ECMA 31380 - Proposal Final Project

Fernando Rocha Urbano

Autumn 2024

1 Goal

The goal of this project is to evaluate the performance of different Double Machine Learning (DML) models in recovering the Average Treatment Effect (ATE) across various causal inference scenarios addressed in simulations.

The desired broader impact of the project is to provide actionable recommendations for selecting the most appropriate DML method based on the causal scenario, data structure, noise levels, and sample size. Additionally, we aim to create a benchmarking framework that practitioners can use to evaluate and compare causal inference estimates derived from the tested methods in different scenarios.

The DML methods compared in this study are:

- LASSO.
- Random Forest.
- Neural Networks.

The causal inference scenarios are:

- Backdoor adjustment.
- Frontdoor adjustment with and without access to mediators.
- Instrumental Variables with varying levels of instrument strength and correct vs. incorrect use of the IV.

We explore how each model implementation performs under varying conditions:

- Relationships as linear and non-linear between covariates and treatment and between covariates and target ($g(\cdot)$ and $m(\cdot)$).
- Noise levels (ε).
- Sample size (n).
- Size of high-dimensional impactful covariates (d_c).

- Size of high-dimensional irrelevant covariates (d_a).

For each of the scenario combinations, we show which of LASSO, Random Forest and Neural Networks provides more accurate ATE estimates based on the following metrics.

- Bias of $\hat{\text{ATE}}$.
- Variance of $\hat{\text{ATE}}$ across repeated simulations for a given scenario.
- Analytical Variance of the $\hat{\text{ATE}}$.
- MSE (Mean Squared Error).

2 Methodology

2.1 Problem Setup

Consider a random sample $(Y_i, T_i, X_i)_{i=1}^n$, where:

- $Y_i \in \mathbb{R}$ is the outcome variable.
- $T_i \in 0, 1$ is a binary treatment indicator.
- $X_i \in \mathbb{R}^p$ is a vector of covariates.

Our goal is to estimate the Average Treatment Effect (ATE), defined as:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)], \tag{2.1}$$

where $Y_i(t)$ denotes the potential outcome for unit i under treatment $T_i = t$.

2.2 Identification via Conditional Expectations

Under the Conditional Independence Assumption (CIA) and overlap conditions, the ATE can be identified as:

$$\tau = \mathbb{E}[\mu_1(X_i) - \mu_0(X_i)], \tag{2.2}$$

where $\mu_t(X_i) = \mathbb{E}[Y_i|T_i = t, X_i]$ is the conditional expectation of the outcome given treatment and covariates.

2.3 Double Machine Learning (DML) Estimator

The DML framework aims to estimate τ while controlling for high-dimensional or complex relationships between Y_i, T_i, and X_i. The key idea is to use machine learning methods to estimate the nuisance parameters and then construct an estimator for τ that is robust to estimation errors in these nuisance parameters.

2.3.1 Nuisance Parameter Estimation

We define the following nuisance functions:

$$m(X_i) = \mathbb{E}[Y_i|X_i], \tag{2.3}$$
$$g(X_i) = \mathbb{E}[T_i|X_i], \tag{2.4}$$
$$\pi(X_i) = \mathbb{P}(T_i = 1|X_i) \quad \text{(propensity score)} \tag{2.5}$$

These functions can be estimated using flexible machine learning methods such as LASSO, Random Forests, or Neural Networks.

Under mild regularity conditions and appropriate rates of convergence for the nuisance estimators, the DML estimator is root-n consistent and asymptotically normal. It provides valid confidence intervals and hypothesis tests even when using complex machine learning methods for $\hat{m}(X)$ and $\hat{g}(X)$.

2.3.2 Orthogonal Score Function

To achieve robustness, we construct an orthogonal score function $\psi(Y_i, T_i, X_i; \eta)$, where $\eta = (m, g)$ represents the nuisance parameters. The orthogonal score satisfies the Neyman orthogonality condition, which ensures that small estimation errors in η have a negligible first-order impact on the estimation of τ.

A common choice for the orthogonal score is:

$$\psi(Y_i, T_i, X_i; \eta) = \left(\frac{T_i - g(X_i)}{\pi(X_i)(1 - \pi(X_i))} \right) (Y_i - m(X_i)) + (m_1(X_i) - m_0(X_i)) - \tau, \tag{2.6}$$

where $g(X_i) = \pi(X_i)$ in case of binary treatment and:

$$m_t(X_i) = \mathbb{E}[Y_i|T_i = t, X_i] \quad \text{for } t \in \{0, 1\} \tag{2.7}$$

2.3.3 Estimation Procedure

The estimation proceeds in several steps:

1. Estimate Nuisance Functions: Use one of the three outline ML models to obtain estimators $\hat{m}(X_i)$ and $\hat{g}(X_i)$ for the nuisance functions.
2. Compute the Score Function: Evaluate the orthogonal score $\psi(Y_i, T_i, X_i; \hat{\eta})$ using the estimated nuisance parameters.
3. Estimate τ: Solve the empirical moment condition which yields $\hat{\tau}$:

$$\frac{1}{n} \sum_{i=1}^{n} \psi(Y_i, T_i, X_i; \hat{\eta}) = 0. \tag{2.8}$$

The orthogonal score function mitigates the impact of errors in $\hat{m}(X)$ and $\hat{g}(X)$ on the estimation of τ while also accomodating the use of modern ML techniques, allowing complex relationships between Y, X and T.

To prevent overfitting and ensure that the estimation error in the nuisance parameters does not bias the estimator of τ we employ cross-fitting.

2.3.4 Cross-Fitting

The steps of cross-fitting are:

1. Split the Sample: Divide the data into K folds $\{\mathcal{I}_k\}k = 1^K$.
2. For Each Fold:
   (a) Train Nuisance Estimators: Use data from all other folds:
   $$\mathcal{I}_{-k} = \bigcup j \neq k\mathcal{I}_j$$
   to estimate $\hat{m}^{(-k)}(X_i)$ and $\hat{g}^{(-k)}(X_i)$.
   (b) Compute Score Function: For observations in fold $\mathcal{I}_k$, compute $\psi(Y_i, T_i, X_i; \hat{\eta}^{(-k)})$ using the nuisance estimates from step (a):
   $$\begin{aligned}\psi(Y_i, T_i, X_i; \hat{\eta}^{(-k)}) = &\left(\frac{T_i - \hat{g}^{(-k)}(X_i)}{\hat{\pi}^{(-k)}(X_i)}\right)(Y_i - \hat{m}^{(-k)}(X_i)) \\ &+ \hat{m}_1^{(-k)}(X_i) - \hat{m}_0^{(-k)}(X_i) - \tau^{(-k)}.\end{aligned} \tag{2.9}$$
3. Aggregate: Combine the estimates from all folds:
   $$\hat{\tau} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\hat{\tau}_i^{(-k)} = \frac{1}{K}\sum_{k=1}^{K}\hat{\tau}^{(-k)} \tag{2.10}$$

3 Causal Inference Scenarios

The three most relevant scenarios of causal inference are the ones that require Instrumental Variables, Backdoor Adjustment, or Front Door Adjustment. A common way to represent the causal relation related to those is through the use of DAGs.

3.1 DAG (Directed Acyclic Graph)

Directed Acyclic Graph (DAG) serves as a representation of causal assumptions and a tool for deriving statistical properties of the variables involved.

It is composed of nodes (vertices) representing random variables or features and directed edges (arrows), indicating a direct influence or causal effect of T on Y.

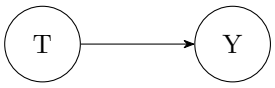

Figure 1: DAG Example

The acyclic characterist is due to absence of directed cycles; that is, there is no path where you can start at a node X and, by following directed edges, return to X.

3.2 Backdoor Adjustment

A backdoor path from treatment T to the outcome Y represents alternative routes through which association can flow from T to Y that are not due to the causal effect of T on Y. In the representation, the confounder C is a common cause of both T and Y, thus, a backdoor path.

In such case, the association between T and Y may be partially or entirely due to their mutual dependence on C rather than a direct causal effect, leading to biased causal estimates of the treatment if C is ignored.

The causal effect of T on Y can be expressed using the backdoor adjustment formula:

$$\mathbb{P}[Y(t)] = \sum_{C} \mathbb{P}[Y \mid T, C]\mathbb{P}[C], \tag{3.1}$$

Which serves a markov factorization, calculating with respect to the DAG structure.

For the backdoor adjustment to be valid, the following conditions must be satisfied:

1. No variable in the adjustment set is a descendant of the treatment T.
2. The adjustment set blocks all backdoor paths from T to Y (backdoor path is any path from T to Y that starts with an arrow into T).

DAG 3.2 illustrates the backdoor path involving the confounder C, treatment T, outcome Y, and features with non-causal association X_a which would not be present in a typical backdoor adjustment DAG, but play a relevant role in the simulations proposed.

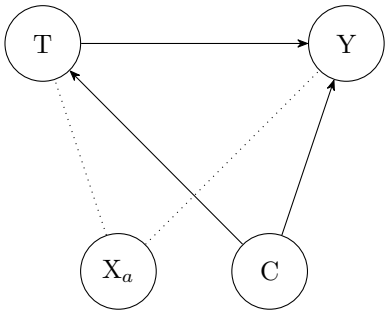

Figure 2: DAG of Backdoor Path

3.2.1 Backdoor Adjustment Scenario in DML

In Backdoor Adjustment with DML, the ideal $m(X_i) = \mathbb{E}[Y_i|X_i]$ and $g(X_i) = \mathbb{E}[T_i|X_i]$ from Equations (2.3) and (2.4) are:

$$m_{\text{BA}}(C_i) = \mathbb{E}[Y_i|C_i], \tag{3.2}$$
$$g_{\text{BA}}(C_i) = \mathbb{E}[T_i|C_i], \tag{3.3}$$

In our simulations we also test the habilit of the ML models to estimate ATE under:

$$m^{\text{w1}}_{\text{BA}}(C_i, X_{a,i}) = \mathbb{E}[Y_i|C_i, X_{a,i}], \tag{3.4}$$
$$g^{\text{w1}}_{\text{BA}}(C_i, X_{a,i}) = \mathbb{E}[T_i|C_i, X_{a,i}], \tag{3.5}$$

Equations (3.4) and (3.5) represent a scenario in which one would not be aware that X_a does not cause T and Y.

3.3 Frontdoor Adjustment

Frontdoor adjustment is a method used to estimate the causal effect of treatment T on outcome Y when there is unmeasured confounding that cannot be addressed using backdoor adjustment. It leverages a mediator M that lies on the causal path from T to Y.

The causal effect of T on Y can be expressed using the frontdoor adjustment formula:

$$\mathbb{P}[Y(t)] = \sum_{M} \mathbb{P}[M \mid T = t] \sum_{t'} \mathbb{P}[Y \mid M, T = t']\mathbb{P}[T = t']. \tag{3.6}$$

For instance, the average treatment effect in case $M, T \in \{0, 1\}$:

$$\tau = [P(M = 1 \mid T = 1) - P(M = 1 \mid T = 0)] \times [E[Y \mid M = 1] - E[Y \mid M = 0]] \tag{3.7}$$

For the frontdoor adjustment to be valid, the following conditions must be satisfied:

1. All causal paths from T to Y pass through M (i.e., there is no direct effect of T on Y bypassing M).
2. There are no unmeasured confounders between T and M.
3. All backdoor paths from M to Y are blocked by T (i.e., there are no unmeasured confounders between M and Y that are not affected by T).

DAG 3.3 illustrates the frontdoor adjustment involving the treatment T, mediator M, outcome Y, observed confounders C and features with no causal association X_a. C and X_a would not be present in a typical frontdoor adjustment DAG, but are included due to their relevance in the proposed simulations.

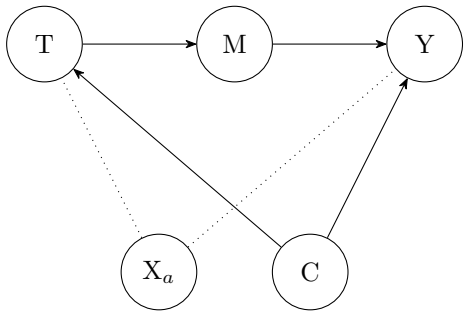

Figure 3: DAG of Frontdoor Path

3.3.1 Frontdoor Adjustment Scenario in DML

In our simulations, we use a binary mediator $M_i \in \{0, 1\}$, similar to the binary treatment T_i.

In the Frontdoor Adjustment scenario with DML, we need to account for the mediator M_i when estimating the ATE. The identification of the causal effect involves modeling the relationships between T_i, M_i, and Y_i.

The ideal nuisance functions for DML in this scenario are:

$$m_{\text{FA}}(M_i, C_i) = \mathbb{E}[Y_i \mid M_i, C_i], \tag{3.8}$$
$$h_{\text{FA}}(T_i) = \mathbb{E}(M_i \mid T_i), \tag{3.9}$$
$$g_{\text{FA}}(C_i) = \mathbb{E}(T_i \mid C_i), \tag{3.10}$$

Here $h_{\text{FA}}(M_i, T_i, C_i)$ is the mediator model, representing the probability of the mediator given treatment.

To adapt the orthogonal score function in the presence of the mediator, we modify Equation (2.6) to incorporate the mediator's effect. The adapted orthogonal score function is:

$$\psi(Y_i, T_i, M_i, C_i; \eta) = \left(\frac{T_i - g_{\text{FA}}(C_i)}{g_{\text{FA}}(C_i)(1 - g_{\text{FA}}(C_i))} \right) (M_i - h_{\text{FA}}(T_i)) (Y_i - m_{\text{FA}}(M_i, C_i)) + \delta_M \delta_Y(C_i) - \tau \tag{3.11}$$

where:

$$\delta_M = h_{\text{FA}}(T_i = 1) - h_{\text{FA}}(T_i = 0), \tag{3.12}$$
$$\delta_Y(C_i) = m_{\text{FA}}(M_i = 1, C_i) - m_{\text{FA}}(M_i = 0, C_i), \tag{3.13}$$

and η represents the collection of nuisance functions.

In this score function:

- The first term adjusts for the treatment assignment, similar to the original score function, but now includes the mediator.

- The product $(M_i - h_{\text{FA}}(T_i))\,(Y_i - m_{\text{FA}}(M_i, C_i))$ captures the interaction between the mediator and the outcome.
- The term $\delta_M \delta_Y(C_i)$ represents the estimated causal effect based on the mediator and outcome models.

In our simulations we also test the hability of the ML models to estimate ATE under the following scenario.

Inclusion of C_i and $X_{a,i}$ on every nuisance function

$$m_{\text{FA}}^{\text{w}_1}(M_i, C_i, X_{a,i}) = \mathbb{E}[Y_i \mid M_i, C_i, X_{a,i}], \tag{3.14}$$
$$h_{\text{FA}}^{\text{w}_1}(T_i, C_i, X_{a,i}) = \mathbb{E}(M_i \mid C_i, X_{a,i}), \tag{3.15}$$
$$g_{\text{FA}}^{\text{w}_1}(C_i, X_{a,i}) = \mathbb{E}(T_i \mid C_i, X_{a,i}), \tag{3.16}$$

Equations (3.14), (3.15), and (3.16) represent a scenario where one is unaware that $X_{a,i}$ has no causal relation to T_i, M_i, or Y_i. Equation (3.15) represents scenarion where one is unaware that C has no causal relation to M_i.

Ignoring the mediator M_i

$$m_{\text{FA}}^{\text{w}_2}(C_i, X_{a,i}) = \mathbb{E}[Y_i \mid C_i, X_{a,i}], \tag{3.17}$$
$$g_{\text{FA}}^{\text{w}_2}(C_i, X_{a,i}) = \mathbb{E}[T_i \mid C_i, X_{a,i}], \tag{3.18}$$

Equations (3.17) and (3.18) represent a scenario where the mediator M_i is unavailable or ignored and one is unaware that $X_{a,i}$ has no causal relationship to T or Y. In this case, the orthogonal score function is the same as in Equation (2.6).

Considering the mediator M_i as a normal covariate

$$m_{\text{FA}}^{\text{w}_3}(C_i, M_i, X_{a,i}) = \mathbb{E}[Y_i \mid C_i, M_i, X_{a,i}], \tag{3.19}$$
$$g_{\text{FA}}^{\text{w}_3}(C_i, M_i, X_{a,i}) = \mathbb{E}[T_i \mid C_i, M_i, X_{a,i}], \tag{3.20}$$

Equations (3.19) to (3.20) also represent scenarios where one is unaware that $X_{a,i}$ has no causal relation to T_i or Y_i, and the scenario where M_i is available but not recognized as a mediator. In this case, the orthogonal score function is the same as in Equation (2.6).

3.4 Instrumental Variable

Instrumental variable (IV) estimation is a method used to estimate the causal effect of a treatment T on an outcome Y when there is unmeasured confounding that cannot be addressed using backdoor or frontdoor adjustments. This method leverages an instrument Z, which influences the treatment T

but has no direct effect on the outcome Y except through T, and is independent of any unmeasured confounders U affecting both T and Y.

The causal effect of T on Y can be estimated using the instrumental variable formula:

$$\hat{\tau} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, T)}. \tag{3.21}$$

Alternatively, in terms of expectations for a binary instrument Z:

$$\hat{\tau} = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[T \mid Z = 1] - \mathbb{E}[T \mid Z = 0]}. \tag{3.22}$$

For the instrumental variable method to be valid, the following conditions must be satisfied:

1. Relevance: The instrument Z is associated with the treatment T (i.e., $\text{Cov}(Z, T) \neq 0$ or $Z \not\perp\!\!\!\perp T$).
2. Exclusion Restriction: The instrument Z affects the outcome Y only through its effect on the treatment T (i.e., there is no direct effect of Z on Y and no other pathways from Z to Y except through T).
3. Independence (Ignorability): The instrument Z is independent of any unmeasured confounders U that affect both T and Y (*i.e.*, $Z \perp\!\!\!\perp U$).

DAG 3.4 illustrates the instrumental variable setup involving the unobserved confounder U, instrument Z, treatment T, outcome Y, observed confounder C, and features with non-causal associations X_a. Again, the last two would not be present in a typical instrumental variable DAG, but are included due to their relevance in the proposed simulations.

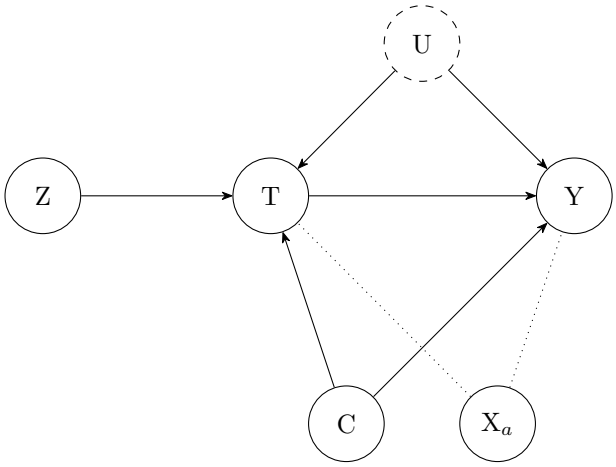

Figure 4: DAG of Instrumental Variable

3.4.1 Instrumental Variable Scenario in DML

This formulation adjusts for the binary nature of T_i and incorporates the propensity scores directly.

In Instrumental Variable Scenario with DML, the ideal $m(X_i) = \mathbb{E}[Y_i|X_i]$ and $g(X_i) = \mathbb{E}[T_i|X_i]$ from Equations (2.3) and (2.4) are:

$$m_{\text{IV}}(C_i) = \mathbb{E}[Y_i|C_i], \tag{3.23}$$
$$g_{\text{IV}}(C_i, Z_i) = \mathbb{E}[T_i|C_i, Z_i], \tag{3.24}$$

The orthogonal score function from (2.6) becomes:

$$\psi_{\text{IV}}(Y_i, T_i, Z_i, X_i; \theta, \eta) = \left(\frac{T_i - g(X_i)}{g(X_i)(1 - g(X_i))} \right) (Y_i - m(X_i) - \tau[T_i - g(X_i)]) \cdot [g(Z_i, X_i) - g(X_i)]. \tag{3.25}$$

In our simulations we also test the habilit of the ML models to estimate ATE under:

$$m^{\text{w}}_{\text{IV}}(C_i, Z_i, X_{a,i}) = \mathbb{E}[Y_i|C_i, Z_i, X_{a,i}], \tag{3.26}$$
$$g^{\text{w}}_{\text{IV}}(C_i, Z_i, X_{a,i}) = \mathbb{E}[T_i|C_i, Z_i, X_{a,i}], \tag{3.27}$$

Equation (3.33) represents a scenario in which one would not be aware that the the instrument Z only has a causal relation to T. Equations (3.33) and (3.35) represent a scenario in which one would not be aware that X_a has no causal relation to T and Y.

3.4.2 Instrumental Variable Scenario in DML

In the Instrumental Variable (IV) scenario with DML, we aim to estimate the causal effect of the treatment T_i on the outcome Y_i using an instrument Z_i that affects T_i but has no direct effect on Y_i except through T_i. The presence of unobserved confounders U_i that affect both T_i and Y_i violates the Conditional Independence Assumption, necessitating the use of instrumental variables.

To adapt the DML framework to the IV setting, we need to define appropriate nuisance functions and construct an orthogonal score function suitable for the IV context.

The ideal nuisance functions in this scenario are:

$$m_{\text{IV}}(X_i) = \mathbb{E}[Y_i \mid X_i], \tag{3.28}$$
$$q_{\text{IV}}(Z_i, X_i) = \mathbb{E}[T_i \mid Z_i, X_i], \tag{3.29}$$
$$g_{\text{IV}}(X_i) = \mathbb{E}[T_i \mid X_i], \tag{3.30}$$

Here:

- $m_{\text{IV}}(X_i)$ is the outcome model, capturing the expected outcome given covariates. - $q_{\text{IV}}(Z_i, X_i)$ is the instrument propensity score, representing the expected treatment given the instrument and covariates. - $g_{\text{IV}}(X_i)$ is the treatment model, representing the expected treatment given covariates.

The orthogonal score function for the IV scenario is different from the standard DML orthogonal score function. An appropriate orthogonal score function in the linear IV context is:

$$\psi_{\mathrm{IV}}(W_i;\theta,\eta) = (Y_i - m_{\mathrm{IV}}(X_i) - \theta[T_i - g_{\mathrm{IV}}(X_i)])\,(q_{\mathrm{IV}}(Z_i,X_i) - g_{\mathrm{IV}}(X_i))\,, \tag{3.31}$$

where:

- $W_i = (Y_i, T_i, Z_i, X_i)$. - θ is the parameter of interest (the causal effect of T_i on Y_i). - η represents the collection of nuisance functions.

This score function leverages the variation in T_i induced by the instrument Z_i while controlling for X_i. It satisfies the Neyman orthogonality condition, making it robust to estimation errors in the nuisance functions.

The estimation procedure involves:

1. **Estimate Nuisance Functions**: Use flexible machine learning methods to estimate $\hat{m}_{\mathrm{IV}}(X_i)$, $\hat{q}_{\mathrm{IV}}(Z_i,X_i)$, and $\hat{g}_{\mathrm{IV}}(X_i)$.
2. **Implement Cross-Fitting**: Split the data into folds, estimate the nuisance functions on training folds, and predict on validation folds to mitigate overfitting biases.
3. **Compute the Orthogonal Score**: Evaluate $\psi_{\mathrm{IV}}(W_i;\theta,\hat{\eta})$ for each observation using the estimated nuisance functions.
4. **Estimate** θ: Solve the empirical moment condition:

$$\frac{1}{n}\sum_{i=1}^{n}\psi_{\mathrm{IV}}(W_i;\hat{\theta},\hat{\eta}) = 0, \tag{3.32}$$

   which yields the estimator $\hat{\theta}$.

In our simulations, we also test the ability of the ML models to estimate the causal effect under different scenarios:

$$m^{\mathrm{w}}_{\mathrm{IV}}(X_i,Z_i,X_{a,i}) = \mathbb{E}[Y_i \mid X_i,Z_i,X_{a,i}], \tag{3.33}$$
$$q^{\mathrm{w}}_{\mathrm{IV}}(Z_i,X_i,X_{a,i}) = \mathbb{E}[T_i \mid Z_i,X_i,X_{a,i}], \tag{3.34}$$
$$g^{\mathrm{w}}_{\mathrm{IV}}(X_i,X_{a,i}) = \mathbb{E}[T_i \mid X_i,X_{a,i}], \tag{3.35}$$

These equations represent scenarios where irrelevant variables $X_{a,i}$, which have no causal relation to T_i or Y_i, are included in the models. Equation (3.33) also represents a scenario where one incorrectly includes the instrument Z_i in the outcome model, violating the exclusion restriction.

For the instrumental variable method to be valid, the following conditions must be satisfied:

1. **Relevance**: The instrument Z_i is associated with the treatment T_i (i.e., $\mathbb{E}[T_i \mid Z_i,X_i] \neq \mathbb{E}[T_i \mid X_i]$).
2. **Exclusion Restriction**: The instrument Z_i affects the outcome Y_i only through its effect on the treatment T_i, and not directly or through any other pathways.

3. **Independence (Ignorability)**: The instrument Z_i is independent of any unmeasured confounders U_i that affect both T_i and Y_i, conditional on covariates X_i.

By appropriately specifying and estimating the nuisance functions, and constructing the orthogonal score function as in Equation (3.31), the DML framework can provide consistent and asymptotically normal estimates of the causal effect in the IV scenario, even in the presence of high-dimensional covariates and complex relationships among variables.