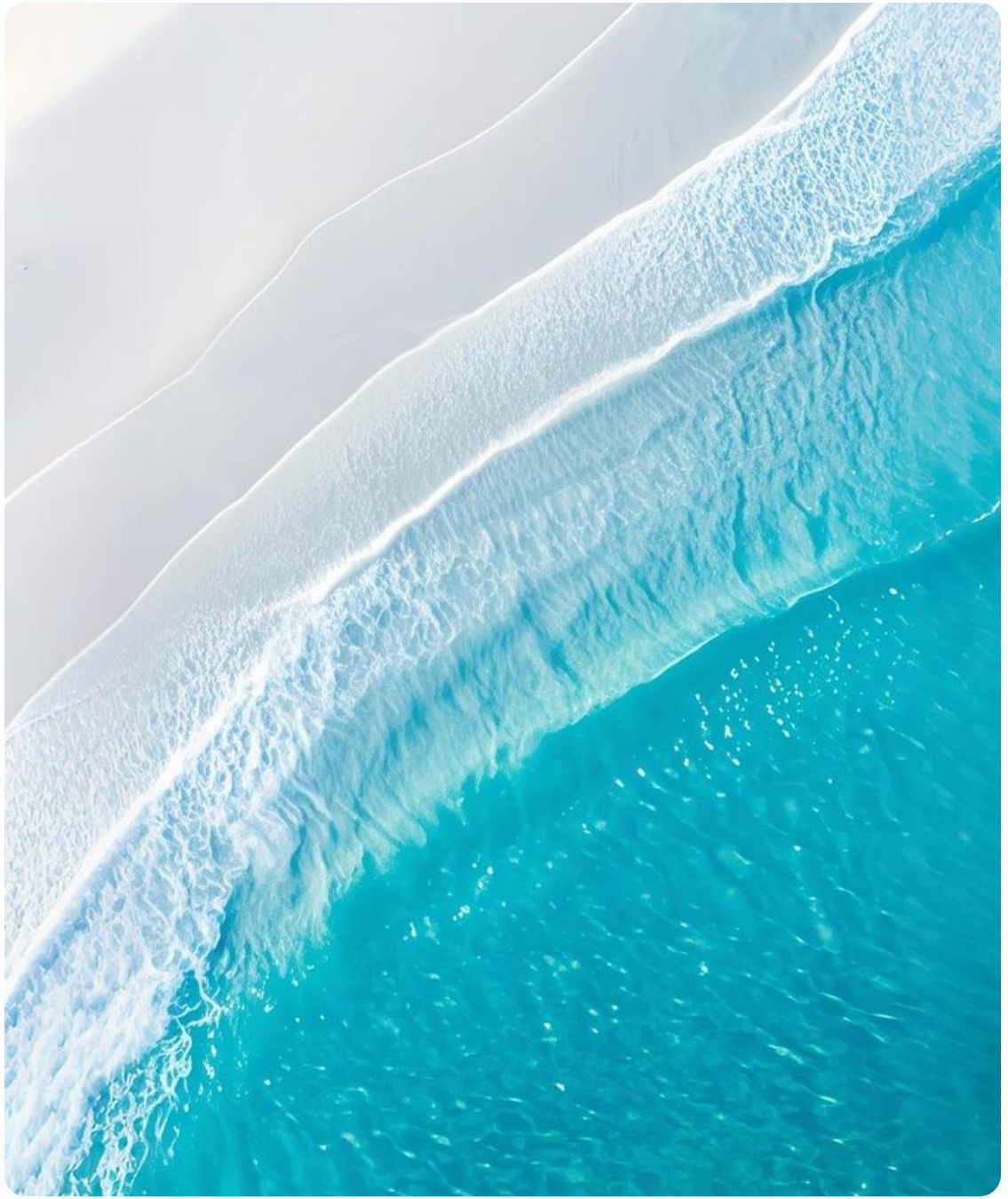
超越混淆因子

Speaker：许文立

wlxu@cityu.edu.mo

August-November, 2025

Faculty of Finance, City University of Macau



CONTENTS

有效的控制变量

- 回归中增加额外的控制变量有助于因果识别
- 混淆因子必须控制
- 将所有可观测变量都作为控制变量纳入回归方程?
- 大数据时代，成百上千个变量，不紧不必要，而且有时候有害
- 例子：在金融机构做数据分析，试图评估和设计账单催款的邮件形式，结果变量是逾期客户的还款金融
- 随机实验：随机抽选5000名客户，抛硬币，为正则向客户发催款邮件，为反面则不发（对照组）

有效的控制变量

	payments	email	opened	agreem~t	credit~t	risk_s~e
1.	740	1	1	0	2348.495	.666752
2.	580	1	1	1	334.112	.2073951
3.	600	1	1	1	1360.661	.5504789
4.	770	0	0	0	1531.829	.5604882
5.	660	0	0	0	979.8557	.4551403

有效的控制变量

- 由于数据是随机的，处理是随机分配，所以满足条件独立性
- 第一种方式：一阶差分估计量FD:

$$ATE = E[Y|T = 1] - E[Y|T = 0]$$

有效的控制变量

```
. * Calculate difference in means
. sum payments if email == 1

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
    payments |      2,454    669.3562    102.0547       330       1140

. scalar mean_treat = r(mean)

.
. sum payments if email == 0

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
    payments |      2,546    669.9764    105.8026       340       1050

. scalar mean_control = r(mean)

.
. display "Difference in means: " %5.2f (mean_treat - mean_control)
Difference in means: -0.62
```

有效的控制变量

- 由于数据是随机的，处理是随机分配，所以满足条件独立性
- 第二种方式：回归

$$\text{Payments} = \beta_0 + \beta_1 \text{Email} + \varepsilon_i$$

payments	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
email	-.6202804	2.941497	-0.21	0.833	-6.386904	5.146343
_cons	669.9764	2.060728	325.12	0.000	665.9365	674.0164

- What弄啥呢？

有效的控制变量

- 给客户发信息催款，居然还会使得客户少还逾期欠款？
- 也有可能，例如，还没到还款日期就打电话给我催催催！烦！
- 此外，p-value也高于0.05，意味着，这一结果可能并没有统计学意义
- 接下来，怎么办？垂头丧气地宣布这个研究没有意义？我们需要更多数据和经费支持？
- 不要轻易放弃！

有效的控制变量

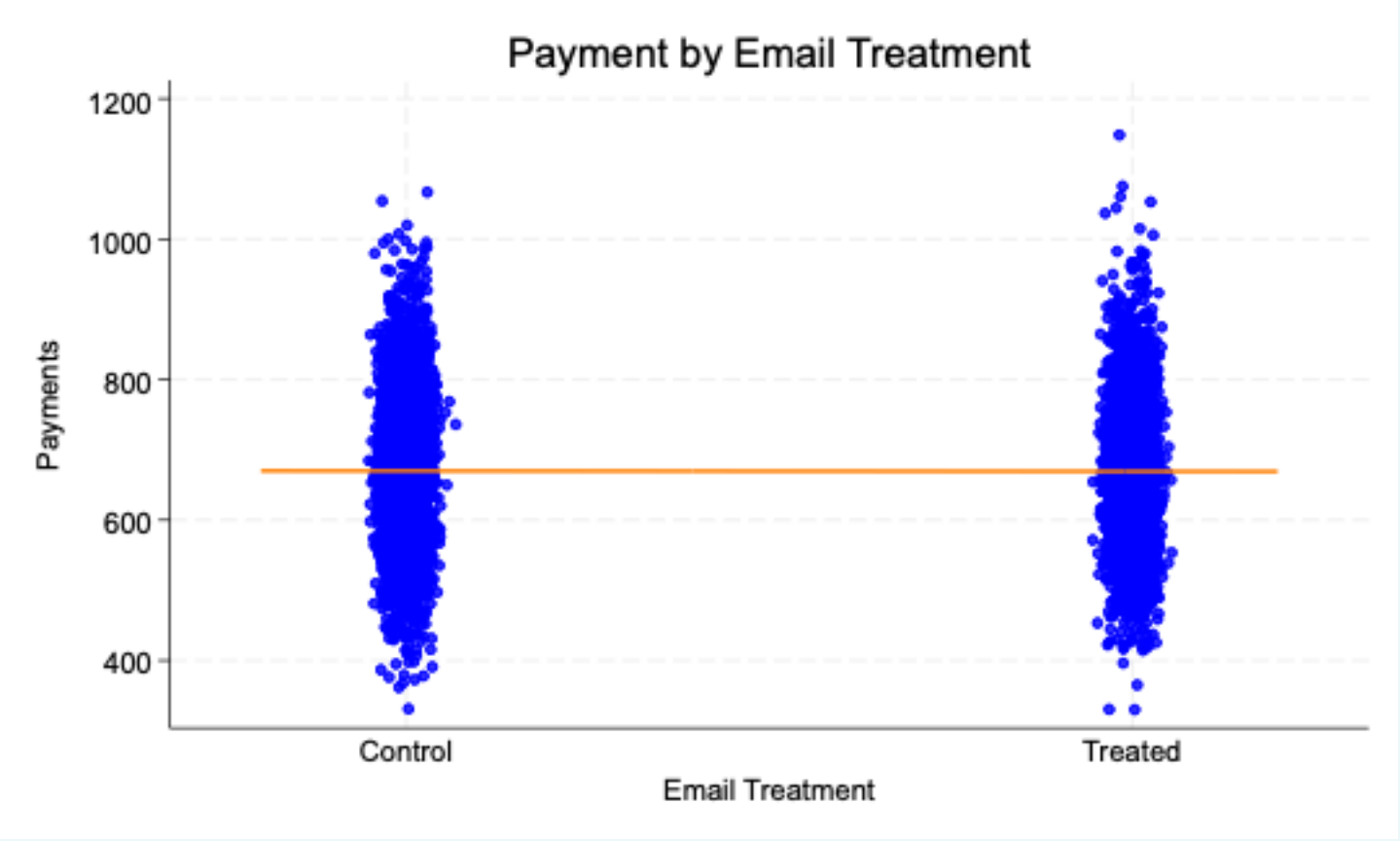
	payments	email	opened	agreem~t	credit~t	risk_s~e
1.	740	1	1	0	2348.495	.666752
2.	580	1	1	1	334.112	.2073951
3.	600	1	1	1	1360.661	.5504789
4.	770	0	0	0	1531.829	.5604882
5.	660	0	0	0	979.8557	.4551403

- 数据里还有其它指标:
- credit_limit 代表客户逾期前的信用额度
- risk_score 对应邮件发送前对客户风险的评估值

有效的控制变量

- 如果处理效应TE存在，为什么无法得到显著的结果?
- 一种可能：效应微乎其微
- 细想一下，促使人们偿还债务的主要因素大多超出了催收部门的控制范围——人们还款是因为找到了新工作、改善了财务状况或收入等
- 用统计学术语来说，**还款行为的变异性更多是由电子邮件之外的其他因素所解释的**

有效的控制变量



- 每个组的还款金融波动非常大
- 如果催款信息的效应仅为5/10块钱，那么，这个效应在[300 1000]的范围就很小

有效的控制变量

- 回归可以帮助降低结果变量的波动性——加入有效的控制变量/协变量
- 若变量能有效预测结果，它就能解释结果的大量方差/波动
- 例如，风险评级和信用额度可以预测还款行为，控制这些变量，能更轻松地识别催款信息对还款金额的效应
- 多元回归的原理：控制住某一变量就是保持这些变量不变时，处理组和控制组的平均结果差异
- 因此，当我们控制风险和信用额度时，其实是在比较具有相同/近风险和信用额度的客户，payments应该也比较接近，波动应当更小

有效的控制变量

- 多元回归的两步回归分解法：
- 第一步：分别用处理变量（发催款邮件）和结果变量（payments）对控制变量（风险和信用额度）回归
- 第二步：用上述两个回归的余值，即结果变量的余值对处理变量的余值回归

有效的控制变量

Variable	Obs	Mean	Std. dev.	Min	Max
payments	5,000	669.672	103.9701	330	1140
email	5,000	.4908	.4999654	0	1
opened	5,000	.2734	.445749	0	1
agreement	5,000	.1608	.3673831	0	1
credit_limit	5,000	1194.845	480.979	193.6956	3882.178
risk_score	5,000	.4808119	.1003763	.1317844	.7734587
email_jitter	5,000	.4907015	.4999052	-.0300313	1.039341
res_email	5,000	5.96e-11	.4992335	-.5323009	.7903931
res_payments	5,000	1.52e-08	75.19032	-261.3965	261.9144

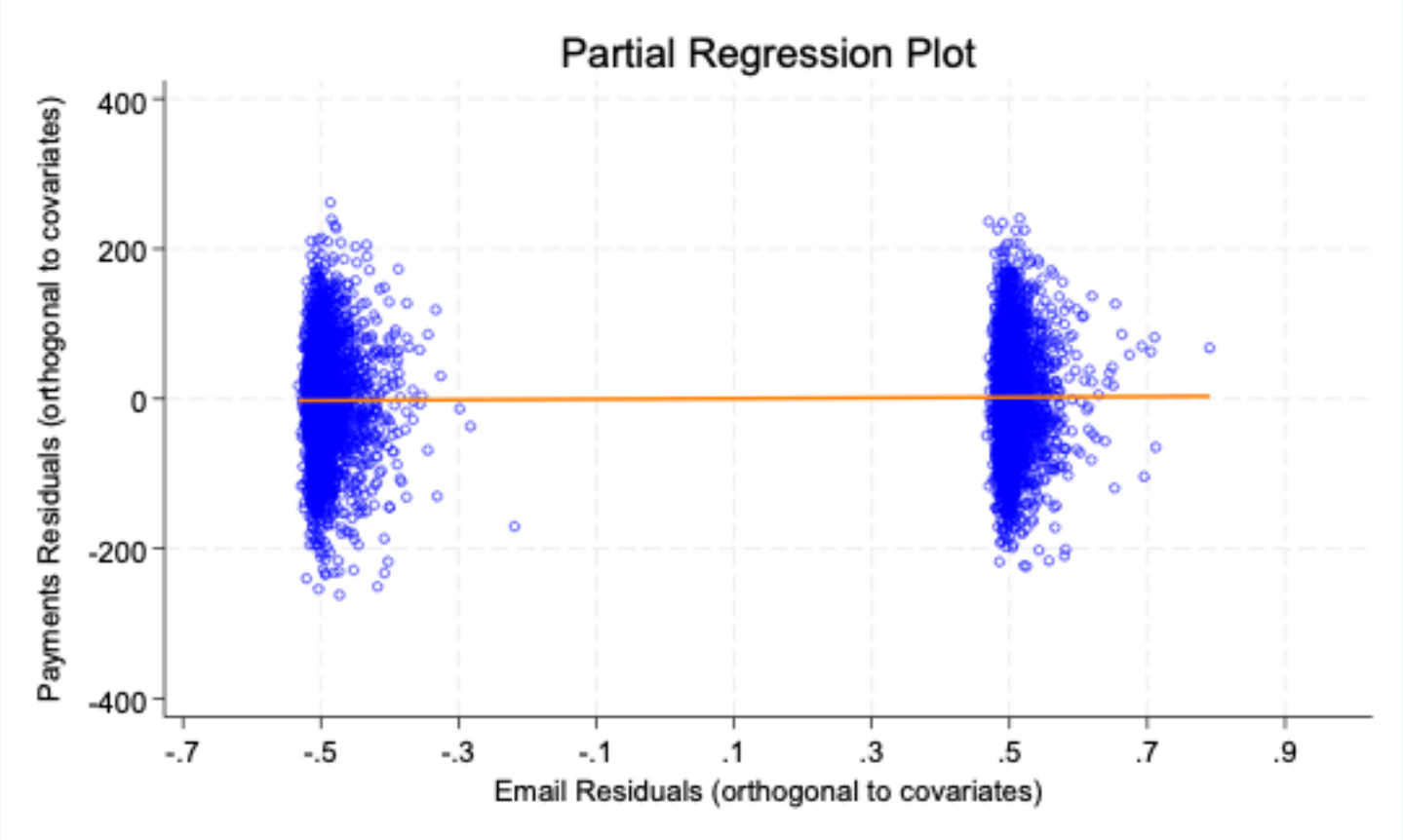
- 控制其它变量后，payments的标准差下降了近一半；
- 催收信息方差没有什么影响

有效的控制变量

- 多元回归的两步回归分解法：
- 第一步：分别用处理变量（发催款邮件）和结果变量（payments）对控制变量（风险和信用额度）回归
- 第二步：用上述两个回归的余值，即结果变量的余值对处理变量的余值回归

res_payments	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
res_email	**4.430355**	**2.129473**	**2.08**	**0.038**	**.255654**	**8.605057**
_cons	**1.49e-08**	**1.062998**	**0.00**	**1.000**	**-2.083942**	**2.083942**

有效的控制变量



- 每个组的还款金融波动非常已经缩小了很多

有效的控制变量

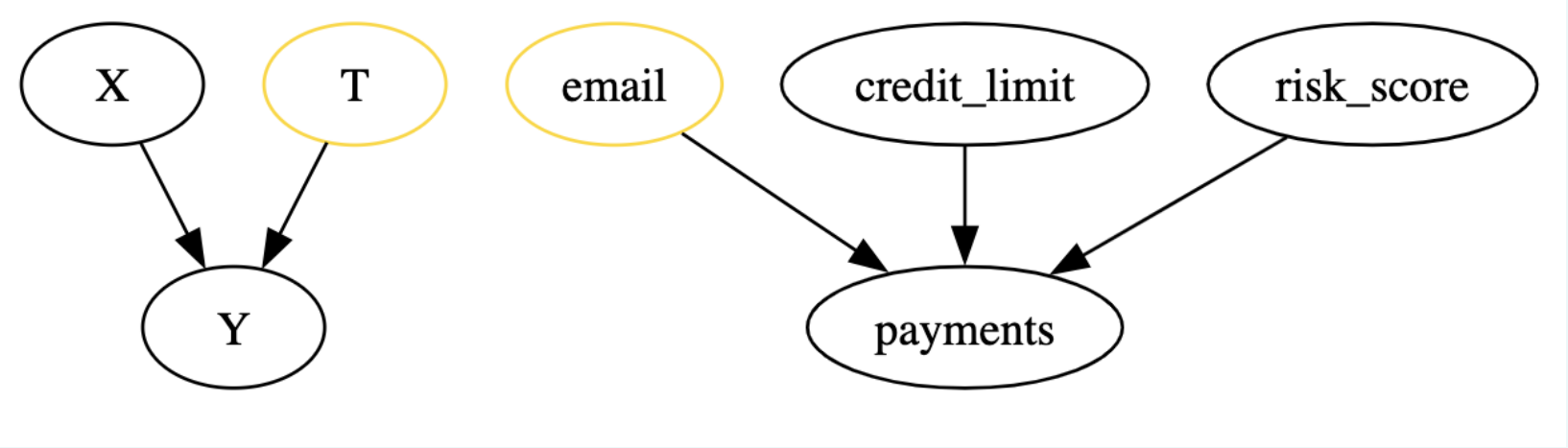
- 实践中，一个回归足以

payments	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
email	4.430355	2.129899	2.08	0.038	.2548181	8.605893
credit_limit	.1510686	.0080216	18.83	0.000	.1353427	.1667946
risk_score	-8.051563	38.42376	-0.21	0.834	-83.37899	67.27586
_cons	490.8653	9.714986	50.53	0.000	471.8196	509.9109

- 所有结果的代码和数据见：https://wenzhe-huang.github.io/python-causality-handbook-zh/07-Beyond-Confounders.html

有效的控制变量

- 任何时候如果我们有一个对结果有良好预测性的控制变量，即便它不是混杂因素，将其纳入模型都是明智之举。
- 这有助于降低我们处理效应估计的方差。



基本有害的控制变量

- 考察新药实验对病人住院天数的影响
- 两家医院：一家医院的处理是向90%的病人提供新药治疗，而10%接受安慰剂；另一家医院则随机向10%病人提供新药，90%为安慰剂
- 还有一个信息/变量：第一家医院接收的病人通常更为严重

	hospital	treatm~t	severity	days
1.	1	1	29.68662	82
2.	1	1	20.05034	57
3.	1	1	20.3024	49
4.	0	0	10.60312	44
5.	0	0	8.332793	15

基本有害的控制变量

- 简单的一阶差分估计量——一元回归?

$$Days = \beta_0 + \beta_1 treatment + \epsilon$$

days	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
treatment	**14.15333**	**3.366811**	**4.20**	**0.000**	**7.450527**	**20.85614**
_cons	**33.26667**	**2.661698**	**12.50**	**0.000**	**27.96763**	**38.5657**

- 新药使得病人住院天数增加
- 注意：两家不同的医院在进行两项随机实验
- 混淆因子：病情严重程度

基本有害的控制变量

- 为了解决这个问题：第一种方法分别对两家医院回归

$$Days = \beta_0 + \beta_1 treatment + \epsilon$$

days	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
treatment	-11.40741	10.92119	-1.04	0.306	-33.81583	11.00102
_cons	30.40741	2.868044	10.60	0.000	24.52267	36.29215

days	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
treatment	-10.39583	6.954515	-1.49	0.141	-24.37146	3.579788
_cons	59	6.746871	8.74	0.000	45.44166	72.55834

基本有害的控制变量

- 上述分组回归减少样本量（不显著）：第二种方法纳入控制变量

$$Days = \beta_0 + \beta_1 treatment + \text{severity} + \epsilon$$

days	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
treatment	-7.591173	2.269234	-3.35	0.001	-12.1098	-3.07255
severity	2.274068	.1537267	14.79	0.000	1.96796	2.580177
_cons	11.66406	2.000134	5.83	0.000	7.681285	15.64684

基本有害的控制变量

- 上述分组回归减少样本量（不显著）：第二种方法纳入控制变量

$$Days = \beta_0 + \beta_1 treatment + \text{severity} + \text{hospital} + \epsilon$$

- 我们思考要不要加入hospital，因为医院类型决定了处理
- 但是，因为我们已经控制了病情程度，因此，医院类型就与住院天数无关了，hospital就不是混淆因子
- 控制hospital可以降低方差，所以应该会有用吧？！
- 注意，降低方差是结果的方差，而不是处理的方差
- 不过，还是想控制它，找了这么久数据，是在忍不住!

基本有害的控制变量

- 上述分组回归减少样本量（不显著）：第二种方法纳入控制变量

$$Days = \beta_0 + \beta_1 treatment + \text{severity} + \text{hospital} + \epsilon$$

days	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
treatment	-5.09447	3.492034	-1.46	0.149	-12.04946	1.860519
severity	2.38653	.1947953	12.25	0.000	1.998561	2.774498
hospital	-4.153548	4.413196	-0.94	0.350	-12.94319	4.636093
_cons	11.01108	2.11845	5.20	0.000	6.791826	15.23034

基本有害的控制变量

基本有害的控制变量

- 控制了病情后，再控制医院类型反而会增加ATE的方差

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$$

$$\mathrm{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

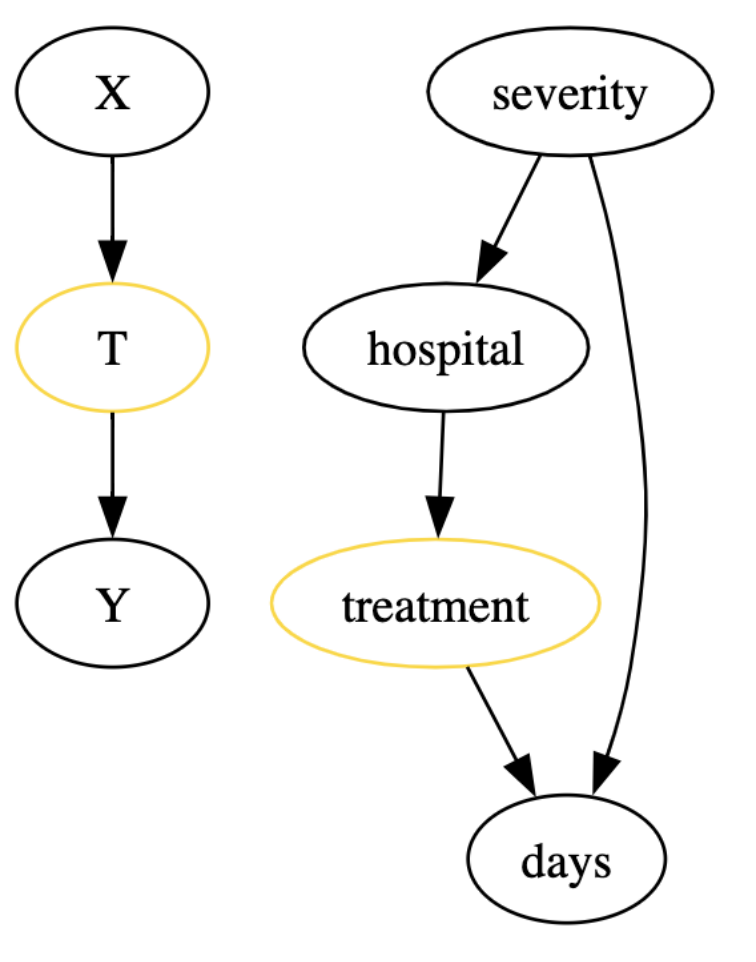
- 上述公式显示，回归系数的标准误与变量X的方差成反比
- 举个极端例子，假设你想评估某种药物的效果，于是对 10000 人进行测试，但其中仅 1 人接受了处理
- 换言之，我们需要处理变量存在大量变异，才能更容易发现其影响。

基本有害的控制变量

Variable	Obs	Mean	Std. dev.	Min	Max
hospital	80	.6375	.4837551	0	1
treatment	80	.625	.4871774	0	1
severity	80	15.47575	7.191461	-4.030356	31.06742
days	80	42.1125	16.04345	0	82
res_treatm~t	80	3.06e-10	.2413655	-.9517725	.9255486
res_days	80	-2.12e-08	7.450029	-22.03868	17.23881

res_days	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
res_treatment	-5.094469	3.446974	-1.48	0.143	-11.95687	1.767928
_cons	-1.97e-08	.8267643	-0.00	1.000	-1.645961	1.645961

基本有害的控制变量



有害的控制变量

- 回到催款信息的例子：邮件是随机发的
- 还有一个虚拟变量：opened、agreement

	payments	email	opened	agreem~t	credit~t	risk_s~e
1.	740	1	1	0	2348.495	.666752
2.	580	1	1	1	334.112	.2073951
3.	600	1	1	1	1360.661	.5504789
4.	770	0	0	0	1531.829	.5604882
5.	660	0	0	0	979.8557	.4551403

有害的控制变量

$$\text{Payments} = \beta_0 + \beta_1 \text{Email} + \text{credit} + \text{risk} + \varepsilon_i$$

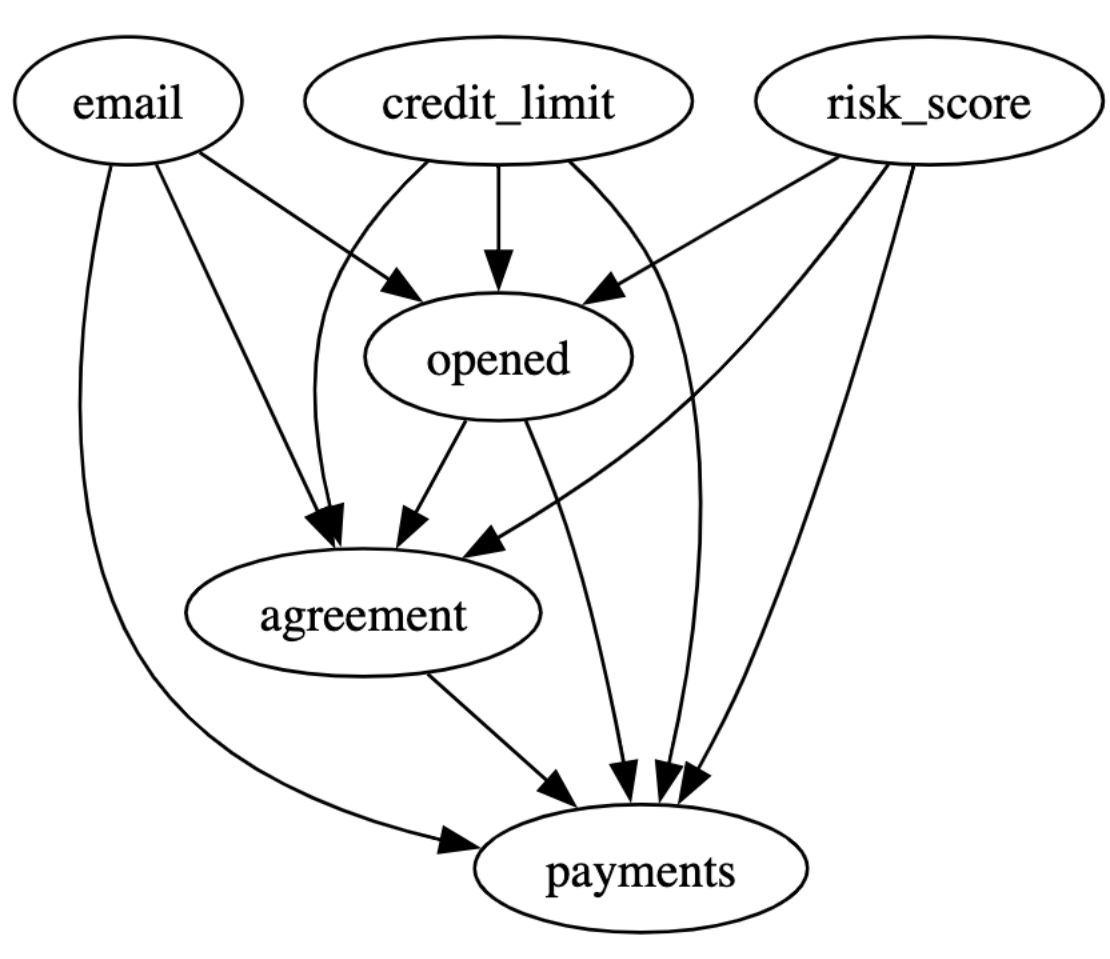
payments	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
email	4.430355	2.129899	2.08	0.038	.2548181	8.605893
credit_limit	.1510686	.0080216	18.83	0.000	.1353427	.1667946
risk_score	-8.051563	38.42376	-0.21	0.834	-83.37899	67.27586
_cons	490.8653	9.714986	50.53	0.000	471.8196	509.9109

有害的控制变量

$$\text{Payments} = \beta_0 + \beta_1 \text{Email} + \text{credit} + \text{risk} + \text{opened} + \text{agreement} + \varepsilon_i$$

payments	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
email	-1.609485	2.723665	-0.59	0.555	-6.949065	3.730095
credit_limit	.1506935	.0080119	18.81	0.000	.1349867	.1664003
risk_score	-2.092863	38.37499	-0.05	0.957	-77.3247	73.13897
opened	3.980847	3.913991	1.02	0.309	-3.692294	11.65399
agreement	11.70934	4.165948	2.81	0.005	3.542252	19.87643
_cons	488.4416	9.715979	50.27	0.000	469.394	507.4892

有害的控制变量



总之

- 包含：混淆因子、结果预测变量
- 排除：处理预测变量、中介变量、共同结果变量

THANK YOU