Causal Estimation with Matrices: 2SLS and DML
Matrix Theory Writing Sample

Zachary Olea

August 11th 2024

1 Introduction

In areas of applied statistics, such as econometrics, understanding and quantifying casual relationships is crucial for determining the effect of one event on another. Generally, causal models seek to identify the effect of a treatment or independent variable on an outcome or dependent variable while controlling for confounding factors that could influence this relation. There are a variety of methodologies used to establish causalities and this paper serves to explore the matrix theory behind two of them, Two Stage Least Squares and Double Machine Learning. The role matricies play in these causal models are within their systems of linear equations and how they facilitate the computation of estimators that quantify the causal effects.

History

In the 40's & 50's, the methodology for Two-Stage Least Squares (2SLS) was developed as a solution to the problem of endogeneity in linear regression models, which occur when an independent variable is correlated with the error term. This leads to biased and inconsistent OLS estimates and is due to various factors like ommited variable bias, measurement error, or simultaneous causality. With the advent of machine learning models and their increased usage, the necessity for establishing causality eventually led to developments such as the methodology called Double Machine Learning (DML) in the 2017 paper "Double/Debiased Machine Learning for Treatment and Causal Parameters[1],

2 Two Stage Least Squares

2.0.1 OLS

Let us begin with the matrix representation of the linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

Where,
$\mathbf{y}$: $n \times 1$ dependent or outcome variable.
$\mathbf{X}$: $n \times k$ matrix of independent or explanatory variables (treatment and control variables).
$\boldsymbol{\beta}$: $k \times 1$ vector of coefficients which we aim to estimate.
$\boldsymbol{\varepsilon}$: $n \times 1$ vector of error terms or residuals.
The goal of the OLS model is to estimate $\boldsymbol{\beta}$, which tells us how changes in the explanatory variables $\mathbf{X}$ impact the outcome $\mathbf{y}$. For causal inference we need to determine whether the estimated relationship is causal or merely correlation. Using matrices, we can construct and manipulate various estimators that can control for confounding variables to determine casual effects.

2.1 2SLS and IV

Two-Stage Least Squares addresses the issue when one or more columns of $\mathbf{X}$ (as in multiple dependent variables) are endogenous, as in correlated with the error term ε. 2SLS mitigates this by using instrumental

variables (IVs) in a two-stage process: first, it projects the endogenous variables onto a space where they are exogenous using the IVs, and then applies OLS to the transformed variables. The chosen instruments influence the endogenous variables but have no direct effect on the dependent variable, isolating the exogenous variation in the explanatory variables.

2.1.1 Stage 1: Instrumental Variable and Projection

Let $\mathbf{Z}$ be an $n \times m$ matrix of IVs, where $m \geq k$. The IVs should be correlated with the variables in $\mathbf{X}$ that are endogenous, but is uncorrelated with ε. For the first stage, we project $\mathbf{X}$ onto the space spanned by $\mathbf{Z}$ to get the predicted values $\hat{\mathbf{X}}$,

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{X} \tag{2}$$

Where,
$\mathbf{Z}^\top \mathbf{Z}$ is an $m \times m$ matrix.
$(\mathbf{Z}^\top \mathbf{Z})^{-1}$ is the inverse of the $m \times m$ matrix (assuming that it is invertible).
$\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top$ is an $n \times n$ projection matrix that maps $\mathbf{X}$ into the space spanned by $\mathbf{Z}$.

2.1.2 Stage 2: Regression on Predicted Values

Next, we regress $\mathbf{y}$ on the predicted values $\hat{\mathbf{X}}$ to obtain the 2SLS estimator for β,

$$\hat{\beta}_{\text{2SLS}} = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^\top \mathbf{y} \tag{3}$$

Substituting $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{X}$ from the first stage we get,

$$\hat{\beta}_{\text{2SLS}} = \left[\mathbf{X}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{X}\right]^{-1} \mathbf{X}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{y} \tag{4}$$

This gives us the 2SLS estimator which is consistent even with endogeneity, as long as the IVs in $\mathbf{Z}$ are valid. The projection matrix $\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top$ from 2SLS is symmetric and idempotent ($\mathbf{P}_Z^2 = \mathbf{P}_Z$). For the 2SLS estimator to be identified, $\mathbf{Z}$ must have full column rank, meaning that $\mathbf{Z}^\top \mathbf{Z}$ is invertible. Under regularity conditions, $\hat{\beta}_{\text{2SLS}}$ is asymptotically normally distributed,

$$\hat{\beta}_{\text{2SLS}} \sim \mathcal{N}\left(\beta, \sigma^2(\mathbf{X}^\top \mathbf{P}_Z \mathbf{X})^{-1}\right) \tag{5}$$

where σ^2 is the variance of the error term ε.

2.1.3 Example

Consider a model with a single endogenous regressor X_1 and a single exogenous regressor X_2,

$$y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{6}$$

Suppose we identified an IV Z for X_1 where,

$$\mathbf{X} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} Z \end{bmatrix} \tag{7}$$

Then the projection of $\mathbf{X}$ onto $\mathbf{Z}$ is,

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{X} \tag{8}$$

Thus the 2SLS estimator for β_1 and β_2 would be,

$$\hat{\beta}_{\text{2SLS}} = \left(\hat{\mathbf{X}}^\top \hat{\mathbf{X}}\right)^{-1} \hat{\mathbf{X}}^\top \mathbf{y} \tag{9}$$

$$\hat{\beta}_{\text{2SLS}} = \left[\left(\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{X}\right)^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{X}\right]^{-1} \left(\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{X}\right)^\top \mathbf{y} \tag{10}$$

$$\hat{\beta}_{\text{2SLS}} = \left[\mathbf{X}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{X}\right]^{-1} \mathbf{X}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{y} \tag{11}$$

By using IV in this way, 2SLS addresses the endogenity problem and ensures that the estimated coefficients represent causal relationships rather than spurious correlation. We can further solve the 2SLS estimator using the SVD of the IV. The SVD of the instrument matrix $\mathbf{Z}$ is,

$$\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^\top \tag{12}$$

Where,
$\mathbf{U}$ is an $n \times n$ orthogonal matrix.
$\mathbf{D}$ is an $n \times m$ diagonal matrix (with singular values on the diagonal).
$\mathbf{V}$ is an $m \times m$ orthogonal matrix.
Next we substitute the SVD of $\mathbf{Z}$ into the projection formula,

$$\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top = \mathbf{U}\mathbf{D}\mathbf{V}^\top \left(\mathbf{V}\mathbf{D}^\top\mathbf{U}^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top\right)^{-1}\mathbf{U}\mathbf{D}\mathbf{V}^\top \tag{13}$$

Since $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$,

$$\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top = \mathbf{U}\mathbf{D}\mathbf{D}^{-2}\mathbf{D}^\top\mathbf{U}^\top \tag{14}$$

$$= \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^\top \tag{15}$$

Substituting the SVD projection back to the 2SLS estimator we have,

$$\hat{\beta}_{\text{2SLS}} = \left[\mathbf{X}^\top\mathbf{U}\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{X}\right]^{-1}\mathbf{X}^\top\mathbf{U}\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{y} \tag{16}$$

SVD is one way to solve for the pseudo-inverse in the case where $\mathbf{Z}^\top\mathbf{Z}$ may not be invertible directly. In particular, SVD is useful for obtaining the 2SLS estimator when there is strong multicollinearity. When there is multicollinearity, some of the columns of matrix $\mathbf{X}$ (or $\mathbf{Z}$ for 2SLS) are nearly linearly dependent. If that is the case then $\mathbf{X}^\top\mathbf{X}$ or $\mathbf{Z}^\top\mathbf{Z}$ are nearly singular, as in the determinant is close to 0 and is thus difficult to invert. SVD decomposes the matrix into singular values and vectors which provide a clear indication of the rank deficiency or the extent of multicollinearity. Very small singular values can be treated as zero in a process called regularization which effectively reduces the rank of the matrix and helps stabilize the solution such as is with the process of L2 Regularization, also known as Ridge Regression.

3 Double Machine Learning

3.1 Background

Next we will cover a high level overview of Double Machine Learning (DML). The point of DML is in its handling of high-dimensional data while ensuring consistent estimation of causal parameters. DML involves constructing orthogonal scores and debiasing the estimates obtained from ML models.

3.1.1 Matrix

Consider the problem of estimating the treatment effect θ in a partially linear model,

$$Y = T\theta + g(X) + \varepsilon \tag{17}$$

Where,
Y is the outcome variable.
T is the treatment variable.
X is a vector of covariates.
$g(X)$ is a non-parametric function of X that captures the confounding effect.
- ε is the error term.
The goal of DML is to estimate θ consistently, even when $g(X)$ is unknown and complex. DML models estimate a nuisance parameter that include,
The outcome model $m(X) = \mathbb{E}[Y|X]$.

The treatment model $p(X) = \mathbb{E}[T|X]$.
ML or non-parametric methods are used to estimate $\hat{m}(X)$ and $\hat{p}(X)$. DML seeks to construct an orthogonal score that is robust to errors in the estimation of the nuisance parameters. The orthogonal score is defined as,

$$\psi(W, \theta) = (Y - \hat{m}(X))\,(T - \hat{p}(X)) - \theta\,(T - \hat{p}(X))^2 \tag{18}$$

where $W = (Y, T, X)$ represents the observed data. A matrix representation for this can be presented as, $\mathbf{Y}$ is the $n \times 1$ vector of outcomes.
$\mathbf{T}$ is the $n \times 1$ vector of treatments.
$\mathbf{X}$ is the $n \times p$ matrix of covariates.
$\hat{\mathbf{m}} = \hat{m}(\mathbf{X})$ is the $n \times 1$ vector of predicted outcomes.
$\hat{\mathbf{p}} = \hat{p}(\mathbf{X})$ is the $n \times 1$ vector of predicted treatments.
The orthogonal score can then be expressed in matrix notation as,

$$\boldsymbol{\psi}(\mathbf{W}, \theta) = (\mathbf{Y} - \hat{\mathbf{m}}) \odot (\mathbf{T} - \hat{\mathbf{p}}) - \theta(\mathbf{T} - \hat{\mathbf{p}}) \odot (\mathbf{T} - \hat{\mathbf{p}}) \tag{19}$$

where $\odot$ denotes the Hadamard (element-wise) product.

3.2 Debias

For the debiased estimation, the DML estimator for θ is obtained with the method of moments estimator,

$$\frac{1}{n}\sum_{i=1}^{n} \psi(W_i, \theta) = 0 \tag{20}$$

In matrix form, this becomes,

$$\frac{1}{n}\mathbf{1}^\top \left[(\mathbf{Y} - \hat{\mathbf{m}}) \odot (\mathbf{T} - \hat{\mathbf{p}})\right] - \theta\frac{1}{n}\mathbf{1}^\top \left[(\mathbf{T} - \hat{\mathbf{p}}) \odot (\mathbf{T} - \hat{\mathbf{p}})\right] = 0 \tag{21}$$

Solving for θ,

$$\hat{\theta}_{\text{DML}} = \frac{\frac{1}{n}\mathbf{1}^\top \left[(\mathbf{Y} - \hat{\mathbf{m}}) \odot (\mathbf{T} - \hat{\mathbf{p}})\right]}{\frac{1}{n}\mathbf{1}^\top \left[(\mathbf{T} - \hat{\mathbf{p}}) \odot (\mathbf{T} - \hat{\mathbf{p}})\right]} \tag{22}$$

This DML estimator adjusts for potential biases in the estimation of the nuisance parameters $m(X)$ and $p(X)$.

3.3 Orthogonal

Lastly, the orthogonality condition ensures that small errors in estimating $\hat{\mathbf{m}}$ and $\hat{\mathbf{p}}$ do not bias the estimation of θ. ML models tend to overfit so this is key when dealing with high-dimensional models. Mathematically this orthogonality is reflected by the partial derivative of the moment function with respect to the nuisance parameters is zero,

$$\frac{\partial \mathbb{E}[\psi(W, \theta)]}{\partial m(X)} = 0 \quad \text{and} \quad \frac{\partial \mathbb{E}[\psi(W, \theta)]}{\partial p(X)} = 0 \tag{23}$$

Thus, even if $\hat{\mathbf{m}}$ or $\hat{\mathbf{p}}$ are slightly misspecified, the impact on $\hat{\theta}_{\text{DML}}$ is minimal. To get a slightly clearer picture of what is happening, I will use DML with L2 regularization (Ridge regression) as discussed previously with 2SLS and SVD.

4 Example: L2 Regularization (Ridge)

As mentioned earlier, SVD can provide a way to compute Ridge Regression efficiently. Ridge regression is a statistical machine learning technique that helps in the presence of multicollinearity by adding an L2 regularization term to the loss function to penalize large coefficients. The objective function is given as,

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\} \tag{24}$$

Where,
$\mathbf{y}$ is the $n \times 1$ vector of observed outcomes,
$\mathbf{X}$ is the $n \times p$ matrix of covariates,
β is the $p \times 1$ vector of coefficients,
$\lambda > 0$ is the regularization parameter.
A closed-form solution for Ridge regression is,

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \tag{25}$$

The SVD of $\mathbf{X}$ allows us to express the Ridge regression solution as:

$$\hat{\beta} = \mathbf{V}(\mathbf{\Sigma}^\top \mathbf{\Sigma} + \lambda \mathbf{I})^{-1} \mathbf{\Sigma}^\top \mathbf{U}^\top \mathbf{y} \tag{26}$$

Given the SVD of $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \tag{27}$$

the Ridge regression solution can be rewritten as:

$$\hat{\beta} = \mathbf{V}(\mathbf{\Sigma}^\top \mathbf{\Sigma} + \lambda \mathbf{I})^{-1} \mathbf{\Sigma}^\top \mathbf{U}^\top \mathbf{y} \tag{28}$$

This expression shows that the coefficients $\hat{\beta}$ are a linear combination of the singular vectors $\mathbf{V}$, with each singular value σ_i (the diagonal elements of $\mathbf{\Sigma}$) adjusted by the regularization parameter λ.

4.0.1 DML with Ridge Regression

The goal is to estimate the treatment effect θ while controlling for the high-dimensional covariates X using DML with L2 regularization. First we start by splitting the data into two pieces, I_1 and I_2. This is to avoid overfitting by using different subsets for estimating nuisance parameters and constructing the orthogonal score.

4.0.2 First Stage - Nuisance Parameters

First estimate the nuisance parameters of the outcome model $m(X)$ and the treatment model $p(X)$. For $m(X)$ use Ridge regression to estimate $m(X) = \mathbb{E}[Y|X]$. For a given data split I_1, the Ridge regression estimator for $m(X)$ is,

$$\hat{m}(X) = \mathbf{X}\hat{\beta}_m \tag{29}$$

where $\hat{\beta}_m$ is obtained by solving:

$$\hat{\beta}_m = \arg\min_{\beta} \left\{ \frac{1}{|I_1|} \sum_{i \in I_1} (Y_i - \mathbf{X}_i^\top \beta)^2 + \lambda \|\beta\|_2^2 \right\} \tag{30}$$

Here, $\lambda > 0$ is the regularization parameter, and $\|\beta\|_2^2 = \beta^\top \beta$ represents the L2 norm (or squared Euclidean norm). We can then estimate the treatment model $p(X) = \mathbb{E}[T|X]$ using Ridge regression,

$$\hat{p}(X) = \mathbf{X}\hat{\beta}_p \tag{31}$$

where $\hat{\beta}_p$ is obtained by solving,

$$\hat{\beta}_p = \arg\min_{\beta} \left\{ \frac{1}{|I_1|} \sum_{i \in I_1} (T_i - \mathbf{X}_i^\top \beta)^2 + \lambda \|\beta\|_2^2 \right\} \tag{32}$$

4.0.3 Second Stage - Constructing the Orthogonal Score

We then construct the orthogonal score using estimates $\hat{m}(X)$ and $\hat{p}(X)$ obtained in the first stage. For each observation in the data split I_2, the orthogonal score is,

$$\psi(W_i, \theta) = (Y_i - \hat{m}(X_i))(T_i - \hat{p}(X_i)) - \theta(T_i - \hat{p}(X_i))^2 \tag{33}$$

where $W_i = (Y_i, T_i, X_i)$ is the data tuple. We can then solve the DML estimator for θ,

$$\frac{1}{|I_2|} \sum_{i \in I_2} \psi(W_i, \theta) = 0 \tag{34}$$

$$\hat{\theta}_{\text{DML}} = \frac{\frac{1}{|I_2|} \sum_{i \in I_2} (Y_i - \hat{m}(X_i))(T_i - \hat{p}(X_i))}{\frac{1}{|I_2|} \sum_{i \in I_2} (T_i - \hat{p}(X_i))^2} \tag{35}$$

We can further improve the estimator's efficiency and reduce over-fitting by repeating the above but reverse the roles of I_1 and I_2, and average the results.

$$\hat{\theta}_{\text{DML}} = \frac{1}{2} \left(\hat{\theta}^{(1)}_{\text{DML}} + \hat{\theta}^{(2)}_{\text{DML}} \right) \tag{36}$$

where $\hat{\theta}^{(1)}_{\text{DML}}$ is the estimate obtained using I_1 for nuisance parameter estimation and I_2 for the orthogonal score, and $\hat{\theta}^{(2)}_{\text{DML}}$ is the reverse. The full form of the final DML estimator is,

$$\hat{\theta}_{\text{DML}} = \frac{1}{2} \left(\frac{\frac{1}{|I_2|} \sum_{i \in I_2} (Y_i - \hat{m}^{(1)}(X_i))(T_i - \hat{p}^{(1)}(X_i))}{\frac{1}{|I_2|} \sum_{i \in I_2} (T_i - \hat{p}^{(1)}(X_i))^2} + \frac{\frac{1}{|I_1|} \sum_{i \in I_1} (Y_i - \hat{m}^{(2)}(X_i))(T_i - \hat{p}^{(2)}(X_i))}{\frac{1}{|I_1|} \sum_{i \in I_1} (T_i - \hat{p}^{(2)}(X_i))^2} \right) \tag{37}$$

5 Conclusion

Overall we have looked at the contextual application for Two Stage Least Squares and Instrumental Variables and the matrix theory involved, further discussing the use of SVD in calculating the psuedo-inverse for it and how that transitions to L2 regularization. From there we establish the fundamental theory of Double Machine Learning and provide an example of DML with L2 regularization. By leveraging the mathematical properties of matrices, these methodologies provide robust solutions to complex problems of endogeneity, multi-collinearity and high-dimensionality in statistical modeling. The analysis demonstrates how matrices facilitate the computation of consistent and unbiased estimators, ultimately contributing to more accurate and reliable causal inferences. Whereas 2SLS and IVs are a fundamental tool used in applied research for Economics, Finance, Accounting, etc. outside of Economics, I have yet to see double machine learning firmly established as a causal methodology; however, I hope to see this methodology used more in all sorts of applied literature in the future.

References

[1] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and causal parameters, 2017.