Machine Learning for Microeconometrics
Part 5: More Causal Inference

A. Colin Cameron
Univ.of California - Davis

.

April 2024

Course Outline

- **1.** Variable selection and cross validation
- **2.** Shrinkage methods
  - ridge, lasso, elastic net
- **3.** ML for causal inference using lasso
  - OLS with many controls, IV with many instruments
- **4**. Other methods for prediction
  - nonparametric regression, principal components, splines
  - neural networks
  - regression trees, random forests, bagging, boosting
- **Part 5: More ML for causal inference**
  - ATE with heterogeneous effects and many controls.
- **6.** Classification and unsupervised learning
  - classification (categorical y) and unsupervised learning (no y).

1. Introduction

- Current microeconometric applications focus on **causal estimation** of a key parameter, such as an average marginal effect, after controlling for confounding factors
  - apply to models with selection on observables only
    - good controls makes this assumption more reasonable
  - and to IV with available instruments
    - good few instruments avoids many instruments problem.
- Machine learning methods determine good controls (or instruments)
  - but valid statistical inference needs to control for this data mining
  - currently extraordinarily active area of econometrics research.
- Previously considered LASSO for partial linear model with **homogeneous** effects.
- Now consider **heterogeneous effects** in **potential outcomes model**.
- This research area is currently exploding
  - these slides will become dated quickly.

Overview

2. Machine Learning for Microeconometrics

- Empirical microeconometrics studies focus on estimating partial effects
  - the effect on y of a change in x_1 controlling for $\mathbf{x}_2$.
- A machine learner would calculate this as follows
  - prediction function is $\widehat{y} = \widehat{f}(x_1, \mathbf{x}_2)$
  - the partial effect of a change of size Δx_1 is then

$$\Delta\widehat{y} = \widehat{f}(x_1 + \Delta x_1, \mathbf{x}_2) - \widehat{f}(x_1, \mathbf{x}_2).$$

- This could be very complicated as $\widehat{f}(\cdot)$ may be very nonlinear in x_1.
- There is difficulty (impossibility?) in obtaining an asymptotic distribution for inference.
- And it requires a correct model $\widehat{f}(x_1, \mathbf{x}_2)$
  - formally the model needs to be consistent
  - i.e. probability that $\widehat{f}(\cdot)$ is correct $\rightarrow 1$ as $n \rightarrow \infty$.

Add Some Structure

- A partially linear control function model specifies

$$y = \alpha x_1 + g(\mathbf{x}_2) + u \text{ where } g(\cdot) \text{ is unknown.}$$

  - for simplicity consider only scalar x_1.
- The partial effect of a change of size Δx_1 is then

$$\Delta\widehat{y} = \alpha \Delta x_1.$$

- Consistent estimator requires $E[y|x_1, \mathbf{x}_2] = \alpha x_1 + g(\mathbf{x}_2)$.
  - more plausible the better the choice of $g(\mathbf{x}_2)$
  - though we still need linear in x_1 and additivity.
- The partially linear model was used initially in semiparametrics
  - typically $\mathbf{x}_1$ and α were high dimension and $\mathbf{x}_2$ low dimension
  - now for causal ML x_1 and α are high dimension and $\mathbf{x}_2$ is high dimension.

How to add the controls

- Biostatistics includes regressors $\mathbf{x}_2$ as controls if $p < 0.05$
  - imperfect selection and also leads to pre-test bias.
- Economists use economics theory and previous studies to include regressors
  - these are included regardless of their statistical significance
  - to guard against omitted variables bias and to avoid pre-test bias.
- Machine learning methods are used to get a good choice of $g(\mathbf{x}_2)$
  - ideally in such a way and/or with assumptions so that standard inference can be used for $\widehat{\alpha}$
    - so data mining has not affected the distribution of $\widehat{\alpha}$.
  - The methods can extend to endogenous x_1.
- The course to date has focused on determine $g(\mathbf{x}_2)$ using the LASSO
  - due to Belloni, Chernozhukov and Hansen and coauthors
  - assumptions including "sparsity" enable use of standard inference for $\widehat{\alpha}_1$.

Alternatively estimate average partial effects

- An alternative to the partially linear model is to use less structure and estimate average partial effects.
- The leading example is the heterogeneous effects literature
  - let x_1 be a binary treatment taking values 0 or 1
  - let $\Delta y/\Delta x_1$ vary across individuals in an unstructured way
  - estimate the average partial effect $E[y|x_1 = 1] - E[y|x_1 = 0]$.
- One method used is propensity score matching
  - machine learning may give a better propensity score estimator.
- Another method used is nearest-neighbors matching
  - machine learning may give a better matching algorithm.
- To control for data mining, however, use an estimator that satisfies the orthogonalization condition.

3. ATE with heterogeneous effects

- Consider the effect of binary treatment d on an outcome y
  - $d = 1$ if treated and $d = 0$ if untreated (control).
- If we could run a **randomized control trial** (RCT)
  - assignment to treatment is completely random (e.g. toss a coin)
  - then the average treatment effect would be simply the difference in means $\bar{y}_1 - \bar{y}_0$.
- Important things to note
  - the treatment effect can differ from individual (we just average)
    - even though mechanically OLS $y_i = \alpha + \beta d_i + u_i$ gives $\widehat{\beta} = \bar{y}_1 - \bar{y}_0$.
  - there is no need to control for $\mathbf{x}'s$ given random assignment
    - though potentially adding controls could improve estimator precision.

Homogeneous effects

- In practice in economics we usually have observational data
  - where individuals may self-select into treatment (d is endogenous).
- The control function approach specifies the **partial linear model**

$$y = \beta d + g(\mathbf{x}) + u.$$

- The key nontestable assumption (**selection on observables only**) is
  - once we include the control function $g(\mathbf{x})$ the treatment variable d can be viewed as if it is randomly assigned (i.e. d is **exogenous**)
  - d is uncorrelated with (or independent of) the error u conditional on $\mathbf{x}$.
- Better control functions $g(\mathbf{x})$ may make this assumption more plausible
  - earlier we used an ML method such as Lasso for flexible $g(\mathbf{x})$.
- This model restricts the treatment effect to be the same β for each individual with the same $\mathbf{x}$
  - called **homogeneous effects**.

3.1 Heterogeneous Effects Model

- The **heterogeneous effects model** allows the treatment effect to differ across individuals
  - it is more flexible
  - and more plausible to believe $\mathbf{x}$ can control for self-selection.
- As for an RCT we wish to estimate the average treatment effect

$$\tau = \text{ATE} = E[y^{(1)} - y^{(0)}]$$

  - where $y^{(1)}$ is the potential outcome if treated $(d = 1)$
  - $y^{(0)}$ is the potential outcome if not treated $(d = 0)$
  - and for any given individual we only observe one of $y_i^{(1)}$ or $y_i^{(0)}$.

Unconfoundedness assumption

- The key nontestable assumption (**unconfoundedness** or **conditional independence**) is
  - once we adjust for control variables $\mathbf{x}$ the treatment variable d can be viewed as if it is randomly assigned (i.e. exogenous)
  - d is independent of the potential outcomes $y^{(0)}$ and $y^{(1)}$ conditional on $\mathbf{x}$.

3.2 Heterogeneous Effects Model Estimators

- Several estimators have been proposed for this model.
  - given in detail in the Appendix.
- Regression model adjustment
  - estimate separate models of y on $\mathbf{x}$ for the treated and the untreated
  - predict y for all individuals using the treated coefficient estimates and predict y for all individuals using the untreated coefficient estimates
  - finally compute the difference in the average predictions.
- Inverse probability weighting (IPW) using the propensity score
  - Use a weighted average of treated y's and untreated y's
  - with weights that adjust for the probability of selection into treatment.
- Augmented IPW combines the preceding two methods.
- Other methods include matching
  - compare the outcome for the treated to the outcome for similar (on $\mathtt{x}$) untreated.

3.2 ATE estimated using Augmented IPW

- Define the following quantities
  - $\mu_1(\mathbf{x}) = E[y^{(1)}|\mathbf{x}]$ the conditional mean of y if treated
  - $\mu_0(\mathbf{x}) = E[y^{(0)}|\mathbf{x}]$ the conditional mean of y if untreated
  - $p(\mathbf{x}) = \Pr[d = 1|\mathbf{x}]$ the conditional probability of treatment (**propensity score**)
- Define corresponding regression estimates for each individual
  - $\widehat{\mu}_1(\mathbf{x}_i)$, $\widehat{\mu}_0(\mathbf{x}_i)$, $\widehat{p}(\mathbf{x}_i)$.
- The doubly-robust method (augmented IPW) uses

$$\widehat{\tau} = \widehat{\text{ATE}} = \begin{array}{l} \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{d_i(y_i-\widehat{\mu}_1(\mathbf{x}_i))}{\widehat{p}(\mathbf{x}_i)} + \widehat{\mu}_1(\mathbf{x}_i)\right\} \\ -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(1-d_i)(y_i-\widehat{\mu}_0(\mathbf{x}_i))}{1-\widehat{p}(\mathbf{x}_i)} + \widehat{\mu}_0(\mathbf{x}_i)\right\} \end{array}$$

- **Doubly-robust** as estimator remains consistent if either
  - the propensity score model $p(\mathbf{x})$ or
  - the regression imputation model $\mu_j(\mathbf{x})$ is misspecified.
- Stata command **teffects aipw** estimates this with OLS and logit.

4. LASSO for the AIPW estimate of the ATE

- For AIPW the moment condition for the ATE parameter τ is
  - $\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{d_i(y_i-\widehat{\mu}_1(\mathbf{x}_i))}{\widehat{p}(\mathbf{x}_i)}+\widehat{\mu}_1(\mathbf{x}_i)-\frac{(1-d_i)(y_i-\widehat{\mu}_0(\mathbf{x}_i))}{1-\widehat{p}(\mathbf{x}_i)}-\widehat{\mu}_0(\mathbf{x}_i)-\tau\right\}=0.$
- In addition to being **doubly-robust**, the **orthogonalization condition** holds (shown below).
- So use
  - LASSO logit to obtain $\widehat{p}(\mathbf{x})$
  - LASSO OLS to obtain $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$.
- The Stata command **telasso** implements this.
- Max Farrell (2015), "Robust Estimation of Average Treatment Effect with Possibly more Covariates than Observations," *Journal of Econometrics*, 189, 1-23.
  - considers multivalued treatment but I present binary d case.

Stata telasso command

- The following code gives lasso AIPW with various methods for the lasso penalty parameter λ where
  - y is log medical expenditures
  - d is whether have supplementary health insurance
  - $\mathbf{x}$ is 176 controls (`$rlist2`)

```
* Plugin values for lambda (the default)
telasso (ltotexp $rlist2) (suppins $rlist2), selection(plugin) vce(robust)
* BIC values for lambda
telasso (ltotexp $rlist2) (suppins $rlist2), selection(bic) vce(robust)
* CV takes a long time
telasso (ltotexp $rlist2) (suppins $rlist2), selection(cv) xfolds(10) ///
   rseed(10101) vce(robust)
```

Stata Results

- Apply to earlier data.
- ATE is effect of supplementary insurance on log medical expenditures

Method	**Stata command**	**ATE**	**se(ATE)**
Partial linear lasso	`poregress`	0.1839	0.0468
Regression adjustment	`teffects, ra`	0.1745	0.0496
IPW	`teffects, ipw`	0.1867	0.0481
Augmented IPW	`teffects, aipw`	0.1713	0.0483
Lasso AIPW plugin	`telasso, sel(plug)`	0.1502	0.0519
Lasso AIPW bic	`telasso, sel(bic)`	0.1428	0.0596
Lasso AIPW CV	`telasso, sel(cv)`	0.1496	

5. Random Forests for Causal ATE

- Random forests predict very well
  - Susan Athey's research emphasizes random forests.
- Stefan Wager and Susan Athey (2018), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," JASA, 1228-1242.
- Standard binary treatment and heterogeneous effects with unconfoundness assumption
  - use random forests to determine the controls.
  - proves asymptotic normality and gives point-wise confidence intervals
    - ★ This is a big theoretical contribution.
- Stefan Wager and Susan Athey (2018), "Estimating treatment Effects with Causal Forests: An Application," Observational Studies 5, September 2019, 21-35 (also https://arxiv.org/pdf/1902.07409)
  - a how-to application (and allows for clustered errors)
  - uses the R package `grf`.

Random Forests for Causal ATE (continued)

- Let L denote a specific leaf in tree b.
- $\tau(\mathbf{x}) = E[y^{(1)} - y^{(0)}|\mathbf{x}]$ in a single regression tree b is estimated by

$$\begin{aligned}\widehat{\tau}_b(\mathbf{x}) &= \tfrac{1}{\#\{i:d_i=1,\mathbf{x}_i\in L\}}\sum_{i:d_i=1,\mathbf{x}_i\in L} y_i - \tfrac{1}{\#\{i:d_i=0,\mathbf{x}_i\in L\}}\sum_{i:d_i=0,\mathbf{x}_i\in L} y_i \\ &= \bar{y}_1 \text{ in leaf } L - \bar{y}_0 \text{ in leaf } L.\end{aligned}$$

- Then a random forest with sub-sample size s gives B trees with

$$\begin{aligned}\widehat{\tau}_b(\mathbf{x}) &= \tfrac{1}{B}\sum_{b=1}^{B}\widehat{\tau}_b(\mathbf{x}) \\ \widehat{Var}[\widehat{\tau}_b(\mathbf{x})] &= \tfrac{n-1}{n}\left(\tfrac{n}{n-2}\right)^2\sum_{i=1}^{n} Cov(\widehat{\tau}_b(\mathbf{x}), d_{ib})\end{aligned}$$

  - where $d_{ib} = 1$ if i^{th} observation in tree b and 0 otherwise
  - and the covariance is taken over all B trees.
- Key is that a tree is honest.
- A tree is honest if for each training observation i it only uses y_i to
  - either estimate $\widehat{\tau}(\mathbf{x})$ within leaf
  - or to decide where to place the splits
  - but not both (analogous to cross fitting in partial linear model).

6. Deep Neural Networks for Causal ATE

- Max Farrell, Tengyuan Liang and Sanjog Misra (2021), ““Deep Neural Networks for Estimation and Inference,” Econometrica.
  - Further detail in “Deep Neural Networks for Estimation and Inference: Application to Causal Effects and Other Semiparametric Estimands,” arXiv:1809.09953v2.
- Obtains nonasymptotic bounds and convergence rates for nonparametric estimation using deep neural networks.
- Then obtain asymptotic normal results for inference on finite-dimensional parameters following first-step estimation using deep neural nets.
- Application is to ATE using doubly robust augmented IPW
  - outcome is consumer spending and interest is in effect of marketing
  - consider effect of three different targeting strategies: (1) never treat; (2) blanket treatment; (3) loyalty policy.

7.1 LATE and local quantile treatment effects

- Belloni, Chernozhukov, Fernandez-Val and Hansen (2015), "Program Evaluation with High-Dimensional Data".
- Binary treatment and heterogeneous effects with endogenous treatment and valid instruments
  - allow for estimation of functions
    - such as local quantile treatment effects over a range of quantiles
  - The paper is very high level as it uses functionals
  - uses LASSO along the way.
- Key is to use an orthogonalization moment condition
  - allows inference to be unaffected by first-stage estimation.

8. Some review articles of ML for Economics

- Susan Athey's website has several wider-audience papers on machine learning in economics.
- Susan Athey (2017), "Beyond Prediction: Using Big Data for Policy Problems," Science 355, 483-485.
  - Off-the shelf prediction methods assume a stable environment
    - ★ includes Kleinberg et al (2015) AER hip replacement.
  - Economics considers causal prediction by
    - ★ adjust for confounders e.g. Belloni et al., Athey et al.
    - ★ designed experiments e.g. Blake et al.
    - ★ excellent references.

Susan Athey (continued)

- Susan Athey (2018), “The Impact of Machine Learning on Economics"
- Lengthy wide-ranging survey paper with no equations.
- Machine learning methods can
  - provide variables to be used in economic analysis (e.g. from images or text)
  - lead to better model selection through e.g. cross-validation
  - provide much quicker computation using stochastic gradient descent
    - ⋆ use gradient at a single data point to approximate average over observations of the gradient
  - lead to better causal estimates
    - ⋆ fundamental identification issues are not solved
    - ⋆ but perhaps make assumptions more credible e.g. unconfoundedness
  - be used whenever semiparametric methods might have been used.
- Paper surveys recent work on ML for causal inference
  - double machine learning (Chernozhukov et al 2018) and orthogonalization are especially promising.

Other Sources

- Dario Sansone (University of Exeter) provides very many good references
  - https://sites.google.com/view/dariosansone/resources/machine-learning
- Susan Athey and Guido Imbens (2019), "Machine Learning Methods Economists Should Know About," Annual Review of Economics.
- This paper provides great detail on the current literature with many references.

9. Appendix: Heterogeneous Effects Model and AIPW

- This appendix provides details for those unfamiliar with heterogeneous effects and associated estimation methods for a binary treatment.
  - Rubin causal model
  - Average treatment effect (ATE)
  - Regression model adjustment estimator
  - Inverse probability-weighted (IPW) estimator
  - Doubly-robust Augmented IPW estimator
  - Proof of orthogonalization condition for AIPW.

9.1 Rubin Causal Model and Potential Outcomes

- Consider a **binary treatment** $d \in \{0, 1\}$
  - for some individuals we observe y only when $d = 1$ (treated)
  - for others we observe y only when $d = 0$ (untreated or control)
  - some methods generalize to multi-valued treatment $d \in \{0, 1, ..., J\}$.
- The **Rubin causal model** defines
  - **potential outcomes** $y^{(1)}$ if $d = 1$ and $y^{(0)}$ if $d = 0$
  - for a given individual we observe only one of $y_i^{(1)}$ and $y_i^{(0)}$
  - we observe $y_i = d_i y_i^{(1)} + (1 - d_i) y_i^{(0)}$.

Average Treatment Effect

- The goal is to estimate the **average treatment effect (ATE)**

$$\tau = \text{ATE} = E[y^{(1)} - y^{(0)}].$$

- Or the conditional treatment effect given $\mathbf{x}$
  - $\tau(\mathbf{x}) = E[y^{(1)} - y^{(0)}|\mathbf{x}]$.
- Also may be interested in the **average treatment effect on the treated (ATET)**

$$\text{ATET} = E[y^{(1)} - y^{(0)}|d = 1].$$

Unconfoundness

- A treatment assignment mechanism is **unconfounded** if assignment to treatment does not depend on the potential outcomes.
- Thus $\Pr[d|y^{(0)}, y^{(1)}, \mathbf{x}] = \Pr[d|y^{*(0)}, y^{*(1)}, \mathbf{x}]$ for all $d, y^{(0)}, y^{(1)}, y^{*(0)}, y^{*(1)}, \mathbf{x}$.
- **This crucial nontestable assumption is also called the conditional independence assumption** and is often written as
  - $d_i \perp \{y_i^{(0)}, y_i^{(1)}\}|\mathbf{x}_i$.
  - conditional on $\mathbf{x}$, treatment is independent of the potential outcome.
- This means once we condition on $\mathbf{x}$
  - the conditional distribution of the potential outcome if treated $(y^{(1)})$ is the same for those who did and did not actually get treatment
    - $y_i^{(1)}|d_i = 1, \mathbf{x}$ has the same distribution as $y_i^{(1)}|d_i = 0, \mathbf{x}$
  - the conditional distribution of the potential outcome if not treated $(y^{(0)})$ is the same for those who did and did not actually get treatment
    - $y_i^{(0)}|d_i = 1, \mathbf{x}$ has the same distribution as $y_i^{(0)}|d_i = 0, \mathbf{x}$.

Heterogeneous Effects Model Estimators

- Several estimators have been proposed for this model
  - given in detail next.
- Regression model adjustment
  - estimate separate models of y on $\mathbf{x}$ for the treated and the untreated
  - predict y for all individuals using the treated coefficient estimates and predict y for all individuals using the untreated coefficient estimates
  - finally compute the difference in the average predictions.
- Inverse probability weighting (IPW) using the propensity score
  - Use a weighted average of treated y's and untreated y's
  - with weights that adjust for the probability of selection into treatment.
- Augmented IPW combines the preceding two methods.
- Other methods include matching
  - compare the outcome for the treated to the outcome for similar (on $\mathtt{x}$) untreated.

9.2 ATE estimated using Regression Model Adjustment

- Regress y on $\mathbf{x}$ for the treated sample, regress y on $\mathbf{x}$ for the untreated sample, predict potential outcomes for a person if treated and for the same person if untreated, average for all individuals and subtract.
- Define the conditional means
  - $\mu_1(\mathbf{x}) = E[y^{(1)}|\mathbf{x}]$ for treated
  - $\mu_0(\mathbf{x}) = E[y^{(0)}|\mathbf{x}]$ for control
  - so $\text{ATE}(\mathbf{x}) = \tau(\mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$.
- Then

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(\mathbf{x}_i) - \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_0(\mathbf{x}_i).$$

- Stata command **teffects ra** does this
  - OLS regression with specified functions $\mu_1(\cdot)$ & $\mu_0(\cdot)$ and specified $\mathbf{x}$
  - equals $\widehat{\beta}_2$ in OLS regression
    $y_i = \beta_1 + \beta_2 d_i + \mathbf{x}_i'\boldsymbol{\beta}_3 + d_i\mathbf{x}_i'\boldsymbol{\beta}_4 + u_i,\ i = 1, ..., n.$

9.3 ATE Estimated using Inverse Probability Weighting

- Adjust for selection into treatment using the propensity score.
- Define the **propensity score** $p(\mathbf{x}) = \Pr[d=1|\mathbf{x}] = E[d|\mathbf{x}]$.
- Under the conditional independence assumption
  - $\mu_1(\mathbf{x}) = E[y^{(1)}|\mathbf{x}] = E[\frac{dy}{p(\mathbf{x})}|\mathbf{x}]$ shown on next slide
  - $\mu_0(\mathbf{x}) = E[y^{(0)}|\mathbf{x}] = E[\frac{(1-d)y}{1-p(\mathbf{x})}|\mathbf{x}]$ by similar proof
  - $\text{ATE}(\mathbf{x}) = \tau(\mathbf{x}) = E[y^{(1)} - y^{(0)}|\mathbf{x}] = E\left[\left(\frac{dy}{p(\mathbf{x})} - \frac{(1-d)y}{1-p(\mathbf{x})}\right)|\mathbf{x}\right]$
    - ★ downweights y for treated with high $p(\mathbf{x})$ and y for untreated with low $p(\mathbf{x})$.
- **Inverse probability weighting** (IPW) uses the sample analog

$$\widehat{\tau} = \widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i y_i}{\widehat{p}_i(\mathbf{x}_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-d_i)y_i}{1-\widehat{p}(\mathbf{x}_i)}.$$

Proof of IPW

- Note that $d^2 = d$ for binary $d = 0$ or 1. Then

$$d \times y = d \times \{dy^{(1)} + (1-d)y^{(0)}\} = d^2 y^{(1)} + (d - d^2)y^{(0)} = dy^{(1)}$$

- So

$$\begin{aligned}
& E_{d,y^{(0)},y^{(1)}}\left[\frac{dy}{p(\mathbf{x})}|\mathbf{x}\right] \\
= \; & E_{d,y^{(0)},y^{(1)}}\left[\frac{dy^{(1)}}{p(\mathbf{x})}|\mathbf{x}\right] \text{ as } d \times y = dy^{(1)} \\
= \; & E_d\left[\frac{d}{p(\mathbf{x})}|\mathbf{x}\right] \times E_{y^{(1)}}\left[y^{(1)}|\mathbf{x}\right] \text{ by unconfoundness assumption} \\
= \; & \frac{p(\mathbf{x})}{p(\mathbf{x})} \times E_{y^{(1)}}\left[y^{(1)}|\mathbf{x}\right] \text{ as } E_d\left[d|\mathbf{x}\right] = p(\mathbf{x}) \\
= \; & E_y\left[y_1^{(1)}|\mathbf{x}\right]
\end{aligned}$$

ATE estimated using propensity scores (continued)

- Stata command **teffects ipw** estimates this
  - logit specification for $p(\mathbf{x})$ with specified $\mathbf{x}$.
- Instead could use ML methods such as LASSO logit to get $\widehat{p}(\mathbf{x})$
  - The conditional independence assumption is more plausible the more $\mathbf{x}'s$ considered.
- This method works best when $\widehat{p}(\mathbf{x})$ is constant as in a randomized trial.
- When $\widehat{p}(\mathbf{x})$ is close to 0 or 1 the weights become very large.
- Then it is better to use a **blocking estimator**
  - partition observations into subclasses based on value of $\widehat{p}(\mathbf{x})$
  - compute the ATE in each subclass as $\bar{y}_1 - \bar{y}_0$
  - then ATE is the average across subclasses (weighted by subclass size).

9.4 ATE estimated using doubly-robust AIPW method

- As before define $\mu_1 = E[y^{(1)}]$ and $\mu_0 = E[y^{(0)}]$ and

$$\mu_1(\mathbf{x}) = E[y^{(1)}|\mathbf{x}];\ \mu_0(\mathbf{x}) = E[y^{(0)}|\mathbf{x}];\ p(\mathbf{x}) = \Pr[d = 1|\mathbf{x}].$$

- The doubly-robust method (augmented IPW) combines the preceding regression adjustment and IPW methods and uses

$$\begin{aligned} \widehat{\tau} &= \widehat{\text{ATE}} = \widehat{\mu}_1 - \widehat{\mu}_0 \\ \widehat{\mu}_1 &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{d_i(y_i - \widehat{\mu}_1(\mathbf{x}_i))}{\widehat{p}(\mathbf{x}_i)} + \widehat{\mu}_1(\mathbf{x}_i)\right\} \\ \widehat{\mu}_0 &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(1-d_i)(y_i - \widehat{\mu}_0(\mathbf{x}_i))}{1-\widehat{p}(\mathbf{x}_i)} + \widehat{\mu}_0(\mathbf{x}_i)\right\} \end{aligned}$$

- **Doubly-robust** as estimator remains consistent if either
  - the propensity score model $p(\mathbf{x})$ or
  - the regression imputation model $\mu_j(\mathbf{x})$ is misspecified.
- Stata command **teffects aipw** estimates this with OLS and logit.

Stata telasso command

- The following code gives standard estimators
  - y is log medical expenditures
  - d is whether have supplementary health insurance
  - $\mathbf{x}$ is 176 controls (`$rlist2`)

```
* Homogeneous effects: partialling-out partial linear model
poregress ltotexp suppins, controls($rlist2)
* Heterogeneous effects: regression adjustment estimate of ATE
teffects ra (ltotexp $rlist2) (suppins), vce(robust)
* Heterogeneous effects: inverse probability weighting estimate of ATE
teffects ipw (ltotexp) (suppins $rlist2), vce(robust)
* Heterogeneous effects: Augmented IPW estimate of ATE
teffects aipw (ltotexp $rlist2) (suppins $rlist2), vce(robust)
```

9.5 Orthogonalization for AIPW

- For simplicity define $\eta_1 = \mu_1(\mathbf{x})$, $\eta_2 = \mu_0(\mathbf{x})$ and $\eta_3 = p(\mathbf{x})$.
- The preceding AIPW estimator τ solves the population moment condition $E[\psi(d, y, \boldsymbol{\eta})] = 0$ where

$$\psi(d, y, \tau, \boldsymbol{\eta}) = \frac{d(y - \eta_1)}{\eta_3} + \eta_1 - \frac{(1-d)(y - \eta_2)}{1 - \eta_3} - \eta_2 - \tau.$$

- Orthogonalization requires $E[\psi(d, y, \tau, \boldsymbol{\eta})/\partial \eta_j|\mathbf{x}] = 0$ for $j = 1, 2, 3$.
- $\eta_1 : E[\psi(d, y, \tau, \boldsymbol{\eta})/\partial \eta_1|\mathbf{x}] = E[\frac{-d}{\eta_3} + 1|\mathbf{x}] = \frac{-E[d|\mathbf{x}]}{\eta_3} + 1$ $= \frac{-\eta_3}{\eta_3} + 1 = 0,$ using $E[d|\mathbf{x}] = \Pr[d = 1|\mathbf{x}] = \eta_3$.
- $\eta_2 : E[\psi(d, y, \tau, \boldsymbol{\eta})/\partial \eta_2|\mathbf{x}] = E[\frac{(1-d)}{\eta_3} - 1|\mathbf{x}] = \frac{1-E[d|\mathbf{x}]}{1-\eta_3} + 1$ $= \frac{1-\eta_3}{1-\eta_3} + 1 = 0,$ using $E[d|\mathbf{x}] = \Pr[d = 1|\mathbf{x}] = \eta_3$.

Othogonalization for AIPW (continued)

- Lastly consider derivative w.r.t. η_3. This is less straightforward.
- $\psi(d, y, \tau, \boldsymbol{\eta}) = \frac{d(y-\eta_1)}{\eta_3} + \eta_1 - \frac{(1-d)(y-\eta_2)}{1-\eta_3} - \eta_2 - \tau$.
- $\eta_3 : E[\psi(d, y, \tau, \boldsymbol{\eta})/\partial\eta_3|\mathbf{x}] = E[-\frac{d(y-\eta_1)}{\eta_3^2} - \frac{(1-d)(y-\eta_2)}{(1-\eta_3)^2}|\mathbf{x}]$.
- This term involves the product dy.
  The conditional independence assumption $y^{(0)}, y^{(1)} \perp d|\mathbf{x}$ implies that $E_{d,y|\mathbf{x}}[dy|\mathbf{x}] = \Pr[d = 1|\mathbf{x}] \times E[y^{(1)}|\mathbf{x}] = \eta_3 \times \eta_1$.
- Then $E[\frac{d(y-\eta_1)}{\eta_3^2}|\mathbf{x}] = \frac{E[dy|\mathbf{x}]-E[d|\mathbf{x}]\eta_1}{\eta_3^2} = \frac{\eta_3\eta_1-\eta_3\eta_1}{\eta_3^2} = 0$.
- And similarly
  $E[\frac{(1-d)(y-\eta_0)}{(1-\eta_3)^2}|\mathbf{x}] = \frac{E[(1-d)y|\mathbf{x}]-E[(1-d)|\mathbf{x}]\eta_2}{(1-\eta_3)^2} = \frac{(1-\eta_3)\eta_2-(1-\eta_3)\eta_2}{(1-\eta_3)^2} = 0$.
- So
  $E[\psi(d, y, \tau, \boldsymbol{\eta})/\partial\eta_3|\mathbf{x}] = E[-\frac{d(y-\eta_1)}{\eta_3^2} - \frac{(1-d)(y-\eta_2)}{(1-\eta_3)^2}|\mathbf{x}] = 0 - 0 = 0$.

10. References

- Achim Ahrens, Christian Hansen, Mark Schaffer (2020), "lassopack: Model selection and prediction with regularized regression in Stata," The Stata Journal, Vol.20, 176–235.
- Susan Athey (2018), "The Impact of Machine Learning on Economics," The economics of artificial intelligence, 507-552.
- Susan Athey and Guido Imbens (2019), "Machine Learning Methods Economists Should Know About," Annual Review of Economics 11, 685-725.
- Alex Belloni, Victor Chernozhukov and Christian Hansen (2011), "Inference Methods for High-Dimensional Sparse Econometric Models," Advances in Economics and Econometrics, ES World Congress 2010, ArXiv 2011.
- Alex Belloni, D. Chen, Victor Chernozhukov and Christian Hansen (2012), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain", Econometrica, Vol. 80, 2369-2429.
- Alex Belloni, Victor Chernozhukov and Christian Hansen (2014), "High-dimensional methods and inference on structural and treatment effects," Journal of Economic Perspectives, Spring, 29-50.

References (continued)

- Alex Belloni, Victor Chernozhukov, Ivan Fernandez-Val and Christian Hansen (2017), “Program Evaluation and Causal Inference with High-Dimensional Data,” Econometrica, 233-299.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins (2018), “Double/debiased machine learning for treatment and structural parameters,” The Econometrics Journal, 21, C1-C68.
- Max Farrell (2015), “Robust Estimation of Average Treatment Effect with Possibly more Covariates than Observations", Journal of Econometrics, 189, 1-23.
- Max Farrell, Tengyuan Liang and Sanjog Misra (2021), “Deep Neural Networks for Estimation and Inference,” Econometrica.
- Stefan Wager and Susan Athey (2018), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests,” JASA, 1228-1242.
- Stefan Wager and Susan Athey (2018), “Estimating Treatment Effects with Causal Forests: An Application,” Observational Studies 5, September 2019, 21-35 (also https://arxiv.org/pdf/1902.07409)