Abstract

I try to give honest statistical background to the CUPED method. This allows us 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 expectation 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 $\theta = \text{Cov}(Y, X)/\text{Var}(X)$ 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 decided 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 1a. Estimate the following regression using OLS:

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

Step 1b. Calculate the semiresidual $r_i = y_i - \hat{\theta} x_i$.
I call this r_i 'semiresudual' as classic residual in econometrics is

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

Step 2a. Estimate the second regression using OLS:

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

Step 2b. Use classical standard errors to build confidence interval for β_2.

Why this two-step procedure is better than just the multivariate regression of the first step with confidence interval for γ_2?

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.

3 Comparison with multivariate regression

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

Theorem 1. *The proposed two step procedure gives exaclty the same estimates for both steps: $\hat{\beta}_1 = \hat{\gamma}_1$, $\hat{\beta}_2 = \hat{\gamma}_2$, both residual vectors are also equal, $y_i - \hat{y}_i = r_i - \hat{r}_i$.*

The residual sums of squares are equal for both steps

$$RSS = \sum (y_i - \hat{y}_i)^2 = \sum (r_i - \hat{r}_i)^2.$$

Proof. Just remark that multivariate regression minimizes $\sum(y_i - \hat{y}_i)^2$ with respect to the three parameters $\hat{\gamma}_1$, $\hat{\gamma}_2$, $\hat{\theta}$ simultaneously.

Two step procedure has the same optimal $\hat{\theta}$ by design and on step two minimizes the same sum with respect to $\hat{\beta}_1$ and $\hat{beta}_2$ with $\hat{\gamma}$ fixed at the optimal value. □

Hence all the difference between the two approaches lies in standard errors and confidence intervals. To discuss the validity of standard errors we'll need some assumptions.

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 $\mathbb{E}(u_i \mid X) = 0$, $\text{Var}(u_i \mid X) = \sigma^2$.

[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?

Here I present a small summary of classic result in econometrics. I find this useful for the completeness of the text.

Let's introduce the matrix of all predictors and its centered version.

$$X = \begin{pmatrix} 1 & w_1 & x_1 \\ 1 & w_2 & x_2 \\ \vdots & \vdots & \vdots \\ 1 & w_n & x_n \end{pmatrix}, X_c = \begin{pmatrix} w_1 - \bar{w} & x_1 - \bar{x} \\ w_2 - \bar{w} & x_2 - \bar{x} \\ \vdots & \vdots & \vdots \\ w_n - \bar{w} & x_n - \bar{x} \end{pmatrix}.$$

Under our assumptions the true variance of estimates is

$$\text{Var}\left(\begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \hat{\theta} \end{pmatrix} \mid X \right) = \sigma^2 (X^T X)^{-1}.$$

The default estimator of unknown σ^2 in multivariate regression with k estimated coefficients is given by

$$\hat{\sigma}^2 = \frac{RSS}{n-k} (X^T X)^{-1}.$$

Hence in our case the default estimate of the variance of $\hat{\beta}_2$ will be found in the matrix

$$\widehat{\text{Var}}\left(\begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \hat{\theta} \end{pmatrix} \mid X \right) = \frac{RSS}{n-3} (X^T X)^{-1}.$$

If we ignore the $\hat{\gamma}_1$ and focus just on $\hat{\gamma}_2$ and $\hat{\theta}$ then our regression is equivalent to regression in centered variables.

$$\text{Var}\left(\begin{pmatrix} \hat{\gamma}_2 \\ \hat{\theta} \end{pmatrix} \mid X \right) = \sigma^2 (X_c^T X_c)^{-1}.$$

This is the FWL theorem or this result may be proven by block-matrix inversion.

Let's look closer on the conditional covariance matrix. I rewrite it as

$$\text{Var}\left(\begin{pmatrix} \hat{\gamma}_2 \\ \hat{\theta} \end{pmatrix} \mid X \right) = \frac{\sigma^2}{n-1} \left(\frac{X_c^T X_c}{n-1} \right)^{-1}.$$

And $\left(\frac{X_c^T X_c}{n-1} \right)$ is exactly the sample covariance matrix of the treatment indicator w_i and covariate x_i,

$$\left(\frac{X_c^T X_c}{n-1} \right) = \text{sVar}\left(\begin{pmatrix} w \\ x \end{pmatrix} \right) = \begin{pmatrix} \text{sVar}(w) & \text{sCov}(w,x) \\ \text{sCov}(w,x) & \text{sVar}(x) \end{pmatrix} = \begin{pmatrix} \frac{\sum(w_i - \bar{w})^2}{n-1} & \frac{\sum(w_i - \bar{w})(x_i - \bar{x})}{n-1} \\ \frac{\sum(w_i - \bar{w})(x_i - \bar{x})}{n-1} & \frac{\sum(x_i - \bar{x})^2}{n-1} \end{pmatrix}.$$

The sample covariance matrix with the growing number of observations will tend in probability to the true covariance matrix

$$\operatorname*{plim}_{n \to \infty} \text{sVar}\left(\begin{pmatrix} w \\ x \end{pmatrix} \right) = \text{Var}\left(\begin{pmatrix} w_i \\ x_i \end{pmatrix} \right) = \begin{pmatrix} \text{Var}(w_i) & \text{Cov}(w_i, x_i) \\ \text{Cov}(w_i, x_i) & \text{Var}(x_i) \end{pmatrix}.$$

What we know about the true covariance matrix?

The treatment is assumed to be independent of the covariate, hence $\text{Cov}(w_i, x_i) = 0$. If the probability of being treated is equal to $1/2$ then $\text{Var}(w_i) = 1/4$.

The classical estimate of covariance matrix is given by

$$\widehat{\text{Var}}\left(\begin{pmatrix}\hat{\gamma}_2\\ \hat{\theta}\end{pmatrix} \mid X\right) = \frac{RSS/(n-3)}{n-1}\left(\frac{X_c^T X_c}{n-1}\right)^{-1}.$$

Let's replaced this classic estimator of covariance matrix by an estimator that uses a-priori information about zero covariance $\text{Cov}(w_i, x_i) = 0$:

$$\widehat{\text{Var}}_{cuped}\left(\begin{pmatrix}\hat{\gamma}_2\\ \hat{\theta}\end{pmatrix} \mid X\right) = \frac{RSS/(n-3)}{n-1}\begin{pmatrix}\text{sVar}(w) & 0\\ 0 & \text{sVar}(x)\end{pmatrix}^{-1}.$$

The inversion of a diagonal matrix is easy and we get

$$\widehat{\text{Var}}_{cuped}(\hat{\gamma}_2) = \frac{RSS/(n-3)}{n-1}\,\text{sVar}(w)^{-1}.$$

How this compares with the CUPED two-step procedure?

The default variance estimate is equal to

$$\widehat{\text{Var}}(\hat{\beta}_2) = \frac{RSS}{n-2}\frac{1}{\sum(w_i - \bar{w})^2} = \frac{RSS/(n-2)}{n-1}\,\text{sVar}(w)^{-1}.$$

And they match almost perfectly! Now we understand what CUPED two-step procudure does!

Ignoring the offset factor $(n-2)/(n-3)$ the CUPED replaces the sample covariance of w and x by zero in the construction of a confidence interval for the effect.

It seems intuitive that using a-priori information is always better than ignoring it. But the reality may seem a little bit surprising.

Imagine that Alice uses a-priori information about zero covariance of treatment and covariate. And Bob uses an old multivariate regression that ignores this fact. They both construct 95% confidence intervals for the effect.

We have proven the point estimates of Alice and Bob are exactly equal. As Alice uses more information her intervals should be shorter. And we are in a contradiction now. If Alice's interval is always shorter than Bob's interval then they cannot both cover unknown parameter with probability 95%.

The use of a-priori information about zero covariance will sometime lead to a wider confidence interval for the effect.

Let's study this counter-intuitive result with a toy problem.

4 Toy problem to really understand the difference

Assume y_i are independent and normally distributed $\mathcal{N}(\mu, 1)$. Alice knows the true variance $\sigma^2 = 1$ and Bob uses the estimate

$$\hat{\sigma}^2 = \text{sVar}(y) = \frac{\sum(y_i - \bar{y})^2}{n-1}.$$

They both build 95% confidence intervals for μ. Alice uses normal distribution and Bob uses t-distribution with $(n-1)$ degrees of freedom.

For simplicity let's assume that $n = 10$.

Alice's confidence interval is

$$[\bar{y} - 1.96 \cdot 1/\sqrt{10}; \bar{y} - 1.96 \cdot 1/sqrt10].$$

Bob's confidence interval is

$$[\bar{y} - 2.26 \cdot \hat{\sigma}/\sqrt{10}; \bar{y} - 2.26 \cdot \hat{\sigma}/sqrt10].$$

Here 1.96 and 2.26 are 0.975 quantiles of standard normal and t distribution with 9 degrees of freedom.

We immediately see that sometimes Bob's interval will be shorter and sometimes Alice's interval will.

When the number of observations will tend to infinity the quantiles will coinced, $\text{plim}\,\hat{\sigma} = 1$, so the intervals will be exactly the same.

5 Heteroscedasticity case

6 Unanswered questions