

MS&E 228: Applied Causal Inference Powered by ML and AI

Lecture 6: Variance of the DR Estimator and Variance Reduction in Experiments

Vasilis Syrgkanis

Stanford University

Winter 2026

Readings: *Applied Causal Inference Powered by ML and AI*, §9-10.

Goals for Today

1. Understand why the **doubly robust (DR) estimator** is not just robust, but also **statistically efficient**.
2. Decompose the **variance of the DR estimator** to see what drives its precision.
3. Learn how **regression adjustment** can be used to **reduce variance** in Randomized Controlled Trials (RCTs).
4. See how **interactive regression adjustment** (Lin 2013) guarantees precision gains in experiments.

Part I: Recap and In-Class Activity

Consolidating our ATE Estimation Toolbox

Recap: The ATE Estimation Landscape

We have built up a rich toolbox of estimators for ATE under conditional exogeneity and overlap:

- ▶ Linear regression adjustment (g -formula with linear model)
- ▶ G-estimator with generic ML for outcome regression (T-learner and S-learner)
- ▶ IPW estimator with generic ML for propensity
- ▶ IPW with un-penalized logistic regression for propensity
- ▶ DR estimator with ML for outcome and propensity (T- and S-learner variants)
- ▶ DR estimator with semi-cross-fitting
- ▶ DR estimator with stacked semi-cross-fitting

In-Class Activity (15 min)

Group Discussion: Algorithm Comparison

In your groups, discuss the estimators from the previous slide. Fill out a table with the following columns:

- ▶ Estimator
- ▶ Pros
- ▶ Cons
- ▶ When to Use

Be ready to share one key insight.

Poll 1

Poll Question

Which estimator would you choose for a setting where you believe the propensity is easy to learn but the outcome regression model might be hard to learn?

- A. IPW with ML
- B. G-formula with ML
- C. Doubly Robust estimator
- D. Linear regression adjustment



(Poll Everywhere)

Part II: Case Study

Aspirin, Pregnancy, and Per-Protocol Effects

In-Class Activity: Reading the Paper (10 min)

Reading Zhong et al. (2022) JAMA Network Open

Discuss in your groups:

1. What is the treatment?
2. What is the outcome?
3. What are the key confounders?
4. Why is this not a simple RCT analysis?
5. What specific DR configuration was used?



Figure: Paper Click Here

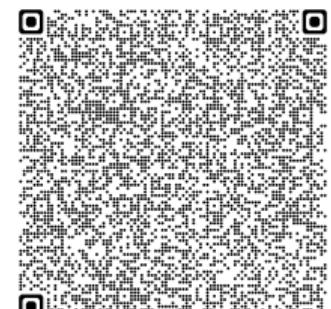


Figure: Supplementary
Material 2: Click Here

The EAGeR Trial: Aspirin and Pregnancy

Study: Effects of Aspirin in Gestation and Reproduction (EAGeR) trial.

- ▶ **Design:** Multicenter, block-randomized, double-blind, placebo-controlled clinical trial.
- ▶ **Participants:** 1,227 women with a history of pregnancy loss.
- ▶ **Intervention:** Daily low-dose aspirin (81 mg) vs. placebo.
- ▶ **Outcome:** hCG-detected pregnancy.
- ▶ **Follow-up:** Up to 6 menstrual cycles for attempted pregnancy.

Why Per-Protocol Analysis?

- ▶ **Intention-to-Treat (ITT):** Effect of *assignment* to treatment.
 - ▶ Preserves randomization.
 - ▶ Can be diluted by non-adherence.
- ▶ **Per-Protocol Effect:** Effect of *adhering* to the treatment protocol.
 - ▶ Often the scientific question of interest.
 - ▶ Breaks randomization—adherence is a choice!
- ▶ **The Challenge:** Per-protocol analysis of an RCT must be treated as an **observational study**, because adherence can be confounded by baseline and post-randomization factors.

EAGeR Trial: Key Results

Analysis	Effect Estimate	95% CI
Intention-to-Treat (ITT)	+4.3 per 100 women	(−1.1, 9.6)
Per-Protocol (DR with ML)	+8.0 per 100 women	(2.5, 13.6)

Key Insight: The per-protocol effect is nearly twice as large and statistically significant. Adherence matters, and flexible ML-based DR methods can uncover effects that ITT misses.

In-Class Activity: Reading the Paper (10 min)

Reading Zhong et al. (2022) JAMA Network Open

Discuss in your groups:

1. What is the treatment? (Adherence: ≥ 5 of 7 days/week for $\geq 80\%$ of follow-up)
2. What is the outcome? (hCG-detected pregnancy)
3. What are the key confounders? (Baseline: age, BMI, prior losses; Post-randomization: bleeding, nausea)
4. Why is this not a simple RCT analysis?
5. What specific DR configuration was used? (AIPW with Super Learner: GLM, MARS, RF, XGBoost)

Part III: DR in Practice

Industry Packages and Applications

EconML: LinearDRLearner (Microsoft)

```
1 from econml.dr import LinearDRLearner
2 from sklearn.ensemble import RandomForestRegressor, RandomForestClassifier
3
4 # S-learner variant by default
5 est = LinearDRLearner(
6     model_propensity=RandomForestClassifier(),
7     model_regression=RandomForestRegressor()
8 )
9 est.fit(Y, T, X=None, W=X)
10
11 # Get ATE and confidence intervals; the `intercept_` parameter
12 est.summary()
```

Note: Uses S-learner API by default. T-learner can be emulated by passing a model that trains separate models for treated/control.

DoubleML: IRM (DoubleML Package)

```
1 from doubleml import DoubleMLIRM, DoubleMLData
2 from sklearn.ensemble import RandomForestRegressor, RandomForestClassifier
3
4 # Prepare data
5 dml_data = DoubleMLData(df, y_col='Y', d_cols='D', x_cols=X_cols)
6
7 # T-learner API: separate models for E[Y|X,D=d]
8 ml_g = RandomForestRegressor()
9 ml_m = RandomForestClassifier()
10
11 dml_irm = DoubleMLIRM(dml_data, ml_g, ml_m, score='ATE')
12 dml_irm.fit()
13
14 print(dml_irm.summary)
15 ci = dml_irm.confint(level=0.95)
```

Note: Uses T-learner API—the model passed will be used to train separate models for treated and control.

Industry Application: Uber

Blog Post: "Using Causal Inference to Improve the Uber User Experience"

<https://www.uber.com/blog/causal-inference-at-uber/>

- ▶ Uber uses DR ML-based estimators for measuring treatment effects in observational studies.
- ▶ Real-world scale: millions of users, complex treatment assignment mechanisms.
- ▶ Key considerations: computational efficiency, robustness to model misspecification.

Takeaway: These methods are not just academic—they are deployed in production at major tech companies and are used by empirical researchers in a variety of fields.

Part IV: Variance of the DR Estimator

Understanding Efficiency

The Semiparametric Efficiency Bound

Key Theorem

In observational settings, the DR estimator achieves the **semiparametric efficiency bound**. This means no other asymptotically unbiased estimator can have a smaller asymptotic variance without making additional assumptions (e.g., linearity of the outcome regression).

Why this matters:

- ▶ DR is not just robust (double robustness), it's also **optimally precise**.
- ▶ If you're willing to make stronger assumptions (e.g., linear CEF), you can do better (e.g., OLS has lower variance under linearity).

Unpacking the Variance Formula (1/5)

Let $\psi_0(Z)$ be the DR score (influence function) at the true nuisance functions g_0, p_0 :

$$\psi_0(Z) = \underbrace{g_0(1, X) - g_0(0, X)}_{\text{G-formula part}} + \underbrace{a_0(D, X) \cdot (Y - g_0(D, X))}_{\text{IPW correction part}}$$

$$\text{where } a_0(D, X) = \frac{D}{p_0(X)} - \frac{1-D}{1-p_0(X)} = \frac{D-p_0(X)}{p_0(X)(1-p_0(X))}.$$

We want to compute $\text{Var}(\psi_0(Z)) = \mathbb{E}[(\psi_0 - \text{ATE})^2]$.

Unpacking the Variance Formula (2/5)

Expanding the variance:

$$\begin{aligned}\text{Var}(\psi_0) &= \underbrace{\mathbb{E}[(g_0(1, X) - g_0(0, X) - \text{ATE})^2]}_{\text{Term A: Variance of CATE}} \\ &\quad + \underbrace{\mathbb{E}[a_0(D, X)^2(Y - g_0(D, X))^2]}_{\text{Term B: Weighted Noise}} \\ &\quad + \underbrace{2\mathbb{E}[(g_0(1, X) - g_0(0, X) - \text{ATE}) \cdot a_0(D, X)(Y - g_0(D, X))]}_{\text{Term C: Cross-term}}\end{aligned}$$

Unpacking the Variance Formula (3/5)

The cross-term C vanishes!

By the law of iterated expectations:

$$\begin{aligned}C &= 2\mathbb{E}\left[(g_0(1, X) - g_0(0, X) - \text{ATE}) \cdot a_0(D, X) \cdot \mathbb{E}[Y - g_0(D, X) | D, X]\right] \\&= 2\mathbb{E}\left[(g_0(1, X) - g_0(0, X) - \text{ATE}) \cdot a_0(D, X) \cdot 0\right] \\&= 0\end{aligned}$$

This is because $\mathbb{E}[Y | D, X] = g_0(D, X)$ by definition of the true conditional expectation function.

Unpacking the Variance Formula (4/5)

Simplifying Term A:

$$A = \mathbb{E}[(\text{CATE}(X) - \text{ATE})^2] = \text{Var}(\text{CATE}(X))$$

where $\text{CATE}(X) = g_0(1, X) - g_0(0, X)$ is the Conditional Average Treatment Effect.

Simplifying Term B:

Let $\epsilon = Y - g_0(D, X)$ be the residual noise. Define $\sigma^2(d, X) = \text{Var}(Y | D = d, X)$ (heteroskedastic noise).

$$B = \mathbb{E}[a_0(D, X)^2 \cdot \mathbb{E}[\epsilon^2 | D, X]] = \mathbb{E}[a_0(D, X)^2 \cdot \sigma^2(D, X)]$$

Unpacking the Variance Formula (5/5)

Expanding $a_0(D, X)^2$ and using $\mathbb{E}[D | X] = p_0(X)$:

$$\begin{aligned} B &= \mathbb{E} \left[\frac{D}{p_0(X)^2} \sigma^2(1, X) + \frac{1 - D}{(1 - p_0(X))^2} \sigma^2(0, X) \right] \\ &= \mathbb{E} \left[\frac{\sigma^2(1, X)}{p_0(X)} + \frac{\sigma^2(0, X)}{1 - p_0(X)} \right] \end{aligned}$$

The Semiparametric Efficiency Bound

$$\text{Var}(\psi_0) = \underbrace{\text{Var}(\text{CATE}(X))}_{\text{CATE Heterogeneity}} + \underbrace{\mathbb{E} \left[\frac{\sigma^2(1, X)}{p_0(X)} + \frac{\sigma^2(0, X)}{1 - p_0(X)} \right]}_{\text{Weighted Noise Variance}}$$

Interpreting the Variance Components

Component 1: $\text{Var}(\text{CATE}(X))$

- ▶ Captures how much the treatment effect varies across individuals.
- ▶ More heterogeneous effects \Rightarrow higher variance.

Component 2: $\mathbb{E} \left[\frac{\sigma^2(1, X)}{p_0(X)} + \frac{\sigma^2(0, X)}{1 - p_0(X)} \right]$

- ▶ Captures how noisy outcomes are, weighted by inverse propensity.
- ▶ When $p_0(X) \approx 0$: we care a lot about $\sigma^2(1, X)$ (few treated).
- ▶ When $p_0(X) \approx 1$: we care a lot about $\sigma^2(0, X)$ (few controls).
- ▶ **Worst case:** We tend to not treat people with noisy outcomes under treatment,
AND we tend to treat people with noisy outcomes under control.

Poll 2

Poll Question

In which scenario would you expect the DR estimator to have the highest variance?

- A. High treatment effect heterogeneity, balanced propensity
- B. Low treatment effect heterogeneity, extreme propensity
- C. High treatment effect heterogeneity, extreme propensity
- D. Low treatment effect heterogeneity, balanced propensity

Part V: Variance Reduction in RCTs

Getting More Precise Estimates

The Main Question in RCTs

If we have pre-treatment variables X in an experiment, how can we use them?

Main idea: Use covariates to **reduce variance** by explaining away predictable variation in the outcome.

Example: In a medical trial, patients with prior conditions will have worse survival. Users with high prior engagement will engage more in the future. Why not “center” outcomes around these explainable parts?

Two-Means Estimator Variance

Consider an experiment where we treat with probability q . The simple two-means estimator is:

$$\hat{\theta}_{\text{TM}} = \bar{Y}_1 - \bar{Y}_0$$

Its variance is:

$$\text{Var}(\hat{\theta}_{\text{TM}}) = \frac{\text{Var}(Y | D = 1)}{q} + \frac{\text{Var}(Y | D = 0)}{1 - q}$$

Problem: We pay for *all* the variance of Y , even the parts that are perfectly predictable from covariates X .

The Ideal: Residual Variance

Ideally, we would want:

$$\frac{\text{Var}(Y - g(1, X) \mid D = 1)}{q} + \frac{\text{Var}(Y - g(0, X) \mid D = 0)}{1 - q}$$

This is the variance of the *residuals* after removing the predictable part.

Recall: The DR variance in an RCT (where $p(X) = q$) simplifies to:

$$\text{Var}(\text{CATE}(X)) + \frac{\mathbb{E}[\sigma^2(1, X)]}{q} + \frac{\mathbb{E}[\sigma^2(0, X)]}{1 - q}$$

Note: $\mathbb{E}[\sigma^2(1, X)] = \mathbb{E}[(Y - g_0(1, X))^2 \mid D = 1]$. So DR achieves roughly this ideal!

DR in RCTs: Key Simplifications (1/2)

In an RCT with treatment probability q :

- ▶ The propensity score is **known and constant**: $p(X) = q$.
- ▶ The Horvitz-Thompson weights simplify:

$$a(D) = \frac{D}{q} - \frac{1-D}{1-q} = \frac{D-q}{q(1-q)}$$

Note: This no longer depends on X !

- ▶ The DR formula becomes:

$$\hat{\theta}_{\text{DR}} = \mathbb{E}_n[\hat{g}(1, X) - \hat{g}(0, X)] + \mathbb{E}_n \left[\frac{D-q}{q(1-q)} (Y - \hat{g}(D, X)) \right]$$

DR in RCTs: Key Simplifications (2/2)

What about our convergence conditions?

- ▶ **Product rate condition:** $\sqrt{n} \cdot \text{RMSE}(\hat{g}) \cdot \text{RMSE}(\hat{p}) \rightarrow 0$
 - ▶ Automatically satisfied since $\text{RMSE}(\hat{p}) = 0$ (we know p exactly).
- ▶ **Propensity consistency:** $\hat{p} \rightarrow p_0$
 - ▶ Trivially satisfied since $\hat{p} = q = p_0$.
- ▶ **Outcome regression consistency:** $\hat{g} \rightarrow g_0$?
 - ▶ Surprisingly, NOT needed! We only need $\hat{g} \rightarrow g_*$ for some limit g_* .

A Surprising Result: No Consistency Needed

Key Insight

In an RCT, we only need \hat{g} to converge to *some* limit g_* , not necessarily the true CEF g_0 .

Examples:

- ▶ Linear regression over low-dimensional features: converges to the population best linear predictor (minimizing RMSE over all linear functions).
- ▶ Lasso over high-dimensional features: under sparsity, also converges to the best linear predictor.

In either case, DR is asymptotically normal and centered at the true ATE, with variance:

$$\text{Var}(g_*(1, X) - g_*(0, X) + a(D)(Y - g_*(D, X)))$$

OLS Adjustment and DR (1/3)

What if we use simple OLS: $Y \sim D + X$?

The estimated model is $\hat{g}(D, X) = \hat{\theta}D + \hat{\beta}'X + \hat{c}$.

The OLS coefficients minimize the empirical squared loss:

$$\mathbb{E}_n[(Y - \theta D - \beta' X - c)^2]$$

Equivalently, they satisfy the **empirical normal equations**:

$$\mathbb{E}_n[\hat{\epsilon} \cdot (1; D; X)] = 0$$

where $\hat{\epsilon} = Y - \hat{\theta}D - \hat{\beta}'X - \hat{c}$ is the OLS residual.

OLS Adjustment and DR (2/3)

The DR estimator with this OLS model is:

$$\hat{\theta}_{\text{DR}} = \mathbb{E}_n[\hat{g}(1, X) - \hat{g}(0, X)] + \mathbb{E}_n \left[\frac{D - q}{q(1 - q)} \hat{\epsilon} \right]$$

The first term is simply $\hat{\theta}$ (the OLS coefficient on D).

The correction term:

$$\begin{aligned} \mathbb{E}_n \left[\frac{D - q}{q(1 - q)} \hat{\epsilon} \right] &= \frac{1}{q(1 - q)} (\mathbb{E}_n[D \cdot \hat{\epsilon}] - q \cdot \mathbb{E}_n[\hat{\epsilon}]) \\ &= \frac{1}{q(1 - q)} (0 - q \cdot 0) = 0 \end{aligned}$$

by the normal equations!

OLS Adjustment and DR (3/3)

Key Result

The DR estimator is **numerically identical** to the OLS coefficient $\hat{\theta}$ when we use OLS for the outcome regression. The debiasing term does nothing because OLS has no regularization bias.

Corollary: Since DR converges to the ATE in an RCT, the coefficient on D in $\text{OLS}(Y \sim D, X)$ is a consistent estimator for the ATE, **even if the true CEF is not linear!**

The Variance of OLS-Adjusted Estimator (1/3)

Since DR = OLS coefficient, we can use the DR variance formula.

For simple linear regression, the limit $g_*(D, X) = \theta D + \beta' X + c$ where θ, β, c minimize the population loss:

$$\min_{\theta, \beta, c} \mathbb{E}[(Y - \theta D - \beta' X - c)^2]$$

These satisfy the **population normal equations**:

$$\mathbb{E}[\epsilon \cdot (1; D; X)] = 0$$

where $\epsilon = Y - \theta D - \beta' X - c$ is the population residual.

Note: We already know $\theta = \text{ATE}$ (since $\hat{\theta} \rightarrow \text{ATE}$).

The Variance of OLS-Adjusted Estimator (2/3)

The limit variance is:

$$V_{\text{DR}} = \text{Var} \left(\theta + \frac{D - q}{q(1 - q)} \epsilon \right) = \text{Var} \left(\frac{D - q}{q(1 - q)} \epsilon \right)$$

since θ is a constant.

Using $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon \cdot D] = 0$ (from normal equations):

$$V_{\text{DR}} = \frac{\mathbb{E}[(D - q)^2 \epsilon^2]}{(q(1 - q))^2}$$

Side note: This is exactly the **HC0 heteroskedasticity-robust variance** for the parameter θ in OLS. When you run OLS and specify `cov_type='HC0'`, this is what you get!

The Variance of OLS-Adjusted Estimator (3/3)

The Two-Means Variance:

$$V_{TM} = \frac{\mathbb{E}[(Y - \theta_1)^2 | D = 1]}{q} + \frac{\mathbb{E}[(Y - \theta_0)^2 | D = 0]}{1 - q}$$

where $\theta_1 = \mathbb{E}[Y | D = 1]$, $\theta_0 = \mathbb{E}[Y | D = 0]$, and $\theta = \theta_1 - \theta_0 = \text{ATE}$.

This can be rewritten as:

$$V_{TM} = \frac{\mathbb{E}[(D - q)^2(Y - \theta D - \theta_0)^2]}{(q(1 - q))^2}$$

Note: Two-means is equivalent to OLS($Y \sim D$), and this is its HC0 variance.

Comparing Two-Means and OLS-Adjusted (1/2)

The residuals are related: $Y - \theta D - \theta_0 = \epsilon + \beta' X + c - \theta_0$.

Substituting:

$$V_{TM} = \frac{\mathbb{E}[(D - q)^2 \epsilon^2]}{(q(1 - q))^2} + \frac{\mathbb{E}[(D - q)^2 (\beta' X + c - \theta_0)^2]}{(q(1 - q))^2}$$
$$+ \frac{2\mathbb{E}[(D - q)^2 \epsilon (\beta' X + c - \theta_0)]}{(q(1 - q))^2}$$

- ▶ First term = V_{DR}
- ▶ Second term ≥ 0
- ▶ Third term = **Cross-term** (can be positive or negative!)

Comparing Two-Means and OLS-Adjusted (2/2)

If the cross-term were zero: $V_{\text{DR}} \leq V_{\text{TM}}$ always!

But it's not necessarily zero. Note that

$(D - q)^2 = D(1/q^2 - 1/(1 - q)^2) + 1/(1 - q)^2$, so the cross-term contains:

$$\beta' \mathbb{E}[\epsilon \cdot X \cdot D] \cdot \left(\frac{1}{q^2} - \frac{1}{(1 - q)^2} \right) + \dots$$

The normal equations give $\mathbb{E}[\epsilon \cdot X] = 0$, but NOT $\mathbb{E}[\epsilon \cdot X \cdot D] = 0$!

Key Finding

The cross-term can be substantially negative, making $V_{\text{TM}} < V_{\text{DR}}$. Simple OLS adjustment can increase variance! (You will see an example in your homework.)

Poll 3

Poll Question

When is simple linear regression adjustment (OLS with $Y \sim D, X$) guaranteed to weakly improve precision over the two-means estimator?

- A. Always
- B. When the CEF of the outcome $\mathbb{E}[Y | D, X]$ is linear
- D. Never

Note: In your homework, you'll also be asked to argue that when the experiment is balanced ($q = 1/2$) then simple OLS also guarantees weak improvement in variance over two-means (problem vanishes).

The Fix: Interactive Regression (Lin, 2013)

The Problem: OLS residuals satisfy $\mathbb{E}[\epsilon \cdot X] = 0$, but not $\mathbb{E}[\epsilon \cdot D \cdot X] = 0$.

The Solution: Add interaction terms to the regression!

Interactive Regression

Run OLS($Y \sim D + X + D \cdot X$).

The estimated model: $\hat{g}(D, X) = \hat{a}_0 D + \hat{a}'_1(D \cdot X) + \hat{\beta}' X + \hat{c}$.

The new normal equations include:

$$\mathbb{E}_n[\hat{\epsilon} \cdot D \cdot X] = 0$$

This is exactly what we need!

Interactive Regression: The ATE Estimator

The CATE from the interactive model is: $\hat{a}_0 + \hat{a}'_1 X$.

The ATE is estimated by the empirical g-formula:

$$\hat{\theta} = \hat{a}_0 + \hat{a}'_1 \mathbb{E}_n[X]$$

Note: This is the same as the DR formula because:

- ▶ We're in an RCT (known propensity).
- ▶ We used OLS (no regularization bias).
- ▶ So the debiasing term is zero: $\mathbb{E}_n \left[\frac{D-q}{q(1-q)} \hat{\epsilon} \right] = 0$.

Interactive Regression: The Variance

The limit variance is:

$$V = \text{Var}(a_0 + a'_1 X) + \frac{\mathbb{E}[(D - q)^2 \epsilon^2]}{(q(1 - q))^2}$$

- ▶ First term: Variance of the best linear approximation to the CATE.
- ▶ Second term: Same form as before, but now ϵ satisfies $\mathbb{E}[\epsilon \cdot D \cdot X] = 0$!

Guaranteed Improvement

The second term is now guaranteed to be $\leq V_{TM}$. Lin (2013) and Negi & Wooldridge (2021) prove that the overall variance is also $\leq V_{TM}$.

Extension: ML-Learned Features (Guo et al., 2021)

We can go further by using ML-learned features $\hat{\phi}(X)$:

1. Train ML models (trees, kernels, neural nets) to predict Y from X (using cross-fitting).
2. Use their predictions as additional features.
3. Run OLS($Y \sim D + X + D \cdot X + \hat{\phi}(X) + D \cdot \hat{\phi}(X)$).

The ATE estimate:

$$\hat{\theta} = \hat{a}_0 + \hat{a}'_1 \mathbb{E}_n[X] + \hat{a}'_2 \mathbb{E}_n[\hat{\phi}(X)]$$

The variance:

$$V = \underbrace{\text{Var}(a_0 + a'_1 X + a'_2 \phi(X))}_{V_1} + \underbrace{\frac{\mathbb{E}[(D - q)^2 \epsilon^2]}{(q(1 - q))^2}}_{V_2}$$

This is used at Meta for variance reduction in A/B testing!

Practical Note: HC0 Standard Errors

How to compute the variance in practice?

- ▶ The second term (V_2) in the variance is the HC0 variance for \hat{a}_0 in OLS, **if you first de-mean the covariates X and $\hat{\phi}(X)$ before passing them to OLS.**
- ▶ Most statistical packages can compute this directly.
- ▶ Add to this variance the variance of the estimated CATE model (V_1).
- ▶ Use this variance when calculating the standard error (i.e. $se = \sqrt{(V_1 + V_2)/n}$).

Standard Errors for Interactive Regression: Python Code

```
1 # De-mean covariates and create interaction terms
2 X_dm = X - X.mean(axis=0)
3 X_interact = patsy.dmatrix('D * (' + '+'.join(X_dm.columns) + ')',
4                             X_dm.assign(D=D), return_type='dataframe')
5 # Run OLS with HC0 standard errors
6 ira = sm.OLS(Y, X_interact)
7 ira_results = ira.fit(cov_type='HC0')
8 ira_est = ira_results.params['D']
9
10 # Account for error in means of X (first-order effect on ATE)
11 interaction_cols = [c for c in X_interact.columns if c.startswith('D:')]
12 cate = X_interact[interaction_cols] @ ira_results.params[interaction_cols]
13 ira_se = np.sqrt(ira_results.HC0_se['D']**2 + np.var(cate)/len(Y))
```

Summary: Four Key Takeaways

1. The DR estimator achieves the **semiparametric efficiency bound**—it's optimally precise in observational settings.
2. DR variance depends on **CATE heterogeneity** and **noise amplified by extreme propensities**.
3. In RCTs, regression adjustment can **reduce variance**, but simple OLS($Y \sim D, X$) is **not guaranteed** to help (except when $q = 1/2$).
4. **Interactive regression** OLS($Y \sim D, X, D \cdot X$) **guarantees** variance reduction (Lin, 2013; Negi & Wooldridge, 2021).

Practical Recommendations

- ▶ **For observational data:** Use DR with ML and cross-fitting. It's both robust and efficient.
- ▶ **For RCTs:** Use interactive regression adjustment $\text{OLS}(Y \sim D, X, D \cdot X)$ to guarantee precision gains.
- ▶ **For large-scale experiments:** Consider ML-augmented adjustment (Guo et al., 2021) for even greater precision.
- ▶ **Always:** Report diagnostics (overlap, model fit) and use HC0 standard errors.

Next Time

- ▶ Continuous Treatments
- ▶ Revisiting OLS from a different perspective
- ▶ HC0 variance formulas beyond the DR connection
- ▶ Partially linear models and the “insensitive” formula

References

- ▶ Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, 7(1), 295-318.
- ▶ Negi, A., & Wooldridge, J. M. (2021). Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, 40(5), 504-534.
- ▶ Guo, Y., et al. (2021). Machine learning for variance reduction in online experiments. *NeurIPS 2021*.
- ▶ Zhong, Y., et al. (2022). Use of machine learning to estimate the per-protocol effect of low-dose aspirin on pregnancy outcomes. *JAMA Network Open*, 5(3), e2143414.