Automatic Debiased Machine Learning Via Reisz Regressions

Whitney K. Newey

European Meeting of the Econometric Society

August 2024

Papers:

Chernozhukov, Newey, Quintas-Martinez, Syrgkanis (2021). "Automatic Debiased Machine Learning via Reisz Regressions," arxiv update, 2024.

Monte Carlo from Chernozhukov, Newey, Quintas-Martinez, Syrgkanis (2022): "RieszNet and ForestRiesz: Automatic Debiased Machine Learning with Neural Nets and Random Forests," *ICML Proceedings*, 2022.

INTRODUCTION

Many interesting parameters depend on regressions.

–Average treatment effect (ATE).

–Dynamic economic models depend on conditional choice probabilities.

–Regressor effects in generalized linear regression (GLR).

The regression may be high dimensional.

This talk is about estimating economic and causal parameters that depend on high dimensional regressions.

Machine learners of regressions provide good predictions in a variety of high dimensional settings.

Methods include neural nets, random forests, Lasso.

Machine learning is good for prediction but biased by regularization and/or model selection.

If "plug-in" machine learner into formula for parameter of interest the bias "passes through" and gives incorrect, very poorly centered, confidence intervals.

Reduce bias by using Neyman orthogonal moment functions and cross-fitting.

See Belloni, et. al. (2012, 2014), Chernozhukov et. al. (2018, "Debiased/Double Machine Learning for Treatment and Structural Parameters," Econometrics Journal) and (2016, 2022, "Locally Robust Semiparametric Estimation," Econometrica).

Neyman orthogonal means zero first order effect of regression on expected moment function.

Cross-fitting is a form of sample splitting where moment function is averaged over different observations than used to estimate regressions; see Bickel (1982), Schick (1986).

Neyman orthogonal moment function depends on unknown debiasing function α_0.

A primary innovation of this paper is a least squares objective function minimized by α_0 that depends only on parameter of interest and so is "automatic."

We refer to this as "Riesz regression" because it estimates the Riesz representer from analysis.

Neural nets, random forests, and other machine learners can be used for Riesz regression.

Another innovation is estimators depending on generalized linear regressions GLR, minimizers of an expected loss.

Examples include conditional means, quantile regressions, and quasi-likelihood maximizers.

Here give weighted Riesz regression to estimate α_0.

Other contributions include finite sample mean square error bounds for Riesz regressions and convergence rates for neural net Riesz regressions.

Give application to effect of race on banks' mortgage denial decisions.

PARAMETERS OF INTEREST

Let W denote a data observation that includes an outcome variable Y regressors X.

Start with θ_0 that depends on conditional mean $\gamma_0(X) = E[Y|X]$.

Example 1: Average Treatment Effect (ATE). $X = (D, Z)$ and $\gamma_0(x) = \gamma_0(d, z)$, where $D \in \{0, 1\}$ is treatment indicator and Z are covariates.

$$\theta_0 = E[\gamma_0(1, Z) - \gamma_0(0, Z)].$$

θ_0 is ATE if potential outcomes mean independent of treatment D conditional on covariates and propensity score $\pi_0(Z) = \Pr(D = 1|Z)$ is never equal to 1 or $0a$ (Rosenbaum and Rubin, 1983).

ATE and other parameters of interest have the form

$$\theta_0 = E[m(W, \gamma_0)],$$

where $m(w, \gamma)$ is a scalar function of a possible data observation w and a possible regression function γ and $m(w, \gamma)$ is linear in γ.

Focus on $m(w, \gamma)$ such that there exists $\alpha_0(X)$ with $E[\alpha_0(X)^2] < \infty$ and

$$E[m(W, \gamma)] = E[\alpha_0(X)\gamma(X)] \text{ for all } \gamma \text{ with } E[\gamma(X)^2] < \infty.$$

This is "Riesz representation" from analysis; includes all root-n consistently estimable $\theta_0 = E[m(W, \gamma_0)]$.

Example 1: Average Treatment Effect (ATE). Here

$$\begin{aligned} m(W, \gamma) &= \gamma(1, Z) - \gamma(0, Z), \\ \alpha_0(X) &= \frac{D}{\pi_0(Z)} - \frac{1 - D}{1 - \pi_0(Z)}; \end{aligned}$$

here $\alpha_0(X)$ is well known.

NEYMAN ORTHOGONALITY

Neyman orthogonal moment function is

$$\psi(w, \gamma, \alpha, \theta) = m(w, \gamma) - \theta + \alpha(x)[y - \gamma(x)].$$

Chernozhukov et al. (2022, "Locally Robust Semiparametric Estimation").

Second term debiases because variation in γ in $m(W, \gamma)$ away from γ_0 is cancelled out by variation of opposite sign in the bias correction term $\alpha_0(X)[Y - \gamma(X)]$.

ESTIMATION

We use estimators of the regression and the debiasing function in the Neyman orthogonal moment function, along with cross-fitting.

Let I_ℓ, $(\ell = 1, ..., L)$, be partition of the observation index set $\{1, ..., n\}$ into L distinct subsets of about equal size; In practice $L = 5$ (5-fold) or $L = 10$ (10-fold) cross-fitting is often used.

Let $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ be estimators constructed from the observations that are not in I_ℓ.

The $\hat{\theta}$ and associated asymptotic variance estimator $\hat{V}$ are

$$\hat{\theta} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \{m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)]\},$$

$$\hat{V} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_{i\ell}^2, \ \hat{\psi}_{i\ell} = m(W_i, \hat{\gamma}_\ell) - \hat{\theta} + \hat{\alpha}_\ell(X_i)[Y_i - \hat{\gamma}_\ell(X_i)].$$

RIESZ REGRESSION

The function α_0 is identified as minimizer of population objective function:

$$\begin{aligned}
\alpha_0 &= \arg\min_{\alpha} E[\{\alpha_0(X) - \alpha(X)\}^2] \\
&= \arg\min_{\alpha} E[\alpha_0(X)^2 - 2\alpha_0(X)\alpha(X) + \alpha(X)^2] \\
&= \arg\min_{\alpha}\{-2E[\alpha_0(X)\alpha(X)] + E[\alpha(X)^2]\} \\
&= \arg\min_{\alpha}\{-2E[m(W,\alpha)] + E[\alpha(X)^2]\} \\
&= \arg\min_{\alpha}\{E[-2m(W,\alpha) + \alpha(X)^2]\}.
\end{aligned}$$

The key is the 4th equality which follows by the Riesz representation

$$E[\alpha_0(X)\alpha(X)] = E[m(W,\alpha)], \text{ for all } \alpha.$$

Automatic in that α_0 minimizes an objective function that depends only on $m(W,\alpha)$ and $\alpha(X)$.

Riesz regression is a primary innovation of this paper.

To estimate α_0 replace the expectation by a sample average and minimize over class $\mathcal{A}_n$ of approximating functions,

$$\hat{\alpha} = \arg\min_{\alpha \in \mathcal{A}} \{\frac{1}{n} \sum_{i=1}^{n} [-2m(W_i, \alpha) + \alpha(X_i)^2] + \hat{P}_\lambda(\alpha)\},$$

where $\mathcal{A}$ is some set of functions of x, $W_1, ..., W_n$ are observations on W, and $\hat{P}_\lambda(\alpha)$ is some possible penalty.

"Automatic" in only using $m(W, \alpha)$ and nothing about form of α_0.

Example 1: ATE, where $m(w, \alpha) = \alpha(1, z) - \alpha(0, z)$,

$$\hat{\alpha} = \arg\min_{\alpha \in \mathcal{A}_n} \{\frac{1}{n} \sum_{i=1}^{n} [-2(\alpha(1, Z_i) - \alpha(0, Z_i)) + \alpha(X_i)^2] + \hat{P}_\lambda(\alpha)\}.$$

The class $\mathcal{A}_n$ could be neural nets, making $\hat{\alpha}$ a neural net estimator of α_0.

THEORY

Assumption 1: *For some* $M > 0$ *it is the case that* $E[m(W,\alpha)^2] \leq ME[\alpha(X)^2]$.

Define

$$\begin{aligned} \text{star}(\mathcal{A} - \alpha_0) &= \{z \to \xi\,(\alpha(x) - \alpha_0(x)) : \alpha \in \mathcal{A},\ \xi \in [0,1]\} \\ \text{star}(m \circ \mathcal{A} - m \circ \alpha_0) &= \{z \to \xi\,(m(w;\alpha) - m(w;\alpha_0)) : \alpha \in \mathcal{A},\ \xi \in [0,1]\} \end{aligned}$$

Assumption 2: $\|f\|_\infty \leq 1$ *for all* $f \in \text{star}(\mathcal{A} - \alpha_0)$ and $f \in \text{star}(m \circ \mathcal{A} - m \circ \alpha_0)$.

Let

$$\alpha^* = \arg\min_{\alpha \in \mathcal{A}} \{E[-2m(W,\alpha) + \alpha(X)^2]\}$$

be the best approximation of α_0 by an element of an approximating set $\mathcal{A}$. Define the critical radius of a set of functions in the usual way and $\|a\| = \sqrt{E[a(X)^2]}$.

Theorem 1: *Let* δ_n *be an upper bound on the critical radius of* $\text{star}(\mathcal{A} - \alpha_0)$ *and* $\text{star}(m \circ \mathcal{A} - m \circ \alpha_0)$. *If Assumptions 1 and 2 are satisfied then for some universal constant* C *it follows that with probability* $1 - \zeta$

$$\|\hat{\alpha} - \alpha_0\|^2 \leq C(M\delta_n^2 + \|\alpha^* - \alpha_0\|^2 + \frac{M \ln(1/\zeta)}{n}).$$

Theorem 2: *If i)* $E[m(W,\gamma)^2] \leq C \|\gamma\|^2$, *ii)* $\alpha_0(X)$ *and* $Var(Y|X)$ *bounded; iii) for each* ℓ, $\|\hat{\gamma}_\ell - \gamma_0\| \xrightarrow{p} 0$, $\|\hat{\alpha}_\ell - \alpha_0\| \xrightarrow{p} 0$, *and* $\sqrt{n}\|\hat{\gamma}_\ell - \gamma_0\|\|\hat{\alpha}_\ell - \alpha_0\| \xrightarrow{p} 0$ *then*

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

Special case of Lemma 18 of "Locally Robust Semiparametric Estimation."

MONTE CARLO FOR ATE

Second paper above compares performance of Reisz regression NN estimator with Dragon Net of Shi et al. (2019) which uses inverse of NN propensity score estimator in $\hat{\alpha}$.

Monte Carlo design is 1000 semi-synthetic data sets based on the Infant Health and Development Program (IHDP).

IHDP a randomized control trial about effect of home visits and attendance at specialized clinics on future developmental and health outcomes for low birth weight, premature infants (Gross, 1993).

Data using NPCI package in R under setting "A" (Dorie, 2016).

747 observations of an outcome Y, a binary treatment T, and 25 continuous and binary confounders X.

Debiasing function estimator $\hat{\alpha}$ is same as Shi et al. (2019) except use Riesz regression instead of inverse of NN estimator of propensity score Pr($D = 1|Z$).

Results of Monte Carlo using NN for outcome and Riesz regression.

Median Absolute Error $\pm$ std. error

Riesz Reg	Dragon Net
$.110 \pm .003$	$.146 \pm .010$

Also coverage probability of nominal 95% confidence interval was .950.

Perhaps NN Riesz regression is better than inverting estimator of propensity score $\mathsf{Pr}(D = 1|Z)$ when Z is high dimensional and the denominator probability can be close to zero.

GENERALIZED LINEAR REGRESSION

Let Γ be linear, mean square closed set.

A GLR is minimizer of expected objetive function over linear set Γ,

$$\gamma_0 = \arg\min_{\gamma \in \Gamma} E[Q(W, \gamma)].$$

First order condition is, for a constant a,

$$E[b(X)\rho(W, \gamma_0)] = 0 \text{ for all } b \in \Gamma,\ \rho(W, \gamma) = -\frac{\partial}{\partial a} Q(W, \gamma + a). \quad (1)$$

Here $\rho(W, \gamma)$ is a residual and equation (1) specifies that the residual is orthogonal to regression set Γ; what follows depends just on this residual.

Includes conditional quantiles and many other γ_0.

Continue with parameter that is linear in γ_0;

$$\theta_0 = E[m(W, \gamma_0)],\ m(W, \gamma) \text{ is linear in } \gamma.$$

Here let $v_m(x) \in \Gamma$ be Riesz representer,

$$E[m(W, \gamma)] = E[v_m(X)\gamma(X)],\ E[v_m(X)^2] < \infty.$$

Example 2: Average Difference in Log-odds.

Here consider a binary outcome of interest $Y \in \{0, 1\}$

$$\theta_0 = E[\gamma_0(1, Z) - \gamma_0(0, Z)],\ \gamma_0(D, Z) = \ln \frac{\Pr(Y = 1|D, Z)}{\Pr(Y = 0|D, Z)}.$$

This γ_0 is a generalized regression where

$$\rho(W, \gamma) = Y - \mu(\gamma(X)),\ \mu \text{ is logit CDF.}$$

Orthogonal moment function for GLR from Ichimura and Newey (2022) is

$$\psi(w, \gamma, \alpha, \theta) = m(w, \gamma) - \theta + \alpha(x)\rho(w, \gamma), \ \gamma \in \Gamma, \ \alpha \in \Gamma.$$

Here $\alpha_0(x)\rho(w, \gamma)$ "cancels out" first-order effect of γ on on $m(w, \gamma)$.

Suppose there is $\bar{v}_\rho(W)$

$$E[\bar{v}_\rho(W)|X] = -\frac{\partial}{\partial a}E[\rho(W, \gamma_0 + a)|X]\Big|_{t=0}.$$

where a is a scalar.

Normalize so that $v_\rho(X) < 0$.

Weighted Riesz regression is

$$\alpha_0 = \arg\min_{a \in L} E[-2m(W, \alpha) - \bar{v}_\rho(W)\alpha(X)^2].$$

Weight $\bar{v}_\rho(W)$ corrects for effect of γ on $E[\alpha_0(X)\rho(W, \gamma)]$.

For $\hat{v}_\rho(W)$ an estimator of $\bar{v}_\rho(W)$, a weighted Riesz regression estimator of α_0 is

$$\hat{\alpha} = \arg \min_{\alpha \in \mathcal{A}_n \subset L} \{\frac{1}{n} \sum_{i=1}^{n} [-2m(W_i, \alpha) - \hat{v}_\rho(W_i)\alpha(X_i)^2] + \hat{P}_\lambda(\alpha)\}.$$

Example 2: Recall that $\rho(W, \gamma) = Y - \mu(\gamma(X))$; we can use $\hat{v}_\rho(W_i) = -\mu_a(\hat{\gamma}(X))$

for $\mu_a(a) = d\mu_a(a)/da$; a weighted Riesz regression estimator of α_0 is

$$\hat{\alpha} = \arg \min_{\alpha \in \mathcal{A}_n} \{\frac{1}{n} \sum_{i=1}^{n} [-2\{\alpha(1, Z_i) - \alpha(0, Z_i)\} + \mu_a(\hat{\gamma}(X_i))\alpha(X_i)^2]\} + \hat{P}_\lambda(\alpha)\}.$$

Estimator with cross-fitting is constructed analogously to before.

$$\hat{\theta} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \{ m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i) \rho(W_i, \hat{\gamma}_\ell(X_i)) \},$$

$$\hat{V} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_{i\ell}^2, \ \hat{\psi}_{i\ell} = m(W_i, \hat{\gamma}_\ell) - \hat{\theta} + \hat{\alpha}_\ell(X_i) \rho(W_i, \hat{\gamma}_\ell(X_i)).$$

Extensions of Theorems 1 and 2 for generalized regression given in the paper.

These account for presence of $\hat{v}_\rho(W_i)$; convergence rate passes through to $\hat{\alpha}$.

EMPIRICAL EXAMPLE

Question: Does race predict mortgage denial decisions of banks.

Following Munnell et al. (1996), we use the publicly available Boston Home Mortgage Disclosure Act (HDMA) dataset.

The dataset of 2,925 mortgage applications for 1990.in the greater Boston metropolitan area.

We restrict attention to black and white applicants, single-family households (excluding other racial minorities and multi-family residences), which reduces our sample size to 2,380 observations.

12 covariates involving individual characteristics and some credit history.

Our outcome of interest is an indicator $Y = 1$ if the mortgage application was denied.

Estimate three parameters of interest.

1. Difference in Probability of Mortgage Denial:

$$\theta_0 = E[\gamma_0(1, Z) - \gamma_0(0, Z)],\ \gamma_0(D, Z) = E[Y|D, Z].$$

This parameter is an average difference in probability of mortgage denial between a black and a white applicant with the same value of covariates Z.

2. Average Difference in Log-Odds of Mortgage Denial:

$$\theta_0 = E[\gamma_0(1, Z) - \gamma_0(0, Z)],\ \gamma_0(D, Z) = \ln \frac{\Pr(Y = 1|D, Z)}{\Pr(Y = 0|D, Z)}.$$

Approximate average percentage difference in odds of mortgage denial between a black and a white applicant with the same value of covariates Z.

3. Average Difference in Odds of Mortgage Denial:

$$\theta_0 = E[\exp(\gamma_0(1, Z)) - \exp(\gamma_0(0, Z))],\ \gamma_0(D, Z) = \ln \frac{\Pr(Y = 1|D, Z)}{\Pr(Y = 0|D, Z)}.$$

Results: Average over covariates of racial differences in probability of denial.

Mortgage Denial Estimated Using Neural Nets

Probability		Log-Odds		Odds	
est	se	est	se	est	se
0.080	(0.021)	0.829	(0.152)	0.157	(0.044)

NONLINEAR FUNCTIONS OF γ

Can also allow for $m(W, \gamma)$ to be nonlinear in γ.

Let $\hat{D}(W, \alpha) = \partial m(W, \hat{\gamma} + t\alpha)/\partial t|_{t=0}$ for scalar t be derivative of $m(W, \gamma)$ with respect to γ in direction α at $\hat{\gamma}$.

Form $\hat{\alpha}$ by replacing $m(W, \alpha)$ by $\hat{D}(W, \alpha)$ in Riesz regression, i.e.

$$\hat{\alpha} = \text{arg} \min_{\alpha \in \mathcal{A}_n} \{\frac{1}{n} \sum_{i=1}^{n} [-2\hat{D}(W_i, \alpha) - \hat{v}_\rho(W_i)\alpha(X_i)^2] + \hat{P}_\lambda(\alpha)\}.$$

Estimators $\hat{\theta}$ and $\hat{V}$ same as above.

SUMMARY

Automatic debiased machine learners using Riesz regression.

Enables debiased machine learning using neural nets, random forests and other machine learners.

Performs well in Monte Carlo examples.

Debiased machine learners for GLR

Empirical examples.