Validate Structural Analysis

Testing GPV Method in First-Price Reverse Auctions

Qingbo Liu

Advisor: Timothy P. Hubbard

December 6, 2020

1 Introduction

Participants at auctions submit bids for items based on their valuations and preferences of the items. There are many ways to think about the transformation from valuations and preferences into bids, and the usual economic approach is to consider the bids as a mathematical function of valuations and preferences, an approach which this paper will take. At the end of auctions, the winner(s) is designated based on certain auction rules and the auction data becomes available. The data includes only bids, the final output of the bidding function, and the inputs, valuations and preferences, remain unknown. Yet interesting and useful analysis of auctions entails the knowledge about valuations and preferences. For example, to evaluate the performance of auctions, it is necessary to have access to bidder's valuations to determine if the particular mechanism has achieved its objective, which leads to structural estimation.

Structural estimation assumes a particular model and uses statistical methods to estimate relevant parameters of the model. The calibrated model can then give predictions regarding objects that are being investigated. For example, in first-price sealed-bid reverse auctions, one of the mechanisms that the government uses for public project procurements, the simple assumption is that all participants independently draw costs from an identical cost distribution. Each bidder submits a sealed bid to the auctioneer and after all bidders have finished submission the winner is the one with the lowest bid. With this bidding model, structural estimation can be used to infer the cost distribution based on observed bids (Guerre, Perrigne, & Vuong, 2000). Compared to regression analysis, structural estimation imposes a stringer set of assumptions on the data. But more than just discovering the correlation between variables as regression analysis does, structural estimation is able to produce stronger counter-factual predictions, such as the cost distribution as mentioned above.

The central assumption underlying structural estimation is that the model used is correct in a particular environment. To overcome such a drawback, alternative studies such as lab experiments are used to confirm the theory's validity. But laboratory environments ultimately differ from the real environments and the conclusions from laboratory studies may not hold in more realistic environments. The goal of this paper is to test model validity in a new different way by borrowing the techniques from statistical learning literature. In statistical learning, multiple models for predictive purposes are trained on part of the dataset and their performance is evaluated on another reserved split of the available dataset. The technique is called cross-validation and is used when there are candidate models and their performance needs to be compared to select the fittest one. In this paper, I will adapt the idea of cross-validation to test the predictiveness of economic models. By splitting data into training data and test data as in statistical learning, I will run structural estimation on the training data, simulate test data using the results from strucutral estimation, and compare the simulated data to actual test data. If the difference is statistically significant, that suggests unfitness of the model in the particular environment. In this paper, I will use data from government procurement auctions to test the first-price sealed-bid equilibrium model (Riley & Samuelson, 1981) for structural estimation [1].

[1] The original goal of this paper also includes comparing the cost distribution of DBE/MBE/WBE (standing respectively for Disadvantaged/Minority/Woman Business Enterprise) to the cost distribution of other firms, which will add to the existing literature on understanding how DBE/MBE/WBE perform relative to other firms. But due to data

The remainder of the paper is organized as follows. Section 2 discusses the characteristics of the environment and derives the model for structural estimation. Section 3 examines the data and presents estimation results. Section 4 runs simulations based on estimation. Section 5 concludes with a discussion on the results and possible improvements.

2 Model

For the model setup, suppose there are n bidders in the auction, with $n \geq 2$, and there is a single, indivisible object being auctioned. Each bidder submits a sealed bid based on an observed cost before the auction. The bidder's cost c is drawn i.i.d from a common distribution $F(\cdot)$ which is continuous with density $f(\cdot)$ on the support $[\underline{c}, \overline{c}] \subset \mathbb{R}^+$. All bidders are quasi-linear utility-maximizers, with the utility function $u(c, b) = b - c$. Bidders are therefore assumed to be risk-neutral. The number of bidders n, the cost distribution F and its support are common knowledge to all bidders in the auction. Besides, each bidder i can only observe her private cost c_i and no one else's.

Different from first-price sealed-bid auctions where the highest bidder wins, in first-price sealed-bid reverse auctions the winner is the one who submits the lowest bid. In the following subsection, I will derive the Bayes-Nash Equilibrium in the FPSB reverse auction environment. After I have characterized the equilibrium, now I will proceed to describe how private costs can be inferred from observed bids based on the equilibrium.

2.1 Bayes-Nash Equilibrium

For bidder i in the auction, she needs to report a bid b_i based on her private cost c_i. The bidder achieves maximal utility by maximizing the function

$$\max_{b_i}(b_i - c_i)P(\text{win}|b_i),$$

availability on these types of firms and time constraints on the project as this is a term paper, this goal is omitted.

which is saying by picking an appropriate bid b_i, the bidder has a good balance between the revenue and probability of winning.

Since this is a first-price reverse auction, where the lowest bidder wins, the probability of winning is

$$\begin{aligned}
P(\text{win}|b_i) &= P(b_1 \geq b_i, ...b_{i-1} \geq b_i, b_{i+1} \geq b_i, ...b_n \geq b_i) \\
&= P(b_1 \geq b_i)...P(b_{i-1} \geq b_i)P(b_{i+1} \geq b_i)...P(b_n \geq b_i) \\
&= [1 - P(b_1 \leq b_i)]...[1 - P(b_{i-1} \leq b_i)][1 - P(b_{i+1} \leq b_i)]...[1 - P(b_n \leq b_i)] \\
&= \prod_{j=1, j\neq i}^{N} [1 - P(b_j \leq b_i)]
\end{aligned}$$

Now, suppose that there is a monotone bidding equilibrium function $s(\cdot)$ such that all bidders abide by it and makes their bids based on it, i.e. $b_i = s(c_i)$. Then

$$\begin{aligned}
P(\text{win}|b_i) &= \prod_{j=1, j\neq i}^{N} [1 - P(s(c_j) \leq b_i)] \\
&= \prod_{j=1, j\neq i}^{N} [1 - P(c_j \leq s^{-1}(b_i))] \\
&= \prod_{j=1, j\neq i}^{N} [1 - F(s^{-1}(b_i))] \\
&= [1 - F(s^{-1}(b_i))]^{n-1},
\end{aligned}$$

where the last equality follows from symmetry – all bidders draw costs (independently) from the same distribution.

The objective function therefore becomes

$$\max_{b_i} (b_i - c_i)[1 - F(s^{-1}(b_i))]^{n-1}$$

and setting the derivative of it with respect to b_i equal to 0 yields

$$\begin{aligned}[1-F(s^{-1}(b_i))]^{n-1} &= (b_i-c_i)(n-1)[1-F(s^{-1}(b_i))]^{n-2}f(s^{-1}(b_i))\frac{ds^{-1}(b_i)}{db_i}\\ [1-F(c_i)]^{n-1} &= (s(c_i)-c_i)(n-1)[1-F(c_i)]^{n-2}f(c_i)\frac{1}{s'(c_i)}\\ 1 &= (s(c_i)-v_i)(n-1)\frac{f(c_i)}{1-F(c_i)}\frac{1}{s'(c_i)} \qquad (1)\end{aligned}$$

The differential equation characterizes the strategy in sealed-bid first-price reverse auction and the solution of it, $s(\cdot)$, gives the equilibrium bidding function as follows (Hubbard & Paarsch, 2009)

$$s(c) = b = c + \frac{\int_c^{\bar{c}}[1-F(u)]^{n-1}du}{[1-F(c)]^{n-1}} \qquad (2)$$

2.2 Cost Inference

Our interest, however, lies not in the particular solution of this differential equation, but in using it as an intermediate step in deriving cost as a function of bids, i.e. $s^{-1}(\cdot)$.

Assume $G(\cdot)$ represents the cumulative distribution of bids, along with the continuous density function $g(\cdot)$ and support $[\underline{b}, \bar{b}] = [s(\underline{c}), s(\bar{c})]$. As in Guerre et al. (2000), by a transformation of random variables, I have the identity $G(b) = P(\tilde{b} < b) = P(\tilde{c} < s^{-1}(b)) = F(s^{-1}(b)) = F(c)$ and $g(b) = f(c)/s'(c)$. The ratio of $g(b)/(1-G(b))$ gives $f(c_i)/(1-F(c_i))(1/s'(c_i))$. From this, equation (1) becomes

$$c_i = b_i - \frac{1}{n-1}\frac{g(b_i)}{1-G(b_i)} \qquad (3)$$

In an environment of n bidders, Equation (3) provides the method to estimate pseudo-costs from observed bids with kernel-estimated bid density and bid distribution. The estimated costs are pseudo because of assumptions that bidders are risk-neutral and play a Bayes-Nash equilibrium bidding strategy.

2.3 Estimation

The estimation consists of two steps. In step 1, I recover costs c based on observed bids b, by first non-parametrically estimating $G(b)$ and $g(b)$ and then using equation (3) to generate pseudo-costs. In the second step, I can non-parametrically estimate the cost density $f(c)$ from pseudo-costs.

The technique used to estimate $G(b)$ and $g(b)$, as well as $f(c)$, is called kernel density estimation and has the following form

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}), \tag{4}$$

and

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x), \tag{5}$$

where I have a set of observables $x_1, x_2, ...x_n$. However, this kernel density estimation (4) method suffers from boundary problem and instead, I will use the boundary-corrected kernel estimation method as proposed in (Hickman & Hubbard, 2015)

This two-step estimation method has the advantage that in order to estimate cost distribution with observed bids, I do not need to assume a prior distribution of $f(c)$ to estimate it (which is a parametric method that works by first assuming the shape of a distribution and then estimating parameters of the distribution from data).

3 Data and Results

The bidding data used in this paper is downloaded from the Colorado Department of Transportation (https://www.codot.gov/business/bidding). I will take a short detour of the data itself before presenting results on cost estimation.

	Mean	Median	Standard Deviation	Max	Min
Total Bid	3,172,164.68	1,366,571.20	5,491,805.02	93,398,000.00	0.00
Percent of Engineer's Estimate (Cost)	106.54	100.02	22.63	750.00	0.00
Number of Bidders	5.47	5	2.5	16	1

Table 1: Basic summary statistics of data

3.1 Data Description

There are in total 1182 auctions in the data. For each auction, I have the number of participants n, bids submitted by participants and the engineer's estimate. The summary statistics of the data are summarized in Table 1. Projects vary from each other in their sizes, types, and many other factors. Even the same type of projects will have different costs. As Table 1 shows, the variance in Total Bid is very large and not an ideal candidate for cost estimates. Instead, I will work with Percent of Engineer's Estimate to infer cost distribution. Percent of Engineer's Estimate is Total Bid divided by Engineer's Estimate, which is also known as normalized bids. Inferred cost distribution constructed this way will be invariant across projects.

Figure 1 plots the frequency of normalized bids. Most bids are centered around 110%, which is slightly higher than the engineer's estimate with respect to which bids are normalized. However, there are some outliers located in the *More* bin, which would cause numeric instability in estimation. For data that is More bins, the kernel-estimated bid distribution $G_B(\cdot)$ is close to 1 and so $1 - G_B(\cdot)$ close to 0. As in the right-hand side of equation (3), the part $\frac{g(b)}{1-G_B b}$ will tend to infinity as $1 - G_B(b)$ approaches zero, resulting in the estimated cost to be negative. The problem will persist over the boundary of the data and is inherent in the kernel distribution estimation. The boundary-correction technique introduced in (Hickman & Hubbard, 2015) applies only to the kernel density estimation, not kernel distribution estimation, and overcoming the boundary problems in kernel distribution estimation is outside the scope of this paper. For simplicity, therefore, I will first use equation (3) to estimate costs and then trim the

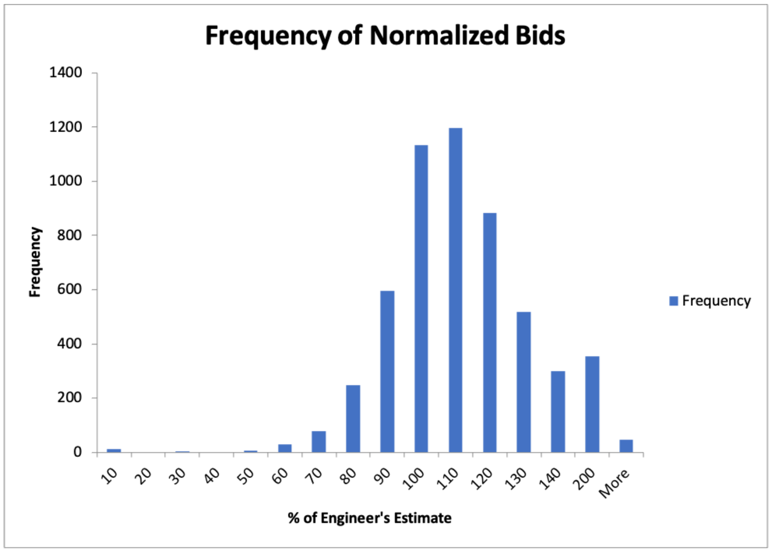

Figure 1: Frequency of normalized bids (bids divided by engineer's estimate)

negative costs.

3.2 Results

With data comprised of auctions that have a different number of participants, the general estimation procedure remains the same. But the equation (3) that assumes a fixed n for the number of participants is modified to accommodate varying n (Gentry, Hubbard, McComb, & Schiller, 2018)

$$c_{in} = b_{in} - \frac{1}{n-1} \cdot \frac{g(b_{in}|N = n)}{1 - G_B(b_{in}|N = n)}, \tag{6}$$

where c_{in} and b_{in} refer to the cost and bid in i-th auction with n participants, and the density $g(\cdot|N = n)$ and distribution $G(\cdot|N = n)$ are estimated on auction data with n participants for each n in the dataset.

Figure 2 plots the estimated cost against bids in the plot. The relationship between costs and bids is almost linear, suggesting that for each bid, the inferred cost is very close in terms of numeric values. This is so mainly due to the bid distribution. As shown in Figure 1, most of the bids are spread

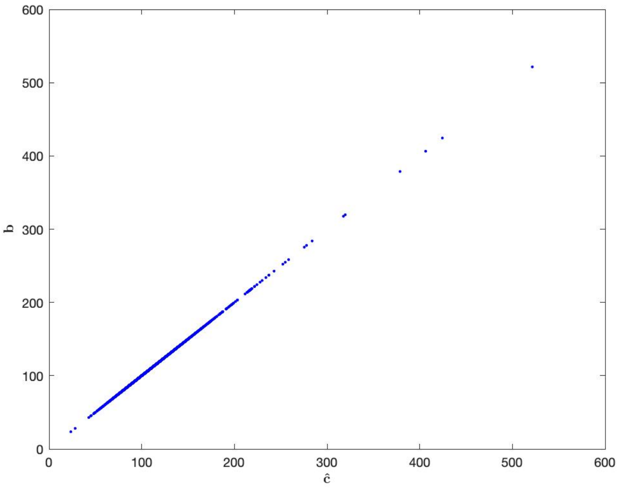

Figure 2: Estimated costs plotted against bids.

in the range [70, 200], which means that for each bid b, the estimated density value at the specific bid value, $f_c(b)$, is small: the peak of $f_c(\cdot)$ being 0.0222. A closer look at the bid distributions of auctions with a different number of participants, n, reveals that the majority of auction data with small n, where the competition is less fierce and bidders are able to have a larger markup over the costs than those having more competitors, bids are mostly distributed around 110%. The auctions that more spread-out data, however, have large n, in which case the markup over costs is decreased due to the competition.

Figure 3 shows the kernel-estimated density and distribution of pseudo-costs. Because pseudo-costs are close to corresponding bids, the density figure on the left of Figure 3 is very similar to the histogram of bids, Figure 1. As both plots in Figure 3 show, the inferred costs also center around the range [70, 200].

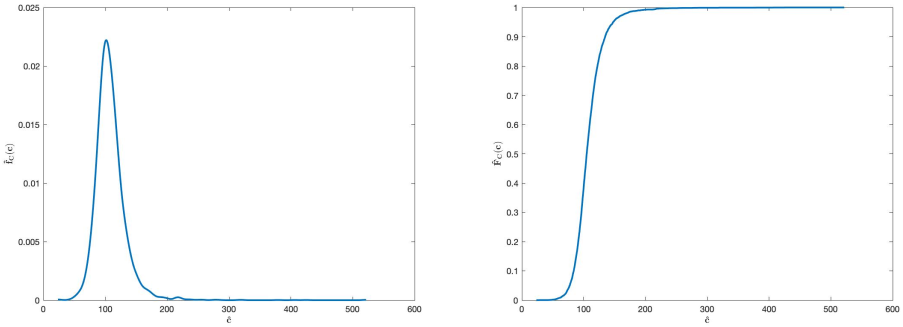

Figure 3: Pseudocost density (on the left) and cost distribution (on the right)

4 Simulation

To test if the estimated cost distribution captures the underlying cost distribution that induces observed bids, I will proceed to validation stage: simulate bids using estimated cost distribution $F_c(\cdot)$ and comparing simulated bids to actual bids. For the test data, I will use an auction from the dataset which has 16 bidders. The data is not used to estimate the cost distribution because the total number of bids with 16 bidders is less than 32.

The simulation technique is known as the inversion method and runs as follows. I will draw 16 random numbers uniformly distributed on the range $(0, 1)$ and use the inverse distribution function, $F_c^{-1}(\cdot)$, to find the corresponding costs. The bids are then computed using equation (2) and the estimated cost distribution F_c. This process is repeated 1000 times and the averages are taken.

Figure 4 plos simulated costs against simulated bids that are inferred from simulated costs using equation (2). The simulated costs and bids center around 110% and the range is small, with the max and min differing by less than 4%. This is due to the estimated unimodal cost distribution concentrating in the range [80, 200], with the peak around 110%. With the inversion method repeated for 1000 times, it is natural that the costs and bids fall into the place where most bids and costs are concentrated. With

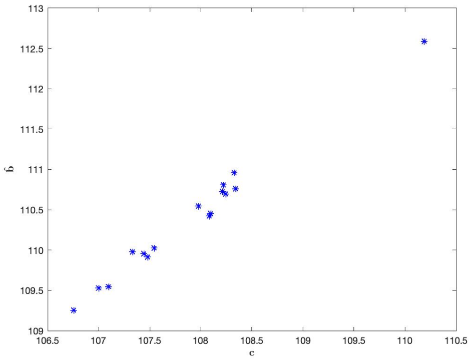

Figure 4: Simulated bids against simulated costs.

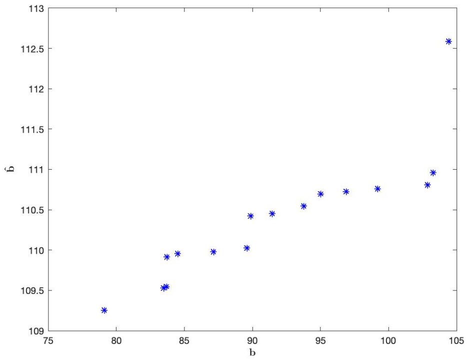

Figure 5: Bids from test data (x-axis) plotted against simulated bids (y-axis).

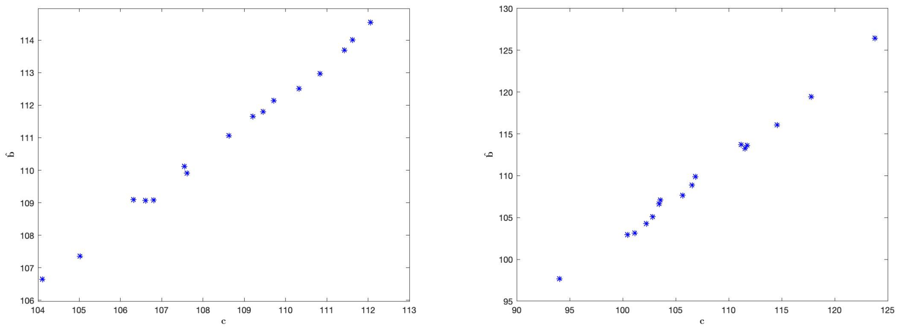

Figure 6: Bid simulation run for 100 times (left) and 10 times(right)

the number of runs curtailed to 100 and 50, as shown in Figure 6, the bids become more spread out and closer to the bid distribution. This shows that the number of runs is an important factor to determine in roder to have a more accurate simulation but it is outside the scope of this paper. For simplicity, I will keep referencing the data generated with 1000 runs.

Figure 5 plots test data against simulated bids. A wide difference between simulation and test data is observed here. This difference, however, is not unexpected for two reasons: (1) the test data, which contains an auciton with 16 bidders, does not conform to the the bid distribution depicted in Figure 1, and (2) I am comparing most likely bids predicted by the estimated cost distribution to a particular auction. Therefore, the difference here not suggest the inaccuracy of the estimated cost distribution or the equilibrium model. Instead, this exercise suggests that we may need to increase the sample size to have a more informative comparison.

5 Conclusion

This paper tries to propose a way to validate models used in structural analysis by cross-validation. The dataset is split into training data and test data. First, this paper has conducted structural estimation on a government procurement auction dataset about highway constructions and estimated the cost distribution on the training data using a two-step nonparametric procedure. Second, using the estimated cost distribution, the paper then constructs simulated bids with the inversion method and compare the simulation to test data. (Even though there are no meaningful results yet, it is very likely that the two will have statistically significant differences). The statistical difference between the simulation and test data suggests that the first-price sealed-bid equilibrium model introduced in Riley and Samuelson (1981) may not be a suitable model to predict bidder's behavior in real-world auctions.

However, in the second step of the cross-validation procedure, comparing bids directly to test theory's validity may not be a good idea, as suggested in Harrison (1989). Harrison (1989) instead proposes to "evaluate subject behavior in the expected payoff space", which can be incorporated into the project's future work.

The idea of cross-validation has been an old idea in the machine learning literature as a way to compare performance across models. I, therefore, consider it a natural idea to introduce it into structural analysis. Structural estimation relies on the model's correctness in modeling agent's behavior, but the verification of models is left to lab experiments, not empirical data which holds the final authority on models. Hopefully, this paper's experiment would shed light on possible directions that structural analysis can head to become more robust.

References

Gentry, M. L., Hubbard, T. P., McComb, R. P., & Schiller, A. R. (2018). Structural economics of auction data: A review. *Foundations and Trends in Econometrics*, *9*(2-4), 79-302.

Guerre, E., Perrigne, I., & Vuong, Q. (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica*, *68*(3), 525-574.

Harrison, G. W. (1989). Theory and misbehavior of first-price auctions. *The American Economic Review*, *79*(4), 749-762.

Hickman, B. R., & Hubbard, T. P. (2015). Replacing sample trimming with boundary correction in nonparametric estimation of first-price auctions. *Journal of Applied Econometrics*, *30*(5), 739-762.

Hubbard, T. P., & Paarsch, H. J. (2009). Investigating bid preferences at low-price, sealed-bid auctions with endogenous participation. *International Journal of Industrial Organization.*, *27*(1), 1-14. doi: 10.1016/j.ijindorg.2008.05.004

Riley, J. G., & Samuelson, W. F. (1981). Optimal auctions. *The American Economic Review*, *71*(3), 381-392.