NOTE FOR CAUSALML

Xiaowei Yin

1 Predictive inference

1.1 (Chap1)Predictive Inference with Linear Regression in Moderately High Dimensions

Linear Regression: Inference about Predictive Effects via partialling out

$$Y = \beta_1 D + \beta_2' W + \varepsilon$$

$$\tilde{Y} = \beta_1 \tilde{D} + \varepsilon, \quad \mathrm{E}[\varepsilon \tilde{D}] = 0$$

Theorem 1.1 (Frisch-Waugh-Lovell, FWL)**.** *The population linear regression coefficient β_1 can be recovered from the population linear regression of $\tilde{Y}$ on $\tilde{D}$:*

$$\beta_1 = \arg\min_{b_1} \mathrm{E}\left[\left(\tilde{Y} - b_1 \tilde{D}\right)^2\right] = \left(\mathrm{E}\left[\tilde{D}^2\right]\right)^{-1} \mathrm{E}[\tilde{D}\tilde{Y}],$$

where we assume D cannot be perfectly predicted by W , i.e., $\mathrm{E}\left[\tilde{D}^2\right] > 0$, so β_1 is uniquely defined.

Theorem 1.2 (Adaptive Inference)**.** *Under regularity conditions and if $p/n \approx 0$, the estimation error in $\check{D}_i$ and $\check{Y}_i$ has no first order effect on the stochastic behavior of $\hat{\beta}_1$. Namely,*

$$\sqrt{n}\left(\hat{\beta}_1 - \beta_1\right) \approx \sqrt{n}\mathbb{E}_n[\tilde{D}\varepsilon]/\mathbb{E}_n\left[\tilde{D}^2\right]$$

and consequently,

$$\sqrt{n}\left(\hat{\beta}_1 - \beta_1\right) \overset{\mathrm{a}}{\sim} N(0,\ \mathrm{V})$$

where

$$\mathrm{V} = \left(\mathrm{E}\left[\tilde{D}^2\right]\right)^{-1} \mathrm{E}\left[\tilde{D}^2\varepsilon^2\right] \left(\mathrm{E}\left[\tilde{D}^2\right]\right)^{-1}$$

1.2 (chap3) Predictive Inference with High Dimentional Covairates– Penalized regression

Lasso: Classical linear regression or least squares fails in these high-dimensional settings because it overfits in finite samples. So Lasso is used in this situation to avoid overfitting.

$$\hat{\beta}_j = 0 \text{ if } \left| \frac{\partial}{\partial \hat{\beta}_j} \sum_i \left(Y_i - \hat{\beta}' X_i \right)^2 \right| < \lambda$$

Note: Lasso estimator is set to 0 if the marginal predictive benefit of changing β away from 0 is smaller than the marginal increase in penalty. Also indicate that the penalty λ should dominate the noise in the measurement of the marginal predictive ability.

Post Lasso:We can use the Lasso-selected set of regressors, those regressors whose Lasso coefficient estimates are non-zero, to refit the model by least squares.

Ridge: penalty term given by the sum of the squared values of the coefficients times a penalty level λ.

Elastic Net: penalty given by a linear combination of the Lasso and Ridge penalties.

Lava: penalty given by a linear combination of the lasso and Ridge. But the parameters are split into a dense part and sparse part that put into different penalty term. It is designed to work well in the sparse + dense settings.

The following theorems provide theoretical understanding of the predictive performance of Lasso.

Definition 1.1 (Approximate sparsity)**.** *The sorted absolute values of the coefficients decay quickly. Specifically, the j^{th} largest coefficient (in absolute value) denoted by $|\beta|_{(j)}$ obeys*

$$|\beta|_{(j)} \leq A j^{-a}, \quad a > 1/2,$$

for each j , where the constants a and A do not depend on the sample sizen.

Theorem 1.3. *Under approximate sparsity as defined in Definition , restricted isometry conditions stated below, choosing λ , and other regularity conditions stated , with probability approaching $1 - \alpha$ as $n \to \infty$, the following bound holds:*

$$\sqrt{\mathrm{E}_X \left[\left(\beta' X - \hat{\beta}' X \right)^2 \right]} \leq \text{const} \cdot \sqrt{\mathrm{E}\left[\epsilon^2\right]} \sqrt{\frac{s \log(\max\{p, n\})}{n}}$$

where E_X denotes expectation with respect to X , and the effective dimension is

$$s = \ const \ \cdot A^{1/a} \cdot n^{\frac{1}{2a}},$$

where constant a is the speed of decay of the sorted coefficient values in the approximate sparsity definition, Definition 3.1.1. Moreover, the number of regressors selected by Lasso is bounded by

$$const \cdot S$$

with probability approaching 1 - a as $n \to \infty$. The constants const are different in different places and may depend on the distribution of (Y, X) and on a.

Note: Under approximate sparsity (s¡¡n) and with appropriate choice of penalty parameters, Lasso and Post-Lasso will approximate the best linear predictor well.

Definition 1.2 (Restricted Isometry)**.** *The following conditions hold:*

$$\text{Uniformly in } Z \subset X : \dim(Z) \leq L = s\log(n),$$

$$\sup_{\|a\|=1} |a' \left(\mathbb{E}_n \left[ZZ'\right] - E\left[ZZ'\right]\right) a| \approx 0$$

$$0 < C_1 \leq \inf_{\|a\|=1} a' E\left[ZZ'\right] a \leq C_2 < \infty,$$

where C_1 and C_2 are constants.

Note: This condition says that "small groups" of regressors are not collinear and are well-behaved.

1.3 (chap4) Statistical Inference on Predictive and Causal Effects in High-Dimensional Linear Regression Models

Double Lasso :

1. We run Lasso regressions of Y_i on W_i and D_i on W_i

$$\hat{\gamma}_{YW} = \arg\min_{\gamma \in \mathbb{R}^p} \sum_i \left(Y_i - \gamma' W_i\right)^2 + \lambda_1 \sum_j \hat{\psi}_j^{\gamma} |\gamma_j|$$

$$\hat{\gamma}_{DW} = \arg\min_{\gamma \in \mathbb{R}^p} \sum_i \left(D_i - \gamma' W_i\right)^2 + \lambda_2 \sum_j \hat{\psi}_j^{D} |\gamma_j|,$$

and obtain the resulting residuals:

$$\check{Y}_i = Y_i - \hat{\gamma}'_{YW} W_i$$

$$\check{D}_i = D_i - \hat{\gamma}'_{DW} W_i.$$

2. We run the least squares regression of $\hat{Y}_i$ on $\hat{D}_i$ obtain the estimator $\hat{\alpha}$:

$$\hat{\alpha} = \arg\min_{a \in \mathbb{R}} \mathbb{E}_n \left[(\check{Y} - a\check{D})^2\right]$$

$$= \left(\mathbb{E}_n \left[\check{D}^2\right]\right)^{-1} \mathbb{E}_n[\check{D}\check{Y}].$$

We can use standard results from this regression, ignoring that the input variables were previously estimated, to perform inference about the predictive effect, α.

Theorem 1.4 (Adaptive Inference with Double Lasso in High-Dimensional Regression)**.** *Under the stated approximate sparsity, the conditions required for Theorem 3.2.1 (e.g. restricted isometry), and additional regularity conditions, the estimation error in $\check{D}_i$ and $\check{Y}_i$ has no first order effect on $\hat{\alpha}$, and*

$$\sqrt{n}(\hat{\alpha} - \alpha) \approx \sqrt{n}\mathbb{E}_n[\tilde{D}\epsilon]/\mathbb{E}_n\left[\tilde{D}^2\right] \overset{\text{a}}{\sim} N(0,\ \mathrm{V}),$$

where

$$\mathrm{V} = \left(\mathrm{E}\left[\tilde{D}^2\right]\right)^{-1} \mathrm{E}\left[\tilde{D}^2\epsilon^2\right] \left(\mathrm{E}\left[\tilde{D}^2\right]\right)^{-1}$$

Neyman Orthogonality of Double Lasso : In the Double Lasso method, we estimate α though the prior estimation of γ_{DW} and γ_{YW} in the partialling out procedure. These are called nuisance parameters with the true value

$$\eta^o = (\gamma'_{DW}, \gamma'_{YW})'$$

and we consider the explicitly dependence of $\hat{\alpha}$ on the nuisance parameters: $\hat{\alpha}(\eta)$

In the double lasso procedure, we constructs the residuals:

$$\check{Y}_i(\eta) = Y_i - \eta'_1 W_i, \quad \check{D}_i(\eta) = D_i - \eta'_2 W_i$$

and solve the population moment equation

$$\mathrm{M}(a, \eta) := \mathrm{E}[(\tilde{Y}(\eta) - a\tilde{D}(\eta))\tilde{D}(\eta)] = 0$$

which again implicitly defines the function $\hat{\alpha}(\eta)$.

The main idea of the Double Lasso approach is that, in the population limit, it corresponds to a procedure for learning the target parameter α that is first-order insensitive to local perturbations of the nuisance parameters around their true values, η^0:

$$\partial_\eta \alpha\left(\eta^o\right) = 0$$

Note: We will call the local insensitivity of target parameters to nuisance parameters Neyman orthogonality of the estimation process. It is important in high dimensional settings where we generally use regularization for estimation. The use of regularization generally results in bias in the nuisance parameters. Neyman orthognoality guarantees that the target parameter is locally insensitive to perturbations of the nuisance parameters around their true value, then ensures that the bias does not transmit to the estimation of the target parameter, at least to the first order.

Next we show the proof of Neyman Orthognality of Double Lasso process:

Proof: Since the function $\alpha(\eta)$ is implicitly defined as the solution to the equation $M(\alpha, \eta) = 0$, by the implicit function theorem and letting $\alpha = \alpha(\eta^0)$

$$\partial_\eta \alpha\left(\eta^o\right) = -\partial_a \mathrm{M}\left(\alpha, \eta^o\right)^{-1} \partial_\eta \mathrm{M}\left(\alpha, \eta^o\right)$$

the second term consists of 2 components:

$$\partial_{\eta_1} \mathrm{M}(\alpha, \eta^o) = \mathrm{E}\left[W\tilde{D}(\eta^o)\right] = \mathrm{E}\left[W\left(D - \gamma'_{DW}W\right)\right] = 0$$

$$\begin{aligned}\partial_{\eta_2} \mathrm{M}(\alpha, \eta^o) &= -\mathrm{E}\left[W\tilde{Y}(\eta^o)\right] + 2\mathrm{E}\left[\alpha W\tilde{D}(\eta^o)\right] \\ &= -\mathrm{E}\left[W\left(Y - \gamma'_{YW}W\right)\right] + 2\mathrm{E}\left[\alpha W\left(D - \gamma'_{DW}W\right)\right] = 0\end{aligned}$$

Therefore we proved $\partial_\eta \alpha(\eta^o) = 0$

1.4 (Chap9) Predictive Inference via Modern Nonlinear Regression

This chapter introduced several modern nonlinear regression methods like **Regression Trees, Random Forests, Boosted Trees, Neural Nets/Deep Neural Networks and Ensemble Learning**. With a special emphasis on their prediction quality.

Assumption 1.1 (Structured Sparsity of Regression Function)**.** *We assume that g is generated as a composition of $q+1$ vector-valued functions:*

$$g = f_q \circ \ldots \circ f_0$$

where the i -th function f_i

$$f_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i+1}},$$

has each of its d_{i+1} components β_i -smooth and depends only on t_i variables, where t_i can be much smaller than d_i .

Assumption 1.2 (Nonparametric Sparsity of a Regression Function with Binary Regressors)**.** *We assume that there exists a subset S of size $|S| = r$, such that the function g can be written as a function of only the variables in S ; i.e. we can write*

$$g(Z) = f(Z_S)$$

where Z_S is the subvector of Z containing only the coordinates in S .

Under these assumptions, we can have learning guarantees for DNN(Under approxiamte sparsity), Shallow Regression Trees and Sub-Sampled Honest Forests.

$$\|\hat{g} - g\|_{L^2(Z)} = \sqrt{\mathrm{E}_Z\left[(\hat{g}(Z) - g(Z))^2\right]} \to 0, \quad \text{as } n \to \infty$$

Note: (The Honest Trees) An honest training approach is as follows: When we train a tree on a sub-sample, we randomly partition the data in half and we use half of the data to find the best splits in a greedy manner, and the other half of the data to construct the estimates at each node of the tree.

Ensemble Learning: It is an aggregated prediction is a linear combination of the basic predictors.

$$\tilde{g}(Z) = \sum_{k=1}^{K} \tilde{\alpha}_k \hat{g}_k(Z)$$

where $\hat{g}_k$ denote basic predictors that computed on the training data.

We can then figure out the coefficients of the optimal linear combination of the rules using test data V by minimizing the sum of prediction errors.

$$\min_{(\alpha_k)_{k=1}^K} \sum_{i \in V} \left(Y_i - \sum_{k=1}^{K} \alpha_k \hat{g}_k \left(Z_i \right) \right)^2$$

If K is large, we can instead use Lasso for aggregation:

$$\min_{(\alpha_k)_{k=1}^K} \sum_{i \in V} \left(Y_i - \sum_{k=1}^{K} \alpha_k \hat{g}_k \left(Z_i \right) \right)^2 + \lambda \sum_{k=1}^{K} |\alpha_k|$$

1.5 (Chap10) Statistical Inference on Predictive and Causal Effects in Modern Nonlinear Regression Models

DML Inference in the Partially Linear Regression Model(PLM)

$$Y = \beta D + g(X) + \epsilon, \quad \mathrm{E}[\epsilon \mid D, X] = 0$$

$$\tilde{Y} = \beta \tilde{D} + \epsilon, \quad \mathrm{E}[\epsilon \tilde{D}] = 0,$$

Procedure :

1. Partition data indices into random folds of approximately equal size: $\{1, \ldots, n\} = \cup_{k=1}^{K} I_k$. For each fold $k = 1, \ldots, K$, compute ML estimators $\hat{\ell}_{[k]}$ and $\hat{m}_{[k]}$ of the conditional expectation functions ℓ and m, leaving out the k -th block of data. Obtain the cross-fitted residuals for each $i \in I_k$:

$$\check{Y}_i = Y_i - \hat{\ell}_{[k]} \left(X_i \right), \quad \check{D}_i = D_i - \hat{m}_{[k]} \left(X_i \right).$$

2. Apply ordinary least squares of $\check{Y}_i$ on $\check{D}_i$. That is, obtain $\hat{\beta}$ as the root in b of the normal equations:

$$\mathbb{E}_n[(\check{Y} - b\check{D})\check{D}] = 0$$

Theorem 1.5 (Adaptive Inference on a Target Parameter in PLM). *Consider the PLM model. Suppose that estimators $\hat{\ell}_{[k]}(X)$ and $\hat{m}_{[k]}(X)$ provide approximations to the best predictors $\ell(X)$ and $m(X)$ that are of sufficiently high-quality:*

$$n^{1/4}\left(\left\|\hat{\ell}_{[k]}-\ell\right\|_{L^2}+\left\|\hat{m}_{[k]}-m\right\|_{L^2}\right)\approx 0.$$

Suppose that $\mathrm{E}\left[\tilde{D}^2\right]$ *is bounded away from zero; that is, suppose $\tilde{D}$ has non-trival variation left after partialling out. Suppose other regularity conditions listed in [2] hold. Then the estimation error in $\check{D}_i$ and $\check{Y}_i$ has no first order effect on $\hat{\beta}$:*

$$\sqrt{n}(\hat{\beta}-\beta)\approx\left(\mathbb{E}_n\left[\tilde{D}^2\right]\right)^{-1}\sqrt{n}\mathbb{E}_n[\tilde{D}\epsilon].$$

Consequently, $\hat{\beta}$ concentrates in a $1/\sqrt{n}$ neighborhood of β with deviations approximated by the Gaussian law:

$$\sqrt{n}(\hat{\beta}-\beta)\overset{\mathrm{a}}{\sim}N(0,\ \mathrm{V}),$$

DML Inference in the Interactive Regression Model (IRM)

$$\begin{aligned} Y &= g_0(D,X)+\epsilon, \quad \mathrm{E}[\epsilon \mid X,D]=0 \\ D &= m_0(X)+\tilde{D}, \quad \mathrm{E}[\tilde{D} \mid X]=0 \end{aligned}$$

APE from the IRM: Under conditional exogeneity, the APE coincides with the average treatment effect (ATE) of the intervention that move D= 0 to D= 1.

$$\theta_0=\mathrm{E}\left[g_0(1,X)-g_0(0,X)\right]$$

ATE from the IRM: Our construction of the efficient estimator for ATE will be based upon the relation.

$$\theta_0=\mathrm{E}\varphi_0(W)$$

where

$$\varphi_0(W)=g_0(1,X)-g_0(0,X)+(Y-g_0(D,X))\,H_0$$

$$H_0=\frac{1(D=1)}{m_0(X)}-\frac{1(D=0)}{1-m_0(X)}$$

Note: This estimator is doubly robust and is constructed by the combination of regression adjusted representation $\theta_0=E[g_0(1,X)-g_0(0,X)]$ and propensity score reweighting representation $\theta_0=E[YH_0]$. While neither of these representation is Neyman Orthogonal, the above estimator that constructed using the combination of them is Neyman Orthogonal

Procedure:

1. Partition sample indices into random folds of approximately equal size: $\{1,\ldots,n\} = \cup_{k=1}^{K} I_k$. For each $k = 1,\ldots,K$, compute estimators $\hat{g}_{[k]}$ and $\hat{m}_{[k]}$ of the conditional expectation functions g_0 and m_0, leaving out the k -th block of data, such that $\epsilon \leq \hat{m}_{[k]} \leq 1-\epsilon$, and for each $i \in I_k$ compute

$$\hat{\varphi}\left(W_i\right) = \hat{g}_{[k]}\left(1, X_i\right) - \hat{g}_{[k]}\left(0, X_i\right) + \left(Y_i - \hat{g}_{[k]}\left(D_i, X_i\right)\right)\hat{H}_i$$

with

$$\hat{H}_i = \frac{1\left(D_i = 1\right)}{\hat{m}_{[k]}\left(X_i\right)} - \frac{1\left(D_i = 0\right)}{1 - \hat{m}_{[k]}\left(X_i\right)}.$$

2. Compute the estimator

$$\hat{\theta} = \mathbb{E}_n[\hat{\varphi}(W)]$$

3. Construct standard errors via

$$\sqrt{\hat{V}/n}, \quad \hat{V} = \mathbb{E}_n[\hat{\varphi}(W) - \hat{\theta}]^2$$

and use standard normal critical values for inference.

Theorem 1.6 (Adaptive Inference on ATE with DML)**.** *Suppose conditions specified in [2] hold. In particular, suppose that the overlap condition holds, namely for some $\epsilon > 0$ with probability 1*

$$\epsilon < m_0(X) < 1 - \epsilon.$$

If estimators $\hat{g}_{[k]}(D, X)$ and $\hat{m}_{[k]}(X)$ are such that $\epsilon \leq \hat{m}_{[k]}(X) \leq 1-\epsilon$ and provide sufficiently high-quality approximations to the best predictors $g_0(D, X)$ and $m_0(X)$ such that

$$\left\|\hat{g}_{[k]} - g_0\right\|_{L^2} + \left\|\hat{m}_{[k]} - m_0\right\|_{L^2} + \sqrt{n}\left\|\hat{g}_{[k]} - g_0\right\|_{L^2}\left\|\hat{m}_{[k]} - m_0\right\|_{L^2} \approx 0,$$

then the estimation error in these nuisance parameter has no first order effect on $\hat{\theta}$:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \approx \sqrt{n}\mathbb{E}_n\left(\varphi_0(W) - \theta_0\right).$$

Consequently, the estimator concentrates in $1/\sqrt{n}$ neigborhood of θ_0 , with deviations controlled by the Gaussian law:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \overset{a}{\sim} N(0,\ \mathrm{V})$$

where

$$\mathrm{V} = \mathrm{E}\left(\varphi_0(W) - \theta_0\right)^2$$

Note: There is a trade-off between the estimator of propensity score m_0 and regression function g_0

DML inference for GATES:

$$\theta_0 = \mathrm{E}\left[g_0(1,X) - g_0(0,X) \mid G = 1\right]$$
$$\theta_0 = \mathrm{E}\left[\varphi_0(X) \mid G = 1\right] = \mathrm{E}\left[\varphi_0(X)G\right]/\mathrm{P}(G = 1)$$

DML inference for ATETS:

$$\theta_0 = \mathrm{E}\left[g_0(1,X) - g_0(0,X) \mid D = 1\right]$$

Generic Debiased (or Double) Machine Learning

A general construction upon which DML estimation and inference can be built relies on a method-of-moments estimator for some low-dimensional target parameter θ_0 based upon the empirical analog of the moment condition.

$$\mathrm{E}\psi\left(W;\theta_0,\eta_0\right) = 0$$

The first key input of the generic DML procedure is using a score function $\psi(W;\theta,\eta)$ such that

$$\mathrm{M}(\theta,\eta) = \mathrm{E}[\psi(W;\theta,\eta)]$$

identifies θ_0 when $\eta = \eta_0$ - that is,

$$\mathrm{M}\left(\theta,\eta_0\right) = 0 \text{ if and only if } \theta = \theta_0-$$

and the Neyman orthogonality condition is satisfied:

$$\partial_\eta \mathrm{M}\left(\theta_0,\eta\right)|_{\eta=\eta_0} = 0$$

score functions:

Scores for Partially Linear Regression Model:

$$\psi(W;\theta,\eta) := \{Y - \ell(X) - \theta(D - m(X))\}(D - m(X)),$$
$$\ell_0(X) = \mathrm{E}[Y \mid X], \quad m_0(X) = \mathrm{E}[D \mid X]$$

Scores for Interactive Regression Model:

$$\psi_1(W;\theta,\eta) := (g(1,X) - g(0,X)) + H(D,X)(Y - g(D,X)) - \theta$$
$$H(D,X) := \frac{D}{m(X)} - \frac{(1-D)}{1-m(X)}$$
$$g_0(D,X) = \mathrm{E}[Y \mid D,X], \quad m_0(X) = \mathrm{P}[D = 1 \mid X]$$

Scores for estimating GATEs:

$$\psi(W;\theta,\eta) := \frac{G}{p}\psi_1(W;\theta,\eta)$$
$$p_0 = \mathrm{P}(G = 1)$$

Scores for estimating ATETs:

$$\psi(W;\theta,\eta) := H(D,X)\frac{m(X)}{p}(Y - g(0,X)) - \frac{D\theta}{p}$$

$$p_0 = \mathrm{P}(D=1)$$

Note: The score function equals zero under the true value of the target parameter θ_0

The second key input is the use of high-quality machine learning estimators of the nuisance parameters. A sufficient condition in the examples given includes the requirement

$$n^{1/4}\left\|\hat{\eta} - \eta_0\right\|_{L^2} \approx 0$$

Procedure:

1. Inputs: Provide the data frame $(W_i)_{i=1}^n$, the Neymanorthogonal score/moment function $\psi(W,\theta,\eta)$ that identifies the statistical parameter of interest, and the name and model for ML estimation method(s) for η .

2. Train ML Predictors on Folds: Take a K-fold random partition $(I_k)_{k=1}^K$ of observation indices $\{1,\ldots,n\}$ such that the size of each fold is about the same. For each $k \in \{1,\ldots,K\}$, construct a high-quality machine learning estimator $\hat{\eta}_{[k]}$ that depends only on a subset of data $(X_i)_{i\notin I_k}$ that excludes the k -th fold. 3. Estimate Moments: Letting $k(i) = \{k : i \in I_k\}$, construct the moment equation estimate

$$\hat{\mathrm{M}}(\theta,\hat{\eta}) = \frac{1}{n}\sum_{i=1}^n \psi\left(W_i;\theta,\hat{\eta}_{[k(i)]}\right)$$

4. Compute the Estimator: Set the estimator $\hat{\theta}$ as the solution to the equation.

$$\hat{M}(\hat{\theta},\hat{\eta}) = 0.$$

5. Estimate Its Variance: Estimate the asymptotic variance of $\hat{\theta}$ by

$$\begin{aligned}\hat{V} =& \frac{1}{n}\sum_{i=1}^n \left[\hat{\varphi}\left(W_i\right)\hat{\varphi}\left(W_i\right)'\right] \\ & - \frac{1}{n}\sum_{i=1}^n \left[\hat{\varphi}\left(W_i\right)\right]\frac{1}{n}\sum_{i=1}^n \left[\hat{\varphi}\left(W_i\right)\right]',\end{aligned}$$

where

$$\hat{\varphi}\left(W_i\right) = -\hat{J}_0^{-1}\psi\left(W_i;\hat{\theta},\hat{\eta}_{[k(i)]}\right)$$

and

$$\hat{J}_0 := \partial_\theta \frac{1}{n}\sum_{i=1}^n \psi\left(W_i;\hat{\theta},\hat{\eta}_{[k(i)]}\right).$$

6. Confidence Intervals: Form an approximate $(1-\alpha)$%confidence interval for any functional $\ell'\theta_0$, where ℓ is a vector of constants, as

$$\left[\ell'\hat{\theta} \pm c\sqrt{\ell'\hat{V}\ell/n}\right]$$

where c is the $(1-\alpha/2)$ quantile of N(0,1) .

Note: The general DML process can be understood using the PLM as an example.

The following theorems show the properties of the general DML estimator

Definition 1.3 (Strong Identification). *We have that* $\mathrm{M}(\theta, \eta_0) = 0$ *if and and only if* $\theta = \theta_0$ *, and that*

$$J_0 := \partial_\theta \mathrm{E}\left[\psi\left(W; \theta_0, \eta_0\right)\right]$$

has singular values that is bounded away from zero.

Theorem 1.7 (Generic Adaptive Inference with DML). *Assume that estimates of nuisance parameters are of sufficiently high-quality. Assume strong identification holds.*

Then, estimation of nuisance parameter does not affect the behavior of the estimator to the first order; namely,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \approx \sqrt{n}\mathbb{E}_n\left[\varphi_0(W)\right],$$

where

$$\varphi_0(W) = -J_0^{-1}\psi\left(W; \theta_0, \eta_0\right), \quad J_0 := \partial_\theta \mathrm{E}\left[\psi\left(W; \theta_0, \eta_0\right)\right],$$

and $J_0 = \mathrm{E}\left[\psi^a\left(W; \eta_0\right)\right]$ *for linear scores. Consequently,* $\hat{\theta}$ *concentrates in a* $1/\sqrt{n}$ *-neighborhood of* θ_0 *and the sampling error* $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ *is approximately normal:*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \overset{\mathrm{a}}{\sim} N(0,\ \mathrm{V}), \quad \mathrm{V} := \mathrm{E}\left[\varphi_0(W)\varphi_0(W)'\right].$$

Theorem 1.8. *Under the same regularity conditions, the interval* $\left[\ell'\hat{\theta} \pm c\sqrt{\ell'\hat{\mathrm{V}}\ell/n}\right]$ *where* c *is the* $(1-\alpha/2)$ *quantile of a N(0,1) contains* $\ell'\theta_0$ *for approximately* $(1-\alpha)\times$ *100 percent of data realizations:*

$$\mathrm{P}\left(\ell'\theta_0 \in \left[\ell'\hat{\theta} \pm c\sqrt{\ell'\hat{\mathrm{V}}\ell/n}\right]\right) \approx (1-\alpha).$$

Selection of the Best ML Methods for DML is used to Minimize Upper Bounds on Bias.

2 Causal Inference

2.1 (Chap2) Causal Inference via Randomized Experiments

Potential Outcome Framework

Assumption 2.1 (Consistency). *we observe* $Y := Y(D)$

$$\mathrm{E}[Y \mid D = d] = \mathrm{E}[Y(d) \mid D = d], \text{ for } d \in \{0, 1\}.$$

Note: It requires that the treatment and control states are well-defined and clearly aligned with the observed treatment status, D.

Assumption 2.2 (Stable Unit-Treatment Value Assumption (SUTVA)). *Potential outcomes for any observational unit depend only on the treatment status of that unit and not on the treatment unit of any other unit.*

Assumption 2.3 (Random Assignment/Exogeneity). *suppose that treatment status is randomly assigned. Namely D is statistically independent of each potential outcome Y(d) for* $d \in \{0, 1\}$, *and* $0 < \mathrm{P}(D = 1) < 1$.

$$D \perp Y(d)$$

Note: This assumption is important in control for selection bias. It shows that D is uninformative about the potential outcome.

Selection bias: difference between ATE (calsal effect) and APE (predictive effect)

$$\text{APE} : \pi = \mathrm{E}[Y \mid D = 1] - \mathrm{E}[Y \mid D = 0]$$

$$\text{ATE} : \delta = \mathrm{E}[Y(1) - Y(0)] = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$$

If, for example, the treatment assignment D is associated the potential outcome Y, then it is likely that the observed APE is a biased estimator of the causal effect ATE.

Theorem 2.1 (Randomization Removes Selection Bias). *Under Random Assignment/Exogeneity Assumption , the average outcome in treatment group d recovers the average potential outcome under the treatment status d :*

$$\mathrm{E}[Y \mid D = d] = \mathrm{E}[Y(d) \mid D = d] = \mathrm{E}[Y(d)],$$

for each $d \in \{0, 1\}$. *Hence the average predictive effect and average treatment effect coincide:*

$$\begin{aligned} \pi &:= \mathrm{E}[Y \mid D = 1] - \mathrm{E}[Y \mid D = 0] \\ &= \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)] =: \delta. \end{aligned}$$

Pre-treatment covariates and heterogeneity

Assumption 2.4 (Random Assignment idenpendent of Covariates)**.** *Suppose that treatment status is randomly assigned. Namely, D is statistically independent of both the potential outcomes and a set of pre-determined covariates.* $0 < \mathrm{P}(D = 1) < 1$

$$D \perp (Y(0), Y(1), W)$$

Theorem 2.2 (Randomization with Covariates)**.** *Under Random Assignment idenpendent of Covariates , the expected value of Y conditional on treatment status D=d and covariates W coincides with the expected value of potential outcome Y(d) conditional on covariates W :*

$$\mathrm{E}[Y \mid D = d, W] = \mathrm{E}[Y(d) \mid D = d, W] = \mathrm{E}[Y(d) \mid W],$$

for each d . Hence the conditional predictive and average treatment effects agree:

$$\pi(W) = \delta(W)$$

Note: This assumption spells out that, if we plan to use covariates in the analysis, randomization has to be made with respect to these covariates as well.

Testing covariate balance: Testing Covariance Balance. The random assignment assumption induces covariate balance. Namely, the distribution of covariates should be the same under both treatment and control:

$$W|D = 1 \sim W|D = 0,$$

and, equivalently,

$$D \mid W \sim D$$

.

A useful implication is that D is not predictable by W :

$$\mathrm{E}[D \mid W] = \mathrm{E}[D].$$

This latter conditions is testable using regression tools. It amounts to saying that the R^2 of a regression of D on W is 0.

Classical Additive Approach: Improving Precision Under Linearity

The projection coefficient α of the folloing function recovers the ATE:

$$Y = D\alpha + \beta' X + \epsilon, \quad \epsilon \perp (D, X)$$

Illustrate this in linearity assumption:

$$\mathrm{E}[Y \mid D, W] = D\alpha + \beta' X$$

We assume that covariates are centered, and there is covariate balance:

$$\mathrm{E}[W] = 0, \quad \mathrm{E}[W \mid D = 1] = \mathrm{E}[W \mid D = 0]$$

Using centered covariates implies that:

$$\mathrm{E}[Y(0)] = \mathrm{E}[\mathrm{E}[Y \mid D = 0, X]] = \beta_1$$

$$\mathrm{E}[Y(1)] = \mathrm{E}[\mathrm{E}[Y \mid D = 1, X]] = \beta_1 + \alpha.$$

$$\delta = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)] = \alpha$$

Statistical inference on the ATE:

$$\begin{pmatrix} \sqrt{n}(\hat{\alpha} - \alpha) \\ \sqrt{n}\left(\hat{\beta}_1 - \beta_1\right) \end{pmatrix} \overset{\text{a}}{\sim} N(0,\ \mathrm{V}),$$

where covariance matrix V has components:

$$\vee_{11} = \frac{E\left[\epsilon^2 \tilde{D}^2\right]}{\left(E\left[\tilde{D}^2\right]\right)^2}, \quad V_{22} = \frac{E\left[\epsilon^2 \tilde{1}^2\right]}{\left(E\left[\tilde{1}^2\right]\right)^2}, \quad V_{12} = V_{21} = \frac{E\left[\epsilon^2 \tilde{D}\tilde{1}\right]}{E\left[\tilde{1}^2\right] E\left[\tilde{D}^2\right]}$$

We consider what happens when we do not include covariates in the regression. In this case, the OLS estimator $\bar{\alpha}$ estimates the projection coefficient α in the BLP using (1, D) alone

$$Y = \alpha D + \beta_1 + U, \quad \mathrm{E}[U] = \mathrm{E}[UD] = 0$$

where the noise

$$U = \beta'(X - E[X]) + \epsilon$$

contains the part of Y that is linearly predicted by X, $\beta'(X - \mathrm{E}[X]) = \beta' X - \beta_1$. We then have that $\bar{\alpha}$ obeys

$$\sqrt{n}(\bar{\alpha} - \alpha) \overset{\text{a}}{\sim} N\left(0, \overline{\mathrm{V}}_{11}\right), \quad \overline{\mathrm{V}}_{11} = \frac{\mathrm{E}\left[U^2 \tilde{D}^2\right]}{\left(\mathrm{E}\left[\tilde{D}^2\right]\right)^2}.$$

Under linear assumption, it follows that $V_{11} \leq \bar{V}_{11}$, with the inequality being strict if $Var(\beta\prime X) > 0$

Note: Under linear assumption, Using pre-determined covariates improves the precision of estimating the ATE

The Interactive Approach: Always Improves Precision and Discovers Heterogeneity

Letting X=(1, W) be an intercept and the pre-treatment covariates W , let us write the BLP of each of Y(0) and Y(1) using X as

$$Y(d) = \beta_d' X + \varepsilon_d, \quad \varepsilon_d \perp X, \quad d = 0, 1.$$

Under linear Assumption, it coincides with the BLP of Y using X in the D=d population. Letting $\varepsilon = D\varepsilon_1 + (1 - D)\varepsilon_0$, we thus have

$$Y = \beta_d' X + \varepsilon, \quad \mathrm{E}[\varepsilon X \mid D = d] = 0, \quad d = 0, 1.$$

$$Y = \beta_0' X + \beta_\delta' XD + \varepsilon, \quad \varepsilon \perp (X, DX)$$

where $\beta_\delta = \beta_1 - \beta_0$

We assume that covariates are centered:

$$\mathrm{E}[W] = 0$$

.

Since X contains an intercept, $\varepsilon_d \perp X$ implies $\mathrm{E}[\varepsilon_d] = 0$. Together with centered covariates, we find that

$$\mathrm{E}[Y(d)] = \mathrm{E}[\beta_d' X + \varepsilon_d] = \beta_{d,1}.$$

This means that the ATE coincides with the coefficient on D in the BLP of Y using (X, DX) . That is, $\beta_{\delta,1} = \delta$.

If we use OLS to estimate the BLP of Y using (X, D X) , then an application of the OLS theory in the previous chapter gives us that, under regularity conditions,

$$\begin{pmatrix} \sqrt{n}\left(\hat{\beta}_{\delta,1} - \delta\right) \\ \sqrt{n}\left(\hat{\beta}_{0,1} - \mathrm{E}[Y(0)]\right) \end{pmatrix} \overset{\mathrm{a}}{\sim} N(0,\ \mathrm{V}),$$

where covariance matrix V has components:

$$V_{11} = \frac{E\left[\epsilon^2 \tilde{D}^2\right]}{\left(E\left[\tilde{D}^2\right]\right)^2}, \quad V_{22} = \frac{E\left[\epsilon^2 \tilde{1}^2\right]}{\left(E\left[\tilde{1}^2\right]\right)^2}, \quad V_{12} = V_{21} = \frac{E\left[\epsilon^2 \tilde{D}\tilde{1}\right]}{E\left[\tilde{1}^2\right] E\left[\tilde{D}^2\right]},$$

where $\tilde{D} = D - E[D]$ is the residual after partialling out linearly $(1, W, DW)$ from D and $\tilde{1} := (1 - D)$ is the residual after partialling out (D, W, DW) from 1

Recall that when we use (1,D) to estimate the ATE, the estimator obeys:

$$\sqrt{n}(\hat{\delta} - \delta) \overset{\mathrm{a}}{\sim} N\left(0, \overline{\mathrm{V}}_{11}\right), \overline{\mathrm{V}}_{11} = \frac{\mathrm{E}\left[U^2 \tilde{D}^2\right]}{\left(\mathrm{E}\left[\tilde{D}^2\right]\right)^2}$$

Since ϵ satisfies the BLP conditions for each of the treatment populations, i.e. $\mathrm{E}[\varepsilon W \mid D = d] = 0$, it then follows that

$$\mathrm{V}_{11} \leq \overline{\mathrm{V}}_{11}.$$

Moreover, the inequality is strict if $\mathrm{Var}\left(\beta'_{0,2}W\right) > 0$ or $\mathrm{Var}\left(\beta'_{1,2}W\right) > 0$

Note: Pre-determined covariates improve the precision of estimating the ATE δ, when using the interactive model, without any linearity assumptions on the CEF.

2.2 (Chap5) Causal Inference via Conditional Ignorability

Here we discuss how average causal effects may be identified using regression when treatment is not randomly assigned but instead depends on observed covariates.

Assumption 2.5 (Conditional Ignorability and Consistency)**.** *Ignorability: Suppose that treatment status D is independent of potential outcomes Y(d) conditional on a set of covariates X: For each d,*

$$D \perp Y(d)|X$$

Consistency: Suppose that Y is generated as Y := Y(D)

Assumption 2.6 (Overlap/Full Support)**.** *The probability of receiving treatment given X, the propensity score*

$$p(x) := P(D = 1|X)$$

is non-degenerate:

$$P(0 < p(x) < 1) = 1$$

Note: Without this condition, there are values x in the support of X where we cannot construct a contrast between treatment and control units. We cannot learn the conditional average treatment effect at these values of X and thus are also unable to learn the unconditional average effect of the treatment.

Theorem 2.3 (Conditioning on X removes selection bias)**.** *Under Conditional Ignorability and Overlap, the conditional expectation function of observed outcome Y given D=d and X recovers the conditional expectation of the potential outcome Y(d) given X :*

$$\mathrm{E}[Y \mid D = d, X] = \mathrm{E}[Y(d) \mid D = d, X] = \mathrm{E}[Y(d) \mid X].$$

Note: note that the overlap assumption makes it possible to condition on the events D = 0, X and D = 1, Xat any value in the support of X and that the second equality holds by ignorability.

Identification by conditioning

Hence, the Conditional Average Predictive Effect (CAPE),

$$\pi(X) = \mathrm{E}[Y \mid D = 1, X] - \mathrm{E}[Y \mid D = 0, X],$$

is equal to the Conditional Average Treatment Effect (CATE),

$$\delta(X) = \mathrm{E}[Y(1) \mid X] - \mathrm{E}[Y(0) \mid X].$$

Thus, the APE and ATE also agree:

$$\delta = \mathrm{E}[\delta(X)] = \mathrm{E}[\pi(X)] = \pi.$$

Under conditioning assumptions, we next illustrate how we con adopt linear regression to retrieve causal estimates. A simple instance is under the linear assumption:

$$\mathrm{E}[Y \mid D, X] = \alpha D + \beta' W,$$

which gives a model

$$Y = \alpha D + \beta' W + \epsilon, \quad E[\epsilon \mid D, X] = 0.$$

Here it is understood that W may include X as well as prespecified nonlinear transformations of X .

In this model, α identifies δ

$$\delta = \alpha$$

Note: The linearity assumption and ignorability assumptions imply that treatment effects are homogeneous; that is, $\delta(x) = \delta$ for all x in the support of X.

We can relax the linear assumption by considering the interactions:

$$\mathrm{E}[Y \mid D, X] = \alpha_1 D + \alpha_2' W D + \beta_1 + \beta_2' W$$

where we also maintain that we are working with centered covariates: EW = 0

We then recover the ATE as

$$\delta = \alpha_1$$

and CATE as

$$\delta(X) = \alpha_1 + \alpha_2' W.$$

Identification by Propensity Scores

The identification by conditioning approach requires being able to accurately model the "outcome process", When the outcome process is hard to model, we might have a much better handle on the "treatment selection process," i.e. the propensity score. An alternative approach, known as the Horvitz-Thompson method , uses propensity score reweighting to recover averages of potential outcomes.

Theorem 2.4 (Horvitz-Thompson: Propensity Score Reweight ing Removes Bias)**.** *Under Conditional Ignorability and Overlap, the conditional expectation of an appropriately reweighted observed outcome Y , given X , identifies the conditional average of potential outcome $Y(d)$ given X :*

$$\mathrm{E}\left[Y\frac{1(D=d)}{\mathrm{P}(D=d\mid X)}\middle|\ X\right]=\mathrm{E}[Y(d)\mid X]$$

Then, averaging over X identifies the average potential outcome:

$$\mathrm{E}\left[Y\frac{1(D=d)}{\mathrm{P}(D=d\mid X)}\right]=\mathrm{E}[Y(d)]$$

Proof:

$$\mathrm{E}\left[Y\frac{1(D=d)}{\mathrm{P}(D=d\mid X)}\middle|\ X\right]=\frac{\mathrm{E}[Y1(D=d)\mid X]}{\mathrm{P}(D=d\mid X)}=\mathrm{E}[Y(d)\mid X]\frac{\mathrm{E}[1(D=d)\mid X]}{\mathrm{P}(D=d\mid X)}=\mathrm{E}[Y(d)\mid X],$$

where we used conditional ignorability in the second equality.

As a consequence, we can identify average treatment effects by simple averaging of transformed outcomes:

$$\delta=\mathrm{E}[YH],\quad \delta(X)=\mathrm{E}[YH\mid X].$$

$$H=\frac{1(D=1)}{\mathrm{P}(D=1\mid X)}-\frac{1(D=0)}{\mathrm{P}(D=0\mid X)},$$

where H is called the Horvitz-Thompson transform.

Remark: Propensity score reweighting is generally not the most efficient approach to estimating treatment effects from a statistical point of view because it ignores any dependence between the outcomes and controls,X, that is not captured by the propensity score. Moreover, estimation based on only propensity score reweighting fails under imbalances that might arise due to imperfect data collection. Later, we will use both regression and reweighting as part of "double machine learning" to operationalize efficient statistical inference on treatment effects in fully nonlinear (nonparametric) models.

We can perform covariate balance check to check if the statified RCT/ Conditional ignorability is valid. Specifically, conditional ignorability implies that $E[H|X]=0$. Thus if covairates predict H, we can conclude that Conditional ignorability does not hold.

group ATE (GATE)

$$\delta_G=\mathrm{E}[Y(1)-Y(0)\mid G=1]$$

$\mathrm{E}[Y(1)-Y(0)\mid G=1]=\mathrm{E}[\mathrm{E}[Y\mid D=1,X]-\mathrm{E}[Y\mid D=0,X]\mid G=1]=\mathrm{E}[HY\mid G=1].$

we can identify GATEs either by taking the difference in regression functions or applying propensity score reweighting of outcomes and then averaging over group G.

average treatment effect on the treated (ATET)

$$\delta_1 = \mathrm{E}[Y(1) - Y(0) \mid D = 1]$$
$$\mathrm{E}[\mathrm{E}[Y \mid D = 1, X] - \mathrm{E}[Y \mid D = 0, X] \mid D = 1]$$

Assumption 2.7 (Ignorability and Overlap for Treated)**.** *(a) Ignorability. Suppose that the treatment status D is independent of Y(0) conditional on a set of covariates X , that is*

$$D \perp Y(0) \mid X.$$

(b) Weak Overlap. Suppose that the propensity score satisfies:

$$\mathrm{P}(p(X) < 1) = 1$$

Theorem 2.5 (Identification of ATET)**.** *Under ignorability and overlap assumption for treated,*

$$\delta_1 = \mathrm{E}[Y \mid D = 1] - \mathrm{E}[\mathrm{E}[Y \mid X, D = 0] \mid D = 1]$$

proof:

$$\begin{aligned}\mathrm{E}[Y(0) \mid D = 1] &= \mathrm{E}[\mathrm{E}[Y(0) \mid D = 1, X] \mid D = 1] \\ &= \mathrm{E}[\mathrm{E}[Y(0) \mid D = 0, X] \mid D = 1] \\ &= \mathrm{E}[\mathrm{E}[Y \mid D = 0, X] \mid D = 1]\end{aligned}$$

Theorem 2.6 (Propensity Score Reweighting for the Treated)**.** *Under ignorability and overlap assumption for treated,*

$$\mathrm{E}[Y\bar{H}] = \delta_1, \quad \bar{H} = Hp(X)/\mathrm{E}[D].$$

proof:

$$\frac{\mathrm{E}[DY]}{\mathrm{E}[D]} = \frac{\mathrm{E}[DY(1)]}{\mathrm{E}[D]} = \mathrm{E}[Y(1) \mid D = 1]$$

$$\begin{aligned}
\frac{\mathrm{E}\left[\frac{(1-D)}{1-p(X)}p(X)Y\right]}{\mathrm{E}[D]} &= \frac{\mathrm{E}\left[\frac{p(X)}{1-p(X)}\mathrm{E}[(1-D)Y \mid X]\right]}{\mathrm{E}[D]} \\
&= \frac{\mathrm{E}\left[\frac{p(X)}{1-p(X)}\mathrm{E}[(1-D)Y(0) \mid X]\right]}{\mathrm{E}[D]} \\
&= \frac{\mathrm{E}\left[\frac{p(X)}{1-p(X)}\mathrm{E}[1-D \mid X]\mathrm{E}[Y(0) \mid X]\right]}{\mathrm{E}[D]} \\
&= \frac{\mathrm{E}[p(X)\mathrm{E}[Y(0) \mid X]]}{\mathrm{E}[D]} \\
&= \frac{\mathrm{E}[\mathrm{E}[D \mid X]\mathrm{E}[Y(0) \mid X]]}{\mathrm{E}[D]} \\
&= \frac{\mathrm{E}[\mathrm{E}[DY(0) \mid X]]}{\mathrm{E}[D]} \\
&= \frac{\mathrm{E}[DY(0)]}{\mathrm{E}[D]} = \mathrm{E}[Y(0) \mid D=1]
\end{aligned}$$

Clever Covariate Regression

estimating the ATE, then it suffices to learn the BLP of the outcome Y using the single covariate

$$\phi(D,X) := H = \frac{1(D=1)}{p(X)} - \frac{1(D=0)}{1-p(X)}$$

We can then use this BLP model as a proxy for the CEF $\mathrm{E}[Y \mid D, p(X)]$. Specifically, we learn a decomposition

$$Y = \beta\phi(D,X) + \epsilon, \epsilon \perp \phi(D,X)$$

by running OLS of Y on $\phi(D,X)$and then use

$$\mathrm{E}[\beta(\phi(1,X) - \phi(0,X))]$$

as the ATE.

proof: Note that the random variable H satisfies

$$\mathrm{E}[f(D,X)H \mid X] = f(1,X) - f(0,X)$$

for any function f(D, X). Then, by orthogonality of ϵ in the BLP decomposition:

$$\begin{aligned}
\mathrm{E}[Y(1) - Y(0)] &= \mathrm{E}[YH] = \mathrm{E}[\beta\phi(D,X)H] \\
&= \mathrm{E}[\beta(\phi(1,X) - \phi(0,X))].
\end{aligned}$$

Rosenbaum-Rubin's Result: Conditioning on the propensity score

Theorem 2.7 (Rosenbaum and Rubin: Conditioning on the Propensity Score Removes Selection Bias)**.** *Under Ignorability and Overlap, D is generated independently of Y(d) for each d , conditional on the propensity score p(X) : For each d ,*

$$D \perp Y(d) \mid p(X).$$

Note: In other words, conditional on p(X)=p, variation in D is as good as randomly assigned. We can identify the conditional average potential outcome as

$$\mathrm{E}[Y(d) \mid p(X)] = \mathrm{E}[Y \mid D = d, p(X)]$$

proof: First we state that the following equivalence relations hold:

$$D \perp X \mid p(X) \quad \Leftrightarrow \quad \mathrm{P}(D = 1 \mid X, p(X)) = \mathrm{P}(D = 1 \mid p(X)).$$

From the LHS to the RHS follows from:

$$\mathrm{P}(D = 1 \mid X, p(X)) = \mathrm{P}(D = 1 \mid X) = p(X)$$

and

$$\mathrm{P}(D = 1 \mid p(X)) = \mathrm{E}[D = 1 \mid p(X)] = \mathrm{E}[\mathrm{E}[D \mid X, p(X)] \mid p(X)] = \mathrm{E}[p(X) \mid p(X)] = p(X)$$

This property underlies covariate balance checks.

From the RHS to the LHS follows from:

$$\begin{aligned}
\mathrm{E}[g(Y(1)) \mid p(X)] =& \mathrm{E}[\mathrm{E}[g(Y(1)) \mid X, p(X)] \mid p(X)] \\
=& \mathrm{E}[\mathrm{E}[g(Y(1)) \mid X] \mid p(X)] \\
=& \mathrm{E}\left[g(Y) \frac{1(D=1)}{p(X)} \middle| \, p(X) \right] \\
=& \mathrm{E}\left[g(Y) \frac{1(D=1)}{p(X)} \middle| \, D = 1, p(X) \right] P(D = 1 \mid p(X)) \\
& + \mathrm{E}\left[g(Y) \frac{1(D=1)}{p(X)} \middle| \, D = 0, p(X) \right] P(D = 0 \mid p(X)) \\
=& \mathrm{E}[g(Y) \mid D = 1, p(X)] \frac{P(D = 1 \mid p(X))}{p(X)} \\
=& \mathrm{E}[g(Y) \mid D = 1, p(X)] \\
=& \mathrm{E}[g(Y(1)) \mid D = 1, p(X)]
\end{aligned}$$

where we use $P(D = 1 \mid p(X)) = p(X)$. We can similarly argue for the case of d=0 . Thus, the conditional distribution of Y(1) does not depend on D , once we condition on p(X) .

2.3 (Chap6) Causal Inference via Linear Structural Equations

triangular structural equation model (TSEM):

$$\begin{aligned}
Y &:= \delta P + X'\beta + \epsilon_Y \\
P &:= X'v + \epsilon_P \\
X &
\end{aligned}$$

where ϵ_Y, ϵ_P and X are mutually independent (or at least uncorrelated) and determined outside of the model

2.4 (Chap7) Causal Inference via Directed Acyclical Graphs and Nonlinear Structural Equation Models

Definition 2.1 (acyclic structural equation model (ASEM)). *The ASEM corresponding to the DAG G=(V, E) is the collection of random variables* $\{X_j\}_{j \in V}$ *such that*

$$X_j := f_j\left(Pa_j, \epsilon_j\right), \quad j \in V,$$

where the disturbances $(\epsilon_j)_{j \in V}$ *are jointly independent.*

Definition 2.2 (d-Separation). *Given a DAG G, a set of nodes S d-separates nodes X and Y if nodes in S block all paths between X and Y . d -separation is denoted as*

$$(Y \perp_d X \mid S)_{\mathrm{G}}.$$

Theorem 2.8 (Conditional Independence from d-Separation). *d -Separation implies conditional independence, Global Markov:*

$$(Y \perp_d X \mid S)_G \Longrightarrow Y \perp X \mid S$$

.

The reverse implication is not true in general, this is regard as unfaithfulness.

$$Y \perp X \mid S \Longrightarrow (Y \perp_d X \mid S)_{\mathrm{G}}$$

2.5 (Chap8) Valid Adjustment Sets from DAGs

Theorem 2.9 (A Complete Criterion for Identification by Conditioning). *Consider any ASEM with DAG G. Let us re-label a policy node* X_j *as D , and let Y , an outcome of interest, be any other descendant of D . Consider a SWIG DAG* $\tilde{\mathrm{G}}(d)$ *which is induced by the fix (D=d) intervention. Consider any other subset of nodes S that appears in both* G *and* $\tilde{\mathrm{G}}(d)$ *, such that Y(d) is d -separated from D by S in* $\tilde{\mathbf{G}}(d)$ *. - Then the following conditional exogeneity/ignorability holds:*

$$Y(d) \perp D \mid S.$$

Then

$$\mathrm{E}[Y(d) \mid S = s] = \mathrm{E}[Y \mid D = d, S = s]$$

holds for all s such that p(d, s)¿0 .

Adjustment strategies

1. Conditioning on one of all parents of Y (that are not descendants of D)

2. Conditioning using the backdoor criterion enables us to find all minimal adjustment sets

3. Conditioning on all common causes of Y and D is also sufficient.