NBER WORKING PAPER SERIES

DESIGN AND ANALYSIS OF CLUSTER-RANDOMIZED FIELD EXPERIMENTS IN PANEL DATA SETTINGS

Bharat K. Chandar
Ali Hortaçsu
John A. List
Ian Muir
Jeffrey M. Wooldridge

Working Paper 26389
http://www.nber.org/papers/w26389

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2019

The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w26389.ack

Design and Analysis of Cluster-Randomized Field Experiments in Panel Data Settings
Bharat K. Chandar, Ali Hortaçsu, John A. List, Ian Muir, and Jeffrey M. Wooldridge
NBER Working Paper No. 26389
October 2019
JEL No. C23,C33,C5,C9,C91,C92,C93,D47

ABSTRACT

Field experiments conducted with the village, city, state, region, or even country as the unit of randomization are becoming commonplace in the social sciences. While convenient, subsequent data analysis may be complicated by the constraint on the number of clusters in treatment and control. Through a battery of Monte Carlo simulations, we examine best practices for estimating unit-level treatment effects in cluster-randomized field experiments, particularly in settings that generate short panel data. In most settings we consider, unit-level estimation with unit fixed effects and cluster-level estimation weighted by the number of units per cluster tend to be robust to potentially problematic features in the data while giving greater statistical power. Using insights from our analysis, we evaluate the effect of a unique field experiment: a nationwide tipping field experiment across markets on the Uber app. Beyond the import of showing how tipping affects aggregate market outcomes, we provide several insights on aspects of generating and analyzing cluster-randomized experimental data when there are constraints on the number of experimental units in treatment and control.

Bharat K. Chandar
Stanford University
579 Serra Mall
Stanford, CA
chandarbharatk@gmail.com

Ali Hortaçsu
Kenneth C. Griffin Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
hortacsu@uchicago.edu

John A. List
Department of Economics
University of Chicago
1126 East 59th
Chicago, IL 60637
and NBER
jlist@uchicago.edu

Ian Muir
Lyft
muir.ian.m@gmail.com

Jeffrey M. Wooldridge
Department of Economics, Michigan State University.
wooldri1@msu.edu

1 Introduction

In the past few decades field experiments in the social sciences, and in particular economics, have grown dramatically in their usage. Today, organizations as varied as firms, governments, and NGOs partner with academics to generate data via field experiments to solve their problems. As field experiments have evolved, the unit of randomization has expanded beyond the individual level. For example, cities, villages, regions, or even countries have been the target for interventions of both scientists and practitioners. Understanding the benefits and costs of design and analysis choices at such levels of randomization is the goal of our study.

This is not entirely original ground. There exists an extensive literature on the analysis of cluster-randomized field experiments and computation of cluster-robust standard errors. Abadie et al. (2017) present clustering as essentially a design problem. They show that clustering is justified when either a subset of clusters are sampled randomly from a broader population of clusters or if assignment is correlated within clusters. We focus on the latter case. In such instances Athey and Imbens (2016) recommend inference based on the cluster as the unit of analysis, arguing that, in practice, including unit-level characteristics generally improves precision by a relatively modest amount compared to including cluster-level averages as covariates in a cluster-level analysis. If the unit-level population treatment effect is the main target of interest, they suggest cluster-level inference to supplement unit-level estimation, arguing that in cases with very different cluster sizes, unit-level estimation may be noisier.

Neither of these papers, however, explicitly considers panel data settings. A seminal paper on cluster-randomized treatments in panel data is Bertrand et al. (2003), which shows through simulation that not clustering standard errors in panel data can lead to systematic overrejection of significance tests in the presence of serial correlation. While Bertrand et al. (2003) demonstrate the importance of clustering standard errors or using randomization-based inference, they do not offer recommendations on unit-level versus cluster-level analysis of treatment effects. In big data settings, the number of units observed may greatly exceed the number of clusters, so such issues merit consideration.

Brewer et al. (2017) argue, in response to Bertrand et al. (2003), that straightforward methods can be used to yield a correctly sized test in diff-in-diff settings. They propose using cluster-randomized standard errors (CRSE) and, instead of using the standard normal approximation to compute p-values, using a t-distribution with degrees of freedom equal to the number of clusters minus one. Through simulation they show that while these methods can give tests that are correctly sized they lead to underpowered estimates. They recommend using the bias-corrected FGLS method derived in Hansen (2007) or FGLS with CRSE to improve power.

Yet, both Brewer et al. (2017) and Hansen (2007) focus on achieving higher power by correcting for serial correlation in the errors, an issue that is typically less important in panels with few time periods. In Monte Carlo simulations, Hansen (2007) shows that serial correlation does not pose a serious problem to inference when there are six time periods in the data. In randomized control trial settings, where measuring outcomes repeatedly might be infeasible or too expensive, having many time periods may not be possible. In such cases, there may be more effective ways to improve the efficiency of estimators than focusing on modeling serial correlation.

In this study, we combine the panel data literature on cluster-robust standard errors (Moulton (1990), Liang and Zeger (1986), Cameron and Miller (2015)) with the finite population causal standard errors literature (Abadie et al. (2017), Abadie, Athey, Imbens, and Wooldridge (2014), Athey and Imbens (2016)), focusing on cluster-randomized experimental settings with short panels. Our main contribution is to compare different methods of estimation—most notably unit-level versus cluster-level estimation of treatment effects—in terms of whether they provide correctly sized tests of the treatment effect.

We conduct extensive simulations on both synthetic data constructed under increasingly complex data generating processes and on novel experimental data described below. Among the factors we consider are the number of clusters, the number of units, the variation in cluster size, the level of intracluster correlation, and the number of time periods.

One key insight is that unit-level estimation with unit fixed effects and cluster-level estimates weighted by the number of observations per cluster can give higher power than unweighted cluster-level estimates in most settings we consider. Importantly, this result holds without sacrificing proper coverage of the treatment effect. This is especially true when there is treatment effect heterogeneity at the cluster level that is correlated with cluster size. In short panels, these simple changes can lead to much larger improvements in precision than properly modeling serial correlation.

Another key insight is that from a design perspective, we show that when there are heterogeneous treatment effects at the cluster level, there are important cases when the analyst should include more treated clusters than control clusters. The intuition is that the added treated sample can be valuable for averaging over the heterogeneity and recovering the population mean effect.

We also evaluate the robustness of computing stratified standard errors, as suggested by Abadie et al. (2014). We find that stratifying standard errors based on observable characteristics that are correlated with the treatment effect can lead to substantial improvements in power, but when the data-generating process is complex—e.g. with differently sized clusters and complex spatial correlation—it may lead to underestimates of the true standard error.

We put our insights into action by evaluating the effect of introducing tipping on Uber's marketplace. In July of 2017 we created a unique opportunity to explore this class of issues when we helped Uber launch optional in-app tipping. We designed a field experiment across 209 operational cities in the United States and Canada where we placed half of the cities in a treatment group, in which tipping started on July 6, 2017, and half of the cities in a control group, where tipping started 10 days later. We use this exogenous variation to evaluate market level changes in earnings, labor supply, demand, and other outcomes of interest (in a companion study, Chandar et al. (2019), we explore individual behaviors). Over 800,000 drivers were included in the experiment. Units were cluster-randomized because of concerns about interactions between drivers within a city and because of business concerns about randomly excluding access across drivers within a city to a popular new feature.

While the main results are not statistically significant at conventional levels, we find improvements in precision from adopting the methods described. Point estimates are imprecise but suggest a short-run increase in labor supply, decrease in demand, fall in utilization, and fall in hourly earnings for drivers. We view our methodological work as the main contribution of our study, but exploring the tipping data at the market level does provide new insights into the tipping literature, which to date has not experimentally examined the questions that we explore.

The remainder of our paper proceeds as follows. Section 2 describes the simulation setting. Section 3 shows simulation results under different data generating processes increasing in complexity to mimic the Uber data. Section 4 considers simplified simulation settings that seek to isolate the effect of various potential features of the data. Section 5 presents results and placebo test simulations in the real data for labor supply. Section 6 summarizes empirical results for the full set of market outcomes. Section 7 concludes.

2 Simulation Set Up

We begin by describing the simulation framework used in the remainder of the paper to compare different methods of estimation. The motivating example for our simulations is the Uber tipping launch. In the Uber tipping case, which we expand on much more patiently in Section 5, the entire population of cities is observed. The uncertainty in the Uber context is not from sampling from the population, but instead from the treatment assignment for each city, which is randomized. In the following simulations we model a scenario where the entire population is observed using the potential outcomes framework.

Let N be the number of individuals in the population. These individuals are split between a set of cities C, with the city that individual i belongs to denoted c_i. Each city c has N_c individuals. Treatment assignment occurs at the city level, so we can separate C into $C = \{C_{\text{treated}}, C_{\text{control}}\}$, where C_{treated} is the set of treated cities and C_{control} is the set of control cities. Let $w_{i,t}$ be the treatment status for individual i in time t, where t goes from 1 to T. In the Uber example, $w_{i,t}$ is whether an individual is in a city with tipping at time t. Since the experiment is cluster-randomized,

$$w_{i,t} = \begin{cases} 1 \text{ if } c_i \in C_{\text{treated}} \text{ and } t \geq T_{\text{treated}} \\ 0 \text{ otherwise} \end{cases}$$

where T_{treated} is the time of the treatment event. We assume that the time of the treatment event does not vary by treated city.

Let $Y_{i,t}$, the outcome for individual i in time t, be

$$Y_{i,t} = \begin{cases} Y_{i,t}(1) \text{ if } w_{i,t} = 1 \\ Y_{i,t}(0) \text{ otherwise} \end{cases}$$

$Y_{i,t}(1)$ is the outcome for individual i in time t if they are treated, and $Y_{i,t}(0)$ is the outcome if they are in the control group. While in the pre-treatment period we observe $Y_{i,t}(0)$ for all i, when $t \geq T_{\text{treated}}$, we observe only $Y_{i,t}(0)$ for some individuals and only $Y_{i,t}(1)$ for others, depending on the treatment status for their city. If we take the population of interest to be active drivers in cities across the United States and Canada, then our application is one with no sampling uncertainty of individuals (since we observe the entire population). However, there is assignment uncertainty since in the post-period we only observe one of their potential outcomes.

We run Monte Carlo simulations to evaluate different estimation strategies under this framework. We first sample $Y_{i,t}(0)$ for all i, t and $Y_{i,t}(1)$ for $t \geq T_{\text{treated}}$ from some chosen data generating process (DGP). These vectors of outcomes constitute the fixed population. Then, for each run of the simulation we randomize treatment assignment to the clusters and estimate the treatment effect using each estimation strategy. By computing the estimated treatment effect many times, we can estimate separately for each estimation strategy the distribution of the treatment effect and standard error over the randomizations of treatment status. Over these randomization distributions, we estimate coverage of the treatment effect (the proportion of times the true population treatment effect falls within the 95% confidence interval on a run) and the rejection frequency (the proportion of times the null hypothesis of no effect is rejected when the alternative is true) for each estimation method. Note that the source of variation for each run of the simulation comes only from the randomization of treatment status; there is no variation in the potential outcomes of the underlying population across runs of the simulation.

The treatment effect in the population is

$$\tau_p = \frac{\sum_{i=1}^{N} \sum_{t=T_{treated}}^{T} [Y_{i,t}(1) - Y_{i,t}(0)]}{N(T - T_{treated} + 1)},$$

that is, the average treatment effect across individuals in the post-treatment period. Note that the treatment effect depends on the population, which is drawn from some data generating process. While in Section 3 we draw only one fixed population for each set of results, in Section 4 we present error bars that show how coverage and rejection frequencies vary depending on the population that is drawn.

Across the simulations we vary several dimensions: the number of time periods, whether there is cluster heterogeneity in outcomes and treatment effects, whether there are individual-level intercepts, and the complexity of the spatial correlation within clusters, in addition to other features of the data-generating process. We also consider cases where the treatment effect is correlated with observable characteristics. Exact parameter values for the data-generating process for each population are presented in Appendix A.

3 Simulation Results

Using the above potential outcomes framework we run Monte Carlo simulations to examine the performance of various models for identifying treatment effects in settings with cluster randomization. In the results that follow we iteratively add complexity to the data generating process, starting with a simple one-period model. In all of the examples in this section there are 20,000 individuals across 212 clusters.

3.1 One-Period Model, Homogeneous Effect

We first consider one-period models with a homogeneous treatment effect. Since there is only one period we drop the time subscript t. Treatment status for individual i is

$$w_i = \begin{cases} 1 \text{ if } c_i \in C_{\text{treated}} \\ 0 \text{ otherwise} \end{cases}$$

The data generating process is

$$[DGP] : Y_i = \tau w_i + \epsilon_i$$

τ is the average treatment effect, and $\epsilon_i \sim N(0, \sigma^2)$ is an error term. We first consider a case in which every city has the same number of observations. In these simulations there is no heterogeneity in the treatment effect across cities. We relax this assumption in subsequent models. Exact parameter values are in Table 27 in the appendix.

We estimate the regression model

$$Y_i = \tau w_i + \epsilon_i$$

across individuals. In this case, standard errors give proper coverage under the null hypothesis that $\tau = 0$, regardless of whether the regression is clustered at the city level using Liang-Zeger (LZ) clustered standard errors (Liang and Zeger, 1986). Table 1 shows the results of a Monte Carlo simulation under this data generating process when each city is the same size. Specifically, for each city c the number of individuals is set to $\frac{N}{|C|}$, rounded to the nearest integer.

In order to compare the individual-level regressions to ones collapsed to the city-level, we also consider regressions of the form:

$$\bar{Y}_c = \tau w_c + \bar{\epsilon}_c$$

with

$$\bar{Y}_c = \frac{1}{N_c} \sum_{i:c_i=c} Y_i, \qquad \bar{\epsilon}_c = \frac{1}{N_c} \sum_{i:c_i=c} \epsilon_i$$

and w_c the treatment status for cluster c.

Row 3 in Table 1 show the results from the city-level regression. In this case there is little difference between the various methods on efficiency of the estimator. Results are similar under the alternative hypothesis, as shown in Table 2.

3.2 One-Period Model: Homogeneous Treatment Effect With Heterogeneous Cluster Sizes

When clusters vary in size, it becomes necessary to weight the cluster-level regression by the number of observations in each cluster to maintain the same power as in the individual-level regression. For Tables 3 and 4 we vary the size of the clusters, with sizes roughly calibrated to the sizes of cities in the Uber field experiment. Let N_c^* be the number of observations in city c in the real data. In the simulated data, we set $N_c = \text{round}(N \times \frac{N_c^*}{\sum_{c'} N_{c'}^*})$, where we set N to 20,000.

Under both the null and alternative hypotheses, the unweighted city-level regression has larger standard errors. This is because errors are heteroskedastic at the city level since they are averages across individuals. Running weighted least squares with weights equal to cluster sizes gives increased efficiency and higher rejection rates when the alternative is true.

3.3 One-Period Model: Cluster Intercepts

We next consider a one-period model with cluster-specific intercepts. The data-generating process is

$$[DGP] : Y_i = \tau w_i + \alpha_{c_i} + \epsilon_i$$

where $\alpha_{c_i} \sim N(0, \sigma_c)$ is a cluster-specific intercept. We again estimate the regression models

$$Y_i = \tau w_i + \epsilon_i$$

and

$$\bar{Y}_c = \tau w_c + \bar{\epsilon}_c$$

Tables 5 and 6 show results from simulation when clusters are equal sizes. Individual-level regressions that are not clustered underestimate the standard error. This is consistent with the findings of Abadie et al. (2017), who show that if clusters vary in mean outcomes, assignment is clustered, and fixed effects are not included in estimation then standard errors should be clustered. The clustered individual-level model and the city-level model give comparably small standard errors.

When clusters vary in size, as in Tables 7 and 8, the unweighted city-level model has higher variance in the estimated coefficient than the clustered individual-level model. Using Eicker-Huber-White (White, 2014) standard errors in the city-level model does not have a considerable effect on the estimated standard errors. The city-level model weighted by size reduces the variance of the estimator to be comparable to the individual-level models, but the standard error is too small, leading to poor coverage of the true treatment effect. Weighting the city-level model by the number of observations and also using EHW standard errors leads to similar performance as the individual-level model with clustering. Weighted least squares is insufficient unless also using robust standard errors because weighting by the size of the cluster ignores the random cluster-specific intercept, which contributes to the variance in the outcome. In settings where clusters vary in size, EHW estimation can lead to conservative standard errors and WLS can lead to standard errors that are too small. Combining the two leads to tests that have better size.

Notwithstanding, the unweighted city-level estimator maintains better coverage of the treatment effect than the other models. This result echoes the recommendations of Athey and Imbens (2016) for a simple one-period model with differing cluster sizes. In subsequent results, we show that the individual-level estimator and the weighted city-level estimator may perform better when having multiple time periods because it allows us to include cluster-fixed effects in the models.

3.4 Two-Period Model: City and Time Fixed Effects

In the Uber tipping context, our goal is to estimate the treatment effect across individuals in a multi-period setting where cities only become treated after the launch of the product. In order to simplify the model while still keeping it informative, we consider a two period model.

Let $Y_{i,t}$ be the outcome for individual i in time period t. Treatment status for individual i at time t is now

$$w_{i,t} = \begin{cases} 1 \text{ if } c_i \in C_t \text{ and } t = 2 \\ 0 \text{ otherwise} \end{cases}$$

Individuals are only treated if they are in a treated city and $t = 2$.

The first DGP we consider is

$$[DGP] : Y_{i,t} = \tau_i w_{i,t} + \alpha_{c_i} + \beta_t + \epsilon_{i,t}$$

where $\tau_i \sim N(\tau, \sigma^2_\tau)$ is an individual-specific treatment effect, $\beta_t \sim N(0, \sigma^2_t)$ is a time effect, and $\epsilon_{i,t} \sim N(0, \sigma^2_\epsilon)$ is an independent and identically distributed error term. Cluster sizes are calibrated to the Uber data. The treatment effect now varies by individual under the alternative hypothesis, but under the null $\tau_i = 0$ for all i.

We consider individual-level regressions of the form

$$Y_{i,t} = \tau w_{i,t} + \alpha_{c_i} + \beta_t + \epsilon_{i,t}$$

and city-level regressions of the form

$$\bar{Y}_{c,t} = \tau w_{c,t} + \alpha_c + \beta_t + \bar{\epsilon}_{c,t}$$

Because we have repeated observations for each cluster over time we can now use fixed effects estimation to account for cluster intercepts. Results are in Tables 9 and 10.

Clustering the standard errors without also including cluster fixed effects leads to underestimates of the standard errors for the individual-level models and the weighted city models. Clustering alone is insufficient to account for the within-cluster correlation between individuals when there are only around 200 clusters. For these models, standard errors are both smaller and more properly sized when fixed effects are included. This shows that including cluster fixed effects can improve both the coverage and power of the treatment effect estimator even when using Liang-Zeger clustered standard errors. We do not see the same result for the unweighted model. While for the individual-level models (and similarly for the weighted city-level models) LZ clustered standard errors need to account for both correlation across individuals and across time, for the unweighted city-level models clustering only needs to correct for correlation across the two time periods. For this simple two-period model, clustering the standard errors is adequate for accounting for such correlation. The fixed effects drop too many degrees of freedom, leading to higher variance in the estimated treatment effect.

The individual-level model and the weighted city-level model with cluster effects have much smaller variance in the estimated treatment effect than the unweighted city-level model, consistent with what we found in the one-period case. However, in contrast to our earlier analysis, the clustered weighted city-level model now also gives proper coverage of the treatment effect after we include cluster effects, meaning it gives higher power than the unweighted model without sacrificing the size of the test.

3.5 Two-Period Model: Individual Fixed Effects and Heterogeneous Cluster Effects

Next we add individual effects and heterogeneous treatment effects at the cluster level. The DGP is

$$[DGP] : Y_{i,t} = \tau_i w_{i,t} + \alpha_{c_i} + \beta_t + h_{c_i} w_{i,t} + \gamma_i + \epsilon_{i,t}$$

where $h_{c_i} \sim \text{Normal}(0, \sigma_h)$ is cluster-level heterogeneity in the treatment effect and $\gamma_i \sim \text{Normal}(0, \sigma_\gamma)$ is an individual-specific intercept term. Under the null we set $h_{c_i} = 0$ for all i. We again consider individual-level regressions of the form

$$Y_{i,t} = \tau w_{i,t} + \alpha_{c_i} + \beta_t + \epsilon_{i,t}$$

and city-level regressions of the form

$$\bar{Y}_{c,t} = \tau w_{c,t} + \alpha_c + \beta_t + \bar{\epsilon}_{c,t}$$

The additional terms in the model increase the variance of the estimated treatment effect. In order to display reasonably large rejection frequencies under the alternative hypothesis, we increase τ to 0.1 from 0.035. Results are in Tables 11 and 12. Under the null for this model, in the individual-level model clustering by city does not improve estimation when city fixed effects are included. Under the alternative hypothesis, when there are heterogeneous treatment effects at the cluster level clustering becomes necessary. The mean standard error for the individual-level model without clustering in Table 12 underestimates the standard deviation of the estimated treatment effect.

Including individual fixed effects instead of city fixed effects leads to a substantial increase in the mean standard error because of reduced degrees of freedom, making the test too conservative. Of all the models that give properly sized standard errors, the clustered individual-level model with city fixed effects has the highest rejection frequency.

3.6 Two-Period Model: City and Time Fixed Effects With Complex Correlation

We now consider situations where the errors are additionally correlated within cities for each time period. Instead of an identity matrix, the covariance matrix of $\epsilon_{i,t}$ is now block-diagonal, with each block corresponding to a city in a given time period. The covariance matrix for each city at time t, $\Sigma_{c,t}$, has the following form:

$$\begin{bmatrix} \sigma^2 & \sigma^c_{12} & \dots & \sigma^c_{1N_c} \\ \sigma^c_{21} & \sigma^2 & \dots & \sigma^c_{2N_c} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \sigma^c_{N_c1} & \sigma^c_{N_c2} & \dots & \sigma^2 \end{bmatrix}$$

For simplicity we again assume that the diagonal elements σ^2 of the error covariance matrix are constant.

The goal is to model a complex, a priori unknown, covariance structure in the errors. For example, in the Uber example it might be that drivers in a city who work in the same areas during the same times of the week have correlated outcomes. Simply including city fixed effects would not correctly model this aspect in the data. To simulate this setting we need a dense covariance matrix with the correlation between drivers increasing with their (geographic or temporal) proximity. We draw the vector of errors for each city in each time period, $\vec{\epsilon}_{c,t}$, from a Markov chain to generate this sort of covariance structure. The sampling algorithm, shown in Algorithm 1, resembles an AR(1) process.

Algorithm 1: Sampling algorithm to generate a dense covariance matrix for each cluster.

```
Initialize $u^{(1)}$;
for $d = 1, ..., D$ do
    - Sample $u^{(d+1)} \sim f(u^{(d+1)}|u^{(d)})$;
end
Collapse $u^{(1)}, ..., u^{(D)}$ to a vector $U$;
Set $\vec{\epsilon}_{c,t}$ to the last $N_c$ elements of $U$
```

In Algorithm 1, $u^{(d)}$ and $u^{(d+1)}$ may represent errors for drivers who work in very similar areas. We assume the joint distribution of $u^{(d)}$ and $u^{(d+1)}$, $f(u^{(d)}, u^{(d+1)})$, is bivariate Normal with diagonal elements equal to σ^2 and off diagonal elements equal to ρ. We set $\vec{\epsilon}_{c,t}$ to only the last N_c elements of the vector U, dropping the first few draws from the chain, so that the variance stabilizes to a constant. As ρ approaches 1 all of the error terms within a city for a given time period approach equivalence, and if $\rho = 0$ then all the error terms will be independent.

To summarize, the DGP is

$$\begin{aligned} [DGP] : Y_{i,t} = \tau_i w_{i,t} + & \alpha_{c_i} + \beta_t + h_{c_i} w_{i,t} + \gamma_i + \epsilon_{i,t} \\ \epsilon_{i,t} \sim & N(0, \Sigma) \text{ for block-diagonal } \Sigma \end{aligned} \tag{1}$$

The estimated models are the same as in our earlier analysis. Results in simulation when ρ is set to 0.5 are in Tables 13 and 14. Under the null it becomes necessary to cluster the standard errors when there is unmodeled correlation in the errors. The individual-level model with individual fixed effects again gives a conservative test relative to the model with city fixed effects.

The standard errors for the weighted city-level model are too small in this case unless clustered by city. Again, the unweighted city-level model has larger standard errors than the other models.

3.7 Heterogeneity from Observables

To model situations where the treatment effect varies with observable driver characteristics, we consider cases with an interaction between observable characteristics z_i and the treatment indicator:

$$\begin{aligned}[DGP]: Y_{i,t} =& \tau_i w_{i,t} + c_i + t + h_c w_{i,t} + \tau_z w_{i,t} \times z_i + \alpha_i + \epsilon_{i,t} \\ &\epsilon_{i,t} \sim N(0, \Sigma) \text{ for block-diagonal } \Sigma\end{aligned}$$

This format is similar to the one used in Abadie et al. (2014). They model the potential outcomes as

$$Y_i(x) = Y_i(0) + (\xi_0 + \xi_1 Z_i + \xi_2 \eta_i) \times x,$$

where ξ_1 and ξ_2 are varied across simulations and ξ_0 is set to 0. $\eta_i \in \{-1, 1\}$ is an unobserved source of heterogeneity in the treatment effect, $x \in \{-1, 1\}$ is the treatment status, and $Z \in \{-1, 1\}$ is a binary attribute of the individual. When a non-negligible proportion of observations in the full population are observed and $\xi_2 \neq 0$, they show standard errors stratified by Z are less conservative and closer to the true variance of θ than EHW standard errors are when estimating the model

$$Y_i = \beta_0 + \beta_1 Z_i + \theta X_i + \epsilon_i$$

The potential outcomes for the DGP in our setting can be expressed as

$$Y_{i,t}(1) = Y_{i,t}(0) + (\tau + \eta_i + h_{c_i} + \tau_z z_i) \times w_{i,t}$$

where $\eta_i \equiv \tau_i - \tau$ is a source of unobserved heterogeneity at the individual level and h_c is a source of unobserved heterogeneity at the cluster level. z_i is the source of heterogeneity in the treatment effect determined by observables.

In simulation, we set the observable attribute z_i to the quintile of the cluster size, subtracting 3 so that z is centered around 0 across clusters. Ties are broken arbitrarily so that an equal number of clusters fall in each quintile, even if clusters are identically sized. Since in our simulations z_i is the cluster size quintile there is no intracluster variation in z. Collapsing to city-level means that the potential outcomes are

$$\begin{aligned}\bar{Y}_{c,t}(1) =&\bar{Y}_{c,t}(0) + (\tau + \bar{\eta}_c + h_c + \tau_z z_c) \times w_{c,t} \\ \approx \bar{Y}_{c,t}(1) = \bar{Y}_{c,t}(0) + (\tau + h_c + \tau_z z_c) \times w_{c,t}\end{aligned}$$

which closely resemble the outcomes in Abadie et al. (2014). The models we fit are again

$$Y_{i,t} = \tau w_{i,t} + \alpha_{c_i} + \beta_t + \epsilon_{i,t}$$

for individuals and

$$\bar{Y}_{c,t} = \tau w_{c,t} + \alpha_c + \beta_t + \bar{\epsilon}_{c,t}$$

when collapsing to city means.

Simulation results are summarized in Table 15 for a case where there is heterogeneity in the treatment effect and clusters are equal sizes. Stratified standard errors are lower on average than unstratified standard errors across all of the models. The individual-level model with city fixed effects has the smallest reduction in standard errors. Table 16 shows results are similar when clusters vary in size. Because under the modeling process large clusters have a larger treatment effect, the mean treatment effect across individuals is greater in the case with unequal clusters.

The city-level models that are not weighted are biased estimates of the individual-level mean treatment effect when clusters are different sizes. Since they weight each cluster equally, they overweight small clusters, which have small treatment effects, and underweight large clusters, which have large treatment effects. The unweighted model estimates a different quantity, treating each city as the unit of interest instead of each individual. The individual-level models and the weighted city-level models give unbiased estimates of the individual-level treatment effect and have comparable standard errors.

To isolate the effect of the heterogeneity from observables, we plot the ratio between the stratified and unstratified standard errors as we vary τ_z. Figure 1 shows the pattern when clusters are of equal sizes. All estimates are clustered at the city level. The two plots in the top panel show that both the stratified and unstratified models become too conservative as the amount of heterogeneity increases, but the plot in the bottom-left panel shows that the stratified model does increasingly better than the unstratified model. Results are similar in Figure 2, where cluster sizes are roughly proportional to Uber cities.

4 Isolating Effects

What causes differences in standard errors across the different regression specifications? In the following simulation results we show the effect of varying the number of clusters, the number of observations per cluster, the amount of dispersion in cluster size, the level of intracluster correlation, and the proportions of treatment and control cities. We also consider models with more than two time periods and serial correlation across those time periods.

Based on the results in the previous section and our review of the literature, we only consider cases where it is most evident that clustering is necessary to gain proper coverage. We examine cases with city heterogeneity in the treatment effect and complex, unmodeled, spatial correlation in the errors within a cluster. For the following figures, the DGP is

$$[DGP] : Y_{i,t} = \tau_i w_{i,t} + \alpha_{c_i} + \beta_t + h_{c_i} w_{i,t} + \gamma_i + \epsilon_{i,t}$$
$$\epsilon_{i,t} \sim N(0, \Sigma) \text{ for block-diagonal } \Sigma$$

The models labeled "Cit. Clustering" are collapsed to the city level and clustered by city. They include time and city fixed effects. The models labeled "Cit. Clustering Weighted" are similar, but weighted by the number of individuals in each cluster. Models marked "Ind. Clustering" are at the individual level, include city and time fixed effects and are clustered by city. The models labeled "Ind. Ind FE City Clus." are at the individual level with individual and time fixed effects, and are clustered by city. These four were chosen because they were the most robust specifications from the previous simulations.

We report coverage of the treatment effect and the probability of rejecting the null hypothesis in the following figures. Coverage and rejection rate results in the trend lines are the average over estimates from 25 different populations drawn from the DGP. For each population, coverage and rejection rate are computed over 200 different re-randomizations of treatment status. We also show error bars that span the minimum to maximum coverage and rejection rate estimates across the 25 different populations to show how much the results vary depending on the population.

4.1 Varying the Number of Observations and Clusters

We start with cases where all clusters are the same size and there are two time periods. First, we vary the number of observations, holding the number of clusters fixed at 200. The number of observations per cluster varies, but in each specification of the DGP the clusters are equally sized. Results are shown in Figure 3. The clustered city-level models and the individual-level model with individual fixed effects are conservative. All have coverage above 95% for each number of observations. The individual-level model with city fixed effects has higher power and has coverage roughly centered around 95%.

In Figure 4 we see that results are similar when increasing the number of clusters, holding fixed the number of total observations at 20,000. When there are few clusters, the individual-level model with city fixed effects has coverage of the treatment effect below 95%. It does not explicitly model idiosyncrasies across individuals, relying on cluster robust standard errors to account for this source of variation in the data-generating process. Without sufficiently many clusters, clustered standard errors may not suffice. All of the other models maintain coverage above 95%. Power increases with the number of clusters across the specifications.

4.2 Varying the Amount of Cluster Dispersion

Next we vary the amount of cluster dispersion. Let $\bar{N}_c$ be the mean cluster size when there are 20,000 observations and 212 clusters, or each cluster's size if observations were equally spread across clusters. With N_c^* set to the number of observations in city c in the real data and N set to 20,000, again let $N_c = \text{round}(N \times \frac{N_c^*}{\sum\limits_{c'} N_{c'}^*})$.

We set the number of observations in cluster c to

$$N_{c,\delta} = \text{ceiling}((1-\delta)\bar{N}_c + \delta N_c), \tag{2}$$

varying δ from 0 to 1. When $\delta = 0$ the clusters are all the same size, while when $\delta = 1$ the clusters match the calibrated Uber data case. All specifications except the individual-level model with city fixed effects maintain proper coverage of the treatment effect as the variation in cluster size increases. The individual-level model with individual fixed effects and the clustered and weighted city-level model have higher average rejection rates than the unweighted city-level model when the variation in cluster sizes is large. This suggests that information about the number of individuals in each cluster can improve precision.

4.3 Varying the Amount of Spatial Correlation

In Figures 6 and 7 we vary the amount of spatial correlation in the errors. In Figure 6 clusters are the same size. The amount of spatial correlation does not seem to have a meaningful effect on the city-level models or the individual-level model with individual fixed effects when clusters are equal. The individual-level model with city fixed effects has reduced coverage when the spatial correlation is very high, even when clusters are equal sizes. Power declines as the correlation increases for all specifications. When clusters are different sizes, all models except the unweighted city-level model witness a small decline in coverage when the correlation parameter is close to 1. The unweighted city-level model is dominated in rejection rates by the other models, though rejection rates converge across the specifications to a low number when the intra-cluster dependence parameter is close to 1. However, we also find that the outcome for the unweighted city-level model is less dependent on the the population that is drawn. In Figure 7 the error bars show that the minimum coverage under the null across the 25 drawn populations is close to 80% for the weighted city-level model and the individual-level model with individual fixed effects. It is close to 70% for the individual-level model with city effects. In contrast, for the unweighted city-level model coverage is 98% or higher across every population that is drawn. In extreme cases, the unweighted city-level model may be more robust to the characteristics of the population.

4.4 Varying the Number of Treated Clusters

In the previous results there were an equal number of treated and control clusters. In Figure 8 we vary the number of treated clusters out of 212 clusters in total, with cluster sizes calibrated to the Uber data. All specifications aside from the unweighted city-level model have an inverted U-shape for coverage and power. This means that when there are too few or too many treated clusters they underestimate the standard error.

Under the null, the inverted U-shape for coverage is symmetric, meaning having too many or too few treated clusters is equally harmful. In contrast, under the alternative hypothesis coverage and rejection rates are highest when there are slightly more treated clusters than control clusters. This is likely because when there are heterogeneous treatment effects at the cluster level, additional treated clusters can be valuable for averaging over them and recovering the mean in the population. For this reason, it may be preferable in certain cases to have more treated clusters than control clusters when designing an experiment if the researcher expects a lot of cluster-level treatment effect heterogeneity.

4.5 Varying the Number of Periods

In Figures 9 and 10 we report results when increasing the number of periods before and after the treatment event, again with an equal number of treated and control clusters and cluster sizes calibrated to the Uber data. We additionally show results when using the bias-corrected FGLS estimator for an AR(1) process in Hansen (2007), though we expect it to give little improvement in rejection rates in this case since there is no serial correlation in the response beyond individual- and city-level fixed effects. In Figure 9 the treatment event occurs in the final period, so in all preceding time periods all of the clusters are untreated. In Figure 10 the treatment event occurs in the second time period.

Rejection rates are increasing across all specifications as the number of periods increases, though the improvement levels off after there are more than five periods. This means that the returns to repeated observations in terms of statistical precision flatten out quickly in cluster-randomized experiments. As the number of time periods increases all of the specifications show a reduction in coverage of the treatment effect. We still find substantially higher rejection rates for the weighted models, regardless of the number of time periods.

In Figure 11 we add additional serial correlation in the errors at the individual level. We assume this follows an AR(1) process. We draw error terms for each city using Algorithm 1 to generate spatial correlation and update them to allow for serial correlation. Algorithm 2 shows the precise process:

Algorithm 2: Algorithm to create serial correlation in the data generating process. ϕ is a scalar that controls the level of serial correlation.

```
Draw a vector $\tilde{v}^{(1)}$ from Algorithm 1;
for $d = 1, ..., D$ do
    - Sample $v^{(d+1)}$ from Algorithm 1;
    - Set $\tilde{v}^{(d+1)} = \phi\tilde{v}^{(d)} + v^{(d+1)}$;
end
Set $(\vec{\epsilon}_{c,1}, \ldots, \vec{\epsilon}_{c,T})$ to equal the last $T$ vectors $(\tilde{v}^{(D-T+1)} \ldots \tilde{v}^{(D)})$
```

We discard the first four time periods of the above process so that the variance of the error terms converge to be roughly constant over time. We vary the serial correlation parameter ϕ and examine its effect on coverage and rejection rates, only considering cases with 4 additional periods before treatment takes place.

The serial correlation parameter seems to have little effect on the estimates, at least for the moderate levels plotted in Figure 11. Coverage appears unchanged, while rejection rates decay slightly. When city or individual effects are included such correlation does not seem to have much effect on estimates. Even in this case, the FGLS estimators have little or no improvement in power. This is consistent with Monte Carlo evidence in Tables 3 and 4 of Hansen (2007), which shows that for short panels potential gains in power from using the FGLS estimator can be quite small. A much larger increase in power comes from using weighted least squares or individual-level estimation.

5 Simulations in Labor Supply Data

We now evaluate the aforementioned approaches on novel experimental data from Uber. We begin by focusing on changes in labor supply resulting from the launch of in-app tipping. We start with a description of the data and results for a balance test in the period before the product launch. We then provide results for the effect of tipping on labor supply along with findings from placebo test simulations similar to ones in Bertrand et al. (2003). We focus on labor supply as an exemplar case to show the complications in the data and the impact this has on estimation of the treatment effect.

5.1 The Data

We worked with Uber to launch optional in-app tipping on its platform in June, 2017. We launched tipping in three waves. On June 20, 2017, tipping launched in Seattle, Houston, and Minneapolis for testing. We refer to this as the alpha launch. On July 6, 2017, tipping launched in half of the remaining operational cities in the U.S. and Canada, with the other half serving as a control group to evaluate changes in market aggregates of interest: earnings, labor supply, demand, and other market outcomes. We refer to this as the beta launch.

Finally, on July 17, tipping launched across the rest of the US and Canada. We refer to this as the full launch. For the period between July 6 and July 17, we have experimental variation across cities to evaluate the impact of the introduction of tipping. While we would have preferred a longer window, firm level constraints on desiring a fast roll out led to our field experimental design.

Cities with business or technological restrictions were excluded from our field experiment. This includes cities such as New York, which launched tipping on July 6 because of business purposes. The remaining 110 largest cities were randomized to treatment and control according to a matched pairs design. First, cities were grouped by whether they had UberPOOL and UberEATS. Then within these groupings, cities were matched with the city that was most similar across several dimensions, including city size, the current state of driver earnings, and the demographic composition of the city. Similarity was determined by taking the Euclidean distance, weighting each of these components equally.

Within these pairs, cities were randomly assigned to treatment or control. Cities that did not have a pair within their group because the group had an odd number of cities were randomly assigned to treatment or control. After the randomization, the assignment of suburbs of large cities were overridden so that they had the same treatment as their main metro area. For instance, Orange County and Los Angeles have the same treatment status. The remaining cities were simply randomly assigned to treatment and control without constructing groupings or pairs. Data from the experiment has 215 cities in total, 109 in treatment and 106 in control. We drop six highly irregular cities from the analysis for reasons we discuss below, leaving 104 cities in treatment and 105 in control.

We consider the four weeks preceding the beta launch and the week after the beta launch to determine whether tipping had an effect on drivers' labor supply choices. Drivers who worked on the platform on any day between June 8 and July 13 were included in the data and were matched to the city in which they worked the most in the 5-week period. For each driver, we aggregated total minutes worked for each week, so the observations in the data are at the driver-week level. The number of minutes worked was coded as 0 for weeks the driver did not log onto the Uber app.

The data present several challenges for analysis. First, the dispersion in city size is large in the real data. Figure 5 suggests that when cluster sizes are very unequal power declines. A plot of each city's size, ordered from smallest to largest is given in Figure 12. The distribution is quite skewed, as more than half of drivers work in the nine largest cities: Los Angeles, Miami, Chicago, San Francisco, Washington D.C., Boston, Philadelphia, Dallas, and Tampa.

The distribution of outcomes in the real data also does not resemble the normally distributed outcomes from the simulations. A driver's labor supply in a given week can take on any integer value greater than or equal to 0 (up to 10,800 minutes in the week in principle at the time of the experiment). 36% of the labor supply observations are 0 because drivers on Uber often take weeks off. This means that the real data is semi-continuous at the individual level. There is also considerable heterogeneity across cities, as seen in Figure 13. Conditional on working, the distribution of minutes worked for drivers is fat-tailed and varies by city. Densities for the number of minutes worked conditional on working are shown for various cities in Figure 14.

While the data is irregular at the driver level, it is somewhat better behaved at the city level. The density of city-level minutes worked per week, shown in Figure 15, has a heavy right tail, but is closer to Normal than the individual-level distribution. Figure 16 shows the mean number of minutes worked in each week for a sample of cities in the data set. Cities differ in week-to-week variation in labor supply, but generally changes are not substantial. Two exceptions are Corpus Christi and Anchorage, shown in Figure 17. Both cities were randomized to the treatment group. Uber resumed operations in Corpus Christi on June 29th when new state regulations were passed after a long hiatus. Similarly, Uber began operating again in Anchorage on June 15 following the passage of municipal legislation. We removed these two cities and four other cities that had weeks in which no drivers worked: Laredo, TX; Gallup, NM; Adirondack, NY; and Eagle Pass, TX. Removing these cities left 105 control cities and 104 treated cities.

One final complication in the data is that there may be correlated outcomes for drivers across cities that are near each other. For instance, there are drivers who work in San Francisco but commute from Sacramento. This case is particularly problematic because San Francisco is in the treatment group and Sacramento is in the control group, meaning there may be contamination between the treatment arms.

5.2 Balance Between Treatment and Control

Prior to the tipping launch, drivers in treated cities tended to work more than drivers in control cities. Drivers in control cities worked about 653 minutes per week on average and drivers in treated cities worked about 683 minutes per week[1]. There are also more than 90,000 fewer treated individuals than control individuals (362,057 in treatment, 452,525 in control). A time series plot of minutes worked by day for the treated and control groups is in Figure 18. A t-test at the individual-level clustered by city fails to reject a test for equality of means across groups at the driver level because the variance in outcomes across cities is so high. Results at the city level are similar. Estimates are in Table 17.

The differences between the treated and control groups are driven by the variation in city sizes and heterogeneity in mean labor supply across cities. If instead of a t-test we conduct a randomization test on the pre-treatment period means we find similar results. The difference in means between the treated and control individuals is about 29 minutes. Over possible randomizations of the treatment, this is at the 34th percentile of absolute value of difference in means. When collapsing to the city-level, the realized absolute value of difference in means, 12.5 minutes, is at the 43rd percentile. Even though the randomization was valid, the lack of balance at the individual-level in the pre-treatment period means a simple difference in means in the treatment period will be an underpowered test statistic

[1] If the tipping launch has an effect along the extensive margin then this could be caused by drivers leaving or coming onto the platform in the week after the launch, which affects whether they are included in the data set for the pre-treatment period. However, the results presented in this section also hold when only considering drivers who drove at some point in the pre-treatment period, suggesting the effect is not driven solely by this mechanism.

for evaluating the effect of the tipping launch. Table 51 in the appendix shows that a t-test at the city-level has a minimum detectable effect size of more than 10%. Estimating the effect of the experiment with sufficient power requires adopting assumptions about the time trends, such as in the panel data framework used in the preceding simulations.

5.3 Labor Supply Results

We run regressions on the labor supply data using insights from the previous simulations. Results for specifications that include city and time fixed effects are shown in Table 18. Tipping does not have a statistically significant effect in any of the specifications. The point estimates differ between the individual- and unweighted city-level models substantially because of heterogeneous mean outcomes across city sizes. To estimate the counterfactual expected outcome we take the mean outcome in the treated period, subtracting the estimated treatment effect for treated individuals. The unweighted city-level models estimate roughly a 1.83% change relative to the counterfactual outcome, while the individual-level and weighted city-level models estimate a 0.39% change. The standard error is higher in all specifications when clustering is done at the city level. We also see that the standard error is lower in the clustered individual-level model than in the clustered city-level models.

To check whether coverage is correct under each specification, we re-randomize treatment across cities to simulate placebo experiments and examine coverage under the null for the randomization distribution. While the results should not be interpreted as coverage of the true treatment effect of launching tipping, especially if the treatment effect is heterogeneous across clusters, they can still give some information about Type I error rates. Results are shown in Table 19 for 250 re-randomizations.

The mean of the estimated coefficient is slightly larger than 0 across the specifications, but assuming normality of the estimates we would fail to reject a test that they are in fact 0 on average. This provides suggestive evidence that bias may not be a major problem in the regression models. Unclustered estimators generally underestimate the true standard deviation of the estimated coefficient, leading to poor coverage of the treatment effect. The individual-level regression with city clustering also underestimates the standard error, leading to less than perfect coverage under the null. The clustered city-level regressions with and without weighting are the only ones for which the mean standard errors exceed the actual standard deviation of the coefficients, suggesting these estimators may be more reliable. Of the two, the weighted model has a lower standard deviation of the estimated coefficients and also lower mean standard errors.

Next we see if more granular fixed effects in the individual-level model assist in estimating the underlying correlation structure in cities and across time. We try individual-level fixed effects and also try to control for where and when drivers typically work. We estimate where drivers usually work by taking the level-4 geohash[2] in which they took the most trips in the period prior to June 8, 2017. There are 1,584 unique geohashes in which drivers took

[2] A geohash is a geocoding system that divides the world into a grid of squares of arbitrary precision. For a more detailed discussion see Cook et al. (2018).

their most trips in the dataset. For context, 153 of them are within Uber's boundaries for Chicago. To estimate when drivers usually work, we take the 6-hour block of the week, offset by 4 hours to more accurately reflect driving patterns, in which they took the most trips prior to June 8, 2017. For example, the first block starts on Sunday at 4:00 am and goes until Sunday at 10:00 am, followed by Sunday at 10:00 am to Sunday at 4:00 pm, etc. To capture both when and where drivers work, we take the geohash × time block in which they took the most trips.

Results are in Table 20. Adding more granular fixed effects seems to have little effect on the point estimate and leads to slightly higher standard errors compared to the model with city fixed effects. Results in placebo test simulations are in Table 21. Including fixed effects for when and where drivers typically work seems to lead to a higher standard deviation in the coefficients compared to using city fixed effects. In contrast, using individual fixed effects instead of city fixed effects results in a lower standard deviation for the estimated treatment effect. The mean standard error is higher than in the case when including city fixed effects instead, but the individual-level model now gives a properly sized test. These findings mimic what we found in our simulations. The clustered city-level models and clustered individual-level model with individual fixed effects seem to be most robust to complex features of the data set.

In Table 22, we collapse to the geohash x city, time x city, and geohash x time x city level rather than the individual or city level to explore if estimation at a level more granular than the city leads to more precise results. All regressions are weighted by the number of drivers and clustered by city. All three specifications have very similar point estimates and lower standard errors than the weighted city-level model clustered by city. Results in simulations over the randomization distribution are given in Table 23. All three models have standard errors that underestimate the standard deviation of the estimated treatment effect.

Finally, we consider estimates where standard errors are stratified. Results from running stratified regressions are in Table 24 for select specifications. The individual-level model with city fixed effects does not see a reduction in standard errors, but the model with individual fixed effects does. The city-level models with and without weighting also see a reduction in standard errors. Results in simulations over the randomization distribution are in Table 25. Over the randomization distribution there is no heterogeneity in the treatment effect explained by observables since the true treatment effect is 0. However, the results indicate that all of the models underestimate the standard error except the unweighted city-level specification. While the unweighted stratified city-level model gives proper standard errors, Table 16 shows that the estimate of the treatment effect may be biased under the alternative hypothesis if the treatment effect varies with city size.

6 Market Response to Tipping

Using the results from previous sections we now present more general findings on the effect of the tipping launch on the Uber marketplace. All analyses use weighted city-level models with city and time fixed effects. Estimates are clustered by city. Alternative specifications are left for the appendix. The metrics we focus on are overall labor supply (discussed in Section 5), rider demand, driver utilization, and hourly earnings.

We measure rider demand using a similar method as was used for labor supply. We include all riders who opened the Uber app and entered a destination at some point in the four weeks prior to the experiment or during the first week of the experiment. Riders are associated with the city in which they had the most unique sessions in the period, where a unique session is defined as a block of time when a rider opens the app, enters a destination, and considers taking a trip of any product type (e.g. UberX or UberPOOL). The number of unique sessions is recorded as 0 in weeks when a rider does not consider taking a trip. We exclude the same cities, most notably New York, as we did for labor supply. There are over 16 million riders in total in the data set. A plot of the number of riders in each city is in Figure 19. The distribution of riders per city is highly skewed.

A driver's utilization is the amount of time "working" divided by the total amount of time spent on the platform. For example, if a driver spends 10 minutes waiting for a dispatch, 10 minutes driving to pick up a passenger, 40 minutes driving that passenger to their destination, and then goes offline, their utilization rate will be 83.33% (50 minutes producing a trip out of 60 minutes online). A complication with analyzing the effect of tipping on utilization is that it can only be measured in weeks in which a driver works. This creates selection in which drivers are observed in each week and makes the panel unbalanced. Even without accounting for selection, changes in utilization are useful for explaining observed changes in hourly earnings.

We analyze three different metrics for hourly earnings: gross earnings before Uber's commission and including earnings from incentives and tips; earnings net of Uber's commission (which does not apply to tips); and gross earnings before the commission but excluding incentives and tips. We refer to these metrics as total hourly earnings, post-commission hourly earnings, and organic hourly earnings, respectively. Post-commission hourly earnings are the driver's revenue from Uber. Organic hourly earnings are what the driver earns exclusively through providing trips, factoring in base fare, time spent on trips, distance traveled, and the surge multiplier.

Weighted city-level estimates are in Table 26. None of the estimates are significant. The point estimates suggest an increase in minutes worked and a decrease in demand, with the combination of these factors leading to a decrease in utilization. This results in a drop in gross and organic earnings. The fall in post-commission earnings is more muted, in part because Uber does not collect commissions on tip earnings. The stability in post-commission earnings seems consistent with the market equilibration found in Hall et al. (2018).

While our estimates are noisy, they provide an advance for the tipping literature. For instance, even though tipping has played an important role for centuries in modern economies and remains a hallmark of service economies, until recently economists have not made deep contributions in the area. Notable recent exceptions include the excellent work of Conlin et al. (2003), Azar (2007), and Lynn (2016). While these authors have made important inroads into the phenomenon of tipping, many open questions remain. For instance, how does tipping influence market outcomes such as demand for the good or service and supply of work hours? All of these issues remain unsettled, but our estimates provide a first indication that the market results are not considerable. More work is necessary.

7 Conclusion

This paper helps to clarify questions surrounding design issues and which methods of estimation give properly sized tests in cluster-randomized settings with panel data. We find that when clusters vary in size, unit-level estimation with unit fixed effects and cluster-level estimates weighted by the number of observations per cluster can give more precise estimates than unweighted cluster-level estimates in short panels. If the interest is in measuring individual-level treatment effects, these methods may also be more robust than unweighted cluster-level estimation to cluster-level treatment heterogeneity. We also verify that stratifying standard errors based on observable characteristics that are correlated with the treatment effect can lead to substantial improvements in precision, but show that when the data-generating process is complex it can lead to underestimates of the true standard error. These findings can be applied to other cluster-randomized experiments with big data, both in the gig economy and in field experiments more generally.

We use the simulation insights to evaluate the effect of introducing tipping on Uber's marketplace. While measured results are imprecise, we find a short-run increase in labor supply, decrease in demand, fall in utilization, and fall in hourly earnings for drivers after the introduction of tipping. In equilibrium, introducing tipping did not have a dramatic effect on the Uber marketplace.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2014): "Finite Population Causal Standard Errors," NBER Working Paper 20325.

——— (2017): "When Should You Adjust Standard Errors for Clustering?" NBER Working Paper 24003.

Athey, S. and G. W. Imbens (2016): "The Econometrics of Randomized Experiments," *arXiv*, https://arxiv.org/abs/1607.00698.

Azar, O. H. (2007): "The Social Norm of Tipping: A Review," *Journal of Applied Social Psychology*, 37, 380–402.

Bertrand, M., E. Duflo, and S. Mullainathan (2003): "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, 119, 249–275.

Brewer, M., T. F. Crossley, and R. Joyce (2017): "Inference with Differences-in-Differences Revisited," *Journal of Econometric Methods*, 7.

Cameron, A. C. and D. L. Miller (2015): "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 50, 317–372.

Chandar, B. K., U. Gneezy, J. A. List, and I. Muir (2019): "The Drivers of Social Preferences: Evidence from a Nationwide Tipping Field Experiment," NBER Working Paper.

Chetverikov, D., B. Larsen, and C. Palmer (2016): "IV Quantile Regression for Group-Level Treatments," *Econometrica*, 84, 809–833.

Conlin, M., M. Lynn, and T. O'Donoghue (2003): "The Norm of Restaurant Tipping," *Journal of Economic Behavior and Organization*, 52, 297–321.

Cook, C., R. Diamond, J. Hall, J. A. List, and P. Oyer (2018): "The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers," NBER Working Paper 24732.

Hall, J. V., J. J. Horton, and D. T. Knoepfle (2018): "Pricing Efficiently in Designed Markets: Evidence from Ride-Sharing," Working Paper, https://john-joseph-horton.com/papers/uber_price.pdf.

Hansen, C. B. (2007): "Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects," *Journal of Econometrics*, 140, 670–694.

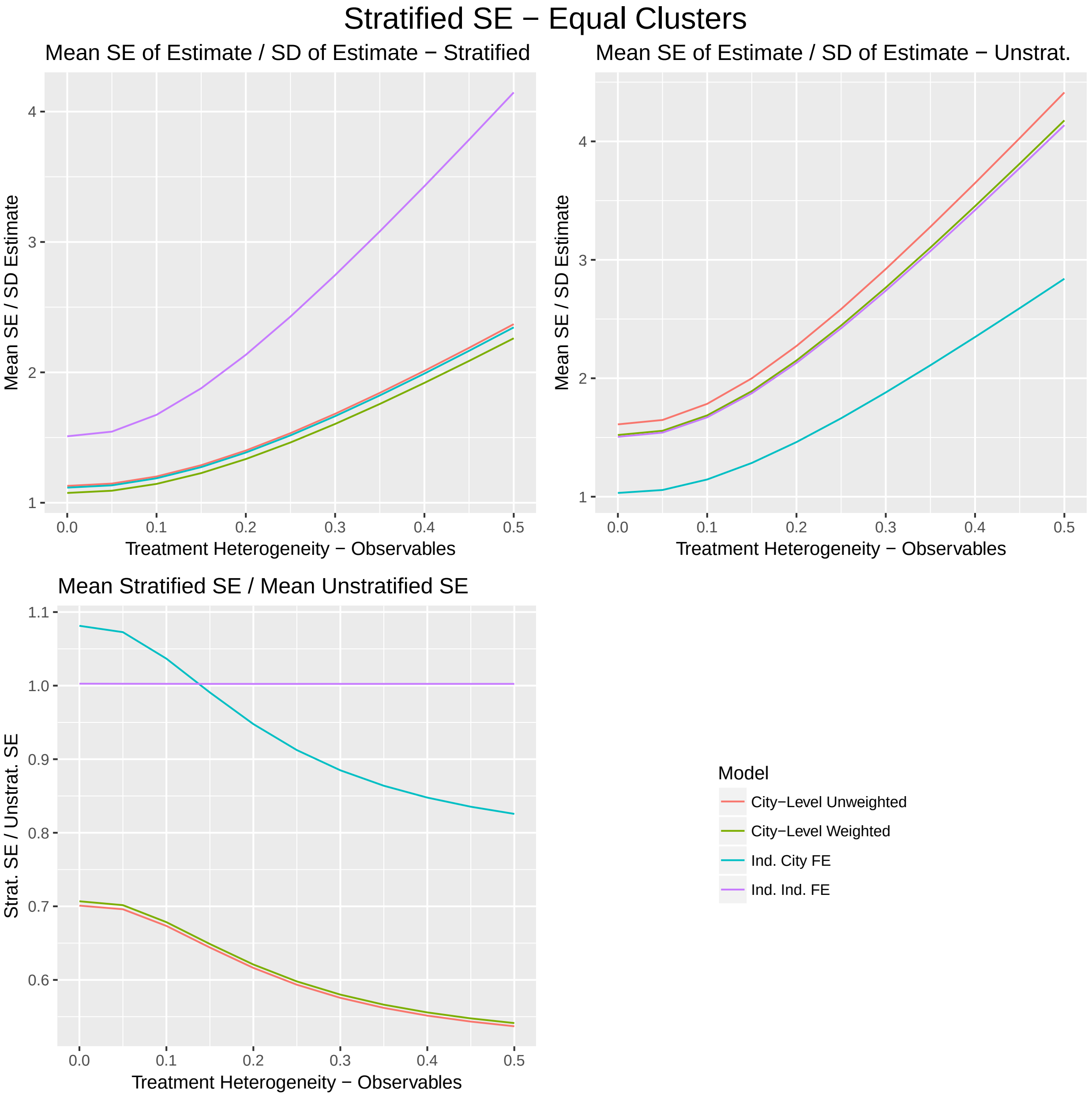
Liang, K.-Y. and S. L. Zeger (1986): "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lynn, M. (2016): "Why Are We More Likely to Tip Some Service Occupations than Others?: Theory, Evidence, and Implications," *Journal of Economic Psychology*, 54, 134–150.

Moulton, B. R. (1990): "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *The Review of Economics and Statistics*, 72, 334–338.

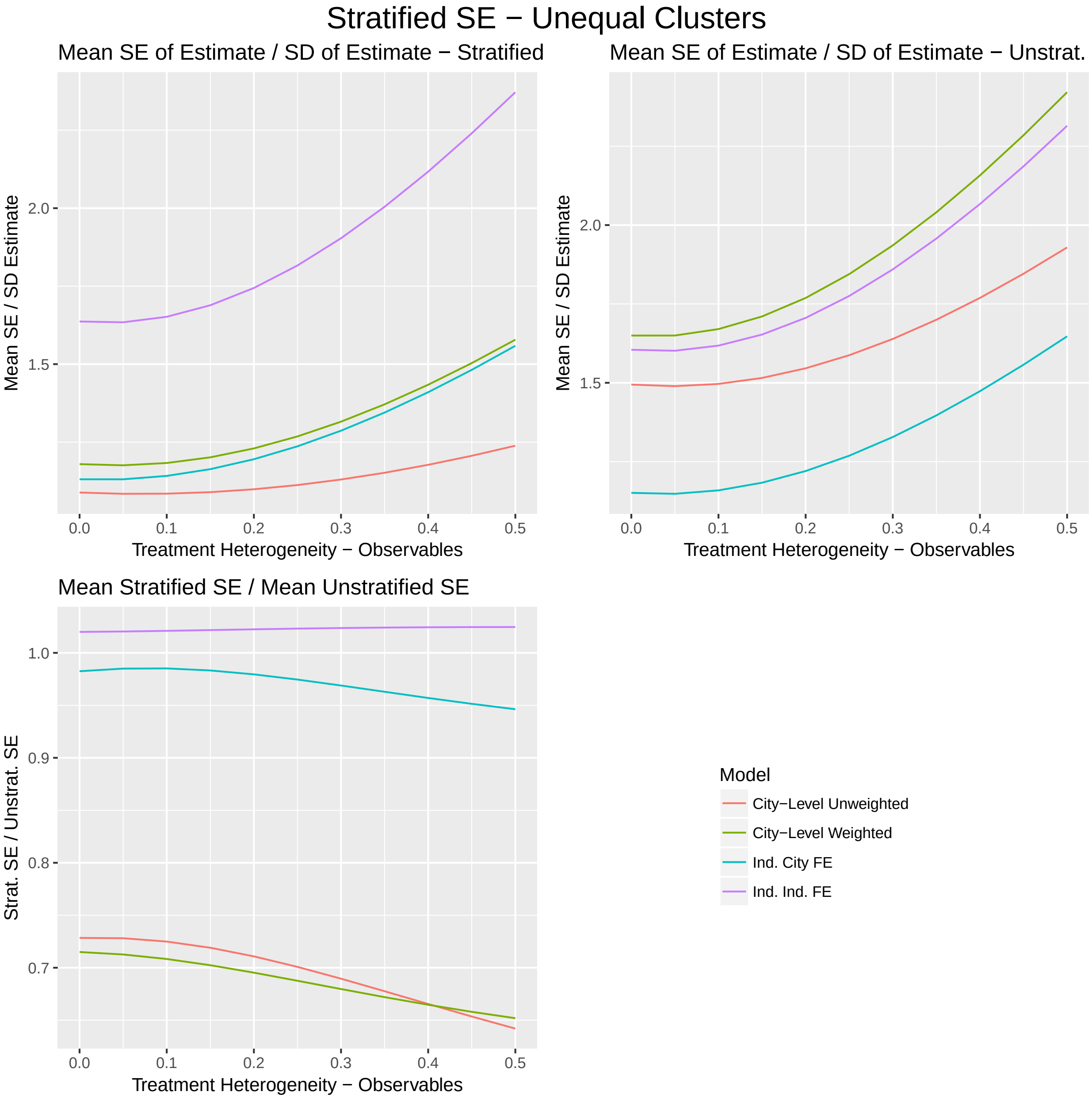
White, H. (2014): *Asymptotic Theory for Econometricians*, Academic Press.

Figure 1



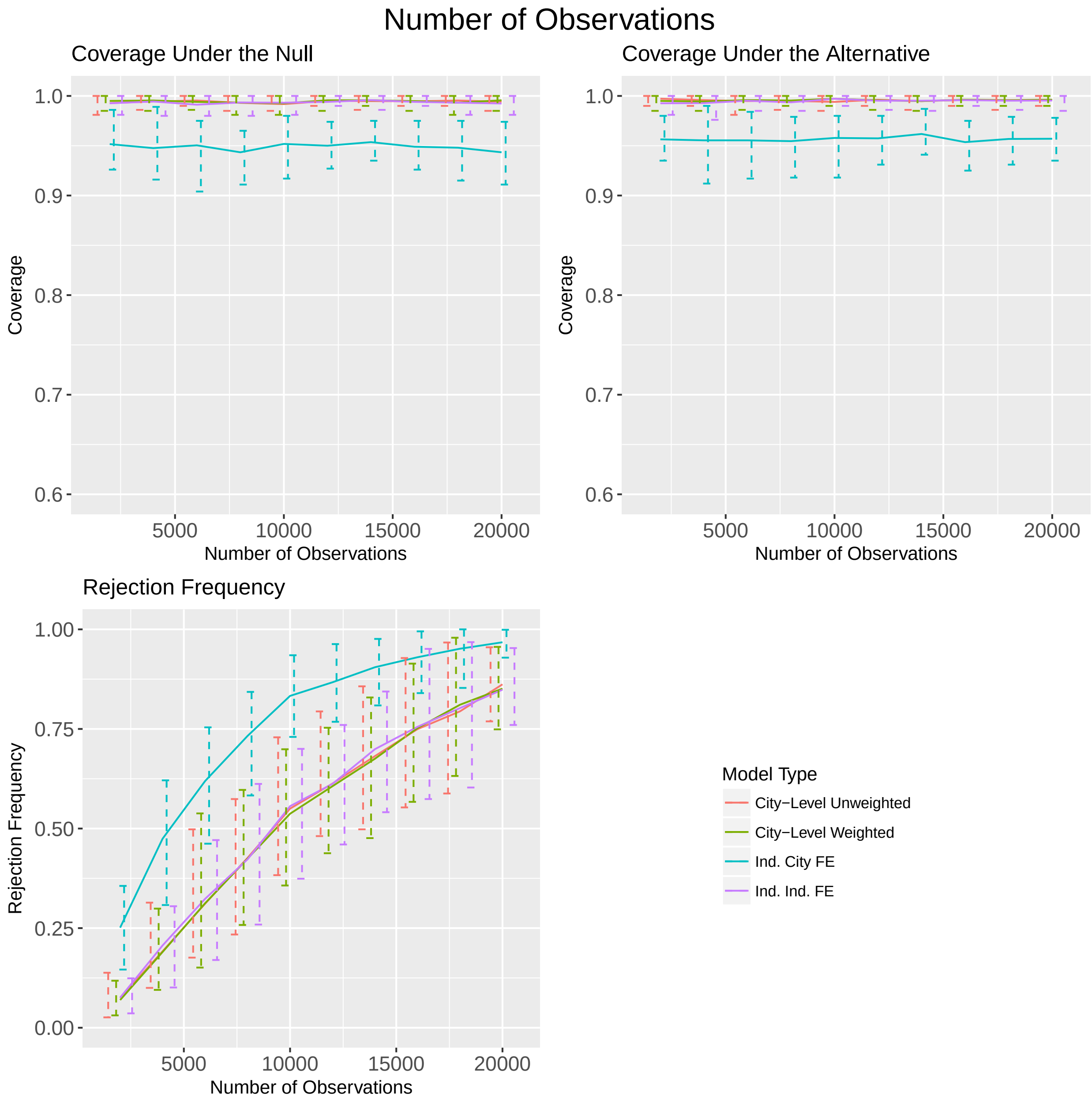
Note: These figures show the effect of varying the observable treatment heterogeneity parameter on estimated standard errors when clusters are all the same size. The top-left plot shows the mean standard error divided by the standard deviation of the estimated treatment effects when standard errors are stratified by the observable characteristics. Averages are taken over 500 different randomizations of treatment status across clusters. If the test is correctly sized, then the ratio of these quantities should equal 1 because the standard error would correctly measure the true variability in the estimated treatment effect (the converse may not be true if the standard error is sometimes an underestimate and sometimes an overestimate of the true variability). The top-right plot shows the same plot but when standard errors are not stratified. The bottom-left figure shows the ratio of the mean stratified standard error to the mean unstratified standard error. If this ratio is less than 1, then stratification may give higher rejection frequencies. Full DGP parameter values are in Table 36.

Figure 2



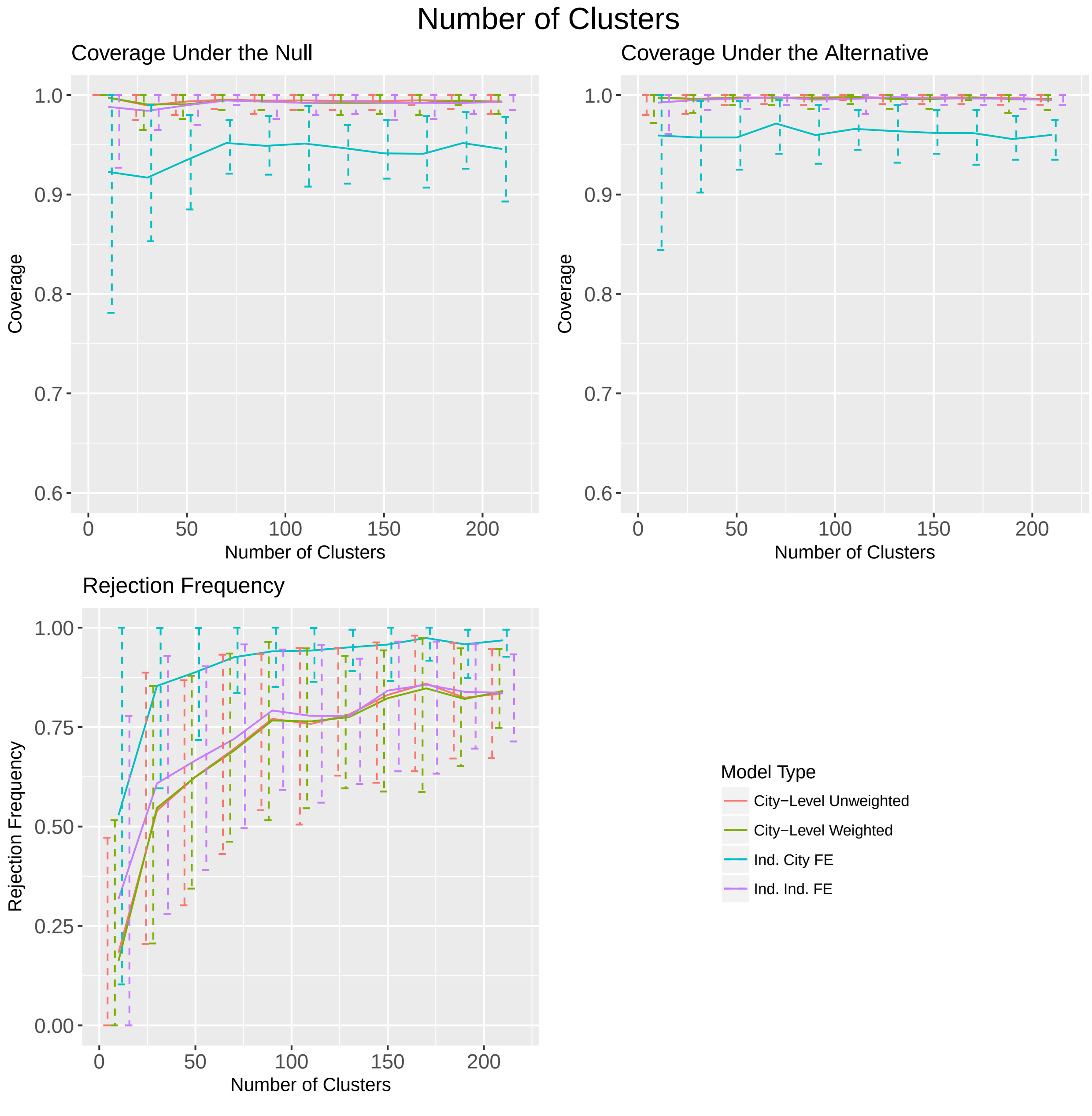
Note: These plots are similar to the ones in Figure 1, but in this case clusters vary in size. Full DGP parameter values are in Table 37.

Figure 3



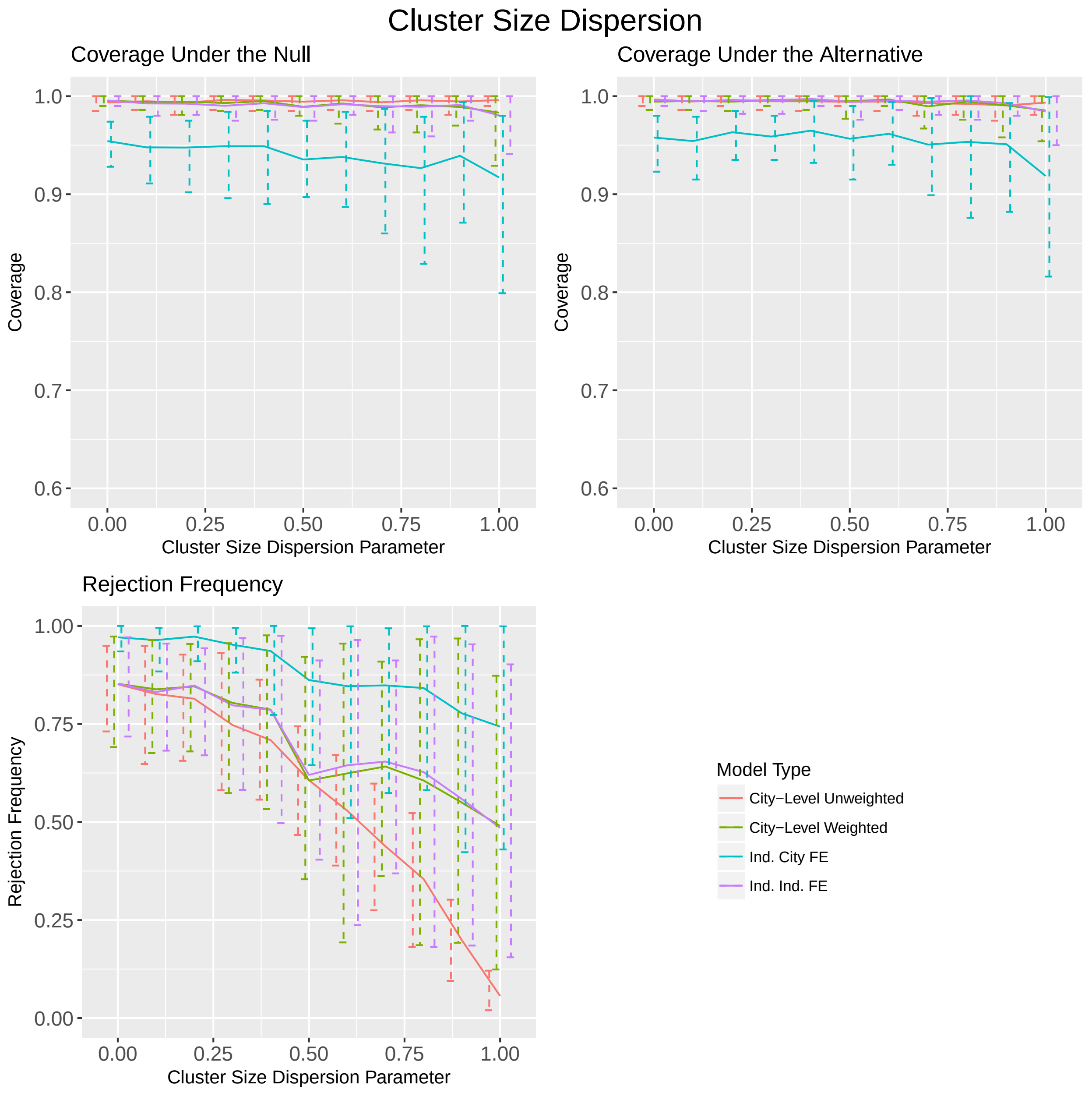
Note: These plots show the effect of varying the number of observations when clusters are all the same size. The top-left plot shows coverage results under the null hypothesis of no treatment effect as the number of observations varies. The top-right plot shows coverage results under the alternative hypothesis. The bottom-left plot shows rejection frequencies under the alternative hypothesis. Full DGP parameter values are in Table 38.

Figure 4



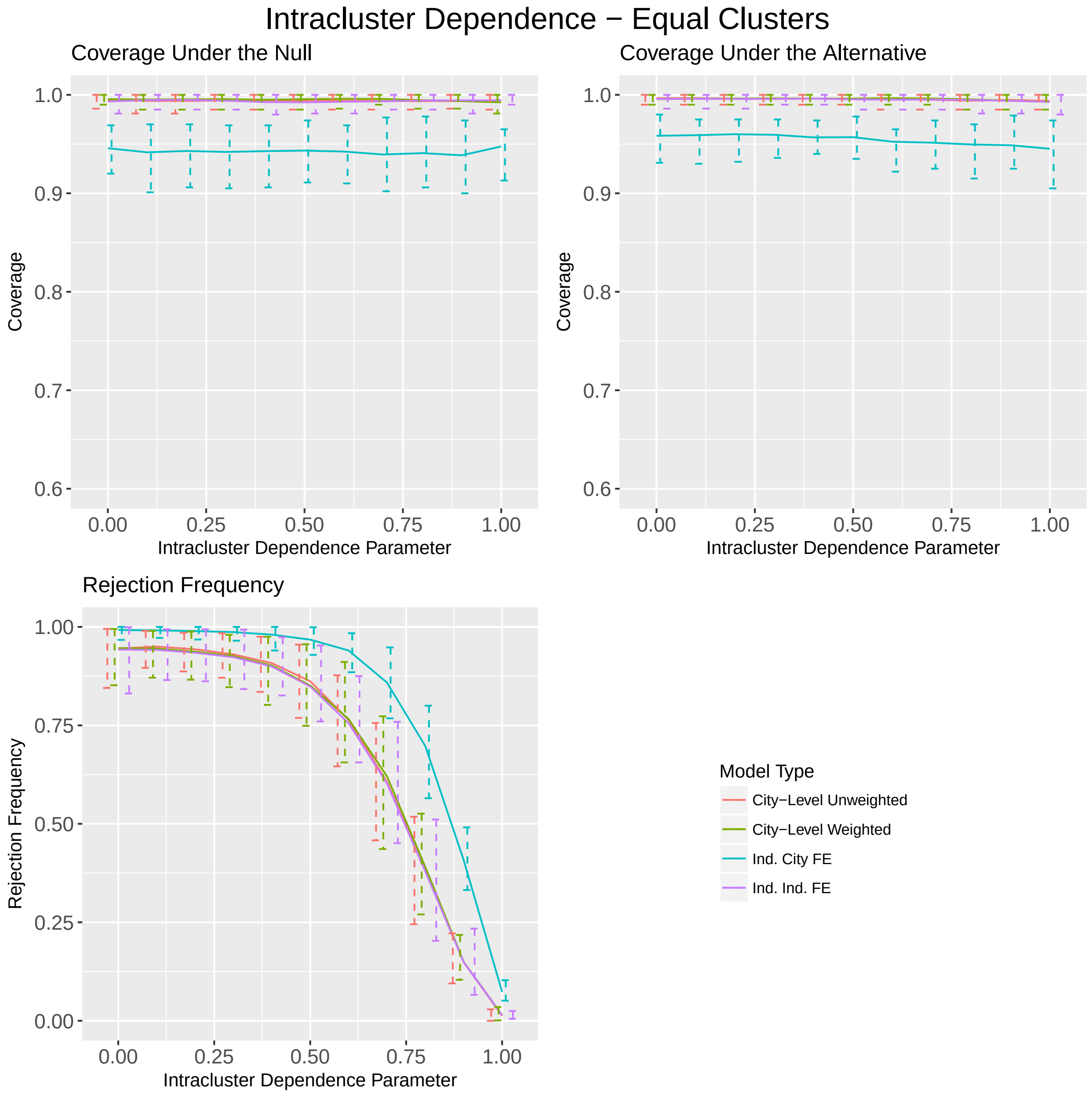
Note: These plots show the effect of varying the number of clusters when clusters are all the same size. Full DGP parameter values are in Table 39.

Figure 5



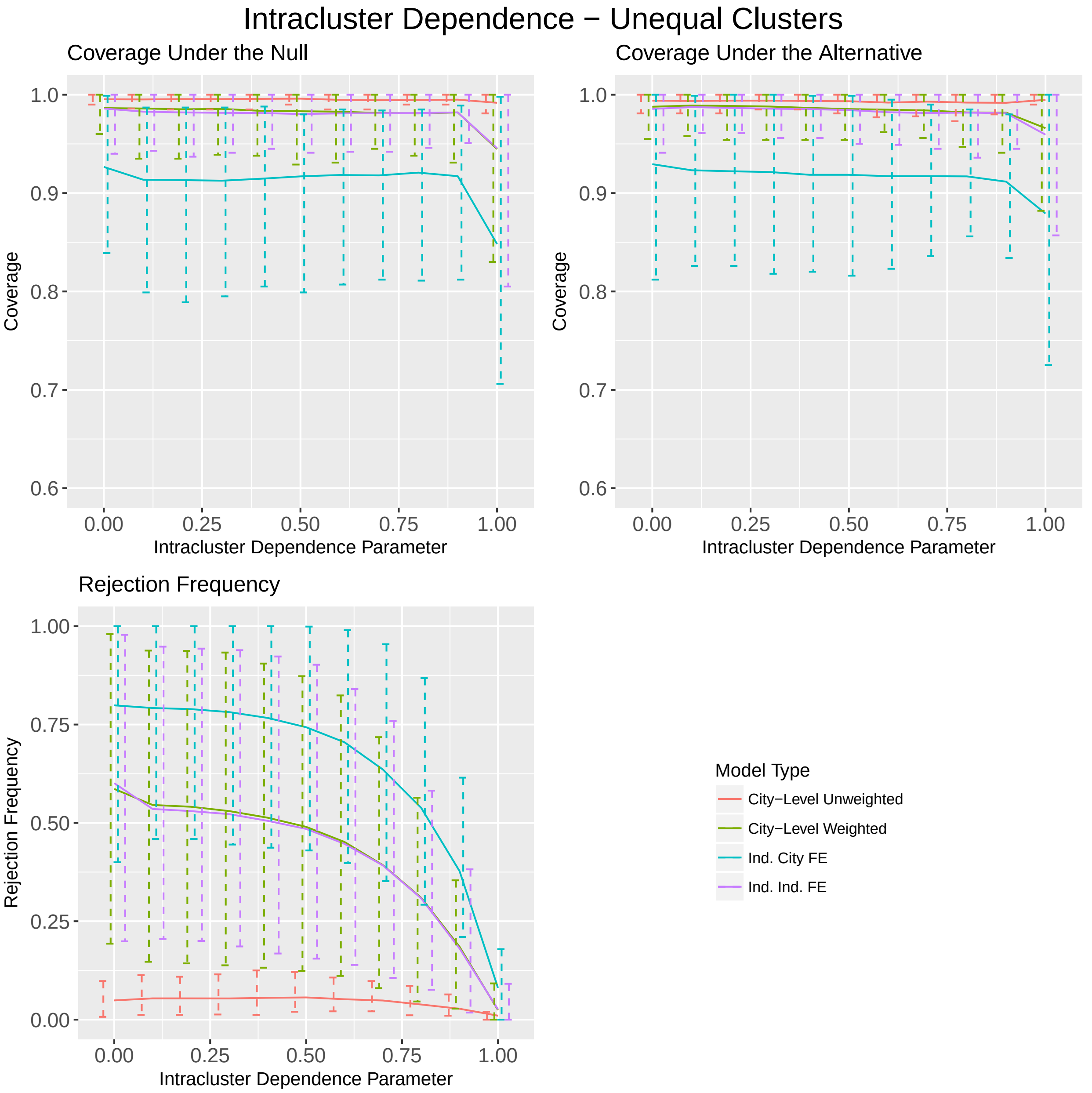
Note: These plots show the effect of varying the dispersion in cluster sizes. Cluster dispersion varies linearly from equality to the Uber data calibration. Full DGP parameter values are in Table 40.

Figure 6



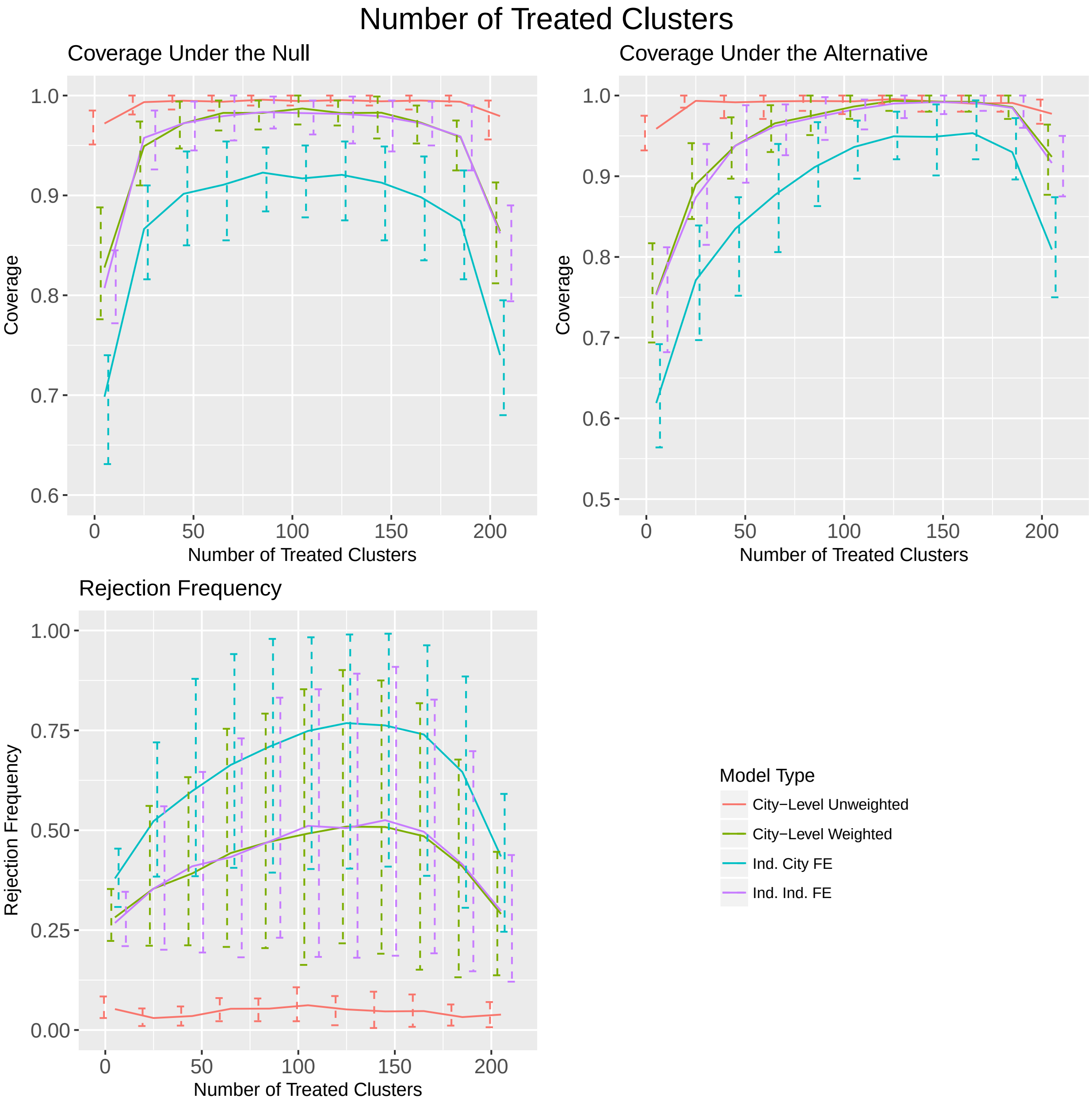
Note: These plots show the effect of varying the amount of spatial correlation in errors for each city-week. Clusters are all the same size. Full DGP parameter values are in Table 41.

Figure 7



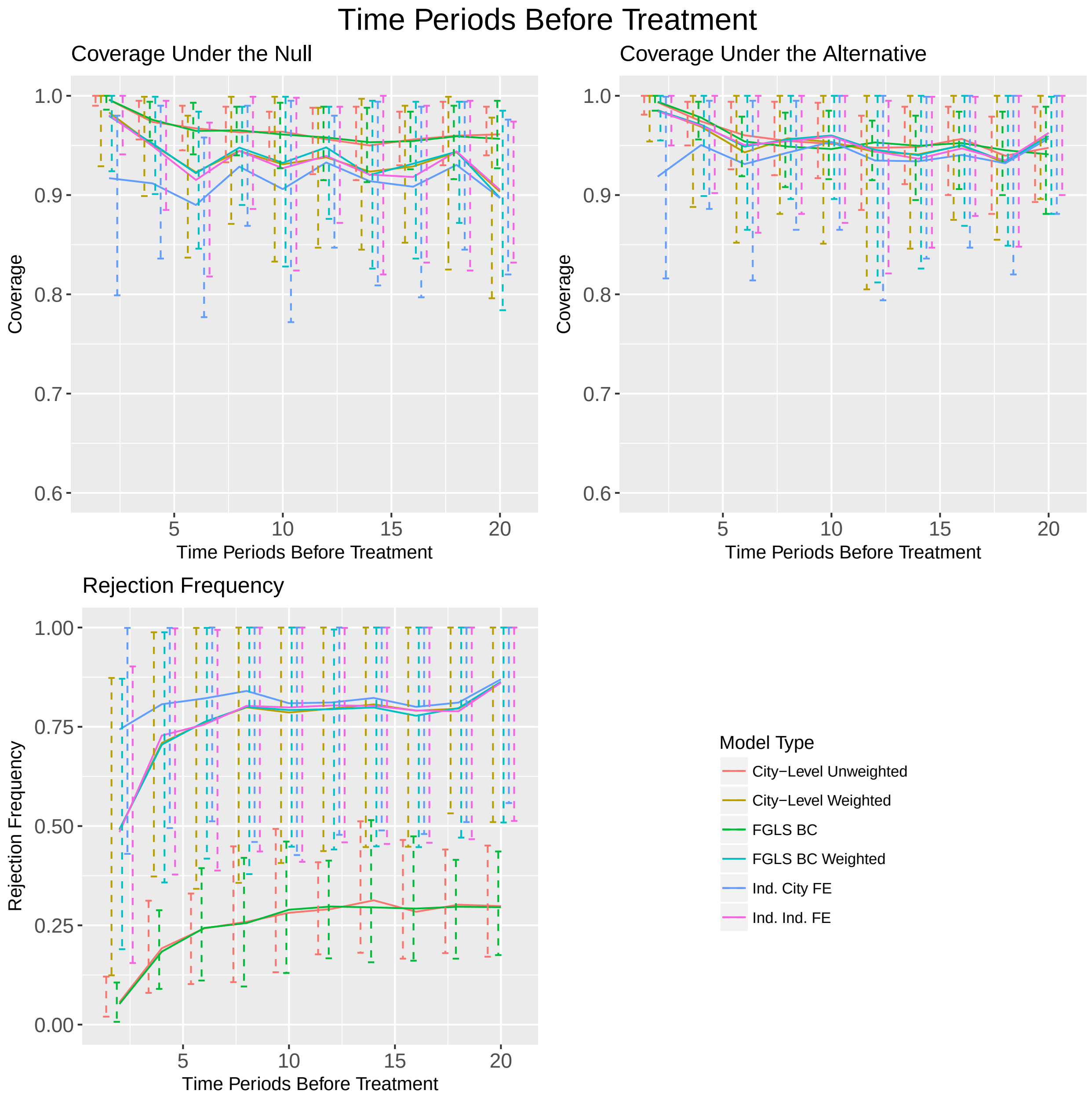
Note: These plots show the effect of varying the amount of spatial correlation in errors for each city-week. Cluster sizes are calibrated to the Uber data. Full DGP parameter values are in Table 42.

Figure 8



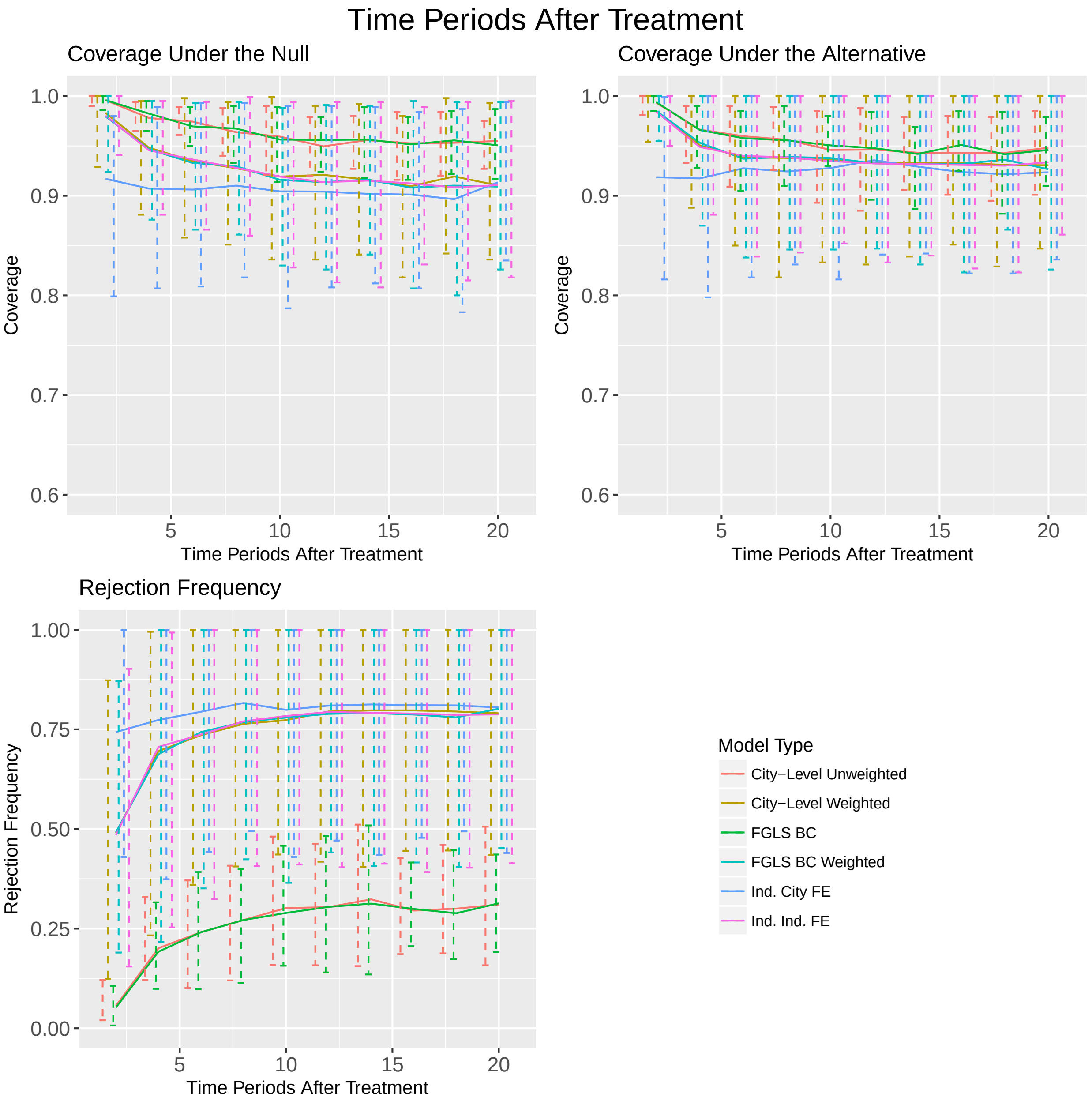
Note: These plots show the effect of varying the number of treated clusters. Cluster sizes are calibrated to the Uber data. Full DGP parameter values are in Table 43.

Figure 9



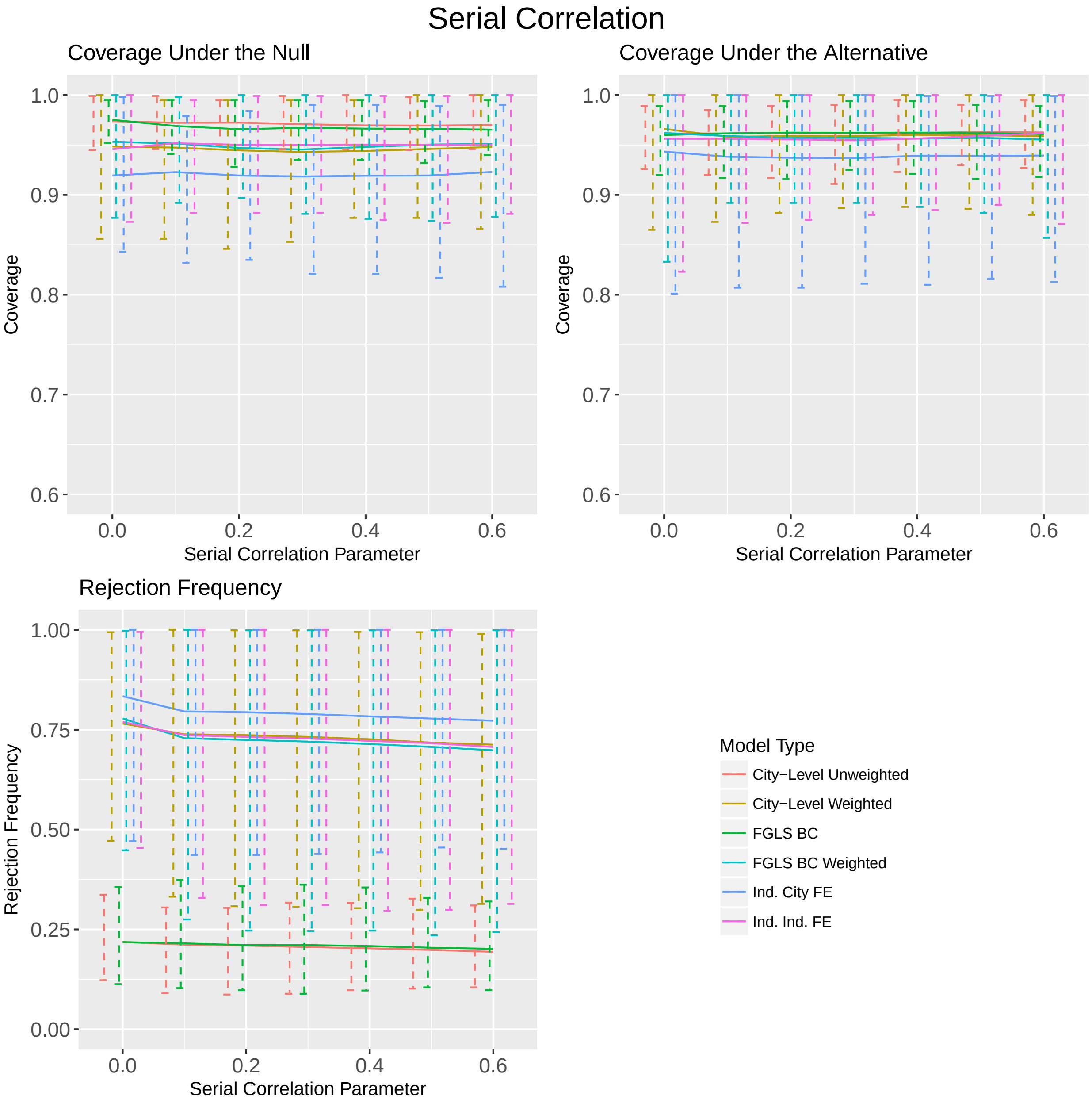
Note: These plots show the effect of varying the number of time periods before the treatment event. Treatment occurs during the last period for all results displayed. Cluster sizes are calibrated to the Uber data. Full DGP parameter values are in Table 44.

Figure 10



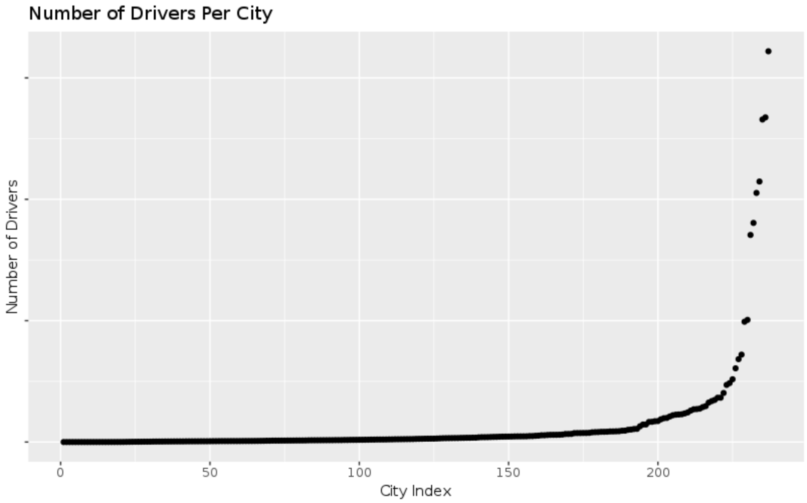
Note: These plots show the effect of varying the number of time periods after the treatment event. Treatment occurs during period 2 for all results displayed. Cluster sizes are calibrated to the Uber data. Full DGP parameter values are in Table 45.

Figure 11



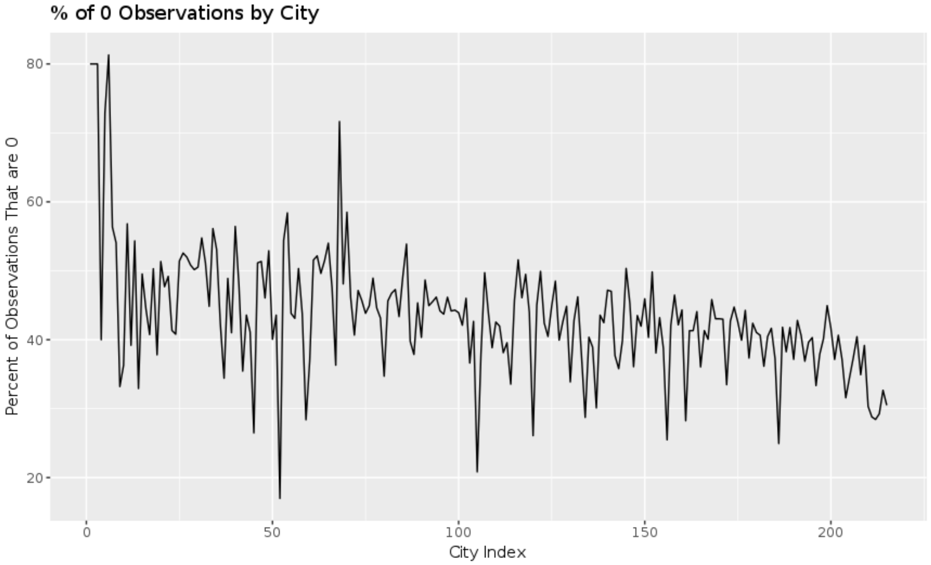
Note: These plots show the effect of varying the amount of serial correlation in the errors. In each case there are five time periods, with treatment occurring during the fifth period. Cluster sizes are calibrated to the Uber data. Full DGP parameter values are in Table 46.

Figure 12



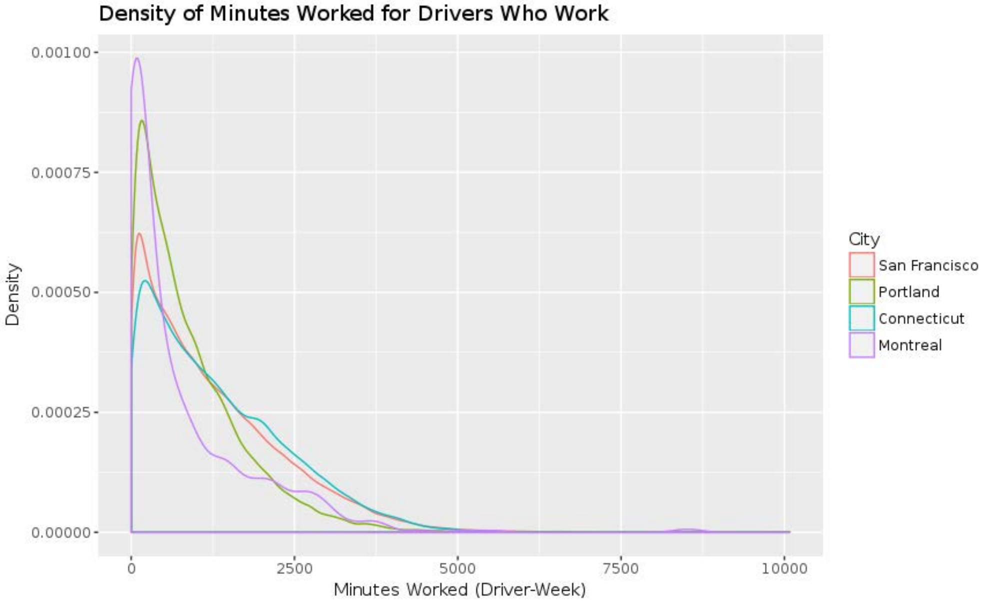
Note: This plot shows the number of drivers across cities in the United States and Canada. Cities are ordered from fewest drivers to most drivers.

Figure 13



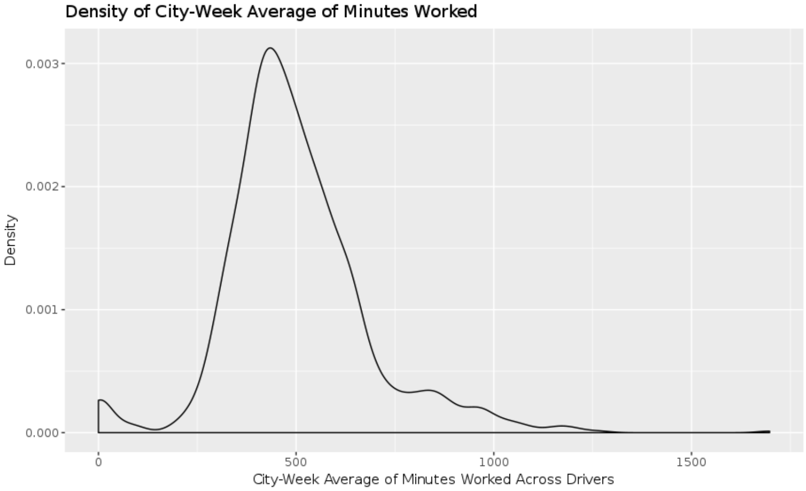
Note: This plot shows the percentage of driver-weeks with 0 minutes worked by city. Cities are ordered from fewest drivers to most drivers.

Figure 14



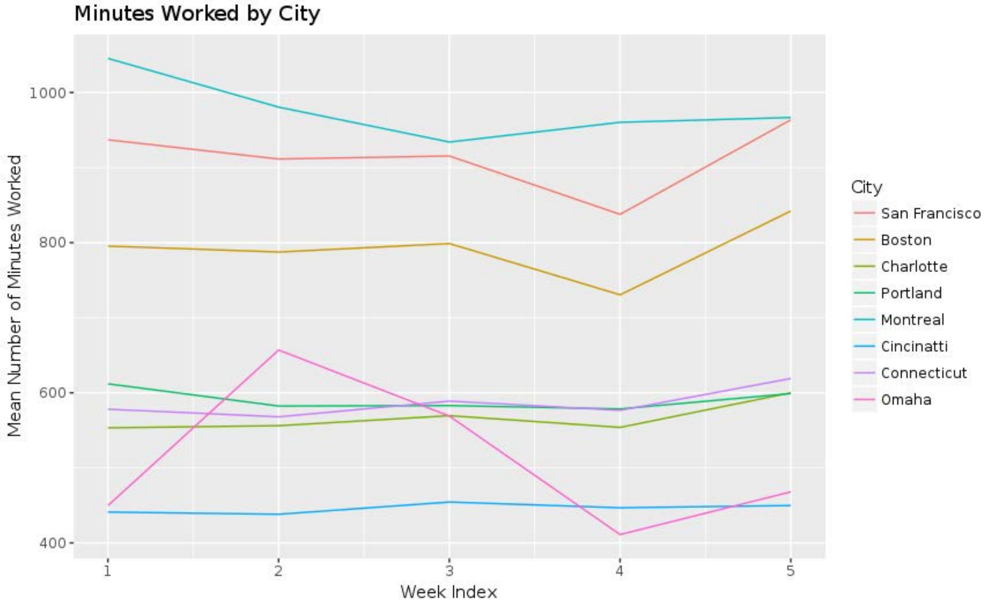
Note: This plot shows the distribution over driver-weeks of minutes worked conditional on working for four different cities: San Francisco, Portland, Montreal, and Connecticut.

Figure 15



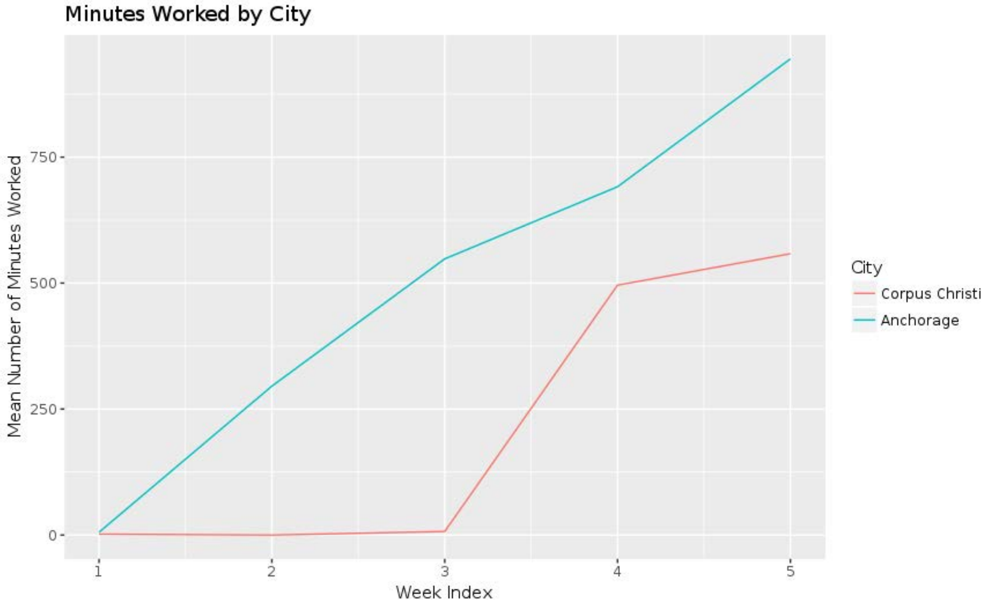
Note: This plot shows the density across city-weeks of the mean number of minutes worked by drivers.

Figure 16



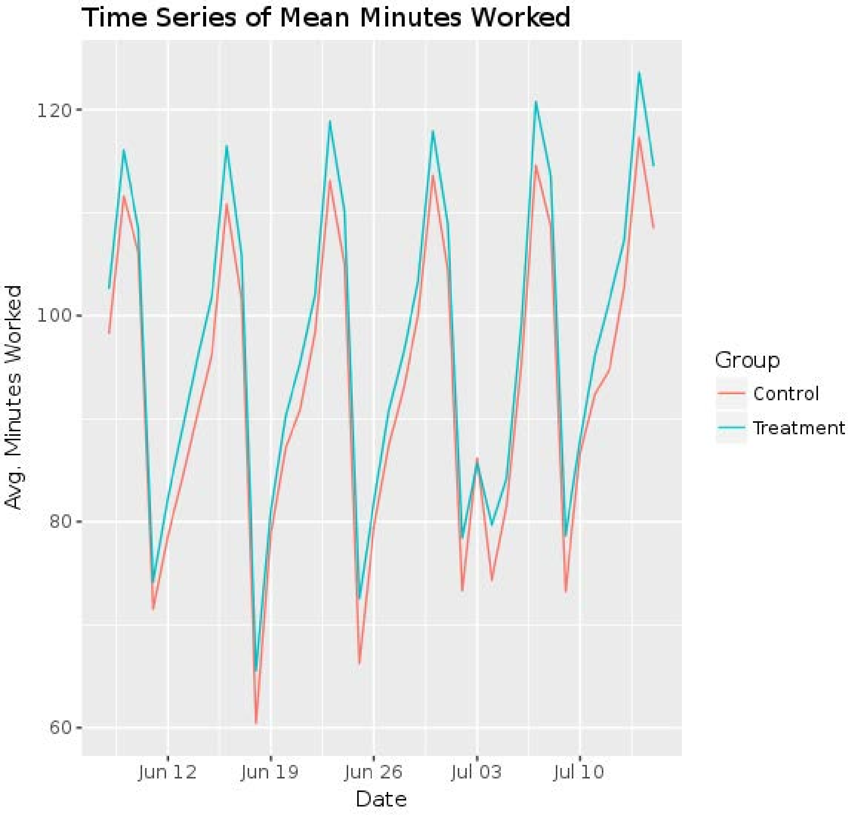
Note: This plot shows the mean number of minutes worked for each week in various cities. Each line in the plot is a different city.

Figure 17



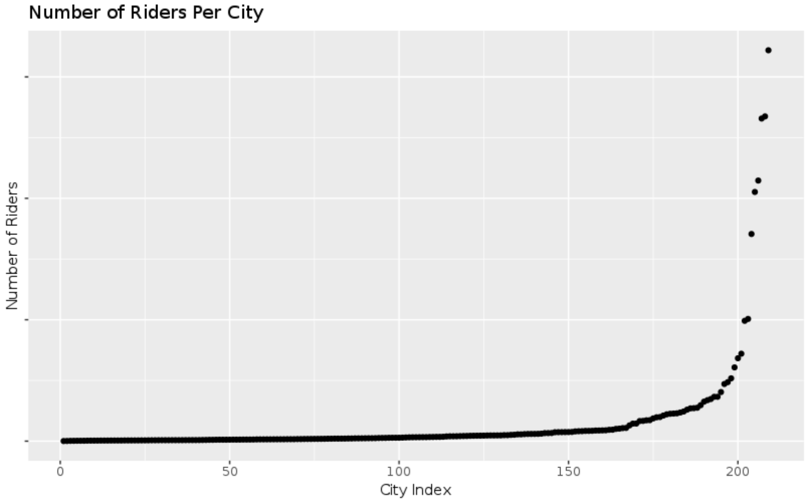
Note: This plot shows the mean number of minutes worked for each week in Anchorage and Corpus Christi, two cities that are removed from the sample used in the analysis. Uber began operating in Anchorage on June 15, 2017 following a hiatus due to regulatory reasons. This date corresponds to the start of week 2. Uber started operating in Corpus Christi on June 29, 2017, or at the start of week 4.

Figure 18



Note: This plot shows the time series for minutes worked for the treated and control groups. Tipping launched in treated cities on July 06.

Figure 19



Note: This plot shows the number of riders across cities in the United States and Canada. Cities are ordered from fewest riders to most riders.

Table 1

	Mean Coef.	SD Coef.	Mean SE	Coverage
Individual, No Clustering	-0.001	0.012	0.014	0.980
Individual, City Clustering	-0.001	0.012	0.013	0.966
City, No Clustering	-0.001	0.012	0.013	0.968

Note: This table shows simulation results under the null hypothesis of no effect for a one-period model with no cluster heterogeneity. All cities are the same size. Full DGP parameter values are in Table 27.

Table 2

	Mean Coef.	SD Coef.	Mean SE	Cov.	Rej. Freq.
Individual, No Clustering	0.035	0.013	0.014	0.968	0.720
Individual, City Clustering	0.035	0.013	0.013	0.958	0.782
City, No Clustering	0.035	0.013	0.013	0.958	0.772

Note: This table shows simulation results under the alternative hypothesis for a one-period model with no cluster heterogeneity. All cities are the same size. $\tau_p = 0.035$. Full DGP parameter values are in Table 27.

Table 3

	Mean Coef.	SD Coef.	Mean SE	Coverage
Individual, No Clustering	0.000	0.015	0.015	0.932
Individual, City Clustering	0.000	0.015	0.015	0.922
City, No Clustering	0.001	0.054	0.053	0.958
City, EHW SE	0.001	0.054	0.053	0.960
City, Size Weighted	0.000	0.016	0.014	0.922
City, Size Weighted, EHW SE	0.000	0.016	0.015	0.924

Note: This table shows simulation results under the null hypothesis of no effect for a one-period model with no cluster heterogeneity. Cluster sizes are calibrated to the Uber data. Full DGP parameter values are in Table 28.

Table 4

	Mean Coef.	SD Coef.	Mean SE	Cov.	Rej. Freq.
Individual, No Clustering	0.035	0.015	0.015	0.932	0.646
Individual, City Clustering	0.035	0.015	0.015	0.922	0.622
City, No Clustering	0.036	0.054	0.053	0.958	0.092
City, EHW SE	0.036	0.054	0.053	0.96	0.096
City, Size Weighted	0.035	0.016	0.014	0.922	0.664
City, Size Weighted, EHW SE	0.035	0.016	0.015	0.924	0.616

Note: This table shows simulation results under the alternative hypothesis for a one-period model with no cluster heterogeneity. Cluster sizes are calibrated to the Uber data. $\tau_p = 0.035$. Full DGP parameter values are in Table 28.

Table 5

	Mean Coef.	SD Coef.	Mean SE	Coverage
Individual, No Clustering	0.002	0.018	0.014	0.896
Individual, City Clustering	0.002	0.018	0.019	0.958
City, No Clustering	0.002	0.018	0.019	0.954

Note: This table shows simulation results under the null hypothesis of no effect for a one-period model with city-specific intercepts in the DGP. All cities are the same size. Full DGP parameter values are in Table 29.

Table 6

	Mean Coef.	SD Coef.	Mean SE	Cov.	Rej. Freq.
Individual, No Clustering	0.037	0.018	0.014	0.896	0.682
Individual, City Clustering	0.037	0.018	0.019	0.958	0.522
City, No Clustering	0.037	0.018	0.019	0.954	0.478

Note: This table shows simulation results under the alternative hypothesis for a one-period model with city-specific intercepts in the DGP. All cities are the same size. $\tau_p = 0.035$. Full DGP parameter values are in Table 29.

Table 7

	Mean Coef.	SD Coef.	Mean SE	Coverage
Individual, No Clustering	-0.002	0.051	0.015	0.380
Individual, City Clustering	-0.002	0.051	0.045	0.862
City, No Clustering	-0.002	0.065	0.066	0.950
City, EHW SE	-0.002	0.065	0.066	0.950
City, Size Weighted	-0.002	0.051	0.021	0.532
City, Size Weighted, EHW SE	-0.002	0.051	0.045	0.876

Note: This table shows simulation results under the null hypothesis of no effect for a one-period model with city-specific intercepts in the DGP. Cluster sizes are calibrated to the Uber data. Full DGP parameter values are in Table 30.

Table 8

	Mean Coef.	SD Coef.	Mean SE	Cov.	Rej. Freq.
Individual, No Clustering	0.033	0.051	0.015	0.380	0.532
Individual, City Clustering	0.033	0.051	0.045	0.862	0.182
City, No Clustering	0.033	0.065	0.066	0.950	0.074
City, EHW SE	0.033	0.065	0.066	0.950	0.072
City, Size Weighted	0.033	0.051	0.021	0.532	0.438
City, Size Weighted, EHW SE	0.033	0.051	0.045	0.876	0.174

Note: This table shows simulation results under the alternative hypothesis for a one-period model with city-specific intercepts in the DGP. Cluster sizes are calibrated to the Uber data. $\tau_p = 0.035$. Full DGP parameter values are in Table 30.

Table 9

	Mean Coef.	SD Coef.	Mean SE	Coverage
Individual, No Clustering	-0.001	0.029	0.021	0.816
Individual, No Clus. FE, City Clustering	0.000	0.041	0.037	0.898
Individual, City Clustering	-0.001	0.029	0.027	0.896
City, No Clustering	0.001	0.078	0.080	0.952
City, City Clustering	0.001	0.078	0.113	0.994
City, Clustered, No Clus. FE	0.001	0.062	0.062	0.952
City, Size Weighted, No Clustering	-0.001	0.029	0.021	0.826
City, Clustered, Size Weighted	-0.001	0.029	0.038	0.988
City, Size Weighted + Clustered, No Clus. FE	0.000	0.041	0.037	0.898

Note: This table shows simulation results under the null hypothesis of no effect for a two-period model with city effects and time effects in the DGP. Cluster sizes are calibrated to the Uber data. All model specifications include time and city fixed effects unless otherwise noted. Full DGP parameter values are in Table 31.

Table 10

	Mean Coef.	SD Coef.	Mean SE	Cov.	Rej. Freq.
Individual, No Clustering	0.033	0.021	0.021	0.95	0.338
Individual, No Clus. FE, City Clustering	0.032	0.034	0.031	0.906	0.19
Individual, City Clustering	0.033	0.021	0.02	0.93	0.35
City, No Clustering	0.026	0.075	0.076	0.952	0.06
City, City Clustering	0.026	0.075	0.107	1	0.004
City, Clustered, No Clus. FE	0.027	0.057	0.058	0.95	0.056
City, Size Weighted, No Clustering	0.033	0.021	0.021	0.948	0.354
City, Clustered, Size Weighted	0.033	0.021	0.029	0.992	0.136
City, Size Weighted + Clustered, No Clus. FE	0.032	0.034	0.032	0.906	0.188

Note: This table shows simulation results under the alternative hypothesis for a two-period model with city effects and time effects in the DGP. The treatment effect varies by individual. Cluster sizes are calibrated to the Uber data. All model specifications include time and city fixed effects unless otherwise noted. $\tau_p = 0.0326$. Full DGP parameter values are in Table 31.

Table 11

	Mean Coef.	SD Coef.	Mean SE	Coverage
Individual, No Clustering	0.000	0.019	0.021	0.966
Individual, City Clustering	0.000	0.019	0.018	0.936
Individual, Ind. FE, City Clustering	0.000	0.019	0.026	0.988
City, No Clustering	0.004	0.072	0.073	0.956
City, City Clustering	0.004	0.072	0.103	0.998
City, Size Weighted, No Clustering	0.000	0.019	0.020	0.962
City, Size Weighted, City Clustering	0.000	0.019	0.026	0.990

Note: This table shows simulation results under the null hypothesis of no effect for a two-period model with city effects, time effects, and individual effects in the DGP. Cluster sizes are calibrated to the Uber data. All model specifications include time and city fixed effects unless otherwise noted. Full DGP parameter values are in Table 32.

Table 12

	Mean Coef.	SD Coef.	Mean SE	Cov.	Rej. Freq.
Individual, No Clustering	0.064	0.026	0.021	0.89	0.808
Individual, City Clustering	0.064	0.026	0.028	0.932	0.548
Individual, Ind. FE, City Clustering	0.064	0.026	0.039	0.982	0.408
City, No Clustering	0.133	0.069	0.071	0.854	0.456
City, City Clustering	0.133	0.069	0.101	0.958	0.166
City, Size Weighted, No Clustering	0.064	0.026	0.021	0.874	0.810
City, Size Weighted, City Clustering	0.064	0.026	0.040	0.982	0.404

Note: This table shows simulation results under the alternative hypothesis for a two-period model with city effects, time effects, and individual effects in the DGP. The treatment effect varies by individual and by cluster. Cluster sizes are calibrated to the Uber data. All model specifications include time and city fixed effects unless otherwise noted. $\tau_p = 0.116$. Full DGP parameter values are in Table 32.

Table 13

	Mean Coef.	SD Coef.	Mean SE	Coverage
Individual, No Clustering	-0.001	0.025	0.018	0.862
Individual, City Clustering	-0.001	0.025	0.024	0.930
Individual, Ind. FE, City Clustering	-0.001	0.025	0.035	0.994
City, No Clustering	0.002	0.088	0.087	0.948
City, City Clustering	0.002	0.088	0.123	0.996
City, Size Weighted	-0.001	0.025	0.023	0.932
City, Size Weighted, City Clustering	-0.001	0.025	0.035	0.994

Note: This table shows simulation results under the null hypothesis of no effect for a two-period model with city effects, time effects, and individual effects in the DGP. Individual error terms within a city-week are drawn from Algorithm 1. Cluster sizes are calibrated to the Uber data. All model specifications include time and city fixed effects unless otherwise noted. Full DGP parameter values are in Table 33.

Table 14

	Mean Coef.	SD Coef.	Mean SE	Cov.	Rej. Freq.
Individual, No Clustering	0.064	0.027	0.019	0.83	0.822
Individual, City Clustering	0.064	0.027	0.029	0.958	0.582
Individual, Ind. FE, City Clustering	0.064	0.027	0.041	0.994	0.36
City, No Clustering	0.132	0.077	0.077	0.854	0.402
City, City Clustering	0.132	0.077	0.109	0.976	0.13
City, Size Weighted	0.064	0.027	0.024	0.92	0.72
City, Size Weighted, City Clustering	0.064	0.027	0.041	0.994	0.356

Note: This table shows simulation results under the alternative hypothesis for a two-period model with city effects, time effects, and individual effects in the DGP. The treatment effect varies by individual and by cluster. Individual error terms within a city-week are drawn from Algorithm 1. Cluster sizes are calibrated to the Uber data. All model specifications include time and city fixed effects unless otherwise noted. $\tau_p = 0.116$. Full DGP parameter values are in Table 33.

Table 15

	Mean Coef.	SD Coef.	Mean SE	Cov.	Rej. Freq.
Individual, City Clustering	0.35	0.026	0.028	0.962	1
Individual, City Clustering, Stratified	0.35	0.025	0.026	0.948	1
Individual, Ind. FE, City Clus.	0.35	0.026	0.039	0.998	1
Individual, Ind. FE, City Clus., Stratified	0.35	0.026	0.026	0.956	1
City, City Clustering	0.35	0.026	0.039	0.998	1
City, City Clustering, Stratified	0.35	0.026	0.026	0.956	1
City, Size Weighted, City Clustering	0.35	0.026	0.039	0.998	1
City, Weighted + Clustered, Stratified	0.35	0.026	0.026	0.956	1

Note: This table shows simulation results under the alternative hypothesis for a two-period model with city effects, time effects, and individual effects in the DGP. The treatment effect varies by individual and by cluster. Individual error terms within a city-week are drawn from Algorithm 1. Further, treatment effects vary by an observable cluster-level parameter. Clusters are equal sizes. All model specifications include time and city fixed effects unless otherwise noted. $\tau_p = 0.351$. Full DGP parameter values are in Table 34.

Table 16

	Mean Coef.	SD Coef.	Mean SE	Cov.	Rej. Freq.
Individual, City Clustering	0.524	0.025	0.025	0.942	1
Individual, City Clustering, Stratified	0.522	0.023	0.025	0.96	1
Individual, Ind. FE, City Clus.	0.524	0.025	0.035	0.988	1
Individual, Ind. FE, City Clus., Stratified	0.524	0.025	0.025	0.942	1
City, City Clustering	0.386	0.077	0.108	0.834	0.986
City, City Clustering, Stratified	0.386	0.077	0.076	0.574	1
City, Size Weighted, City Clustering	0.524	0.025	0.035	0.988	1
City, Weighted + Clustered, Stratified	0.524	0.025	0.025	0.942	1

Note: This table shows simulation results under the alternative hypothesis for a two-period model with city effects, time effects, and individual effects in the DGP. The treatment effect varies by individual and by cluster. Individual error terms within a city-week are drawn from Algorithm 1. Further, treatment effects vary by cluster size. Cluster sizes are calibrated to the Uber data. All model specifications include time and city fixed effects unless otherwise noted. $\tau_p = 0.534$. Full DGP parameter values are in Table 35.

Table 17

Level	Outcome	T	P	Control Mean	Treat. Mean
Individual	Minutes Worked	0.440	0.660	653.498	682.553
City	Minutes Worked	0.553	0.580	500.417	512.871
Individual	Did Work	0.501	0.617	0.636	0.646
City	Did Work	-0.114	0.909	0.569	0.568
Individual	Minutes Worked $\vert$Worked	0.393	0.694	$1,027.669$	$1,056.076$
City	Minutes Worked $\vert$Worked	1.019	0.308	863.138	889.840

Note: This table shows tests for equality of means between treatment and control in the period before the tipping launch. All estimates are clustered by city. The first row compares minutes worked between treatment and control at the driver level. The second row compares minutes worked after first collapsing to means across drivers for each city-week. Rows 3 and 4 show t-test estimates for whether the driver worked at least one minute. Rows 5 and 6 show estimates for how much drivers work conditional on working at least one minute.

Table 18

	Dependent variable:					
	Minutes Worked					
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	2.725 (2.408)	2.725 (8.691)	9.637 (7.188)	9.637 (10.498)	2.725 (4.579)	2.725 (9.742)
Unit	Driver	Driver	City	City	City Weighted	City Weighted
City Clustering	No	Yes	No	Yes	No	Yes
Counterfactual	697	697	527	527	697	697
% Change	0.39%	0.39%	1.83%	1.83%	0.39%	0.39%
Observations	4,072,910	4,072,910	1,045	1,045	1,045	1,045
R^2	0.032	0.032	0.941	0.941	0.978	0.978
Adjusted R^2	0.031	0.031	0.925	0.925	0.972	0.972
Note:						*p<0.1; **p<0.05; ***p<0.01

Note: This table shows regression estimates for the effect of the tipping launch on labor supply. All model specifications include time and city effects. In columns 1 and 2, observations are at the driver-week level. In columns 3 and 4, observations are at the city-week level, and regressions are unweighted. In columns 5 and 6, observations are at the city-week level, and regressions are weighted by the number of drivers for each city-week.

Table 19

	Mean Coef.	SD Coef.	Mean SE	Coverage
Individual, No Clustering	0.064	8.343	2.435	0.420
Individual, City Clustering	0.064	8.343	7.707	0.900
City, No Clustering	0.239	9.221	7.188	0.883
City, Size Weighted	0.239	9.221	10.481	0.983
City, Clustered	0.064	8.343	4.621	0.727
City, Clustered + Weighted	0.064	8.343	8.638	0.940

Note: This table shows results over placebo randomizations in the real Uber labor supply data. All model specifications include time and city fixed effects.

Table 20

	Dependent variable:			
	Minutes Worked			
	(1)	(2)	(3)	(4)
Treated	2.725 (9.717)	2.725 (8.693)	2.725 (8.692)	2.725 (8.712)
Unit	Driver	Driver	Driver	Driver
FE	Driver	Geohash	Time of Week	Geohash x Time of Week
City Clustering	Yes	Yes	Yes	Yes
Counterfactual	697	697	697	697
% Change	0.39%	0.39%	0.39%	0.39%
Observations	4,072,910	4,072,910	4,072,910	4,072,910
R^2	0.808	0.085	0.075	0.109
Adjusted R^2	0.760	0.085	0.075	0.105
Note:				*p<0.1; **p<0.05; ***p<0.01

Note: This table shows regression results for the effect of the tipping launch on labor supply when using more granular fixed effects in addition to city and week effects. Observations are at the driver-week level. All estimates are clustered by city. The specification in column 1 uses driver effects. Column 2 uses more granular location effects based on where drivers work, while column 3 uses more granular time effects based on when drivers typically work. Column 4 includes effects for the interaction of the location and time effects.

Table 21

	Mean Coef.	SD Coef.	Mean SE	Coverage
Individual FE	0.222	8.035	8.669	0.948
Top Geohash FE	0.858	8.547	7.663	0.892
Top Week Time FE	0.858	8.547	7.661	0.892
Geohash x Time	0.858	8.547	7.679	0.892

Note: This table shows results over placebo randomizations in the real Uber labor supply data with more granular fixed effects. All model specifications include time and city fixed effects.

Table 22

	Dependent variable:		
	Minutes Worked		
	(1)	(2)	(3)
Treated	2.725	2.725	2.725
	(8.715)	(8.725)	(8.697)
Unit	City x Geohash	City x Time	City x Geohash x Time
City Clustering	Yes	Yes	Yes
Counterfactual	537	580	577
% Change	0.51%	0.47%	0.47%
Observations	38,400	27,740	160,580
R^2	0.329	0.355	0.242
Adjusted R^2	0.325	0.350	0.241
Note:			*p<0.1; **p<0.05; ***p<0.01

Note: This table shows regression results for the effect of the tipping launch on labor supply when using observations more granular than city-weeks. All estimates are clustered by city and weighted by the number of drivers. All model specifications include time and city fixed effects. The specification in column 1 collapses to the city x geohash level. Column 2 collapses to the city x typical time block level, while column 3 collapses to the city x geohash x time block level.

Table 23

	Mean Coef.	SD Coef.	Mean SE	Coverage
Geohash x City	0.064	8.343	7.728	0.900
Week Time x City	0.064	8.343	7.736	0.900
Geohash x Week Time x City	0.064	8.343	7.712	0.900

Note: This table shows results over placebo randomizations in the Uber labor supply data collapsing to level more granular than cities. All model specifications include time and city fixed effects.

Table 24

	Dependent variable:			
	Minutes Worked			
	(1)	(2)	(3)	(4)
Treated	2.725 (8.734)	2.725 (8.734)	9.637 (9.394)	2.725 (8.734)
Unit	Individual	Individual	City	City Weighted
FE	City	Individual	City	City
City Clustering	Yes	Yes	Yes	Yes
Stratification	Yes	Yes	Yes	Yes
Counterfactual	697	697	527	697
% Change	0.39%	0.39%	1.83%	0.39%
Observations	4,072,910	4,072,910	1,045	1,045
Note:				*p<0.1; **p<0.05; ***p<0.01

Note: This table shows regression results for the effect of the tipping launch on labor supply when stratifying standard errors by cluster size tercile. All estimates are clustered by city and include time and city fixed effects.

Table 25

	Mean Coef.	SD Coef.	Mean SE	Coverage
Individual, Driver FE	0.124	8.278	7.725	0.920
Individual, City FE	0.187	8.353	7.787	0.900
City, Clustered	-0.526	9.220	9.351	0.968
City, Clustered + Weighted	0.124	8.278	7.725	0.920

Note: This table shows results over placebo randomizations in the real Uber labor supply data with standard errors stratified by city size tercile.

Table 26

	Dependent variable:					
	Minutes Worked	Uniq. Rider Sess.	Utilization	Total	Post-Comm.	Organic
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	2.725	−0.013	−0.004	−0.227	−0.086	−0.173
	(9.742)	(0.028)	(0.010)	(0.554)	(0.436)	(0.498)
Counterfactual	697	0.997	0.489	16.22	12.42	15.03
% Change	0.39%	-1.320%	-0.750%	-1.40%	-0.69%	-1.15%
Observations	1,045	1,045	1,045	1,045	1,045	1,045
R^2	0.978	0.964	0.964	0.966	0.969	0.953
Adjusted R^2	0.972	0.955	0.954	0.958	0.961	0.941

Note: *p<0.1; **p<0.05; ***p<0.01

Note: This table shows regression estimates of the effect of tipping on labor supply, demand, utilization, and earnings. Observations are at the city-week level and weighted by the number of drivers (or weighted by the number of riders in the case of sessions). Standard errors are clustered by city. Minutes worked estimates are the same as in Column 6 in Table 18. Unique rider sessions are blocks of time when a rider opens the app, enters a destination, and considers taking a trip of any product type (e.g. UberX or UberPOOL). Gross hourly earnings include tips and non-trip driving incentives but are calculated before the commission is deducted. Post-commission hourly earnings are the same as gross hourly earnings minus the commission paid to Uber. Organic hourly earnings are calculated before the commission is deducted but exclude incentives and tips.

Appendix A Simulation Calibrations

Table 27

Parameter	Meaning	Value
τ	Treatment Effect Parameter	0.035
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	1
$T_{treated}$	Time of Treatment	1
R	Number of Iterations in Simulation	500

Note: DGP parameter settings for Tables 1 and 2. Clusters are equal sizes.

Table 28

Parameter	Meaning	Value
τ	Treatment Effect Parameter	0.035
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	1
$T_{treated}$	Time of Treatment	1
R	Number of Iterations in Simulation	500

Note: DGP parameter settings for Tables 3 and 4. Cluster sizes are calibrated to the Uber data.

Table 29

Parameter	Meaning	Value
τ	Average Treatment Effect	0.035
σ_c	SD of City FE	0.1
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	1
$T_{treated}$	Time of Treatment	1
R	Number of Iterations in Simulation	500

Note: DGP parameter settings for Tables 5 and 6. Clusters are equal sizes.

Table 30

Parameter	Meaning	Value
τ	Average Treatment Effect	0.035
σ_c	SD of City FE	0.1
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	1
$T_{treated}$	Time of Treatment	1
R	Number of Iterations in Simulation	500

Note: DGP parameter settings for Tables 7 and 8. Cluster sizes are calibrated to the Uber data.

Table 31

Parameter	Meaning	Value
τ	Average Treatment Effect	0.035
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	2
$T_{treated}$	Time of Treatment	2
R	Number of Iterations in Simulation	500

Note: DGP parameter settings for Tables 9 and 10. Cluster sizes are calibrated to the Uber data.

Table 32

Parameter	Meaning	Value
τ	Average Treatment Effect	0.1
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.1
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	2
$T_{treated}$	Time of Treatment	2
R	Number of Iterations in Simulation	500

Note: DGP parameter settings for Tables 11 and 12. Cluster sizes are calibrated to the Uber data.

Table 33

Parameter	Meaning	Value
τ	Average Treatment Effect	0.1
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.1
ρ	Spatial Correlation Parameter	0.5
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	2
$T_{treated}$	Time of Treatment	2
R	Number of Iterations in Simulation	500

Note: DGP parameter settings for Tables 13 and 14. Cluster sizes are calibrated to the Uber data.

Table 34

Parameter	Meaning	Value
τ	Average Treatment Effect	0.35
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.02
τ_z	Observable Het. Param.	0.1
ρ	Spatial Correlation Parameter	0.5
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	2
$T_{treated}$	Time of Treatment	2
R	Number of Iterations in Simulation	500

Note: DGP parameter settings for Table 15. Clusters are equal sizes.

Table 35

Parameter	Meaning	Value
τ	Average Treatment Effect	0.35
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.02
τ_z	Observable Het. Param.	0.1
ρ	Spatial Correlation Parameter	0.5
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	2
$T_{treated}$	Time of Treatment	2
R	Number of Iterations in Simulation	500

Note: DGP parameter settings for Table 16. Cluster sizes are calibrated to Uber data.

Table 36

Parameter	Meaning	Value
τ	Average Treatment Effect	0.1
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.1
τ_z	Observable Het. Param.	Varies
ρ	Spatial Correlation Parameter	0.5
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	2
$T_{treated}$	Time of Treatment	2
R	Number of Iterations in Simulation	500

Note: DGP parameter settings for Figure 1. Clusters are equal sizes.

Table 37

Parameter	Meaning	Value
τ	Average Treatment Effect	0.1
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.1
τ_z	Observable Het. Param.	Varies
ρ	Spatial Correlation Parameter	0.5
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	2
$T_{treated}$	Time of Treatment	2
R	Number of Iterations in Simulation	500

Note: DGP parameter settings for Figure 2. Cluster sizes are calibrated to the Uber data.

Table 38

Parameter	Meaning	Value
τ	Average Treatment Effect	0.1
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.1
τ_z	Observable Het. Param.	0
ρ	Spatial Correlation Parameter	0.5
C	Number of Clusters	200
N	Number of Individuals	Varies
σ	SD of Error Term	1
T	Number of Time Periods	2
$T_{treated}$	Time of Treatment	2
R	Number of Re-randomizations	200
P	Number of Populations Drawn	25

Note: DGP parameter settings for Figure 3. Clusters are equal sizes.

Table 39

Parameter	Meaning	Value
τ	Average Treatment Effect	0.1
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.1
τ_z	Observable Het. Param.	0
ρ	Spatial Correlation Parameter	0.5
C	Number of Clusters	Varies
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	2
$T_{treated}$	Time of Treatment	2
R	Number of Re-randomizations	200
P	Number of Populations Drawn	25

Note: DGP parameter settings for Figure 4. Clusters are equal sizes.

Table 40

Parameter	Meaning	Value
τ	Average Treatment Effect	0.1
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.1
τ_z	Observable Het. Param.	0
ρ	Spatial Correlation Parameter	0.5
C	Number of Clusters	212
N	Number of Individuals	20,000
δ	Cluster Size Variation	Varies
σ	SD of Error Term	1
T	Number of Time Periods	2
$T_{treated}$	Time of Treatment	2
R	Number of Re-randomizations	200
P	Number of Populations Drawn	25

Note: DGP parameter settings for Figure 5.

Table 41

Parameter	Meaning	Value
τ	Average Treatment Effect	0.1
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.1
τ_z	Observable Het. Param.	0
ρ	Spatial Correlation Parameter	Varies
C	Number of Clusters	200
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	2
$T_{treated}$	Time of Treatment	2
R	Number of Re-randomizations	200
P	Number of Populations Drawn	25

Note: DGP parameter settings for Figure 6. Clusters are equal sizes.

Table 42

Parameter	Meaning	Value
τ	Average Treatment Effect	0.1
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.1
τ_z	Observable Het. Param.	0
ρ	Spatial Correlation Parameter	Varies
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	2
$T_{treated}$	Time of Treatment	2
R	Number of Re-randomizations	200
P	Number of Populations Drawn	25

Note: DGP parameter settings for Figure 7. Cluster sizes calibrated to the Uber data.

Table 43

Parameter	Meaning	Value
τ	Average Treatment Effect	0.1
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.1
τ_z	Observable Het. Param.	0
ρ	Spatial Correlation Parameter	0.5
C	Number of Clusters	212
N	Number of Individuals	20,000
$\|C_{treated}\|$	Number of Treated Clusters	Varies
σ	SD of Error Term	1
T	Number of Time Periods	2
$T_{treated}$	Time of Treatment	2
R	Number of Re-randomizations	200
P	Number of Populations Drawn	25

Note: DGP parameter settings for Figure 8. Cluster sizes calibrated to the Uber data.

Table 44

Parameter	Meaning	Value
τ	Average Treatment Effect	0.1
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.1
τ_z	Observable Het. Param.	0
ρ	Spatial Correlation Parameter	0.5
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	Varies
$T_{treated}$	Time of Treatment	Same as T
R	Number of Re-randomizations	200
P	Number of Populations Drawn	25

Note: DGP parameter settings for Figure 9. Cluster sizes calibrated to the Uber data.

Table 45

Parameter	Meaning	Value
τ	Average Treatment Effect	0.1
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.1
τ_z	Observable Het. Param.	0
ρ	Spatial Correlation Parameter	0.5
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	Varies
$T_{treated}$	Time of Treatment	2
R	Number of Re-randomizations	200
P	Number of Populations Drawn	25

Note: DGP parameter settings for Figure 10. Cluster sizes calibrated to the Uber data.

Table 46

Parameter	Meaning	Value
τ	Average Treatment Effect	0.1
σ_τ	SD of Ind. Treatment Effect	0.5
σ_c	SD of City FE	0.1
σ_t	SD of Time FE	0.1
σ_γ	SD of Ind. FE	0.1
σ_h	SD of City Treatment Effect Het.	0.1
ϕ	Serial Correlation Parameter	Varies
τ_z	Observable Het. Param.	0
ρ	Spatial Correlation Parameter	0.5
C	Number of Clusters	212
N	Number of Individuals	20,000
σ	SD of Error Term	1
T	Number of Time Periods	5
$T_{treated}$	Time of Treatment	5
R	Number of Re-randomizations	200
P	Number of Populations Drawn	25

Note: DGP parameter settings for Figure 11. Cluster sizes calibrated to the Uber data.

Appendix B Alternative Regression Specifications and Outcomes

Table 47

	Var	Coef.	SE	tstat	pval	dof	mean	% change
1	Unique Sessions	-0.013	0.028	-0.461	0.645	831	0.997	-1.319
2	Riders w/ Sessions	-0.002	0.006	-0.331	0.741	831	0.342	-0.605
3	Unique Sessions Cond. >0	0.034	0.034	0.985	0.325	831	2.834	1.199

Note: This table shows regression results for rider-level demand outcomes. Regressions are estimated at the city level and weighted by the number of riders in each city-week. Standard errors are clustered by city. Row 1 shows estimates for the number of unique rider sessions. Row 2 shows estimates for an indicator for whether the rider had at least on session. Row 3 shows estimates for the number of unique sessions conditional on the rider having at least one unique session for the week.

Table 48

	Var	Coef.	SE	tstat	pval	dof	mean	% change
1	Total Earnings	-1.940	6.819	-0.285	0.776	831	209.688	-0.925
2	Organic Earnings	-3.953	6.835	-0.578	0.563	831	192.500	-2.054
3	Minutes Worked	2.725	9.742	0.280	0.780	831	697.208	0.391
4	Incentive Earnings	-0.161	0.910	-0.177	0.859	831	6.006	-2.685
5	Surge Earnings	-2.283	3.185	-0.717	0.474	831	5.228	-43.665
6	Post-Commission Earn.	-0.486	5.261	-0.092	0.926	831	161.280	-0.301
7	Hourly Earnings	-0.227	0.554	-0.410	0.682	831	16.219	-1.399
8	Organic Hourly	-0.173	0.498	-0.347	0.729	831	15.030	-1.151
9	Utilization	-0.004	0.010	-0.378	0.706	831	0.489	-0.746
10	Incentive Hourly	-0.039	0.051	-0.752	0.452	831	0.462	-8.348
11	Surge Hourly	-0.173	0.212	-0.816	0.415	831	0.459	-37.731
12	Post-Commission Hourly	-0.086	0.436	-0.196	0.844	831	12.423	-0.690
13	Minutes Worked	2.725	9.742	0.280	0.780	831	697.208	0.391
14	Min. Worked - Intensive	-3.339	13.065	-0.256	0.798	831	1,054.207	-0.317
15	Did Work	0.004	0.003	1.225	0.221	831	0.662	0.623
16	Completed Trips	-0.128	0.252	-0.508	0.612	831	28.323	-0.452
17	Mean Fare	-0.317	0.277	-1.144	0.253	831	12.905	-2.457
18	Trip Distance	0	0.034	0.003	0.998	831	6.625	0.002
19	Trip Duration	5.510	8.625	0.639	0.523	831	972.191	0.567
20	Trip Surge	-0.007	0.011	-0.650	0.516	831	1.027	-0.721

Note: This table shows regression results for driver-level and trip-level outcomes. Regressions are estimated at the city level and weighted by the number of drivers in each city-week. Standard errors are clustered by city. Rows 1-6 refer to aggregate earnings and labor supply measures for each driver-week. Rows 7-12 report hourly measures, excluding driver-weeks in which the driver did not work. Rows 13-15 report estimates for minutes worked, minutes worked conditional on working at least one minute, and an indicator for whether or not the driver worked at least one minute, respectively. Rows 16-20 report results for trip-level outcomes.

Table 49

	Var	Coefficient	SE	tstat	pval	dof	mean	% change
1	Unique Sessions	-0.013	0.026	-0.515	0.607	201	0.997	-1.319
2	Riders w/ Sessions	-0.002	0.006	-0.369	0.712	201	0.342	-0.605
3	Unique Sessions Cond. >0	0.034	0.031	1.099	0.273	201	2.834	1.199

Note: This table shows regression results for rider-level demand outcomes. Regressions are estimated at the city level and weighted by the number of riders in each city-week. Standard errors are clustered by city and stratified by the tercile of the number of riders in each city.

Table 50

	Var	Coef.	SE	tstat	pval	dof	mean	% change
1	Total Earnings	-1.940	6.113	-0.317	0.751	201	209.688	-0.925
2	Organic Earnings	-3.953	6.127	-0.645	0.520	201	192.500	-2.054
3	Minutes Worked	2.725	8.734	0.312	0.755	201	697.208	0.391
4	Incentive Earnings	-0.161	0.816	-0.198	0.844	201	6.006	-2.685
5	Surge Earnings	-2.283	2.855	-0.800	0.425	201	5.228	-43.665
6	Post-Commission Earn.	-0.486	4.717	-0.103	0.918	201	161.280	-0.301
7	Hourly Earnings	-0.228	0.496	-0.459	0.647	201	16.219	-1.404
8	Organic Earnings	-0.174	0.447	-0.389	0.697	201	15.031	-1.157
9	Utilization	-0.004	0.009	-0.423	0.672	201	0.489	-0.750
10	Incentive Hourly	-0.039	0.046	-0.839	0.403	201	0.462	-8.343
11	Surge Hourly	-0.173	0.190	-0.910	0.364	201	0.459	-37.715
12	Post-Commission Hourly	-0.086	0.391	-0.221	0.825	201	12.423	-0.695
13	Minutes Worked	2.725	8.734	0.312	0.755	201	697.208	0.391
14	Min. Worked - Intensive	-3.291	11.719	-0.281	0.779	201	1,054.186	-0.312
15	Did Work	0.004	0.003	1.368	0.173	201	0.662	0.623
16	Completed Trips	-0.129	0.227	-0.571	0.569	201	28.324	-0.457
17	Mean Fare	-0.316	0.248	-1.273	0.204	201	12.904	-2.452
18	Trip Distance	0	0.031	0.016	0.987	201	6.624	0.007
19	Trip Duration	5.533	7.730	0.716	0.475	201	972.181	0.569
20	Trip Surge	-0.007	0.010	-0.725	0.469	201	1.027	-0.721

Note: This table shows regression results for driver-level and trip-level outcomes. Regressions are estimated at the city level and weighted by the number of drivers in each city-week. Standard errors are clustered by city and stratified by the tercile of the number of drivers in each city.

Appendix C City Level t-test

Since labor supply seems to be balanced in the pre-treatment period at the city level without any weighting, we can try a t-test to check for differences in the experiment period. Results are in Table 51. The minimum detectable effect size is about 9%.

Table 51

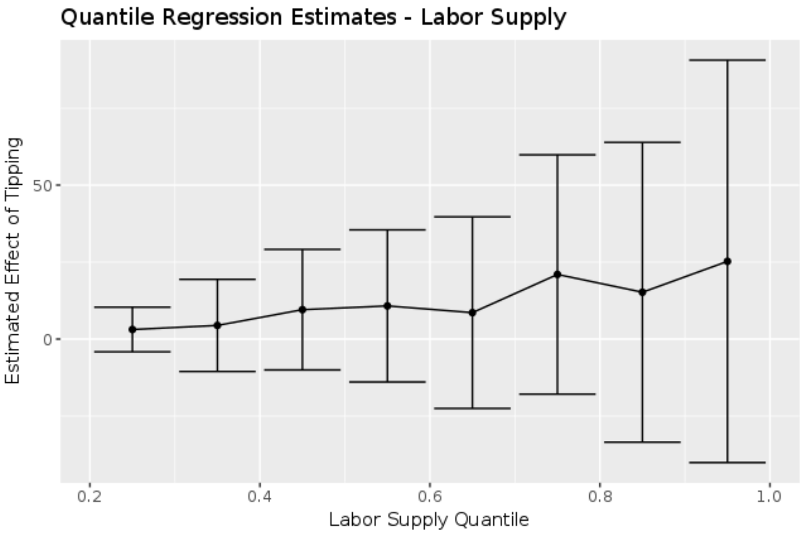
Mean Treated	Mean Control	T Stat	P Value	CI
542.70	520.61	0.88	0.38	(-27.33, 71.51)

Note: This table mean labor supply outcomes across cities in the control vs treatment group in the post-treatment period. Cities are not weighted by the number of drivers. The minimum detectable effect size for the t-test is about 9%.

Appendix D Quantile Regressions

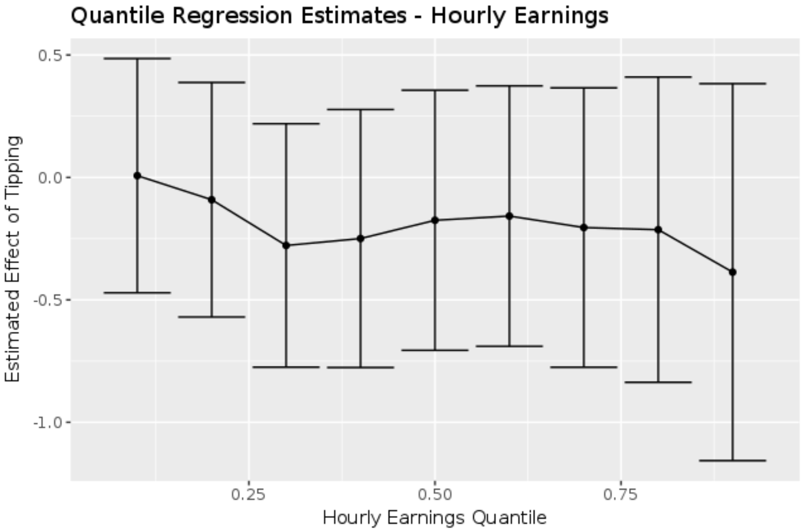
As a robustness check we report quantile regression estimates using the methods in Chetverikov et al. (2016). These methods can help estimate differential effects across drivers and may be more robust to outliers. In this application, estimating the quantile effects can be performed by simply computing the quantile of the outcome variable within each city-by-week cell and treating this computed quantile as the dependent variable in an OLS regression, using the same right-hand side specification from our city-level models. Estimates along with confidence intervals are shown in Figures 20 through 25. We do not plot quantiles that have little to no variation across city-weeks. For example, the tenth percentile of the labor supply is 0 across all city-weeks in the data and is not shown in Figure 20. As before, results are underpowered and should be interpreted with caution. The point estimates would suggest full-time drivers increased labor supply the most. The drivers with highest hourly gross earnings and utilization may have seen a larger drop in these metrics, though post-commission earnings are roughly flat, perhaps offset by tip income. Changes in demand do not seem to vary much by quantile.

Figure 20



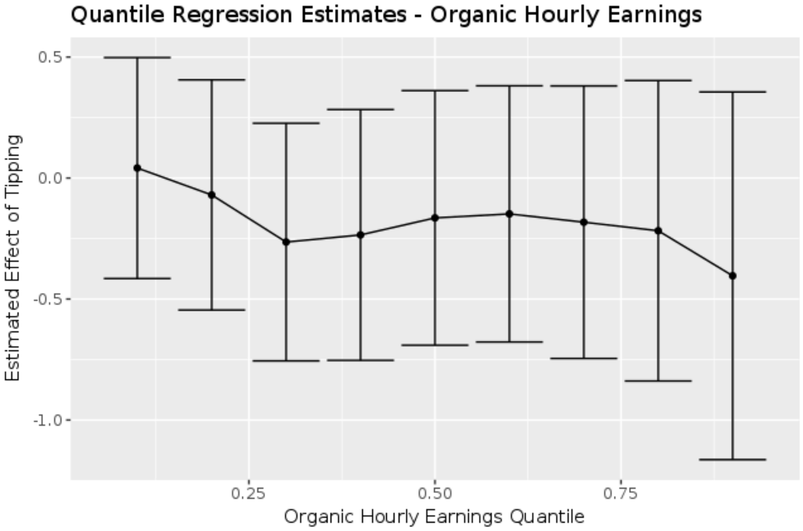
Note: This plot shows quantile regression estimates for labor supply. We exclude percentiles under 25% because below this threshold labor supply is zero for every city-week. Quantile regression estimates are computed as per Chetverikov et al. (2016).

Figure 21



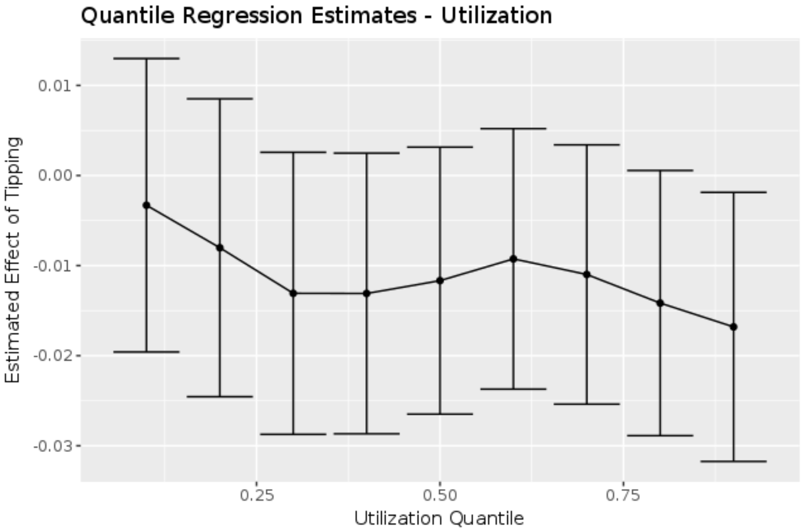
Note: This plot shows quantile regression estimates for hourly earnings. Quantiles for each city-week are estimated over only drivers who worked in that week.

Figure 22



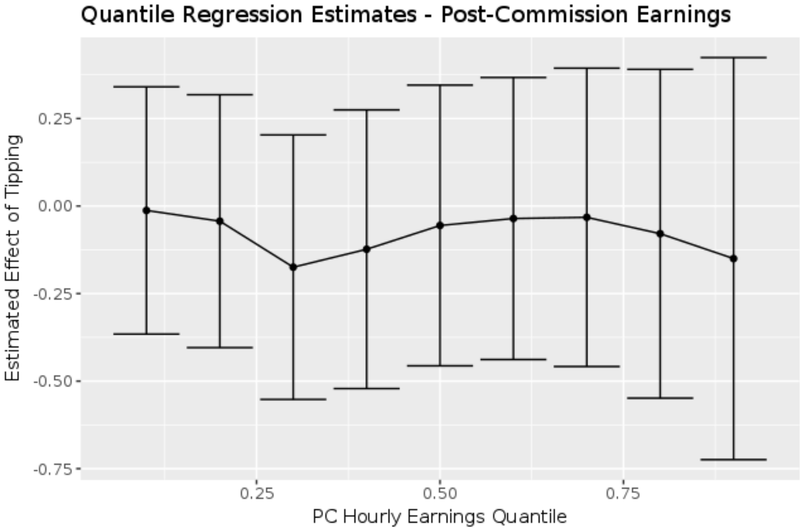
Note: This plot shows quantile regression estimates for organic hourly earnings. Quantiles for each city-week are estimated over only drivers who worked in that week.

Figure 23



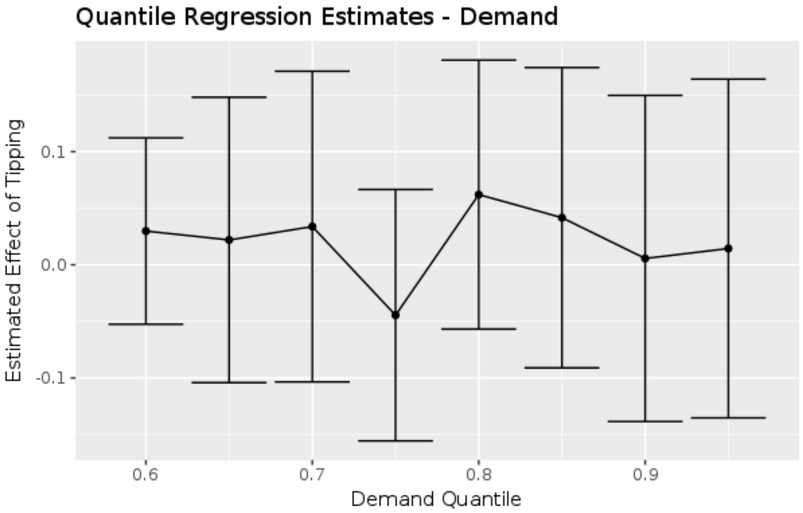
Note: This plot shows quantile regression estimates for utilization. Quantiles for each city-week are estimated over only drivers who worked in that week.

Figure 24



Note: This plot shows quantile regression estimates for post-commission hourly earnings. Quantiles for each city-week are estimated over only drivers who worked in that week.

Figure 25



Note: This plot shows quantile regression estimates for the number of unique sessions for each rider. A rider records a unique session when they open the app and search for a potential destination, whether or not they book and complete a trip. We exclude percentiles under 60% because below this threshold the number of unique sessions is zero for every city-week.