Control Variates

David S. Rosenberg

NYU: CDS

February 17, 2021

Contents

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

Additive control variate: simplest setting

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

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

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

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.

Simple extreme cases

- f predicts Y perfectly: $f(X) = Y$

$$\begin{aligned} \hat{\mu} &= Y - f(X) + \mathbb{E}f(X) = \mathbb{E}f(X) \\ \text{Var}(\hat{\mu}) &= \text{Var}(Y - f(X)) = 0. \end{aligned}$$

- $f(X) = -Y$

$$\text{Var}(\hat{\mu}) = \text{Var}(Y - f(X)) = \text{Var}(2Y) = 4\text{Var}(Y).$$

  which is much worse than just using Y.

Simple extreme cases

- f predicts Y perfectly: $f(X) = Y$

$$\begin{aligned} \hat{\mu} &= Y - f(X) + \mathbb{E}f(X) = \mathbb{E}f(X) \\ \mathrm{Var}(\hat{\mu}) &= \mathrm{Var}(Y - f(X)) = 0. \end{aligned}$$

- $f(X) = -Y$

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}(Y - f(X)) = \mathrm{Var}(2Y) = 4\mathrm{Var}(Y).$$

which is much worse than just using Y.

- These cases are only possible if Y is some deterministic function of X.

Ideal case

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

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

- The projection-residual decomposition of variance gives:

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

- So $\hat{\mu}$ has smaller variance than Y by an amount $\text{Var}(\mathbb{E}[Y \mid X])$.
- 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. [Owe13, 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

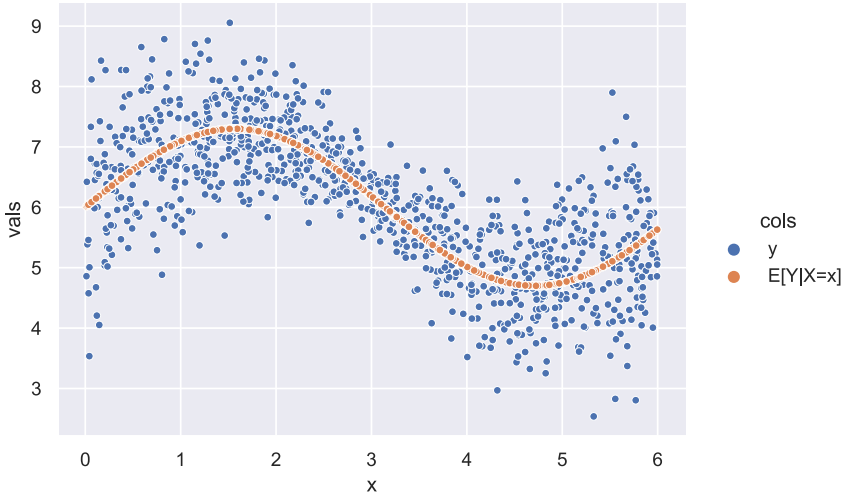
Control variate experiment

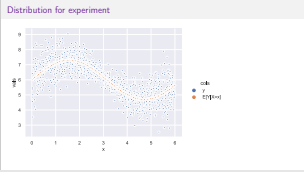
- Consider the following joint distribution of (X, Y):

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

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

Distribution for experiment





- 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

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

.

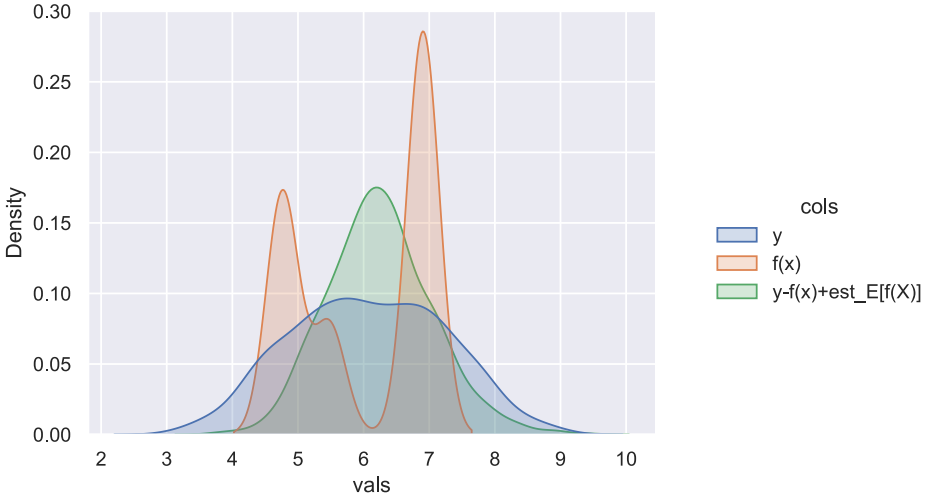
Distribution with our regression fit

Estimators for $\mathbb{E}Y$

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

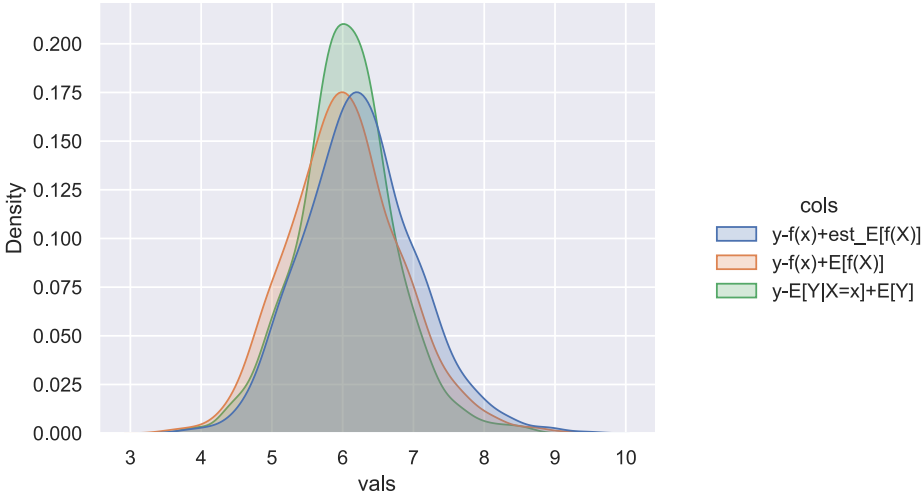
- Y (our baseline)
- $f(X)$ (benefits from the preliminary sample of size $n = 100$)
- $Y - f(X) + \hat{\mathbb{E}}[f(X)]$ (our estimate with a control variate)
- $Y - f(X) + \mathbb{E}[f(X)]$ (the actual $\mathbb{E}[f(X)]$ rather an estimate from the $n = 100$ sample)
- $Y - \mathbb{E}[Y \mid X = x]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

Estimator distributions

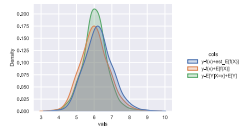


- These how the densities of the sampling distribution for three estimators of $\mathbb{E}Y$.
- The way to think about it is when we take our sample from the data generating distribution (here, just a sample of size $n = 1$), and then use that to instantiate our estimator, the estimator value we get 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 $f(X)$.

Estimator compared to idealized estimators



Estimator compared to idealized estimators



- Visually, not a lot of difference when we replace $\hat{E}[f(X)]$ by $\mathbb{E}[f(X)]$. And that difference is dominated by the SD of the sampling distributions.
- There seems to be some improvement using $\mathbb{E}[Y \mid X = x]$, but doesn't seem to be huge.
- Next we'll look at numeric results.

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
$y - \mathbb{E}[Y \mid X = x] + \mathbb{E}Y$	6.0090	0.7080	0.0007	0.0000	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 $f(X)$.

Control Variates in Practice

In practice...

- In our previous experiment,
  - assumed we already had $f(x)$ before we got our sample
  - assumed we already knew or had an estimate for $\mathbb{E}[f(X)]$.
- Sometimes this is reasonable,
  - e.g. $f(x)$ can be from a previous experiment.
  - We can think of $f(x)$ as like a prior
- But often we only get a single sample, and we don't have an $f(x)$ to start with.

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 θ using least squares:

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

- 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 $\text{Var}(\hat{\mu}) = \frac{1}{n}\text{Var}\left(Y - f(X;\hat{\theta})\right)$.
- 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.

Cross-fitting

- Simulate knowing $f(x)$ and $\mathbb{E}[f(X)]$ by cross-fitting.
- Randomly split data into two halves: $\mathcal{D}_1, \mathcal{D}_2$
- Fit $\hat{f}^{\mathcal{D}_1}$ on $\mathcal{D}_1$ and use $\hat{\mathbb{E}}^{\mathcal{D}_1}\hat{f}^{\mathcal{D}_1}(X) = \frac{1}{|\mathcal{D}_1|}\sum_{i\in\mathcal{D}_1}\hat{f}^{\mathcal{D}_1}(X_i)$ as mean estimate.
- Then $\hat{\mu}^{\mathcal{D}_2} = \hat{\mathbb{E}}^{\mathcal{D}_1}\hat{f}^{\mathcal{D}_1}(X) + \frac{1}{|\mathcal{D}_2|}\sum_{i\in\mathcal{D}_2}\left(Y_i - \hat{f}^{\mathcal{D}_1}(X_i)\right)$.
- and $\hat{\mu}^{\mathcal{D}_1} = \hat{\mathbb{E}}^{\mathcal{D}_2}\hat{f}^{\mathcal{D}_2}(X) + \frac{1}{|\mathcal{D}_1|}\sum_{i\in\mathcal{D}_1}\left(Y_i - \hat{f}^{\mathcal{D}_2}(X_i)\right)$
- Then define

$$\hat{\mu} = \frac{|\mathcal{D}_1|}{n}\hat{\mu}^{\mathcal{D}_1} + \frac{|\mathcal{D}_2|}{n}\hat{\mu}^{\mathcal{D}_2}.$$

- This is called **cross-estimation**.

$f(x)$ 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:

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

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 $p(x)$.

Optimal scaling to improve variance

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 [Owe13, Ch 8.9].
- The variance is

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

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

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

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

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 [Owe13, Ch 8.9].
- The variance is
$$\begin{aligned} \mathrm{Var}\left(\hat{\mu}_\beta\right) &= \mathrm{Var}(Y - \beta f(X)) \\ &= \mathrm{Var}(Y) + \beta^2 \mathrm{Var}(f(X)) - 2\beta \mathrm{Cov}(Y, f(X)) \end{aligned}$$
- If we know $\mathrm{Var}(Y)$, $\mathrm{Var}(f(X))$, and $\mathrm{Cov}(Y, f(X))$, then the β that minimizes the variance is
$$\beta_{opt} = \rho \frac{\mathrm{SD}(Y)}{\mathrm{SD}(f(X))},$$
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 β is simple differential calculus:

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

Optimal β and optimal variance

- The resulting variance is

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

- 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 $\hat{\beta}_{\mathsf{opt}}$ instead of β_{opt} will lead to a slight bias [Owe13, Ch 8.9].

Optimal β and optimal variance
- The resulting variance is
$$\mathrm{Var}\left(\hat{\mu}_{\hat{\beta}_{\mathsf{opt}}}\right) = (1-\rho^2)\mathrm{Var}(Y).$$
- 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 $\hat{\beta}_{\mathsf{opt}}$ instead of β_{opt} will lead to a slight bias [Owe13, Ch 8.9].

- Derivation is

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

- In the video I said that using $\hat{\beta}_{\mathsf{opt}}$ is the recommended approach. I was trying to capture Owen's statement that "In general $\mathbb{E}\left[\hat{\mu}_{\hat{\beta}_{\mathsf{opt}}}\right] \neq \mathbb{E}Y$, but this bias is usually small" [Owe13, 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.

References

Resources

- A good reference for control variates is a section in Owen's book [Owe13, Ch 8.9].
- In general, most longer discussions of Monte Carlo methods get into control variates as a way to reduce variance.
- The short Wikipedia article is actually quite readable, and has its own useful list of references.

References I

[Owe13] Art B. Owen, *Monte carlo theory, methods and examples*, 2013.