Data Science Lab - 4

Academic year 2018-2019

Lecturer Falco J. Bargagli-Stoffi

IMT - School for Advanced Studies Lucca & KU Leuven

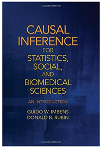
Causal Inference - Introduction

Why Causal Inference?

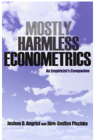
- Correlation is not causation: simple correlations can lead to misguided policies
- Among many different options, important to choose the *most effective* intervention
- Accurate cost-benefit analysis

Causality Frameworks

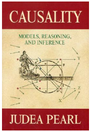
- Rubin Causal Model (Imbens & Rubin, 2015)



- Angrist & Pischke (2009)



- Pearl (2000)



RCM (1980): Potential Outcomes Framework

RCM (1980): Set Up

- Rubin's potential outcome framework (1974):
  - Given a set of N units, indexed by $i = 1, ..., N$. Let W_i be the binary indicator of the reception of the treatment:

$$W_i \in \{0, 1\}$$

  - Given this notation and SUTVA we can postulate the existence of a pair of potential outcomes for each unit:

$$Y_i^{obs} = Y_i(W_i) = \begin{cases} Y_i(0) & if\ W_i = 0 \\ Y_i(1) & if\ W_i = 1 \end{cases}$$

  - We can define the Causal Effect as a simple difference between the potential outcome under treatment and under control:

$$\tau_i = Y_i(1) - Y_i(0)$$

RCM (1974): Science World

- Imagine that we want to assess the effect (*causal effect*) of a job training (*treatment*) on a pool of students (*units*)

ID	Education X_i	Treated W_i	No job training $Y_i(0)$	Job training $Y_i(1)$	Treatment effect $\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	1	1
2	High school	1	0	1	1
3	High school	1	1	1	0
4	College	1	1	1	0
5	College	0	1	1	0
6	College	0	0	1	1

- Average Treatment Effect (ATE):

$$
\begin{aligned}
\bar{\tau} &= \bar{Y}(1) - \bar{Y}(0) \\
&= 1 - 0.5 \\
&= 0.5
\end{aligned}
$$

RCM (1974): Real World

ID	Education X_i	Treated W_i	No job training $Y_i(0)$	Job training $Y_i(1)$	Treatment effect $\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	1	?	1	?
3	High school	1	?	1	?
4	College	1	?	1	?
5	College	0	1	?	?
6	College	0	0	?	?

- Average Treatment Effect:

$$\bar{\tau} = 0.66$$

$$\downarrow$$

32% bigger: why this bias?

Selection Bias (intuition)

- People do not randomly select into various programs which we would like to evaluate

ID	Education X_i	Treated W_i	No job training $Y_i(0)$	Job training $Y_i(1)$	Treatment effect $\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	1	?	1	?
3	High school	1	?	1	?
4	College	1	?	1	?
5	College	0	1	?	?
6	College	0	0	?	?

Higher treatment rate & higher treatment effects: $W_i \not\perp Y_i(0), Y_i(1)$

Selection Bias (mathematical intuition)

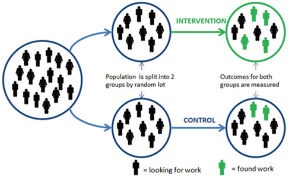
- As noted above, simply comparing those who are and are not treated may provide a misleading estimate of a treatment effect
- This problem can be efficiently described by using mathematical expectation notation to denote population averages:

$$\begin{aligned}\bar{\tau} &= \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] \\ &= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1]}_{\text{Average Treatment Effect on the Treated}} + \underbrace{\left[\mathbb{E}[Y_i(0)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0]\right]}_{\text{Selection bias}}\end{aligned}$$

- Thus, the naive contrast can be written as the sum of two components, ATET, plus Selection Bias
- Average earnings of non-trainees, $\mathbb{E}[Y_i(0)|W_i = 0]$, may not be a good standing for the earnings of trainees had they not been trained, $\mathbb{E}[Y_i(0)|W_i = 1]$

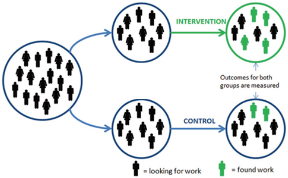
Possible solutions

- The problem of selection bias motivates the use of:
  1. Random assignment (ex-ante) → experimental set-up



The basic design of a randomised controlled trial (RCT)

  2. Unconfoundedness (ex-post) → observational studies



  3. Instrumental variable (ex-post) → observational studies

Random Assignment

- Random assignment ensures that the potential earnings of trainees had they not been trained are well-represented by the randomly selected control group

- Formally, when W_i is randomly assigned, then:

  $$\mathbb{E}[Y_i|W_i = 1] - \mathbb{E}[Y_i|W_i = 0] = [Y_i(1) - Y_i(0)|W_i = 1] = E[Y_i(1) - Y_i(0)]$$

- Replacing $E[Y_i|W_i = 1]$ and $E[Y_i|W_i = 0]$ with the corresponding sample analog provides a consistent estimate of ATE

Unconfoundedness (or CIA)

- The Unconfoundedness assumption states that conditional on observed characteristics, the selection bias disappears
- Formally, we overcome the problem that we have seen at slide 9, because: $W_i \perp\!\!\!\perp Y_i(0), Y_i(1) | X_i$
  This holds true even if conditioning just on:
  $e(x) = P(W = 1 | X_i = x)$
- Given unconfoundedness, comparison of average effects of job training have a causal interpretation:

$$\bar{\tau} = \mathbb{E}[Y_i(1)|W_i = 1, X_i] - \mathbb{E}[Y_i(0)|W_i = 0, X_i] = \mathbb{E}[Y_i(1) - Y_i(0)|X_i]$$

- This can be generalized to the case of a continuous treatment variable (i.e effects of education on employment): $s_i \perp\!\!\!\perp Y_{s_i} | X_i$
- Conditional on X_i, what is the average causal effect of a one-year increase in collage attendance?

$$\mathbb{E}[Y_i|s_i = s, X_i] - \mathbb{E}[Y_i|s_i = s - 1, X_i] = \mathbb{E}[f_i(s) - f_i(s - 1)|X_i]$$

Using regression to summarize

- Since s_i takes on many values, there are as many average causal effects as the possible increments in s_i
- Regression is a weighted average of the individual specific difference: $f_i(s) - f_i(s-1)$
- Then if we assume that:

$$f_i(s) = Y_i = \underbrace{\beta_0 + \beta_1 s + \eta_i}_{\eta_i = X_i^T\gamma + \epsilon_i \text{ where } \mathbb{E}[\eta_i] = X_i^T\gamma}$$

- If unconfoundedness holds:

$$\mathbb{E}[f_i(s)|X_i, s_i] = \mathbb{E}[f(s_i)|X_i] = \beta_0 + \beta_1 s + \mathbb{E}[\eta_i] = \beta_0 + \beta_1 s + X_i^T\gamma$$

- If we assume ϵ_i to be uncorrelated with s_i and X_i then β_1 is the *causal effect*
- Yet, note that we assume that observable characteristics X_i are the only reason why s_i and η_i are correlated

Omitted Variable Bias

- Let $Y_i = \beta_0 + \beta_1 s_i + A_i^T \gamma + e_i$ where the vector A^T stands for *ability* (factors such as intelligence, motivation etc)
- If unconfoundedness holds given A_i then β_1 is the coefficient of a linear causal model
- But what happens if we cannot measure A_i i.e. we leave it out of the regression?

$$\frac{Cov(Y_i, s_i)}{V(s_i)} = \frac{Cov\left(\left[\beta_0 + \beta_1 s_i + A_i^T \gamma + e_i\right], s_i\right)}{Var(s_i)} = \underbrace{\beta_1}_{\text{causal effect}} + \underbrace{\gamma^T \delta_{A_s}}_{\text{OV bias}}$$

  where δ_{A_s} is the vector of coefficients from the regression of A_i on s_i
- Take home message: the more regressor your include (longer regression), the less your causal estimand is biased

Machine Learning and Causality

Using CART to estimate heterogenous causal effect

Machine Learning and Causality

- `Econometrics/ Statistics/ Social Science`
  - Formal theory of causality
    - Potential outcomes methods (Rubin) maps onto economic approaches
  - Well-developed and widely used tools for estimation and inference of causal effect in experimental and observational studies
    - Used by social science, policy-makers, development organizations, medicine, business, experimentation
  - Weaknesses
    - Non-parametric approaches fail with many covariates
    - Model selection unprincipled

Motivations

- Experiments and Data-Mining
  - Concerns about ex-post "data-mining"
    - In medicine, scholars are required to pre-specify analysis plan (similar in economic field experiments)
- How is it possible to deal with sets of treatment effects among subsets of the entire population?
- Estimate of treatment effect heterogeneity needed for optimal decision-making

Definition 1 (Athey and Imbens, 2015; 2016)

1. Estimating heterogeneity by features in causal effects in experimental or observational studies
2. Conduct inference about the magnitude of the differences in the treatment effects across subsets of the population

Causal Inference Framework

- Causal inference in observational studies:
  - As we saw previously, assuming unconfoundedness to hold, we can treat observations as having come from a randomized experiment
  - Therefore we can define the conditional average treatment effect (CATE) as follows:

$$\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$$

  - The population average treatment effect then is:

$$\tau^p = E[Y_i(1) - Y_i(0)] = E[\tau(X_i)]$$

Why is CATE important?

- There are a variety of reasons that reasercher wish to conduct estimation and inference on $\tau(x)$:

  1. It my be used to assign future units to their optimal treatment (in presence of different levels of the treatment):

  $$W_i^{opt} = max\, \tau(X_i)$$

  2. If we don't pre-specify the sub-populations it can be the case that the overall effect is negative, but it can be positive on subpopulations, then:

  $$W_i^{PTE} = \mathbf{1}_{\tau(X_i) \geq 0}$$

  e.g.: treatment is a drug $\rightarrow$ prescribe it just to those who benefit from it

Using Trees to Estimate Causal Effects

Athey and Imbens (2015; 2016) propose 3 different approaches:

- Approach I: Analyze two groups separately:
  - Estimate $\hat{\mu}(1, x)$ using dataset where W_i=1
  - Estimate $\hat{\mu}(0, x)$ using dataset where W_i=0
  - Preform within group cross-validation to choose tuning parameters
  - Predict $\hat{\tau} = \hat{\mu}(1, x) - \hat{\mu}(0, x)$
- Approach II: Estimate $\mu(w, x)$ using just one tree:
  - Estimate $\hat{\mu}(1, x)$ and $\hat{\mu}(0, x)$ using just one tree
  - Preform within tree cross-validation to choose tuning parameters
  - Predict $\hat{\tau} = \hat{\mu}(1, x) - \hat{\mu}(0, x)$
  - Estimate is zero for x where tree does not split on w

The CATE Transformation of the Outcome

- The authors' goal is to develop an algorithm that generally leads to an accurate approximation of $\hat{\tau}$ the Conditional Average Treatment Effect.
  1. Ideally we would measure the quality of the approximation in terms of goodness of fit using the MSE:
  $$Q^{infeas} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0) - \hat{\tau}(X_i))^2$$
  2. We can address this problem of infeasibiliy by transforming the outcome using the treatment indicator W_i and $e(X)$:
  $$Y_i^* = Y_i^{obs} \cdot \frac{W_i - e(X_i)}{(1 - e(X_i)) \cdot e(X_i)}$$
  3. Then:
  $$E[Y_i^* | X_I = x] = \tau(x)$$

How to estimate the In-Sample Goodness of fit?

- The ideal goodness of fit measure would be:

$$Q^{infeas}(\hat{\tau}) = \mathbb{E}[(\tau_i - \hat{\tau}(X_i))^2].$$

- A useful proxy that can be used for the goodness of fit measure is:

$$\mathbb{E}[\tau_i^2 | X_i \in S_j] = \frac{1}{N} \sum_i \hat{\tau}(x_i)^2.$$

This leads to our In-sample goodness of fit function:

$$Q^{is} = -\frac{1}{N} \sum_i \hat{\tau}(x_i)^2.$$

Transformed Outcome Tree Model

- Approach 3:

1. Model and Estimation
    - Model Type: Tree structure
    - Estimator $\hat{\tau}_i^{TOT}$: sample average treatment effect within leaf
2. Criterion function (for fixed tuning parameter λ)
    - In-sample Goodness-of-fit function:

$$Q^{is} = -MSE = -\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i^{TOT})^2$$

    - Structure and use of criterion:

$$Q^{crit} = Q^{is} - \lambda \times leaves$$

    - Select member of set of candidate estimators that maximizes Q^{crit}, given λ
3. Cross-validation approach
    - Out-of-Sample Goodness-of-fit function:

$$Q^{oos} = -MSE = -\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i^{TOT} - Y_i^*)^2$$

    - Approach: select tuning parameter λ with highest Q^{os}

Critique to the TOT approach

- Transformation of the Outcome in a randomized set-up:

$$Y_i^* = Y_i^{obs} \cdot \frac{W_i - p}{(1-p) \cdot p} = \begin{cases} \frac{1}{p} \cdot Y_i^{obs} & if \;\; W_i = 1 \\ -\frac{1}{1-p} \cdot Y_i^{obs} & if \;\; W_i = 0 \end{cases}$$

- Within a leaf the sample average of Y_i^* is not the most efficient estimator of treatment effect
- The proportion of treated units within the leaf is not the same as the overall sample proportion
- We use a weighted estimator similar to the Hirano, Imbens and Ridder (2003) estimator

Causal Tree Approach

- In details the Treatment Effect in a generic leaf $\mathbb{X}_j$ is:

$$\tau^{CT}(X_i) = \frac{\sum_{j:X_j \in \mathbb{X}_j} Y_i^{obs} \cdot \frac{W_i}{\hat{e}(X_i)}}{\sum_{j:X_j \in \mathbb{X}_j} \frac{W_i}{\hat{e}(X_i)}} - \frac{\sum_{j:X_j \in \mathbb{X}_j} Y_i^{obs} \cdot \frac{(1-W_i)}{(1-\hat{e}(X_i))}}{\sum_{j:X_j \in \mathbb{X}_j} \frac{(1-W_i)}{(1-\hat{e}(X_i))}}$$

- This estimator is a consistent estimator of:

$$\tau_{\mathbb{X}_j} = \mathbb{E}[Y_i(1) - Y_i(0) | X_i \in \mathbb{X}_j]$$

- The variance can be estimated the Neyman estimator:

$$\hat{\mathbb{V}}_{Neyman} = \frac{s_t^2}{N_t} + \frac{s_c^2}{N_c}$$

These two quantities can be estimated as:

$$s_{t,j}^{te,2} = \frac{1}{N_t - 1} \sum_{i:W_i=1} (Y_i(1) - \overline{Y}_t^{obs})^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} (Y_i - \overline{Y}_t^{obs})^2$$

$$s_{c,j}^{te,2} = \frac{1}{N_c - 1} \sum_{i:W_i=0} (Y_i(0) - \overline{Y}_c^{obs})^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} (Y_i - \overline{Y}_c^{obs})^2$$

Attractive features of Causal trees

1. Can easily separate tree construction from treatment effect estimation
2. Tree costructed on training sample is indipendent of sampling variation in the test sample
3. Holding tree from training sample fixed, can use standard methods to conduct inference within each leaf of the tree on test sample
4. Can use any valid method for treatment effect estimation, not just the methods used in training
5. Simulations run by the authors show that the Causal Tree Algorithm overperforms the ST, TT and TOT approaches

Case Study

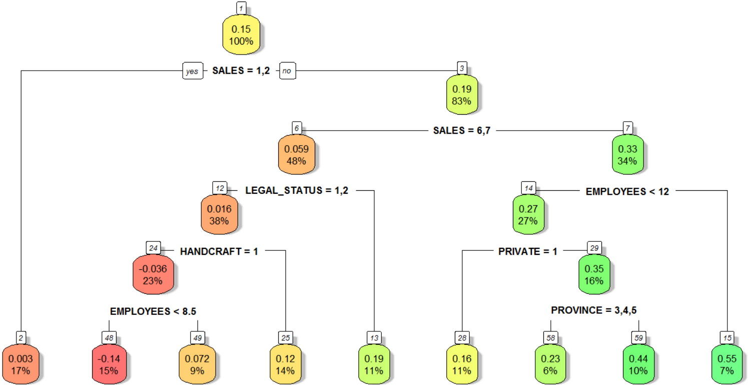


Figure: Bargagli-Stoffi & Gnecco (2019)

Conclusions

1. The main problem to face is the absence of a *ground truth* when we deal with causal inference problems
2. The approach developed is strongly data-driven: selection of subpopulation is optimized by the algorithm
3. The appoach is tailored for applications where:
    1. there may be many attribute relative to the number of units observed (*fat-data*)
    2. the functional form of the relationship between treatment effecs and the attributes of units ins not known

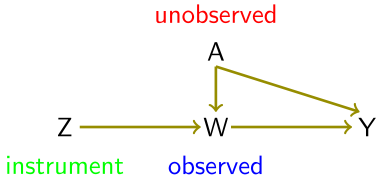
Further Readings

- S.Athey, G.Imbens *Machine learning methods for estimating heterogeneous causal effects,* 2015
- S.Athey, S.Wager *Estimation and Inference of Heterogeneous Treatment Effects using Random Forest,* 2015
- L. Breiman. *Random Forest,* Machine learning, 24:123-140, 2001
- L. Breiman, J.H. Olshen, C.J. Stone. *Classification and Regression Trees,* CRC press, 1984
- T.J. Hastie, R.J. Tibshirani, J.H. Friedman. *The Elements of Statistical Learning.* Speringer, New York, 2009
- K.P. Murphy. *Machine Learning. A Probabilistic Perspective.* The MIT Press, Cambridge, Massachusetts, 2012

Machine Learning and Causality
Using LASSO to select instruments

Instrumental Variable intuition

- Instrumental variable regression is a very powerful tool in causal inference since it gives the researcher the possibility to solve at the same time 3 issues:
  1. Omitted Variable Bias
  2. Simultaneous equation models (aka reverse causality)
  3. Measurement error
- The idea behind IV



- IV regression can isolate the causal effect of W by means of an instrument Z, which detect movement in W uncorrelated to D

Formalization and definitions

- $Y_i = \beta^T X_i + \rho s + \eta_i$ where $\eta_i = A_i^T \gamma + \epsilon_i$ and $\mathbb{E}[\eta_i] = A_i^T \gamma$
- $s_i = X_i^T \pi_{10} + \pi_{11} Z_i + \epsilon_{1i}$ is the First Stage
- $Y_i = X_i^T \pi_{20} + \pi_{21} Z_i + \epsilon_{2i}$ is the Reduced Form
- s_i and Y_i are the endogenous variables
- X_i and Z_i are the exogenous variables (X_i are the exogenous covariates)
- From the first stage and the reduced form we have:

$$\rho = \frac{\pi_{21}}{\pi_{11}} = \frac{Cov(Y_i, \tilde{z}_i)}{Cov(s_i, \tilde{z}_i)}$$

Angrist and Krueger (1991): on economic return to education

- Most states want student to enter school in the calendar year in which they turn 6
- Group A: children born in the 4th quarter enter school shortly before they turn 6
- Group B: children born in the 1st quarter enter school at around age 6.5
- Law requires students to remain in school until their 16th birthday
- Therefore, A and B will be in different grades, or have a different length of schooling, when the turn 16

First Stage

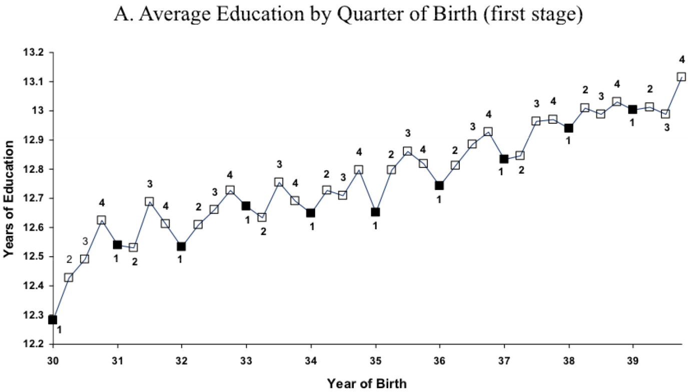


Figure: Average education by quarter of birth. Men born earlier in the calendar year tend to have lower average schooling levels

Reduced Form

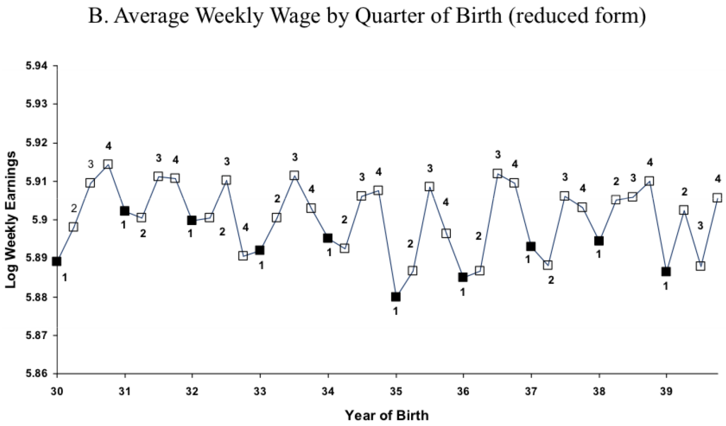


Figure: Average earning by quarter of birth. Man born in earlier quarters earn less, on average, than those born later in the year

Two-Stage Least Squares

- You can obtain the reduced form by substituting the 1st stage into the causal relationship

$$\begin{aligned} Y_i &= \alpha^T X_i + \rho\big[X_i\pi_{10} + \pi_{11}z_i + \epsilon_{1i}\big] + \eta_i \\ &= X_i^T\big[\alpha + \rho\pi_{10}\big] + \rho\pi_{11}z_i + \big[\rho\epsilon_{1i} + \eta_i\big] \\ &= X_i^T\pi_{20} + \pi_{21}z_i + \epsilon_{2i} \end{aligned}$$

- Note that this shows again $\rho = \frac{\pi_{21}}{\pi_{11}}$
- As we usually work with samples, we compute
  - $\hat{s}_i = X_i^T\hat{\pi}_{10} + \hat{\pi}_{11}z_i$: first-stage fitted values
  - $Y_i = \alpha^T X_i + \rho\hat{s}_i + \big[\eta_i + \rho(s_i - \hat{s}_i)\big]$: second-stage equation
- The resulting estimator is consistent for ρ because both X_i and $\hat{s}_i$ are uncorrelated with η_i as well as with $(s_i - \hat{s}_i)$
- Intuition: 2SLS retains only the variation in s_i that is generated bt exogenous quasi-experimental variation

Multiple Instrument Case

- Say that we have three instruments z_{1i}, z_{2i}, z_{3i} e.g., dummies for quarter of birth
- The 1st stage is then $s_i = X_i^T \pi_{10} + \pi_{11} z_{1i} + \pi_{12} z_{2i} + \pi_{31} z_{3i} + \epsilon_{1i}$
- The 2nd stage is the same, though $\hat{s}_i$ is now from the above
- Angrist and Krueger (1991) also include interaction terms

$$\begin{aligned} s_i &= X_i^T \pi_{10} + \pi_{11} z_{1i} + \pi_{12} z_{2i} + \pi_{13} z_{3i} \\ &+ \sum_j (B_{ij} z_{1i}) k_{1j} + \sum_j (B_{ij} z_{2i}) k_{2j} + \sum_j (B_{ij} z_{3i}) k_{3j} + \epsilon_{1i} \end{aligned}$$

where B_{ij} is a dummy for year of birth, for $j = 1931 - 39$

Using multiple instruments

Table 4.1.1: 2SLS estimates of the economic returns to schooling

	OLS		2SLS					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education	0.075 (0.0004)	0.072 (0.0004)	0.103 (0.024)	0.112 (0.021)	0.106 (0.026)	0.108 (0.019)	0.089 (0.016)	0.061 (0.031)
Covariates:								
Age (in quarters)								✓
Age (in quarters) squared								✓
9 year of birth dummies		✓			✓	✓	✓	✓
50 state of birth dummies		✓			✓	✓	✓	✓
Instruments:			dummy for QOB=1	dummy for QOB=1 or QOB=2	dummy for QOB=1	full set of QOB dummies	full set of QOB dummies int. with year of birth dummies	full set of QOB dummies int. with year of birth dummies

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the the Angrist and Krueger (1991) 1980 Census sample. This sample includes native-born men, born 1930-1939, with positive earnings and non-allocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses.

Problems with (multiple) instruments

- Bound, Jaeger, and Baker (1995) note, that a possible problem with IV is caused by the selection of *weak* instruments
- Namely, weak instruments are poor predictors of the endogenous question predictor in the first-stage equation
- Weak instrument lead to bias and very large variance in the IV causal estimands
- Moreover,the weak instrument bias tends to get worse as we add more (weak instruments)
- In other words, the bias gets worse when there are many over-identifying restrictions (many instruments compared to endogenous regressors)

Selection problem

We need to select, among the different possible instruments, the ones that have the higher *predictive power* in the First Stage regression

Possible solution: LASSO regression

- Example taken from (Belloni and Chernozhukov, 2011) using data from (Angrist and Krueger, 1991)
- Model of the form

$$\begin{aligned} y_i &= \theta_0 + x_i\theta_1 + \underline{c}_i^T\underline{\gamma} + u_i\,, \quad \mathbb{E}\left\{u_i|\underline{c}_i, \underline{z}_i\right\} = 0 && (1) \\ x_i &= \underline{z}_i^T\underline{\beta} + \underline{c}_i^T\underline{\delta} + v_i\,, \quad \mathbb{E}\left\{v_i|\underline{c}_i, \underline{z}_i\right\} = 0 && (2) \end{aligned}$$

  where, for each person i, y_i indicates wage, x_i denotes education, $\underline{c}_i$ indicates a vector of control variables, and $\underline{z}_i$ denotes a vector of instrumental variables that affect education but do not directly affect the wage
- u_i and v_i are error terms
- In the specific problem, x and u are correlated, hence the OLS estimate of θ_1 from equation (1) - which does not use the vector of instrumental variables - is biased

Application of the LASSO to instrumental variable selection

- The vector of instrumental variables can be used to obtain an unbiased estimate of θ_1, e.g., through the following two-stage regression procedure:
  - first stage: regression of the x_i's from equation (2), using the $\underline{c}_i$'s and the $\underline{z}_i$'s
  - second stage: regression of the y_i's from equation (1), using the $\underline{c}_i$'s and the estimates of the x_i's obtained in the first stage
- In this context, the LASSO can be used to do instrumental variable selection, possibly improving the estimate of θ_1 (see the numerical results in (Belloni and Chernozhukov, 2011))
- A similar application of LASSO - this time in control variable selection - is done in (Belloni et al., 2014), using data from (Acemoglu et al. 2001), to do control variable selection when estimating the effect of institutions on output, using mortality rates for early European settlers as an instrument for institution quality

Other possible applications of ML in Causal Inference

- Combining Bayesian inference, machine learning and causal inference (Jennifer Hill, 2011; Hahn et al. 2017; Nethery et al., 2018; Bargagli-Stoffi et al., 2019+)
- Causal Inference with random forests in randomized experiments (Wager and Athey, 2018) and in observational studies (Athey et al., 2019)
- Heterogeneous effects in IV settings (Guber and Farbmacher, 2018; Bargagli-Stoffi and Gnecco; 2018; 2019; Johnson et al., 2019)
- Heterogeneous effects with network interference (Forastiere et al., 2019+)
- Causal rules (Lee et al., 2018; 2019)
- Personalized treatment (Kallus, 2017; 2018)

Bibliography

- Alexandre Belloni and Victor Chernozhukov, "High Dimensional Sparse Econometric Models: An Introduction," in Inverse Problems and High-Dimensional Estimation, vol. 203 of the series Lecture Notes in Statistics, pp. 121-156, 2011
- Trevor Hastie, Robert Tibshirani, and Martin Wainwright, "Statistical Learning with Sparsity: The Lasso and Generalizations," CRC Press, 2015 (Chapters 2 and 11)
- Ryan Tibshirani and Larry Wasserman, "Sparsity and the LASSO," Lecture notes for the course "Statistical Machine Learning," Carnegie Mellon University, Spring 2015
- Joshua D. Angrist and Alan B. Krueger, "Does Compulsory School Attendance Affect Schooling and Earnings?," Quarterly Journal of Economics, vol. 106, pp. 979-1014, 1991
- Robert Tibshirani, "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society, Series B, volume 58, pp. 267-288, 1996