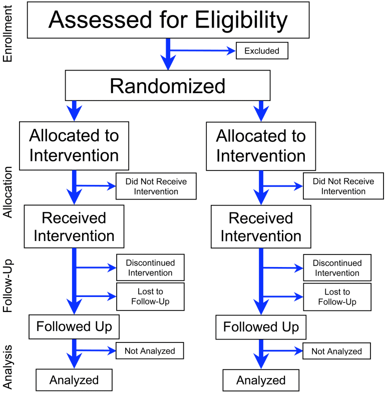
Instrumental Variables, DeepIV, and Forbidden Regressions

Aaron Mishkin

UBC MLRG 2019W2

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

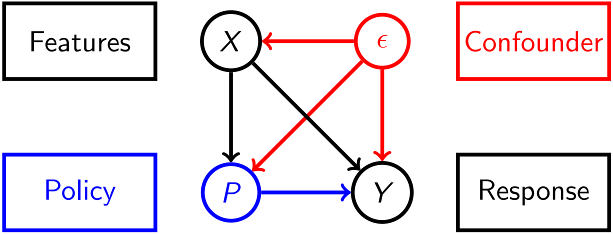
Introduction: Motivation

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!

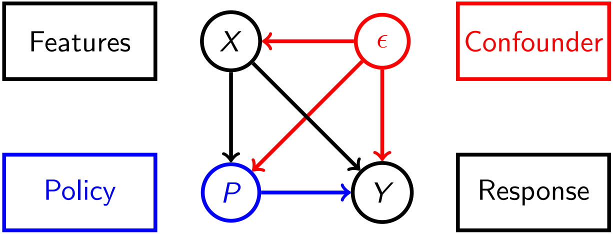
Introduction: Graphical Model



We will graphical models to represent our learning problem.

- X: observed *features* associated with a trial.
- ϵ: unobserved (possibly unknown) *confounders*.
- P: the *policy* variable we will to control.
- Y: the *response* we want to predict.

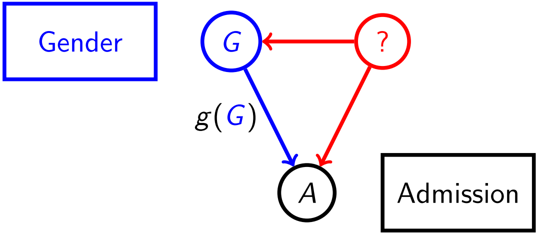
Introduction: Answering Causal Questions



- **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?

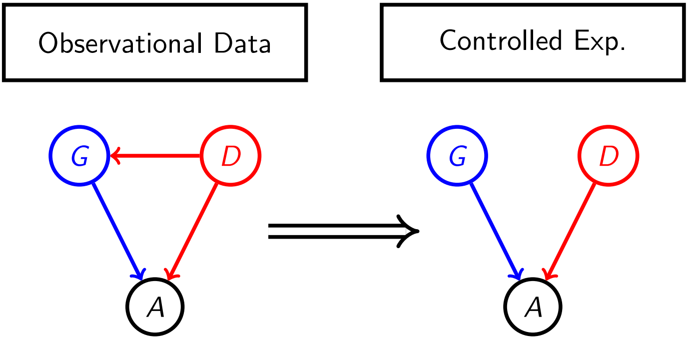
Introduction: Berkeley Gender Bias Study

S: Gender causes admission to UC Berkeley [Bickel et al., 1975].
A: Estimate mapping $g(p)$ from 1973 admissions records.



Men		Women	
Applications	Admitted	Applications	Admitted
8442	44%	4321	35%

Introduction: Berkeley with a Controlled Trial

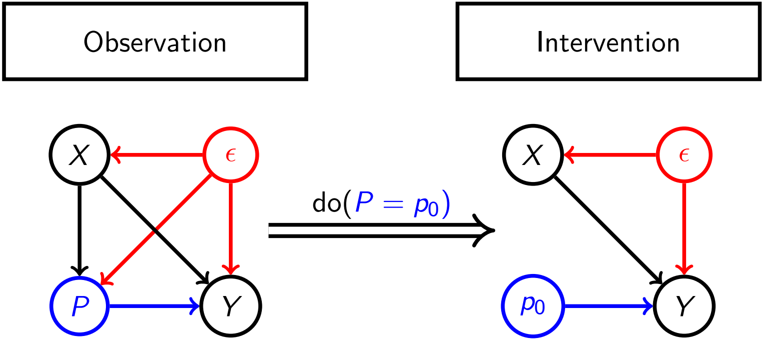


Simpson's Paradox: Controlling for the effects of D shows "small but statistically significant bias in favor of women" [Bickel et al., 1975].

Part 1: "Intervention Graphs"

Intervention Graphs

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



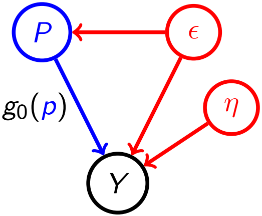
Intuition: effects of forcing $P = p_0$ vs "natural" occurrence.

Intervention Graphs: Supervised vs Causal Learning

Setup

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

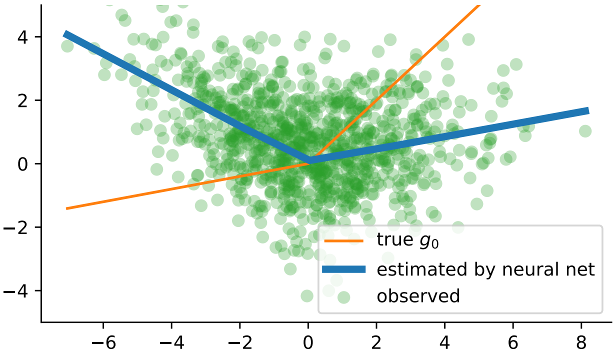
Graphical Model



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

Synthetic example introduced by Bennett et al. [2019]

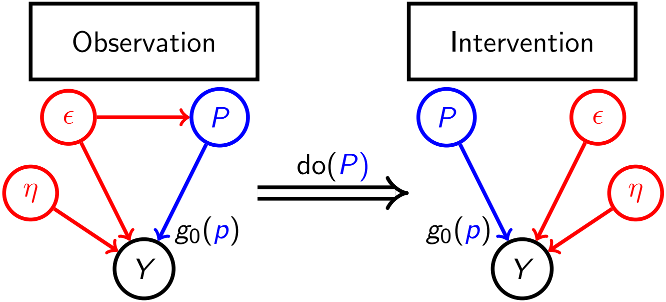
Intervention Graphs: Supervised Failure



Supervised learning fails because it assumes $P \perp\!\!\!\perp \epsilon$!

Taken from https://arxiv.org/abs/1905.12495

Intervention Graphs: Supervised vs Causal Learning



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

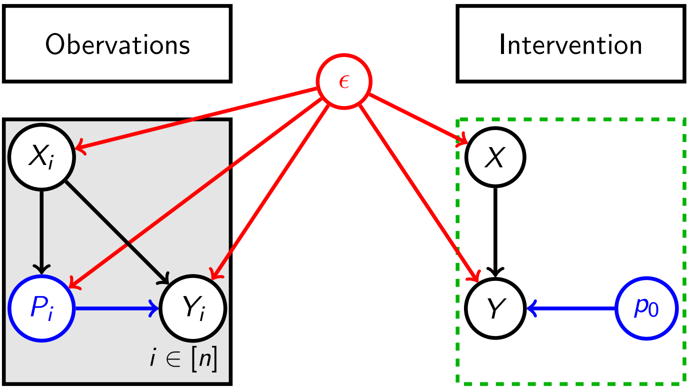
- **Supervised Learning** estimates the conditional

$$\mathbb{E}[Y \mid P] = g_0(P) - 2\mathbb{E}[\epsilon \mid P]$$

- **Causal Learning** estimates the conditional

$$\mathbb{E}[Y \mid \mathrm{do}(P)] = g_0(P) - 2\underbrace{\mathbb{E}[\epsilon]}_{=0}$$

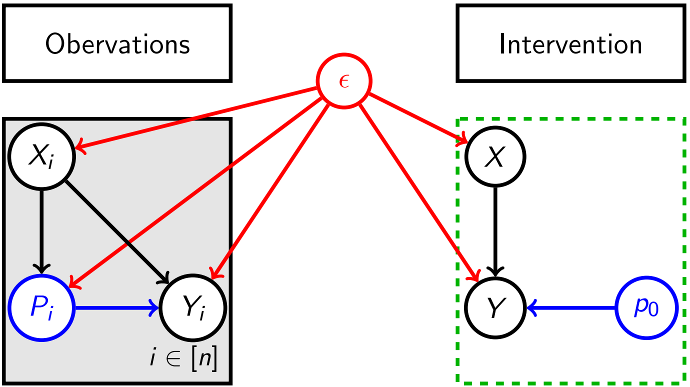
Intervention Graphs: Known Confounders



What if

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

Intervention Graphs: Inference



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 $\mathrm{do}(P = p_0)$.
3. **Prediction**: compute $P(Y \mid \mathrm{do}(P = p_0), \{x_i, p_i, y_i\}_{i=1}^{n})$.

Intervention Graphs: Limitations

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**.

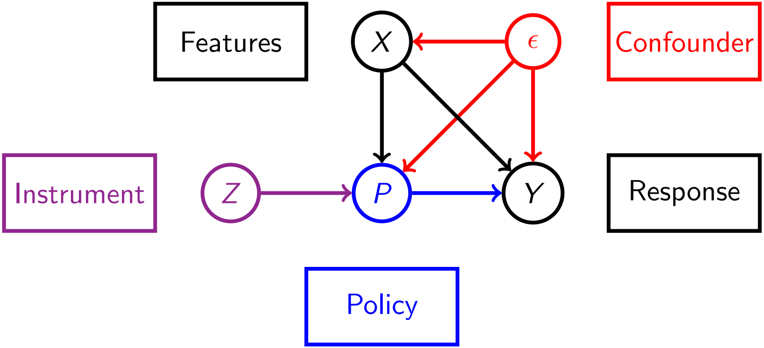
What we really want is to

- 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]

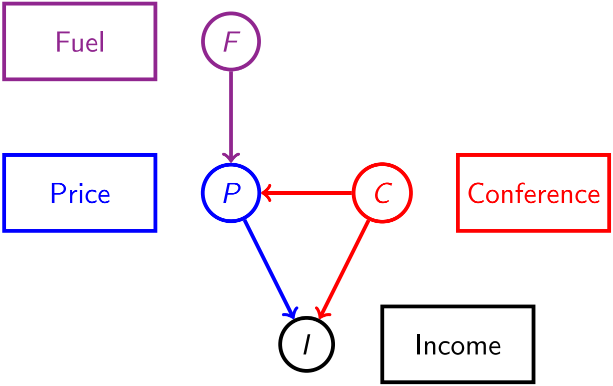
IV: Expanded Model



We augment our model with an *instrumental variable* Z that

- affects the distribution of P;
- only affects Y through P;
- is conditionally independent of ϵ.

IV: Air Travel Example



Intuition: "[F is] as good as randomization for the purposes of causal inference"— Hartford et al. [2017].

IV: Formally

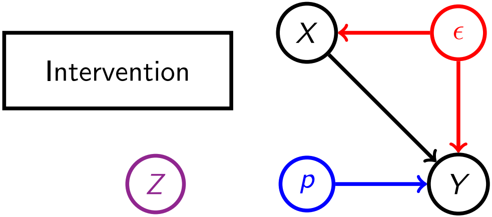
Goal: counterfactual predictions of the form

$$\mathbb{E}\left[Y \mid X, \mathrm{do}(P = p_0)\right] - \mathbb{E}\left[Y \mid X, \mathrm{do}(P = p_1)\right].$$

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**: $p(P \mid X, Z)$ is not constant in Z.
   2.2 **Exclusion**: $Z \perp\!\!\!\perp Y \mid P, X, \epsilon$.
   2.3 **Unconfounded Instrument**: $Z \perp\!\!\!\perp \epsilon \mid P$.

IV: Model Learning Part 1



$Y = g(P, X) + \epsilon$

Under the do operator:

$$\mathbb{E}[Y \mid X, \text{do}(P = p_0)] - \mathbb{E}[Y \mid X, \text{do}(P = p_1)] = g(p_0, X) - g(p_1, X) + \underbrace{\mathbb{E}[\epsilon - \epsilon \mid X]}_{=0}.$$

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

IV: Model Learning Part 2

Want: $h(P, X) = g(P, X) + \mathbb{E}[\epsilon \mid X]$.

Approach: Marginalize out confounded policy P.

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

Key Trick: $\mathbb{E}[\epsilon \mid X]$ is the same as $\mathbb{E}[\epsilon \mid P, X]$ when marginalizing.

IV: Two-Stage Methods

$$\textbf{Objective}: \quad \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(y_i, \int_P h(P, x_i)dp(P \mid z_i)\right).$$

Two-stage methods:

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

$$
\begin{aligned}
h(P, X) &= \mathbf{w}_0^\top P + \mathbf{w}_1^\top X + \epsilon \\
\mathbb{E}[P \mid X, Z] &= \mathbf{A}_0 X + \mathbf{A}_1 Z + r(\epsilon)
\end{aligned}
$$

Then we have the following:

$$
\begin{aligned}
\mathbb{E}[Y \mid X, Z] &= \int_P h(P, X) dp(P \mid X, Z) \\
&= \int_P \left(\mathbf{w}_0^\top P + \mathbf{w}_1^\top X \right) dp(P \mid X, Z) \\
&= \mathbf{w}_1^\top X + \mathbf{w}_0^\top \int_P P dp(P \mid X, Z) \\
&= \mathbf{w}_1^\top X + \mathbf{w}_0^\top (\mathbf{A}_0 X + \mathbf{A}_1 Z).
\end{aligned}
$$

No need for density estimation!

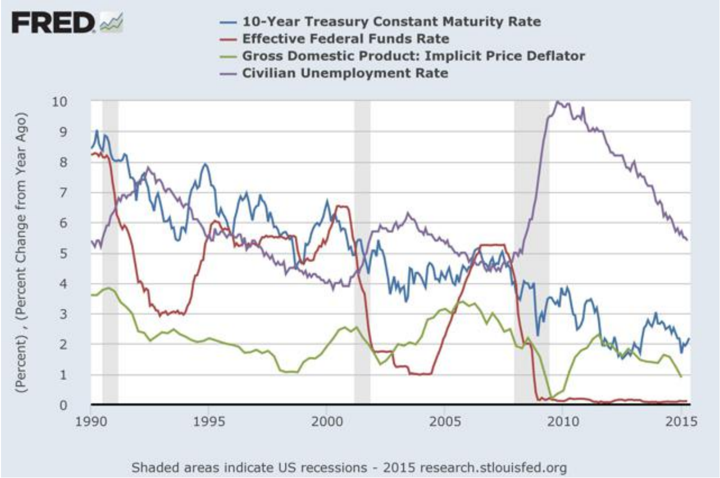
See Angrist and Pischke [2008].

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?

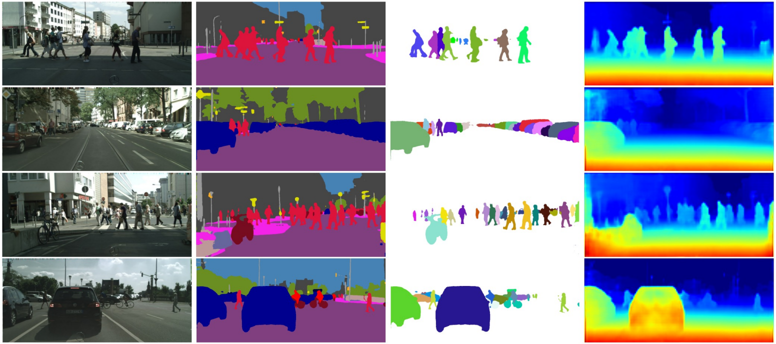


Federal Reserve Economic Research, Federal Reserve Bank of Saint Louis. https://fred.stlouisfed.org/

Deep IV: Problems with 2SLS

Problem: Linear models aren't very expressive.

- How about complex image data?



(a) Input image (b) Segmentation output (c) Instance output (d) Depth output

https://alexgkendall.com/computer_vision/bayesian_deep_learning_for_safe_ai/

Deep IV: Approach

Remember our objective function:

$$\textbf{Objective}: \quad \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(y_i, \int_P h(P, x_i)dp(P \mid z_i)\right).$$

Deep IV: Two-stage method using deep neural networks.

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

$$\hat{h}_\theta(P, X) \approx h(P, X).$$

Autogressive 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:

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

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

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

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

where $p_j \sim \hat{p}\left(P \mid \phi(x_i, z_i) \right)$.

Deep IV: Biased and Unbiased Gradients

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

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

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

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

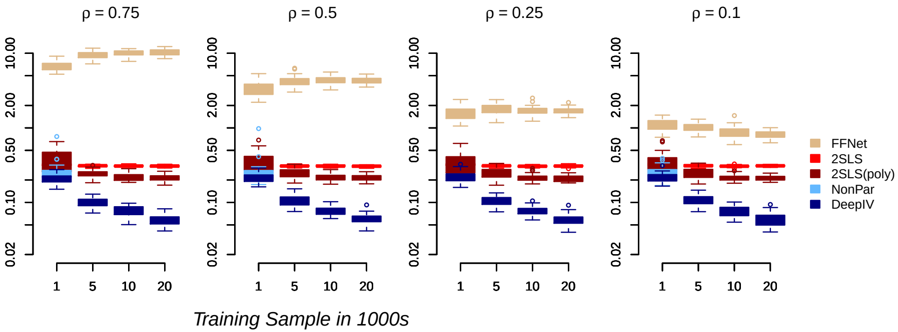
by Jensen's inequality.

Part 4: Experimental Results and Forbidden Techniques

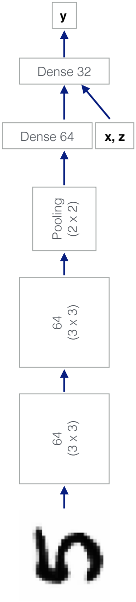
Results: Price Sensitivity

Synthetic Price Sensitivity: $\rho \in [0, 1]$ tunes confounding.

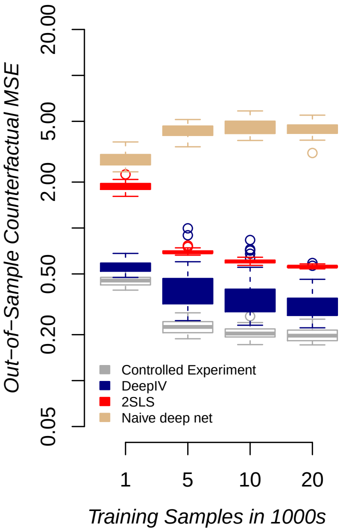
- Customer Type: $S \in \{1, \dots, 7\}$; Price Sensitivity: ψ_t
- $Z \sim \mathcal{N}(0, 1)$, $\quad \eta \sim \mathcal{N}(0, 1)$
- $\epsilon \sim \mathcal{N}(\rho * \eta, 1 - \rho^2)$. $\Longleftarrow$ Important!
- $P = 25 + (Z + 3)\psi_t + \eta$
- $Y = 100 + (10 + P)S\psi_t - 2P + \epsilon$



Results: Price Sensitivity with Image Features



What if S is an **MNIST digit?**



Did we do something wrong?

A Forbidden Regression

"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]

Forbidden Regression: 2SLS vs DeepIV

Let f be some (non-linear) function and consider

$$h(P, X) = \mathbf{w}_0^\top P + \mathbf{w}_1^\top X + \epsilon$$
$$\mathbb{E}[P \mid X, Z] = f\,(X, Z, \epsilon),$$

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**.

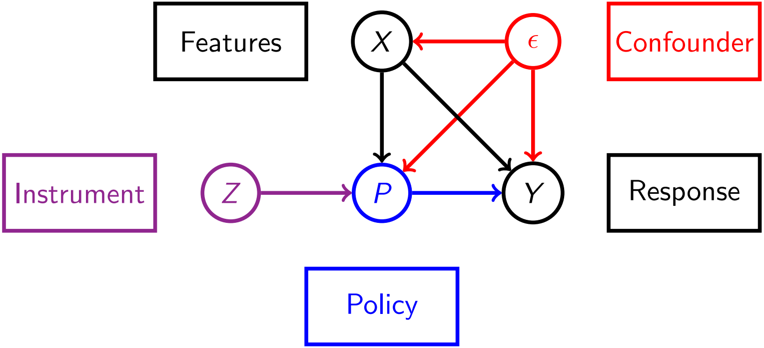
See this PDF for a hint on how to proceed.

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?



Features
X
ϵ
Confounder
Instrument
Z
P
Y
Response
Policy

References I

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems*, pages 3559–3569, 2019.

Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187 (4175):398–404, 1975.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889, 2015.

References II

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1414–1423. JMLR. org, 2017.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

References III

Kenneth F Schulz, Douglas G Altman, and David Moher. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine*, 8(1):18, 2010.

Benigno Uria, Iain Murray, and Hugo Larochelle. Rnade: The real-valued neural autoregressive density-estimator. In *Advances in Neural Information Processing Systems*, pages 2175–2183, 2013.