BUS41100 Applied Regression Analysis

Week 5: Causal Inferece

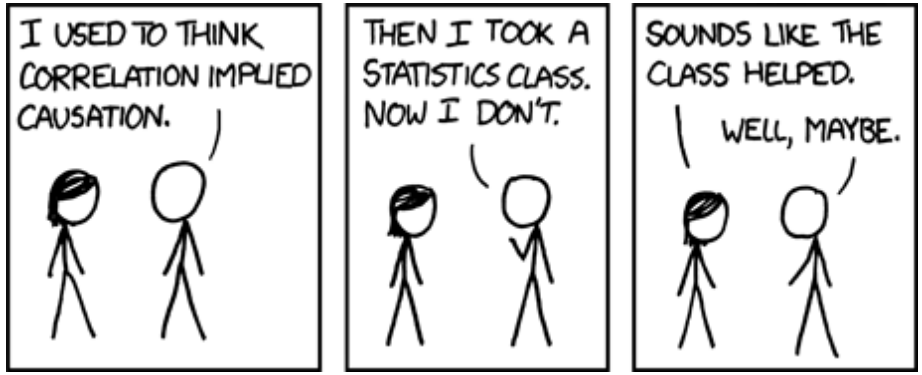
Randomized Experiments, Causation in Observational Data

Max H. Farrell
The University of Chicago Booth School of Business

Causality

When does

correlation $\Rightarrow$ causation?



- We have been careful to never say that X *causes* Y ...
- ... but we've really wanted to.
- We want to find a "real" underlying mechanism:

  What's the change *in* Y *as* T *moves*

  independent of **all** *other influences?*

But how can we do this in regression?

- First we'll look at the Gold Standard: experiments
- Watch out for multiple testing
- Then see how this works in regression

Randomized Experiments

We want to know the effect of treatment T on outcome Y

What's the problem with "regular" data? **Selection.**

- People choose their treatments
  - Eg: *(i)* Firm investment & tax laws; *(ii)* people & training/education; *(iii)* ....

Experiments are the **best** way to find a true causal effect.

Why? The key is **randomization**:

- No systematic relationship between units and treatments
  - T moves independently *by design*.
- T is discrete, usually binary.
  - Classic: drug vs. placebo
  - Newer: Website experience (A/B testing)
- Experiments are important (& common) in their own right

The fundamental question: Is Y better on average with T?

$$\mathbb{E}[Y \mid T = 1] \; > \; \mathbb{E}[Y \mid T = 0] \; \textbf{?}$$

We need a model for $\mathbb{E}[Y \mid T]$

- T is just a special X variable:

$$\mathbb{E}[Y \mid T] = \beta_0 + \beta_T T$$

- β_T is the Average Treatment Effect (ATE)
- This is not a **prediction** problem, . . .
- . . . it's an **inference** problem, about a single coefficient.

Estimation:

$$b_T = \hat{\beta}_T = \bar{Y}_{T=1} - \bar{Y}_{T=0}$$

Can't usually do better than this. (Be wary of any claims.)

Why do we care about the average Y?

First, we might care about Y directly, for an individual unit:

- Does $Y = \texttt{earnings}$ increase after $T = \texttt{training}$?
  - e.g. does getting an MBA increase earnings?
- Do firms benefit from consulting?
- Do people live longer with a medication/procedure?
- Do people stay longer on my website with the new design?

Or, we might care about aggregate measures:

- $Y = \texttt{purchase}$ $\texttt{yes/no}$, then profit is $P = \texttt{price} \times Y$
  - Average profit per customer: $\mathbb{E}[P \times Y]$
  - Total profit: (No. customers)$\times \mathbb{E}[P \times Y]$
- Higher price means fewer customers, but perhaps more profit overall? *(Ignore Giffen goods)*

Profit Maximization

Data from an online recruiting service

- Customers are **firms** looking to hire
- Fixed price is charged for access
  - Post job openings, find candidates, etc

Question is: what price to charge?

Profit at price P = *Quantity*(P) $\times$ (P - *Cost*)

Arriving customers are shown a random price P

- P is our treatment variable T
- How to randomize matters:
  - Why not do: P_1 in June, P_2 in July, . . . ? What's wrong?

Data set includes

- $P =$ `price` – price they were shown, \$99 or \$249
- $Y =$ `buy` – did this firm sign up for service: yes/no

Let's see the data

```
> price.data <- read.csv("priceExperiment.csv")
> summary(price.data)
> head(price.data)
```

Note that $Y = \text{buy}$ is binary. That's okay!

$$\mathbb{E}[Y] = \mathbb{P}[Y = 1]$$

Computing the ATE and Profit:

```
> purchases <- by(price.data$buy, price.data$price, mean)
> purchases[2] - purchases[1]
-0.1291639
> 249*purchases[2] - 99*purchases[1]
4.311221
```

-0.13 what? 4.31 what? For whom? How many?

Regression version: computing ATE

```
> summary(lm(price.data$buy ~ price.data$price))

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.3284017  0.0195456  16.802   <2e-16 ***
price.data$price -0.0008611  0.0001039  -8.287   <2e-16 ***
```

careful with how you code the variables!

```
> summary(lm(price.data$buy ~ (price.data$price==249)))

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   0.24315    0.01091  22.285   <2e-16 ***
price.data$price == 249TRUE -0.12916    0.01559  -8.287   <2e-16 ***
```

What's so special about $T = 0/1$?

Regression version: computing profit

```
> profit <- price.data$buy*price.data$price
> summary(lm(profit ~ (price.data$price==249)))

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                   24.072      1.820  13.226   <2e-16 ***
price.data$price == 249TRUE    4.311      2.600   1.658   0.0974 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.18 on 2361 degrees of freedom
Multiple R-squared:  0.001163,  Adjusted R-squared:  0.0007402
F-statistic:  2.75 on 1 and 2361 DF,  p-value: 0.09741
```

- Same profit estimate, thanks to transformed Y variable
- Tiny R^2! Why?
- What's 24.072?

What about variables other than Y and T?

We usually have information (some X's) other than Y and T

- Key: when was the information recorded?
- Useful other X variables are "pre-treatment":
  not affected by treatment or even treatment assignment
- Useful for targeting, heterogeneity (see homework)

Important idea:

Randomized means randomized for every value of X

```
> table(price.data$customerSize)

   0    1    2
1897  216  250
```

$\Rightarrow$ Nothing wrong with

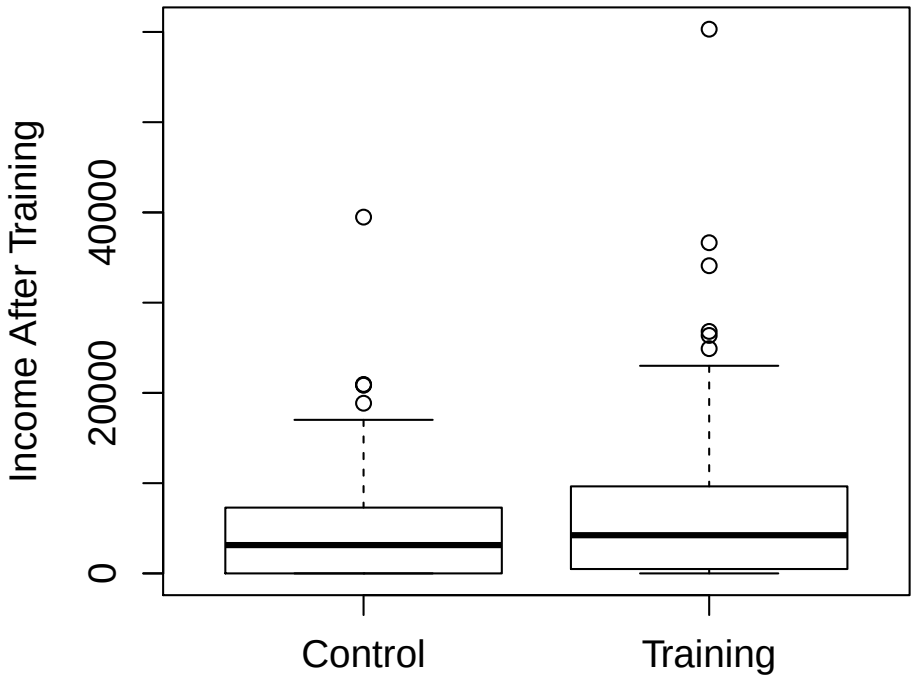
```
> summary(lm(buy~(price==249),data=price.data[price.data$customerSize==2,]))
```

Heterogeneous Treatment Effects

Example: Job Training Program & Income

Un(der)employed men were randomized: 185 received job training, 260 didn't

```
> nsw <- read.csv("nsw.csv")
> nsw.outcomes <- by(nsw$income.after, nsw$treat, mean)
> nsw.outcomes[2] - nsw.outcomes[1]
1794.342
```



- Outliers?
- Bunching?
- Income = 0?

Today, we'll ignore these problems.

What do we learn from the regression output?

```
> summary(nsw.reg <- lm(nsw$income.after ~ nsw$treat))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4554.8      408.0  11.162  < 2e-16 ***
nsw$treat     1794.3      632.9   2.835  0.00479 **

Residual standard error: 6580 on 443 degrees of freedom
Multiple R-squared:  0.01782,   Adjusted R-squared:  0.01561
F-statistic: 8.039 on 1 and 443 DF,  p-value: 0.004788
```

- Training increases earnings
- Tiny R^2! Why?

How does the TE vary over $\mathbf{X}$? Useful for targeting

$$\mathbb{E}[Y \mid T = 1, \texttt{HSDegree}] - \mathbb{E}[Y \mid T = 0, \texttt{HSDegree}] = \textbf{?}$$

```
> summary(lm(income.after ~ treat, data=nsw[nsw$hsdegree==1,]))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     4854       1132   4.287 4.35e-05 ***
treat           3192       1518   2.103   0.0381 *
```

Subpopulation	n	ATE	p-value
Black	371	2029	0.004
Not Black	74	803	0.549
Hispanic	39	793	0.708
Not Hispanic	406	1960	0.003
Married	75	3709	0.017
Unmarried	370	1373	0.049
HS Degree	97	3192	0.038
No High School	348	1154	0.098
Black + Unmarried	307	1548	0.046
Black + No HS	293	1129	0.139
Unmarried + No HS	292	795	0.308
Black, Unmarried, No HS	244	644	0.448

Watch out for multiple testing and multicollinearity

Using **regression**

$\mathbb{E}[Y \mid T, \texttt{HSDegree}]$ is just $\mathbb{E}[Y \mid T, X]$: just a MLR

First try: $\mathbb{E}[Y \mid T, \texttt{HSDegree}] = \beta_0 + \beta_T T + \beta_1 \texttt{HSDegree}$

- $\mathbb{E}[Y \mid T = 1, \texttt{HSDegree}] - \mathbb{E}[Y \mid T = 0, \texttt{HSDegree}] = \beta_T$

```
> summary(lm(income.after ~ treat + hsdegree, data=nsw))

             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4321.5      426.0  10.144   <2e-16 ***
treat         1615.9      638.5   2.531   0.0117 *
hsdegree      1410.4      762.1   1.851   0.0649 .
```

Not the same! Why?

- Interpret conditional on the model
- Same β_T for `Not black`, why?

Recall dummy variables: different intercepts, same slope.

A better/correct model includes interactions:

$$\mathbb{E}[Y \mid T, \texttt{black}] = \beta_0 + \beta_T T + \beta_1 \texttt{HSDegree} + \beta_2 \big(T \times \texttt{HSDegree}\big)$$

$$\Rightarrow \quad \mathbb{E}[Y \mid T = 1, \texttt{HSDegree}] - \mathbb{E}[Y \mid T = 0, \texttt{HSDegree}] = \beta_T + \beta_2$$

```
> summary(lm(income.after ~ treat*black, data=nsw))

               Estimate Std. Error t value Pr(>|t|)
(Intercept)      4495.4      445.0  10.101   <2e-16 ***
treat            1154.0      725.3   1.591    0.112
hsdegree          359.1     1094.3   0.328    0.743
treat:hsdegree   2038.0     1523.6   1.338    0.182
```

- $\beta_T + \beta_2 = 3192$
- Nothing is significant, but F-test: `0.004354`

Continuous variables are fine too!

```
> summary(lm(income.after ~ treat*education, data=nsw))
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3803.01    2568.91   1.480   0.1395
treat           -4585.18    3601.53  -1.273   0.2036
education          74.52     251.45   0.296   0.7671
treat:education   614.77     347.28   1.770   0.0774 .
```

But now we're making a **functional form** assumption!

- This isn't a "subset" at all
- We are assuming the model for $\mathbb{E}[Y \mid T = 1, \texttt{education}]$ is linear in education. Why?
- See `week5-Rcode.R` for $\texttt{education}^2$. What does that mean?

Causality Without Randomization

We want to find:

> *The change in* Y **caused by** T *moving* **independently** *of all other influences.*

Our MLR interpretation of $\mathbb{E}[Y \mid T, X]$:

> *The change in* Y **associated with** T*, holding fixed all X variables.*

$\Rightarrow$ We need T to be randomly assigned given X

- X must include enough variables so T is random.
  - Requires a lot of knowledge!
- No systematic relationship between units and treatments, conditional on X.
  - It's OK if X is predictive of Y.

The model is the same as always:

$$\mathbb{E}[Y \mid T, X] = \beta_0 + \beta_T T + \beta_1 X_1 + \cdots \beta_d X_d.$$

But the assumptions change:

- This is a *structural* model: it says something true about the real world.
- Need X to control for **all** sources of non-randomness.
  - Even possible?

Then the interpretation changes:

$$\beta_T \textit{ is the average treatment effect}$$

- Continuous "treatments" are easy.
- Not a "conditional average treatment effect"
  - What happens to β_T as the variables change? To b_T?
- No $T \times X$ interactions, why? What would these mean?

Example: Bike Sharing & Weather: does a change in humidity **cause** a change in bike rentals?

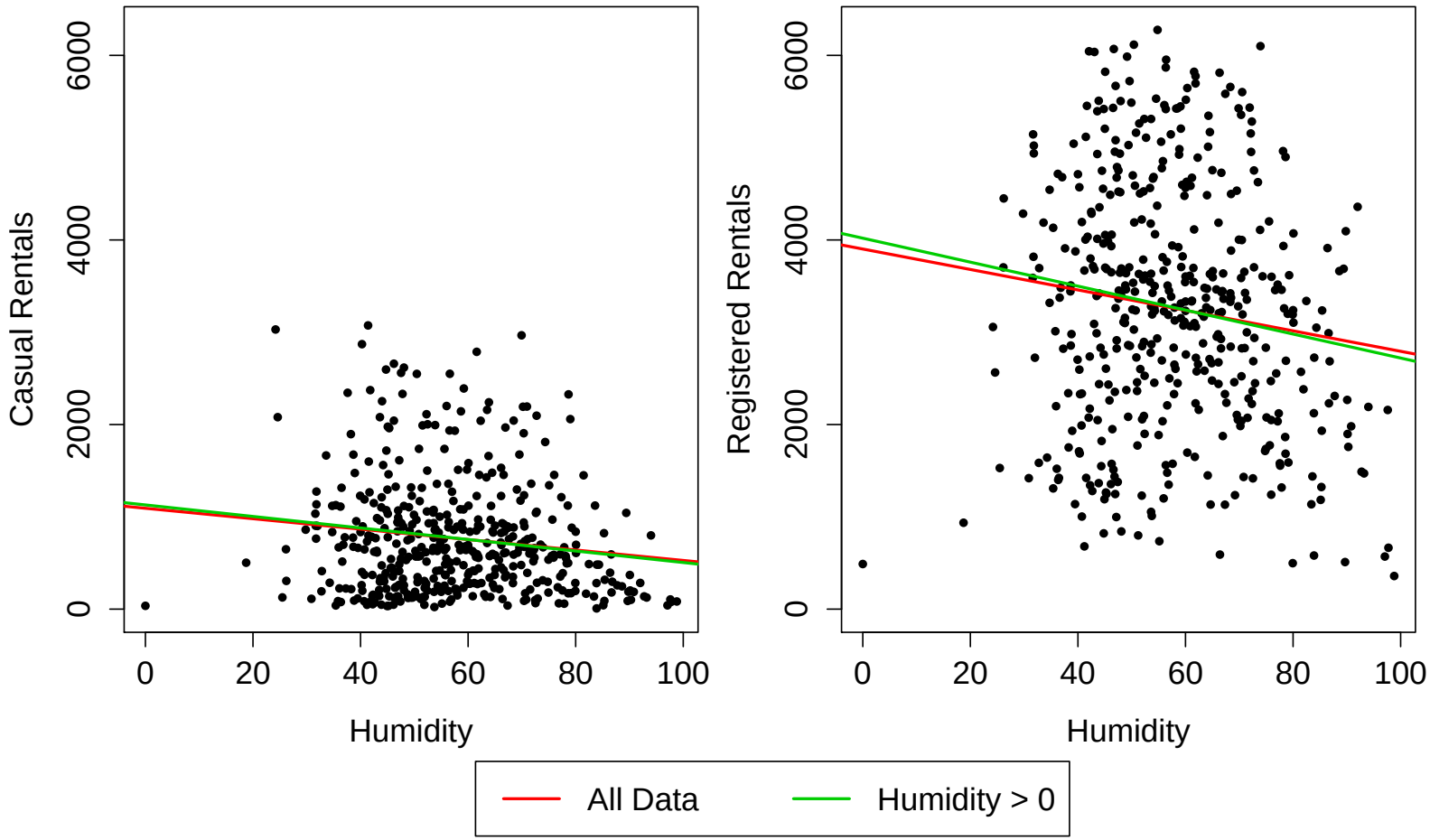
From Capital Bikeshare (D.C.'s Divvy) we have daily bike rentals & weather info.

- $Y_1 =$ `registered` – # rentals by registered users
- $Y_2 =$ `casual` – # rentals by non-registered users
- $T =$ `humidity` – relative humidity (*continuous!*)

Possible controls/confounders:

- `season`
- `holiday` – Is the day a holiday?
- `workingday` – Is it a work day (not holiday, not weekend)?
- `weather` – coded 1=nice, 2=OK, 3=bad
- `temp` – degrees Celsius
- `feels.like` – "feels like" in Celsius
- `windspeed`

Is humidity randomly assigned to days?



humidity $\uparrow \Rightarrow$ rentals $\downarrow$!

Or is this because of something else?

The "randomized experiment" coefficient

```
> summary(casual.reg <- lm(casual ~ humidity, data=bike))

            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1092.719    114.116   9.576  < 2e-16 ***
humidity      -5.652      1.912  -2.957  0.00327 **
```

. . . is pretty similar to the effect with controls. So what?

```
> summary(casual.reg.main <- lm(casual ~ humidity + season + holiday +
 workingday + weather + temp + windspeed, data=bike))

             Estimate Std. Error t value Pr(>|t|)
(Intercept)   716.964    203.273   3.527 0.000464 ***
humidity       -6.845      1.496  -4.574  6.2e-06 ***
seasonspring  -94.041     82.189  -1.144 0.253152
seasonsummer  182.964     53.249   3.436 0.000646 ***
seasonwinter   57.194     68.849   0.831 0.406578
holiday      -285.327    103.757  -2.750 0.006203 **
workingday   -796.933     37.381 -21.319  < 2e-16 ***
weathernice   308.495    100.633   3.066 0.002305 **
weatherok     264.843     92.695   2.857 0.004475 **
temp           39.430      4.045   9.747  < 2e-16 ***
windspeed     -10.912      3.071  -3.554 0.000420 ***
```

The bottom line:

You only get causal effects with **strong** assumptions.

- Real-world concerns take precedence over statistics.
- Is there an economic/business/etc justification for your choice of X?
- The data/computer cannot help. Only assumptions.

Causal inference from observational data may be the hardest problem in statistics.

- We are just scratching the surface in terms of ideas, methods, applications, ...
- Still an active area of research in econometrics, statistics, & machine learning.