Causal Effect Estimation - Part 2

KP

2025-06-15

```r
data <- read.csv('data/Processed_Data.csv')
head(data)
```

```
##   Driver_ID       Date Driver_City Driver_ExperStartDate AppliedDate
## 1         1 2016-03-04          26            2016-03-30            
## 2         1 2016-03-05          26            2016-03-30            
## 3         1 2016-03-06          26            2016-03-30            
## 4         1 2016-03-07          26            2016-03-30            
## 5         1 2016-03-08          26            2016-03-30            
## 6         1 2016-03-09          26            2016-03-30            
##   EnrolledDate DaysSinceStart Earnings_Dollars DistanceDriven_Miles
## 1                           0               86                   69
## 2                           0              103                   54
## 3                           0              185                  124
## 4                           0               21                   18
## 5                           0               59                   50
## 6                           0                0                    0
##   TimeDriving_Minutes Drove Treated After Applied Enrolled EverApplied
## 1                 257     1       0     0       0        0           0
## 2                 212     1       0     0       0        0           0
## 3                 414     1       0     0       0        0           0
## 4                  69     1       0     0       0        0           0
## 5                 257     1       0     0       0        0           0
## 6                   0     0       0     0       0        0           0
##   EverEnrolled City_Cohort_FE    Date_FE      City_Cohort_Date_FE
## 1            0  26_2016-03-30 2016-03-04 26_2016-03-30_2016-03-04
## 2            0  26_2016-03-30 2016-03-05 26_2016-03-30_2016-03-05
## 3            0  26_2016-03-30 2016-03-06 26_2016-03-30_2016-03-06
## 4            0  26_2016-03-30 2016-03-07 26_2016-03-30_2016-03-07
## 5            0  26_2016-03-30 2016-03-08 26_2016-03-30_2016-03-08
## 6            0  26_2016-03-30 2016-03-09 26_2016-03-30_2016-03-09
```

IV 2SLS Estimation

TOT - Causal Effect of EverApplied on TimeDriving_Minutes

The following Two-Stage Least Squares (2SLS) specification is adopted for estimation on the causal effect of having ever applied throughout all time periods (`EverApplied`) on time spent driving per day (`TimeDriving_Minutes`).

First stage:

$$\text{EverApplied}_i \cdot \text{After}_{it} = \pi_1(\text{Treated}_i \cdot \text{After}_{it}) + \mu_{\text{city}_i,\text{expstart}_i,\text{date}_t} + \nu_{it}$$

Second Stage:

$$\text{TimeDriving\_Minutes}_{it} = \beta_1 \cdot (\text{EverApplied}_i \cdot \text{After}_{it})_{predicted} + \mu_{\text{city}_i,\text{expstart}_i,\text{date}_t} + \epsilon_{it}$$

Recall that the $\mu_{\text{city}_i,\text{expstart}_i,\text{date}_t}$ (Cohort x Time) FE is perfectly colinear with `After`. Hence, it is not included in the second stage. We have also omitted `EverApplied` in our second stage, as it is an endogenous variable and by adding it, it would violate the exclusion restriction that `Treated` x `After` only affect `TimeDriving_Minutes` through `EverApplied` x `After`.

```
library(fixest)

data$City_Cohort_FE <- as.factor(data$City_Cohort_Date_FE)

model <- feols(
  TimeDriving_Minutes ~ 1 | City_Cohort_Date_FE | EverApplied:After ~ Treated:After,
  data = data
)
summary(model, stage = 1)
```

```
## TSLS estimation - Dep. Var.: EverApplied:After
##                   Endo.    : EverApplied:After
##                   Instr.   : Treated:After
## First stage: Dep. Var.: EverApplied:After
## Observations: 11,002,548
## Fixed-effects: City_Cohort_Date_FE: 57,155
## Standard-errors: Clustered (City_Cohort_Date_FE)
##               Estimate Std. Error t value  Pr(>|t|)
## Treated:After 0.139438   0.001181 118.069 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.097327     Adj. R2: 0.128432
##                  Within R2: 0.074237
## F-test (1st stage): stat = 882,290.0, p < 2.2e-16, on 1 and 11,002,546 DoF.
```

The above shows the first stage of the 2SLS estimation. As shown, the P-value is incredibly small, indicating that `Treated` x `After` is a strong instrument for `EverApplied` x `After`.

```
summary(model)
```

```
## TSLS estimation - Dep. Var.: TimeDriving_Minutes
##                   Endo.    : EverApplied:After
##                   Instr.   : Treated:After
## Second stage: Dep. Var.: TimeDriving_Minutes
## Observations: 11,002,548
## Fixed-effects: City_Cohort_Date_FE: 57,155
## Standard-errors: Clustered (City_Cohort_Date_FE)
##                       Estimate Std. Error t value  Pr(>|t|)
## fit_EverApplied:After  24.0518    2.74864 8.75045 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 205.5     Adj. R2:  0.071757
##               Within R2: -1.419e-4
## F-test (1st stage), EverApplied:After: stat = 882,290.0, p < 2.2e-16, on 1 and 11,002,546 DoF.
##                           Wu-Hausman: stat =     124.6, p < 2.2e-16, on 1 and 10,945,391 DoF.
```

```r
level_effect <- coef(model)[["fit_EverApplied:After"]]

control_pre_mean <- mean(
  data$TimeDriving_Minutes[data$EverApplied == 0 & data$After == 0],
  na.rm = TRUE
)
percent_effect <- 100 * (level_effect / control_pre_mean)

sprintf(
  "TOT Estimate (IV): Taking up the program increases time driven by %.2f minutes per day on average (%
  level_effect, percent_effect, control_pre_mean
)
```

```
## [1] "TOT Estimate (IV): Taking up the program increases time driven by 24.05 minutes per day on avera
```

TOT - Causal Effect of EverApplied on TimeDriving_Minutes

```r
model <- feols(
  TimeDriving_Minutes ~ 1 | City_Cohort_Date_FE | EverEnrolled:After ~ Treated:After,
  data = data
)
summary(model, stage = 1)
```

```
## TSLS estimation - Dep. Var.: EverEnrolled:After
##                   Endo.    : EverEnrolled:After
##                   Instr.   : Treated:After
## First stage: Dep. Var.: EverEnrolled:After
## Observations: 11,002,548
## Fixed-effects: City_Cohort_Date_FE: 57,155
## Standard-errors: Clustered (City_Cohort_Date_FE)
##                Estimate Std. Error t value  Pr(>|t|)
## Treated:After 0.116002   0.001003 115.616 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.089958     Adj. R2: 0.105693
##                  Within R2: 0.061001
## F-test (1st stage): stat = 714,762.0, p < 2.2e-16, on 1 and 11,002,546 DoF.
```

Again, strong instrument.

```r
summary(model)
```

```
## TSLS estimation - Dep. Var.: TimeDriving_Minutes
##                   Endo.    : EverEnrolled:After
##                   Instr.   : Treated:After
## Second stage: Dep. Var.: TimeDriving_Minutes
## Observations: 11,002,548
## Fixed-effects: City_Cohort_Date_FE: 57,155
## Standard-errors: Clustered (City_Cohort_Date_FE)
##                        Estimate Std. Error t value  Pr(>|t|)
## fit_EverEnrolled:After   28.911    3.32248 8.70163 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 205.5     Adj. R2:  0.07173
##                Within R2: -1.705e-4
```

```
## F-test (1st stage), EverEnrolled:After: stat = 714,762.0, p < 2.2e-16, on 1 and 11,002,546 DoF.
##                           Wu-Hausman: stat =     121.3, p < 2.2e-16, on 1 and 10,945,391 DoF.
```

```
level_effect <- coef(model)[["fit_EverEnrolled:After"]]

control_pre_mean <- mean(
  data$TimeDriving_Minutes[data$EverEnrolled == 0 & data$After == 0],
  na.rm = TRUE
)
percent_effect <- 100 * (level_effect / control_pre_mean)

sprintf(
  "TOT Estimate (IV): Enrolling (applying and enrolling successfully) in the program increases time dri
  level_effect, percent_effect, control_pre_mean
)
```

```
## [1] "TOT Estimate (IV): Enrolling (applying and enrolling successfully) in the program increases time
```

Turning Uptake Into Impact: What It Means for Platform Strategy

Comparing the TOT estimate of a 17.9% increase in minutes driven per day to the ITT estimate of 2.09%, the IV regression identifies the causal effect of enrolling in the program — specifically for those drivers who enrolled because they were offered the program (compliers). This suggests the incentive program meaningfully increases driver activity on the platform among participants.

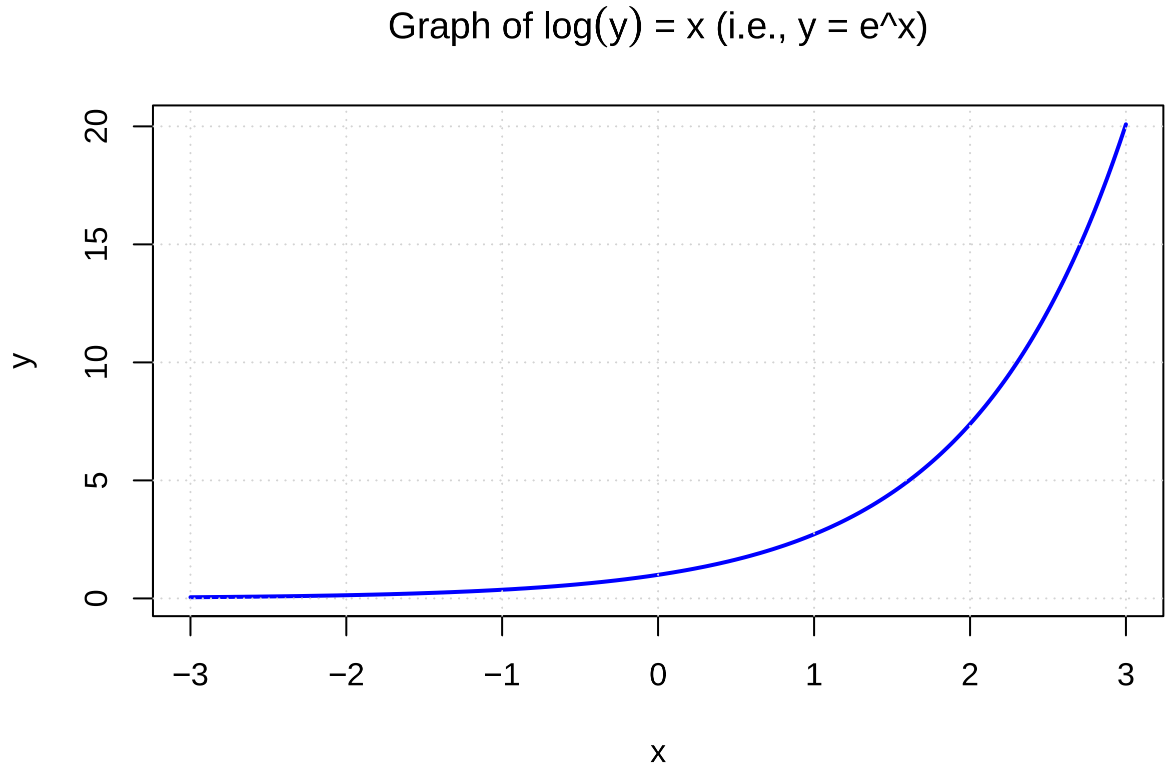
This finding supports the idea that **the program can be rolled out more broadly, especially in supply-constrained cities**, to improve labor availability. Since take-up is critical to achieving these gains, it's important for the firm to also invest in strategies that boost enrollment.

While not shown here for brevity of this analysis, a log-linear specification (e.g., $\log(1 + y) \sim x$) would allow for **direct interpretation in percentage terms**, and the addition of 1 prevents issues when the outcome is zero. As shown below, for $\log(y) \sim x$, when $\log(y) = 0$, x is undefined.

```
#(log(y) = x  →  y = exp(x))
x <- seq(-3, 3, length.out = 300)
y <- exp(x)

plot(x, y,
     type = "l",
     col = "blue",
     lwd = 2,
     xlab = "x",
     ylab = "y",
     main = expression(paste("Graph of ", log(y), " = x (i.e., ", y, " = e^x)"))
)

grid()
```



Combined with estimates of **labor elasticity** (how much 1% more labor supply affects KPIs like rides or profits), the platform can **estimate the value of one additional enrollee** and benchmark that against acquisition costs. This could support decisions based on marginal cost = marginal revenue logic for marketing spend per city.

Lastly, since this is based on early rollout, it's important to monitor effect sizes and take-up rates over time, as both may decline due to saturation or selection. Ongoing reevaluation will help maximize return on investment.

Thank you for taking the time to read this analysis. I'm especially grateful to Professor Keith Chen for his incredibly captivating teaching and the inspiration he provided, which motivated me to explore and build upon the homework assignment from his class, where this dataset was originally provided.