第7章 模型设定与数据问题

遗漏变量

- 由于某些数据难以获得，遗漏变量现象几乎难以避免。
- 假设真实的模型(true model)为

$$y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon$$

其中，解释变量x_1和x_2与扰动项ε不相关。

- 实际估计的模型(estimated model)为

$$y = \alpha + \beta x_1 + \mu$$

遗漏变量(omitted variable)x_2被归入扰动项$\mu = \gamma x_2 + \varepsilon$。

- 遗漏变量是否一定导致不一致的估计?

- 考虑两种情形：
- （1）遗漏变量x_2与解释变量x_1**不相关**，即$\mathrm{Cov}(x_1, x_2) = 0$。
- 扰动项$\mu = \gamma x_2 + \varepsilon$与解释变量$x_1$不相关，因为

  $$\mathrm{Cov}(x_1, \mu) = \mathrm{Cov}(x_1, \gamma x_2 + \varepsilon) = \mathrm{Cov}(x_1, x_2) + \mathrm{Cov}(x_1, \varepsilon) = 0 + 0 = 0$$

- 虽然存在遗漏变量，但 OLS 依然可一致地估计回归系数。
- 由于遗漏变量x_2被归入扰动项中，可能增大扰动项的方差，影响 OLS 估计的精确度。

- (2) 遗漏变量x_2与解释变量x_1**相关**，即$\mathrm{Cov}(x_1, x_2) \neq 0$。
- 根据大样本理论，OLS 估计不一致，称为“遗漏变量偏差”(omitted variable bias)。

- 这种偏差在实践中较常见，成为某些实证研究的致命伤。
- 比如，研究教育投资回报时，个人能力因无法观测而遗漏，但能力与教育年限正相关。
- 存在遗漏变量本身并不要紧；关键在于，遗漏变量不能与方程的解释变量相关。

- 解决遗漏变量偏差的方法主要有：
- (i) 加入尽可能多的控制变量(control variable)；
- (ii) 随机实验与自然实验；
- (iii) 工具变量法；
- (iv) 使用面板数据；
- 第(i)种方法“加入尽可能多的控制变量”着眼于直接解决遗漏变量问题。
- 首先从理论出发，列出所有可能对被解释变量有影响的变量，然后尽可能去收集数据。
- 如果有些相关变量无法获得，则需从理论上说明，遗漏变量不会与解释变量相关，或相关性很弱。

- **例** 李宏彬等(2012)通过就业调查数据，研究“官二代”大学毕业生的起薪是否高于非官二代。
- 由于可能存在遗漏变量，该文包括了尽可能多的控制变量，比如年龄、性别、城镇户口、父母收入、父母学历、高考成绩、大学成绩、文理科、党员、学生会干部、兼职实习经历、拥有技术等级证书等。

- 解决遗漏变量偏差的第(ii)种方法为随机实验或自然实验。
- 物理学常使用“控制实验”(controlled experiment)来研究x对y的因果关系，即给定影响y的所有其他因素，单独让x变化，然后观察y如何变化。
- 控制实验的方法在其他学科未必可行
- 比如，医学对新药x疗效的实验。
- 由于参加实验者的体质与生活方式不同，不可能完全控制所有其他因素，故无法进行严格的控制实验。

- 统计学之父费舍尔(Ronald Fischer)提出了随机(控制)实验(randomized controlled experiment)的概念。
- 考虑回归模型：

$$y = \alpha + \beta x + \varepsilon$$

x完全随机地决定(比如，通过抛硬币或电脑随机数)，故x与任何其他变量都相互独立。

- 由于x独立于ε，故$\mathrm{Cov}(x, \varepsilon) = 0$，无论遗漏多少解释变量，OLS都一致。
- 如果x是取值为0或1的虚拟变量，可将样本数据分为两组。
- 称“$x = 1$”的组为“实验组”或“处理组”(treatment group)，比如医学实验中吃药的那组。
- 称“$x = 0$”的组为“控制组”(control group)或“对照组”，比如医学实验中不吃药的那组。

- 随机实验的核心思想是，将实验人群(或个体)随机分为两组，即实验组与控制组，则这两组除了x不同外，在所有其他方面都没有系统差别。
- **例** 农学中将地块随机分成三组(很难找到土壤条件完全一样的地块)，分别给予不同的施肥量，然后考察施肥的效果。
- **例** 班级规模是否影响学习成绩？由于遗漏变量的存在，观测数据很难回答此问题。比如，规模较小的班级可能位于好学区，师资好，家庭也富有。
- 美国田纳西州进行了为期四年的随机实验(Project STAR，即Student-Teacher Achievement Ratio)。

- 将幼儿园至小学三年级的学生随机分为三组。
- 第一组为普通班，每班 22-25 名学生；
- 第二组为小班，每班 13-17 名学生；
- 第三组也为小班，但配备一名教学助理(teacher’s aide)。
- 教师也随机分到这三类班级。
- 结果发现，班级规模对学习成绩的影响在统计上显著，但在经济上并不显著(即此效应本身比较小，普通班与小班的成绩差距类似于男生与女生的成绩差距)。
- 随机实验说服力强，但通常成本高。

- 另一实验方法为“自然实验”(natural experiment)或“准实验”(quasi experiment)，即由于某些并非为了实验目的而发生的外部突发事件，使得当事人**仿佛**被随机分在实验组或控制组。
- **例** 最低工资对就业的影响。提高法定最低工资(minimum wage)
- 在多大程度上会影响对低技能工人的需求?
- Card and Krueger (1994)考虑一个自然实验。
- 1992 年，新泽西州通过法律将最低工资从每小时$4.25 提高到$5.05，但相邻的宾夕法尼亚州最低工资保持不变。这两个州的雇主仿佛被随机分配到实验组(新泽西州)与控制组(宾夕法尼亚州)。
- Card and Kruger (1994)收集两个州的快餐店在实施新法前后雇佣人数的数据；结果发现，提高最低工资对低技能工人的就业几乎没有影响。

- **例**(一个失败的例子) 京杭大运河流经省份的人均 GDP 平均而言高于其他省份。这是否可以归功于京杭大运河对区域经济增长的促进作用?
- 隋炀帝确定京杭大运河的位置时，他是在地图上随机画了一条线吗?
- 由于影响被解释变量的因素往往很多，而局限于数据的可得性(availability)，故在任何实证研究中几乎总存在遗漏变量。
- 一篇专业水准的实证论文几乎总要说明，如何在存在遗漏变量的情况下避免遗漏变量偏差。

无关变量

- 与遗漏变量相反的情形是，在回归方程中加入了与被解释变量无关的变量。
- 假设真实的模型为

$$y = \alpha + \beta x_1 + \varepsilon$$

其中，$Cov(x_1, \varepsilon) = 0$

- 实际估计的模型为

$$y = \alpha + \beta x_1 + \gamma x_2 + (\varepsilon - \gamma x_2)$$

- 由于真实参数$\gamma = 0$(x_2对y无影响)，故可将模型写为

$$y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon$$

- 由于x_2与y无关，根据“无关变量”的定义，x_2也与y的扰动项ε无关，即$Cov(x_2, \varepsilon) = 0$
- 扰动项ε与所有解释变量均无关，故 OLS 一致，即$\operatorname{plim}_{n\to\infty} \hat{\beta} = \beta$，$\operatorname{plim}_{n\to\infty} \hat{\gamma} = \gamma = 0$
- 引入无关变量后，受到无关变量的干扰，估计量$\hat{\beta}$ 的方差一般会增大。
- 对于解释变量的选择最好遵循经济理论的指导。

建模策略：“由小到大”还是“由大到小”

- “由小到大”(specific to general)的建模方式首先从最简单的小模型开始，逐渐增加解释变量。
- 比如，先将被解释变量y对关键解释变量x回归，然后再加入其他控制变量z。
- 但小模型很可能存在遗漏变量偏差，系数估计不一致，t 检验、F 检验都失效，很难确定如何取舍变量。

- “由大到小”(general to specific)的建模方式从尽可能大的模型开始，收集所有可能的解释变量，再逐步剔除不显著的解释变量(可依次剔除最不显著，即 p值最大的变量)。
- 虽然冒着包含无关变量的危险，但危害性没有遗漏变量严重。
- 但在实际操作上，很难找到与被解释变量相关的所有解释变量。
- 实证研究中，常采用以上两种策略的折中方案。比如，总是包括核心解释变量，而变换控制变量，来考察回归结果的稳健性。

解释变量个数的选择

- 好的经济理论使用简洁的模型来很好地描述复杂的经济现实。
- 但这两个目标常常矛盾。加入过多的解释变量可提高模型的解释力(比如增大拟合优度R^2)，但也牺牲了模型的简洁性(parsimony)。
- 需在模型的解释力与简洁性之间找到最佳平衡。
- 在时间序列模型里，常需选择解释变量滞后的期数(比如，确定自回归模型的阶数)。
- 更一般地，要确定解释变量的个数。

- 可供选择的权衡标准如下：
- （1）校正可决系数$\bar{R}^2$：选择解释变量的个数K以最大化$\bar{R}^2$。
- （2）“赤池信息准则”(Akaike Information Criterion，AIC)：选择解释变量的个数K，使得目标函数最小化：

$$\min_K AIC \equiv \underbrace{ln(\frac{SSR}{n})}_{\text{奖励}} + \underbrace{\frac{2}{n}K}_{\text{惩罚}}$$

- 其中，SSR为残差平方和$\sum_{i=1}^{n} {e_i}^2$
- 第一项为对模型拟合度的奖励(减少残差平方和 SSR)。
- 第二项为对解释变量过多的惩罚(为解释变量个数 K 的增函数)。
- 当K 上升时，第一项下降而第二项上升

- (3)“贝叶斯信息准则”(Bayesian Information Criterion，BIC)或“施瓦茨信息准则”(Schwarz Information Criterion，SIC 或 SBIC)：选择解释变量的个数 K，使得目标函数最小化：

$$\min_{K} BIC \equiv \underbrace{ln(\frac{SSR}{n})}_{\text{奖励}} + \underbrace{\frac{lnn}{n}K}_{\text{惩罚}}$$

- BIC 准则与 AIC 准则仅第二项有差别。
- 一般来说，$\ln n > 2$，故 BIC 准则对于解释变量过多的惩罚比 AIC 准则更为严厉。
- BIC 准则更强调模型的简洁性。在时间序列模型中，常用信息准则确定滞后阶数。

- 考虑p阶自回归模型，滞后阶数 p可通过信息准则来确定。

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t$$

- 根据 BIC 准则计算的$\hat{p}$是真实滞后阶数 p的一致估计量
- 但根据 AIC 计算的$\hat{p}$却不一致，即使在大样本中也可能高估 p。
- 由于现实样本通常有限，而 BIC 准则可能导致模型过小(对解释变量过多的惩罚太严厉)，故 AIC 准则依然很常用。

- (4) “由大到小的序贯 t 规则”(general-to-specific sequential t rule)。
- 这种方法常用于时间序列模型，比如 AR(p)。
- 首先，设最大滞后期p_{max}，令$\hat{p} = p_{max}$进行估计，并对最后一阶系数的显著性进行t检验。
- 如果接受该系数为 0，则令$\hat{p} = p_{max} - 1$，重新进行估计，再对(新的)最后一阶系数的显著性进行t检验，如果显著，则停止；否则，令$\hat{p} = p_{max} - 2$，以此类推。

对函数形式的检验

- 很多经济关系是非线性的。多元线性回归可看作是非线性关系的一阶线性近似。
- 如果存在非线性项，但被遗漏，则会导致遗漏变量偏差，这是模型设定误差的一种形式。
- 假设真实模型为

$$y = \alpha + \beta x + (\gamma x^2 + \varepsilon)$$

- 其中，$Cov(x, \varepsilon) = 0$，而平方项x^2被遗漏。
- 解释变量与扰动项相关：

$$Cov(x, \gamma x^2 + \varepsilon) = \gamma Cov(x, x^2) + Cov(x, \varepsilon) = \gamma Cov(x, x^2) \neq 0$$

- 遗漏高次项会导致遗漏变量偏差。

- “Ramsey’s RESET检验”(Regression Equation Specification Error Test)(Ramsey, 1969)的基本思想：

如果怀疑遗漏非线性项，就把非线性项引入方程，检验其系数是否显著。

- 假设线性回归模型为

$$y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon$$

- 记此回归的拟合值为

$$\hat{y} = \hat{\alpha} + \hat{\beta} x_1 + \hat{\gamma} x_2$$

- 既然$\hat{y}$是解释变量的线性组合，$\hat{y}^2$就包含了解释变量二次项(含平方项与交叉项)的信息，$\hat{y}^3$就包含了解释变量三次项的信息，以此类推。

- 考虑辅助回归

$$y = \alpha + \beta x_1 + \gamma x_2 + \delta_2 \hat{y}^2 + \delta_3 \hat{y}^3 + \delta_4 \hat{y}^4 + \varepsilon$$

- 对$H_0: \delta_2 = \delta_3 = \delta_4 = 0$作F检验。
- 如果拒绝H_0，说明模型中应有高次项；
- 如果接受H_0，则可使用线性模型。
- RESET 检验的缺点是，在拒绝H_0的情况下，并不提供具体遗漏哪些高次项的信息。
- 也可直接将解释变量x_1和x_2的高次项放入辅助回归中，比如

$$y = \alpha + \beta x_1 + \gamma x_2 + \delta_2 {x_1}^2 + \delta_3 {x_2}^2 + \delta_4 x_1 x_2 + \varepsilon$$

然后检验$H_0: \delta_2 = \delta_3 = \delta_4 = 0$

- 关于如何确定回归方程的函数形式，最好从经济理论出发。
- 在缺乏理论指导的情况下，可先从线性模型出发，然后进行RESET 检验，看是否应加入非线性项。

多重共线性

- 如果在解释变量中，某一解释变量可由其他解释变量线性表出，则存在“严格多重共线性”(strict multicollinearity)。
- 此时数据矩阵$\boldsymbol{X}$不满列秩，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$不存在，无法定义OLS估计量 $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
- 比如，解释变量x_2正好是解释变量x_3的两倍，则无法区分量x_2与x_3对被解释变量y的影响。
- 严格多重共线性是对数据的最低要求，现实中较少出现。
- stata会自动处理。

- 在实践中更为常见的为近似(非严格)的多重共线性，简称“多重共线性”(multicollinearity)或“共线性”。
- 多重共线性的主要表现是，如果将第 k 个解释变量x_k对其余解释变量$\{x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_k\}$进行回归，所得可决系数（记为$R_k^2$）较高。
- 在多重共线性的情况下，OLS 仍是 BLUE，因为高斯-马尔可夫定理并未排除多重共线性的情形。
- 但 BLUE 只保证 OLS 估计量在所有线性无偏估计量中相对而言方差最小，并不意味着 OLS 估计量的方差在绝对意义上小。

- 多重共线性的后果?
- 如果存在严格多重共线性，则矩阵$(\boldsymbol{X}'\boldsymbol{X})$不可逆。
- 在多重共线性的情况下，矩阵$(\boldsymbol{X}'\boldsymbol{X})$变得“几乎不可逆”，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$变得很“大”，致使方差$Var(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$增大，系数估计变得不准确。
- $\boldsymbol{X}$中元素轻微变化就会引起$(\boldsymbol{X}'\boldsymbol{X})^{-1}$很大变化，导致OLS估计值$\hat{\boldsymbol{\beta}}$发生很大变化。
- 多重共线性的通常症状是，虽然整个回归方程的R^2较大，F检验也很显著，但单个系数的t检验却不显著。
- 另一症状是，增减解释变量使得系数估计值发生较大变化(比如，加入的解释变量与已有解释变量构成多重共线性)。

- 如何度量?
- 如果两个(或多个)解释变量之间高度相关，则不易区分它们对被解释变量的单独影响力。
- 在严格多重共线性的极端情况下，一个变量刚好是其他变量的倍数，则完全无法区分。
- R_k^2越高，解释变量x_k与其他解释变量的共线性越严重，则x_k的系数估计量$\hat{\beta}_k$的方差越大。可以证明:

$$Var(\hat{\beta}_k|\boldsymbol{X}) = \frac{\sigma^2}{(1-R_k^2)S_k}$$

- 其中，$\sigma^2 \equiv Var(\varepsilon)$为扰动项的方差。

$S_k \equiv \sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2$为$x_k$的离差平方和，反映$x_k$的变动幅度。

- 如果x_k变动很少，很难准确地估计x_k对y的作用。
- 在极端情况下，x_k完全不变（为常数），$S_k=0$，则完全无法估计$\hat{\beta}_k$（x_k与常数项构成严格多重共线性）。
- 方差$Var(\hat{\beta}_k|\boldsymbol{X})$与$(1-R_k^2)$成反比。
- 定义解释变量x_k的“方差膨胀因子”(Variance Inflation Factor，VIF)为

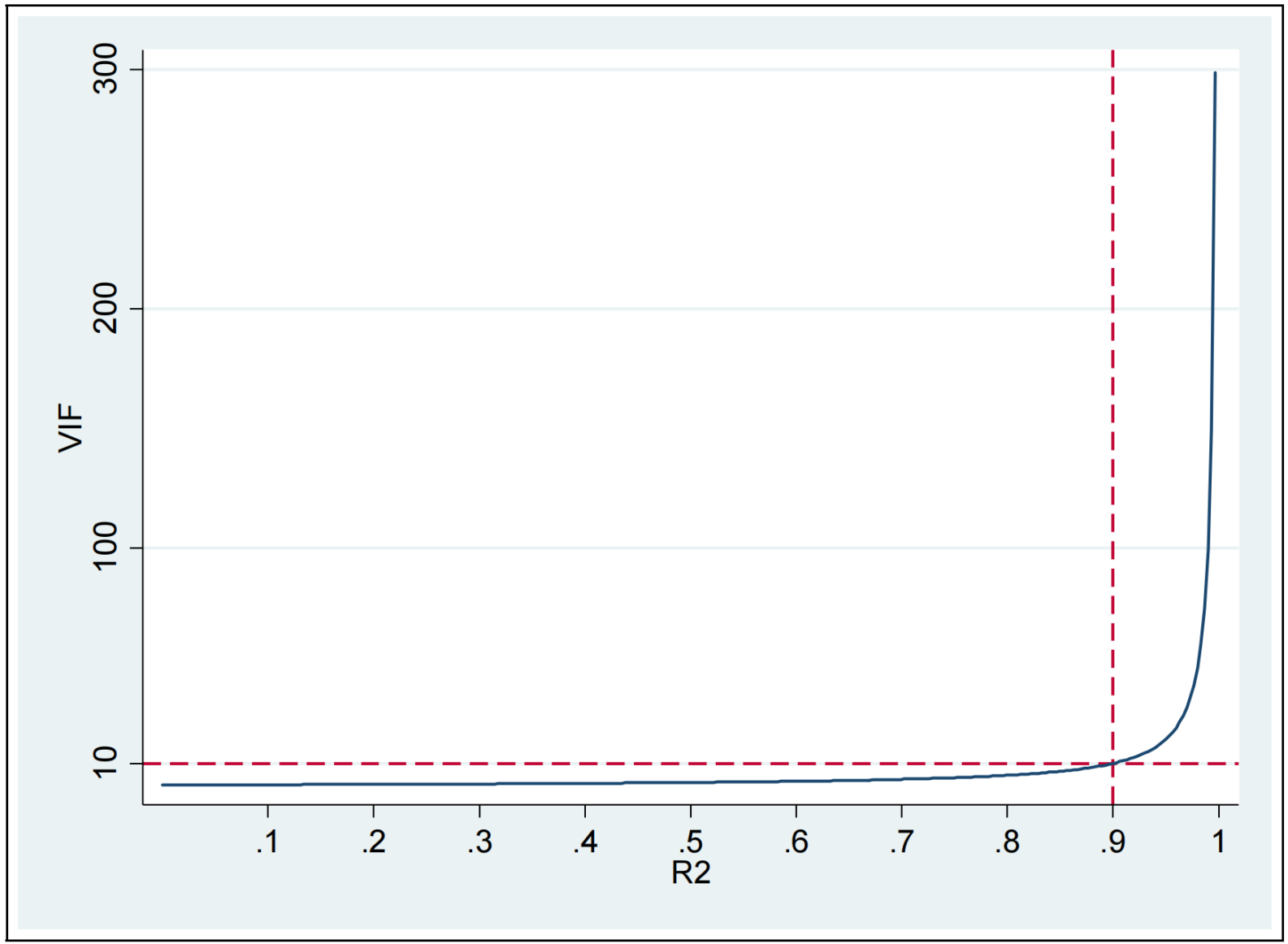
$$VIF_k \equiv \frac{1}{1-R_k^2}$$

可将方差写为

$$Var(\hat{\beta}_k|\boldsymbol{X}) = VIF_k \cdot \frac{\sigma^2}{S_k}$$

- 方差膨胀因子VIF_k越大，x_k的多重共线性问题越严重，其方差$Var(\hat{\beta}_k|\boldsymbol{X})$将变得更大。

- 对于K个解释变量$\{x_1,\cdots,x_k\}$,可计算相应的方差膨胀因子$\{VIF_1,\cdots,VIF_k\}$。
- 判断是否存在多重共线性的经验规则是，$\{VIF_1,\cdots,VIF_k\}$的最大值不应超过10。
- 求解$10=\frac{1}{1-R_k^2}$可知，相应的R_k^2不应超过0.9。
- 解释变量x_k的多重共线性越严重，R_k^2越接近于1，方差膨胀因子VIF_k将急剧上升。


VIF
300
200
100
10
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
R2

- 如发现多重共线性，可采取以下处理方法。
- (1) 如不关心具体的回归系数，只关心整个方程的预测能力，可不必理会多重共线性(假设整个方程显著)。多重共线性的主要后果是使得对单个变量的贡献估计不准，但所有变量的整体效应仍可较准确地估计。
- (2) 如关心具体的回归系数，但多重共线性并不影响所关心变量的显著性，也可不必理会。

在方差膨胀的情况下，系数依然显著；如没有多重共线性，只会更显著。

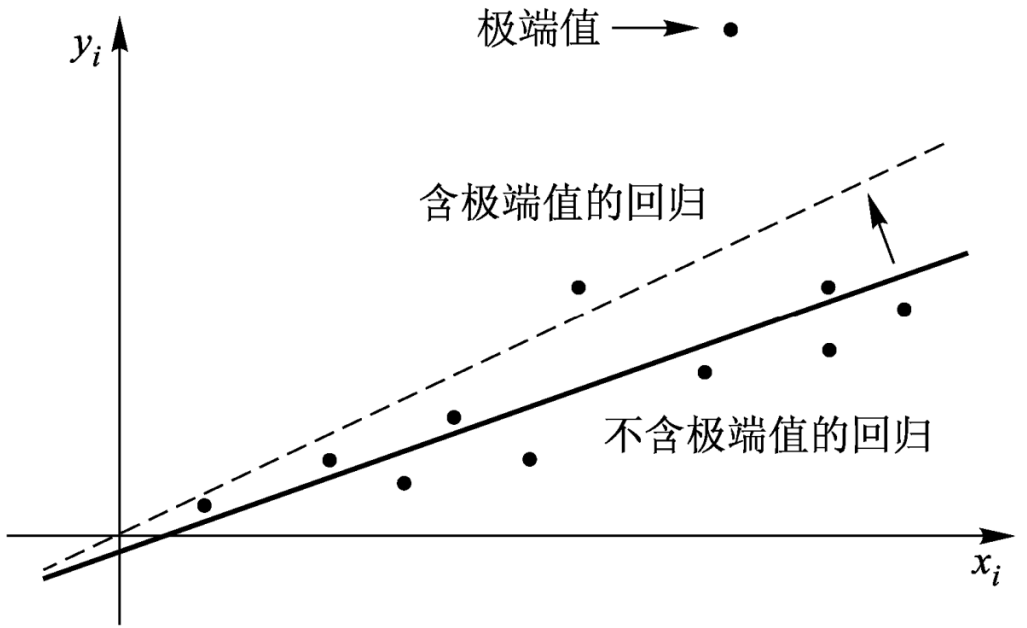
- (3) 如多重共线性影响所关心变量的显著性，应设法进行处理。

比如，增大样本容量，剔除导致严重共线性的变量，将变量标准化，或对模型设定进行修改。

- 解释变量之间的相关性普遍存在，在一定程度上也是允许的。处理多重共线性的最常见方法是“无为而治”(do nothing)。

极端数据

- 如果样本数据中的少数观测值离大多数观测值很远，可能对OLS的回归系数产生很大影响。
- 这些数据称为“极端观测值”(outliers) 或“高影响力数据”(influential data)。



- 以数据集 nerlove.dta 为例。首先，进行回归；然后人为构造极端值，再回归，并比较结果。

```
. use nerlove.dta,clear
. reg lntc lnq lnpl lnpk lnpf
```

Source	SS	df	MS		
Model	269.524728	4	67.3811819		
Residual	21.5420958	140	.153872113		
Total	291.066823	144	2.02129738		

Number of obs	=	145			
F(4, 140)	=	437.90			
Prob > F	=	0.0000			
R-squared	=	0.9260			
Adj R-squared	=	0.9239			
Root MSE	=	.39227			

lntc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnq	.7209135	.0174337	41.35	0.000	.6864462	.7553808
lnpl	.4559645	.299802	1.52	0.131	-.1367602	1.048689
lnpk	-.2151476	.3398295	-0.63	0.528	-.8870089	.4567136
lnpf	.4258137	.1003218	4.24	0.000	.2274721	.6241554
_cons	-3.566513	1.779383	-2.00	0.047	-7.084448	-.0485779

- 将第一个观测值的产量对数(lnq)乘以 100，再次进行回归。

```
. replace lnq=lnq*100 if _n==1
(1 real change made)
```

其中，“_n”表示第n个观测值，故“_n==1”表示第 1 个观测值。

```
. reg lntc lnq lnpl lnpk lnpf
```

Source	SS	df	MS		
Model	7.4424142	4	1.86060355	Number of obs =	145
Residual	283.624409	140	2.02588864	F(4, 140) =	0.92
				Prob > F =	0.4551
Total	291.066823	144	2.02129738	R-squared =	0.0256
				Adj R-squared =	-0.0023
				Root MSE =	1.4233

lntc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnq	.0156026	.0218326	0.71	0.476	-.0275616	.0587668
lnpl	1.214014	1.093264	1.11	0.269	-.9474289	3.375456
lnpk	-1.096614	1.233079	-0.89	0.375	-3.534477	1.341248
lnpf	-.2427032	.3641193	-0.67	0.506	-.9625867	.4771803
_cons	7.230075	6.388544	1.13	0.260	-5.400419	19.86057

将此人造极端值去掉，再对比回归结果。

```
. reg lntc lnq lnpl lnpk lnpf if _n>1
```

Source	SS	df	MS	Number of obs =	144
				F(4, 139) =	406.10
Model	251.560166	4	62.8900416	Prob > F =	0.0000
Residual	21.5261194	139	.154864168	R-squared =	0.9212
				Adj R-squared =	0.9189
Total	273.086286	143	1.90969431	Root MSE =	.39353

lntc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnq	.7225462	.0182135	39.67	0.000	.6865348	.7585576
lnpl	.4445846	.3028466	1.47	0.144	-.1541969	1.043366
lnpk	-.2219834	.3415869	-0.65	0.517	-.8973614	.4533947
lnpf	.4311295	.1019964	4.23	0.000	.2294645	.6327944
_cons	-3.55227	1.78566	-1.99	0.049	-7.082837	-.0217022

去掉极端值后的回归结果又“恢复正常”。

- 如何发现极端数据?
- 对于一元回归，可通过画 (x, y)的散点图来考察，但对多元回归行不通。
- 某个观测值的影响力可通过去掉此观测值对回归系数的影响来衡量。
- 记$\widehat{\boldsymbol{\beta}}$为全样本的OLS估计值，$\widehat{\boldsymbol{\beta}}^{(i)}$为去掉第$i$个观测值后的OLS估计值。
- 关心$(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(i)})$的变化幅度及如何决定。

- 定义第i个观测值对回归系数的“影响力”或“杠杆作用”(leverage)为

$$lev_i = \boldsymbol{x}_i^{'}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$$

其中，$\boldsymbol{x}_i \equiv (1\ x_{i2}\ \cdots\ x_{iK})'$为包含个体i的全部解释变量，而$\boldsymbol{X} = (\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n)'$数据矩阵。

- lev_i与$(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(i)})$存在如下关系：

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(i)} = (\frac{1}{1 - lev_i})(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i e_i$$

- lev_i越大，则$(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(i)})$的变化越大。

- 可以证明，所有观测数据的影响力lev_i满足：
- (i)$0 \leq lev_i \leq 1, (i = 1, \cdots, n)$;
- (ii)$\sum_{i=1}^{n} lev_i = K$(解释变量个数)。
- 因此，影响力lev_i的平均值为(K/n)。
- 如果某些数据的lev_i比平均值(K/n)高很多，则可能对回归系数有很大影响。

- 如何处理极端数据?
- 首先，应检查是否因数据输入有误导致极端观测值。
- 其次，对极端观测值的个体进行背景调查，看是否由与研究课题无关的特殊现象所致，必要时可删除极端数据。
- 最后，比较稳健的做法是同时汇报“全样本”(full sample)与删除极端数据后的“子样本”(subsample)的回归结果，让读者自己做判断。

地方公共债务增长的制度基础*

——兼顾财政和金融的视角

毛　捷　刘　潘　吕冰洋

摘　要：财政金融体制和地方公共债务密不可分。理论研究发现，在隐性金融分权创造的弱融资约束环境下，税收分成对地方公共债务依存度和负债率均有正向刺激作用。背后的机制是，税收分成比例的提高增强了地方政府扩大债务规模的能力和动机，而隐性金融分权导致地方政府的融资约束变弱，使得扩大债务规模的动机更易实现。实证结果支持理论发现，税收分成比例与地方公共债务依存度和负债率均显著正相关，税收分成比例与金融分权度的交互项系数显著为正，且税收分成与地方公共债务规模之间的关系受宏观税率和增长效果等影响。财政金融体制的激励作用是地方公共债务持续增长的制度原因，这为科学管控债务规模、防范化解债务风险提供了思路。

关键词：税收分成　财政分权　金融分权　地方公共债务

三、实证方法和数据说明

（一）计量模型设定

依据前述理论分析，并参考相关文献，②我们使用适用于面板数据的固定效应模型估计方法，实证检验分税（即税收分成）对地方公共债务增长的影响。计量方程具体如下：

$$Debt_{it}=\alpha+\beta\cdot Taxshare_{it}+X'_{it}\cdot\gamma+\mu_i+\upsilon_t+\varepsilon_{it} \quad (26)$$

其中，i 表示地区（地级市），t 表示时间（年份），Debt 代表地方公共债务变量，Taxshare 为税收分成变量，X 是一个由控制变量组成的向量（即影响地方公共债务的其他因素）；α 是常数项，μ_i 为地区固定效应，υ_t 是时间固定效应，ε_{it} 为误差项；β 和 γ 分别是税收分成和其他控制变量的回归系数。

在式（26）里，我们主要关注系数 β 的符号和显著程度：如果 β 显著为正，表明税收分成对地方公共债务增长具有正向影响，符合前述理论分析结论；如果 β 不显著或显著为负，表明税收分成对地方公共债务增长无显著影响或有显著的负向影响，不符合本文的理论预期。

（二）稳健性检验

基准回归未考虑异常值（outlier）、官员变动以及财政体制改革等因素，这些因素是否会干扰基准实证结论，需要进一步检验。①

1. 考虑异常值影响

基准回归中虽然排除了直辖市可能产生的影响，但不能排除某些省份的地级市②的数据存在异常值进而影响估计结果的可能性。为此，我们在回归中逐一剔除某一省份进行回归估计，通过观察税收分成变量的回归系数是否发生显著变化，排除基准回归结果受异常值影响的可能性，结果显示：逐一剔除某一省份的估计结果与基准回归结果基本相同（回归系数未超出基准回归系数的置信区间），表明基准回归结果不受异常值影响，即税收分成对地方公共债务的正向刺激作用在全国各地区普遍存在。

① 限于篇幅，稳健性检验和延伸分析中未报告实证图表，如有需要，可向作者索要相关图表。

② 考虑到直辖市的特殊性（经济社会各方面与地级市存在明显差异），样本中不包括直辖市；之所以不逐一剔除地级市而是逐一剔除某一省份进而排除异常值的影响，主要是因为政府间税收收入弹性分成主要体现为省际差异，同一省份下不同地市之间的税收收入分成率差异不大。

激励与平衡:中国经济增长的财政动因

吕冰洋　李　钊　马光荣*

内容提要　中国以分税和转移支付为代表的财税体制是近30年来经济增长的重要动力,分税激发了地方发展经济的积极性,而中央政府将分税用于转移支付又会影响受助地区的经济增长。因此有必要将分税和转移支付纳入统一的分析框架,研究其对经济增长的综合效果。理论分析说明,分税的逻辑是寓平衡于激励,转移支付的逻辑是寓激励于平衡,分税与转移支付对经济增长均同时存在激励效应和收入效应,两种效应的总和决定了总体增长效应。本文基于2000-2016年面板数据,证实了市县政府财政收入分成率与转移支付均等化力度对地区经济增长的促进作用。本文首次考察了分税与转移支付对经济增长的综合作用效果,能够为解释中国经济增长之谜提供新的视角与经验证据。

五　计量模型与数据说明

（一）计量模型

为检验财政收入分成率与转移支付均等化力度对地区经济增长的影响，本文以GDP增长率作为被解释变量构造计量模型。市级样本的基准回归模型为：

$$gdpgrowth_{pit} = \alpha + \beta_1 taxshare_{pt} + \beta_2 equalization_{pt} + \gamma Z_{pit} + \mu_i + \varphi_t + \varepsilon_{pit} \quad (11)$$

$$gdpgrowth_{pit} = \alpha + \beta_1 taxshare_{pt} + \beta_2 equalization_{pt} + \beta_3 dev_{pit} + \beta_4 dev_{pit} \times equalization_{pt} + \gamma Z_{pit} + \mu_i + \varphi_t + \varepsilon_{pit} \quad (12)$$

其中，被解释变量 $gdpgrowth_{pit}$ 是 p 省份地级市 i 在 t 年的GDP实际增长率；两个核心解释变量 $taxshare_{pt}$、$equalization_{pt}$ 分别表示地级市 i 所在的 p 省份在 t 年的市县政府财政收入分成率和转移支付均等化力度。基准模型中，我们采用市县分得的一般公共预算收入比例衡量分成率，计算公式为：

$$taxshare_{pt} = \frac{\text{省内市县获得的一般公共预算收入}}{\text{该省实现的全部一般公共预算收入}} \quad (13)$$

1. 改变转移支付均等化力度计算方法。本文以泰尔指数和基尼系数重新计算各省财力差距，表 6 的回归结果显示，无论采取哪种计算指标，转移支付均等化力度系数均显著为正，其与相对经济发展水平的交互项系数均显著为负，说明转移支付均等化力度的作用效果是稳健的，不受计算指标的影响。

2. 变换回归样本。计划单列市在行政层级上属于副省级城市，在财政上虽然不一定与所在省的财政完全脱钩，但直接与中央财政结算。为排除计划单列市的影响，我们将大连、青岛、宁波、厦门、深圳 5 个计划单列市从回归样本中剔除，重复基准回归。回归结果显示，市县财政收入分成率和转移支付均等化力度对经济增长仍具有显著正向影响，且地区相对经济发展水平越高，转移支付对经济增长的促进作用越小，与基准回归结果一致①。

虚拟变量

- 如使用“定性数据”(qualitative data)或“分类数据”(categorical data)，需要引入“虚拟变量”，即取值为 0或1 的变量。
- 比如，性别分男女，可定义“男性虚拟变量”

$$D = \begin{cases} 1, & \text{男} \\ 0, & \text{女} \end{cases}$$

- 但不需要再定义“女性虚拟变量”。如果数据分两类，则仅需要一个虚拟变量。
- 对于全球的五大洲，则需要四个虚拟变量，即

$$D_1 = \begin{cases} 1, & \text{亚洲} \\ 0, & \text{其他} \end{cases}, \quad D_2 = \begin{cases} 1, & \text{美洲} \\ 0, & \text{其他} \end{cases},$$

$$D_3 = \begin{cases} 1, & \text{欧洲} \\ 0, & \text{其他} \end{cases}, \quad D_4 = \begin{cases} 1, & \text{非洲} \\ 0, & \text{其他} \end{cases}$$

- 如果$D_1 = D_2 = D_3 = D_4$，则表明为大洋洲。

- 在有常数项的模型中，如果定性指标共分M类，则最多只能在方程中放入$(M-1)$个虚拟变量。
- 如果方程中包含M个变量，会产生严格多重共线性，因为如果将这M个虚拟变量在数据矩阵$\boldsymbol{X}$中对应的列向量相加，就会得到与常数项完全相同的向量，即$(1 \cdots 1)'$（因为M类中必居其一）。
- 称为“虚拟变量陷阱”(dummy variable trap)。
- 如模型中没有常数项，可放入M 个虚拟变量。

- **例** 假设样本中只有四位个体，分别属于三类。其中，前两位个体属于第一类，第三位个体属于第二类，而第四位个体属于第三类。相应地，定义三个虚拟变量D_1，D_2，D_3，则其取值分别为：

$$D_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix},\ D_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix},\ D_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

- 考虑将这三个虚拟变量同时放入回归方程，并包含常数项：

$$y = \alpha + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon$$

- 这三个虚拟变量之和正好就是常数项，因为

$$D_1 + D_2 + D_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

- 故方程存在严格多重共线性，即虚拟变量陷阱。
- 解决方法之一是去掉一个虚拟变量：

$$y = \alpha + \beta_2 D_2 + \beta_3 D_3 + \varepsilon$$

- 解决方法之二是去掉常数项（不常用）：

$$y = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon$$

- 在模型中引入虚拟变量，会带来什么影响呢?
- 考虑有关中国经济的时间序列模型:

$$y_t = \alpha + \beta x_t + \varepsilon_t \ \ (t = 1950, \cdots, 2000)$$

经济结构可能在 1978 年后有变化，引入虚拟变量:

$$D_t = \begin{cases} 1, & \text{若} t \geq 1978 \\ 0, & \text{其他} \end{cases}$$

根据引入虚拟变量的方式，考虑两种情况。

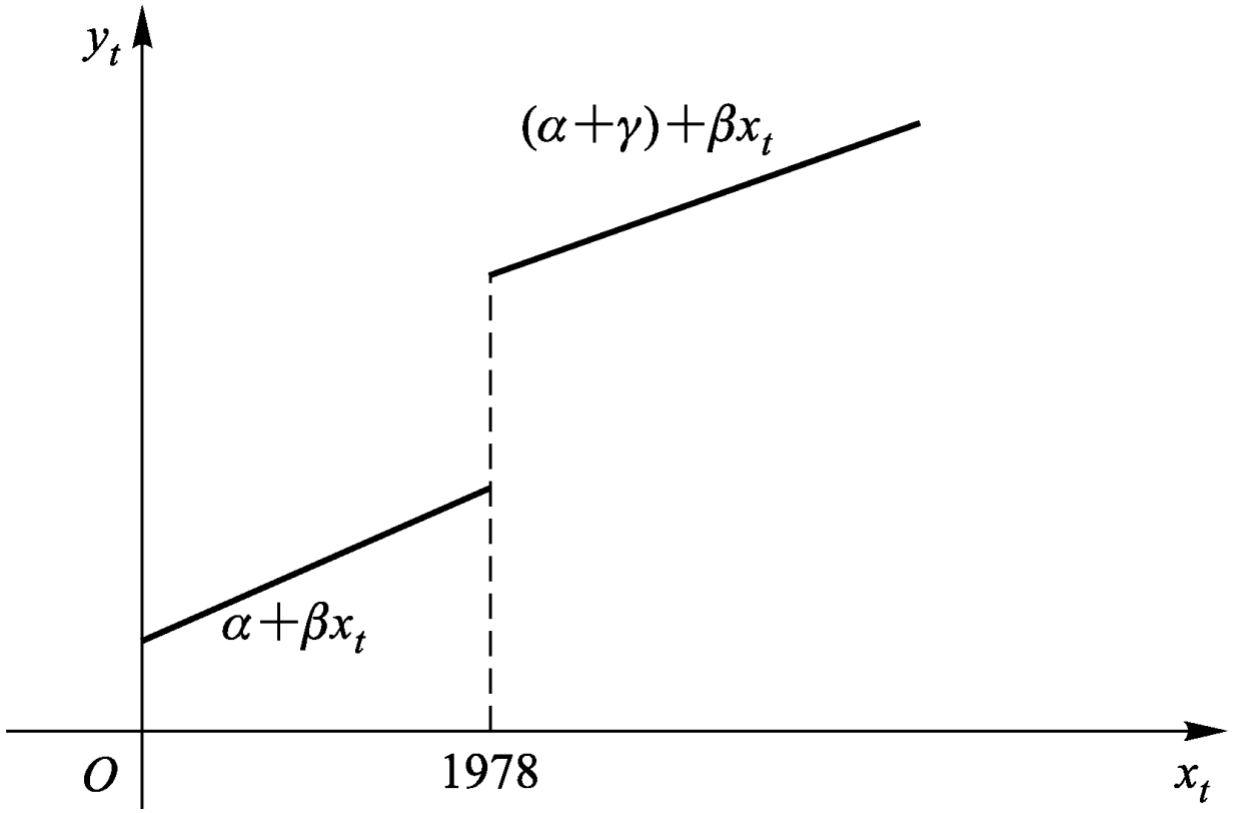
（1） 仅仅引入虚拟变量D_t本身，回归方程为

$$y_t = \alpha + \beta x_t + \gamma D_t + \varepsilon_t$$

该模型等价于

$$y_t = \begin{cases} \alpha + \beta x_t + \varepsilon_t, & \text{若} t < 1978 \\ (\alpha + \gamma) + \beta x_t + \varepsilon_t, & \text{若} t \geq 1978 \end{cases}$$

- 仅引入虚拟变量相当于在不同时期给予不同截距项



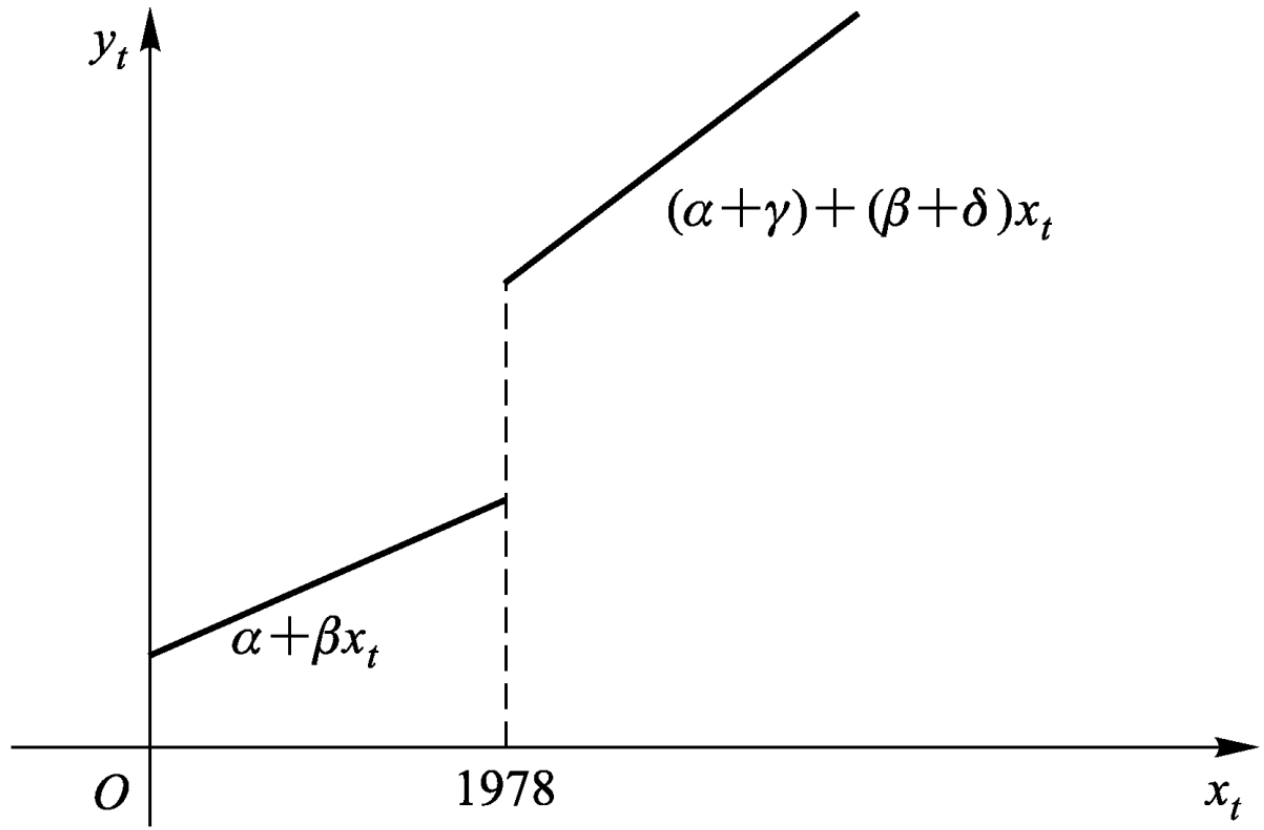
- (2) 引入虚拟变量D_t，以及虚拟变量与解释变量的“互动项”(interaction term)$D_t x_t$，回归方程为

$$y_t = \alpha + \beta x_t + \gamma D_t + \delta D_t x_t + \varepsilon_t$$

该模型等价于

$$y_t = \begin{cases} \alpha + \beta x_t + \varepsilon_t, & \text{若} t < 1978 \\ (\alpha + \gamma) + (\beta + \delta) x_t + \varepsilon_t, & \text{若} t \geq 1978 \end{cases}$$

- 引入虚拟变量及其互动项，相当于在不同时期使用不同的**截距项**与**斜率**。
- 如果仅仅引入互动项，则仅改变斜率(比较少见)。


y_t
$(\alpha+\gamma)+(\beta+\delta)x_t$
$\alpha+\beta x_t$
O
1978
x_t

分税与税率:从政府到企业*

吕冰洋　马光荣　毛　捷

内容提要:分税体现着政府间财政关系,税率体现着政府与市场的关系,通过对分税与税率关系的研究,可以观察政府间财政关系对政府与市场关系的影响。研究认为,中国政府间实行税收弹性分成、企业面临着差异化税率是两个重要典型事实;理论分析发现,地方政府偏好、税收分成比例分别影响着企业实际税率;实证分析发现,市县级政府面临的企业所得税和增值税分成比例上升,导致企业逃税减少和实际税率上升。上述结论从分税的角度揭示了政府与市场的关系如何受财政体制影响的独特机制,解释了中国企业面临的差异化税率的来源,为下一步财税体制改革提供了参考。

关键词:弹性分成　差异化税率　财政分权　政府与市场

$$ETR_{ipt} = \beta_1 VATShare_{pt} + \beta_2 X_{ipt} + \beta_3 Z_{pt} + \rho_i + \alpha_t + \zeta_{ict} \quad (25)$$

其中，因变量 ETR_{ipt} 是企业的增值税实际税率，即企业缴纳的增值税与增值额之比，下标 i、p、t 分别代表企业、省份和年份。关键解释变量 $VATShare_{pt}$ 是省份 p 的市县政府增值税分成率。X_{ipt} 是企业层面的控制变量，包括企业规模（增值额的对数）、增值率（增值额与总产出之比）①、销售利润率（利润与销售额之比）、年龄（取对数）、贷款能力、出口额占总产值的比重②、是否为国有企业、是否为外资企业等。Z_{pt} 是省级层面的控制变量，同样包括人均 GDP、市场化指数和财政自主度。ρ_i 和 α_t 分别是企业固定效应和年份固定效应，通过控制企业固定效应，可以很大程度上控制因企业所在行业、所生产产品类型导致的增值税实际税率差异。ζ_{ict} 是随机扰动项，我们将标准误群聚（clustering）到省层面。

表 6　　增值税分成对税率的影响：不同类型企业

	被解释变量：增值税实际税率（ETR）		
	（1）	（2）	（3）
增值税分成率	0.104**	0.101**	0.084**
	（0.044）	（0.043）	（0.039）
增值税分成率×大企业	0.007		
	（0.01）		
增值税分成率×国有企业		0.002	
		（0.023）	
增值税分成率×外资企业			0.133***
			（0.042）
企业控制变量	控制	控制	控制
县级控制变量	控制	控制	控制
企业固定效应	控制	控制	控制
年份固定效应	控制	控制	控制
Observations	1406204	1406204	1406204
R-squared	0.619	0.619	0.619

注：括号内为标准误，标准误均群聚（cluster）在省份层面。*、**、*** 分别代表在 10%、5% 和 1% 水平下显著。

经济结构变动的检验

- 对于时间序列而言，模型系数的稳定性(stability)很重要。
- 如存在“结构变动”(structural break)，但未加考虑，也是模型设定误差。
- 考虑结构变动日期已知的情形。
- 检验中国经济是否在 1978 年发生结构变动。
- 定义第 1 个时期为$1950 \leq t < 1978$，第2个时期为$1978 \leq t \leq 2010$。
- 原假设为，经济结构在这两个时期内没有变化，即“$H_0: \alpha_1 = \alpha_2, \beta_1 = \beta_2$”,共有两个约束。

- 如有K个解释变量(包含常数项)，则H_0共有K个约束。
- 在无约束的情况下，对两个时期，分别进行回归。
- 在有约束的情况下，可将模型合并为

$$y = \alpha + \beta x_t + \varepsilon_t \quad (1950 \leq t \leq 2010)$$

- 其中，$\alpha = \alpha_1 = \alpha_2$，$\beta = \beta_1 = \beta_2$。
- 可将所有样本数据合在一起回归
- 传统的“邹检验”(Chow, 1960)通过作三个回归来检验“无结构变动”的原假设。
- (1) 回归整个样本，$1950 \leq t \leq 2010$得到残差平方和，记为SSR^*。
- (2) 回归第 1 部分子样本，$1950 \leq t \leq 1978$,得到残差平方和SSR_1。
- (3)回归第 2 部分子样本，$1978 \leq t \leq 2010$,得到残差平方和SSR_2。

- 将整个样本一起回归为“有约束 OLS”，其残差平方和为SSR*。
- 将样本一分为二，分别进行回归为“无约束 OLS”，其残差平方和为
$$SSR = SSR_1 + SSR_2$$
- $SSR^* \geq SSR = SSR_1 + SSR_2$，因为有约束OLS的拟合优度比无约束OLS更差。
- 如H_0成立(无结构变动），则$(SSR^* - SSR_1 - SSR_2)$应较小；施加约束后，不应使残差平方和上升很多。
- 如$(SSR^* - SSR_1 - SSR_2)$很大，则倾向于认为H_0不成立，存在结构变动。

- 在对m个线性约束进行联合检验时，似然比检验原理的F统计量为

$$F = \frac{(SSR^* - SSR)/m}{SSR/(n-K)} \sim F(m, n-K)$$

- 其中，SSR为无约束的残差平方和，SSR^*为有约束的残差平方和，n为样本容量，而K为无约束回归的参数个数。
- 回到检验结构变动的情形，如果有K个解释变量（含常数项），则共有K个约束条件，而无约束回归的参数个数为$2K$。
- 检验结构变动的F统计量：

$$F = \frac{(SSR^* - SSR_1 - SSR_2)/K}{(SSR_1 + SSR_2)/(n-2K)} \sim F(K, n-2K)$$

- 其中，n为样本容量，K 为有约束回归的参数个数(含常数项)。
- 对于一元回归的例子，$K = 2$。

- 检验结构变动的另一方法是引入虚拟变量，并检验所有虚拟变量以及其与解释变量交叉项的系数的联合显著性。
- 对于$K = 2$的情形，可进行如下回归：

$$y_t = \alpha + \beta x_t + \gamma D_t + \delta D_t x_t + \varepsilon_t$$

- 检验联合假设“$H_0: \gamma = \delta = 0$”。
- 所得F 统计量与邹检验完全相同，故虚拟变量法与邹检验等价。

- 与邹检验相比，虚拟变量法的优点包括：
- (1) 只需生成虚拟变量即可检验，十分方便；
- (2) 邹检验在“球形扰动项”(同方差、无自相关)的假设下得到，并不适用于异方差或自相关的情形。
- 在异方差或自相关的情况下，仍可使用虚拟变量法，只要在估计方程时，使用异方差自相关稳健的 HAC 标准误即可。
- (3) 如发现结构变动，邹检验不提供究竟是截距项还是斜率变动的信息，虚拟变量法可提供这些信息。

缺失数据与线性插值

- 有时会出现某些时期数据缺失(missing data)的情形，尤其是历史较久远的数据。
- 缺失的观测值在 Stata 中以“.”来表示。
- 在运行 Stata 命令时(比如 reg)，会自动将缺失的观测值从样本中去掉，导致样本容量损失。
- 在数据缺失不严重的情况下，为保持样本容量，可采用“线性插值”(linear interpolation)的方法补上缺失数据。

- 已知x_{t-1}与x_{t+1}，但缺失x_t的数据，则x_t对时间t的线性插值为

$$\widehat{x_t} = \frac{x_{t-1} + x_{t+1}}{2}$$

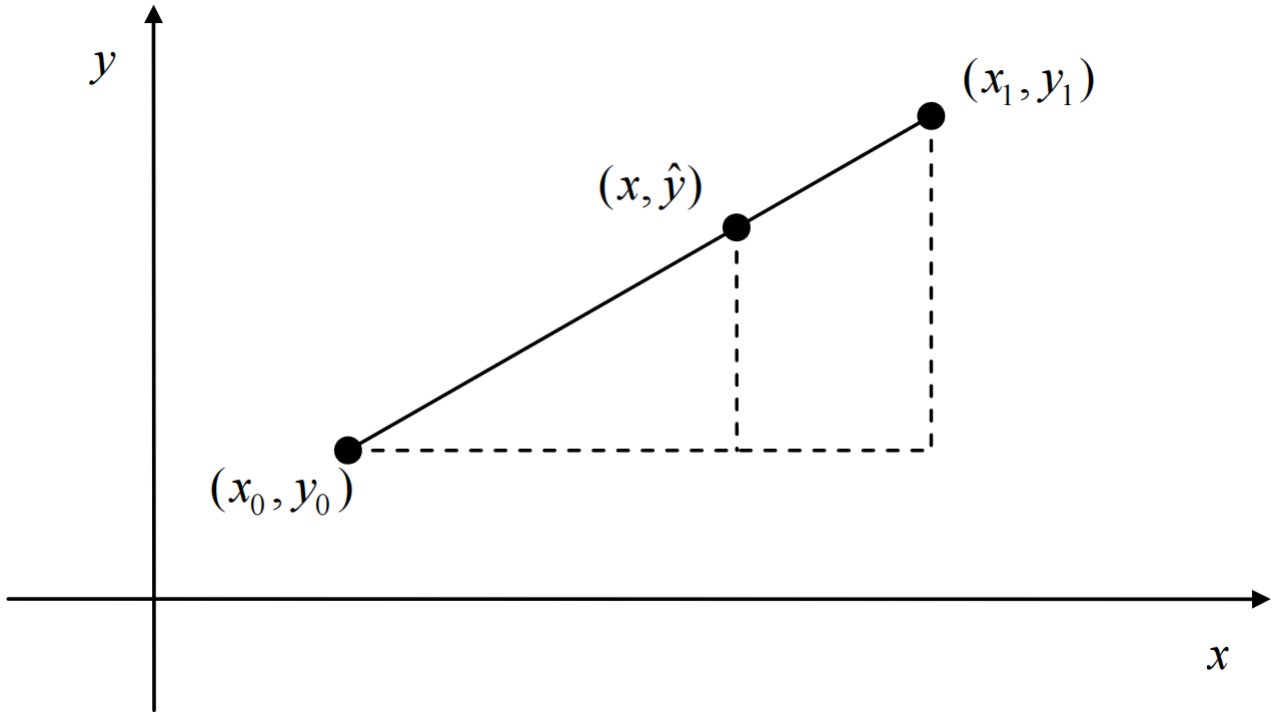
- 更一般地，假设与x(通常为时间)对应的y缺失，而最临近的两个点分别为(x_0, y_0)与(x_1, y_1)，且$x_0 < x < x_1$，则y对x的线性插值$\hat{y}$满足：

$$\frac{\hat{y} - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0}$$

- 整理可得

$$\hat{y} = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + y_0$$

- 线性插值的基本假设是变量以线性速度均匀变化。
- 如变量y有指数增长趋势势(比如 GDP)，应先取对数，再用lny进行线性插值，以避免偏差。
- 如需要以原变量y进行回归，可将将线性插值的对数值$\widehat{lny}$取反对数(antilog)，即计算$exp(\widehat{lny})$.



变量单位的选择

- 在选择变量单位时，应尽量避免变量间的数量级差别过于悬殊，

以免出现计算机运算的较大误差。

- 比如，通货膨胀率通常小于 1，而如果模型中有 GDP 这个变量，则 GDP 应该使用亿或万亿作为单位。
- 否则，变量 GDP 的取值将是通货膨胀率的很多倍，即数据矩阵$\boldsymbol{X}$中某列的数值是另一列的很多倍，可能使计算机在对$(\boldsymbol{X}'\boldsymbol{X})^{-1}$进行数值计算时出现较大误差。
- 电脑的存储空间有限，实际上只能作近似计算，即精确到小数点后若干位。

谢谢！