Better Lucky Than Rich? Welfare Analysis of Automobile License Allocations in Beijing and Shanghai

Shanjun Li†

Forthcoming, Review of Economic Studies

Abstract

Economists often favor market-based mechanisms over non-market based mechanisms to allocate scarce public resources on grounds of economic efficiency and revenue generation. When the usage of the resources in question generates type-dependent negative externalities, the welfare comparison can become ambiguous. Both types of allocation mechanisms are being implemented in China’s major cities to distribute limited vehicle licenses as a measure to combat worsening traffic congestion and air pollution. While Beijing employs non-transferable lotteries, Shanghai uses an auction system. This paper empirically quantifies the welfare consequences of the two mechanisms by taking into account both allocation efficiency and automobile externalities post-allocation. Our analysis shows that different allocation mechanisms lead to dramatic differences in social welfare. Although Beijing’s lottery system has a large advantage in reducing automobile externalities over auction, the advantage is offset by the significant allocative cost from misallocation. The lottery system in Beijing resulted in a social welfare loss of 30 billion Yuan (nearly $5 billion) in 2012 alone. A uniform-price auction would have generated nearly 20 billion Yuan to Beijing municipal government, more than covering all its subsidies to the local public transit system.

Keywords: Allocation Efficiency, Automobile Externalities, Auction, Lottery

JEL classification: Q58, R48

*I thank George Akerlof, Panle Jia Barwick, Antonio Bento, Sebastien Houde, Sabrina Howell, Mark Jacobsen, Sarah Jacobson, Matthew Kahn, Neha Khanna, Benjamin Leard, Arik Levinson, Greg Lewis, Joshua Linn, Yimin Liu, Lucija Muehlenbachs, Mark Roberts, Stuart Rosenthal, Emily Wang, Catherine Wolfram, Jun Yang, Jérôme Adda (editor), three anonymous referees, and seminar participants at Binghamton University, Colorado School of Mines, Ford Motor, Georgetown University, Shanghai University of Finance and Economics, Texas A&M, University of Alabama, University of California-Berkeley, University of Calgary, University of Connecticut, University of Massachusetts, University of Sao Paulo, 2014 NBER Workshop on China’s Environment, 2014 NBER Summer Institute, and 2014 NBER Workshop on Chinese Economy for helpful suggestions. Financial support from the Lehman fund and the Institute for the Social Sciences at Cornell University is acknowledged.

†Dyson School of Applied Economics and Management, Cornell University and NBER, 405 Warren Hall, Ithaca, NY 14853, USA. Email: SL2448@cornell.edu. Phone: 1-(607)255-1832.
1 Introduction

Market-based mechanisms (e.g., auction) have often been advocated for allocating scarce public resources on grounds of economic efficiency and revenue generation, as opposed to non-market based mechanisms (e.g., lottery). Both have been used widely in practice, often for different types of resources. For example, auctions are routinely used to sell mineral rights, timber, and radio spectrum while lotteries are employed in distributing hunting licenses, charter school admissions, and jury duty.\footnote{1} Market-based mechanisms have the potential to achieve efficiency by using price signals to distribute the scarce resource to those with the highest willingness to pay (WTP), whereas lotteries with no-transferability can lead to misallocation and welfare loss. However, when the usage of the resources generates negative externalities that are increasing in WTP, the welfare comparison between the two mechanisms can become ambiguous. In this case, the social benefit (consumer welfare net of externalities), the basis for measuring social welfare, diverges from and may even be decreasing with WTP, the basis for resource allocation under the market-based mechanisms.

The resource in question here is a license or permit to register a vehicle. Major cities in China have been experiencing the world’s worst traffic congestion and air pollution as a result of rapid economic growth as well as vehicle ownership outpacing transport infrastructure and environmental regulation.\footnote{2} Beijing, the second largest city in China, has been routinely ranked as one of the cities with the worst traffic condition in the world, with the average traffic speed during peak hours below 15 miles per hour. The daily concentration of PM2.5, a key measure of air quality, frequently reaches over 10 times of the daily health limit recommended by the World Health Organization (WHO). To ease gridlock and improve air quality, Beijing municipal government introduced a quota system for vehicle licenses in 2011 in order to limit the growth of vehicle ownership. About 20,000 new licenses are distributed each month through non-transferable lotteries during 2011 and 2013 and the monthly quota was reduced to about 12,000 after 2013. Winning the lottery has become increasingly difficult: the odds decreased from 1:10 in January 2011 to 1:100 by the end of 2013 and further to less than 1/700 in August 2016.

\footnote{1}{The Federal Communications Commission used both administrative process and lotteries to allocate radio spectrum in the early years, both of which led to wide discontent due to long delays. The FCC finally adopted auctions in 1994 after many decades of persistent arguments by economists. The first auction yielded ten times of revenue predicted by the Congressional Budget Office (Cramton 2002).}

\footnote{2}{New automobile sales in China grew from 2.4 million units in 2001 to 22 million in 2013, surpassing the U.S. to become the largest market in 2009. Zheng and Kahn (2013) offer a comprehensive review on China’s urban pollution challenges and government policies to deal with them.}
Shanghai, the largest city in China, also has a vehicle license quota system but its license allocation is through an auction system instead of lotteries. In fact, an auction system has been in place since 1994 but the goal of reducing vehicle ownership did not emerge until much more recently. In 2008, Shanghai government adopted a form of multi-unit, discriminatory and dynamic auction held monthly online to distribute about 10,000 licenses. In 2012, the auction system generated over 6.7 billion Yuan (over $1 billion) to Shanghai municipal government.\(^3\) The average bid for a license reached over 92,000 Yuan in March 2013, higher than the price of many entry-level vehicle models.

The main objective of this study is to empirically quantify the welfare consequences of the two allocation mechanisms in distributing vehicle licenses taking into account both allocation efficiency and externalities associated with vehicle usage. This is an important question for at least two reasons. First, traffic congestion and air pollution impose major costs to the society (Parry et al. 2007). Creutzig and He (2008) estimate that the external costs from automobile usage in Beijing amount to over 7.5% of its GDP. Under the endorsement of the central government, many large cities in China are adopting the license quota system and employing different allocation mechanisms.\(^4\) However, the impacts and welfare consequences of these policies are unknown.

Second, there is theoretical ambiguity \textit{a priori} in welfare comparison between the two mechanisms because the usage of license (registering and ultimately driving a vehicle) is associated with negative externalities such as congestion and pollution. These type-dependent externalities are likely to be increasing in consumers’ WTP for a license in that consumers with high WTP tend to have high income. At the same time, high-income households on average buy less fuel-efficient vehicles and drive more than low-income households as household travel survey data from both China and U.S. show (see Table 1 and discussion in Section 2). Whether the lottery or the auction system leads to higher social welfare depends critically on the level of consumer heterogeneity in WTP and its relationship with automobile externalities. Therefore, the efficiency comparison and the magnitude of welfare impacts are ultimately empirical questions.

Since licenses are allocated through lotteries and trade is not allowed in Beijing, we do

\(^3\)The exchange rate between US dollar and Chinese Yuan was 1 to 6:23 at the end of 2012. It ranged from 6.06 to 6.96 during 2013-2016.

\(^4\)The National Plan on Air Pollution Control published by the State Council, the chief administrative authority in China, in September 2013 set goals and ten broad strategies for reducing PM2.5 for the first time in China. One of the strategies calls for controlling vehicle ownership in major cities by using the quota system. Cities of Guiyang and Guangzhou adopted license lotteries in July 2011 and August 2012. Tianjin, Hangzhou, and Shenzhen started to implement a hybrid system in January, March, and December 2014.
not observe consumers’ WTP. In Shanghai, licenses are auctioned in a non-standard format where the bids may not reflect value and the equilibrium bidding function is difficult to characterize as shown in an experimental study by Liao and Holt (2013). To recover consumers’ WTP for a license, we recognize that consumer surplus from owning a vehicle should inform WTP for a license, the additional cost that consumers have to bear in order to own a vehicle. To estimate consumer surplus from vehicle ownership, we estimate a random coefficients discrete choice model of vehicle demand that takes into account consumer preference heterogeneity and unobserved product attributes developed by Berry, Levinsohn, and Pakes (1995) (henceforth BLP). Similar to Petrin (2002), our estimation strategy employs both aggregate market-level data and information from household surveys to form moment conditions. From household surveys on new vehicle buyers, we obtain the share of buyers among different income groups. We use them to form (micro-) moment conditions and they are critical in identifying consumer preference heterogeneity.

An important departure in our identification strategy from the literature is that we do not rely on the maintained exogeneity assumption that unobserved product attributes are uncorrelated with observed product attributes. Taking advantage of the multi-market nature of the data, we include vehicle model fixed effects to control for unobserved product attributes. In addition, we explicitly employ the common-trend assumption implicit in the difference-in-differences (DID) framework to better isolate the effect of the policy shock from unobserved demand shocks: the specification without the common trend assumption tends to attribute the large sales drop in 2011 as coming from a large (unobserved) negative demand shock. Our market-level data include data for four cities: Beijing, Nanjing, Shanghai, and Tianjin. Nanjing and Tianjin are two large cities next to Shanghai and Beijing, respectively. They did not have the license quota system during the data period and as we show through graphs and regressions, the automobile markets in these two cities exhibit similar trends to Beijing and Shanghai in the absence of the policies.

Several important findings emerge from our analysis. First, both policies in Beijing and Shanghai significantly limited new vehicle sales: the lottery system in Beijing reduced new vehicle sales by over 1.2 million in 2011 and 2012 while the Shanghai auction reduced sales by nearly two million from 2008 to 2012. These reductions are substantial and a reflection of the stringency of the quota system. Second, the welfare consequences of different allocation mechanisms are huge in magnitude. While Beijing’s lottery system has a non-trivial advantage of three billion Yuan over an auction system in reducing automobile externalities, its allocative cost due to misallocation is over 33 billion Yuan in 2012, implying a welfare loss of nearly 30 billion in Beijing in 2012 alone. The significant allocative cost from the
lottery system is driven by the large consumer heterogeneity in WTP for licenses. Third, a uniform price auction would have generated nearly 20 billion revenue to Beijing municipal government in 2012, more than enough to cover all its subsidies to the local public transit system. Fourth, based on a range of plausible assumptions, the optimal level of quota in Beijing was lower than the quota used in 2012 and further reducing the quota would increase net social welfare.

This study contributes to the literature in the following aspects. First, although theoretical literature on allocation mechanisms are abundant, there are very few empirical studies on quantifying welfare outcomes of different mechanisms. Glaeser and Luttmer (2003) study housing market rationing under the rent control in New York city and provide evidence of significant misallocation of houses, without explicitly estimating the allocative cost. Davis and Kilian (2011) find significant allocative costs from misallocation under the price ceiling in the U.S. residential market for natural gas. In the context of a quantity regulation, this paper empirically examines allocative outcomes by exploring a rare opportunity where both market-based and non-market mechanisms are used for the same type of resources.\(^5\)

Second, as discussed further above, in the presence of type-dependent externalities that is increasing in WTP, the efficiency comparison between the lottery and auction systems could be ambiguous due to the fact that externalities introduce a wedge between net social benefit, the basis for measuring social welfare, and the private benefit, the basis of resource allocation under auction. The implication of type-dependent externalities on optimal allocation mechanisms has been examined in several theoretical studies in the context of firm competition.\(^6\) Our analysis empirically highlights type-dependent externalities in the context of consumer goods and showcases the advantage of the lottery system in reducing externalities.

Third, our analysis adds to the emerging literature on China’s environmental and energy policies. China is by far the largest emitter of greenhouse gases, accounting for nearly 30 percent of world emissions in 2013. It is the largest energy consumer and is also the top importer of crude oil in the world. China’s domestic policies could have global impacts but

\(^5\)An emerging literature on the efficient targeting of public programs in the development context compares allocation outcomes of different mechanisms and finds that properly designed auctions can reduce program costs by better targeting on private information relative to lottery (see for example Jack (2013)).

\(^6\)Brocas (2013) analyzes optimal auction design with type-dependent negative externalities in the context of multiple firms competing for the acquisition of a technology license or a procurement contract. The winner induces a negative externality on the loser since the winner will have a competitive advantage over the loser in downstream competition. Similarly, the usage of licenses by firms may have implications on market power. This has been discussed in the context of airport slot allocations and spectrum auctions where different collection of winners may subsequently lead to different market structure (Borenstein 1988; Cramton 2002; Hazlett and Munoz 2009).
our understanding on the impacts of these policies is very limited.  

Before we proceed, it is helpful to further clarify the scope of our welfare analysis. First, our analysis focuses on the welfare consequences of different allocation mechanisms within the policy framework of a quota system. Automobile usage generates multiple types of externalities including congestion, air pollution, traffic accidents and noise. A vehicle license quota system in theory is not the first-best instrument to internalize these externalities since it does not directly address the source of externalities such as driving. Second, our analysis focuses on welfare impacts in the short run. The allocation of vehicle licenses are likely to have impacts on household location, job decision and schooling choices, all of which could have important welfare implications in the long term. Although part of these welfare impacts such as the desire to live in a large apartment outside of the city center and hence the need to have a vehicle are captured by the WTP that we estimate, we are not investigating the policy impacts in these dimensions. The broader welfare comparisons between the quota system with other potentially more efficient policies such as congestion pricing and fuel taxes as well as long-run analysis will necessitate additional information and modelling efforts.

2 Allocation Mechanisms and Externalities

The purpose of this section is to offer an illustration on the welfare comparison between lottery and auction systems in the presence of type-dependent negative externalities. In allocating scarce public resources, governments or resource managers have relied on both market-based mechanisms and non-market based mechanisms often for different types of goods. Market-based mechanisms such as a well-designed auction can allocate the resource to those with the highest value and hence achieve efficiency while non-market based mechanisms such as an administration process or lottery do not. It is argued that non-market based mechanisms such as a lottery are chosen often out of concern of fairness or for political convenience (Taylor et al. 2003).

Various federal and local policies are being adopted in China to address urban congestion and air pollution including fuel economy standards, emission standards, gasoline tax, and driving restrictions. Chen et al. (2013) study the environmental impacts of measures such as driving restriction and plant closure by the Beijing government to clear up the air before the 2008 Olympic Games. Xiao and Heng (2013) evaluate and compare the effects of favorable consumption tax treatment for small vehicles and the gasoline tax.

In the first-best world, multiple instruments should be used to correct for these externalities. For example, real-time congestion pricing is the first-best instrument to deal with congestion externality while fuel taxes is the first-best instrument to internalize the externality from CO\textsubscript{2} emissions and other pollutants that are proportional to fuel use. Nevertheless, these policy instruments would require different administrative costs. The quota system has become a popular policy option among local policy makers perhaps for its lower administrative costs or due to inertia from traditional command-and-control regulatory approach.
However, when the usage of resources generates market failures such as externalities, market-based mechanisms such as an auction may not yield efficient allocation because the private value and the social benefit diverge. We illustrate this point within the context of allocating vehicle licenses using lotteries and auctions. Consider the following environment: (1) there are $Q$ licenses to be allocated; (2) there are $N$ agents, each demanding at most one license; (3) each agent $i$ has a private value (or WTP) $V_i$ and the value is independent and identically distributed with a support of $[0, \bar{V}]$; (4) the usage of the license imposes an external cost of $E_i$, which is increasing in $V_i$. The fourth assumption of type-dependent externalities is a key departure from a standard model of resource allocation. As we document below, consumers with high WTP for a license tend to have higher income. On average, they drive larger and less fuel-efficient vehicles and they drive more relative to those with a low WTP. So the usage of licenses by those with a higher WTP is likely to generate larger external costs. Our estimates below show that neither the magnitude of external costs nor the difference in external costs across households is trivial. That is, the discussion below does have empirical relevance in our context.

We compare two allocation mechanisms: a non-transferable lottery where all agents can participate and have an equal chance of winning (random allocation); and a uniform price auction where the $Q$ highest bidders each gets one license and pays a price equal to the highest rejected bid. Harris and Raviv (1981) show that each agent bids her value in this auction and therefore the licenses will be allocated to the $Q$ agents with the highest value. We use the uniform price auction for exposition and the point is not lost with other types of auctions such as discriminatory auctions that can achieve the same allocation outcome.

We start with a simple case of allocation mechanisms with constant external costs from the usage of the license and ultimately automobiles as depicted in the top panel of Figure 1. To ease exposition, we focus on a linear WTP schedule. Line $Q(p)$ is the (smoothed) WTP schedule. Line EC is the external cost for the agents with the corresponding WTP and it is constant in this case. Assuming the quota $Q$ is lower than the optimal cap $Q^*$, the area BCD is the conventional deadweight loss (DWL) from quantity constraint. Total consumer surplus from the auction system is given by the area $ABQ$ but the lottery system can only realize $Q/Q_1$ of consumer surplus from the random allocation, which is given by $AQO$. Therefore, the allocative cost, i.e., welfare loss from misallocation, of the lottery system is $ABQ$, which could even be larger than the DWL from quantity constraint. The empirical importance of allocative cost from random allocation is highlighted in Davis and Kilian (2011) in the context of price ceiling in the U.S. residential natural gas market.
The middle panel in Figure 1 shows a case where the external costs are positively correlated with WTP. As in the top panel, BCD is the DWL from quantity constraint and ABQ is the allocative cost from the lottery system. However, the presence of type-dependent external costs implies that the lottery system leads to a reduction in external costs depicted by EDF, offsetting part of the allocative cost. The net welfare loss from lottery (versus auction) is given by the area $ABQ - EDF$, which is positive in this case. The advantage in reducing externalities from the lottery system is dominated by the welfare loss from misallocation, implying that the auction system still produces a better welfare outcome.

The welfare comparison between the two systems is reversed in the bottom panel where the external costs are “strongly” increasing in WTP. To ease exposition, we assume that the quota $Q$ is at the intersection of the WTP schedule and the corresponding external cost curve. In this case, the reduction in external costs from the lottery given by $EBQ_3$ dominates the allocative cost given by $ABQ$. Since agents with the highest private value generate the smallest social value, the auction will allocate the licenses to the wrong hands from the efficiency perspective. This result is driven by the fact that the wedge introduced by the external costs between the private value and the social value leads to a negative correlation between the two. Note that the external cost line does not have to surpass the WTP schedule as shown in the graph to make the point. The result could still hold if the EC line is below the WTP schedule but is steeper.

Which of the two scenarios is playing out in reality depends on the WTP schedule as well as its relationship with the external cost curve. Although one can get a good sense on the external costs from driving from different households based on household survey data, the WTP schedule is not readily available. The empirical goal is to develop a method to estimate the WTP schedule and conduct welfare comparison taking into account the external costs. The latent nature of WTP prevents me from showing the relationship between WTP and externalities from automobile usage directly off of the data. Appendix Table 2 based on household surveys for new vehicle buyers (to be discussed in the data section) clearly shows that high-income households account for a disproportionately large share of new vehicle buyers, suggesting a positive correlation between income and WTP for a new vehicle. Because of these two reasons (not observing WTP directly and the positive correlation between WTP and income), we use income as a proxy for WTP in showing the potential relationship between WTP and externalities to motivate our analysis.

Table 1 shows the average engine size and annual vehicle miles travelled (VMT) by seven income groups from 2010 Beijing Household Travel Survey. For vehicle owners, the
<table>
<thead>
<tr>
<th>Annual Income (Yuan)</th>
<th>Engine Size(liter)</th>
<th>Annual VMT(km)</th>
<th>Number of Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 50k</td>
<td>1.58</td>
<td>13907</td>
<td>4827</td>
</tr>
<tr>
<td>50-100k</td>
<td>1.68</td>
<td>14922</td>
<td>5379</td>
</tr>
<tr>
<td>100-150k</td>
<td>1.78</td>
<td>16554</td>
<td>1327</td>
</tr>
<tr>
<td>150-200k</td>
<td>1.83</td>
<td>17882</td>
<td>417</td>
</tr>
<tr>
<td>200-250k</td>
<td>1.87</td>
<td>17661</td>
<td>133</td>
</tr>
<tr>
<td>250-300k</td>
<td>1.94</td>
<td>18923</td>
<td>70</td>
</tr>
<tr>
<td>≥ 300k</td>
<td>2.18</td>
<td>20382</td>
<td>98</td>
</tr>
<tr>
<td><strong>Full Sample</strong></td>
<td>1.67</td>
<td>14896</td>
<td>12251</td>
</tr>
</tbody>
</table>

Notes: The data source is 2010 Beijing Household Travel Survey. There are 12,251 vehicles surveyed from 10,577 households. 1,440 households own two vehicles while 93 households own three vehicles.

The table shows that high-income households own vehicles with larger engines and drive more on average. These positive correlations are consistent with evidence of vehicle ownership and travel behavior in the U.S. (Small and Van Dender 2007; Li et al. 2009, 2014). Since the major part of externalities from automobile usage in urban areas comes from congestion and air pollution and these externalities are directly linked to travel distance and gasoline usage (Parry et al. 2007). This implies that households with a higher income are likely to produce larger externalities from automobile usage than households with a lower income. The welfare analysis will return to this point and provide estimates of externalities for different households with different WTP.

### 3 Policy and Data Description

In this section, we first describe the background for the quota systems. We then discuss the lottery and auction policies in Beijing and Shanghai and present our data thereafter.

---

9The survey is a random sample of Beijing households with a sample size (vehicle owner and non-owner) of 46900 households. The survey however does not have information on vehicle model, vintage, and fuel economy information, preventing us from using the data directly for estimating vehicle demand. Vehicles with a smaller engine tend to have better fuel economy. The correlation coefficient between engine size and fuel economy (liter per 100 km) is 0.77 from our vehicle attribute data discussed below.

10The National Household Travel Survey in the U.S. has information on household income (18 categories), vehicle fuel economy (MPG) and VMT. Based on 75,148 households from three waves (2001, 2005 and 2008) used in Li, Linn and Muehlegger (2014), the correlation coefficient between household income (in 18 categories) and fuel economy is -0.09 and that between household income and VMT is 0.25.
3.1 Background

During the past three decades, China has embarked on an extraordinary journey of economic growth with its GDP growing at about 10 percent a year. As household income grows, luxury good consumption such as automobiles started to pick up dramatically at the turn of the century. Annual sales of new passenger vehicles increased from 2.4 million units in 2001 to nearly 22 million in 2013 as shown in Appendix Figure 1. China surpassed the U.S. to become the largest auto market in 2009 and in 2013, it accounted for 26 percent of world’s total new auto sales. Large cities in China are ahead of the curve in both economic growth and vehicle ownership. Beijing with about seven million households has gone from a city on bikes to a city on cars during this period: the stock of passenger vehicles increased from 1.6 million units to nearly 5 million units as shown in Appendix Figure 1.\(^\text{11}\)

The rapid growth in vehicle ownership leads to serious traffic congestion, despite significant efforts in expanding roads and public transit systems and a whole range of other traffic management policies such as driving restrictions and reducing public transit fares. The city is often ranked as one of the worst cities in traffic conditions. The traffic speed on arterial roads within the 5th ring road during morning peak hours (7:00-9:00) on work days averaged 14.7 miles/hour (MPH) in 2011, reduced from 22.8 MPH in 2005. The average speed was 13.4 MPH during afternoon peak hours (17:00-19:00) in 2011, compared with 20.2 MPH in 2005. During the same period, air quality has worsened dramatically, and the air quality index is frequently above the level above which the U.S. EPA recommends that everyone should avoid all outdoor physical activities.\(^\text{12}\)

Shanghai, with 8.5 million households, has slightly better traffic conditions, on par with those in Los Angeles. The traffic speed on arterial roads averages 21.2 MPH and 22.3 MPH during morning and evening peak. As shown in the bottom panel of Appendix Figure 1, the total number of vehicles in Shanghai is less than half of that in Beijing, despite having more households and higher household income. This is largely due to vehicle purchase restrictions put in place before vehicle ownership took off as we discuss below. The air quality in Beijing is consistently worse than Shanghai due to more vehicles, worse traffic, and winter heating coupled with unfavorable topography. In a ranking of air quality among 28 major cities on

\(^{11}\)Among the 5 million vehicles, about 4 million are owned by households. The household vehicle ownership rate is 0.58 in Beijing, comparing to 0.46 in New York city and 1.16 in the U.S. from 2010 U.S. Census.

\(^{12}\)The average daily concentration of PM2.5 frequently reaches over 250 micrograms per cube meter, compared to the recommended daily level of 20 by WHO. According to Beijing Environmental Protection Bureau, automobiles are the largest source of PM2.5, accounting for 22 percent in the whole city and about one third in the urban core. The second largest source of PM2.5 is coal burning (17 percent) followed by construction site dusts (16 percent) in 2012.
the east coast, Beijing ranked at the bottom with an annual average PM10 concentration of 121 micrograms per cubic meter while Shanghai ranked at 8th with an average PM10 concentration of 79 in 2010, still significantly higher than the WHO’s health limit.

3.2 Policy Description

To address traffic congestion and air pollution, Beijing municipal government announced the policy of capping new vehicle licenses which are necessary to register a vehicle, on December 23, 2010. Since January 2011, lotteries have been used to allocate about 20,000 licenses each month during 2011-2013 and the annual quota was reduced to 150,000 after 2013. A license is needed for first-time buyers, and those who purchase an old vehicle, accept a gifted vehicle, or transfer out-of-state registration to Beijing. Vehicle owners who scrap the used vehicle can transfer the old license to the new vehicle and do not need a new license. The eligible participants include Beijing residents and non-residents who have been paying income tax for at least five years in Beijing. The licenses are assigned to winners through random drawings. The winners can then use the license to register their vehicles.

Transferring a license from a winner to other people is prohibited. Although there is anecdotal evidence that some transferring occurred by having vehicle registered under the winner but paid and used by another person, this is not known to be widespread because the legal owner (the winner) not only has the liabilities in paying annual registration fee, traffic fines and emission inspections, but also is liable for damages and injuries in accidents. In addition, barriers are in place to prevent Beijing residents from registering vehicles in neighboring provinces. First, a temporary driving permit is needed to be able to drive an out-of-state vehicle in Beijing. More importantly, these vehicles are banned from entering the 5th ring road (within which the vast majority of business and population are located) during rush hours. So this avoidance behavior is also not likely to be widespread.

Among all the licenses allocated, about 88% (or 17,600 each month) are for private vehicles and the rest are for institutions. The winners are determined in two different pools for these two categories. While the private licenses are allocated monthly, the institutional licenses are done bi-monthly. The first lottery was held on January 26th, 2011 and 17,600 private licenses were allocated among 187,420 participants. The winning odds reduced to 1:100 by the end of 2013 and further to 1/725 in August 2016, due to the cumulation of pent-up demand over time as well as future buyers entering into the lottery pool.\textsuperscript{13} The

\textsuperscript{13}With the shrinking odds of winning, there is increasing discontent with the system. According to a survey of 800 residents in 2013 by Beijing Statistical Bureau, nearly 70 percent of the participants agree that the system needs to be improved. Among them, 42.2 percent would like to abolish the system and 41.5
top panel of Appendix Figure 2 shows monthly licenses allocated and new vehicle sales. The dramatic decrease in vehicle sales since the start of the policy reflects the stringency of the policy relative to the demand for new vehicles. The difference between vehicle sales and the number of licenses allocated is due to the sales that do not need a license (vehicle replacement after scrappage). The winners have six months to register a new vehicle before the winning certificates become expired. Once expired, the license quota recycles back for distribution in future lotteries. The winners who allow their licenses to expire will not be permitted to participate in the lottery within the next three years.

Shanghai is the first city to implement a vehicle license quota system in China and it auctioned its first license in 1986. Although the market for private vehicles was very small at that time, traffic congestion was a big problem due to insufficient road infrastructure. In fact, Shanghai experimented with vehicle driving restrictions even before 1995. The auction system has evolved over time. Initially, it was a sealed-bid auction where reservation prices and quota levels varied across vehicles produced in Shanghai, non-Shanghai produced vehicles, and imports. In 2003, a unified auction system without a reservation price was put in place for domestic vehicles and imports.

The online auction format during 2008 to 2012 can be characterized as a multi-unit, discriminatory (pay as you bid), and dynamic auction. The auction is held monthly during a 90-minute period and bidders observe the current lowest accepted bid prior to submitting a bid. In the first hour, bidders can submit a single initial bid and in the last 30 minutes, each bidder can revise their bids up to two times. The revised bid however has to be within a window of 300 Yuan below and above the current lowest accepted bid. The purpose of the bid revision period and the restriction on bid revisions is to reduce price volatility. The bottom panel in Appendix Figure 2 shows the average and lowest accepted bid in each month from 2008 to 2012. The average bid price increased from 23,370 to 69,346 Yuan during this period. The plot shows that the average and the lowest winning bids are very

\[14\] In 1998, Shanghai government set the reservation price of 20,000 Yuan for vehicles produced in Shanghai while the reservation price was set at 100,000 Yuan for vehicles produced elsewhere. In protest, Hubei province which has a large automobile industry initiated an additional charge of 70,000 Yuan for those who purchase brands produced in Shanghai. The trade war ended in 2000 when Shanghai removed the reservation price in auctions for domestically produced vehicles.

\[15\] Liao and Holt (2013) use experiments to study the relationship between bid and WTP under the Shanghai auction and compare it with other formats such as the first price auction. They show that initial bids in the first stage of the auction tend to be much lower than those in the first price auction which themselves are lower than WTP. Bids are revised up in the second stage but they are still much lower than WTP and the difference is larger among high WTP bidders. In addition, the auction design suppresses heterogeneity in bids, which will be a key factor in our welfare comparison.
close: the difference is usually less than 500 Yuan or one to two percent of the average bid.\textsuperscript{16} The winners are required to purchase a new vehicle within three months before the license expires. The vehicle and the license cannot be transferred within the first year of registration. Similar to Beijing, vehicles registered outside of Shanghai are not allowed to use the major roads during rush hours. Although there is anecdotal evidence that some households choose to register their vehicles in neighboring provinces due to high license prices, this phenomenon is not believed to be widespread.

3.3 Data Description

Our analysis focuses on policies in Beijing and Shanghai and we bring two nearby cities, Nanjing and Tianjin into analysis to facilitate identification. Nanjing is about 300km away from Shanghai and is the capital city of Jiangsu province that shares boarder with Shanghai. Tianjin is about 150km from Beijing. The characteristics of these four cities are shown in Appendix Table 1. Shanghai and Beijing are the two largest cities in China by population. Tianjin is the sixth and Nanjing is the 11th.\textsuperscript{17} Shanghai has the highest average household income while Tianjin has the lowest. Appendix Table 1 shows that Beijing has the smallest increase in average nominal income of 42\% during the 5-year period while Shanghai has the largest increase of 57\%, with the inflation being 13.7\% during this period.

We rely on four main data sets together with a variety of auxiliary data for our analysis. The first main data set contains monthly vehicle sales by model (vintage-nameplate) in each city from 2008 to 2012. There are 84,912 observations (city-model-year-month) with 1,769 distinct models.\textsuperscript{18} Figure 2 plots monthly sales (in log) in each city and displays two important features. First, sales in all four cities grew over time and tracked each other well before 2011 and the trend is more consistent across Beijing, Nanjing and Tianjin, reflecting the fact that Shanghai has an auction policy in place throughout the data period. Second, there was strong seasonality which is largely driven by holidays. The sales in December 2010 went up dramatically in all cities but then dropped significantly in February 2011. This is due to the fact that Chinese New Year was on February 3rd in 2011.\textsuperscript{19} Third, the sales

\textsuperscript{16}During the first month of the new auction system in January 2008, the lowest price was 8,100 Yuan compared with the average price of 23,370 Yuan. This anomaly was due to a computer glitch. The abrupt change in price in December 2010 was due to the speculation that the policy was about to be phased out.

\textsuperscript{17}Beijing, Shanghai and Tianjin are three of the four province-level cities. They are at the same level of administrative subdivision as provinces and are right below the central government.

\textsuperscript{18}The data source is the Polk & co. (now a part of IHS) which provides market consulting services for a variety of industries including the automobile industry across the globe.

\textsuperscript{19}During the holiday season that starts from at least one week before the Chinese New Year and ends two weeks after, people are occupied with visiting families and friends. Big-item purchases such as buying a car
increase in December 2011 appeared to be stronger in Beijing than in other cities. This is due to the anticipation and more importantly the fact that the quota policy was announced in December 23, 2010. Very little if any discussion on the policy was made public before the announcement. However, once the policy was announced, many who planned to buy a vehicle in the near future moved their purchase forward into the last week of December. In the main specification of our analysis, we remove the last two months in 2010 and the first two months in 2011 in Beijing to deal with the anticipation and more importantly announcement effects.

The second data set contains vehicle characteristics of each model in the sales data. These characteristics include price, fuel economy (liters/100km), vehicle size, engine size, vehicle type, and vehicle segment. The summary statistics are presented in Table 2. Vehicle prices are computed based on the Manufacturer Suggested Retail Prices (MSRPs) and the sales tax.\textsuperscript{20} The sales tax is normally set at 10\% but was reduced to 5 and 7.5\% percent for vehicles with engine displacement no more than 1.6 liter in 2009 and 2010, respectively. The average MSRP in the sample is over 300,000 Yuan and the median price is over 190,000, both significantly higher than the average household income in these cities. During the data period, the quality-adjusted prices have been decreasing by 3\% each during year.\textsuperscript{21}

MSRPs are set by manufacturers and are generally constant across locations and within a model year. There could be potential pitfalls in using MSRPs when they are different from the transaction prices due to promotions. Different from the promotion-heavy environment in the U.S., China’s auto market has infrequent promotions from manufactures or dealers and retail prices are often very close to or the same as MSRPs, a phenomenon commonly seen for luxury products in China. Our analysis uses MSRPs with the implicit assumption that the allocation mechanisms in Shanghai and Beijing are not likely to affect firms price-setting behavior or local dealer incentives. This assumption should not be a driving factor in our results given the small market shares of these two cities and the marketing environment for automobiles in China.\textsuperscript{22}

The third data set is income distributions by city by year. Income is arguably the most
tend to occur before the holiday season.

\textsuperscript{20}MSRPs already include two types of taxes: consumption tax ranging from 1\% for vehicles with small engines to 40\% for large engines, and value-added tax of 17\% of before-tax price. These high taxes partly contribute to the high vehicle prices in China relative to those in the United States.

\textsuperscript{21}This is obtained from a price regression with vehicle nameplate fixed effects. The salient feature of price decline over time has been studied in Li et al. (2015) using data from 2004-2009. The study suggests that both the intensified competition and cost reduction contribute to the price drop over time.

\textsuperscript{22}Minimum Retail Price Maintenance (RPM) where automakers prohibit dealers from selling below a preset price was a common practice before 2014. This drew attention from China’s Antitrust authority in 2014 when several automakers were fined for this practice.
important determinant in vehicle purchase decisions. China’s National Bureau of Statistics conducts a census every 10 years but the income data at the household level are not publicly available. Instead, we obtain the average income by income quantiles in each city in each year from the statistical yearbooks of each city. We construct household income distribution based on these aggregate information together with Chinese Household Income Survey (2002), a national representative survey, conducted by researchers at the University of Michigan. We adjust the income in the household survey (14,971 observations) proportionally and separately for each of the quantiles so that the interpolated income distributions in a given year are consistent with the annual income statistics from the yearbooks.

The fourth set of data contains aggregate information from an annual national representative household survey among new vehicle buyers from Ford Motor Company. We were provided household shares by four income groups among new vehicles buyers in each city in each year. Appendix Table 2 presents these shares along with the shares of different income group among all households, which are constructed based on the income distributions from each city. The table shows that high income groups account for a disproportionally large share of vehicle buyers. While the highest income group (annual household income over 144,000) accounts for only 4% of all households in 2008 in Beijing, this group accounts for over 20% of new vehicle buyers. These information will be used to form additional moment conditions that are crucial to identify consumer preference parameters.

4 Empirical Model of Vehicle Demand

In order to conduct welfare analysis, we need to recover consumers’ WTP schedule for vehicle licenses. As shown in Section 2, the efficiency outcomes under different allocation mechanisms hinge on the heterogeneity of consumer WTP: different mechanisms would have the same allocation efficiency in the absence of any heterogeneity. We do not observe license prices in Beijing since the licenses are allocated through lottery for free and transfers are not allowed. Our strategy to recover consumers’ WTP for a license is to recognize that the WTP for a license reflects the additional cost a consumer is willing to bear in order to own a vehicle. Therefore, consumer surplus from owning a vehicle, i.e., the WTP for the vehicle minus the market price of the vehicle, should inform the WTP for the license itself. In particular, the WTP for the license itself should be equal to the consumer surplus from the most preferred vehicle if the license lasts as long as the vehicle (10-15 years).23

23 Although the current policy does not have an expiration date on the license and could last multiple vehicle lifetimes in theory, it is unlikely that consumers would take such a long-term perspective in valuing a
In order to estimate consumer surplus for different vehicles models, we set up and estimate a random coefficient discrete choice model of vehicle demand. In this section, we first specify the utility function, the basis of individual choices. We then discuss the aggregation process to obtain the market demand. We focus on the mechanism of the model in this section and leave the discussion on model estimation and identification to the next section.

### 4.1 Utility Function Specification

Let $m = \{1, 2, 3, 4\}$ denote a market (i.e., Beijing, Nanjing, Shanghai, Tianjin) and a year-month by $t$ from 2008 to 2012. Let $i$ denote a household and $j \in J$ denote a model (i.e., vintage-nameplate) where $J$ is the choice set. Household $i$’s utility from product $j$ is a function of household demographics and product characteristics. A household chooses one product from a total of $J$ models and an outside alternative in a given month. The outside alternative captures the decision of not purchasing any new vehicle in the current month. The indirect utility of household $i$ from product $j$ in market $m$ at time $t$ is defined as

$$u_{mtij} = \bar{u}(p_j, b_{mti}, X_j, \xi_{mtj}, y_{mti}, Z_{mti}) + \epsilon_{mtij},$$

where the first term on the right, $\bar{u}(.)$, denotes the deterministic component of the utility as a function of vehicle attributes and consumer characteristics. $p_j$ is the tax-inclusive price of product $j$ and it does not vary across markets and months within a year as we discussed in the data section. In Shanghai, for consumers who need to obtain a license through auction, $b_{mti}$ is the price paid (i.e., the bid); and it is zero for consumers who do not need a new license (e.g., after scraping a used vehicle).\(^{24}\) In other cities, $b_{mti}$ is always zero for everyone. The importance of variation in $b_{mti}$ for identification is discussed in Section 5.2. As shown above, the winning bids in the discriminatory multi-unit auction in Shanghai have very small spread. Given that we do not observe the distribution of the bids, we use the average bid as the price paid by all the winners. The effect of this measurement error on our results should be small since the difference between the average winning bid and the lowest winning bid is generally less than two percent of the average bid.

$X_j$ is a vector of observed product attributes (other than price) including a constant term, license given significant policy uncertainty and government efforts to improve and expand the public transit system. Our analysis assumes that a license last one vehicle lifetime and hence provides a theoretical lower bound on consumer WTP. Nevertheless, as we show below, the WTP estimates we obtained using this assumption are reasonable in magnitude based on several pieces of evidence.

\(^{24}\)As Figure 2 in Appendix shows, the difference between the maximum winner bid and the minimum winning bid is only about one to two percent of the average bid. In the estimation, I assume that all consumers who need to participate the auction pay the average winning bid in the corresponding month which ranges from 23,370 to 69,460 Yuan during the data period.
vehicle size, engine size and fuel cost. $\xi_{mtj}$ includes unobserved product attributes such as product quality and unobserved demand shocks to be specified below. $y_{mti}$ is the income of household $i$ and $Z_{mti}$ is a vector of (unobserved) household demographics. $\epsilon_{mtij}$ is an i.i.d. random taste shock and is assumed to follow the type I extreme value distribution. The utility from the outside good is defined as $u_{mti0} = \epsilon_{mti0}$, where $\epsilon_{mti0}$ also follows the type I extreme value distribution. Following the literature, we specify the first part of $u_{mtij}$, the deterministic utility to be:

$$\bar{u}_{mtij} = \alpha_{mti} \ln(p_j + b_{mti}) + \sum_{k=1}^{K} X_{mtjk} \tilde{\beta}_{mtik} + \xi_j. \quad (2)$$

$\alpha_{mti}$ measures consumer $i$’s preference or distaste for price and it is defined as:

$$\alpha_{mti} = \alpha_0 + \alpha_1 \ln y_{mti} + \sigma \nu_{mti},$$

where $\alpha_i$ will be negatively related to income if $\alpha_1$ is negative. One would expect high income households to be less sensitive to price due to diminishing marginal utility of income. $\alpha_i$ is also affected by unobserved household attributes captured by $\nu_{mti}$. We assume that $\nu_{mti}$ has a standard normal distribution in the benchmark specification and $\sigma$ is the standard deviation of a normal distribution.

A note on the functional form of consumer preference on price is in order. The literature on vehicle demand has used different specifications for the price and income interactions. BLP and Petrin (2002) use $\ln(y_{i} - p_j)$ and the term has a natural explanation as the utility from the composite good. As discussed above, the median vehicle price in China are higher than the average income of most households, hence this specification does not lend itself well to our context. One might argue that we should use the current payment on the vehicle rather than vehicle price in the utility function. However, nearly 90% of vehicle buyers make full cash payment upfront on their purchases in China. Goldberg (1995) specifies the price and income interactions as $\alpha_i(y_i-p_j)$ where $\alpha_i$ varies across income categories. She argues that one can view the price and income variables to be proxies for vehicle capital cost and the lifetime wealth of the household, respectively. Berry et al. (1999) specify $\alpha_i p_j$ where $\alpha_i$ is inversely related to income. In our context, both of these specifications do not lead to the intuitive pattern of price elasticity where more expensive products have less elastic demand. In order to generate that pattern, these specifications require that the increase in

---

25The median price of a new vehicle in China during our data period was about 200k Yuan and the average price was about $315k Yuan while only 16% of household income Beijing whose annual household income is above 144k Yuan. Among the new vehicle buyers, only 32% of buyers have an annual household income over 144k.
household income among the buyers has to be faster than the price increase if we compare two products with different prices. The current specification is chosen so as to allow income to affect consumer preference on price in a way that leads to the intuitive pattern of price elasticity. In obtaining consumer surplus and welfare analysis, we rely on the price variable rather than the income variable because of the difficulty in directly interpreting the income variable in our context.\footnote{Household income can be a rather imprecise proxy of wealth in these cities where housing value can account for the majority of the wealth. Housing values have increased several folds during the past 10 years in these cities. Many residents inherited housing from their parents and many others especially those who work at government agencies were given subsidized housing. These people do not necessarily have as high income as those who purchase their houses from the market. Alternatively, one can treat the income variable as another variable (such as education) that does not have a monetary unit.}

\[ X_{mtjk} \] is the \( k \)th attribute of product \( j \). \( \tilde{\beta}_{ik} \) is the random taste parameter of household \( i \) over product attribute \( k \). It is a function of unobserved household demographics captured by \( \nu_{ik} \), which is assumed to have a standard normal distribution.

\[ \tilde{\beta}_{mtik} = \bar{\beta}_k + \sigma_k \nu_{mtik}. \quad (3) \]

The preference parameters defined above underscore consumer heterogeneity that our model tries to capture. The heterogeneity will translate into heterogeneity in consumers’ WTP for a new vehicle, which is crucial for our welfare analysis. With all the components defined above, the utility function can be fully written out as the following:

\[ u_{mtij} = (\alpha_0 + \alpha_1 \ln y_{mti} + \sigma \nu_{mti}) \ln (p_j + b_{mti}) + \sum_{k=1}^{K} X_{mtjk} (\bar{\beta}_k + \sigma_k \nu_{mtik}) + \xi_{mtj} + \epsilon_{mtij}. \quad (4) \]

### 4.2 Choice Probabilities and Aggregate Demand

Based on the i.i.d. type I extreme value distribution of \( \epsilon_{mtij} \) and \( \epsilon_{mti0} \), the choice probability of household \( i \) for product \( j \) is

\[ Pr_{mtij}(p_j, b_{mti}, X_j, \xi_{mtj}, y_{mti}, Z_{mti}) = \frac{\exp(\bar{u}_{mtij})}{1 + \sum_h \exp(\bar{u}_{mthh})}. \quad (5) \]

When there is no quantity constraint, \( b_{mti}=0 \). \( Z_{mti} \) includes unobserved household demographics \( \nu_{mti} \) that affects consumer preference on price and \( \nu_{mtik} \) that affects consumer preference on other vehicle attributes such as size and fuel cost. This equation can be used directly to generate aggregate sales in the market when there is no quantity constraint such as in Nanjing and Tianjin as well as in pre-policy Beijing. Without quantity constraint, the
market share of product \( j \) in market \( m \) at time \( t \) is:

\[
S_{mtj} = \int Pr_{mtij}(p_j, X_j, \xi_{mtj}, y_{mti}, Z_{mti})dF(y_{mti}, Z_{mti}).
\]  

(6)

In the case of quantity constraint, the aggregation needs to take into account the allocation mechanisms. There are two types of households: those who need to acquire a new license to register a vehicle and those who do not. Under both lottery and auction policies, households who scrap a used vehicle can use the old license to register a new vehicle. We do not have household level data and therefore do not observe the type of households in this regard. Instead, we explicitly model the type probabilities as a function of vehicle ownership rate in Beijing and Shanghai. Denote \( L_{mt} \) as the probability that a household in city \( m \) and time \( t \) would need to obtain a new license to register a vehicle. We parameterize \( L_{mt} \) as a logistic function of observed vehicle ownership rate across markets and over time, \( o_{mt} \):

\[
L_{mt} = \frac{1}{1 + \exp(\gamma_0 + o_{mt} \gamma_1)}.
\]  

(7)

A positive coefficient \( \gamma_1 \) would imply that as vehicle ownership rate rises, the probability of a household needing a new license will decrease.\(^{27}\) In theory, we could allow the function \( L_{mt} \) to be individual specific by allowing it to depend on household income \( y_{mti} \). The separate identification of the two income coefficients (one in this function and the other in the utility function) comes from two sources. First, Nanjing and Tianjin do not have a quota system and the income coefficient in the utility function can be identified solely based on the sales data and the micro-moments to be discussed below from these two cities. Second, there is a difference between total new vehicle sales and the number of permits allocated in Beijing and Shanghai. The difference is driven by the consumers who buy a new vehicle to replace a used vehicle hence do not need a new license. The variation in this difference over time and across Beijing and Shanghai helps identify the \( L_{mt} \) function. When both \( o_{mt} \) and \( y_{mti} \) are included in the \( L_{mt} \) function in a robustness check, neither is statistically significant though the preference parameters in the utility function barely change. This suggests that the separate identification of the coefficients of \( o_{mt} \) and \( y_{mti} \) in the \( L_{mt} \) function would necessitate more detailed household-level data than what we have such as income differences among those who purchase to replace a used vehicle and other buyers.

In Beijing, the households who need a new license must obtain it through the lottery system. Denote the odds of winning a lottery in Beijing in a given month among all potential buyers from 2011 (\( t > 36 \)) as \( \rho \). Denote \( c_{mt} \) as a random draw from a Bernoulli distribution

\(^{27}\)Assuming \( L_{mt} \) to be constant over time during our data period in a robustness check does not significantly change our welfare analysis.
with probability of $L_{mt}$ being 1. The market share of product $j$ in Beijing ($m=1$) is:

$$S_{1tj|t>36} = \int \mathbb{1}(c_{1t} = 1) \Pr_{1tij}(\cdot) \ast \rho + \mathbb{1}(c_{1t} = 0) \Pr_{1tij}(\cdot) dF(y_{1ti}, Z_{1ti}, c_{1t}), \quad (8)$$

where $\mathbb{1}(\cdot)$ is the indicator function. The choice probability $\Pr_{1tij}$ is defined by equation (5) where the winning bid $b_{mti}=0$ since Beijing uses a lottery system. The first part of the integrand defines the case of buyers needing a new license and hence participating the lottery and the second part of the integrand for the case of buyers who do not need a new license (e.g., buy a new vehicle to replace a used one).

In Shanghai, vehicle buyers who need a new license have to pay the winning bid $b_{mti}$ through auction while those who buy a new vehicle to replace a used one do not need to pay ($b_{mti}=0$). The market share of product $j$ in Shanghai ($m=3$) is defined as:

$$S_{3tj} = \int \mathbb{1}(c_{3t} = 1) \Pr_{3tij}(b_{mti} > 0, \cdot) + \mathbb{1}(c_{3t} = 0) \Pr_{3tij}(b_{mti} = 0, \cdot) dF(y_{3ti}, Z_{3ti}, c_{3t}), \quad (9)$$

where the choice probability $\Pr_{3tij}(b_{mti} > 0, \cdot)$ is defined by equation (5) where the winning bid $b_{mti}$ is positive for those who need a new license.

The static model above follows closely with the BLP demand literature on automobiles but ignoring dynamics could be an important limitation given that automobiles are durable goods and transaction costs could be nontrivial. Recent literature has tried to estimate the demand for durable goods in a dynamic setting where dynamically optimizing consumers can choose among a set of available vehicle models or wait to purchase in the future by taking into account changes in product attributes, the choice set and the industry structure. To address the issue of large state space in this type of settings, Gowrisankaran and Rysman (2012) propose an empirical strategy based on the assumption that the logit-inclusive values (i.e., ex-ante discounted lifetime utility of buying a preferred product) are sufficient statistics for capturing the evolution of state variables that affect consumers’ dynamic decision. This alleviates the need to track a large set of state variables and makes the computation more tractable.28

Adopting this approach in my context is challenging for at least two reasons. First, our data spans only five years and estimating the process of the logit-inclusive value (e.g., a AR(1) process) based on the short time-series is unlikely to yield credible results, which is a crucial piece in this approach. Second, the approach does not lend itself well when there is

---

28Gowrisankaran and Rysman (2012) estimate a dynamic demand model for camcorders that incorporate rich product differentiation, consumer heterogeneity, rational expectations as well as repeated purchases. This method has been applied by Schiraldi (2011) and Li and Wei (2016) to study vehicle scrappage subsidies in Italy and US.
a need to include a large set of fixed effects as this will essentially increase the state space. In our BLP framework, our analysis shows that it is important to include year-month fixed effects as well as city-vehicle segment fixed effects to control for the national trend in vehicle sales and city-specific demand for different types of vehicles.\textsuperscript{29}

The BLP literature typically assumes that for each time period, a set of potential consumers from the same population arrive at the market to make choices. This assumption is strong in our context where vehicle ownership increases dramatically over time and the majority of vehicle owners are first-time buyers. In addition, the quota policy could also affect the pool of potential buyers from one period to another. In our main analysis, we allow the pool of the potential buyers to evolve over time by drawing them from the population without replacement. In estimation, consumers who are more likely to make a purchase in one period are less likely to appear in the pool of potential buyers in the next period.\textsuperscript{30}

Equation (7) and the simulation of consumer pool without replacement over time capture two different aspects of the market. Equation (7) captures vehicle replacement where the (second-time) buyers replace their used vehicles and are not subject to the quota system. The simulation of consumer pool without replacement over time allows the pool of new buyers (who are subject to the quota system) to evolve due to the impact of the quota policy. If we had household level data on vehicle ownership such as the age of the vehicle and vehicle purchase decisions, these two aspects of the market could be modelled to incorporate household-level heterogeneity in the simulation of potential buyers. For example, one could structure the simulation of consumer pool such that: (1) households who own vehicles say age 10 years or older have a low probability of needing a license when buying a new vehicle; and (2) households who recently bought a new vehicle have a low probability of being a buyer. Without detailed household-level information on vehicle ownership, we instead model the first aspect using vehicle ownership data at the city level through Equation (7).

We define the market size $N_{mt}$ to be half of the number of households in the city in a given year in the benchmark specification. To check the sensitivity of the result to this definition, we estimate a model where the market size is the total number of households as

\textsuperscript{29}A model without these fixed effects would put a large burden on other key model parameters to fit the data. For example, as one of the robustness checks below shows, without enforcing the common trend assumption which necessitates including time fixed effects and city-segment fix effects, the model would predict unusually low sales in the absence of the quota policy in Beijing because it would mistakenly attribute the low sales in 2011 to some unobserved negative demand shock, underestimating the welfare impact of the policy.

\textsuperscript{30}In the previous version, our analysis follows the BLP literature by drawing potential consumers from the same population in each period. The specification slightly over-estimates consumer price sensitivity by 11 percent and welfare results by 28 percent, although the qualitatively findings are the same.
has been traditionally done in the literature for the U.S. market.\textsuperscript{31} The results do not differ in any significant way as we will show below.

Furthermore, we make the conceptual distinction between the market size and the lottery pool. The market is composed of all potential buyers who first make lottery participation decisions. Those who decide to participate form the lottery pool and the winners then make vehicle purchase decisions. Our specification and aggregation method lump these two decisions together. The parameter $\rho$ in equation (8) captures both participation decisions and the winning odds of the lottery. Therefore, it should not be compared against the observed odds in the data. That is, our specification does not explicitly model the increase of the lottery pool due to the cumulation of unmet demand as well as strategic participation behavior, which explains the drop of the winning odds over time. As one of the robustness checks, we estimate the model without using the post-policy data in Beijing (2011-2012) and equation (8). The analysis shows that this simplified modelling choice does not have qualitative impacts on the welfare outcomes.

5 Identification and Estimation

5.1 Constructing Moment Conditions

Our goal is to recover the preference parameters in equation (4) in order to estimate consumers’ WTP for a license. The identification challenge comes from the fact that there are unobserved product attributes as well as demand shocks $\xi_{mtj}$ in the utility function. The unobserved product attributes such as product quality are likely to be correlated with prices. Previous studies show that ignoring these unobserved product attributes biases the price coefficient toward zero and leads to wrong welfare calculations. This challenge in fact motivated the methodology in BLP. An additional issue in our context is that unobserved demand shocks in Shanghai are likely to be correlated with average bids and hence render them endogenous. Ignoring this can also bias the price coefficient toward zero.

To facilitate the discussion on identification and estimation below, notice that the utility function in equation (4) contains terms that vary by households as well as terms that do not. We separate these two categories and rewrite the utility function as the following:

$$u_{mtij} = \delta_{mtj} + \mu_{mtij} + \epsilon_{mtij}, \tag{10}$$

\textsuperscript{31}In the U.S., about 13 percent of households purchased a new vehicle each year before the economic downturn in 2008. In China, the number was about 5 percent in 2012.
where \( \delta_{mtj} \) is the household-invariant utility or the mean utility of product \( j \) in market \( m \) at time \( t \). Based on equation (4), it is specified as follows

\[
\delta_{mtj}(\theta_1) = X_j \beta + \xi_{mtj} \\
= X_j \beta + \xi_j + \eta_t + \mathbf{1}(m = 3) \eta'_t + \zeta_{ms} + \kappa_{myr_t} + e_{mtj} \\
= \delta_j + \eta_t + \mathbf{1}(m = 3) \eta'_t + \zeta_{ms} + \kappa_{myr_t} + e_{mtj},
\]

where we write \( \xi_{mtj} \) into several terms in the second line. \( \xi_j \) is unobserved product attributes such as quality and safety features that do not vary over time and across markets. \( \eta_t \) captures time (year-month) fixed effects that control for common demand shocks and seasonalities across cities. \( \mathbf{1}(m = 3) \eta'_t \) captures Shanghai-specific time effects. \( \zeta_{ms} \) is city-specific preferences for different vehicle segments where \( s \) is an index for segments. \( yr_t \) is year (1 to 5) and \( \kappa_{myr_t} \) captures city-specific time trend. \( e_{mtj} \) is time-varying and city-specific demand shocks. The last line combines \( X_j \beta + \xi_j \) into product dummies \( \delta_j \), absorbing the utility that is constant for all households across the markets. The parameters in the mean utility function are denoted as \( \theta_1 = \{ \delta_j, \eta_t, \eta'_t, \zeta_{ms}, \kappa_{m} \} \).

The second part in equation (10), \( \mu_{mtij} \), is household-specific utility defined as:

\[
\mu_{mtij}(\theta_2) = [\alpha_0 + \alpha_1 \ln y_{mti} + \sigma \nu_{mti}] \ln (p_{mtj} + b_{mti}) + \sum_k x_{mjk} \nu_{mtik} \sigma_k^u.
\]

The parameters in the household-specific utility are denoted as \( \theta_2 = \{ \alpha_0, \alpha_1, \sigma, \sigma^u \} \). With this specification, we can rewrite the choice probabilities in equation (5) as follows

\[
\text{Pr}_{mtij}(p_j, b_{mti}, X_j, \xi_{mtj}, y_{mti}, Z_{mti}) = \frac{\exp[\delta_{mtj}(\theta_1) + \mu_{mtij}(\theta_2)]}{1 + \sum_h \{\exp[\delta_{mth}(\theta_1) + \mu_{mth}(\theta_2)]\}}.
\]

The market shares can be written as \( S_{mtj}(\delta_{mtj}, \theta_2, \theta_3) \), where \( \theta_3 = \{ \gamma_0, \gamma_1, \rho \} \) which characterizes the license allocation processes described above.

In the choice probabilities, unobserved product attributes and demand shocks are absorbed in \( \delta_{mtj} \) while the price and bid variables are in \( \mu_{mtij} \). If we could include market-time-product fixed effects subsuming \( \delta_{mtj} \), we can control for unobserved product attributes and demand shocks. However, this is impractical in this nonlinear model. BLP develop a methodology to back out \( \delta_{mtj} \). Under mild regularity conditions, for given vectors of \( \theta_2 \) and \( \theta_3 \), a unique vector of \( \delta_{mt} \) for each market that equalizes the predicted market shares with observed market shares can be recovered through a contraction mapping algorithm:

\[
\delta_{mt}^{n+1} = \delta_{mt}^n + \ln(S^o_{mt}) - \ln(\hat{S}(\delta_{mt}^n, \theta_2, \theta_3)),
\]

where \( n \) is the number of iterations. \( S^o \) is a vector of observed market shares while \( \hat{S}(\cdot) \) is
predicted market share. With the recovered $\delta_{mt}$ for given vectors of $\theta_2$ and $\theta_3$, $\theta_1$ can be estimated using a linear framework following equation (11).

To estimate the model, we follow BLP by using simulated GMM with the nested contraction mapping. The GMM is based on three sets of moment conditions. The first set is formed based on the city-year specific demand shocks in equation (11). The identification assumption is that these demand shocks are mean independent of city-year dummy variables, i.e., having zero mean at the city-year level:

$$E\left[ e_{mtj}(\theta_2, \theta_3) | d_{mt}\right] = 0,$$

where $d_{mt}$ are city-year dummies. This assumption amounts to that time-varying demand shocks have a common trend across cities and the common trend is controlled by time fixed effects. What is left from the time trend $e_{mtj}$ is not systematically different across cities. Note that we have also included city-segment fixed effects and these control for difference in levels in demand shocks. Since Beijing implemented the lottery in 2011 and 2012, this assumption implies that the lottery policy is exogenous to the time-varying demand-shocks in Beijing.

This common trend assumption (in the absence of the policy) is motivated by the graphical evidence in Figure 2 and it is a key assumption needed in the DID analysis for policy evaluation. Although one cannot test this assumption directly given that we do not observe the counterfactual of no policy for the treatment group (Beijing in our case), we have three years of data before the policy and we can examine if the pre-policy time trends are the same across the cities in a reduced-form framework. If they are the same before the policy, we would be more comfortable with the assumption (Heckman and Hotz 1989).

Since Shanghai implements an auction system throughout our data period, we do not have a pre-policy period for comparison. To allow for the possibility of different time trend between Shanghai and other cities, we include Shanghai-specific time effects in equation (11) in the benchmark specification as a conservative measure.\(^{32}\) Recall we have city-segment dummies (which swaps city fixed effects) and time fixed effects in the equation. This leaves us eight exclusion restrictions in the first set including city-year dummy variables for Beijing and Nanjing from 2009 to 2012 (Tianjin is the base group and year 2008 is the base year).

The second set of moment conditions is constructed based on the aggregate information from the household survey presented in the right panel of Appendix Table 2. We match\(^{32}\) A joint test on Shanghai-specific time effects (based on Wald-statistics) rejects the null in the benchmark specification.
the predicted shares of households by income group by city among new vehicle buyers to those in the table. We use the fourth group as the base group and this gives us 12 moment conditions (four cities each with three income groups):

\[
E_t \left[ \tilde{S}_{mg|buyers}(\theta_2, \theta_3) - S_{mg|buyers} \right] = 0, \tag{16}
\]

where \( g \) is a income group and \( \tilde{S}_{mg|buyers} \) is the predicted share of income group \( g \) among vehicle buyers while \( S_{mg|buyers} \) is the observed counterpart. The former is calculated as:

\[
\tilde{S}_{mg|buyers}(\theta_2, \theta_3) = \frac{\sum_{i=1}^{N_{mt}} d(y_{mti} \in INC_g) \sum_{j=1}^{J} Pr_{mtij}}{\sum_{i=1}^{N_{mt}} \sum_{j=1}^{J} Pr_{mtij}}, \tag{17}
\]

where \( d(.) \) is an indicator function being 1 for household \( i \) whose income \( (y_{mti}) \) falls into the income range of group \( g \), \( INC_g \). \( N_{mt} \) is the market size, the number of potential buyers. These moment conditions turn out to be crucial in recovering the curvature of the WTP schedule (i.e. heterogeneity in WTP).

The third set of moment conditions matches the predicted quantity of licenses to the observed quota in each month.\(^{33}\)

\[
E_t[\tilde{Q}_{mt}(\theta_2, \theta_3) - Q_{mt}] = 0, \tag{18}
\]

where \( \tilde{Q}_{mt} \) is predicted quantity of licenses and it is calculated for Beijing (\( m=1 \) and \( t > 36 \)) and Shanghai (\( m=3 \)) as the following:

\[
\tilde{Q}_{1t} = \sum_i \sum_{j=1}^{J} [I(c_{1t} = 1) Pr_{1tij}(\cdot) \ast \rho],
\]

\[
\tilde{Q}_{3t} = \sum_i \sum_{j=1}^{J} [I(c_{3t} = 1) Pr_{3tij}(b_{3ti} > 0, \cdot)], \tag{19}
\]

where \( I(.) \) is the indicator function and \( c_{mt} \) are random draws from a Bernoulli distribution as discussed in Section 4.2. There could be a time gap between winning a license and purchasing a vehicle. In Beijing, winners have six months to purchase a new vehicle while in Shanghai, winners have three months before the license expires. Many consumers indeed take their time to purchase their vehicles. In the estimation, \( Q_{mt} \) is not the quota observed

\(^{33}\)In theory, if we can characterize the bidding strategy (bids as a function of WTP) in Shanghai auction, we could then write down the marketing clearing conditions and match the predicted marketing-cleaning bids to observed winning bids. This could be added as an additional set of moment conditions. However, as discussed in Section 3.2 the Shanghai auction is a non-standard format where the equilibrium bidding function is difficult to characterize as shown in an experimental study by Liao and Holt (2013).
in that particular month; rather it is the average of the last six months and three months for Beijing and Shanghai, respectively.

We form the objective function by stacking these three sets of moment conditions. The procedure involves iteratively updating $\theta_2$ and $\theta_3$ and then $\delta_{m_j}$ from the inner loop of contraction mapping to minimize the objective function. The estimation starts with an initial weighted matrix to obtain consistent initial estimates of the parameters and optimal weighting matrix. The model is then re-estimated using the new weighting matrix.

5.2 Identification Sources and Strategy

In order to address the potential endogeneity in vehicle prices due to unobserved product attributes, the maintained identification assumption in the BLP literature is that unobserved product attributes are mean independent of those observed ones. This allows researchers to use the exclusion restrictions given by the product attributes of other products within the firm and outside the firm to form moment conditions (macro-moments). This identification assumption could be violated if firms choose product attributes (observed and unobserved) jointly (Klier and Linn 2012). Although our empirical setup follows closely the BLP literature, our identification strategy does not rely on this assumption. Because we have data from four cities, we include vehicle model fixed effects which would subsume both unobserved product attributes and household-invariant observed product attributes in the mean utility function in equation (11).

Our identification strategy is made possible by the fact that different households are paying different effective prices (price plus bid) for the same vehicle in Shanghai depending on whether they need a new license or not. This allows us to have all the price variables in the household-specific utility and be isolated from unobserved product and demand shocks. To understand this, imagine if we do not have data for Shanghai, we would have $\alpha_0 \ln(p_j)$ entering the mean-utility term. We would then need to estimate $\alpha_0$ for welfare analysis. Since the price variable and the unobserved product attributes would both appear in the mean utility, one would need to evoke some type of exogeneity assumption such as the one maintained in the literature to deal with price endogeneity.\(^{34}\)

The identification of the WTP schedule critically hinges on the coefficients on the price

\(^{34}\)Alternatively, one can assume away $\alpha_0 \ln(p_j)$ from the utility specification so that the price variable is always interacted with household demographics such as income and hence appear in the household specific utility alone as in Berry et al. (1999) and Beresteau and Li (2011). This could be a restrictive functional form and our estimation results do not support this form in our context as the coefficient estimate $\alpha_0$ is large in magnitude and statistically significant.
and income variables in equation (4). Because the price variables enter the household-specific utility, both cross-model and within model variation in prices are helpful for identification. There are several important sources of price variation. First, during our data period, MSRP's range from 36,000 to over 1.1 million Yuan with the quality-adjusted average price decreases by about three percent each year. Second, the sales tax, which affects consumer prices, reduced from 10% to 5% and 7.5% for vehicles with engine displacement no more than 1.6 Liter in 2009 and 2010, respectively. For other vehicles, the sales tax was kept at 10%. Third, there is considerable variation over time in average winning bids in Shanghai auction. The average winning bid ranges from 23,370 to 69,346 during the data period. This then affects effective price (price + winning bid) for consumers who need a license in Shanghai (for vehicle replacement, no license is needed). Intuitively, the discontinuity in purchase costs (prices plus bids) between Shanghai and other cities due to the auction policy helps identify the price coefficient. Higher purchase costs for buyers who need a new license in Shanghai lead to lower vehicle sales than in other cities, due to consumer disutility from high prices. The magnitude in sales reduction regulates the magnitude of the price coefficient.

The micro-moments on the share of different income groups among new vehicle buyers help identify heterogeneity in consumer price sensitivity and WTP. These micro-moments in Appendix Table 2 show that high income groups account for disproportionally larger shares among new vehicle buyers, suggesting that households with higher income are less price sensitive. The disproportionality is more salient in Shanghai. This can be explained by the fact that higher purchase costs in Shanghai make new vehicle purchase more out of reach for low income households. These micro-moments help pin down the coefficient on the income variable and hence the curvature of the WTP schedule. We will come back to this when we present our WTP estimates in Section 7.2.

Another source of heterogeneity in consumer WTP is captured by the random coefficient on the price variable and the coefficient estimate turns out to be small in magnitude, suggesting that the heterogeneity in WTP is largely driven by income. Our sensitivity analysis below shows that the large heterogeneity in the WTP schedule is not driven by the infinite support of the normal distribution used to characterize the unobserved heterogeneity.

5.3 Further Discussions on Computation

Before concluding this section, we offer some additional details for estimation. First, the estimation starts from generating a set of households in each year-month and in each market. Each of the households is defined by a vector of household demographics including income.
from the income distribution and unobserved household attributes from the standard normal. When generating the random draws, we use randomized Halton sequences to improve efficiency (Train 2003). Our results below are all based on 250 households in each year-month and market.\footnote{Using the benchmark specification, we tried 200 random draws but that made very little difference. The convergence criterion for the simulated GMM (outer loop) is 10e-8 while that for the contraction mapping (inner loop) is set to 10e-14 to minimize the approximation error.}

Second, we speed up the estimation process through a combination of two techniques. The first technique is to parallelize the computation across the four markets. The time savings from the parallel process more than offset the additional overhead time. The second technique is to modify equation (14) for the contraction mapping by employing Newton’s method where the update is based on the derivate of the market share with respect to the mean utility $\delta_{mt}$:

$$\delta_{mt}^{n+1} = \delta_{mt}^n + \left[ \frac{\partial \ln(\hat{S}(\delta_{mt}^n, \theta_2, \theta_3))}{\partial \delta_{mt}^n} \right]^{-1} \left\{ \ln(S_{mt}^\circ) - \ln(\hat{S}(\delta_{mt}^n, \theta_2, \theta_3)) \right\}. \quad (20)$$

Although additional time is needed to calculate the derivatives, there is still considerable savings from fewer iterations due to the quadratic convergence rate of Newton’s method.

### 6 Estimation Results

In this section, we first present evidence from the reduced-form regressions on the common trend assumption and the sales impact of the lottery policy in Beijing. Then we discuss the parameter estimates from the random coefficients discrete choice model.

#### 6.1 Evidence from Reduced-form Regressions

To examine the validity of the common trend assumption across the cities, we estimate the following regression based on data from 2008 to 2010 (pre-policy period).

$$\ln(S_{mtj}) = \alpha \ln(p_j + b_{mt}) + \delta_j + \lambda_{mt} + \eta_t + \mathbb{1}(m = 3)\eta_t' + \zeta_{ms} + \epsilon_{mtj}, \quad (21)$$

where the dependent variable is the log market shares. $b_{mt}$ is the average monthly winning bid in Shanghai and zero for other cities (i.e., $m \neq 3$).\footnote{As discussed above, in the structural analysis, we distinguish two groups of consumers in Shanghai: those who need a license and those who do not in buying a new vehicle. Those who buy a new vehicle to replace their used vehicles do not need a new license and hence the bid would be zero for them. In this reduce-from analysis, the aggregate nature of the model prevents us from having this distinction.} $\delta_j$ is model (vintage-nameplate) dummies. $\lambda_{mt}$ is city-year fixed effects to capture city-specific and time-varying demand
shocks. The common trend assumption assumes that these shocks are the same across cities in a given year. The other terms are defined the same as in equation (11): we include time (year-month) fixed effects, Shanghai-specific time effects and city-segment fixed effects.

Table 3 presents the regression results for three specifications. The first two use all observations while the third one drops the data in the last two months of 2010 in Beijing to remove the anticipation effect and more importantly the announcement effect in December 2010. In all specifications, the base group is Tianjin and the base year is 2008. The first specification does not include Shanghai-specific time fixed effects but include Shanghai-year fixed effects. The coefficient estimate on ln(price+bid) suggests a price elasticity of -4.33, which is a plausible magnitude. All the city-year interactions are small in magnitude and not statistically different from zero, suggesting a similar time trend across the four cities. Note that the income variable, arguably the most important determinant of vehicle purchase is only controlled for through fixed effects. A model with more flexible specification of income heterogeneity such as our demand model should be able to better control for the effect of more subtle changes in income (such as different rate of change among different income groups) on vehicle demand.

The second specification include Shanghai-specific time fixed effects to control for monthly demand shocks in Shanghai that are different from the base group and could be correlated with the average bid. The price coefficient reduces to -4.53, consistent with the conjecture that unobserved demand shocks that are correlated with the average bid can bias the price coefficient toward zero. Nevertheless, the difference in the price coefficient estimates is quite small. The city-year interactions again have small and insignificant coefficient estimates. The third specification produces very similar results to the second one, suggesting that the anticipation and announcement effects are not large enough perhaps due to the short notice. The evidence from Figure 2 and these results support the common trend assumption, the basis of our first set of moment conditions in the structural estimation.

We next use a DID framework to examine the sales impact of the lottery policy. These results will be compared with those from the structural demand model. The equation for DID is very similar to equation (21) except replacing city-year fixed effects with lottery policy dummies for Beijing in 2011 and 2012. The results are presented in Appendix Table 3. The first specification uses all observations while the other two drop observations in the last two months in 2010 and the first two months in 2011 in Beijing. While the first two specifications include city-specific time trend (up to second-order polynomials), the third one does not.
Using the full data set, the lottery policy is estimated to have reduced sales by 60.6% in 2011 and 50.7% in 2012. This implies that without the policy, the sales would have been 847,000 units in 2011 and 1.05 million units in 2012, compared with a pre-policy sales of 804,000 in 2010. The second specification produces slightly smaller sales impacts: 54.1% and 40.4% in 2011 and 2012, respectively. This is intuitive since we drop the last two months in 2010 where the increase in sales was partly due to the fact that people moved their purchase forward from the future. So without the policy, the sales would have been smaller in 2010. The sales impacts of the policy would have been smaller in 2011 and 2012 to be consistent with growth in other cities. These estimates imply that the sales would have been about 728,000 and 873,000 in the absence of the policy in 2011 and 2012. The third specification leads to slightly larger sales impacts than those from Specification 2. We will come back to these estimates for comparison once we obtain estimates from the structural model.

6.2 Parameter Estimates from the Demand System

Table 4 shows parameter estimates from the GMM estimation for five specifications. The first panel represent parameters in $\theta_2$ which appear in the household-specific utility function in equation (12). The three parameters in the second panel are the auxiliary parameters $\theta_3$ that are needed to incorporate the policies into the calculation of market shares as shown in equations (8) and (9). We do not present parameter estimates for $\theta_1$ since they are not needed to perform our policy simulations and welfare analysis: $\theta_2$, $\theta_3$, and $\delta_{mtj}$, the mean utilities from equation (20) suffice.

The first specification is the benchmark model and our preferred specification. Before we discuss the coefficient estimates and compare results across different specifications, we note that the magnitude of the preference parameters by themselves are hard to interpret and we defer much of the discussion on the comparison across specifications to the next two sections where we simulate sales and conduct welfare analysis using these parameters.

In the benchmark specification, the coefficient estimate on ln(price+bid) is negative while that on the interaction between this price variable and ln(income) is positive. This suggests that households with a higher income are less price sensitive. Given the range of ln(income) from 0.55 to 6.72, the first two coefficient estimates suggest that all households dislike high prices. The variable ln(income) in the specification is to capture the fact that the utility difference between a new vehicle and the outside good varies by income. The second and third coefficient estimates imply the partial effect of ln(income) is positive for about 95% of the vehicle models, suggesting that the utility difference increases with income.
The next three coefficients are random coefficients, representing the standard deviation estimates of the normal distribution for preferences on each vehicle characteristic. The random coefficient on constant captures the variation (due to unobserved household demographics) in the utility difference between a new vehicle and the outside good. The random coefficients are statistically significant, adding consumer heterogeneity to what is implied by income heterogeneity.

To get a sense of the magnitude of coefficient estimates on price variables, we calculate price elasticities based on model estimates. The average own price elasticity is -9.49 with a range of -7.80 to -14.53. Models with a higher price tend to have a smaller elasticity in magnitude (see Appendix Figure 3) as buyers of more expensive vehicles tend to be of higher income and less price sensitive. The average elasticity is somewhat larger in magnitude than the estimates obtained for the U.S. market which range from -3 to -8.4 (BLP, Goldberg 1995, Petrin 2002, and Beresteanu and Li 2011).37 We believe our estimates are reasonable and the difference could be attributed to at least the following two reasons. First, the income level in these four cities in China is less than one half of the U.S. income level during the data period of 1981 to 1993 used in Petrin (2002). To the extent that higher income would reduce price sensitivity, the differences in income could lead to the differences in price elasticities. Second, vehicle prices in our data are much higher than MSPRs in the U.S. for the same brand.38 For example, a Hyundai Sonata GLS Sedan with 2.4 Liter engine with base options had a MSRP of $19,695 in the U.S., and a similar model produced in China had an MSPR of 178,800 Yuan (over $28,000). That is, one needs to adjust our price elasticities downward (in magnitude) in order to compare them with the elasticities in the U.S. market.39

The first auxiliary parameter \( \rho \) is the ratio of total license allocated over the number of potential vehicle buyers (without the quota constraint) that would need a new license (e.g., first-time buyers) under the quota system. It measures the stringency of the quota system: the smaller it is, the more stringent the quota is. It is a very important parameter in estimating the counterfactual sales under the policy. The parameter is estimated to be 0.229 in the benchmark specification, implying that only one out of five potential buyers that

---

37 Petrin (2002) based on data from 1981 to 1993 in the U.S. market and Beresteanu and Li (2011) based on data from 1999 to 2006 both use micro-moments for estimation, yielding an average price elasticity of -6 and -8.4, respectively.

38 Imports account for less than 3% of the auto market in China. Most brands sold in U.S. are available in China but they are produced there by joint ventures between foreign and domestic auto makers. Please see Li et al. (2015) for a discussion on China’s auto industry.

39 High vehicle prices in China are in part due to the fact that they include three types of taxes on top of the prices that dealers get: value-added tax, consumption tax, and sales tax. For an average vehicle, these three amount to about one third of the vehicle price.
need a license are able to obtain a license through the lottery. As discussed above, this should not be compared with the observed odds because the observed lottery pool includes not only those who enter the market for a new vehicle in the current month but also unmet demand in the past months and future buyers. Our empirical model lumps lottery participation decision and vehicle choices together. Nevertheless, as we show below, the estimate of 0.229 (together with other parameters) leads to reasonable counterfactual sales without the policy.

The second and third auxiliary parameters define the probability of a buyer needing a new license given by equation (7). The positive coefficient $\gamma_1$ suggests that as vehicle ownership goes up, the share of potential buyers who need a new license decreases. This is intuitive since as vehicle ownership increases, more and more households would need to replace their old vehicles with new vehicles and hence do not need a new license. These two parameter estimates imply that about 80% of potential buyers in Shanghai and 78% in Beijing in 2012 would need a license should they decide to buy a vehicle.

### 6.3 Robustness Checks

To examine the importance of the first set of moment conditions based on the common trend assumption, we estimate the model without these moment conditions under Specification 2. The coefficient estimates on $\ln(\text{price}+\text{bid})$ and its interaction with income are both larger in magnitude. The average own price elasticity is -8.44 with a range from -6.40 to -13.11. These are similar to the estimates from the benchmark specification in magnitude. A key difference is that the estimate of $\rho$ is almost twice as large as that from the benchmark model. This estimate implies that the quota is much less stringent and as we show below, the model estimates from this specification predict unreasonably small sales under the counterfactual scenario of no policy and under-estimate consumers WTP for licenses as discussed below.

In another robustness check not shown in the table, the estimation does not utilize the second set of moment conditions (micro-moments) that are based on shares of new vehicle purchases by income group. These moment conditions are important in recovering the heterogeneity in WTP for a new car and price sensitivity across income groups. Without these moments, the parameter estimates imply that high income households are more price sensitive, running against our intuition as well as the results from the benchmark model. In addition, price elasticity estimates are larger in magnitude for more expensive vehicles, opposite to the results from the first two specifications. This highlights the importance of the micro-moments in identifying price coefficients and WTP schedule.

Table 4 also presents results from three additional specifications as further robustness
checks. These robustness checks show that our estimation results are robust to: (1) the
definition of market size; (2) the usage of licenses for buying used vehicles; and (3) the
distribution of the random coefficients. Specification 3 examines the sensitivity of the results
to the definition of market size. In the benchmark model, we assume half of the households
in a city participate in the market for new vehicles, implying the market size of a month is
the number of households divided by 24. Specification 3 assumes the market size to be the
number of all households as is often done in the studies on the U.S. market. The coefficients
by and large are very similar to those in the benchmark specification. This implies that the
mean utilities must be smaller in this specification in order to generate the same number of
new vehicle sales as in the benchmark specification.

Specification 4 assumes that 85% of the lottery winners use the license to purchase new
vehicles instead of used vehicles in Beijing and the number in Shanghai is 95%. As specified
by the policy, buyers of used vehicles need a license if they do not already have one (i.e.,
from scrapping an old vehicle). We do not have detailed statistics on the ultimate usage of
the licenses and we use these two numbers as the upper bound. We choose a lower number
for Shanghai because the buyers there tend to have higher income than in Beijing.\footnote{In May 2011, about 1600 licenses from the lottery were used for used vehicle purchases in Beijing according to information from used vehicle dealers: \url{http://auto.sina.com.cn/news/2011-06-21/0752790152.shtml}. In June and July 2011, the ratio was 11.7% and 14%, respectively: \url{http://auto.sohu.com/20110816/n316431063.shtml}.}

In the first four specifications, draws for random coefficients are from the standard normal
distribution which is unbounded. To remove the impacts of extreme values, Specification
5 uses random draws from a truncated normal distribution by removing the draws below
2.5 percentile and above 97.5 percentile. The results are again close to the benchmark
specification and so are the simulation results and welfare analysis shown below.

### 6.4 Impacts on Vehicle Sales

Table 5 presents the simulated sales under the counterfactual of no policy for various speci-
fications. Under the benchmark specification, the counterfactual sales in Beijing are 833,431
and 1,274,954 in 2011 and 2012 (with standard errors 144,810 and 136,000), relative to ob-
served sales of 334,308 and 520,442 under the policy. This suggests that the lottery policy
reduced sales by 60 and 59 percent, respectively in 2011 and 2012. The estimated sales
impact in 2011 is very close to those from the DID analysis in Appendix Table 3. The
sales impact in 2012 from the structural model is somewhat larger than those from the DID
analysis.
The model estimates predict the sales would have increased by 53% from 2011 to 2012 in Beijing without the quota system. This is quite a bit larger than what was observed in Tianjin and Nanjing (13% and 5%), our control groups. This larger uptake in sales, relative to other cities, could be due to the combination of the following two reasons. First, the number of households in Beijing increased by 2.8% from 2011 to 2012, compared to 0.6% in Nanjing, 1.4% in Shanghai, and 4.3% in Tianjin. Second, the average income of the top quantile increased from 158,233 to 189,213 from 2011 to 2012 in Beijing, a 19.6% growth. This is much larger than the other three cities: 6.5%, 10.9% and 7.9% in Nanjing, Shanghai and Tianjin, respectively. Given that the high-income group contributes disproportionally more to new vehicle purchases as shown in Appendix Table 2, the larger income growth in this group could lead to a larger increase in vehicle sales.

Under Specification 2 without the first set of moment conditions, the counterfactual sales in Beijing without the policy are 612,654 and 940,851 in 2011 and 2012, respectively. The 31% drop in sales in 2011 from 804,355 in 2010 without the policy is hard to explain given that the sales decrease during the same period in Nanjing and Tianjin was 6.6% and 0.02%. Since the common trend assumption is not enforced, this specification attributes part of the large sales drop from 2010 to 2011 from the quota system to a negative demand shock in 2011 in Beijing relative to other cities. This then leads to a larger estimate on $\rho$ (i.e., less stringent quota) and predicts a large decrease in sales without the policy.

It is interesting to note that although Specification 2 without the micro-moments does not provide sensible pattern of price elasticities, the simulated sales impacts (not shown in Table 5) are close to those from the benchmark model. This is because the policy impacts are mainly identified through the first set of moment conditions that utilize common trend assumption, just as in a DID analysis. The micro-moment conditions, on the other hand, are very important for identifying consumer WTP and subsequent welfare comparisons.

The bottom panel of Table 5 shows the counterfactual sales without the auction policy from 2008 to 2012 in Shanghai. The results from the benchmark specification suggest that the policy reduced the sales by 50-70% during this period. Interestingly, although Specification 2 leads to unreasonably low predictions in the absence of the policy for Beijing, its predictions for Shanghai are close to those from the benchmark specification. This is because the identification of the policy impacts in Shanghai is through the changes in treatment intensity (i.e., average bids) over time. That is, the estimation of the policy impacts in Shanghai is largely based on the parameter estimates on the price variables.

The last three columns of Table 5 presents the sales impacts from the three robustness
checks based on different assumptions on the market size, license usage for used cars, and
distribution of random draws. As we have discussed in the above section, the parameter
estimates from these specifications are very similar to those in the benchmark specification.
Consistent with that, this table shows that the sales impacts are by and large similar to the
estimates from the benchmark specification.

7 Welfare Analysis

The purpose of this section is to compare welfare consequences under the lottery and auction
systems and we focus on 2012 for illustration. The comparison is performed on both allocative
efficiency and the external costs associated with automobile usage.

7.1 WTP Schedules

We first derive consumer WTP schedule for a license in Beijing and Shanghai in 2012 based
on parameter estimates from the benchmark specification in Table 4. The WTP for a license
by household $i$ is given by the expected consumer surplus (CS) from the most preferred
vehicle:

$$E(CS_i) = E_i \left[ \max_{j=1,...,J} \frac{\bar{u}_{ij} + \epsilon_{ij}}{MU_{ij}} \right] = \frac{1}{MU} \ln \left[ \sum_{j=1,...,J} \exp (\bar{u}_{ij}) \right] \text{ if } MU_{ij} \text{ is constant},$$

where $MU_{ij}$ is marginal utility of money. The last equality holds only if the marginal utility of
money is constant. However, because our utility specification is nonlinear in price, we cannot
use the closed-form expression in the last equality to estimate consumer surplus. Instead,
we follow the simulation method developed by Herriges and Kling (1999) to estimate the
expected CS from the most preferred vehicle.\footnote{Independent of the estimation, we separately generate 2000 households for each market for each month
in 2012. Each household is characterized by a vector of income and random draws of unobserved household
attributes. For each household, we draw 200 vectors of $\epsilon_i$ to achieve better precision.}

Figure 3 depicts the WTP schedule or the demand curve for licenses in Beijing (top panel)
and Shanghai (bottom panel) in 2012. he WTP schedules can be viewed as populated by
households with different WTP since each household demands no more than one vehicle in
a period. The red vertical line denotes the level of quota: 259,800 in Beijing and 108,100
in Shanghai. The WTP schedules exhibit two salient features. First, there is considerable
heterogeneity with some households willing to pay more than 200,000 Yuan for a license in
both markets while many others (marginal consumers) are willing to pay very little for a
license. Among those with positive WTP in Beijing (would-be vehicle buyers without the quota system), the average WTP is about 46,020 Yuan with a standard deviation of about 60,041 Yuan. The average WTP in Shanghai is 23,278 Yuan with a standard deviation of 24,798 Yuan. Our random coefficients utility framework is well suited to capture consumer heterogeneity, which is critical for welfare comparison under different allocation mechanisms as we will show below.

The second feature is that the WTP schedule in Beijing is higher than that in Shanghai, implying that households in Beijing on average are willing to pay a higher price for a license. This is a reflection that consumers in Shanghai have better access to the public transit and hence have less strong demand for a new vehicle. This is also consistent with the simulated sales under the counterfactual scenario of no quota systems in Table 5: the counterfactual sales in Beijing and Shanghai in 2012 would have been 1,274,954 and 869,915, respectively. Although Shanghai has more households and a higher average income than Beijing, the sales are lower, reflecting a lower WTP for a new vehicle. As discussed above, Shanghai has been carrying out the auctions to distribute licenses for nearly three decades and the auction revenue has been used to build road infrastructure and improve public transit. Shanghai has a more extensive public transit network including both subway and buses and at the same time, the public transit system is less crowded in Shanghai. Therefore, consumers in Shanghai have a better outside option, implying a lower WTP for a new vehicle.

7.2 Supportive Evidence

Given the importance of the WTP estimates in our welfare analysis, we further discuss the identification of the WTP curves and provide additional supportive evidence on the magnitude of these estimates. As we mentioned in Section 5.2, the curvature of the WTP schedule is largely identified based on the micro-moments which show that high income groups accounts for a disproportionally large share among new vehicle buyers. For example, the lowest income group (annual income less than 48,000 Yuan) in Beijing in 2012 accounted for 16.5% among all households and 7.9% among new vehicle buyers. Meanwhile, the highest income group (annual income more than 144,000 Yuan) accounted for 15.9% among all households and 31.9% among new vehicle buyers. This suggests that the likelihood of vehicle

42 Anecdotal evidence from the internet suggests that the price of a vehicle license in the black market in Beijing was about 200,000 Yuan in 2012 even though there is a large legal risk in such transactions. Truncating the maximum WTP to 200,000 Yuan do not qualitatively affect the conclusion on welfare comparison in as to be discussed in Section 7.4.

43 For example, the total length of Beijing’s city rail system is 465km with 277 stations while that in Shanghai is 538km with 329 stations by the end of 2013.
purchase among households in the highest income group is four times as high as that among households in the lowest income group. The ratio is even larger in other cities, especially in Shanghai where it is more than 30, likely due to the selection outcome of the auction policy. The contrasts of these shares for the four income groups reflect the strong underlying effect of income on consumer WTP for new vehicles. Our empirical method recovers this underlying relationship by explicitly enforcing the micro-moment conditions based on the shares of different income groups among new vehicle buyers. In another word, the curvature of the estimated WTP curve is consistent with these micro-moment conditions.

We now offer two pieces of additional evidence to support our estimates. First, the intersection between the vertical lines and the WTP schedules in Figure 3 gives the market clearing prices under a uniform price auction, the highest rejected bid in such a auction. The clearing price would have been 75,890 Yuan in Beijing and 55,928 Yuan in Shanghai based on simulations. Although we do not observe license transactions and hence market prices in Beijing, we do observe the average and lowest winning bids in Shanghai. Among the 12 monthly auctions held in 2012, the lowest winning bid was 52,800 Yuan, only slightly below our model prediction. Since our estimation does not have any explicit constraints on WTP and auction clearing prices, this provides a favorable validity check on the estimation.

Our second piece of supportive evidence comes from a recent auction of impounded vehicles held by People's High Court in Beijing. On November 26, 2014, a 12-hour online auction was held to auction off 11 vehicles confiscated by the court and the auction was announced two weeks earlier and the information including photos about these vehicles were posted online. The vehicles are auctioned together with their license so this offers a way for those who have not win the vehicle license lottery to own a vehicle. The participants have to be eligible for entering the lottery (such as being a Beijing resident) and each participant can only bid on one vehicle. The highest bid wins and pays the winning bid. In the event of a tie, the participant who failed the lottery the most times wins. The court set a reserve price on each vehicle and quite curiously a maximum allowable bid (cap) as well, which is 150% of the reserve price.

Appendix Table 4 presents a summary on these vehicles and auction outcomes. According to the auction website, the reserve prices were set at the market values of these vehicles.

---

44 The average winning bids ranged from 53,195 to 69,346 Yuan during the 12 monthly auctions in 2012 while the lowest winning bids ranged from 52,800 to 68,900 Yuan.
45 The information on these vehicles as well as the auction results (all bids) are posted at http://ssxkc.cbex.com.cn/jcas/auctionlist.
46 For example, the vehicle with the lowest reserve price is a 2011 Peugeot 508 (manual transmission, 2.0 Liter) with a 111,989 Kilometers. The reserve price is 104,300 Yuan. The market value is 104,850 according
Therefore, the difference between the cap and reserve price, ranging from 52,150 Yuan (a Peugeot 508) to 155,000 Yuan (a Jaguar XF) can be viewed as the implicit asking price for the license itself. During the auction, the cap was reached for all vehicles within three minutes. For the cheapest model with a reserve price of 104,300 Yuan, the cap was reached in 6 seconds and there were in total 1596 bids at the cap of 156,450. This implies that all these 1,596 bidders have a WTP at least as high as 52,150 Yuan on the license itself. For the most expensive vehicle, the cap was reached at 1 minutes and 50 seconds and in the end 24 participating bidder at the cap. This implies that these 24 participants have a WTP of at least 155,000 Yuan for a license. Among the 4,100 participants, 3,637 bid at the cap, implying the vast majority of them have a WTP of at least 52,150 Yuan for a license.

Appendix Figure 4 plots the 11 implicit license prices against the number of participants whose WTP for the license itself is at least as high as the corresponding prices. Given how fast the caps were reached, it stands to reason that these implicit prices (cap and reserve price differences) could be much lower than the WTP of the participants who bid at the cap. In addition, although the reserve prices are set at the market values, it is unlikely these vehicles are the ideal vehicles (among all vehicles models in the auto market) for these bidders. If they keep these vehicles, they would suffer a loss in consumer surplus (relative to the most preferred vehicle). If they sell these vehicles in Beijing or elsewhere, there would be a transaction cost. Either way, this would reduce the WTP for the license itself compared with a situation where the most preferred vehicle is available in the bundle. The evidence from the auction suggests that our WTP estimates are plausible in both the magnitude and the range.

7.3 Externalities from Automobile Usage

After estimating the WTP curve, we now turn to the externalities for automobile usage post allocation under the two mechanisms. Vehicle usage generates multiple externalities including congestion, local air pollution, CO₂ emissions, and traffic accident (Parry et al. 2007). Congestion externality captures the time lost from travel delays due to increased vehicles on the road while the externality from local air pollution is due to human mortality and morbidity. The externality from CO₂ emissions captures the damages from climate change and accident externality arises from the extra risk of fatality, injuries and property damages imposed on other road users. These externalities could impose a huge burden on the society especially in rapidly developing countries where they are not adequately controlled.

to a large used-car website http://2sc.sohu.com.
for. There exists a large literature on measuring each component of the total external costs. Parry and Small (2005) develop a framework to aggregate these components on a basis of per-unit gasoline consumption and estimate the external costs of one gallon of gasoline consumption to be around $0.83 and $1.23 (in 2000$) in the US and UK, respectively. Parry et al. (2014) estimate the external costs at the country level for different type of fuels including motor gasoline for over 100 countries.

As we argued in Section 2, although auctions can achieve more efficient allocations than lottery in general, this may not be true when the resources being allocated generate negative externalities that increase with WTP. Table 1 in Section 2 provides evidence on why the externalities from automobile usage are positively correlated with consumer WTP for a vehicle license. First, households with a high WTP tend to be those with high household income. They tend to drive larger and more luxurious vehicles that have lower fuel economy and hence burn more gasoline for the same travel distance. Second, households with high income tend to drive more. Small and Van Dender (2007) estimate the elasticity of annual vehicle miles traveled (VMT) with respect to income to be 0.11 using the U.S. state-level data. The positive relationship between income and travel distance holds among households in Beijing according to the 2010 Beijing Household Travel Survey as discussed in Section 2.

Based on our demand estimation and simulations, Table 6 shows that the average fuel economy of new vehicles sold under the lottery is 9.14 Liters/100km compared to 9.55 under the auction. This is consistent with our discussion above: the auction system will lead to a market with more high income households than the lottery system. In calculating VMT for households in each income group, we use the average VMT in the corresponding income group based on the 2010 Beijing Household Travel Survey as discussed in Section 2. The average annual VMT of 16,350 km under the lottery is also smaller than that from the auction, 17,750 km. To quantify automobile externalities under the two mechanisms, we need to make two additional assumptions.

The first is with regard to the time horizon over which the external costs accrue. Although a license can theoretically stay with the households over multiple vehicle lifetimes, consumers are unlikely to take such a long-term view in valuing a license. In response to the fast-changing environment, policies in China tend to change frequently. In addition, there has

---

47 Larger and heavier vehicles also impose larger accident externalities as shown in (Anderson and Auffhammer 2014; Jacobsen 2013). However, accident externalities in general are found to be small relative to congestion and air pollution in the slow-moving urban traffic.

48 A robustness check is performed where we draw VMT from the empirical distribution in the corresponding income group based on the 2010 Beijing Household Travel Survey. Incorporating the heterogeneity in VMT within the same income group does not significantly affect the welfare results.
been significant efforts in improving and expanding the public transit system, which will affect demand for vehicles. How long the quota system will last and how binding it will be is uncertain. We carry out our analysis based on two scenarios: 15 years, the maximum legal age of usage, and 10 years. To the extent that the license can last multiple vehicle lifetimes and consumers do take this into account, our analysis would provide a lower bound on the welfare consequences. Nevertheless, reasonable variations of time horizons would change our welfare estimates but would not affect the qualitative comparison of the two mechanisms.

The second assumption is on the external costs associated with gasoline consumption. The literature has generally translated externalities on a per-unit of gasoline basis in the discussion on the optimal gasoline tax as a (second-best) policy instrument.\(^49\) We take the results from Creutzig and He (2008) that estimate the externalities from automobile usage including congestion, local air pollution, traffic accident and \(\text{CO}_2\) emissions in Beijing. The most plausible estimate from their study is 0.85 Yuan/km (in 2012 terms) where congestion and air pollution account for about 80%. This translates to 9.7 Yuan per liter of gasoline consumed at the average level of fuel economy, implying an external cost of $6.02 per gallon of gasoline.\(^50\) Parry and Timilsina (2009) estimate the external costs from automobile usage to be about $6.05 (in 2005 $) per gallon of gasoline in Mexico city. Mexico city probably offers a reasonable comparison to Beijing in that both have very severe congestion and air pollution. In addition, the average household income in these two cities is roughly the same, about one fourth of that in the US. We note that although reasonable variations on the unit external cost affect the magnitude of welfare estimates, they do not qualitatively affect the efficiency comparison between the two mechanisms in Beijing.

Figure 4 plots the external cost curve (the dashed downward sloping curves) for the 15-year time horizon in the top panel and 10-year time-horizon in the bottom panel. The external cost curves depict the total external costs from the automobile usage during a given time horizon for the corresponding households on the WTP curve. The estimates are obtained based on a five percent annual discount rate. Since the external costs increase with WTP, the auction system leads to larger externalities than the lottery system. As shown in Table 6, the total external costs during a 15-year time horizon is estimated to be 29.97 billion

\(^{49}\)Congestion is directly related to travel distance while \(\text{CO}_2\) emissions is proportional to gasoline consumption. Air pollution is more complicated as it relates to engine running time even for the same distance. In addition, it also has to do with vehicle age as old vehicles pollute disproportionately more (Kahn 1996).

\(^{50}\)For final calculation of external costs, we recognize that China currently imposes a gasoline tax of one Yuan per liter to deal with the externalities from automobile usage. So we take 8.7 Yuan/km as the (un-internalized) external costs. Parry et al. (2014) estimates the national-average external cost of $0.55 (3.4 Yuan) from per liter of motor gasoline in China. The externalities should be much larger in urban areas like Beijing given its much more severe traffic congestion and air pollution than most parts of the country.
Yuan under the lottery system and 32.88 under the auction system, implying a non-trivial difference of nearly three billion Yuan. With a 10-year time horizon, the estimates are 22.29 billion Yuan and 24.46 billion Yuan, respectively.

7.4 Social Welfare and Comparison

We now compare consumer surplus from license allocations and externalities associated with automobile usage under the two mechanisms. We first examine allocative efficiency in Beijing from lottery and auction mechanisms under the observed quota level of 259,800 units in 2012. We use a uniform price auction rather than the Shanghai auction format as the basis for comparison. As shown in Liao and Holt (2013), the Shanghai auction is non-standard and may not yield the efficient outcome. Table 6 presents the simulation results for two different scenarios regarding the time-span for the external cost analysis to be discussed below, bearing in mind that consumer surplus does not vary between the two scenarios.

Under a uniform price auction, the clearing price would have been 75,890 Yuan in Beijing in 2012. The total consumer surplus given by the area to the left of the vertical line in Figure 3 would have been 37.5 billion, but only 11.5 percent of it (or 4.31 billion) is realized under the lottery system, resulting in a welfare loss of nearly 33.2 billion.\footnote{This welfare loss can be decomposed into that from observed vehicle attributes (20.9 billion or 63\%) and the logit error term (12.3 billion or 37\%). The welfare impact from observed attributes is slightly lower than the 69\% figure in Petrin (2002) where the choice set is about 150 compared to over 400 in this context.} This striking result is due to: (1) there is a great deal of heterogeneity in WTP, and (2) the lottery allocates the license randomly. The significant welfare loss is robust to the three assumptions (market size, license usage for used vehicles, and extreme random draws) examined under the last three specifications in Table 4: it is 31.3, 29.3 and 30.1 billion Yuan, respectively.

Does the advantage of externality reduction from the lottery system change the efficiency ranking of the two mechanisms? The answer is no in our context because this advantage is dominated by the allocative cost from the lottery. The net social welfare (total surplus minus total external costs) from the lottery system is estimated to be -25.66 billion, compared with 4.62 billion from the auction. The net social welfare from the lottery system being negative is due to: (1) the external costs are larger than consumer surplus for a larger number of households with relatively low WTP as shown in Figure 4; and (2) the level of quota is set too high relative to the optimal level as we discuss below.

The difference in net social welfare implies that by using the lottery system instead of a uniform price auction, Beijing municipal government left 30.3 billion Yuan on the table in 2012 alone. If we use a 10-year time horizon to calculate total external costs, the difference in
net social welfare is about 31 billion. These findings suggest that under the model estimates and a set of reasonable assumptions, the lottery system performs poorly in efficiency relative to a uniform price auction. In terms of government revenue, a uniform price auction would have generated 19.7 billion Yuan for Beijing government in 2012. If the revenue were to be used for additional support to the public transit as Shanghai does, it would have more than doubled the existing subsidy level of 17 billion. By improving and subsidizing public transit, the auction revenue could be used to address the distributional or equity concern that is considered as a shortcoming of the auction system. In Shanghai, the auction revenue is designated for improving public transit and as a result, Shanghai has larger subway and bus networks than Beijing. With good public transit, consumers who were outbid in the auction have decent travel mode alternatives to satisfy their travel demand.

The above welfare analysis considers the auction revenue as a mere transfer from the high WTP consumers to the government. Nevertheless, the extent to which the government revenue would be used efficiently has important implications for social welfare. If (some of) the auction revenue were to be wasted, the welfare gain from the auction system would be diminished. In the extreme case where all the revenue from the auction were to be wasted, the auction system would still achieve a welfare gain of 10 billion Yuan relative to the lottery system under the benchmark specification. If the auction revenue were to be used to reduce pre-existing distortionary taxes such as the income tax, it could potentially generate the so-called double-dividend as discussed in the context of environmental taxes (Bovenberg and de Mooij 1994; Goulder et al. 1997). Alternatively, the revenue could be directly returned to consumers through lump-sum transfers. If the transfer is uniform across households, this would amount to a transfer from (high-income) buyers to (low-income) non-buyers.

Appendix Table 5 presents two robustness checks on the welfare comparison. Both deal with the the high end of the WTP distribution. Although the welfare estimates change, the

---

52 In 2013, the auction system yielded a revenue of 8.79 billion Yuan for Shanghai municipal government. The government in turn used 7.4 billion in 2013 to support and improve the public transit system and the remainder went into the surplus fund that they have been accumulating from the auction system (about 6 billion by the end of 2013). Among the 7.4 billion they used in 2013, 1.96 billion went to the purchase of new buses; 1.6 billion was for subsidy of bus fare; 0.58 billion for public transit maintenance; 3 billion for subway upgrade; and 0.35 billion for other service expenditure.

53 How would Beijing municipal government use the auction revenue and could it use the revenue as efficiently as in Shanghai? One concern for how efficiently the revenue would be used by the government has to do with corruption. Measurement of corruption is inherently difficult. Data from Beijing and Shanghai Statistical Yearbooks from 1998 to 2014 shows that Beijing had a lower number of corruption prosecutions per 10,000 government employees than Shanghai. The Enterprise Survey Data by the World Bank contain questions on the business environment including corruption in different cities in China. The survey of business owners and top managers in firms in 2012 does not show a significant difference between Beijing and Shanghai in questions related to corruption.
welfare comparison between the two allocation mechanisms stay the same qualitatively. The first panel assumes that the maximum WTP is 200,000 Yuan and truncates the distribution at that level. As a result, total consumer surplus dropped from 37.5 billion to 35.2 billion and the welfare loss from the lottery compared to auction decreased from 30.3 billion to about 28.2 billion in Beijing in 2012. The second panel assumes that households with a WTP over 200,000 Yuan would be able to get a license and own a vehicle under the lottery system, perhaps through other channels than winning the lottery. This increases both total realized consumer surplus and total external costs. Overall, the welfare loss from the lottery in this robustness check to 20.4 billion Yuan (nearly $2.5 billion), still a significant amount.

The last exercise we conduct is to examine the optimal level of quota in Beijing. Our analysis is not meant to devise the optimal policy for dealing with congestion and air pollution. In fact, real-time congestion pricing and gasoline taxes should be more effective policies than the quota system since they directly target the source of externalities: driving and gasoline consumption. Nevertheless, we investigate another margin of improvement within the framework of the quota system. The optimal level of quota is determined by the intersection of the WTP and external cost curves as shown in Figure 4. Under the 15-year time horizon, the optimal level of quota is 111,404, about half of the quota used in 2012. As shown in the second panel of Table 6, the clearing price from a uniform price auction would be 143,500 Yuan, nearly twice as high as it would be with the existing quota. Consumer surplus would be about 30 billion, less than 37.5 billion under the existing quota. However, the total external costs would be 14.4 billion with the smaller quota, less than half of those from the existing quota. As a result, the net social welfare would be 15.6 billion, compared with 4.6 billion under the existing quota. Government revenue would be 16 billion, only slightly smaller. If we use a 10-year time horizon, the optimal quota would be higher since the unit external costs would be smaller. The net social welfare would increase to 23.5 billion Yuan.

8 Conclusion

Air pollution and traffic congestion are arguably two of the most pressing issues for China’s urban residents. To combat these problems, transit authorities in several major cities in China are implementing vehicle license quota systems to curb the growth in vehicle ownership. How should they distribute the limited number of vehicle licenses? One might be quick to point out that market-based mechanisms that distribute the resources to those with

54Interestingly, the annual quota in Beijing is reduced to 150,000 in 2014 from about 250,000 in previous years. This new quota includes 20,000 licenses designated for electric vehicles.
the highest value would be more efficient. However, if the usage of the resources generates negative externalities that are positively correlated with the willingness to pay, the efficiency advantage of market-based mechanisms may disappear because the usage of the resources by those with the highest private benefit could lead to smaller social benefit.

Although there is a large and distinguished theoretical literature on resource allocation mechanisms since Coase (1959), with the exception of the auction literature, empirical studies that quantify resource misallocation and its welfare impacts are sparse. This study offers the first empirical analysis on the welfare outcomes of different mechanisms to distribute vehicle licenses in urban China. Our analysis shows that different allocation mechanisms under a quota system lead to dramatic differences in welfare outcomes. The lottery system offers a large advantage over the auction system in reducing automobile externalities. Nevertheless, this advantage is dominated by the significant allocative cost from misallocation. Taking into account both consumer surplus from allocation and automobile externalities post-allocation, the estimated welfare loss from Beijing’s lottery system is about 30 billion Yuan (nearly $5 billion) in 2012 alone. This underscores the importance of allocation efficiency in the presence of large consumer heterogeneity. The significant loss is especially concerning given that the central government has endorsed the quota system as a policy to combat air pollution in major urban areas and that Beijing is often a leader in setting local policies. Five other major cities have adopted quota systems since 2011: two are using lotteries while the others a hybrid system. Although the rationale for choosing the lottery system has not been made public, one can postulate that the decision reflects concerns for equality. We find that a uniform-price auction could have generated nearly 20 billion Yuan to Beijing municipal government in 2012, which could be used to double its subsidies to the public transit system or to address the distributional concern using other measures.

The vehicle license quota system being adopted by urban transit authorities in China is a blunt instrument that does not directly attack the source of externalities. Economists have long been advocating market-based mechanisms such as Pigouvian polices to correct for externalities. Among real-world implementations, congestion pricing schemes have been adopted in London, Singapore and Stockholm and have been found effective in reducing congestion (Anas and Lindsey 2011). Fuel taxes offer a second-best and administratively simple policy to deal with multiple automobile externalities (Parry et al. 2007). In this paper, we focus on the welfare consequences of the two allocation mechanisms within the framework of a license quota system. The comparison between the quota system itself with other policy alternatives such as road pricing and gasoline taxes is worthy of future research.
References


Figure 1: Welfare Comparison of Lottery and Auction with Externalities
Figure 2: New Vehicle Monthly Sales (in logarithm) in the Four Cities

- Beijing
- Shanghai
- Tianjin
- Nanjing
Figure 3: WTP Schedules for Licenses in Beijing and Shanghai

Notes: The top panel depicts the demand curve or the WTP schedule for licenses in Beijing in 2012 while the graph in the bottom panel is for Shanghai in 2012. The vertical lines are the quotas, 259,800 and 108,100 in Beijing and Shanghai, respectively. The clearing prices in the graph denote the clearing prices that would have emerged under a uniform price auction.
Notes: The external costs depicted in the top panel (the downward sloping dashed line) are calculated over a 15-year time span of using the license while those in the bottom panel are based on a 10-year time span. The solid vertical lines denote the observed quota level of 259,800. The dashed vertical lines are the optimal levels of quota of 111,404 and 179,407 for the top and bottom, respectively.
Table 2: Summary Statistics of Vehicle Sales and Characteristics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>S.D.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price (in 2012 Yuan)</td>
<td>311.31</td>
<td>193.92</td>
<td>277.75</td>
<td>36.49</td>
<td>1148.38</td>
</tr>
<tr>
<td>Monthly sales by model in Beijing</td>
<td>125.93</td>
<td>31.00</td>
<td>240.66</td>
<td>1.00</td>
<td>3464.00</td>
</tr>
<tr>
<td>Monthly sales by model in Nanjing</td>
<td>34.71</td>
<td>9.00</td>
<td>65.36</td>
<td>1.00</td>
<td>964.00</td>
</tr>
<tr>
<td>Monthly sales by model in Shanghai</td>
<td>57.01</td>
<td>11.00</td>
<td>131.23</td>
<td>1.00</td>
<td>3036.00</td>
</tr>
<tr>
<td>Monthly sales by model in Tianjin</td>
<td>57.33</td>
<td>12.00</td>
<td>125.75</td>
<td>1.00</td>
<td>2295.00</td>
</tr>
<tr>
<td>Vehicle size (m²)</td>
<td>8.01</td>
<td>8.09</td>
<td>1.01</td>
<td>4.20</td>
<td>10.97</td>
</tr>
<tr>
<td>Displacement (liter)</td>
<td>2.09</td>
<td>1.87</td>
<td>0.77</td>
<td>0.90</td>
<td>5.70</td>
</tr>
<tr>
<td>Liters per 100 kilometers</td>
<td>9.10</td>
<td>8.80</td>
<td>2.09</td>
<td>2.90</td>
<td>17.20</td>
</tr>
<tr>
<td>Yuan per 100 kilometers</td>
<td>66.17</td>
<td>63.62</td>
<td>16.10</td>
<td>20.61</td>
<td>129.95</td>
</tr>
<tr>
<td>Passenger car dummy</td>
<td>0.63</td>
<td>1.00</td>
<td>0.48</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>SUV dummy</td>
<td>0.22</td>
<td>0.00</td>
<td>0.41</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Minivan dummy</td>
<td>0.15</td>
<td>0.00</td>
<td>0.36</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Large dummy</td>
<td>0.14</td>
<td>0.00</td>
<td>0.35</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Luxury dummy</td>
<td>0.02</td>
<td>0.00</td>
<td>0.16</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Medium dummy</td>
<td>0.35</td>
<td>0.00</td>
<td>0.48</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Mini dummy</td>
<td>0.03</td>
<td>0.00</td>
<td>0.17</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Small dummy</td>
<td>0.21</td>
<td>0.00</td>
<td>0.41</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Upper medium dummy</td>
<td>0.24</td>
<td>0.00</td>
<td>0.43</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Notes: The observation is at the vehicle model-year-month level. Prices include vehicle sales tax which is 10 percent in 2008, 2011 and 2012. The tax varied across vehicles with different engine size in 2009 and 2010. There are 21,228 observations from 2008 to 2012 with 1,769 models (vintage-nameplate) in the data set.
Table 3: Pre-policy Trend Analysis

<table>
<thead>
<tr>
<th>Variables</th>
<th>Specification 1</th>
<th>Specification 2</th>
<th>Specification 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>S.E.</td>
<td>Coef.</td>
</tr>
<tr>
<td>Ln(price+bid)</td>
<td>-4.329</td>
<td>0.310</td>
<td>-4.530</td>
</tr>
<tr>
<td>Beijing*2009</td>
<td>-0.033</td>
<td>0.085</td>
<td>-0.033</td>
</tr>
<tr>
<td>Beijing*2010</td>
<td>0.021</td>
<td>0.080</td>
<td>0.021</td>
</tr>
<tr>
<td>Nanjing*2009</td>
<td>0.056</td>
<td>0.071</td>
<td>0.056</td>
</tr>
<tr>
<td>Nanjing*2010</td>
<td>0.081</td>
<td>0.069</td>
<td>0.081</td>
</tr>
<tr>
<td>Shanghai*2009</td>
<td>-0.062</td>
<td>0.095</td>
<td>No</td>
</tr>
<tr>
<td>Shanghai*2010</td>
<td>-0.105</td>
<td>0.088</td>
<td>No</td>
</tr>
<tr>
<td>Vintage-model fixed effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>City-segment fixed effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Year-month fixed effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Shanghai year-month fixed effects</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Notes: The dependent variable is ln(market shares). Specifications 1 and 2 use all the 47,232 observations from 2008 to 2010. Specification 3 drops observations in Nov. and Dec. of 2010 in Beijing to remove anticipation effect and has 46,462 observations. Tianjin is the base group and 2008 is the base year. The standard errors are clustered at the model level.
Table 4: Parameter Estimates from GMM

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.  S.E.</td>
<td>Coef. S.E.</td>
<td>Coef. S.E.</td>
<td>Coef. S.E.</td>
<td>Coef. S.E.</td>
</tr>
<tr>
<td><strong>Parameters in the household-specific utility ((\theta_2))</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ln(price+bid)</td>
<td>-16.810 0.225</td>
<td>-17.080 0.200</td>
<td>-14.515 0.185</td>
<td>-17.323 0.173</td>
<td>-16.571 0.173</td>
</tr>
<tr>
<td>Ln(price+bid)*Ln(income)</td>
<td>1.620 0.033</td>
<td>1.888 0.031</td>
<td>1.443 0.054</td>
<td>1.620 0.026</td>
<td>1.620 0.037</td>
</tr>
<tr>
<td>Ln(income)</td>
<td>-6.394 0.366</td>
<td>-7.619 0.352</td>
<td>-5.807 0.409</td>
<td>-6.404 0.301</td>
<td>-6.418 0.287</td>
</tr>
<tr>
<td>(\sigma) for Ln(price+bid)</td>
<td>0.067 0.014</td>
<td>0.073 0.026</td>
<td>0.073 0.016</td>
<td>0.069 0.012</td>
<td>0.067 0.027</td>
</tr>
<tr>
<td>(\sigma) for constant</td>
<td>0.655 0.133</td>
<td>0.698 0.415</td>
<td>0.714 0.152</td>
<td>0.583 0.184</td>
<td>0.643 0.304</td>
</tr>
<tr>
<td>(\sigma) for engine displacement</td>
<td>0.645 0.289</td>
<td>0.708 0.041</td>
<td>0.708 0.250</td>
<td>0.671 0.276</td>
<td>0.648 0.347</td>
</tr>
<tr>
<td><strong>Auxiliary parameters for allocation mechanism ((\theta_3))</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\rho)</td>
<td>0.229 0.030</td>
<td>0.384 0.087</td>
<td>0.236 0.039</td>
<td>0.220 0.034</td>
<td>0.231 0.030</td>
</tr>
<tr>
<td>(\gamma_0)</td>
<td>-1.426 0.203</td>
<td>-1.111 0.406</td>
<td>-1.422 0.262</td>
<td>-1.428 0.185</td>
<td>-1.429 0.179</td>
</tr>
<tr>
<td>(\gamma_1)</td>
<td>0.259 0.020</td>
<td>0.139 0.061</td>
<td>0.250 0.026</td>
<td>0.259 0.035</td>
<td>0.260 0.038</td>
</tr>
</tbody>
</table>

Notes: Specification 1 is the benchmark and the preferred model. Specification 2 does not include the first set of moment conditions (common-trend moments). Specification 3 assumes all the households as the potential buyers in a year so the market size in each month is the total number of households divided by 12. Specification 4 assumes a certain percentage of licenses are used to buy used vehicles: 15 percent in Beijing and 5 percent in Shanghai. Specification 5 takes the random draws for unobserved household attributes from a standard normal distribution removing draws below 2.5 and above 97.5 percentiles.
Table 5: Policy Impacts on Sales in Beijing and Shanghai

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Beijing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>334,308</td>
<td>833,431</td>
<td>612,655</td>
<td>843,796</td>
<td>848,434</td>
<td>830,453</td>
</tr>
<tr>
<td>2012</td>
<td>520,442</td>
<td>1,274,954</td>
<td>940,851</td>
<td>1,289,578</td>
<td>1,297,333</td>
<td>1,269,863</td>
</tr>
<tr>
<td></td>
<td>Shanghai</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>165,298</td>
<td>373,569</td>
<td>319,957</td>
<td>356,631</td>
<td>385,944</td>
<td>369,692</td>
</tr>
<tr>
<td>2009</td>
<td>208,570</td>
<td>482,096</td>
<td>413,194</td>
<td>461,496</td>
<td>498,211</td>
<td>478,276</td>
</tr>
<tr>
<td>2010</td>
<td>264,232</td>
<td>662,500</td>
<td>567,299</td>
<td>635,184</td>
<td>684,922</td>
<td>659,805</td>
</tr>
<tr>
<td>2011</td>
<td>277,119</td>
<td>744,558</td>
<td>629,219</td>
<td>715,686</td>
<td>769,095</td>
<td>739,892</td>
</tr>
<tr>
<td>2012</td>
<td>295,047</td>
<td>869,915</td>
<td>723,721</td>
<td>830,736</td>
<td>897,528</td>
<td>865,674</td>
</tr>
</tbody>
</table>

Notes: The five counterfactual outcomes correspond to the five specifications in Table 4. Specification 1 is the benchmark and preferred model. Specification 2 does not include the first set of moment conditions (common-trend moments). Specification 3 assumes all the households as the potential buyers in a year so the market size in each month is the total number of households divided by 12. Specification 4 assumes a certain percentage of licenses are used to buy used vehicles: 15 percent in Beijing and 5 percent in Shanghai. Specification 5 takes the random draws for unobserved household attributes from a standard normal distribution removing draws below 2.5 and above 97.5 percentiles. The sales in Beijing from 2008 to 2010 were 411,936, 602,219, and 804,355, respectively.
Table 6: Welfare Comparison Between Lottery and Auction in Beijing in 2012

<table>
<thead>
<tr>
<th></th>
<th>15-year Horizon</th>
<th>10-year Horizon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lottery (1)</td>
<td>Auction (2)</td>
</tr>
<tr>
<td></td>
<td>Lottery (3)</td>
<td>Auction (4)</td>
</tr>
<tr>
<td><strong>Observed Quota Level</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quota in 2012</td>
<td>259,800</td>
<td>259,800</td>
</tr>
<tr>
<td>Clearing price (in 1000 Yuan)</td>
<td>0.00</td>
<td>75.89</td>
</tr>
<tr>
<td>Realized consumer surplus (in billion)</td>
<td>4.31</td>
<td>37.50</td>
</tr>
<tr>
<td>Unrealized consumer surplus (in billion)</td>
<td>33.19</td>
<td>0.00</td>
</tr>
<tr>
<td>Average fuel economy (liters per 100km)</td>
<td>9.14</td>
<td>9.55</td>
</tr>
<tr>
<td>Annual VMT (1000 kms)</td>
<td>16.35</td>
<td>17.75</td>
</tr>
<tr>
<td>Total external costs (in billion)</td>
<td>29.97</td>
<td>32.88</td>
</tr>
<tr>
<td>Government revenue (in billion)</td>
<td>0.00</td>
<td>19.72</td>
</tr>
<tr>
<td>Net social welfare (in billion)</td>
<td>-25.66</td>
<td>4.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Optimal Quota Level</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal quota</td>
<td>1,114,04</td>
<td>1,794,07</td>
</tr>
<tr>
<td>Clearing price (in 1000)</td>
<td>143.50</td>
<td>102.50</td>
</tr>
<tr>
<td>Consumer surplus (in billion)</td>
<td>30.04</td>
<td>40.50</td>
</tr>
<tr>
<td>Total external costs (in billion)</td>
<td>14.40</td>
<td>17.04</td>
</tr>
<tr>
<td>Government revenue (in billion)</td>
<td>15.99</td>
<td>18.39</td>
</tr>
<tr>
<td>Net social welfare</td>
<td>15.64</td>
<td>23.47</td>
</tr>
</tbody>
</table>

Notes: All monetary variables are in Yuan ($1 for about 6.1 Yuan). The results are based on 2000 households in each month. For each household, 200 random draws of the logit error term are used to generate the average WTP for a license for that household. The top panel presents welfare comparison between a lottery system and a uniform price auction for Beijing in 2012. The bottom panel shows welfare outcomes with the optimal levels of quota. The optimal quotas are defined by the intersections between the vertical dashed line and the downward sloping dashed line in the two panels of Figure 4. Net social welfare is equal to realized consumer surplus minus total external costs. Columns (1) and (2) use a 15-year time span for calculating total external costs, i.e., total discounted external costs during 15 years with a discount rate of 5%. Columns (3) and (4) are results based on a 10-year time span for external costs.