Notes on the Propensity Score and Regression Models

Yusufhan Balci
www.chickpeainference.com

November 7, 2025

1 Introduction

The propensity score is one of the key building blocks of any causal inference model. It denotes the likelihood that an individual will be treated given key characteristics like age, socio-economic status, and gender. For example, individuals who grow up in better socio-economic environments might be more likely to pursue higher education. The propensity score allows us to adjust for such cases of selection bias. Now, in the literature, we can find a myriad of different methods to adjust for this bias using the propensity score: ranging from including the fitted propensity in a regression model for the potential outcomes, which I focus on my first PhD project, to inverse probability of treatment weighting (IPTW) methods. The aim of this overview is to demystify the various ways in which the propensity score can be used for regression models in causal inference.

2 Regression Models and Endogeneity

In causal inference, we might like to postulate a model that can both predict potential outcomes and provide valid inference for the causal estimates of interest, like the average treatment effect (ATE) or conditional average treatment effect (CATE).

We specify the following model that caters to these needs:

$$y_i = x_i^\top \gamma + t_i \eta + \varepsilon_{i,y} \tag{1}$$

$$t_i = x_i^\top \zeta + \varepsilon_{i,t} \tag{2}$$

For simplicity, we assume a continuous treatment t_i and a continuous outcome y_i. η denotes the average effect of treatment, while γ and ζ denote, respectively, the relationships between the covariates and y_i and t_i. The structural error terms for the outcome and treatment assignment functions are denoted respectively by $\varepsilon_{i,y}$ and $\varepsilon_{i,t}$. We assume selection on observables, which means that structural errors are not correlated: $\text{Cov}(\varepsilon_{i,y}, \varepsilon_{i,t}) = 0$.

The intuition is that, if we estimate Eq. (1) without properly accounting for Eq. (2), we will run into the problem of endogeneity. This problem stems from two issues. Firstly, due to $x_i^\top \zeta$, there will be multicollinearity, i.e., the different independent variables (x_i and t_i) are correlated with each other. Although this might lead to wider confidence intervals for our

estimate of η, it will not, by itself, bias the OLS estimate of η. However, if we estimate our model with regularization, this can introduce Regularization-Induced Confounding (RIC) and bias our estimates (Hahn et al., 2018).

Secondly, the main culprit for OLS bias is the relationship between t_i and the composite error term of the model. We see this when we substitute Eq. (2) into Eq. (1):

$$y_i = x_i^\top \gamma + (x_i^\top \zeta + \varepsilon_{i,t})\eta + \varepsilon_{i,y} = x_i^\top (\gamma + \zeta\eta) + (\varepsilon_{i,t}\eta + \varepsilon_{i,y}) \tag{3}$$

If we estimate the simple model $y_i = x_i^\top \tilde{\gamma} + t_i\eta + \text{error}$, the full error term is $(\varepsilon_{i,t}\eta + \varepsilon_{i,y})$. We can clearly see that the covariance between the treatment t_i and this error term is non-zero:

$$\text{Cov}(t_i, \varepsilon_{i,t}\eta + \varepsilon_{i,y}) = \text{Cov}(x_i^\top \zeta + \varepsilon_{i,t}, \varepsilon_{i,t}\eta + \varepsilon_{i,y}) \tag{4}$$

Assuming $\text{Cov}(\varepsilon_{i,t}, \varepsilon_{i,y}) = 0$ (selection on observables) and $\text{Cov}(x, \varepsilon) = 0$, this simplifies to:

$$= \text{Cov}(\varepsilon_{i,t}, \varepsilon_{i,t}\eta) \neq 0, \tag{5}$$

which violates a key assumption of OLS. The violation of this assumption leads to biased estimates of η. This is no different from having random measurement errors in our independent variables, which would lead to the same problems.

Now that we have a brief overview what kind of issues are at play, we might look at possible solutions. As the problem of endogeneity is a classical problem in econometrics and statistics, the tools to solve this problem can also be applied in causal inference.

2.1 Instrumental Variables

One of the first solutions that might pop into your head if you are from an economics or econometrics background is instrumental variable (IV) regression. Indeed, the instrumental variable is one of the tools we have against endogeneity and has been widely employed in the literature (Imbens and Klaauw, 1995; Angrist et al., 1996; Duflo, 2001). An instrument Z is chosen such that $\text{Cov}(T, Z)$ is sufficiently large, but $\text{Cov}(Z, \varepsilon_{i,y}) = 0$. Here, T, Z, and $\varepsilon_{i,y}$ denote vectors with the treatment allocation, the instrument, and the outcome residuals. To put it simply, our aim is to find an instrument Z that is not related to $\varepsilon_{i,y}$, but which is sufficiently correlated with the endogenous covariate T. For example, T an indicator whether or how much individuals comply with the treatment regime they are assigned to, while Z an indicator whether an individual for a certain intervention. Another famous example is to use birth year or a lottery draw as an instrument, when analyzing the effect of military conscription Imbens and Klaauw (1995).

We can then use this instrument in a two-stage least squares (2SLS) process. In the first stage, we regress the treatment T onto the instrument Z (and covariates X) to get fitted values:

$$\hat{t}_i = \hat{\xi} z_i + x_i^\top \hat{\delta} \tag{6}$$

After which we can plug $\hat{t}_i$ into Eq. (1) to obtain unbiased estimates for η, the average treatment effect. Although 2SLS has nice properties, finding a suitable instrument in practice is hard.

2.2 The Propensity Score as Additional Covariate

An alternative, when there is no suitable instrument, is to include the fitted treatment regime $\hat{t}_i$ as a covariate in the outcome function, by fitting Eq. (12) and obtaining $\hat{t}_i$, similar to what we do in Eq. (6). The difference here is that we do not have an instrument, and construct this projection of t_i only with the covariates x_i. Linero (2024) shows that in a Bayesian setting, assuming that the priors for γ and ζ are independent leads to a strong prior belief that there is no confounding, which is unrealistic. Including the treatment model fit in the outcome model helps address this. This also helps mitigate Regularization-Induced Confounding (RIC) (Hahn et al., 2018). To solve these aforementioned issues, Hahn et al. (2020); Linero (2024) propose to include the propensity score in their models.

This idea of using the treatment model (Eq. (2)) to help the outcome model (Eq. (1)) is the core of doubly robust methods. To see this more clearly, it is helpful to look at the more common case of a binary treatment, where the treatment model is the propensity score, $\hat{\pi}(x) = P(T = 1|X = x)$. Often the treatment assignment function is considered a nuisance function, and is fitted using flexible machine learning methods.

In that setting, including the propensity score as a covariate shares properties with Targeted Maximum Likelihood Estimation (TMLE) and doubly robust machine learning (Chernozhukov et al., 2017; Van Der Laan and Rubin, 2006). This means that the estimator for η can be consistent as long as either the outcome model (Eq. (1)) or the treatment model (Eq. (2)) is correctly specified. In TMLE, we employ the propensity score to adjust for any bias in the outcome function (Gruber and Van Der Laan, 2009). The intuition is that once an initial estimate for the outcome is obtained, we can "update" it using the propensity score. For a simple model with continuous outcome and *binary* treatment $\tilde{t}_i \in (0, 1)$:

$$y_i = \hat{y}_i + \phi H(\tilde{t}_i, x_i) \tag{7}$$

and

$$H(\tilde{t}_i, x_i) = \frac{1(\tilde{t}_i = 1)}{\hat{\pi}(x)} - \frac{1(\tilde{t}_i = 0)}{1 - \hat{\pi}(x)} \tag{8}$$

, where we have $\hat{y}_i$ the initial estimate of the outcome, and a clever covariate $H(\tilde{t}_i, x_i)$ depending on the inverse propensity weights. TMLE is based on the efficient influence function and the example given above can be generalized for continuous treatments as well. Please see Vansteelandt and Dukes (2022) for further technical details.

2.3 Structural Equation Modeling (BDML)

While including the fitted treatment model has benefits, the derivations are not always straightforward for complex machine learning methods (Souto and Louzada, 2024). A more integrated solution is to model Eq. (1) and Eq. (2) as one structural equation model (SEM) with a multivariate normal distribution.

In the Bayesian Framework, DiTraglia and Liu (2025) present this approach as *Bayesian Double Machine Learning (BDML)*. Instead of estimating η in the outcome function directly, we first re-parameterize the model. We substitute the treatment model (Eq. (2)) into the

outcome model (Eq. (1)) to create a new, reduced-form system:

$$y_i = x_i^\top \gamma + (\overbrace{x_i^\top \zeta + \varepsilon_{i,t}}^{t_i})\eta + \varepsilon_{i,y} \tag{9}$$

$$y_i = x_i^\top (\gamma + \eta\zeta) + (\eta\varepsilon_{i,t} + \varepsilon_{i,y}) \tag{10}$$

Define new reduced-form parameters $\delta \equiv \gamma + \eta\zeta$ and a new reduced-form error $U_i \equiv \eta\varepsilon_{i,t} + \varepsilon_{i,y}$. This gives the bivariate reduced-form model:

$$y_i = x_i^\top \delta + U_i \tag{11}$$

$$t_i = x_i^\top \zeta + \varepsilon_{i,t} \tag{12}$$

We now have a standard Bayesian multivariate regression, where we estimate the coefficients (δ, ζ) and the 2×2 covariance matrix $\mathbf{\Sigma}$ of the reduced-form errors $(U_i, \varepsilon_{i,t})$.

The key insight is that our parameter of interest, η, is now recoverable from the components of this new covariance matrix $\mathbf{\Sigma}$. Let $\sigma_t^2 = \mathrm{Var}(\varepsilon_{i,t})$ and $\sigma_{Ut} = \mathrm{Cov}(U_i, \varepsilon_{i,t})$.

$$\sigma_{Ut} = \mathrm{Cov}(\eta\varepsilon_{i,t} + \varepsilon_{i,y}, \varepsilon_{i,t}) \tag{13}$$

$$= \eta\,\mathrm{Cov}(\varepsilon_{i,t}, \varepsilon_{i,t}) + \mathrm{Cov}(\varepsilon_{i,y}, \varepsilon_{i,t}) \tag{14}$$

$$= \eta\,\mathrm{Var}(\varepsilon_{i,t}) + 0 \tag{15}$$

The last step holds because our initial selection-on-observables assumption states that the structural errors are uncorrelated, $\mathrm{Cov}(\varepsilon_{i,y}, \varepsilon_{i,t}) = 0$.

This gives us the simple relationship: $\sigma_{Ut} = \eta\sigma_t^2$. Therefore, we can recover the ATE η simply by dividing the covariance of the reduced-form errors by the variance of the treatment-model error:

$$\eta = \frac{\mathrm{Cov}(U_i, \varepsilon_{i,t})}{\mathrm{Var}(\varepsilon_{i,t})} = \frac{\sigma_{Ut}}{\sigma_t^2} \tag{16}$$

The BDML approach, then, is to place priors on the parameters of the reduced-form model $(\delta, \zeta, \mathbf{\Sigma})$, run MCMC to get posterior samples for $\mathbf{\Sigma}$, and then for each sample, compute the posterior draw for η using this ratio. This avoids the problem of RIC and prior dogmatism that can arise from estimating the structural model directly. Potentially, this approach could be extended such that the covariance matrix is estimated conditional on the underlying covariates x_i, this would result in an expression of η conditional on the covariates allowing us to describe the CATE.

3 Conclusion

Navigating causal inference in the presence of confounding is a complex task. While simple OLS is biased by endogeneity, and instrumental variables are often hard to find, a new class of methods has emerged to tackle the problem directly.

We have seen that the challenge deepens with regularization, which can introduce its own biases (RIC). The solution lies in modeling both the outcome and the treatment process. This can be done through a two-step control function approach, as seen in the work of Hahn et al. (2018) and Linero (2024).

However, a more integrated approach, as proposed by DiTraglia and Liu (2025), is to jointly model the system using a reduced-form re-parameterization. By modeling the bivariate system for (y_i, t_i) and estimating their reduced-form error covariance, we can cleverly recover the structural causal effect η (our ATE). This Bayesian Double Machine Learning (BDML) approach not only aligns with the principles of doubly robust estimation but also avoids the pitfalls of prior dogmatism and RIC that can plague simpler high-dimensional models.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265.

DiTraglia, F. J. and Liu, L. (2025). Bayesian double machine learning for causal inference. *arXiv preprint arXiv:2508.12688*.

Duflo, E. (2001). Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American economic review*, 91(4):795–813.

Gruber, S. and Van Der Laan, M. J. (2009). Targeted maximum likelihood estimation: A gentle introduction.

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. (2018). Regularization and confounding in linear regression for treatment effect estimation.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056.

Imbens, G. and Klaauw, W. V. D. (1995). Evaluating the cost of conscription in the netherlands. *Journal of Business & Economic Statistics*, 13(2):207–215.

Linero, A. R. (2024). In nonparametric and high-dimensional models, bayesian ignorability is an informative prior. *Journal of the American Statistical Association*, 119(548):2785–2798.

Souto, H. G. and Louzada, F. (2024). Ablation studies for novel treatment effect estimation models. *arXiv preprint arXiv:2410.15560*.

Van Der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1).

Vansteelandt, S. and Dukes, O. (2022). Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):657–685.