Causal Learning in Economics

Mingli Chen

University of Warwick

May 2, 2025

Roadmap

(Prediction & Inference; Causal/Counterfactual Prediction & Inference)

- Supervised Learning
  - (post)-Lasso (Ridge, Elastic Net), (prune/boosted)Tree, Random Forest, Deep Neural Nets, etc
  - Ensemble learning/Aggregation and Cross-Breading of the ML methods.
  - Inference, partial linear model (RCT) & IV regression with selection/regularization
    - double machine learning, double partialling out (FWL), Neyman-orthogonalization
    - sample splitting, cross-fitting
  - extension, triple machine learning (heterogeneous treatment effect)
  - empirical applications: Mincer equations, Barro-Lee, AJR etc
  - Quantile regression: ℓ_1-QR
- Unsupervised Learning
  - Kernel density
  - PCA, factor model
  - Clustering, Kmeans
  - Topic modelling; Text Analysis

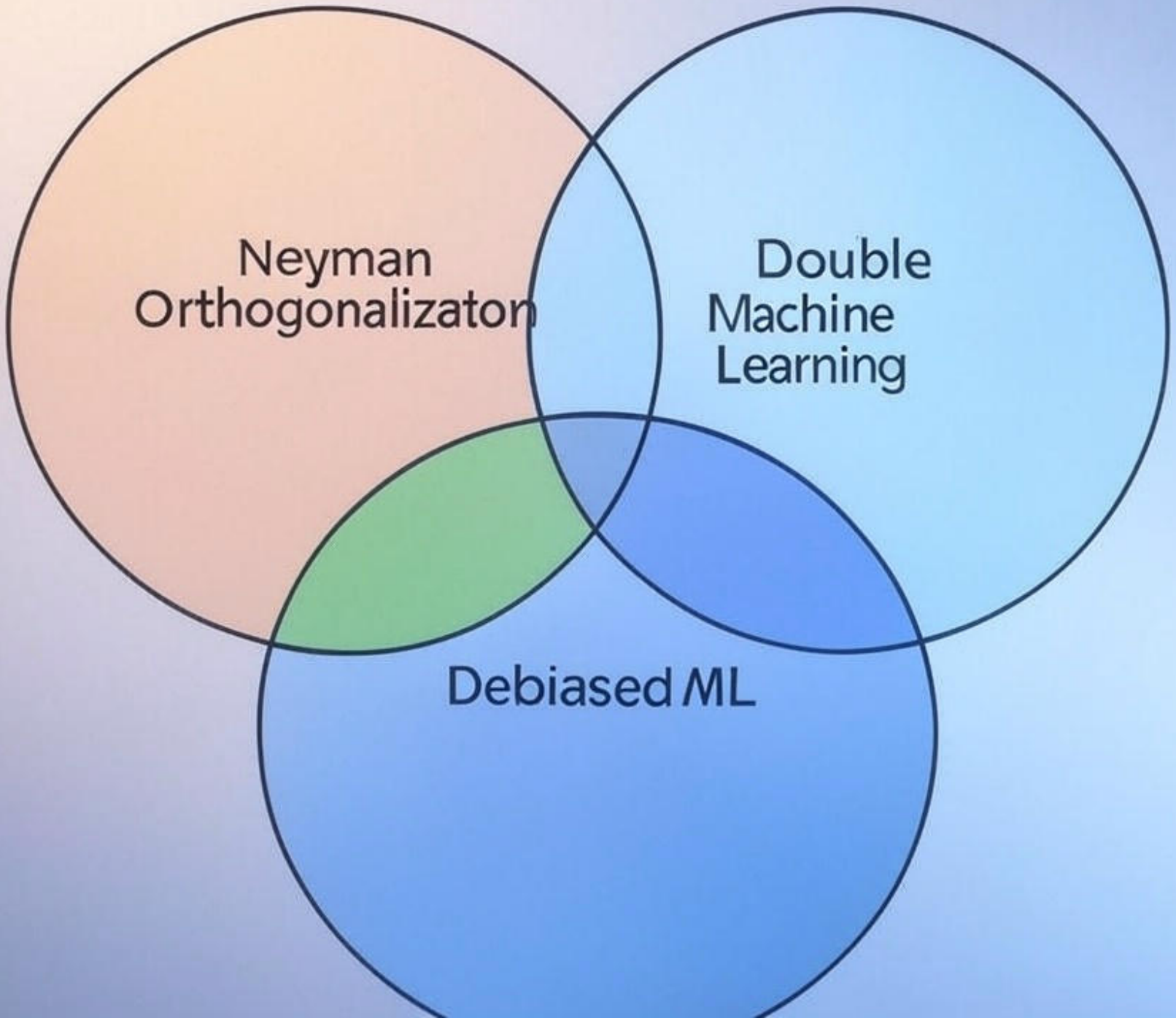
Roadmap (cont.)

- Supervised Learning + Unsupervised Learning
  - Causal Panel: Synthetic Control, Factor Model, and Nuclear Norm penalisation
    - ℓ_1-norm regularization on singular value of a matrix
  - High Dimensional Panel Quantile Models
    - many interesting work to be done
    - more difficult, challenges due to non-differentiability and individual and time effects
- (Deep) Reinforcement Learning & AI agents

Part I: Supervised Learning

- Machine Learning (ML) and Causal Inference
  - prediction VS counterfactual prediction

- Causal ML framework, e.g. ML with nuisance functions/parameters/models
  - Neyman Orthogonalization
  - Double Machine Learning
  - Debiased ML


Neyman
Orthogonalizaton
Double
Machine
Learning
Debiased ML

Part I: Supervised Learning

- Machine Learning (ML) and Causal Inference
  - prediction VS counterfactual prediction
- Causal ML framework, e.g. ML with nuisance functions/parameters/models
  - Neyman Orthogonalization
  - Double Machine Learning
  - Debiased ML
- Causal Inference in High-Dimensional Approximately Sparse Structural Linear Models
- Graphical model

Part I: Supervised Learning

- Machine Learning (ML) and Causal Inference
  - prediction VS counterfactual prediction
- Causal ML framework, e.g. ML with nuisance functions/parameters/models
  - Neyman Orthogonalization
  - Double Machine Learning
  - Debiased ML
- Causal Inference in High-Dimensional Approximately Sparse Structural Linear Models
- Graphical model

"Lasso is the new OLS"

Part I: Supervised Learning

- Machine Learning (ML) and Causal Inference
  - prediction VS counterfactual prediction

- Causal ML framework, e.g. ML with nuisance functions/parameters/models
  - Neyman Orthogonalization
  - Double Machine Learning
  - Debiased ML

- Causal Inference in High-Dimensional Approximately Sparse Structural Linear Models

- Graphical model

"Lasso is the new OLS"
R-package "hdm"

Model 1: High Dimensional Approximately Sparse Model

$$Y = \alpha_0 D + f_0(X) + U$$

If nuisance function f_0 is estimable at $O(n^{-1/2})$ rate then so is α_0

Problem: accurate nuisance estimates often unachievable when f_0 is non-parametric or linear and high-dimensional

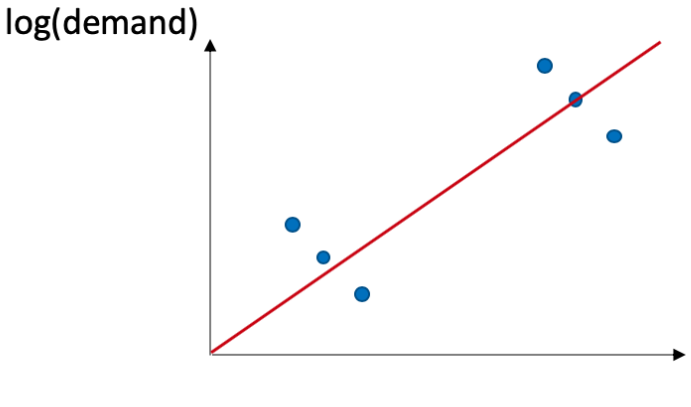
Model 2: Linear Endogenous Model

$$\underbrace{y_i}_{\text{outcome}} = \underbrace{d_i}_{\text{treatment}} \overbrace{\alpha_0}^{\text{effect}} + \underbrace{\sum_{j=1}^{p} x_{ij}\beta_{0j}}_{\text{controls}} + \underbrace{u_i}_{\text{noise}},$$

$$\mathrm{E}[u_i | \underbrace{x_i, z_i}_{\text{exogenous vars}}] = 0$$

Example 1: Estimating Price Elasticity of Demand

Goal: Estimate elasticity, the effect of a change in price on demand



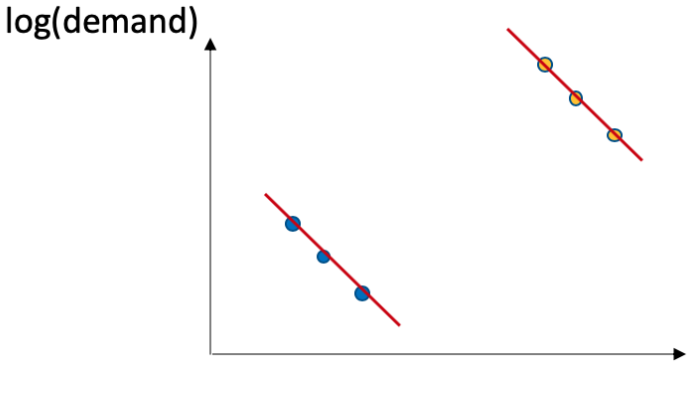
$$\underbrace{Y}_{\log(\text{demand})} = \underbrace{\alpha_0}_{\text{elasticity}} \cdot \underbrace{D}_{\log(\text{price})} + \underbrace{U}_{\text{noise}}$$

Conclusion: Increasing price increases demand!

Problem: Demand increases in winter and price anticipates demand

Example 1: Estimating Price Elasticity of Demand (Cont.)

Goal: Estimate elasticity, the effect of a change in price on demand

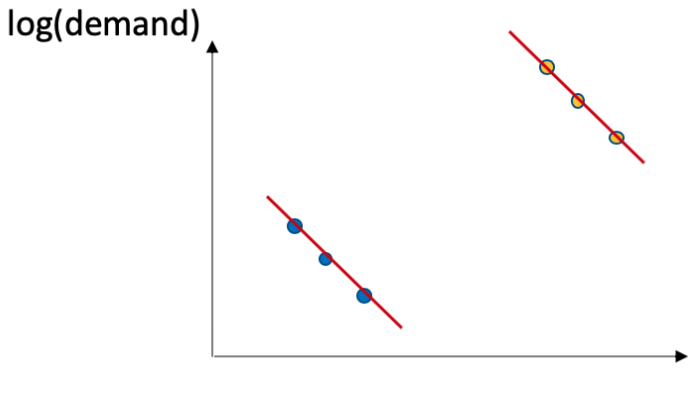


$$\underbrace{Y}_{\text{log(demand)}} = \underbrace{\alpha_0}_{\text{elasticity}} \cdot \underbrace{D}_{\text{log(price)}} + \beta_0 \cdot \underbrace{X}_{\substack{\text{season} \\ \text{indicator}}} + \underbrace{U}_{\text{noise}}$$

Idea: Introduce confounder (the season) into regression

Example 1: Estimating Price Elasticity of Demand (Cont.)

Goal: Estimate elasticity, the effect of a change in price on demand



$$\underbrace{Y}_{\text{log(demand)}} = \underbrace{\alpha_0}_{\text{elasticity}} \cdot \underbrace{D}_{\text{log(price)}} + \beta_0 \cdot \underbrace{X}_{\substack{\text{season} \\ \text{indicator}}} + \underbrace{U}_{\text{noise}}$$

Problem: What if there are 100s or 1000s of potential confounders?

Example 1: Estimating Price Elasticity of Demand (Cont.)

Problem: What if there are 100s or 1000s of potential confounders?

- Time of day, day of week, month, purchase and browsing history, other product prices, demographics, weather, …

One Option: Estimate effect of all potential confounders really well

$$\underbrace{Y}_{\text{log(demand)}} = \underbrace{\alpha_0}_{\text{elasticity}} \cdot \underbrace{D}_{\text{log(price)}} + \underbrace{f_0(X)}_{\text{effect of potential confounders}} + \underbrace{\epsilon}_{\text{noise}}$$

- If nuisance function f_0 is estimable at $O\left(n^{-1/2}\right)$ rate, then so is α_0

Problem: Accurate nuisance estimates often unachievable when f_0 is non-parametric or linear and high-dimensional

Example 1: Estimating Price Elasticity of Demand (Cont.)

Comments:

- Endogeneity; unobserved confounders; missing data
- i). Unobserved factors
  - although, adding interaction terms can be viewed as an approximation
- ii). Instrumental Variables (IVs), Z
  - many IVs => selection on Z

Example 2: Wage

- Scalar outcome variable Y, wage of a worker
- Vector of regressor variables or features CX:

$$CX = (CX_1, ..., CX_p)'$$

which contains worker's characteristics, e.g. education, experience, gender. We assume that a constant of 1 is included as a component.

- When needed, we may partition CX into

$$CX = (D, X')',$$

where D is the target regressor or treatment (e.g. gender indicator), whose impact is of interest, and X are other regressors that usually serve as controls.

Two Main Questions:

- The purpose of regression analysis is to characterise the statistical relation of Y with X:

  1. The **Prediction Question:** How can we use CX to predict Y well?

  2. The **Inference Question:** How does the predicted value of Y change if we change the component D of CX, holding X (the other components of X) fixed?

- We will address the prediction question first and the inference question second.

Two Main Questions in the Wage Example

- The Prediction Question: how to use job-relevant characteristics, such as education and experience, to best predict wages?
- The Inference Question: what is the difference in predicted wages between men and women with the same job-relevant characteristics?

Preview of a Wage Case Study

- Case study using data from the U.S. Current Population Survey (CPS) in 2012 for single (never married) workers.
- We shall
  - Construct a prediction rule for hourly wage, which depends linearly on the job-relevant characteristics.
  - Assess the quality of the prediction rule using out-of-sample prediction performance.
  - Find that on average women are paid about 2 dollars less per hour than men with the same experience and other recorded characteristics.
- This estimate is called the *gender wage gap* in labor economics, and measures (in part) gender pay discrimination.

Quality of Prediction: Intuition

- The best linear prediction rule is $\alpha_0 D + \beta_0' X$. Does $\hat{\alpha} D + \hat{\beta}' X$ approximate $\alpha_0 D + \beta_0' X$?
- We are trying to estimate p parameters $\alpha_0, \beta_{01}, ..., \beta_{0,p-1}$, without imposing any assumptions on these parameters.
- Intuitively, to estimate each parameter well, we we need many observations per parameter.
- This means that n/p must be large, or, equivalently p/n must be small.

Quality of Prediction: Theory

Denote $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$, $\theta_0 = (\alpha_0, \beta_0)$.

Theorem

Under regularity conditions

$$\sqrt{\mathrm{E}_X(\theta_0' CX - \hat{\theta}' CX)^2} \leq \mathrm{const} \cdot \sqrt{\mathrm{E}u^2} \sqrt{\frac{p}{n}},$$

where E_X is expectation with respect to X and the inequality holds w.p.a. 1 as $n \to \infty$.

If n is large and p is much smaller than n, for nearly all realizations of data, the sample linear regression can come close to the population linear regression.

Summary 1

- We define linear regression in population and in sample thought the best linear prediction problems solved in population and in the sample
- The sample linear regression (best linear predictor) approximates the population linear regression (best linear predictor) when the ratio p/n is small
- We will discuss the assessment of prediction performance in practice next.

Case Study: Predicting Wages

Our goals are

1. Predict wages using various characteristics of workers.

2. Assess the predictive performance using adjusted MSE and R^2, and out-of-sample MSE and R^2.

Data

- Data is from the March Supplement of the U.S. Current Population Survey, year 2012.
- Focus on the single (never married) workers with education levels equal to high-school, some college, or college graduates.
- The sample is of size $n \approx 4,000$
- The outcome Y is hourly wage, and CX are various characteristics of workers.

Descriptive Statistics

	Mean
Wage	15.53
Female	0.42
Experience	13.35
College graduate	0.38
Some college	0.32
High school graduate	0.30
Midwest	0.29
South	0.24
West	0.21
Northeast	0.26

Predictive Models

- **Basic Model:** CX consists of the female indicator (D) and other controls X, which contain a constant, experience, experience squared, experience cubed, education indicators, and regional indicators. CX includes $p = 10$ regressors.

- **Flexible Model:** CX consists of D as well as X, which contains all of the components of X in the basic model plus their two-way interactions. An example of a regressor created through a two-way interaction is experience times the indicator of having a college degree; another example is the indicator of having a high-school diploma times the indicator of working in the "north-east" region. CX includes $p = 33$ regressors.

Performance of Predictive Models

- Since p/n is small, the sample linear regression should approximate the population linear regression well.
- We expect the sample R^2 to agree with adjusted R^2 and be a good measure of out-of-sample performance.

Assessing Predictive Performance

	p	R^2_{sample}	R^2_{adj}	MSE_{adj}
basic reg	10	0.09	0.09	165.68
flex reg	33	0.10	0.10	165.12

We conclude that the performance of the basic and flexible model are about the same, with the flexible model being just slightly better (slightly higher R^2_{adj} and lower MSE_{adj}).

- Using a real example, we have assessed predictive performance of two linear prediction rules.

- Next we will proceed to discuss the Inference Problem.

The Inference Question: Introduction

- We partition the vector of regressors CX into two components:

$$CX = (D, X')'.$$

  where D represents the "target" regressors of interest, and W represents the other regressors, sometimes called the controls.

- In the wage example, D is the female indicator and W include experience, educational, and geographic characteristics.

- Accordingly, write

$$Y = \underbrace{\alpha_0 D + \beta' X}_{\text{Predicted value}} + \underbrace{U}_{\text{error}}$$

- The **Inference Question**: How does the predicted value of Y change if we increase D by a unit, holding X fixed?

The Inference Question: Introduction (Cont.)

- In the wage example: what is the difference in predicted wages between men and women with the same job-relevant characteristics?
- The answer is the population regression coefficient

$$\alpha_0$$

  corresponding to the target regressor D.
- In the wage example, D is the female indicator and α_0 is the Gender Wage Gap.

Understanding α_0 via "Partialling-Out"

- "Partialling-out" is an important tool that provides conceptual understanding of the regression coefficient α_0.

- In the *population*, define the partialling-out operation as a procedure that takes a random variable V and creates a "residual" $\tilde{V}$ by subtracting the part of V that is linearly predicted by X :

$$\tilde{V} = V - \gamma_{VX}' X, \quad \gamma_{VX} = \underbrace{\arg\min}_{\gamma} \mathrm{E}(\mathrm{V} - \gamma' \mathrm{X})^2$$

- When V is a vector, we apply the operation to each component.

- It can be shown that the partialling-out operation is linear:

$$Y = V + U \Rightarrow \tilde{Y} = \tilde{V} + \tilde{U}.$$

- We apply the partialling-out to both sides of our regression equation $Y = \alpha_0 D + \beta' X + U$ to get:

$$\tilde{Y} = \alpha_0 \tilde{D} + \beta' \tilde{X} + \tilde{U},$$

which simplifies to the decomposition:

$$\tilde{Y} = \alpha_0 \tilde{D} + U, \quad \mathrm{E} U \tilde{D} = 0.$$

- This follows because partialling-out takes out $\beta' X$, since $\tilde{X} = 0$, and leaves U untouched, $\tilde{U} = U$, since U is linearly unpredictable by CX and therefore by X .

- Moreover, $\mathrm{E} U \tilde{D} = 0$ since $\tilde{D}$ is a linear function of CX.

Frisch-Waugh-Lovell Theorem

The decomposition implies that $\mathbb{E}U\tilde{D} = 0$ are the Normal Equations for the population regression of $\tilde{Y}$ on $\tilde{D}$. Thus:

Theorem (FWL)

The population linear regression coefficient α_0 can be recovered from the population linear regression of $\tilde{Y}$ on $\tilde{D}$:

$$\alpha_0 = \arg\min_{b_1} \mathrm{E}(\tilde{Y} - b_1\tilde{D})^2 = (\mathrm{E}\tilde{D}^2)^{-1}\mathrm{E}\tilde{D}\tilde{Y},$$

where α_0 is uniquely defined if D cannot perfectly predicted by X, i.e. $\mathrm{E}\tilde{D}^2 > 0$.

This is a remarkable fact. It asserts that α_0 can be interpreted as a (univariate) regression coefficient of **residualized** Y on **residualized** D, where the residuals are defined by partialling-out the linear effect of X from Y and D.

How to do Estimation?

- In the sample, we will mimic the partialling-out in the population.
- When p/n is small, we can do this by sample linear regression.
- When p/n is not small, using sample linear regression for partialling-out is not a good idea. Can do penalized regression and dimension reduction instead. More on this later.

Inference Result: Theory

Theorem (Inference)

If p/n is small, then the estimation error in $\check{D}_i$ and $\check{Y}_i$ has no first order effect on $\hat{\alpha}$, and

$$\hat{\alpha} \sim N(\alpha_0, V/n)$$

where

$$V = (\mathrm{E}\tilde{D}^2)^{-1}\mathrm{E}(\tilde{D}U^2)(\mathrm{E}\tilde{D}^2)^{-1}$$

Comments:

- We interpreted α_0 as the regression coefficient in the bivariate regression of the response variable on the target variable, after we have removed the linear effect of the other variables.
- This result is useful for interpretation and understanding of the regression coefficient.
- It will also be useful for setting up inference in modern high-dimensional settings.
- Next, we will carry out a case study for the Wage Example.

Case Study: Inference about Gender Wage Gap

What is the difference in predicted wages between men and women with the same job-relevant characteristics?

CPS 2012 Data: Summary

	Male means	Female means
Wage	16.12	14.72
Experience	13.58	13.04
College graduate	0.35	0.41
Some college	0.30	0.35
High school graduate	0.34	0.24
Midwest	0.28	0.29
South	0.24	0.26
West	0.22	0.20
Northeast	0.26	0.26

Specifications

- We estimate the linear regression model:

$$Y = \alpha_0 D + \beta' X + U$$

- D is the indicator of being a female (1 if female and 0 otherwise). X's are controls.
- Basic model: X 's consist of education and regional indicators, experience, experience squared, and experience cubed.
- Flexible model: X's consist of controls in the basic model plus all of their two-way interactions.

Results

	Estimate	Std Error	Confidence Interval
	$\widehat{\beta}_1$	$\sqrt{\widehat{V}/n}$	95%
Basic reg	-1.83	0.42	[-2.66 -0.99]
Flex reg	-1.88	0.42	[-2.71 -1.05]

- The estimated gender gap in hourly wage is about $-2\$$ with a confidence interval that ranges from about $-2.7\$$ to $-1\$$.
- This means that women get paid $2\$$ less per hour on average than men, controlling for experience, education, and geographical region.

Comparison of Estimates Based on Full Regression and Partialling-Out

	Estimate	Standard Error
basic reg	-1.83	0.42
flex reg	-1.88	0.42
basic reg with partialling out	-1.83	0.42
flex reg with partialling out	-1.88	0.42

We now turn to the ultra/high dimensional setting:

- Double Machine Learning
- Triple Machine Learning
- Neyman Orthogonalization for
  - policy learning;
  - graph learning (graphical model), e.g. one of Belloni, Chen, Chernozhukov 2016's contributions is providing inference for graphical models

LASSO

- Least Absolute Shrinkage and Selection Operator
- Add an ℓ_1 penalty to the regression objective. Used in high-dimensional nonparametric regression to select relevant basis functions or features.

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

where y_i is the response, λ is the tuning parameter, and β_j are coefficients.

Regression Trees

- Partition the feature space into regions based on feature thresholds, estimating a constant or simple function in each region. Commonly used in regression and classification via recursive binary splitting.
- For regression trees, the predicted value in a region R_m is:

$$\hat{f}(x) = \sum_{m=1}^{M} \hat{\beta}_m \mathbb{I}(x \in R_m)$$

where $\mathbb{I}(x \in R_m)$ is the indicator for region R_m, and $\hat{\beta}_m$ is the predicted value.

Regression Trees

- The predicted values $\beta = (\beta_1, ..., \beta_M)$ are obtained by minimizing the sample MSE:

$$\beta = \arg \min_{b_1,...,b_M} \sum_{i=1}^{n} \left(y_i - \sum_{m=1}^{M} b_m \mathbb{I}(x_i \in R_m) \right)^2$$

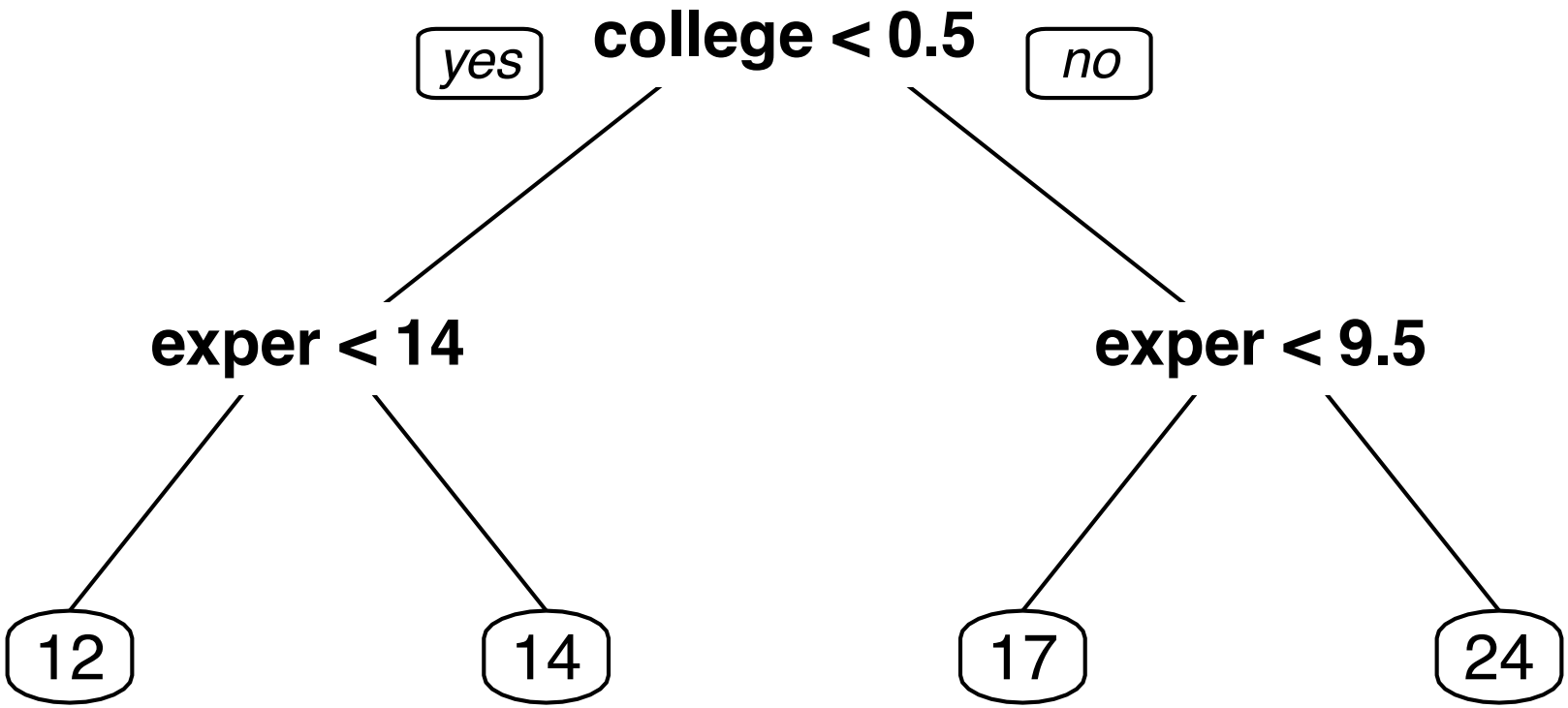
so that

$$\beta_m = \text{average of } y_i \text{ where } x_i \in R_m$$

- The regions $R_1, ..., R_M$ are called nodes, and each node R_m has a predicted value $\hat{\beta}_m$ associated with it.

Figure that illustrates Idea:

A nice feature of the regression trees is that you get to draw cool pictures. Consider the Wage Example, where Y is hourly wage, and Z include experience, geographic, and educational characteristics.



In this tree the predicted hourly wage for college graduates (college $=1$) with more than 9.5 years of experience is 24 dollars, and otherwise is 17; the predicted wage for non-college graduates with more than 14 years of experience is 14 and otherwise is 12.

Figure 1:“To prune a tree”

Neural Networks

- Use **parameterized nonlinear transformations** of **linear combinations** of **the raw regressors** as constructed regressors (called neurons), and produce the predicted value as a **linear function** of these regressors.
- The method and the name "neural networks" were loosely inspired by the mode of operation of the human brain, and developed by scientists working on the Artificial Intelligence. Difference between DRL and DL.
- They can be represented by cool graphs and diagrams that we will discuss shortly, so please stay tuned.
- Here we focus on single layer neural network to discuss the idea.

Single Layer Neural Networks

- The estimated prediction rule will take the form:

$$\hat{g}(X) = \sum_{m=1}^{M} \hat{\beta}_m Z_m(\hat{\alpha}_m)$$

where the $Z_m(\hat{\alpha}_m)$'s are the **constructed regressors** called **neurons**

- The M neurons are generated by

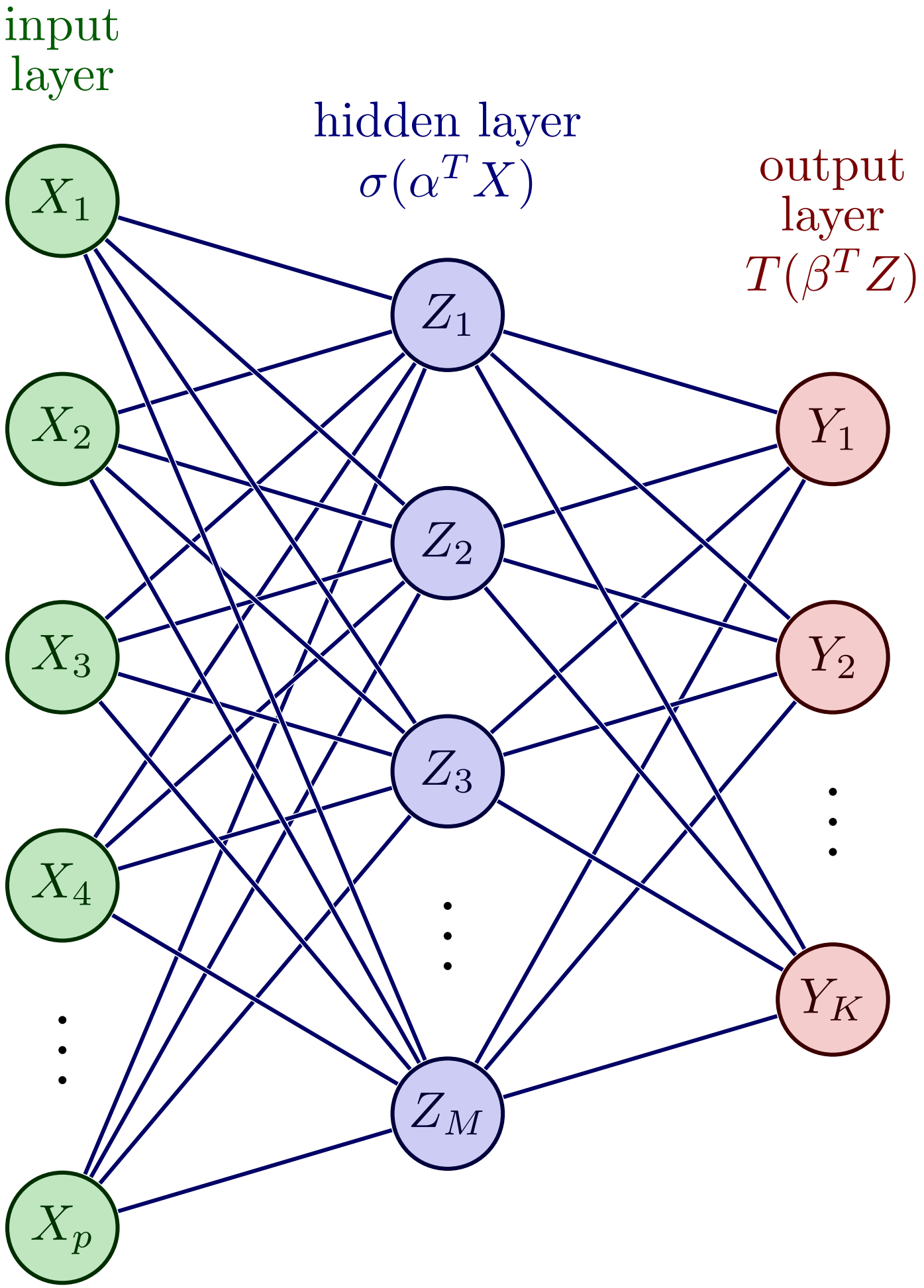
$$Z_m(\alpha_m) = \sigma(\alpha_m^T X), \quad m = 1, ..., M,$$

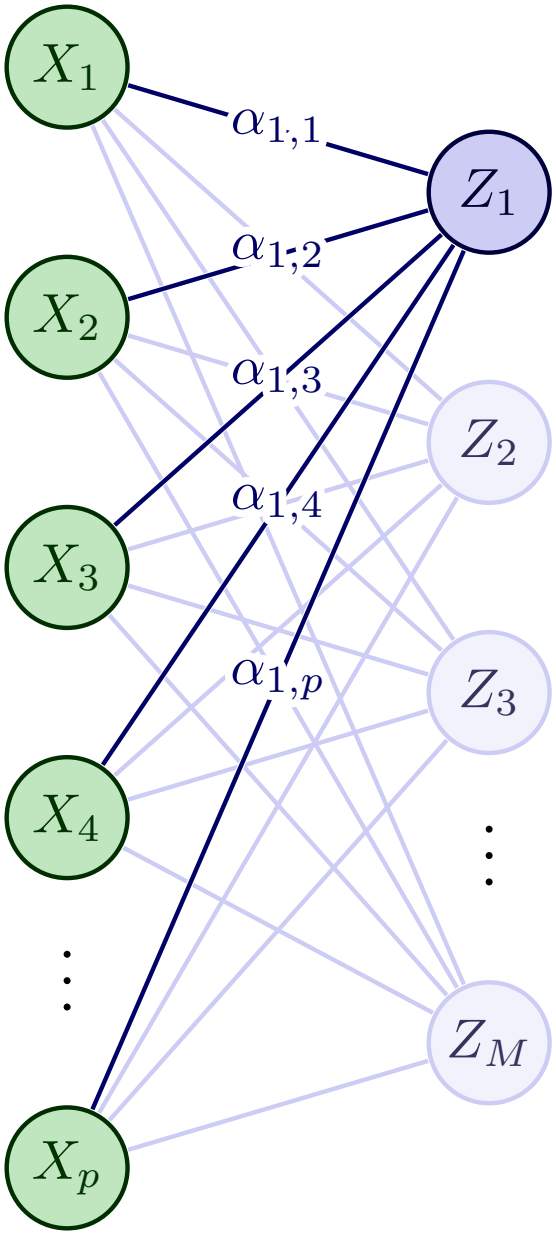
where α_m's are neuron-specific vectors of parameters called weights, and σ is the activation function, for example:
  - the sigmoid function:

$$\sigma(X) = \frac{1}{1 + e^{-X}}$$

  - the rectified linear unit function (ReLU):

$$\sigma(X) = \max(0, X)$$


input
layer
hidden layer
$\sigma(\alpha^T X)$
output
layer
$T(\beta^T Z)$
X_1
X_2
X_3
X_4
X_p
Z_1
Z_2
Z_3
Z_M
Y_1
Y_2
Y_K



$$= \sigma\left(\alpha_{1,1}X_1 + \alpha_{1,2}X_2 + \ldots + \alpha_{1,p}X_p + b_1^{(0)}\right)$$

$$= \sigma\left(\sum_{i=1}^{p} \alpha_{1,i}X_i + b_1^{(0)}\right)$$

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_M \end{pmatrix} = \sigma\left[\begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,p} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{M,1} & \alpha_{M,2} & \cdots & \alpha_{M,p} \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} + \begin{pmatrix} b_1^{(0)} \\ b_2^{(0)} \\ \vdots \\ b_M^{(0)} \end{pmatrix}\right]$$

$$\mathbf{Z} = \sigma\left(\alpha^{(0)}\mathbf{X} + \mathbf{b}^{(0)}\right)$$

Estimation

The estimators $\hat{\alpha}_m$ and $\hat{\beta}_m$, for $m = 1, ..., M$, are obtained as the solution to the penalised the nonlinear least squares problem:

$$\min_{\{\alpha_m\},\{\beta_m\}} \sum_i \left(Y_i - \sum_{m=1}^{M} \beta_m^T Z_{im}(\alpha_m) \right)^2 + \lambda \left(\sum_m \sum_j |\alpha_{mj}| + \sum_m |\beta_m| \right)$$

In this formula we use Lasso type penalty, but we can also use Ridge and other type of penalties.

Double ML for Treatment Effect

[Belloni, Chernozhukov, Hansen 2014] [Chernozhukov et al 2017a,b]

1. Lasso/Regress $Y \sim X$, learn $q(X) = \hat{E}[Y|X]$
2. Lasso/Regress $D \sim X$, learn $p(X) = \hat{E}[D|X]$ (mean treatment policy)
3. Linear Regression on residuals: $Y - \hat{E}[Y|X] \sim D - \hat{E}[D|X]$

$$\min_{\alpha} \frac{1}{n} \sum_{i} \left(Y_i - q(X_i) - \alpha \cdot (D_i - p(X_i)) \right)^2$$

coefficients in final regression is treatment effect α

- Neyman orthogonal estimator of α_0 robust to first-order errors in nuisance estimates; yields $\sqrt{n}$-consistent and asymptotically normal estimate of α_0
- Nuisance estimates can be fitted by arbitrary ML methods, subject to achieving RMSE consistency at the slow rate of $O(n^{-1/4})$

Other Examples

Effect of Institutions on the Wealth of Nations

- Acemoglu, Johnson, Robinson (2001)
- Impact of institutions on wealth

$$\underbrace{y_i}_{\text{log gdp per capita}} = \underbrace{d_i}_{\text{quality of institutions}} \overbrace{\alpha_0}^{\text{effect}} + \underbrace{\sum_{j=1}^{p} x_{ij}\beta_{0j}}_{\text{geography controls}} + u_i,$$

- Instrument z_i : the early settler mortality (200 years ago)
- Sample size $n = 67$
- Specification of controls:
  - Basic: constant, latitude ($p = 2$)
  - Flexible: + cubic spline in latitude, continent dummies ($p = 16$)
- R-package **hdm**

Some remarks

- The current theorem applied to the estimation of a constant treatment effect α; for personalized decisions we want a heterogeneous treatment effect $\alpha(X)$
- We want to build a complex ML model for $\alpha(X)$

Triple ML

[Chernozhukov et al 2017a,b] [Nie and Wager, 2017]

1. Lasso/Regress $Y \sim X, W$, learn $q(X, W) = \hat{E}[Y|X, W]$
2. Lasso/Regress $D \sim X, W$, learn $p(X, W) = \hat{E}[D|X, W]$
3. minimize residual square loss

$$\min_{\alpha(\cdot)\in\Theta} \frac{1}{n}\sum_i \left(Y_i - q(X_i, W_i) - \alpha(X_i)\cdot(D_i - p(X_i, W_i))\right)^2$$

error in final regression is of the same order as if we knew the nuisance functions

Comments: reverse causal problem: causal policy learning

Triple ML for Treatment Effect

Binary Treatment [Foster, Syrgkanis, 2019], [Oprescu, Wu, Syrgkanis, 2018]

1. Regress $Y \sim D, X, W$, learn $h_t(X, W) = \hat{E}[Y|D = t, X, W]$
2. Regress $D \sim X, W$, learn $p_t(X, W) = \hat{P}[D = t|X, W]$ (prob of treatment)
3. Doubly Robust Target

$$Y^{(t)}_{i,DR} = h_t(X_i, W_i) + \frac{(Y_i - h_t(X_i, W_i)) \cdot 1\{D_i = t\}}{p_t(X_i, W_i)}$$

4. Regress $Y^{(1)}_{i,DR} - Y^{(0)}_{i,DR} \sim X$

$$\min_{\alpha(\cdot)\in\Theta} \frac{1}{n} \sum_i \left(Y^{(1)}_{i,DR} - Y^{(0)}_{i,DR} - \alpha(X_i) \right)^2$$

Quantile Graphical Model

(Approximating) Conditional Independence [Belloni, Chen, Chernozhukov 2016]

We consider a non-Gaussian (high-dimensional) setting.

Quantile Graphical Model

(Approximating) Conditional Independence [Belloni, Chen, Chernozhukov 2016]

We consider a non-Gaussian (high-dimensional) setting.

$$X_a \perp X_b | X_{V \setminus \{a,b\}}$$

if and only if

$$Q_{X_a}(\tau | X_{V \setminus \{a\}}) = Q_{X_a}(\tau | X_{V \setminus \{a,b\}}) \text{ for all } \tau \in (0,1), \text{ and } X_{V \setminus \{a\}} \in \mathcal{X}_{V \setminus \{a\}}$$

Quantile Graphical Model

(Approximating) Conditional Independence [Belloni, Chen, Chernozhukov 2016]

We consider a non-Gaussian (high-dimensional) setting.

$$X_a \perp X_b | X_{V \setminus \{a,b\}}$$

if and only if

$$Q_{X_a}(\tau | X_{V \setminus \{a\}}) = Q_{X_a}(\tau | X_{V \setminus \{a,b\}}) \text{ for all } \tau \in (0,1), \text{ and } X_{V \setminus \{a\}} \in \mathcal{X}_{V \setminus \{a\}}$$

For a set of quantile indices $\mathcal{T} \subset (0,1)$, we say that

$$X_a \perp_{\mathcal{T}} X_b | X_{V \setminus \{a,b\}}$$

X_a and X_b are $\mathcal{T}$-conditionally independent given $X_{V \setminus \{a,b\}}$

Estimation of CIQGM

- For each $a \in V$,

$$Q_{X_a}(\tau | X_{V\setminus\{a\}}) = CX^a \beta_{a\tau} + r_{a\tau}, \quad \beta_{a\tau} \in \mathbb{R}^p, \quad \text{for all } \tau \in \mathcal{T}$$

  - the p-dimensional vector $CX^a = CX^a(X_{V\setminus\{a\}}) \in \mathbb{R}^p$ is based on transformations of the original covariates $X_{V\setminus\{a\}}$
  - $r_{a\tau}$ denotes a small approximation error

- For $b \in V\setminus\{a\}$,

$$I_a(b) := \{j : CX^a_j \text{ depends on } X_b\}$$

  - under correct specification, if X_a and X_b are conditionally independent, we have $\beta_{a\tau,j} = 0$ for all $j \in I_a(b)$, $\tau \in (0,1)$

Estimation of CIQGM (Algorithm 1)

For each $a \in V$, and $j \in [p]$, and $\tau \in \mathcal{T}$, perform the following:

1. Run Post-ℓ_1-quantile regression of X_a on CX^a
2. Run Post-Lasso of $f_{a\tau} CX^a_j$ on $f_{a\tau} CX^a_{-j}$
3. Construct the score function

$$\hat{\psi}_i(\alpha) = (\tau - 1\{X_{ia} \leq CX^a_{ij}\alpha + CX^a_{i,-j}\widetilde{\beta}_{a\tau,-j}\})f_{ia\tau}(CX^a_{ij} - CX^a_{i,-j}\widetilde{\gamma}^j_{a\tau})$$

for $L_{a\tau j}(\alpha) = |\mathbb{E}_n[\hat{\psi}_i(\alpha)]|^2/\mathbb{E}_n[\hat{\psi}^2_i(\alpha)]$, set

$$\check{\beta}_{a\tau,j} \in \arg\min_{\alpha \in \mathcal{A}_{a\tau j}} L_{a\tau j}(\alpha)$$

Quantile Graphical Model

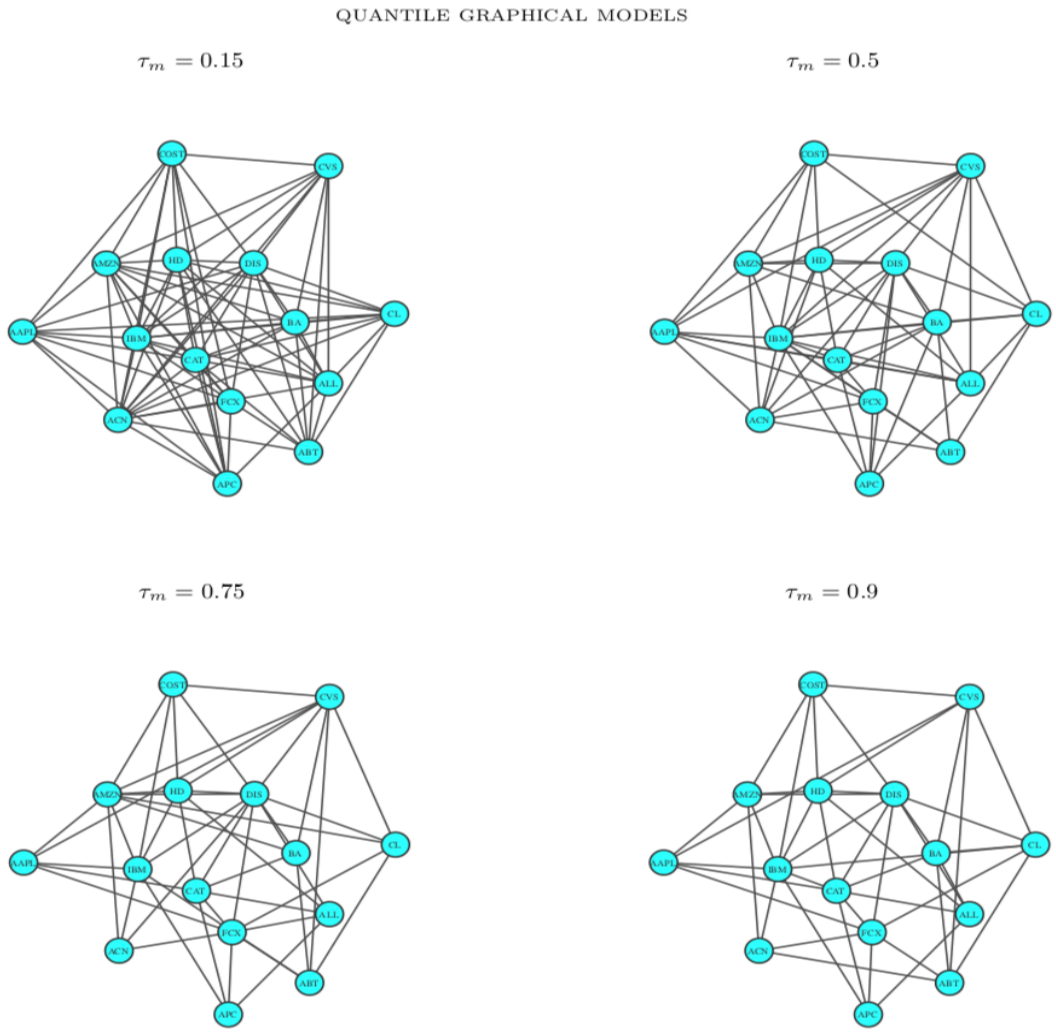


FIGURE 1. Stock Returns Interdependence under Different Market Conditions. Note: Presence of edges between nodes indicate that these two nodes (or stocks) are conditionally dependent.

Neyman Orthogonality in a Nutshell

Directional derivative:

$$D_\alpha L_D(\alpha; g)[v_\alpha] = \frac{d}{dt} L_D(\alpha + t \cdot v_\alpha; g)$$

A loss $L_D(\alpha; g)$ is Neyman Orthogonal if

$$D_g D_\alpha L_D(\alpha_0; g_0)[v_\alpha, v_g] = 0$$

Intuition:
Small perturbation of nuisance g around its true value, do not change the gradient information of the loss with respect to target

Take-Away

- Neyman orthogonality can improve ML theory for causal problems
- Robustness to nuisance errors and improved quality of target parameters while maintaining causal interpretation
- Enables asymptotically valid confidence interval construction
- Nicely blends with modern and classical statistics/econometrics