Instrumental Variables, DeepIV, and Forbidden Regressions

Aaron Mishkin

UBC MLRG 2019W2
Introduction

Goal: Counterfactual reasoning in the presence of unknown confounders.

From the CONSORT 2010 statement [Schulz et al., 2010];
https://commons.wikimedia.org/w/index.php?curid=9841081
Can we draw causal conclusions from observational data?

- **Medical Trials**: Is the new sunscreen I’m using effective?
  - **Confounder**: I live in my laboratory!

- **Pricing**: Should airlines increase ticket prices next December?
  - **Confounder**: NeurIPS 2019 was in Vancouver.

- **Policy**: Will unemployment continue to drop if the Federal Reserve keeps interest rates low?
  - **Confounder**: US shale oil production increases.

We cannot control for confounders in observational data!
Introduction: Graphical Model

We will graphical models to represent our learning problem.

- $X$: observed features associated with a trial.
- $\epsilon$: unobserved (possibly unknown) confounders.
- $P$: the policy variable we will to control.
- $Y$: the response we want to predict.
Introduction: Answering Causal Questions

- **Causal Statements**: Y is caused by P.
- **Action Sentences**: Y will happen if we do P.
- **Counterfactuals**: Given (x, p, y) happened, how would Y change if we had done P instead?
**Introduction: Berkeley Gender Bias Study**

**S:** Gender causes admission to UC Berkeley [Bickel et al., 1975].

**A:** Estimate mapping \( g(p) \) from 1973 admissions records.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Admission</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>G</strong></td>
<td>？</td>
</tr>
<tr>
<td>( g(G) )</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Men</th>
<th>Applications</th>
<th>Admitted</th>
<th>Women</th>
<th>Applications</th>
<th>Admitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>8442</td>
<td>44%</td>
<td></td>
<td>4321</td>
<td>35%</td>
<td></td>
</tr>
</tbody>
</table>
**Introduction: Berkeley with a Controlled Trial**

Through Simpson’s Paradox, controlling for the effects of $D$ shows “small but statistically significant bias in favor of women” [Bickel et al., 1975].
Part 1: “Intervention Graphs”
Intervention Graphs

The \texttt{do(\cdot)} operator formalizes this transformation [Pearl, 2009].

\textbf{Intuition}: effects of forcing $P = p_0$ vs “natural” occurrence.
**Setup**

- $\epsilon, \eta \sim \mathcal{N}(0, 1)$.
- $P = p + 2\epsilon$.
- $g_0(P) = \max \left\{ \frac{P}{5}, P \right\}$.
- $Y = g_0(P) - 2\epsilon + \eta$.

Can supervised learning recover $g_0(P = p_0)$ from observations?

---

Synthetic example introduced by Bennett et al. [2019]
Supervised learning fails because it assumes $P \perp \epsilon$!
Intervention Graphs: Supervised vs Causal Learning

Given dataset $\mathcal{D} = \{p_i, y_i\}_{i=1}^n$:

- **Supervised Learning** estimates the conditional
  \[
  \mathbb{E} [Y \mid P] = g_0(P) - 2\mathbb{E} [\epsilon \mid P]
  \]

- **Causal Learning** estimates the conditional
  \[
  \mathbb{E} [Y \mid \text{do}(P)] = g_0(P) - 2\mathbb{E} [\epsilon]
  \]

  $\Rightarrow 0$
What if

1. all confounders are known and in $\epsilon$;
2. $\epsilon$ persists across observations;
3. the mapping $Y = f(X, P, \epsilon)$ is known and persists.
Steps to inference:

1. **Abduction**: compute posterior $P(\epsilon \mid \{x_i, p_i, y_i\}_{i=1}^n)$
2. **Action**: form subgraph corresponding to $\text{do}(P = p_0)$.
3. **Prediction**: compute $P(Y \mid \text{do}(P = p_0), \{x_i, p_i, y_i\}_{i=1}^n)$. 
Our assumptions are unrealistic since

- identifying all confounders is hard.
- assuming all confounders are “global” is unrealistic.
- characterizing $Y = f(X, P, \epsilon)$ requires expert knowledge.

What we really want is to

- allow any number and kind of confounders!
- allow confounders to be “local”.
- learn $f(X, P, \epsilon)$ from data!
Part 2: Instrumental Variables
...the drawing of inferences from studies in which subjects have the final choice of program; the randomization is confined to an indirect *instrument* (or assignment) that merely encourages or discourages participation in the various programs.

― Pearl [2009]
We augment our model with an *instrumental variable* $Z$ that
- affects the distribution of $P$;
- only affects $Y$ through $P$;
- is conditionally independent of $\epsilon$. 
Intuition: “[F is] as good as randomization for the purposes of causal inference”— Hartford et al. [2017].
IV: Formally

**Goal**: counterfactual predictions of the form

$$E [Y | X, do(P = p_0)] - E [Y | X, do(P = p_1)].$$

Let’s make the following assumptions:

1. the additive noise model $Y = g(P, X) + \epsilon$,

2. the following conditions on the IV:
   2.1 **Relevance**: $p(P | X, Z)$ is not constant in $Z$.
   2.2 **Exclusion**: $Z \perp Y | P, X, \epsilon$.
   2.3 **Unconfounded Instrument**: $Z \perp \epsilon | P$. 
IV: Model Learning Part 1

\[ Y = g(P, X) + \epsilon \]

Under the do operator:

\[
E [Y \mid X, \text{do}(P = p_0)] - E [Y \mid X, \text{do}(P = p_1)] = g(p_0, X) - g(p_1, X) + E [\epsilon - \epsilon \mid X].
\]

\[
\text{\underbrace{E [\epsilon - \epsilon \mid X]}_{=0}}.
\]

So, we only need to estimate \( h(P, X) = g(P, X) + E [\epsilon \mid X]! \)
Want: \( h(P, X) = g(P, X) + \mathbb{E}[\epsilon \mid X] \).

**Approach:** Marginalize out confounded policy \( P \).

\[
\mathbb{E}[Y \mid X, Z] = \int_P (g(P, X) + \mathbb{E}[\epsilon \mid P, X]) \, dp(P \mid X, Z)
\]
\[
= \int_P (g(P, X) + \mathbb{E}[\epsilon \mid X]) \, dp(P \mid X, Z)
\]
\[
= \int_P h(P, X) \, dp(P \mid X, Z).
\]

**Key Trick:** \( \mathbb{E}[\epsilon \mid X] \) is the same as \( \mathbb{E}[\epsilon \mid P, X] \) when marginalizing.
Objective: \[
\frac{1}{n} \sum_{i=1}^{n} \mathcal{L} \left( y_i, \int_{P} h(P, x_i) dp(P \mid z_i) \right).
\]

Two-stage methods:

1. **Estimate Density**: learn \( \hat{\rho}(P \mid X, Z) \) from \( \mathcal{D} = \{ p_i, x_i, z_i \}_{i=1}^{n} \).

2. **Estimate Function**: learn \( \hat{h}(P, X) \) from \( \bar{\mathcal{D}} = \{ y_i, x_i, z_i \}_{i=1}^{n} \).

3. **Evaluate**: counterfactual reasoning via \( \hat{h}(p_0, x) - \hat{h}(p_1, x) \).
IV: Two-Stage Least-Squares

Classic Approach: two-stage least-squares (2SLS).

\[ h(P, X) = w_0^\top P + w_1^\top X + \epsilon \]

\[ \mathbb{E}[P | X, Z] = A_0X + A_1Z + r(\epsilon) \]

Then we have the following:

\[ \mathbb{E}[Y | X, Z] = \int_P h(P, X) dp(P | X, Z) \]

\[ = \int_P \left( w_0^\top P + w_1^\top X \right) dp(P | X, Z) \]

\[ = w_1^\top X + w_0^\top \int_P P dp(P | X, Z) \]

\[ = w_1^\top X + w_0^\top (A_0X + A_1Z) \]

No need for density estimation!

See Angrist and Pischke [2008].
Part 3: Deep IV
Deep IV: Problems with 2SLS

**Problem:** Linear models aren’t very expressive.
- What if we want to do causal inference with time-series?

---

Deep IV: Problems with 2SLS

**Problem**: Linear models aren’t very expressive.
- How about complex image data?

https://alexgkendall.com/computer_vision/bayesian_deep_learning_for_safe_ai/
Deep IV: Approach

Remember our objective function:

\[
\text{Objective} : \quad \frac{1}{n} \sum_{i=1}^{n} \mathcal{L} \left( y_i, \int_{P} h(P, x_i) dP \big| z_i \right) .
\]


1. **Treatment Network**: estimate \( \hat{p}(P \mid \phi(X, Z)) \).

   - **Categorical** \( P \): softmax w/ favourite architecture.
   - **Continuous** \( P \): autoregressive models (MADE, RNADE, etc.), normalizing flows (MAF, IAF, etc) and so on.

2. **Outcome Network**: fit favorite architecture \( \hat{h}_\theta(P, X) \approx h(P, X) \).

---

Autoregressive models: [Germain et al., 2015, Uria et al., 2013], Normalizing Flows: [Rezende and Mohamed, 2015, Papamakarios et al., 2017, Kingma et al., 2016]
Deep IV: Training Deep IV Models

1. **Treatment Network** “easy” via maximum-likelihood:

\[
\phi^* = \arg \max_{\phi} \left\{ \sum_{i=1}^{n} \log \hat{p}(p_i | \phi(x_i, z_i)) \right\}
\]

2. **Outcome Network**: Monte Carlo approximation for loss:

\[
L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L} \left( y_i, \int_{P} \hat{h}_\theta(P, X) \, d\hat{p}(P | \phi(x_i, z_i)) \right)
\]

\[
\approx \frac{1}{n} \sum_{i=1}^{n} \mathcal{L} \left( y_i, \frac{1}{M} \sum_{j=1}^{m} \hat{h}_\theta(p_j, x_i) \right) := \hat{L}(\theta),
\]

where \( p_j \sim \hat{p}(P | \phi(x_i, z_i)) \).
Deep IV: Biased and Unbiased Gradients

When $L(y, \hat{y}) = (y - \hat{y})^2$:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \int_{P} h(P, x_i) dp(P | z_i) \right)^2.$$ 

If we use a single set of samples to estimate $\mathbb{E}_\hat{p}[\hat{h}_\theta(P, x_i)]$:

$$\nabla \hat{L}(\theta) \approx -2 \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\hat{p} \left[ y_i - \hat{h}_\theta(P, x_i) \nabla_\theta \hat{h}_\theta(P, x_i) \right]$$

$$\geq -2 \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\hat{p} \left[ y_i - \hat{h}_\theta(P, x_i) \right] \mathbb{E}_\hat{p} \left[ \nabla_\theta \hat{h}_\theta(P, x_i) \right] = \nabla_\theta L(\theta),$$

by Jensen’s inequality.
Part 4: Experimental Results and Forbidden Techniques
Results: Price Sensitivity

**Synthetic Price Sensitivity:** $\rho \in [0, 1]$ tunes confounding.
- Customer Type: $S \in \{1, \ldots, 7\};$ Price Sensitivity: $\psi_t$
- $Z \sim \mathcal{N}(0, 1), \quad \eta \sim \mathcal{N}(0, 1)$
- $\epsilon \sim \mathcal{N}(\rho \cdot \eta, 1 - \rho^2)$.
- $P = 25 + (Z + 3)\psi_t + \eta$
- $Y = 100 + (10 + P)S\psi_t - 2P + \epsilon$

$\rho = 0.75 \quad \rho = 0.5 \quad \rho = 0.25 \quad \rho = 0.1$

**Training Sample in 1000s**

Out-of-Sample MSE (log scale)

FFNet 2SLS 2SLS(poly) NonPar DeepIV
Results: Price Sensitivity with Image Features

What if $S$ is an MNIST digit?

![Diagram showing the network structure with $y$, $x$, and $z$ as inputs, and pooling and dense layers with 64 and 32 units.]

![Box plot showing out-of-sample counterfactual MSE for different models: Controlled Experiment, DeepIV, 2SLS, and Naive deep net. The x-axis represents training samples in 1000s, ranging from 1 to 20.]
Results: Any Issues?

Did we do something wrong?
“Forbidden regressions were forbidden by MIT Professor Jerry Hausman in 1975, and while they occasionally resurface in an under-supervised thesis, they are still technically off-limits.”

—Angrist and Pischke [2008]
Let $f$ be some (non-linear) function and consider

$$h(P, X) = w_0^T P + w_1^T X + \epsilon$$

$$\mathbb{E}[P \mid X, Z] = f(X, Z, \epsilon),$$

**Amazing Property:** 2SLS is consistent if $h$ is linear even if $f$ isn’t!

- Prove using **orthogonality** of residual and prediction.

**Deep IV:** bias from $\hat{p}(P \mid \phi(X, Z))$ propagates to $\hat{h}_\theta(P, X)$.

- Asymptotically OK if density estimation is **realizable**.

See [this PDF](#) for a hint on how to proceed.
Recap

Today:

- Our goal was counterfactual reasoning from observations.
- Naive supervised learning can fail catastrophically due to confounders.
- Probabilistic counterfactuals are possible with persistent confounders.
- Instrumental variables allow counterfactual inference when confounders are unknown.
- Deep IV uses instrumental variables with neural networks for flexible counterfactual reasoning.
Questions?


References II


