Causal Regressions For Unstructured Data

Amandeep Singh
University of Washington
Seattle, WA
amdeep@uw.edu

Bolong Zheng
University of Washington
Seattle, WA
bolongzh@uw.edu

Abstract

The focus of much recent research in economics and marketing has been (1) to allow for unstructured data in causal studies and (2) to flexibly address the issue of endogeneity with observational data and perform valid causal inference. Directly using machine learning algorithms to predict the outcome variable can help deal with the issue of unstructured data; however, it is well known that such an approach does not perform well in the presence of endogeneity in the explanatory variables. On the other hand, extant methods catered towards addressing endogeneity issues make strong parametric assumptions and hence are incapable of "directly" incorporating high-dimensional unstructured data. In this paper, we propose an estimator, which we term "RieszIV" for carrying out estimation and inference with high-dimensional observational data without resorting to parametric approximations. We demonstrate our estimator exhibits asymptotic consistency and normality under a mild set of conditions. We carry out extensive Monte Carlo simulations with both low-dimensional and high-dimensional unstructured data to demonstrate the finite sample performance of our estimator. Finally, using app downloads and review data for apps on Google Play we demonstrate how our method can be used to conduct inference over counterfactual policies over rich text data. We show how large language models can be used as a viable counterfactual policy generation operator. This represents an important advance in expanding counterfactual inference to complex, real-world settings.

1 Introduction

A lot of recent studies in economics and related fields have seen a growing use of unstructured data in their econometric analysis. Further, there has been a growing need to more flexibly model the relationship between outcome and the explanatory variables more. One way to address both these issues could be to use flexible machine learning frameworks. However, it is well known that standard machine learning algorithms do not perform well in presence of endogeneity in the data.

A common approach to correct for the endogeneity bias is to use instrumental variables (IVs). Nonparametric instrumental variables (NPIV) methods have regained popularity among applied researchers over the last decade as they do not require imposing (possibly) implausible parametric assumptions on the target function. However, existing nonparametric estimation techniques still require the researcher to specify a target function approximation (ideally driven by some ex-ante understanding of the data generating process), e.g. a sieve space, which in turn drives the choice of unconditional moment restrictions (or simply put, the choice of IV basis functions). It has been widely demonstrated, that if the approximation is bad, it leads to misspecification issues, and if the IVs are "weak" , most likely the standard NPIV asymptotic techniques will no longer be valid.

Recent studies have made important progress in developing estimators to correct for endogeneity bias in machine learning models without requiring ex-ante specification of basis functions (Hartford et al.

Accepted to the NeurIPS 2023 Workshop on Causal Representation Learning.

2016; Muandet et al. 2019; R. Singh, Sahani, and Gretton 2019; Dikkala et al. 2020; Bakhitov and A. Singh 2022). Termed "MLIV" estimators, these methods enable valid inference in settings with endogeneity. However, some key challenges remain. First, carrying out statistically valid inference in endogenous settings is still an unexplored area. Existing MLIV estimators do not provide tools for constructing confidence intervals. Second, applying counterfactual analysis to unstructured data like images and text also remains largely unexplored in empirical settings.[1] To this end, we –

- We propose an estimator we term *RieszIV*, to carry out estimation and inference of causal effects in endogenous settings. We also extend the extant theoretical results and demonstrate our estimator is consistent and asymptotically normal.
- We conduct extensive Monte-Carlo analysis to demonstrate the small sample properties of our estimator. We find our estimator has superior performance compared to existing methods and showcases almost optimal coverage.
- Finally, we demonstrate how RieszIV can be applied to rich, unstructured data like app reviews. Using data on app downloads and reviews from Google Play, we conduct counterfactual analysis to evaluate potential policy changes.

2 Preliminary Definitions

In this paper, we broadly will consider the estimation of the following non-parametric regression

$$y_i = f_0(x_i) + \epsilon_i, \tag{1}$$

where unlike traditional machine learning regression problems, $\mathbb{E}[\epsilon \cdot x] \neq 0$. This is known as the standard endogeneity issue and is very rampant with archival and observational data. In such contexts, directly using machine learning algorithms to estimate f leads to a substantial bias (see Bakhitov and A. Singh 2022). However, interestingly, in most empirical contexts where such a scenario occurs, the economic object (say θ_0) of interest is not f directly but is some moment function of f which in most cases can be represented in the following form –

$$\theta_0 = \mathbb{E}[m(w, f_0)], \tag{2}$$

where θ_0 is the object of interest, w denotes all the observable data, f is the function that captures the relationship between the outcome variable and explanatory variables, and m is some known functional.

Remark 1. *(Average Treatment Effects): Consider $W = (Y, T, X)$, where $T \in \{0, 1\}$ refers to the discrete treatment variable, $X \in \mathbb{R}^d$ refers to other exogenous characteristics, and $Y_i \in \mathbb{R}$ refers to the outcome variable. In such settings, one is often interested in measuring the average treatment of giving the treatment i.e., the effect of $T = 1$ vs $T = 0$.*

$$m(w_i, f) = f(\{1, X_i\}) - f(\{0, X_i\})$$
$$\theta_0 = \mathbb{E}[m(w_i, f)] = \mathbb{E}[f(\{1, X_i\}) - f(\{0, X_i\})]$$

Remark 2. *(Average Policy Effects): Consider $W_i = (Y_i, X_i)$, where $X_i \in \mathbb{R}^d$ refers to some potentially endogenous variables, and $Y_i \in \mathbb{R}$ refers to the outcome variable. Now consider a policy evaluation function $\pi : \mathbb{R}^d \to \mathbb{R}^d$. In such settings, one is often interested in measuring the average treatment of applying a known policy operator π.*

$$m(w_i, f) = f(\pi(X_i)) - f(X_i)$$
$$\theta_0 = \mathbb{E}[m(w_i, f)] = \mathbb{E}[f(\pi(X_i)) - f(X_i)]$$

Even though non-parametric regression functions are estimable at a slower rate than parametric regressions, i.e., it might not be possible to construct confidence intervals directly on the estimated $\hat{f}$, however, it is possible to do inference and construct valid confidence intervals on the economic object $\hat{\theta}$. Prior literature (Chernozhukov, W. K. Newey, et al. 2021, Chernozhukov, W. Newey, Quintas-Martınez, et al. 2022) has proposed methods to construct such valid confidence intervals when X is exogenous i.e., $\mathbb{E}[\epsilon \cdot x] = 0$. In the next sections, we extend the existing literature to allow for endogeneity in the explanatory variables.

[1] Extant work has been limited to looking low-dim representations of the high-dim data to conduct counterfactuals

3 Estimation of Average Moments

In this section, we will look at the estimation of average moments as defined in equation 2. As one can notice, average moments are a function of f, so we will first discuss the estimation of f. In recent years, many methods have been proposed (Hartford et al. 2016, R. Singh, Sahani, and Gretton 2019, Muandet et al. 2019, Dikkala et al. 2020, Bakhitov and A. Singh 2022) to estimate f_0. To allow for identification of f_0, we assume there exists a set of variables z (also known as instrumental variables) such that –

$$\mathbb{E}[\epsilon \mid z] = 0, \tag{3}$$

$$\mathbb{E}[y - f(x) \mid z] = 0, \tag{4}$$

Among all the MLIV methods proposed AGMM estimator of Dikkala et al. 2020 has demonstrated superior performance with unstructured data. Thus, we will demonstrate the performance of our estimator with the AGMM. However, the same method can be applied to other MLIV methods as well. Thus, to estimate f, we construct the criterion function as proposed in Dikkala et al. 2020. Specifically, we take as our criterion function the maximum moment deviation over the set of instrument functions, where the set of instrument functions is potentially infinite.

$$f_0 = \arg\inf_{f \in \mathcal{F}} \sup_{h \in \mathcal{H}} \mathbb{E}[(y - f(x))h(z)] =: \Psi(f, h) \tag{5}$$

Dikkala et al. 2020 shows that as long as the set of instrument functions $\mathcal{H}$ contains all functions of the form $h(z) = \mathbb{E}\left[f(x) - f'(x) \mid z\right]$ for $f, f' \in \mathcal{F}$, then such an estimator achieves a projected MSE rate that scales with the critical radius of the function classes $\mathcal{H}, \mathcal{F}$ and their tensor product class (i.e. functions of the form $h(z) \cdot f(x)$, with $h \in \mathcal{H}$ and $f \in \mathcal{F}$).

Theorem 1 (Dikkala et al. 2020)**.** *Consider a set of test functions $\mathcal{H} := \cup_{i=1}^{d} \mathcal{H}^i$, that is decomposable as a union of d symmetric test function spaces $\mathcal{H}^i$ and let $\mathcal{H}^i_U = \left\{h \in \mathcal{H}^i : \|f\|^2_{\mathcal{H}} \leq U\right\}$. Consider the estimator:*

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sup_{h \in \mathcal{H}_U} \Psi_n(f, h) + \lambda \|f\|_{\mathcal{F}}$$

Let $f_0 \in \mathcal{F}_B$ be any fixed (independent of the samples) hypothesis that satisfies the Conditional Moment (4). Let $\delta_{n,\zeta} := 2\max_{i=1}^{d} \mathcal{R}\left(\mathcal{H}^i_U\right) + c_0\sqrt{\frac{\log(c_1 d/\zeta)}{n}}$, for some universal constants c_0, c_1 and $B_{n,\lambda,\zeta} := \|f_0\|_{\mathcal{F}} + \delta_{n,\zeta}/\lambda$. Suppose that:

$$\forall f \in \mathcal{F}_{B_{n,\lambda,\zeta}} : \frac{T\left(f - f_0\right)}{\left\|T\left(f - f_0\right)\right\|_2} \in \operatorname{span}_{\kappa}\left(\mathcal{H}_U\right)$$

Then if $\lambda \geq \delta_{n,\zeta}$, $\hat{f}$ satisfies for some universal constants c_0, c_1, that w.p. $1 - \zeta$:

$$\left\|T\left(f_0 - \hat{f}\right)\right\|_2 \leq \kappa\left(2(B+1)\mathcal{R}\left(\mathcal{F}_1\right) + \delta_{n,\zeta} + \lambda\left(\|f_0\|_{\mathcal{F}} - \|\hat{f}\|_{\mathcal{F}}\right)\right)$$

such that

$$\left\|T\left(\hat{f} - f_0\right)\right\|_2 := \sqrt{\mathbb{E}\left[\left(\mathbb{E}\left[\hat{f}(x) - f_0(x) \mid z\right]\right)^2\right]}$$

The above theorem establishes relatively tight fast rates for the projected root mean square errors on f, however non-parametric regression function f, is still estimatable at a slower rate than parameteric regression rates, and thus we cannot carry out valid inference directly with functionals f. Having said that, one can still carry out valid inference on the economic objects θ_0. In this section, we discuss the procedure to conduct estimation and inference over θ_0.

One intuitive way to estimate θ_0, would be to compute its empirical analog using the estimated $\hat{f}$ in the first stage, i.e.,

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} m(w_i, \hat{f}) \tag{6}$$

However, this estimator might not be $\sqrt{n}-$ consistent if the first order bias does not vanish at $\sqrt{n}$. Irrespective of the method used to estimate f, this is often the case, as flexible estimation of f always requires some form of regularization and/or model selection.

The non-parametric regression function f is generally estimable at a slower rate than its parametric counterpart, which poses challenges when performing valid causal inference with functionals of f. Although the non-parametric regression function is typically estimable only at slower rates than parametric regression, it is often possible to achieve parametric rates for the average moment functional.

However, obtaining these rates is not as straightforward as plugging a non-parametric regression estimate into the moment formula and averaging. Instead, debiasing techniques are required to mitigate the effects of regularization and/or model selection when learning the non-parametric regression.

We will focus on problems where there exists a square-integrable random variable $\alpha_0(Z)$ such that –

$$\mathbb{E}[m(W, f)] = \mathbb{E}[\alpha_0(Z)f(X)]$$
$$\forall f \text{ with } \mathbb{E}[f(X)^2] < \infty$$

Orthogonal moment functions are those where the expected moment functions have zero derivative with respect to the first step. Observe that to construct an $\sqrt{n}-$ consistent estimator of θ_0, one also needs to estimate α_0. Further, recent nn based Riesz estimators do not allow for endogeneity.

4 Riesz Representor: RieszIV

To flexibly estimate Riesz estimator, we propose an adversarial estimation of the Riesz representor we term RieszIV. We consider the estimation of the Riesz representer within some function space $\mathcal{A}$ and propose an adversarial estimator based on regularized variants of the following min-max criterion -

$$\hat{\alpha} = \arg\min_{\alpha\in\mathcal{A}} \max_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}[m(w_i; f) - \alpha(z_i)\cdot f(x_i) - f(x_i)^2] + \lambda||f||^2_{\mathcal{F}} + \mu||\alpha||^2_{\mathcal{A}} \tag{7}$$

We now provide fast convergence rates of our regularized minimax estimator, parameterized by the critical radii of the function classes in a similar fashion to Dikkala et al. 2020 and Chernozhukov, W. Newey, R. Singh, et al. 2020:

$$\mathcal{F}_B := \left\{f \in \mathcal{F} : \|f\|^2_{\mathcal{A}_\perp} \leq B\right\}$$
$$m \circ \mathcal{F}_B := \{m(\cdot; f) : f \in \mathcal{F}_B\}$$

for some appropriately defined constant B and $\mathcal{A}_\perp$ refers to the projected functional class such that $\mathcal{A}_\perp := \{K\alpha : \alpha \in \mathcal{A}\}$. The critical radius of a function class $\mathcal{F}$ with range in $[-1, 1]$ is defined as any solution δ_n to the inequality:

$$\mathcal{R}(\delta; \mathcal{F}) \leq \delta^2 \quad \text{with: } \mathcal{R}(\delta; \mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}:\|f\|_2\leq\delta} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right]$$

with $\epsilon_{1:n}$ are independent Rademacher random variables drawn equiprobably in $\{-1, 1\}$. For

Proposition 1. *Assume that mean-squared continuity holds for some constant $M \geq 1$ and that for some $B \geq 0$, the functions in $\mathcal{F}_B$ and $m \circ \mathcal{F}_B$ have uniformly bounded ranges in $[-1, 1]$. Let:*

$$\delta := \delta_n + \epsilon_n + c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$$

for universal constants c_0, c_1, where δ_n upper bounds the critical radii of $\mathcal{F}_B, m \circ \mathcal{F}_B$ and ϵ_n upper bounds the bias $\min_{\alpha\in\mathcal{A}} \|K(\alpha - \alpha_0)\|_2$. Let $\alpha_* = \arg\min_{\alpha\in\mathcal{A}} \|K(\alpha - \alpha_0)\|_2$. Then the estimator in Equation (7), with $\mu \geq 6\lambda \geq 12\delta^2/B$, satisfies w.p. $1 - \zeta$:

$$\|K(\hat{\alpha} - \alpha_0)\|_2 \leq O\left(M^2\delta + \frac{\mu}{\delta}\|K\alpha_*\|^2_{\mathcal{F}}\right)$$

For $\mu \leq C\delta^2/B$, for some constant C, the latter is: $O\left(\delta \max\left\{M^2, \frac{\|a_*\|^2_{\mathcal{A}}}{B}\right\}\right)$.

This immediately follows from Theorem 1 of Chernozhukov, W. Newey, R. Singh, et al. 2020, such that $K : \mathcal{A} \to \mathcal{F}$ is the linear operator defined as $K\alpha := \mathbb{E}[\alpha(Z) \mid X = \cdot]$

5 Theoretical Properties

In this section, we demonstrate that our estimator is asymptotically consistent and normal. Our results build on the work by Chernozhukov, Escanciano, et al. 2016, Chernozhukov, W. K. Newey, et al. 2021, and Ichimura and W. K. Newey 2017.

Assumption 1. *i)* $\alpha_0(X)$ *and* $E[(Y - f_0(X))^2|X]$ *are bounded; ii)* $E[m(W, f_0)^2] < \infty$

These assumptions are standard regularity conditions used in the automatic machine learning literature.

Assumption 2. *i)* $||\hat{f} - f_0|| \xrightarrow{p} 0$ *and* $||\hat{\alpha} - \alpha_0|| \xrightarrow{p} 0$*; ii)* $\sqrt{n}||\hat{\alpha} - \alpha||||T(\hat{f} - f_0)|| \xrightarrow{p} 0$ *or* $\sqrt{n}||K(\hat{\alpha} - \alpha)||||(\hat{f} - f_0)|| \xrightarrow{p} 0$*; iii)* $\hat{\alpha}(Z)$ *is bounded.*

Intuitively these assumptions mean that (i) the estimator of both f and α should be consistent, which is the case for all MLIV estimators and RieszIV estimator. Further, it requires that the product of projected mean square error of $\hat{\alpha}$ and mean square error of f should vanish at $\sqrt{n}-$ rate.

Assumption 3. $m(w, f)$ *is linear in* f *and there is* $C > 0$ *such that*

$$|E\left[m(w, f) - \theta_0 + \alpha_0(z)(y - f(x))\right]| \leq C \left\|f - f_0\right\|^2$$

Theorem 2. *If Assumptions 1-3 are satisfied then for* $V = E[\{m(w, f_0) - \theta_0 + \alpha_0(Z)(y - f_0(x))\}^2]$*,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, V), \hat{V} \xrightarrow{p} V.$$

This result follows immediately from Theorem 4 of Chernozhukov, W. K. Newey, et al. 2021. See Appendix for details. However, unlike Chernozhukov, W. K. Newey, et al. 2021, we require a fast enough rate for the projected root mean squared error rather than the root mean squared error.

6 Simulation Studies

In this section, we apply our approach to simulated data to demonstrate if our framework can provide valid inference. We will consider two scenarios – (A) Both x and z are low dimensional, and (B) x and z are high dimensional images.

6.1 Low-dimensional Case

Specifically, we consider the following data-generating process -

$$\begin{aligned} y &= f(x_{[:,0]}) + e + \delta, \quad \delta \sim \mathcal{N}(0, 0.1) \\ x &= \gamma z + e, \quad z \sim \mathcal{N}(0, 2I_d), e \sim \mathcal{N}(0, 2) \end{aligned}$$

where we have an equal number of treatments n_x and instruments n_z, such that $d = n_x = n_z > 1$. γ is the strength of instruments.

We select the following functional forms as our ground truth model:

sin: $f(x) = \sin(\frac{x}{2})$ 2dpoly: $f(x) = -1.5x + 0.9x^2$ abs: $f(x) = |x|$ linear: $f(x) = x$

We let the $\theta_0 = \mathbb{E}[m(w, f)]$, where $m(w, f) = f(x + 1) - f(x)$. Thus, we want to estimate the average treatment effect (ATE) of increasing the input x by 1. We follow the estimation outline **??**, and present the aggregate results over 100 iterations in Table 1, compared against classical regression approaches. In the simulations, we observe that RieszIV is very effective on non-linear functional forms such as 2dpoly and abs. On more linear functional forms such as sin and linear, its performance is on par with 2SLS. Also, it consistently outperforms standard polynomial linear regression models without interaction terms. See Appendix B for details about simulation setups, model architectures, and hyperparameters.

Table 1: Simulation results (low-dimensional and high-dimensional)

	True Instruments				Polynomial Regression		2SLS		RieszIV		
f	n	n_t	γ	θ	Bias	MSE	Bias	MSE	Bias	MSE	Coverage
sin	1500	3	0.6	0.243	0.297	0.088	0.011	0.002	0.010	0.003	0.86
2dpoly	1500	3	0.6	-0.600	0.291	0.097	-0.919	0.898	-0.031	0.023	0.93
abs	1500	3	0.6	0.168	0.304	0.093	-0.170	0.032	0.001	0.004	0.88
linear	1500	3	0.6	1.000	0.470	0.222	-0.003	0.001	-0.004	0.003	0.94

n	γ	Bias	MSE	Coverage
1500	0.7	-0.068198	0.256432	0.92
3000	0.7	-0.001368	0.158810	0.87

6.2 High-dimensional Case

We now move on to the case of high-dimensional data to demonstrate the ability of our framework to perform valid inference with unstructured data. For this purpose, we will make use of the MNIST dataset . To generate the unstructured dataset we consider the following data generating process –

$$
\begin{aligned}
y &= f(x) + e \\
x^{(1)} &= z^{(1)} + e + \gamma \\
z^{(1)} &\sim \text{Uniform}\left([-3, 3]\right), \quad e \sim \mathcal{N}(0, 1)
\end{aligned}
$$

We consider the following scenario:
- $\text{MNIST}_{\mathbf{x},\mathbf{z}}$: $x \leftarrow \text{RandomImage}\left(\sigma\left(x^{(1)}\right)\right), z \leftarrow \text{RandomImage}\left(\sigma\left(z^{(1)}\right)\right)$.

Here, we follow some experiment setups in Bennett, Kallus, and Schnabel 2019. and Dikkala et al. 2020. We let the true average policy effect $\theta_0 = \mathbb{E}[m(w, f)]$, where f maps the input to its attribute value (the integer-valued digit in an image in this case) and $m(w, f) = f(\pi(x)) - f(x)$.

In our simulation, the choice of policy maps the input image to a randomly sampled image whose attribute value is increased by 1 (i.e. $\pi(x) = \text{RandomImage}(\sigma(f(x)) + 1)$, where σ is a piece-wise linear function that transforms the input into an integer between 0 to 9 by clipping, RandomImage takes an integer between 0 to 9, and outputs a randomly sampled image whose attribute value is equal to its input). Thus, we want to estimate the average policy effect (APE) of applying π.

We present the simulation results of 3-fold cross-fitting on Table 1. We can see that the estimator achieves a desirable bias and MSE.

See Appendix B for details about simulation setups, model architectures, and hyperparameters.

7 Conclusion

Recent literature has established the inability of standard machine learning methods to perform well in endogenous settings. Fortunately, some methods have been proposed to flexibly estimate target functions under such contexts, however, conducting inference and constructing valid confidence intervals with such methods has remained an unexplored territory. In this paper, we propose a method to conduct causal inference and construct valid confidence intervals in endogenous settings. We believe our method could be widely adopted and be very useful for applied researchers who actively use instrumental variables to identify causal effects. In this paper, we establish the efficacy of our method through extensive Monte Carlo design. We believe extensive future research can look into applying this estimator to existing studies with publicly available datasets and compare results and implications with existing literature.

References

Chernozhukov, Victor, Juan Carlos Escanciano, et al. (2016). "Locally robust semiparametric estimation". In: arXiv preprint arXiv:1608.00033.

Hartford, Jason et al. (2016). "Counterfactual prediction with deep instrumental variables networks". In: arXiv preprint arXiv:1612.09596.

Ichimura, Hidehiko and Whitney K Newey (2017). "The influence function of semiparametric estimators". In: arXiv preprint arXiv:1508.01378.

Bennett, Andrew, Nathan Kallus, and Tobias Schnabel (2019). "Deep generalized method of moments for instrumental variable analysis". In: Advances in Neural Information Processing Systems, pp. 3564–3574.

Muandet, Krikamol et al. (2019). "Dual IV: A Single Stage Instrumental Variable Regression". In: arXiv preprint arXiv:1910.12358.

Sanh, Victor et al. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: ArXiv abs/1910.01108.

Singh, Rahul, Maneesh Sahani, and Arthur Gretton (2019). "Kernel instrumental variable regression". In: Advances in Neural Information Processing Systems, pp. 4595–4607.

Chernozhukov, Victor, Whitney Newey, Rahul Singh, et al. (2020). "Adversarial estimation of riesz representers". In: arXiv preprint arXiv:2101.00009.

Dikkala, Nishanth et al. (2020). "Minimax estimation of conditional moment models". In: Advances in Neural Information Processing Systems 33, pp. 12248–12262.

Chernozhukov, Victor, Whitney K Newey, et al. (2021). "Automatic debiased machine learning via neural nets for generalized linear regression". In: arXiv preprint arXiv:2104.14737.

Bakhitov, Edvard and Amandeep Singh (2022). "Causal gradient boosting: Boosted instrumental variable regression". In: Proceedings of the 23rd ACM Conference on Economics and Computation, pp. 604–605.

Chernozhukov, Victor, Whitney Newey, Vıctor M Quintas-Martınez, et al. (2022). "RieszNet and ForestRiesz: Automatic Debiased Machine Learning with Neural Nets and Random Forests". In: International Conference on Machine Learning. PMLR, pp. 3901–3914.

Appendix A

Remark 3. *(Population limit). Consider the population limit of our criterion where $n \to \infty$ and $\lambda, \mu \to 0$. Then our criterion is:*

$$\max_{f \in \mathcal{F}} \mathbb{E}[m(w; f) - a(z) \cdot f(X)] - \|f\|_2^2$$

By the definition of the Riesz representer we thus have:

$$\begin{aligned}\max_{f \in \mathcal{F}} \mathbb{E}[m(w; f) - \alpha(z) \cdot f(x)] - \|f\|_2^2 &= \max_{f \in \mathcal{F}} \mathbb{E}\left[(\alpha_0(z) - \alpha(z)) \cdot f(x) - f(x)^2\right] \\ &= \frac{1}{4}\mathbb{E}\left[(\alpha_0(z) - \alpha(z))^2\right] =: \frac{1}{4}\|\hat{\alpha} - \alpha_0\|_2^2\end{aligned}$$

Thus our empirical criterion converges to the mean-squared-error criterion in the population limit.

Corollary 3. *We now get a bound on the RMSE of $\hat{\alpha}$, i.e. $\|\hat{\alpha} - \alpha_0\|_2$, utilizing the bound on the projected RMSE. First, we bound the quantity*

$$\tau^*(\delta) = \sup_{\alpha \in \mathcal{A}: \|K(\alpha - \alpha_*)\|_2 \leq \delta} \|\alpha - \alpha_*\|_2$$

This is also known as the measure of ill-posedness of the operator K. Consider the bound on this measure, defined as:

$$\tau := \sup_{\alpha \in \mathcal{A}} \frac{\|\alpha - \alpha_*\|_2}{\|K(\alpha - \alpha_*)\|}.$$

Then observe that Theorem 2 implies that:

$$\|\hat{\alpha} - \alpha_0\|_2 \leq \tau \|T(\hat{\alpha} - \alpha_0)\|_2 \leq O\left(\tau M^2 \delta + \tau \frac{\mu}{\delta} \|a_*\|_{\mathcal{A}}^2\right)$$

which by a triangle inequality also implies that:

$$\|\hat{\alpha} - \alpha_0\|_2 \leq O\left(\tau \delta_n + (\tau + 1) \|\alpha_* - \alpha_0\|_2\right)$$

Choosing $\alpha_* = \arg\min_{\alpha \in \mathcal{A}: \|\alpha\|_{\mathcal{A}} \leq B} \|\alpha_* - \alpha_0\|$, yields the bound:

$$\|\hat{\alpha} - \alpha_0\|_2 \leq O\left(\tau \delta_n + (\tau + 1) \inf_{\alpha \in \mathcal{A}: \|\alpha\|_{\mathcal{A}}} \|\alpha - \alpha_0\|_2\right)$$

Proof. To show the asymptotic normality we will first verify the Assumptions 1-3 of Chernozhukov, Escanciano, et al. 2016, from now on CEINR, with $g(w, f, \theta) = m(w, f) - \theta$ and $\phi(w, f, \alpha, \theta) = \alpha(z) \cdot (Y - f(x))$. Using Assumption 1 and $||\hat{f} - f_0|| \xrightarrow{p} 0$ we have,

$$\begin{aligned}\int ||g(w, \hat{f}, \theta_0) - g(w, f_0, \theta_0)||^2 \mathcal{P}_0(dw) &= \int ||m(w, \hat{f}) - m(w, f_0)||^2 \mathcal{P}_0(dw) \\ &\leq C \int ||\hat{f} - f_0||^2 \mathcal{P}_0(dw) \xrightarrow{p} 0\end{aligned} \tag{8}$$

Also by Assumption 1 i) and ii), and $||\hat{f} - f|| \xrightarrow{p} 0$,

$$\begin{aligned}\int ||\phi(w, \hat{f}, \alpha_0, \theta_0) - \phi(w, f_0, \alpha_0, \theta_0)||^2 \mathcal{P}_0(dw) &= \int ||\alpha_0(z)(f_0(x) - \hat{f}(x))||^2 \mathcal{P}_0(dw) \\ &\leq C \int ||(f_0(x) - \hat{f}(x))||^2 \mathcal{P}_0(dw) \\ &\leq C||\hat{f} - f_0||^2 \xrightarrow{p} 0\end{aligned} \tag{9}$$

Also by Assumption 1 i) and $||\hat{\alpha} - \alpha_0|| \xrightarrow{p} 0$, we have,

$$\begin{aligned}\int ||\phi(w, f_0, \hat{\alpha}, \tilde{\theta}) - \phi(w, f_0, \alpha_0, \theta_0)||^2 \mathcal{P}_0(dw) &= \int ||\left(\hat{\alpha}(z) - \alpha_0(z)\right)\left(y - f_0(x)\right)||^2 \mathcal{P}_0(dw) \\ &\leq C \int ||\hat{\alpha} - \alpha_0||^2 \mathcal{P}_0(dw) \\ &\leq C||\hat{\alpha} - \alpha_0||^2 \xrightarrow{p} 0\end{aligned} \tag{10}$$

This satisfies Assumption 1 parts i), ii), and iii) of CEINR.

Next, consider

$$\begin{aligned}\hat{\Delta}(w) &:= \phi\left(w, \hat{f}, \hat{\alpha}, \tilde{\theta}\right) - \phi\left(w, f_0, \hat{\alpha}, \tilde{\theta}\right) - \phi\left(w, \hat{f}, \alpha_0, \theta_0\right) + \phi\left(w, f_0, \alpha_0, \theta_0\right) \\ &= -\left[\hat{\alpha}(z) - \alpha_0(z)\right]\left[\hat{f}(x) - f(x)\right].\end{aligned}$$

Then by iterated expectations, the Cauchy-Schwartz inequality, and Assumptions 2 i) and ii)

$$\begin{aligned}E\left[\hat{\Delta}(w)\right] &= E\left[E\left[\hat{\Delta}(w)|z\right]\right] \\ &= \int -\left[\hat{\alpha}(z) - \alpha_0(z)\right]\left[T(\hat{f}(x) - f(x))\right]\mathcal{P}_0(dz) \\ &\leq \|\hat{\alpha} - \alpha_0\|\left\|T(\hat{f}(x) - f(x))\right\| = o_p\left(\frac{1}{\sqrt{n}}\right)\end{aligned} \tag{11}$$

Also since $\hat{\alpha}(z)$ and $\alpha(z)$ is bounded, we have

$$\begin{aligned}\int\left\|\hat{\Delta}(w)\right\|^2\mathcal{P}_0(dw) &= \int\left[\hat{\alpha}(z) - \alpha_0(z)\right]^2\left[T(\hat{f}(x) - f(x))\right]^2\mathcal{P}_0(dz) \\ &\leq C\int\left[T(\hat{f}(x) - f(x))\right]^2\mathcal{P}_0(dz) \xrightarrow{p} 0\end{aligned} \tag{12}$$

Thus Equation 11 and Equation 12 verify Assumption 2 i) of CEINR.

Next Assumption 3 of CEINR is satisfied through Assumption 3. Thus Assumptions 1-3 of CEINR are satisfied. Thus asymptotic normality follows by Lemma 15 of CEINR and the Lindberg-Levy central limit theorem.

Finally, we know $\theta \xrightarrow{p} \theta_0$. And thus we have, $\int\left\|g\left(w, \hat{f}, \tilde{\theta}\right) - g\left(w, \hat{f}, \theta_0\right)\right\|^2\mathcal{P}_0(dw) \xrightarrow{p} 0$

To show the second conclusion, we closely follow Chernozhukov, W. K. Newey, et al. 2021. Let $\psi_i = \psi_0\left(W_i\right)$. Then for $i \in I_\ell$,

$$\left(\hat{\psi}_{i\ell} - \psi_i\right)^2 \leq C\left(\sum_{j=1}^{3} R_{ij} + R\right), R_{i1} = \left[m\left(W_i, \hat{f}_\ell\right) - m\left(W_i, f_0\right)\right]^2, R_{i2} = \hat{\alpha}_\ell\left(Z_i\right)^2\left\{\hat{f}_\ell\left(X_i\right) - \bar{f}\left(X_i\right)\right\}^2$$

$$R_{i3} = \left\{\hat{\alpha}_\ell\left(Z_i\right) - \bar{\alpha}\left(Z_i\right)\right\}^2\left\{Y_i - \bar{f}\left(X_i\right)\right\}^2, R = (\hat{\theta} - \bar{\theta})^2$$

We just showed $R \xrightarrow{p} 0$. Let $\mathcal{W}_{-\ell}$ denote the observations not in I_ℓ. By Assumption 10 ,

$$\mathrm{E}\left[R_{i1} \mid \mathcal{W}_{-\ell}\right] = \int\left[m\left(w, \hat{\gamma}_\ell\right) - m(w, \bar{\gamma})\right]^2 F_W(dw) = o_p(1)$$

By Assumption 4 and Lemma A9, uniformly in x

$$|\hat{\alpha}_\ell(x)| \leq \sum_{j=1}^{p}|b_j(x)|\,|\hat{\rho}_{\ell j}| \leq C\left\|\hat{\rho}_\ell\right\|_1 = O_p(1)$$

Then by Assumption 11 ,

$$\mathrm{E}\left[R_{i2} \mid \mathcal{W}_{-\ell}\right] \leq C\left\|\hat{\rho}_\ell\right\|_1^2\int\left\{\hat{\gamma}_\ell(x) - \bar{\gamma}(x)\right\}^2 F_W(dw) = C\left\|\hat{\rho}_\ell\right\|_1^2\left\|\hat{\gamma}_\ell - \bar{\gamma}\right\|^2 \leq O_p(1)o_p(1) = o_p(1)$$

Also by Assumption 9 and iterated expectations

$$\begin{aligned}\mathrm{E}\left[R_{i3} \mid \mathcal{W}_{-\ell}\right] &\leq \int\left\{\hat{\alpha}_\ell(x) - \bar{\alpha}(x)\right\}^2\mathrm{E}\left[(Y - \bar{\gamma}(x))^2 \mid X = x\right]F_X(dx) \\ &\leq C\int\left\{\hat{\alpha}_\ell(x) - \bar{\alpha}(x)\right\}^2 F_X(dx) = C\left\|\hat{\alpha}_\ell - \bar{\alpha}\right\|^2 = o_p(1).\end{aligned}$$

By the triangle inequality,

$$\mathrm{E}\left[\frac{1}{n}\sum_{i\in I_\ell}\sum_{j=1}^{3} R_{ij} \mid \mathcal{W}_{-\ell}\right] \leq \mathrm{E}\left[R_{i1} \mid \mathcal{W}_{-\ell}\right] + \mathrm{E}\left[R_{i3} \mid \mathcal{W}_{-\ell}\right] + \mathrm{E}\left[R_{i3} \mid \mathcal{W}_{-\ell}\right] = o_p(1)$$

It then follows by the conditional Markov inequality that $\sum_{i\in I_\ell}\sum_{j=1}^{3} R_{ij}/n = o_p(1)$. The triangle inequality and adding up over ℓ then gives $\left(\hat{\psi}_{i\ell} - \psi_i\right)^2$

$$\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\left(\hat{\psi}_{i\ell} - \psi_i\right)^2 = o_p(1).$$

Note also that by Assumptions 9 and 10,

$$\mathrm{E}\left[\psi_i^2\right] \leq C\left(1 + \mathrm{E}\left[m(W,\bar{f})^2\right] + \mathrm{E}\left[\bar{\alpha}(Z)^2\{Y - \bar{f}(X)\}^2\right]\right) < \infty.$$

Then

$\hat{V} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\psi}_{i\ell}^2 = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\left(\hat{\psi}_{i\ell} - \psi_i + \psi_i\right)^2 = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\left(\hat{\psi}_{i\ell} - \psi_i\right)^2 + 2\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\left(\hat{\psi}_{i\ell} - \psi_i\right)\psi_i + \frac{1}{n}\sum_{i=1}^{n}\psi_i^2;$

Furthermore by the Cauchy-Schwartz and Markov inequalities we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_i - \psi_i\right)\psi_i\right| \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\psi}_i - \psi_i\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\psi_i^2} \xrightarrow{p} 0.$$

Then $\hat{V} \xrightarrow{p} V$ follows by the triangle inequality and the law of large numbers. □

Appendix B

A Estimation Outline

In this section, we outline the estimation procedure step by step. In essence, to estimate θ_0, we will use cross-fitting to calculate the orthogonal moment function

- Consider the dataset $\{y_i, x_i, z_i\}_{i=1}^n$ is independently and identically distributed. Now we randomly split the data into L folds such that the data $D_l := \{y_i, x_i, z_i\}_{i\in I_l}$, where I_l denotes the l^{th} partition.
- In the second stage for each fold I_l, we estimate both the AGMM estimator and the RieszIV estimator on the left out data $D_{l^c} := \{y_i, x_i, z_i\}_{i\notin I_l}$

$$\hat{f}_l = \arg\min_{f\in\mathcal{F}}\max_{h\in\mathcal{H}}\frac{1}{|D_{l^c}|}\sum_{i\in D_{l^c}}[(y_i - f(x_i))h(z_i)] \tag{13}$$

$$\hat{\alpha}_l = \arg\min_{\alpha\in\mathcal{A}}\max_{f\in\mathcal{F}}\frac{1}{|D_{l^c}|}\sum_{i\in D_{l^c}}[m(w_i;f) - \alpha(z_i)\cdot f(x_i) - f(x_i)^2] \tag{14}$$

- Now for every I_l, we will estimate $\hat{\theta}_l$ by using the estimators estimated on D_l^c. And finally, to estimate θ, we average it out across all folds. Thus the estimator for θ_0 and its variance can be given as follows –

Figure 1: Examples of policy functions

In the first example, the policy function takes in an image of a clothing model and outputs the same image but with the facial expression transformed to a smile. In the second example, the policy function applies a word limit of 30 to a description of a second-hand laptop. The examples are from Amazon and Mercari. In e-commerce settings, such features like product thumbnails and descriptions are associated with econometric variables that are of interest to platforms and sellers. We are interested in how policies of this kind would affect y: the demand of products

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in D_\ell}\left\{m\left(w_i, \hat{f}_\ell\right) + \hat{\alpha}_\ell\left(z_i\right)\left(y_i - \hat{f}_\ell(x_i)\right)\right\}$$

$$\hat{V} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in D_\ell}\hat{\psi}_{i\ell}^2, \hat{\psi}_{i\ell} = m\left(w_i, \hat{f}_\ell\right) - \hat{\theta} + \hat{\alpha}_\ell\left(z_i\right)\left(y_i - \hat{f}_\ell(x_i)\right)$$

A.1 Low-dimensional Case

Parameter space: For low-dimesional simulations, we estimate Average Treatment Effect on different functional forms using the Cartesian products of the following parameter choices as setups.

$n \in \{1000, 1500\}$

$d = n_z = n_x \in \{2, 3\}$

$\gamma \in \{0.3, 0.6\}$

$l \in \{2\}$

where n is the sample size, n_x is the number of treatments, n_z is the number of instruments, γ is the strength of instruments, and l is the number of folds.

Ground truth: For the true average treatment effect θ, we perform Monte-Carlo simulations for different functional forms with varying instrument strength, and pre-compute an estimated true average treatment effect. This estimated θ is used in the calculations of bias and MSE.

Alternatively, exact values can be computed.

Baseline: For the baselines, we test our estimator against third degree polynomial regression (without interactions), and 2sls.

Coverage: For the coverage, it is obtained by calculating the proportion of times the true average treatment effect falling within the 95% confidence interval of our estimator.

Architecture: In general, for our proposed RieszIV estimator, we use shallow feed-foward neural networks with dropouts and LeakyReLu non-linearity as learners and adversaries.

Specifically:

AGMM learner: it is a sequential model that starts with a dropout layer with probability 0.1, followed by a linear layer with 100 neurons, a LeakyReLu activation with 0.01 negative slope, another dropout layer with probability 0.1, another linear layer with 100 neurons, a ReLu activation, a dropout layer with probability 0.1, and finally a linear layer whose output dimension is 1.

The architectures for AGMM adversary, Reisz learner, and Reisz adversary are identical.

AGMM adversary: it is a sequential model that starts with a dropout layer with probability 0.1, followed by a linear layer with 100 neurons, a LeakyReLu activation with 0.01 negative slope, another dropout layer with probability 0.1, and finally a linear layer whose output dimension is 1.

Optimization: The optimization is performed by optimistic Adam with weight decay.

Hyperparameters: A batch size of 100 is used and run for 1000 epochs. The learning rates, and regularization strengths for both learners and adversaries are tuned.

A.2 High-dimensional Case

Parameter space: For high-dimesional simulations, we estimate Average Treatment Effect on different functional forms using the Cartesian products of the following parameter choices as setups.

$n \in \{1000, 1500, 2000\}$

$d = n_z = n_x \in \{1\}$

$\gamma \in \{0.3, 0.5, 0.7\}$

$l \in \{3\}$

Problem formulation: Due to the formulation of the data generating process, we pose the modeling process as a regression problem instead of a classification problem. Hence, MSE loss is used for gradient update rather than cross-entropy loss, which is typically used for the dataset.

Moment function: For the policy function, we sample with replacement random images whose attribute value is incremented by one from the entire MNIST dataset. For the digit 9, since there is no target within the distribution such that the policy effect is 1, we mask out 9 in the generated dataset in the training of RieszIV networks and calculation of the estimates.

Due to the masking, the number of samples that are used for calculating the estimators can be inconsistent across folds. We note this and try to correct the standard error by using the reduced sample size after masking.

Ground truth: For the true average policy effect, since we know under the current setting that the true effect is consistently 1, we use 1 in the calculations of the performance metrics.

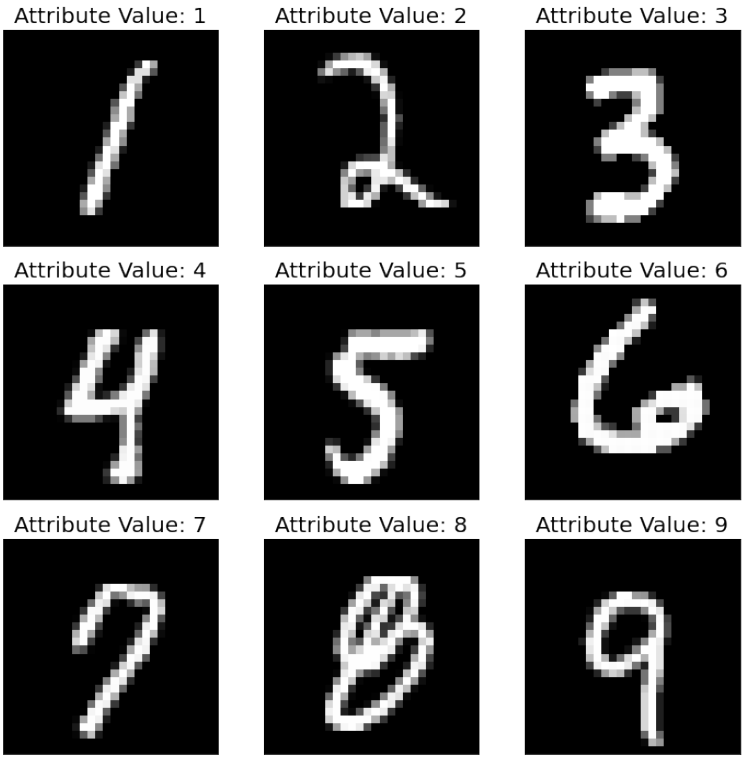
Coverage: For the coverage, it is obtained by calculating the proportion of times the true policy effect falling within the 95% confidence interval of our estimator.

Architecture: We adopt pretrained ResNet-18 (from PyTorch) as the backbone and change the output shape of the last fully connected layer to 1 (so that it generates a one dimensional regression output) for the adversaries and learners.

The architecture for the learners and adversaries is identical using this ResNet-18 based structure.

Pre-processing: The preprocessing pipeline transforms the gray-scale MNIST images to 3-channel images that are compatible with ResNet, and then normalizes the input across channels.

Figure 2: Attribute Values of Various Products



Note. The figure reports various images from the MNIST dataset that are treated as different products in our simulation. We assume that contribution of each image to the utility is equal to the handwritten digit.

Optimization: The optimization is performed by Adam with weight decay.

Hyperparameters: Batch size of 50 is used and run for 1000 epochs. The learning rates, and regularization strengths for both learners and adversaries are tuned.

A.3 Causal Natural Language Processing

We now apply our estimator to a setting with rich text data. We use app download data from Google Play. We use panel data on daily app downloads in the U.S from January 2013 to December 2014. The app data come from App Annie, a firm that collects data from both Google Play and the Apple app store. We observe the list of top 500 apps downloaded (across all genres) on the Google Play U.S. store each day from 2013-2014, the same two-year period for which we observe handset sales. This process results in $6,950$ unique Google Play apps. For each of these apps, we observe the daily price, displayed rating, daily review activity, and versioning activity.

In this experiment, we will be focusing on paid apps, which we define as apps that had at least one day of non-zero price during the two year span.

To estimate app demand we use the standard logit model of demand –

$$\ln(s_{jt}/s_{0t}) = f(x_{jt}) + \xi_{jt}, \tag{15}$$

The explanatory variables x_{jt} include the app's current version, current version's age, displayed rating, and most recent 5 displayed text reviews (see **??**). App developers continuously update their apps. The apps' current version number captures this development and is a good measure of app maturity. To capture the decay in the quality we use version age. Finally, to understand how text reviews affect consumer demand, we also directly include raw review text.

	Google-Play	
	Mean	Std
Price	3.28	4.48
Rate	19.91	36.80
Average Customer Review Rating (Quality)	4.05	1.07
Average Version Age (months)	4.19	5.27
# of Versions	17	33
# of words (Review Text)	93	56

Note: The summary statistics are computed across a sample of 30 paid Google Play apps over the 2-year period. This subset of data has 11618 rows. Rate refers to the exchange rate of the currency in the country of app firm headquarters. # of Versions refer to the number of versions released for each app as observed in December 2014. Average Customer Review Rating refers to the average valence (out of 5) of customer reviews received by the app. # of words (Review Text) refers to the total word count of the last 5 reviews received by apps on each date.

Table 2: App Characteristics.

For the instruments z_{jt}, we include the exchange rate η_{jt} of the currency of the countries apps' firm headquarters are located in, and features in x_{jt} except price. That is, $z_{jt} = \{\eta_{jt}, \{i : i \in x_{jt}, i \neq p_{jt}\}\}$.

We first present some results of 2SLS on the sampled dataset. Here, we replace the raw reviews by the total word counts of reviews. We also artificially generate a machine learning instrumental variable(MLIV) by applying XGBoost regression of p_{jt} on η_{jt}. See results on Figure 2.

A.4 Counterfactual Analysis

Effect of Review Length:

We use the GPT large language model from OpenAI to aid the measurement of the effect of review length.

In particular, we want to measure the policy effect of summarizing a review over 30 words to within 30 words. Reviews under 30 words are invariant to this policy.

We prompt the LLM ("gpt-3.5-turbo" engine) with the following instruction and append a newline along with raw reviews:

> "Step into the shoes of a typical app reviewer.Your original review states: *original_review*. Now, follow a 30-word limit policy as you revise your review. Share only the revised version."

Below, we present some examples of the reviews and their shortened counterparts in Table 3.

We observe that LLM concisely summarizes the reviews, and fixes grammar and punctuations as side-effects, making the comments more readable.

We run the AGMM model and RieszIV model on the subset of data containing information about 30 randomly selected paid apps across 5 different markets (based on currency). The networks and moment functions take in the raw reviews and reviews after treatment as language embeddings. We also experiment on a price increase policy in which the price feature $p_{jt} \in x_{jt}$ is incremented by 10% upon treatment for comparison.

The AGMM model outputs an estimate average policy effect of 0.0245 for the word limit policy, and -0.0019 for the price increase policy.

The RieszIV model ouputs an estimate of 0.0097 for the word limit policy, and -0.0046 for the price increase policy. The 95% confidence intervals are [-0.0259, 0.0452], and [-0.0079, -0.0013] respectively.

We can see that the word limit scheme has a positive effect on the demand, and the price increase policy has a negative effect on the demand.

In an attempt to explain this effect, we apply an off-the-shelf sentiment analysis model to examine the score difference before and after summarization. We observe that the average sentiment score of the reviews improved. To test a change in mean, We conducted a paired t-test to compare the sentiment score before and after the policy implementation. The analysis revealed a highly significant difference. Specifically, we observed a t-statistic of -14.654 with 58089 degrees of freedom (df), and a significant p-value where $p < 2.2 \times 10^{-16}$.

The mean difference is estimated to be -0.01, accompanied by a 95% confidence interval ranging from -0.0115 to -0.00877.

This result suggests a significant shift in average sentiment. Also, the increase in average sentiment score aligns with the observation on the counterfactual policy effect, in that higher sentiment score could indicate improved impressions of the apps, resulting in higher demand.

Review Before	**Review After**
I dont even play any more I want money its the eorst game I EVER PLAYED EVER RRRRRRRRRRRRRR THIS IS A SCAM THEY TRICK US AND I WANT A GOOD SIMS 3 LIKE ONE THE COMPUTER	App is terrible and a scam. It's not worth playing; I just want a good Sims 3-like game on my computer.
As others have said, RSS is the best way to get your news. I use gReader's off-line function to take my feeds with me in the morning, and I use the built in mobilizer to eliminate the web-bloat. Love the app. And I am already concerned about what is going to happen to gReader Pro when Google kills Reader in July.	gReader is the top RSS app. I rely on its offline feature and mobilizer to streamline my news. Worried about its future after Google Reader's discontinuation.
Loved it when it was on the original xbox you people are geniuses and heroes to me for making this mobile bards tale its by far one of my favorite games ever keep up the great work hope to see more games come out	Loved the mobile version of The Bard's Tale! You guys are geniuses for making it. One of my all-time favorite games. Keep up the great work and I hope to see more games soon!
Used to sleep with the fan on, now I don't have to. Multiple sounds and more available for download, and they all sound natural and smooth, without skips or obvious repetition. Was pleasantly surprised that it has an outstanding alarm clock feature that allows the screen to be dimmed enough so it isn't intrusive. Wonderful app and worth every penny.	This app is fantastic! It offers various natural sounds for sleeping, has a great alarm clock with a dimming screen, and is worth the price.
Love this app. Added all my music n videos except for my purchased movies and albums which sucks. Dont know if maybe I'm doing something wrong. But definitely the first itune sync app that does what it says it will do. Thank you. Switching from apple to android should not be so difficult. Hopefully other apps can learn from this one.	Great app, except for not being able to add purchased movies and albums. Unsure if it's user error. Finally found an iTunes sync app that actually works. Thanks for making switching from Apple to Android easier. Hope other apps take notes.

Table 3: Text Revisions: word limit policy

	Logit 1	Logit 2	Logit + IV
log(1 + no_words)	0.29^{***}		1.28^{***}
	(0.00)		(0.05)
R_t	0.19^{***}	0.27^{***}	0.91^{***}
	(0.00)	(0.06)	(0.06)
prices	0.01^{***}	-0.36^{***}	-0.30^{***}
	(0.00)	(0.01)	(0.01)
version	0.32^{***}	0.45^{***}	0.32^{***}
	(0.00)	(0.01)	(0.01)
log(1 + version_age)	0.09^{***}	0.09^{***}	0.19^{***}
	(0.00)	(0.01)	(0.01)
R^2	0.13	0.00	0.01
Adj. R^2	0.13	-0.00	0.01
Num. obs.	159405	159405	159405

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4: Statistical models

A.5 Counterfactual Analysis: A Few Details

The review length is defined by the length of the output after applying Python's strip method and then the split method on target strings.

For language features, we query OpenAI's text embedding API ("text-embedding-ada-002") to obtain language embeddings for the last 5 reviews for each row. Then we concatenate these embeddings as the textual feature. If a record receives less than 5 reviews, we pad the features with the embedding of empty reviews (indicated by " " (a single space)). In the presence of missing values (denoted by nan in the data), we impute them as " " also.

For the sentiment analysis, we use a pretrained DistilBERT model described in Sanh et al. 2019, which outputs the probability of a text being positive.

The model architectures follow the same setup as the low-dimensional case.