浙江大學
ZHEJIANG UNIVERSITY

Review
Instrumental Variable
In Causal Inference (1)

Reporter： Anpeng Wu

Diligence / Polite / Insight / Honest

What is cause inference?

Causal Inference: precisely predicting the potential outcomes under ’what-if’ scenarios.

What is cause inference?

Causal Inference: precisely predicting the potential outcomes under 'what-if' scenarios.

If we do …(see a picture, conduct a treatment)… , then …(what would happen)… .

What is cause inference?

Causal Inference: precisely predicting the potential outcomes under 'what-if' scenarios.

If we do …(see a picture, conduct a treatment)… , then …(what would happen)… .

Cause Variables (Causes/Treatments/Interventions)

Effect Variables (Outcomes/Responses/Results)

Identifying Emotions

Angry!!!

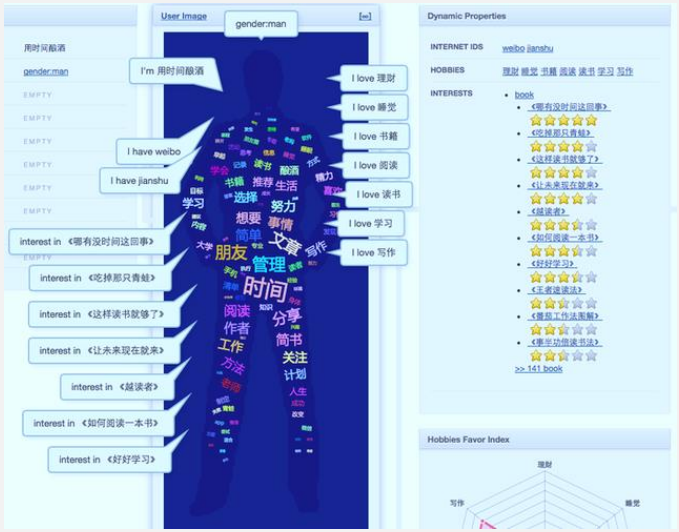
What is cause inference?

Causal Inference: precisely predicting the potential outcomes under ’what-if’ scenarios.

If we do …(see a picture, conduct a treatment)… , then …(what would happen)… .

Cause Variables (Causes/Treatments/Interventions)

Effect Variables (Outcomes/Responses/Results)



Recommend Movies

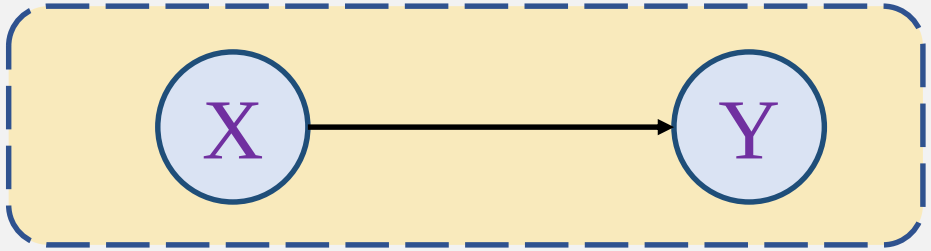
《Spirited Away》

What is cause inference?

Causal Inference: precisely predicting the potential outcomes under 'what-if' scenarios.

If we do …(see a picture, conduct a treatment)… , then …(what would happen)… .

Cause Variables (Causes/Treatments/Interventions) **Effect Variables (Outcomes/Responses/Results)**



Causal Graphs (Structural Causal Models)

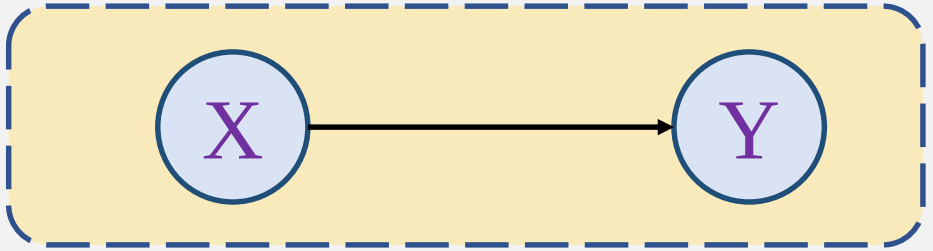
Traditional Supervised Learning: Using Input to Predict Outcome.

What is cause inference?

Causal Inference: precisely predicting the potential outcomes under 'what-if' scenarios.

If we do …(see a picture, conduct a treatment)… , then …(what would happen)… .

Cause Variables (Causes/Treatments/Interventions) **Effect Variables (Outcomes/Responses/Results)**



Causal Graphs (Structural Causal Models)

Are there any Wrongs!!!

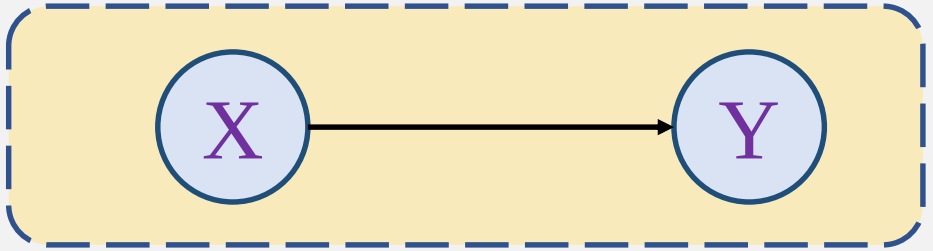
Traditional Supervised Learning: Using Input to Predict Outcome.

Next Question: What is Correlation-based?

Causal Inference: precisely predicting the potential outcomes under 'what-if' scenarios.

If we do …(see a picture, conduct a treatment)… , then …(what would happen)… .

Cause Variables (Causes/Treatments/Interventions) **Effect Variables (Outcomes/Responses/Results)**



Causal Graphs (Structural Causal Models)

Totally Correct!

Traditional Supervised Learning: Using Input to Predict Outcome.
Pattern recognition: If there are no confounding issue, **Correlation = Causality.**

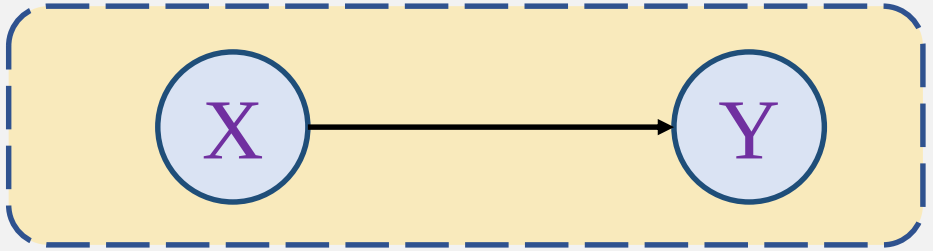
Next Question: What is Correlation-based?

Causal Inference: precisely predicting the potential outcomes under 'what-if' scenarios.

If we do …(see a picture, conduct a treatment)… , then …(what would happen)… .

Cause Variables (Causes/Treatments/Interventions) **Effect Variables (Outcomes/Responses/Results)**

Causality denotes the direct effect.

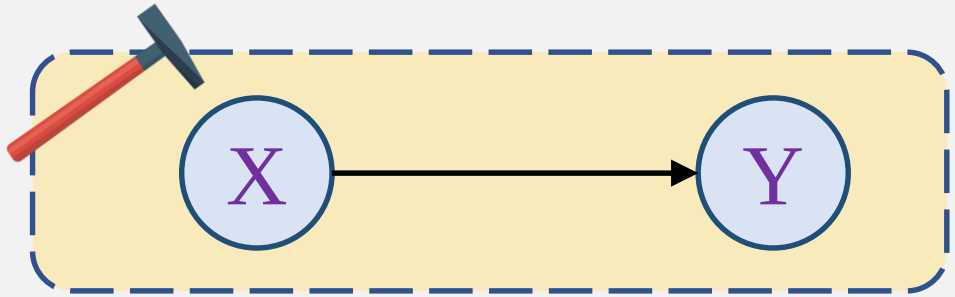


Causal Graphs (Structural Causal Models)

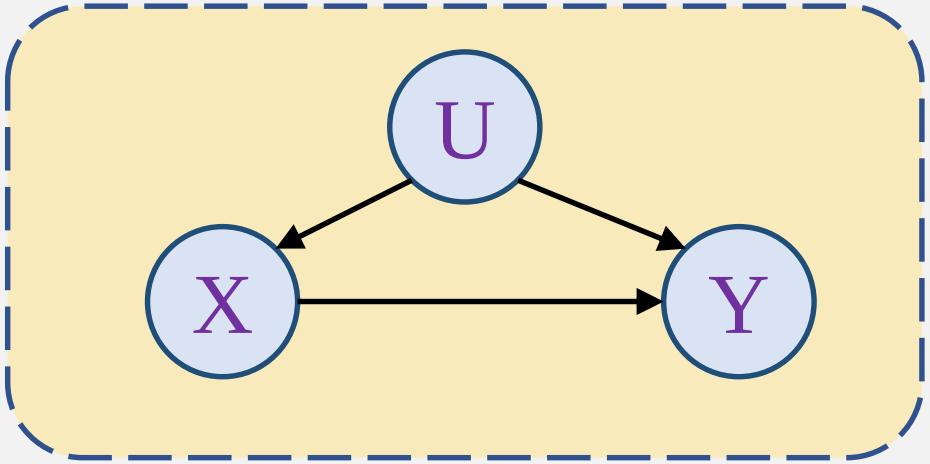
Correlation denotes the total effect.

Traditional Supervised Learning: Using Input to Predict Outcome.
Pattern recognition: If there are no confounding issue, **Correlation = Causality.**

So, why do we study causal inference?



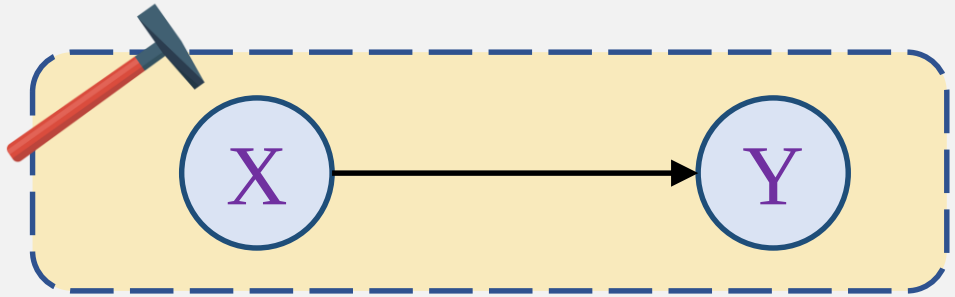
Direct Causal Effect (**Causality**)
All variables has measured.
We can directly use Supervised Learning.



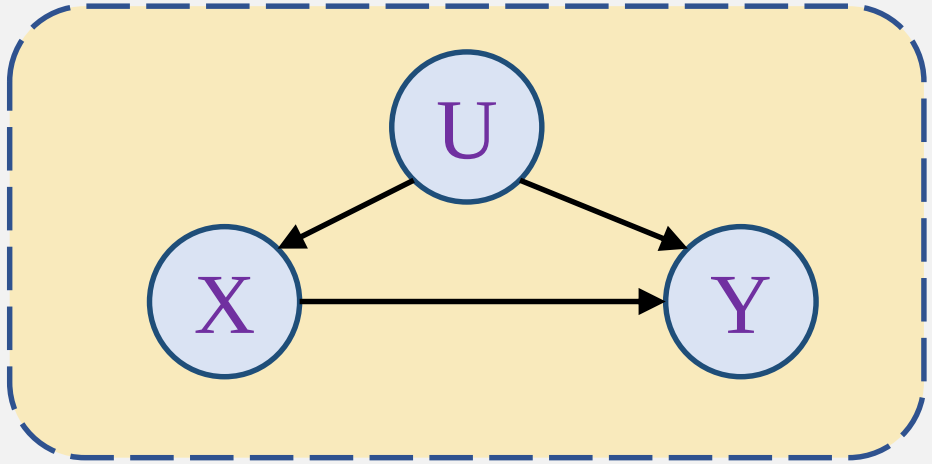
Confounding Causal Effect (**Correlation**)
Some Key Variables is not under Control.

Traditional Supervised Learning: Using Input to Predict Outcome.
Pattern recognition: If there are no confounding issue, **Correlation = Causality.**
Confounding Bias: When Some Key Variables is not under Control, **Correlation ≠ Causality.**

So, why do we study causal inference?



Direct Causal Effect (**Causality**)
All variables has measured.
We can directly use Supervised Learning.

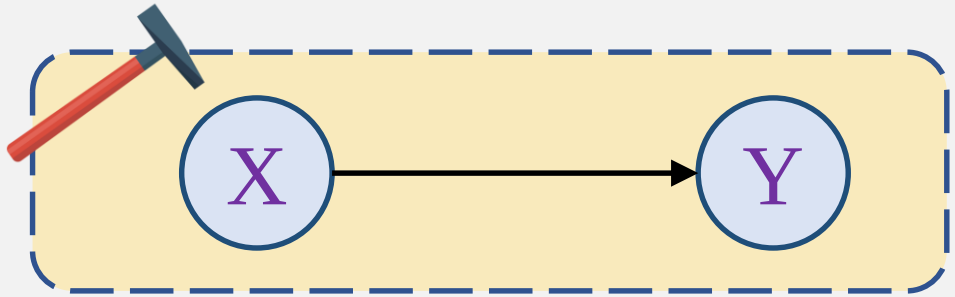


Confounding Causal Effect (**Correlation**)
Some Key Variables is not under Control.

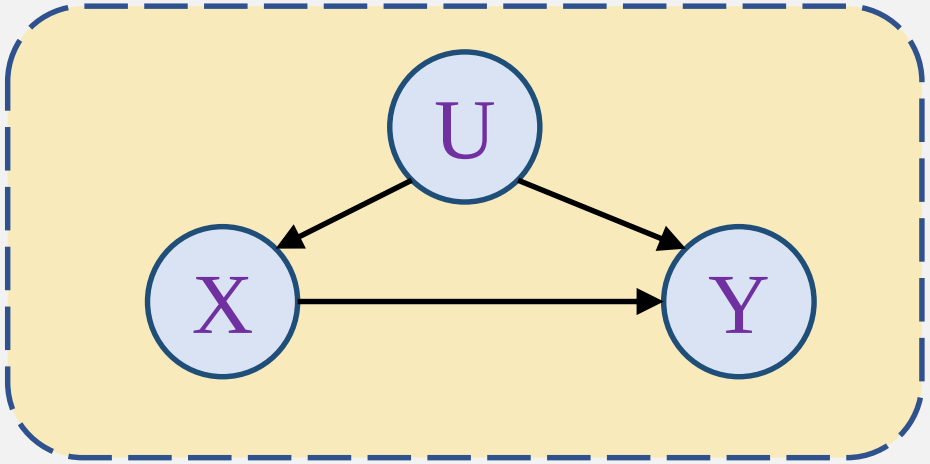
What would happen???

Traditional Supervised Learning: Using Input to Predict Outcome.
Pattern recognition: If there are no confounding issue, **Correlation = Causality.**
Confounding Bias: When Some Key Variables is not under Control, **Correlation ≠ Causality.**

So, why do we study causal inference?



Direct Causal Effect (**Causality**)
All variables has measured.
We can directly use Supervised Learning.



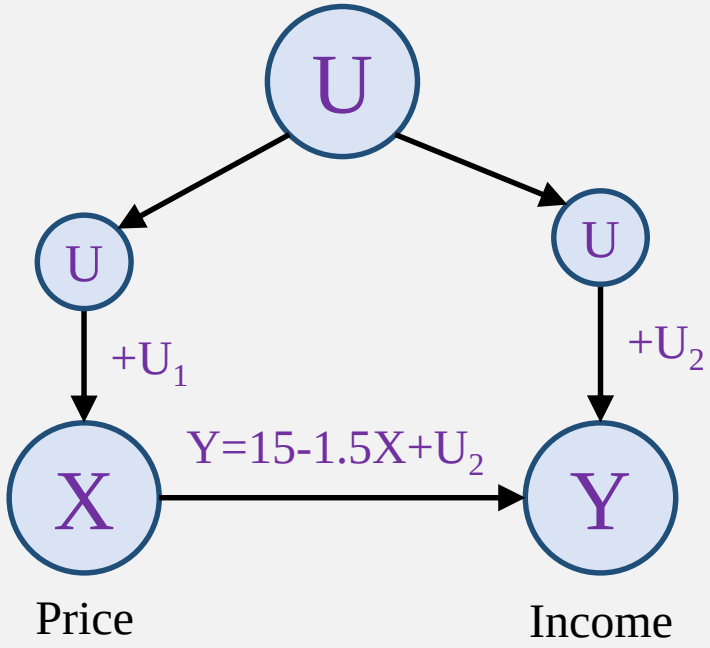
Confounding Causal Effect (**Correlation**)
Some Key Variables is not under Control.

What does the confounding effect or selection effect mean???

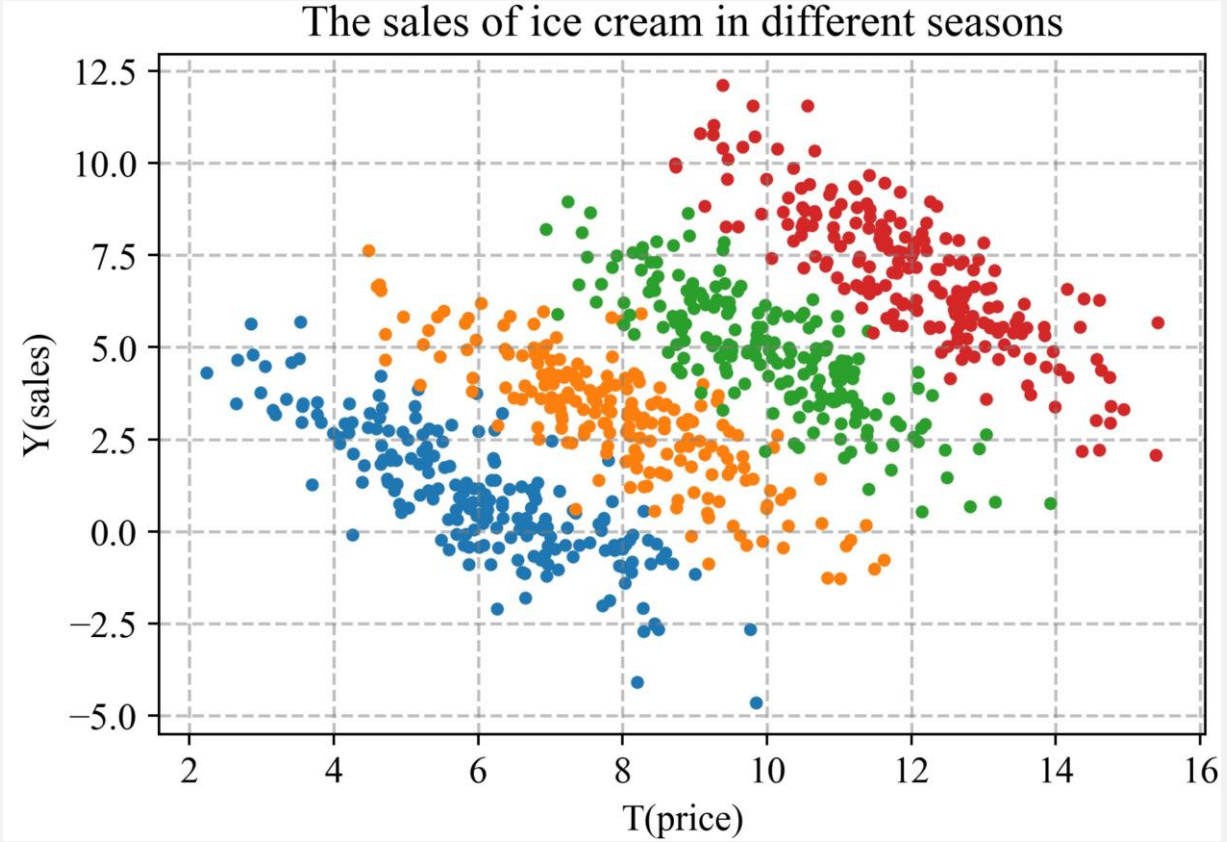
What would happen???

Traditional Supervised Learning: Using Input to Predict Outcome.
Pattern recognition: If there are no confounding issue, **Correlation = Causality.**
Confounding Bias: When Some Key Variables is not under Control, **Correlation ≠ Causality.**

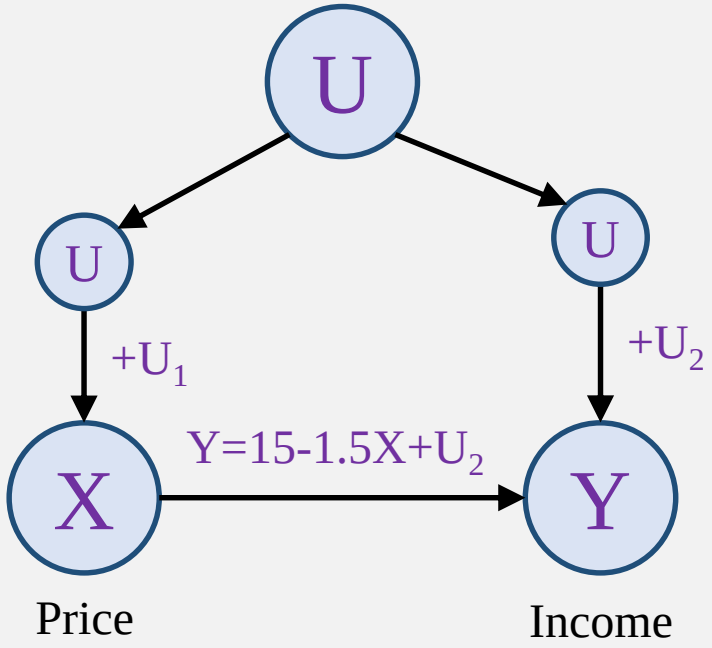
Firstly, confounding effect.



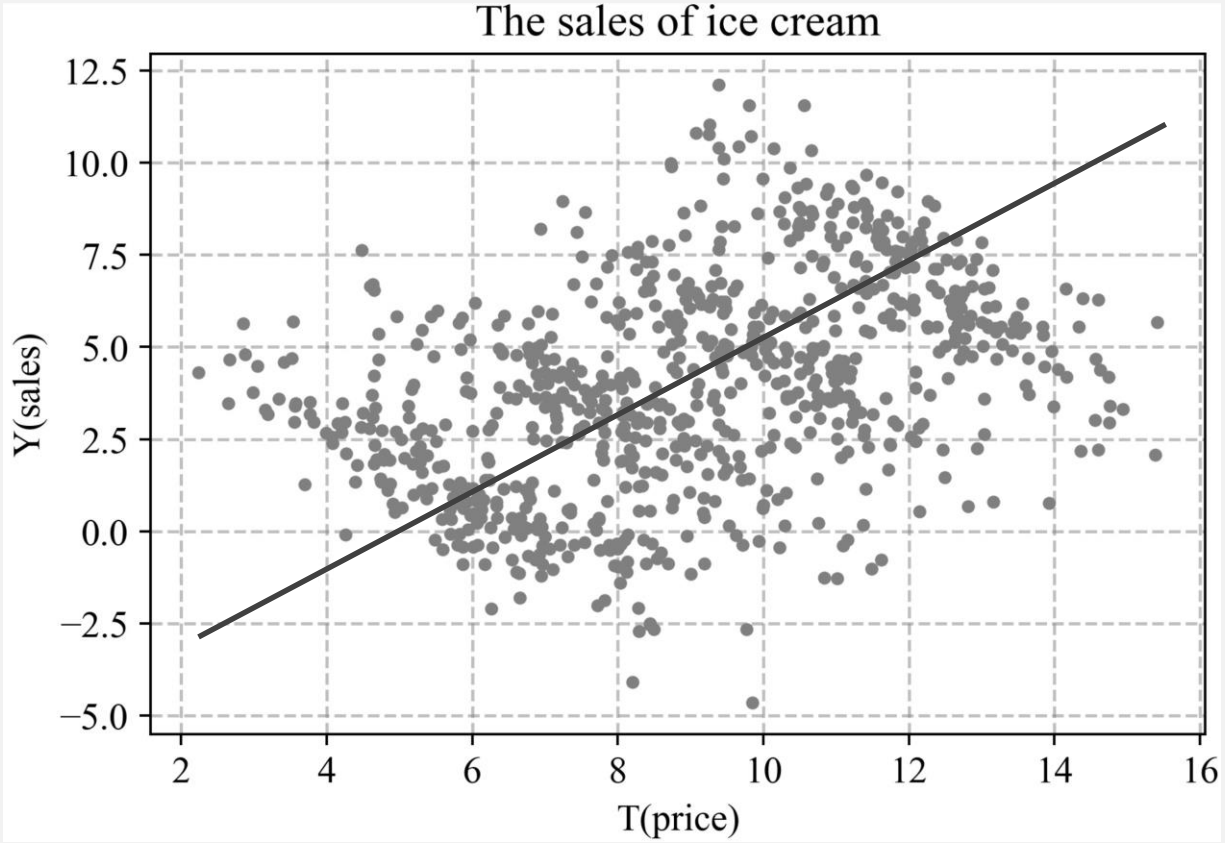
If U = Spring, then Y=10-1.5X



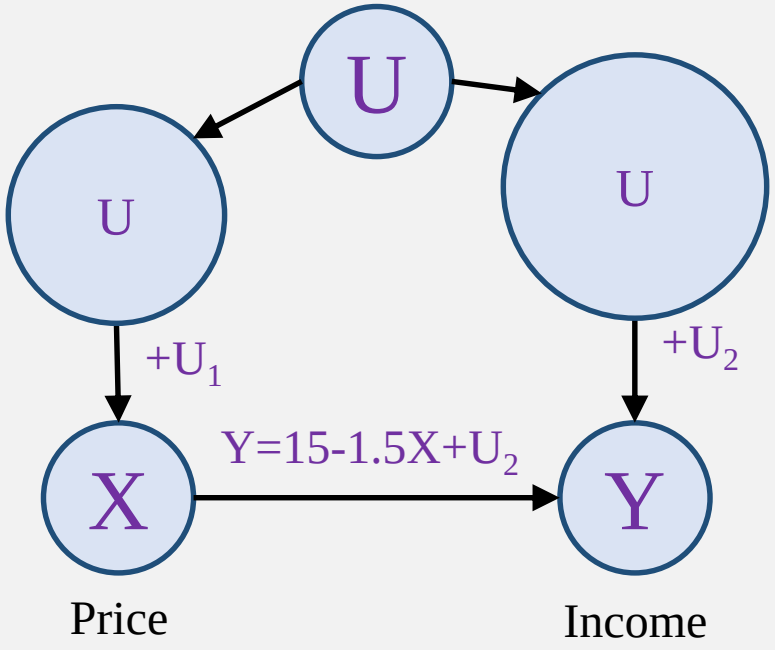
Firstly, confounding effect.



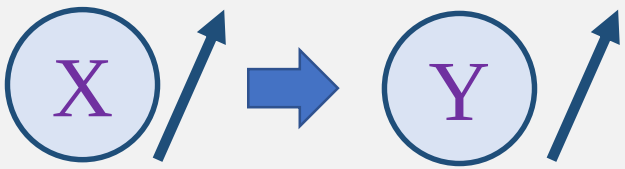
If U = Spring, then Y=10-1.5X

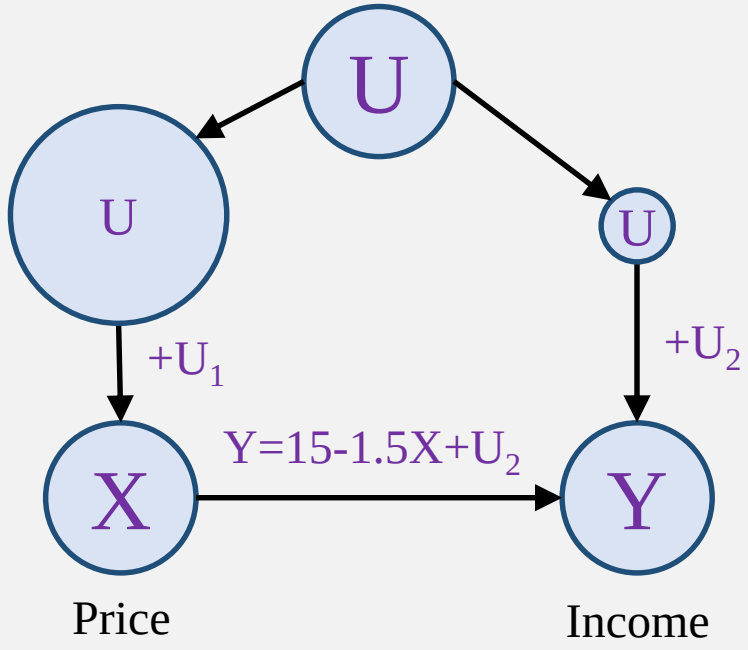


Firstly, confounding effect.

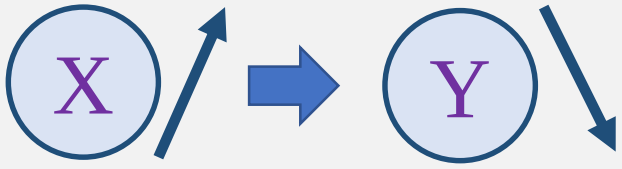


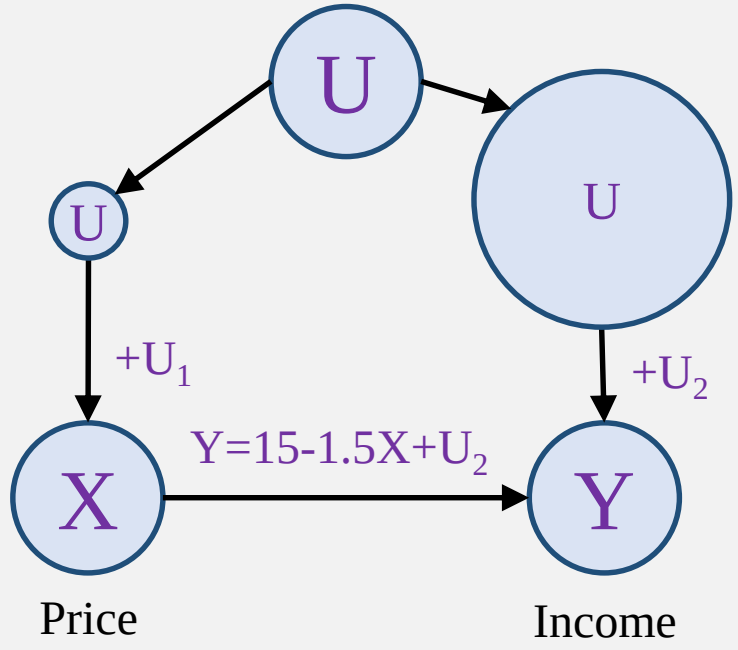
Increase in price, increase in income.



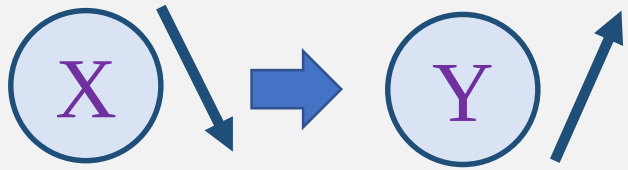


Increase in price, decrease in income.

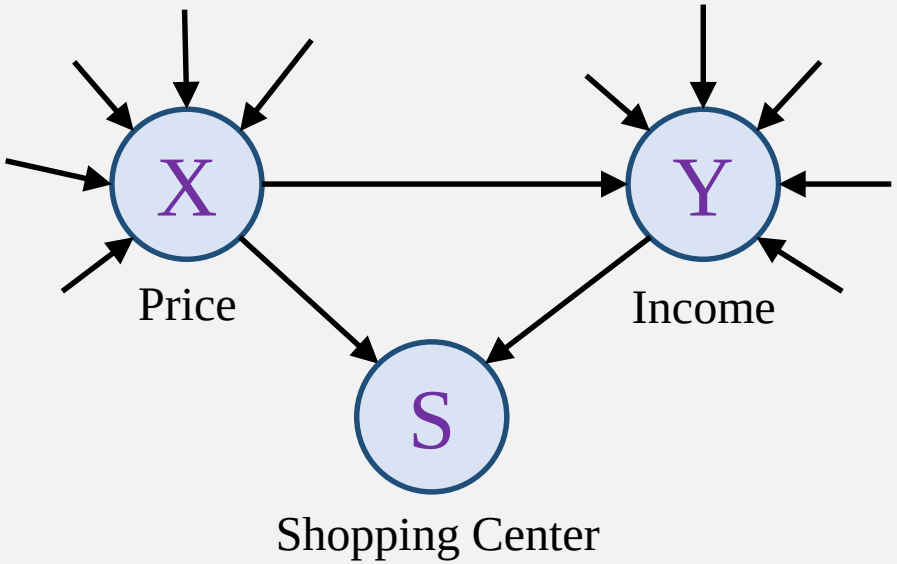




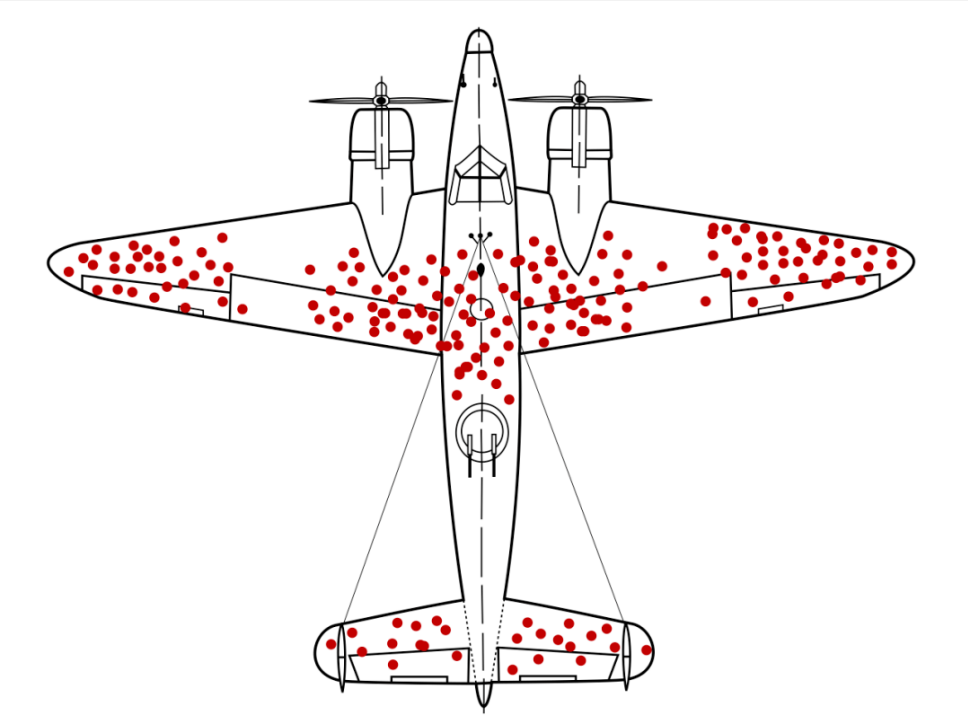
Decrease in price, increase in income.



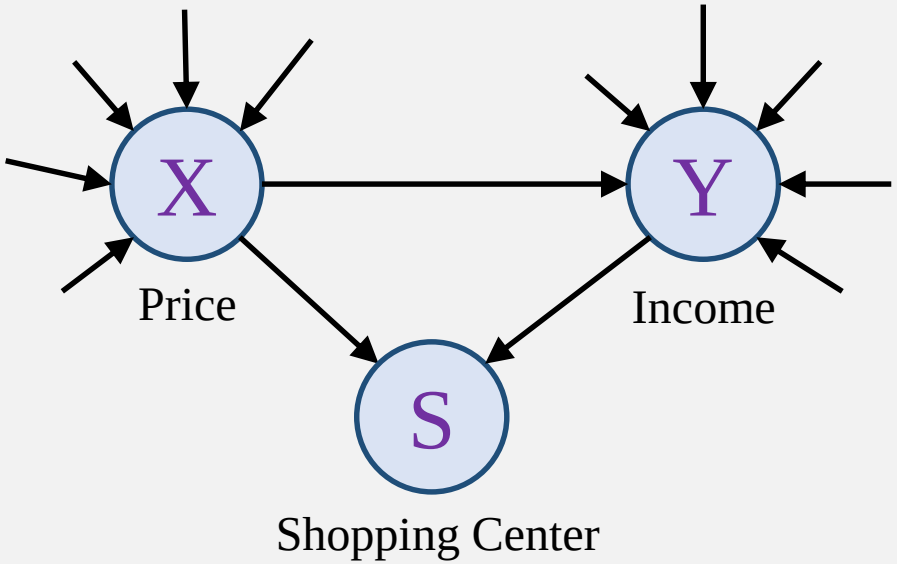
Secondly, Selection Bias



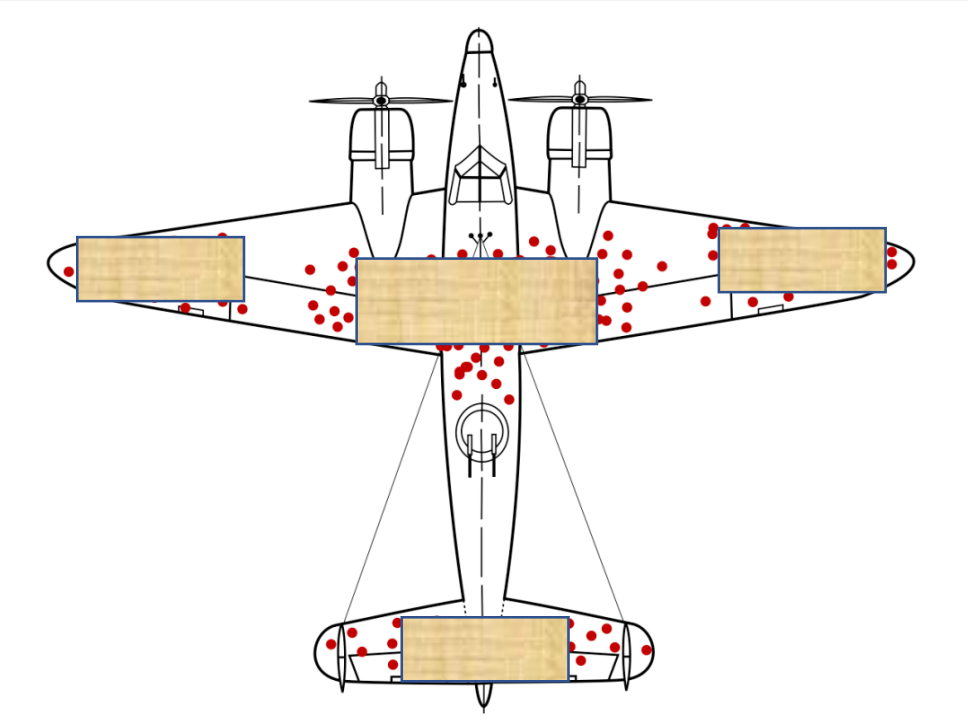
Due to data missing, privacy policy, only partial data are recorded.



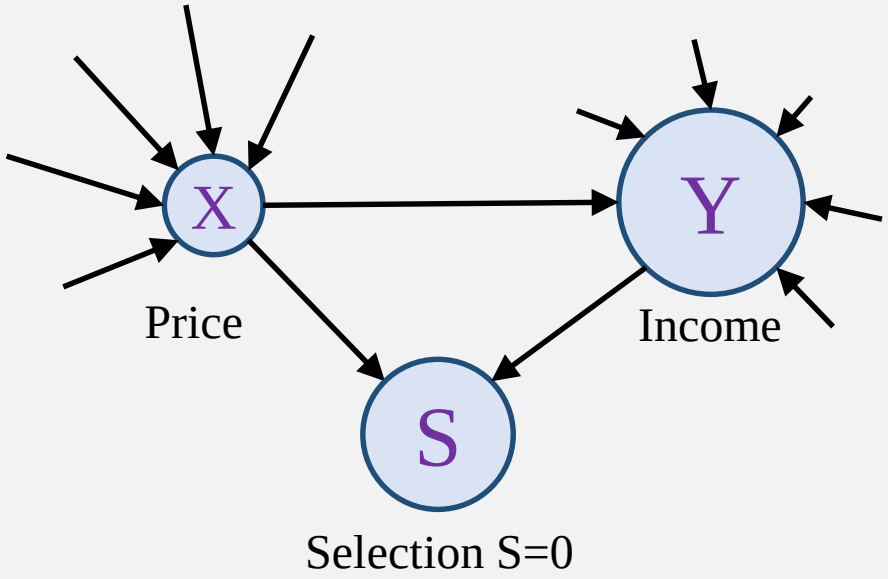
Secondly, Selection Bias

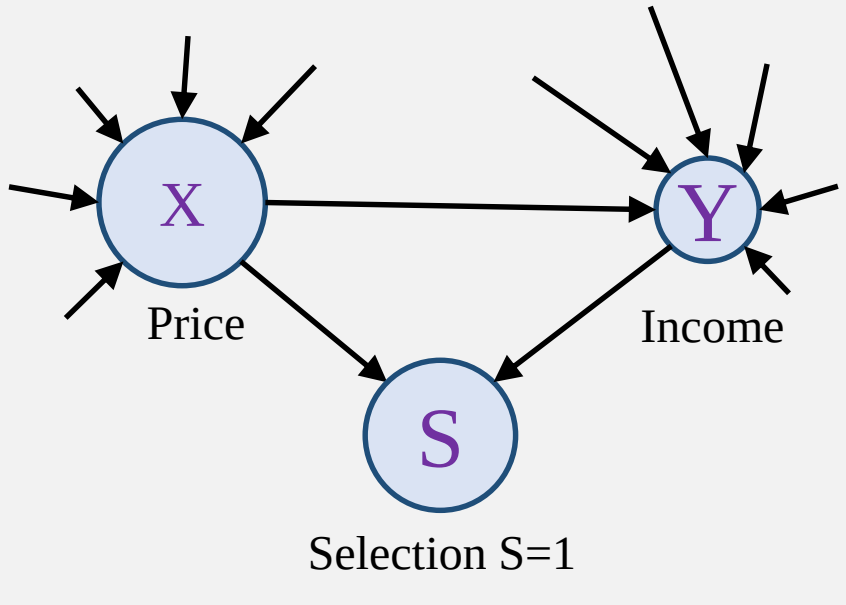


Due to data missing, privacy policy, only partial data are recorded.

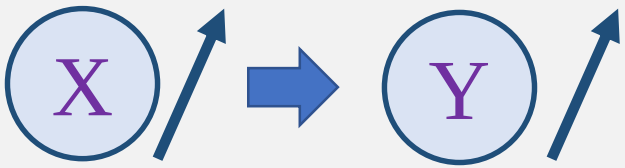


Secondly, Selection Bias

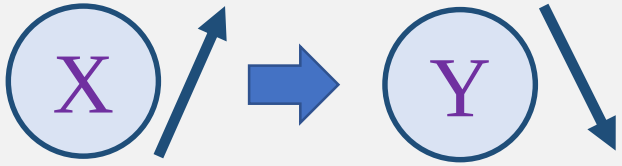




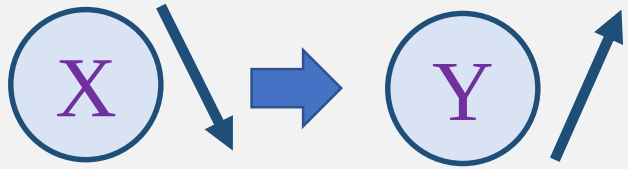
Increase in price, increase in income.



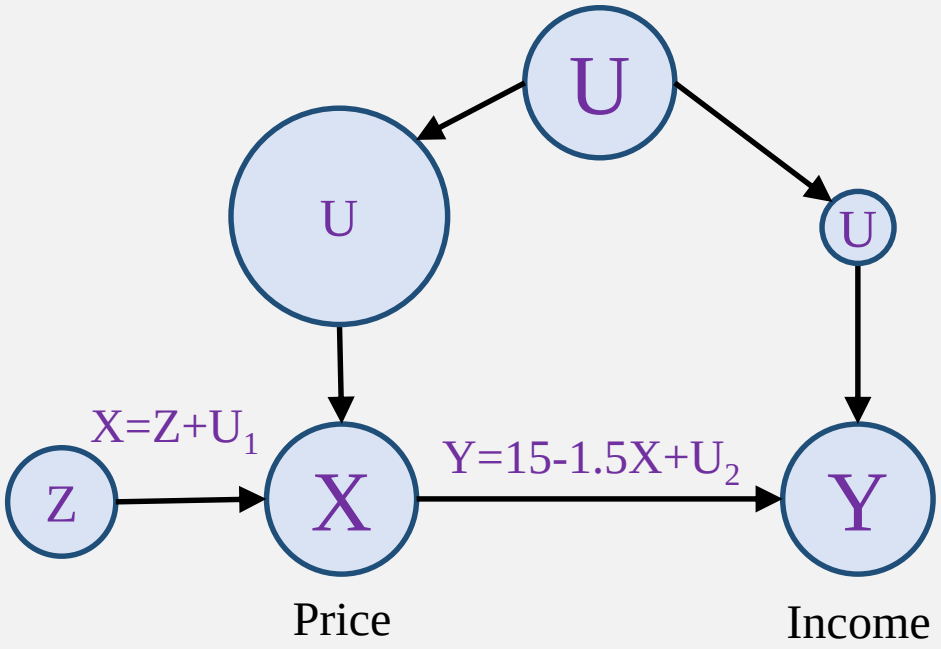
Increase in price, decrease in income.

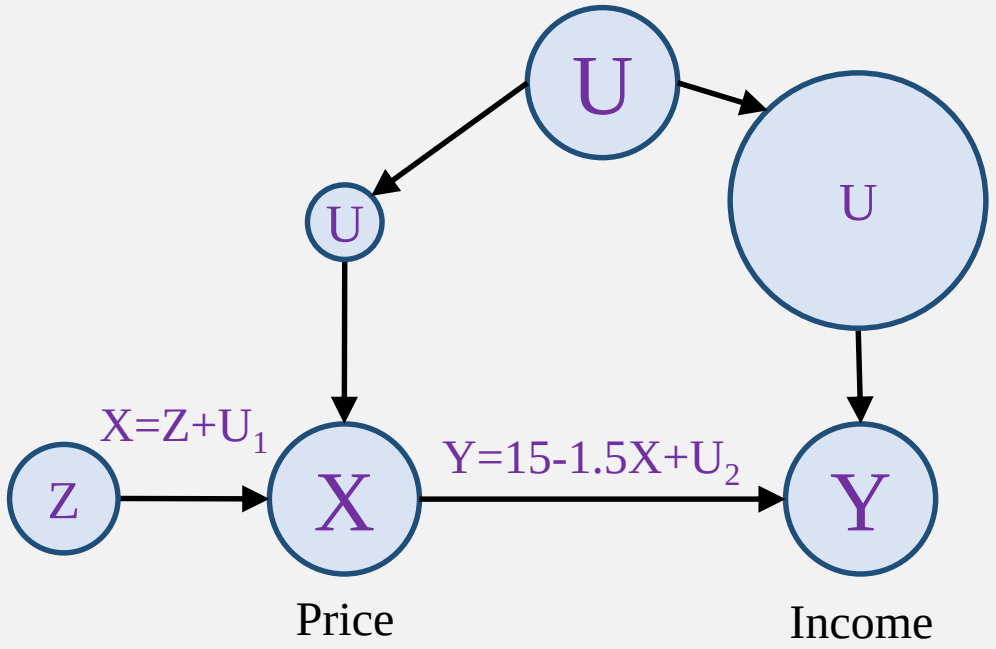


Decrease in price, increase in income.

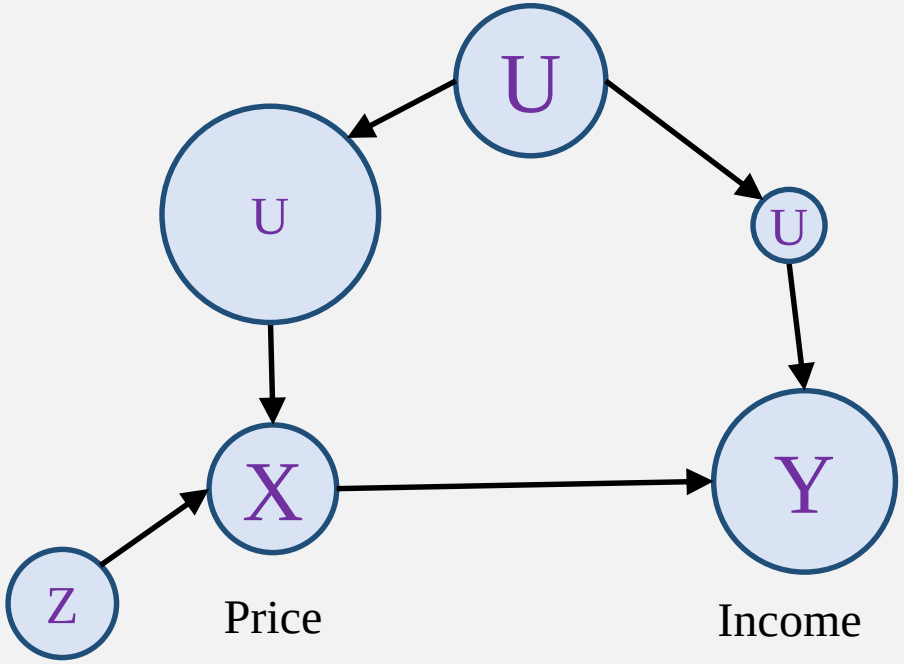


Due to time limited, we directly introduce Instrumental Variables.

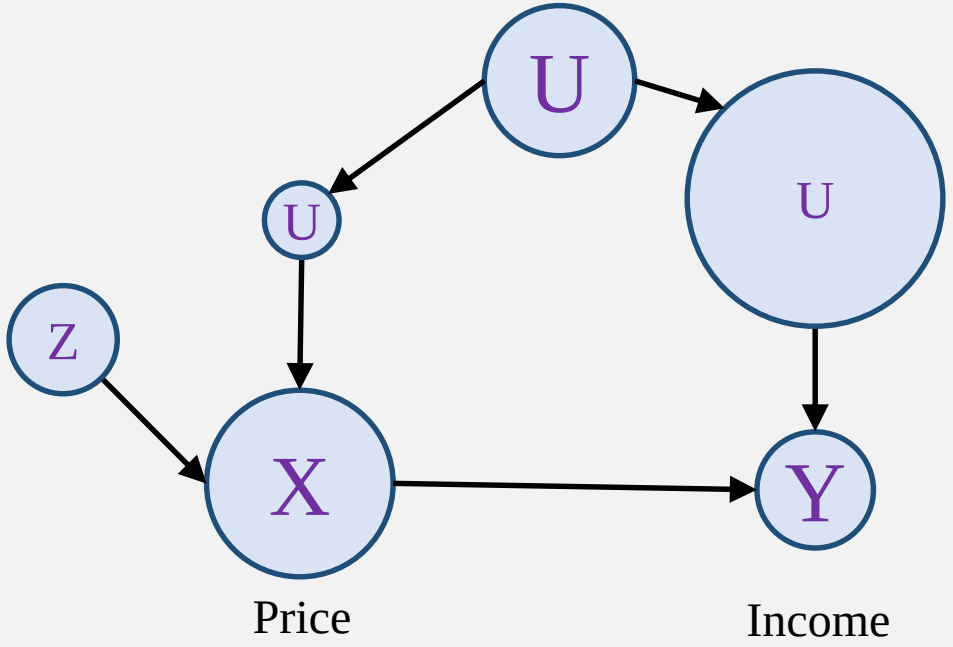




Due to time limited, we directly introduce Instrumental Variables.



Z=1, X: -1*Z, Y: +1.5*Z

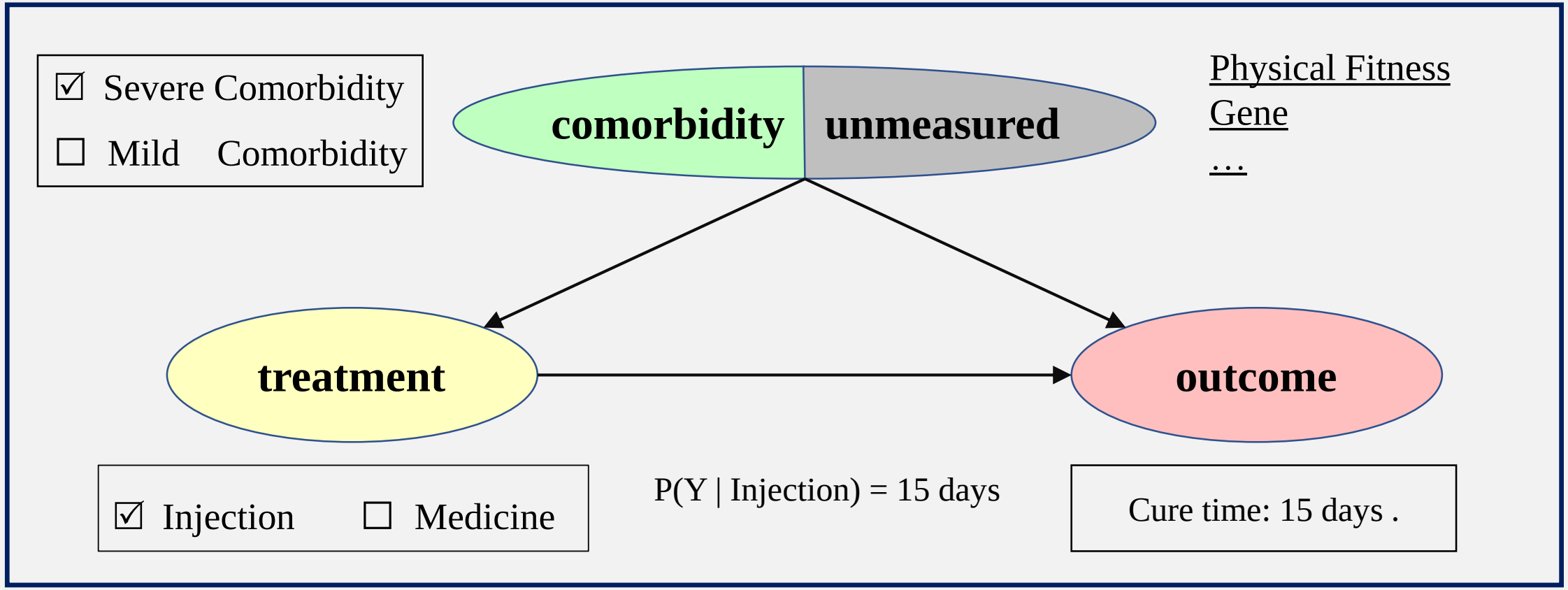


Z=-1, X: 1*Z, Y: -1.5*Z

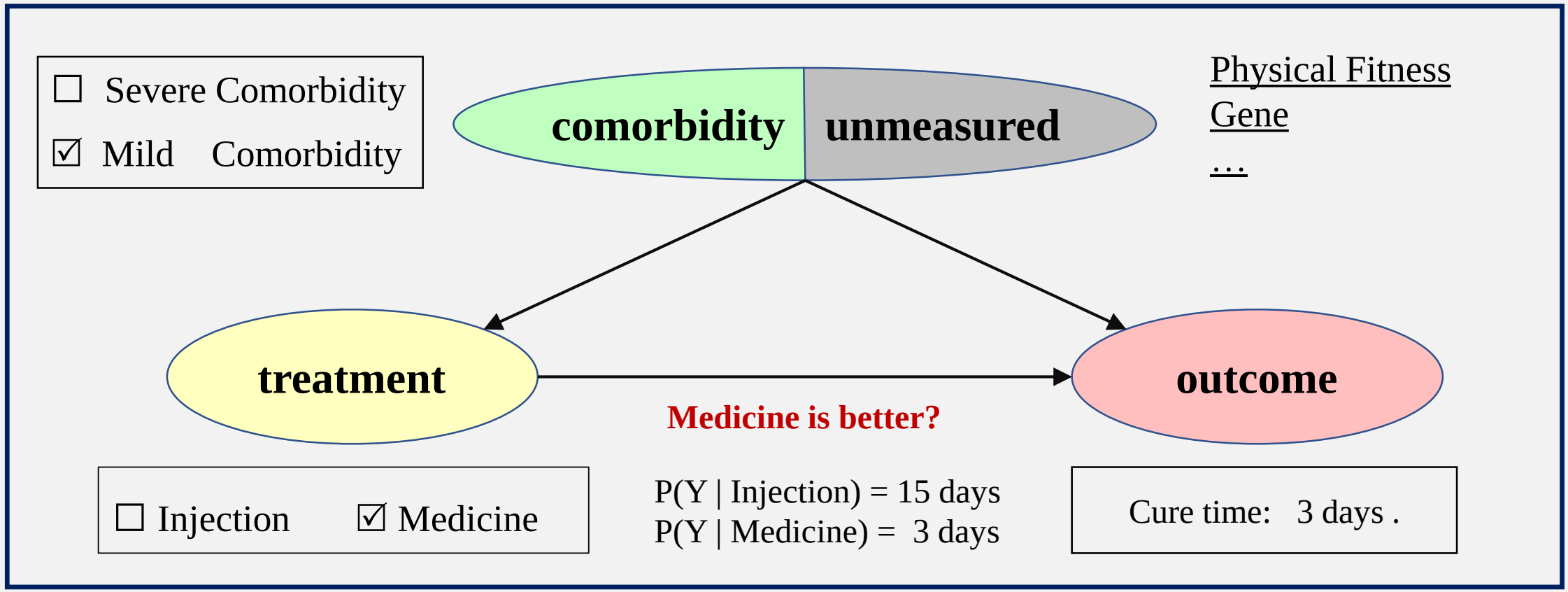
Instrumental variable, identifying the invariant instrumental effect from the variant total effect.

Toward ATE with Unmeasured Confounders
Need & Challenge

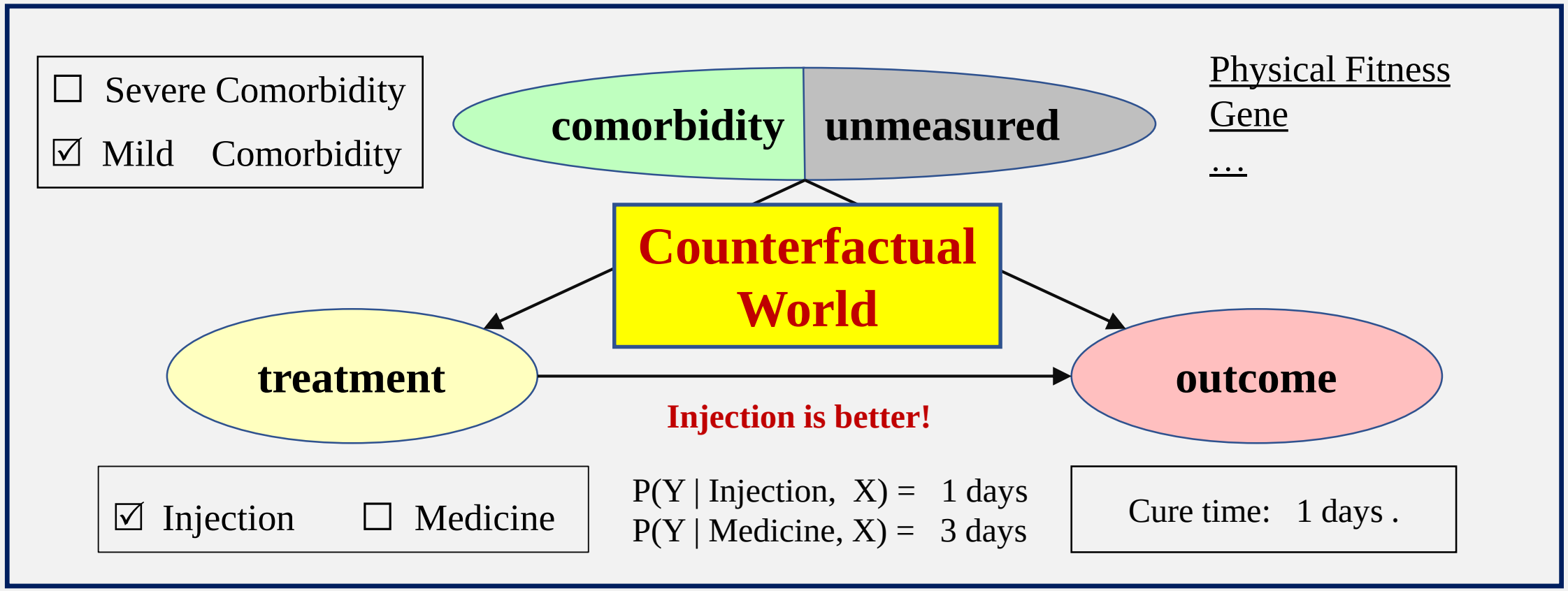
Real-World Examples



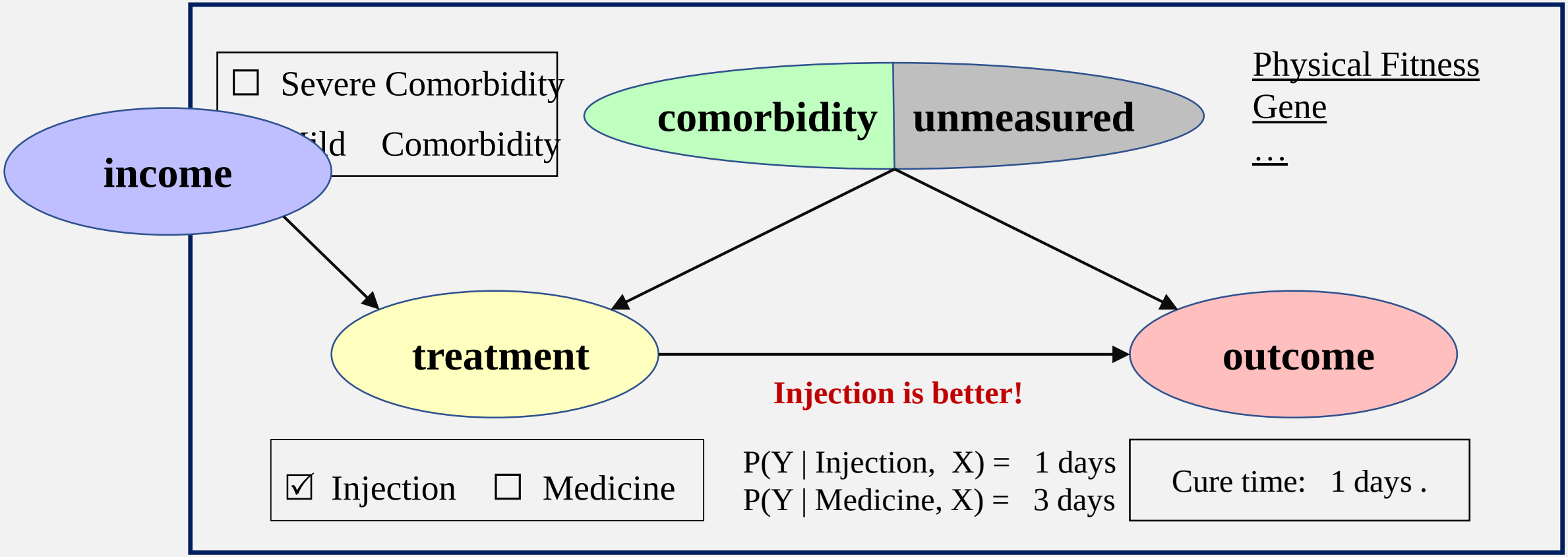
Toward ATE with Unmeasured Confounders
Need & Challenge



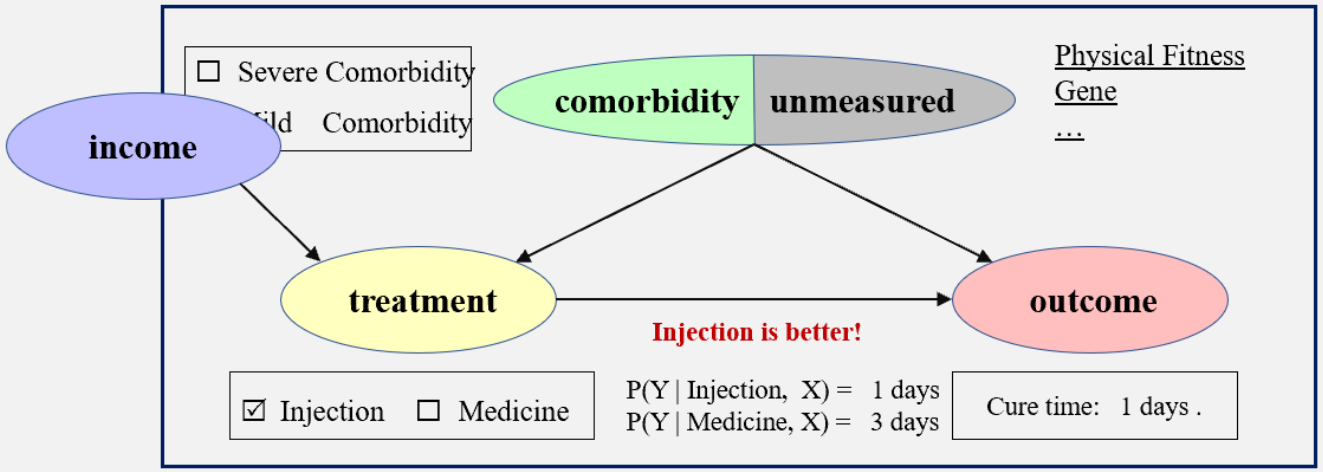
Toward ATE with Unmeasured Confounders
Need & Challenge

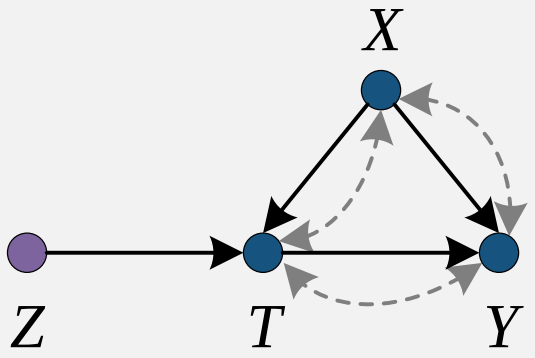


Instrumental Variable & Assumption



Instrumental Variable & Assumption

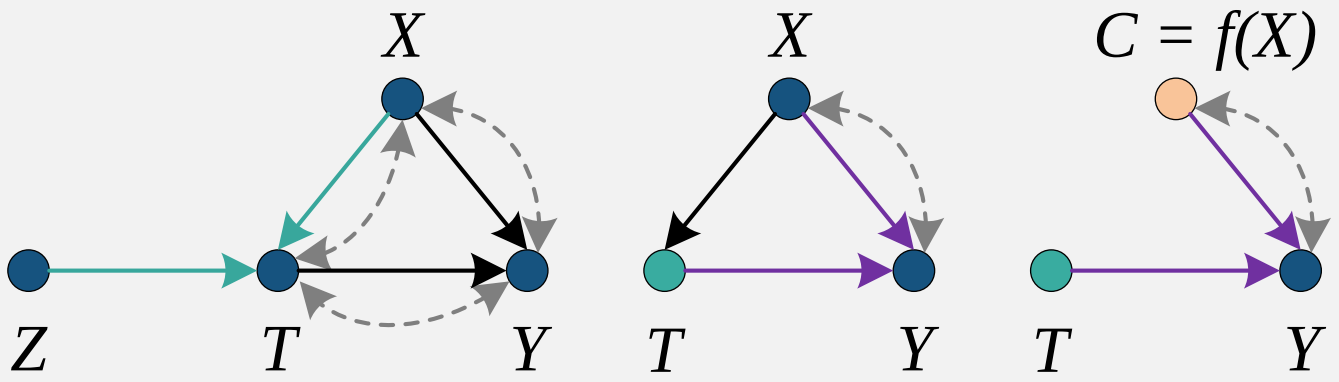




Definition 3.2. An Instrument Variable Z is an exogenous variable that affects the treatment T, but does not directly affect the outcome Y. Besides, an valid instrument variable satisfies the following three assumptions:

Relevance: Z is a cause of T, i.e., $\mathbb{P}(T \mid Z) \neq \mathbb{P}(T)$.
Exclusion: Z does not directly affect the outcome Y, i.e., $Z \perp Y \mid T, X, U$.
Unconfounded: Z is independent of all confounders, including X and U, i.e., $Z \perp X, U$

Method



X: Covariate;
Z: Instrument (Seasons);
T: Treatment (Price);
Y: Outcome (Sales);
→: Stage 1; →: Stage 2.

Instrumental Variable Regression with Confounder Balancing (CB-IV)

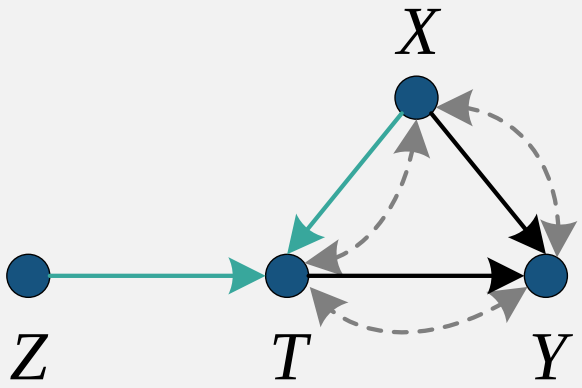
- Treatment Regression (Stage 1)
- Confounder Balancing (Stage 2)
- Outcome Regression (Stage 2)

Method - Stage 1

Binary Cases

Treatment Regression: In this part, we propose to regress treatment T with IVs Z and observed confounders X directly, as the treatment regression stage did in the previous nonlinear IV-based method. Specifically, we estimate the conditional probability distribution of the treatments $\hat{P}(T|Z,X)$ with a logistic regression network $\pi_\mu(z_i, x_i)$ with learnable parameter μ for each unit i, and optimize the following loss function for treatment regression:

$$\begin{aligned} \mathcal{L}_T &= -\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(t_i \log\left(\pi_\mu(z_i, x_i)\right) \\ &+ \left(1-t_i\right)\left(1-\log\left(\pi_\mu(z_i, x_i)\right)\right)) \end{aligned} \tag{4}$$



Continuous Cases

Besides, to reduce computational complexity, we can set $\sigma_\psi = c$ as constant for low uncertainty models, and simplify the distribution estimation as a regression problem:

$$\mathcal{L}_T = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(t_i - \hat{t}_i^j\right)^2, \hat{t}_i^j \sim \hat{P}(t_i|z_i, x_i), \tag{27}$$

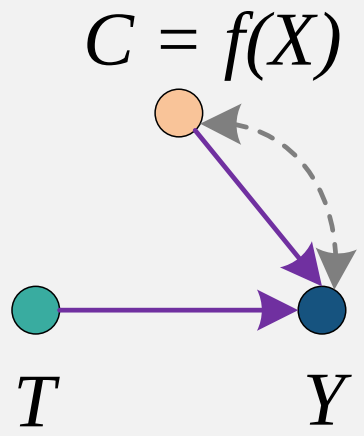
we sample m (the larger the better) treatment $\{\hat{t}_i^j\}_{j=1,\ldots,m}$ for each unit $\{z_i, x_i\}$ to approximate the true treatment t_i. Empirically, the above objective (Eq. (27)) is sufficient to accurately estimate causal effects in continuous CB-IV framework.

Method - Stage 2

Binary Cases

Confounder Balancing: After treatment regression, we can obtain the causal graph as shown in the figure 1(b), where the observed variables X would become the confounders for outcome regression. To address this problem, we propose to learn a representation of X (i.e., $C = f_\theta(X)$) with a representation network $f_\theta(\cdot)$ with learnable parameter θ, and minimize the discrepancy of distributions for different treatment arms to achieve $C \perp \hat{T}$ for confounder balancing:

$$\begin{aligned}\text{disc}(\hat{T}, f_\theta(X)) = \text{IPM}(&\{f_\theta(x_i)\hat{P}(t_i = 0 \mid z_i, x_i)\}_{i=1}^n, \\ &\{f_\theta(x_i)\hat{P}(t_i = 1 \mid z_i, x_i)\}_{i=1}^n) \quad (5)\end{aligned}$$



Continuous Cases

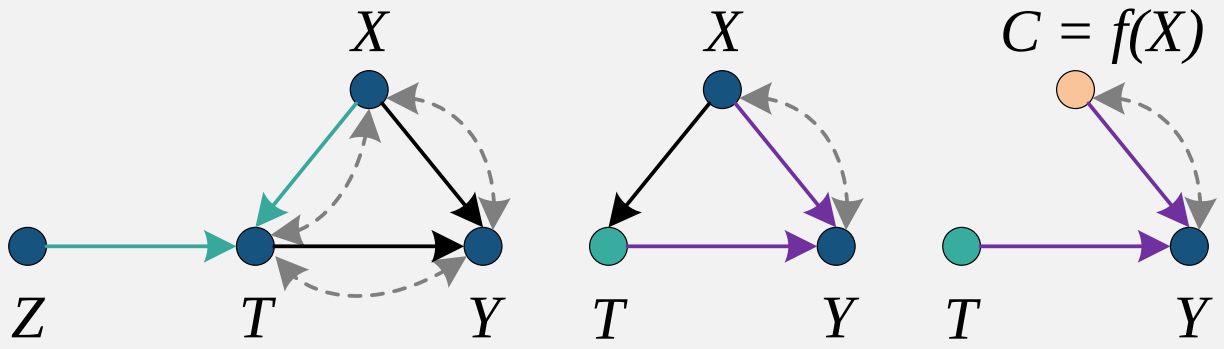
Confounder Balancing: For continuous treatment T, we learn a "balanced" representation (i.e., C) of the observed confounders X as $C = f_\theta(X)$ via mutual information (MI) minimization constraints (Cheng et al., 2020): firstly, we use variational distribution $Q_\psi(\hat{T} \mid C) = \mathcal{N}(\mu_\psi(C), \sigma_\psi(C))$ parameterized by neural networks $\{\mu_\psi, \sigma_\psi\}$ to approximate the true conditional distribution $P(\hat{T} \mid C)$; then, we minimize the log-likelihood loss function of variational approximation $Q_\psi(\hat{T} \mid C)$ with n samples to estimate MI:

$$\text{disc}(\hat{T}, C) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\log Q_\psi\left(\hat{t}_i \mid c_i\right) - \log Q_\psi\left(\hat{t}_j \mid c_i\right)\right]. \quad (28)$$

where, $C = f_\theta(X)$. We adopt an alternating training strategy to iteratively optimize $Q_\psi(\hat{T} \mid C)$ and the network $C = f_{\theta(X)}$ to implement balanced representation in the Confounder Balancing.

Method - Stage 2

Final,



X: Covariate;
Z: Instrument (Seasons);
T: Treatment (Price);
Y: Outcome (Sales);
→: Stage 1; →: Stage 2.

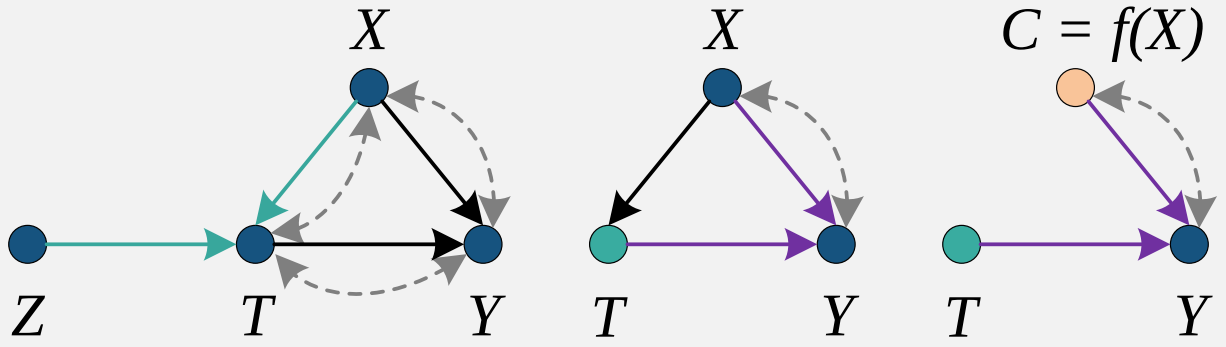
Outcome Regression: Finally, we propose to regress the outcome with the estimated treatment $\hat{T} \sim P(T|Z, X)$ obtained in treatment regression module and the representation of confounders $C = f_\theta(X)$ obtained in confounder balancing module:

$$\mathcal{L}_Y = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - h_\xi(\hat{t}_i, f_\theta(x_i))\right)^2 + \alpha \, \text{disc}(\hat{T}, f_\theta(X)) \tag{29}$$

where α is a trade-off hyper-parameter, and $\hat{t}_i \sim \hat{P}(T|Z, X)$ and $f_\theta(x_i)$ are derived from treatment regression module and confounder balancing module, respectively.

Method - Stage 2

Final,



X: Covariate;
Z: Instrument (Seasons);
T: Treatment (Price);
Y: Outcome (Sales);
→: Stage 1; →: Stage 2.

Outcome Regression: Finally, we propose to regress the outcome with the estimated treatment $\hat{T} \sim P(T|Z,X)$ obtained in treatment regression module and the representation of confounders $C = f_\theta(X)$ obtained in confounder balancing module:

$$\mathcal{L}_Y = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - h_\xi(\hat{t}_i, f_\theta(x_i))\right)^2 + \alpha\,\mathrm{disc}(\hat{T}, f_\theta(X)) \tag{29}$$

where α is a trade-off hyper-parameter, and $\hat{t}_i \sim \hat{P}(T|Z,X)$ and $f_\theta(x_i)$ are derived from treatment regression module and confounder balancing module, respectively.

Conditional average treatment effect: $ATE_i = h(T = t_i, f(x_i)) - h(T = 0, f(x_i))$

Experimental Results

Evaluation Measure

The conditional average treatment effect (CATE):

$$ATE(t) = \mathbb{E}[Y \mid do(T = t), X] - \mathbb{E}[Y \mid do(T = 0), X]$$

Bias of the conditional average treatment effect:

$$ATE_i = h(T = t_i, f(x_i)) - h(T = 0, f(x_i))$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(ATE_i - ATE_i)^2$$

The lower is the better.

Benchmarks

Binary Cases

- Systematically varied the dimensions of Z, X and U: m_Z, m_X, m_U.
- Naming convention: Syn – m_Z – m_X – m_U
- 10 runs for each trial with 10000 samples

Continuous Cases

- Demand is a common benchmark used in IV Regressions (Hartford et al., 2017, Singh et al., 2019, Muandet et al., 2020, Xu et al., 2021).
- Systematically varied the importance of instrumental variables and confounders: **γ, λ**
- Naming convention: Demand – **γ** – **λ**

Real-World Datasets

- IHDP/Twins – m_Z – m_X – m_U

Experimental Results

The results of ATE estimation, including bias (mean(std)), in

Out-of-Sample				
Method	**Syn-1-4-4**	**Syn-2-4-4**	**Syn-2-10-4**	**Syn-2-4-10**
DeepIV-LOG	1.055(0.010)	1.057(0.008)	1.093(0.009)	1.020(0.008)
DeepIV-GMM	0.933(0.011)	0.874(0.019)	0.768(0.023)	0.925(0.017)
KernelIV	0.495(0.055)	0.458(0.052)	0.765(0.028)	0.625(0.063)
DualIV	1.472(0.079)	1.467(0.076)	1.732(0.072)	1.513(0.066)
OneSIV	0.822(0.076)	0.661(0.095)	0.690(0.053)	0.851(0.073)
DFIV	0.851(0.009)	0.860(0.007)	0.851(0.007)	0.886(0.009)
DFL	0.840(0.002)	0.851(0.002)	0.838(0.002)	0.831(0.004)
DirectRep	0.172(0.016)	0.164(0.009)	0.116(0.015)	0.199(0.014)
CFR	0.172(0.015)	0.159(0.018)	0.103(0.019)	0.198(0.016)
DRCFR	0.151(0.055)	0.137(0.035)	**0.062(0.045)**	0.154(0.032)
CB-IV	**0.037(0.075)**	**0.017(0.046)**	0.075(0.040)	**0.010(0.064)**

(a) Binary Cases

Out-of-Sample			
Method	**Demand-0-1**	**Demand-0-5**	**Demand-5-1**
DeepIV-LOG	-	-	-
DeepIV-GMM	1006(313.7)	2829(724.6)	1151(284.1)
KernelIV	994.9(146.2)	5435(435.2)	1004(216.7)
DualIV	>5000	>5000	>5000
OneSIV	>5000	>5000	>5000
DFIV	190.5(8.977)	668.3(566.7)	196.2(16.66)
DFL	182.9(11.52)	597.6(622.1)	189.7(7.422)
DirectRep	193.9(7.380)	689.6(692.1)	489.9(121.1)
CFR	192.0(8.932)	417.3(123.5)	469.7(140.7)
DRCFR	532.4(199.5)	497.3(26.37)	470.5(143.4)
CB-IV	**172.9(5.340)**	**224.3(18.06)**	**165.8(7.142)**

(b) Continuous Cases

Out-of-Sample				
Method	**IHDP-2-6-0**	**IHDP-2-4-2**	**Twins-5-8-0**	**Twins-5-5-3**
DeepIV-LOG	2.876(0.055)	2.623(0.069)	0.014(0.021)	0.024(0.011)
DeepIV-GMM	3.777(0.035)	3.739(0.042)	0.019(0.005)	0.022(0.004)
KernelIV	3.070(0.306)	3.023(0.440)	-	-
DualIV	0.564(0.266)	0.715(0.355)	-	-
OneSIV	1.729(0.372)	1.735(0.343)	0.008(0.019)	0.008(0.017)
DFIV	3.554(0.090)	3.623(0.106)	0.027(0.001)	0.026(0.000)
DFL	3.204(0.050)	3.199(0.038)	0.062(0.058)	0.085(0.005)
DirectRep	0.061(0.082)	0.457(0.076)	0.016(0.018)	0.019(0.025)
CFR	0.079(0.081)	0.480(0.069)	0.011(0.016)	0.022(0.018)
DRCFR	0.045(0.095)	0.432(0.067)	0.011(0.022)	0.012(0.017)
CB-IV	**0.015(0.393)**	**0.158(0.254)**	**0.006(0.027)**	**0.002(0.025)**

* Most confounders are discrete variables and the outcome is binary variable in Twins data. The results of kernel-based IV methods in Twins are NaN. We use '-' to denote it.

(c) Real-World Datasets

Conclusion

Experimental Results

The results of ATE estimation, including bias (mean(std)), in

Out-of-Sample				
Method	Syn-1-4-4	Syn-2-4-4	Syn-2-10-4	Syn-2-4-10
DeepIV-LOG	1.055(0.010)	1.057(0.008)	1.093(0.009)	1.020(0.008)
DeepIV-GMM	0.933(0.011)	0.874(0.019)	0.768(0.023)	0.925(0.017)
KernelIV	0.495(0.055)	0.458(0.052)	0.765(0.028)	0.625(0.063)
DualIV	1.472(0.079)	1.467(0.076)	1.732(0.072)	1.513(0.066)
OneSIV	0.822(0.076)	0.661(0.095)	0.690(0.053)	0.851(0.073)
DFIV	0.851(0.009)	0.860(0.007)	0.851(0.007)	0.886(0.009)
DFL	0.840(0.002)	0.851(0.002)	0.838(0.002)	0.831(0.004)
DirectRep	0.172(0.016)	0.164(0.009)	0.116(0.015)	0.199(0.014)
CFR	0.172(0.015)	0.159(0.018)	0.103(0.019)	0.198(0.016)
DRCFR	0.151(0.055)	0.137(0.035)	**0.062(0.045)**	0.154(0.032)
CB-IV	**0.037(0.075)**	**0.017(0.046)**	0.075(0.040)	**0.010(0.064)**

(a) Binary Cases

Out-of-Sample			
Method	Demand-0-1	Demand-0-5	Demand-5-1
DeepIV-LOG	-	-	-
DeepIV-GMM	1006(313.7)	2829(724.6)	1151(284.1)
KernelIV	994.9(146.2)	5435(435.2)	1004(216.7)
DualIV	>5000	>5000	>5000
OneSIV	>5000	>5000	>5000
DFIV	190.5(8.977)	668.3(566.7)	196.2(16.66)
DFL	182.9(11.52)	597.6(622.1)	189.7(7.422)
DirectRep	193.9(7.380)	689.6(692.1)	489.9(121.1)
CFR	192.0(8.932)	417.3(123.5)	469.7(140.7)
DRCFR	532.4(199.5)	497.3(26.37)	470.5(143.4)
CB-IV	**172.9(5.340)**	**224.3(18.06)**	**165.8(7.142)**

(b) Continuous Cases

Out-of-Sample				
Method	IHDP-2-6-0	IHDP-2-4-2	Twins-5-8-0	Twins-5-5-3
DeepIV-LOG	2.876(0.055)	2.623(0.069)	0.014(0.021)	0.024(0.011)
DeepIV-GMM	3.777(0.035)	3.739(0.042)	0.019(0.005)	0.022(0.004)
KernelIV	3.070(0.306)	3.023(0.440)	-	-
DualIV	0.564(0.266)	0.715(0.355)	-	-
OneSIV	1.729(0.372)	1.735(0.343)	0.008(0.019)	0.008(0.017)
DFIV	3.554(0.090)	3.623(0.106)	0.027(0.001)	0.026(0.000)
DFL	3.204(0.050)	3.199(0.038)	0.062(0.058)	0.085(0.005)
DirectRep	0.061(0.082)	0.457(0.076)	0.016(0.018)	0.019(0.025)
CFR	0.079(0.081)	0.480(0.069)	0.011(0.016)	0.022(0.018)
DRCFR	0.045(0.095)	0.432(0.067)	0.011(0.022)	0.012(0.017)
CB-IV	**0.015(0.393)**	**0.158(0.254)**	**0.006(0.027)**	**0.002(0.025)**

* Most confounders are discrete variables and the outcome is binary variable in Twins data. The results of kernel-based IV methods in Twins are NaN. We use '-' to denote it.

(c) Real-World Datasets

Conclusion

- The traditional IV-based methods would suffer from the confounding bias from the observed confounders, if the outcome model is mis-specified and covariates are imbalanced;
- Considering confounder balancing in IV regression, our CB-IV improves considerably over the traditional IV-based methods and achieves better performance than confounder balancing methods in most settings.
- Extensive experimental results supports the promise of the proposed method and perspective.

The implementation of CB-IV is publicly available at:
https://github.com/anpwu/CB-IV

Thanks

Acknowledgement

National Natural Science Foundation of China (No. 62006207, No. 62037001, No.72171131), Key R \& D Projects of the Ministry of Science and Technology (2020YFC0832500), Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), Project by Shanghai AI Laboratory (P22KS00111), the Fundamental Research Funds for the Central Universities (226-2022-00142). Tsinghua University Initiative Scientific Research Grant (No. 2019THZWJC11); Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grants 2020AAA0108400 and 2020AAA0108403.