# Data Generating Process, Variation, and Identification

AECN 396/896-002

# Before we start

## Learning objectives

Understand

- what data generating process is
- variation in a variable
- identification of the impact of a variable

## Table of contents

1. Data Generating Process and Clean Variations to Use
2. Identification

## Reference

The contents of this lecture borrow heavily from "The Effect" by Nick Huntington-Klein (book available for free here).

"Huntington-Klein, N. (2021). The effect: An introduction to research design and causality. Chapman and Hall/CRC."

# Data Generating Process and Clean Variations

# Data Generating Process

**Definition**

The set of underlying laws that determine how the data we observed is created

**Features**

- We cannot see them directly (at least for economic phenomenon)
- But, we get to observe data generated from it

# Example (Non-economic)

$$F = \frac{G \times m_1 \times m_2}{r^2}$$

- $G$: gravitational constant
- $m_1$: mass of object 1
- $m_2$: mass of object 2
- $r$: distance between the two objects
- $F$: force pullinf the two objects together
- This is the physical law (data generating process) that governs how an object (say a ball) fall to the ground when it is let go of your hand (ignoring wind). The observation that an object has fallen is data.

**Key**

We did not know the data generating process until Newton discovers it. By looking at the data, he learned that the underlying process has to be the one above. We are trying to do the same.

# A toy example

### Data Generating Process

- Income is log-normally distributed
- Being brown-haired gives you a 10% income boost
- 20% of people are naturally brown-haired
- Having a college degree gives you a 20% income boost
- 30% of people have college degrees
- 40% of people who don't have brown hair or a college degree will choose to dye their hair brown

### Research Goal

You are interested in learning the impact of having brown-hair on income.

## The data generating process in code

```r
set.seed(89403)

N <- 10000 #* number of observations

data <-
  tibble(
    brown_haired = runif(N) < 0.2, # 1 if naturally brown haired
    college = runif(N) < 0.3, # 1 if have college degrees
    error = 0.1 * rnorm(N), #* error term
  ) %>%
  mutate(
    dye_to_brown = runif(N) < 0.4, # whether to dye hair to brown or not
    brown_haired = ifelse(
      dye_to_brown == TRUE & college == FALSE,
      TRUE,
      brown_haired
    )
  ) %>%
  mutate(
    income = exp(0.1 * brown_haired + 0.2 * college + error)
  )
```

## Your Model

$$log(income) = \alpha + \beta \text{brown-haired} + v$$

## Interpretation

$\beta$ represents the percentage difference in income between brown-hared and non-brown-haired people (baseline is the non-brown-haired people).

## Diagnostics of the zero conditional mean assumption

$E[v|\text{brown-haired}] = 0$?

## Direction of bias?

- Correlation between `brown-haired` and `college`?
- The sign of the impact of `college`?
- So, the bias is (negative or positive)?

## Naive model

```
feols(log(income) ~ i(brown_haired), data = data) %>% tidy()
```

```
## # A tibble: 2 × 5
##   term               estimate std.error statistic  p.value
##   <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)          0.0831   0.00176      47.3 0
## 2 brown_haired::TRUE   0.0497   0.00272      18.3 1.75e-73
```

Okay, as we expected, we are severely underestimating the impact of `brown_haired`.

## Question

What should we do to get the estimation right?

## Sensible model

```
feols(log(income) ~ i(brown_haired) + i(college), data = data) %>% tidy()
```

```
## # A tibble: 3 × 5
##   term                estimate std.error statistic p.value
##   <chr>                  <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)         -0.00201   0.00163     -1.23   0.217
## 2 brown_haired::TRUE   0.104     0.00213     49.1    0
## 3 college::TRUE        0.201     0.00227     88.6    0
```

Okay, as we expected, we are good now.

But, let's think of another way to recover a good estimate of the impact of being `brown-haired` using the information we have about the data generating process while still using the naive model of just regressing `log(income)` on `brown-haired`. Any idea?

- Income is log-normally distributed
- Being brown-haired gives you a 10% income boost ( pretend you do not know this, as this is the objective )
- 20% of people are naturally brown-haired
- Having a college degree gives you a 20% income boost
- 30% of people have college degrees
- 40% of people who don't have brown hair or a college degree will choose to dye their hair brown

## Solution

Notice that those with college degrees do not dye their hair to brown. So, if we just use the observations for those people, we can cleanly identify the impact of `brown-haired`.

```
feols(
  log(income) ~ i(brown_haired),
  data = filter(data, college == TRUE)
) %>%
  tidy()
```

```
## # A tibble: 2 × 5
##   term               estimate std.error statistic   p.value
##   <chr>                 <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)           0.199   0.00202      98.4 0
## 2 brown_haired::TRUE    0.104   0.00448      23.1 1.84e-109
```

### Variation (Definition)

How a variable changes from observation to observation

### clean and bad variations

- Bad variations: variations in `brown_haired` for the entire sample

`brown_haired` is correlated with `college`, which made us confound (mix) the impact of `brown_haired` and `college`, when the naive model is used.

- Clean variations: variations in `brown_haired` only for the samples with college degrees

No ones with college degrees do not dye their to brown. So, if we just focus on (limit ourselves to) those people, variations in `brown_haired` is not correlated with `college`. So, we were able to estimate the impact of `brown_haired` even with the simple model.

### Note

- This is just a toy example and we could have just included `college` as a covariate.
- But, this is just to get you start thinking about different types of variations there are in the dataset.
- There are "clean" and "dirty" variations. Limiting ourselves to only the "clean" variation is very important.

# Key message through this toy example

### Message 1

By understanding the data generating process, we know why we cannot trust the naive estimation of the impact of `brown_haired` on `income`.

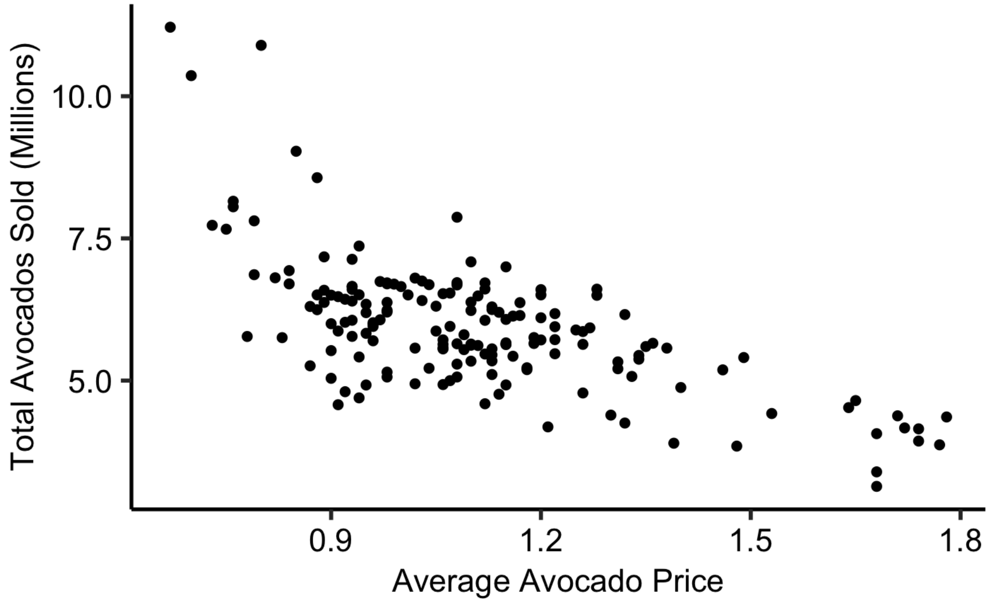
### Message 2

We make use of our knowledge about a part of the data generating process to identify the imapact of `brown_haired` credibly:

"40% of people who don't have brown hair or a college degree will choose to dye their hair brown"

Of course, in real world applications, we almost always do not have such a clean and crucial information. But, knowing the context of your study lets you make credible "assumptions" that will let you find clean "variations" that we can harness to estimated the impact of our variable of interest credibly.

# A further example of looking for clean variations



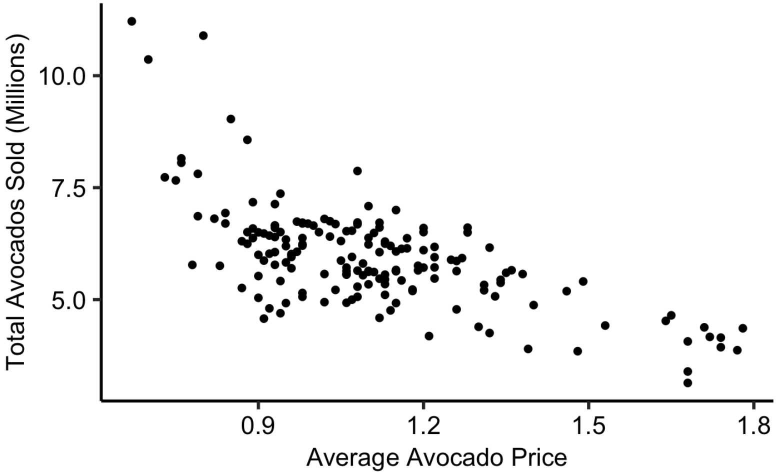
Data from Hass Avocado Board
c/o https://www.kaggle.com/neuromusic/avocado-prices/

Weekly Sales of Avocados in California, Jan 2015 - Match 2018

You are interested in understanding the impact of avocado price on its consumption.

## Questions

Can you answer the research question from this figure?

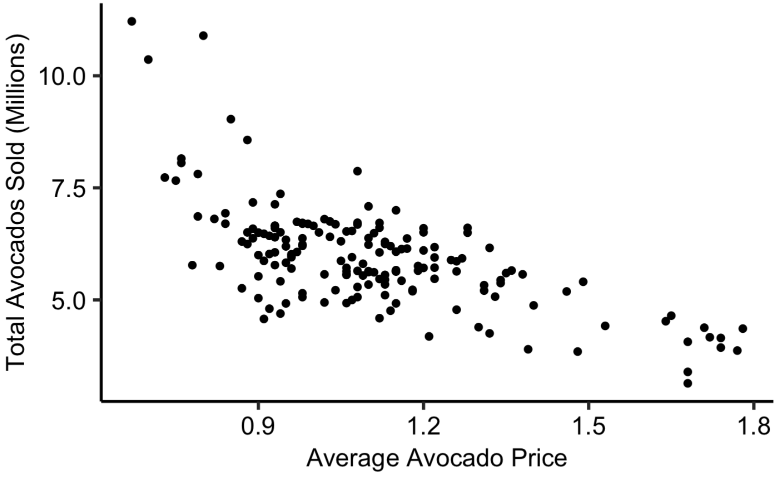


Data from Hass Avocado Board
c/o https://www.kaggle.com/neuromusic/avocado-prices/

- They are negatively associated with each other
    - Avocado sales tend to be lower in weeks where the price of avocados is high.
    - Prices tend to be higher in weeks where fewer avocados are sold
- You cannot make a causal statement like this:

"An increase in avocado price make consumers buy less avocado."

- Reverse causality
    - price affects demand
    - demand affects price



Data from Hass Avocado Board
c/o https://www.kaggle.com/neuromusic/avocado-prices/

### Problem

Reverse Causality: Price affects demand and demand affects price.

### Question

Suppose your can run an experiment on the avocado market (ideal situation). If we want to identify the impact of price on demand free of confusing with the impact of demand on price, what would you do?

### Special contextual knowledge

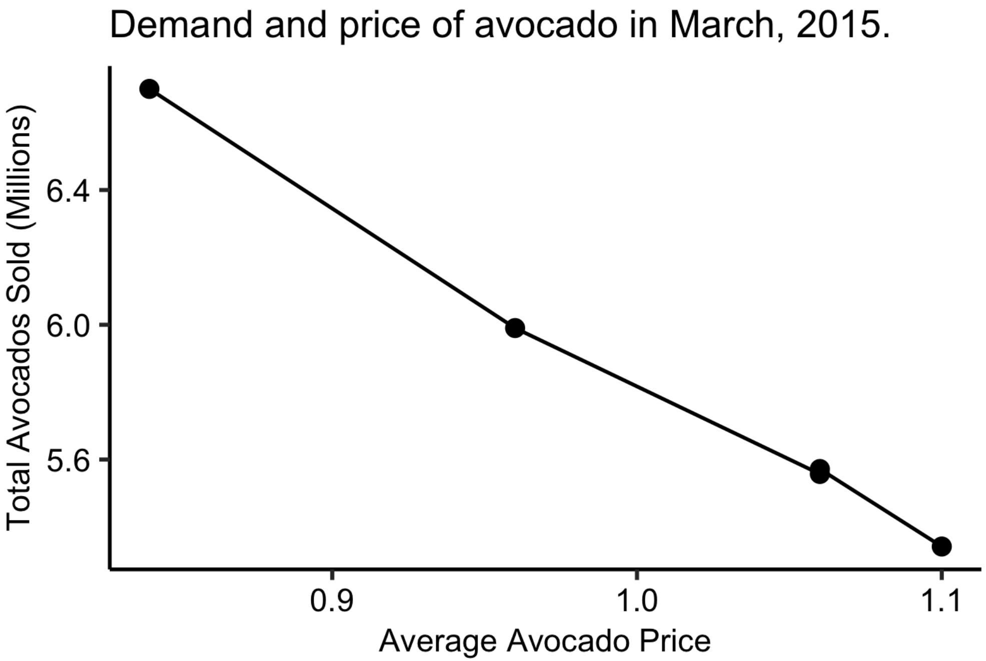
Now, suppose you learned the following fact after studying the supply and purchasing mechanism on the avocado market:

At the beginning of each month, avocado suppliers make a plan for what avocado prices will be each week in that month, and never change their plans until the next month.

This means that within the same month changes in avocado price every week is not a function of how much avocado has been bought in the previous weeks, effectively breaking the causal effect of demand on price.

So, our estimation strategy would be to just look at the variations in demand and price within individual months, but ignore variations in price between months.

An example of clean variations in price and its impact on demand.


Demand and price of avocado in March, 2015.

Data from Hass Avocado Board
c/o https://www.kaggle.com/neuromusic/avocado-prices/

We will talk about how we can use only the within-month variations in avocado price, but leave out the between-month variations in avocado price econometrically using R.

# Key message through this example

**Message 1**

By understanding the data generating process (knowing how any economic market works), we recognize the problem of simply looking at the relationship between the avocado price and demand to conclude the causal impact of price on demand (reverse causality).

**Message 2**

We study the context very well and how the avocado market works in California (of course it is not really how CA avocado market works in reality) and make use of the information to identify the "clean" variations in avocado price to identify its impact on demand.