Lasso in Causal Analysis

- Selection among many controls.
    - Double Selection
    - Partialing-out Lasso
    - Cross-fit partialing-out
- Selection among many instruments.

Selection among many controls

Consider a linear model where a treatment variable, d_i, is taken as exogenous after conditioning on control variables

$$y_i = \alpha d_i + \underbrace{x_i'\theta_y + r_{yi}}_{g(x_i)} + \varsigma_i,$$

where the parameter of interest is α, the effect of the treatment on the outcome, $E[\varsigma_i | d_i, x_i, r_{yi}] = 0$, r_{yi} is an approximation error. Further

$$d_i = \underbrace{x_i'\theta_d + r_{di}}_{m(x_i)} + v_i,$$

where $E[v_i | x_i, r_{di}] = 0$.

- Sparsity: Including only s non-zero coefficients make approximation errors small as n increases:

$$E\left[r_{yi}^2\right]^{\frac{1}{2}}, E\left[r_{di}^2\right]^{\frac{1}{2}} \leq c\sqrt{\frac{s}{n}}, \text{ for some } s.$$

Double selection

Double selection

1. Use a lasso of y on x to select covariates $\tilde{x}_y$ that predict y.
2. Use a lasso of d on x to select covariates $\tilde{x}_d$ that predict d.
3. Regress y on d and the union of the covariates in $\tilde{x}_y$ and $\tilde{x}_d$.to get estimate and standard error for α.

- Conditions for consistent estimate of α in Belloni et al. (2013).
- Using both selection steps also enhances efficiency by finding variables that are strongly predictive of the outcome and may remove residual variance.
- Additional variables can be added for robustness.
- Other regularization methods can be used to find included regressors (as long as these satisfy the sparsity condition).

Partialing-out (PO) Lasso

1. Use a lasso of y on x to select covariates $\widetilde{x}_y$ that predict y.
2. Regress y on $\widetilde{x}_y$, and let $\widetilde{y}$ be residuals from this regression.
3. Use a lasso of d on x to select covariates $\widetilde{x}_d$ that predict d.
4. Regress d on $\widetilde{x}_d$, and let $\widetilde{d}$ be residuals from this regression.
5. Regress $\widetilde{y}$ on $\widetilde{d}$ to get estimate and standard error for α.

Cross-fit partialing-out (XPO): Chernozhukov et al (2018)

1. Split sample into folds. Exclude one fold and
   1.1 Use a lasso of y on x to select covariates $\widetilde{x}_y$ that predict y.
   1.2 Regress y on $\widetilde{x}_y$, and let $\widetilde{\beta}^A$ be the estimated coefficients.
   1.3 Use a lasso of d on x to select covariates $\widetilde{x}_d$ that predict d.
   1.4 Regress d on $\widetilde{x}_d$, and let $\widetilde{\delta}^A$ be the estimated coefficients.
2. For the excluded fold:
   2.1 Fill in the residuals for $\widetilde{y} = y - \widetilde{x}_y \widetilde{\beta}^A$.
   2.2 Fill in the residuals for $\widetilde{d} = d - \widetilde{x}_d \widetilde{\delta}^A$
3. When the residuals are filled in for the whole sample, regress $\widetilde{y}$ on $\widetilde{d}$ to estimate α.

- The functions $g(x_i)$ and $m(x_i)$, can be nonlinearly estimated in steps 1a and 1c using, e.g. regression trees.
- Chernozhukov et al.."Double/debiased machine learning for treatment and structural parameters". Econometrics Journal, 2018.

Comparison

- XPO$\succ$DS$\succ$PO
- XPO
  - has better large- and finite-sample properties than DS and PO,
  - takes longer than PO and DS because of its fold-level computations.
- DS performs better than PO in small sample (Belloni et atl, 2016). Same asymptotic properties.

Tuning parameter λ

- Plug-in method
  - PO, DS, and XPO estimators have proven large-sample properties, as discussed by (Belloni et atl, 2016).
- Cross-validation
  - May not provide good performance when prediction is not the end goal. Use for robustness.

Selection among many instruments

Consider the standard IV setting

$$\begin{aligned} y_i &= \alpha d_i + \varepsilon_i \\ d_i &= z_i'\Pi + r_i + v_i, \end{aligned}$$

where

$$E[\varepsilon_i | z_i] = E[v_i | z_i, r_i] = 0$$

but

$$E[\varepsilon_i v_i] \neq 0.$$

Including a small number of exogenous variables is straightforward. Suppose that there are many valid instruments with varying strenght. Then the set of instruments in the first stage can be selected via e.g. Lasso to minimize test error.

- This works because
  - There is no selection over d_i, only over the first stage purely predictive problem.
  - Model selection among valid first stage instruments will not bias the second stage estimate of α.

Legalized Abortion and Crime (Donohue and Levitt, 2001)

- Differences-in-differences estimation for state-level crime rates 1985-1997.

$$y_{cit} = \alpha_c a_{cit} + w'_{it}\beta_c + \delta_{ci} + \gamma_{ct} + \varepsilon_{cit}$$

  - y_{cit} : crime-rate for crime type $c \in \{violent, property, murder\}$ in state i in year t
  - a_{it} : abortion rate relevant for type of crime c (as determined by the ages of criminals when they tend to commit crimes)
  - δ_{ci}, γ_{ct} : state and year-fixed effects
  - w_{it} : log of lagged prisoners per capita, the log of lagged police per capita, the unemployment rate, per-capita income, the poverty rate, the generosity of the Aid to Families with Dependent Children (AFDC) welfare program at time t - 15, a dummy for having a concealed weapons law, and beer consumption per capita.

- Paper presents baseline results based on this formulation as well as results from different models which vary the sample and set of controls in their tables IV and V.

- We will now check whether the results are robust to including nonlinear trends interacted with observed state-specific characteristics.
- Pre-selection of control variables z_{itc}:
  - 284 variables made up of
    - the levels, differences, initial level, initial difference, and within-state average of the eight state-specific time-varying observables, the initial level and initial difference of the abortion rate relevant for crime type c,
    - quadratics in each of the preceding variables,
    - interactions of all the aforementioned variables with t and t^2, and the main effects t and t^2 .
- Use Lasso to automatically select controls to include in regression.
  - Select variables that explain year-to-year changes in crime and abortion:

$$
\begin{aligned}
\Delta y_{cit} &= \alpha_c \Delta a_{cit} + z'_{cit}\beta_c + \widetilde{\gamma}_{ct} + \Delta \varepsilon_{cit} \\
\Delta a_{cit} &= z'_{cit}\Pi_c + \widetilde{\kappa}_{ct} + \Delta v_{cit},
\end{aligned}
$$

    where $\widetilde{\gamma}_{ct}$ and $\widetilde{\kappa}_{ct}$ are time-fixed effects.

- For violent crime, eight variables are selected in the abortion equation,and no variables are selected in the crime equation.
    - lagged prisoners per capita, the lagged unemployment rate, the initial change in beer consumption interacted with a linear trend, the initial change in income squared interacted with a linear trend, the within-state mean of income, the within-state mean of lagged prisoners per capita interacted with a linear trend, the within-state mean of income interacted with a linear trend, and the initial level of the abortion rate.
- For property crime, nine variables are selected in the abortion equation, and three are selected in the crime equation.
- For murder, nine variables are selected in the abortion equation, and none were selected in the crime equation.

Effect of Abortion on Crime

	Type of crime					
	Violent		Property		Murder	
Estimator	*Effect*	*Std. error*	*Effect*	*Std. error*	*Effect*	*Std. err*
First-difference	−.157	.034	−.106	.021	−.218	.068
All controls	.071	.284	−.161	.106	−1.327	.932
Double selection	−.171	.117	−.061	.057	−.189	.177

Notes: This table reports results from estimating the effect of abortion on violent crime, property crim and murder. The row labeled "First-difference" gives baseline first-difference estimates using the contro from Donohue and Levitt (2001). The row labeled "All controls" includes a broad set of controls mear to allow flexible trends that vary with state-level characteristics. The row labeled "Double selection reports results based on the double selection method outlined in this paper and selecting among th variables used in the "All controls" results.

- Results are not robust to the inclusion of fairly parsimonious nonlinear trends.
    - NB. Controls mostly in abortion equation. Maximizing prediction of abortion = maximizing multicollinearity problem.

Lasso in Stata 16

- Example: Donohue and Levitt, 2004.

```
. * Violence equation with selected controls ;
. reg Dyviol Dviol `vDS' `tdums' , cluster(statenum) ;

Linear regression                               Number of obs     =        600
                                                F(21, 49)         =      31.19
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2712
                                                Root MSE          =      .0713

                              (Std. Err. adjusted for 50 clusters in statenum)

             |               Robust
      Dyviol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Dviol |  -.1711086   .1169667    -1.46   0.150    -.40616
       viol0 |   .0802067   .1059545     0.76   0.453    -.13271
    Lxxprison|   .0098219    .008758     1.12   0.268     -.0077
    Lxxpolice|  -.0207962    .042971    -0.48   0.631    -.10714
    Mxxincome|   5.818115   6.795879     0.86   0.396    -7.8387
  Dxxincome0 |  -22.96238   39.82136    -0.58   0.567    -102.98
  LxxpoliceXt|  -.0475144   .0577602    -0.82   0.415    -.16358
  MxxincomeXt|  -6.052352   9.589547    -0.63   0.531    -25.323
Dxxincome0Xt |    21.1692   62.79434     0.34   0.737    -105.02
  Dxxbeer0Xt |    1.27701   .5919436     2.16   0.036     .08745
   _Iyear_87 |  -.0267729   .0831013    -0.32   0.749    -.19377
   _Iyear_88 |   .0016363   .1567621     0.58   0.562    -.22338
```

Effect of Abortion on Crime

	Violent	
Estimator	*Effect*	*Std. error*
First-difference	−.157	.034
All controls	.071	.284
Double selection	−.171	.117

```
* Violence Outcome ;
lassoShooting Dyviol `AllViol' , controls(`tdums') lasiter(100) verbose(0) fdisplay(0) ;
local yvSel `r(selected)' ;
di "`yvSel'" ;

* Violence Abortion ;
lassoShooting Dviol `AllViol' , controls(`tdums') lasiter(100) verbose(0) fdisplay(0) ;
local xvSel `r(selected)' ;
di "`xvSel'" ;

* Get union of selected instruments ;
local vDS : list yvSel | xvSel ;

* Violence equation with selected controls ;
reg Dyviol Dviol `vDS' `tdums' , cluster(statenum) ;
```

Lasso in Stata

```
* Double Selection
dsregress Dyviol Dviol, controls(($tdums) $AllViol) cluster(statenum);
```

```
Double-selection linear model         Number of obs               =        600
                                      Number of controls          =        323
                                      Number of selected controls =         19
                                      Wald chi2(1)                =       3.67
                                      Prob > chi2                 =     0.0553
```

Dyviol	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Dviol	-.1974046	.103006	-1.92	0.055	-.3992927	.0044835

```
. lassoinfo

    Estimate:  active
     Command:  dsregress
```

Variable	Model	Selection method	lambda	No. of selected variables
Dyviol	linear	plugin	.1816933	11
Dviol	linear	plugin	.1816933	19

```
. lassocoef ( ., for(Dyviol)) ( ., for(Dviol))
```

	Dyviol	Dviol
_Iyear_87	x	x
_Iyear_88	x	x
_Iyear_89	x	x
_Iyear_90	x	x
_Iyear_91	x	x
_Iyear_92	x	x
_Iyear_93	x	x
_Iyear_94	x	x
_Iyear_95	x	x
_Iyear_96	x	x
_Iyear_97	x	x
viol0		x
Lxxprison		x
Lxxunemp		x
Lxxpolice2		x
Mxxpolice		x
Mxxincome		x
Dxxincome0Xt		x
Dxxbeer0Xt		x
_cons	x	x

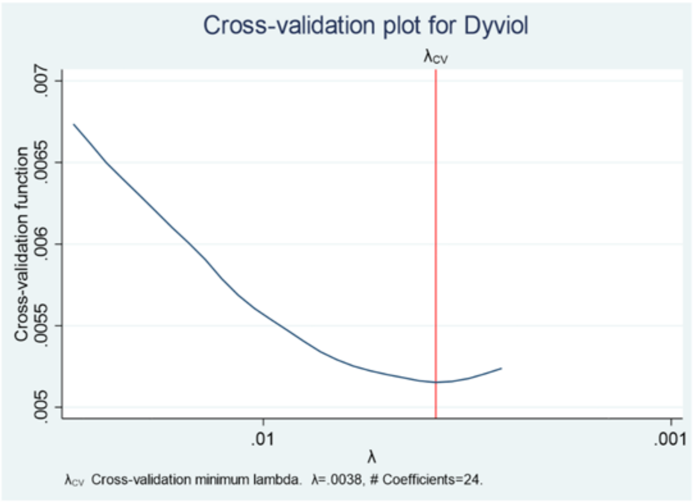
Select lambda:
plug-in, cv, adaptive for causal analysis

```
* Double Selection, lambda selected by cross validation
dsregress Dyviol Dviol, controls( $AllViol $tdums) selection(cv)  cluster(statenum);
```

```
Double-selection linear model         Number of obs               =        600
                                      Number of controls          =        323
                                      Number of selected controls =        136
                                      Wald chi2(1)                =       0.00
                                      Prob > chi2                 =     0.9951

------------------------------------------------------------------------------
             |               Robust
      Dyviol |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Dviol |   .0012863   .2113531     0.01   0.995    -.4129582    .4155308
------------------------------------------------------------------------------
```

```
* Hand select lambda
cvplot, for(Dyviol);
lassoknots , for(Dyviol);
lassoselect id = 18, for(Dyviol);
cvplot, for(Dyviol);
```



Cross-fit partialing-out (xpo)

```
* Cross-fit partialing-out
xporegress Dyviol Dviol, controls( $AllViol $tdums) cluster(statenum)
```

```
Cross-fit partialing-out          Number of obs                 =        600
linear model                      Number of controls            =        323
                                  Number of selected controls   =         22
                                  Number of folds in cross-fit  =         10
                                  Number of resamples           =          1
                                  Wald chi2(1)                  =       4.53
                                  Prob > chi2                   =     0.0333
```

Dyviol	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Dviol	-.2066832	.0971296	-2.13	0.033	-.3970538	-.0163127

```
. lassoinfo

    Estimate:  active
     Command:  xporegress
```

Variable	Model	Selection method	No. of selected variables min	median	max
Dviol	linear	plugin	17	18	20
Dyviol	linear	plugin	11	11	11

Selection of controls in IV: Institutions and Output
Acemoglu, Johnson, and Robinson (2001)

Three equation system:

$$log(GDPpercapita_i) = \alpha \bullet ProtectionfromExpropriation_i + x_i'\beta + \varepsilon_i.$$

$$ProtectionfromExpropriation_i = \pi_1 \bullet SettlerMortality_i + x_i'\Pi_2 + v_i$$

$$SettlerMortality_i = x_i'\gamma + u_i,$$

Reduced form

$$log(GDPpercapita_i) = x_i'\widetilde{\beta} + \widetilde{\varepsilon}_i.$$

$$ProtectionfromExpropriation_i = x_i'\widetilde{\Pi}_2 + \widetilde{v}_i$$

$$SettlerMortality_i = x_i'\gamma + u_i.$$

- Paper controls for Latitude.
- Pre-selection of control variables x_i:
  - Latitude, latitude^2 , latitude^3 , (latitude-.08)$_+$, (latitude-.16)$_+$, (latitude-.24)$_+$, ((latitude-.08)$_+$)^2 , ((latitude-.16)$_+$)^2 , ((latitude-.24)$_+$)^2 , ((latitude-.08)$_+$)^3 , ((latitude-.16)$_+$)^3 , and ((latitude-.24)$_+$)^3 where latitude denotes the distance of a country from the equator normalized to be between 0 and 1, the breakpoints in the latitude function were chosen by taking round numbers near the quartiles of latitude, and (a)$_+$ is shorthand notation for (a)1(a $>$ 0) where 1($\cdot$) is the indicator function that returns 1 when the expression inside the parentheses is true and 0 otherwise.
- Use Lasso to automatically select controls to include in regression.
  - NB. Only the Africa-dummy selected for both equations. Results not much affected.

IV-Example: Estimating the Impact of Eminent Domain on House Prices (Belloni et al EMA, 2012)s

- Eminent domain refers to the government's taking of private property.
- Endogeneity between takings law decisions and economic variables: for example, a taking may be less likely if real estate prices are low and sellers are eager to unload property.
- Solution: random assignment of judges to federal appellate panels.
    - The identity of the judges and their demographics are randomly assigned conditional on the distribution of characteristics of federal circuit court judges in a given circuit-year.
    - Thus the judge's characteristics will plausibly satisfy the instrumental variable exclusion restriction.
- All judges' characteristics satisfy the instrumental variables exclusion restriction.
    - Use Lasso to select strong instruments in this set.

- Estimated equation

$$log(Case - Shiller_{ct}) = \alpha \cdot TakingsLaw_{ct} + \beta_c + \beta_t + \gamma_c t + W'_{ct}\delta + \varepsilon_{ct},$$

$$TakingsLaw_{ct} = x'_{ct}\theta + +v_{ct}.$$

  - $Case - Shiller_{ct}$ = average Case–Shiller home price index within circuit court c at time t;
  - $TakingsLaw_{ct}$ = # of pro-plaintiff (government taking of land was unlawful) appellate takings decisions in federal circuit court c and year t
  - W_{ct} : exogenous variables incl. a dummy variable for whether there were relevant cases in that circuit-year, the number of takings appellate decisions, and controls for the distribution of characteristics of federal circuit court judges in a given circuit-year;
  - β_c, β_t, and $\gamma_c t$ are respectively circuit- and time-specific effects, and circuit-specific time trends.
  - α = the effect of an additional decision upholding individual property rights on a measure of property prices.
  - x_{ct} are characteristics of the judges in a court circuit-year.

- Specification:

$$\widehat{\theta} \in \arg\min E\left[\left(TakingsLaw_{ct} - x'_{ct}\theta\right)^2\right] + \frac{\lambda}{n}\left|\widehat{\Upsilon}_l\theta\right|$$

  where

$$\widehat{\Upsilon}_l = diag\left(\widehat{\gamma}_{l1}, ..., \widehat{\gamma}_{lp}\right)$$

  is a diagonal matrix specifying penalty loadings.

  - λ and $\widehat{\Upsilon}_l$ are set to obtain sharp convergence results for the Lasso estimator.

- Fast convergence under the condition that the log of the number of regressors p is small relative to $n1/3$, that is, log(p) = o(n1/3).
- Sample size is 183.

- Pre-selection of 147 variables
  - Use economic intuition to select candidate variables: Gender, race, religion, party affifiliation, source of academic degrees (BA from in-state university, BA from a public university, JD from a public university, has an LLM or SJD), and whether the judge had been elevated from a district court.
  - For each, three new variables constructed: counting the number of panels with one, two or three members with each characteristic, and three members with each characteristic.
  - First-order interactions between all of the previously mentioned variables, a cubic polynomial in the number of panels with at least one Democrat, a cubic polynomial in the number of panels with at least one member with a JD from a public university, and a cubic polynomial in the number of panels with at least one member elevated from within the district.
  - Additional pre-processing to remove instrument with extremely small standard deviation was extremely small and one instrument from any pair of instruments that had a bivariate correlation exceeding .99 in absolute value.
- Among these, use LASSO to identify instruments that strongly predict $\mathit{TakingsLaw}_{ct}$.
  - Selects one instrument only: the number of panels with one or more members with JD from a public university squared

- Results:
    - first-stage coeffificient of 0.45 with standard error of 0.05
    - second stage estimate of 0.065 with estimated standard error of 0.024.
    - a single additional judicial decision reinforcing individual property rights is associated with between 2 and 11 percent higher property prices with an average number of pro-plaintiff decisions per year of 0.19.
- Main benefit of selection is to find stronger instrument in first stage.
    - I added 1000 random (normal) potential instruments. The procedure still identified the same instrument 100 times of 100.
    - I also removed the one instrument selected by the procedure. The procedure then identified linear version of the same variable, and gives similar results.

These methods are a complement to sensitivity analysis, not a substitute

- Sparsity is a strong and untestable assumption.
  - Sparsity: Including only s non-zero coefficients make approximation errors small as n increases:

$$E\left[r_{yi}^2\right]^{\frac{1}{2}}, E\left[r_{di}^2\right]^{\frac{1}{2}} \leq c\sqrt{\frac{s}{n}}, \text{ for some } s.$$

- Angrist and Frandsen (2021): when the sparsity condition is violated, IV-estimates using ML selection of instruments are biased.
- Use to identify controls that should be added to the regression. It is not sufficient to add these controls for robust identification.

Problem set: cross-fit partial out (xporegress in Stata)

Effect of unemployment insurance on unemployment duration

In this problem set you will apply the cross-fit partial out method to study the effect of unemployment insurance..The Pennsylvania Reemployment Bonus experiment was conducted by the US Department of Labor in the 1980s. UI claimants were randomly assigned either to a control group or to one of five treatment groups. Individuals in the treatment groups were offered a cash bonus if they found a job within some pre-specified period of time (qualification period), provided that the job was retained for a specified duration. In the control group, the standard rules of the UI system applied.

The data is provided on Athena as penn_jae.dta. Our treatment variable, D, is an indicator variable for being assigned treatment tg=4 (drop tg=1,2,3), and the outcome variable, Y, is the log of duration of unemployment for the UI claimants (log(inuidur1)). The vector of covariates consists of age group dummies, gender, race, the number of dependents, quarter of the experiment, location within the state, existence of recall expectations and type of occupation (variables female- husd). We use a set of potential control variables X formed from the raw set of covariates and all second-order terms (i.e. all squares and first-order interactions).

1a. Use the cross-fit partialing out estimator to estimate

$$Y = \theta_0 D + g_0(X) + U,$$

where $g_0(X)$ is estimated using Lasso. Using 5 folds and set the tuning parameter using the plugin formula. Resample 15 times. Set a random seed to allow replication.

Compare your results to Table 1, column 1 of Chernozhukov et al (2018). How many control variables are selected (with non-zero coefficients) in the regression for Y. The selected variables vary by fold and sample, which are the selected variables for a specific fold and sample?

b. Set the raw covariates so that they are always included in the regression and use Lasso to select additional controls among the second-order terms. How many control variables are selected (with non-zero coefficients),and which are these variables?

c. Use the specification in 1a but now select the tuning parameter by cross-validation. How many variables are included, why is this number higher than in 1a?

Table 1 Chernozhukov et al. (2018)

Table 1. Estimated effect of cash bonus on unemployment duration.

	Lasso	Reg. tree	Random forest	Boosting	Neural network	Ensemble	Best
Panel A: interactive regression model							
ATE	−0.081	−0.084	−0.074	−0.079	−0.073	−0.079	−0.078
(twofold)	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]
	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
ATE	−0.081	−0.085	−0.074	−0.077	−0.073	−0.078	−0.077
(fivefold)	[0.036]	[0.036]	[0.036]	[0.035]	[0.036]	[0.036]	[0.036]
	(0.036)	(0.037)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
Panel B: partially linear regression model							
ATE	−0.080	−0.084	−0.077	−0.076	−0.074	−0.075	−0.075
(twofold)	[0.036]	[0.036]	[0.035]	[0.035]	[0.035]	[0.035]	[0.035]
	(0.036)	(0.036)	(0.037)	(0.036)	(0.036)	(0.036)	(0.036)
ATE	−0.080	−0.084	−0.077	−0.074	−0.073	−0.075	−0.074
(fivefold)	[0.036]	[0.036]	[0.035]	[0.035]	[0.035]	[0.035]	[0.035]
	(0.036)	(0.037)	(0.036)	(0.035)	(0.036)	(0.035)	(0.035)

Note: Estimated ATE and standard errors from a linear model (Panel B) and heterogeneous effect model (Panel A) based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Results are based on 100 splits with point estimates calculated the median method. The median standard errors across the splits are reported in brackets and standard errors calculated using the median method to adjust for variation across splits are provided in parentheses. Further details about the methods are provided in the main text.

Note non-linear specifications of $g_0(X)$: regression tree, random forest, etc.