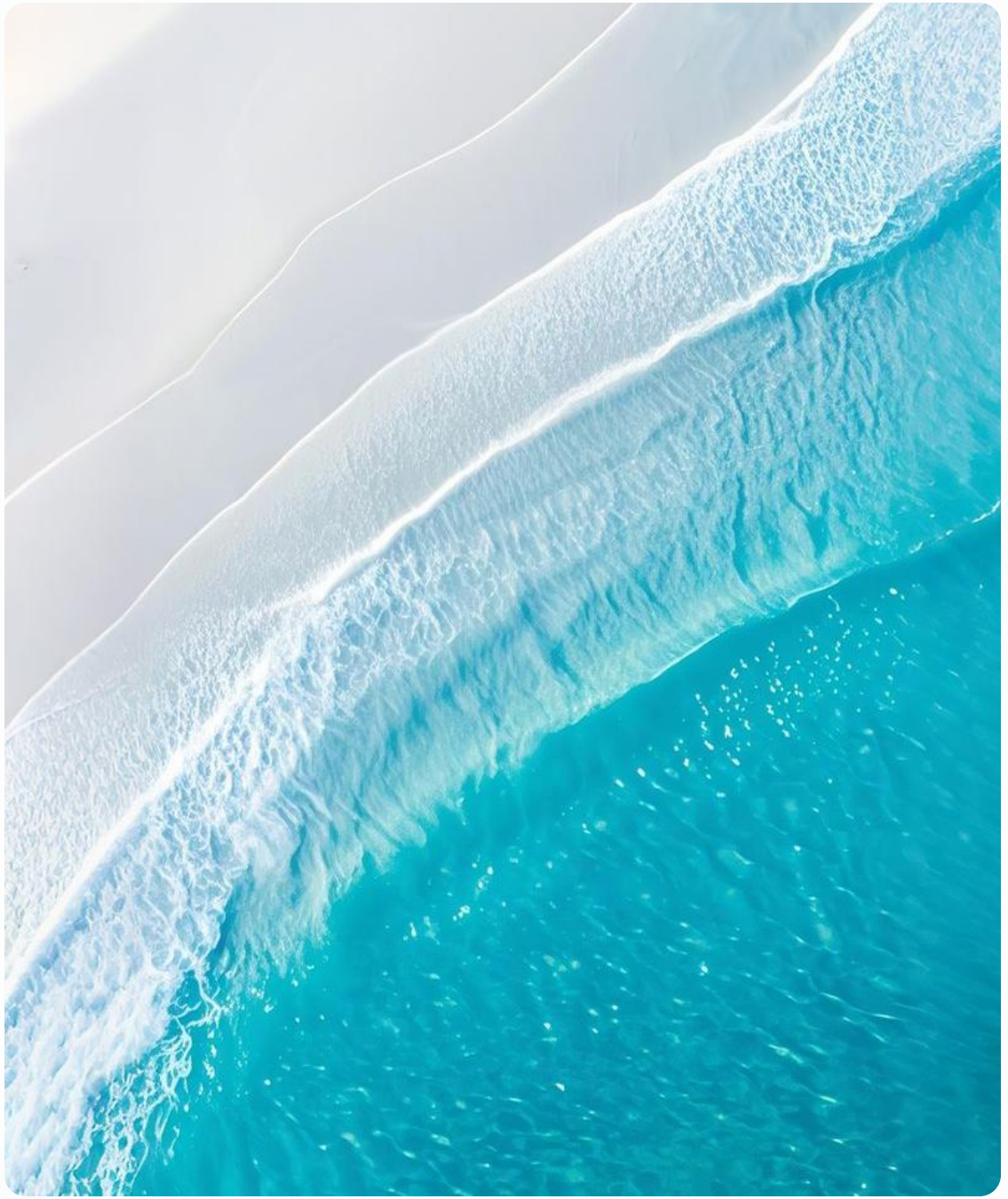
线性回归的非凡威力

Speaker：许文立

wlxu@cityu.edu.mo

August-November, 2025

Faculty of Finance, City University of Macau



CONTENTS

Part. 01

回归为何有效

线性回归的不可思议

线性回归的低估现象

在因果推断领域，线性回归常被低估，其“不合理有效性”并非魔法，而是源于对条件期望的有效刻画。这种现象在实际应用中屡见不鲜，却往往被忽视。

课程动机与疑问

本课程旨在探讨线性回归为何如此有效，其背后的理论基础是什么。通过深入分析，我们将揭开这一现象背后的秘密，为后续深入研究奠定基础。

万物皆可回归

- 两种状态：0和1
- 两个潜在结果：Y0和Y1
- 无法知晓：$\tau_i = Y_{\{1i\}} - Y_{\{0i\}}$
- 观测结果：

$$Y_i = Y_{\{0i\}} + T_i\left(Y_{\{1i\}} - Y_{\{0i\}}\right)$$
$$= (1 - T_i)Y_{\{0i\}} + T_i Y_{\{1i\}}$$

- **平均**处理效应ATE

$$ATE = E[Y_1 - Y_0]$$

- **恒定的**处理效应 κ :

$$\mathrm{Y}_{\{1\mathrm{i}\}} = \mathrm{Y}_{\{0\mathrm{i}\}} + \kappa$$

万物皆可回归

- 用均值差分E[Y|T=1] - E[Y|T=0]估计ATE会有偏误

$$E[Y|T=1]-E[Y|T=0]=$$
$$\underbrace{E[Y_1-Y_0|T=1]}_{ATET}+\underbrace{\{E[Y_0|T=1]-E[Y_0|T=0]\}}_{BIAS}$$

- 随机控制实验（RCT）可以消除偏误
- 一种方式：一阶差分估计量，计算置信区间，进行假设检验

- 另一种方式：线性回归

上网课的效应：回归

- 线性回归方程：

$$exam_i = \beta_0 + \kappa\, Online_i + u_i$$

- Ui包括所有其它影响成绩的因素；

$$E[Y|T=0] = \beta_0$$

$$E[Y|T=1] = \beta_0 + \kappa$$

- 因此，κ就是ATE

上网课的效应：回归

$$exam_i = \beta_0 + \kappa\, Online_i + u_i$$

falsexam	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
format_ol	-4.912222	1.679574	-2.92	0.004	-8.223027	-1.601417
_cons	78.54748	1.113156	70.56	0.000	76.35321	80.74176

- 代码、数据和结果见
- https://wenzhe-huang.github.io/python-causality-handbook-zh/05-The-Unreasonable-Effectiveness-of-Linear-Regression.html

Part. 02

理论回顾：识别假设

条件期望与最小二乘

最小二乘法的核心

线性回归通过最小化均方误差来逼近条件期望E[Y|X]，这是其有效性的关键。只要模型设定正确，最小二乘法就能得到一致估计。

$$\beta^* = \underset{\beta}{argmin}\, E[(Y_i - X_i'\beta)^2]$$

样本估计形式（见代码和数据）：$\hat{\beta} = (X'X)^{-1}X'Y$

条件期望与最小二乘

一元回归：

$$\beta_1 = \frac{Cov(Y_i, T_i)}{Var(T_i)}$$

见代码和数据

https://wenzhe-huang.github.io/python-causality-handbook-zh/05-The-Unreasonable-Effectiveness-of-Linear-Regression.html

条件期望与最小二乘

多元回归：

$$y_i = \beta_0 + \kappa T_i + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_i$$

κ 可通过以下公式求得

$$\kappa = \frac{Cov(Y_i, \tilde{T}_i)}{Var(\tilde{T}_i)}$$

直觉

- 如果我们能通过其他变量预测 T，意味着它并非随机
- 使用线性回归基于其他变量进行预测，并取其回归残差
- 残差与协变量X无关
- https://wenzhe-huang.github.io/python-causality-handbook-zh/05-The-Unreasonable-Effectiveness-of-Linear-Regression.html

非随机数据：教育的回报

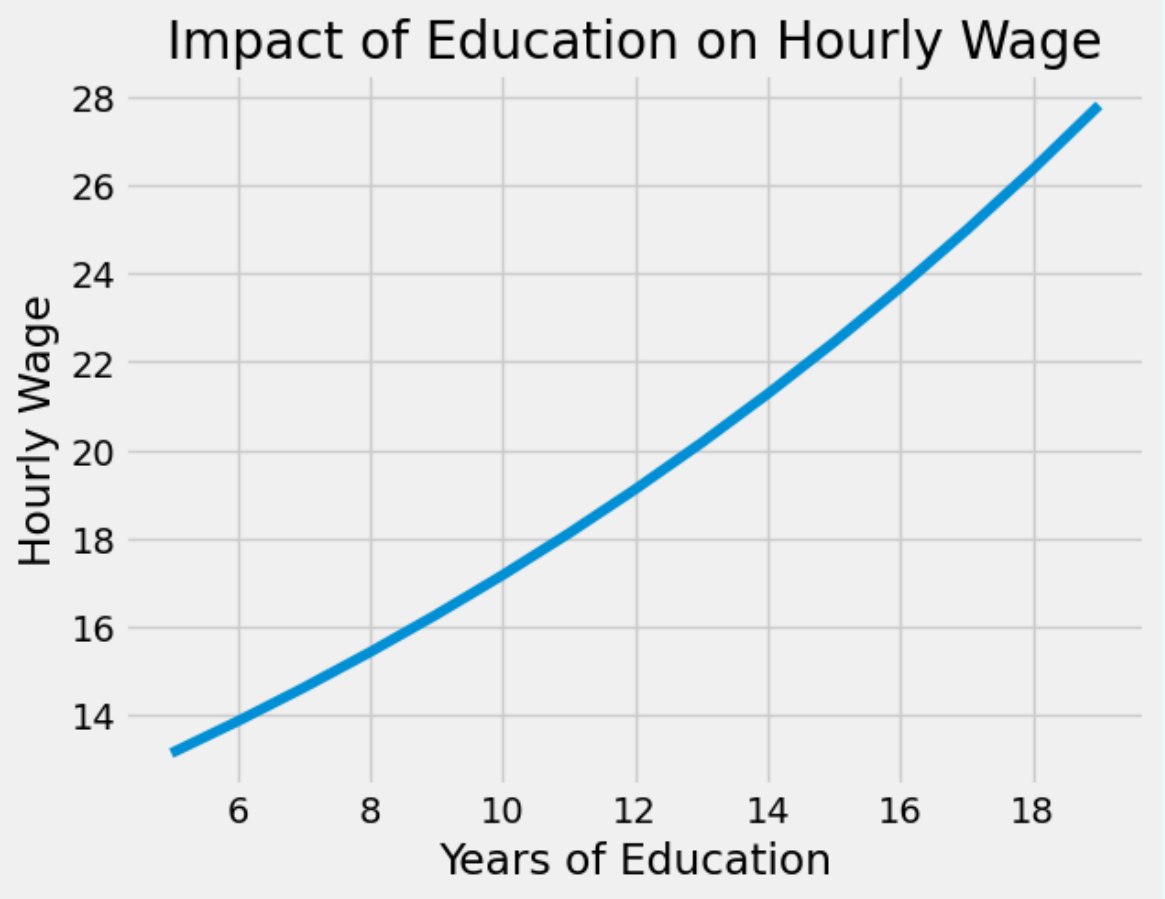
- **教育年限是**处理变量/自变量，工资对数是结果变量
- 注意：工资对数的理解应该是百分比，即教育每增加一年，工资增加x%
- **回**归方程为：

$$log(hwage)_i = \beta_0 + \beta_1 educ_i + u_i$$

lnhwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	.0529472	.0065313	8.11	0.000	.0401295	.065765
_cons	2.295423	.0891276	25.75	0.000	2.120509	2.470337

非随机数据：教育的回报

- **教育年限是**处理变量/自变量，工资对数是结果变量



教育的回报 5.3%（相关还是因果）？

回归方程为：

$$log(hwage)_i = \beta_0 + \beta_1 educ_i + u_i$$

lnhwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	.0529472	.0065313	8.11	0.000	.0401295	.065765
_cons	2.295423	.0891276	25.75	0.000	2.120509	2.470337

- 偏误：研究生与本科生不可比
- 受教育年限可能与IQ有关：即使都是本科，更聪明的人收入也会更高

教育的回报 5.3%（相关还是因果）？

- 控制变量：父母教育程度、IQ、工作年限、工作经验、其它变量
- 回归方程

$$log(hwage)_i = \beta_0 + \kappa\, educ_i + \boldsymbol{\beta} \boldsymbol{X}_i + u_i$$

$$\kappa = \frac{Cov(Y_i, \tilde{T}_i)}{Var(\tilde{T}_i)}$$

1. 该公式表明，我们可以通过父母的教育程度、智商、经验等因素来预测 educ
2. “受教育年限更长的人之所以如此，是因为他们拥有更高的智商。教育并不会带来更高的工资，只是与智商相关，而智商才是驱动工资的因素”

教育的回报 5.3%（相关还是因果）？

- 在回归中包括IQ：保持IQ不变的情况下，每多受一年教育，工资的增长
- 也就是说，回归中的控制变量，就是实现保持处理组和对照组的其它因素在相同的水平（不变）

$$\kappa = \frac{Cov(Y_i, \tilde{T}_i)}{Var(\tilde{T}_i)}$$

```
. // Display the final result
. display "Local average treatment effect (kappa) = " kappa
Local average treatment effect (kappa) = .04114719
```

教育的回报 5.3%（相关还是因果）？

lhwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	.0411472	.0100982	4.07	0.000	.0213181	.0609763
iq	.003791	.0013568	2.79	0.005	.0011267	.0064553
exper	.0153412	.0050593	3.03	0.003	.0054067	.0252757
tenure	.0093548	.003298	2.84	0.005	.0028786	.0158309
age	.0085528	.0062718	1.36	0.173	-.0037626	.0208683
married	.1795054	.0525565	3.42	0.001	.0763041	.2827068
black	-.0800658	.063392	-1.26	0.207	-.204544	.0444124
south	-.0396773	.035137	-1.13	0.259	-.1086732	.0293187
urban	.1925515	.0355389	5.42	0.000	.1227664	.2623365
sibs	.0064728	.0089611	0.72	0.470	-.0111235	.0240691
brthord	-.0079794	.0132029	-0.60	0.546	-.0339049	.0179461
meduc	.0089114	.0070456	1.26	0.206	-.0049235	.0227462
feduc	.0068715	.0061742	1.11	0.266	-.0052523	.0189954
_cons	1.115604	.2323293	4.80	0.000	.6593963	1.571812

遗漏变量与混淆因子

- 教育是工资的原因吗？教育会提高工资4.1%吗？
- 不确定
- 假设教育影响工资的真实模型为：

$$Wage_i = \alpha + \kappa\, Educ_i + A_i'\beta + u_i$$

- 如果回归遗漏了A，那么：

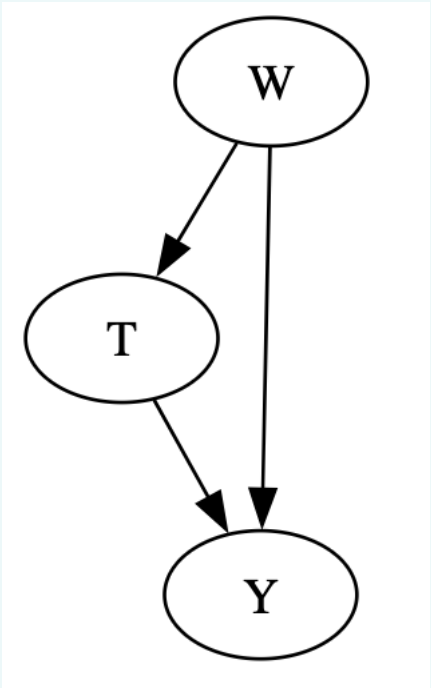
$$\frac{Cov(Wage_i, Educ_i)}{Var(Educ_i)} = \kappa + \beta'\delta_{Ability}$$

遗漏变量与混淆因子

- 首先，如果被忽略的变量对因变量没有影响，那么偏误项将为零。
- 这完全合乎逻辑——在试图理解教育对工资影响时，无需控制与之无关的因素（比如田野百合的高度）。
- 其次，若被忽略的变量对处理变量也无影响，偏误项同样为零。
- 这一点也直观易懂：如果模型中已包含所有影响教育的因素，那么教育对工资的估计影响就不可能混杂着教育与其他同样影响工资变量之间的相关性。
- 简而言之，**若模型中已纳入所有混淆变量，则可认为不存在遗漏变量偏误（OVB）**

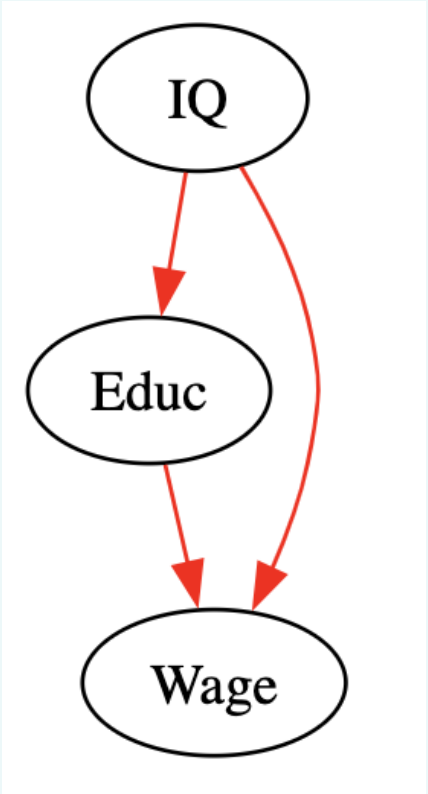
遗漏变量与混淆因子

- 混淆变量是指**同时影响处理变量和结果的变量**



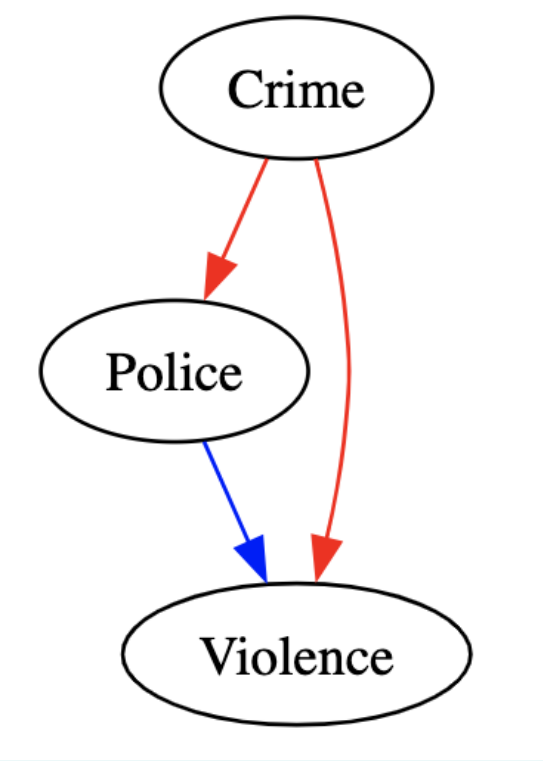
遗漏变量与混淆因子

- 正偏误



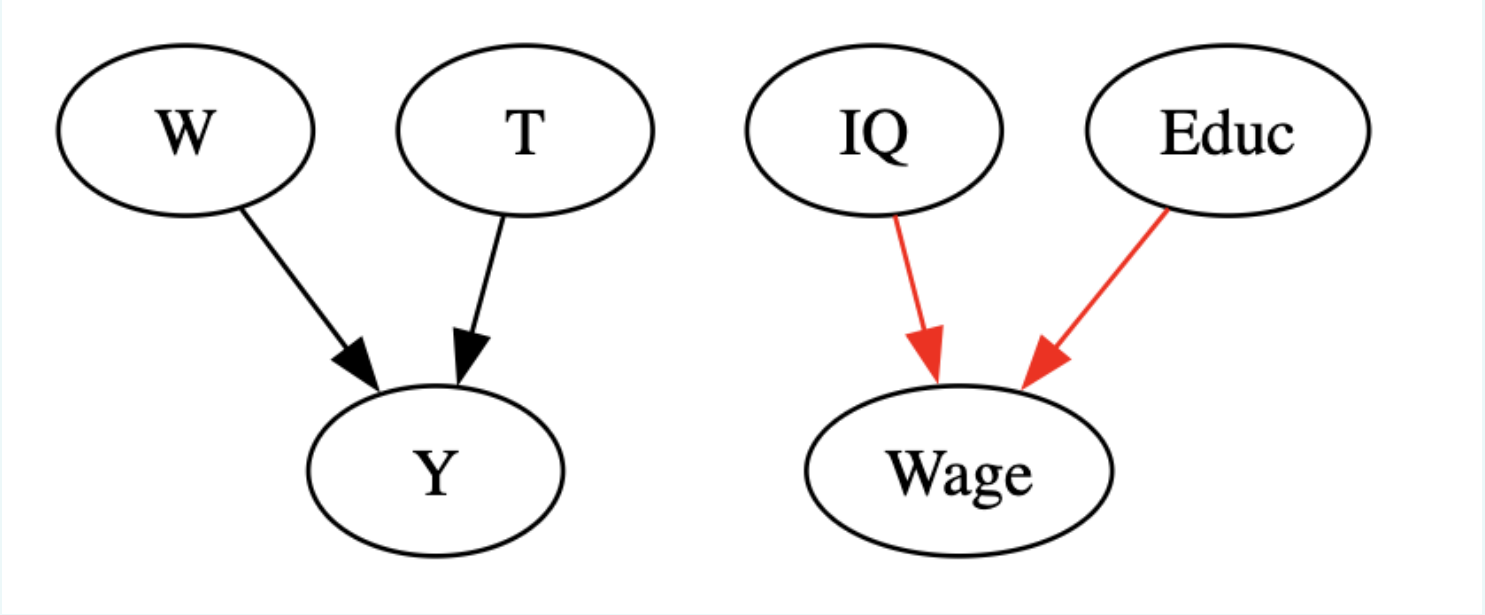
遗漏变量与混淆因子

- 负偏误



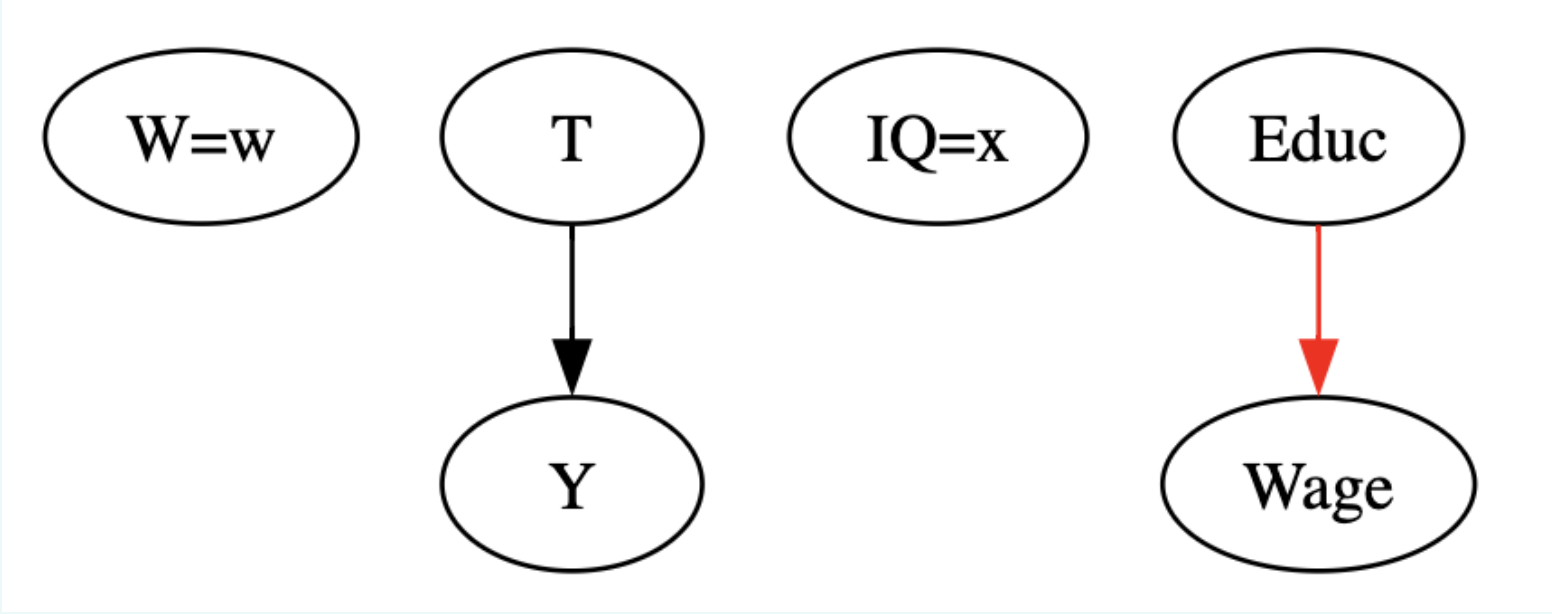
遗漏变量与混淆因子

- 随机实验

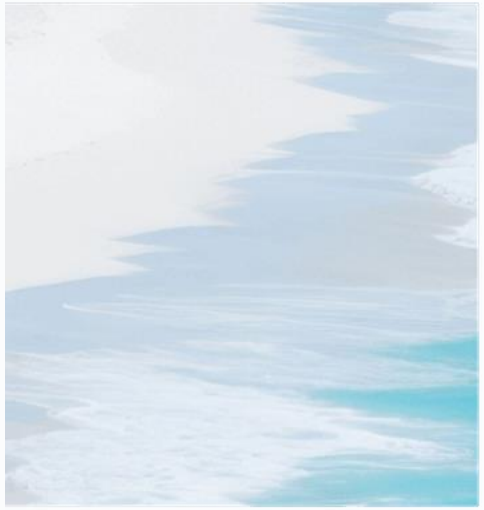


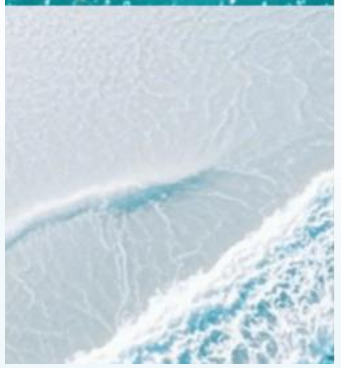
遗漏变量与混淆因子

- 回归



- 这表明，**对于非随机或观察性数据的因果推断，我们应始终保持审慎态度**。
- 永远无法确定所有混杂因素是否已被完全控制。





条件独立假设

条件独立假设的核心

条件独立假设（CIA）是线性回归因果推断的关键。它要求在给定协变量X后，处理变量T与潜在结果独立。这一假设确保了回归系数具有因果解释。

重叠与无混淆

01

重叠假设的作用

重叠假设要求处理组和对照组在协变量的分布上存在重叠，否则会导致权重极端化，影响因果估计的准确性。

02

无混淆假设的重要性

无混淆假设确保所有相关变量都被纳入模型，避免遗漏变量导致的偏差。这一假设是因果推断的基础。

03

违反假设的后果

当重叠或无混淆假设不成立时，因果估计可能出现极端权重和偏差，导致结果不可靠。因此，检验这些假设至关重要。

04

检验策略

可以通过检查协变量的分布和权重的极端值来检验重叠假设。对于无混淆假设，可以利用领域知识和敏感性分析来评估。

Part. 03

模型误设影响

函数形式误设后果

误设下的近似因果效应

即使真实关系非线性或存在交互，线性回归仍可能给出近似的因果效应。这种近似偏差与方差的权衡使得线性回归在有限误设下依然可信。

遗漏变量与代理变量

01 遗漏变量的偏差方向

遗漏混淆变量会导致因果估计的偏差，其方向取决于遗漏变量与处理变量和结果变量的关系。这种偏差可能误导研究结论。

02 代理变量的作用

当无法观测全部混淆变量时，可以通过合理的代理变量来部分恢复因果效应。代理变量的选择需要结合领域知识。

03 提高模型稳健性

引入代理变量可以有效提高模型的稳健性，减少因遗漏变量导致的偏差，从而提高因果估计的可靠性。

04 实际应用中的挑战

在实际应用中，选择合适的代理变量是一个挑战，需要研究者具备深厚的领域知识和对数据的深入理解。

Part. 04

实践与稳健策略

诊断与可视化工具

残差图的作用

残差图可以帮助识别模型假设的偏离，通过观察残差的分布，研究者可以发现潜在的问题并及时调整模型。

杠杆值与影响点诊断

杠杆值和影响点诊断工具可以帮助识别极端权重和异常值，这些值可能对模型估计产生重大影响。

指导实践调整

通过这些诊断工具，研究者可以及时调整模型或采集更多数据，确保模型假设的合理性，从而提高结论的可靠性。

正则化与双机器学习

01 正则化方法

Lasso和Ridge等正则化方法在高维场景下可以有效降低方差，提高模型的稳定性和预测能力。

02 双机器学习框架

双机器学习框架通过交叉拟合nuisance参数，进一步提升因果估计的稳健性和效率，适用于复杂的因果推断问题。

Part. 05

总结与展望

线性回归的边界与超越

线性回归的优势

线性回归在因果推断中具有简单易用、解释性强等优势，适用于许多实际问题。

01

线性回归的局限

然而，线性回归也有其局限性，例如对模型假设的依赖较强，当假设严重违背时，结果可能不可靠。

02

正确假设的重要性

正确识别和检验模型假设是确保线性回归有效性的关键，研究者需要结合领域知识进行判断。

03

超越线性回归

当线性回归的假设无法满足时，研究者可以考虑非参数或机器学习方法，以获得更准确的因果估计。

04

课程回顾与思考

核心概念回顾

本课程回顾了条件期望、识别假设、加权视角、模型误设四大核心概念，强调了在实践中持续检验假设和结合领域知识的重要性。

THANK YOU