Improved Double Machine Learning Estimation for Multivariate Outcomes
Outline
Kyle Colangelo

I. Introduction

A. Goals of the paper:

i. Develop a framework and theory for analyzing multiple outcomes simultaneously with double machine learning

ii. Develop a new estimator to improve efficiency when considering multiple outcomes simultaneously, including the case of missing data.

iii. Double Machine Learning is a very general framework so we consider a select few special cases. The binary average treatment effect and the partial linear model.

B. Motivation

i. Analyzing multiple outcomes jointly can improve efficiency

ii. Multi-output algorithms and transfer learning are already in wide use for predictive problems, there are likely advantages to using them for causal inference.

iii. Testing of joint hypotheses for multiple outcomes

iv. Andrew Ng said in 2016 that transfer learning will be the next driver of commercial success in machine learning. It may be useful to develop a framework for causal inference that uses similar methods.

v. In research where specific outcomes have substantial missing data, the power of any analysis might be so low as to be very uninformative. There are various areas where algorithms have been adapted to this kind of setting to compensate for missing data in predictive problems. This is particularly the case for nonparametric problems where power is reduced.

C. Examples of advantages of multi-output ML for predictive problems

i. Examples of when/why they are used

D. Examples of applications where the new approach make sense and can have potential advantages

E. Contributions

i. Introduce a way to estimate a fully nonparametric SUR model with DML

ii. Construct a novel new estimator which improves efficiency over equation-by-equation DML

iii. Improve statistical power when some outcomes have substantial missing data

iv. The development of restricted DML as a by-product of this paper may be of independent interest

v. The development of a stein-type estimator for GMM may be of independent interest

F. Literature Review

   i. Double machine learning. Why use it?
   ii. SUR advantages/disadvantages
   iii. Multi-output machine learning. Advantages to doing it
   iv. Transfer Learning
   v. Stein-type averaging

G. How is the paper organized

   i. Framework and theory generalizing DML to multiple outcomes
   ii. Theory introducing the averaging estimator to improve efficiency for given nuisance estimators.
   iii. Theory for the case of missing data
   iv. Simulations
   v. Empirical Application
   vi. Summary/Conclusion

II. Theory Part I: Multiple Outcomes

A. Goals of this section:

   i. Establish notation and baseline estimators for each case considered
   ii. Precisely state assumptions and theorems for joint normality of causal parameters from all equations
   iii. Establish an estimator for the causal parameter covariance matrix for each case considered

B. Problem Framework

   i. Model

$$\begin{aligned} Y_1 &= g_1(T, X) + \varepsilon_1, \\ Y_2 &= g_2(T, X) + \varepsilon_2, \\ &\vdots \\ Y_m &= g_m(T, X) + \varepsilon_m, \end{aligned}$$

$$Y = \boldsymbol{g}(T, X) + \varepsilon$$

   ii. Special Cases:
      a. Fully linear model (SUR)
      b. Partial linear model (SUR-PLM)

C. Main Assumptions

Assumption 1 (Random Sample)**.** (Y_i, T_i, X_i) *are i.i.d.*

Assumption 2 (Exogeneity)**.** $E(\varepsilon|X) = 0$

Assumption 3 (Error Variance)**.** *$E(\varepsilon_i\varepsilon_j') = \sigma_{ij}I_m$ $(E(\varepsilon\varepsilon') = \Sigma \otimes I_m)$*

where $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ & & \vdots & \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{bmatrix}$

Assumption 4 (Efficiency)**.** *$Y_{1i}, Y_{0i} \perp T_i | X_i$*

D. ATE (binary)

  i. Model

    a. Additionally assume $T = m(X) + V$

  ii. Causal Objects of Interest:

    a. ATE: $E[\boldsymbol{g}(1, X) - \boldsymbol{g}(0, X)]$

  iii. Estimator

    a. Split the sample into L subgroups. Defined I_l as the set of observations corresponding to the l^{th} subgroup and I_l^C as the complement set of observations (all observations not in l)

    b. Estimate $\mathbf{g}$ *jointly* using multi-output ML for each subgroup l (along with m), by fitting the models on I_l^C and evaluating them on I_l

    c. ATE:

$$\hat{\boldsymbol{\beta}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\phi}(W_i, \hat{\eta}),$$

which denotes a vector of ATE's, one for each of the m equations considered. Where $W_i = (Y_i, T_i, X_i)$ and $\hat{\eta} = (\hat{\boldsymbol{g}}, \hat{m})$, and $\boldsymbol{\phi}(W_i; \hat{\eta}_i) = (\phi_1(W_i; \hat{\eta}_i), ..., \phi_m(W_i; \hat{\eta}_i))' = \hat{\boldsymbol{g}}(1, X_i) - \hat{\boldsymbol{g}}(0, X_i) + \frac{T_i(Y_i - \hat{\boldsymbol{g}}(1,X_i))}{\hat{m}(X_i)} - \frac{(1-T_i)(Y_i - \hat{\boldsymbol{g}}(0,X_i))}{1-\hat{m}(X_i)}$.

    d. Further define $\boldsymbol{\psi}(W; \boldsymbol{\beta}, \eta) = \boldsymbol{\phi}(W; \eta) - \boldsymbol{\beta}$

    e. Alternatively we could minimize the Neyman-orthogonal moment conditions with a GMM type estimator. The moment conditions would be $E[\boldsymbol{\psi}(W_i; \boldsymbol{\beta}, \eta_i)] = 0$

  iv. Asymptotic theory

    a. Given the same assumptions to Chernozhukov et al. (2018) consistency and asymptotic normality are established for each individual ATE. The only difference so far is that we are suggesting estimating the nuisance functions jointly. If the new nuisance estimators still satisfies the necessary convergence properties, the results should still hold.

    b. With assumptions 1-3 we have the following result for joint normality:

Theorem 1 (Joint Normality-ATE))**.** *Under assumptions 1-3, and the assumptions in Chernozhukov et al. (2018) we have:*

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, V)$$

Where

$$V = \begin{bmatrix} E[\psi_{1i}^2] & E[\psi_{1i}\psi_{2i}] & \cdots & E[\psi_{1i}\psi_{mi}] \\ E[\psi_{2i}\psi_{1i}] & E[\psi_{2i}^2] & \cdots & E[\psi_{2i}\psi_{mi}] \\ & & \vdots & \\ E[\psi_{mi}\psi_{1i}] & E[\psi_{mi}\psi_{2i}] & \cdots & E[\psi_{mi}^2] \end{bmatrix}$$

Proof. From previous results in Chernozhukov et al. (2018) we have the following asymptotic linear representation:

$$\begin{aligned}\sqrt{N}(\hat{\beta} - \beta) &= \sqrt{N}(\frac{1}{N}\sum_{i=1}^{N}\psi_i) + o_p(1) \\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_i + o_p(1)\end{aligned}$$

Given assumption 1 we can apply the Lindberg Multivariate Central Limit Theorem

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_i + o_p(1) \xrightarrow{d} N(0, V)$$

where

$$V = \begin{bmatrix} Var(\phi_{1i}) & Cov(\phi_{1i}, \psi_{2i}) & \cdots & Cov(\phi_{1i}, \phi_{mi}) \\ Cov(\phi_{2i}, \phi_{1i}) & Var(\phi_{2i}) & \cdots & Cov(\phi_{2i}, \phi_{mi}) \\ & & \vdots & \\ Cov(\phi_{mi}, \phi_{1i}) & Cov(\phi_{mi}, \phi_{2i}) & \cdots & Var(\phi_{1i}) \end{bmatrix}$$

From previous results we know the functional form of $Var(\psi_{1i})$. We will now derive the functional form of $Cov(\psi_1\psi_2)$. By the law of total variance we have:

$$\begin{aligned} V = Var(\boldsymbol{\phi}_i) &= Var(E(\boldsymbol{\phi}_i|T,X)) + E(Var(\boldsymbol{\phi}_i|T,X)) \\ &= Var\Big(E\big(\mathbf{g}(1,X) - \mathbf{g}(0,X) + T\frac{Y - g(1,X)}{m(X)} \\ &- (1-T)\frac{Y - g(0,X)}{1 - m(X)}|T,X\big)\Big) + E\Big(Var\big(\mathbf{g}(1,X) - \mathbf{g}(0,X) \\ &+ T\frac{Y - g(1,X)}{m(X)} - (1-T)\frac{Y - g(0,X)}{1 - m(X)}|T,X\big)\Big) \\ &= Var\Big(g(1,X) - g(0,X)\Big) + E\Big(Var\big(\frac{TY}{m(X)} - \frac{(1-T)Y}{1 - m(X)}|T,X\big)\Big) \\ &= E\Big((g(1,X) - g(0,X) - \beta)^2\Big) \\ &+ E\Big(\big(\frac{T}{m(X)} - \frac{(1-T)}{(1 - m(X))}\big)^2 Var(Y|T,X)\Big) \end{aligned}$$

Taking the 2nd term we note that $T = 1_{\{T=1\}}$ and $(1-T) = 1_{\{T=0\}}$, furthermore distributing the square we get:

$$= E\Big(\Big(\frac{1_{\{T=1\}}}{m(X)^2} + \frac{1_{\{T=0\}}}{(1-m(X))^2}\Big)Var(Y|T,X)\Big)$$

When applying the square, the indicator functions are unchanged and the product term is always equal to zero. We can apply the law of iterated expectations, and then apply the definition of the inner expectation, multiplying

$$\begin{aligned}
&= E\Big(E\Big(\Big(\frac{1_{\{T=1\}}}{m(X)^2} + \frac{1_{\{T=0\}}}{(1-m(X))^2}\Big)Var(Y|T,X)|X\Big)\Big)\\
&= E\Big(m(X)\Big(\frac{1}{m(X)^2}\Big)\Sigma_1(X) + (1-m(X))\Big(\frac{1}{(1-m(X))}\Big)\Sigma_0(X)|X\Big)
\end{aligned}$$

Therefore finally we have:

$$V = E\Big(\Big(\frac{\boldsymbol{\Sigma}_1(X)}{m(X)} + \frac{\boldsymbol{\Sigma}_0(X)}{(1-m(X))}\Big) + (\mathbf{g}(1,X) - \mathbf{g}(0,X) - \beta)^2\Big)$$

□

v. Covariance matrix estimator

a. Sample Analog

$$\hat{V} = \begin{bmatrix}
\frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{1i}^2 & \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{1i}\hat{\psi}_{2i} & \cdots & \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{1i}\hat{\psi}_{mi}\\
\frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{1i}\hat{\psi}_{2i} & \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{2i}^2 & \cdots & \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{2i}\hat{\psi}_{mi}\\
 & & \vdots & \\
\frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{1i}\hat{\psi}_{mi} & \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{2i}\hat{\psi}_{mi} & \cdots & \frac{1}{N}\sum_{i=1}^{N}\hat{\psi}_{mi}^2
\end{bmatrix}$$

E. PLM

i. Adjusted Model:

$$\begin{aligned}
Y_1 &= \beta_1 T + g_1(X) + \varepsilon_1,\\
Y_2 &= \beta_2 T + g_2(X) + \varepsilon_2,\\
&\vdots\\
Y_m &= \beta_m T + g_m(X) + \varepsilon_m,
\end{aligned}$$

ii. Causal Object of Interest: $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)'$

iii. ***Need to work out PLM, will wait until ATE is completely fleshed out**

iv. In the PLM case i believe a GLS type approach would be applicable given how the DML estimation is performed in this case. The treatment of PLM may end up being entirely different than for the ATE.

F. General Case

   i. Model: Any arbitrary model from which sufficient neyman-orthogonal moment conditions can be constructed.

   ii. Estimator

      a. Define $\bar{\psi}_N = \frac{1}{N}\sum_{i=1}^{N} \psi(W;\beta,\eta)$

      b. $\hat{\beta} = \arg\min_{\beta} \; \bar{\psi}_N' \bar{\psi}_N$

G. Special Case: Strictly linear SUR model

III. Theory Part II: Restricted Estimator

A. Goals of this section:

   i. Propose a restricted DML estimator for each case and derive their properties.

   ii. Precisely define the constraint

B. ATE (Binary)

   i. GMM version of estimator

      a. It is helpful to think of the DML estimators as GMM type estimators for the purpose of placing restrictions on the estimators

      b. $\hat{\beta} = \arg\min_{\beta} \; \bar{\psi}_N' W \bar{\psi}_N$

      c. In the normal DML case the weighting matrix is irrelevant.

      d. Note that $\bar{\psi}_N = \bar{\phi}_N - \boldsymbol{\beta}$

      e. Define $\Omega = E\big[\psi_i \psi_i'\big]$

      f. Constrained estimation problem:

$$\begin{aligned}\tilde{\beta} =& \arg\min_{\beta} \; (\bar{\phi}_N - \boldsymbol{\beta})'\Omega^{-1}(\bar{\phi}_N - \boldsymbol{\beta}) \\ &\text{s.t.} \quad R'\boldsymbol{\beta} = 0\end{aligned}$$

   ii. The constraint:

      a.

$$\beta_1 = \beta_2 = \cdots = \beta_m = \bar{\beta}$$

Where $\bar{\beta}$ is a weighted average of the β's, following Mehrabani and Ullah (2020)

$$\bar{\beta} = \left(J'\Omega^{-1}J\right)^{-1} J'\Omega^{-1}\beta$$

In general we may consider other constraints which can result in differing properties of the final estimator.

b. We can precisely define this constraint in matrix notation as:

$$\begin{bmatrix} \beta_1 - \bar{\beta} \\ \beta_2 - \bar{\beta} \\ \vdots \\ \beta_m - \bar{\beta} \end{bmatrix} = R'\boldsymbol{\beta} = 0$$

Where $R = I - J\big(J'\Omega^{-1}J\big)^{-1}J'\Omega^{-1}$ is idempotent.

iii. Performing the optimization:

$$\begin{aligned} &\min_{\beta} \ \ (\bar{\phi}_N - \boldsymbol{\beta})'\Omega^{-1}(\bar{\phi}_N - \boldsymbol{\beta}) \\ &\text{s.t.} \quad R'\boldsymbol{\beta} = 0 \end{aligned}$$

$$\begin{aligned} &\mathcal{L}(\beta, \lambda) = \frac{1}{2}\big(\bar{\phi}_N'\Omega^{-1}\bar{\phi}_N - 2\bar{\phi}_N'\Omega^{-1}\beta + \beta'\Omega^{-1}\beta\big) + \lambda'(R'\beta) \\ &\frac{\partial L}{\partial \beta} = -\Omega^{-1}\bar{\phi}_N + \Omega^{-1}\tilde{\beta} + R\tilde{\lambda} = 0 \\ &\frac{\partial L}{\partial \lambda} = R'\tilde{\beta} = 0 \end{aligned}$$

Multiplying the derivative with respect to β by $R'\Omega$ we get

$$\begin{aligned} &R'\hat{\beta} + R'\tilde{\beta} + R'\Omega R\tilde{\lambda} = 0 \\ &\tilde{\lambda} = (R'\Omega R)^{-1}R'\hat{\beta} \end{aligned}$$

Plugging in and solving for $\tilde{\beta}$ we get:

$$\begin{aligned} \tilde{\beta} &= \hat{\beta} - \Omega R(R'\Omega R)^{-1}R'\hat{\beta} \\ &= (I - \Omega R(R'\Omega R)^{-1}R')\hat{\beta} \\ &= (I - R)\hat{\beta} \end{aligned}$$

Which is infeasible, so we can construct a feasible version of this estimator:

$$\tilde{\beta}_F = (I - \hat{R})\hat{\beta}$$

In general we could set any linear constraint $R'\beta = c$ and derive a very similar estimator.

iv. Properties:

C. PLM case:

i. ***needs work**

IV. Theory Part III: Averaging Estimator

A. General Concept: Balance bias and variance by constructing a weighted average of the constrained and unconstrained estimators.

B. Averaging Estimator

$$\hat{\beta}_A = \left(1 - \frac{\tau}{D}\right)\hat{\boldsymbol{\beta}} - \frac{\tau}{D}\tilde{\boldsymbol{\beta}}$$

Where $D = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'W(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})$, which measures the distance between the constrained and unconstrained estimators.

C. Bias

D. MSEM

V. Simulations ***Code has been written for the estimators. I am determining precisely what I want to do and then I can run them***

A. Graphs for RMSE vs. correlation

B. Coverage Rate for Joint tests

C. Missing data

VI. Empirical Application ***Still need to determine where i want to apply the new approach***

A. Application without missing data

B. Application with missing data

VII. Summary/Conclusion

References

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, "Double/debiased machine learning for treatment and structural parameters," 2018.

Mehrabani, Ali and Aman Ullah, "Improved Average Estimation in Seemingly Unrelated Regressions," *Econometrics*, 2020, *8* (2), 15.