Problem Set 8 (Answer Key)

Joe Ornstein

Problem 1

Perhaps there's still a backdoor path through education. Countries with a more educated population may be more democratic and less corrupt on average, biasing estimate of the effect of democracy on corruption. I downloaded data on the primary school completion rate by country here.

First, let's load and join the data

```
library(tidyverse)
library(here)
library(countrycode)

d <- read_csv( here('data/week-09/corruption-data.csv') )

d
```

```
# A tibble: 180 x 6
   country     iso3  cpi_score gdp_per_capita polity2 democracy
   <chr>       <chr>     <dbl>          <dbl>   <dbl>     <dbl>
 1 Denmark     DNK          88         62134.      10         1
 2 New Zealand NZL          88         44814.      10         1
 3 Finland     FIN          85         53159.      10         1
 4 Singapore   SGP          85        101649.      -2         0
 5 Sweden      SWE          85         56668.      10         1
 6 Switzerland CHE          85         72372.      10         1
 7 Norway      NOR          84         70006.      10         1
 8 Netherlands NLD          82         61243.      10         1
 9 Germany     DEU          80         57558.      10         1
10 Luxembourg  LUX          80        124569.      10         1
# ... with 170 more rows
```

```r
education <- read_csv( here('data/week-09/API_SE.PRM.CMPT.ZS_DS2_en_csv_v2_4685031.csv'),
                       skip = 4) |>
  select(iso3 = `Country Code`,
         primary_education = `2019`)

education
```

```
# A tibble: 266 x 2
   iso3  primary_education
   <chr>             <dbl>
 1 ABW                NA
 2 AFE                NA
 3 AFG                84.3
 4 AFW                NA
 5 AGO                NA
 6 ALB               103.
 7 AND                NA
 8 ARB                86.5
 9 ARE               112.
10 ARG                98.5
# ... with 256 more rows
```

```r
d <- left_join(d, education, by = 'iso3')

d
```

```
# A tibble: 180 x 7
   country     iso3  cpi_score gdp_per_capita polity2 democracy primary_educat~1
   <chr>       <chr>     <dbl>          <dbl>   <dbl>     <dbl>             <dbl>
 1 Denmark     DNK          88         62134.      10         1             102.
 2 New Zealand NZL          88         44814.      10         1              NA
 3 Finland     FIN          85         53159.      10         1             101.
 4 Singapore   SGP          85        101649.      -2         0              98.1
 5 Sweden      SWE          85         56668.      10         1             105.
 6 Switzerland CHE          85         72372.      10         1              96.4
 7 Norway      NOR          84         70006.      10         1             101.
 8 Netherlands NLD          82         61243.      10         1              NA
 9 Germany     DEU          80         57558.      10         1              99.0
10 Luxembourg  LUX          80        124569.      10         1              82.2
# ... with 170 more rows, and abbreviated variable name 1: primary_education
```

Notice how I used the `skip = 4` option, since the data doesn't start until after the first 4 rows of that spreadsheet.

The estimated effect of democracy on corruption *without* conditioning on education was:

```
lm(cpi_score ~ democracy + gdp_per_capita, data = d)
```

```
Call:
lm(formula = cpi_score ~ democracy + gdp_per_capita, data = d)

Coefficients:
   (Intercept)       democracy  gdp_per_capita
     2.367e+01       8.873e+00       6.238e-04
```

Now, conditioning on education:

```
lm(cpi_score ~ democracy + gdp_per_capita + primary_education, data = d)
```

```
Call:
lm(formula = cpi_score ~ democracy + gdp_per_capita + primary_education,
    data = d)

Coefficients:
      (Intercept)          democracy     gdp_per_capita  primary_education
        21.019857           8.628256           0.000525           0.068171
```

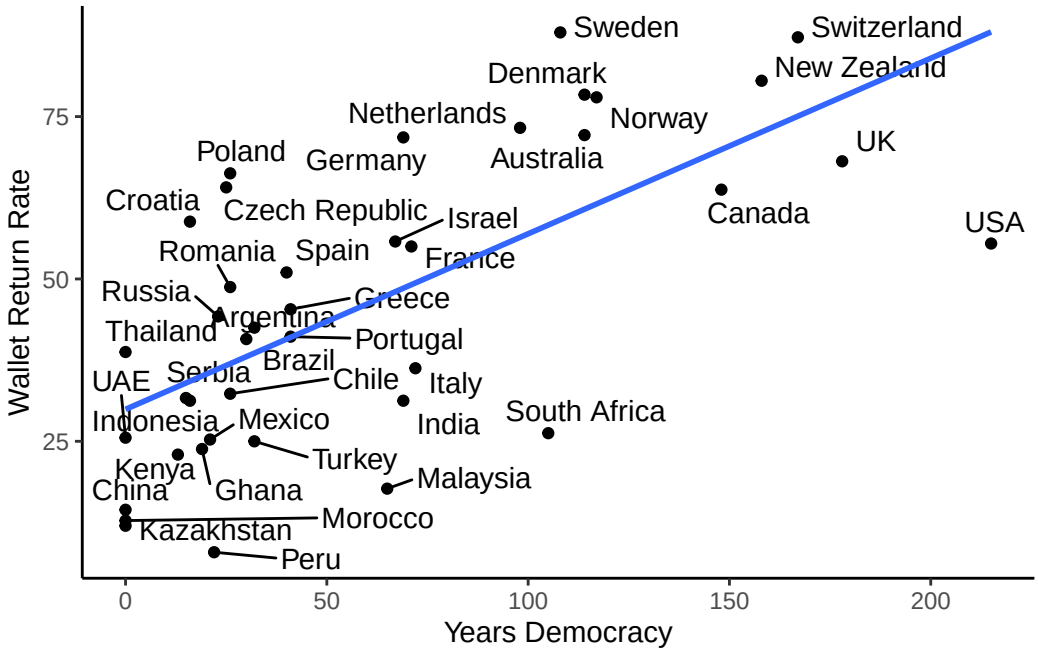
This suggests that (conditional on GDP per capita), education was not confounding the observed relationship between democracy and corruption.

Problem 2

Let's load the wallet data and, keep only the wallets left at public institutions, and summarize each country's wallet return rate compared to its democratic history.

```
cohn <- read_csv( here('data/cohn-2019/behavioral data (csv file).csv'))

d <- cohn |>
  # keep only the wallets left at public institutions
```

```
  filter(public == 1) |>
  group_by(Country) |>
  summarize(wallet_return_rate = mean(response),
            years_democracy = mean(c_PIV_years_democracy))

library(ggrepel)

ggplot(data = d,
       mapping = aes(x = years_democracy,
                     y = wallet_return_rate,
                     label = Country)) +
  geom_text_repel() +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  theme_classic() +
  labs(x = 'Years Democracy',
       y = 'Wallet Return Rate')
```




```
lm(wallet_return_rate ~ years_democracy, data = d)
```

```
Call:
lm(formula = wallet_return_rate ~ years_democracy, data = d)

Coefficients:
    (Intercept)  years_democracy
        29.9250           0.2703
```

For each extra year as a democracy, the average wallet return rate increases roughly 0.27 percentage points. But we can't interpret this as causal without doing additional work! We measured the two variables differently, but there are still a large number of potential back door paths that could be confounding the relationship between democracy and public-sector wallet stealing.

Problem 3

First, compute the wallet return rate in each country by treatment condition.

```
d <- cohn |>
  # keep just the Money and No Money conditions
  filter(cond %in% c(0,1)) |>
  # relabel those conditions as "Money" and "NoMoney"
  mutate(cond = case_when(cond == 1 ~ 'Money',
                          cond == 0 ~ 'NoMoney')) |>
  # compute reporting rate by country and treatment condition
  group_by(Country, cond) |>
  summarize(pct_reported = mean(response)) |>
  ungroup()

d
```

```
# A tibble: 80 x 3
   Country   cond    pct_reported
   <chr>     <chr>          <dbl>
 1 Argentina Money           49
 2 Argentina NoMoney         46
 3 Australia Money           69
 4 Australia NoMoney         52.3
 5 Brazil    Money           48.7
 6 Brazil    NoMoney         34
 7 Canada    Money           63.5
 8 Canada    NoMoney         46.5
```

```
 9 Chile     Money           36.9
10 Chile     NoMoney         35.4
# ... with 70 more rows
```

Next, it's useful to *pivot* the data so that each column represents the return rate for a treatment condition.

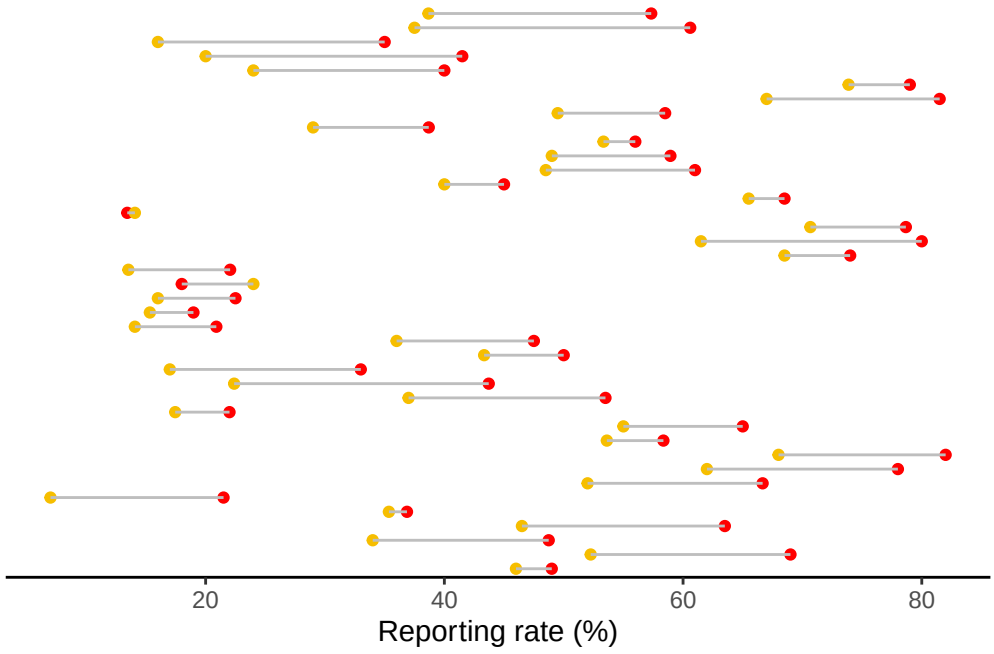
```
d <- d |>
  pivot_wider(names_from = cond,
              values_from = pct_reported)

d
```

```
# A tibble: 40 x 3
   Country        Money NoMoney
   <chr>          <dbl>   <dbl>
 1 Argentina       49      46
 2 Australia       69      52.3
 3 Brazil          48.7    34
 4 Canada          63.5    46.5
 5 Chile           36.9    35.4
 6 China           21.5     7
 7 Croatia         66.7    52
 8 Czech Republic  78      62
 9 Denmark         82      68
10 France          58.4    53.6
# ... with 30 more rows
```

Then plot it, using the `geom_point()` and `geom_segment()` aesthetics.

```
# begin ggplot
ggplot(data = d) +
  # red points for the Money condition
  geom_point(mapping = aes(x=Money, y=Country),
             color = 'red') +
  # yellow points for the NoMoney Condition
  geom_point(mapping = aes(x=NoMoney, y=Country),
             color = '#F6BE00') +
  # a gray line segment in between them
  geom_segment(mapping = aes(x=Money, xend=NoMoney,
                             y=Country, yend=Country),
```

```
              color = 'gray', size = 0.5) +
  labs(x = 'Reporting rate (%)', y = '', color = 'Condition') +
  theme_classic() +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        axis.line.y = element_blank())
```



They're all out of order and they don't have those nice floating labels. We can fix that like so:

```
d <- d |>
  # reorder Country by the NoMoney reporting rate
  mutate(Country = fct_reorder(Country, NoMoney)) |>
  # compute label position, a bit left of the minimum reporting rate
  mutate(label_position =
           pmin(Money, NoMoney) - nchar(as.character(Country))/3.5 - 1)

# then ggplot, adding a geom_text() layer
# begin ggplot
ggplot(data = d) +
  # red points for the Money condition
```

```
geom_point(mapping = aes(x=Money, y=Country),
           color = 'red') +
# yellow points for the NoMoney Condition
geom_point(mapping = aes(x=NoMoney, y=Country),
           color = '#F6BE00') +
# a gray line segment in between them
geom_segment(mapping = aes(x=Money, xend=NoMoney,
                           y=Country, yend=Country),
             color = 'gray', size = 0.5) +
labs(x = 'Reporting rate (%)', y = '', color = 'Condition') +
theme_classic() +
theme(axis.text.y = element_blank(),
      axis.ticks.y = element_blank(),
      axis.line.y = element_blank()) +
geom_text(mapping = aes(x=label_position,
                        y=Country,
                        label = Country),
          size = 2)
```