Can Big Data Resolve Selection Biases?

Evidence from Wholesale Used Car Auctions

Michael Fogarty

March 20, 2019

Abstract

Can big data help overcome problems with making causal inferences from observational data? I use a large dataset of wholesale used car auctions to test whether incidental truncation, a form of sample selection where the dependent variable is only observed for a subset of observations, biases a hedonic pricing model for used cars. I estimate OLS and sample selection-corrected models to assess the extent of the sample selection bias. I employ a novel approach that uses plausibly exogenous variation in the ability of the auctioneers to identify a two-step sample-selection model. I find that, in the base specification with no controls, there is a significant degree of sample selection bias. However, adding progressively finer fixed effects (for make, model, body type, and model year) attenuates and eventually eliminates the sample selection bias. This example illustrates that big datasets with rich details can potentially help researchers mitigate sample selection bias and make credible causal claims from observational data.

1 Introduction

I show how big data can potentially help overcome obstacles to making causal inferences from observational data. Using a large dataset of wholesale used car auction results, I explore whether incidental truncation biases a hedonic regression of auction sale price on observable car characteristics such as age and odometer mileage. I find that the standard OLS predictions converge to the bias-corrected predictions after controlling for detailed car-level characteristics.

Empirical economists are often concerned with drawing causal claims from observational data. An important impediment is that with "observational data, correlations are almost certainly not reflecting a causal relationship because the variables are endogenously chosen."[1] Endogeneity and sample selection both threaten causal inference by biasing the estimated effects from the observed data. Economists have developed a variety of techniques to overcome these hurdles, including difference-in-differences models, instrumental variables, and regression discontinuity designs; big data are another potentially useful tool for this purpose.

Big data, or data collected "as the byproduct of some other business activity,"[2] provide an opportunity for economists to test their theories on very large, detailed datasets. However, it is not obvious *ex-ante* that using these new, bigger datasets reduces the endogeneity and sample selection problems that complicate causal inference from observational data. It is conceivable that using larger datasets would instead provide more and more precise estimates of the biased effects. For big data to be useful for researchers the data must provide more information for each observation, not merely more observations.

Varian (2014) suggests that big data can be useful for identifying causal relationships by better modeling "both the observed difference in outcome and the selection bias."[3] Einav

[1] Cunningham (2018, p. 18)

[2] Cunningham (2018, p. 18)

[3] Varian (2014, p. 22)

and Levin (2014) concur with this assessment, arguing that "the use of highly granular data to find targeted variation that plausibly allows for causal estimates."[4] In this paper I do just that; I explore sample selection bias in a hedonic pricing model of used cars, using a large and detailed dataset of wholesale used car auctions.

Used car auctions are an integral part the used car market in the U.S., which is substantially larger than the new car market in terms of the number of vehicles sold. Over 15 million used cars are sold at auction each year, which accounts for more than one third of the roughly 40 million used cars sold in 2017. Wholesale used car auctions provide liquidity to dealers and the used car market as a whole. The number of used cars sold at auction each year is roughly equivalent to the number of new car sales.[5]

Just over 50% of the cars in my sample successfully sell at auction. Therefore, I only observe the key dependent variable, the sale price, for half of the sample. The observation of sale price itself depends on the outcome of another variable — whether or not the car sells. This is a type of sample selection called incidental truncation.[6] If the auction outcomes are randomly assigned, this is not problematic. However, whether or not the car sells at auction is determined endogenously in the auction process. Incidental truncation can bias the coefficient estimates in a hedonic regression of car characteristics on price because the regression is run on a non-random subsample of the data.

I take advantage of a large, rich dataset of wholesale used car auctions to test my hypotheses. The data consist of individual car characteristics (make, model, age, mileage) and details of the auction environment, including unique auctioneer identifiers. The data span five years and total 20 million unique auction observations. The data come from one of the two main wholesale used car auction houses in the United States.

[4] Einav and Levin (2014, p. 4)

[5] Cox Automotive (2018)

[6] Wooldridge (2012, p. 615)

I test whether the incidental truncation biases the coefficient estimates in a standard OLS hedonic regression of price on the observable characteristics of the cars. I employ a novel approach that uses variation in the ability of the auctioneers as the exclusion restriction to identify a two step sample-selection model. I take advantage of the (conditional) random assignment of cars to auctioneers, which creates plausibly exogenous variation in the probability that a car sells at auction. I then compare sample selection-corrected model to the the uncorrected OLS model to assess the extent of the sample selection bias.

I find that, in the base specification with no fixed effects, there is a significant degree of sample selection bias. Failing to account for the unobserved correlation between whether or not a car sells and the sale price biases the coefficient estimates. Strikingly, adding progressively finer fixed effects (for make, model, body type, and model year) attenuates and eventually eliminates the sample selection bias present in the baseline specification. In this case, controlling for individual-level characteristics eliminates the practical consequences of selection bias, which is consistent with Varian's and Einav and Levin's arguments that big data can facilitate controlling for sample selection bias and drawing causal inference from observational data.

The rest of the paper is organized as follows. In Section 2, I review the literature on used cars, auctions, and wholesale used car auctions. In Section 3 I describe my theoretical model and lay out my identification strategy in more detail. In Section 4 I describe the auction environment and my data. In Section 5 I present and discuss my empirical results. I then conclude.

2 Literature Review

Used car sales are a classic example of a market with information asymmetries. In his famous paper "The Market for Lemons," Akerlof (1970) argues that private information on the part of the seller can unravel the market and prevent otherwise efficient exchanges. Buyers have the incentive to misrepresent low-quality cars, or “lemons”, as high quality cars. This adverse selection problem drives sellers of high-quality cars out of the market and leads buyers to infer that used cars advertised as high quality are actually lemons. The auction house serves as an intermediary in the used car market that helps overcome the adverse selection problem. The auction house also provides ex-post arbitration if a buyer believes they have been misled.

There is a moderately-sized literature that tests auction theory and auction mechanics in the wholesale used car auction environment. Most of these papers use hedonic pricing models to test the impact of various factors, such as the auctioneer heterogeneity or information disclosure, on auction outcomes. I add to this literature on wholesale auto auctions by addressing potential sample selection bias resulting from the fact that a large fraction of cars fail to successfully sell at auction.

Lacetera and Sydnor (2014) use similar wholesale used car auction data to evaluate the impact of the country in which a car is manufactured on sale price. They find that for older model years, cars manufactured in Japan sell at a small premium to observationally identical cars built in the United States, but that the difference disappears after 2002.

Tadelis and Zettelmeyer (2015) show in an experimental setting that imperfect information plays a key role in the wholesale used car market, and is an important friction preventing cars from selling at auction. They perform a field experiment at a used car auction in which they randomly disclose information about the car’s quality to bidders in some auctions. Interestingly, information disclosure had a positive impact on the probability of sale *for all* types of cars, regardless of whether they are in good or bad condition. Tadelis and Zettelmeyer argue that this information signal allows bidders with heterogenous preferences over cars to better sort on the type of car they wish to purchase at the auction.

Larsen (2018) explores the efficiency of the ex-post bargaining system employed when cars do not sell at auction. He finds that incomplete information is not the only force that drives a large portion of auctions to end with no exchange. His key finding is that "8–14% of feasible trades (cases where the buyer indeed values the good more than the seller) fail."[7] While a portion of this inefficiency is due to incomplete information and adverse selection (as shown by Tadelis and Zettelmeyer's information disclosure experiment), some of the inefficiency is driven by the auction and ex-post bargaining mechanisms themselves.

Lacetera et al. (2016) consider the effect of the auction environment, specifically the auctioneer, on auction outcomes. They find that in the oral ascending auction used at the wholesale used car auction, differences in auctioneer ability do have an important impact on auction outcomes, specifically the "auctioneer's conversion rate, defined as the fraction of auctions that end in a sale."[8] They find substantial variation in auctioneer ability, and that ability has a substantive impact on auction outcomes. The authors conclude that "the most successful auctioneers tend to be those who can best manipulate the pace of their auctions, with faster paced auctions resulting in better conversion rates."

Auctioneer's are incentivized to maximize their conversion rate. The auctioneers are independent contractors paid a flat daily wage, but "the auction house periodically uses small bonus incentives tied to targets like the fraction of cars sold in a lane per day."[9] Much like a stock exchange, the auction house cares about maximizing the number of successful transactions, rather than the sale price. The general manager of one of the auction houses emphasized this when he wrote to the authors that "conversion rate pays the bills."[10]

The main threat to causal inference is that cars are not assigned randomly to different auctioneers, so the auctioneer's raw conversion rates, while correlated with their underlying skill, do not perfectly reflect the auctioneer effect on the probability of sale. "Fleet/lease" sellers (rental car companies, leasing companies, and corporate fleets) frequently bring their

[7] Larsen (2018, p. 5)

[8] Lacetera et al. (2016, p. 196)

[9] Lacetera et al. (2016, p. 200)

[10] Lacetera et al. (2016, p. 202)

own auctioneers and also set very low reserve prices, so a large fraction of their cars sell at auction compared to those sold by car dealers.

Lacetera et al. restrict their sample to cars sold by other used car dealers, and argue that cars are randomly assigned to auctioneers conditional on a variety of car and auction environment specific factors. Their preferred specification controls for mileage as well as seller, time-of-day, auction house, lane, and make, model, age, and body type fixed effects. The auctioneer effects are substantial: "a one standard deviation increase in auctioneer performance corresponds to an increase in the probability of sale of 2.3 percentage points (about a 4.3 percent increase over an average conversion rate of 0.53)"[11]

Additionally, the auction company produces blue-book style price estimates based on hedonic regressions, so sample selection bias could impact a relatively important part of their business, as these blue book prices help auctioneers, buyers, and sellers set expectations about a reasonable sale price.

3 Sample Selection Model

Figure 1 below illustrates visually the sample selection problem. In the figure, X are the car-level characteristics: make, model, age, mileage, etc. There is a direct effect of the car characteristics on the main outcome variable Y, the sale price. However, X also affects S, an indicator variable for whether or not the car sells. If there is no relationship between S and Y, then regressing Y on X will give the causal effect of the car characteristics on the sale price.

However, if there are unobserved factors U that influence both the sale price and whether or not the car sells, then regressing Y on X will produce a biased estimate of the effect of car characteristics on price. If, as I have assumed in the diagram, there is a relationship between

[11] Lacetera et al. (2016, p. 197)

[12] Cunningham (2018, ch. 4)

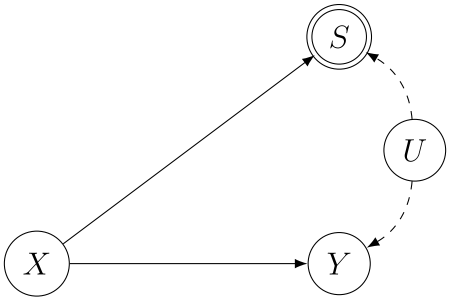


Figure 1: DAG Representation of Sample Selection Model[12]

X and S (i.e. $\text{Cov}(X, S) \neq 0$) as well as between S and Y, then omitting this from an OLS regression would bias the estimate of β, the true effect of X on Y, as seen in Equation 1.

$$\hat{\beta} = \beta + \rho \frac{\text{Cov}(X, S)}{\text{Var}(X_1)} \tag{1}$$

In order to get a clean estimate of the effect of X on Y when there is unobserved correlation between inclusion in the sample and the sale price, I need to identify some other factor A that influences Y only through its influence on S. Figure 2 illustrates this requirement visually; the key assumption is that there is no direct causal path from A to Y. Then, taking into account the relationship between S and Y caused by changes in A, I can identify the effect of X on Y.

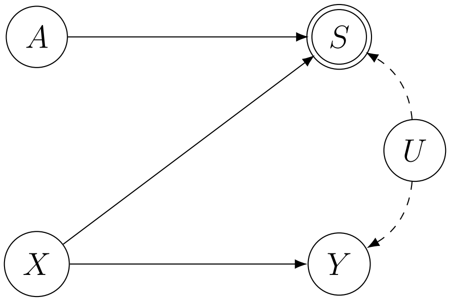


Figure 2: Sample Selection with Exclusion Restriction

3.1 Heckman Sample Selection Model

Heckman (1979) introduced a statistical method to correct for the type of sample selection bias illustrated in the diagrams above. A simple hedonic regression for the price of the wholesale used cars is:

$$\hat{Y}_i = \alpha + \beta_1(Age_i) + \beta_2(Miles_i) + X_i\delta + u_i \tag{2}$$

Where X is a vector of fixed effects dummies for make, model, model year, etc. Equation 2 can only be estimated if $S = 1$, when the car successfully sells at auction.

Heckman proposed a two-step procedure to correct for sample-selection bias. In the first step, I estimate a probit regression on the likelihood of inclusion in the sample (i.e. $S = 1$).

$$P(S = 1)_i = \Phi(Z_i\gamma) \tag{3}$$

Where i indexes individual cars and $Z \supset X$. There must be an additional regressor (A) that affects the sale price only through its effect on probability of sale. This is the exclusion restriction.

In the second stage, I estimate a linear regression on the subsample for which I observe the dependent variable with the inverse Mills ratio, $\frac{\phi}{\Phi}$, evaluated at $Z\gamma$, the linear predictions of the probit model, included as a covariate. The inverse Mills ratio is decreasing in the probability of inclusion in the sample. The second stage regression is:

$$\hat{Y}_i = \alpha_1 + \beta_1(Age) + \beta_2(Miles_i) + X_i\delta + \rho\hat{\lambda}\,(Z\gamma) + u_i \tag{4}$$

If ρ, the coefficient of the inverse Mills ratio, is statistically different from zero, then there is evidence of a sample selection problem, because there is a correlation between the probability of inclusion in the sample and the regressors in the main model.

3.2 Semi-parametric Sample Selection Model

Newey et al. (1990) and Newey (2009) suggest a semi-parametric version of the Heckman model that relaxes the assumption of joint normality of the error terms in the first and second stage regressions. Instead, they propose a two-step estimator in which "a non-parametric approximation to [the sample selection correction] is used in the second-step regression rather than the inverse Mills ratio."[13] Newey (2009) suggests a polynomial or spline of the predicted probabilities of sample inclusion as the non-parametric approximation to the inverse Mills ratio.

3.3 Identification

As I discussed in greater detail above, Lacetera et al. (2016) find that that auctioneers are not identical in their abilities, and that as a consequence some have higher conversion rates than others. I argue that the (conditionally) random assignment of cars to auctioneers is a plausible instrument for the probability of sale. That is, that there is an effect of auctioneer ability on a car's inclusion in the sample, but no direct effect of the auctioneer on the sale price.

Lacetera et al. do not analyze the effect of auctioneer ability on sale price directly, but instead consider the "residual price", the difference between the sale price and the auction house's estimated blue-book price. They find a small positive effect of auctioneer ability on residual price – $41.80 compared to the mean sale price in their sample of $15,141. In percentage terms, this effect is an order of magnitude smaller than the auctioneer's effect on the probability of sale (0.28% compared to 4.3%).

I use two alternative measures of the auctioneer effects as instruments to identify the model: the raw conversion rates calculated from the data and the auctioneer effects recovered

[13] Newey (2009, p. 219)

from a regression of car and auction characteristics on whether or not a car sells.

$$Y_{ik} = \alpha + \beta_k + X_i\gamma + \epsilon_{ik} \tag{5}$$

Where i indexes individual cars and k indexes auctioneers. X_i is a vector of car characteristics and auction environment fixed effects. The β_ks are the auctioneer effects, the estimates of interest. In Equation 5, I use Lacetera et al.'s preferred specification, which controls for seller, day of the week, time of day, lane, make×model×age×body, and mileage. Of the two measures, I argue that the recovered fixed effects are a better exclusion restriction because they are a cleaner estimate of auctioneers' true influence on the probability of sale.

4 Auction Environment and Data Description

The entire dataset consists of 55 million observations across seven years, 2002-2008. I observe rich details about each car that goes up for auction: make, model, model year, and body type (e.g. 2000 Toyota Camry LE). I also observe important details about the auction environment and auction results, including the auction date (which allows me to calculate the age of the car when it is sold), auction location, the specific lane at the auction house in which it is run, the unique auctioneer identifier, as well as the reserve price, the high bid, and the sale price.

Throughout my analysis, I will restrict my focus to dealer-sellers. The cars sold by dealers are typically trade-ins that the dealer does not want to sell themselves. Another large segment of the wholesale used car market are fleet/lease sellers: rental car companies, leasing companies and corporate fleets. These sellers want to get the cars off their hands; they set low reserve prices (or do not set them at all), and consequently sell almost all of their cars. Additionally, fleet/lease sellers frequently bring their own auctioneers to the auction house.

I also drop the first two years of data, 2002 and 2003, because they are missing the unique auctioneer identifiers. In addition, I drop outliers on the observable variables: cars that sell for over $75,000 or less than $100, cars with over 250,000 miles, and cars that were over 25 years old when they sold. Additionally, I dropped trailers, boats, and other recreational vehicles that weren't consumer cars or trucks. After the data cleaning process about 20 million observations remain, spread across 5 years.

The auctions are oral, ascending-price auctions, or English auctions.[14] On average, the auctions last for about one minute. Each auction house conducts auctions multiple times a week, and at each auction there are multiple lanes of cars. Bidders are free to wander from lane to lane. Although I do not observe the number of bidders or the bid history in the data, Genesove (1995) observes that there are typically between five and ten bidders at each auction lane, and Lacetera et al. reaffirm this observation.

4.1 Summary Statistics

Table 1 displays summary statistics for my sample of used cars, broken down by whether or not the cars successfully sell at auction. 51.8% of cars in my sample sell at auction. The average sale price for cars that do sell is just over $8,000. The average car that goes up for auction is about 5.5 years old and has around 80 thousand miles on the odometer. Interestingly, cars that sell are, on average, slightly older and have more mileage than cars that do not sell at auction, although there is significantly more variance in the age of cars that do not sell at auction. The fact that cars that do sell tend to be older may reflect the adverse selection problems described by Akerlof (1970).

[14] The English auction is strategically equivalent to a second-price sealed bid auction, where the good goes to the bidder that values the item the most but the winner only pays the valuation of the second-highest bidder

Table 1: Summary Statistics by Sale Status

	Unsold	Sold	Total
Share of Cars Sold	–	–	0.518
Sale Price	0	$8,284.15	$4,289.63
		($7404.9)	($6747.4)
Age	5.28	6.01	5.66
	(10.85)	(4.11)	(8.10)
Miles	76.49	83.49	80.11
(Thousands)	(46.77)	(47.18)	(47.12)

Note: Standard deviations in parentheses

4.2 Auctioneer Cutoff

It is important to choose a cutoff value for auctioneers by the number of auctions they have conducted. There is much higher variance in the conversion rate and estimated auctioneer effects among auctioneers with a lower number of auctions conducted. Lacetera et al. choose to restrict their sample to auctioneers who have conducted at least 5,000 auctions. I explore different potential cutoff values and their effects on both the sample size and the distribution of auctioneer effects and conversion rates. Table 2 reports the effect of different potential cutoff values on the number of observations left in the sample. Figure 5 displays the distribution of conversion rates at different cutoff values; as the cutoff increases, the mass points at 0 and 1 disappear and the conversion rates appear to be normally distributed around the mean sample value. Figure 6 displays the correlation between the conversion rates and estimated auctioneer effects; as the cutoff increases the correlation weakens as the auctioneers with extreme values for either measure are dropped from the sample. I follow Lacetera et al. (2016) and restrict my sample to auctioneers who have at least 5,000 auction observations.

Table 2: Auctioneer Cutoffs

Cutoff	CDF Value	Observations Omitted	Observations Left
0	0	0	19,989,986
100	8e-04	15,575	19,974,411
250	0.0024	47,414	19,942,572
500	0.0048	95,043	19,894,943
1000	0.0095	190,204	19,799,782
5000	0.0695	1,389,372	18,600,614

5 Results and Discussion

5.1 Estimation

Due to the size of the data and the high-dimensional fixed effects, I cannot feasibly estimate the sample selection models using the full maximum-likelihood method. Instead, I implemented both the parametric estimator presented in Heckman (1979) and the semi-parametric model from Newey (2009) using the `lfe` and `bife` packages in R. These packages implement pseudeo-demeaning algorithms to estimate linear and binary response models with high-dimensional fixed effects in a computationally efficient manner.[15] This method is computationally feasible and produces consistent point estimates and predicted values.

For each level of fixed effects, I estimated OLS, Heckman, and semi-parametric sample selection hedonic regressions of the sale price of used cars on a quadratic of the car's age and a cubic of the car's odometer mileage. I estimated the following OLS regression:

$$\hat{Y}_i = \beta_0 + \beta_1(Age_i) + \beta_2(Age_i^2) + \beta_3(Miles_i) + \beta_4(Miles_i^2) + \beta_5(Miles_i^3) + X\delta + u_i \qquad (6)$$

Where $Miles$ is the odometer mileage in 1000's of miles, and X is the vector of fixed effects for each specification. For the first stage of both the Heckman estimator and the semi-parametric

[15] Gaure (2013) and Stammann et al. (2016), respectively

estimator, I estimate:

$$\begin{aligned}P(S=1)_i = \Phi(\gamma_0 + \gamma_1(Age_i) + \gamma_2(Age_i^2) + \gamma_3(Miles_i) + \gamma_4(Miles_i^2) + \gamma_5(Miles_i^3) \\ + \gamma_6(\alpha_k) + X\delta + \nu_i \quad (7)\end{aligned}$$

Where α_k are the auctioneer effects recovered from Equation 5.

In the second step of the Heckman model, I estimate the inverse Mills ratio $\left(\frac{\phi}{\Phi}\right)$ at the linear predictions of the probit model, then estimate the following OLS regression:

$$\hat{Y}_i = \beta_0 + \beta_1(Age_i) + \beta_2(Age_i^2) + \beta_3(Miles_i) + \beta_4(Miles_i^2) + \beta_5(Miles_i^3) + \beta_6(\hat{\lambda}_i) + X'\delta + u_i \quad (8)$$

For the second stage of the semi-parametric model, I estimate the following OLS regression:

$$\begin{aligned}\hat{Y}_i = \beta_0 + \beta_1(Age_i) + \beta_2(Age_i^2) + \beta_3(Miles_i) + \beta_4(Miles_i^2) + \beta_5(Miles_i^3) \\ + poly(Z'\hat{\gamma}, 3)\eta + X'\delta + u_i \quad (9)\end{aligned}$$

Where $poly(Z'\hat{\gamma}, 3)$ is a 3rd degree orthogonal polynomial of the predictions from Equation 6, the first stage regression.[16]

5.2 Results

Table 3 presents the estimates from the baseline OLS regressions as well as the both of the sample selection-corrected models. Model 1 has no fixed effects, Model 2 has make and model fixed effects, Model 3 has make, model, and body type fixed effects, and Model 4 has make, model, body type and model year fixed effects. Controlling for model year in addition to age captures the fact that a five year old Toyota Camry sold in 2002 is not the same as a five

[16] I also tried higher-degree polynomials as well as polynomial splines as the non-parametric approximation of the inverse Mills ratio, but the results were not substantively different.

year old Toyota Camry sold in 2007. Adding finer degrees of fixed effects adds substantially more explanatory power to the the model; the adjusted R^2 of Model 4 is 0.94, compared to 0.51 in the Model 1 with no fixed effects.

Table 5 in the Appendix presents the results from the first stage probit regressions for all four models. As expected, the auctioneer effect has a consistently strong and positive effect on the predicted probability of sample inclusion across all four models. For the first model with no fixed effects, the average marginal effect of a standard deviation increase in auctioneer ability is 4.12 percentage points; the average marginal effect of auctioneer ability is 4.25 percentage points in Model 4 with the comprehensive fixed effects. The magnitude and robustness of the effect suggest that auctioneer ability is a strong instrument. These predictions closely match the results of Lacetera et al. (2016).

All of the models use robust standard errors clustered around the unit of fixed effects. That is, the standard errors in Model 1 are not clustered, the standard errors in Model 2 are clustered at the make-model level, and so on. Across all models, the coefficient estimates for all of the age and miles terms are highly statistically significant, which is unsurprising given the large number of observations in my dataset.

Substantively, the coefficient estimates suggest that both age and miles have negative effects on the sale price, which makes intuitive sense. All else equal, people generally prefer newer cars with less miles on them. The marginal effect of both age and miles on the sale price are decreasing in the level of age and miles. Again, this agrees with my *a priori* expectations; the difference between a eight-year-old car and a nine-year-old car is smaller than the difference between and one-year-old car and a two-year-old car. This broad pattern holds across all specifications, although the individual point estimates vary somewhat between models.

My preferred specification is the semi-parametric sample selection model. As I described in subsection 3.2, the semi-parametric estimator does not require the assumption of joint normality of the two error terms as in the Heckman model. Additionally, it allows the correction term to enter the hedonic regression in a more flexible manner. This can be seen from the marginal improvement in the adjusted R^2 when comparing the Heckit and semi-parametric estimates in for Model 1. For Models 2-4, the differences between the Heckit and semi-parametric estimates are inconsequential.

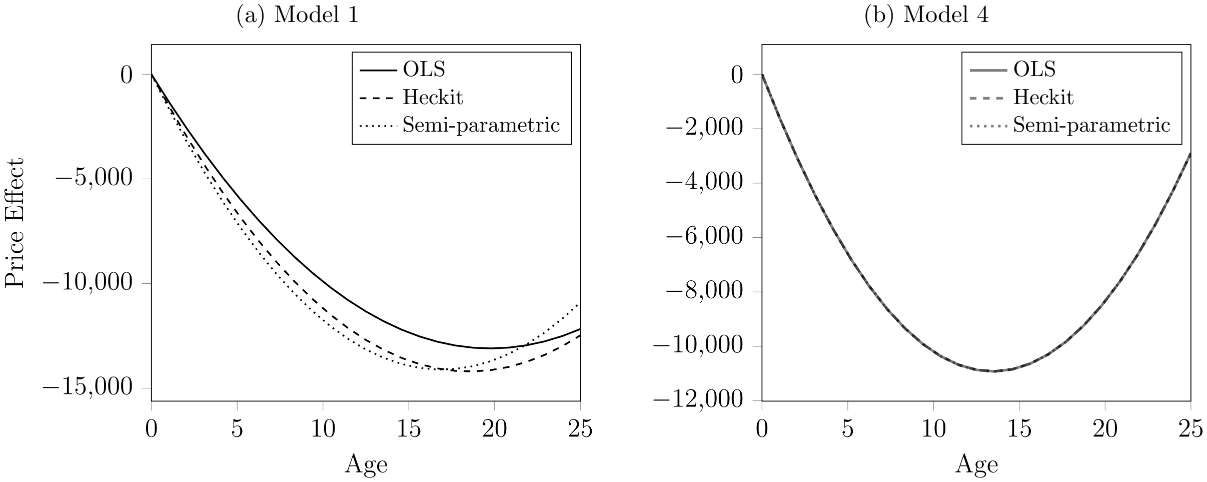
5.3 Discussion

I used several different methods to assess whether or not there is sample selection bias in each of the four models. The first and most straightforward test is to check whether the coefficients on the selection correction are statistically different from zero.[17] Each of the models pass this test; $\hat{\lambda}$ is highly statistically significant in all four of the Heckit models, and the coefficients on the all of the polynomial terms are also highly significant. Additionally, in each case a Wald test comparing the unrestricted (selection corrected) and restricted (OLS) models also strongly rejects the null hypothesis that excluding the selection correction from the model does not impact the explanatory power of the model. Both of these tests suggest that there is evidence of sample selection bias in all four models.

However, the statistical significance of the correction term(s) and the Wald tests may be more a reflection of the sample size than the true extent of the sample selection problem. In Model 1, there are clearly substantive differences between the estimated coefficients for both age and miles. However, in Models 2-4, the coefficient estimates between the OLS and both sample selection models hardly differ at all.

Figure 3 plots the predicted effect of age on sale price for Model 1 and Model 4. There is a substantial difference in the predictions between the OLS, Heckit, and semi-parametric models for the first specification with no fixed effects, but the curves overlap almost perfectly

[17] Wooldridge (2012)


(a) Model 1

(b) Model 4

Figure 3: Effect of Selection Bias on Age Coefficients

for the specification with the most controls. These plots illustrate visually how the sample selection bias disappears after adding the fixed effects for car characteristics.

To assess the practical impact of the sample selection problem on the predictions of the hedonic pricing model, I compared the predicted values of the OLS and semi-parametric models for each of the four specifications using two measures: the correlation between the predicted values of the two models, and the mean of the absolute value of the pairwise differences of the predicted values. The correlation between the predicted values for the OLS and semi-parametric sample selection models was very high across all specifications, but the correlation coefficient increased as I included finer and finer fixed effects.

Equation 10 gives the formula for the mean absolute pairwise difference in predicted values.

$$\frac{1}{N}\sum_{i=1}^{N} |\hat{y}_{OLS_i} - \hat{y}_{SP_i}| \tag{10}$$

Where $\hat{y}_{OLS}$ are the predicted values from the OLS model and $\hat{y}_{SP}$ are the predicted values form the semi-parametric model. The average difference in predicted values between the OLS and semi-parametric estimations for Model 1 is more than \$185. The average difference falls to about $10 for each of Models 2-4. Figure 4 plots the the statistic for each of the four models.

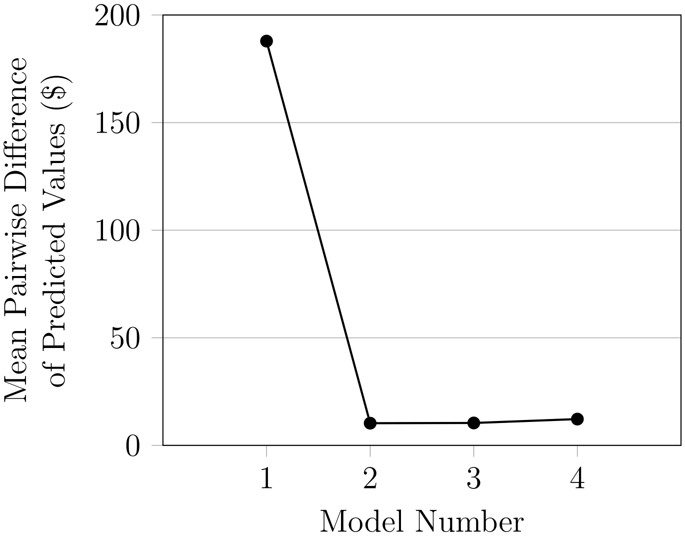


Figure 4: Mean Absolute Pairwise Difference of Predictions by Model

The average difference in predicted values of $185 is economically significant; it is about 2.25% of the mean sale price of cars in the sample. However, the substantive differences in predicted values disappear when I add fixed effects for make and model, and then remain the about same for the other specifications. This suggests that controlling for make and model is sufficient to eliminate the practical consequences of the sample selection bias. However, controlling for finer-degree fixed effects is still advisable if the goal is to maximize predictive accuracy, as they continue to increase the adjusted R^2 of the model.

Table 3: OLS and Heckit Model Results[†]

	Model 1			Model 2			Model 3			Model 4		
	OLS	Heckit	Semi-parametric	OLS	Heckit	Semi-parametric	OLS	Heckit	Semi-parametric	OLS	Heckit	Semi-parametric
Age	−1,326.419*** (1.471)	−1,531.749*** (2.198)	−1,668.178*** (2.391)	−1,317.914*** (58.054)	−1,317.851*** (58.053)	−1,317.848*** (58.053)	−1,359.115*** (37.555)	−1,359.038*** (37.553)	−1,359.035*** (37.553)	−1,622.897*** (22.492)	−1,622.681*** (22.489)	−1,622.660*** (22.489)
Age^2	33.577*** (0.070)	41.307*** (0.093)	49.293*** (0.108)	32.573*** (2.469)	32.572*** (2.469)	32.572*** (2.469)	40.421*** (1.529)	40.418*** (1.528)	40.418*** (1.528)	60.289*** (1.073)	60.277*** (1.073)	60.276*** (1.073)
Miles	−155.351*** (0.304)	−148.597*** (0.308)	−144.059*** (0.311)	−140.230*** (8.043)	−140.231*** (8.043)	−140.231*** (8.043)	−126.132*** (3.833)	−126.134*** (3.833)	−126.134*** (3.833)	−115.420*** (1.849)	−115.415*** (1.849)	−115.413*** (1.849)
Miles^2	0.761*** (0.003)	0.719*** (0.003)	0.693*** (0.003)	0.748*** (0.068)	0.748*** (0.068)	0.748*** (0.068)	0.651*** (0.033)	0.651*** (0.033)	0.651*** (0.033)	0.551*** (0.016)	0.551*** (0.016)	0.551*** (0.016)
Miles^3	−0.001*** (0.00001)	−0.001*** (0.00001)	−0.001*** (0.00001)	−0.001*** (0.0002)	−0.001*** (0.0002)	−0.001*** (0.0002)	−0.001*** (0.0001)	−0.001*** (0.0001)	−0.001*** (0.0001)	−0.001*** (0.00004)	−0.001*** (0.00004)	−0.001*** (0.00004)
$\hat{\lambda}$		−3,152.552*** (25.081)			−74.031*** (10.241)			−70.406*** (5.408)			−73.036*** (3.576)	
$poly(Z'\hat{\gamma}, 3)$			✓			✓			✓			✓
Constant	21,798.630*** (7.534)	24,729.280*** (24.549)	22,537.260*** (8.840)									
Fixed Effects		None			make×model			make×model×body			make×model×body×model year	
Observations	9,594,894	9,594,894	9,594,894	9,594,894	9,594,894	9,594,894	9,594,894	9,594,894	9,594,894	9,594,894	9,594,894	9,594,894
Adjusted R^2	0.512	0.513	0.514	0.867	0.867	0.867	0.919	0.919	0.919	0.941	0.941	0.941

Note: Clustered standard errors in parentheses *p<0.1; **p<0.05; ***p<0.01

[†] Models 2-4 do not report a constant because it is lost in the fixed-effects demeaning algorithm.

6 Conclusion

In this paper, I explore the extent of sample selection bias in the wholesale used auto market, as well as the larger implications for the ability for researchers to use big data to overcome sample selection bias. Using a dataset of about 20 million unique wholesale used car auctions across five years, I build a hedonic pricing model that predicts the sale price of used cars as a function of observable characteristics: age, mileage, make, model, and body type. I test whether the incidental truncation of the dependent variable, sale price, biases the coefficient estimates of the hedonic pricing model.

I use a two step sample-selection model to correct for the incidental truncation of the dependent variable. I identify a plausible exclusion restriction: the ability or skill of the auctioneer to which the car is randomly assigned at auction. In the baseline model with no fixed effects, I find evidence for substantial sample selection bias. On average, the predictions of the naïve model differ from the sample selection-corrected model by $185; this difference is small but economically meaningful compared to the average sale price of just over $8,000

Adding finer fixed effects to the model eliminates the practical effects of the sample selection bias, even while statistical tests strongly reject the null hypothesis that there is no sample selection bias, which is likely a consequence of the large sample size. In the model with the largest set of controls, the average difference in predicted prices is just $10.

My results suggest that controlling for sufficiently fine-grained fixed effects can eliminate the practical consequences of unobserved correlation between a car's probability of sale and its sale price; the naïve model performs just as well as the bias corrected model, without the need to identify a plausible exclusion restriction. However, it is impossible to know whether you have a problem with sample selection if you do not have a valid instrument, so in practice the best course of action is to control for as many factors as possible. While not a replacement for lab or field experiments, big datasets that allow researchers to control for fine details can potentially provide researchers with another useful tool for eliminating sample selection bias from analyses of observational data.

References

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics 84*(3), 488–500.

Cox Automotive (2018). Cox Automotive 2018 Used Car Market Report & Outlook. Technical report, Cox Automotive.

Cunningham, S. (2018). *Causal Inference: The Mixtape* (1.7 ed.). http://scunning.com/cunningham_mixtape.pdf.

Einav, L. and J. Levin (2014). Economics in the age of big data. *Science 346*(6210), 1243089.

Gaure, S. (2013). OLS with multiple high dimensional category variables. *Computational Statistics & Data Analysis 66*, 8–18.

Genesove, D. (1995). Search at wholesale auto auctions. *The Quarterly Journal of Economics 110*(1), 23–49.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica 47*(1), 153–161.

Lacetera, N., B. J. Larsen, D. G. Pope, and J. R. Sydnor (2016). Bid Takers or Market Makers? The Effect of Auctioneers on Auction Outcome. *American Economic Journal: Microeconomics 8*(4), 195–229.

Lacetera, N. and J. Sydnor (2014). Would You Buy a Honda Made in the United States? The Impact of Production Location on Manufacturing Quality. *The Review of Economics and Statistics 97*(4), 855–876.

Larsen, B. (2018). The Efficiency of Real-World Bargaining: Evidence from Wholesale Used-Auto Auctions. *NBER Working Paper*, 1–83.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal 12*(S1), S217–S229.

Newey, W. K., J. L. Powell, and J. R. Walker (1990). Semiparametric Estimation of Selection Models: Some Empirical Results. *The American Economic Review 80*(2), 324–328.

Stammann, A., F. Heiss, and D. McFadden (2016). Estimating Fixed Effects Logit Models with Large Panel Data. Kiel und Hamburg: ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft.

Tadelis, S. and F. Zettelmeyer (2015). Information Disclosure as a Matching Mechanism: Theory and Evidence from a Field Experiment. *American Economic Review 105*(2), 886–905.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives 28*(2), 3–28.

Wooldridge, J. M. (2012). *Introductory Econometrics* (5th ed.). South-Western.

Image Appendix

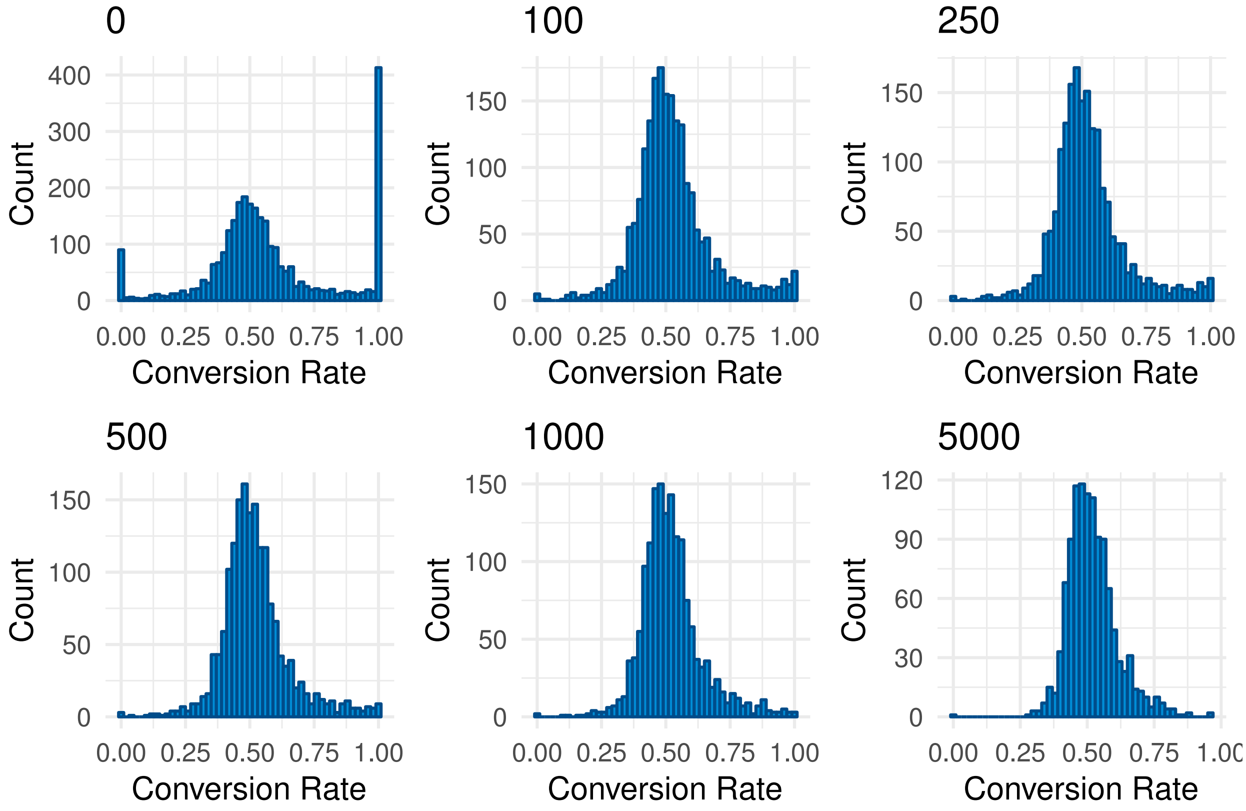


Figure 5: Distribution of Conversion Rates by Auctioneer Cutoff

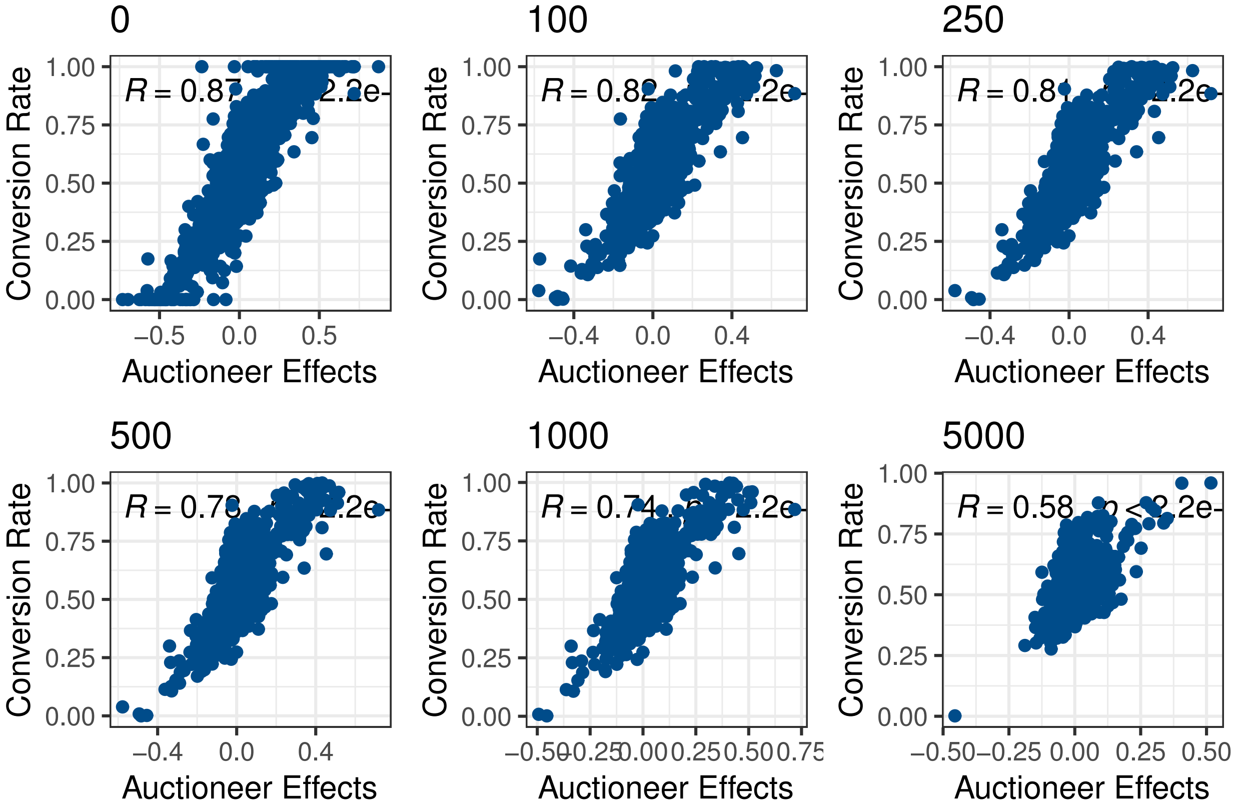


Figure 6: Correlation between Conversion Rate and Auctioneer Effect by Auctioneer Cutoff

Table Appendix

Table 4: Baseline OLS Regression Results

	Model 1	Model 2	Model 3	Model 4
Age	−1,218.045***	−1,281.933***	−1,305.374***	−1,621.759***
	(1.266)	(54.287)	(36.831)	(22.463)
Age^2	27.532***	30.377***	36.902***	60.166***
	(0.055)	(2.030)	(1.510)	(1.069)
Miles	−162.353***	−142.472***	−129.225***	−115.420***
	(0.300)	(7.824)	(3.952)	(1.851)
Miles^2	0.793***	0.761***	0.672***	0.551***
	(0.003)	(0.066)	(0.034)	(0.016)
Miles^3	−0.001***	−0.001***	−0.001***	−0.001***
	(0.00001)	(0.0002)	(0.0001)	(0.00004)
Constant	21,777.620***			
	(7.544)			
Fixed Effects	None	make×model	make×model×body	make×model×body×model year
Observations	9,616,977	9,616,977	9,616,977	9,616,977
Adjusted R^2	0.510	0.867	0.918	0.941

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 5: First Stage Probit Regression Results

	Model 1	Model 2	Model 3	Model 4
Age	0.099^{***}	0.120^{***}	0.138^{***}	0.118^{***}
	(0.0003)	(0.0003)	(0.0004)	(0.0006)
Age^2	-0.004^{***}	-0.0045^{***}	-0.005^{***}	-0.004^{***}
	(0.00001)	(0.00002)	(0.00002)	(0.00004)
Miles	-0.003^{***}	-0.004^{***}	-0.005^{***}	-0.006^{***}
	(0.0001)	(0.00006)	(0.00006)	(0.00006)
Miles^2	0.00002^{***}	0.00002^{***}	0.00003^{***}	0.00003^{***}
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Miles^3	-0.00000^{***}	-0.00000^{***}	-0.00000^{***}	-0.00000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Auctioneer Effect	1.678^{***}	1.666^{***}	1.649^{***}	1.776^{***}
	(0.005)	(0.005)	(0.005)	(0.005)
Fixed Effects	None	make×model	make×model×body	make×model×body×model year
Observations	18,526,965	18,526,965	18,526,965	18,526,965

Note: *p<0.1; **p<0.05; ***p<0.01