# Week 4 Recap

David S. Rosenberg

NYU: CDS

February 24, 2021

# Contents

1. Where are we?
2. Neyman-Rubin potential outcome framework
3. Various settings for investigating treatment effects
4. Control variates
5. Control variate experiment

# Where are we?

# Techniques and applications so far

| | Techniques | Applications |
|---|---|---|
| So far | Inverse propensity weighting (IPW) | Missing data / response bias |
| | Self-normalization | |
| | Regression imputation | |
| | Importance sampling / weighting | Covariate shift |
| This week | Control variates | Average treatment effect estimation |
| | Doubly robust estimators | Conditional ATE estimation |
| Next few weeks | Policy gradient | Bandit optimization |
| | Thompson sampling | Offline bandit optimization |
| | REINFORCE | Reinforcement learning |

# Neyman-Rubin potential outcome framework

# Treatments

- Suppose we want to know
    - whether a new medicine improves outcomes
    - whether a new webpage layout keeps people on the page longer
    - whether sending somebody a particular postcard increases their probability of donating money
- We can think of each of these as a **treatment**.
- Our **goal** is to understand the effect of a treatment on an outcome measure.

# Individual Treatment Effects

- Let $Y_i(1) \in \mathbb{R}$ be the "**potential outcome**" if we **give** the treatment to individual $i$.
- Let $Y_i(0) \in \mathbb{R}$ be the "**potential outcome**" if we **do not give** the treatment to $i$
- The **individual treatment effect** for individual $i$ is defined as

$$D_i = Y_i(1) - Y_i(0).$$

- The problem is, we never observe $Y_i(1)$ and $Y_i(0)$ for the same person!
- Some call this the **fundamental problem of causal inference**.

# Treatment assignment indicator

- $W_i \in \{0,1\}$ is the **treatment indicator** for individual $i$:

$$W_i = \begin{cases} 0 & \text{if individual } i \text{ does } \textbf{not} \text{ receive the treatment} \\ 1 & \text{if individual } i \text{ receives the treatment} \end{cases}$$

- When $W_i = 1$, we observe $Y_i(1)$ but not $Y_i(0)$.
- When $W_i = 0$, we observe $Y_i(0)$ but not $Y_i(1)$.
- The group of individuals with $W_i = 0$ is called the **control group**.
- The group of individuals with $W_i = 1$ is called the **treatment group**.

# What we observe

- We'll write the **observed data** $\mathcal{D}$ as

$$(W_1, Y_1), \ldots, (W_n, Y_n),$$

where

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

# What we observe

| Id | $W$ | $Y(0)$ | $Y(1)$ | $D = Y(1) - Y(0)$ |
|---|---|---|---|---|
| 1 | 0 | 1.2 | ? | ? |
| 2 | 0 | 2.3 | ? | ? |
| 3 | 1 | ? | 8.6 | ? |
| 4 | 0 | .7 | ? | ? |
| 5 | 1 | ? | 3.4 | ? |

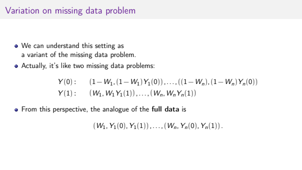
# Variation on missing data problem

- We can understand this setting as a variant of the missing data problem.
- Actually, it's like two missing data problems:

$$\begin{aligned} Y(0): &\quad (1-W_1,(1-W_1)Y_1(0)),\ldots,((1-W_n),(1-W_n)Y_n(0)) \\ Y(1): &\quad (W_1,W_1Y_1(1)),\ldots,(W_n,W_nY_n(1)) \end{aligned}$$

- From this perspective, the analogue of the **full data** is

$$(W_1,Y_1(0),Y_1(1)),\ldots,(W_n,Y_n(0),Y_n(1)).$$



Variation on missing data problem

- We can understand this setting as
  a variant of the missing data problem.
- Actually, it's like two missing data problems:

$$\begin{aligned} Y(0): &\quad (1-W_1,(1-W_1)Y_1(0)),\ldots,((1-W_n),(1-W_n)Y_n(0)) \\ Y(1): &\quad (W_1,W_1Y_1(1)),\ldots,(W_n,W_nY_n(1)) \end{aligned}$$

- From this perspective, the analogue of the **full data** is

$$(W_1,Y_1(0),Y_1(1)),\ldots,(W_n,Y_n(0),Y_n(1)).$$

- We'll refer to the two missing data problems as "$Y(0)$" and "$Y(1)$".
- The response indicators for these problems are $1-W$ and $W$, respectively.

# Various settings for investigating treatment effects

# General idea

- Randomly sample individuals $i = 1, \ldots, n$ from a population.
- Each individual $i$ is assigned to one of two groups:

  control group: individuals do not receive the treatment

  treatment group: individuals do receive the treatment
- Individuals **assigned randomly** to treatment and control groups
  - Simplest: Individuals assigned by a flipping a **fair coin** (RCT)
  - Simple: Individuals assigned by a flipping a **biased coin** (RCT)
  - Less simple: bias of coin depends on **covariates** (unconfoundedness)

# Randomized control trial (RCT)

In a randomized control trial (RCT), we assume
$(W, Y(0), Y(1)), (W_1, Y_1(0), Y_1(1)), \ldots, (W_n, Y_n(0), Y_n(1))$
are i.i.d. subject to the following assumption:

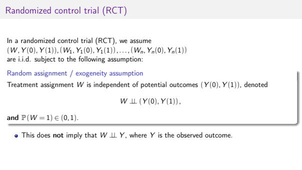
### Random assignment / exogeneity assumption

Treatment assignment $W$ is independent of potential outcomes $(Y(0), Y(1))$, denoted

$$W \perp\!\!\!\perp (Y(0), Y(1)),$$

**and** $\mathbb{P}(W = 1) \in (0, 1)$.

- This does **not** imply that $W \perp\!\!\!\perp Y$, where $Y$ is the observed outcome.



Randomized control trial (RCT)

In a randomized control trial (RCT), we assume
$(W, Y(0), Y(1)), (W_1, Y_1(0), Y_1(1)), \dots, (W_n, Y_n(0), Y_n(1))$
are i.i.d. subject to the following assumption:

Random assignment / exogeneity assumption

Treatment assignment $W$ is independent of potential outcomes $(Y(0), Y(1))$, denoted

$$W \perp\!\!\!\perp (Y(0), Y(1)),$$

**and** $\mathbb{P}(W = 1) \in (0,1)$.

- This does **not** imply that $W \perp\!\!\!\perp Y$, where $Y$ is the observed outcome.

- Note that we've added a generic individual $(W, Y(0), Y(1))$ that has the same distribution as any one in our sample. This is a common trick in probability and statistics to clean up notation. Since individuals are i.i.d., we could talk about $\mathbb{E}Y_1(0)$ or $\mathbb{E}Y_i(0)$ and they're both the same. So by introducing $(W, Y(0), Y(1))$ we can just drop the subscript.

- Recall that the **observed outcome** is $Y = (1-W)Y(0) + WY(1)$.
- Should be clear why we want $\mathbb{P}(W = 1) \in (0,1)$?
- Making the connection to the missing data setting, the exogeneity assumption implies that the $Y(0)$ and $Y(1)$ missing data problems are both MCAR, since exogeneity implies $(1-W) \perp\!\!\!\perp Y(0)$ and $W \perp\!\!\!\perp Y(1)$ and .
- In words, exogeneity says that even if we know an individual's potential outcomes to treatment and control (i.e. $Y(0)$ and $Y(1)$), this would give no information on whether the individual was assigned the treatment. This precludes doctors giving treatments to individuals who they believe are more likely to benefit (assuming the doctors' predictions are not completely independent of reality)

# Connection to MCAR

- Consider the exogeneity assumption $W \perp\!\!\!\perp (Y(0), Y(1))$.
- Exogeneity implies
    - $Y(0)$ missing data problem is MCAR (i.e. $(1-W) \perp\!\!\!\perp Y(0)$)
    - $Y(1)$ missing data problem is MCAR (i.e. $W \perp\!\!\!\perp Y(1)$)
- Where $1-W$ and $W$ are the respective response indicators.

# Average treatment effect

- Define the **average treatment effect** as

$$\text{ATE} := \mathbb{E}\left[Y(1) - Y(0)\right].$$

- If we had full data, we could use the natural estimator:

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i(1) - Y_i(0)\right).$$

- Unfortunately, either $Y_i(1)$ or $Y_i(0)$ is missing in every summand.

# Difference-in-means estimator

- Estimating ATE is like two missing data problems in the MCAR setting:
  - estimate $\mathbb{E}Y(1)$ from observations of $Y(1)$
  - estimate $\mathbb{E}Y(0)$ from observations of $Y(0)$
- Let's define two complete case mean estimators:

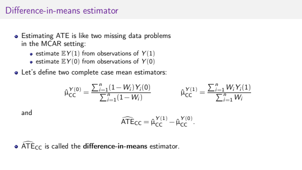
$$\hat{\mu}_{\text{CC}}^{Y(0)} = \frac{\sum_{i=1}^{n}(1-W_i)Y_i(0)}{\sum_{i=1}^{n}(1-W_i)} \qquad \hat{\mu}_{\text{CC}}^{Y(1)} = \frac{\sum_{i=1}^{n}W_iY_i(1)}{\sum_{i=1}^{n}W_i}$$

  and

$$\widehat{\text{ATE}}_{\text{CC}} = \hat{\mu}_{\text{CC}}^{Y(1)} - \hat{\mu}_{\text{CC}}^{Y(0)}.$$

- $\widehat{\text{ATE}}_{\text{CC}}$ is called the **difference-in-means** estimator.

## └─Difference-in-means estimator



### Difference-in-means estimator

- Estimating ATE is like two missing data problems in the MCAR setting:
  - estimate $\mathbb{E}Y(1)$ from observations of $Y(1)$
  - estimate $\mathbb{E}Y(0)$ from observations of $Y(0)$
- Let's define two complete case mean estimators:

$$\hat{\mu}_{\mathrm{CC}}^{Y(0)} = \frac{\sum_{i=1}^{n}(1-W_i)Y_i(0)}{\sum_{i=1}^{n}(1-W_i)} \qquad \hat{\mu}_{\mathrm{CC}}^{Y(1)} = \frac{\sum_{i=1}^{n}W_iY_i(1)}{\sum_{i=1}^{n}W_i}$$

and

$$\widehat{\mathrm{ATE}}_{\mathrm{CC}} = \hat{\mu}_{\mathrm{CC}}^{Y(1)} - \hat{\mu}_{\mathrm{CC}}^{Y(0)}.$$

- $\widehat{\mathrm{ATE}}_{\mathrm{CC}}$ is called the **difference-in-means** estimator.

- We only use $Y_i(1)$ when $W_i = 1$ and $Y_i(0)$ and $W_i = 0$. So we only need the observed data to use these estimators.
- We could also have written these estimators by replacing both $Y_i(1)$ and $Y_i(0)$ with the observed outcome $Y_i$.

# Relaxing random assignment / exogeneity

- [?] speaks of
  "assignment to treatment group on the basis of a covariate."
- Think of assigning an individual to treatment or control by a coin toss
  - but the coin has a different bias depending on the covariates / features of the individual
- We'll refer to it as **ignorability** or **unconfoundedness.**
- Sometimes referred to as "**no hidden confounders**"

# Introducing a covariate

- For each individual $i$,
  - we'll associate a covariate $X_i \in \mathcal{X}$.
- Then our **full data** is

$$(X, W, Y(0), Y(1)), (X_1, W_1, Y_1(0), Y_1(1)), \ldots, (X_n, W_n, Y_n(0), Y_n(1)),$$

  which we assume are i.i.d.

# Ignorability / unconfoundedness assumption

### Ignorability / unconfoundedness assumption

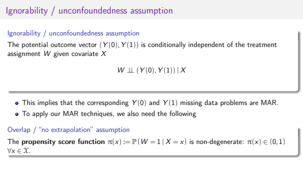
The potential outcome vector $(Y(0), Y(1))$ is conditionally independent of the treatment assignment $W$ given covariate $X$

$$W \perp\!\!\!\perp (Y(0), Y(1)) \mid X$$

- This implies that the corresponding $Y(0)$ and $Y(1)$ missing data problems are MAR.
- To apply our MAR techniques, we also need the following

### Overlap / "no extrapolation" assumption

The **propensity score function** $\pi(x) := \mathbb{P}(W = 1 \mid X = x)$ is non-degenerate: $\pi(x) \in (0,1)$ $\forall x \in \mathcal{X}$.



Ignorability / unconfoundedness assumption

Ignorability / unconfoundedness assumption

The potential outcome vector $(Y(0), Y(1))$ is conditionally independent of the treatment assignment $W$ given covariate $X$

$$W \perp\!\!\!\perp (Y(0), Y(1)) \mid X$$

- This implies that the corresponding $Y(0)$ and $Y(1)$ missing data problems are MAR.
- To apply our MAR techniques, we also need the following

Overlap / "no extrapolation" assumption

The **propensity score function** $\pi(x) := \mathbb{P}(W = 1 \mid X = x)$ is non-degenerate: $\pi(x) \in (0,1)$ $\forall x \in \mathcal{X}$.

- In words: are once we observe $X$, knowing the potential outcomes $Y(0)$ and $Y(1)$ would give no additional information about treatment assignment $W$.
- By analogy with the MAR terminology, we might want to call the ignorability assumption "assignment at random", and by analogy with MCAR we might call exogeneity / random assignment "assignment completely at random." Unfortunately, nobody uses these terms, so we won't either.
- The overlap assumptions will imply that the propensity score is strictly positive for both the $Y(0)$ and the $Y(1)$ missing data problems.

# Implications

- Under the assumptions of
    1. ignorability / unconfoundedness and
    2. overlap / “no extrapolation”
- We can treat ATE estimation as two missing data problems under MAR.
- All of our estimators in the missing data setting can be applied as ATE estimators.

# The IPW estimator

- Let's estimate $\mathbb{E}Y(1)$ and $\mathbb{E}Y(0)$ using missing data strategies.
- Let's try the IPW mean estimator:

$$\hat{\mu}_{\text{ipw}}^{Y(1)} := \frac{1}{n}\sum_{i=1}^{n}\frac{W_i Y_i(1)}{\pi(X_i)} \qquad \hat{\mu}_{\text{ipw}}^{Y(0)} := \frac{1}{n}\sum_{i=1}^{n}\frac{(1-W_i)Y_i(0)}{1-\pi(X_i)}$$

- Putting this together gives us

$$\widehat{\text{ATE}}_{\text{ipw}} := \hat{\mu}_{\text{ipw}}^{Y(1)} - \hat{\mu}_{\text{ipw}}^{Y(0)}.$$

- This is **unbiased** for ATE since the estimators for $\mathbb{E}Y(1)$ and $\mathbb{E}Y(0)$ are unbiased.
- This is **consistent** for ATE since the estimators for $\mathbb{E}Y(1)$ and $\mathbb{E}Y(0)$ are consistent.

# And so on...

- We can build ATE estimators using
  - self-normalized IPW estimators (haven't seen this in the literature, but surely somebody has done it)
  - regression imputation estimators
  - augmented IPW estimators (informally referred to as "the doubly robust estimator")
- To summarize, the common assumptions made in ATE estimation allow us to reduce ATE estimation to two missing data problems.
- This works well for us, since we now have a pretty thorough understanding of missing data problems :).

# Control variates

# Simplest setting

- Suppose we observe $X \in \mathcal{X}$ and $Y \in \mathbb{R}$.
- $(X, Y)$ have some unknown joint distribution.
- **Goal:** Estimate $\mathbb{E}Y$.
- $Y$ is a simple estimator for $\mathbb{E}Y$.
- It's even unbiased.
- Can we use $X$ to improve our estimate of $Y$?
- In particular, to reduce the variance of our estimate?

# Suppose we have a regression function

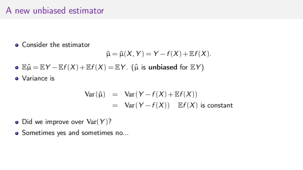
- Suppose somehow we have a function $f$.
- And **we think** $f(X) \approx Y$. No guarantees.
- Suppose we also know $\mathbb{E}f(X)$.
- Can we use $f(X)$ to get a better estimate of $\mathbb{E}Y$?

# A new unbiased estimator

- Consider the estimator
$$\hat{\mu} = \hat{\mu}(X, Y) = Y - f(X) + \mathbb{E}f(X).$$
- $\mathbb{E}\hat{\mu} = \mathbb{E}Y - \mathbb{E}f(X) + \mathbb{E}f(X) = \mathbb{E}Y$. ($\hat{\mu}$ is **unbiased** for $\mathbb{E}Y$)
- Variance is

$$\begin{aligned} \operatorname{Var}(\hat{\mu}) &= \operatorname{Var}(Y - f(X) + \mathbb{E}f(X)) \\ &= \operatorname{Var}(Y - f(X)) \quad \mathbb{E}f(X) \text{ is constant} \end{aligned}$$

- Did we improve over $\operatorname{Var}(Y)$?
- Sometimes yes and sometimes no...



## A new unbiased estimator

- Consider the estimator
$$\hat{\mu} = \hat{\mu}(X,Y) = Y - f(X) + \mathbb{E}f(X).$$
- $\mathbb{E}\hat{\mu} = \mathbb{E}Y - \mathbb{E}f(X) + \mathbb{E}f(X) = \mathbb{E}Y$. ($\hat{\mu}$ is **unbiased** for $\mathbb{E}Y$)
- Variance is
$$\begin{aligned} \mathrm{Var}(\hat{\mu}) &= \mathrm{Var}(Y - f(X) + \mathbb{E}f(X)) \\ &= \mathrm{Var}(Y - f(X)) \quad \mathbb{E}f(X) \text{ is constant} \end{aligned}$$
- Did we improve over $\mathrm{Var}(Y)$?
- Sometimes yes and sometimes no...

- How should we think about this estimator intuitively?
- We can think that we're starting our estimate of $\mathbb{E}Y$ with $\mathbb{E}f(X)$.
- Then we want to correct $\mathbb{E}f(X)$ by how much it's off by. Ideally that would be $\mathbb{E}(Y - f(X))$.
- We don't know $\mathbb{E}(Y - f(X))$, but $Y - f(X)$ is an unbiased estimate that we'll use instead.

# Ideal case

- Suppose $f(x) = \mathbb{E}[Y \mid X = x]$.
- The best approximation for $Y$ given $X = x$ (in MSE)
- Then

$$\operatorname{Var}(\hat{\mu}) = \operatorname{Var}(Y - \mathbb{E}[Y \mid X]).$$

- The projection-residual decomposition of variance gives:

$$\begin{aligned} \operatorname{Var}(Y) &= \operatorname{Var}(Y - \mathbb{E}[Y \mid X]) + \operatorname{Var}(\mathbb{E}[Y \mid X]) \\ &= \operatorname{Var}(\hat{\mu}) + \operatorname{Var}(\mathbb{E}[Y \mid X]) \end{aligned}$$

- So $\hat{\mu}$ has smaller variance than $Y$ by an amount $\operatorname{Var}(\mathbb{E}[Y \mid X])$.
- This is the amount of variation in $Y$ that we can account for with $X$.

# Control variates

### Definition

A **control variate** is a random variable with **known expectation** used to reduce the variance of an estimator. [?, Sec 8.9].

- In the context described above, with

$$\hat{\mu} = Y - f(X) + \mathbb{E}f(X),$$

- $f(X)$ is called a **control variate**.
- Intuitively, to be effective in creating lower-variance estimators of $\mathbb{E}Y$, we need $f(X) \approx Y$,
  - but we'll make that more precise later.

# Control variate experiment

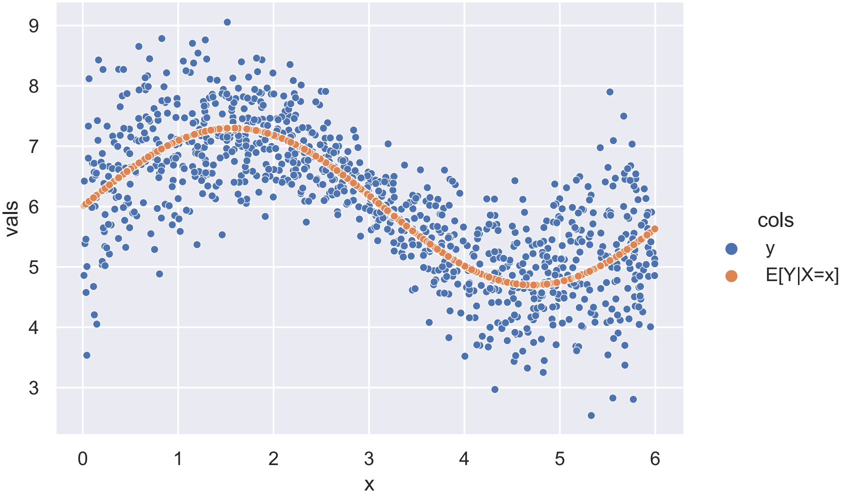
# Control variate experiment

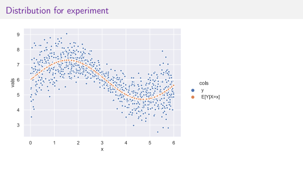
- Consider the following joint distribution of $(X, Y)$:

$$
\begin{aligned}
X &\sim \text{Unif}[0,6] \\
Y \mid X &\sim \mathcal{N}\left(6 + 1.3\sin(X), \left[.3 + \frac{1}{4}|3 - X|\right]^2\right)
\end{aligned}
$$

- Goal: Estimate $\mathbb{E}Y$.

# Distribution for experiment





Distribution for experiment

- Note that the variance is much smaller around $x = 3$ than near $x = 0$ and $x = 6$.
- The objective is to estimate $\mathbb{E}Y$ given a sample of size $n = 1$, i.e. just one of those blue points.
- Can you roughly estimate, by eyeball, what $\mathrm{SD}(\mathbb{E}[Y \mid X])$ is? Looks to me roughly about 1. [In fact, it's 0.91.]
- Can you roughly estimate, by eyeball, what $\mathrm{SD}(Y)$ is? To me, it looks roughly like 2. [In fact, it's 1.18 – I was pretty off.]
- The hope is that with a control variate, we can eliminate a lot of the variance in $Y$ that's due to $\mathbb{E}[Y \mid X]$. So we're hoping that our control-variate estimator will have variance in the ballpark of $\mathrm{Var}(Y) - \mathrm{Var}(\mathbb{E}[Y \mid X])$, which is about $(1.18)^2 - (.91)^2 = .55$. Of course, we can't get that low – that's what we would get if we knew $\mathbb{E}[Y \mid X = x]$. Perhaps we can come close with an estimate of $\mathbb{E}[Y \mid X = x]$

# Getting our control variate

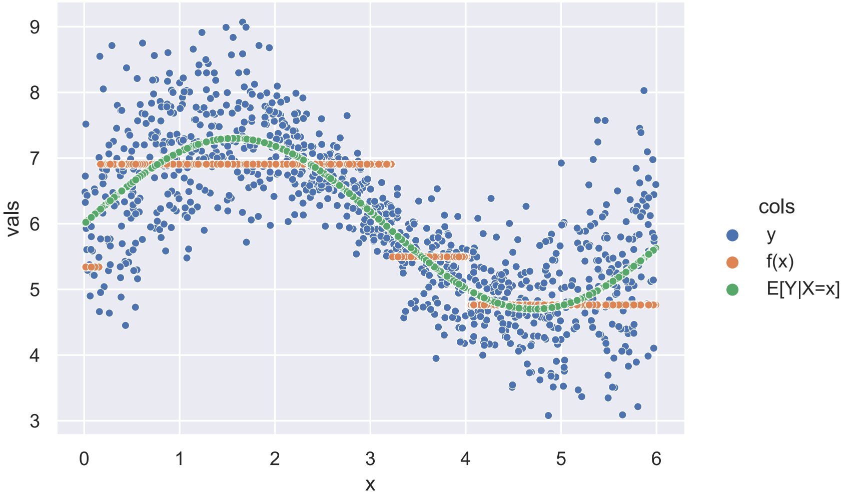
To get our control variate:

- Take a preliminary sample of size $n = 100$
- Fit a simple regression tree model to get $f(x)$
- Estimate $\mathbb{E}f(X)$ using the same preliminary sample, which we'll denote

$$\hat{\mathbb{E}}f(X) = \frac{1}{100}\sum_{i=1}^{100} f(X_i).$$

.

# Distribution with our regression fit

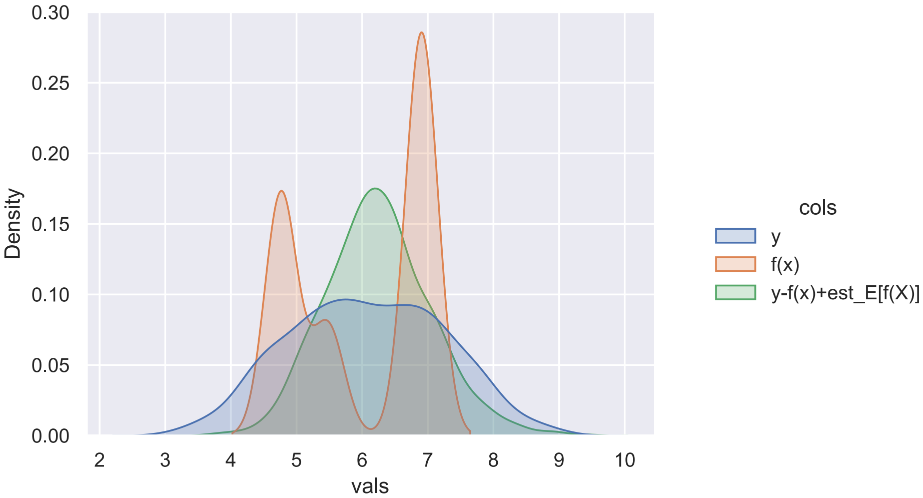


# Estimators for $\mathbb{E}Y$

Given $(X,Y)$, a sample of size $n=1$,
we'll consider the following estimators for $\mathbb{E}Y$:

- $Y$ (our baseline)
- $f(X)$ (benefits from the preliminary sample of size $n=100$)
- $Y-f(X)+\hat{\mathbb{E}}[f(X)]$ (our estimate with a control variate)

# Estimator distributions

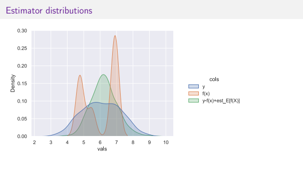


2021-02-24

Week 4 Recap

└Control variate experiment

└Estimator distributions

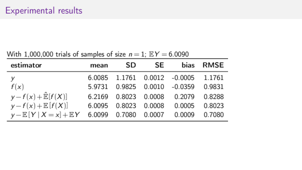

Estimator distributions

- These how the densities of the sampling distribution for three estimators of $\mathbb{E}Y$.
- The way to think about it is when we take our sample from the data generating distribution (here, just a sample of size $n = 1$), and then use that to instantiate our estimator, the estimator value we get is like being randomly drawn from one of the densities visualized above.
- Visually we can see that the control variate adjusted estimator has significantly lower variance than $Y$ and $f(X)$.

# Experimental results

With 1,000,000 trials of samples of size $n = 1$; $\mathbb{E}Y = 6.0090$

| estimator | mean | SD | SE | bias | RMSE |
|---|---|---|---|---|---|
| $y$ | 6.0085 | 1.1761 | 0.0012 | -0.0005 | 1.1761 |
| $f(x)$ | 5.9731 | 0.9825 | 0.0010 | -0.0359 | 0.9831 |
| $y - f(x) + \hat{\mathbb{E}}[f(X)]$ | 6.2169 | 0.8023 | 0.0008 | 0.2079 | 0.8288 |
| $y - f(x) + \mathbb{E}[f(X)]$ | 6.0095 | 0.8023 | 0.0008 | 0.0005 | 0.8023 |
| $y - \mathbb{E}[Y \mid X = x] + \mathbb{E}Y$ | 6.0099 | 0.7080 | 0.0007 | 0.0009 | 0.7080 |



Experimental results

With 1,000,000 trials of samples of size $n=1$; $\mathbb{E}Y=6.0090$

| estimator | mean | SD | SE | bias | RMSE |
|---|---|---|---|---|---|
| $y$ | 6.0085 | 1.1761 | 0.0012 | -0.0005 | 1.1761 |
| $f(x)$ | 5.9731 | 0.9825 | 0.0010 | -0.0359 | 0.9831 |
| $y-f(x)+\hat{\mathbb{E}}[f(X)]$ | 6.2169 | 0.8023 | 0.0008 | 0.2079 | 0.8288 |
| $y-f(x)+\mathbb{E}[f(X)]$ | 6.0095 | 0.8023 | 0.0008 | 0.0005 | 0.8023 |
| $y-\mathbb{E}[Y\mid X=x]+\mathbb{E}Y$ | 6.0099 | 0.7080 | 0.0007 | 0.0009 | 0.7080 |

- The estimator $y-f(x)+\hat{\mathbb{E}}[f(X)]$ uses the small ($n=100$) preliminary sample to determine both $f(x)$ and to get the estimate $\hat{\mathbb{E}}[f(X)]$. If we use the true value $\mathbb{E}[f(X)]$ in the estimator, the bias reduces significantly (from .21 to .00). However, the SD is so much larger than the bias, than the improvement in RMSE is relatively small.
- We certainly have more improvement from using the ideal control variate $\mathbb{E}[Y\mid X]$ over $f(X)$, but the majority of the RMSE improvement was already achieved by using $f(X)$.

# Practical setup for estimator with control variate

- $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.
- Goal: Estimate $\mathbb{E}Y$.
- Parameterized functions: $f(x; \theta) : \mathcal{X} \to \mathbb{R}$, for $\theta \in \mathbb{R}^d$
- Fit $\theta$ using least squares:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} \left(f(X_i; \theta) - Y_i\right)^2$$

- Use $f(X; \hat{\theta})$ as control variate.

# Estimate with control variate

- Consider the following estimator:

$$\hat{\mu} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - f(X_i;\hat{\theta}) + \mathbb{E}_X\left[f(X;\hat{\theta})\right]\right)$$

- This is the mean of $n$ control variate adjusted estimates of $Y$.
- $\mathbb{E}\hat{\mu} = \mathbb{E}Y$ and $\mathrm{Var}(\hat{\mu}) = \frac{1}{n}\mathrm{Var}\left(Y - f(X;\hat{\theta})\right)$.
- But how to get $\mathbb{E}_X\left[f(X,\hat{\theta})\right]$? (expectation only over $X$).

# How to get $\mathbb{E}\left[f(X,\hat{\theta})\right]$?

- It's either easy, or it's hard.
- If we know $p(x)$ and/or we can sample from $p(x)$,
  - then we can get $\mathbb{E}\left[f(X,\hat{\theta})\right]$ to whatever precision we need.

### Example (Survey from a voter file)

Suppose we have a "voter file", which has covariate information ($X$'s) about 200M eligible voters. We survey 1000 individuals to get $(X_1, Y_1), \ldots, (X_n, Y_n)$, and we want to estimate $\mathbb{E}Y$. We fit $f(x,\hat{\theta})$ with this sample. Then estimate $\mathbb{E}f(X,\hat{\theta})$ using the full voter file.

- If all we know about $p(x)$ is from our sample $(X_1, Y_1), \ldots, (X_n, Y_n)$,
  - then it's going to be hard.

# Can we estimate $\mathbb{E}\left[f(X,\hat{\theta})\right]$ from the sample?

- Estimate $\mathbb{E}_X\left[f(X;\hat{\theta})\right]$ as $\hat{\mathbb{E}}_X\left[f(X;\hat{\theta})\right] = \frac{1}{n}\sum_{i=1}^{n} f(X_i;\hat{\theta})$.
- If we plug this into our estimator, we get

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - f(X_i;\hat{\theta}) + \hat{\mathbb{E}}_X\left[f(X;\hat{\theta})\right]\right) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - f(X_i;\hat{\theta})\right) + \frac{1}{n}\sum_{i=1}^{n} f(X_i;\hat{\theta}) \\
= \; & \frac{1}{n}\sum_{i=1}^{n} Y_i
\end{aligned}$$

- This leaves us back where we started.

# $f(x)$ and $\mathbb{E}[f(X)]$ known: summary

- Conditions:
  - We know $f(x)$ and $\mathbb{E}[f(X)]$ before we get our sample.
  - We get a sample $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$
- Estimator:

$$\hat{\mu} = \mathbb{E}f(X) + \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))$$

- $\mathbb{E}\hat{\mu} = \mathbb{E}Y$. (unbiased estimator)
- Interpretation: We start by estimating $\mathbb{E}Y$ using $\mathbb{E}f(X)$, and then correct it by the average residual, which is an unbiased estimator for $\mathbb{E}Y - \mathbb{E}f(X)$, the residual of $\mathbb{E}f(X)$ as an estimate for $\mathbb{E}Y$.

# Unlimited samples from $p(x)$: summary

- Conditions:
  - We know $p(x)$ and/or can can get unlimited samples from it
  - We get a sample $(X,Y),(X_1,Y_1),\ldots,(X_n,Y_n)$
- Estimator:

$$\begin{aligned}
\hat{\theta} &= \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i;\hat{\theta})-Y_i\right)^2 \\
\hat{\mu} &= \hat{\mathbb{E}}_X f(X;\hat{\theta}) + \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - f(X_i;\hat{\theta})\right),
\end{aligned}$$

where $\hat{\mathbb{E}}_X f(X;\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N} f(X_i;\hat{\theta})$, where $X_1,\ldots,X_N$ is a large i.i.d. sample from $p(x)$.

# Regression estimator with control variate

- The following estimator is also unbiased for $\mathbb{E}Y$, for any $\beta \in \mathbb{R}$:

$$\hat{\mu}_\beta = Y - \beta f(X) + \beta \mathbb{E} f(X).$$

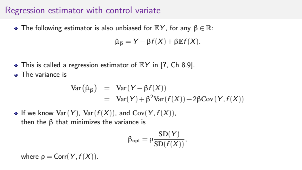
- This is called a regression estimator of $\mathbb{E}Y$ in [?, Ch 8.9].
- The variance is

$$\begin{aligned}\mathrm{Var}\left(\hat{\mu}_\beta\right) &= \mathrm{Var}\left(Y - \beta f(X)\right) \\ &= \mathrm{Var}(Y) + \beta^2 \mathrm{Var}\left(f(X)\right) - 2\beta \mathrm{Cov}\left(Y, f(X)\right)\end{aligned}$$

- If we know $\mathrm{Var}(Y)$, $\mathrm{Var}(f(X))$, and $\mathrm{Cov}(Y, f(X))$, then the $\beta$ that minimizes the variance is

$$\beta_{\text{opt}} = \rho \frac{\mathrm{SD}(Y)}{\mathrm{SD}(f(X))},$$

where $\rho = \mathrm{Corr}(Y, f(X))$.



## Regression estimator with control variate

- The following estimator is also unbiased for $\mathbb{E}Y$, for any $\beta \in \mathbb{R}$:

$$\hat{\mu}_\beta = Y - \beta f(X) + \beta \mathbb{E} f(X).$$

- This is called a regression estimator of $\mathbb{E}Y$ in [?, Ch 8.9].
- The variance is

$$\begin{aligned} \mathrm{Var}\left(\hat{\mu}_\beta\right) &= \mathrm{Var}\left(Y - \beta f(X)\right) \\ &= \mathrm{Var}(Y) + \beta^2 \mathrm{Var}\left(f(X)\right) - 2\beta \mathrm{Cov}\left(Y, f(X)\right) \end{aligned}$$

- If we know $\mathrm{Var}(Y)$, $\mathrm{Var}(f(X))$, and $\mathrm{Cov}(Y, f(X))$, then the $\beta$ that minimizes the variance is

$$\beta_{\mathrm{opt}} = \rho \frac{\mathrm{SD}(Y)}{\mathrm{SD}(f(X))},$$

where $\rho = \mathrm{Corr}(Y, f(X))$.

- Recall that the **variance of a sum** is

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i,j=1}^{n} \mathrm{Cov}\left(X_i, X_j\right) = \sum_{i=1}^{n} \mathrm{Var}\left(X_i\right) + 2 \sum_{i \neq j} \mathrm{Cov}\left(X_i, X_j\right)$$

- Also that the **correlation** is defined as

$$\rho = \mathrm{Corr}\left(X, Y\right) = \frac{\mathrm{Cov}\left(X, Y\right)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}\left(X, Y\right)}{\mathrm{SD}(X)\mathrm{SD}(Y)}.$$

- Finding the optimum $\beta$ is simple differential calculus:

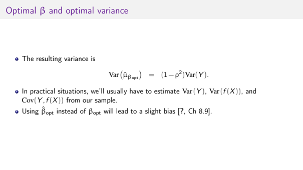
$$\begin{aligned} 2\beta \mathrm{Var}\left(f\left(X\right)\right) - 2\mathrm{Cov}\left(Y, f\left(X\right)\right) &= 0 \\ \iff \beta &= \frac{\mathrm{Cov}\left(Y, f\left(X\right)\right)}{\mathrm{Var}\left(f\left(X\right)\right)} = \rho \frac{\mathrm{SD}(Y)}{\mathrm{SD}(f(X))} \end{aligned}$$

# Optimal $\beta$ and optimal variance

- The resulting variance is

$$\mathrm{Var}\left(\hat{\mu}_{\beta_{\text{opt}}}\right) \quad = \quad (1-\rho^2)\mathrm{Var}(Y).$$

- In practical situations, we'll usually have to estimate $\mathrm{Var}(Y)$, $\mathrm{Var}(f(X))$, and $\mathrm{Cov}(Y, f(X))$ from our sample.
- Using $\hat{\beta}_{\text{opt}}$ instead of $\beta_{\text{opt}}$ will lead to a slight bias [?, Ch 8.9].



Optimal $\beta$ and optimal variance

- The resulting variance is

$$\mathrm{Var}\left(\hat{\mu}_{\beta_{\text{opt}}}\right) = (1-\rho^2)\mathrm{Var}(Y).$$

- In practical situations, we'll usually have to estimate $\mathrm{Var}(Y)$, $\mathrm{Var}(f(X))$, and $\mathrm{Cov}(Y,f(X))$ from our sample.
- Using $\hat{\beta}_{\text{opt}}$ instead of $\beta_{\text{opt}}$ will lead to a slight bias [?, Ch 8.9].

- Derivation is

$$\begin{aligned}
\mathrm{Var}\left(\hat{\mu}_{\beta_{\text{opt}}}\right) &= \mathrm{Var}(Y) + \rho^2\mathrm{Var}(Y) - 2\rho\frac{\mathrm{SD}(Y)}{\mathrm{SD}(f(X))}\mathrm{Cov}\left(Y,f(X)\right) \\
&= \mathrm{Var}(Y) + \rho^2\mathrm{Var}(Y) - 2\rho\frac{\mathrm{Var}(Y)}{\mathrm{SD}(f(X))}\frac{\mathrm{Cov}\left(Y,f(X)\right)}{\mathrm{SD}(Y)} \\
&= \mathrm{Var}(Y) + \rho^2\mathrm{Var}(Y) - 2\rho^2\mathrm{Var}(Y) \\
&= (1-\rho^2)\mathrm{Var}(Y)
\end{aligned}$$

- In the video I said that using $\hat{\beta}_{\text{opt}}$ is the recommended approach. I was trying to capture Owen's statement that "In general $\mathbb{E}\left[\hat{\mu}_{\hat{\beta}_{\text{opt}}}\right] \neq \mathbb{E}Y$, but this bias is usually small" [**?**, Ch 8.9]. However, whether this approach improves things or not can be situation dependent. We only demonstrated improvement when the variances and covariance are known – when these are estimated, no guarantees. Besides introducing a bias, we may also inflate the variance, as we plug these random estimates into a ratio.