DOORDASH ENGINEERING

Experiment Rigor for Switchback Experiment Analysis

February 20, 2019 14 Minute Read Machine Learning 40

 **Carla Sneider**

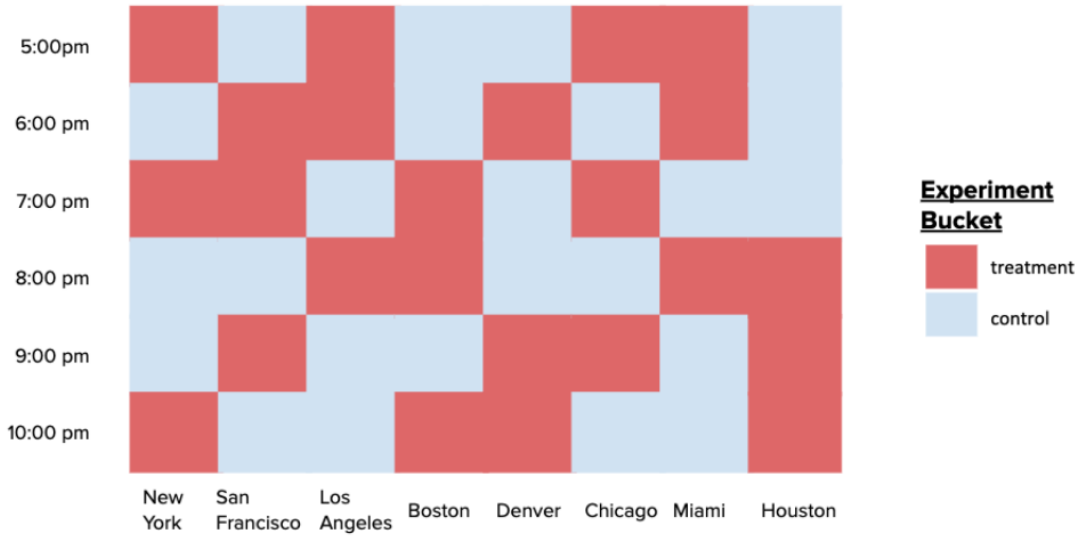
 **Yixin Tang**

At DoorDash, we believe in learning from our marketplace of Consumers, Dashers, and Merchants and thus rely heavily on experimentation to make the data-driven product and business decisions. Although the majority of the experiments conducted at DoorDash are A/B tests or difference-in-difference analyses, DoorDash occasionally relies on a type of experimentation internally referred to as "switchback testing". Switchback tests randomize experiment buckets on geographic region and time "units" rather than users to reduce the impact of dependencies in observations on our experiment results. Although the implementation of switchbacks is similar to A/B tests, two characteristics of their datasets add complexity to their analysis: (1) the nested data structure and (2) the small number of more independent "units" available for analysis. This blog post will discuss how we determined the most accurate approach to analyzing switchback experiments and improved the statistical power of our experiments. In so, we have been able to iterate on products decisions more confidently and 30% more quickly.

Introduction to Switchbacks

While A/B testing is commonly used at DoorDash, it is not optimal for testing our assignment algorithm because the assignment of one delivery to a Dasher depends heavily on the outcome of another delivery's assignment to a Dasher. For example, say there are two deliveries that need to be assigned and only one Dasher is available. If we apply an algorithm change to one delivery which will assign it quicker than the standard, we risk impacting the second "control" delivery's assignment since the "treated" delivery would

most likely be assigned to the only Dasher. Switchback testing mitigates the risk of these network effects by randomizing our experiment into buckets on regions and times of day rather than Dashers or deliveries. By randomizing on these regional-time "units", all deliveries and Dashers in each unit are exposed to the same type of algorithm, in turn reducing the impact of dependencies among deliveries in two different experiment buckets. An illustration of this randomization structure is shown below.



Once randomization happens on the region-time unit level, each delivery is bucketed into to a treatment or control group based on that of its region and time, and as a result, we get a nested data structure: multiple deliveries are part one of one 'unit'.

Key Considerations in Switchback Analysis

The most straightforward way to analyze our switchback delivery data might appear to be a two-sample t-test, which directly compares the average delivery duration in our treatment experiment group versus our control group, as represented by the following regression:

$$Duration \sim bucket$$

However, using deliveries as observations violates the assumption of independence for which we tried to correct by randomizing on region-time units in the first place. We therefore tended to aggregate our results on the *unit of randomization* (i.e. regional-time unit) prior to running t-tests, as we found this provided a more accurate estimate of the average effect and variance of our experiment's treatment. More on "unit-level" analysis can be found in the prior blog post *Switchback Tests and Randomized Experimentation Under Network Effects at DoorDash*.

Still, unit-level summarization has two drawbacks. First, by first aggregating delivery durations by region-time unit, it is difficult to obtain statistically significant results due to limited sample size. For example, let's assume one regional-time unit has 20 deliveries. If we analyze our experiment on regional and time units rather than deliveries, the 20 deliveries available for a power analysis become only 1 unit from which we can get statistical power. Second, unit-level analysis does not correctly account for instances in which our algorithm change might have distinct effects on delivery times in regional-time units with few deliveries (i.e. 1am) versus those with many deliveries (i.e. 5pm). In these situations, it is difficult for us to confidently conclude whether an algorithm change reduced or increased our delivery durations on average, as unit-level and delivery-level results can directionally diverge.

Therefore, we wanted to test if we could improve our treatment effect and variance estimates by analyzing our experiment results using two other methods: unit-level t-tests with variance reduction covariates and multilevel modeling (MLM), both briefly discussed below.

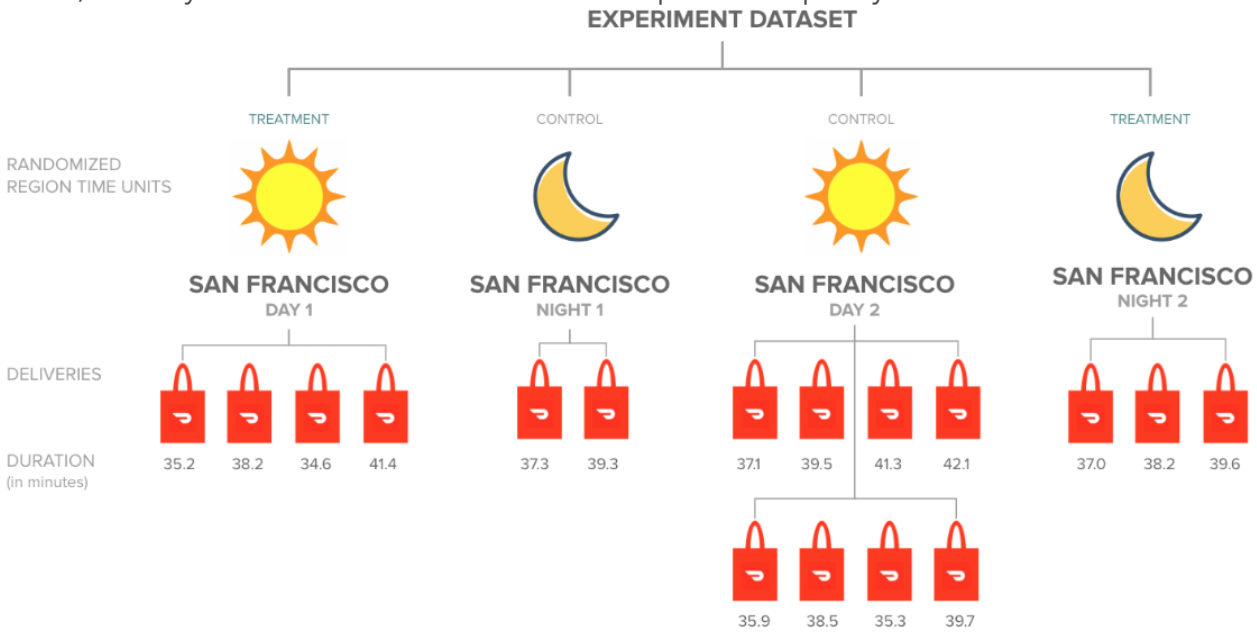
Unit-level t-tests with variance reduction were appealing because they would address our statistical power concern of unit-level analysis. By adding covariates X_i that satisfy the conditional independence requirements (expressed below), we looked to explain away some of the variation in delivery times unrelated to the experiment bucket in our experiments. Some covariates we used included time of day, week number, region, and total order value.

$$\{Duration_{0i}, Duration_{1i}\} \perp\!\!\!\perp Bucket_i \mid X_i$$

$$Duration \sim bucket + X$$

MLM was appealing because our dataset is naturally nested (meaning correlations exist between delivery durations from the same day, same time of day, and same region), and MLM is a tried and tested approach for interpreting nested datasets.

To help understand MLM, let's look at an example of a switchback experiment run only in San Francisco, illustrated below. Note, although switchbacks randomize across regions and times, we only randomize on times in this example for simplicity's sake:



Running MLM can be viewed as a combination of two stages. The first stage runs a linear regression on each region-time unit to explain delivery-level variation using the equation $Duration \sim 1 + bucket + \varepsilon$ for every region and time . In the example above, this would mean running four of such regressions, two of which are illustrated below:

San Francisco, Day 1

$$\begin{pmatrix} 35.2 \\ 38.2 \\ 34.6 \\ 41.4 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_{0\_SF,\ Day1} \\ \beta_{1\_SF,\ Day1} \end{pmatrix} + \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}$$

Delivery Duration *Intercept* *expt_bucket* *[1 = treatment]* *Error term*

San Francisco, Night 1

$$\begin{pmatrix} 35.2 \\ 38.2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_{0\_SF,\ Night1} \\ \beta_{1\_SF,\ Night1} \end{pmatrix} + \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \end{pmatrix}$$

Delivery Duration *Intercept* *expt_bucket* *[0 = control]* *Error term*

The second-stage regression uses the coefficients determined in the first-stage regression to explain the variability between the region-time units, as shown in the following regressions where β_0 and β_1 represent the average intercept and average treatment effect respectively and þ_0_RegionTime and þ_1_RegionTime represent random-effects. The random effects make it possible for each region-time unit to have different intercepts to explain delivery durations.

To Calculate Average Treatment Effect

$$\begin{bmatrix} \beta_{1\_SF,\ Day1} \\ \beta_{1\_SF,\ Night1} \\ \beta_{1\_SF,\ Day2} \\ \beta_{1\_SF,\ Night2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ b_{1\_SF,\ Day1} \\ b_{1\_SF,\ Night1} \\ b_{1\_SF,\ Day2} \\ b_{1\_SF,\ Night2} \end{bmatrix} + \begin{bmatrix} u_0 \\ u_1 \\ u_3 \\ u_4 \end{bmatrix}$$

Expt Coefficient for Region-Times — *Intercept* — *Dummy-variables for region-time units* — *Error term*

To Calculate Average Intercept

$$\begin{bmatrix} \beta_{0\_SF,\ Day1} \\ \beta_{0\_SF,\ Night1} \\ \beta_{0\_SF,\ Day2} \\ \beta_{0\_SF,\ Night2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ b_{0\_SF,\ Day1} \\ b_{0\_SF,\ Night1} \\ b_{0\_SF,\ Day2} \\ b_{0\_SF,\ Night2} \end{bmatrix} + \begin{bmatrix} u_0 \\ u_1 \\ u_3 \\ u_4 \end{bmatrix}$$

Expt Coefficient for Region-Times — *Intercept* — *Dummy-variables for region-time units* — *Error term*

Although MLM is not exactly a two-staged regression, it is very similar. In combining the two stages into one formula (as shown below), MLM can account for individual delivery durations in each unit when estimating unit level average intercept β_0 and treatment effect β_1.

$$Duration_{Region\_Time,\,i} \sim (\beta_0 + b_{0\_Region\_Time}) + (\beta_1 + b_{1\_Region\_Time}) * expt\_bucket_{Region\_Time,\,i} + \varepsilon_{Region\_Time,\,i}$$

Region_Time represents the region-time unit (i.e. SF Day 1 in the above example) and *i* represents a delivery. When choosing MLM as an alternative approach, we hypothesized MLM would not only be more statistically powerful than unit-level analysis because it includes individual deliveries in its calculations of treatment effects and variance; we also believed MLM would account for dependencies among deliveries when calculating variance, as will be proven in the results.

Objective For Improving Analysis of Switchbacks

In improving our analysis of switchbacks, we set out to reduce the false positives and the false negatives associated with our experiment analysis because we want to trust that (A) we do not incorrectly conclude a treatment effect exists when none exists, and (B) detect a change in our key metric when our treatment changes that metric.

		ACTUAL VALUE	
PREDICTED VALUE		**Positive**	**Negative**
	Positive	True Positive, Power	False Positive, Type I Error
	Negative	False Negative, Type II Error	True Negative

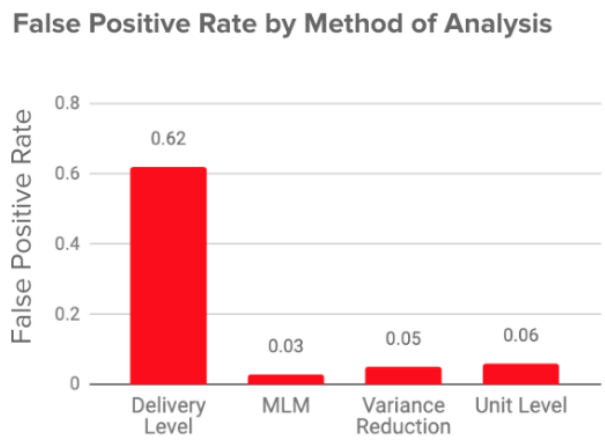
Specifically, **false positives** occur if we conclude a treatment effect exists when in reality there does not. For example, if we run an AA test data and see a p-value less than 0.05, we would incorrectly reject the "null hypothesis" that there was no treatment effect where in fact the null hypothesis is true. **False negatives** occur if we fail to conclude a treatment effect exists when one does exist. For example, if we run an A/B test and, given we have enough sample size to based on our power calculations, the p-value is greater than 0.05 when a known treatment effect exists, we incorrectly accept the null hypothesis. Notice how sample size is a component here, as false negatives relate to statistical power.

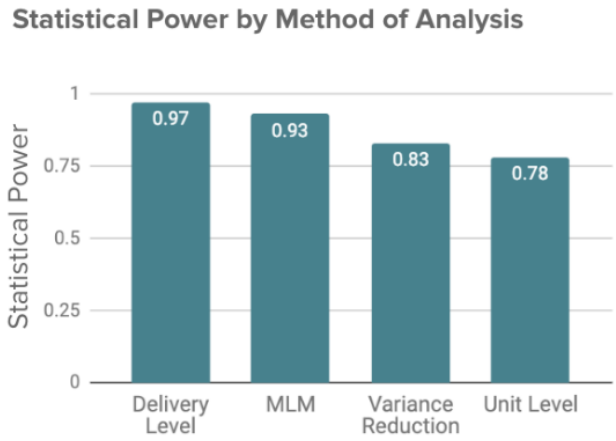
Comparing Analysis Alternatives Through Data Simulation

We evaluated our models on switchback data using frequentist statistics perspective, meaning we assumed (A) there exists a fixed treatment effect (i.e. 0 for AA tests) and (B) the estimate of that treatment effect relies on what data is available. To implement our approach, we took a few weeks of company-wide delivery data and simulated many versions of an artificial experiment on the same data. These simulations were done by grouping our delivery data into time-region units and randomizing the units into treatment and control, as would be done in a switchback, meaning every delivery would get the experiment group of its unit. For our AA tests, we kept everything as is, and for our A/B tests, we added a known normally distributed treatment effect to all deliveries in the treatment group units; the normal distribution was a simplifying assumption. Next, we tested our 4 methods of experiment analyses (delivery-level analysis, unit-level analysis, unit-level with variance reduction, and multilevel modeling) on each of the simulated AA and AB tests and recorded the mean treatment effect and p-value for each of our simulations. Finally, we graphed the distributions of p-values and treatment effects for our final step: calculating false positive rate, false negative rate, and bias.

Simulation Findings: MLM versus other methods

From our simulations, we determined MLM was the optimal method of analysis. As shown below, MLM had the lowest false positive rate (0.03) and the second highest statistical power (0.93).





There are several advantages to applying MLM for switchback experiment analysis. First, although delivery-level t-tests have the highest statistical power, MLM provides a more dependable treatment effect estimate given delivery-level's unfavorably high false

positive rate (0.62). The high delivery-level false positive rate stems from the nested structure of our switchback datasets: the more correlated our deliveries are within the same regional-time unit, the more we underestimate the variance of our clustered dataset when using traditional t-tests. To concretely understand why delivery-level t-tests underestimate the variance of our switchback experiment results, it's helpful to compare the t-test variance estimate to that of the data's true variance using the variance formula for ordinary least squares (OLS): *Duration ~ β_0 + β_hat + bucket*. Assuming *Var(β_hat)* is the true variance of the treatment effect, *n* is our sample size, and *ϱ_e* is our intraclass correlation coefficient (ICC) (detailed below), the variance for our treatment effect estimate is:

$$Var_c(\hat{\beta}) = \frac{Var(\hat{\beta})}{1 + (n-1)\rho_e}$$

It is easy to see from the above equation that we correctly estimate the variance of our treatment effect only when the ICC (denoted by *ϱ_e*) equals 0. This is not possible with our switchback data because when data has a nested structure, the ICC is always greater than 0. This is because the ICC quantifies the proportion of total variance that comes from the two components of variance in a nested dataset: between-group variance (i.e. the variance in regional-time units) and within-group variance (i.e. the variance in delivery durations within regional-time units). More concretely, the ICC is calculated as follows, where the numerator denotes the between-group variance and denominator denotes the total variance**:**

$$\rho_e = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2}$$

If the numerator is zero in the above equation, that means we see no between-group variance exists, meaning our data is not nested and we can in theory use t-tests to analyze our data. However, when any between-group correlations exists, as in the case of our switchback dataset, we know our dataset is nested and t-tests are no longer appropriate. Therefore, by using a t-test to interpret our switchback dataset, it makes sense that we get a large t-statistic and an incorrectly small p-value:

$$t = \frac{\hat{\beta}}{\sqrt{Var(\hat{\beta})}}$$

MLM corrects for the underestimation of variance in nested datasets by incorporating both between-group variance and within-group variance in its variances estimates, as seen by MLM's low false positive rate. Second, although unit-level analysis improve our estimation of variance by eliminating the correlation among delivery observations, MLM has much higher statistical power than unit-level analysis. In fact, by using MLM, we cut the time necessary to get statistically significant results by ~30% compared to unit-level analyses. This is because the sample size available for unit-level analysis is substantially lower than that available in MLM, due to the inclusion of delivery-level and unit-level data in MLM. With fewer observations, we get lower t-statistics and thus p-values higher than 0.05 for unit-level analysis when treatment effects exist. Third, while we can improve the power of the unit-level test by adding more covariates such as time of day (i.e. lunch or dinner) and regional area, we could not find covariates that reduced the variance of the treatment effect estimate by enough to compensate for the sample size differences between unit-level analysis and MLM. Additionally, with variance reduction regressions, we risk introducing bias to our experiment results if adding the wrong covariates (i.e. covariates that correlate to the treatment effect) or omitting necessary covariates. The details of bias are included in the APPENDIX. For example, if we include unit_id's as a dummy variables, we substantially reduce the degrees of freedom and increases variance in our estimates, whereas MLM does not have the issue when it considers unit_id as a

random effect. By including unit_id as random effect, we can model the difference in average delivery times per regional-time unit compared to that of the entire dataset in what is essentially a two-stage regression. All of these findings strongly suggest we should use MLM to analyze our switchbacks with unit_id as a random effect.

Conclusion

Obtaining experiment results in which we can be confident is a key component for shipping products that will ultimately improve the user experience for our Consumers, Dashers, and Merchants. Analyzing switchback experiments using MLM is a big step forward in iterating on our marketplace experiments more confidently and more quickly than ever before. This is just the beginning. There is still room for improvement on variance reduction for switchback data to get results more quickly through (1) using random effects in other variance reduction techniques such as sequential testing or CUPED, (2) adding additional predictors and fixed and random coefficients to MLM (using the build-up strategy to help determine which effects to include and the complexity of model), (3) using other robust variance estimation, such as cluster standard errors, Huber–White standard errors, Generalized Estimating Equations and/or (3) other solutions like Weighted least squares (WLS) on unit-level data or Block bootstrap. As we look to iterate more quickly on marketplace experiments with small treatment effects, we plan to expand upon the simulation processes used here to find a better MLM model for our experiment analysis in the future. Want to help us improve experimentation? Check out open positions on our careers page!

APPENDIX

Further Explanation of Unit-Level Variance-Reduction Results

To better understand why unit-level results with variance reduction did not perform as well as MLM, we need to recall (1) the conditional independence requirements and (2) unit-level sample size concerns which were both mentioned above in the "Key Considerations in Switchback Analysis" section of this blog. Regarding conditional independence, if we can find a set of covariates that satisfies the conditional independence requirements, this will make the selection bias zero, and hence observed difference in delivery durations will be the average treatment effect, as shown below:

$$\mathbb{E}\left[Duration_{0i} \mid Bucket_i = 1, X_i\right] - \mathbb{E}\left[Duration_{0i} \mid Bucket_i = 0, X_i\right] = 0$$

$$\begin{aligned}&\mathbb{E}\left[Duration_i \mid Bucket_i = 1, X_i\right] - \mathbb{E}\left[Duration_i \mid Bucket_i = 0, X_i\right] \\ = \quad &\mathbb{E}\left[Duration_{1i} - Duration_{0i} \mid Bucket_i = 1, X_i\right]\end{aligned}$$

where *Duration_0i* is the duration of delivery *i* had it been assigned to the control bucket, irrespective of whether it actually was assigned to control, and *Duration_1i* is the duration of delivery had it been assigned to treatment.

$$Potential\ Duration = \begin{cases} Duration_{0i}, & \text{if } Bucket_i = 0 \\ Duration_{1i}, & \text{if } Bucket_i = 1 \end{cases}$$

However, generally we do not know the perfect set of covariates satisfying the conditional independence and we will likely omit variables that are correlated with the delivery duration and one or more of its explanatory covariates in the model. This is known as the omitted variable bias. We can quantify this bias using the Omitted Variables Bias Formula

to get a sense how the result changes by adding new covariates.

$$\frac{Cov(bucket, Duration)}{V(bucket)} = \rho + \gamma\delta_{(hourOfDay, Bucket)}$$

However, even if we reduce the omitted variable bias, unit-level variance results still start from fewer observations from which to get statistical significance than MLM. Therefore, the explanatory power of covariates added to the model will have to compensate for the advantage of having both within-group and between-group variances.

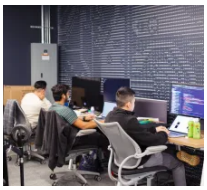
Comments

Share on:

Popular Posts

Your Deep Links Might Be Broken: Web Intents and Android 12

10 Minute Read



How to detect iOS memory leaks and retain cycles using Xcode's memory graph debugger

10 Minute Read

How DoorDash is Scaling its Data Platform to Delight Customers and Meet our Growing Demand

26 Minute Read