Abstract

This articles tries to give honest statistical background to the CUPED method. Statistical background allows to correctly include multiple predictors and use heteroscedasticity robust standard errors.

CUPED: statistician viewpoint

Boris Demeshev

October 20, 2021

1 Déjà vu

On the third page Deng writes ‘the linear model makes strong assumptions that are usually not satisfied in practice, i.e., the conditional ex- pectation of the outcome metric is linear in the treatment assignment and covariates. In addition, it also requires all residuals to have a common variance’.

As I am teaching statistics and econometrics I was eager to read further. But than I encounter $\theta = \mathrm{Cov}(Y, X) / \mathrm{Var}(X)$ in equation 4 which is a theoretical counterpart of slope estimate in simple regression. And later t-test is applied to δ_{cv} that is again equivalent to a second simple regression. Regression is replaced by something similar to two regressions. Déjà vu.

So I diceded to expose the CUPED method using old boring regression language. Let’s see what will hapen!

2 Old regression friend

To simplify the use of regression language I will start with one dataset of n observations with three variables:

- w_i the indicator of treatment: $w_i = 1$ for the treated group and $w_i = 0$ for the untreated group.
- x_i any covariate that is a-priori independent with treatment indicator w_i.
- y_i the target variable that is probably dependent both with w_i and x_i.

Using regression language CUPED is a two step procedure:

Step one. Estimate the regression

$$\hat{y}_i = \hat{\gamma}_1 + \hat{\gamma}_2 w_i + \hat{\theta} x_i$$

using OLS.

Calculate semiresidual $r_i = y_i - \hat{\theta} x_i$. I call this r_i ’semiresudual’ as classic residual in econometrics is

$$\hat{u}_i = y_i - \hat{y}_i = y_i - (\hat{\gamma}_1 + \hat{\gamma}_2 w_i + \hat{\theta} x_i).$$

Honestly speaking Deng is not very explicit which regression should be used in the first step. On the page three the theoretical unknown θ is used.

So one may also consider a simplier alternative regression

$$\hat{y}_i = \hat{\gamma}_1 + \hat{\theta} x_i.$$

I will discuss why I prefer the inclusion of w_i as regressor in the first step.

Step two. Estimate regression

$$\hat{r}_i = \hat{\beta}_1 + \hat{\beta}_2 w_i$$

using OLS.

Use classical standard errors to build confidence interval for β_2.

Why this two-step procedure is better than just plain old multiple regression

$$\hat{y}_i = \hat{\gamma}_1 + \hat{\gamma}_2 w_i + \hat{\theta} x_i$$

with confidence interval for γ_2 build with classic standard errors?

3 Comparison with multivariate regression

Let's talk about numeric estimates withoud assumptions at all.

[the proof of equality]

Let's start easy first. No heteroscedasticity and no interaction between treatment w_i and covariate x_i. Correctly specified linear model.

Assume that the true model is

$$y_i = \gamma_1 + \gamma_2 w_i + \theta x_i + u_i.$$

The observations are independent and identically distributed with finite forth moments. The error term u_i satisfies $\mathbb{E}(u_i \mid X) = 0$, $\mathrm{Var}(u_i \mid X) = \sigma^2$.

[here goes the picture]

It is well known in econometrics that OLS estimator $\hat{\gamma}_2$ is unbiased and consistent in this case. As $\hat{\beta}_2$ estimate from second step is exactly equal to $\hat{\gamma}_2$ the same result applies.

And what about standard errors?

4 Toy problem to understand the difference

5 Heteroscedasticity case

6 Unanswered questions