Machine Learning Methods in Economics

A. Colin Cameron
U.C.-Davis

.

Presented at University of California - Riverside

April 24 2019

Introduction

- Begin with an economics application of machine learning to straight prediction.
- Then focus on the microeconometrics literature where concerned especially with
  - causal inference for a treatment effect
  - with valid inference controlling for data mining.
- These slides are a companion to my slides
  - Machine Learning Overview Part 1: Basics - selection, shrinkage, dimension reduction
  - Machine Learning Overview Part 2: Flexible methods
  - (and to the more abbreviated Machine Learning Overview)

Causal Inference with Machine Learning

- 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.
- Consider both homogeneous effects and heterogeneous effects.
- This research area is currently exploding
    - these slides may become dates quickly.

1. Prediction for economics
2. Machine learning for microeconometrics
3. Causal homogeneous effects
    1. Focus on LASSO methods
4. Causal heterogeneous effects
    1. LASSO for doubly-robust ATE
    2. Random forests for conditional treatment effects
5. Double/debiased machine learning
6. Some review articles of ML for Economics

1. Prediction for Economics

- Microeconometrics focuses on estimation of β or of partial effects.
- But in some cases we are directly interested in predicting y
  - probability of one-year survival following hip transplant operation
    - if low then do not have the operation.
  - probability of re-offending
    - if low then grant parole to prisoner.
- Mullainathan and Spiess (2017)
  - consider prediction of housing prices
  - detail how to do this using machine learning methods
  - and then summarize many recent economics ML applications.
- So summarize Mullainathan and Spiess (2017).

1.1 Summary of Machine Learning Algorithms

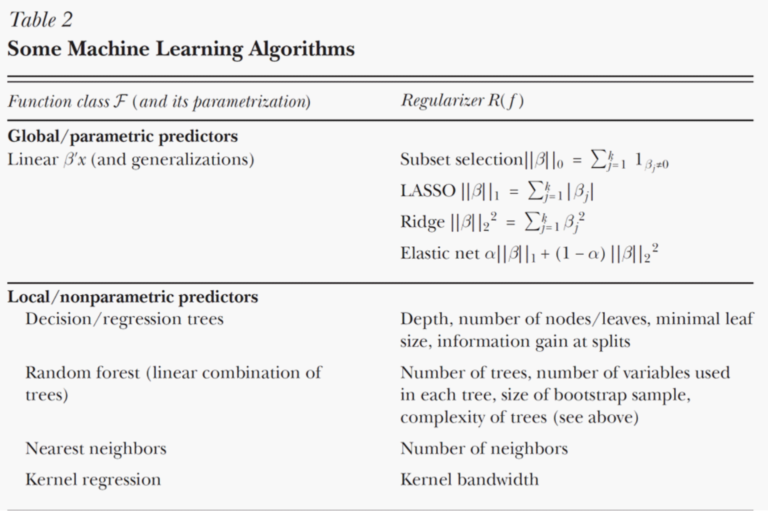
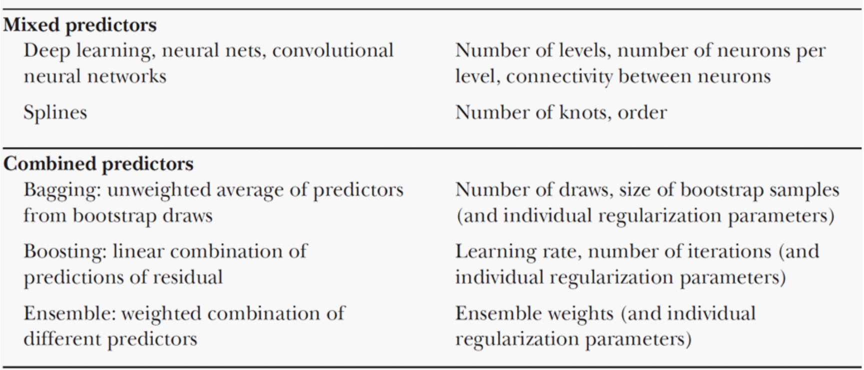


Table 2

Some Machine Learning Algorithms

Function class $\mathcal{F}$ (and its parametrization)	*Regularizer $R(f)$*
Global/parametric predictors	
Linear $\beta'x$ (and generalizations)	Subset selection$\|\|\beta\|\|_0 = \sum_{j=1}^{k} 1_{\beta_j \neq 0}$ LASSO $\|\|\beta\|\|_1 = \sum_{j=1}^{k} \|\beta_j\|$ Ridge $\|\|\beta\|\|_2^2 = \sum_{j=1}^{k} \beta_j^2$ Elastic net $\alpha\|\|\beta\|\|_1 + (1-\alpha)\|\|\beta\|\|_2^2$
Local/nonparametric predictors	
Decision/regression trees	Depth, number of nodes/leaves, minimal leaf size, information gain at splits
Random forest (linear combination of trees)	Number of trees, number of variables used in each tree, size of bootstrap sample, complexity of trees (see above)
Nearest neighbors	Number of neighbors
Kernel regression	Kernel bandwidth

Table 2 (continued)



Mixed predictors	
Deep learning, neural nets, convolutional neural networks	Number of levels, number of neurons per level, connectivity between neurons
Splines	Number of knots, order
Combined predictors	
Bagging: unweighted average of predictors from bootstrap draws	Number of draws, size of bootstrap samples (and individual regularization parameters)
Boosting: linear combination of predictions of residual	Learning rate, number of iterations (and individual regularization parameters)
Ensemble: weighted combination of different predictors	Ensemble weights (and individual regularization parameters)

1.2 Predict housing prices

- y is log house price in U.S. 2011
    - $n = 51,808$ is sample size
    - $p = 150$ is number of potential regressors.
- Predict using
    - OLS (using all regressors)
    - regression tree
    - LASSO
    - random forest
    - ensemble: an optimal weighted average of the above methods.
- 1. Train model on 10,000 observations using 8-fold CV.
- 2. Fit preferred model on these 10,000 observations.
- 3. Predict on remaining 41,808 observations
    - and do 500 bootstraps to get 95% CI for R^2.

- Random forest (and subsequent ensemble) does best out of sample.

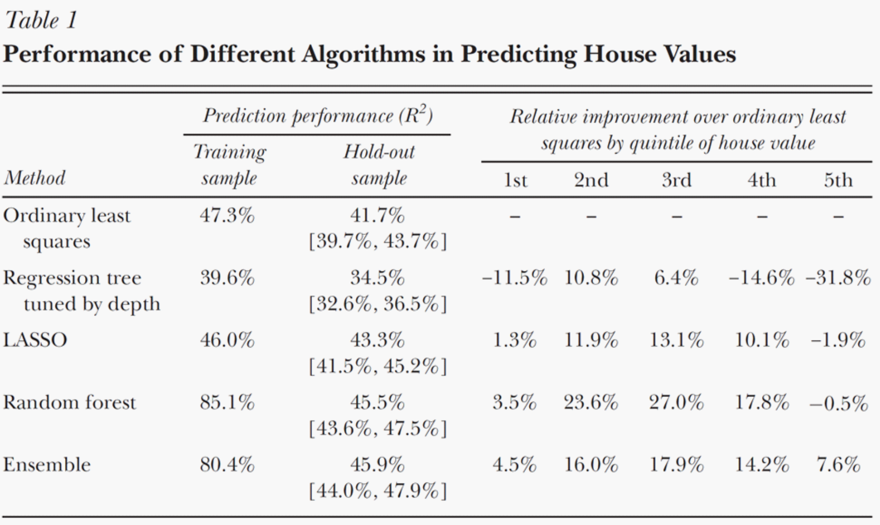


Table 1

Performance of Different Algorithms in Predicting House Values

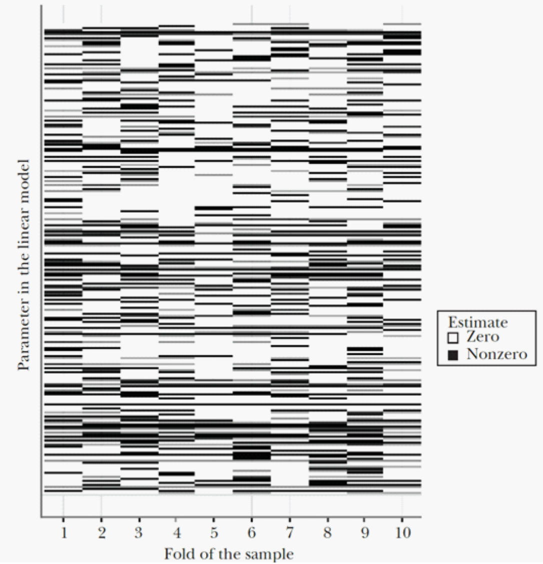
	Prediction performance (R^2)		*Relative improvement over ordinary least squares by quintile of house value*				
Method	*Training sample*	*Hold-out sample*	1st	2nd	3rd	4th	5th
Ordinary least squares	47.3%	41.7% [39.7%, 43.7%]	–	–	–	–	–
Regression tree tuned by depth	39.6%	34.5% [32.6%, 36.5%]	–11.5%	10.8%	6.4%	–14.6%	–31.8%
LASSO	46.0%	43.3% [41.5%, 45.2%]	1.3%	11.9%	13.1%	10.1%	–1.9%
Random forest	85.1%	45.5% [43.6%, 47.5%]	3.5%	23.6%	27.0%	17.8%	–0.5%
Ensemble	80.4%	45.9% [44.0%, 47.9%]	4.5%	16.0%	17.9%	14.2%	7.6%

Further details

- Downloadable appendix to the paper gives more details and R code.
- 1. Divide into training and hold-out sample.
- 2. On the training sample do 8-fold cross-validation to get tuning parameter(s) such as λ.
  - If e.g. two tuning parameters then do two-dimensional grid search.
- 3. The prediction function $\widehat{f}(x)$ is estimated using the entire sample with optimal λ.
- 4. Now apply this $\widehat{f}(x)$ to the hold-out sample and can compute R^2 and MSE.
- 5. A 95% CI for R^2 can be obtained by bootstrapping hold-out sample.
- Ensemble weights are obtained by 8-fold CV in the training sample.

LASSO

- LASSO does not pick the "correct" regressors
  - it just gets the correct $\widehat{f}(x)$ especially when regressors are correlated with each other.
- Diagram on next slide shows which of the 150 variables are included in separate models for 10 subsamples
  - there are many variables that appear sometimes but not at other times
    - appearing sometimes in white and sometimes in black.


Parameter in the linear model
Fold of the sample
1
2
3
4
5
6
7
8
9
10
Estimate
□ Zero
■ Nonzero

1.3 Some Thoughts on ML Prediction

- Clearly there are many decisions to make in implementation
  - how are features converted into x's
  - tuning parameter values
  - which ML method to use
  - even more with an ensemble forecast.
- For commercial use this may not matter
  - all that matters is that predict well enough.
- But for published research we want reproducibility
  - At the very least document exactly what you did
  - provide all code (and data if it is publicly available)
  - keep this in mind at the time you are doing the project.
- For public policy we prefer some understanding of the black box
  - this may be impossible.

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 a 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 → 1 as $n \to \infty$.

Add Some Structure

- A partially linear control function model specifies

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

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

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

- Consistent estimator requires $E[y|x_1, \mathbf{x}_2] = \beta x_1 + g(\mathbf{x}_2)$.
  - 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 $\boldsymbol{\beta}$ were high dimension and $\mathbf{x}_2$ low dimension
  - now for causal ML x_1 and β 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 $\widehat{\beta}$
    - ★ so data mining has not affected the distribution of $\widehat{\beta}$.
  - The methods can extend to endogenous x_1.
- We focus on use of the LASSO to determine $g(\mathbf{x}_2)$
  - due to Belloni, Chernozhukov and Hansen and coauthors
  - assumptions including "sparsity" enable use of standard inference for $\widehat{\beta}_1$
  - use the `lassopack` Stata package of Ahrens, Hansen and Schaffer (2019).

Alternatively estimate average partial effects

- An alternative to the partially lienr 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.
- In fact better methods than matching methods are used.

3. LASSO for causal homogeneous effects

- Belloni, Chernozhukov and Hansen and coauthors have many papers
  - focus on the following three papers.
- Belloni, Chernozhukov and Hansen (2014), "High-dimensional methods and inference on structural and treatment effects," *Journal of Economic Perspectives*, Spring, 29-50
  - accessible paper with three applications.
- Ahrens, Hansen and Schaffer (2019), "lassopack: Model selection and prediction with regularized regression in Stata," arXiv:1901.05397
  - more detail on LASSO methods as well as on Stata commands.
- Belloni, Chernozhukov and Hansen (2011), "Inference Methods for High-Dimensional Sparse Econometric Models," *Advances in Economics and Econometrics*, ES World Congress 2010, ArXiv 2011
  - even more detail and summarizes several of their subsequently published papers.

3.1 Standard Lasso

- Begin with basic LASSO before use for causal effects.
- Consider a variant of LASSO with variable weights
  - useful for extension to heteroskedastic and clustered errors.
- The **LASSO estimator** $\widehat{\boldsymbol{\beta}}_\lambda$ of $\boldsymbol{\beta}$ minimizes

$$Q_\lambda(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \frac{\lambda}{n}\sum_{j=1}^p \psi_j|\beta_j|$$

  - where y_i and x_{ij} are demeaned so $\bar{y} = 0$ and $\bar{x}_j = 0$
  - and $\lambda \geq 0$ is a tuning parameter to be determined.
- For homoskedastic errors $\psi_j = \sqrt{\frac{1}{n}\sum_{i=1}^n x_{ij}^2}$
  - this is the same as fixed weight $(\psi_j = 1)$ LASSO on standardized x_{ij}.
- For heteroskedastic errors $\psi_j = \sqrt{\frac{1}{n}\sum_{i=1}^n x_{ij}^2 \widehat{u}_{i,0}^2}$
  - **rlasso** obtains $\widehat{u}_{i,0}$ from OLS of y on the five x_j most correlated with y.
- For errors clustered within t, $\psi_j = \sqrt{\frac{1}{nT}\sum_{i=1}^n \widehat{u}_{ij}^2}$ where $\widehat{u}_{ij} = \sum_{t=1}^T x_{ijt}\widehat{u}_{it,0}^2$.

Determination of tuning parameter lambda

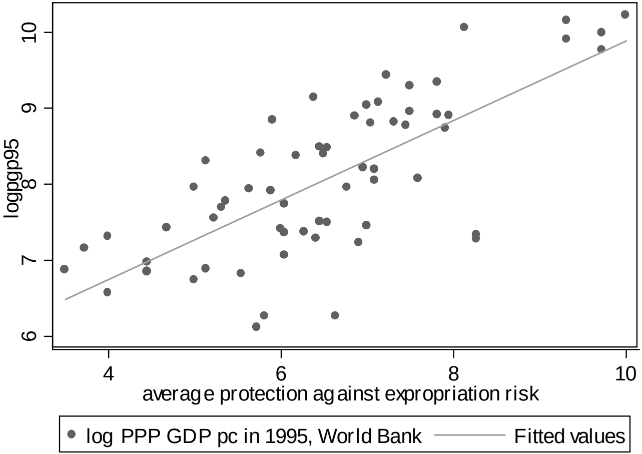
- K-fold cross-validation
  - `cvlasso` command in `lassopack` package.
- Penalized goodness-of-fit (AIC, BIC, AICC, EBIC)
  - `lasso2` command in `lassopack` package.
- User-specified value "theory-driven" or "rigorous"
  - `rlasso` command in `lassopack` package.
- For `rlasso` the theory is asymptotic that gives appropriate rates but entails specification of two constants c and γ
  - homoskedastic: $\lambda = 2c\sigma\sqrt{n}\Phi^{-1}(1-\gamma/2p)$ where $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_{i,0}^2$.
  - heteroskedastic: $\lambda = 2c\sqrt{n}\Phi^{-1}(1-\gamma/2p)$
  - defaults are $c = 1.1$ and $\gamma = 0.1/\log(n)$
    - ★ these defaults can be changed.

3.2 Example

- Example from Acemoglu, Johnson and Robinson (2001).
- Relationship between
  - $y =$ `logpgp95` (log GDP per capita in 1995 at PPP)
  - $x_1 =$ `avexpr` (average protection against expropriation risk)
- When `avexpr` is treated as endogenous use as instrument
  - $z =$ `logem4` (log settler mortality)
- Theory: better institutions (`avexpr`) lead to higher GDP
  - but causation could be other way
  - so instrument `avexpr` with log settler mortality (`logem4`)
    - initial settlers would invest in institutions if they thought they'd survive.

- $n = 64$ with $p = 24$ potential controls
  - latitude, ethnicity, temperature, humidity, ...
- These slides
  - first straight LASSO to predict `logpgp95`
  - then control function with exogenous `avexpr` and LASSO to get controls
  - then endogenous `avexpr` with instrument `logem4`.

Scatterplot of y on x1



```
variable name   type    format     label      variable label
---------------------------------------------------------------------------------
logpgp95        float   %9.0g                 log PPP GDP pc in 1995, World Bank
avexpr          float   %9.0g                 average protection against
                                                expropriation risk
logem4          float   %9.0g                 log settler mortality
lat_abst        float   %9.0g                 Abs(latitude of capital)/90
edes1975        float   %9.0g                 % of European descent in 1975
avelf           float   %9.0g                 ethno fract avg 5indic east_lev
temp1           float   %9.0g                 first of 5 temperature indicators
temp2           float   %9.0g
temp3           float   %9.0g
temp4           float   %9.0g
temp5           float   %9.0g
humid1          float   %9.0g                 first of four humidity indicators
humid2          float   %9.0g
humid3          float   %9.0g
humid4          float   %9.0g
steplow         float   %9.0g                 first of six soil indicators
deslow          float   %9.0g
stepmid         float   %9.0g
desmid          float   %9.0g
drystep         float   %9.0g
drywint         float   %9.0g
landlock        float   %9.0g                 =1 if landlocked
goldm           float   %9.0g                 first of five mineral indicators
iron            float   %9.0g
silv            float   %9.0g
zinc            float   %9.0g
oilres          float   %9.0g
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    logpgp95 |         64    8.062237    1.043359   6.109248   10.21574
      avexpr |         64    6.515625    1.468647        3.5         10
      logem4 |         64    4.657031    1.257984   2.145931   7.986165
    lat_abst |         64    .1811028    .1326669          0   .6666667
    edes1975 |         64    18.06719    29.62672          0        100
       avelf |         64    .4100421    .3148905          0   .8902469
       temp1 |         64    22.85938    5.206428          4         29
       temp2 |         64    29.26563    5.237963         12         40
       temp3 |         64    37.96875    5.371792         24         48
       temp4 |         64     6.46875    10.28941        -37         20
       temp5 |         64    15.64063    6.338011          1         24
      humid1 |         64    72.14063    17.89818         18         97
      humid2 |         64    87.20313    8.138054         54         98
      humid3 |         64      50.375    16.75358         10         78
      humid4 |         64    70.65625    9.229385         41         92
     steplow |         64      .28125    .4531635          0          1
      deslow |         64      .21875    .4166667          0          1
     stepmid |         64      .03125    .1753681          0          1
      desmid |         64     .015625        .125          0          1
     drystep |         64      .09375    .2937848          0          1
     drywint |         64     .015625        .125          0          1
    landlock |         64      .09375    .2937848          0          1
       goldm |         64    1.046875    6.027529          0         47
        iron |         64    .6265625    2.393393          0         16
        silv |         64     .859375    3.017763          0         13
        zinc |         64      .96875     3.03926          0         15
      oilres |         64    105852.2    394573.8          0    3040000
```

LASSO of y on x1 and x2

- Here just LASSO of y on all $x's$
  - Post-LASSO is from OLS on the three selected variables.
- rlasso results

```
. // Straight lasso using rlasso with default c and gamma
. // Picks three regressors - aveexpr edes1975 avelf
. rlasso logpgp95 avexpr $x2list

──────────────────────────────────────────────────────
            Selected │           Lasso   Post-est OLS
─────────────────────┼────────────────────────────────
              avexpr │       0.2735483       0.3866856
            edes1975 │       0.0057595       0.0087536
               avelf │      -0.4630296      -1.0130167
               _cons │ *     6.3657022       5.7999650
──────────────────────────────────────────────────────
*Not penalized
```

- cvlasso with 5 folds and seed(10101) picks four regressors
  - avexpr edes1975 avelf humid3
- lasso2 with AIC penalty picks nine regressors
  - preceding four + temp5 steplow deslow goldm silv

Theory

- Belloni, Chernozhukov and Hansen (2011)
- Data generating process
  - $y_i = f(\mathbf{x}_i) + u_i$ where $u_i \sim N(0, \sigma^2)$
  - i.i.d. normal error for simplicity.
- Model for the conditional mean
  - $f(\mathbf{x}_i) = \mathbf{w}_i' \boldsymbol{\beta}_0 + r_i$ where $\mathbf{w}_i$ is $p \times 1$
  - the $\mathbf{w}_i's$ are (demeaned) transformations of $\mathbf{x}_i$.
- 1. Large p: There are many $\mathbf{w}_i's$.
- 2. Sparsity: $\boldsymbol{\beta}_0$ has at most s non-zero terms with $s << n$.
- 3. Approximation error r_i declines with n
  - $\sqrt{E[r_{ij}^2]} \leq k\sigma\sqrt{s/n}$ for some fixed k.

3.3 Control function and exogenous regressor

- Belloni, Chernozhukov and Hansen, REStud (2014).
- Estimate β and choose controls in partially linear model
  - $y = \beta_0 x_1 + g(\mathbf{x}_2) + u$, $E[u|x_1, \mathbf{x}_2] = 0$
  - $x_1 = m(\mathbf{x}_2) + v$, $E[v|\mathbf{x}_2] = 0$.
- Approximate by (where $\mathbf{w}$ is rich transformations of $\mathbf{x}_2$)
  - $y = \beta_0 x_1 + \mathbf{w}'\delta_0 + r_g + u$
  - $x_1 = \mathbf{w}'\boldsymbol{\pi}_0 + r_m + v$.
- Assume that
  - sparsity: δ_0 and $\boldsymbol{\pi}_0$ have at most $s = s_n << n$ non-zero entries
    - ★ and $s^2 \log^2\{\min(p, n)\}/n \leq \delta_n \rightarrow 0$
    - ★ (if correct variables were known we need $s^2/n \rightarrow 0$)
  - Approximation error r_i declines with n
    - ★ $\sqrt{\bar{E}[r_{gi}^2]} \leq c\sqrt{s/n}$ and $\sqrt{\bar{E}[r_{mi}^2]} \leq c\sqrt{s/n}$ some constant c.

Control function and exogenous regressor (continued)

- The reduced forms (subbing out x_1 in the y_1 equation are)
  - $y = \mathbf{w}'\overline{\delta}_0 + \bar{r}_g + \bar{u}$
  - $x_1 = \mathbf{w}'\boldsymbol{\pi}_0 + r_m + v$
- Double selection method
  - LASSO selects controls separately in the y and x_1 reduced forms
  - do OLS of y on x_1 and the union of selected controls
    - ★ double selection reduces the chance of omitted variables bias
    - ★ and can use standard inference (given the assumptions made).
- We obtain
  - 1. LASSO of y on $\mathbf{w}$
    - ★ selected `edes1975 avelf`
  - 2. LASSO of x_1 on $\mathbf{w}$
    - ★ selected `edes1975 zinc`
- Post Lasso OLS of y on x_1 with controls those in union of 1. and 2.
  - `edes1975 avelf zinc`

First half of output

```
. // Basic usage: select from high-dim controls. OLS control function.
. // Expect same result as rlasso above
. // as using same defaults and rlasso included aveexpr
. pdslasso logpgp95 avexpr ($x2list)
1.  (PDS/CHS) Selecting HD controls for dep var logpgp95...
Selected: edes1975 avelf
2.  (PDS/CHS) Selecting HD controls for exog regressor avexpr...
Selected: edes1975 zinc


Estimation results:

Specification:
Regularization method:                 lasso
Penalty loadings:                      homoskedastic
Number of observations:                64
Exogenous (1):                         avexpr
High-dim controls (24):                lat_abst edes1975 avelf temp1 temp2 temp3
                                       temp4 temp5 humid1 humid2 humid3 humid4
                                       steplow deslow stepmid desmid drystep
                                       drywint landlock goldm iron silv zinc
                                       oilres
Selected controls (3):                 edes1975 avelf zinc
Unpenalized controls (1):              _cons
```

Second half of output

- OLS was $\widehat{\beta} = 0.372$ and $se(\widehat{\beta}) = 0.064$.
- Note: inference only valid on $\widehat{\beta}_{Postlasso}$

```
Structural equation:

OLS using CHS lasso-orthogonalized vars
------------------------------------------------------------------------------
     logpgp95 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avexpr |   .4262511   .0540552     7.89   0.000     .3203049    .5321974
------------------------------------------------------------------------------

OLS using CHS post-lasso-orthogonalized vars
------------------------------------------------------------------------------
     logpgp95 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avexpr |    .391257   .0574894     6.81   0.000     .2785799     .503934
------------------------------------------------------------------------------

OLS with PDS-selected variables and full regressor set
------------------------------------------------------------------------------
     logpgp95 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avexpr |   .3913455   .0561862     6.97   0.000     .2812225    .5014684
    edes1975 |   .0091289    .003184     2.87   0.004     .0028883    .0153694
       avelf |  -.9974943   .2474453    -4.03   0.000    -1.482478   -.5125104
        zinc |  -.0079226   .0280604    -0.28   0.778    -.0629201    .0470748
       _cons |   5.764133   .3773706    15.27   0.000     5.024501    6.503766
------------------------------------------------------------------------------
Standard errors and test statistics valid for the following variables only:
    avexpr
------------------------------------------------------------------------------
```

JEP (2014) Application

- Effect of abortion policy on crime (Donohue and Levitt)
  - y_{st} = crime rate (Violent, property or murder)
  - d_{st} = abortion rate for state i at time t ($n = 50,\ T = 12$)
  - $y_{st} = \beta d_{st} + \mathbf{x}_{st}'\boldsymbol{\delta} + \delta_s + \gamma_t + \varepsilon_{st}$
- Analyze first-differences with state FE's model

  $\Delta y_{st} = \beta \Delta d_{st} + \Delta \mathbf{x}_{st}'\boldsymbol{\delta} + \gamma_s + u_{st}$

  $\Delta d_{st} = \Delta \mathbf{x}_{st}'\boldsymbol{\pi} + \lambda_s + v_{st}$
  - $p = 284$ possible variables in $\mathbf{x}$ due to interactions (see paper)
- Find $s = 10$ with chosen controls being
  - lagged prisoners, lagged police×t, initial income difference, initial income difference×t, initial beer consumption difference×t, average income, average income×t, initial abortion rate.
- Find similar $\widehat{\alpha}$ to Donohue-Levitt who have many more controls
  - but post-Lasso OLS standard errors are one-third the size!

3.4 Endogenous regressor

- Consider IV estimation with many instruments of the model
  - $y = \beta x_1 + \mathbf{x}_2'\delta + u$ where x_1 (scalar for simplicity) is endogenous.
- In theory we should use all instruments
  - as the asymptotic efficiency of IV improves with more instruments.
- But **asymptotic theory works poorly** if include too many instruments.
  - finite sample bias of IV may not disappear even in large samples
  - and standard hypothesis tests have the wrong size.
- Use the LASSO to pick just a few of the potential instruments
  - assume sparsity: only a few of the potential instruments are valid.

How can we have many instruments?

- Case 1: there are naturally many instruments
  - often due to economic theory such as a no arbitrage condition.
- Case 2: there is a single instrument z
  - but the optimal instrument need not be z
  - in the i.i.d. case the optimal instrument for x_1 is $E[x_1|\text{exogenous variables}]$
  - so additional instruments such as powers and interactions may be better.
- Case 3: $y = \beta x_1 + g(\mathbf{x}_2) + u$
  - we don't know $g(\mathbf{x}_2)$
  - so we use $g(\mathbf{x}_2) = \mathbf{w}'\delta$ as in the exogenous x_1 case
  - so z and all $\mathbf{w}$ may be considered as instruments for x_1.

First part of output

```
. // Select controls; specify that logem4 is an unpenalized instrument ///
. //   to be partialled out.
. ivlasso logpgp95 (avexpr=logem4) ($x2list), partial(logem4)
1.  (PDS/CHS) Selecting HD controls for dep var logpgp95...
Selected: lat_abst edes1975 temp3 humid2 humid3
3.  (PDS) Selecting HD controls for endog regressor avexpr...
Selected: lat_abst edes1975 temp3 humid1 humid2 humid3 humid4 iron zinc
4.  (PDS) Selecting HD controls for IV logem4...
Selected: avelf temp2 temp5 humid2
5.  (CHS) Selecting HD controls and IVs for endog regressor avexpr...
Selected:
Also inc: logem4
6a. (CHS) Selecting lasso HD controls and creating optimal IV for endog regressor
> avexpr...
Selected: lat_abst edes1975 temp3 humid2 humid3
6b. (CHS) Selecting post-lasso HD controls and creating optimal IV for endog regre
> ssor avexpr...
Selected: lat_abst edes1975 temp3 humid2 humid3
7.  (CHS) Creating orthogonalized endogenous regressor avexpr...
```

Last part of output

- Exogenous case had $\widehat{\beta}_{\text{avexpr}} = 0.39$ with $se = 0.056$.

```
IV with PDS-selected variables and full regressor set
------------------------------------------------------------------------------
    logpgp95 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avexpr |    .765767   .2020636     3.79   0.000     .3697296    1.161804
    lat_abst |  -1.348873   1.369259    -0.99   0.325    -4.032571    1.334824
    edes1975 |   .0019131   .0049826     0.38   0.701    -.0078525    .0116788
       avelf |  -1.131221    .337874    -3.35   0.001    -1.793442   -.4690002
       temp2 |  -.0249502   .0326909    -0.76   0.445    -.0890232    .0391229
       temp3 |    .021731    .026079     0.83   0.405    -.0293829     .072845
       temp5 |  -.0183963   .0227653    -0.81   0.419    -.0630154    .0262228
      humid1 |  -.0432292   .0184622    -2.34   0.019    -.0794144   -.007044
      humid2 |   .0651965   .0290922     2.24   0.025     .0081767    .1222162
      humid3 |   .0458411   .0170148     2.69   0.007     .0124928    .0791895
      humid4 |   -.043253   .0205977    -2.10   0.036    -.0836239   -.0028822
        iron |  -.1437898   .0835631    -1.72   0.085    -.3075705    .0199909
        zinc |   .0531055   .0584786     0.91   0.364    -.0615104    .1677214
       _cons |   2.157911    2.00102     1.08   0.281    -1.764016    6.079839
------------------------------------------------------------------------------
Standard errors and test statistics valid for the following variables only:
    avexpr
------------------------------------------------------------------------------
```

JEP (2014) Application

- Effect of endogenous court decisions on house prices
  - y_{ct} = home price index within court circuit c at time t
  - d_{ct} = # of takings appellate decisions that rule that a taking was unlawful
  - $y_{ct} = \alpha_c + \alpha_t + \alpha_c t + \beta d_{ct} + \mathbf{w}'_{ct}\delta + \varepsilon_{ct}$
- Frisch-Waugh partial out fixed effects, time trends and $\mathbf{w}_{ct}$
  - $\widetilde{y}_{ct} = \alpha + \beta\widetilde{d}_{ct} + error$
  - $p = 183$ possible instruments (due to interactions)
- Find $s = 1$ (JEP survey paper) or $s = 2$ (Econometrica paper)
  - the JEP instrument is the square of the number of panels with one or more members with JD from a public university.

3.5 Caution

- The LASSO methods are easy to estimate using the `lassopack` program
  - they'll be (blindly) used a lot.
- However in any application
  - is the underlying assumption of sparsity reasonable?
  - has the asymptotic theory kicked in?
  - are the default values of c and γ reasonable?

4. ATE with heterogeneous effects

- Consider the effect of treatment d on an outcome y
  - where individuals may self-select into treatment.
- The preceding control function approach assumes that $E[u|d, \mathbf{x}_2]$ in the model

$$y = \beta d + \mathbf{x}_2'\delta + u$$

  - an untestable unconfoundedness assumption
  - or if d is still endogenous that we can do IV.
- For binary treatment the heterogeneous effects model is more flexible
  - and hence more plausible in controlling for self-selection.

4.1 Heterogeneous effects model

- 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 $d \in \{0, 1, ..., J\}$.
- Denote 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 $y_i^{(0)}$.
- The goal is to estimate the average treatment effect
  - ATE$= E[y_i^{(1)} - y_i^{(0)}]$
- Or the conditional treatment effect given $\mathbf{x}$
  - $\tau(\mathbf{x}) = E[y_i^{(1)} - y_i^{(0)}|\mathbf{x}]$
- The key assumption is the conditional independence assumption
  - $d_i \perp \{y_i^{(0)}, y_i^{(1)}\}|\mathbf{x}_i$.
  - conditional on $\mathbf{x}$, treatment is independent of the potential outcome.

4.2 ATE estimated using regression adjustment

- 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 $\tau(\mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$.
- Use machine learning methods such as LASSO to get $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$
  - $\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}_1(\mathbf{x}_i) - \frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}_0(\mathbf{x}_i)$.
- Problem: this does not take into account any correlation of regressors with the treatment variable d_i.

4.3 ATE estimated using propensity scores

- Define the propensity score $p(\mathbf{x}) = \Pr[d = 1|\mathbf{x}]$.
- Under the conditional independence assumption
  - $\mu_1(\mathbf{x}) = E[y^{(1)}|\mathbf{x}] = E[\frac{d}{p(\mathbf{x})}y|\mathbf{x}]$
  - $\mu_0(\mathbf{x}) = E[y^{(0)}|\mathbf{x}] = E[\frac{1-d}{1-p(\mathbf{x})}y|\mathbf{x}]$
  - $\tau(\mathbf{x}) = E[y^{(1)} - y^{(0)}|\mathbf{x}] = E\left[\left(\frac{d}{p(\mathbf{x})} - \frac{1-d}{1-p(\mathbf{x})}\right)y|\mathbf{x}\right]$
- So use $\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i}{\widehat{p}_i(\mathbf{x})}y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{1-d_i}{1-\widehat{p}(\mathbf{x}_i)}y_i$
  - and use e.g. LASSO for $\widehat{p}_i(\mathbf{x})$.
- The conditional independence assumption is more plausible the more $\mathbf{x}'s$ considered.
- Aside:
  $E_{d,y}\left[\frac{d}{p(\mathbf{x})}y|\mathbf{x}\right] = E_{d,y}\left[\frac{dy_1}{p(\mathbf{x})}|\mathbf{x}\right] = E_d\left[\frac{d}{p(\mathbf{x})}|\mathbf{x}\right] \times E_y\left[y_1|\mathbf{x}\right] = E_y\left[y_1|\mathbf{x}\right]$
  - second last equality uses conditional independence
  - last equality uses $E_d\left[d|\mathbf{x}\right] = \Pr[d = 1|\mathbf{x}] = p(\mathbf{x})$.

4.4 ATE estimated using doubly-robust method

- 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.
- As before $\mu_1(\mathbf{x}) = E[y^{(1)}|\mathbf{x}]$, $\mu_0(\mathbf{x}) = E[y^{(0)}|\mathbf{x}]$ and $p(\mathbf{x}) = \Pr[d = 1|\mathbf{x}]$.
- Farrell uses the doubly-robust method
  - $\widehat{\text{ATE}} = \widehat{\mu}_1 - \widehat{\mu}_0$ where $\mu_1 = E[y^{(1)}]$ and $\mu_0 = E[y^{(0)}]$
  - $\widehat{\mu}_j = \frac{1}{n}\sum_{i=1}^n \left\{ \frac{1[d_i=j](y_i - \widehat{\mu}_j(\mathbf{x}_i))}{\widehat{p}_j(\mathbf{x}_i)} + \widehat{\mu}_j(\mathbf{x}_i) \right\}$ for $j = 0, 1$
    - where $\widehat{p}_1(\mathbf{x}_i) = \widehat{p}(\mathbf{x}_i)$ and $\widehat{p}_0(\mathbf{x}_i) = 1 - \widehat{p}_1(\mathbf{x}_i)$
  - doubly-robust as estimator remains consistent if either
    - the propensity score model $p(\mathbf{x})$ or
    - the regression imputation model $\mu_j(x)$ is misspecified.
- The LASSO is used to obtain $\widehat{p}(\mathbf{x})$ and $\widehat{\mu}_1(x)$ and $\widehat{\mu}_2(x)$.
  - simulation and apply to Dehejia-Wahba data.

4.5 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.
    - more on this in section on double/debiased machine learning.

4.6 Heterogeneous Effects using Random Forests

- 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.

Heterogeneous Effects using Random Forests (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

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

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

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

  - 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.

4.7 Treatment Effects using Deep Neural Networks

- Max Farrell, Tengyuan Liang and Sanjog Misra (2018), "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.

5. Double or Debiased Machine Learning

- Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2018), "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*.
- Interest lies in estimation of key parameter(s) controlling for high-dimensional nuisance parameters.
- Two components to double ML or debiased ML and subsequent inference
  - Work with orthogonalized moment conditions to allow consistent estimation of parameter(s) of interest.
  - Use sample splitting (cross fitting) to remove bias induced by overfitting.

Double or Debiased Machine Learning (continued)

- Then get asymptotic normal confidence intervals for parameters of interest
  - where a variety of ML methods can be used
    - random forests, lasso, ridge, deep neural nets, boosted trees, ensembles
  - that don't necessarily need sparsity
  - and theory does not require Donsker properties
- Can apply to
  - partial linear model (with exogenous or endogenous regressor)
    - orthogonality conditions are presented below
  - ATE and ATET under unconfoundedness
    - orthogonality conditions are presented below for ATE
  - LATE in an IV setting.

5.1 Partially linear model

- Consider partially linear model

$$
\begin{aligned}
y &= \beta x_1 + g(\mathbf{x}_2) + u \\
x_1 &= m(\mathbf{x}_2) + v
\end{aligned}
$$

- Naive approach is
  - use ML method such as LASSO to get $\widehat{\beta} x_1 + \widehat{g}(\mathbf{x}_2)$ in a training sample
  - then compute $\widetilde{\beta}$ from regress y on x_1 and $\widehat{g}(\mathbf{x}_2)$ in a different sample
  - but can show that the bias in $\widehat{g}(\mathbf{x}_2)$ leads to bias in $\widetilde{\beta}$.
- Instead partial out the effect of $\mathbf{x}_2$ on x_1 (an example of orthogonalization)
  - use ML method for regress x_1 on $\mathbf{x}_2$ and form $\widehat{m}(\mathbf{x}_2)$ in the training sample
  - then $\widetilde{\beta}$ is coefficient in OLS of $(y - \widehat{g}(\mathbf{x}_2))$ on $(x_1 - \widehat{m}(\mathbf{x}_2))$ in a different sample
  - the distribution of $\sqrt{n}(\widetilde{\beta} - \beta)$ involves a multiple of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\widehat{m}(\mathbf{x}_2) - m(\mathbf{x}_2))(\widehat{g}(\mathbf{x}_2) - g(\mathbf{x}_2))$ that disappears
  - the ML errors disappear as they appear as a product.

5.2 Orthogonalization defined

- Define $\boldsymbol{\beta}$ as parameters of interest and $\boldsymbol{\eta}$ as nuisance parameters.
- Estimate $\widehat{\boldsymbol{\beta}}$ is obtained following first step estimate $\widehat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$
  - First stage: $\widehat{\boldsymbol{\eta}}$ solves $\sum_{i=1}^{n} \omega(\mathbf{w}_i, \boldsymbol{\eta}) = \mathbf{0}$
  - Second stage: $\widehat{\boldsymbol{\beta}}$ solves $\sum_{i=1}^{n} \psi(\mathbf{w}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\eta}}) = \mathbf{0}$.
- The distribution of $\widehat{\boldsymbol{\beta}}$ is usually affected by the noise due to estimating $\boldsymbol{\eta}$
  - e.g. Heckman's two-step estimator in selection models.
- But this is not always the case
  - e.g. the asymptotic distribution of feasible GLS is not affected by first-stage estimation of variance model parameters to get $\widehat{\Omega}$.
- Result: The distribution of $\widehat{\boldsymbol{\beta}}$ is unaffected by first-step estimation of $\boldsymbol{\eta}$ if the function $g(\cdot)$ satisfies
  - $E[\partial\psi(\mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\eta})/\partial\boldsymbol{\eta}] = \mathbf{0}$; see next slide.
- So choose functions $\psi(\cdot)$ that satisfy the orthogonalization condition

$$E[\partial\psi(\mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\eta})/\partial\boldsymbol{\eta}] = \mathbf{0}.$$

Orthogonalization (continued)

- Why does this work?

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(\mathbf{w}_i,\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\eta}})\\
= \quad &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(\mathbf{w}_i,\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)+\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial\psi(\mathbf{w}_i,\boldsymbol{\beta},\boldsymbol{\eta})}{\partial\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}_0,\boldsymbol{\eta}_0}\times\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial\psi(\mathbf{w}_i,\boldsymbol{\beta},\boldsymbol{\eta})}{\partial\boldsymbol{\eta}'}\right|_{\boldsymbol{\beta}_0,\boldsymbol{\eta}_0}\times\sqrt{n}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)
\end{aligned}
$$

- By a law of large numbers $\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial\psi(\mathbf{w}_i,\boldsymbol{\beta},\boldsymbol{\eta})}{\partial\boldsymbol{\eta}}\right|_{\boldsymbol{\beta}_0,\boldsymbol{\eta}_0}$ converges to its expected value which is zero if $E[\partial\psi(\mathbf{w}_i,\boldsymbol{\beta},\boldsymbol{\eta})/\partial\boldsymbol{\eta}]=\mathbf{0}$.
- So the term involving $\widehat{\boldsymbol{\eta}}$ drops out.
- For more detail see Cameron and Trivedi (2005, p.201).

5.3 Orthogonalization in partially linear model: method 1

- Consider the partially linear model and manipulate

$$\begin{array}{rrcll} & y & = & \beta x_1 + g(\mathbf{x}_2) + u & \text{where } E[u|x_1, \mathbf{x}_2] = 0 \\ \Rightarrow & E[y|\mathbf{x}_2] & = & \beta E[x_1|\mathbf{x}_2] + g(\mathbf{x}_2) + u & \text{as } E[u|\mathbf{x}_2] = 0 \\ & y - E[y|\mathbf{x}_2] & = & \beta(x_1 - E[x_1|\mathbf{x}_2]) + u & \text{subtracting} \end{array}$$

- Robinson (1988) differencing estimator
  - use kernel methods to get $\widehat{E}[y|\mathbf{x}_2]$ and $\widehat{E}[x_1|\mathbf{x}_2]$
  - $\widehat{\beta}$ from OLS regress $(y - \widehat{E}[y|\mathbf{x}_2])$ on $(x_1 - \widehat{E}[x_1|\mathbf{x}_2])$
- Instead here use machine learning methods for $\widehat{E}[y|\mathbf{x}_2]$ and $\widehat{E}[x_1|\mathbf{x}_2]$.
- Recall that OLS of y on $\mathbf{x}$ has f.o.c. $\sum_i x_i u_i = 0$
  - so is sample analog of population moment condition $E[xu] = 0$.
- Robinson estimator therefore solves population moment condition
  - $E[(x_1 - E[x_1|\mathbf{x}_2])\{y - E[y|\mathbf{x}_2] - (x_1 - E[x_1|\mathbf{x}_2])\beta\}] = 0.$

Orthogonalization in partially linear model (continued)

- Robinson estimator solves population moment condition $E[\psi(\cdot)] = 0$ where
  - $\psi(\cdot) = (x_1 - E[x_1|\mathbf{x}_2])\{y - E[y|\mathbf{x}_2] - (x_1 - E[x_1|\mathbf{x}_2])\beta\}$.
- Define $\eta_1 = E[x_1|\mathbf{x}_2]$ and $\eta_2 = E[y|\mathbf{x}_2]$, so
  - $\psi(w, \beta, \boldsymbol{\eta}) = (x_1 - \eta_1))\{y - \eta_2 - (x_1 - \eta_1)\beta\}$
- This satisfies the orthogonalization condition
  - $E[\partial\psi(\mathbf{w}, \beta, \boldsymbol{\eta})/\partial\eta_1] = E[2(x_1 - \eta_1)\beta] = 0$ as $\eta_1 = E[x_1|\mathbf{x}_2]$
  - $E[\partial\psi(\mathbf{w}, \beta, \boldsymbol{\eta})/\partial\eta_2] = E[-(x_1 - \eta_1)] = 0$ as $\eta_1 = E[x_1|\mathbf{x}_2]$.

Orthogonalization in partially linear model: method 2

- Again $y = \beta x_1 + g(\mathbf{x}_2) + u$ and $x_1 = m(\mathbf{x}_2) + v$.
- An alternative asymptotically equivalent method given earlier
  - $\widetilde{\beta}$ is coefficient in OLS of $(y - \widehat{g}(\mathbf{x}_2))$ on $(x_1 - \widehat{m}(\mathbf{x}_2))$.
- This estimator solves population moment condition
  - $E[\psi(\cdot)] = E[(x_1 - m(\mathbf{x}_2))\{y - g(\mathbf{x}_2) - (x_1 - m(\mathbf{x}_2)\beta\}] = 0.$
- Define $\eta_1 = m(\mathbf{x}_2) =$ and $\eta_2 = g(\mathbf{x}_2)$, so
  - $\psi(w, \beta, \boldsymbol{\eta}) = (x_1 - \eta_1))\{y - \eta_2 - (x_1 - \eta_1)\beta\}$
- This satisfies the orthogonalization condition
  - $E[\partial\psi(\mathbf{w}, \beta, \boldsymbol{\eta})/\partial\eta_1] = E[2(x_1 - \eta_1)\beta] = 0$ as $\eta_1 = E[x_1|\mathbf{x}_2]$
  - $E[\partial\psi(\mathbf{w}, \beta, \boldsymbol{\eta})/\partial\eta_2] = E[-(x_1 - \eta_1)] = 0$ as $\eta_1 = E[x_1|\mathbf{x}_2]$.
- Aside: method 1 equals method 2 asymptotically as
  - $E[y|\mathbf{x}_2] = E[x_1|\mathbf{x}_2]\beta + g(\mathbf{x}_2)$ and manipulate.
- Method 2 is derived in the paper using a general result on how to obtain a moment condition that satisfies the orthogonalization condition.

5.4 Orthogonalization for doubly robust ATE

- We used $\widehat{\text{ATE}} = \widehat{\tau} = \widehat{\mu}_1 - \widehat{\mu}_0$ where $\widehat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbf{1}[d_i = j](y_i - \widehat{\mu}_j(\mathbf{x}_i))}{\widehat{p}_j(\mathbf{x}_i)} + \widehat{\mu}_j(\mathbf{x}_i)\right\}$.
- The ATE solves the population moment condition

$$E\left[\frac{\mathbf{1}[d=1](y-\mu_1(\mathbf{x}))}{p(\mathbf{x})} - \mu_1(\mathbf{x}) - \frac{\mathbf{1}[d=0](y-\mu_o(\mathbf{x}_i))}{1-p(\mathbf{x})} + \mu_0(\mathbf{x}) + \tau\right] = 0.$$

- The parameter of interest is τ (β in above notation)
- Define the nuisance parameters
  - $\eta_1 = \mu_1(\mathbf{x}) = E[y_1|\mathbf{x}]$, $\eta_2 = \mu_0(\mathbf{x}) = E[y_0|\mathbf{x}]$ and $\eta_3 = p(\mathbf{x})$.
- Then $E[\psi(w, \tau, \boldsymbol{\eta})] = 0$ for
  - $\psi(w, \tau, \boldsymbol{\eta}) = \frac{\mathbf{1}[d=1](y-\eta_1)}{\eta_3} + \eta_1 - \frac{\mathbf{1}[d=0](y-\eta_2)}{1-\eta_3} - \eta_2 + \tau.$

Orthogonalization for doubly robust ATE (continued)

- We have $\psi(w,\tau,\boldsymbol{\eta})=\frac{\mathbf{1}[d=1](y-\eta_1)}{\eta_3}+\eta_1-\frac{\mathbf{1}[d=0](y-\eta_2)}{1-\eta_3}-\eta_2+\tau$.
- This satisfies the orthogonalization condition
  - $E[\partial\psi(\mathbf{w},\tau,\boldsymbol{\eta})/\partial\eta_1]=E[-\frac{\mathbf{1}[d=1])}{\eta_3}+1]=0$
    - as $E[\mathbf{1}[d=1]]=p(\mathbf{x})=\eta_3$
  - $E[\partial\psi(\mathbf{w},\tau,\boldsymbol{\eta})/\partial\eta_2]=E[\frac{\mathbf{1}[d=0])}{1-\eta_3}-1]=0$
    - as $E[\mathbf{1}[d=0]]=1-p(\mathbf{x})=1-\eta_3$
  - $E[\partial\psi(\mathbf{w},\tau,\boldsymbol{\eta})/\partial\eta_2]=E[-\frac{\mathbf{1}[d=1](y-\eta_1)}{\eta_3^2}-\frac{\mathbf{1}[d=0](y-\eta_2)}{(1-\eta_3)^2}]=0-0=0$
    - as $E[\mathbf{1}[d=1](y-\eta_1)]=E[y_1|\mathbf{x}]-\eta_1=0$
    - and $E[\mathbf{1}[d=0](y-\eta_0)]=E[y_0|\mathbf{x}]-\eta_0=0$.

Sample Splitting

- For the kth of K partitions formed by K-fold splitting (e.g. $K = 10$)
  - use ML to get $\widehat{\boldsymbol{\eta}}_k$ using data in $(K-1)$ folds (most of the data)
  - then obtain $\widehat{\beta}_k$ that solves $E[g(\mathbf{w}_i, \beta, \widehat{\boldsymbol{\eta}}_k)] = 0$ using remaining data in the kth fold
  - and form $\widehat{\beta} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\beta}_k$.
- So most data is used to obtain $\widehat{\boldsymbol{\eta}}_k$.
- We then get K separate $\widehat{\beta}_k^{\prime}s$
  - these are obtained using K distinct (independent) samples
  - so there is little loss in efficiency due to breaking into K pieces,
- Asymptotically this has a normal distribution with usual $Var[\widehat{\beta}] = A^{-1}BA^{-1}$
  - where $A = \partial E[g(\mathbf{w}_i, \beta, \boldsymbol{\eta})]/\partial\beta$ and $B = E[g(\mathbf{w}_i, \beta, \boldsymbol{\eta})g(\mathbf{w}_i, \beta, \boldsymbol{\eta})^{\prime}]$.

Sample Splitting (continued)

- The sample-splitting adds noise.
- To control for this
  - S times repeat the sample splitting method (e.g. $S = 500$)
  - each time get a $\widehat{\beta}_s$ (from averaging the K $\widehat{\beta}'_{ks}$) and $\widehat{\sigma}^2_s = Var[\widehat{\beta}_s]$
- Then $\overline{\widehat{\beta}} = \frac{1}{S}\sum_{s=1}^{S}\widehat{\beta}_s$
- And $Var[\widehat{\beta}] = \frac{1}{S}\sum_{s=1}^{S}\widehat{\sigma}^2_s + \frac{1}{S}\sum_{s=1}^{S}(\widehat{\beta}_s - \overline{\widehat{\beta}})^2$.

6.1 Hal Varian

- Hal Varian (2014), "Big Data: New Tricks for Econometrics," JEP, Spring, 3-28.
- Surveys tools for handling big data
  - file system for files split into large blocks across computers
    - Google file system (Google), Hadoop file system
  - database management system to handle large amounts of data across many computers
    - Bigtable (Google), Cassandra
  - accessing and manipulating big data sets across many computers
    - MapReduce (Google), Hadoop.
  - language for Mapreduce / Hadoop
    - Sawzall (Google), Pig
  - Computer language for parallel processing
    - Go (Google - open source)
  - simplified structured query language (SQL) for data enquiries
    - Dremel, Big Query (Google), Hive, Drill, Impala.

Hal Varian (continued)

- Surveys methods
  - article discusses k-fold CV, trees, lasso, ...
  - small discussion of causality and prediction
  - (note that a classic fail is Google flu trends)
  - for references mentions ESL and ISL.

6.2 Susan Athey

- 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.

Susan Athey and Guido Imbens

- Susan Athey and Guido Imbens (2019), "Machine Learning Methods Economists Should Know About."
- This paper provides great detail on the current literature with many references.

7. References

- Achim Ahrens, Christian Hansen, Mark Schaffer (2019), "lassopack: Model selection and prediction with regularized regression in Stata," arXiv:1901.05397
- Susan Athey (2018), "The Impact of Machine Learning on Economics". http://www.nber.org/chapters/c14009.pdf
- Susan Athey and Guido Imbens (2019), "Machine Learning Methods Economists Should Know About."
- 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.

References (continued)

- 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.
- 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 (2018), “Deep Neural Networks for Estimation and Inference: Application to Causal Effects and Other Semiparametric Estimands,” arXiv:1809.09953v2.

References (continued)

- Jon Kleinberg, H. Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan (2018), “Human decisions and Machine Predictions”, *Quarterly Journal of Economics*, 237-293.
- Sendhil Mullainathan and J. Spiess: “Machine Learning: Am Applied Econometric Approach”, *Journal of Economic Perspectives*, Spring 2017, 87-106.
- Hal Varian (2014), “Big Data: New Tricks for Econometrics”, *Journal of Economic Perspectives*, Spring, 3-28.
- Stefan Wager and Susan Athey (2018), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests,” JASA, 1228-1242.