



Analyzing Switchback Experiments by Cluster Robust Standard Error to Prevent False Positive Results

📅 September 11, 2019

🕒 10 Minute Read

📖 Machine Learning

38



Yixin Tang



Caixia Huang

Within the dispatch team of DoorDash, we are making decisions and iterations every day ranging from business strategies, products, machine learning algorithms, to optimizations. Since all these decisions are made based on experiment results, it is critical for us to have an experiment framework with rigor and velocity. Over the last few years, we have established Switchback Framework as the foundation for most of our dispatch experiments. On top of that, we explored ex-post methods like Multilevel modeling (MLM) to improve experiment rigor under certain assumptions.

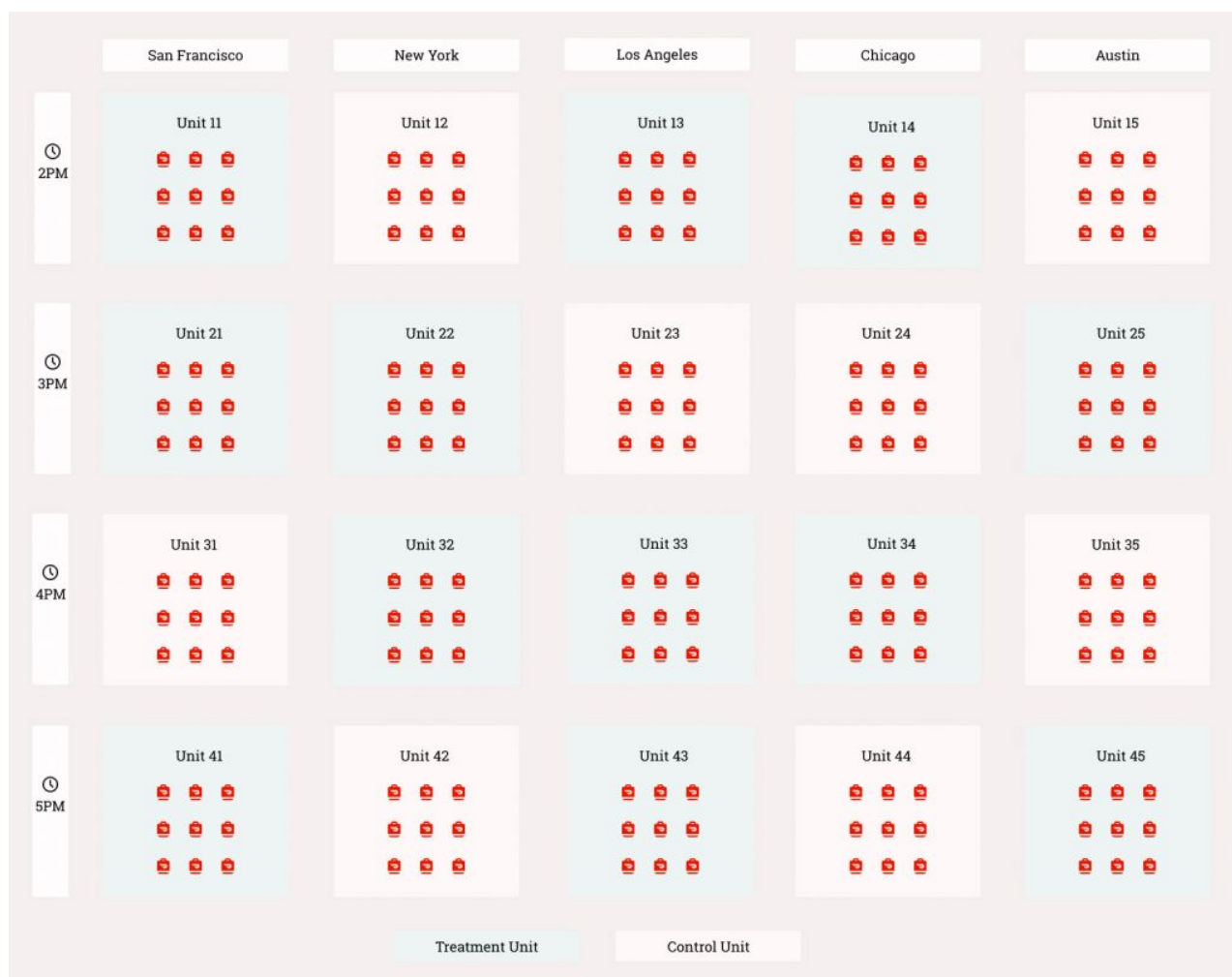
In this blog post, we will talk about how we use another statistical method: Cluster Robust Standard Error (CRSE) in Switchback Framework. We present the problem of within-cluster correlation in data, and show MLM can be biased when certain assumptions do not hold. Then we discuss the different types of robust standard error estimations given error correlations, how we use them in our switchback testing, and evaluation results based on cluster-bootstrap simulations. Finally, we discuss how we use CRSE in Diff-in-Diff to improve rigor and some tips and caveats we found valuable in practice.

Problem of clustering

Introduction

Clustering is a phenomenon that individuals are grouped into clusters and individuals within the same cluster are correlated. As a result, when applying regression model in practice, errors are independent across clusters but correlated within clusters. One classic example of clustering is geographic region cluster where individuals in the same city/state are correlated. At DoorDash, clustering is very common. For example, deliveries are clustered on different regions or time as deliveries in the same region/hour would share similar characteristics like delivery duration, dasher efficiency, etc. Orders from the same merchant can form a cluster because they have similar food preparation time. Deliveries completed by the same dashers can form a cluster because they might have similar travel or parking time.

In dispatch experiment, we use switchback testing that randomizes on regional-time “units”, i.e. all deliveries and Dashers in each unit are exposed to the same type of algorithm. The main reason we use switchback is to deal with network effects which have been elaborated in the prior blog post [Switchback Tests and Randomized Experimentation Under Network Effects at DoorDash](#).



When the desired estimated treatment effect is usually at delivery level, clustering introduces problems in the regression model, as the error terms are correlated within each regional-time unit. The correlation within unit would lead to underestimated standard error and p-value, and hence higher false positive rate. In one of our previous post [Experiment Rigor for Switchback Experiment Analysis](#), we quantitatively measured how much the standard error is underestimated under OLS. The results show that the false positive rate is as high as 0.6 in our switchback world. To mitigate this issue, we will need to resolve the clustering problem and thus correctly estimate the standard error of the treatment effect.

Pitfall of using MLM

To solve the clustering issue in the past, we applied MLM model on some switchback experiments. MLM, also known as linear mixed-effect model

is a statistical model whose parameters can be either fixed or random effect, and can vary more than one level. Although, simulation result shows that it is much more efficient compared to most other models, MLM does not apply to all of the experiment scenarios at DoorDash. For some experiments, we found that MLM can give us contradictory results compared to OLS estimates. For example, MLM estimator produced a statistical significant result of -0.22 treatment effect, while the difference between average treatment and average control is 0.26 , which is hard to interpret.

Method	Treatment Effect	p-value
MLM	-0.22	$6e-5$
OLS	0.26	$< 2e-16$

The main reason why MLM can give us a biased result is due to the misspecification of model. In MLM, we assume normal distributed randomness of heterogeneity, which is not always the case in reality. When this assumption does not hold, the result can be biased. A more broad representation of this class of methods can be described as two stages: First by adding some additional constraints and assumptions when estimating the error correlation, then using FGLS to estimate the original model. The success depends on if we can consistently estimate the error. If all assumptions hold for the data in an experiment, then these model based approaches usually have a high power.

At DoorDash, dispatch experiments are quite diverse, ranging from time-delayed effects and route optimization, to parameter tuning, so there is usually no unique assumption that holds for all cases. Hence, for a new experiment that we do not have much prior knowledge, we need a “model free” method that can give a correct standard error estimation while not requiring any specification of error correlation.

Cluster Robust Standard Error

Introduction to CRSE

Cluster robust standard error (CRSE) can account for heteroskedasticity and correlations within clusters, while not making strong assumptions for error correlation. In this section, we will walk you through the development of CRSE from the OLS “nominal” variance that we are most familiar with.

From the well-known formula of the solution to OLS, we can write the beta and variance of beta as:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$Var(\hat{\beta}) = (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}$$

where Ω is the covariance matrix of the errors $V(\varepsilon|X)$.

When we have the assumption that errors are independent and identically distributed (iid), becomes a diagonal matrix with all elements equal σ^2 . Then we can get the variance of estimated treatment effect:

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

When the errors are heteroskedastic, matrix becomes a diagonal matrix with all elements different. We can write the “meat” of the “sandwich” as below, and the variance is called heteroscedasticity-consistent (HC) standard errors.

$$X^T \Omega X = \sum_{i=1}^n \sigma_i^2 X_i^T X_i$$

When it comes to cluster standard error, we allow errors can not only be heteroskedastic but also correlated with others within the same cluster. Given such structure, Ω becomes a block-diagonal matrix, where ϵ_i is the error vector in each cluster.

$$\Omega = \begin{bmatrix} E(\epsilon_1 \epsilon_1^T | X_1) & & \\ & \ddots & \\ & & E(\epsilon_g \epsilon_g^T | X_g) \end{bmatrix}$$

And the “meat” of “sandwich” becomes:

$$X^T \Omega X = \sum_{g=1}^G X_g^T E(\epsilon_g \epsilon_g^T | X_g) X_g$$

Where G is the number of clusters. To better illustrate, below is a visualization of the variance matrix Ω , taking switchback experiment at DoorDash as an example. In this simplified example, we have three regional-time units: San Francisco 1pm, New York 3PM, and Los Angeles 4AM. In each of the three units, there are only three deliveries. From the definition of CRSE above, errors of deliveries are correlated within units but independent across units.

		San Francisco 🕒 1PM	New York 🕒 3PM	Los Angeles 🕒 4AM
		🛒 🛒 🛒	🛒 🛒 🛒	🛒 🛒 🛒
San Francisco 🕒 1PM	🛒	ϵ_{11}^2 $\epsilon_{11}\epsilon_{12}$ $\epsilon_{11}\epsilon_{13}$	0 0 0	0 0 0
	🛒	$\epsilon_{12}\epsilon_{11}$ $\epsilon_{12}\epsilon_{12}$ $\epsilon_{12}\epsilon_{13}$	0 0 0	0 0 0
	🛒	$\epsilon_{13}\epsilon_{11}$ $\epsilon_{13}\epsilon_{12}$ ϵ_{13}^2	0 0 0	0 0 0
New York 🕒 3PM	🛒	0 0 0	ϵ_{21}^2 $\epsilon_{21}\epsilon_{22}$ $\epsilon_{21}\epsilon_{23}$	0 0 0
	🛒	0 0 0	$\epsilon_{22}\epsilon_{21}$ $\epsilon_{22}\epsilon_{22}$ $\epsilon_{22}\epsilon_{23}$	0 0 0
	🛒	0 0 0	$\epsilon_{23}\epsilon_{21}$ $\epsilon_{23}\epsilon_{22}$ ϵ_{23}^2	0 0 0
Los Angeles 🕒 4AM	🛒	0 0 0	0 0 0	ϵ_{31}^2 $\epsilon_{31}\epsilon_{32}$ $\epsilon_{31}\epsilon_{33}$
	🛒	0 0 0	0 0 0	$\epsilon_{32}\epsilon_{31}$ $\epsilon_{32}\epsilon_{32}$ $\epsilon_{32}\epsilon_{33}$
	🛒	0 0 0	0 0 0	$\epsilon_{33}\epsilon_{31}$ $\epsilon_{33}\epsilon_{32}$ ϵ_{33}^2

As we mentioned earlier, OLS, by neglecting within-cluster correlation, severely underestimates the variance. The formula below provides a useful approximation of how much the default OLS variance estimate should be inflated:

$$\tau_k \simeq 1 + \rho_{x_k} \rho_u (\bar{N}_g - 1)$$

where the first ρ is a measure of within-cluster correlation of regressor, the second ρ is the within-cluster error correlation, and N_g is the average cluster size.

In dispatch switchback experiment, since the regressor is experiment bucket and it remains constant within a cluster, they are perfectly correlated. The model errors also have high correlation because deliveries within the regional-time units have very similar characteristics. As for cluster size, it is considerably large in popular region during peak time. Therefore, in our case, the default OLS variance estimator is severely downward biased and much smaller than CRSE.

Simulation

To validate that cluster robust standard error correctly estimate the standard error in dispatch experiments, we ran through a simulation procedure where we assign normal treatment effect to 50% of randomly selected deliveries from bootstrapped data and apply multiple different methods. The methods we used are:

1. Delivery level regression
2. Regional-time unit level regression
3. Delivery level with CRSE on regional-time unit
4. Delivery level with CRSE on regional-time unit and added market as fixed effects

Simulation Results

Here are the simulation results using the above mentioned methods. We use WithinCI, the percentage which computed the confidence interval that covers the true mean, to measure the validity of method; and Power, the percentage which we actually detect the difference with statistical significance when there is any, to evaluate and compare across methods.

Simulation Result		
Method	WithinCI	Power
$y_{u,d} \sim T_u$	0.66	0.89
$y_u \sim T_u$	0.96	0.26
$y_{u,d} \sim T_u(\text{cluster} = u)$	0.95	0.52
$y_{u,d} \sim T_u + M_{u,d}(\text{cluster} = u)$	0.95	0.64
$\log(y_{u,d}) \sim T_u + M_{u,d}(\text{cluster} = u)$	0.94	0.55

$y_{u,d}$: duration of delivery d in unit u

y_u : average duration of deliveries in unit u

T_u : treatment or control assignment of an experiment unit

$M_{u,d}$: market of delivery d in unit u

From the table, we can see that when we conduct test on delivery level without CRSE, WithinCI is much smaller than 0.95, which means it severely underestimates variance and confidence interval, and hence cannot be used. Unit level test has good validity from the evidence that WithinCI is close to 0.95. However, the power is very low and the sample size becomes much smaller after taking average over each unit. More importantly, taking average over each unit will weigh each units equally. From business consideration, however, we would want to put equal weight on each delivery instead of each region-time unit. After using CRSE on region-time unit at delivery level, the simulation result shows that the standard error is correctly estimated, with an improvement on power. We also experimented adding region fix effect or transform the metric on top of CRSE application, the result shows a large power improvement with fix effect. Although the result is not shown here, we also simulated using MLM on the same data. CRSE again proves that it is a more robust method on our switchback experiments.

Other Applications and Implementation

Implementation Caveats of CRSE

An important assumption of cluster robust standard error is that the number of clusters goes to infinity. Adjustment is common on finite

cluster scenarios. For example, in stata, Instead of using `ug`, `cug` in formula (2) can be used rather than `ug`, where

$$c = \frac{G}{G-1} \frac{N-1}{N-K}$$

There are lots of software packages and libraries that implement CRSE and they can be slightly different. When we applied CRSE, in order to check if the specific implementation is suitable, we use cluster bootstrap to obtain a “true” cluster robust standard error and compare it with the one we implemented. In cluster bootstrap, re-sampling is done on the cluster level.

- For i in number of bootstrap samples N :
 - Generate m clusters $\{(X_1, y_1), (X_2, y_2), \dots (X_m, y_m)\}$ by resampling with replacement m times from the original data
 - Compute estimator β from generated data i
- Collect $\{\beta_i, i=1,2,3,\dots,n\}$ and compute the variance

We expect our implementation of CRSE to have a close enough value to cluster bootstrapped result. One major reason we do not directly apply cluster bootstrap variance in experiment is the speed. Generating CRSE by bootstrap enough times on big dataset can take a fairly long time.

Application of CRSE in Diff-in-Diff

As mentioned earlier, clustering is a very common phenomenon in experiments at DoorDash, so cluster robust standard error can be used in many experimentation analysis. Particularly, cluster robust standard error is used in Diff-in-Diff experiments. At DoorDash, Diff-in-Diff experiment is usually applied when we measure treatment effect at aggregate geographic level. For example, we experiment on the effect of a marketing campaign by assigning the marketing campaign to some treatment states and use some states as control.

$$y_{it} = \gamma_{s(i)} + \lambda_t + \delta I(s(i) \text{ is the treatment group, } t \text{ is in the after period}) + \varepsilon_{it}$$

Where i is individual, t is time, and $s(i)$ is the market that individual i is in. The errors are highly correlated with each other over time within one market. Therefore, CRSE is necessary in diff-in-diff to obtain a correct estimate of standard error. In this example, since the standard error is clustered on market level, CRSE should be applied at market level.

Conclusion

In switchback experiments where data are grouped into clusters, CRSE is a robust and straightforward way to get unbiased statistical inference. As we tackled this problem on the dispatch team, we were able to find applications of CRSE on many other experiments in Consumers, Dashers, and Merchants as well. Success to control for cluster error correlation is a big step forward in iterating on our marketplace experiments with more confidence. As we move on, we will keep improving our switchback experiment framework, especially on interaction effects, sequential testing, and joint test.

Special thanks to Professor Navdeep S.Sahni, Sifeng Lin, Richard Hwang, and the whole dispatch team at DoorDash for their help in publishing this post.

References

- A. Colin Cameron & Douglas L. Miller, (2015). A Practitioner's Guide to Cluster-Robust Inference. Journal of Human Resources, University of Wisconsin Press, vol. 50(2), 317-372.
- Freedman, D. A. (2008). On regression adjustments to experimental data. Advances in Applied Mathematics, 40(2), 180-193.

Comments

