Assignment 2

Answer Key

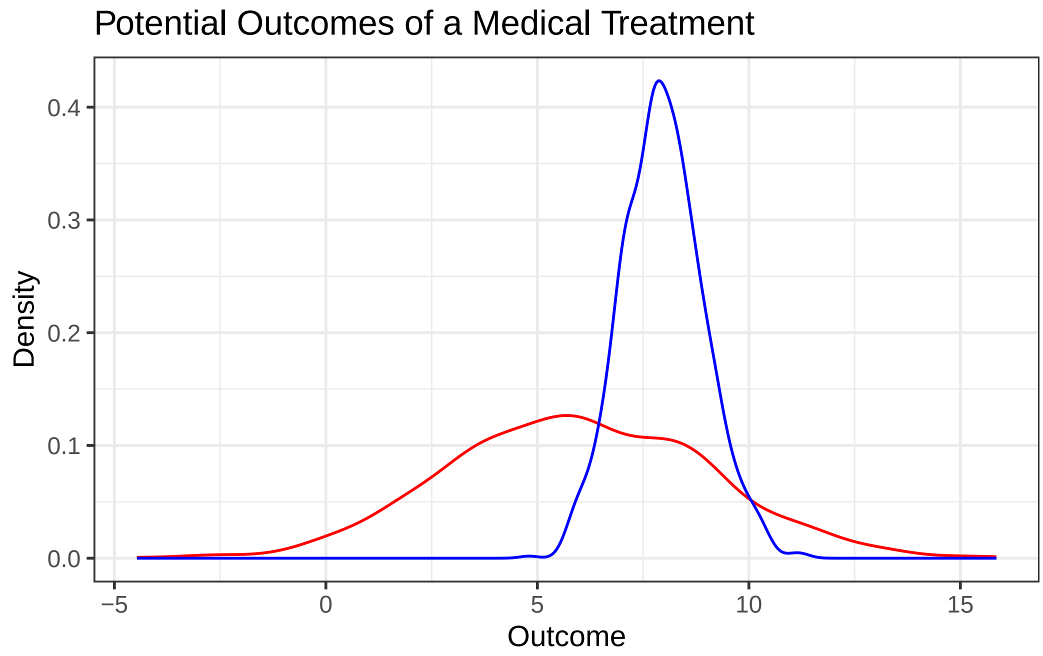
2023-11-06

Questions

1. Suppose that we are interested in the effect of a medical treatment on an outcome y. The code below creates potential outcome with treatment y_1 and the potential outcome y_0 without treatment for 1000 fictional people.

Plot the kernel density estimate of y_0 and y_1 on the same plot. On your plot, ensure that the two density estimates have different colours, the axes are labelled intuitively, and the plot has a title.

```r
potential <- tibble(
  y0 = 6 + rnorm(1000,0,3),
  y1 = 8 + rnorm(1000,0,1)
)

ggplot(potential) +
  geom_density(aes(x = y0), color = "red", alpha = 0.5) +
  geom_density(aes(x = y1), color = "blue", alpha = 0.5) +
  labs(x = "Outcome", y = "Density", title = "Potential Outcomes of a Medical Treatment")
  theme_bw()
```



2. Pretend that the doctor who assigns people to treatment is evil, and puts people in the treatment group only if they don't benefit from it. Add the treatment variable $w = 1$ for the treatment group and $w = 0$ for the control to the dataset based on this scenario, and compute the Average Treatment Effect on the Treated (ATT), the Average Treatment Effect on the Non-Treated (ATNT), and the Average Treatment Effect (ATE). Comment on the differences between them an explain why you think they are different.

The ATE, ATT, and ATE are listed in the table below. Remember that the ATE is the average of $y_1 - y_0$ in the population (i.e. it ignores which treatment group you are in). In this case, it equals 1.8 and the average person in the population benefits from the treatment. The ATT is the average of $y_1 - y_0$ *among people who get treated*. Once you assign the treatement group based on the evil doctor, the average is -2. This makes sense because people only get treated if they don't benefit from it, and so among this group the ATT is negative. Conversely, the ATNT is the the average of $y_1 - y_0$ *among people who do not get treated*, and equals 3.2. Again, people only end up in the control group if they benefit from the treatment, and so among this group the ATNT is positive.

```r
potential <- potential %>%
  mutate(w = ifelse(y0 > y1, 1, 0))

treateffects <- tibble(ate = mean(potential$y1 - potential$y0),
                       att = mean(potential$y1[potential$w == 1] -
                                    potential$y0[potential$w == 1]),
```

Table 1: Summary Statistics

Variable	Mean
ate	2
att	-1.9
atnt	3.4

```
                       atnt = mean(potential$y1[potential$w == 0] -
                                     potential$y0[potential$w == 0]))

sumtable(treateffects, summ=c('mean(x)'),summ.names= c('Mean'))
```

3. Compute the observed outcome $y = y_0 + (y_1 - y_0)w$ based on the actual treatment status. Estimate the regression of y on w, and report the results in a professional-looking output table (for example stargazer, but you can do it any way you like). Explain why it does not match any of the average treatment effects computed above.

The slope estimate in the regression is 2.579, which is not equal to any of the treatment effects. This is because people are selected into treatment based on both y_0 and y_1, and when this happens we have selection bias. Another way to say this is that we have no randomization, no independence, and no mean independence of the potential outcomes, which means bias. Intuitively, a completely non-random set of people end up in both treatment and control, so it is no surprise that the observed differences between treatment and control are biased.

```
potential <- potential %>%
  mutate(y = y0 + (y1 - y0)*w)

model <- lm(y ~ w, data = potential)

stargazer(model, type = "text")
```

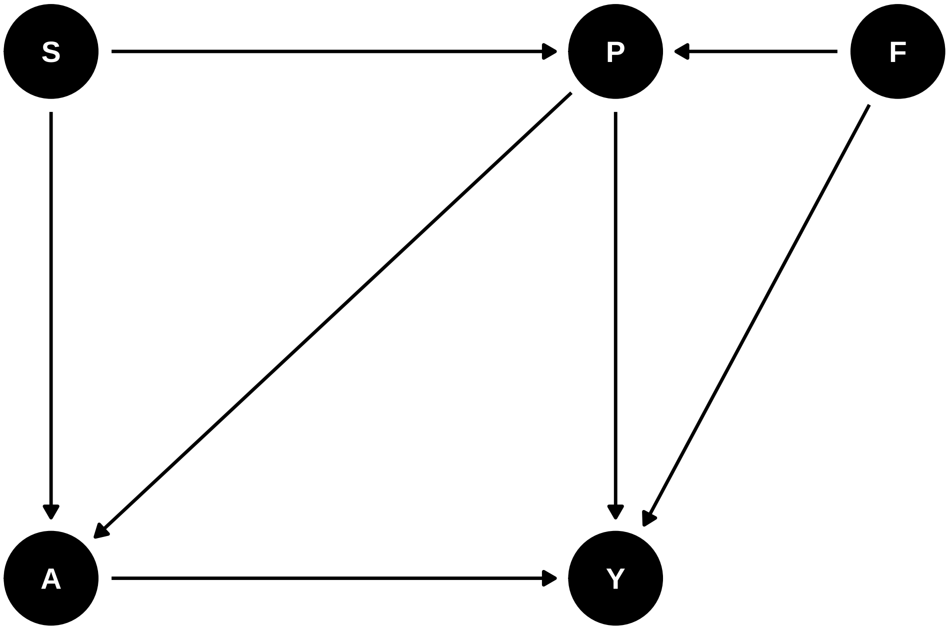
```
===============================================
                        Dependent variable:    
                    ---------------------------
                                 y             
-----------------------------------------------
w                             3.020***         
                               (0.140)         
                                               
Constant                      4.635***         
```

```
                              (0.074)

-----------------------------------------------
Observations                   1,000
R2                             0.317
Adjusted R2                    0.316
Residual Std. Error      1.981 (df = 998)
F Statistic          462.406*** (df = 1; 998)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

4. You are a research assistant helping an instructor on a study that tries to measure the effect of Parental Involvement (P) on Student Test Scores (Y). The DAG below relates these two variables with other factors including Socioeconomic Status (S), Extracurricular Activities (A), and Family Stress (F) [NOTE: compile the document to see the DAG]. List all of the paths that connect P to Y, and indicate whether they are front door or back door paths.

Path	Type
$P \rightarrow Y$	Front Door
$P \rightarrow A \rightarrow Y$	Front Door
$P \leftarrow F \rightarrow Y$	Back Door
$P \leftarrow S \rightarrow A \rightarrow Y$	Back Door

```r
coord_dag<-list(x = c(A = 0, S = 0, Y = 1, P = 1, F = 1.5),
                y = c(A = 0, S = 1, Y = 0, P = 1, F = 1))
dag <- dagify(Y~P + A + F, P ~ F + S, A~ P + S, coords = coord_dag) %>%
  tidy_dagitty()

ggplot(dag,aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_dag_point() +
  geom_dag_edges() +
  geom_dag_text() +
  theme_dag()
```



5. Assume that the relationships above are all linear. Write down a linear regression model that would allow you estimate the direct effect of P on Y, and explain why the model identifies the causal effect.

In the DAG in the previous question, there are two front door paths from P to Y: one that is direct, and another that goes through a mediator A. Thus to get the direct effect of P on Y, we will need to close down the path that goes through the mediator by controlling for A. There are also two backdoor paths that have no colliders, so we also need to control for variables along those paths. Controlling for A already closes one of them, and to close the other we need to control for F. With that, we can identify the causal effect with the model

$$Y = \beta_0 + \beta_1 P + \beta_2 A + \beta_3 F + u$$

where β_1 is the direct effect of P on Y.

6. You are considering having children and you want to know whether it will affect your sleep. You decide to use your econometric skills to estimate the relationship with some data. The dataset `sleep75` loaded below has information on sleep time in minutes for a set of 706 individuals, in addition to demographic and other information [for variable descriptions, load the `wooldridge` package and type `?sleep75` in the console]. Create a single professional-looking table for variables *minutes of sleep at night per week, age, gender, spousal pay, health status*, where the table has the mean of each variable separately by whether or not the person has young kids. Ensure your table has an intuitive title and variable names. Are there any notable differences in these variables?

Table 3: Summary Statistics

Young Kids	No	Yes
Variable	Mean	Mean
Mins Sleep/wk	3269	3251
Age	40	29
Male	0.55	0.7
Spousal Pay	5357	3703
Health Status	0.89	0.91

Surprisingly, there is not much of a difference in the mean amount of sleep per week between people with and without kids: it amounts to 18 minutes. There are some notable differences on other variables, including the fact that people with kids in the sample are much younger, more likely to be male, have low-paid spouses, and be in good health.

```
data <- sleep75

data2 <- select(data, sleep, age, male, spsepay, gdhlth, yngkid) %>%
  mutate(yngkid = as.logical(yngkid))

labs <- data.frame(cbind(names(data2),
                         c("Mins Sleep/wk", "Age", "Male",
                           "Spousal Pay", "Health Status", "Young Kids")))

sumtable(data2, summ=c('mean(x)'),summ.names= c('Mean'), labels =labs,
         group = 'yngkid', logical.labels = c("No", "Yes"),
         title = "Summary Statistics")
```

7. Estimate the relationship between minutes of sleep at night per week and the presence of young kids, controlling for the remaining variables listed in the previous question. Is there any evidence that children affect sleep? When drawing these conclusions, comments on the coefficient estimate and perform an appropriate hypothesis test using robust standard errors.

The results show a negative relationship between having young kids and sleep, but the coefficient is not statistically significant (i.e. we do not reject the null that this slope is zero at conventional significance levels). The coefficient is also quite small, so the effect is not very large. Whether or not this is a causal effect is unclear, since there are many other differences between people with and without kids, and we have no way to control for them.

```
model2 <- lm(sleep ~ age + male + spsepay + gdhlth + yngkid, data = data)
modelsummary(list("OLS" = model2),
```

	OLS
(Intercept)	3260.796***
	(95.825)
age	3.489*
	(1.570)
male	−30.248
	(36.253)
spsepay	0.000
	(0.002)
gdhlth	−132.926*
	(58.640)
yngkid	28.769
	(53.338)
Num.Obs.	706
R2	0.019
F	2.837
RMSE	439.94
Std.Errors	HC1

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
             gof_omit = "IC|Log|Adj|p\\.value|statistic|se_type",
             stars = TRUE, metrics = "all", vcov = "HC1")
```

8. Pretend that you now think the effect of young kids on sleep depends on your age. Alter your model to incorporate this new information, report the results in a professional-looking table, and test whether kids affect sleep using an appropriate hypothesis test with robust standard errors. Comment on the results.

To allow the effect of kids to depend on age, we need to interact the two variables. The results show that the coefficient on the interaction term is positive and statistically insignificant. Whether or not kids affect sleep overall depends on the joint test between the coefficient on the interaction term and the coefficient on the dummy variable for kids. The joint test is also statistically insignificant with the p-value of the F-test being 0.8292, so we do not reject the null that kids have no effect on sleep.

```
model3 <- lm(sleep ~ age + male + spsepay + gdhlth + yngkid + yngkid:age, data = data)
modelsummary(list("OLS1" = model2, "OLS2" = model3),
             gof_omit = "IC|Log|Adj|p\\.value|statistic|se_type",
             stars = TRUE, metrics = "all", vcov = "HC1")
```

	OLS1	OLS2
(Intercept)	3260.796***	3264.339***
	(95.825)	(97.241)
age	3.489*	3.428*
	(1.570)	(1.593)
male	−30.248	−31.057
	(36.253)	(36.322)
spsepay	0.000	0.000
	(0.002)	(0.002)
gdhlth	−132.926*	−133.609*
	(58.640)	(58.817)
yngkid	28.769	−64.441
	(53.338)	(299.731)
age × yngkid		3.153
		(10.057)
Num.Obs.	706	706
R2	0.019	0.019
F	2.837	2.389
RMSE	439.94	439.92
Std.Errors	HC1	HC1

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

```
linearHypothesis(model3, c("yngkid = 0", "age:yngkid = 0"), white.adjust = "hc1")
```

```
Linear hypothesis test

Hypothesis:
yngkid = 0
age:yngkid = 0

Model 1: restricted model
Model 2: sleep ~ age + male + spsepay + gdhlth + yngkid + yngkid:age

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F Pr(>F)
1    701
2    699  2 0.1874 0.8292
```