MS&E 226: Fundamentals of Data Science

Lecture 15: Additional topics in causal inference

Ramesh Johari

Regression analysis of experiments

Regression analysis of an experiment

Recall using OLS to fit the following model:

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_1 W_i,$$

where $W_i \in \{0, 1\}$ is the assignment in a randomized experiment (0 is control, 1 is treatment), and Y_i is the corresponding observed outcome for individual i.
As we showed:

- $\hat{\beta}_0$ is the average outcome in the control group.
- $\hat{\beta}_0 + \hat{\beta}_1$ is the average outcome in the treatment group.
- $\hat{\beta}_1 = \widehat{\mathsf{ATE}}$.

Going further

In this lecture we consider what happens when we have additional covariates we can exploit in our analysis.

Suppose in addition to $Y(0), Y(1)$, and W, each individual also has a vector of observed covariates $\vec{X}$.

Going further

In this lecture we consider what happens when we have additional covariates we can exploit in our analysis.

Suppose in addition to $Y(0), Y(1)$, and W, each individual also has a vector of observed covariates $\vec{X}$.

There are two ways in which the regression approach to experimental analysis is powerful:

Going further

In this lecture we consider what happens when we have additional covariates we can exploit in our analysis.

Suppose in addition to $Y(0), Y(1)$, and W, each individual also has a vector of observed covariates $\vec{X}$.

There are two ways in which the regression approach to experimental analysis is powerful:

- Controlling for observed covariates helps improve estimation of the ATE.

Going further

In this lecture we consider what happens when we have additional covariates we can exploit in our analysis.

Suppose in addition to $Y(0), Y(1)$, and W, each individual also has a vector of observed covariates $\vec{X}$.

There are two ways in which the regression approach to experimental analysis is powerful:

- Controlling for observed covariates helps improve estimation of the ATE.
- Interactions with the treatment effect allow us to see how the treatment effect varies among individuals with different covariate vectors.

Going further

In this lecture we consider what happens when we have additional covariates we can exploit in our analysis.

Suppose in addition to $Y(0), Y(1)$, and W, each individual also has a vector of observed covariates $\vec{X}$.

There are two ways in which the regression approach to experimental analysis is powerful:

- Controlling for observed covariates helps improve estimation of the ATE.
- Interactions with the treatment effect allow us to see how the treatment effect varies among individuals with different covariate vectors.

Warning: The covariates $\vec{X}$ must be observed *pre-treatment*!

Controlling for observables: An example

I created a synthetic experiment where $n_0 = n_1 = 150$.

For each individual i, $X_i \sim \mathcal{N}(0, 1)$ is a pre-existing covariate, and W_i is the treatment indicator.

I contructed Y_i as:

$$Y_i = 10 + 0.5 \times W_i + X_i + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, 1)$.

In this example:

- The true ATE is 0.5—it does not vary depending on X.
- However, some of the variation in Y_i's is explained the X's as well.

Controlling for observables: An example

Suppose we regress Y on the treatment indicator W alone:

```
Call:
lm(formula = Y ~ 1 + W, data = df)

...

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.9807     0.1168   85.45  < 2e-16 ***
W1            0.4608     0.1652    2.79  0.00561 **
```

Controlling for observables: An example

Now suppose we include the covariate X in the regression:

```
Call:
lm(formula = Y ~ 1 + W + X, data = df)

...

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.91498    0.08688 114.123  < 2e-16 ***
W1           0.61827    0.12314   5.021 8.88e-07 ***
X            1.01032    0.06483  15.584  < 2e-16 ***
```

Notice that the standard error is smaller on the coefficient of W.

Controlling for observables: Interpretation

In the specification `Y ~ 1 + W + X`, we still interpret the coefficient on W as an estimate of the population-level ATE.

The point is that adding X to the regression gives us a better estimate of the baseline $Y(0)$ for each individual.

Essentially, this regression says that for an individual with covariate X:

- $Y(0) \approx \hat{\beta}_0 + \hat{\beta}_2 X$.
- $Y(1) \approx \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 X$.

Controlling for observables

Controlling for observed covariates has another effect as well:

If the randomization was less than perfect, controlling for observed covariates can reduce the sampling bias.

How this works:

- Suppose, e.g., individuals with higher X were more likely to receive the treatment.

Controlling for observables

Controlling for observed covariates has another effect as well:

If the randomization was less than perfect, controlling for observed covariates can reduce the sampling bias.

How this works:

- Suppose, e.g., individuals with higher X were more likely to receive the treatment.
- Ignoring this fact will lead to a biased estimate of the ATE: part of the variation in the observed Y's is explained by variation in the X's, *not* by the variation in the treatment. (This is an omitted variable bias.)

Controlling for observables

Controlling for observed covariates has another effect as well:

If the randomization was less than perfect, controlling for observed covariates can reduce the sampling bias.

How this works:

- Suppose, e.g., individuals with higher X were more likely to receive the treatment.
- Ignoring this fact will lead to a biased estimate of the ATE: part of the variation in the observed Y's is explained by variation in the X's, *not* by the variation in the treatment. (This is an omitted variable bias.)
- Controlling for X removes the omitted variable bias.

Controlling for observables

Controlling for observed covariates has another effect as well:

If the randomization was less than perfect, controlling for observed covariates can reduce the sampling bias.

How this works:

- Suppose, e.g., individuals with higher X were more likely to receive the treatment.
- Ignoring this fact will lead to a biased estimate of the ATE: part of the variation in the observed Y's is explained by variation in the X's, *not* by the variation in the treatment. (This is an omitted variable bias.)
- Controlling for X removes the omitted variable bias.

What are the limitations to this process?

Interactions

The preceding slides suggest one limitation of merely controlling for observed covariates:

What if the treatment effect itself varies depending on the covariates observed?

To address this issue we employ interactions with the treatment indicator.

Interactions

Suppose given a covariate X, we add the interaction term $W \times X$ to the model:

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_W W_i + \hat{\beta}_X X_i + \hat{\beta}_{WX} W_i X_i.$$

With the addition of this term we can interpret the model as follows:

Interactions

Suppose given a covariate X, we add the interaction term $W \times X$ to the model:

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_W W_i + \hat{\beta}_X X_i + \hat{\beta}_{WX} W_i X_i.$$

With the addition of this term we can interpret the model as follows:

For an individual with covariate X,

Interactions

Suppose given a covariate X, we add the interaction term $W \times X$ to the model:

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_W W_i + \hat{\beta}_X X_i + \hat{\beta}_{WX} W_i X_i.$$

With the addition of this term we can interpret the model as follows:

For an individual with covariate X,

- $Y(0) \approx \hat{\beta}_0 + \hat{\beta}_X X$.

Interactions

Suppose given a covariate X, we add the interaction term $W \times X$ to the model:

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_W W_i + \hat{\beta}_X X_i + \hat{\beta}_{WX} W_i X_i.$$

With the addition of this term we can interpret the model as follows:

For an individual with covariate X,

- $Y(0) \approx \hat{\beta}_0 + \hat{\beta}_X X.$
- $Y(1) \approx \hat{\beta}_0 + \hat{\beta}_W + (\hat{\beta}_X + \hat{\beta}_{WX})X.$

Interactions

Suppose given a covariate X, we add the interaction term $W \times X$ to the model:

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_W W_i + \hat{\beta}_X X_i + \hat{\beta}_{WX} W_i X_i.$$

With the addition of this term we can interpret the model as follows:

For an individual with covariate X,

- $Y(0) \approx \hat{\beta}_0 + \hat{\beta}_X X$.
- $Y(1) \approx \hat{\beta}_0 + \hat{\beta}_W + (\hat{\beta}_X + \hat{\beta}_{WX})X$.
- The estimated causal effect is $\approx \hat{\beta}_W + \hat{\beta}_{WX} X$.

Interactions

Suppose given a covariate X, we add the interaction term $W \times X$ to the model:

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_W W_i + \hat{\beta}_X X_i + \hat{\beta}_{WX} W_i X_i.$$

With the addition of this term we can interpret the model as follows:

For an individual with covariate X,

- $Y(0) \approx \hat{\beta}_0 + \hat{\beta}_X X$.
- $Y(1) \approx \hat{\beta}_0 + \hat{\beta}_W + (\hat{\beta}_X + \hat{\beta}_{WX})X$.
- The estimated causal effect is $\approx \hat{\beta}_W + \hat{\beta}_{WX} X$.

This allows us to measure *heterogeneous treatment effects* across the population.

Interactions: Example

In the earlier example, there should be no meaningful change in the treatment effect across individuals with different X's.

```
Call:
lm(formula = Y ~ 1 + W + X + X * W, data = df)

...

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.91161    0.08705 113.860  < 2e-16 ***
W1           0.61730    0.12323   5.009  9.4e-07 ***
X            1.06204    0.09365  11.340  < 2e-16 ***
W1:X        -0.09945    0.12986  -0.766    0.444
```

Interactions: Example

Now suppose we change the model so that in the population, changing X also changes the treatment effect.

In particular, suppose:

$$Y_i = 10 + (0.5 + X_i)W_i + X_i + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, 1)$.

What happens when we estimate a model with interactions on the resulting experimental data?

Interactions: Example

The result:

```
Call:
lm(formula = Y ~ 1 + W + X + X * W, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2.32757 -0.73146  0.05078  0.62216  2.85012

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.02695    0.08510 117.827  < 2e-16 ***
W1           0.51447    0.12047   4.271 2.63e-05 ***
X            1.06476    0.09155  11.630  < 2e-16 ***
W1:X         0.86899    0.12695   6.845 4.40e-11 ***
```

SUTVA and interference

Interference

Implicitly throughout our discussion of causal inference, we have assumed there is no *interference* between treatment and control:

Whether or not individual i receives treatment or control has *no impact* on the causal effect of treatment on another individual j.

When might this fail?

Interference

Suppose Airbnb decides to A/B test a new feature that dramatically simplifies the booking process for a guest.

In the test, guests are randomized at when they start the booking process; control is the old experience, treatment is the new experience.

It is found that customers with the new experience book much more frequently than customers with the old experience, but the estimated $\widehat{\text{ATE}}$ is an *overestimate*. Why?

Interference

Both treatment and control see the *same* inventory of host listings!

So if treatment individuals book more often, that *reduces* the inventory available to control individuals, and implies their booking rates will be lower.

SUTVA

If interference is present, the "potential outcomes" for an individual are much more complicated: they depend on not just the treatment a single individual received, but also on the treatment *other* individuals received.

With n individuals, this is 2^n potential outcomes for each individual!

SUTVA

If interference is present, the "potential outcomes" for an individual are much more complicated: they depend on not just the treatment a single individual received, but also on the treatment *other* individuals received.

With n individuals, this is 2^n potential outcomes for each individual!

The assumption that there is no interference between treatment and control is part of the *stable unit treatment value assumpton* (SUTVA) in econometrics and causal inference.

The other part of SUTVA is that there is only one form of treatment or control: e.g., if treatment is "taking a drug", there should be no variation in the treatment group as to *how much* of the drug is taken.

A paradox

A puzzle

A new vaccine for the flu is introduced, and compared against no vaccination (control).

Let $W = 0, 1$ denote control or treatment, respectively.

Let $Y = 0, 1$ denote the outcome of flu infection or no flu infection, respectively.

Let Z denote whether the individual is an adult (A) or child (C).

A puzzle

You run an experiment with a large sample size, and equal numbers of adults and children.

	Adults		Children	
	No Flu ($Y = 1$)	Flu ($Y = 0$)	No Flu ($Y = 1$)	Flu ($Y = 0$)
Treatment ($W = 1$)	0.1500	0.2250	0.1000	0.0250
Control ($W = 0$)	0.0375	0.0875	0.2625	0.1125

(Here the numbers are the fractions of individuals in each category.)

A puzzle

Analyzing the results:

- On average, $\mathbb{P}(Y = 1|W = 1) = 0.5$ while $\mathbb{P}(Y = 1|W = 0) = 0.6$, so the vaccine appears detrimental.

A puzzle

Analyzing the results:

- On average, $\mathbb{P}(Y = 1|W = 1) = 0.5$ while $\mathbb{P}(Y = 1|W = 0) = 0.6$, so the vaccine appears detrimental.
- On the other hand, $\mathbb{P}(Y = 1|W = 1, Z = A) = 0.4$, while $\mathbb{P}(Y = 1|W = 0, Z = A) = 0.3$, so the vaccine appears to be beneficial to adults.

A puzzle

Analyzing the results:

- On average, $\mathbb{P}(Y = 1|W = 1) = 0.5$ while $\mathbb{P}(Y = 1|W = 0) = 0.6$, so the vaccine appears detrimental.
- On the other hand, $\mathbb{P}(Y = 1|W = 1, Z = A) = 0.4$, while $\mathbb{P}(Y = 1|W = 0, Z = A) = 0.3$, so the vaccine appears to be beneficial to adults.
- In addition, $\mathbb{P}(Y = 1|W = 1, Z = C) = 0.8$, while $\mathbb{P}(Y = 1|W = 0, Z = C) = 0.7$, so the vaccine appears to also be beneficial to children as well!

What happened? (This is called *Simpson's paradox*.)

Potential outcomes, causal effects, and sampling bias

Each adult and child has two potential outcomes $Y(0)$ and $Y(1)$, associated to control and treatment, respectively.

If we presume there was no sampling bias among adults, (so W is uncorrelated with Y given $Z = A$) then the average causal effect among adults is:

$$\begin{aligned}
&\mathbb{E}[Y(1) - Y(0)|Z = A] \\
&\quad = \mathbb{E}[Y(1)|Z = A, W = 1] - \mathbb{E}[Y(0)|Z = A, W = 0] \\
&\quad = \mathbb{P}(Y = 1|Z = A, W = 1) - \mathbb{P}(Y = 1|Z = A, W = 0) \\
&\quad = 0.1
\end{aligned}$$

Potential outcomes, causal effects, and sampling bias

Each adult and child has two potential outcomes $Y(0)$ and $Y(1)$, associated to control and treatment, respectively.

If we presume there was no sampling bias among adults, (so W is uncorrelated with Y given $Z = A$) then the average causal effect among adults is:

$$\begin{aligned}
&\mathbb{E}[Y(1) - Y(0)|Z = A] \\
&\quad = \mathbb{E}[Y(1)|Z = A, W = 1] - \mathbb{E}[Y(0)|Z = A, W = 0] \\
&\quad = \mathbb{P}(Y = 1|Z = A, W = 1) - \mathbb{P}(Y = 1|Z = A, W = 0) \\
&\quad = 0.1
\end{aligned}$$

Similarly the average causal effect among children is $\mathbb{E}[Y(1) - Y(0)|Z = C] = 0.1$.

Potential outcomes, causal effects, and sampling bias

So what is the average causal effect overall?

$$\begin{aligned}
&\mathbb{E}[Y(1) - Y(0)] \\
&\quad = \mathbb{E}[Y(1) - Y(0) | Z = A]\mathbb{P}(Z = A) \\
&\qquad + \mathbb{E}[Y(1) - Y(0) | Z = C]\mathbb{P}(Z = C) \\
&\quad = 0.1
\end{aligned}$$

So there is no paradox: if the causal effect for adults and children is separately positive, it must be positive overall.

Potential outcomes, causal effects, and sampling bias

The issue is that in this example:

$$\mathbb{E}[Y(1) - Y(0)] \neq \mathbb{E}[Y(1)|W = 1] - \mathbb{E}[Y(0)|W = 0].$$

The reason is that if we *ignore* age, there *is* a sampling bias:

- *Children* are more likely to be in control than treatment; adults are more likely to be in treatment than control.
- And children have *higher* potential outcomes on average than adults: the average outcome of a child in treatment (resp. control) is 0.8 (resp., 0.7), while the same for an adult in treatment is 0.4 (resp., 0.3).
- This combination of effects lowers the average outcome in the treatment group relative to the overall population (since the treatment group is primarily adults), and raises the average outcome in the control group relative to the overall population (since the control group is primarily children).

Potential outcomes, causal effects, and sampling bias

The preceding analysis shows that ignoring age creates an *omitted variable bias* in our estimate of the average treatment effect.

Note that we assumed no further sampling bias beyond age; the example makes clear that any such bias would only further cloud the true causal effect.

Another example: Berkeley admissions

Berkeley was sued for gender bias in admissions based on 1973 statistics: 44% of men were admitted, while only 35% of women were admitted.

But based on individual departments' admissions statistics, there did not appear to be statistically significant gender-based discrimination (in fact if anything, some departments tended to *favor* women).

What happened is that there was a sampling bias: women were systematically applying to majors that were much more competitive.

The moral

This example is meant to illustrate how to use potential outcomes to carefully describe the causal effect of interest.

Perfect randomization makes up for a lot of deficiencies, but sometimes things are less than perfect.

Taking care to think through potential outcomes and sampling bias carefully can help avoid incorrect inference!

Observational data

Natural experiments

How can we make causal inferences *without* randomized experiments?

As the preceding lecture shows, we need to find other ways to eliminate sampling bias.

The phrase "natural experiment" refers to the fact that we look for structure in the data we are given that "mimics" an experiment we would have wanted to conduct.

Examples

Some examples include:

- Regression discontinuity analysis
- Propensity score matching
- Instrumental variables

Regression discontinuity analysis

Example: Suppose a scholarship is awarded to all students who score above a threshold t. What is the value of the scholarship, in terms of recipients' educational and professional outcomes?

Problem: Receiving the scholarship is based on merit, which means students receiving scholarships are more successful partly due to innate merit alone.

Solution: Compare students just *above* the threshold t to those just *below* the threshold t. Conditional on scoring near t, the presumption is that students are similar to each other. For this group, it is as if assignment to the scholarship is *random*.

Generalizability: Do you believe the causal effect estimated this way applies to the entire population of students?

Matching and propensity scores

Matching refers to the process of finding a "match" for each observation in treatment with one in control (i.e., an observation that is similar in every respect except for assignment to treatment).

Propensity score matching refers to a process where individuals are "matched" on the basis of their predicted probability of being assigned treatment. (The resulting matching is checked to make sure covariates are balanced across treatment and control.)

What are the benefits of these approaches? What are some potential pitfalls?

Instrumental variables

In econometrics, a common technique to causal inference from observational data is the use of *instrumental variables*.

The problem: When selection bias is present, then assignment W is *correlated* with the outcome Y.

The solution: "Instruments" are choices of variables Z that are presumed to have an effect on assignment W, but are otherwise independent of the outcome Y. (Key assumption!)

In this case, can use variation of outcome Y with instrument Z as a proxy for exogenous variation of the treatment assignment W.

Example: Encouragement to use a new service (e.g., Carta)