ZHEJIANG UNIVERSITY

TSINGHUA UNIVERSITY

ICML | 2022

Thirty-ninth International Conference on Machine Learning

CB-IV: Instrumental Variable Regression with Confounder Balancing

Anpeng Wu[1]✉, Kun Kuang[1*], Bo Li[2], Fei Wu[134]

1 Department of Computer Science and Technology, Zhejiang University, Hangzhou, China

2 School of Economics and Management, Tsinghua University, Beijing, China

3 Shanghai Institute for Advanced Study, Zhejiang University, Shanghai, China

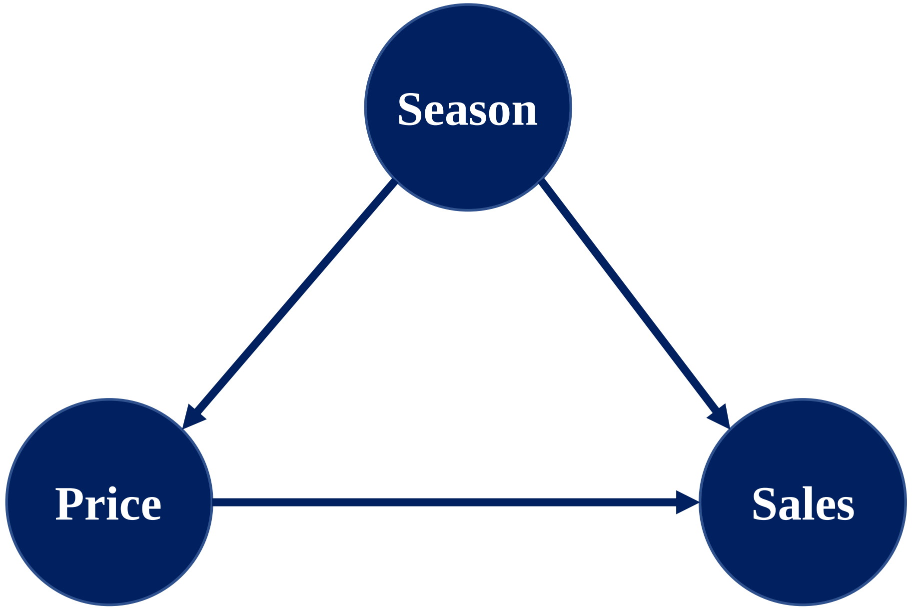
4 Shanghai AI Laboratory, Shanghai, China

✉ anpwu@zju.edu.cn

*Corresponding author

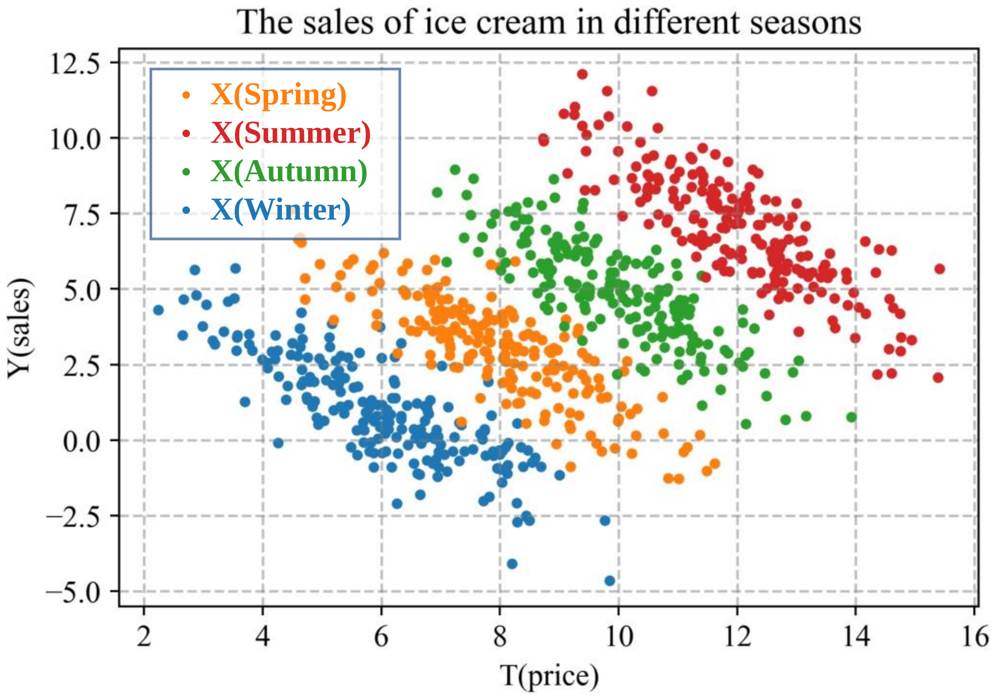
Introduction

Confounder / Confounding Variable

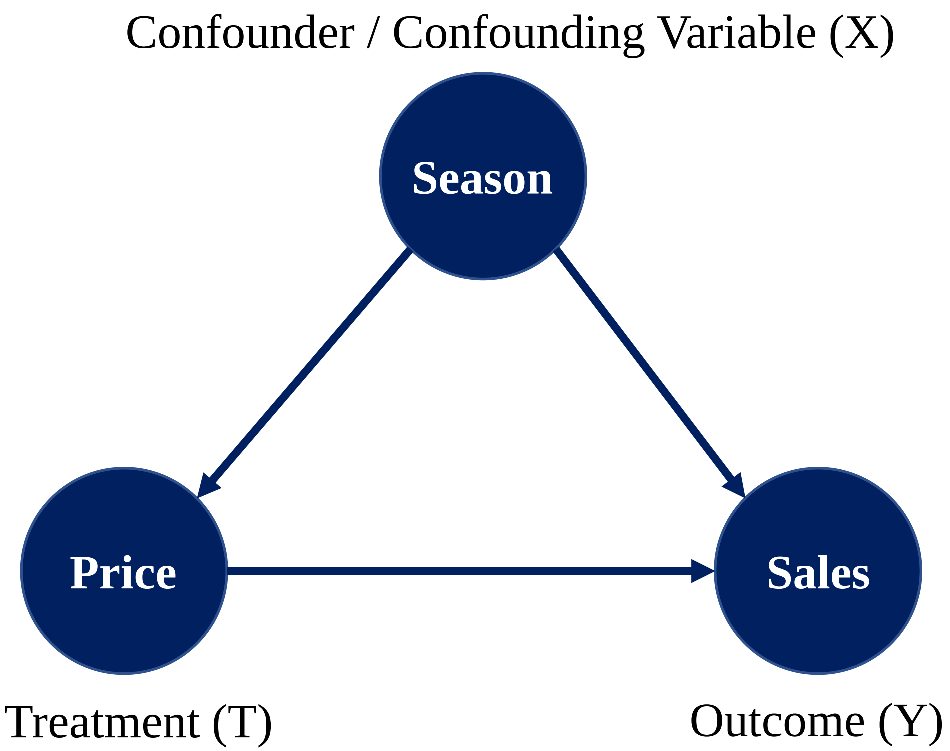


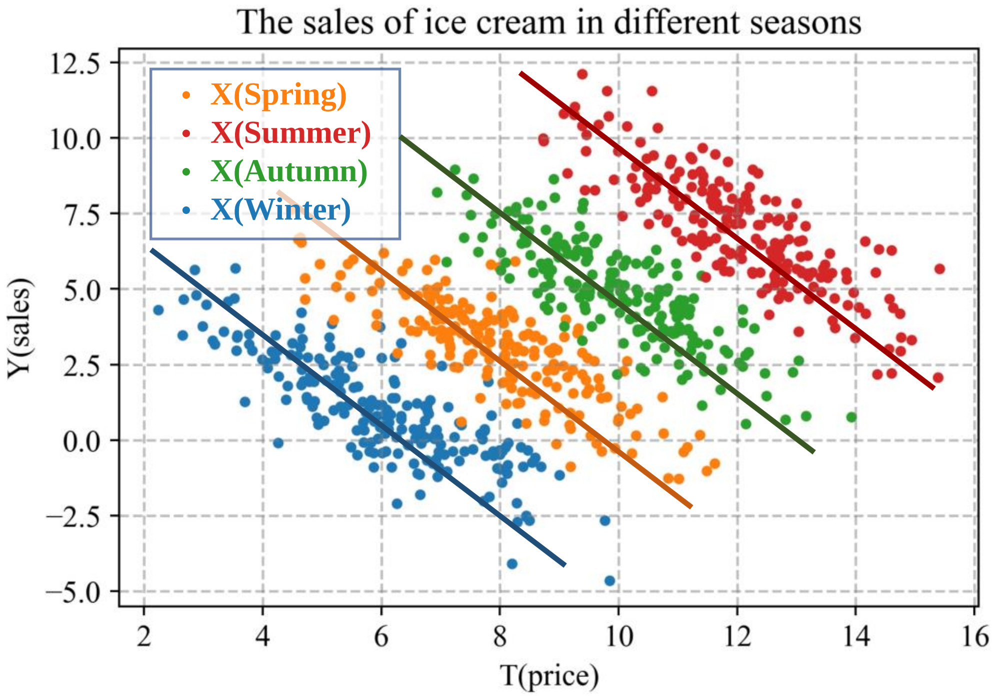
Treatment

Outcome



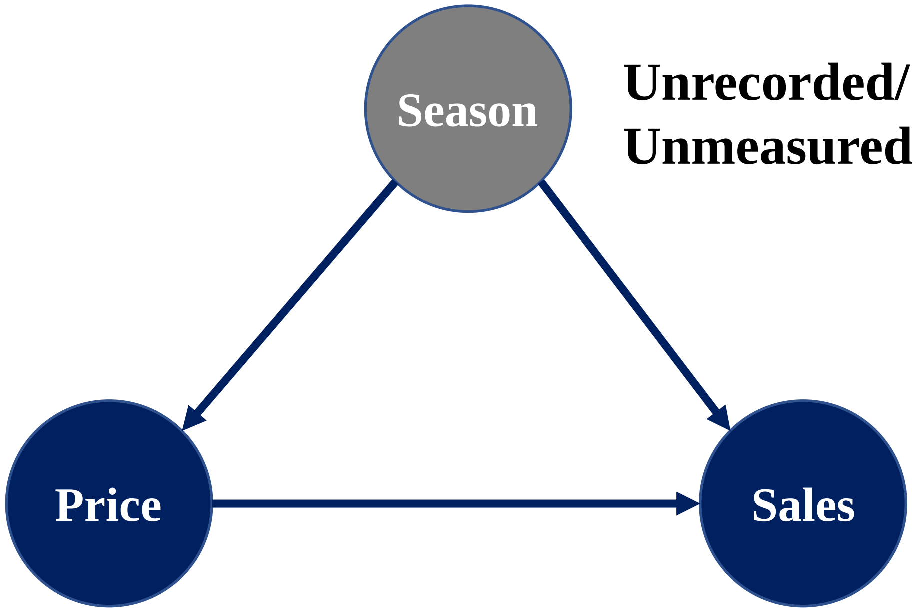
Introduction





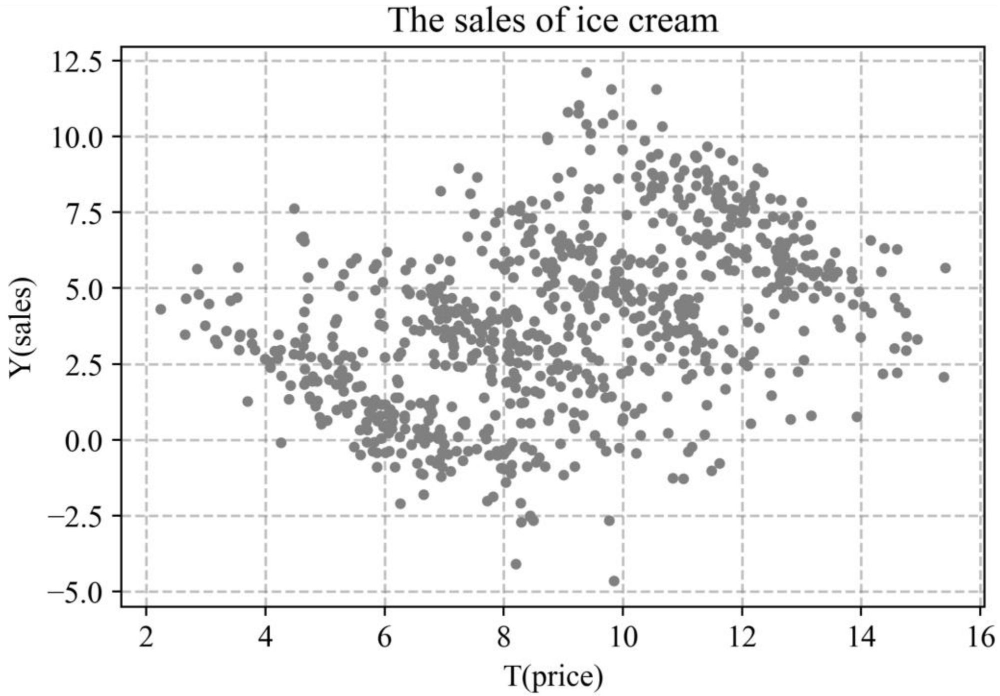
Introduction

Confounder / Confounding Variable (X)

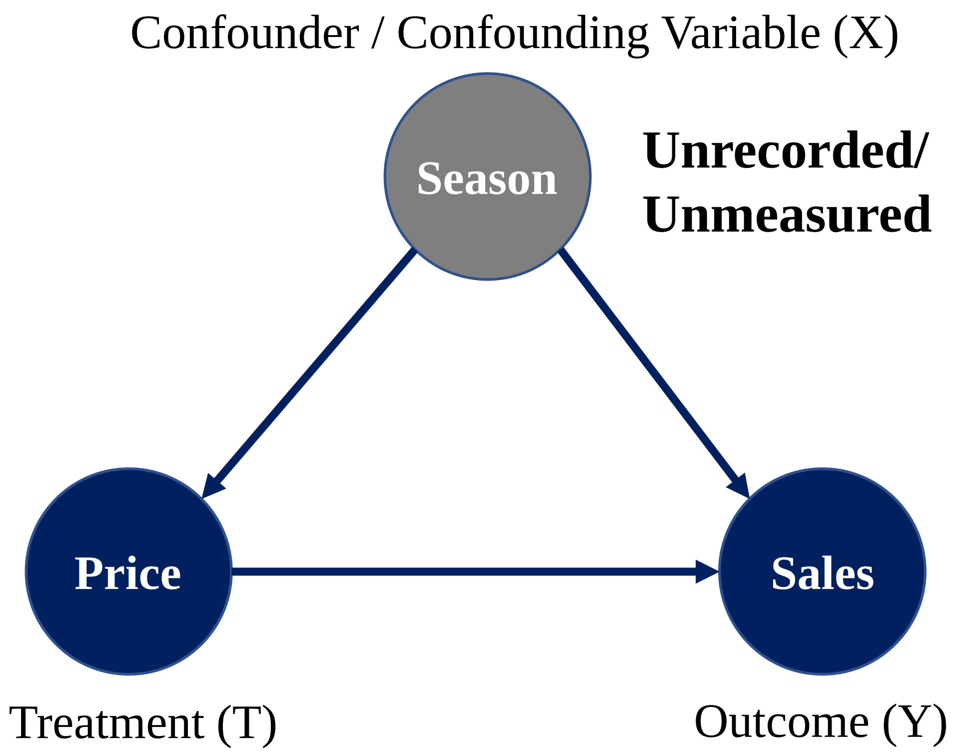


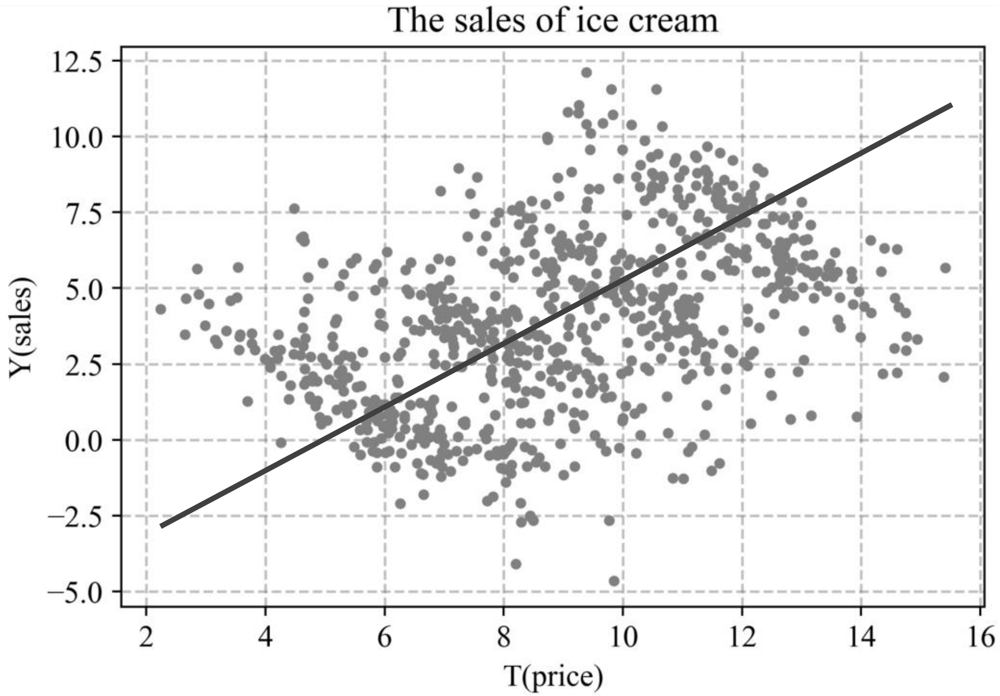
Treatment (T)

Outcome (Y)

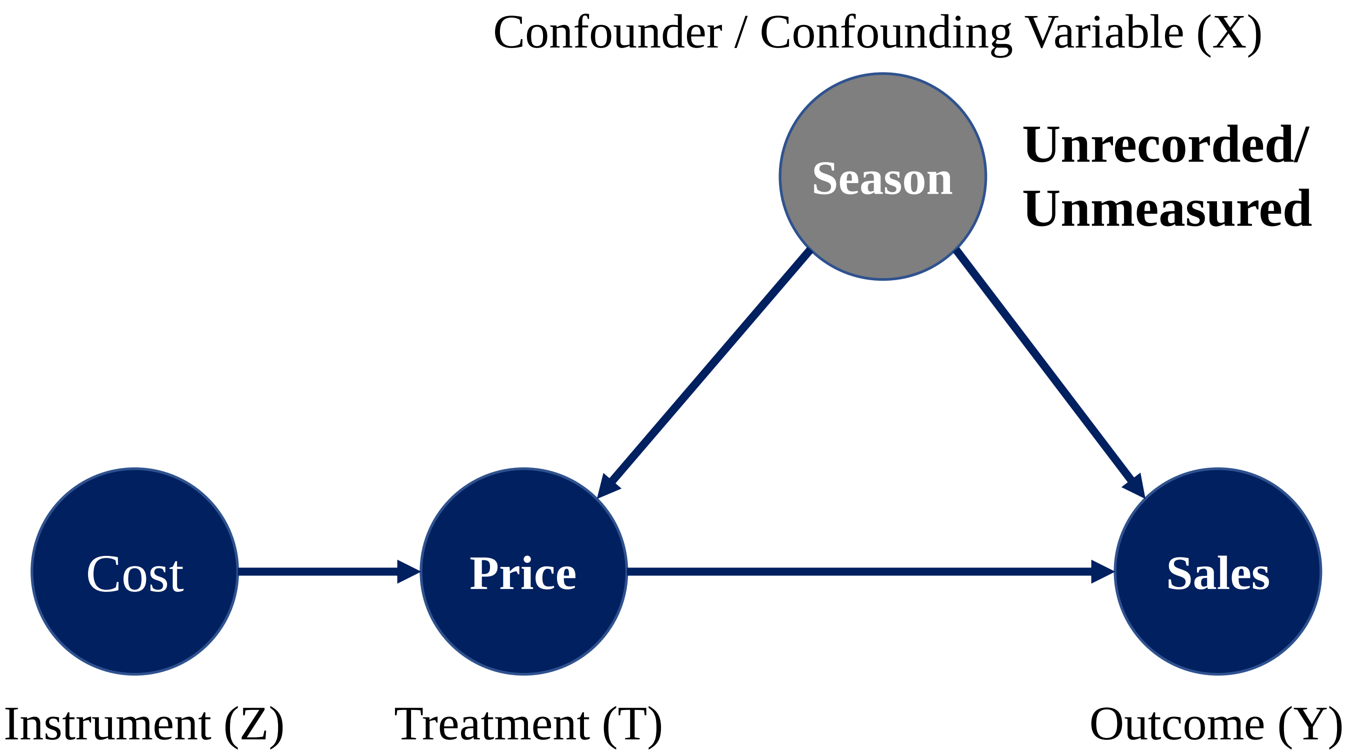


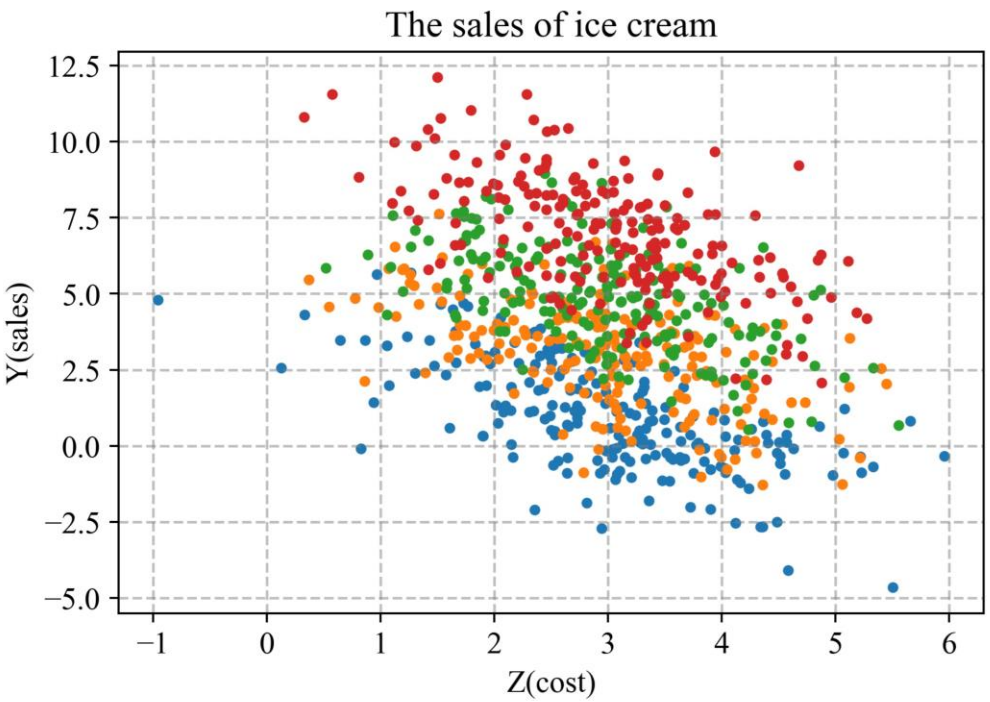
Introduction



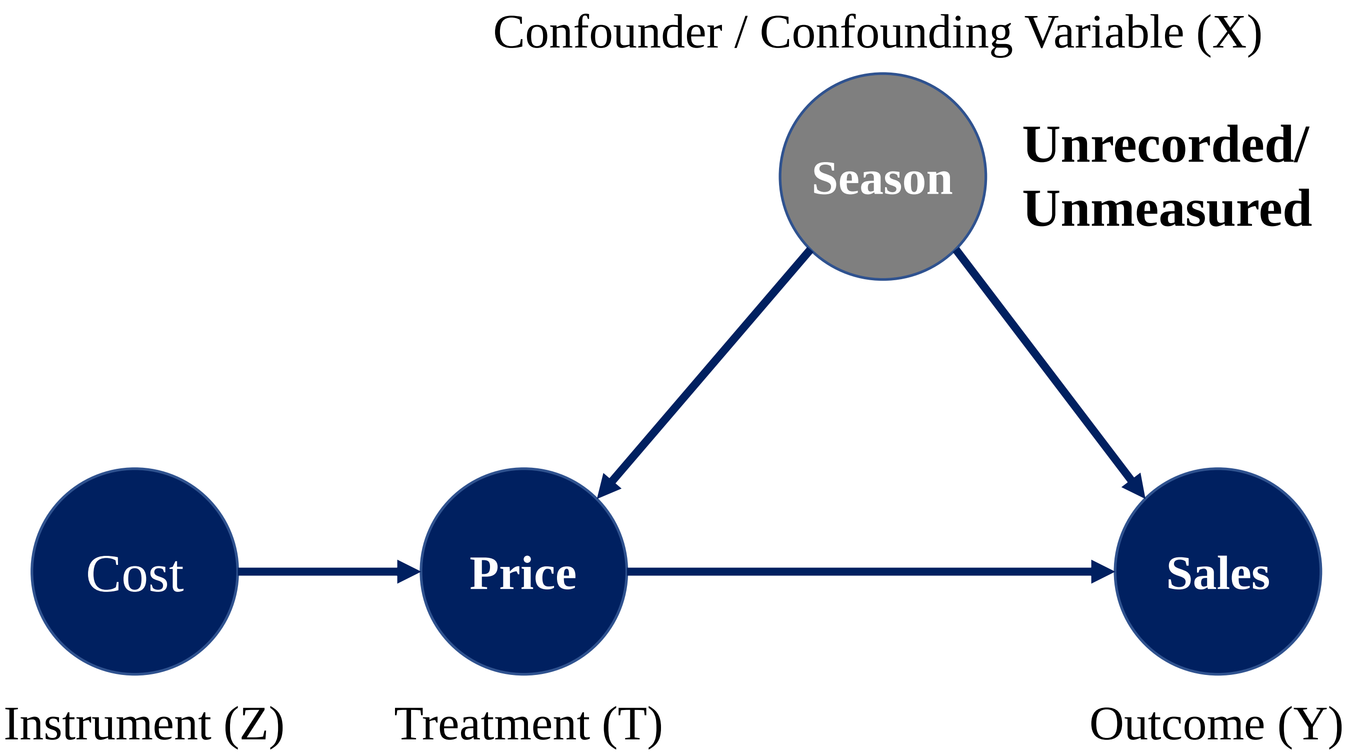


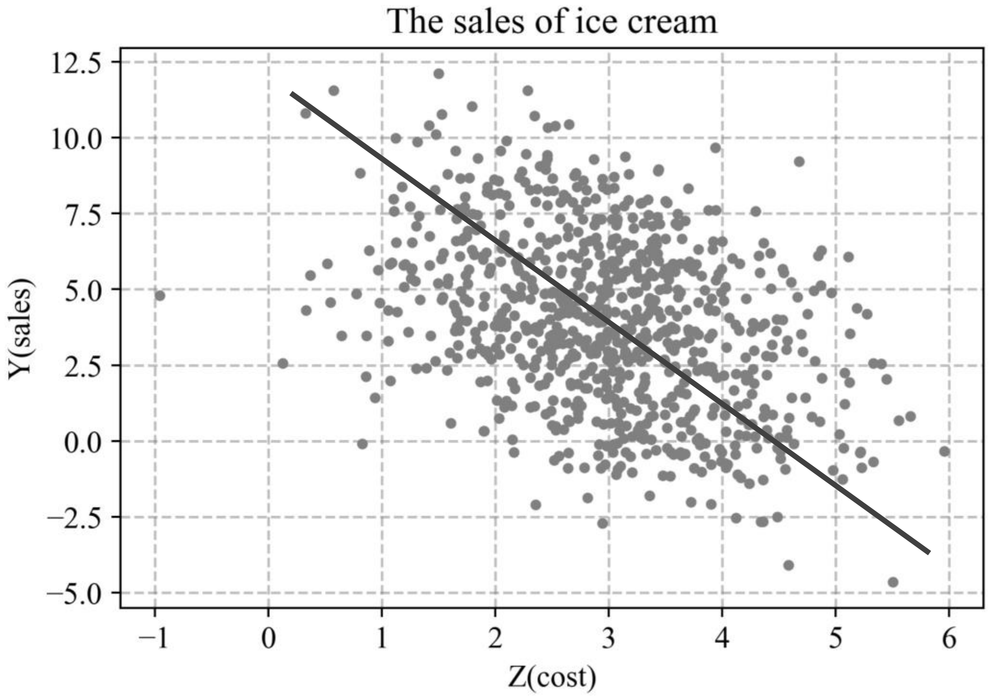
Introduction





Introduction

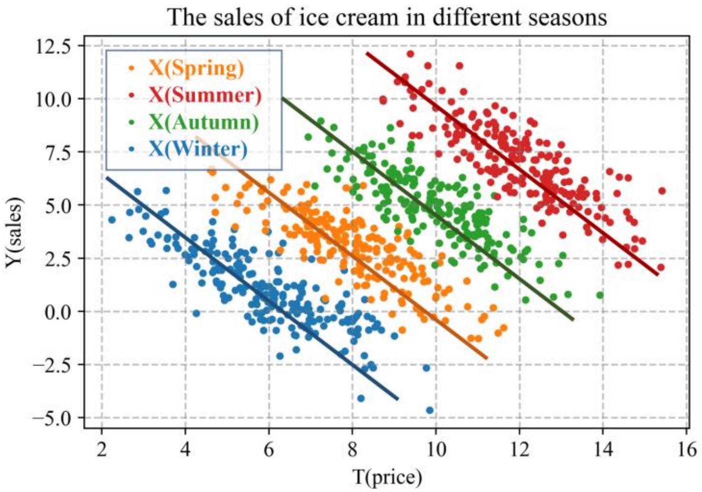


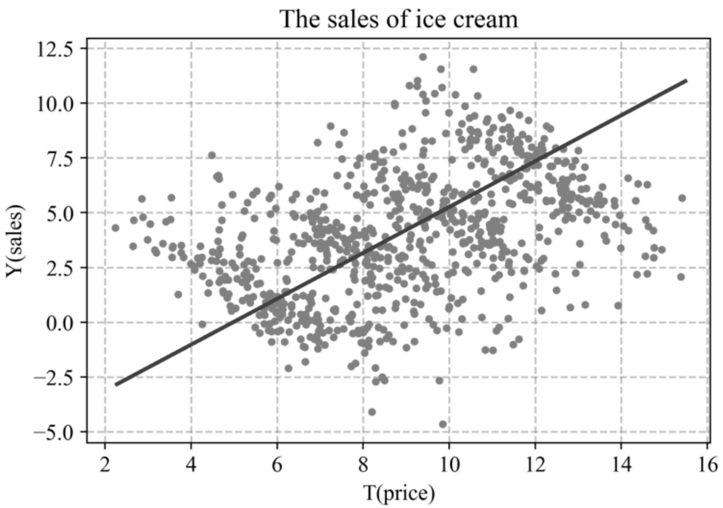


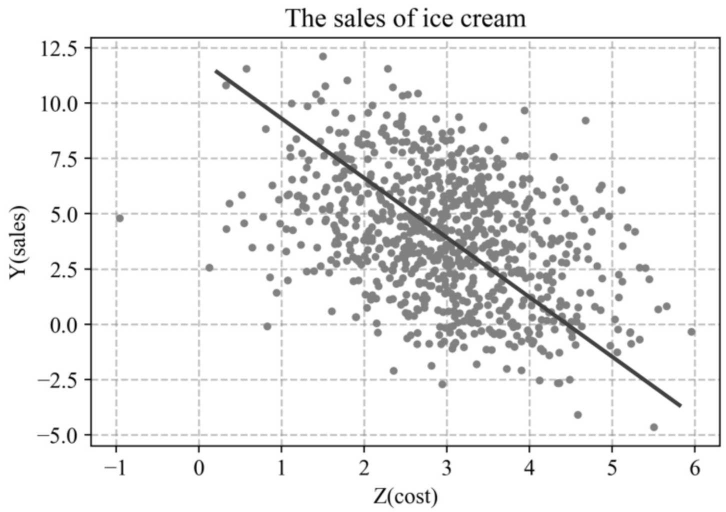
An Instrument Variable *Z* is an exogenous variable that affects the treatment *T*, but does not directly affect the outcome *Y*.

- **Relevance: Z is a cause of T, i.e., $P(T \mid Z) \neq P(T)$;**
- **Exclusion: Z does not directly affect the outcome Y, i.e., $Z \perp Y \mid T, X$;**
- **Unconfounded: Z is independent of all confounders X, i.e., $Z \perp X$.**

Identifying Assumption



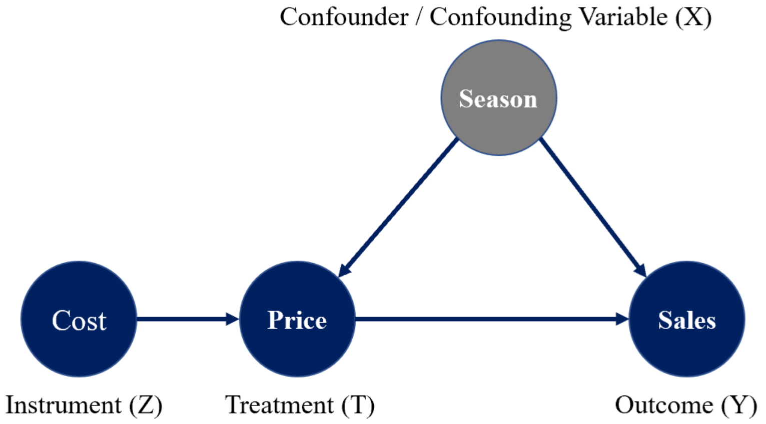




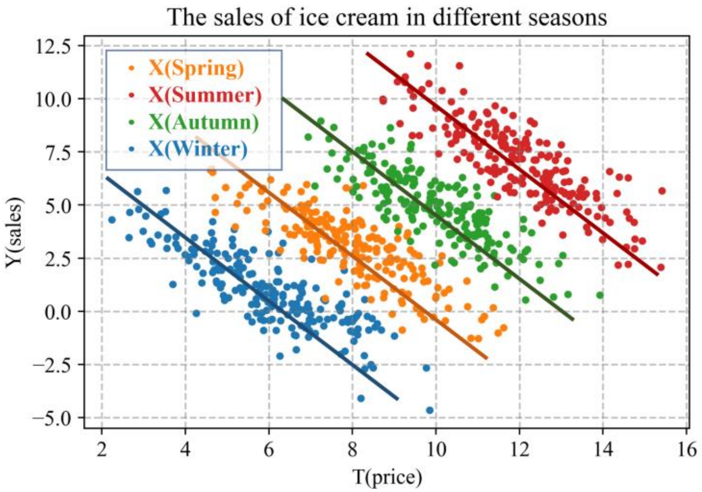
(1) Linear Assumption: $Y = \beta T + \theta_1 X + \theta_2,\ T = \alpha Z + \theta_3 X + \theta_4$

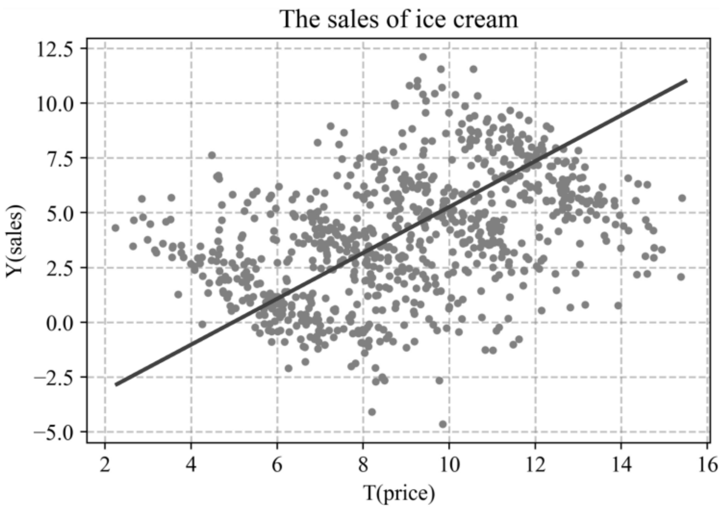
Instrumental Variable Regression:

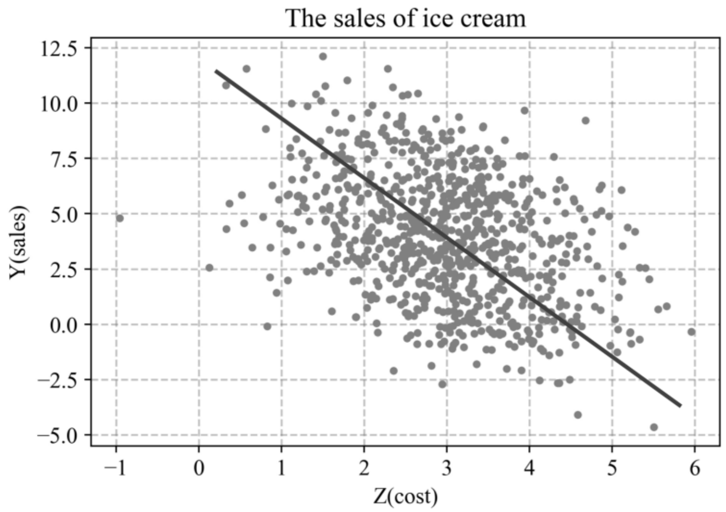
- **Stage 1:** $\hat{\alpha} = \arg\min_{\alpha} \sum_{i=1}^{n} (t_i - \alpha z_i)^2$
- **Stage 2:** $\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta \hat{t}_i)^2 = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta \hat{\alpha} z_i)^2$



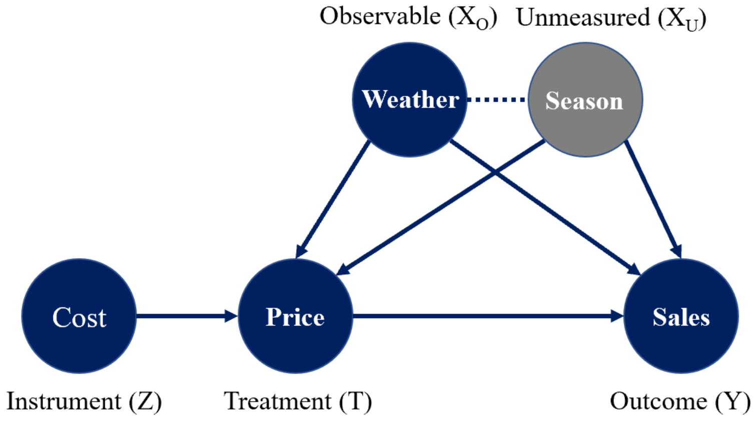
Identifying Assumption



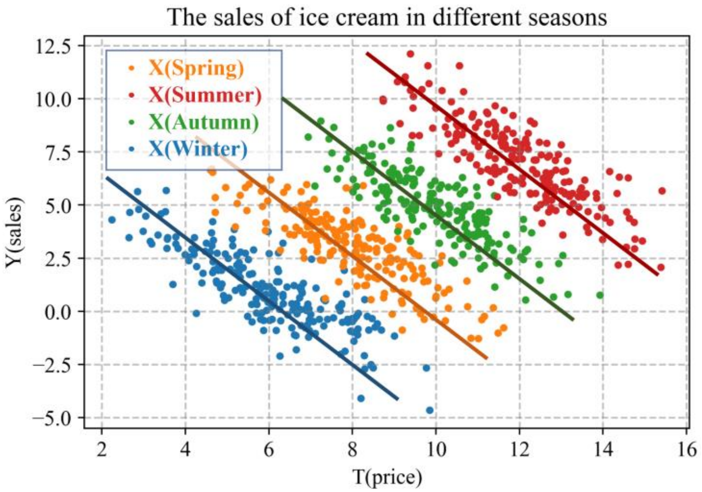


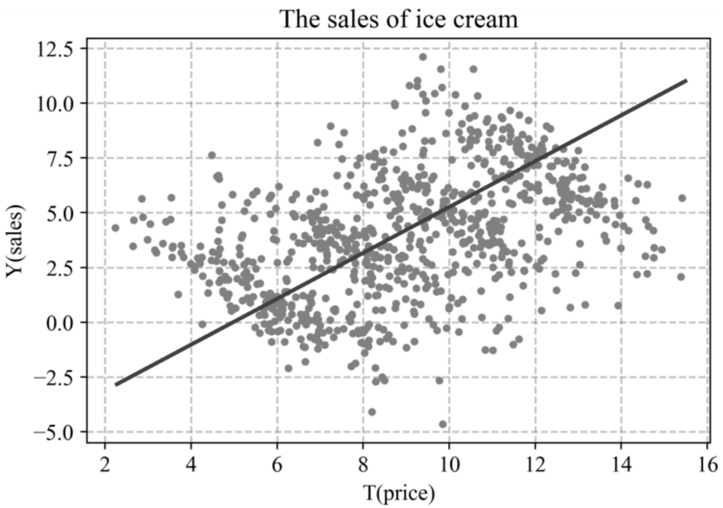


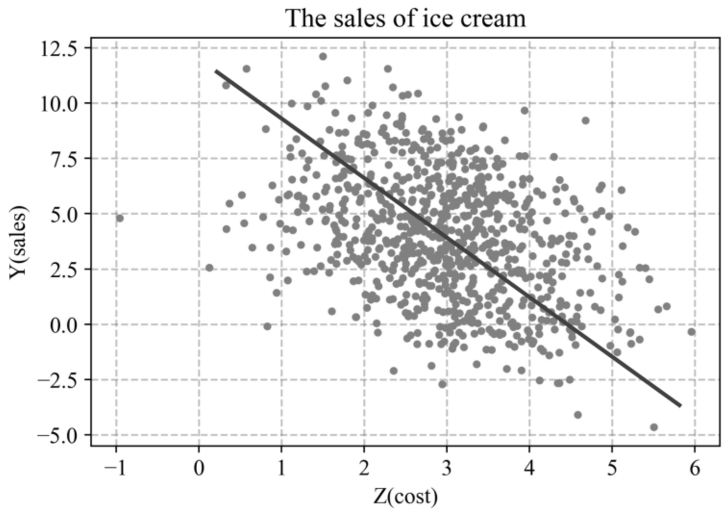
(2) Additive Noise Assumption: $Y = g(T, X_O) + X_U$



Identifying Assumption



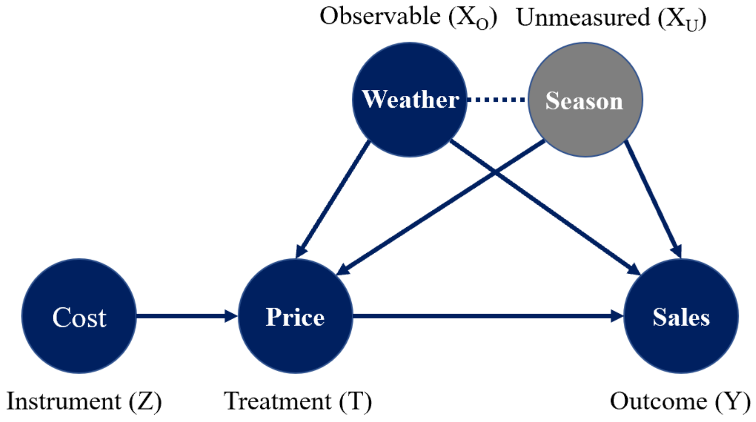




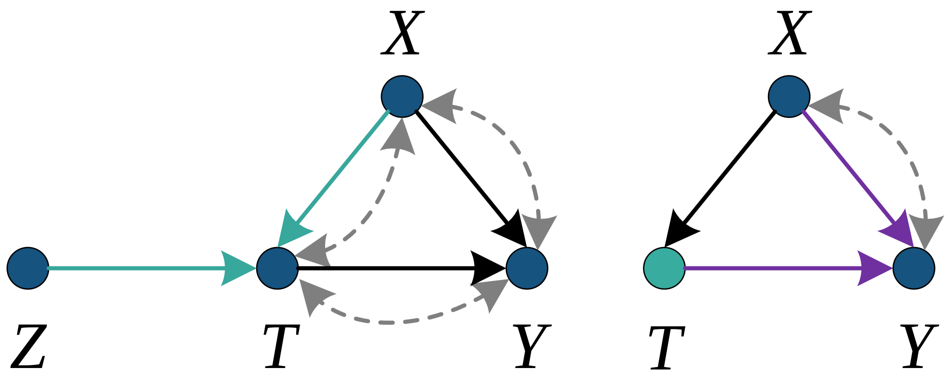
(2) Additive Noise Assumption: $Y = g(T, X) + U$

Instrumental Variable Regression:

- **Stage 1:** $\hat{f} = \arg\min_f \sum_{i=1}^{n} (t_i - P_{f(z_i, x_i)}(t_i \mid z_i, x_i))^2, \hat{t}_i \sim P_{\hat{f}(z_i, x_i)}(t \mid z_i, x_i)$
- **Stage 2:** $\hat{g} = \arg\min_g \sum_{i=1}^{n} (y_i - g(\hat{t}_i, x_i))^2, \hat{t}_i$ is obtained from stage 1.

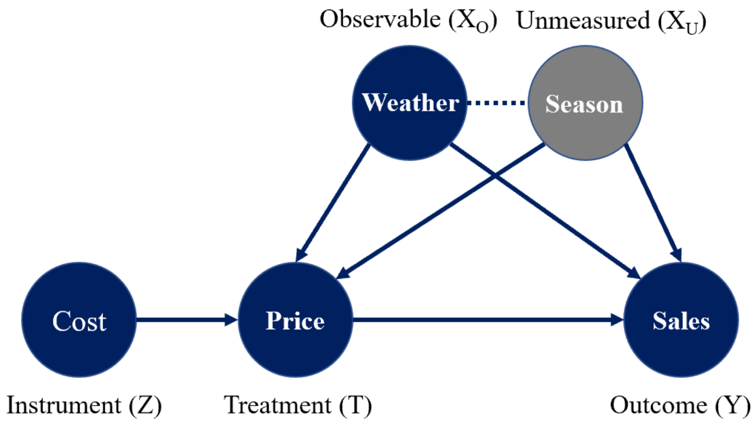


Motivation



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

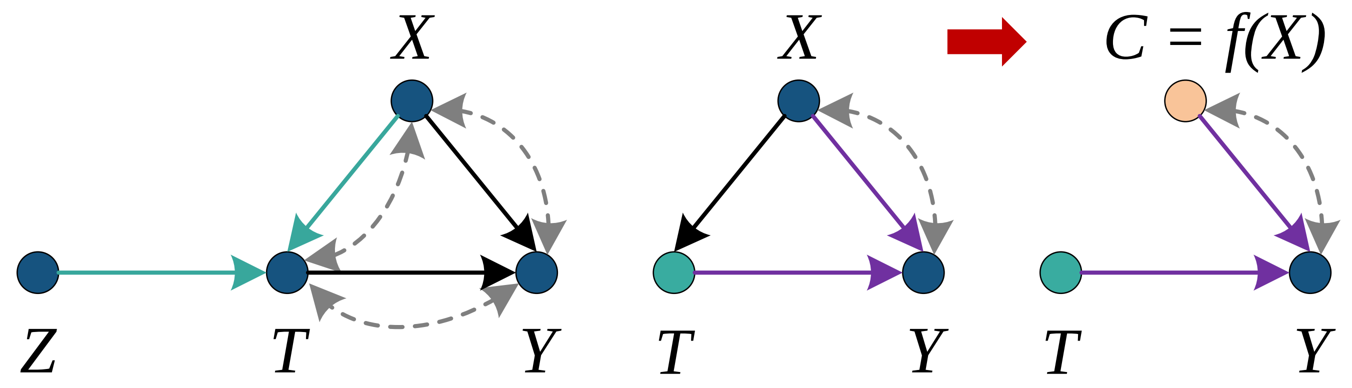
(2) Additive Noise Assumption: $Y = g(T, X) + U$



Instrumental Variable Regression:

- **Stage 1:** $\hat{f} = \arg\min_f \sum_{i=1}^{n} (t_i - P_{f(z_i, x_i)}(t_i \mid z_i, x_i))^2, \hat{t}_i \sim P_{\hat{f}(z_i, x_i)}(t \mid z_i, x_i)$
- **Stage 2:** $\hat{g} = \arg\min_g \sum_{i=1}^{n} (y_i - g(\hat{t}_i, x_i))^2, \hat{t}_i$ is obtained from stage 1.

Motivation

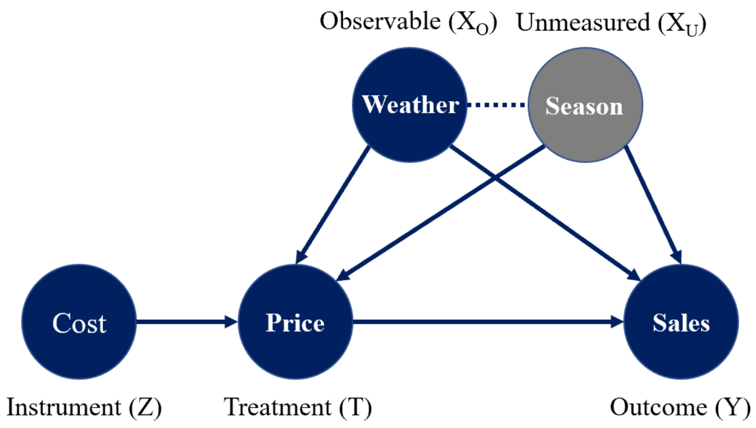


X: Covariate;
Z: Instrument (Seasons);
T: Treatment (Price);
Y: Outcome (Sales);
: Stage 1; : Stage 2.

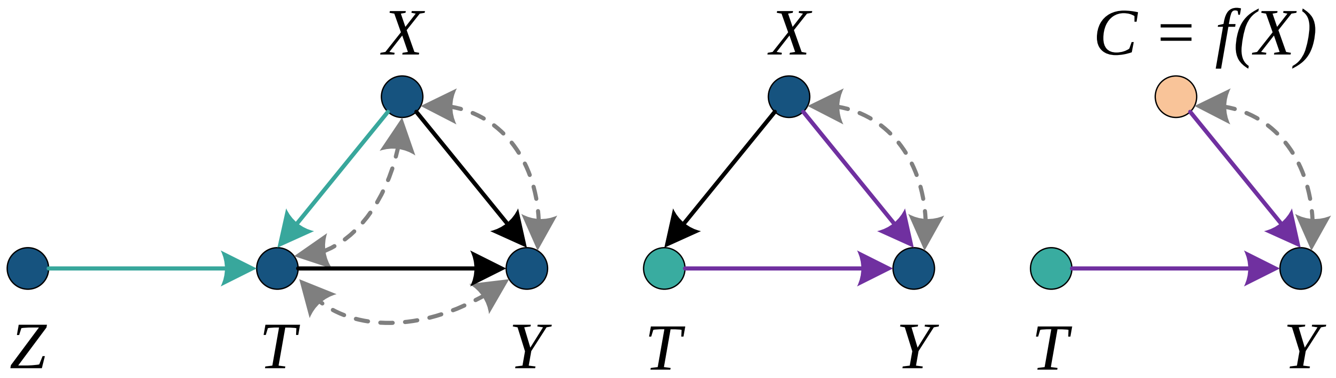
(2) Additive Noise Assumption: $Y = g(T, X) + U$

Instrumental Variable Regression:

- **Stage 1:** $\hat{f} = \arg\min_f \sum_{i=1}^{n} (t_i - P_{f(z_i,x_i)}(t_i \mid z_i, x_i))^2, \hat{t}_i \sim P_{\hat{f}(z_i,x_i)}(t \mid z_i, x_i)$
- **Stage 2:** $\hat{g} = \arg\min_g \sum_{i=1}^{n} (y_i - g(\hat{t}_i, c_i))^2, \hat{t}_i$ is obtained from stage 1.



Theory



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

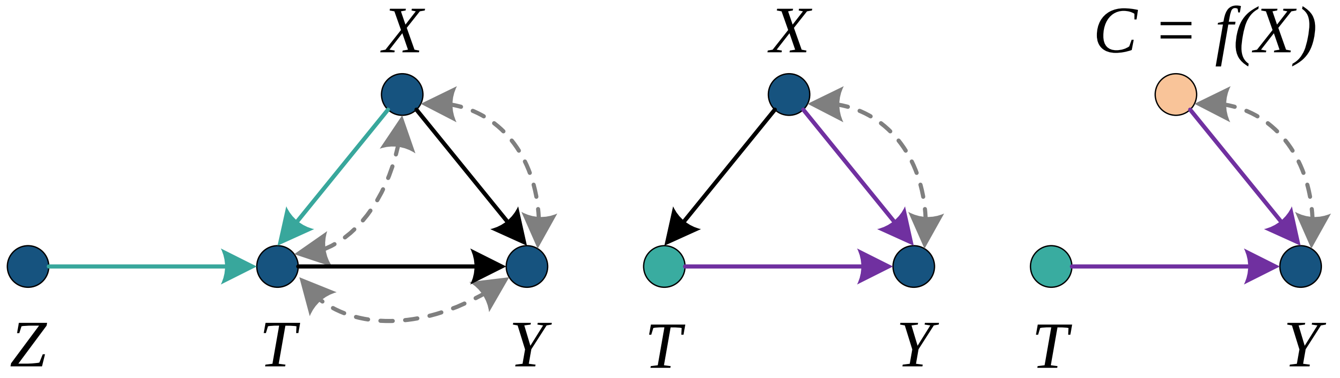
Homogeneity Assumptions: (3) $\mathrm{E}[T \mid Z=a, U] - \mathrm{E}[T \mid Z=b, U] = \mathrm{E}[T \mid Z=a] - \mathrm{E}[T \mid Z=b]$

(4) $\mathrm{E}[Y \mid T=a, U] - \mathrm{E}[Y \mid T=b, U] = \mathrm{E}[Y \mid T=a] - \mathrm{E}[Y \mid T=b]$

Theorem 4.1. ***(Inverse Relationship of Eq. (13)).*** *If the learned representation of observed confounders $C = f_\theta(X)$ is independent with the estimated treatment $\hat{T}$, then the counterfactual prediction function $h(T, C)$ can be identified with instrumental variables Z and representation C. Then, we can establish an inverse relationship for $h(T, C)$ given $\mathbb{E}[Y \mid Z, C, X]$ and $P(T \mid Z, X)$, as follow:*

$$\mathbb{E}[Y \mid Z, C, X] = \int [h(T, C)] \, dP(T \mid Z, X) \tag{15}$$

Method



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

Instrumental Variable Regression with Confounder Balancing (CB-IV)

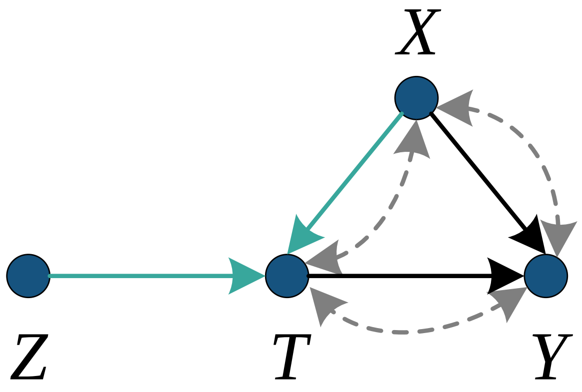
- Treatment Regression (Stage 1)
- Confounder Balancing (Stage 2)
- Outcome Regression (Stage 2)

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:

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



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:

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

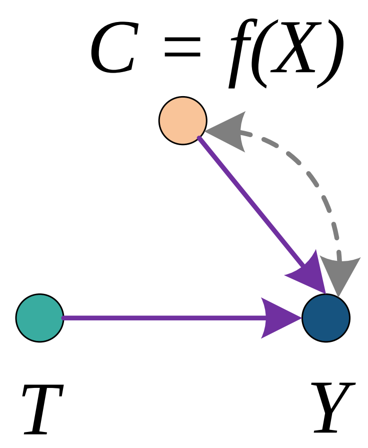
we sample m (the larger the better) treatment $\{\hat{t}_i^j\}_{j=1,\dots,m}$ for each unit $\{z_i,x_i\}$ to approximate the true treatment t_i. Empirically, the above objective (Eq. (27)) is sufficient to accurately estimate causal effects in continuous CB-IV framework.

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:

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



Continuous Cases

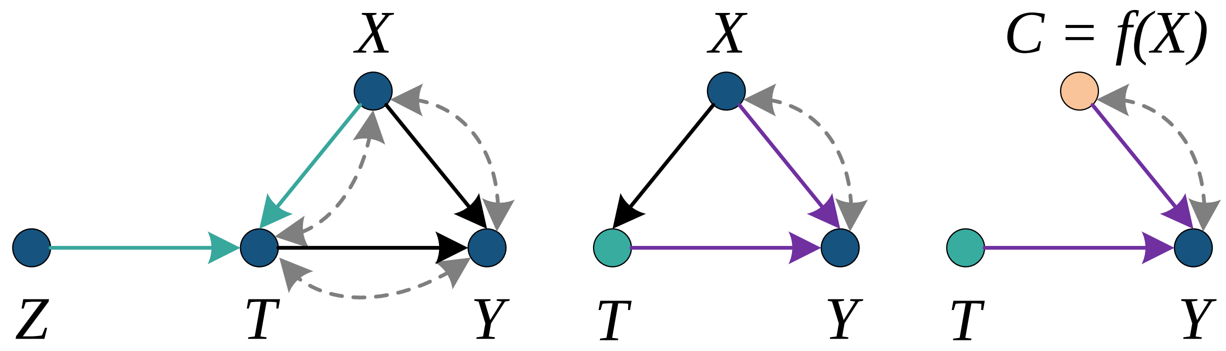
Confounder Balancing: For continuous treatment T, we learn a ”balanced” representation (i.e., C) of the observed confounders X as $C = f_\theta(X)$ via mutual information (MI) minimization constraints (Cheng et al., 2020): firstly, we use variational distribution $Q_\psi(\hat{T} \mid C) = \mathcal{N}(\mu_\psi(C), \sigma_\psi(C))$ 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:

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

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.

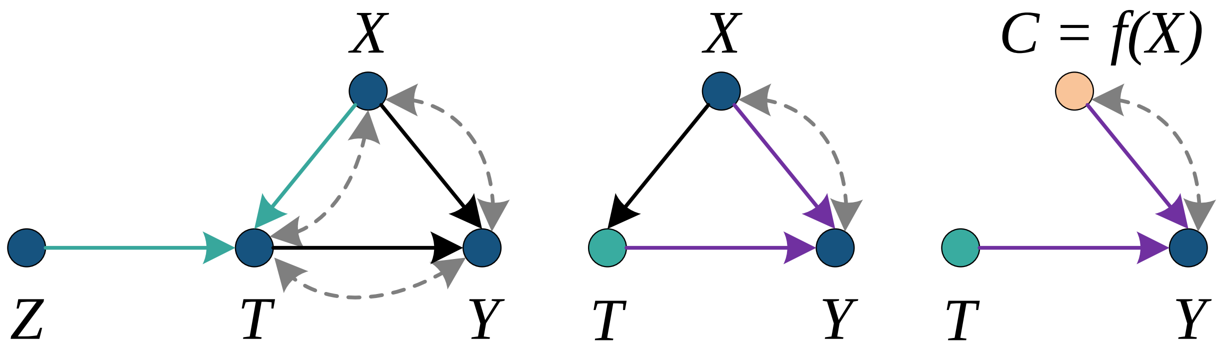
Outcome Regression: Finally, we propose to regress the outcome with the estimated treatment $\hat{T} \sim P(T|Z, X)$ obtained in treatment regression module and the representation of confounders $C = f_\theta(X)$ obtained in confounder balancing module:

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

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.

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:

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

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):

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

Bias of the conditional average treatment effect:

$$\hat{ATE}_i = h(T=t_i, f(x_i)) - h(T=0, f(x_i))$$

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

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

Method	Syn-1-4-4	Syn-2-4-4	Syn-2-10-4	Syn-2-4-10
	Out-of-Sample			
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

Method	Demand-0-1	Demand-0-5	Demand-5-1
	Out-of-Sample		
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

Method	IHDP-2-6-0	IHDP-2-4-2	Twins-5-8-0	Twins-5-5-3
	Out-of-Sample			
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	[illegible]	>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(8.952)	417.5(123.[illegible])	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)**

IV-based Methods

Confounder Balancing Methods

(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

Method	Out-of-Sample 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

Method	Out-of-Sample 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

Method	Out-of-Sample 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

Experimental Results

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

Method	Out-of-Sample 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

Method	Out-of-Sample 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

Method	Out-of-Sample 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;

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.

Experimental Results

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

Method	Out-of-Sample 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

Method	Out-of-Sample 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

Method	Out-of-Sample 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.

Our 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.