# Biostats 730: Project Proposal

Joshua Rosen

## Project Title: Bikeshare and Traffic Injuries: A Case Study in Bayesian Instrumental Variables

## Outcome of Interest

I have two primary outcomes of interest:

(1) A successfully designed and implemented model utilizing Bayesian Instrumental Variables

(2) Traffic injuries

## Data set

The merged/cleaned data containing bikeshare trips and traffic injuries by city/month can be accessed via my Github here. This data was collected from the Massachusetts Crash Data Portal and the Bluebikes Public Repository.

The full link to my public Github repository for all elements of this project moving forward can be found here.

## Reference Paper

The primary reference paper for this project is Lopes and Polson (2012), due to the thorough (but possibly outdated) literature review on Bayesian instrumental variables. I have also referenced McKeigue et al. (2010) for a slightly different approach. In terms of smaller tutorials, Gelman, Hill and Vehtari (2020), and two additional online posts have been referenced (Magnusson), (Savage).

## Description of the Project

I plan to examine the effect of the introduction of a bikeshare system on a city's total traffic injuries. Lightweight micromobility systems, such as dock-to-dock bikeshare and dockless scooters, are ubiquitous in a growing number of cities, but impacts have not been well studied. I therefore plan to utilize a Bayesian quasi-experimental approach to study the Boston metropolitan area's launch and initial expansion of its Bluebikes bikeshare system between 2011 and 2014.

Instrumental variables allow us to estimate the impact of an endogenous treatment $x_i$ on an outcome $y_i$ by 'instrumenting' through an exogenous variable $z_i$. The form is as follows:

(1)

$$x_i = \gamma + z_i\delta + \eta_i$$

$$y_i = \alpha + \hat{x_i}\beta + \epsilon_i$$

Where the two equations represent the first and second stages respectively. Here, $\alpha$ and $\gamma$ are intercepts, and $\delta$ and $\beta$ are regression coefficients. We also note that in the second stage, $x_i$ is replaced with $\hat{x_i}$ to indicate the fitted value of our endogenous treatment variable generated by estimating the first stage with instrument $z_i$. Critically, the errors $\eta_i$ and $\epsilon_i$ are correlated, and we are thus unable to only estimate the second stage regression.

In this paper, I also iterate on the classic 'difference-in-difference' technique in order to estimate the causal impact of changes in the number of bikeshare rides on traffic injuries. In frequentist language, a hypothetical difference-in-differences specification could take the form presented below,

(2)

$$y_{it} = \alpha_0 + \alpha_i T_i + \Sigma_{k \neq 0}[\delta_k D_k] + \gamma_i + \gamma_t + \epsilon_{it}$$

where $y_{it}$ denotes the number of traffic injuries that occurred in city $i$ of month $t$, $T_i$ denotes an indicator for being a treated city, $D_k$ denotes an indicator for time since treatment $= k$, $\gamma_i$ denotes city-level fixed effects, and $\gamma_t$ denotes calendar month dummies variables. In this four year span surveyed, Bluebikes slowly expanded into four cities: Boston, Brookline, Cambridge, and Somerville, while in the following six years the system would launch into six more cities. I therefore construct the treatment group using the four cities that experienced the initial expansion, and the control group from five of the six cities on the later expansionary phase.

However, a straightforward generalized DiD approach even in the Bayesian framework is unlikely to capture the true relationship between bikeshare's introduction and a possible reduction in traffic injuries. I therefore construct an instrumented difference-in-difference approach in order to estimate the causal impact of changes in the number of bikeshare rides on traffic injuries. The addition of the IV estimation also necessitates a further identifying assumption via the exclusion restriction. Here, I assume that the introduction of the bikeshare system only impacts the number of traffic injuries suffered through the number of rides taken. Ultimately, an instrumented difference-in-differences estimator allows us to replace any implicit assumptions over how the introduction of a bikeshare system would impact traffic safety, with a new explicit assumption. As a result, we now assume that the only effect a bikeshare system will have on traffic injuries is through the aggregate rides it manages to induce.

In the frequentist framework, we could represent a first stage regression as such:

(3)

$$TotalTrips_{it} = \alpha_0 + \alpha_1 T_i + \delta T_i * After_t + \gamma_i + \gamma_t + \epsilon_{it}$$

where $T_i * After_t$ represents an interaction term between indicator variables for if a city is a member of the treatment group, and if the city has received the treatment. I therefore allow this term to instrument for the number of total trips taken per city and month.

After instrumenting for the number of trips recorded by month and city, we could then establish the following second stage regression:

(4)

$$y_{it} = \alpha_0 + \alpha_1 T_i + \beta TotalTrips_{it} + \gamma_i + \gamma_t + \epsilon_{it}$$

where $y_{it}$ again represents the total number of traffic injuries recorded at the city-month level. The coefficient $\beta$ therefore reports the causal effect of variation in bikeshare trips on total traffic injuries.

Nevertheless, Bayesian inference is the obvious choice for a number of reasons. First, a hierarchical approach is necessary in order to estimate parameters in a setting where there are nested populations. In this example, data is represented at two levels: one level consisting of cities/months, and the other of individuals within the cities/months. Second, random effects are especially useful in situations where there is uneven sampling across levels. Due to the staggered nature of the roll out, time-since-treatment is not uniform across the treated cities (or months) and our available data varies by two group-levels. As a result, partial pooling may be a useful tool for estimating treatment effects.

In the Bayesian framework, I plan to estimate the following first stage regressions:

(5)

$$Trips_i = \phi + \nu_{0,cm[i]} + \gamma_0 T_i + (\gamma_1 + \nu_{1,m[i]})T_i + (\delta_0 + \nu_{2,c[i]})T_i \times After_i + (\delta_1 + \nu_{3,m[i]})T_i \times After_i$$
$$\nu_{0,cm[i]}|\sigma_{\nu_0} \quad i.i.d \quad N(0, \sigma^2_{\nu_0})$$
$$\nu_{1,m[i]}|\sigma_{\nu_2} \quad i.i.d \quad N(0, \sigma^2_{\nu_2})$$
$$\nu_{2,c[i]}|\sigma_{\nu_3} \quad i.i.d \quad N(0, \sigma^2_{\nu_3})$$
$$\nu_{3,c[i]}|\sigma_{\nu_1} \quad i.i.d \quad N(0, \sigma^2_{\nu_4})$$

where $T_i \times After_t$ again is used to represent an interaction term between indicator variables for if a city is a member of the treatment group, and if the city has received the treatment. Similarly, $\nu_{c[i]}$ and $\nu_{m[i]}$ allow the relationships between both $T_i$ and $T_i \times After_i$ to vary by city-level and month-level, respectively. Also note that $T_i$ is not modeled hierarchically with respect to city, since the treatment indicator does not vary by city.

The second stage is then as follows:

(6)

$$y_i = \alpha + \eta_{0,mc[i]} + (\beta_1 + \eta_{1,c[i]})Trips_i + (\beta_2 + \eta_{2,m[1]})Trips_i$$
$$\eta_{0,cm[i]}|\sigma_{\eta_0} \quad i.i.d \quad N(0, \sigma^2_{\eta_0})$$
$$\eta_{1,c[i]}|\sigma_{\eta_1} \quad i.i.d \quad N(0, \sigma^2_{\eta_1})$$
$$\eta_{2,m[i]}|\sigma_{\eta_2} \quad i.i.d \quad N(0, \sigma^2_{\eta_2})$$

The primary goal here is to capture both city-specific $\eta_{1,c[i]}$ and month-specific $\eta_{2,m[i]}$ deviations in the relationship between the instrumented total bikeshare trips and the outcome variable total injuries. Like the frequentist specifications, the first stage is used to instrument for total bikeshare rides via an interaction term between indicator variables for if a city is a member of the treatment group, and if the city has received the treatment. The instrumented variable is then used in the second stage specification in order to estimate how variation in the number of bikeshare rides impacts total traffic injuries.

The final step of the Bayesian specification is to define the prior. While I have not yet selected priors, I plan to rely heavily on the suggestions addressed by Lopes and Polson (2012). Further, I may adapt the likelihood functions in response to a deeper literature review.

# Action Plan

Moving forward, I will group my intended next steps into four categories: (1) review the literature on model formulation (2) begin constructing the model in Brms and (r)stan; (3) begin testing models and running sensitivity checks; (4) begin writing the formal paper. Step (1) therefore provisions time to formulate sensible priors and likelihood functions by reviewing previous literature, as well as time to adapt the project due to any Professor feedback. In order to optimize the remaining time to craft a strong project, prerequisite tasks have been completed prior to the proposal's submission (data cleaned, GitHub public repository started).