0116   1.8346   9.5598

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.05137    0.04114   74.17   <2e-16 ***
D_i          1.84377    0.08684   21.23   <2e-16 ***
D_tilde      4.72955    0.11488   41.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.721 on 9997 degrees of freedom
Multiple R-squared:  0.4472,    Adjusted R-squared:  0.4471
F-statistic:  4044 on 2 and 9997 DF,  p-value: < 2.2e-16
```

```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
-11.254  -1.965  -0.020   1.943  10.719
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.98493    0.04456   66.98   <2e-16 ***
D_i          3.43945    0.09407   36.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.948 on 9998 degrees of freedom
Multiple R-Squared: 0.389,  Adjusted R-squared: 0.389
Wald test:  1337 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
-12.3145  -2.4747   0.1354   2.3866  13.7776

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.11782    0.05375  58.004   <2e-16 ***
D_i          0.24808    0.11346   2.186   0.0288 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.555 on 9998 degrees of freedom
Multiple R-Squared: 0.03252,    Adjusted R-squared: 0.03243
Wald test: 4.781 on 1 and 9998 DF,  p-value: 0.02881
```

```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
```

```
)
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
-11.27465  -2.11405   0.02026   2.13714  11.19224

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.00548    0.04649   64.64   <2e-16 ***
D_i          2.94582    0.06596   44.66   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.146 on 9998 degrees of freedom
Multiple R-Squared: 0.2248, Adjusted R-squared: 0.2247
Wald test:  1994 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```
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
-11.678019  -2.149812  -0.004922   2.113936  13.201871

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

```
(Intercept)  3.09280    0.04788    64.6   <2e-16 ***
D_i          0.84887    0.06793    12.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.239 on 9998 degrees of freedom
Multiple R-Squared: 0.04213,    Adjusted R-squared: 0.04203
Wald test: 156.2 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 0),
    pC = 1 - pAT - pNT
  )
length(df$compliers)
```

```
[1] 0
```

```r
cor(df$Z_i, df$D_i)
```

```
[1] 0.6547057
```

```r
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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 0),
    pC = 1 - pAT - pNT
  )
length(df$compliers)
```

```
[1] 0
```

```r
cor(df$Z_i, df$D_i)
```

```
[1] 0.9544518
```

```r
set.seed(4)

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
  )
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
-11.4046  -1.9854  -0.1565   1.9254  10.3646

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13905    0.03606   59.32   <2e-16 ***
D_i          4.49902    0.06058   74.26   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.898 on 9998 degrees of freedom
Multiple R-squared:  0.3555,    Adjusted R-squared:  0.3554
F-statistic:  5515 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-12.300464  -2.116453   0.006042   2.120659  11.762486

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.03490    0.04691   64.70   <2e-16 ***
D_i          1.97053    0.09837   20.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.14 on 9998 degrees of freedom
Multiple R-Squared: 0.2432, Adjusted R-squared: 0.2431
Wald test: 401.3 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-13.6293  -2.4288   0.1159   2.4444  12.5395

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0349     0.0531   57.16   <2e-16 ***
```

```
D_hat         1.9705     0.1113   17.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.554 on 9998 degrees of freedom
Multiple R-squared:  0.03038,   Adjusted R-squared:  0.03028
F-statistic: 313.2 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-11.3063  -1.7988   0.0274   1.7855  10.2777

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.03490    0.04015   75.59   <2e-16 ***
D_i          1.97053    0.08419   23.41   <2e-16 ***
D_tilde      4.55943    0.11305   40.33   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.688 on 9997 degrees of freedom
Multiple R-squared:  0.4457,    Adjusted R-squared:  0.4456
F-statistic:  4019 on 2 and 9997 DF,  p-value: < 2.2e-16
```

```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
-11.4138  -1.9213  -0.0336   1.9296  11.2524

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.97866    0.04319   68.96   <2e-16 ***
D_i          3.53683    0.09057   39.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.891 on 9998 degrees of freedom
Multiple R-Squared: 0.399,  Adjusted R-squared: 0.3989
Wald test:  1525 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
-13.35670  -2.37179   0.08975   2.40051  12.27255

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.09114    0.05172  59.771  < 2e-16 ***
D_i          0.40422    0.10844   3.728 0.000194 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.462 on 9998 degrees of freedom
Multiple R-Squared: 0.05264,    Adjusted R-squared: 0.05254
Wald test: 13.89 on 1 and 9998 DF,  p-value: 0.0001944
```

```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
)
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
-12.2839  -2.0965   0.0221   2.1158  11.7496

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.99717    0.04544   65.96   <2e-16 ***
D_i          3.02112    0.06451   46.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.099 on 9998 degrees of freedom
Multiple R-Squared: 0.2302, Adjusted R-squared: 0.2301
Wald test:  2193 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
```

```
-12.27719  -2.13016  -0.01604   2.10461  11.75629

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.07191    0.04658   65.94   <2e-16 ***
D_i          0.93971    0.06614   14.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.177 on 9998 degrees of freedom
Multiple R-Squared: 0.04528,    Adjusted R-squared: 0.04518
Wald test: 201.9 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
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
```

```r
cor(df$Z_i, df$D_i)
```

```
[1] 0.6674092
```

```r
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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 1),
    pC = 1 - pAT - pNT,
  )
length(df$compliers)
```

```
[1] 10000
```

```
cor(df$Z_i, df$D_i)
```

```
[1] 0.9613202
```

Part 13: Now, recreate DGP with $N = 500$.

```
set.seed(1)

df <- data.frame(
  id = 1:500,
  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
  )
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
```

```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
```

```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
```

```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
```

```r
df <- df %>%
  mutate(
    Y_i = 3 + 2 * D_i + 1 * Z_i + 6 * U_i + epsilon_Y
)
```

```r
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
```

```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
)
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
```

```r
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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 0),
    pC = 1 - pAT - pNT
  )
summarize(df, pC = mean(pC), pNT = mean(pNT), pAT = mean(pAT))
```

```
      pC    pNT    pAT
1 0.3658 0.1579 0.4763
```

```r
cor(df$D_i, df$Z_i)
```

```
[1] 0.6694964
```

```r
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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 0),
    pC = 1 - pAT - pNT
  )
summarize(df, pC = mean(pC), pNT = mean(pNT), pAT = mean(pAT))
```

```
      pC    pNT    pAT
1 0.5224 0.0013 0.4763
```

```
cor(df$D_i, df$Z_i)
```

```
[1] 0.9570938
```

```
set.seed(2)

df <- data.frame(
  id = 1:500,
  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
  )
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
-10.1670  -2.0674  -0.1108   1.9913  10.4347
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.05792    0.03628   56.72   <2e-16 ***
D_i          4.64392    0.06083   76.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.912 on 9998 degrees of freedom
Multiple R-squared:  0.3683,    Adjusted R-squared:  0.3682
F-statistic:  5828 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-11.07161  -2.13429  -0.02852   2.17754  12.07252

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.96252    0.04739   62.52   <2e-16 ***
D_i          2.10146    0.09933   21.16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.156 on 9998 degrees of freedom
Multiple R-Squared: 0.2579, Adjusted R-squared: 0.2578
Wald test: 447.6 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-12.4818  -2.4946   0.1184   2.4921  13.0604

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.96252    0.05408   54.77   <2e-16 ***
D_hat        2.10146    0.11338   18.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.603 on 9998 degrees of freedom
Multiple R-squared:  0.03322,   Adjusted R-squared:  0.03312
F-statistic: 343.6 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-10.2498  -1.8127   0.0082   1.8597  10.5775

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.96252    0.04060   72.96   <2e-16 ***
D_i          2.10146    0.08511   24.69   <2e-16 ***
D_tilde      4.54446    0.11379   39.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.705 on 9997 degrees of freedom
Multiple R-squared:  0.4552,    Adjusted R-squared:  0.4551
F-statistic:  4176 on 2 and 9997 DF,  p-value: < 2.2e-16
```

```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
-10.35525  -1.95761  -0.02863   1.97623  11.55490

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.90669    0.04363   66.62   <2e-16 ***
D_i          3.67491    0.09146   40.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.906 on 9998 degrees of freedom
Multiple R-Squared: 0.4111, Adjusted R-squared: 0.4111
Wald test:  1614 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
-12.12745  -2.41072   0.07622   2.40888  12.59013
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.01835    0.05225  57.764  < 2e-16 ***
D_i          0.52802    0.10954   4.821 1.45e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.481 on 9998 degrees of freedom
Multiple R-Squared: 0.06862,    Adjusted R-squared: 0.06853
Wald test: 23.24 on 1 and 9998 DF,  p-value: 1.453e-06
```

```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
)
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
-11.142994  -2.137132  -0.006763   2.200596  12.102601

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92684    0.04624   63.29   <2e-16 ***
D_i          3.10706    0.06529   47.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.139 on 9998 degrees of freedom
Multiple R-Squared: 0.233,  Adjusted R-squared: 0.2329
Wald test:  2265 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
-11.13680  -2.16387  -0.03453   2.17185  12.10879

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.00065    0.04733   63.40   <2e-16 ***
D_i          1.02707    0.06683   15.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.213 on 9998 degrees of freedom
Multiple R-Squared: 0.04927,    Adjusted R-squared: 0.04918
Wald test: 236.2 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```
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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 1),
    pC = 1 - pAT - pNT
  )
summarize(df, pC = mean(pC), pNT = mean(pNT), pAT = mean(pAT))
```

```
      pC    pNT    pAT
1 0.8166 0.1658 0.0176
```

```
cor(df$D_i, df$Z_i)
```

```
[1] 0.6637294
```

```r
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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 1),
    pC = 1 - pAT - pNT
  )
summarize(df, pC = mean(pC), pNT = mean(pNT), pAT = mean(pAT))
```

```
      pC    pNT    pAT
1 0.9809 0.0015 0.0176
```

```r
cor(df$D_i, df$Z_i)
```

```
[1] 0.962287
```

```r
set.seed(3)

df <- data.frame(
  id = 1:500,
  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
  )
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
-10.3604  -2.0804  -0.1357   2.0174  10.5062

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.09125    0.03665   57.06   <2e-16 ***
D_i          4.54605    0.06148   73.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.942 on 9998 degrees of freedom
Multiple R-squared:  0.3535,    Adjusted R-squared:  0.3535
F-statistic:  5467 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-11.32054  -2.15633   0.01342   2.15176  12.24840

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.05137    0.04859   62.79   <2e-16 ***
D_i          1.84377    0.10258   17.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.214 on 9998 degrees of freedom
Multiple R-Squared: 0.2286, Adjusted R-squared: 0.2285
Wald test: 323.1 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-12.4803  -2.5186   0.1553   2.4288  14.0154

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.05137    0.05464   55.85   <2e-16 ***
D_hat        1.84377    0.11533   15.99   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.614 on 9998 degrees of freedom
Multiple R-squared:  0.02493,   Adjusted R-squared:  0.02483
F-statistic: 255.6 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```
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
-11.1236  -1.7829   0.0116   1.8346   9.5598

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.05137    0.04114   74.17   <2e-16 ***
D_i          1.84377    0.08684   21.23   <2e-16 ***
D_tilde      4.72955    0.11488   41.17   <2e-16 ***
---
```

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

Residual standard error: 2.721 on 9997 degrees of freedom
Multiple R-squared:  0.4472,    Adjusted R-squared:  0.4471
F-statistic:  4044 on 2 and 9997 DF,  p-value: < 2.2e-16
```

```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
-11.254  -1.965  -0.020   1.943  10.719

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.98493    0.04456   66.98   <2e-16 ***
D_i          3.43945    0.09407   36.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.948 on 9998 degrees of freedom
Multiple R-Squared: 0.389,  Adjusted R-squared: 0.389
Wald test:  1337 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
-12.3145  -2.4747   0.1354   2.3866  13.7776

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.11782    0.05375  58.004   <2e-16 ***
D_i          0.24808    0.11346   2.186   0.0288 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.555 on 9998 degrees of freedom
Multiple R-Squared: 0.03252,    Adjusted R-squared: 0.03243
Wald test: 4.781 on 1 and 9998 DF,  p-value: 0.02881
```

```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
)
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
-11.27465  -2.11405   0.02026   2.13714  11.19224

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.00548    0.04649   64.64   <2e-16 ***
D_i          2.94582    0.06596   44.66   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.146 on 9998 degrees of freedom
Multiple R-Squared: 0.2248, Adjusted R-squared: 0.2247
```

```
Wald test:  1994 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
-11.678019  -2.149812  -0.004922   2.113936  13.201871

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.09280    0.04788    64.6   <2e-16 ***
D_i          0.84887    0.06793    12.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.239 on 9998 degrees of freedom
Multiple R-Squared: 0.04213,    Adjusted R-squared: 0.04203
Wald test: 156.2 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 1),
    pC = 1 - pAT - pNT
  )
summarize(df, pC = mean(pC), pNT = mean(pNT), pAT = mean(pAT))
```

```
      pC   pNT    pAT
1 0.8132 0.166 0.0208
```

```r
cor(df$D_i, df$Z_i)
```

```
[1] 0.6547057
```

```r
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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 1),
    pC = 1 - pAT - pNT
  )
summarize(df, pC = mean(pC), pNT = mean(pNT), pAT = mean(pAT))
```

```
      pC    pNT    pAT
1 0.9769 0.0023 0.0208
```

```r
cor(df$D_i, df$Z_i)
```

```
[1] 0.9544518
```

```r
set.seed(4)

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
  )
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
-11.4046  -1.9854  -0.1565   1.9254  10.3646

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13905    0.03606   59.32   <2e-16 ***
D_i          4.49902    0.06058   74.26   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.898 on 9998 degrees of freedom
Multiple R-squared:  0.3555,    Adjusted R-squared:  0.3554
F-statistic:  5515 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-12.300464  -2.116453   0.006042   2.120659  11.762486

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.03490    0.04691   64.70   <2e-16 ***
D_i          1.97053    0.09837   20.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.14 on 9998 degrees of freedom
Multiple R-Squared: 0.2432, Adjusted R-squared: 0.2431
Wald test: 401.3 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-13.6293  -2.4288   0.1159   2.4444  12.5395

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0349     0.0531   57.16   <2e-16 ***
D_hat         1.9705     0.1113   17.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.554 on 9998 degrees of freedom
Multiple R-squared:  0.03038,   Adjusted R-squared:  0.03028
F-statistic: 313.2 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```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
-11.3063  -1.7988   0.0274   1.7855  10.2777

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.03490    0.04015   75.59   <2e-16 ***
```

```
D_i          1.97053    0.08419   23.41   <2e-16 ***
D_tilde      4.55943    0.11305   40.33   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.688 on 9997 degrees of freedom
Multiple R-squared:  0.4457,    Adjusted R-squared:  0.4456
F-statistic:  4019 on 2 and 9997 DF,  p-value: < 2.2e-16
```

```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
-11.4138  -1.9213  -0.0336   1.9296  11.2524

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.97866    0.04319   68.96   <2e-16 ***
D_i          3.53683    0.09057   39.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.891 on 9998 degrees of freedom
Multiple R-Squared: 0.399,  Adjusted R-squared: 0.3989
Wald test:  1525 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
-13.35670  -2.37179   0.08975   2.40051  12.27255

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.09114    0.05172  59.771  < 2e-16 ***
D_i          0.40422    0.10844   3.728 0.000194 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.462 on 9998 degrees of freedom
Multiple R-Squared: 0.05264,    Adjusted R-squared: 0.05254
Wald test: 13.89 on 1 and 9998 DF,  p-value: 0.0001944
```

```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
)
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
-12.2839  -2.0965   0.0221   2.1158  11.7496

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.99717    0.04544   65.96   <2e-16 ***
D_i          3.02112    0.06451   46.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.099 on 9998 degrees of freedom
Multiple R-Squared: 0.2302, Adjusted R-squared: 0.2301
Wald test:  2193 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
-12.27719  -2.13016  -0.01604   2.10461  11.75629

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.07191    0.04658   65.94   <2e-16 ***
D_i          0.93971    0.06614   14.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.177 on 9998 degrees of freedom
Multiple R-Squared: 0.04528,    Adjusted R-squared: 0.04518
Wald test: 201.9 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 1),
    pC = 1 - pAT - pNT
  )
summarize(df, pC = mean(pC), pNT = mean(pNT), pAT = mean(pAT))
```

```
      pC    pNT   pAT
1 0.8196 0.1624 0.018
```

```r
cor(df$D_i, df$Z_i)
```

```
[1] 0.6674092
```

```r
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,
    pNT = mean(Z_i == 1 & D_i == 0),
    pAT = mean(Z_i == 0 & D_i == 1),
    pC = 1 - pAT - pNT
  )
summarize(df, pC = mean(pC), pNT = mean(pNT), pAT = mean(pAT))
```

```
      pC    pNT   pAT
1 0.9804 0.0016 0.018
```

```r
cor(df$D_i, df$Z_i)
```

```
[1] 0.9613202
```