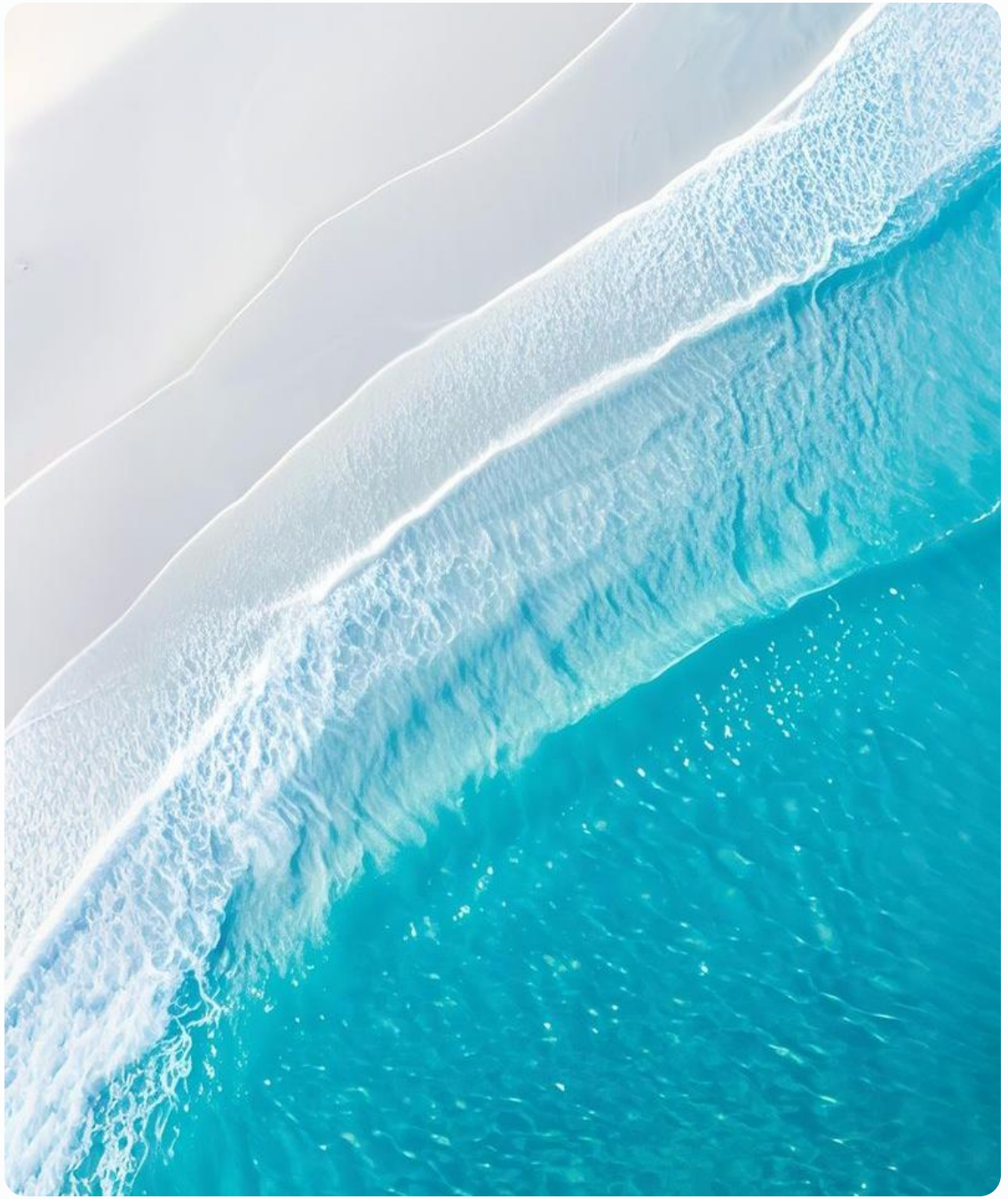
分组回归与虚拟变量

Speaker：许文立

wlxu@cityu.edu.mo

August-November, 2025

Faculty of Finance, City University of Macau



CONTENTS

课程导入 01

分组回归原理 02

虚拟变量回归 03

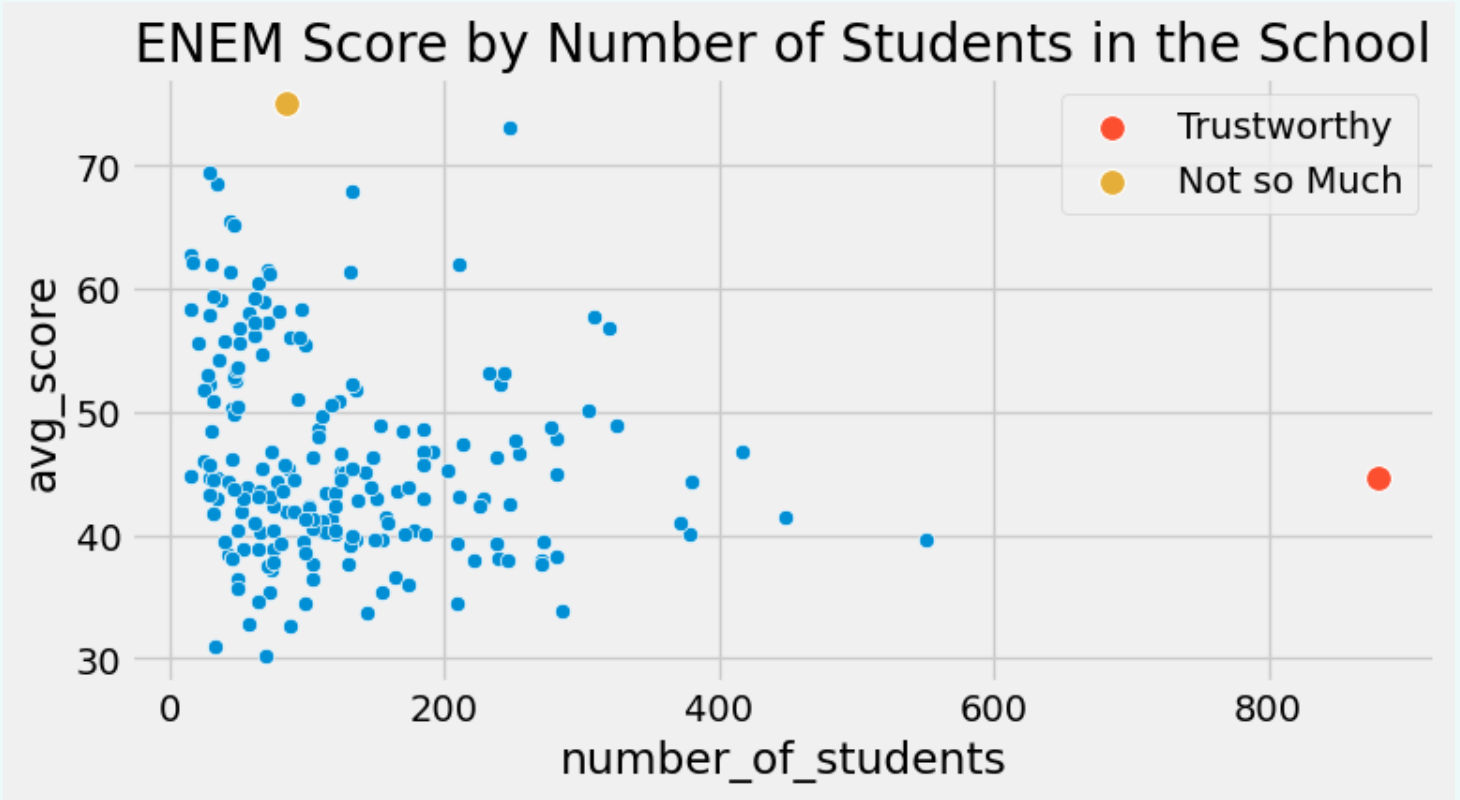
总结与展望 04

Part. 01

课程导入

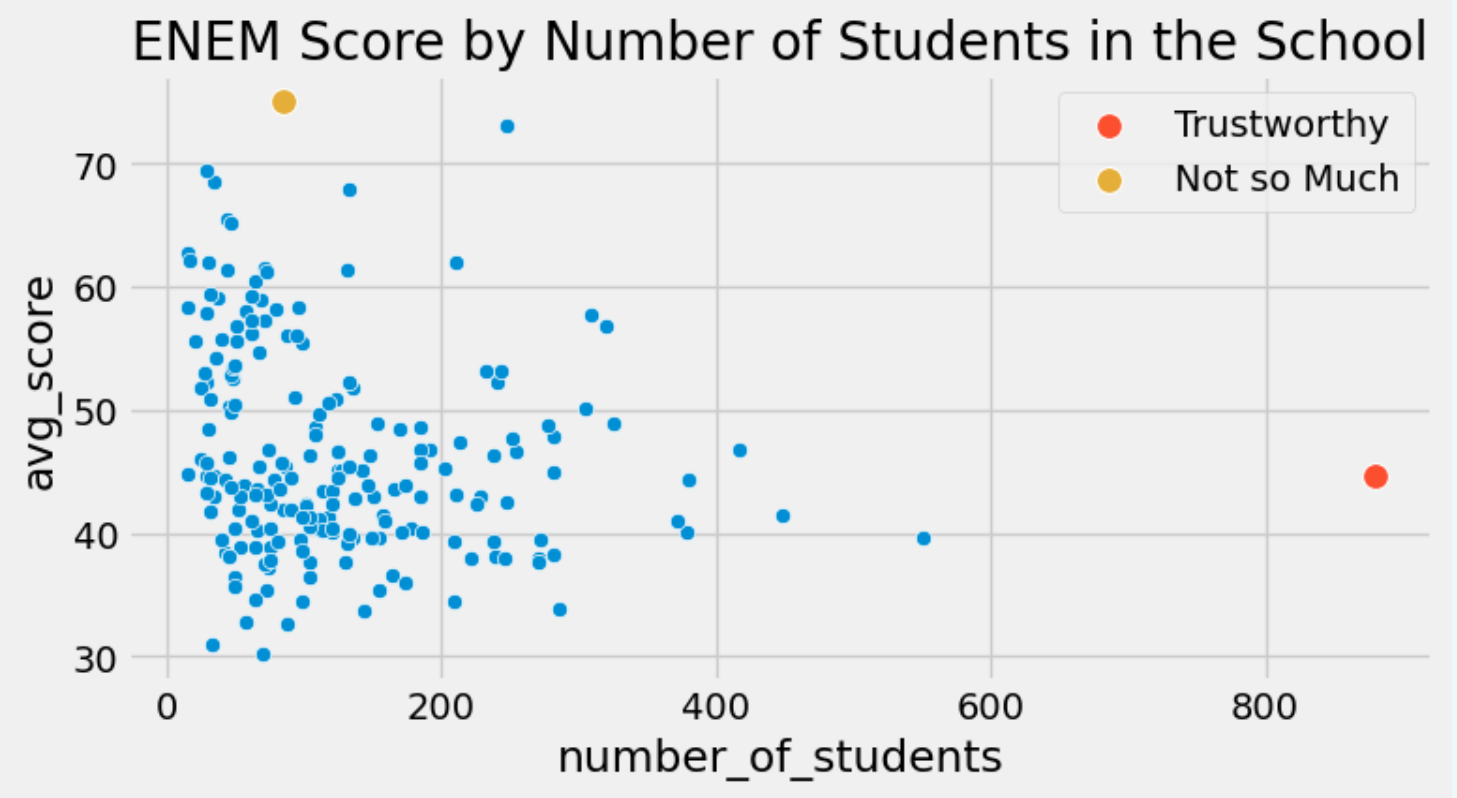
分组数据回归

- 再次看看巴西ENEM数据集：
- 相较于小规模学校的平均成绩，更相信大规模学校的成绩



分组数据回归

- 存在一个低方差区域与另一个高方差区域并存的情况，被称为异方差性
- 导致异方差最常见的原因仍是分组数据



课程目标与核心问题

课程意义

通过分组回归与虚拟变量，比较不同子群体的因果效应，回答‘政策是否对所有人都一样有效’这一核心问题。

课程特色

本课程将用在线学习平台实验数据贯穿始终，帮助学生建立从理论到代码的完整认知。

未分组数据回归：教育的回报

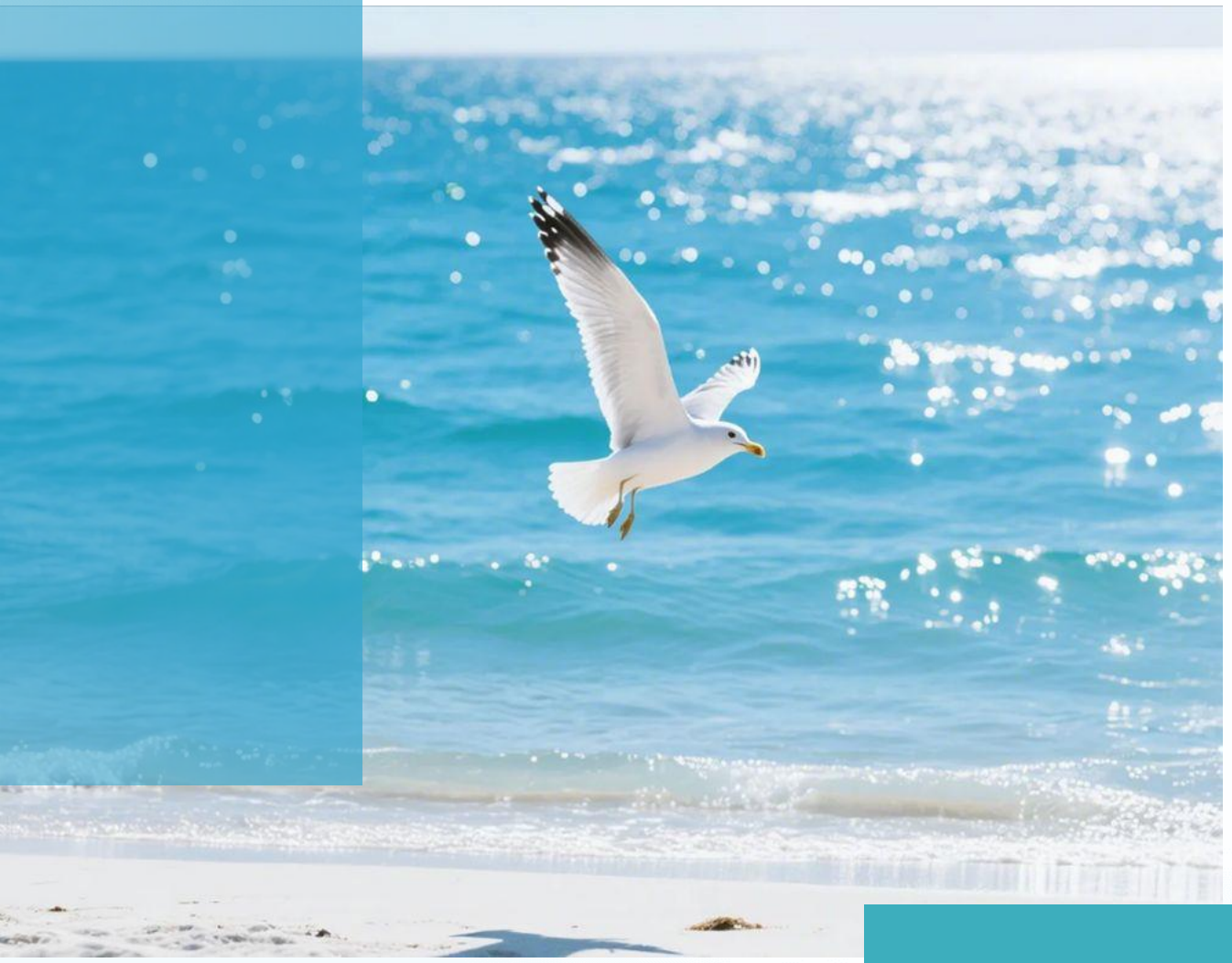
- **教育年限是**处理变量/自变量，工资对数是结果变量
- 注意：工资对数的理解应该是百分比，即教育每增加一年，工资增加x%
- **回**归方程为：

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

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

Part. 02

分组回归原理



分组回归定义与公式

定义与公式

分组回归是按协变量切分样本后分别估计Y~T的系数。公式为Y=α_g+τ_g·T+ε，τ_g即组内ATE，直观展示异质效应。

分组数据回归：教育的回报

- 受数据隐私限制，不能提供个体层面的信息，例如，中国人口抽样调查数据
- 按受教育年限分组，仅提供每组的平均小时工资对数及组内人数

	educ	lhwage	count
0	9	2.856475	10
1	10	2.786911	35
2	11	2.855997	43
3	12	2.922168	393
4	13	3.021182	85
5	14	3.042352	77
6	15	3.090766	45
7	16	3.176184	150
8	17	3.246566	40
9	18	3.144257	57

回归加权视角

- 别担心!
- 回归分析并不依赖大数据也能发挥作用!
- 我们可以为线性回归模型赋予权重，这样它会更重视样本量较大的组别，而非小样本组

回归即加权估计量

加权平均的视角

最小二乘法（OLS）可以被重写为加权平均的形式，其中权重由协变量与处理变量的关系决定。这种视角帮助我们理解线性回归的稳健性。

隐含的权重分配

线性回归隐含地对样本赋予不同的权重，这种权重分配方式使得模型能够更好地拟合数据，但也可能导致极端权重的出现。

分组数据回归：加权最小二乘

	(1)	(2)
	OLS	WLS
educ	**0.0529*** **	**0.0529*** **
	(8.11)	**(9.23)**
_cons	**2.295*** **	**2.295*** **
	(25.75)	**(29.33)**
N	**935**	**10**
r2_a	**0.065**	**0.903**

t statistics in parentheses

* p<0.05, ** p<0.01, *** p<0.001

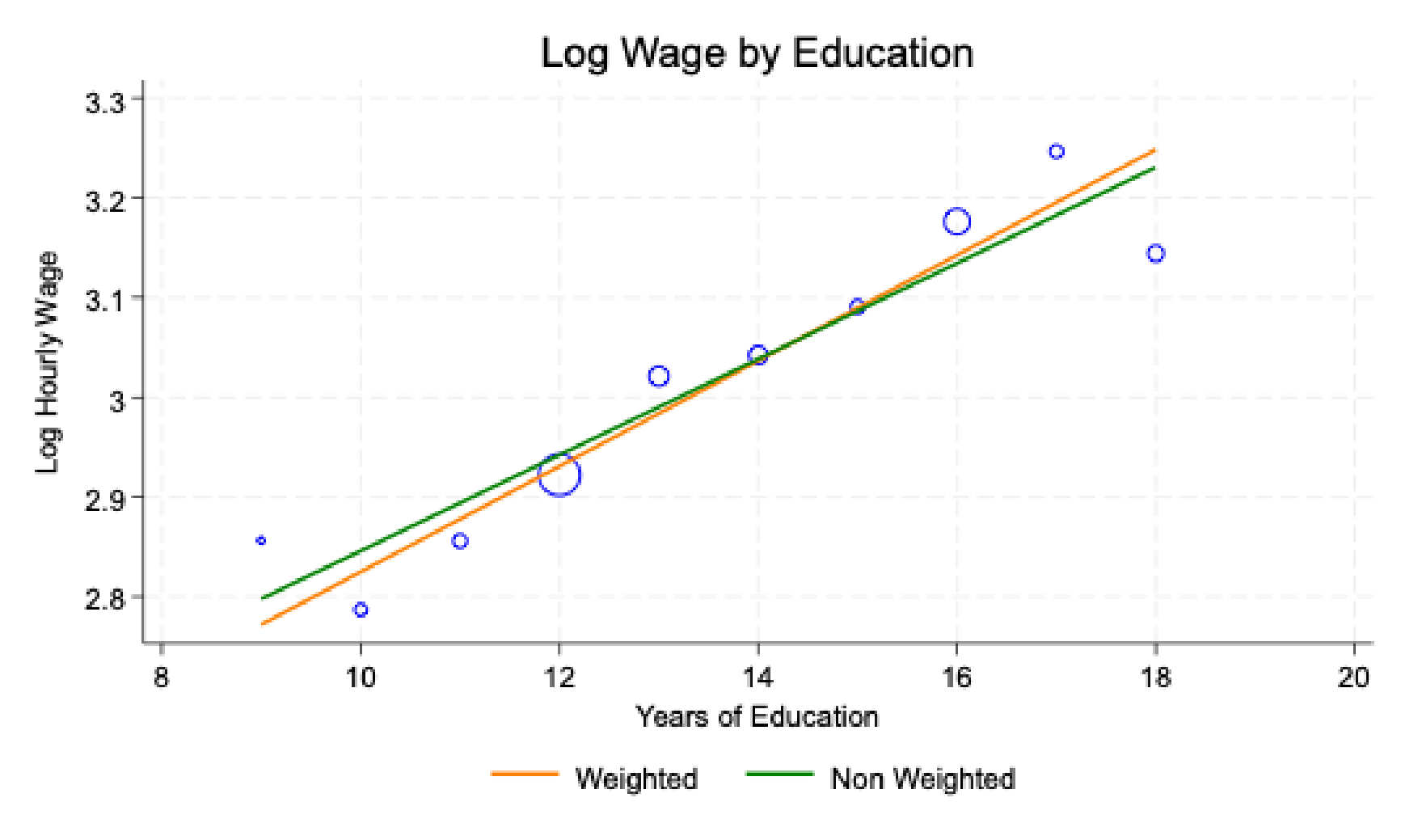
分组数据回归：加权最小二乘

	(1) Original OLS	(2) WLS	(3) Grouped OLS
educ	**0.0529*** **	**0.0529*** **	**0.0481*** **
	(8.11)	**(9.23)**	**(8.14)**
_cons	**2.295*** **	**2.295*** **	**2.365*** **
	(25.75)	**(29.33)**	**(28.99)**
N	**935**	**10**	**10**
r2	**0.066**	**0.914**	**0.892**

t statistics in parentheses

* p<0.05, ** p<0.01, *** p<0.001

分组数据回归：加权最小二乘



分组数据回归：加权最小二乘

	(1) Original OLS	(2) WLS	(3) Grouped OLS	(4) With Covar~e
educ	**0.0529*** **	**0.0529*** **	**0.0481*** **	**0.0257**
	(8.11)	**(9.23)**	**(8.14)**	**(1.20)**
iq				**0.00770**
				(1.31)
_cons	**2.295*** **	**2.295*** **	**2.365*** **	**1.882*** **
	(25.75)	**(29.33)**	**(28.99)**	**(5.80)**
N	**935**	**10**	**10**	**10**
r2	**0.066**	**0.914**	**0.892**	**0.931**

t statistics in parentheses

* p<0.05, ** p<0.01, *** p<0.001

权重极端化风险

1 协变量分布偏移的影响

当协变量的分布发生偏移时，权重可能会趋向极端，导致方差膨胀和估计不稳定。这种现象在实际应用中需要特别注意。

2 极端权重的风险

极端权重可能导致模型对某些样本过度依赖，从而影响估计的准确性和可靠性。因此，检查最大权重和杠杆值是必要的。

3 避免误导性结论

为了避免误导性结论，研究者需要在实践中检查权重的分布，及时调整模型或采集更多数据，以确保结果的可靠性。

分组回归优缺点

01

优点

直观、易解释、可直接观测异质性。

02

缺点

样本量折损导致方差增大。

03

缺点

组多时多重检验风险。

04

缺点

无法直接检验组间差异显著性。

Part. 03

虚拟变量回归

虚拟变量

- 假设你有一个希望纳入模型的性别变量，该变量被编码为两类：男性、女性

gender（性别）
male
female
female
other
male

虚拟变量

- 假设你有一个希望纳入模型的性别变量，该变量被编码为两类：男性、女性

gender（性别）	female	other
male	0	0
female	1	0
female	1	0
other	0	1
male	0	0

虚拟变量

- 读大学对工资的影响

	hwage	IQ	T
0	19.225	93	0
1	16.160	119	1
2	20.625	108	1
3	16.250	96	0
4	14.050	74	0

虚拟变量

- 读大学对工资的影响

hwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
T	**4.90438**	**.626382**	**7.83**	**0.000**	**3.675099**	**6.133661**
_cons	**19.94048**	**.436477**	**45.69**	**0.000**	**19.08389**	**20.79707**

在此案例中，当个体未完成 12 年级学业（虚拟变量关闭）时，其平均收入为 19.9。若完成 12 年级学业（虚拟变量开启），预测值即平均收入则为 24.8449（19.9405 + 4.9044）。因此，虚拟变量的系数捕捉了均值差异，本例中该差异值为 4.9044。

虚拟变量

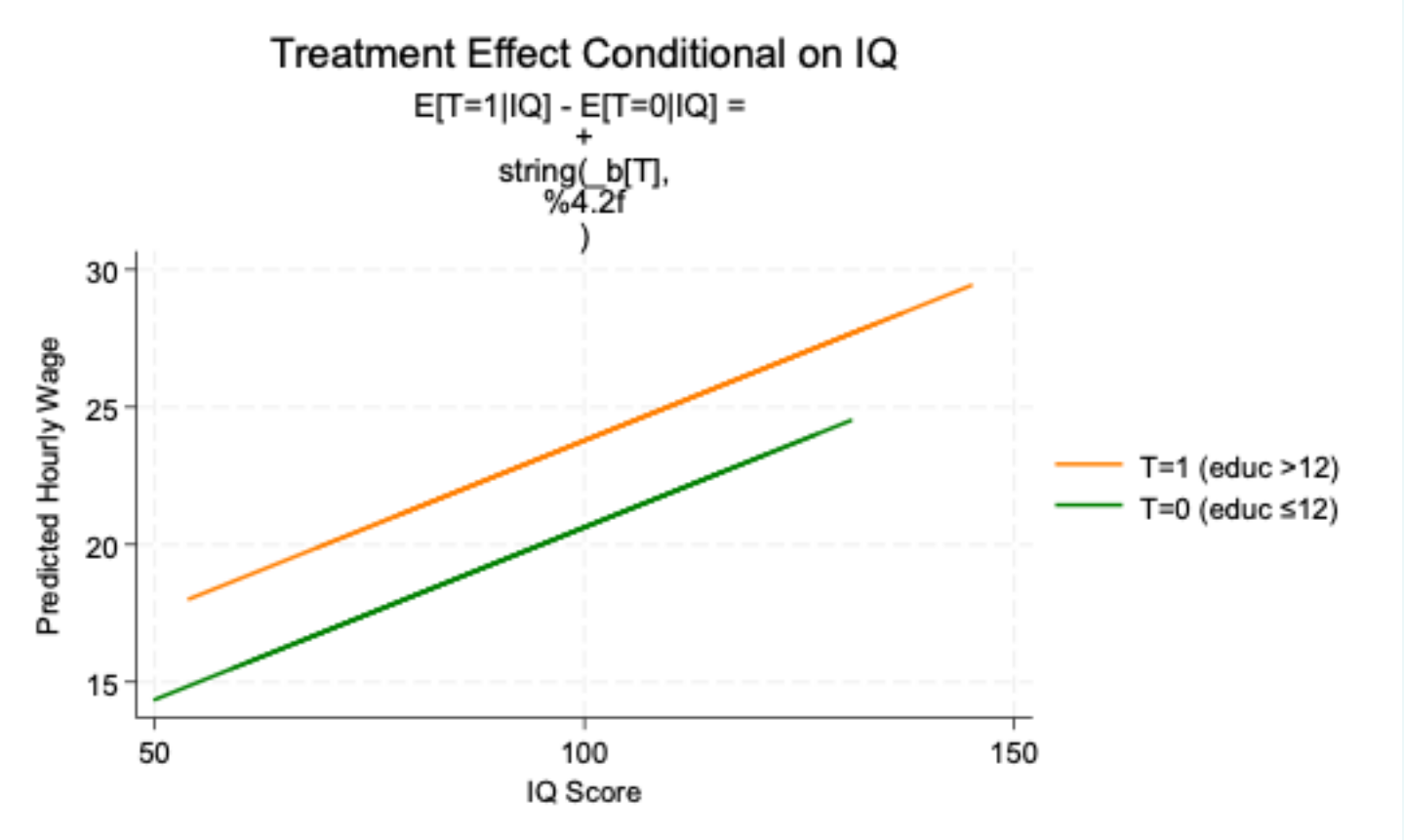
更正式地说，当自变量为二分变量时（如处理指标常见的情形），回归能完美捕捉平均处理效应（ATE）。这是因为回归是对条件期望函数（CEF）$E[Y|X]$ 的线性近似，而在此特定情境下，CEF 本身就是线性的。具体而言，我们可以定义 $E[Y_i|T_i=0]=\alpha$ 和 $E[Y_i|T_i=1]=\alpha+\beta$，从而导出如下 CEF 表达式

$$E[Y_i|T_i]=E[Y_i|T_i=0]+\beta T_i=\alpha+\beta T_i$$

而 β 在随机数据情形下即为均值差异或平均处理效应（ATE）

$$\beta=[Y_i|T_i=1]-[Y_i|T_i=0]$$

虚拟变量+协变量


Treatment Effect Conditional on IQ
E[T=1|IQ] - E[T=0|IQ] =
+
string(_b[T],
%4.2f
)
Predicted Hourly Wage
30
25
20
15
50
100
150
IQ Score
T=1 (educ >12)
T=0 (educ ≤12)

虚拟变量+协变量

- 回归方程：

$$wage_i = \beta_0 + \beta_1 T_i + \beta_2 IQ_i + e_i$$

hwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
T	3.157328	.6962329	4.53	0.000	1.790962	4.523694
iq	.1253244	.0231293	5.42	0.000	.0799328	.1707161
_cons	8.095629	2.227923	3.63	0.000	3.723302	12.46796

虚拟变量+协变量

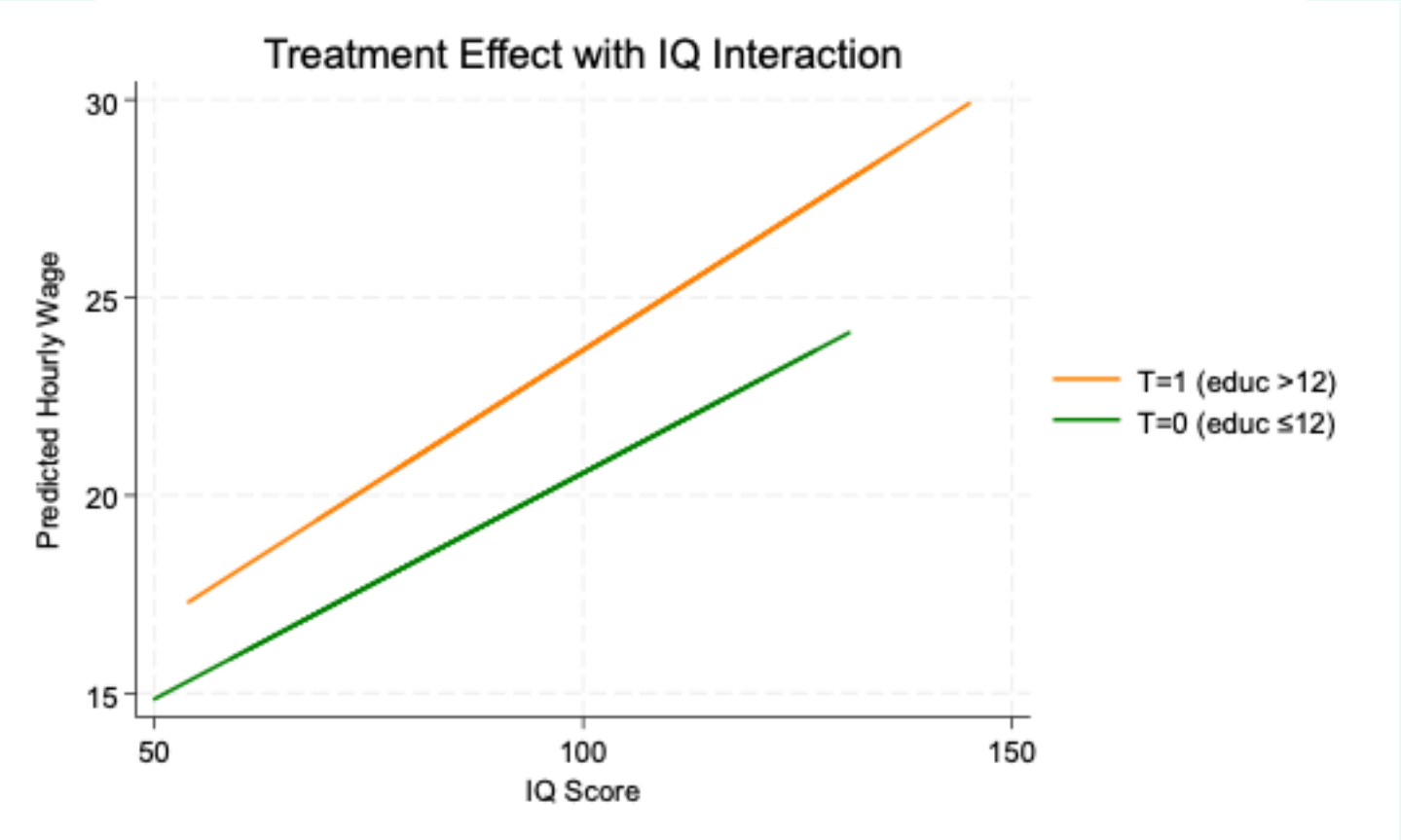
- **加入交互**项的回归方程:

$$wage_i = \beta_0 + \beta_1 T_i + \beta_2 IQ_i + \beta_3 IQ_i * T_i + e_i$$

hwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
T	.6860437	4.79099	0.14	0.886	-8.716348	10.08844
iq	.1142338	.031431	3.63	0.000	.05255	.1759176
c.T#c.iq	.0242121	.0464404	0.52	0.602	-.0669279	.1153521
_cons	9.143843	3.001636	3.05	0.002	3.253086	15.0346

虚拟变量+协变量

- **加入交互**项的回归方程:



两个虚拟变量交乘模型设定

模型设定

虚拟变量回归在单一模型中加入组虚拟变量及其与处理的交互项，公式为$Y=\alpha+\beta\cdot T+\gamma\cdot G+\delta\cdot(T\times G)+\varepsilon$。

模型意义

δ即组间处理效应差异，可直接检验组间差异显著性。

虚拟变量×虚拟变量

- 我们将智商（IQ）离散化为 4 个区间，并将受教育年限视为类别变量

$$wage_i = \beta_0 + \beta_1 T_i + \beta_2 IQ_i + \beta_3 IQ_i * T_i + e_i$$

hwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ						
10	-.7874062	3.414083	-0.23	0.818	-7.487654	5.912841
11	.1083727	3.342763	0.03	0.974	-6.451906	6.668652
12	1.747866	3.049005	0.57	0.567	-4.235903	7.731636
13	4.32896	3.183129	1.36	0.174	-1.918032	10.57595
14	4.088807	3.200488	1.28	0.202	-2.192252	10.36987
15	6.301331	3.328718	1.89	0.059	-.2313836	12.83405
16	7.222462	3.109684	2.32	0.020	1.119608	13.32532
17	9.590472	3.366332	2.85	0.004	2.983939	16.19701
18	7.368068	3.264389	2.26	0.024	.9616013	13.77453
_cons	18.56	3.010938	6.16	0.000	12.65094	24.46906

虚拟变量×虚拟变量

- 我们将智商（IQ）离散化为 4 个区间，并将受教育年限视为类别变量

$$wage_i = \beta_0 + \beta_1 T_i + \beta_2 IQ_i + \beta_3 IQ_i * T_i + e_i$$

hwage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ						
10	-1.214945	3.392133	-0.36	0.720	-7.872142	5.442252
11	-.4686802	3.331899	-0.14	0.888	-7.007666	6.070306
12	.3400133	3.059154	0.11	0.912	-5.663699	6.343725
13	2.410296	3.206084	0.75	0.452	-3.881773	8.702364
14	1.804044	3.237931	0.56	0.578	-4.550526	8.158614
15	3.859915	3.369186	1.15	0.252	-2.752249	10.47208
16	4.405956	3.170708	1.39	0.165	-1.816686	10.6286
17	6.747029	3.422486	1.97	0.049	.0302629	13.46379
18	4.346324	3.332478	1.30	0.192	-2.193798	10.88645
iq_bins						
Q2	1.421552	.897667	1.58	0.114	-.3401556	3.18326
Q3	2.97169	.9301948	3.19	0.001	1.146145	4.797235
Q4 (Highest)	3.787875	1.021506	3.71	0.000	1.783128	5.792622
_cons	18.41784	2.991103	6.16	0.000	12.54768	24.28801

Part. 04

总结与展望

关键要点与未来扩展

01 课程回顾

回顾分组与虚拟变量回归的核心逻辑、实现步骤与选择标准。

02 未来展望

提示当协变量连续或多维时可考虑交互项或机器学习方法，鼓励学员将异质效应思维迁移到更复杂的因果推断场景。

THANK YOU