Experiments

Motivation: Which customers should xBox advertise to?

Hypothetical Summary data:

Segment	Share of Sales	Click-through rate	Conversion rate
Men	90%	0.10	0.001
Women	10%	0.01	0.0005

Which segment should we target with our advertising? Does this tell us?

Motivation: Which customers should xBox advertise to?

Hypothetical detailed data for fitting predictive model:

Clicked on Ad	Age	Past browsing behiavor	Gender
0	17	...	M
0	37	...	M
1	25	...	F
0	19	...	M
...	...	...	...

Which segment should we target with our advertising? Does this tell us?

Motivation: Which customers should xBox advertise to?

- Assume consumer the decision process is unknown to marketers:
  - Men are influenced by
    - What their friends recommend
    - Their friends make recommendations randomly
  - Women are influenced by:
    - Reviews of the product
    - Price of the product
    - Ads
    - Their friends
- Should xBox advertise to men or women?
  - Typical data will, shown in the previous slides, will not answer this question.
  - Only an experiment can answer it for certain.

Predictive and Causal Analytics

- Predictive analytics
  - Finds associations between variables that naturally occur in the world
  - Machine learning is a sophisticated form of predictive analytics
    - It ensures that the associations are robust by using regularization, bagging, boosting.
- Causal analytics
  - Answer that questions: what will happen if I change X?
  - Core question to marketers. What happens if I change:
    - Advertising
    - Prices
    - Promotions

Why not run experiments?

Why not run experiments?

- Direct Costs: have to set them up and execute them.
  - May require special code
  - May require additional labor to implement
  - Advertising platforms may not have infrastructure or willingness to implement experiments

Why not run experiments?

- Direct Costs: have to set them up and execute them.
  - May require special code
  - May require additional labor to implement
  - Advertising platforms may not have infrastructure or willingness to implement experiments
- Opportunity Cost:
  - The policy currently in place might be effective
  - The people in the control group would not experience this policy and may be less inclined to buy as a result
  - The people in the control group may have a bad experience, never return to the product and tell everyone they know about their bad experience

Why not run experiments?

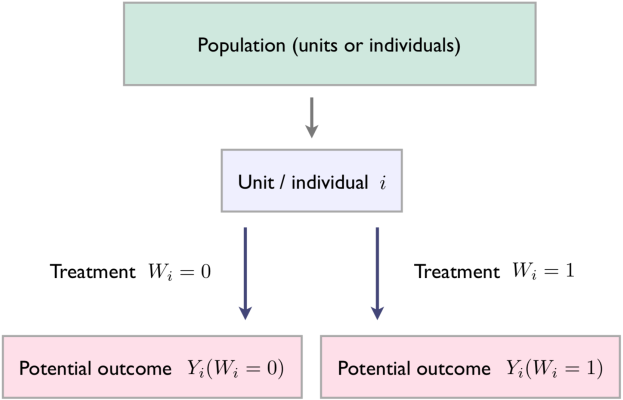
- Direct Costs: have to set them up and execute them.
  - May require special code
  - May require additional labor to implement
  - Advertising platforms may not have infrastructure or willingness to implement experiments
- Opportunity Cost:
  - The policy currently in place might be effective
  - The people in the control group would not experience this policy and may be less inclined to buy as a result
  - The people in the control group may have a bad experience, never return to the product and tell everyone they know about their bad experience
- Managers Not Convinced
  - Managers who have not used them before may be skeptical about their benefit.
  - Revealing the truth does not benefit everyone. What if a top manager's pet project is shown completely ineffective?

Experiments: How do they work?

Experiments: How do they work?

- How do experiments recover causality?
  - Randomization
  - This generates different populations of people who are expected to be identical
  - If the populations are expected to be identical, but we treat one and not the other then we know the treatment is the only difference between the populations.
- Names of experiments
  - Randomized Control Trials (RCT)
  - A/B testing
  - Field experiments

Formally, what do we mean by causal effects?

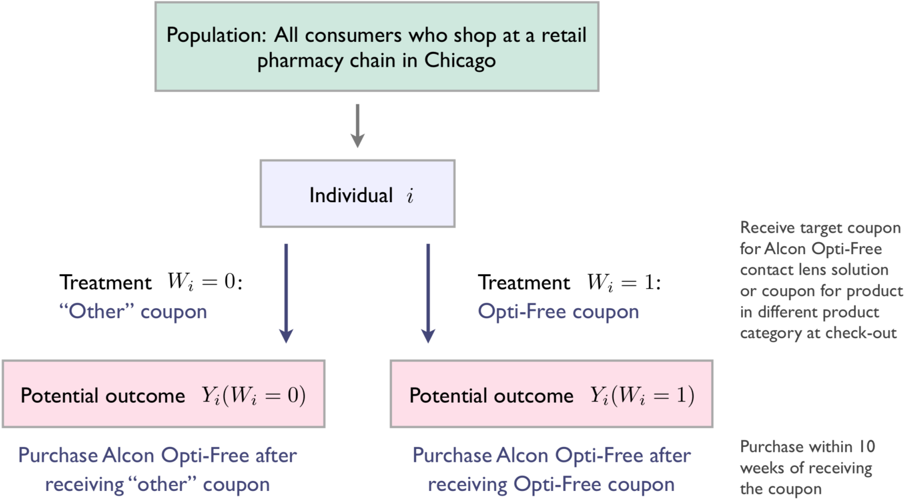


Populations are groups of units (consumers, patients, stores, markets, ...)

Two treatment levels: 0,1

Can be generalized to multiple treatments

Example: Receiving a target coupon and purchase rate



Causal effect (treatment effect)

- **Causal effect (treatment effect)** of the treatment W_i on individual i:
  - Difference in potential outcomes $= Y_i(W_i = 1) - Y_i(W_i = 0)$
  - Potential outcomes framework is a model of parallel worlds, where the treatment is the only (!) difference across the two worlds
  - Examples:
    - ⋆ Difference in purchase rate when receiving vs. not receiving Opti-Free target coupon

Potential outcomes vs. actually observed data

- In the data we only observe the *realized outcome* for each individual
  - Only observe one, never both of the following outcomes:

$$Y_i^{obs} = \begin{cases} Y_i(W_i = 0) & \text{if } W_i = 0 \text{ Purchase if individual did not receive coupor} \\ Y_i(W_i = 1) & \text{if } W_i = 1 \text{ Purchase if individual did receiv the coupon} \end{cases}$$

- This means we cannot directly estimate the individual causal effect (treatment effect)
  - One of the potential outcomes to estimate the causal effect is not in the data

$$Y_i(W_i = 1) - Y_i(W_i = 0)$$

Average treatment effect

- Instead of trying to estimate individual treatment effects, maybe we can estimate the **average treatment effect** in the population:

$$\text{ATE}=\mathbb{E}[Y_i(W_i=1)-Y_i(W_i=0)]$$

- Examples:
  - Difference in purchase rate when receiving vs. not receiving Opti-Free target coupon

Estimating the average treatment effect

- Remember what data we observe:

$$Y_i^{obs} = \begin{cases} Y_i(W_i = 0) & \text{if } W_i = 0 \\ Y_i(W_i = 1) & \text{if } W_i = 1 \end{cases}$$

- Can we estimate the ATE using the observed data?
  - ATE$=\mathbb{E}[Y_i(W_i = 1) - Y_i(W_i = 0)]$

Confounds

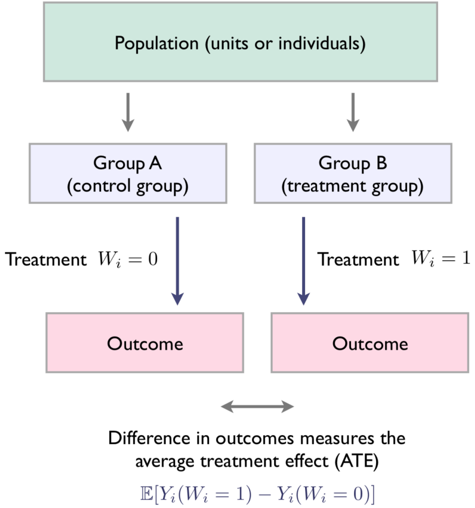
- Confounds — "*correlation does not imply causation*"
  - Individuals who exercise vigorously may generally be healthier (allowing them to exercise vigorously) than those who exercise moderately
  - Suppose Alcon Opti-Free coupon is targeted to individuals who buy the product at the current purchase occasion. Then individuals with a coupon for Alcon Opti-Free will generally be more likely to buy in future than those who receive a different coupon.
- In general, $\tau = \mathbb{E}[Y_i(W_i = 1) - Y_i(W_i = 0)]$ is not the average treatment effect
- We never observe both parallel worlds, and hence *cannot generally estimate the average treatment* effect from observed data

Randomized controlled trials (RCT's) — A/B testing

Random assignment to groups

Intervention (treatment)

- Exercise regime
- Target coupon



Average causal effect

Intuition

- Random assignment creates *two replicas of the same population of units*
- *Only difference* between the two populations is the treatment assignment — the experimental *manipulation*
- Implies that the difference in outcomes must be *caused* by the treatment

Successful design of an experiment (RCT's or A/B testing)

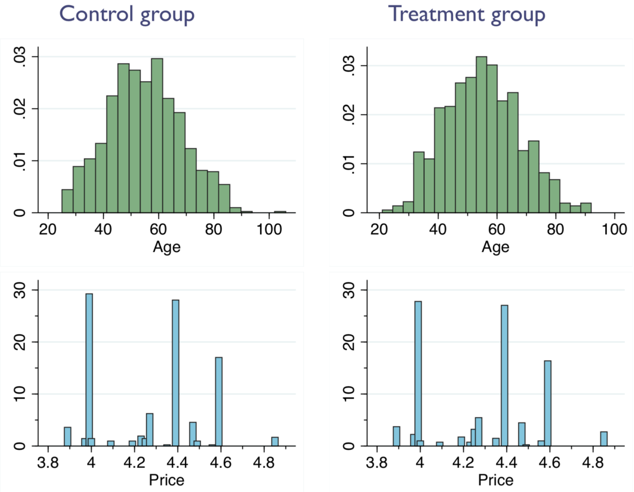
- Randomization → control and treatment groups are replicas of the population

Subject characteristics

- Demographics (age, income, ...)
- Location

Marketing environment

- Prices
- Shelf layout in stores, ...



Estimating the average treatment effect using regression

- Assume:
  - The treatment indicator, $W_i = 0, 1$ is a dummy variable.
  - W_i was assigned randomly for each i
- Analysis:
  - To obtain the causal estimate, τ, can estimate:

$$Y_i = \beta_0 + \tau W_i + \varepsilon_i$$

  - No other covariates are required. The random assignment of W_i prevents omitted variable bias.

Estimating the average treatment effect using regression

- Assume:
  - The treatment indicator, $W_i = 0, 1$ is a dummy variable.
  - W_i was assigned randomly for each i
- Analysis:
  - To obtain the causal estimate, τ, can estimate:

    $$Y_i = \beta_0 + \tau W_i + \varepsilon_i$$

  - No other covariates are required. The random assignment of W_i prevents omitted variable bias.
  - However, covariates may still be useful. We may prefer to estimate τ using:

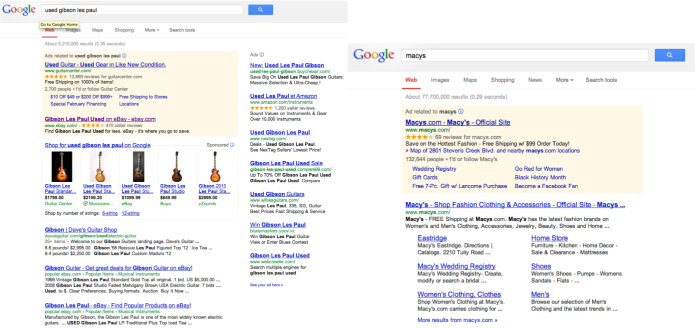
    $$Y_i = \beta_0 + \tau W_i + (\beta_1 X_{i1} + \ldots + \beta_k X_{ik}) + \varepsilon_i$$

  - In some cases the additional covariates $(X_{i1} \ldots X_{ik})$ may help the *precision* of our estimate for τ.
    - i.e., the covariates may reduce the standard error on τ

eBay RCT: Paid Search Effectiveness

- Paper by Blake, Nosko and Tadelis, Econometrica (2015)
- "Search Engine Marketing" (SEM)
  - Google searches produce
    - Organic results
    - Sponsored results based on keywords
- "Branded Search"—includes own company name
- "Non-branded search"—doesn't include company name

Figure 1: Google Ad Examples



(a) Used Gibson Les Paul (b) Macys

eBay RCT: Do clicks tell us about how effective SEM is?

- Endogeneity problem:

eBay RCT: Do clicks tell us about how effective SEM is?

- Endogeneity problem:
  - People who searched for eBay are probably going there anyway
  - What value does the ad banner for eBay add?
  - How do we test whether it adds value?
- Solution: Experiment

eBay RCT: Experiment Set up

- Google advertising allows vendors to bid on ads by Designated Marketing Area (DMA)
  - Unlike FB, Google cannot track individual users searches
  - Randomization by DMA, not user—limits statistical power

eBay RCT: Experiment Set up

- Google advertising allows vendors to bid on ads by Designated Marketing Area (DMA)
  - Unlike FB, Google cannot track individual users searches
  - Randomization by DMA, not user—limits statistical power
- eBay had concerns about lost sales if the ads were effective
  - only 30% of DMAs could have ads turned off.

eBay RCT: Experiment Set up

- Google advertising allows vendors to bid on ads by Designated Marketing Area (DMA)
    - Unlike FB, Google cannot track individual users searches
    - Randomization by DMA, not user—limits statistical power
- eBay had concerns about lost sales if the ads were effective
    - only 30% of DMAs could have ads turned off.
- DMA selection for RCT (on "non-branded" search, excluding "ebay" from the keywords)
    - Control and test groups generated with algorithm so they had similar historic time trends
    - Selection of the Test/Control groups was not random!
    - Used difference-in-differences analysis to control for size differences and changes in sales over time
    - Similar historic time trends suggests that the "parallel trends" assumption is valid.

eBay RCT: Experiment Set up

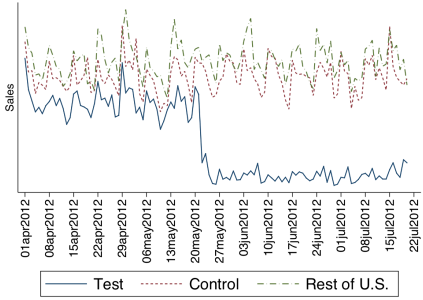
- Google advertising allows vendors to bid on ads by Designated Marketing Area (DMA)
  - Unlike FB, Google cannot track individual users searches
  - Randomization by DMA, not user—limits statistical power
- eBay had concerns about lost sales if the ads were effective
  - only 30% of DMAs could have ads turned off.
- DMA selection for RCT (on "non-branded" search, excluding "ebay" from the keywords)
  - Control and test groups generated with algorithm so they had similar historic time trends
  - Selection of the Test/Control groups was not random!
  - Used difference-in-differences analysis to control for size differences and changes in sales over time
  - Similar historic time trends suggests that the "parallel trends" assumption is valid.
- DMA groups:
  - Test group: 68 DMAs where SEM stopped
  - Control group: 65 DMAs with similar historic time trends to the Test group
  - Rest of U.S: Remaining DMAs

eBay RCT: Experimental Conditions

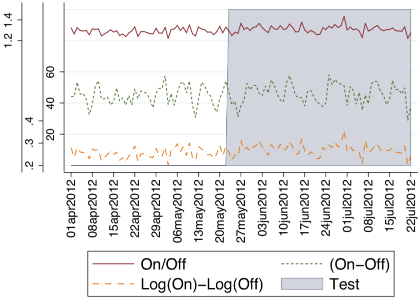
- Left Panel: Total Sales by users that clicked on an eBay ad
  - Test group should be zero, but there are errors in determining user location

eBay RCT: Experimental Conditions

- Left Panel: Total Sales by users that clicked on an eBay ad
  - Test group should be zero, but there are errors in determining user location
- Right Panel: Daily Average Sales: difference between control region and test region: simple difference, ratio and log difference.



(a) Attributed Sales by Region



(b) Differences in Total Sales

eBay RCT: How to analyze these data?

(1) Naive specification:

$$\log(\text{Sales}_{it}) = \beta_0 + \beta_1 \log(\text{AdSpend}_{it}) + \varepsilon_{it}$$

(2) Better controls: Panel data allows controls for time and location effects:

$$\log(\text{Sales}_{it}) = \beta_0 + \beta_1 \log(\text{AdSpend}_{it}) + \gamma_i + \alpha_t + \varepsilon_{it}$$

(3) Experimental variation (AdsOn_{it} is a dummy for being in a DMA that continued receiving ads during the test period):

$$\log(\text{Sales}_{it}) = \beta_0 + \beta_1 \text{AdsOn}_{it} + \gamma_i + \alpha_t + \varepsilon_{it}$$

AdsOn_{it} indicates that the DMA-day is a DMA that continued receiving advertising and the day is during the test period.

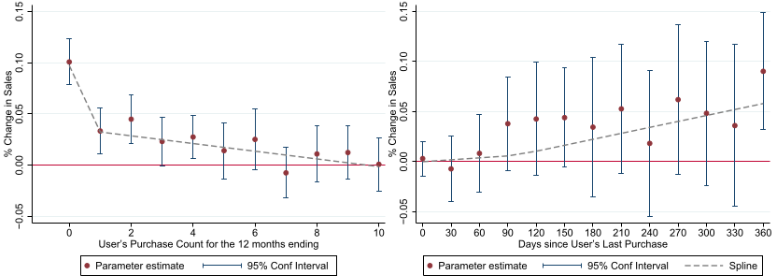
eBay RCT: Estimates

	(1)	(2)	(3)
Estimated Coefficient	0.885	0.126	0.0066
(Std Err)	(0.0143)	(0.040)	(0.0056)
DMA Fixed Effects		Yes	Yes
Date Fixed Effects		Yes	Yes
N	10500	10500	23730
ROI (Implied)	4173%	1632%	-63%

eBay RCT: Are SEM effective for any users?

Infrequent users are most influenced by SEM. Let m index user segments:

$$\log(\text{Sales}_{it}) = \beta_0 + \beta_1 \text{AdsOn}_{it} \times \theta_m + \gamma_i + \alpha_t + \varepsilon_{it}$$



(a) User Frequency (b) User Recency

Summary

- Potential outcomes model ("parallel worlds")
  - Causality: Difference across potential outcomes
- Measuring causality – average treatment effect
  - Randomized control trials (RCT) or A/B testing
  - RCT's always more powerful than observational studies, because we know they give an unbiased estimate
- Applications: Large-scale experiment
  - The impact of the experiment on business can be reduced by running it on a smaller subset of markets
  - The groups that can be randomized affect the statistical power: randomizing by users is much more informative than randomizing by DMA.
  - Often linear regression estimates overstate the effect of advertising

Practice questions

Which of the following is true about experiments?
* Ans: They provide causal estimates.
* They help most with predictive analytics.
* They make machine learning algorithms make more accurate predictions.
* They help with segmentation analysis.

Practice questions

How do experiments work?
* Ans: They use randomization.
* They predict which subjects should be treated.
* They don't typically impose any cost on a business.
* They require sophisticated techniques to analyze.

Practice questions

Your company wants to advertise more in the south because it predicts people there buy more of your products. Each of the following is a valid concern with this justification, **except**:
* Ans: They didn't test whether the product is more popular in the south of other countries.
* They don't have experimental evidence that people in the south are more responsive to ads.
* People in the south might be more familiar with the product because your company started in the south.
* Correlation is not causation.