Payment Choice using Big Data: New York Taxis

Krzysztof Wozniak
Federal Reserve Board

August 2019, Helsinki

Overview

- Paper uses data on taxi trips in NYC to investigate payment choice
- Research questions:
  - Can merchants impact end users' payment choice without explicitly charging more for some payment methods?
  - If so, what are the likely drivers of merchants ability to do so?
- Research and policy implications:
  - Should theoretical models capture merchants' payment preferences?
  - What potential bias may come from empirical work not capturing merchants' ability to impact payment choice?
  - Do high interchange fees on card payments slow down shift away from cash?
- Paper examines consumer payment choices using a novel panel data set of taxi trips

Contributions

- New data source for studying payment choice: taxi trips
  - Enables estimation of heterogeneity in preferences for payment methods among merchants
- Frontier IT and statistical tools used to perform the analysis
  - Parallel processing on a high-performance cluster environment
  - Distributed file systems to reduce memory requirements
  - Implement a two-stage estimation procedure using modern iterative sparse least squares solver (LSMR).
- Findings contribute to payments literature
  - Payment choice is significantly impacted by merchants' preferences
  - Uncertainty could be a key driver of merchants' ability to steer customers' payment choice

Literature

- Taxi data
  - Farber (2014), Thakral and Tô (2017), and Hall et al. (2017) investigate labor market outcomes and responses using driver level data
  - Haggag and Paci (2014) look at the impact of suggested tip amounts, on the in-cab payment screen, on the realized tip amount
  - Buchholz (2018) and Fréchette et al. (2019) use dynamic equilibrium models to study matching frictions, regulations, and other features in the market
- Payment choice
  - Klee (2008) studies payment choice using scanner data from grocery stores
  - Wang Wollman (2016) test “threshold” theoretical framework of payment choice using retailer scanner data with 2 billion transactions
  - Cohen Rysman Wozniak (wp) study payment choice using home scanner data, focusing on heterogeneity between households and transaction amount endogeneity

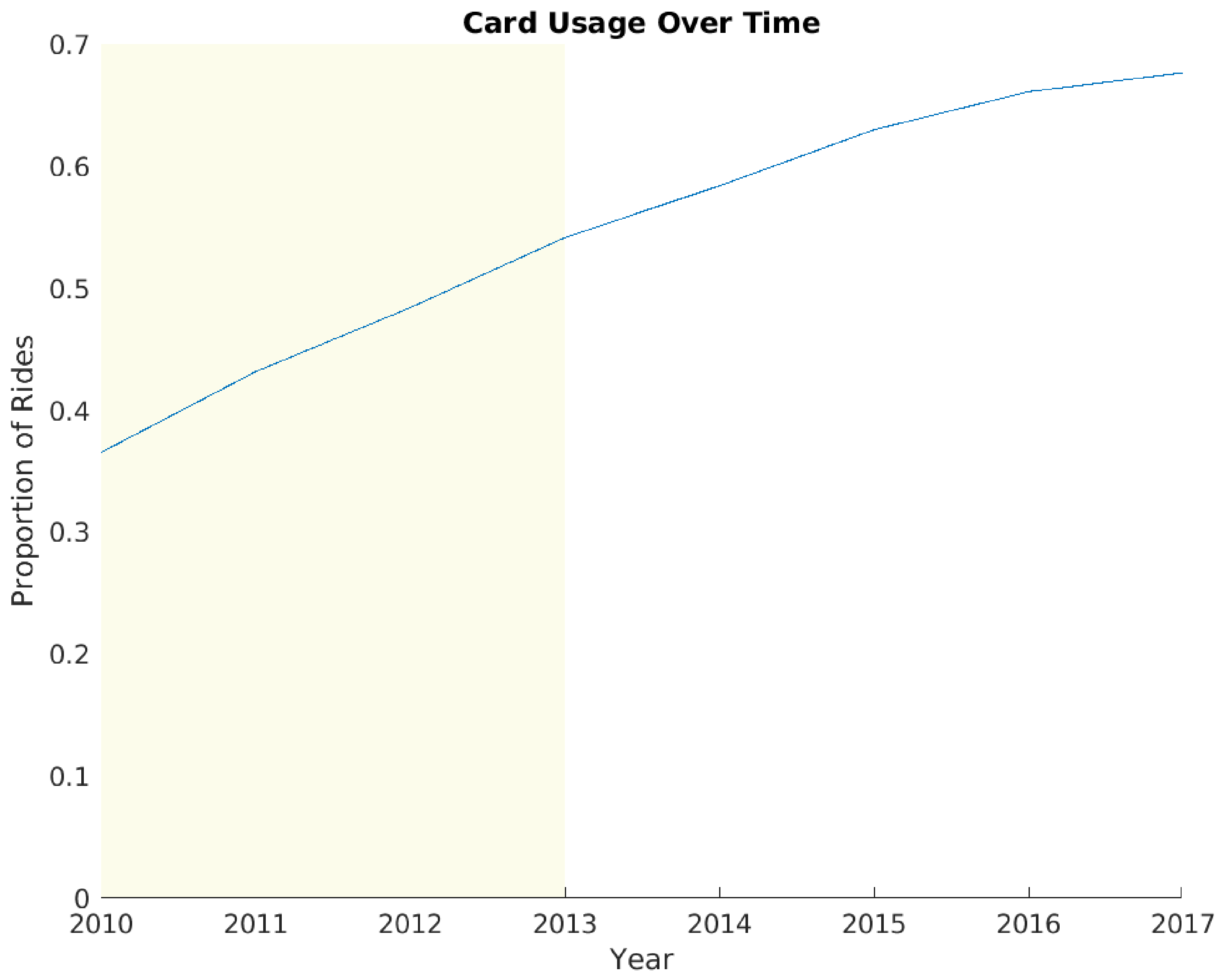
Institutional details

- Taxi and Limousine Commission (TLC) of New York City mandated:
  - 2004: electronic records of all taxi trips
  - 2009: choice between cash and card payments for all taxi trips, no surcharging
- Exogenous pricing for most trips
  - Fare determined through a combination of time and distance
  - Transparent pricing rules
- Negotiated pricing for trips to New Jersey

Data

- Source: TLC Trip Record Data
  - Four years of data (2010-13), over 700 million trips
  - Unbalanced panel, around 35,000 drivers
  - Key variables: payment choice (cash/card), driver ID, trip details (# of passengers, duration, distance, cost, date, pickup/drop-off time and location)
  - Data enhanced with local demographic information (from 2010 census)
- Advantages
  - Panel nature allows for the use of fixed effects to capture unobserved heterogeneity in drivers' ability to influence payment choice
  - 'Bigness' of data allows for very accurate estimation
- Limitations
  - Only two payment choices available to consumers
  - Unable to link driver IDs across years
  - Panel begins after card payments became mandatory

Summary statistics: payment trend over time



Summary statistics: differences between drivers

	Variable Name	Mean	Percentile 10th	25th	50th	75th	90th
	Number of trips	1,913	186	920	1,911	2,773	3,533
	Number of trips paid with card	945	70	397	857	1,409	1,907
Across drivers	Total amount ($)	15.57	12.27	13.13	14.49	16.37	19.07
	Trip distance (miles)	3.7	2.8	3.1	3.4	3.9	4.8
	Trip time (min)	15.2	12.7	13.5	14.5	15.9	17.9
	Number of passengers	1.7	1.0	1.0	1.3	1.6	3.2
Across trips	Total amount ($)	15.05	7.50	9.00	11.80	16.70	26.30
	Trip distance (miles)	3.3	1.2	1.6	2.3	4.0	7.3
	Trip Time (min)	14.6	6.1	8.6	12.1	18.0	25.3
	Number of passengers	1.6	1	1	1	2	3

Regression analysis: modelling approach

Linear probability model with fixed effects

$$Y_i = \mathbf{X}_i\beta + \alpha_{d(i)} + \boldsymbol{\gamma}_{t(i)} + \varepsilon_i$$

where

- *i* is the taxi trip
- Y_i is an indicator for whether the customer paid with *card*
- X_i is a set of observable characteristics for trip *i*, including the transaction value
- $\alpha_{d(i)}$ is the individual effect for driver *d(i)*
- $\boldsymbol{\gamma}_{t(i)}$ is a set of time controls
- $\varepsilon_i \sim N(0,1)$ is the error term

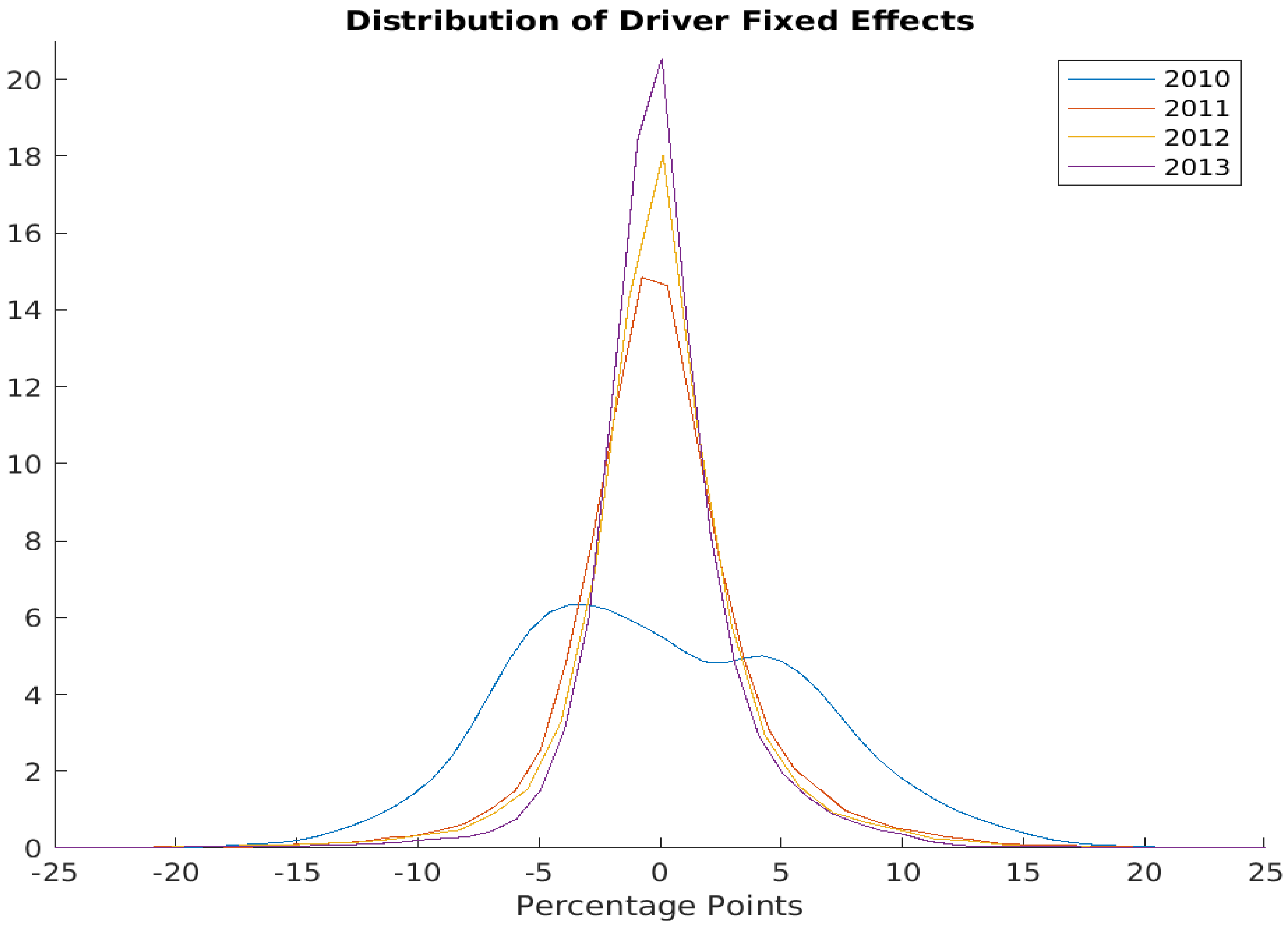
Identification and estimation challenges

- Identification challenges
  - For cash transactions, we do not directly observe tip amount, so there might be a slight endogenity concern
- Estimation challenges
  - With hundreds of millions of observations and forty thousand individual effects, OLS can take a week to estimate
  - We are interested in the value of the coefficients of the license fixed effects, so we cannot use some of the traditional panel regression “tricks”
  - Need to find a way to have accurate estimates for the fixed effects AND get p-values/standard errors for a subset of the controls

Variable	(1)	(2)	(3)	(4)	(5)
Total amount ($)	0.91 ***	0.84 ***	0.89 ***	1.94 ***	4.09 ***
Passenger count	-1.49 ***	-1.35 ***	-3.17 ***	-2.73 ***	-5.41 ***
Weekday am, 12-3		6.43 ***	6.20 ***	0.57 *	4.49 ***
Weekday am, 3-6		1.97 ***	1.95 ***	2.31 ***	1.13
Weekday am, 6-9		-4.41 ***	-1.75 ***	1.44 ***	7.37 ***
Weekday am, 9-12		0.44 *	0.14	0.91 ***	1.16 ***
Weekday pm, 3-6		-0.80 ***	-0.61 ***	-0.45 *	-1.14 **
Weekday pm, 6-9		2.71 ***	2.66 ***	1.77 ***	4.77 ***
Weekday pm, 9-12		-3.59 ***	-3.51 ***	-2.39 ***	-8.03 ***
% male		-3.04 ***	-3.44 ***		
Age (years)		-0.01 ***	-0.01 ***		
Income (1000s)		3.24 ***	3.72 ***		
Population (1000s)		0.54 ***	0.65 ***		
Driver FEs			X	X	X
Location FEs				X	
Zone-to-zone FEs					X
N (million)	271.1	271.1	271.1	270.5	268.1
Adjusted R^2	.035	.066	.111	.159	.192

Note: all values are in percentage points; weekend time, driver, location, and zone-to-zone fixed effects are not reported

Regression results



Conclusions (so far)

- *Card* usage has risen significantly over time
- Transaction value is a key driver of payment choice
- Commuters prefer to pay with *card*
- <u>Contribution</u>: drivers have a significant impact on payment choice
- <u>Contribution</u>: drivers' impact on payment choice falls as customers' uncertainty regarding payment choice falls

Potential next steps

- Investigate when drivers' have the biggest impact on payment choice
  - Tourists?
  - Probably least for regular commuters
- Try to link driver IDs across years
- Model payment choice when price is negotiable
  - Prices for trips between New York and New Jersey are determined through a process of bargaining between customer and taxi driver
  - Evidence of taxi drivers using price incentives to steer customers' payment choice?
- Implement discrete choice estimation procedure?
  - Linear probability model used could be the only procedure tractable enough for the size of the data
  - But, do we need to use all the data?

THANK YOU