Building a Robot Judge: Data Science for Decision-Making

8. Instrumental Variables

Recap: Machine Learning Pitfalls

https://padlet.com/eash44/7lqkwn5zezyo9ata

Learning Objectives

1. Implement and evaluate machine learning pipelines.
2. **Implement and evaluate causal inference designs.**
   - **Today: Instrumental Variables**
3. Understand how (not) to use data science tools (ML and CI) to support expert decision-making.

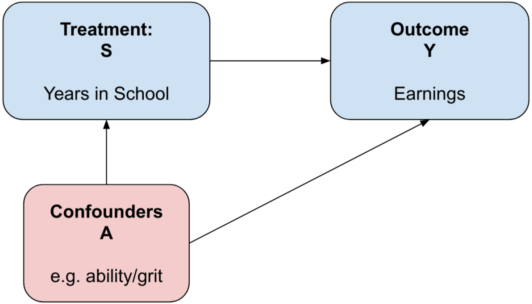
Objectives in an Empirical Project

1. Research question
2. Data
3. Econometrics:
    - Articulate a research design and the identification assumptions for procuring causal estimates.
    - Run regressions to produce the estimates.
    - Run identification checks and specification checks to enhance confidence in results.

Outline

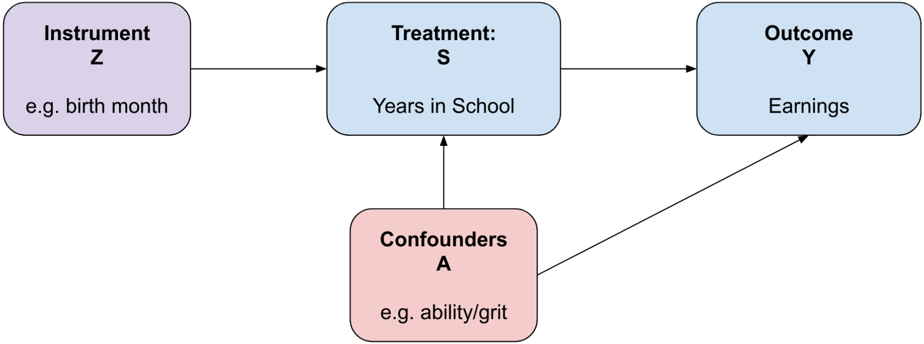
- Example from Week 2: Causal effect of schooling S_i on earnings Y_i.
- There is an unobserved confounder (say ability A_i) correlated with schooling and earnings

$$Y_i = \alpha + \rho S_i + \epsilon_i$$
$$Y_i = \alpha + \rho S_i + \phi \underbrace{A_i}_{\text{unobs}} + \eta_i$$

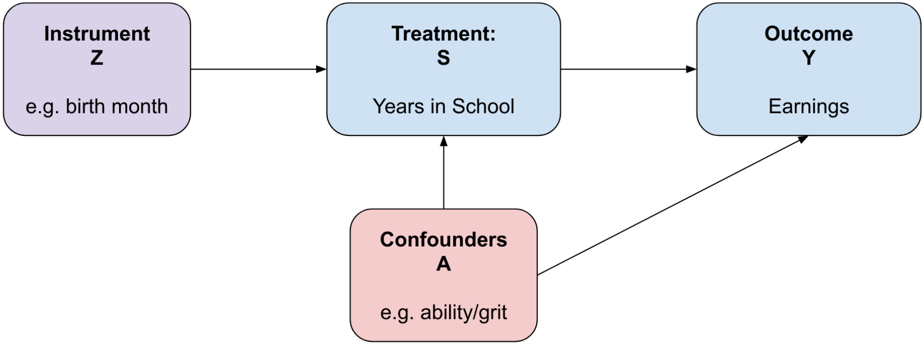


- OLS estimates for $\hat{\rho}$ will be biased.

Instrumental Variable (IV): a variable Z_i, that is correlated with S_i, but not correlated with anything else affecting Y_i.



Instrumental Variable (IV): a variable Z_i, that is correlated with S_i, but not correlated with anything else affecting Y_i.



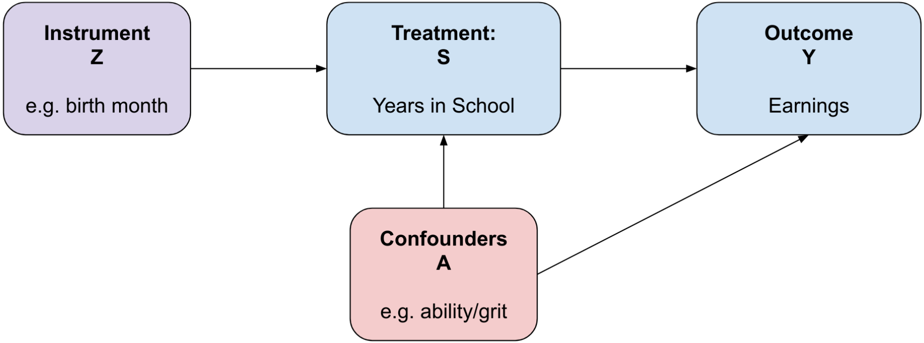
$$Y_i = \alpha + \rho S_i + \underbrace{(+\phi A_i)}_{\text{unobserved}} + \eta_i$$

- Valid instrument Z_i means

$$\text{Cov}[Z_i, S_i] \neq 0, \text{Cov}[Z_i, A_i] = 0$$

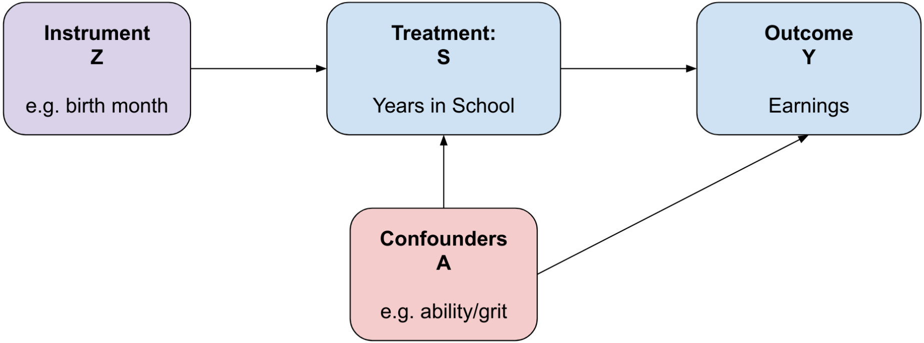
  - With a valid instrument, can procure causal estimates for $\hat{\rho}$

Instrumental Variables: Main Intuition



- We identify a source of variation in treatment assignment that is as good as random – orthogonal to any relevant unobserved confounder.
- We compare individuals that, due to the instrument, are shifted between the control group and treatment group.

What is a valid instrumental variable?



1. Correlated with the causal variable, e.g. S_i:

$$\text{Cov}[Z_i, S_i] \neq 0$$

2. Uncorrelated with any other determinants of outcome Y:

$$\text{Cov}[Z_i, \epsilon_i] = 0$$

IV Identification requirement has two dimensions:

(1) Exogeneity: No unobserved factors affect both the outcome and the instrument:

$$\epsilon_i \nrightarrow Z_i$$

- **No "Z-confounders"**

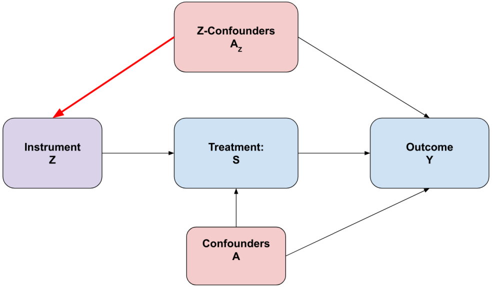
IV Identification requirement has two dimensions:

(1) Exogeneity: No unobserved factors affect both the outcome and the instrument:

$$\epsilon_i \nrightarrow Z_i$$

- **No "Z-confounders"**

Violation of exogeneity:



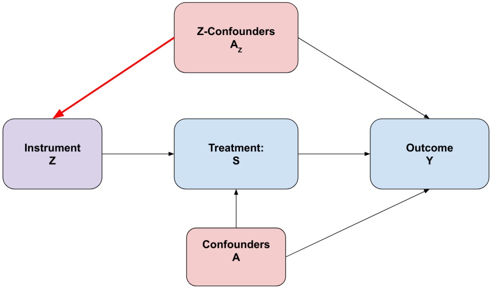
IV Identification requirement has two dimensions:

(1) Exogeneity: No unobserved factors affect both the outcome and the instrument:

$$\epsilon_i \nrightarrow Z_i$$

- **No "Z-confounders"**

Violation of exogeneity:



(2) Exclusion: Instrument only affects outcome through treatment variable:

$$Z_i \nrightarrow \epsilon_i$$

- **"single mediator" condition**

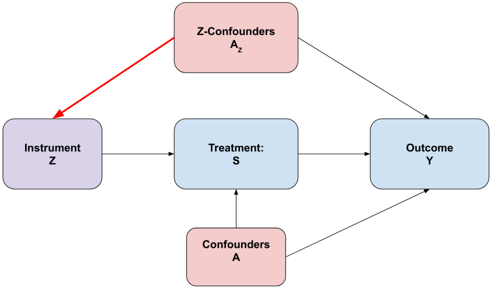
IV Identification requirement has two dimensions:

(1) Exogeneity: No unobserved factors affect both the outcome and the instrument:

$$\epsilon_i \nrightarrow Z_i$$

- **No "Z-confounders"**

Violation of exogeneity:

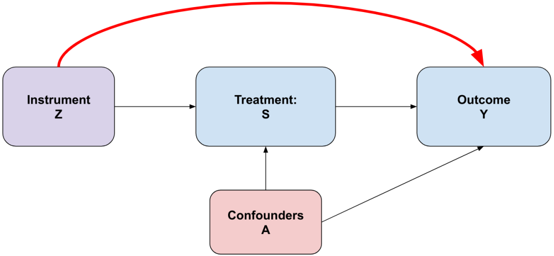


(2) Exclusion: Instrument only affects outcome through treatment variable:

$$Z_i \nrightarrow \epsilon_i$$

- **"single mediator" condition**

Violation of exclusion:



Good instruments are hard to find

- Good instruments come from a combination of three ingredients:
  - Good institutional knowledge
  - Economic theory
  - Last but not least: Originality

Good instruments are hard to find

- Good instruments come from a combination of three ingredients:
  - Good institutional knowledge
  - Economic theory
  - Last but not least: Originality
- Some usual sources of instruments:
  - Nature (e.g. genes, weather)
  - Assignment rules (e.g. random assignment of judges to cases)
  - 'Natural' experiments (e.g. the quarter of birth, conscription lottery, electoral timing...)

Good instruments for schooling

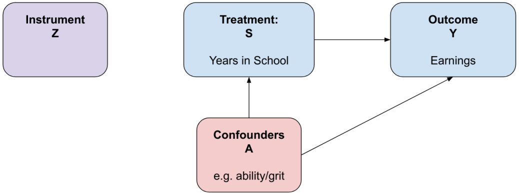
Say we want to estimate the effect of schooling on earnings. Which of the following would make a good instrument for schooling? Explain why or why not.

1. Winning a lottery-assigned scholarship increases the likelihood of attendance.
2. Randomness in the weather reduces years spent in school.
3. Higher standardized test scores increase your chance of getting into college.
4. Being conscripted into the army by lottery reduces years in school.
5. Higher geographical proximity to college increases chances of going to college.
6. Month of birth $\rightarrow$ being born right before the age cutoff increases years in school.

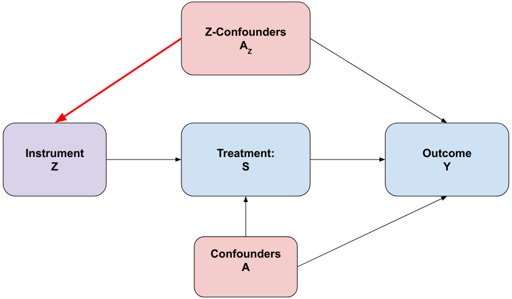
Good instruments for schooling

Good instruments for schooling

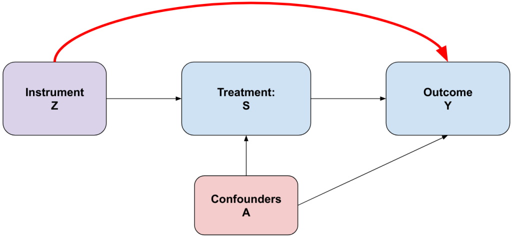
Violation of relevance:



Violation of exogeneity:



Violation of exclusion:



IV estimator

We have

$$Y_i = \alpha + \rho S_i + \epsilon_i$$

and an instrument Z_i where $\text{Cov}[Z_i, S_i] \neq 0$ and $\text{Cov}[Z_i, \epsilon_i] = 0$.

IV estimator

We have

$$Y_i = \alpha + \rho S_i + \epsilon_i$$

and an instrument Z_i where $\text{Cov}[Z_i, S_i] \neq 0$ and $\text{Cov}[Z_i, \epsilon_i] = 0$.

- We can write ρ in terms of the population moments

$$\text{Cov}[Z_i, Y_i] = \rho\text{Cov}[Z_i, S_i] + \underbrace{\text{Cov}[Z_i, \epsilon_i]}_{=0}$$

IV estimator

We have

$$Y_i = \alpha + \rho S_i + \epsilon_i$$

and an instrument Z_i where $\text{Cov}[Z_i, S_i] \neq 0$ and $\text{Cov}[Z_i, \epsilon_i] = 0$.

- We can write ρ in terms of the population moments

$$\text{Cov}[Z_i, Y_i] = \rho\text{Cov}[Z_i, S_i] + \underbrace{\text{Cov}[Z_i, \epsilon_i]}_{=0}$$

- Thus:

$$\rho = \frac{\text{Cov}[Z_i, Y_i]}{\text{Cov}[Z_i, S_i]}$$

with sample estimate

$$\hat{\rho}_{\text{IV}} = \frac{\sum_{i=1}^{n} Z_i Y_i}{\sum_{i=1}^{n} Z_i S_i}$$

```
from linearmodels.iv import IV2SLS
eq = "wages ~ 1 + [schooling ~ instrument] + C(fixed_effect)"
iv = IV2SLS.from_formula(eq,data=df).fit()
```

Examples

Look at papers if curious

- Immigration
  - Networks of immigrants (Card 1991)
- Does police decrease crime?
  - Electoral cycles (Levitt 1997)
- The impact of violent movies on crime
  - Blockbuster movies (Dahl and DellaVigna 2009)
- The effect of preschool television exposure on standardized test scores during adolescence:
  - Gentzkow and Shapiro 2008
- The Potato's Contribution to Population and Urbanization:
  - Nunn and Nancy Qian 2011
- Influence of mass media on U.S. government response to natural disasters
  - Eisensee and Strömberg 2007

Practice: Adding Instruments to Custom Causal Graphs

http://bit.ly/BRJ-W7-graphs-doc

Two-Stage Least Squares (2SLS)

IV estimates are equivalent to running two separate OLS regressions:

1. Estimate"first stage", regressing treatment on instrument:

$$S_i = \gamma Z_i + \nu_i$$

Two-Stage Least Squares (2SLS)

IV estimates are equivalent to running two separate OLS regressions:

1. Estimate"first stage", regressing treatment on instrument:

$$S_i = \gamma Z_i + \nu_i$$

2. Form prediction $\hat{S}_i = \hat{\gamma} Z_i$ and estimate the"second stage", regressing outcome on first-stage-predicted treatment:

$$Y_i = \rho \hat{S}_i + \epsilon_i$$

2SLS Matrix Notation compared to OLS

- With model $Y = X'\beta + U$ and instrument Z, we have

$$\beta_{OLS} = (X'X)^{-1}(X'Y)$$
$$\beta_{IV} = (Z'X)^{-1}(Z'Y)$$

2SLS Matrix Notation compared to OLS

- With model $Y = X'\beta + U$ and instrument Z, we have

$$\beta_{OLS} = (X'X)^{-1}(X'Y)$$
$$\beta_{IV} = (Z'X)^{-1}(Z'Y)$$

$$\mathbb{E}[\beta_{OLS}] = \mathbb{E}[(X'X)^{-1}(X'Y)] = \mathbb{E}[(X'X)^{-1}(X'(X'\beta + \underbrace{U}_{\text{confounders}}))]$$
$$= \beta + \underbrace{\mathbb{E}[(X'X)^{-1}(X'U)]}_{\text{OLS bias}}$$

$$\mathbb{E}[\beta_{IV}] = \mathbb{E}[(Z'X)^{-1}(Z'Y)] = \mathbb{E}[(Z'X)^{-1}(Z'(X'\beta + U))]$$
$$= \beta + \underbrace{\mathbb{E}[(Z'X)^{-1}(Z'U)]}_{\text{2SLS bias}}$$

2SLS Matrix Notation compared to OLS

- With model $Y = X'\beta + U$ and instrument Z, we have

$$\beta_{OLS} = (X'X)^{-1}(X'Y)$$
$$\beta_{IV} = (Z'X)^{-1}(Z'Y)$$

$$\mathbb{E}[\beta_{OLS}] = \mathbb{E}[(X'X)^{-1}(X'Y)] = \mathbb{E}[(X'X)^{-1}(X'(X'\beta + \underbrace{U}_{\text{confounders}}))]$$
$$= \beta + \underbrace{\mathbb{E}[(X'X)^{-1}(X'U)]}_{\text{OLS bias}}$$

$$\mathbb{E}[\beta_{IV}] = \mathbb{E}[(Z'X)^{-1}(Z'Y)] = \mathbb{E}[(Z'X)^{-1}(Z'(X'\beta + U))]$$
$$= \beta + \underbrace{\mathbb{E}[(Z'X)^{-1}(Z'U)]}_{\text{2SLS bias}}$$

- which estimate is more biased?

$$\mathbb{E}[(X'X)^{-1}(X'U)] \gtrless \mathbb{E}[(Z'X)^{-1}(Z'U)]?$$

Can we test validity of IV?

- Is Z_i correlated with causal variable of interest, S_i?
  - YES: check for significance of first stage (first-stage F-statistic)

Can we test validity of IV?

- Is Z_i correlated with causal variable of interest, S_i?
  - YES: check for significance of first stage (first-stage F-statistic)
- Is Z_i uncorrelated with any other determinants of Y_i?
  - Not directly testable – relies on institutional knowledge
  - but often indirect ways to probe exogeneity and exclusion

Weak Instruments

The bias of 2SLS can be written as:

$$\text{plim}\hat{\rho} = \rho + \frac{\text{Corr}[Z, \epsilon]}{\text{Cov}[S, Z]} \cdot \frac{\sigma_\epsilon}{\sigma_S}$$

- When the instrument is weakly correlated with the endogenous regressor, the bias increases.
- Kleibergen-Paap First-stage F-statistic should be higher than 10.

Reduced Form

"Reduced Form" (RF) means regressing the outcome directly on the instrument:

$$Y_i = \alpha + \phi Z_i + \epsilon_i$$

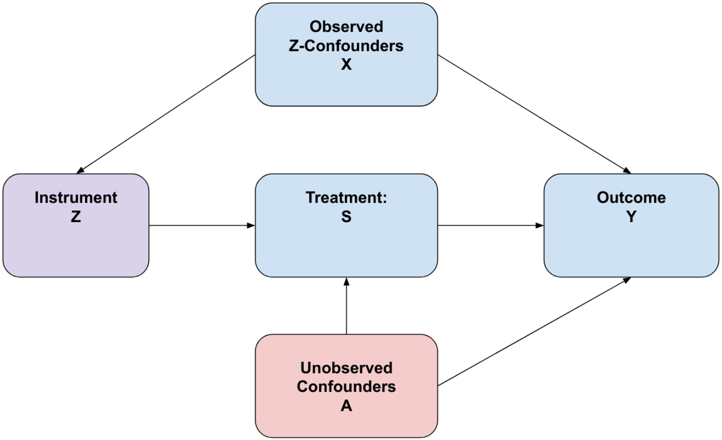
- papers will normally report this along with 2SLS estimates.
- for causal interpretation, RF requires exogeneity but not exclusion.

Instruments with Observed Confounders

- Recall that with OLS, observed confounders are not a problem because we can adjust for them.

Instruments with Observed Confounders

- Recall that with OLS, observed confounders are not a problem because we can adjust for them.
- With Z-confounders, we have the same property.



- IV independence assumption can be written as $\text{Cov}[Z_i, \epsilon_i | X] = 0$.

Fuzzy RD = IV

- **Sharp RD (regression discontinuity):** treatment status is **deterministic/discontinuous** function of running variable (x_i), with cutoff c:

$$Y_i = \alpha + \rho \mathbb{I}[x_i > c] + f(x_i)'\beta + \epsilon_i$$

```
eq = "death_rate ~ above_21 + age + age_squared"
rdd = smf.ols(formula=eq, data=df).fit()
```

Fuzzy RD = IV

- **Sharp RD (regression discontinuity):** treatment status is **deterministic/discontinuous** function of running variable (x_i), with cutoff c:

$$Y_i = \alpha + \rho \mathbb{I}[x_i > c] + f(x_i)'\beta + \epsilon_i$$

```
eq = "death_rate ~ above_21 + age + age_squared"
rdd = smf.ols(formula=eq, data=df).fit()
```

- **Fuzzy RD:** being above threshold increases **probability** of receiving treatment, rather than deterministically changing treatment. Use RD as first stage in 2SLS:

$$D_i = \alpha + \gamma \mathbb{I}[x_i > c] + \eta_i$$
$$Y_i = \alpha + \rho D_i + \epsilon_i$$

  - instrument is a dummy variable for being above cutoff
  - endogenous variable is whether treatment is actually assigned.
  - include polynomials in running variable as covarates.

```
eq = "death_rate ~ age + age_squared + [drinker ~ above_21]"
iv = IV2SLS.from_formula(eq,data=df).fit()
```

Outline

Lasso IV with Weak Instruments

Consider the problem of a sparse first stage:

$$S_i = \alpha + \boldsymbol{Z}_i' \boldsymbol{\phi} + \nu_i$$

- $\boldsymbol{Z}_i$ is a high-dimensional vector
- many elements of $\boldsymbol{\phi} = (\phi_1, ...\phi_{n_z})$ are zero, $\phi_k \approx 0$
- but we don't know which.

Lasso IV with Weak Instruments

Consider the problem of a sparse first stage:

$$S_i = \alpha + \boldsymbol{Z}_i'\boldsymbol{\phi} + \nu_i$$

- $\boldsymbol{Z}_i$ is a high-dimensional vector
- many elements of $\boldsymbol{\phi} = (\phi_1, ...\phi_{n_z})$ are zero, $\phi_k \approx 0$
- but we don't know which.

Solution:

- Train lasso (or elastic net), $S \sim \text{Lasso}(\boldsymbol{Z})$
  - use CV grid search across the whole dataset to select L1 penalty
  - get subset of instruments with non-zero coefficients, Z_{Lasso}.
- Run 2SLS with Z_{Lasso} as instrument(s).
- This is the "optimal" set of instruments under sparsity (Belloni et al 2014).

"Mostly harmless machine learning"

Chen, Chen, and Lewis (2021)

- Can also use double machine learning in the first stage.
  - Form cross-fitted prediction $\hat{S} = F(\boldsymbol{Z})$
  - instrument S with $\hat{S}$
  - $F()$ has to be linear to avoid spurious identification

Heterogeneous Instrument Compliance

- Instruments do not usually affect all individuals equally.
  - e.g., some people won't go to school even if they win a scholarship.

Heterogeneous Instrument Compliance

- Instruments do not usually affect all individuals equally.
  - e.g., some people won't go to school even if they win a scholarship.
  - first stage is driven by "compliers" (responders to instrument).

Heterogeneous Instrument Compliance

- Instruments do not usually affect all individuals equally.
  - e.g., some people won't go to school even if they win a scholarship.
  - first stage is driven by“compliers” (responders to instrument).
- Standard 2SLS estimates give a “local average treatment effect” on the complier population.

Estimating Heterogeneous First Stage

- Can use machine learning to estimate treatment effect heterogeneity in the first stage:

$$S = \gamma(X)Z + \nu$$

Estimating Heterogeneous First Stage

- Can use machine learning to estimate treatment effect heterogeneity in the first stage:
$$S = \gamma(X)Z + \nu$$
- E.g., if instrument is binary, use T-Learner Method (any machine learning model):
  - Learn $\eta_0(X) = \mathbb{E}(S|X, Z = 0)$
  - Learn $\eta_1(X) = \mathbb{E}(S|X, Z = 1)$
- Conditional first stage effect estimate is $\hat{\gamma}(X) = \eta_1(X) - \eta_0(X)$.

Estimating Heterogeneous First Stage

- Can use machine learning to estimate treatment effect heterogeneity in the first stage:
$$S = \gamma(X)Z + \nu$$
- E.g., if instrument is binary, use T-Learner Method (any machine learning model):
  - Learn $\eta_0(X) = \mathbb{E}(S|X, Z = 0)$
  - Learn $\eta_1(X) = \mathbb{E}(S|X, Z = 1)$
- Conditional first stage effect estimate is $\hat{\gamma}(X) = \eta_1(X) - \eta_0(X)$.
- Can be used to analyze complier population, or to re-weight regressions to get closer to an average treatment effect (Coussens and Spiess 2021).

Deep IV = IV + Neural Nets (Hartford, Lewis, Leyton-Brown, and Taddy 2017)

- method uses deep learning to extend 2SLS to high-dimensional settings (many instruments and many endogenous treatment variables).

Deep IV = IV + Neural Nets (Hartford, Lewis, Leyton-Brown, and Taddy 2017)

- method uses deep learning to extend 2SLS to high-dimensional settings (many instruments and many endogenous treatment variables).
- first stage is a multi-outcome neural net, learning a high-dimensional function $\boldsymbol{s} = g(\boldsymbol{z})$ predicting each of the endogenous variables $\boldsymbol{s}$ with the high-dimensional instruments $\boldsymbol{z}$.
- second stage is also a neural net, learning a function $y = f(\boldsymbol{s})$, where $\hat{\boldsymbol{s}}$ is predicted from the first stage neural net.
- implemented by Microsoft's econml package.

Critical Reading Activity: Fox News and COVID-19 Social Distancing

`http://bit.ly/BRJ-W7-FNC-doc`