Human Capital Simulation

Alex Alekseev

```
library(tidyverse)
library(modelsummary)
```

Setup

Let's play around with simulating our own data generating process and looking at what happens when we try to estimate effects naively. Imagine the following DGP, inspired by the human capital theory, that describes the relationship between ability, education, and earnings.

1. Ability is uniformly distributed on [0, 1]
2. Each unit of ability increases your income by 20 thousand euros per year
3. To decide whether you go to college or not, you first flip a coin (heads = go to college, tails = do not go to college). However, if it is tails, you then check your ability. If it is higher than a threshold of 0.5, you go to college anyway.
4. College degree increases your income by 10 thousand euros per year
5. The base income (no college, zero ability) is normally distributed with a mean of 50 thousand euros per year and a s.d. of 20 thousand
6. We observe a sample of 5000 individuals

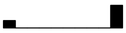
Simulation

First, let's set up the parameters of the DGP.

```
n <- 5000

thresh <- 0.5
educ_prob <- 0.5
ability_eff <- 20
education_eff <- 10
```

Let's simulate some data according to our DGP.

```
set.seed(42)
df <- tibble(
  ability = runif(n)
  , education = 1*(runif(n) <= educ_prob +
                     (1 - educ_prob)*(ability >= thresh))
  , income = rnorm(n, mean = 50, sd = 10) +
    ability_eff*ability +
```

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
ability	5000	0	0.50	0.29	0.00	0.50	1.00	
education	2	0	0.75	0.43	0.00	1.00	1.00	
income	5000	0	67.52	13.36	20.78	67.94	108.60	

	naive
(Intercept)	54.9
	[54.3, 55.6]
education	16.7
	[15.9, 17.4]
Num.Obs.	5000

```
    education_eff*education
)

datasummary_skim(df, fmt = 2)
```

Regression

Let's pretend that we do not know anything about the DGP and try to estimate the college effect.

```
reg1 <- lm(income ~ education, data = df)
msummary(
  list("naive" = reg1)
  , statistic = "[{conf.low}, {conf.high}]"
  , gof_map = "nobs"
  , fmt = 1
)
```

Wow! Our estimate is clearly biased upwards. We know that the true effect of college is 10, but we get 16.67. What is going on?

Let's have a look at the people with a college degree and the people without it. Is the college degree the only difference between them?

```
datasummary(
  (factor(education))*(ability) ~ N + mean + median + sd + min + max
  , data = df
)
```

factor(education)		N	mean	median	sd	min	max
0	ability	1229	0.25	0.25	0.14	0.00	0.50
1	ability	3771	0.59	0.63	0.28	0.00	1.00

	naive	smart
(Intercept)	54.9	50.0
	[54.3, 55.6]	[49.3, 50.6]
education	16.7	9.9
	[15.9, 17.4]	[9.2, 10.7]
ability		20.0
		[18.9, 21.1]
Num.Obs.	5000	5000

Now we can see what is going on. The people with the college degree are, on average, smarter than the people without the college degree. We are not comparing the same people. We need to somehow get rid of the variation in ability to identify the causal effect of college. From our DGP, we know that ability affects both your education and earnings. Education is not exogenous. Simply running a regression on education does not identify the true causal effect, because the estimated coefficient also picks up the variation in ability.

Let's run a regression controlling for ability.

```
reg2 <- lm(income ~ education + ability, data = df)

msummary(
  models = list("naive" = reg1, "smart" = reg2)
  , statistic = "[{conf.low}, {conf.high}]"
  , gof_map = "nobs"
  , fmt = 1
)
```

That is much better! Controlling for ability helped us eliminate the variation that we did not need and focus on the variation that we do need. The estimates are fairly close to the true values.

Just as an exploration, can we identify the effect of ability without controlling for education?

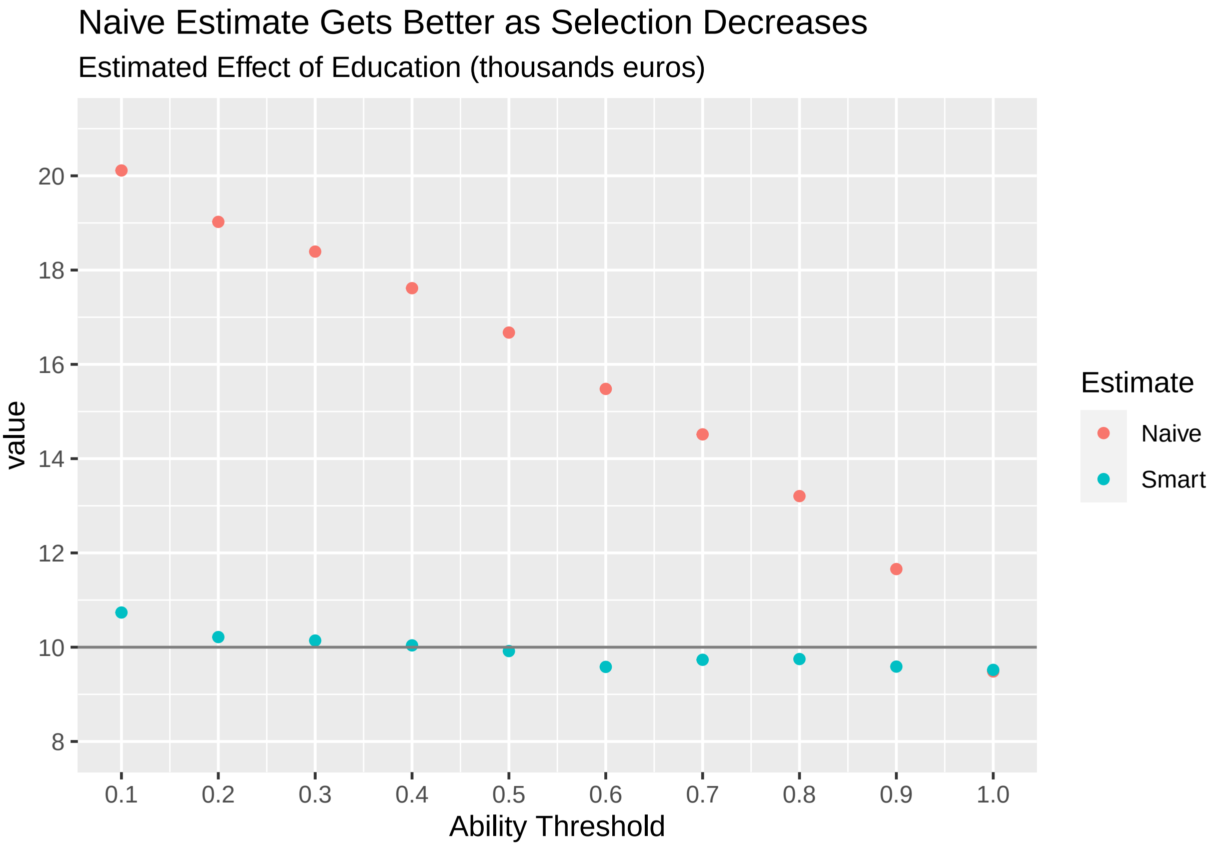
```
reg3 <- lm(income ~ ability, data = df)

msummary(
  models = list("naive" = reg1, "smart" = reg2, "curious" = reg3)
  , statistic = "[{conf.low}, {conf.high}]"
  , gof_map = "nobs"
  , fmt = 1
)
```

No, because part of the effect of ability goes through education. Unless we control for education, we do not isolate the causal effect of ability on earnings.

We can explore how the selection problem, as proxied by the ability threshold, affects our naive estimate of the college effect.

	naive	smart	curious
(Intercept)	54.9	50.0	53.8
	[54.3, 55.6]	[49.3, 50.6]	[53.2, 54.4]
education	16.7	9.9	
	[15.9, 17.4]	[9.2, 10.7]	
ability		20.0	27.3
		[18.9, 21.1]	[26.3, 28.4]
Num.Obs.	5000	5000	5000

```r
effect_est <- function(thresh) {

  set.seed(42)
  df <- tibble(
    ability = runif(n)
    , education = 1*(runif(n) <= educ_prob +
                       (1 - educ_prob)*(ability >= thresh))
    , income = rnorm(n, mean = 50, sd = 10) +
      ability_eff*ability + education_eff*education
  )

  reg1 <- lm(income ~ education, data = df)
  reg2 <- lm(income ~ education + ability, data = df)

  res <- list("naive" = coef(reg1)[2], "smart" = coef(reg2)[2])

  return(res)

}

tibble(
  thresh = seq(0.1, 1, 0.1)
) %>%
  mutate(
    Naive = map_dbl(
      thresh
      , ~{effect_est(.x) %>% .[["naive"]]}
    )
    , Smart = map_dbl(
      thresh
      , ~{effect_est(.x) %>% .[["smart"]]}
    )
  ) %>%
  pivot_longer(cols = c(Naive, Smart), names_to = "Estimate") %>%
  ggplot(aes(thresh, value, color = Estimate)) +
  geom_point() +
  geom_hline(yintercept = education_eff, color = "gray50") +
  scale_x_continuous(n.breaks = 10) +
  scale_y_continuous(limits = c(8, 21), n.breaks = 8) +
  labs(
    title = "Naive Estimate Gets Better as Selection Decreases"
    , subtitle = "Estimated Effect of Education (thousands euros)"
    , x = "Ability Threshold"
  )
```



A