

A Semi-Parametric Two-Stage Estimator with Machine Learning to Solve Endogeneity *

[Preliminary version]

[Click here for the latest version](#)

Yichun Song[†] Adam Dearing[‡]

March 31, 2023

Abstract

Based on the fast development in machine learning techniques, we propose a semi-parametric two-stage estimator to solve the endogeneity in models when implementing the usual linear-IV is impossible. Those models include binary discrete-choice with endogenous regressors and multinomial Logit model with zero market share. Our first stage is a flexible estimation of the expected outcomes, conditional on the regressors and instruments, whereas the second stage is a simple linear-IV regression with first-stage results inserted. After adding an analytical-calculated correction term, the first-stage error up to $o(N^{-1/4})$ does not affect the asymptotic distribution of the second-stage structural parameter estimates, allowing for a variety of non-parametric methods in the first stage. We demonstrate the behavior of our estimator through Monte Carlo simulations. Moreover, we demonstrate the performance of our method in an application using the publicly-available 2011 to 2015 transaction-level data between coal mining companies and coal-fired power plants in the US.

Keywords: Neyman Orthogonality, Machine Learning, Endogeneity, Binary Choice.

JEL classification codes: C14, C31, C36.

*All errors are our own.

[†]Department of Economics, The Ohio State University, Columbus OH 43210-1120. Email: song.1399@osu.edu.

[‡]Cornell University and NBER, SC Johnson College of Business, Ithaca NY 14853-6201. Email: aed237@cornell.edu

1 Introduction and Literature Review

Two-stage or two-step estimator is a type of estimator where the second-step estimation of structural parameters depends on some preliminary results from the first step. As a well-developed tool, it has gained wide application in various subjects. Newey (1994) summarizes both the parametric and non-parametric first stage. They demonstrate that the second stage estimation can still be \sqrt{N} -consistent even when the first stage converges slower than \sqrt{N} .

However, this advantage fails with the bias-variance trade-off in a typical machine learning estimator. The bias introduced into the model to reduce the probability of overfitting in the first stage may be carried into the second stage and generate inconsistent estimates. Seminal work by Chernozhukov et al. (2018) solves this problem by implementing Neyman Orthogonality conditions and sample-splitting. After this, debiased machine learning has forced its way into economic analysis and becomes a valid alternative in the first stage. The widespread usage of their method starts by allowing a large number of potential explanatory variables in the treatment effect estimation. Empirical works like Dube et al. (2020), Burlig et al. (2020), and Farronato et al. (2020) have demonstrated the power of the method in Chernozhukov et al. (2018).

Succeeded by theoretical improvements on bias correction, like Chernozhukov et al. (2021), Chernozhukov et al. (2022a), and Chernozhukov et al. (2022b), the debiased machine learning gradually gains higher adaptability to various models. For example, Bakhitov (2022) incorporates the automatic debiased machine learning into demand estimation and provides the debiased estimation of own-price derivatives. Cha et al. (2021) combines the orthogonal greedy algorithm and high-dimensional AIC and then applies the method to the production analysis.

High-dimensional techniques are also involved in estimations with endogeneity. Some researchers attempt to use machine learning as the first-stage prediction in 2SLS to recover more correlation between endogenous variables and instruments. However, replacing 2SLS with biased predictions from machine learning would usually aggravate the performance. For instance, Lennon et al. (2021), through a simple decomposition of the bias, demonstrates that directly plugging in machine-learning predictions leads to a more significant bias than ignoring the endogeneity. Instead, Belloni et al. (2011), Singh et al. (2019), and Chen et al. (2021) focus on using machine learning to generate the instruments. Chen et al. (2021) decides to use the first-stage machine-learning results

as an instrumental variable, which they indicate as the most harmless machine-learning in 2SLS.

In this study, we propose a novel two-step semi-parametric estimator to settle the endogeneity when the direct usage of linear-IV fails. This two-step method combines with debiased machine learning whenever model selection or regularization occurs. We do not search for a flexible approximation for the first stage or a selection of the optimal instruments. Alternatively, we target the problems that 2SLS or other linear GMM cannot tackle. We focus on nonlinear models that can be inverted conditional on some unobserved variables. For example, suppose there is a linear relationship $y_i^* = x_i'\beta_x + w_i'\beta_w + \eta_i$, with w_i the endogenous variables. But we do not observe y_i^* – instead, we only know y_i , an imperfect measure of y_i^* , which cannot be inserted in the place of y_i^* directly. These model types include binary choice with endogenous regressors, where we do not observe the latent linear structure or choice probability. Instead, we observe a binary response variable y_i .

Another example is the Logit model for market share, but we observe zero purchased quantity in the data. The logit model itself would not produce zero market share, which impedes the inversion (Nevo, 2000) by generating a negative infinite y_i^* . Usually, researchers deal with this by deleting those zero observations, imputing a fixed small number, or aggregating the purchase data into high-level until no zero quantity is left. All of them may generate bias.

This study shows that it is possible to construct a pseudo value \tilde{y}_i of y_i^* from flexible machine-learning methods. Both \tilde{y}_i and y_i^* are unobservables, but we may be able to estimate \tilde{y}_i . For instance, in the case with zero market share, we have the purchase of a commodity subject to a sampling error (Gandhi et al., 2019), from which \tilde{y}_i can be retrieved. By inserting the first-stage machine-learning estimates \tilde{y} into the original model, we could use the linear-IV estimation to deal with the endogeneity. However, a naive plug-in estimator may introduce a serious econometric problem, especially when we use model selection techniques like Lasso – this is where debiased machine learning techniques come in as a bias-correction method. We show that the bias-correction can be achieved by adding an additional analytical term in the second stage.

Two methods are closely related to ours. The most closely related method is the control function, advocated by Petrin and Train (2010) and Wooldridge (2015). In the first stage, the researcher runs an OLS regression of endogenous variables w_i on the instruments (x_i, z_i) . In the second stage, the parameters of interest are estimated via maximum likelihood or OLS, using the residuals from

the first stage as an additional regressor. When the form of $E[w_i|x_i, z_i]$ is accurately specified in a parametric first stage, the control function estimator is \sqrt{N} -consistent.

Our method, on the other hand, contains the usual linear control function as a special case. After degenerating our first stage through replacing w_i with its linear representation, we can consistently achieve a comparable estimator. Nevertheless, our setting permits more flexibility. Whereas the control function puts pressure on the joint distribution of the structural error term with the residuals in the first stage, ours depends on the correct retrieval of \tilde{y}_i . By exploiting the debiased machine learning techniques, we allow for a complicated first stage with the error up to $o(N^{-1/4})$. Besides, our method can easily accommodate multiple endogenous variables since only a single variable, \tilde{y}_i , must be estimated in our first stage.

Another closely related method is the production function estimation with endogenous variables proposed in Levinsohn and Petrin (2003) and Akerberg et al. (2015). Their method is an instrumental-free control function method, where the structural error term is inverted as an unknown function of capital and intermediate input. They overcome the endogeneity by obtaining a proxy of the structural error term in the first stage. Although both share the attribute of inversion, their method is more likely a control function method, where the first-stage result enters as an extra regressor.

We organize the rest of the paper as follows. Section 2 shows the basic model and the algorithm. We illustrate it via two typical examples – the binary choice with endogenous regressors and the Logit model with zero market shares. These two cases are the most critical applications of our method, and we will revisit them throughout this study. Section 3 demonstrates that our algorithm is consistent with Chernozhukov et al. (2022a) and Chernozhukov et al. (2021). Hence, our estimator has the desirable asymptotic properties. Section 4 contains Monte Carlo simulations to demonstrate the performance. Section 5 then discusses the advantages and potential problems. Section 6 exhibits a possible application for the binary choice model with the publicly-available individual transaction data between US power plants and coal mines. We study plants' purchase patterns in the spot market. Section 7 contains the conclusion and future research direction.

2 Model and Estimation

Here we restrict our attention to the i.i.d data only. Consider a linear function,

$$y_i^* = x_i' \beta_x + w_i' \beta_w + \eta_i \quad (1)$$

with exogenous variables x_i and endogenous variables w_i . Usually, we can solve this problem by 2SLS with excluded instruments z_i , while in certain circumstances, we may not observe y_i^* . Alternatively, we observe a y_i as an imperfect measure of y_i^* , and unfortunately, y_i cannot be expressed as a linear function of x_i and w_i . This study shows that a pseudo-outcome variable \tilde{y}_i can be constructed through a nonlinear, nonparametric method with model selection or machine learning. Then we plug in the \tilde{y}_i in the place of the unknown y_i^* and proceed with the usual linear-IV setting. Two examples of this method are illustrated below.

Case 1 – binary choice model. *Suppose a consumer faces two alternatives – the inside good 1 and the outside option, good 0, which would always be purchased. In other words, a consumer decides whether to purchase a particular product while keeping other daily consumption. The utility consumer i derives from good 1 and good 0 are,*

$$\begin{aligned} u_{i,1} &= x_{i,1}' \beta_x + w_{i,1}' \beta_w + \eta_{i,1} \\ u_{i,0} &= \eta_{i,0} \end{aligned}$$

The outside option follows the general normalization that its non-random utility is 0. Consumer i would purchase good 1 whenever there is,

$$u_{i,1} \geq u_{i,0} \Rightarrow \eta_{i,0} - \eta_{i,1} \leq x_{i,1}' \beta_x + w_{i,1}' \beta_w$$

We observe the binary choice y_i , which equals 0 and 1, representing the purchase decision. Simply replacing y_i^ in (1) by y_i leads to the linear probability model (LPM). LPM is feasible for 2SLS but suffers from a bunch of flaws¹. To overcome the drawbacks of LPM, threshold-crossing Probit or Logit becomes mainstream in estimation. If there is no endogeneity between w_i and η_i , fully-*

1. As summarized in Lewbel et al. (2012), the most commonly recognized drawback of LPM is that LPM may generate fitted probability below zero or above one. Or, LPM can only approximate the probability for a limited set of regressors.

parametrized maximum likelihood is the first choice². However, with endogeneity, the threshold-crossing nature of y_i precludes the direct application of the linear-IV method.

Although the binary choice model seems an elementary setting, a variety of other models can be built on this basic logic. For instance, consider the multiple discrete-continuous models proposed in Bhat (2005) and Bhat (2008), where they consider a situation with a constant-return outside option. A good j will never be purchased if the marginal utility of the first unit consumption is lower than the constant benefit from consuming the outside option. This property breaks down the discrete part of multiple discrete-continuous models into a binary choice problem.

Case 2 – Logit with zero market share and aggregate data. Suppose there are T markets. Each market has M_t individuals. In market $t \in T$, individuals purchase one unit from alternatives $j \in \{0, 1, 2, \dots, J\}$. 0 is the outside option. The random utility for good j is $u_{ijt} = x_{ijt}\beta_x + w_{ijt}\beta_w + \eta_{ijt}$. A consumer will choose the good j whenever u_{ijt} is the maximum. Conditional on the observables, with a random part η_{ijt} , a homogeneous consumer's optimal choice ends up in a choice probability s_{jt} . The probability of observing the quantities purchased $(Q_{0t}, Q_{1t}, Q_{2t}, \dots, Q_{Jt})$ inside a market t is,

$$Pr_t(Q_{0t}, Q_{1t}, Q_{2t}, \dots, Q_{Jt}) = \frac{M_t!}{Q_{0t}!Q_{1t}!Q_{2t}! \dots Q_{Jt}!} s_{0t}^{Q_{0t}} s_{1t}^{Q_{1t}} \dots s_{Jt}^{Q_{Jt}} \quad (2)$$

Equation (2) gives $Q_{jt} = 0$ with positive probability unless $M_t \rightarrow \infty$. However, a model with an error η that has support $[-\infty, +\infty]$ would not allow $Q_{jt} = 0$, since no matter how low $x_{ijt}\beta_x + w_{ijt}\beta_w$ becomes, there is still possibility for a realization of $\eta_{it} = (\eta_{i0t}, \eta_{i1t}, \dots, \eta_{iJt})'$ that makes u_{ijt} the optimal choice.

To deal with the endogenous w_{ijt} in such a model, researchers like Nevo (2000) invert the system by taking \ln of the observed shares. This inverting scheme makes it vulnerable to the potential zero market share since $\ln(x) \rightarrow -\infty$ as $x \rightarrow 0$.

Intuitively, our goal is to retrieve a \tilde{y}_i from y_i , with \tilde{y}_i mimicking the behavior of y_i^* in Equation (1). Nevertheless, y_i^* is a function of the unobserved η_i . Therefore, the first set of assumptions, Assumption 1 below, unveils the condition under which the construction of \tilde{y}_i is possible. Assump-

2. If full parameterization is too restrictive, Cosslett (1983) also provides a distribution-free estimator.

tion $\mathbf{1}(i)$ mainly indicates when the systematic part³ of η_i can be approximated well by a function of the observables.

Assumption 1. (i) $\eta_i = h_\zeta(\zeta_i) + \varepsilon_i$, and ζ_i is the error term that can be defined implicitly by $d(x_i, w_i, z_i, \zeta_i) = 0$. Besides, $E[h_\zeta(\zeta_i)|z_i, x_i] = 0$. (ii) ε_i is orthogonal to $x_i, z_i, h_\zeta(\zeta_i)$. The CDF of ε_i , $\Lambda_\varepsilon(\cdot)$ is known and strictly increasing for all $\varepsilon_i \in \mathcal{R}$, with finite first and second moments.

Assumption $\mathbf{1}(i)$ refers to the case where given (x_i, w_i, z_i) , ζ_i is pinned down uniquely by a deterministic function $d(x_i, w_i, z_i, \zeta_i) = 0$. Assumption $\mathbf{1}(i)$ is restrictive but may not be as restrictive as it appears. The restriction may preclude a more complicated determination of w_i . For instance, if w_i is the price, usually, we cannot allow for a Nash Bertrand competition with an additional error term in the marginal cost unless researchers are willing to risk estimating it semi-parametrically first following Berry and Haile (2014). However, as elucidated by Berry and Haile (2014) and Kim and Petrin (2019), estimating cost shock and a flexible pricing function is feasible but could be computationally intensive.

Assumption $\mathbf{1}(i)$ implies that we can construct a \tilde{y} such that $E[(\tilde{y}_i - x_i'\beta_x - w_i'\beta_w) \otimes (x_i', z_i)'] = 0$, at the true value (β_x^0, β_w^0) . Assumption $\mathbf{1}(i)$ contains the linear control function as a special case, where $\zeta_i = w_i - x_i'\tau_x - z_i'\tau_z$, and the joint distribution of (η_i, ζ_i) is assumed normal. Correspondingly, η_i can be decomposed into two parts as $\eta_i = a\zeta_i + \varepsilon_i$, with ε_i uncorrelated with observables. Nevertheless, Assumption $\mathbf{1}(i)$ allows more flexible patterns than a typical control function. w_i can be determined by an unknown nonlinear function without many parametric assumptions.

As for the multi-product Nash Bertrand competition we commonly encounter in industrial organization literature, Kim and Petrin (2019) has studied the invertibility. First, they notice that one can invert out the cost shock nonparametrically by Berry and Haile (2014)'s identification results. Hence, they treat the supply side as if the η_i is the only error term and then naturally invert η_i out by a flexible function of observed characteristics. However, although they start with an ambitious plan, their application focuses only on the low-dimensional additively separable case, without adjusting for the possibly high-dimensional machine learning first stage, where the convergence speed could be slower than \sqrt{N} .

Assumption $\mathbf{1}(ii)$ claims that we need to assume the distribution of ε_i to proceed with the inversion. One may consider the specification Λ_ε as a normalization for identification, which is common in

3. Systematic part refers to the part in η_i that correlates w .

either binary choice model or market share with Logit errors. The parametric assumption in ε_i may be less restrictive after considering the extra cost we encounter for having a distribution-free scheme. Theoretically, one may refer to the distribution-free maximum likelihood proposed by Cosslett (1983) or the special regressor method advocated in Dong and Lewbel (2015) to have the distribution estimated simultaneously with the parameters. Both of them suffer from non-trivial large-support requirements. Cosslett (1983) calculates the cumulative probability for the observed points only. Similar requirements apply to the probability density in Lewbel (2012) and Dong and Lewbel (2015)⁴. Besides, Cosslett (1983)'s distribution-free estimator still needs to have parameters in front of some other observables normalized.

Then we continue our analysis with those two examples.

Case 1 continue. Suppose we observe $y_i = \mathbf{1}\{\eta_{i,0} - \eta_{i,1} \leq x'_{i,1}\beta_x + w_{i,1}\beta_w\} = \mathbf{1}\{h_\zeta(\zeta_{i,0}) + \varepsilon_{i,0} - h_\zeta(\zeta_{i,1}) - \varepsilon_{i,1} \leq x'_{i,1}\beta_x + w'_{i,1}\beta_w\} = \mathbf{1}\{h(x_{i,0}, w_{i,0}, z_{i,0}) + \varepsilon_{i,0} - h(x_{i,1}, w_{i,1}, z_{i,1}) - \varepsilon_{i,1} \leq x'_{i,1}\beta_x + w'_{i,1}\beta_w\}$. The second inequality comes directly from Assumption 1, where we can re-write the $h_\zeta(\zeta_i)$ as a function $h(x_i, w_i, z_i)$ from $d(x_i, w_i, z_i, \zeta_i) = 0$.

Usual normalization of the outside option gives $h(x_{i,0}, w_{i,0}, z_{i,0}) = 0$. Assume that the CDF of the distribution of $\varepsilon_{i,0} - \varepsilon_{i,1}$ follows Λ_ε . Therefore, the conditional mean of the binary outcome variable is,

$$\begin{aligned} E[y_i|w_i, x_i, z_i] &= E[y_i|w_i, x_i, h(x_i, w_i, z_i)] \\ &= Pr_i(y_i = 1|w_i, x_i, h(x_i, w_i, z_i)) = \Lambda_\varepsilon(x'_i\beta_x + w'_i\beta_w + h(x_i, w_i, z_i)) \end{aligned} \quad (3)$$

By Assumption 1(ii), Λ_ε is known and invertible. If we know the conditional expectation $E[y_i|w_i, x_i, z_i]$, following (3), we can invert the system to obtain,

$$\Lambda_\varepsilon^{-1}(E[y_i|w_i, z_i, x_i]) = x'_i\beta_x + w'_i\beta_w + h(x_i, w_i, z_i)$$

Denote $W_i = (x_i, w_i, z_i)$. Then following Assumption 1(i), we have the moment condition m_i such that,

$$m(W_i, \beta) = [\Lambda_\varepsilon^{-1}(E[y_i|w_i, x_i, z_i]) - x'_i\beta_x - w'_i\beta_w] \otimes (x'_i, z'_i)', \quad \text{with } E[m(W_i, \beta)] = 0. \quad (4)$$

4. Although, according to Dong and Lewbel (2015), a symmetric-support condition can be an alternative, the importance of their large-support conditions has been re-emphasized by a bunch of simulations in Bontemps and Nauges (2017).

In this case, $\tilde{y}_i = \Lambda_\varepsilon^{-1}(E[y_i|w_i, x_i, z_i])$.

Case 2 continue. Similar to Case 1, we then replace the η_{ijt} in the random utility by its expression and have $u_{ijt} = x_{ijt}\beta_x + w_{ijt}\beta_w + h_i(x_{jt}, w_{jt}, z_{jt}) + \varepsilon_{ijt}$. Following the routine assumption for the discrete choice model with aggregate data, we have $\varepsilon_{ijt} \sim T1EV$ and the aggregate market share with the homogeneous consumer i becomes,

$$s_{jt} = \begin{cases} \frac{\exp(x_{jt}\beta_x + w_{jt}\beta_w + h(x_{jt}, w_{jt}, z_{jt}))}{1 + \sum_{\ell=1}^J \exp(x_{\ell t}\beta_x + w_{\ell t}\beta_w + h(x_{\ell t}, w_{\ell t}, z_{\ell t}))}, & j \in \{1, 2, \dots, J\} \\ \frac{1}{1 + \sum_{\ell=1}^J \exp(x_{\ell t}\beta_x + w_{\ell t}\beta_w + h(x_{\ell t}, w_{\ell t}, z_{\ell t}))}, & j = 0 \end{cases}$$

If we know the true s_{jt} , we can invert it (Nevo, 2000) by

$$\log\left(\frac{s_{jt}}{s_{0t}}\right) = x_{jt}\beta_x + w_{jt}\beta_w + h(x_{jt}, w_{jt}, z_{jt})$$

And the implied moment condition m_i is,

$$m(W_i, \beta) = \left[\log\left(\frac{s_{jt}}{s_{0t}}\right) - x_{jt}\beta_x - w_{jt}\beta_w \right] \otimes (x'_i, z'_i)' \quad \text{with } E[m(W_i, \beta)] = 0. \quad (5)$$

However, the moment conditions in equation (4) and (5) are infeasible. In (4), we do not observe the $E[y_i|w_i, x_i, z_i]$. What we observe is y_i . We do not know the true s_{jt} in the second case. Our calculated $\tilde{s}_{jt} = \sum_{i=1}^{M_t} Q_{ijt} / \sum_{j=1}^J \sum_{i=1}^{M_t} Q_{ijt}$ may not be ready to use, since \tilde{s}_{jt} can be 0 while the true s_{jt} cannot. Nevertheless, we may estimate them in the first stage and plug them in to use the linear-in-parameters moment condition with the known distribution of ε_i . Appendix A. shows through a simple example that a typical machine learning method would follow the observed data pattern and obtain the estimated $E[y_i|w_i, x_i, z_i]$, which is what we want in equation (4) and (5).

The performance of this two-step estimator depends on the first-stage estimation of \tilde{y}_i . Newey (1994) demonstrates that the second-stage structural parameters β can still be \sqrt{N} -consistent even when the first-stage converges slower than \sqrt{N}^5 . However, if we have a high dimensional space in which model a bias-variance trade-off is required to prevent overfitting, the original moment condition would fail to be \sqrt{N} -consistent unless using the orthogonal moment and sample-split (Chernozhukov et al., 2018). In this case, we can achieve the orthogonal moment condition by

5. In their examples, kernel estimation can be adopted for a candidate of the first stage.

adding an analytical term ϕ , which can be constructed as a deterministic function of the estimated propensity score $\hat{\pi}_i$.

Suppose the moment condition is $m_i(W_i, \gamma, \beta)$. γ is an invertible function that maps π into \tilde{y}_i . $\gamma_i = \gamma_i(\pi_i) = \tilde{y}_i$. π_i is the probability of choosing option 1 in Case 1, and market share \hat{s}_{jt} in Case 2. Follow the notation in Chernozhukov et al. (2022a), the correction term ϕ equals,

$$\phi(W_i, \gamma) = \alpha(W_i)\rho(W_i, \gamma) = \left[\frac{\partial \gamma(\pi_i)}{\partial \pi_i} \otimes (x'_i, z'_i)' \right] \times (y_i - \pi_i) \quad (6)$$

$\rho(W_i, \gamma) = (y_i - \pi_i) = (y_i - \Lambda(\gamma_i))$. Adding ϕ in (6) to the moment condition, we obtain an orthogonal moment condition ψ such that,

$$\psi(W_i, \gamma, \beta, \gamma) = m(W_i, \gamma, \beta) + \phi(W_i, \gamma)$$

Theorem 1 shows the orthogonality of $\psi(W_i, \gamma, \beta, \gamma)$ regarding the first stage estimates π_i .

Theorem 1. *Adding the ϕ term to m provides the asymptotic orthogonality of moment conditions regarding the first-stage nuisance parameter π_i . And it has $\int_W \phi(W, \gamma_0, \alpha_0) F_0(dW) = 0$.*

Proof. With $\tilde{y}_i = \gamma(\pi_i)$, the population moment condition equals

$$\begin{aligned} \psi(W_i, \gamma, \beta, \pi_i) &= E \left[\left(\frac{\partial \gamma(\pi_i)}{\partial \pi_i} \times (y_i - \pi_i) + \gamma(\pi_i) - x'_i \beta_x - w'_i \beta_w \right) \otimes (x'_i, z'_i)' \right] \\ &= E \left[E \left[\left(\frac{\partial \gamma(\pi_i)}{\partial \pi_i} \times (y_i - \pi_i) + \gamma(\pi_i) - x'_i \beta_x - w'_i \beta_w \right) \middle| W \right] \otimes (x'_i, z'_i)' \right] \\ &= \int_W \left(\frac{\partial \gamma(\pi_i)}{\partial \pi_i} \times (E[y_i | W_i] - \pi_i) + \gamma(\pi_i) - x'_i \beta_x - w'_i \beta_w \right) \middle| W \otimes (x'_i, z'_i)' dF(W) \\ &= \int_W \left(\frac{\partial \gamma(\pi_i)}{\partial \pi_i} \times (\pi_i^0 - \pi_i) + \gamma(\pi_i) - x'_i \beta_x - w'_i \beta_w \right) \middle| W \otimes (x'_i, z'_i)' dF(W) \end{aligned}$$

Take the derivatives with respect to π_i and evaluate at the true value (β^0, π_i^0) . We obtain,

$$\frac{\partial \psi(W_i, \gamma, \beta, \pi_i)}{\partial \pi_i} = \frac{\partial}{\partial \pi_i} \left[\frac{\partial \gamma(\pi_i)}{\partial \pi_i} (\pi_i^0 - \pi_i) + \gamma(\pi_i) \right] = \frac{\partial \gamma^2(\pi_i)}{\partial \pi_i^2} (\pi_i^0 - \pi_i)$$

which equals to 0 when evaluated at $\pi_i = \pi_i^0$, for all possible $(x_i, w_i, z_i) \in \mathcal{X} \times \mathcal{W} \times \mathcal{Z}$. It implies that the Gateaux derivative of moment condition $\psi(\cdot)$ with respect to π_i is zero. Moreover, since $E[y - \pi_0 | W_i] = E[y - E[y|W]|W] = 0$ and $\partial \gamma(\pi)/\partial \pi$ is a function of W , we have $E[\phi(W, \gamma_0)] = \int_W \phi(W, \gamma_0, \alpha_0) F_0(dW) = 0$. \square

Case 1 continue. For the binary choice model, $\pi_i = Pr_i(y_i = 1|W_i)$, and γ is $\Lambda_\varepsilon^{-1}(\cdot)$. Thus, we have the ϕ as,

$$\phi(W_i, \gamma) = \frac{y_i - E[y_i|W_i]}{\Lambda'_\varepsilon(\Lambda_\varepsilon^{-1}(E[y_i|W_i]))} \otimes (x'_i, z'_i)' \quad (7)$$

Then the orthogonal moment condition ψ becomes,

$$E \left\{ \left(\frac{y_i - E[y_i|W_i]}{\Lambda'_\varepsilon(\Lambda_\varepsilon^{-1}(E[y_i|W_i]))} + \Lambda_\varepsilon^{-1}(E[y_i|W_i]) - x'_i\beta_x - w'_i\beta_w \right) \otimes (x'_i, z'_i)' \right\} = 0$$

Case 2 continue. For the discrete choice, we assume the outside option is a “safe good” defined in Gandhi et al. (2019). s_{0t} is always positive and far from 0⁶. $y_{jt} = s_{jt}$ and $\pi_{jt} = \tilde{s}_{jt}$. Then we have the correction term,

$$\phi(W_{jt}, \gamma) = \frac{1}{\tilde{s}_{jt}} (s_{jt} - \tilde{s}_{jt}) \otimes (x'_i, z'_i)'$$

Then the moment condition ψ in the second stage becomes,

$$E \left\{ \left[\frac{1}{\tilde{s}_{jt}} (s_{jt} - \tilde{s}_{jt}) + \log \left(\frac{\tilde{s}_{jt}}{s_{0t}} \right) - x_{jt}\beta_x - w_{jt}\beta_w \right] \otimes (x'_i, z'_i)' \right\} = 0$$

In practice, we can implement our estimation in a few straightforward steps through standard statistical software. Furthermore, sample splitting would further reduce the overfitting bias as indicated in Chernozhukov et al. (2018) and Cattaneo et al. (2019). To implement, we divide the sample into L different sets, $\ell \in \{1, 2, \dots, L\}$ ⁷. I_ℓ contains the data in ℓ and $\hat{\gamma}_\ell$ are the estimates using all sample points not contained in I_ℓ . We summarize our estimation procedure in Algorithm 1 below.

Algorithm 1. We present the algorithm as:

Step 1. For each splitted sample, $\ell \in \{1, 2, \dots, L\}$. Obtain an estimate $\hat{\pi}_\ell$. Usually $\hat{\pi}_\ell$ is probability or propensity score and γ is a deterministic function of $\hat{\pi}$. $\hat{\pi}_\ell$ can be estimated using a flexible nonparametric machine learning or flexible parametric model.

6. This is a reasonable assumption. Usually, the outside option is defined as a composite good or the all-other goods and accounts for more than 50% of the observed market share.

7. A full leave-one-out jackknife estimator like Cattaneo et al. (2019) could be too time-consuming in this case. As pointed out by Chernozhukov et al. (2022a), $L = 5$ or $L = 10$ would be a reasonable choice to balance the computational burden and bias reduction. $L = 10$ works better for the smaller data sets.

Step 2. Construct the moment condition by,

$$\hat{m}(W, \gamma, \beta) = \frac{1}{n} \sum_{\ell=1}^L \sum_{i \in I_\ell} m(W_i, \hat{\gamma}_\ell, \beta)$$

where $\hat{\gamma}_i$ serves as \tilde{y}_i . If there is model selection or parameter penalization for bias-variance trade-off in Step 1, then one needs to refer to the debiased moment condition such that,

$$\hat{\psi}(W, \gamma, \beta) = \hat{m}(W, \gamma, \beta) + \hat{\phi}(W, \gamma) = \frac{1}{n} \sum_{\ell=1}^L \sum_{i \in I_\ell} m(W_i, \hat{\gamma}_\ell, \beta) + \frac{1}{n} \sum_{\ell=1}^L \sum_{i \in I_\ell} \phi(W_i, \hat{\gamma}_\ell)$$

Step 3. Solve for the $\hat{\beta}$, as the minimizer of,

$$\hat{\beta} = \arg \min_{\beta_x, \beta_w} \hat{m}(W, \gamma, \beta)' \hat{\Upsilon} \hat{m}(W, \gamma, \beta) \quad (8)$$

or, with bias-correction,

$$\hat{\beta} = \arg \min_{\beta_x, \beta_w} \hat{\psi}(W, \gamma, \beta)' \hat{\Upsilon} \hat{\psi}(W, \gamma, \beta). \quad (9)$$

$\hat{\Upsilon}$ is the weighting matrix. If we have linear m , then the 2SLS have a simple analytical solution. Additionally, we can have $\hat{\Upsilon} = \hat{\Psi}^{-1}$ to improve efficiency.

The usual GMM standard errors are asymptotically valid using the orthogonal moments. However, they may have worse finite-sample performance. More discussion of asymptotic results can be found in Section 3.

Case 1 continue Consider a special case with ε_i follows *i.i.d* normal standard normal distribution Φ . Then, we calculate the $\hat{\gamma}$ in step 1 of Algorithm 1, obtain the predicted probabilities $\hat{\pi}_i$, and insert it back to get $\tilde{y}_i = \hat{\gamma}(\hat{\pi}_i) = \Phi^{-1}(\hat{\pi}_i)$. In step 2, if model selection techniques are adopted, the pseudo-outcome variable \tilde{y}_i then becomes $\tilde{y}_i = (y_i - \hat{\pi}_i) / (\phi(\Phi^{-1}(\hat{\pi}_i))) + \Phi^{-1}(\hat{\pi}_i)$.

One noteworthy aspect of our proposed technique is that our method is different from the so-called forbidden regression. The name “forbidden regression” in either J. Wooldridge (2001), Angrist and Krueger (2001) or Chen et al. (2021), resorts to regressing y_i on the \hat{w}_i , where \hat{w}_i is estimated through either a nonlinear transformation of linear-OLS or nonlinear regression. They point out that this type of construction leads to inconsistent estimators. However, all those claims focus on

fitted values entering as regressors, which differs from our proposal. Besides, our primary focus does not lie in improving the linear-IV performance. Instead, we want to extend the linear-IV method to those that previously could not use linear-IV by constructing a pseudo-dependent variable \tilde{y}_i .

3 Asymptotic Results for the Debiased Two-Stage Estimator

Firstly, we state several assumptions. Basically, they are the similar set of conditions required in Chernozhukov et al. (2022a) for the method of moments and Chernozhukov et al. (2021) for the generalized linear model.

Assumption 2. Bounded π_i : π_i^0 is between $(\bar{c}, 1 - \bar{c})$ for some \bar{c} uniformly over the support of $W \in \mathcal{X} \times \mathcal{W} \times \mathcal{Z}$.

Assumption 3.

- (i) the weighting matrix, $\hat{\Upsilon} \xrightarrow{P} \Upsilon_0$, and Υ_0 is positive definite;
- (ii) $E[m(W_i, \gamma_0, \beta)] = 0$ holds if and only if $\beta = \beta_0$;
- (iii) $\beta \in \mathcal{B}$ and \mathcal{B} is a compact set, and $E[\sup_{\beta \in \mathcal{B}} |m(W, \gamma_0, \beta)|] < \infty$;
- (iv) there is a $C > 0$ and $\hat{M} = O_p(1)$ such that for $\|\gamma - \gamma_0\|$ small enough and all $\tilde{\beta}, \beta \in \mathcal{B}$, we have

$$\|\hat{m}(W, \gamma, \tilde{\beta}) - \hat{m}(W, \gamma, \beta)\| \leq \hat{M} \|\tilde{\beta} - \beta\|^{1/C}$$

Assumption 4.

- (i) $E[|m(W_i, \gamma_0, \beta_0)|^2] < \infty$ and there exists $C > 0$ such that $E[|m(W, \gamma, \beta)|^2] \leq C \|\rho\|^2$;
- (ii) $\alpha_0(W_i)$ and $E[|\rho(W_i, \pi_0)|^2 | W]$ are bounded;
- (iii) There is a $C > 0$ such that for all $\|\pi_\ell - \pi_0\|$ small enough, such that (1) $E[|\hat{\alpha}_\ell(W_i, \hat{\pi}_\ell) - \alpha_0(W_i, \pi_0)|^2] \leq C \|\hat{\pi}_\ell - \pi_0\|$; (2) $\|\hat{\gamma}(W, \hat{\pi}_\ell) - \gamma_0(W, \pi_0)\| \leq C \|\hat{\pi}_\ell - \pi_0\|$; (3) For the α corresponding to j -th moment condition, $\|\hat{\alpha}_{j,\ell}(W, \hat{\pi}_\ell) - \alpha_{j,0}(W, \pi_0)\| \leq C \|\hat{\pi}_\ell - \pi_0\|$.
- (iv) Faster than $N^{-1/4}$ convergence of estimated π , or $\|\hat{\pi}_\ell - \pi_0\| = o_p(N^{-1/4})$.
- (v) $E[|\psi(W_i, \gamma, \alpha_0, \beta_0)|^2] < \infty$. For all γ with $\|\gamma - \gamma_0\|$ small enough.

$$\|E[\psi(W_i, \gamma, \alpha_0, \beta_0)]\| = \left\| \int [m(W_i, \gamma, \beta) + \alpha(W_i)\rho(W_i, \pi)] F_0(dW) \right\| \leq C \|\gamma - \gamma_0\|^2$$

- (vi) $\int \|m(W_i, \hat{\gamma}_\ell, \hat{\beta}) - m(W_i, \hat{\gamma}_\ell, \beta_0)\|^2 F_0(dW) \xrightarrow{P} 0$

Assumption 2 requires the propensity score π_i to be bounded from below and above. Assumption 2 is a common technical regularization condition for literature that involves propensity score inversion since otherwise, the function would blow up as it approaches 0 or 1. Assumption 2 helps regularize the function γ and α and is necessary even in the work that does not explicitly have propensity score entering into the estimation, like Singh and Sun (2021).

Assumptions 3(i), 3(ii), and 3(iii) are typically identification conditions for the extreme estimator. We need these for consistency proof. According to Newey (1991), Assumption 3(iv) can be replaced by \mathcal{B} is convex, $m(W_i, \gamma, \beta)$ is continuously differentiable and $\partial \hat{m}(W, \beta) / \partial \beta$ is dominated by a stochastically bounded sequence. Both Assumption 3(iv) and the alternative are easily satisfied whenever we have a moment condition linear in β .

Assumption 4 is more specific. Assumption 4(i) requires the finite second moment of m . For this case, it refers to $E[||m(W, \gamma_0, \beta_0)||^2] = E[||h_i \otimes (x_i, z_i)'||^2] < \infty$, a typical assumption for any linear-IV estimator to obtain the asymptotic distribution. Assumption 4(ii) is purely technical, and we maintain it for simplicity.

Recall that both γ and α are deterministic functions of π , Assumption 4(iii) requires the mean square continuous of α and Lipschitz continuity of α and γ , which hold with Assumption 2. In the binary choice case, we have the quantile function (inversed CDF) $\gamma = \tilde{y}_i = \Lambda_\varepsilon^{-1}(\pi_i)$ and the inverse propensity score $\alpha = (1/(\Lambda'_\varepsilon(\Lambda_\varepsilon^{-1}(\pi_i)))) \otimes (x'_i, z'_i)$. By Assumption 2, the continuity follows directly after the elimination of the points where $\pi_i \rightarrow 0$ or $\pi_i \rightarrow 1$, or both γ and α diverge rapidly. A similar situation applies to the zero-share case. We need the true market share to be bounded away from 0 and 1.

Assumption 4(v) bounds the expectation of the debiased moment condition. Assumption 4(vi) is equivalent to $\hat{\beta}_\ell$ is a consistent estimator for the linear-IV model with $\hat{\gamma}_\ell$ inserted, and the linear moment m is mean-square continuous in β – a condition that is easy to meet with linear moment m .

Assumption 4(iv) requires the rate of convergence for the estimated π_i is faster than $N^{-1/4}$. As illustrated in Chernozhukov et al. (2021), unfavorable nonlinearity in ρ regarding γ eliminates the possibility of double robustness⁸. Moreover, it requires a faster convergence rate in the first stage,

8. Double robustness is an encouraging property widely existing in treatment-effect estimation. One needs to have both γ and ρ entering linearly for the double robustness to hold, which is impossible in our setting. For our setting,

which is easily stated but may be hard to verify in practice. Fortunately, the $o(N^{-1/4})$ convergence is theoretically attainable in nonparametric settings without endogenous variables.

The major problem in showing the convergence rate is that the estimated $\hat{\gamma}$ may not be structurally specified correctly. For instance, higher-order polynomials can approximate the underlying γ_0 while the γ_0 itself needs not to be a linear combination of the exact polynomials. With this, the traditional consistency is no longer applicable. Instead, after assuming an optimal solution inside the restricted function set, the error term can be bounded by oracle inequalities. For the lasso-type estimator, after the influential work by Bickel et al. (2009), further research like Caner and Kock (2019) has shown the oracle inequalities in linear-GMM with Lasso. Their results have been used in Bakhitov (2022) to demonstrate the mean square convergence of α . For the deep neural nets, the bounds provided in Farrell et al. (2021) are then used for the error in Chernozhukov et al. (2021).

While a variety of papers provide well-developed results for linear regression, only a few focus on generalized linear models or maximum likelihood. Alquier et al. (2019) give a version of tight oracle inequality for penalized maximum likelihood, but their analysis is mainly restricted to the case that the regressors follow a Gaussian distribution. Blazère et al. (2014) show the asymptotic oracle inequality of the generalized linear model for a grouped Lasso, incorporating the binary response model with known parametric error. Neural nets are considered a sieve estimation but with sieves determined within the data. Chen and White (1999) studies the shallow neural networks with smooth activation functions and provides bounds for the estimation error. Farrell et al. (2021) demonstrate the performance of deep neural nets with ReLU activation function and Lipschitz loss functions in generalized linear regression, including logistic⁹. They show a non-asymptotic oracle inequality, which attains a convergence rate of $o(N^{-1/4})$ under some smoothness conditions, rendering it a reasonable alternative for the first stage.

Then we reach the central theorem for this study. Theorem 2 states that under the assumptions mentioned above, we would have \sqrt{N} -consistent and asymptotic normality of β , with the propensity score estimated by high-dimensional techniques.

Theorem 2. *After all the Assumptions 1 - 4 and the zero-derivative property illustrated in Theorem 1 are satisfied¹⁰, we would have $\hat{\beta} \xrightarrow{P} \beta_0$ and the asymptotic normality of the structural parameters*

either γ or ρ can be made linearly in m or ϕ , but not both.

9. And they mention that their results can even cover multinomial logistic regression with multiclass classification.

10. For Case 2, we need an extra assumption that the number of individual consumers in each market would also go to infinity with N .

β as,

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \mathbf{V}), \quad \hat{\mathbf{V}} \xrightarrow{p} \mathbf{V}, \quad \hat{\mathbf{V}} = (\hat{G}' \hat{\Upsilon} \hat{G})^{-1} \hat{G}' \hat{\Upsilon} \hat{\Psi} \hat{\Upsilon} \hat{G} (\hat{G}' \hat{\Upsilon} \hat{G})^{-1}$$

$$\hat{G} = \partial \hat{m}(W, \gamma, \hat{\beta}) / \partial \beta.$$

$\hat{\Psi} = \frac{1}{n} \sum_{\ell=1}^L \sum_{i \in I_\ell} \hat{\psi}_{i\ell} \hat{\psi}'_{i\ell}$. Υ serves as the weighting matrix for GMM type estimators, and equals to $\hat{\Upsilon} = (Z'Z)^{-1}$ in the 2sls.

Proof. See [Appendix B.](#) □

4 Monte Carlo Simulations

In this section, we evaluate the finite-sample performance of Algorithm 1 using various data-generating processes for both the binary choice and zero market shares.

4.1 Case 1 Continue

We assume that our binary choice model as the discrete part of the multiple discrete-continuous choice model in Bhat (2018)¹¹. In a market t , there are J inside goods. Suppose the error term η_{ijt} can be decomposed into two parts, $\eta_{ijt} = \xi_{ijt} + \varepsilon_{ijt}$. ξ_{ijt} is the systematical error that causes endogeneity, and ε_{ijt} is the random part. A consumer makes separate decision regarding purchasing each individual good j . Specifically, a consumer i would choose to purchase good $j \in J$ at market t if and only if,

$$x'_{ijt}\beta_x + w'_{ijt}\beta_w - \log(h_{ijt}) + \xi_{ijt} + \varepsilon_{ijt} \leq x'_{i0t}\beta_x + w'_{i0t}\beta_w - \log(h_{i0t}) + \varepsilon_{i0t}$$

In other words, whether or not good j is purchased by consumer i would depends only on the valuation regarding the outside good 0, but not on $j' \in J, j' \neq j$ ¹². Precisely, we focus on the dis-utility rather than the utility. h_{i0t} comes from the capacity constraint of a multiple discrete-continuous optimization. If consumers and goods are homogeneous inside a market t , the subscript i can be omitted. Our goal in first stage is to provide a flexible estimation of the choice probability, or, $\Lambda_\varepsilon[(x_{i0t} - x_{ijt})'\beta_x + (w_{i0t} - w_{ijt})'\beta_w - [\log(h_{i0t}) + \log(h_{ijt})] - \xi_{ijt}]$. For this example, we test the performance of different nonparametric and machine learning methods, including polynomial

11. In Bhat (2018), the baseline utility governing the discrete choice can be different from the continuous part; thus, we may consider them separately. Overall, this does not deviate from a binary discrete choice model.

12. Bhat (2008) shows that the shares for $j \in J$ would be determined in the continuous part of model, which is temporarily omitted in this setting.

approximation, elastic nets (abbreviated as EN in tables), deep neural nets (DNN), etc.

4.1.1 w_{jt} is generated linearly

In the most basic case, we have w_{1t} determined by a linear separable function,

$$w_{jt} = x'_{jt}\tau_x + z'_{jt}\tau_z + \xi_{jt} + v_{jt} \quad (10)$$

v_{jt} follows a Normal distribution with $\mu = 0$ and $\sigma = 0.2$. $x_{jt} = (1, x_{jt,1}, x_{jt,2}, w_{jt})'$. For k -th repetition of simulated sample, we generate $\bar{x}_{j,1}^k \sim U(2, 4)$ and $\bar{x}_{j,2}^k \sim N(0, 1)$. Then, $x_{jt,1}^k = \bar{x}_{j,1}^k + 0.5 \times U(0, 1)$, $x_{jt,2}^k = \min_{j \in J}(\bar{x}_{j,2}^k) + \bar{x}_{j,2}^k + 0.1 + 0.5 \times U(0, 1)$. U denotes the uniform distribution. $\xi_{jt} \sim U(-1, 1)$, and ε follows a T1EV distribution with mean normalized to 0. $z_{i,j}$ are the excluded instruments. $z_{jt} = (z_{jt,1}, z_{jt,2})$ enters as the determinants for w_i and also as exclusive moment conditions in the linear-IV part. $z_{jt,1}^k = z_{j,1}^k = \{0, 1\}$. $\bar{z}_{jt,2}^k \sim U(0, 2)$, and $z_{jt,2}^k = \bar{z}_{jt,2}^k + 0.2 \times x_{jt,1}^k + \chi^2(1)$.

Issues regarding numerical stability need to be emphasized. In the binary response model with endogenous regressors, our correction term has probability density in the denominator. Even without correction, we still have the inverse of the CDF. When implemented, we adopt the ad hoc trimming and delete the estimated $\hat{\pi}_i$ with extreme values. These are standard empirical practices in economic research, though coming with limited theoretical justifications¹³. Moreover, as a common implicit requirement for moments involving inverse propensity scores, we must have enough variation in the data, where we cannot have most choice probabilities clustering at around 0 and 1.

Table 1 to 2 illustrate the simulation with w_{jt} determined by (10). The standard errors are empirical from the 1000 repetitions of simulations. “Raw” indicates using the moment condition (8), while “debiased” refers to the moment condition in (9). While not accounting for endogeneity results in bias, all other methods in Table 1 work reasonably well. Both bias and standard errors slightly increase when we switch from linear control function or Logistic regression in the first stage to the DNN. In general, all the performances are fairly acceptable.

13. For instance, Chernozhukov et al. (2018) drops the observations in group $Z = 0$ with propensity score lying outside the range in group $Z = 1$. Trimming has often been criticized for its randomness, loss of information, and lack of theoretical justification. Li et al. (2019) down-weights the observations with extreme propensity scores in a treatment effect estimation framework. Singh and Sun (2021) refers to the automatic debiased machine learning to directly approximate the propensity score function. Nevertheless, what they have proposed seems not applicable to our setting.

Table 2 includes the case where we have 100 more invalid instruments generated as i.i.d standard Normal series, or $z_{jt,s} \sim N(0, 1)$, $s = 3, 4, \dots, 102$. The invalidity of instruments refers to the fact that they do not enter into the determination of w_{jt} . Table 2 shows the estimation when we adopt Lasso penalization to tackle many instruments. Table 2 indicates that using orthogonal moment accompanied by Lasso would decrease the bias at the cost of a slight variance increase. If the sparsity pattern holds, we can extend it to a high-dimensional case where the number of potential instruments grows higher than the sample size.

Table 3 focus on the case where we have slightly misspecification. In this case, the w_{jt} is determined by

$$w_{jt} = x'_{jt}\tau_x + z'_{jt}\tau_z + 0.5\xi_{jt} + z_{jt,3} \times \xi_{jt} + 0.5\nu_{jt} + 0.5\nu_{jt} \times \xi_{jt}$$

With this misspecification, control function generates more bias for β_w than our debiased estimator. Nevertheless, the level of bias fluctuates across different DNN settings, indicating the sensitivity of our model toward the precision of first-stage prediction.

We need to emphasize that our method may be sensitive to overfitting. We want the machine learning to pick up all the variations in the error term correlated by the observables while not following too closely to the data. Sticking to a reasonable tuning parameter λ would be crucial. For example, in a typical elastic net estimation, we fix the turning parameter in front of the Ridge term, λ_R , at a minimal value. Ten-fold cross-validation to choose the tuning parameter for the Lasso penalty, λ_L , could be too time-consuming. Instead, we adopt the Generalized Information Criterion (GIC) developed in Zhang et al. (2010) and Fan and Tang (2013). Given candidates λ_L , GIC calculates $GIC(\lambda_L) = 1/n\{-2\sum_{i=1}^n \ell_i(y_i, x_i, w_i, z_i) + a_n \times \# \text{ of df}\}$, where ℓ_i is the log-likelihood for sample point i , $\# \text{ of df}$ means the degree of freedom¹⁴ and $a_n = \ln(\ln(n)) \times \ln(\# \text{ coefficients})$, for a variety of maximum likelihood models, as in Fan and Tang (2013). We iterate over all candidates of λ_L and choose the λ_L that gives the smallest GIC after all the regressors are standardized¹⁵.

14. Zou et al. (2007) points out that the degree of freedom for Lasso can be estimated consistently by the number of non-zero parameters. The degree of freedom of elastic net takes a slightly complicated form following Zou and Hastie (2005).

15. GIC-based tuning parameter selection would be time-saving since we only need to solve for the corresponding coefficients once for each λ_L . In our application, when we stick to a very small λ_r , though Zou et al. (2007) points out that the number of non-zero parameters is not an unbiased estimator for the elastic net, the discrepancy is neglectable. So we use the number of non-zero parameters for simplicity.

	Not Consider Endogeneity	Linear Control Function	Raw L = 1	Raw L = 5	Raw DNN L = 1	Raw DNN L = 5	Debiased L = 1	Debiased L = 5	Debiased DNN L = 1	Debiased DNN L = 5
β_0	Mean Bias	0.0088	0.0065	0.0062	-0.0270	-0.0257	-0.0020	0.0062	-0.0136	0.0090
	Std	0.1600	0.1500	0.1502	0.1926	0.2255	0.1627	0.1642	0.1630	0.1582
	Median Bias	0.0000	0.0027	0.0024	-0.0321	-0.0270	-0.0056	0.0035	-0.0184	0.0081
	Coverage	0.9540	0.9460	0.9500	0.9480	0.9510	0.9500	0.9470	0.9480	0.9470
β_1	Mean Bias	-0.0020	-0.0034	-0.0059	0.1649	0.2663	-0.0521	-0.0278	-0.0500	0.0420
	Std	0.2274	0.2244	0.2234	0.2559	0.2609	0.2776	0.2902	0.2790	0.2956
	Median Bias	-0.0093	-0.0057	-0.0095	0.1764	0.2657	-0.0697	-0.0471	-0.0543	0.0267
	Coverage	0.4300	0.9440	0.9490	0.9040	0.8370	0.9400	0.9420	0.9460	0.9440
β_2	Mean Bias	0.0010	0.0013	0.0018	-0.0140	-0.0279	0.0126	0.0073	0.0154	-0.0068
	Std	0.0718	0.0706	0.0705	0.0812	0.0864	0.0855	0.0888	0.0838	0.0859
	Median Bias	0.0016	0.0019	0.0035	-0.0150	-0.0265	0.0158	0.0120	0.0179	-0.0057
	Coverage	0.9490	0.9480	0.9440	0.9420	0.9380	0.9390	0.9500	0.9390	0.9420
β_w	Mean Bias	0.0021	0.0032	0.0049	-0.0160	-0.0270	0.0307	0.0157	0.0348	-0.0252
	Std	0.0993	0.0988	0.0982	0.1119	0.1105	0.1249	0.1313	0.1271	0.1365
	Median Bias	0.0044	0.0063	0.0104	-0.0197	-0.0288	0.0336	0.0223	0.0291	-0.0274
	Coverage	0.9480	0.9450	0.9480	0.9540	0.9380	0.9440	0.9480	0.9420	0.9470

Table 1: w_i is determined by a simple linear regression, $n = 2280$, true $\beta_0 = (-0.2, -1, 0.2, 0.7)'$, true $\tau_0 = (3, 2, -0.5, 0.5, 0.5)'$ for $(1, x_i, z_i)$.

	Not Consider Endogeneity	Linear Control Function	Raw Lasso L = 1	Raw Lasso L = 5	Raw Lasso L = 10	Debiased Lasso L = 1	Debiased Lasso L = 5	Debiased Lasso L = 10
β_0	Mean Bias	0.0015	-0.1040	-0.1160	-0.1096	-0.0534	-0.0659	-0.0618
	Std	0.4732	0.2049	0.2260	0.2136	0.1451	0.1433	0.1441
	Median Bias	0.0198	-0.1043	-0.1143	-0.1088	-0.0574	-0.0689	-0.0659
	Coverage	0.9660	0.9280	0.9290	0.9290	0.9300	0.9250	0.9300
β_1	Mean Bias	-0.5455	-0.1972	-0.2002	-0.2002	-0.0682	-0.0833	-0.0868
	Std	0.2715	0.0783	0.0775	0.0771	0.1956	0.1870	0.1917
	Median Bias	-0.5285	-0.1977	-0.1997	-0.2014	-0.0819	-0.0947	-0.0956
	Coverage	0.5090	0.2500	0.2590	0.2460	0.9390	0.9280	0.9260
β_2	Mean Bias	0.1371	0.1886	0.1933	0.1913	0.0473	0.0538	0.0522
	Std	0.1292	0.0435	0.0431	0.0428	0.0754	0.0733	0.0747
	Median Bias	0.1444	0.1914	0.1952	0.1936	0.0491	0.0539	0.0532
	Coverage	0.8290	0.0150	0.0110	0.0100	0.9040	0.8820	0.8910
β_w	Mean Bias	0.2754	-0.0511	-0.0894	-0.0677	-0.0149	-0.0296	-0.0195
	Std	0.1106	0.0609	0.0626	0.0612	0.0992	0.0954	0.0977
	Median Bias	0.2758	-0.0510	-0.0888	-0.0660	-0.0105	-0.0271	-0.0167
	Coverage	0.2800	0.8730	0.6870	0.7920	0.9520	0.9470	0.9510

Table 2: w_i is determined by a simple linear regression, but we observe 100 more irrelevant candidates for instruments, $n = 2280$, true $\beta = (0.8, 0.2, -0.2, 0.7)'$, true $\tau = (3, 2, -0.5, 0.5, 0.5, 0, \dots, 0)'$ for $(1, x_i, z_i)$.

	Not Consider Endogeneity	Control Function	Debiased L = 1	Raw L = 1	Debiased DNN L = 1	Raw DNN L = 1	Debiased DNN L = 5	Raw DNN L = 5	Debiased DNN L = 228	Raw DNN L = 228
β_0	Mean Bias	-0.0072	0.0045	0.0028	0.0031	-0.0012	-0.0372	0.0025	-0.0196	-0.0015
	Std	0.1546	0.1566	0.1437	0.1563	0.1450	0.1596	0.1464	0.1542	0.1438
	Median Bias	-0.0090	0.0023	0.0080	0.0023	0.0023	-0.0389	0.0023	-0.0214	0.0023
	Coverage	0.9440	0.9490	0.9450	0.9480	0.9460	0.9430	0.9410	0.9470	0.9440
β_1	Mean Bias	-0.0083	-0.0049	-0.0016	-0.0056	0.0068	-0.0411	0.0116	-0.0240	0.0054
	Std	0.1709	0.1877	0.1693	0.1863	0.1728	0.2068	0.1660	0.1872	0.1687
	Median Bias	-0.0114	-0.0047	-0.0023	-0.0063	-0.0063	-0.0442	0.0069	-0.0235	0.0033
	Coverage	0.9430	0.9440	0.9470	0.9440	0.9510	0.9590	0.9510	0.9420	0.9440
β_2	Mean Bias	0.0076	0.0019	0.0025	0.0022	0.0035	0.0197	0.0021	0.0113	0.0038
	Std	0.0640	0.0738	0.0634	0.0730	0.0653	0.0814	0.0635	0.0745	0.0637
	Median Bias	0.0091	0.0018	0.0042	0.0014	0.0014	0.0224	0.0039	0.0115	0.0056
	Coverage	0.9290	0.9390	0.9440	0.9350	0.9400	0.9450	0.9370	0.9370	0.9420
β_w	Mean Bias	0.2373	0.0004	-0.0057	-0.0013	-0.0113	-0.0655	-0.0020	-0.0356	-0.0105
	Std	0.1126	0.1531	0.1367	0.1523	0.1423	0.1750	0.1370	0.1542	0.1394
	Median Bias	0.2412	0.0005	-0.0076	-0.0018	-0.0018	-0.0629	-0.0105	-0.0339	-0.0149
	Coverage	0.4410	0.9520	0.9500	0.9510	0.9440	0.9490	0.9480	0.9470	0.9510

Table 3: Linear, but there is slightly misspecification in linear regression model. $n = 2280$. True $\beta = (0.5, 0.4, -0.2, 0.7)'$, and true $\tau = [1.5, 1, -0.25, 0.25, 0.05, 0.05, \dots, 0.05, 0, \dots, 0]'$ for $(1, x_i, z_{i,1}, z_{i,2}, z_{i,3})$

4.1.2 w_{jt} is generated nonlinearly

Suppose we have a nonlinear function for w_{jt} , with interaction terms between observables and unobservables. We do not know the exact functional form for w_{jt} when estimating. Instead, we assume that we know that $(x_{i,1}, z_{i,2}, w_i)'$ would enter as a nonlinear function that a series of basis functions can approximate. We use sieve maximum likelihood with an elastic-net penalty in the first stage.

The endogenous w_i is generated following,

$$w_{jt} = \frac{1}{3}x_{jt,1} + \frac{1}{3}z_{jt,2} + 0.1z_{jt,2}^2 + \frac{1}{6} \times z_{jt,2} \times \xi_{jt} + \frac{1}{3}x_{jt,1} \times \xi_{jt} + \frac{2}{3} + \frac{1}{15} \times N(0,1) \quad (11)$$

Even though (11) contains a nonlinear part, ξ_{jt} is still an implicit function of all other observables. With this, we may have misspecification in the linear control function (Petrin and Train, 2010). We use B-spline as an approximation for the unknown nonlinear approximation on $(x_{i,1}, z_{i,2}, w_i)'$. Appendix C. contains detailed information for implementation.

Table 4 to 5 displays the performance under different sample sizes. Degree-2 B spline generates 216 nonlinear parameters. The GIC selects a large λ_L , approximately between 30 to 80 for most of the simulated samples when $n = 2880$, and an even larger λ_L for $n = 570$. As expected, the linear control function generates biased estimates due to misspecification, especially for β_w . With the elastic net, the results from B-spline are generally close to the true value when $n = 2280$, demonstrating our estimator's superiority in capturing the underlying nonlinearity and irregularity in constructing w_{jt} . The debiased machine learning combined with sample splitting gives the best estimates across all the cases. However, the encouraging performance exacerbates when the sample size decreases to approximately two times the number of parameters. Although the bias of the debiased sample-splitting estimator is still smaller than the linear control function for β_w , the discrepancy reduces, which is unsurprising since those big-data techniques typically enjoy a good performance with sufficient information. A large sample size is required for the underlying pattern to be detected.

As is well-known in the sieve approximation literature, the curse of dimensionality is the main issue preventing the wide application of nonparametric techniques. Consider the case in Equation (11). The need for inverting ξ_{jt} renders all the variables inside as nonlinear. To reduce the large

dimension, we can adopt a partial-linear scheme. Specifically, for a partial linear model like $w_{jt} = x_{jt}^{(1)}\theta_x + z_{jt}^{(1)}\theta_z + f_N(x_{jt}^{(2)}, z_{jt}^{(2)}, \xi_{jt})$, we can rewrite our polynomial approximation as a function on $(w_{jt} - x_{jt}^{(1)}\theta_x - z_{jt}^{(1)}\theta_z, x_{jt}^{(2)}, z_{jt}^{(2)}, \xi_{jt})$, conditional on the $(\hat{\theta}_x, \hat{\theta}_z)$ off the current iteration. Moreover, [Appendix D](#) contains simulations that are omitted in the main text, including the case where we have very weak instrumental variables or misspecification in the model determining w_{jt} .

4.2 Case 2 Continue

Recall from previous analysis that the settings we use in the simulations for zero market shares originate from Gandhi et al. (2019), where sampling error causes the zero shares observed in the aggregated quantities¹⁶. The multinomial Logit gives the underlying choice probabilities, but the observed quantities come from purchases from a finite number of individuals. As discussed in Section 2, zero market shares would prevent linear-IV’s inversion from dealing with endogeneity.

We test a variety of machine-learning algorithms as the first stage, including elastic net for random-coefficient Logit discussed in Horowitz and Nesheim (2021), and the newly-developed ResLogit in Wong and Farooq (2021). Specifically, Wong and Farooq (2021) provides a partial-linear model with multi-layer neural nets approximating the nonlinear part. In their algorithm, the product characteristics of good $j' \in \{0, 1, 2, \dots, J\}$ join together to determine the choice probability of good j . Wong and Farooq (2021) refers to this as a valid and computationally-feasible method to approximate unknown nested structure. Details for implementation can be found in [Appendix C](#). The subsections below contain different settings and tables for simulated results.

4.2.1 w_{jt} is generated linearly

First, let us consider the most basic case where w_i is a linear function, $w_{jt} = x'_{jt}\tau_x + z'_{jt}\tau_z + \xi_{jt} + \nu_{jt}$. $|J| = 5$, with another always-purchased outside option. Both x_{jt} and z_{jt} are generated in a random manner. $x_{jt} = (x_{jt,1}, x_{jt,2}, x_{jt,3})'$ ¹⁷, follows a joint normal distribution, and $z_{jt} = (z_{jt,1}, z_{jt,2}, z_{jt,3})'$ ¹⁸.

16. In contrast to the situation where there is an underlying structure (Dubé et al., 2021) that determines the purchase propensity.

17. The mean of the random distribution x comes from are $x_{j,1} \Rightarrow (3.3086, 3.2674, 3.7254, 4.3582, 3.3889)'$, $x_{j,2} \Rightarrow (4.5168, 4.1152, 4.0296, 3.9356, 3.1744)'$, $x_{j,3} \Rightarrow (3.5973, 3.0627, 4.3560, 4.8070, 4.0289)'$, $j \in J$. The variance are $x_{j,1} \Rightarrow (0.4075, 0.3924, 0.2520, 0.0887, 0.1580)'$, $x_{j,2} \Rightarrow (0.1092, 0.2599, 0.1331, 0.1235, 0.1172)'$, $x_{j,3} \Rightarrow (0.0865, 0.4609, 0.0999, 0.3079, 0.5459)'$, $j \in J$. For each good j , x_j are correlated, and the correlation are the same across j , where $\sigma_{12} = -0.0780$, $\sigma_{13} = 0.0249$, and $\sigma_{23} = -0.0547$.

18. The mean of the random distribution for z_j are all equal to 0, and the variance are, $z_{j,1} \Rightarrow (0.2552, 0.2728, 0.4306, 0.2932, 0.2597)'$, $z_{j,2} \Rightarrow (0.3503, 0.3894, 0.2610, 0.3969, 0.3714)'$, and $z_{j,3} \Rightarrow$

Approximately 35% of the markets experience zero purchased quantities of inside goods under the current parameters. The performance of different estimators can be found in Table 6 and 7. While ignoring the endogeneity generates the largest bias, eliminating the zero shares may also produce some imprecision. All other methods in Table 6 land in reasonable estimates. Bias-correction does not matter since all the first-stage provides \sqrt{N} -consistency.

Table 7 contains a setting similar to Table 2, where we have 300 repetitions of simulations and 100 irrelevant excluded instruments. Even the linear-IV method based on the true market share would give bias up to 0.0439 for β_p . Dropping the zero market share or not bothering with endogeneity generates a larger bias than in Table 6. MLE with control function, on the other hand, performs the best since linearly separability is satisfied. Among all the methods with machine learning as the first stage, biases would be corrected whenever we switch from Raw to Debaised and have sample splits $L > 1$. Their bias is similar to the true market share but with a more significant standard error, as expected¹⁹.

4.2.2 w_{jt} is generated by Bertrand-Nash competition

In industrial organization, the most popular model for the endogenous prices w_{jt} is the Bertrand-Nash competition inside a market. In each market t , the price vector $\mathbf{p}_t = (w_{1t}, w_{2t}, \dots, w_{Jt})'$ is the solution of a nonlinear system,

$$\mathbf{p}_t = \underbrace{\mathbf{z}_t \tau + \boldsymbol{\iota}_t}_{\text{Marginal Cost}} - \left[\boldsymbol{\Omega}_t \circ \left(\frac{\partial \mathbf{S}_t}{\partial \mathbf{p}_t} \right)' \right]^{-1} \mathbf{S}_t$$

where $\mathbf{S}_t = (s_{1t}, s_{2t}, \dots, s_{Jt})'$, $\mathbf{z}_t = (z_{1t}, z_{2t}, \dots, z_{Jt})'$, $\boldsymbol{\iota}_t = (\iota_{1t}, \iota_{2t}, \dots, \iota_{Jt})'$. x_{jt} and z_{jt} follows the similar generating process as the linear case²⁰. $\boldsymbol{\Omega}_t$ is the $J \times J$ ownership matrix with $\Omega_{t,ij} = 1$ if product i and j belong to the same company. $J = 4$ with an outside option. In this setting, approximately 25% of the markets has zero shares. Since we only have three inside goods, zero-

(0.3696, 0.2857, 0.3830, 0.2928, 0.3642)'. For each good j , z_j are correlated, and the correlation are the same across j , where $\sigma_{12} = 0.1205$, $\sigma_{13} = 0.1459$, and $\sigma_{23} = 0.0170$.

19. Because we have linear-IV as the second-stage estimator, the best we can achieve would be the bias when the true share is used. Further studies may extend this to incorporate some instrumental variable selection techniques.

20. The mean of the random distribution x comes from are $x_{j,1} \Rightarrow (3.9195, 2.6808, 3.1705)'$, $x_{j,2} \Rightarrow (3.5025, 2.5102, 3.0119)'$, $x_{j,3} \Rightarrow (3.7818, 3.9186, 3.0944)'$, $j \in J$. The variance are $x_{j,1} \Rightarrow (0.1584, 0.1908, 0.3658)'$, $x_{j,2} \Rightarrow (0.3578, 0.1866, 0.3923)'$, $x_{j,3} \Rightarrow (0.1725, 0.1889, 0.2984)'$, $j \in J$. For each good j , x_j are correlated, and the correlation are the same across j , where $\sigma_{12} = 0.0381$, $\sigma_{13} = 0.0342$, and $\sigma_{23} = 0.0040$. The mean of the random distribution for z_j are all equal to 0, and the variance are, $z_{j,1} \Rightarrow (0.1859, 0.1815, 0.2769)'$, $z_{j,2} \Rightarrow (0.3575, 0.1967, 0.2845)'$, and $z_{j,3} \Rightarrow (0.1731, 0.2381, 0.2031)'$. For each good j , z_j are correlated, and the correlation are the same across j , where $\sigma_{12} = 0.1261$, $\sigma_{13} = 0.0710$, and $\sigma_{23} = 0.1804$.

share on either of them could be an important concern and dropping any point may hurt the performance. The equilibrium prices are generated by fixed-point iteration.

Kim and Petrin (2019) shows that under multi-product Nash Bertrand competition, the price of good j at market t can be expressed as a good-specific function such that,

$$p_{jt} = p_j(\mathbf{x}_t, \mathbf{z}_t^{(2)}, \boldsymbol{\nu}_t + \mathbf{z}_t^{(1)}, \nu_j(\boldsymbol{\xi}_t)) \quad (12)$$

if there exists a $z_{jt}^{(1)}$ as a perfect substitute to cost shocks $\boldsymbol{\nu}_t$. The function p_j is specific to a commodity j . Kim and Petrin (2019) illustrates that by (12), one can obtain $\tilde{\xi}_{jt}$ by inverting the commodity-specific p_j that $\tilde{\xi}_{jt} = \nu_j(\boldsymbol{\xi}_t) = p_j^{-1}(\mathbf{x}_t, \mathbf{z}_t^{(2)}, \boldsymbol{\nu}_t + \mathbf{z}_t^{(1)}, p_{jt})$.

For the simulations below, we assume to know the cost shocks $\boldsymbol{\nu}_t + \mathbf{z}_t^{(1)}$. This simplified assumption comes from the supply-side identification results in Berry and Haile (2014), where they point out that the cost shock can be identified without specifying a supply-side model under the presence of an index variable $z_{jt}^{(1)}$ ²¹. Due to the lack of a computationally-feasible method for supply-side NPTV estimation, for simplicity, we omit the procedure of estimating and treat the $\boldsymbol{\nu}_t + \mathbf{z}_t^{(1)}$ as observed.

We attempt several methods for approximating the market shares. First, we use the ResLogit (Wong and Farooq, 2021). Because all the characteristics for the $|J| - 1$ commodities launch into the function p_j , a sieve approximation based on tensor-product becomes almost infeasible even in a parsimonious setting²². Hence, we use the first and second-order interactions, which may not give a good approximation but is computationally feasible.

Table 8 contains the simulation results. All the methods except the case with true market share have some misspecification. The control function incorrectly retains the separable linearity. ResLogit restricts how the nonlinearity enters into the model. Polynomials generated simply from interaction lose the usual good properties in B-spline or Chebyshev polynomials. In any case, considering endogeneity would never make things worse. If our machine learning estimation is not flexible enough, the estimated β would still be biased and may be worse than just ignoring the zero

21. $\boldsymbol{\nu}_t + \mathbf{z}_t^{(1)}$ can be expressed as a function of s_t and p_t . Then, one can construct moment conditions and estimate via NPIV methods.

22. Consider we have the one-dimension x_{jt} , $z_{jt}^{(2)}$ and $\boldsymbol{\nu}_t + \mathbf{z}_t^{(1)}$. With $J = 4$, p_j is a flexible function of $(x_{1t}, x_{2t}, x_{3t}, z_{1t}^{(2)}, z_{2t}^{(2)}, z_{3t}^{(2)}, \boldsymbol{\nu}_t + \mathbf{z}_t^{(1)}, \boldsymbol{\nu}_t + \mathbf{z}_t^{(1)}, \boldsymbol{\nu}_t + \mathbf{z}_t^{(1)}, p_{jt})$. For degree-2 Chebyshev polynomials, there are $3^{10} = 59049$ bases.

market shares. Methods like the Res-Logit with Layer 1 could be worse than the linear control function, though Res-Logit with Layer 1 itself is more flexible than the linear control function. Nevertheless, the debiased moments would unambiguously improve performance, especially for the polynomial case with the Lasso penalty.

Overall, machine learning itself may not be a panacea. We regard a careful implementation of those methods as a necessity for the success of our estimator. Moreover, researchers need to propose a feasible method to solve the high-dimensional problem before the identification results in Kim and Petrin (2019) can be applied empirically to a supply-side with Bertrand-Nash competition.

5 Advantages and Problems

In this section, we discuss our estimator’s advantages and potential issues. Our method treats the control function method (Petrin and Train, 2010) as a special case. Assuming that w_i is determined by a low-dimensional linear function, our estimated probability is \sqrt{N} -consistent, and the plug-in estimator would be enough, similar to the control function. On the other hand, our method does not stipulate the exact formula of w_{jt} . w_{jt} can even be determined by an implicit function. All we need is a reasonable probability estimation. The analytical construction of orthogonality moments equips the first-stage probability estimation with flexible machine-learning methods, which may be hard to achieve in the control function.

As for the problems, numerical instability is the first issue we encounter. Our method includes inverting a system of probabilities, shares, or propensity scores. If probabilities are close to 0 or 1, they would have enormous leverage in the second stage. For instance, consider the case where ε follows a standard normal distribution, and the deterministic part $x\beta$ takes a very large or small value. Then the true probabilities would be concentrated near 0 or 1, generating unstably large inverted values of both $\Lambda^{-1}(\cdot)$ and $\Lambda'(\Lambda^{-1}(\cdot))$. This problem cannot be relieved in a large sample.

This numerical instability leads to a data-dependent performance of our method, which is common in the well-developed inverse propensity score weighted estimators²³. Thus, our estimator works better when the difference in product characteristics enters as x rather than the characteristics

23. This is why the Assumption 2 is always a pre-requirement for these types of estimators. Besides, Singh and Sun (2021) refer to automatic debiased machine learning techniques and double robustness to circumvent the need to estimate the propensity score directly. However, their methods are not applicable in our case. We still need to bear the risk of inserting estimated probabilities.

themselves. Also, performance may improve if more structures are included in the first stage.

Secondly, as discussed in section 3, our estimator does not have double-robustness as in Bakhitov (2022), which requires linearity in γ and the correction term simultaneously. This lack of double-robustness triggers a high-standard requirement in the first stage of propensity score estimation. Not enough flexibility may cause it to be inadequate for endogeneity, while too much flexibility may result in overfitting. For the sieve-based polynomial approximation, tensor-based parameters may quickly render it impractical. Besides, researchers usually resort to nonparametric or machine learning methods when they are reluctant to impose model restrictions, making justifying those techniques' performance hard.

In summary, flexibility can lead to instability, especially when comparing our estimator to the linear control function. The first-stage machine learning method needs special attention and careful tuning, in order to obtain good performance.

6 Illustrative Application

Now a simple application of our method is considered in the context of coal purchases by US power plants from 2011 to 2015. We use the public data available on the EIA website. EIA form 923²⁴ contains the monthly transaction-level coal procurement data for all the regulated power plants in the US. A transaction record carries some detailed information, for instance,

“Colbert, a power plant operated by Tennessee Valley Authority, on January 2012, purchased in the spot market coals from the Black Thunder mine at Gillette, WY, owned by Arch Coal. The purchase involves 297.0 short tons, with average heat at 18.000 MMBtu per short ton, 0.23% sulfur, and 3.8% ash content. In total, it paid 247.8 cents per MMBtu for the coal to deliver to its site by railroad.”

Note that the fuel cost is aggregated, including the free-on-board price and the transportation fees. Coals are traded either by long-term contracts or in the spot market²⁵. Train and Wilson (2011) shows that power plants' purchasing behavior falls into multiple discrete-continuous choice models²⁶. For simplicity, we only focus on the discrete-choice part – power plants' binary decision

24. EIA Form 923 is available here: <https://www.eia.gov/electricity/data/eia923/>.

25. Long-term contracts and spot market co-exist in the coal transactions. Long-term contracts last at least a year, and power plants may order coal delivery every month. Delivery prices are adjusted following a pre-determined scheme in the contract. Further information can be found in Joskow (1985).

26. Coal-fired power plants decide which source of coals they purchase from and how much to purchase from each source, which perfectly fits in the scope of multiple discrete-continuous choices (Bhat, 2005)

on purchasing from a specific coal source j in the spot market, given the long-term contracts they currently hold as the outside options. Moreover, we aggregate a large number of coal varieties into six composite products grounded on their place of origin – (i) tier 1 SPRB; (ii) tier 2 SPRB; (iii) Other PRB coal²⁷; (iv) Appalachia; (v) Other western; (vi) Mid-continent²⁸. The choice set of plant i in the spot market is assumed constant across time and constituted by the purchase history. Specifically, if plant i purchased j in April 2014, then j is in the choice set for plant i for the entire period.

Besides EIA-923, we also refer to EIA-860 for plant-related attributes, like the nameplate capacity and the installation of capital-intensive environmental equipment²⁹. Table 9 contains the averaged observable characteristics, showing that coals from distinct regions have different characteristics. For instance, tier 1 SPRB, compared with the rest of the PRB region, has higher energy content and lower sulfur content. Correspondingly, the delivery price for tier 1 SPRB coal is higher.

For each plant-month pair, we model the behavior of each plant as a binary choice, with j denoting the inside option from the spot market and the outside option with subscript 0 referring to the quantity-averaged characteristics from the currently-hold long-term contract. We construct a model following a Probit setting where a power plant’s purchase decision y_{ijt} ,

$$y_{ijt} = 1\{X_{ijt}\beta + \xi_{ijt} + \varepsilon_{ijt} \geq X_{i0t}\beta + \xi_{i0t} + \varepsilon_{i0t}\} \quad (13)$$

If $\varepsilon_{i0t} - \varepsilon_{ijt}$ follows a standard normal distribution, then the probability of j is being chosen given the outside option is $\Phi((X_{ijt} - X_{i0t})\beta + (\varepsilon_{ijt} - \varepsilon_{i0t}))$, which depends on the difference in characteristics. One important drawback of this individual transaction data is that we do not observe the proposed trading prices whenever the transaction breaks down since the power plants only need to report their successful purchases. While the physical characteristics of coals stay stable, the spot price needs to

27. PRB coal refers to the low-sulfur content subbituminous coal produced in WY or MT along the powder river basin. Tier 1 SPRB includes the largest coal mines in SPRB, like North Antelope Rochelle, Black Thunder, Antelope, and Jacobs Ranch. Tier 2 SPRB locates north of Tier 1, indicates coal mines like Caballo, Belle Ayr, Coal Creek, and Cordero Rojo. Tier 1 and 2 SPRB coal have higher quality than other PRB coal.

28. Appalachia refers to the high-sulfur bituminous coal produced along the Appalachia mountain. Other western mainly points to the coal produced in Southern Wyoming and Rockies. Coals produced in Illinois Basin are ascribed as Mid-continent.

29. The environmental equipment mainly refers to the Wet Scrubber Flue Gas Desulfurization (abbreviated as scrubber), an apparatus used to remove the sulfur in the gas waste, helping the plant to meet the requirement of the Clean Air Act. Scrubber installation is expensive, while its operation is less costly (Cicala, 2015). We expect a power plant better equipped with scrubbers to be less sensitive to sulfur content.

be imputed, following the appendix in Jha (2022)³⁰. Simply speaking, the price is approximately calculated by a reduced-form forecast problem of log price on other characteristics and the price from the closest plant in physical distance. Denote that plant using superscript D , and then the reduced-form regression takes the form below³¹,

$$\ln(p_{ijt}) = \kappa_{jt} + \mu_i + \ln(p_{ijt}^D)\varrho_1 + X_{2,ijt}\varrho_2 + \sum_{r=1}^4 [X_{2,ijt,r} - X_{2,ijt,r}^D]^2 \varrho_{3,r} + v_{ijt} \quad (14)$$

Next, we implement several models and compare them with the same coal-purchase data set. x_{ijt} contains transaction-specific characteristics, like the difference in heat content, ash content, sulfur content, and sulfur content interacting with the scrubber installation. The potentially endogenous variable is the ln price per ton. The excluded instruments we refer to are the physical distance calculated by the Haversine formula from the power plant to the mine county and the proportion of coal generating capacity inside a plant.

In Table 10, Model (1) is the Probit without considering the endogeneity of p_{ijt} ; Model (2) is the linear control-function procedure³²; Model (3) is our estimator, where the first stage machine learning is a tensor-product-based degree-2 Chebyshev polynomials, and $\lambda_R = 0.000001$, λ_L comes from GIC as in the simulation; Model (4), instead, has the first stage as the DNN with three layers and (30, 10, 2) nodes in each layer. $L = 10$ for both Model (3) and (4). Both of them are truncated to eliminate the extremely large correction terms.

Table 10 shows the estimated results from different methods. The price coefficient would be positive without explicitly accounting for the endogeneity, while all other methods give a negative price coefficient. Model (2) and Model (3) gives similar estimated parameters for $\ln(p_{ijt})$, which indicates that there may not be much nonlinearity in the current model. Model (4) gives a larger estimate on the price coefficient³³. As expected, the standard error increases due to the introduction of machine learning first stage.

30. When a full supply-side model is available, those missing prices can be calculated using the equilibrium of a structural model (Song, 2022). However, we do not specify a supply model in this illustration, so the missing price must be imputed separately and outside the system (Jha, 2022).

31. where $X_{2,ijt}$ contains heat, sulfur, ash content, and Haversine distance.

32. The procedure is the same as <https://www.stata.com/manuals/rivprobit.pdf>

33. Note that the variables in Table 10 are all the difference of product j with the outside option, like $\ln(p_{ijt}) - \ln(p_{i0t})$. So even a large coefficient may not result in a significant effect.

7 Conclusions

In this study, we propose a new machine-learning-based semi-parametric estimator. We demonstrate the applicability of the binary discrete choice model with endogenous variables and the Logit model with zero market shares. We show that in those cases, the correction term for Neyman orthogonality can be calculated analytically with the estimated probabilities from the first stage. By construction, first-stage error up to $o(N^{-1/4})$ would not affect the asymptotic distribution of the second-stage parameter estimates. Moreover, it can be easily implemented in standard statistical software (Stata, Python, R, etc.). Simulations under a collection of settings demonstrate that our estimator is a flexible alternative to the control function, especially when there is unknown non-linearity in determining endogenous variables. Furthermore, we provide an illustrative application using the individual coal transaction data from 2011 to 2015, where the decision of power plants is simplified as a binary choice inside the spot market in light of the currently-hold long-term contracts.

Further research directions may include the possibility of finding this type of analytical correction term for a broader class of models; or the extension of this type of method into a more structural model like the original multiple discrete-continuous models.

	Not Consider Endogeneity	Linear Control Function	Raw EN L = 1	Raw EN L = 5	Raw EN L = 10	Debiased EN L = 1	Debiased EN L = 5	Debiased EN L = 10
β_0	Mean Bias	0.0081	0.0189	0.0226	0.0215	-0.0019	0.0249	0.0213
	Std	0.4661	0.2156	0.2176	0.2132	0.1904	0.2234	0.2192
	Median Bias	0.0347	0.0179	0.0303	0.0270	0.0025	0.0245	0.0178
	Coverage	0.9660	0.9580	0.9460	0.9470	0.9590	0.9600	0.9530
β_1	Mean Bias	-0.1161	-0.1551	-0.1616	-0.1568	-0.0560	-0.0580	-0.0584
	Std	0.1979	0.1447	0.1315	0.1256	0.1303	0.1665	0.1599
	Median Bias	-0.1041	-0.1556	-0.1696	-0.1612	-0.0574	-0.0611	-0.0634
	Coverage	0.8900	0.9050	0.7720	0.7580	0.9610	0.9570	0.9440
β_2	Mean Bias	0.0049	0.1039	0.1160	0.1086	0.0193	0.0191	0.0197
	Std	0.1161	0.0717	0.0749	0.0738	0.0727	0.0827	0.0827
	Median Bias	0.0014	0.1038	0.1206	0.1128	0.0208	0.0256	0.0232
	Coverage	0.9450	0.6750	0.6340	0.6580	0.9570	0.9610	0.9510
β_w	Mean Bias	0.2128	-0.0120	-0.0297	-0.0163	0.0370	0.0105	0.0190
	Std	0.0849	0.1278	0.1394	0.1350	0.1370	0.1518	0.1496
	Median Bias	0.2090	-0.0101	-0.0245	-0.0119	0.0348	0.0065	0.0166
	Coverage	0.2990	0.9580	0.9410	0.9540	0.9640	0.9540	0.9490

Table 4: Nonlinear Generalization of w_i . With GIC choice of λ . $n = 2280$. Polynomial degree = 2, and number of parameters in the machine learning first stage = 216. True $\beta = (-0.3, -0.4, -0.2, 0.7)'$.

	Not Consider Endogeneity	Linear Control Function	Raw EN L = 1	Raw EN L = 5	Raw EN L = 10	Debiased EN L = 1	Debiased EN L = 5	Debiased EN L = 10
β_0	Mean Bias	-0.0133	0.0506	0.0501	0.0563	0.0144	0.0313	0.0404
	Std	0.5192	0.3059	0.3970	0.3235	0.0190	0.3625	0.2897
	Median Bias	-0.0203	0.0458	0.0638	0.0595	-0.2856	0.0450	0.0479
	Coverage	0.9480	0.9440	0.9710	0.9480	0.9510	0.9760	0.9470
β_1	Mean Bias	-0.1690	-0.2436	-0.2659	-0.2517	-0.0728	-0.0785	-0.0740
	Std	0.2713	0.1438	0.1736	0.1561	-0.0849	0.2116	0.2017
	Median Bias	-0.1918	-0.0777	-0.2809	-0.2662	0.3272	-0.0854	-0.0830
	Coverage	0.9260	0.9430	0.6930	0.6070	0.9420	0.9660	0.9580
β_2	Mean Bias	0.0051	0.1710	0.1788	0.1749	0.0441	0.0465	0.0451
	Std	0.1660	0.0674	0.0637	0.0688	0.0497	0.0971	0.1008
	Median Bias	0.0051	0.1815	0.1860	0.1835	-0.1559	0.0498	0.0492
	Coverage	0.9560	0.9530	0.1420	0.1980	0.9190	0.9200	0.9240
β_w	Mean Bias	0.2390	0.1628	-0.2398	-0.2136	-0.0963	-0.1352	-0.1191
	Std	0.1633	0.1561	0.1796	0.1652	-0.0911	0.1870	0.1728
	Median Bias	0.2365	0.1616	-0.2467	-0.2149	0.6037	-0.1376	-0.1133
	Coverage	0.6950	0.8290	0.7490	0.7530	0.9030	0.9180	0.9050

Table 5: Nonlinear Generalization of w_i . With high-dimensional parameter space in the machine learning first stage, compared with the sample size $n = 570$.

	True Share	Drop 0	MLE No endogeneity	MLE with control function	Linear MLE first stage	Debiased Res-Logit Layer 1 ¹	Raw Res-Logit Layer 1
β_0	Mean Bias	-0.1477	-0.8129	-0.0080	-0.0078	-0.0013	-0.0014
	Std	0.0357	0.0322	0.0306	0.0375	0.0540	0.0378
	Median Bias	-0.0019	-0.8130	-0.0079	-0.0085	-0.0011	-0.0013
	Coverage	0.9460	0.2380	0.9360	0.9470	0.9488	0.9488
β_1	Mean Bias	-0.1065	-0.8847	-0.0043	-0.0042	-0.0008	0.0005
	Std	0.0353	0.0519	0.0268	0.0370	0.0497	0.0373
	Median Bias	-0.0026	-0.1097	-0.0037	-0.0044	-0.0022	-0.0012
	Coverage	0.9580	0.4420	0.9510	0.9580	0.9488	0.9549
β_2	Mean Bias	-0.0004	-0.0448	0.0010	0.0008	-0.0005	0.0001
	Std	0.0195	0.0214	0.0192	0.0198	0.0243	0.0199
	Median Bias	0.0005	-0.0449	0.0010	0.0009	-0.0008	-0.0001
	Coverage	0.9520	0.7540	0.9460	0.9460	0.9448	0.9478
β_p	Mean Bias	0.0006	0.4046	0.0030	0.0031	0.0011	0.0006
	Std	0.0148	0.0087	0.0101	0.0156	0.0218	0.0157
	Median Bias	0.0011	0.4046	0.0036	0.0036	0.0013	0.0011
	Coverage	0.9470	0.1970	0.9380	0.9510	0.9559	0.9478

¹ Out of 1000 repetitions, there are three simulated samples that do not have convergence for the Res-Logit first stage. Those are excluded from the outcomes in the table.

Table 6: Zero market share with linear pricing. True $\beta = (3, 2, -0.6, -1.2)'$, $\tau = (\tau'_x, \tau'_z)' = (1, 0.4, -0.3, 0.5, 1.2, 0.4)$.

	True Share	Drop 0	MLE No endogeneity	MLE with control function	Debiased MLE Lasso L = 1	Raw MLE Lasso L = 1	Debiased MLE Lasso L = 10	Raw MLE Lasso L = 10
β_0	Mean Bias	-0.5240	-0.7214	-0.0560	-0.0913	-0.1071	-0.0896	-0.1023
	Std	0.0571	0.0306	0.0310	0.0533	0.0393	0.0534	0.0393
	Median Bias	-0.5180	-0.7224	-0.0558	-0.0868	-0.1063	-0.0845	-0.1015
	Coverage	0.0000	0.0000	0.5767	0.6267	0.2067	0.6333	0.2400
β_1	Mean Bias	-0.4352	-0.8013	-0.0241	-0.0959	-0.1108	-0.0944	-0.1034
	Std	0.0534	0.0236	0.0264	0.0509	0.0396	0.0511	0.0397
	Median Bias	-0.4345	-0.8015	-0.0231	-0.0948	-0.1085	-0.0936	-0.1005
	Coverage	0.0000	0.0000	0.8500	0.5300	0.1933	0.5400	0.2933
β_2	Mean Bias	0.0534	-0.0730	0.0227	-0.0159	-0.0180	-0.0159	-0.0153
	Std	0.0181	0.0216	0.0187	0.0231	0.0195	0.0232	0.0196
	Median Bias	0.0520	-0.0734	0.0227	-0.0167	-0.0179	-0.0172	-0.0158
	Coverage	0.1567	0.0833	0.7500	0.9200	0.8433	0.9167	0.8767
β_p	Mean Bias	0.1855	0.3722	0.0155	0.0490	0.0561	0.0485	0.0526
	Std	0.0248	0.0086	0.0108	0.0234	0.0182	0.0235	0.0182
	Median Bias	0.1851	0.3720	0.0145	0.0499	0.0550	0.0491	0.0515
	Coverage	0.0000	0.0000	0.6933	0.4300	0.1567	0.4367	0.1833

Table 7: Zero market share with linear pricing, and with a bunch of irrelevant instruments. True $\beta = (3, 2, -0.6, -1.2)'$, $\theta = (\tau'_x, \tau'_z)'$ = $(1, 0.4, -0.3, 0.5, 1.2, 0.4, 0, \dots, 0)'$.

	True Share	Drop 0	MLE No endogeneity	MLE control function	Debiased Res-Logit Layer 1 L = 1	Raw Res-Logit Layer 1 L = 1
β_1	Mean Bias	-0.2283	-1.0394	-0.1627	-0.4302	-1.0975
	Std	0.2214	0.0088	0.1111	0.1515	0.1007
	Median Bias	-0.2554	-1.0400	-0.1768	-0.4429	-1.1048
	Coverage	0.8100	0.0000	0.6533	0.1967	0.0000
β_p	Mean Bias	0.2569	1.2073	0.3906	0.5645	1.2334
	Std	0.2201	0.0084	0.1044	0.1427	0.0954
	Median Bias	0.2846	1.2075	0.4045	0.5778	1.2400
	Coverage	0.7700	0.0000	0.0433	0.0300	0.0000
	Debiased Res-Logit Layer 5 L = 1	Raw Res-Logit Layer 5 L = 1	Debiased Res-Logit Layer 5 L = 10	Raw Res-Logit Layer 5 L = 10	Debiased Poly L = 1	Raw Poly L = 1
β_1	Mean Bias	-0.1498	-0.1648	-0.1712	-0.1776	-1.4436
	Std	0.1806	0.1975	0.1469	0.3291	0.1359
	Median Bias	-0.1674	-0.1772	-0.1789	-0.2064	-1.4253
	Coverage	0.8733	0.8896	0.7726	0.9367	0.0000
β_p	Mean Bias	0.2461	0.2644	0.2688	0.2946	1.4768
	Std	0.1716	0.1878	0.1401	0.3122	0.1300
	Median Bias	0.2608	0.2759	0.2737	0.3153	1.4598
	Coverage	0.6900	0.4033	0.6890	0.9000	0.0000

Table 8: Market share, pricing by Bertrand Competition. True $\beta = (1, -2)'$, and true $\tau = (\tau'_x, \tau'_z) = (1, 1, -0.5)'$.

	Tier 1 SPRB	Tier 2 SPRB	Other PRB	Other Western	Appalachia	Mid-continent
MMBtu/ton	17.6613 (0.1834)	16.8586 (0.1767)	16.9809 (0.6729)	24.5660 (1.2782)	18.5145 (3.6250)	21.9418 (3.2108)
Sulfur % weight	0.2504 (0.0499)	0.3058 (0.0208)	0.3365 (0.0939)	1.9655 (1.0480)	0.6898 (0.2092)	2.7152 (0.6864)
Ash % weight	4.9914 (0.4347)	5.4268 (0.3428)	5.0946 (1.1784)	11.0948 (3.7780)	12.4929 (5.2866)	9.9531 (3.7480)
Cents MMBtu (Real)	34.0770 (7.2455)	28.9759 (7.6250)	28.0718 (8.2127)	74.7183 (19.8008)	37.4142 (22.3650)	53.2300 (13.3230)
Tons(000)	138.2633 (147.3759)	141.4755 (139.1950)	113.2860 (130.9553)	82.5121 (108.1117)	187.2692 (195.8328)	107.8122 (128.1342)
Proportion of Scrubber	0.4931 (0.5277)	0.4796 (0.5005)	0.5834 (0.5005)	0.6453 (0.4888)	0.7861 (0.5853)	0.7404 (0.4233)
# of deliveries	6261	1713	3102	6117	2434	4814
Long-term	0.81	0.83	0.81	0.67	0.82	0.74
Spot	0.19	0.17	0.19	0.33	0.18	0.26

Standard errors in parentheses.

Table 9: Summary statistics for the coals originated from different regions.

	Model (1)	Model (2)	Model (3)	Model (4)
Constant	0.6776*** (0.0982)	-0.1983*** (0.0105)	0.2092 (0.1583)	0.0348 (0.1857)
Heat	0.0925*** (0.0091)	0.5969*** (0.0225)	0.6913*** (0.1247)	0.8420*** (0.1400)
Sulfur	-0.3489*** (0.0229)	-0.2931*** (0.0225)	-0.4633*** (0.0527)	-0.4959*** (0.0610)
Ash	0.0029 (0.0044)	0.0757*** (0.0039)	0.0819*** (0.0181)	0.1236*** (0.0209)
$\ln(p_{ijt})$	0.1177* (0.0691)	-6.2658*** (0.1621)	-6.6400*** (1.4089)	-8.5312*** (1.5853)
Sulfur \times Scrubber	0.0409** (0.0208)	-0.0510*** (0.0162)	-0.0596 (0.0370)	-0.0402 (0.0465)
$\ln(Q_{i0t})$	-0.1143*** (0.0086)	-0.0419*** (0.0090)	-0.1133*** (0.0114)	-0.1093*** (0.0133)
N	18149	18149	18141	18141

Standard errors in parentheses.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 10: Coefficients estimated in the illustrative application.

References

- Akerberg, D., K. Caves, and G. Frazer. 2015. “Identification Properties of Recent Production Function Estimators.” *Econometrica* 83 (6): 2411–2451.
- Alquier, P., V. Cottet, and G. Lecue. 2019. “Estimation Bounds and Sharp Oracle Inequalities of Regularized Procedures with Lipschitz Loss Functions.” *The Annals of Statistics* 47 (4): 2117–2144.
- Angrist, J., and A. Krueger. 2001. “Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments.” *Journal Of Economic Perspectives* 15 (4): 69–85.
- Bakhitov, E. 2022. “Automatic Debiased Machine Learning in Presence of Endogeneity.” Working Paper, https://edbakhitov.com/assets/pdf/jmp_edbakhitov.pdf.
- Belloni, A., V. Chernozhukov, and C. Hansen. 2011. “Lasso Methods for Gaussian Instrumental Variables Models.” Working Paper, <https://arxiv.org/pdf/1012.1297.pdf>, February.
- Berry, S., and P. Haile. 2014. “Identification in Differentiated Products Markets Using Market Level Data.” *Econometrica* 82 (5): 1749–1797.
- Bhat, C. 2005. “A Multiple Discrete–Continuous Extreme Value Model: Formulation and Application to Discretionary Time-Use Decisions.” *Transportation Research Part B* 39 (8): 274–303.
- . 2008. “The Multiple Discrete-Continuous Extreme Value (MDCEV) Model: Role of Utility Function Parameters, Identification Considerations, and Model Extensions.” *Transportation Research Part B* 42 (3): 274–303.
- . 2018. “A New Flexible Multiple Discrete–Continuous Extreme Value (MDCEV) Choice Model.” *Transportation Research Part B: Methodological* 110:261–279.
- Bickel, P., Y. Ritov, and A. Tsybakov. 2009. “Simultaneous Analysis of Lasso and Dantzig Selector.” *The Annals of Statistics* 37 (4): 1705–1732.
- Blazère, M., J. Loubes, and F. Gamboa. 2014. “Oracle Inequalities for a Group Lasso Procedure Applied to Generalized Linear Models in High Dimension.” *IEEE Transactions on Information Theory* 60 (4): 637–660.

- Bontemps, C., and C. Nauges. 2017. “Endogenous Variables in Binary Choice Models: Some Insights for Practitioners.” Working Paper, https://www.tse-fr.eu/sites/default/files/TSE/document/doc/wp/2017/wp_tse_855.pdf, October.
- Burlig, F., C. Knittel, D. Rapson, M. Reguant, and C. Wolfram. 2020. “Machine Learning from Schools about Energy Efficiency.” *Journal of the Association of Environmental and Resource Economists* 7 (6).
- Caner, M., and A. Kock. 2019. “High Dimensional Linear GMM.” Working Paper, <https://arxiv.org/pdf/1811.08779.pdf>.
- Cattaneo, M., M. Jansson, and X. Ma. 2019. “Two-Step Estimation and Inference with Possibly Many Included Covariates.” *The Review of Economic Studies* 86 (3): 1095–1122.
- Cha, J., H. Chiang, and Y. Sasaki. 2021. “Inference in High-dimensional Regression Models without the Exact or L^p Sparsity.” Working Paper, <https://arxiv.org/pdf/2108.09520.pdf>, August.
- Chen, J., D. Chen, and G. Lewis. 2021. “Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models.” Working Paper, <https://arxiv.org/pdf/2011.06158.pdf>.
- Chen, X., and H. White. 1999. “Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators.” *IEEE Transactions on Information Theory* 45 (2): 682–691.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. 2018. “Double/debiased Machine Learning for Treatment and Structural Parameters.” *The Econometrics Journal* 21 (1): C1–C68.
- Chernozhukov, V., J. Escanciano, H. Ichimura, W. K. Newey, and J. Robins. 2022a. “Locally Robust Semiparametric Estimation.” *Econometrica* 90 (4): 1501–1535.
- Chernozhukov, V., W. Newey, V. Quintas-Martinez, and V. Syrgkanis. 2021. “Automatic Debiased Machine Learning via Neural Nets for Generalized Linear Regression.” Working Paper, <https://arxiv.org/pdf/2104.14737.pdf>.
- Chernozhukov, V., W. Newey, and R. Singh. 2022b. “Automatic Debiased Machine Learning of Causal and Structural Effects.” *Econometrica* 90 (3): 967–1027.
- Cicala, S. 2015. “When Does Regulation Distort Costs? Lessons from Fuel Procurement in US Electricity Generation.” *American Economic Review* 105 (1): 411–444.

- Cosslett, S. 1983. "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model." *Econometrica* 51 (3): 765–782.
- Dong, Y., and A. Lewbel. 2015. "A Simple Estimator for Binary Choice Models with Endogenous Regressors." *Econometric Reviews* 34 (1-2): 82–105.
- Dube, A., J. Jacobs, S. Naidu, and S. Suri. 2020. "Monopsony in Online Labor Markets." *American Economic Review: Insights* 2 (1): 33–46.
- Dubé, J., A. Hortaçsu, and J. Joo. 2021. "Random-Coefficients Logit Demand Estimation with Zero-Valued Market Shares." *Marketing Science* 40 (4): 637–660.
- Dung, V., and T. Tjahjowidodo. 2017. "A Direct Method to Solve Optimal Knots of B-Spline Curves: An Application for Non-uniform B-spline Curves Fitting." *PLoS ONE* 12 (3): 1–24.
- Fan, Y., and C. Tang. 2013. "Tuning Parameter Selection in High Dimensional Penalized Likelihood." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 75 (3): 531–552.
- Farrell, M., T. Liang, and S. Misra. 2021. "Deep Neural Networks for Estimation and Inference." *Econometrica* 89 (1): 181–213.
- Farronato, C., A. Fradkin, B. Larsen, and E. Brynjolfsson. 2020. "Consumer Protection in an Online World: An Analysis of Occupational Licensing." NBER Working Paper, https://www.nber.org/system/files/working_papers/w26601/w26601.pdf, January.
- Gandhi, A., Z. Lu, and X. Shi. 2019. "Estimating Demand for Differentiated Products with Zeros in Market Share Data." Working Paper, https://www.ssc.wisc.edu/~xshi/research/gandhi_lu_shi.pdf.
- Horowitz, J., and L. Nesheim. 2021. "Using Penalized Likelihood to Select Parameters in a Random Coefficients Multinomial Logit model." *Journal of Econometrics* 222 (1): 44–55.
- Jha, A. 2022. "Regulatory Induced Risk Aversion in Coal Contracting at US Power Plants: Implications for Environmental Policy." *Journal of the Association of Environmental and Resource Economists* 9 (1): 51–78.
- Joskow, P. 1985. "Vertical Integration and Long Term Contracts: The Case of Coal Burning Electric Generating Plants." *Journal of Law, Economics, and Organization* 1 (1): 33–89.

- Kim, K., and A. Petrin. 2019. “Control Function Corrections for Unobserved Factors in Differentiated Product Models.” Working Paper, <https://drive.google.com/file/d/1oOuegNyH2QtvUi6X3UfyPiXGjGyN01C3/view>.
- Lennon, C., E. Rubin, and G. Waddell. 2021. “What Can We Machine Learn (too much of) in 2SLS? Insights from a Bias Decomposition and Simulation.” Working Paper, <http://edrub.in/Papers/draft-mliv.pdf>, May.
- Levinsohn, J., and A. Petrin. 2003. “Estimating Production Functions Using Inputs to Control for Unobservables.” *Review of Economic Studies* 70 (2): 317–341.
- Lewbel, A. 2012. “An Overview of the Special Regressor Method.” Technical Report, <http://fmwww.bc.edu/EC-P/wp810.pdf>.
- Lewbel, A., Y. Dong, and T. Yang. 2012. “Comparing Features of Convenient Estimators for Binary Choice Models with Endogenous Regressors.” *Canadian Journal of Economics* 45 (3): 809–829.
- Li, F., L. E. Thomas, and F. Li. 2019. “Addressing Extreme Propensity Scores via the Overlap Weights.” *American Journal of Epidemiology* 188 (1): 250–257.
- Lin, W., and J. Wooldridge. 2017. “Binary and Fractional Response Models with Continuous and Binary Endogenous Explanatory Variables.” Working Paper, http://www.weilinmetrics.com/uploads/5/1/4/0/51404393/chap1_10302017_weilin.pdf, November.
- Nevo, A. 2000. “A Practitioner’s Guide to Estimation of Random-Coefficients Logit Models of Demand.” *Journal of Economics & Management Strategy* 9 (4): 513–548.
- Newey, W. 1991. “Uniform Convergence in Probability and Stochastic Equicontinuity.” *Econometrica* 59 (4): 1161–1167.
- . 1994. “The Asymptotic Variance of Semiparametric Estimators.” *Econometrica* 62 (6): 1349–1382.
- Petrin, A., and K. Train. 2010. “A Control Function Approach to Endogeneity in Consumer Choice Models.” *Journal of Marketing Research* 47 (1): 3–13.
- Singh, A., K. Hosanagar, and A. Gandhi. 2019. “Machine Learning Instrument Variables for Causal Inference.” Working Paper, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3352957, April.

- Singh, R., and L. Sun. 2021. “Automatic Kappa Weighting for Instrumental Variable Models of Complier Treatment Effects.” Working Paper, <https://arxiv.org/pdf/1909.05244.pdf>.
- Song, Y. 2022. “Bargaining, Merger and Endogenous Network Formation: the case of power plants and coal companies in the US.” Working Paper, , August.
- Train, K., and W. Wilson. 2011. “Coal Demand and Transportation in the Ohio River Basin: Estimation of a Continuous/Discrete Demand System with Numerous Alternatives.” Working Paper, <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.491.4959&rep=rep1&type=pdf>.
- Wong, M., and B. Farooq. 2021. “ResLogit: A residual neural network logit model for data-driven choice modelling.” *Transportation Research Part C: Emerging Technologies* 126 (103050).
- Wooldridge. 2015. “Control Function Methods in Applied Econometrics.” *The Journal of Human Resources* 50 (2): 420–445.
- Wooldridge, J. 2001. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.
- Zhang, Y., R. Li, and C. Tsai. 2010. “Regularization Parameter Selections via Generalized Information Criterion.” *Journal of the American Statistical Association* 105 (489): 312–323.
- Zou, H., and T. Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67 (2): 301–320.
- Zou, H., T. Hastie, and R. Tibshirani. 2007. “On the “degrees of freedom” of the lasso.” *The Annals of Statistics* 35 (5): 2173–2192.

Appendix

Appendix A. A simple example showing the flexible machine learning method

Here we give two simple examples to demonstrate how flexible machine learning methods are in capturing the underlying data patterns. The first example is inherited from Bakhitov (2022). The DGP follows,

$$\xi_i = 2 \text{Uniform}(0, 1) - 1, \quad Z_i = N(0, 1), \quad X_i = 0.5Z_i + \xi_i + N(0, 0.1)$$

$$Y_i = \log(|16X_i - 8| + 1) \text{sign}(X - 0.5) + \xi_i + N(0, 0.1)$$

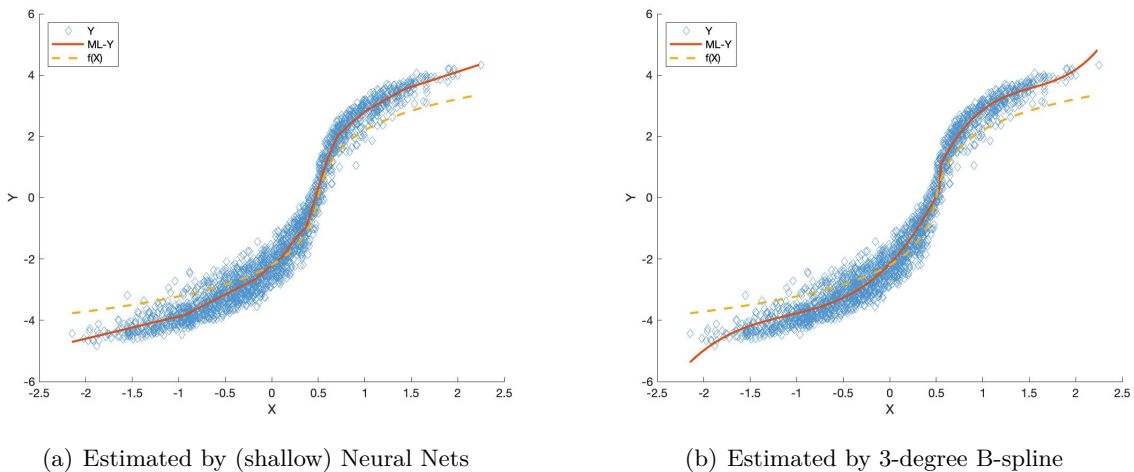


Figure 1: Flexible approximation of function value from simple machine learning methods.

From Figures (1) above, we can tell that the machine learning method would flexibly capture the influence of the unobserved part ξ_i . Machine learning methods may not be able to retrieve the underlying structural pattern, but they can provide a good approximation, as required for our first stage. Another example builds on the propensity score estimation with the DGP,

$$\xi_i = 2 \text{Uniform}(0, 1) - 1, \quad Z_i = N(0, 1), \quad X_i = 0.5Z_i + \xi_i + N(0, 0.1)$$

$$Y_i^* = \ln(|16X_i - 8| + 1) \text{sign}(X - 0.5) + \xi_i + N(0, 1), \quad Y_i = \mathbf{1}\{Y_i^* > 0\}$$

In Figure 2, the red dots represent the probability of $Y_i = 1$, after accounting for the unobserved structural ξ_i . The pattern can be approximated well by some flexible machine learning methods, as Figure 2 suggests.

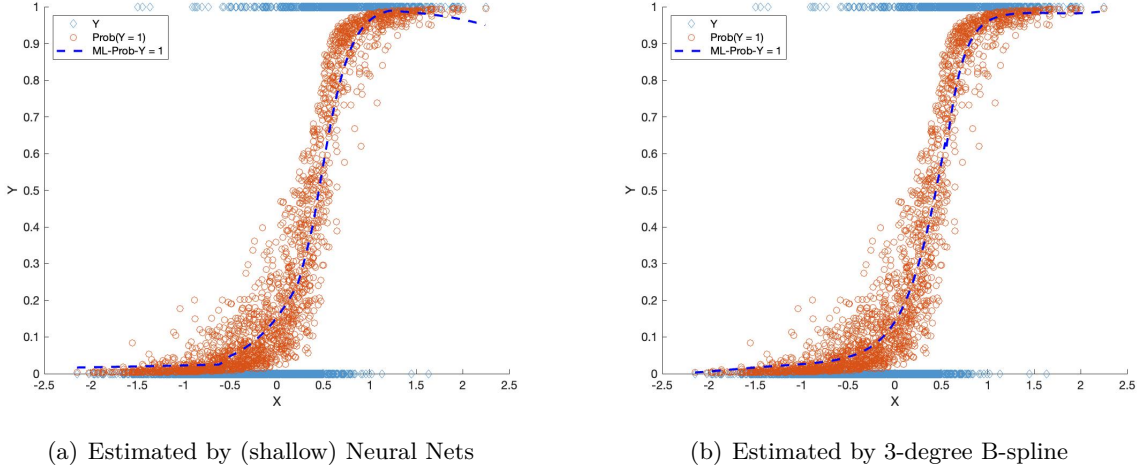


Figure 2: Flexible approximation of choice probability from simple machine learning methods.

Appendix B. Proofs of Results

Proof of Theorem 2

Proof. Here we verify that the assumptions in Chernozhukov et al. (2022a) are satisfied in our setting. Appendix F in Chernozhukov et al. (2022a) offers the conditions sufficient for the consistency of β . Conditions for the consistency and asymptotic normality are overlapping, so here we try to verify them together. First, Our Assumptions 3 are exactly the same as Theorem A3 i), ii), and iii) in Chernozhukov et al. (2022a), Appendix F, while Theorem A3 v) is a sufficient condition to our Assumption 3(iv). Assumptions 4(v) leads to $E[|\phi(W, \gamma_0, \alpha_0, \theta_0)|] < \infty$.

Theorem 1 gives the orthogonal moment condition based on ψ . Next, we want to show that our assumptions are sufficient to guarantee Assumption 1 in Chernozhukov et al. (2022a). For simplicity, we write $\hat{\alpha}_\ell(W_i) = \hat{\alpha}_\ell(W_i, \hat{\pi}_\ell)$, $\hat{\rho}_\ell(W_i) = \hat{\rho}_\ell(W_i, \hat{\pi}_\ell)$, and $\hat{\gamma}_\ell(W_i) = \hat{\gamma}_\ell(W_i, \hat{\pi}_\ell)$ for individual data point. Similar notations apply to α_0 , γ_0 and ρ_0 .

By Assumptions 4(i), and γ enters into the original moment condition m linearly,

$$\int_W \|m(W_i, \hat{\gamma}_\ell, \beta) - m(W_i, \gamma_0, \beta)\|^2 F_0(dW) \leq C \|\hat{\rho}_\ell - \rho_0\|^2 \xrightarrow{P} 0$$

The last inequality follows from the definition of ρ such that $\|\hat{\rho}_\ell - \rho_0\| = \|y_\ell - \hat{\pi}_\ell - y_\ell + \pi_0\| = \|\pi_0 - \hat{\pi}_\ell\|$. Hence, the norm convergence defined in π is the same as in ρ . By Jensen's inequality, $0 \leq (\int_W \|m(W, \hat{\gamma}_\ell, \beta) - m(W, \gamma_0, \beta)\|)^2 \leq \int_W \|m(W, \hat{\gamma}_\ell, \beta) - m(W, \gamma_0, \beta)\|^2 \xrightarrow{P} 0$ – the Assumption iv) in Appendix F, Chernozhukov et al. (2022a) is satisfied.

Then by Assumptions 4(ii), each element inside $\alpha_0(W)$ is bounded.

$$\begin{aligned} \int_W \|\phi(W_i, \hat{\gamma}_\ell, \alpha_0) - \phi(W_i, \gamma_0, \alpha_0)\|^2 F_0(dW) &= \int_W \|\alpha_0(W)\|^2 [\hat{\rho}_\ell(W_i) - \rho_0(W_i)]^2 F_0(dW) \\ &\leq C \int_W [\hat{\rho}_\ell(W_i) - \rho_0(W_i)]^2 F_0(dW) = C \|\hat{\rho}_\ell - \rho_0\|^2 \xrightarrow{P} 0 \end{aligned}$$

Also by Assumption 4(ii), 4(iii), the law of iterated expectation gives,

$$\begin{aligned} \int_W \|\phi(W_i, \gamma_0, \hat{\alpha}_\ell) - \phi(W_i, \gamma_0, \alpha_0)\|^2 F_0(dW) &= \int_W \|\hat{\alpha}_\ell(W_i) - \alpha_0(W_i)\|^2 \rho_0(W_i)^2 F_0(dW) \\ &= \int_W \|\hat{\alpha}_\ell(W_i) - \alpha_0(W_i)\|^2 E[\rho(W_i)^2 | W] F_0(dW) \leq C \int_W \|\hat{\alpha}_\ell(W_i) - \alpha_0(W_i)\|^2 F_0(dW) \xrightarrow{P} 0 \end{aligned}$$

Therefore, the Assumption 1 (i), (ii), and (iii) of Chernozhukov et al. (2022a) are satisfied. As for Assumption 2 in Chernozhukov et al. (2022a) and Assumption vi) in Appendix F of Chernozhukov et al. (2022a) for consistency, the interaction term $\hat{\Delta}_\ell(W) = \phi(W_i, \hat{\gamma}_\ell, \hat{\alpha}_\ell) - \phi(W_i, \gamma_0, \hat{\alpha}_\ell) - \phi(W_i, \hat{\gamma}_\ell, \alpha_0) + \phi(W_i, \gamma_0, \alpha_0) = [\hat{\alpha}_\ell(W_i) - \alpha_0(W_i)][\hat{\rho}_\ell(W_i) - \rho_0(W_i)]$. Note that because we have multiple moment conditions, $\hat{\Delta}_\ell(W_i)$ is a vector. For the j -th moment condition, by Assumptions 4(iii), 4(iv) and Cauchy-Schwarz inequality, there are,

$$\begin{aligned} \int_W \hat{\Delta}_{j,\ell}(W_i) F_0(dW) &= \int_W [\hat{\alpha}_{j,\ell}(W_i) - \alpha_{j,0}(W_i)] [\hat{\rho}_\ell(W_i) - \rho_0(W_i)] F_0(dW) \\ &\leq \|\hat{\alpha}_{j,\ell} - \alpha_{j,0}\| \times \|\hat{\rho}_\ell - \rho_0\| \leq C \|\hat{\pi}_\ell - \pi_0\| \times \|\hat{\pi}_\ell - \pi_0\| = o_p(1/\sqrt{N}) \end{aligned}$$

$$\begin{aligned} \int_W \|\hat{\Delta}_\ell(W_i)\|^2 F_0(dW) &= \int_W \|\hat{\alpha}_\ell(W_i) - \alpha_0(W_i)\|^2 [\hat{\rho}_\ell(W_i) - \rho_0(W_i)]^2 F_0(dW) \\ &\leq C \int_W [\hat{\rho}_\ell(W_i) - \rho_0(W_i)]^2 F_0(dW) = C \|\hat{\rho}_\ell - \rho_0\|^2 \xrightarrow{P} 0 \end{aligned}$$

Moreover, Assumption 3 and 4 in Chernozhukov et al. (2022a) is exactly the same with Assumption 4(iv), 4(v) and 4(vi). We do not explicitly account for Assumption 5 of Chernozhukov et al. (2022a)

since our moment condition $m(W, \gamma, \beta)$ is linear in β . We have $m(W_i, \gamma, \beta)$ is differentiable with respect to $\beta \in \mathcal{B}$, and $\partial m(W, \gamma, \beta)/\partial \beta_j - \partial m(W, \gamma, \beta_0)/\partial \beta_j = 0$ by the linear moment construction, so that ii) and iii) of Assumption 5 in Chernozhukov et al. (2022a) are satisfied.

With this, we have all the assumptions required in Chernozhukov et al. (2022a) for the consistency and \sqrt{N} -asymptotic normality of $\hat{\beta}$, which leads directly to the main conclusion proposed in Theorem 2. \square

Appendix C. Numerical Implementation

C.1 ResLogit Algorithm from Wong and Farooq (2021)

Wong and Farooq (2021) proposed a partial linear multinomial Logit regression with the nonlinear part approximated by an underlying multi-layer neural network. Following the notations in Wong and Farooq (2021), the probability of choosing a good j in market t for a group of homogeneous consumers can be calculated with a linear part, $V_{jt} = x'_{jt}\beta_x + w'_{jt}\beta_w = W'_{jt}\beta$ and a nonlinear part g_{jt} , such that, $Pr_{jt} = \exp(V_{jt} + g_{jt}) / \sum_{\ell \in J} \exp(V_{\ell t} + g_{\ell t})$, where $\mathbf{g}_t = (g_{0t}, g_{1t}, \dots, g_{Jt})'$ is defined by an M -layer neural nets,

$$\mathbf{g}_t = - \sum_{m=1}^M \ln(1 + \exp(\theta^{(m)} h_t^{m-1})) \quad (15)$$

θ is a large matrix with $M \times J$ rows and J columns. The first J rows denote as $\theta^{(1)}$, and other $\theta^{(m)}$ defined following the similar way. For each layer, we have $J \times J$ parameters to estimate. The function h is defined recursively by,

$$\begin{aligned} h_t^{(0)} &= \mathbf{V}_t \\ h_t^m &= h_t^{m-1} - \sum_{m'=1}^m \ln(1 + \exp(\theta^{(m')} h_t^{m'-1})) \end{aligned}$$

For instance, if we want to have $M = 3$, then consequently, equation (15) becomes $\mathbf{g}_t = -[\ln(1 + \exp(\theta^{(1)} h^{(0)})) + \ln(1 + \exp(\theta^{(2)} h^{(1)})) + \ln(1 + \exp(\theta^{(3)} h^{(2)}))]$, with h

$$\begin{aligned} h^{(0)} &= \mathbf{V}_t \\ h^{(1)} &= h^{(0)} - \ln(1 + \exp(\theta^{(1)} h^{(0)})) \\ h^{(2)} &= h^{(1)} - [\ln(1 + \exp(\theta^{(2)} h^{(1)})) + \ln(1 + \exp(\theta^{(1)} h^{(0)}))] \end{aligned}$$

Wong and Farooq (2021) studies this basically for approximating a nested Logit structure when the underlying nesting structure is unknown. Although a large number of parameters, which grows with both the number of layers and the number of goods J inside a market, are needed to construct \mathbf{g}_t , unlike the usual exponentially-expanding sieve-based method, it provides a flexible approximation with only polynomially growth. It even implicitly allows different functions for different goods. In this case, even though the asymptotic properties still require further research, we choose the ResLogit as one of the alternatives because it provides a manageable estimation of the interaction behaviors across different goods inside a market, which is principal in industrial organization.

Below we provide the analytical derivatives for Pr_{jt} regarding both β and θ .

$$\begin{aligned}\frac{\partial Pr_{jt}}{\partial \beta} &= \frac{\exp(V_{jt} + g_{jt}) \left\{ \left[\frac{\partial V_{jt}}{\partial \beta} + \frac{\partial g_{jt}}{\partial \beta} \right] \times \left[\sum_{\ell=1}^J \exp(V_{jt} + g_{jt}) \right] - \sum_{\ell=1}^J \exp(V_{\ell t} + g_{\ell t}) \times \left[\frac{\partial V_{\ell t}}{\partial \beta} + \frac{\partial g_{\ell t}}{\partial \beta} \right] \right\}}{\left(\sum_{\ell=1}^J \exp(V_{jt} + g_{jt}) \right)^2} \\ \frac{\partial g_{jt}}{\partial \beta} &= - \sum_{m=1}^M \frac{1}{1 + \exp(\theta^{(m)} h^{(m-1)})} \times \exp(\theta^{(m)} h^{(m-1)}) \times \theta^{(m)} \times \frac{\partial h^{(m-1)}}{\partial \beta}, \quad \frac{\partial h_{jt}^{(0)}}{\partial \beta} = \frac{\partial V_{jt}}{\partial \beta} = W_{jt} \\ \frac{\partial h_{jt}^{(m)}}{\partial \beta} &= \frac{\partial h_{jt}^{(m-1)}}{\partial \beta} - \left\{ \sum_{m'=0}^{m-1} \frac{1}{1 + \exp(\theta_j^{(m'+1)} h_t^{(m')})} \times \exp(\theta_j^{(m'+1)} h_t^{(m')}) \times \theta_j^{(m'+1)} \times \frac{\partial h_t^{(m')}}{\partial \beta} \right\} \\ \frac{\partial Pr_{jt}}{\partial \theta_j^{(m)}} &= \frac{\exp(V_{jt} + g_{jt}) \left\{ \frac{\partial g_{jt}}{\partial \theta_j^{(m)}} \times \left[\sum_{\ell=1}^J \exp(V_{jt} + g_{jt}) \right] - \sum_{\ell=1}^J \exp(V_{\ell t} + g_{\ell t}) \times \frac{\partial g_{\ell t}}{\partial \theta_j^{(m)}} \right\}}{\left(\sum_{\ell=1}^J \exp(V_{jt} + g_{jt}) \right)^2}\end{aligned}$$

$\partial g_t / \partial \theta_j^{(m')}$ is a matrix with the element in j th row, j' th column equals to

$$\begin{aligned}\frac{\partial g_{jt}}{\partial \theta_{jj'}^{(m)}} &= - \left\{ \frac{\exp(\theta_j^{(m)} h^{(m-1)}) \times h_{j'}^{(m-1)}}{1 + \exp(\theta_j^{(m)} h^{(m-1)})} + \sum_{m+1}^M \frac{\exp(\theta_j^{(m')} h^{(m'-1)}) \times \theta_{jj'}^{m'} \times \partial h_{j'}^{(m'-1)} / \partial \theta_{jj'}^{(m)}}{1 + \exp(\theta_j^{(m')} h^{(m'-1)})} \right\} \\ \frac{\partial h^{(m')}}{\partial \theta_{jj'}^{(m)}} &= \begin{cases} [0, \dots, 0]', & \text{if } m' < m \\ \frac{\exp(\theta_j^{(m)} h^{(m-1)}) \otimes [0, \dots, h_{j'}^{m-1}, \dots, 0]'}{1 + \exp(\theta_j^{(m)} h^{(m-1)})}, & \text{if } m' = m \\ \frac{\partial h^{(m')}}{\partial \theta_{jj'}^{(m)}} = \frac{\partial h^{(m'-1)}}{\partial \theta_{jj'}^{(m)}} - \sum_{\tilde{m}=m}^{m'-1} \left[\frac{\exp(\theta_j^{(\tilde{m}+1)} h^{(\tilde{m})}) \times [\theta_j^{(\tilde{m}+1)} \otimes \partial h^{(\tilde{m})} / \partial \theta_{jj'}^{(m)}]}{1 + \exp(\theta_j^{(\tilde{m}+1)} h^{(\tilde{m})})} \right] - \frac{\partial h^{(m)}}{\partial \theta_{jj'}^{(m)}}. & \text{if } m' > m \end{cases}\end{aligned}$$

C.2 Other Machine Learning methods

We use the B-spline for low-dimensional first-stage approximation. Suppose we have a vector X . We normalize it by $\tilde{X} = (X - \min(X)) / (\max(X) + 1 - \min(X))$. Then the B-spline for X_i is defined

recursively following Dung and Tjahjowidodo (2017), with knots ϱ as $[0, 0, 0, 0.5, 0.5, 0.5, 1, 1, 1]'$. A d -degree B-spline is given by $S(\tilde{X}_i) = \sum_{i=1}^{m-d-1} N_{i,d}(\tilde{X}_i)a_i$, where $N_{i,d}(\tilde{X}_i)$ is the i -th basis function for d -degree B-splines. With knot points $\varrho_0 = \varrho_1 = \dots = \varrho_d \leq \varrho_{d+1} \leq \varrho_{d+2} \leq \dots \leq \varrho_{m-d-1} < \varrho_{m-p+1} = \dots = \varrho_m$, basis function are then defined recursively,

$$N_{i,0}(\tilde{X}_i) = \begin{cases} 1, & \text{if } \varrho_i \leq \tilde{X}_i < \varrho_{i+1} \\ 0, & \text{otherwise} \end{cases}, \quad N_{i,j}(\tilde{X}_i) = \frac{\tilde{X}_i - \varrho_i}{\varrho_{i+j} - \varrho_i} N_{i,j-1}(\tilde{X}_i) + \frac{\varrho_{i+j+1} - \tilde{X}_i}{\varrho_{i+j+1} - \varrho_i} N_{i+1,j-1}(\tilde{X}_i).$$

Overall, the number of basis function for \tilde{X}_i equals to $m - d - 1$. So here, for degree = 2, there would be totally $9 - 2 - 1 = 6$ basis for a single \tilde{X}_i . The entire \tilde{X} comes by taking the tensor product across each element. The total number of parameters would grow exponentially with the dimension of X included.

Appendix D. Other Monte Carlo Simulations

This section contains several other Monte Carlo simulations. For the binary choice model (Case 1), suppose, instead of a strong instrumental variable that enters linearly into the system of determining w_i , we may have a weak instrument $z_{jt,2}$,

$$w_{jt} = 0.1 \times z_{jt,2} \times \xi_{jt} + 0.2\xi_{jt} + 0.04 \times N(0, 1) + 2 + 2 \times \Phi \left[\frac{z_{jt,2}^2}{2} - 1 + \xi_i + \sqrt{0.5 + 0.05z_{jt,2}^4 + \ln((1 + z_{jt,2}^4)/10)} \times 2 \right] \quad (16)$$

Table 11 and Table 12 contain the Monte Carlo results for this weak-instrument setting for sample sizes = 4560 and 28800, respectively. Although the estimation of β_w is contaminated by the near-zero correlation between $z_{jt,2}$ and w_{jt} , leading to overwhelming variance, the performance of the moments in the debiased method is still reasonable. This demonstrates the robustness of our method. The absolute bias of β_w decreases from 0.5399 in the no-endogeneity case to around 0.05 in our machine-learning method. On the other hand, the randomness in estimation increases owing to the high standard error, and the differences among those machine-learning algorithms become indistinguishable.

Another simulation in Table 13 have w_{jt} as a discrete-valued response variable with $w_{jt} = 0.4z_{jt,1} + \mathbf{1}(0.7 + 0.3x_{jt,1} + z_{jt,2} \geq \xi_{jt})$, where $z_{jt,1}$ takes either 0 or 1, and the overall w_{jt} can take four values,

$\{0, 0.4, 1, 1.4\}$. Single discrete endogenous variables usually cannot be handled with the control function. Research like Lin and Wooldridge (2017) uses the control function to deal with the case where they simultaneously have a discrete and a continuous endogenous variable. The system is not invertible for ξ_{jt} because the inequalities lead to a range than a point identification. So there is misspecification for the polynomial expansions based on $(x_{jt,1}, z_{jt,2})$ ³⁴. Table 13 shows the basic simulation results.

Table 13 then shows that not accounting for the possible endogeneity would give very biased estimations. The debiased machine learning resolves the majority of endogeneity, despite this misspecification of binary variable. On the other hand, one can see the necessity of the debiased process. The estimator without the orthogonal moment generates a much higher bias than the case that ignores the endogeneity.

34. Due to this misspecification, there is no guarantee that the first-stage machine learning would be a consistent estimator. One could construct some counterexamples resulting in catastrophic estimations.

	Not Consider Endogeneity	Linear Control Function	Raw Lasso L = 1	Raw Lasso L = 5	Raw Lasso L = 10	Debiased Lasso L = 1	Debiased Lasso L = 5	Debiased Lasso L = 10
β_0	Mean Bias	-0.0304	-0.0581	-0.0712	-0.0640	-0.0159	-0.0230	-0.0201
	Std	0.1115	0.2679	0.2977	0.2830	0.1820	0.1806	0.1807
	Median Bias	-0.0312	-0.0535	-0.0648	-0.0585	-0.0192	-0.0226	-0.0220
	Coverage	0.9450	0.9510	0.9490	0.9490	0.9540	0.9510	0.9520
β_1	Mean Bias	0.0023	-0.1509	-0.1628	-0.1567	-0.0144	-0.0177	-0.0159
	Std	0.1012	0.1053	0.1115	0.1078	0.0922	0.0897	0.0907
	Median Bias	0.0048	-0.1559	-0.1657	-0.1614	-0.0128	-0.0158	-0.0140
	Coverage	0.9640	0.7240	0.7340	0.7340	0.9450	0.9480	0.9440
β_2	Mean Bias	0.0001	0.1099	0.1296	0.1191	0.0090	0.0131	0.0109
	Std	0.0550	0.0737	0.0780	0.0756	0.0618	0.0602	0.0607
	Median Bias	-0.0011	0.1111	0.1310	0.1207	0.0093	0.0133	0.0111
	Coverage	0.9480	0.7030	0.6640	0.6640	0.9660	0.9630	0.9640
β_w	Mean Bias	-0.1925	-0.0628	-0.0350	-0.0516	0.0544	0.0575	0.0530
	Std	1.9560	2.4934	2.7787	2.6331	1.8235	1.7987	1.8003
	Median Bias	-0.1196	-0.1589	-0.1698	-0.1723	0.0633	0.0565	0.0624
	Coverage	0	0.9650	0.9530	0.9530	0.9690	0.9650	0.9680

Table 11: Nonlinear generation of w_i , by B-spline polynomial approximation, with degree of B-spline = 2, $n = 4560$. True $\beta = (0.8, 0.2, -0.2, 0.7)'$.

	Not Consider Endogeneity	Linear Control Function	Raw Lasso L = 1	Raw Lasso L = 5	Raw Lasso L = 10	Debiased Lasso L = 1	Debiased Lasso L = 5	Debiased Lasso L = 10
β_0	Mean Bias	-0.0356	0.0126	0.0084	0.0140	-0.0040	-0.0022	-0.0016
	Std	0.4300	0.0916	0.1035	0.1014	0.0828	0.0892	0.0904
	Median Bias	0.0210	0.0064	0.0042	0.0070	-0.0070	-0.0080	-0.0086
	Coverage	0.9720	0.9520	0.9530	0.9530	0.9590	0.9630	0.9630
β_1	Mean Bias	0.0050	-0.0644	-0.0796	-0.0718	-0.0023	-0.0042	-0.0034
	Std	0.0359	0.0433	0.0471	0.0462	0.0412	0.0418	0.0438
	Median Bias	0.0274	-0.0629	-0.0780	-0.0695	-0.0031	-0.0052	-0.0036
	Coverage	0.9450	0.7030	0.6230	0.6820	0.9490	0.9490	0.9570
β_2	Mean Bias	0.0000	0.0412	0.0510	0.0452	0.0019	0.0026	0.0020
	Std	0.0970	0.0281	0.0310	0.0295	0.0245	0.0252	0.0253
	Median Bias	-0.0092	0.0384	0.0478	0.0417	0.0017	0.0024	0.0021
	Coverage	0.9450	0.7300	0.6770	0.7070	0.9470	0.9470	0.9480
β_w	Mean Bias	0.5114	-0.5004	-0.5095	-0.5363	-0.0036	-0.0432	-0.0488
	Std	0.0278	0.7246	0.7908	0.8181	0.8860	0.9339	0.9867
	Median Bias	0.5111	-0.4742	-0.4625	-0.4704	0.0111	-0.0090	0.0022
	Coverage	0	0.9090	0.9140	0.9240	0.9640	0.9580	0.9610

Table 12: Nonlinear generation of w_i , by B-spline polynomial approximation, with degree of B-spline = 2, $n = 28800$. True $\beta = (0.8, 0.2, -0.2, 0.7)'$.

		Not Endogeneity	Raw Lasso L = 1	Deabised Lasso L = 1
β_0	Mean Bias	0.2599	0.4097	0.0475
	Std	0.4858	0.3641	0.2317
	Median Bias	0.2879	0.3793	0.0408
	Coverage	0.9390	0.8160	0.9370
β_1	Mean Bias	0.0275	-0.1975	-0.0291
	Std	0.1855	0.1312	0.0890
	Median Bias	0.0283	-0.2023	-0.0295
	Coverage	0.9420	0.6590	0.9280
β_2	Mean Bias	0.0027	0.1262	0.0155
	Std	0.1071	0.0691	0.0484
	Median Bias	0.0010	0.1338	0.0164
	Coverage	0.9590	0.5040	0.9340
β_w	Mean Bias	-0.3043	-0.4815	-0.0778
	Std	0.1894	0.2879	0.2192
	Median Bias	-0.3149	-0.4842	-0.0819
	Coverage	0.6300	0.6140	0.9280

Table 13: Misspecification with binary endogenous variable. With true $\beta = (-0.7, 0.4, -0.2, 1.2)'$.