Regression Adjustment and CUPED

Justin S. Eloriaga

September 12, 2025

Agenda

Regression Adjustment and CUPED

Methodology: Regression Adjustment and CUPED

Regression Adjustment and CUPED

Overview

- In this example, we will use Jonathan Roth's DGP with heterogenous effects.
- You are a data scientist at Udemy looking at the effects of taking a professional development (D) certificate on earnings (Y).
- You randomly assign a sample of individuals to get the certificate or not.
- Let Z indicate how many online courses a person has taken in the past and Y_{t-1} be their earnings last year.
- Suppose that taking online courses causes lower earnings $Y(0)$ in jobs that don't require any certificates, but higher earnings $Y(1)$ in jobs that do require certificates.

Regression Adjustment

- Classical 2-sample approach, no adjustment (CL)
- Classical linear regression adjustment (CRA)
- Interactive regression adjustment (IRA)
- Let's compare standard errors
- Observe that CRA delivers estimates that are less efficient than CL (pointed out by Freedman), whereas IRA delivers estimates that are more efficient (pointed out by Lin).
- In order for CRA to be more efficient than CL, we need the linear model to be a correct model of the conditional expectation function of Y given D and X, which is not the case here.

Code: Regression Adjustment (1)

```
CL = smf.ols("np.log(Y) ~ D", data=data).fit(cov_type='HC1')
CL.summary().tables[1]
```

Code: Regression Adjustment (2)

```
CRA = smf.ols("np.log(Y) ~ D + Z + np.log(Ypre)", data=data).fit(
    cov_type='HC1')
CRA.summary().tables[1]
```

Code: Regression Adjustment (3)

```
# Demean Z and Ypre
data['Z_dm'] = data['Z'] - data['Z'].mean()
data['Ypre_dm'] = np.log(data['Ypre']) - np.log(data['Ypre']).mean()

# Interactive regression adjusment (IRA)
IRA = smf.ols("np.log(Y) ~ D + Z_dm + Z_dm*D + Ypre_dm + Ypre_dm*D",
    data=data).fit(cov_type='HC1')
IRA.summary().tables[1]
```

CUPED: Controlled-Experiment using Pre-Experiment Data

- This is a very popular technique in business settings to increase the power of RCTs.
- For a recent perspective on CUPED, see - [A New Look at CUPED in 2023](https://arxiv.org/pdf/2312.02935) - [Powering Experiments with CUPED](https://towardsdatascience.com/powering-experiments-with-cuped-and-double-machine-learning-34dc2f3d3284) - [Understanding...
- Steps to implement CUPED:
  - Regress Y on $X \equiv [Z, Y_{t-1}]$ and obtain the residuals $\hat{Y}_{\text{cuped}} = Y - \hat{\beta}X$.
  - Regress $\hat{Y}_{\text{cuped}}$ on D and obtain the treatment effect
- However, this implementation might not work here since we have heterogeneous treatment effect.

Code: CUPED: Controlled-Experiment using Pre-Experiment Data (1)

```
# Compute residuals
data['Y_tilde'] = smf.ols("np.log(Y) ~ Z_dm + Ypre_dm", data=data).fit
    ().resid
cuped = smf.ols("Y_tilde ~ D", data=data).fit(cov_type='HC1')
cuped.summary().tables[1]
```

Methodology: Regression Adjustment and cuped

Unconfoundedness & Identification

Assume unconfoundedness $Y(a) \perp A \mid X$ and positivity $0 < P(A = 1 \mid X) < 1$. Then the ATE is identified as

$$\text{ATE} \;=\; \mathbb{E}\big[\,\mathbb{E}[Y \mid A{=}1, X] - \mathbb{E}[Y \mid A{=}0, X]\,\big].$$

G-Computation / Regression Adjustment

Fit outcome models $\hat{m}_a(x) \approx \mathbb{E}[Y \mid A{=}a, X{=}x]$ for $a \in \{0, 1\}$ and average:

$$\widehat{\mathrm{ATE}}_{RA} \;=\; \frac{1}{n}\sum_{i=1}^{n}\{\hat{m}_1(X_i) - \hat{m}_0(X_i)\}.$$

Use linear models, polynomial features, or ML (RF/GBM); cross-fitting helps reduce overfitting bias.

cuped (Variance Reduction)

Construct $Y^\star = Y - \theta(\hat{Y}_0 - \overline{\hat{Y}_0})$, where $\hat{Y}_0$ predicts a baseline outcome. Estimate the effect with $Y^\star$ (same estimand, smaller variance).

- Fit baseline $\hat{g}(X)$ on pre-treatment data, get $\hat{Y}_0$.
- Estimate θ via regression of Y on $\hat{Y}_0$ (controlling for A).
- Recompute outcome $Y^\star$ and compare means by A (or regress Y on A and $\hat{Y}_0$ directly).