Exercise: Understanding Potential Outcomes

Let's see how the presence of **non-compliance** affects our treatment effect estimates in some simple 'fake' data.

1. First, let's generate an income variable for 10,000 people. The data should be drawn randomly from the normal distribution with mean 500 and standard deviation 50.

```
set.seed(54321)
N <- 10000
d <- tibble(income=rnorm(N,500,50))
```

2. Now let's simulate potential outcomes (let's say they are 'attitudes to redistribution') for each person that depend on their income. Assume $y_0 = N(10, 2) + \frac{income}{100}$ and $y_1 = y_0 + 2$ so there is a constant treatment effect of 2.

```
d <- d %>% mutate(y_0=rnorm(N,10,2) + income/100,
                  y_1=y_0+2)
```

3. Remember that the key issue with non-compliance is that there is a difference between **treatment assignment** and actual **treatment**. Treatment assignment in our case will be completely random, so create a random binomial variable with 50% chance of being assigned to treatment.

```
d <- d %>% mutate(D_assign=rbinom(N,1,0.5))
```

4. For most people, treatment will be the same as treatment assignment, so make another 'treatment' variable that directly copies 'treatment assignment'.

```
d <- d %>% mutate(D=D_assign)
```

5. To introduce non-compliance, let's adjust this 'treatment' variable so that rich people with an income above 570 are **'Never-takers'** - regardless of their value for treatment assignment, they never receive treatment (so treatment=0).

```
d <- d %>% mutate(D=ifelse(income>570,0,D))
```

6. Now calculate the observed outcome based on potential outcomes and actual treatment status.

```
d <- d %>% mutate(y_obs=case_when(D==0~y_0,
                                  D==1~y_1))
```

7. Now let's calculate the standard Average Treatment Effect by running a regression of the observed outcomes on treatment (actual treatment, not treatment assignment). What is your estimate of the average treatment effect? How does this compare to the treatment effect we specified earlier?

```
d %>% lm(y_obs ~ D, data=.) %>% stargazer(single.row=T, header=F, title="Q7")
```

8. Now let's imagine that people with an income below 430 are all **Always-Takers**, so regardless of treatment assignment they always receive treatment (treatment=1). Remember to re-calculate observed outcomes afterwards.

```
d <- d %>% mutate(D=ifelse(income<430,1,D),
                  y_obs=case_when(D==0~y_0,
                                  D==1~y_1))
```

9. Re-run the regression from Q7 on our dataset that includes both never-takers and always-takers. How does this change our estimates of the Average Treatment Effect?

Table 1: Q7

	Dependent variable:
	y_obs
D	1.927*** (0.041)
Constant	15.022*** (0.028)
Observations	10,000
R^2	0.179
Adjusted R^2	0.179
Residual Std. Error	2.057 (df = 9998)
F Statistic	2,181.670*** (df = 1; 9998)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
d %>% lm(y_obs ~ D, data=.) %>% stargazer(single.row=T, header=F, title="Q9")
```

Table 2: Q9

	Dependent variable:
	y_obs
D	1.742*** (0.041)
Constant	15.119*** (0.029)
Observations	10,000
R^2	0.153
Adjusted R^2	0.152
Residual Std. Error	2.053 (df = 9998)
F Statistic	1,799.341*** (df = 1; 9998)
Note:	*p<0.1; **p<0.05; ***p<0.01

10. Given these biases, we can try to use treatment assignment as an instrumental variable for actual treatment. To do this, we first need to run the **First Stage** regression to show that treatment assignment explains treatment. ($D_i \sim Z_i$) Is treatment assignment a good instrument for treatment? Why?

```
d %>% lm(D ~ D_assign, data=.) %>% stargazer(single.row=T, header=F, title="Q10")
```

Table 3: Q10

	Dependent variable:
	D
D_assign	0.837*** (0.005)
Constant	0.088*** (0.004)
Observations	10,000
R^2	0.700
Adjusted R^2	0.700
Residual Std. Error	0.274 (df = 9998)
F Statistic	23,366.810*** (df = 1; 9998)
Note:	*p<0.1; **p<0.05; ***p<0.01

11. Save the fitted values from this regression as a new column in your dataset.

```
d <- d %>% mutate(First_stage_fitted=lm(D ~ D_assign, data=.)$fitted.values)
```

12. Now for the second stage of our instrumental variables analysis, use a regression to estimate how these fitted values explain the observed outcomes. ($y_{obs,i} \sim \hat{D}_i$) How does the result compare to our initial assumption about the size of the treatment effect?

```
d %>% lm(y_obs ~ First_stage_fitted, data=.) %>% stargazer(single.row=T, header=F, title="Q12")
```

Table 4: Q12

	Dependent variable:
	y_obs
First_stage_fitted	2.061*** (0.049)
Constant	14.957*** (0.032)
Observations	10,000
R^2	0.150
Adjusted R^2	0.149
Residual Std. Error	2.057 (df = 9998)
F Statistic	1,758.126*** (df = 1; 9998)
Note:	*p<0.1; **p<0.05; ***p<0.01

13. How should we interpret this estimate? What group does it apply to?

14. The only thing wrong with our 2-Stage Least Squares Regression is that the standard errors are too small. To correct this, we can use an all-in-one Instrumental Variables estimator, eg. *ivreg* in the *AER* package in R or *ivregress* in Stata. How do the standard errors change?

```
library(AER)
d %>% ivreg(y_obs ~ D|D_assign, data=.) %>% stargazer(single.row=T, header=F, title="Q14")
```

Table 5: Q14

	Dependent variable:
	y_obs
D	2.061*** (0.049)
Constant	14.957*** (0.032)
Observations	10,000
R^2	0.147
Adjusted R^2	0.147
Residual Std. Error	2.059 (df = 9998)
Note:	*p<0.1; **p<0.05; ***p<0.01

15. Finally, let's so how this changes when we introduce some defiers: Change your data so that anyone with income less than 480 who was assigned to control actually receives treatment AND so that anyone assigned to treatment actually receives control. Calculate observed outcomes again. Then run the Instrumental Variables regression (2SLS or the all-in-one) and interpret the results.

```
d <- d %>% mutate(D=case_when(income<480 & D_assign==0~1,
                        income<480 & D_assign==1~0,
                        income>=480~D),
```

```
            y_obs=case_when(D==0~y_0,
                            D==1~y_1))

d %>% ivreg(y_obs ~ D|D_assign, data=.) %>% stargazer(single.row=T, header=F, title="Q15")
```

Table 6: Q15

	Dependent variable:
	y_obs
D	2.217*** (0.174)
Constant	14.888*** (0.082)
Observations	10,000
R^2	0.180
Adjusted R^2	0.180
Residual Std. Error	2.061 (df = 9998)
Note:	*p<0.1; **p<0.05; ***p<0.01