Notes on the Balanced NYC 2013 Taxi Data set

Walter W. Zhang & Øystein Daljord

April 2, 2020

1 Overview

The data sets are built from the comprehensive NYC taxi trip data and fare data for 2013 that are publicly available online. The taxi trip and fare data are balanced to one observation per driver, per hour, for all days of the year, and we have constructed some variables describing labor supply. The unbalanced data have observations on about 500,000 thousand drivers. We built two balanced data sets, one with a random selection of 5,000 drivers (500 MB) and one with a random selection of 100,000 drivers (8 GB). The code used to construct these data is publicly available at https://github.com/walterwzhang/NYC-Taxi-2013-Data. Kalouptsidi et al. (2019) uses a subset of these data to estimate the labor supply of the taxi drivers using a dynamic discrete choice model.

2 Data cleaning

The variables are described in Table 1.

The drivers are identified by the `medallion` and `hack_license` variables. The mappings can be found in Table 2. The mapping can be merged back in by `medallion_index` and `driver_index` to recover the original `medallion` and `hack_license`.

Before aggregating and processing the data, we drop observations that seem to be incorrectly coded or stored. We first drop all observations that have a trip time of 0 or a trip distance of 0. We drop all trips that either have a pick up or drop off outside of a geographic box with a northwest point of (latitude = 40.917577, longitude = −74.259090)

Table 1: Data Description

Variable	Description	Object
medallion_index	Medallion integer index	Integer
driver_index	Driver integer index	Integer
working	1 if working that hour, 0 otherwise	Binary
fare	Base fare earned that hour	Numeric
tip	Tip for that hour	Numeric
totalearnings	Total earnings in that hour (fare + tip)	Numeric
start_45m	1 if the driver started a shift that hour, 0 otherwise (45m)	Binary
start_6h	1 if the driver started a shift that hour, 0 otherwise (6h)	Binary
quit_45m	1 if the driver quit a shift that hour, 0 otherwise (45m)	Binary
quit_6h	1 if the driver quit a shift that hour, 0 otherwise (6h)	Binary
shift_45m	Number of cumulative shifts the driver is on (45m)	Integer
shift_6h	Number of cumulative shifts the driver is on (6h)	Integer
shift_type_45m	Shift type (45m)	Integer
shift_type_6h	Shift type (6h)	Integer
cumulative_driving_time_sec_45m	Cumulative driving time current *shift* (45m)	Numeric
cumulative_driving_time_sec_6h	Cumulative driving time for current *shift* (6h)	Numeric
dt	The datetime in hour increments NYT Time	Datetime object

Table 2: Medallion Mapping

Variable	Description	Object
medallion	The taxi's medallion	String
hack_license	The driver's hack license	String
medallion_index	Medallion integer index	Integer
driver_index	Driver integer index	Integer
company_owned	1 if company owned, 0 otherwise	Binary

and southeast point of (latitude = 40.477399, longitude = −73.700272). This is a large box covering the the tri-state area surrounding NYC.

The driving time is imputed from the driver's logged end time and the initial time. The reported time travel from the raw data is noisier and occasionally record trips that last over tens of days.

There are also two known oddities in the data. First, there are some negative fares in the data. The earnings from these fares are usually non-negative, which suggests that the earnings may be correct but either the tip or the fares may be miscoded. Second, there are also some trips that last a only few seconds but generate positive fares for the driver. Both have been retained in the data.

3 Balancing

The `working`, `fare`, `tip`, `totalearnings`, and `driving_time_sec` variables are constructed so they reflect the hourly wage of the driver. For trips that start in one hour and end in another hour, we distribute the fare between those two hours. For example, if one trip starts at 8:55 AM and ends at 9:05 AM, then half that fare contributes to the 8 AM hourly wage and the other half of that fare contributes to the 9 AM hourly wage.

For the variables `shift_type_45m` and `shift_type_6h`, the categories are:

- 0 if the driver has not started a shift yet
- 1 if the driver has started a "Day Shift" (4 AM to 9:59 AM)
- 2 if the driver has started a "Night Shift" (2 PM to 7:59 PM)
- 3 if the driver has started a "Bohemian Shift" (10 AM to 1:59 PM or 8 PM to 3:59 AM)

The variables `shift`, `shift_type`, and `cumulative_driving_time_sec` are cumulative counts. In other words, after a driver ends a shift, the driver's `shift`, `shift_type`, and `cumulative_driving_time_sec` columns will still reflect the last values despite the driver not working. These variables will be carried over to the next hour if the driver continues to not work. When the driver starts a new shift, the `shift` variable will increment by one, the `shift_type` variable will reset to reflect what type of shift it is, and the `cumulative_driving_time_sec` column will also reset and reflect the cumulative driving time for the current shift.

References

Kalouptsidi, M., P. Scott, and E. Souza-Rodrigues (2019). Linear iv regression estimators for structural dynamic discrete choice models. Discussion paper, Forthcoming Journal of Econometrics. 1