MY457: Solution Set 3 - Instrumental Variables

Wed/19/Mar

This problem set is formative and will not contribute to your final grade. However, it is designed to build and deepen your conceptual understanding of the material and to practice applying the material in R. Using tools and resources such as ChatGPT and Stack Overflow is allowed and encouraged. Bear in mind that you are ultimately responsible for the work you submit, and that these tools often provide incorrect solutions. Make sure that however you use these tools it aligns with your best interests, and enhances your learning in this class.

This problem set must be submitted on Moodle by 5pm on Wed/26/Mar. You must also use the provided `.Rmd` template to produce a .pdf with your answers. If your submission is late, is not a .pdf, or is not appropriately formatted, you will not receive feedback on your work.

1 Concepts

This question reviews some of the concepts covered in class. Mathematical notation can be a useful tool to explain concepts, but it's important that you understand and can explain the concepts clearly and concisely. If you want to support your explanations with mathematical notation, this page provides a tutorial on including mathematical notation in Rmarkdown.

Consider an encouragement design with $i \in [1, ..., N]$ units. We have an instrument/encouragement $Z_i \in [0, 1]$, but we cannot guarantee compliance with the actual treatment of interest $D_i \in [0, 1]$. We are interested in the effect of treatment D_i on Y_i.

1.1 What is the difference between encouragement and treatment? How do these concepts relate the Intent to Treat Effect (ITT) and the Local Average Treatment Effect (LATE)?

Encouragement refers to a mechanism that works as an incentive for individuals to take the treatment. Treatment, on the hand, refers to the actual treatment in the design.

In this setting, not all individuals that are encouraged to take treatment take it.

1.2 Given SUTVA, what are the underlying assumptions necessary for the identification of the LATE the above IV setting?

1- Relevance: The instrument has an effect on treatment status.

2- Exogeneity: The instrument is independent of the potential outcomes, both those of (i) D and those of (2) Y.

3- Exclusion: The instrument only affects the outcome via any effect on the treatment.

4- Monotonicity: The sign on the effect of treatment is the same for all units (or is zero). This means we rule out defiers.

1.3 Define the compliance types based on the different scenarios that may occur in this setting. Explain in words what each of the types means.

There are 4 compliance types in an IV setting. People whose treatment status follows their encouragement status (compliers); people that regardless of their encouragment will take treatment (always takers); individuals who will never get treatment regardless of their

encouragement status (never takers); and individuals who would do the opposite of their encouragement status (defiers).

2 Simulations

In this question we will use simulated data to test some of our intuitions about instrumental variables. The advantage of using a simulated dataset is that we have explicit control over the data generating process, and know the 'true' answer to any question we pose.

2.1. Explain the code below and relate it to an instrumental variables data generating process. Be sure to calculate both the true ATE and the true ITT and include that in your answer.

```
set.seed(123)

n_obs <- 1000

U <- rbinom(n_obs, 1, .75)

c_type <- ifelse(U == 1,
                 sample(1:3, n_obs, prob = c(0.7,0.1,0.2), replace=T),
                 sample(1:3, n_obs, prob = c(0.35,0.3,0.35), replace=T))

tau <- ifelse(c_type == 1, 5000, 1000)
tau <- ifelse(c_type == 3, 2500, tau)

Z <- rbinom(n_obs, 1, .5)

D <- ifelse(Z == 1 & c_type == 1, 1, NA)
D <- ifelse(Z == 0 & c_type == 1, 0, D)
D <- ifelse(c_type == 2, 1, D)
D <- ifelse(c_type == 3, 0, D)

Y0 <- rnorm(n_obs, mean = 50000, sd = 2500) + 25000*U
Y1 <- Y0 + tau

Y <- ifelse(D == 1, Y1, Y0)

data <- data.frame(
  cbind(
    Z,
    D,
    Y0,
    Y1,
    Y
  )
)
```

```
true_ate <- prop.table(table(c_type))[1] * 5000 + prop.table(table(c_type))[2]*1000 + prop.table(table(
true_ate <- mean(Y1 - Y0)

#true_itt <- [figure this out]
```

2.2 Using Ordinary Least Squares (OLS), naively estimate the treatment effect using only D as a regressor. What do you find? Does this estimator identify the ATE? Why?

```
lm(Y~D, data = data) %>% summary()
```

```
##
## Call:
## lm(formula = Y ~ D, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -26298  -1314   4984   7591  17189
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  69572.8      487.3 142.777 <0.0000000000000002 ***
## D             1791.3      729.6   2.455              0.0143 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11470 on 998 degrees of freedom
## Multiple R-squared:  0.006003,   Adjusted R-squared:  0.005007
## F-statistic: 6.027 on 1 and 998 DF,  p-value: 0.01426
```

2.3 Repeat the above analysis using only Z as a regressor. What do you find? What estimand does this estimator identify?

```
lm(Y~Z, data = data) %>% summary()
```

```
##
## Call:
## lm(formula = Y ~ Z, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28358  -2541   5064   7669  16776
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  68932.6      513.6 134.213 < 0.0000000000000002 ***
## Z             2844.2      722.0   3.939            0.0000875 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11420 on 998 degrees of freedom
## Multiple R-squared:  0.01531,    Adjusted R-squared:  0.01432
## F-statistic: 15.52 on 1 and 998 DF,  p-value: 0.00008746
```

2.4 Now, estimate the Local Average Treatment Effect (LATE) for the compliers, using the plug-in Wald estimator, considering both D and Z. Do you find any differences when you compare this result to your previous estimates? Explain what you find.

```
cov(Y, Z)/cov(D, Z)
```

```
## [1] 4698.147
```

2.5 Using the Two Stage Leasts Squares (2SLS) estimator, re-estimate the LATE. Do this both manually (using two `lm()` commands) and using the `AER::ivreg` command. How does your result compare to the previous result? How do the two approaches (`lm` and `ivreg`) differ, if at all?

```r
reg1  <- lm(D~Z)
D_hat <- predict(reg1)
summary(reg1)
```

```
##
## Call:
## lm(formula = D ~ Z)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7451 -0.1397 -0.1397  0.2549  0.8603
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  0.13968    0.01776   7.865 0.00000000000000952 ***
## Z            0.60538    0.02496  24.250 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3947 on 998 degrees of freedom
## Multiple R-squared:  0.3708, Adjusted R-squared:  0.3701
## F-statistic:   588 on 1 and 998 DF,  p-value: < 0.00000000000000022
```

```r
reg2 <- lm(Y~D_hat)
summary(reg2)
```

```
##
## Call:
## lm(formula = Y ~ D_hat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28358  -2541   5064   7669  16776
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  68276.3      642.9 106.207 < 0.0000000000000002 ***
## D_hat         4698.1     1192.7   3.939           0.0000875 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11420 on 998 degrees of freedom
## Multiple R-squared:  0.01531,    Adjusted R-squared:  0.01432
## F-statistic: 15.52 on 1 and 998 DF,  p-value: 0.00008746
```

```r
ivreg(Y~D | Z) %>% summary()
```

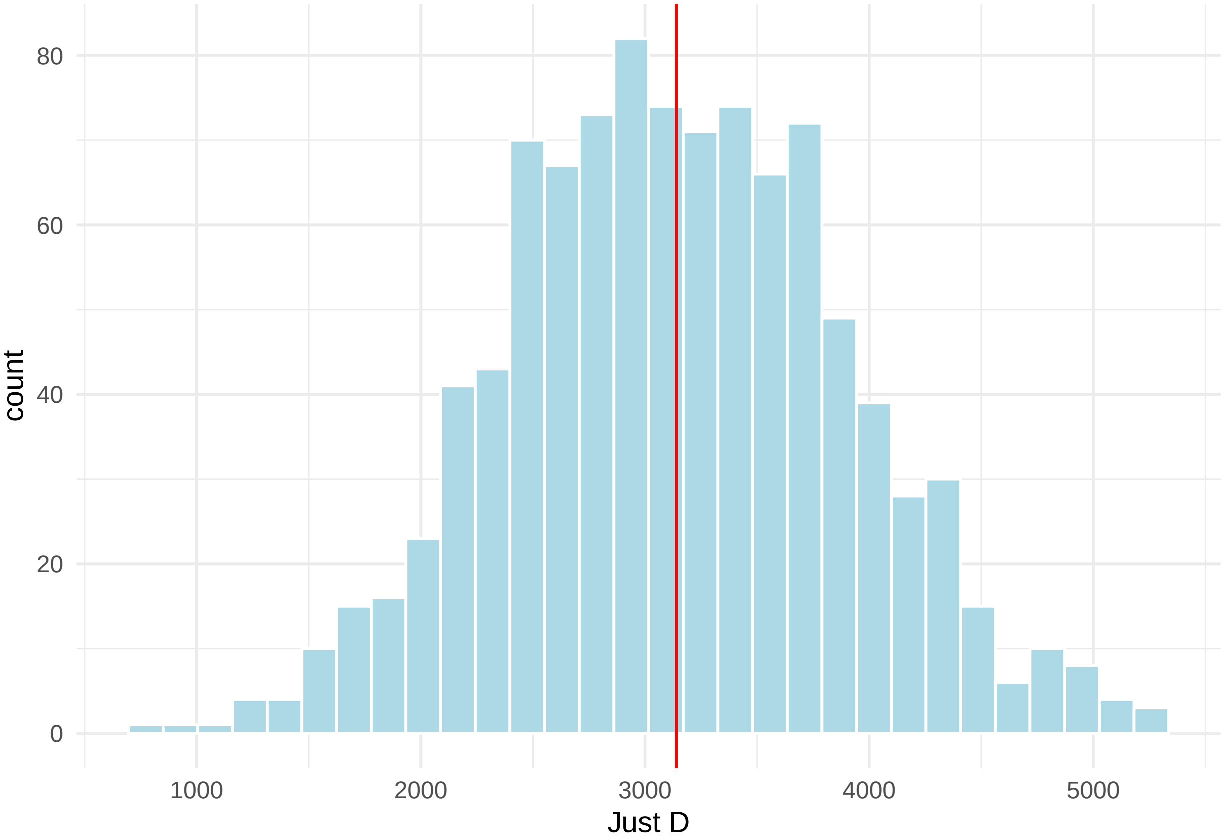
```
##
## Call:
## ivreg(formula = Y ~ D | Z)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27909  -2327   5392   7636  15578
##
```

```
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)    68276        651  104.88 < 0.0000000000000002 ***
## D               4698       1208    3.89             0.000107 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11560 on 998 degrees of freedom
## Multiple R-Squared: -0.009806,   Adjusted R-squared: -0.01082
## Wald test: 15.13 on 1 and 998 DF,  p-value: 0.000107
```

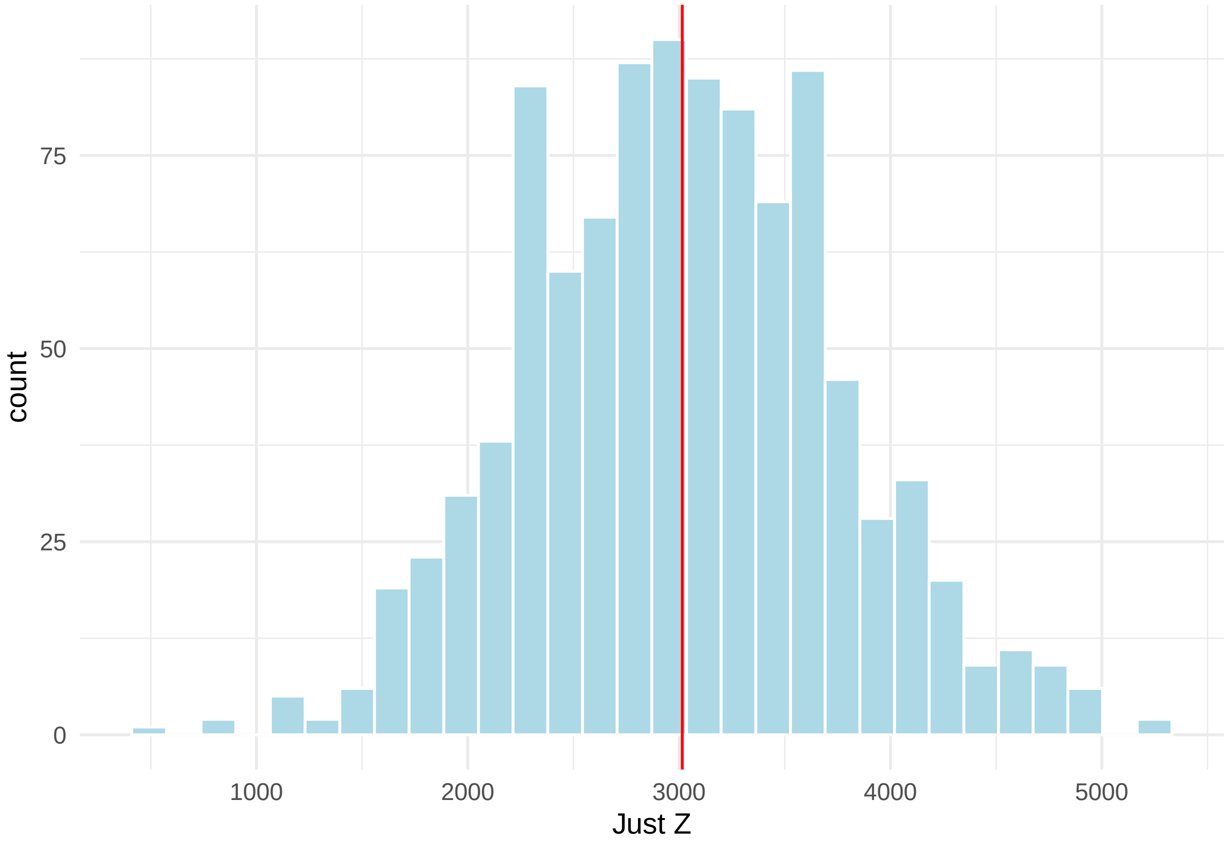
2.6 (Extra credit): Show that your answers to Questions 2.2, 2.3, and 2.4 were not due to chance. Over at least 1,000 repeat simulations (note: you will want to remove the `set.seed` step), generate a fresh draw of the data and estimate the sample-specific ATE, ITT, and LATE using the three different regression specifications. Calculate the difference between the estimated quantities and what you know to be the true values of these parameters, and store those differences. Finally, produce a histogram that shows the distributions of the differences over your repeated samples, along with their means. What do you conclude?

```r
set.seed(321)

n_obs <- 1000

U <- rbinom(n_obs, 1, .75)

first <- NULL
second <- NULL
wald_ <- NULL

for (i in 1:1000) {
  c_type <- ifelse(U == 1,
                  sample(1:3, n_obs, prob = c(0.7,0.1,0.2), replace=T),
                  sample(1:3, n_obs, prob = c(0.35,0.3,0.35), replace=T))

  tau <- ifelse(c_type == 1, 5000, 1000)
  tau <- ifelse(c_type == 3, 2500, tau)

  Z <- rbinom(n_obs, 1, .5)

  D <- ifelse(Z == 1 & c_type == 1, 1, NA)
  D <- ifelse(Z == 0 & c_type == 1, 0, D)
  D <- ifelse(c_type == 2, 1, D)
  D <- ifelse(c_type == 3, 0, D)

  Y0 <- rnorm(n_obs, mean = 50000, sd = 2500) + 25000*U
  Y1 <- Y0 + tau

  Y <- ifelse(D == 1, Y1, Y0)

  data <- data.frame(
    cbind(
      Z,
      D,
      Y0,
      Y1,
      Y
    )
  )

  reg1 <- lm(Y~D, data = data)
  reg2 <- lm(Y~Z, data = data)
  wald <- cov(Y, Z)/cov(D, Z)

  first[i] <- coef(reg1)[2]
  second[i] <- coef(reg2)[2]
  wald_[i] <- wald
}
```

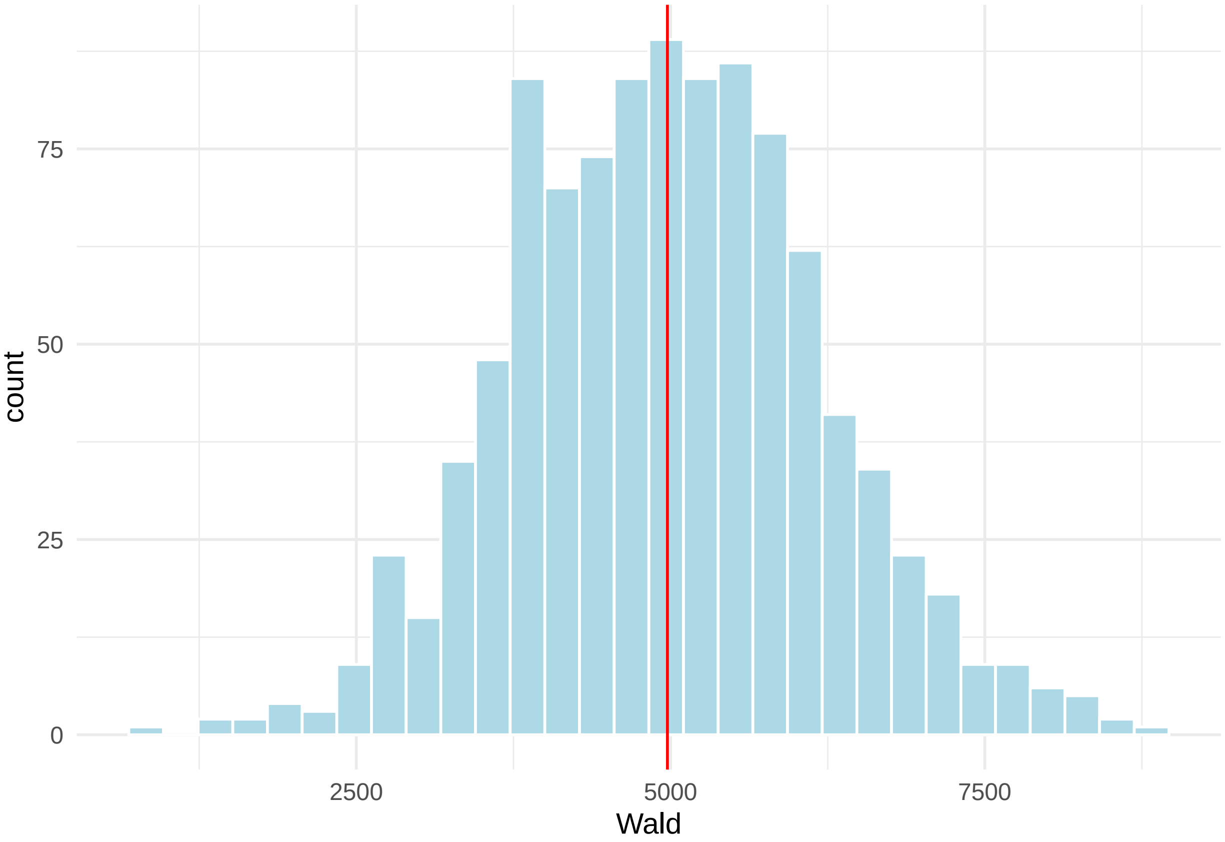
```
ggplot(data.frame(first)) + aes(x = first) +
  geom_histogram(color = "white", fill = "lightblue") +
  geom_vline(xintercept = mean(first), color = "red") +
  xlab("Just D") +
  theme_minimal()
```



```
ggplot(data.frame(second)) + aes(x = second) +
  geom_histogram(color = "white", fill = "lightblue") +
  geom_vline(xintercept = mean(second), color = "red") +
  xlab("Just Z") +
  theme_minimal()
```



```
ggplot(data.frame(wald_)) + aes(x = wald_) +
  geom_histogram(color = "white", fill = "lightblue") +
  geom_vline(xintercept = mean(wald_), color = "red") +
  xlab("Wald") +
  theme_minimal()
```



3 Replication

In this section, we will use real-world data to reinforce what we have learned. We will analyse the dataset employed in *Foreign Aid, Human Rights and Democracy Promotion: Evidence from a Natural Experiment.*

Many countries, especially lower- and middle-income countries, are provided foreign aid with the intention that this will improve the living conditions of those in need. This foreign aid might have some spillover effects by encouraging the protection of human rights and entrenching democratic institutions. Does foreign aid improve human rights and democracy?

The authors use instrumental variables to try and studying the effect of foreign aid, measured as overseas development assistance (`ODA`) on the the CIRI Empowerment Index (`CIRI`).

3.1 Explain in your own words the instrumental variable (IV) design that the authors use to answer the research question. Do you have any concerns about the identifying assumptions?

This study aims to determine whether increased foreign aid improves human rights and democracy. Due to the numerous confounding variables that influence both the receipt of foreign aid and the state of human rights and democracy in a country, a simple comparison is not feasible. To address this, the authors identify an “as-if random” instrument to eliminate confounders: whether a country’s former colonizer holds the presidency of the Council of the European Union. The presidency of the Council rotates among member countries in an “essentially random” manner, according to the authors. They demonstrate the instrument’s relevance by showing a positive association between a country’s former colonizer holding the presidency and the level of EU aid the country receives.

3.2 Read into `R` the replication data set (`final.dta`). Each row in the data is a country-year observation, and note that aside from `CIRI`, `ODA`, and `Colony`, all remaining variables are year or country dummy variables. Nai/“vely estimate the effect of aid (`ODA`) on the CIRI Empowerment Index (`CIRI`), controlling for two-way fixed effects. Be sure to present your results neatly, showing only relevant statistics for ODA and not for all your fixed effects. What do you find?

Hint: to include all variables in a data frame in one regression you can write y~. and to later on exclude some variables you can write y~.-var.

```
data <- read_dta("./foreign_aid_human_rights_and_democracy_promotion.dta")

coef(lm(CIRI~.-ODA, data = data))[2]
```

```
##   Colony
## 0.302316
```

If we only include Colony, we estimate an effect of 0.3, which is significant at the 95% level. This is the equivalent to estimating the reduced form (encouragement on outcomes).

3.3 Estimate the first stage, the second stage, and the ITT (also called the reduced form), again controlling for two-way fixed effects. Again, present the results neatly. What do you find?

```
# First stage
reg1 <- lm(ODA~.-CIRI, data = data)
coef(reg1)[2]
```

```
##    Colony
## 0.1603633
```

```
# Second stage
reg2 <- lm(CIRI~reg1$fitted.values+.-ODA-Colony, data = data)
coef(reg2)[2]
```

```
## reg1$fitted.values
##           1.885195
```

```
# Reduced form
reg3 <- lm(CIRI~Colony+.-ODA, data = data)
coef(reg3)[2]
```

```
##   Colony
## 0.302316
```

The first stage confirms that the relevance assumption holds. We estimate an effect of 0.1 of Z on D.

The second stage shows a treatment effect of 1.88 which is significant at the 95% level. These results are consistent with the findings of the authors.

The reduced form results in the same estimate as in the previous exercise.

3.4 Using the AER package, use the ivreg function and estimate the LATE. Are your results the same as in 3.3? Why?

```
coef(ivreg(CIRI~.-Colony | .-ODA, data = data))[2]
```

```
##      ODA
## 1.885195
```

We get the same point estimate, however, the standard errors and the p-value change. This is because we know that the error from the first stage have to be taken into account and the AER package corrects the standard errors for us.

3.5 Estimate the LATE using the plug-in estimator, this time without using any fixed effects. What do you find? Are your results different from before? Why?

```
cov(data$CIRI, data$Colony)/cov(data$ODA, data$Colony)
```

```
## [1] -4.869788
```

We get a different estimate. This is because in our regression we are using other control variables that account for variance that we do not include in the Wald estimate.