第七讲　功夫计量

樊潇彦

复旦大学经济学院

本讲主要内容

1. 功夫计量之混沌初开

2. 功夫计量之盖世五侠

2.1 随机实验（random experiment）

2.2 匹配回归（matching regression）

2.3 工具变量（instrumental variables, IV）

2.4 断点回归（discontinuity regression, RD）

2.5 双差分（differences in differences, DID）

3. 功夫计量之神龙大侠

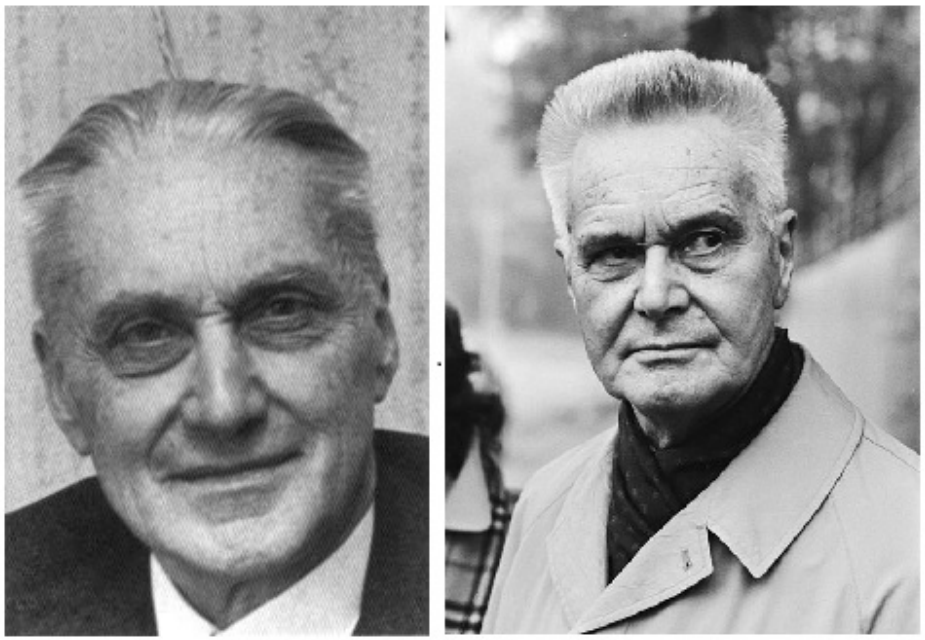
3.1 计量回归模型分类

3.2 古典线性回归模型：OLS估计

3.3 古典线性回归模型的拓展：IV和GLS估计

什么是计量经济学？

Wiki: Econometrics is the application of mathematics, statistical methods, and computer science to economic data and is described as the branch of economics that aims to give empirical content to economic relations. ...allowing economists "to sift through mountains of data to extract simple relationships."



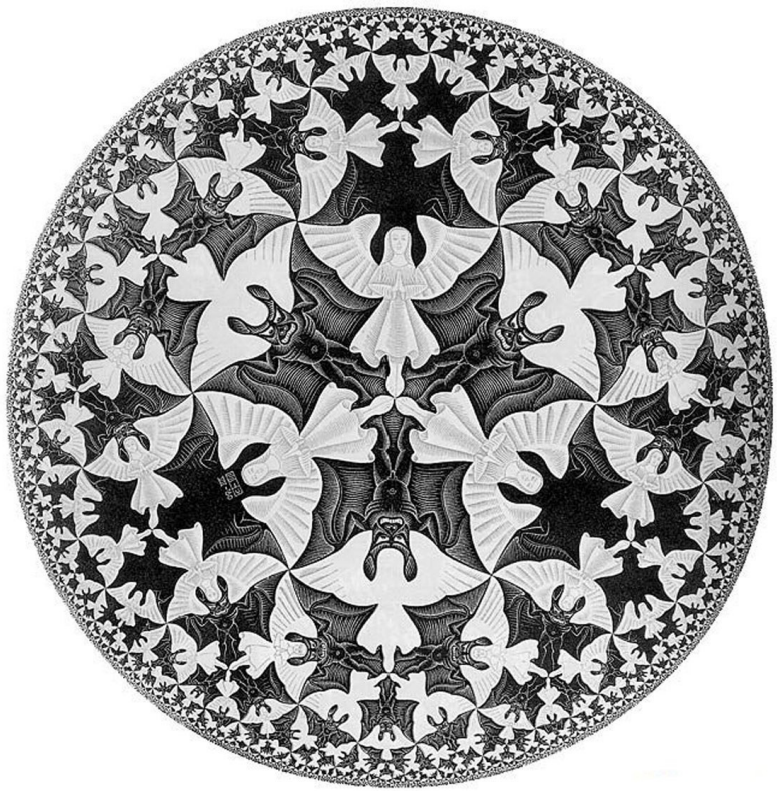
注：计量经济学的开创者，1969年第一届诺贝尔经济学奖获得者拉格纳·弗里希（左）和简·丁伯根（右）。

颁给计量经济学的诺奖

	获奖者	主要贡献
1969	弗里希、丁伯根	提出用计量方法研究经济周期和宏观经济，开创了计量经济学。
1980	克莱因	建立和估计宏观经济联立方程模型。
1989	哈维默	区分理论结构方程和计量简化方程，指出回归分析中的内生性和不可识别问题。
2000	赫克曼、麦克法登	提出微观数据中存在的逆向选择和离散变量回归的处理方法。
2003	恩格尔、格兰杰	提出时间序列数据的建模和分析方法。
2011	西姆斯	提出向量自回归方法，分析时序变量对冲击的动态响应行为。

从数理经济学到数量经济学……

……最难的是什么？



注：M.C.埃舍尔1941年作品《天使与魔鬼》。

计算积分或者进行线性回归，用计算机就能完成，但是，判断所得结果是否有意义，或者判断所采用的方法是否正确，则离不开人的智慧。我们在教授数学时，应该告诉学生如何应用人的智慧，否则，我们培养出来的学生从本质上就会与微软的Excel程序没什么两样，而且反应迟钝、漏洞百出。

—— J. 艾伦伯格，《魔鬼数学》

功夫计量之盖世五侠（Furious Five）

1. 随机实验（random experiment）
2. 匹配回归（matching regression）
3. 工具变量（instrumental variables, IV）
4. 断点回归（discontinuity regression, RD）
5. 双差分（differences in differences, DID）

随机实验：医保让你的健康变差了吗？

Outcomes and treatments for Khuzdar and Maria

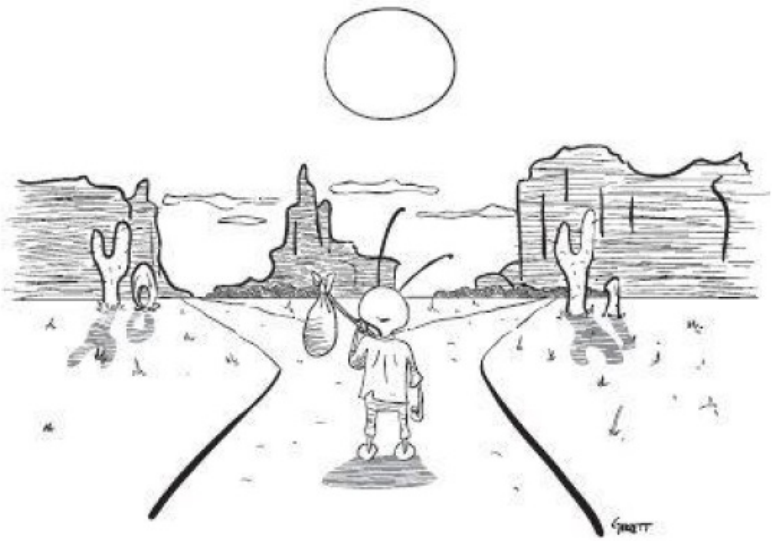
	Khuzdar Khalat	Maria Moreño
Potential outcome without insurance: Y_{0i}	3	5
Potential outcome with insurance: Y_{1i}	4	5
Treatment (insurance status chosen): D_i	1	0
Actual health outcome: Y_i	4	5
Treatment effect: $Y_{1i} - Y_{0i}$	1	0

$$Y_{1,K} - Y_{0,M} = \underbrace{Y_{1,K} - Y_{0,K}}_{\text{因果效应（causal effect）}} + \underbrace{Y_{0,K} - Y_{0,M}}_{\text{选择偏误（selection bias）}}$$

随机实验：消除自选择

假定对个人而言参加医保对健康有好处 $Y_{1,i} - Y_{0,i} = \kappa$，通过随机指定，使是否被保险与个人初始健康状况无关：

$$E[Y_{0,i}|D_i = 1] = E[Y_{0,i}|D_i = 0]$$



Breaking the Deadlock: Just RANDomize

因此从社会整体来看：

$$\begin{aligned}
&E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 0] \\
=&E[Y_{0,i} + \kappa|D_i = 1] - E[Y_{0,i}|D_i = 0] \\
=&\kappa + E[Y_{0,i}|D_i = 1] - E[Y_{0,i}|D_i = 0] \\
=&\kappa
\end{aligned}$$

随机实验：OREGON州的例子

In an awesome social experiment, the state of Oregon recently offered Medicaid to thousands of randomly chosen people in a publicly announced health insurance lottery.

...Winners won the opportunity to apply for the state-run Oregon Health Plan (OHP).

	Oregon		Portland area	
Outcome	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
A. Health indicators				
Health is good	.548	.039 (.008)		
Physical health index			45.5	.29 (.21)
Mental health index			44.4	.47 (.24)
Cholesterol			204	.53 (.69)
Systolic blood pressure (mm Hg)			119	−.13 (.30)
B. Financial health				
Medical expenditures >30% of income			.055	−.011 (.005)
Any medical debt?			.568	−.032 (.010)
Sample size	23,741		12,229	

匹配回归：读私立大学会提高收入吗？

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

$$E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 0] = ?$$

匹配回归：先匹配、再回归

- 匹配（matching）：将学生按照学校录取情况分组，被相同的学校录取（或拒绝）的同学为一组，所在组别 G_{ij} 可以表示学生的个人能力等不可观测的因素；
- 回归（regression）：被解释变量 Y_i 为收入，解释变量包括是否就读私立大学 $P_i = 0, 1$（treatment variable）和所在组别 G_{ij} 等控制变量（control variables）。

以 1-5 号同学为样本，回归以下方程，问 $\hat{\alpha}, \hat{\beta}, \hat{\gamma} =?$

$$Y_i = \alpha + \beta P_i + \gamma A_i + \varepsilon_i$$

匹配回归：COLLEGE AND BEYOND 数据

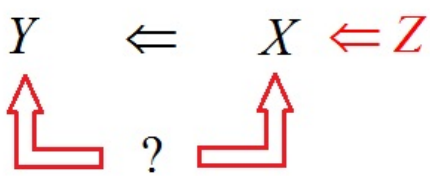
$$lnY_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j G_{ij} + \delta X_i + \varepsilon_i$$

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
			(.018)			(.021)
Black			.005			−.040
			(.041)			(.042)
Hispanic			.062			.032
			(.072)			(.070)

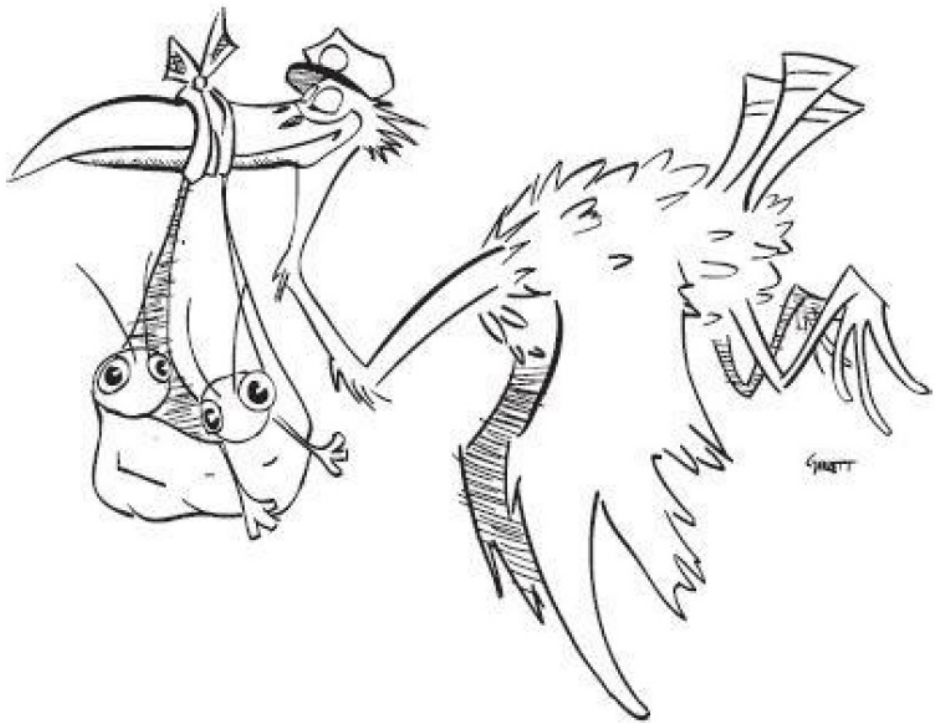
(continued)	(3)	(6)
Asian	.170	.145
	(.074)	(.068)
Other/missing race	−.074	−.079
	(.157)	(.156)
High school top 10%	.095	.082
	(.027)	(.028)
High school rank missing	.019	.015
	(.033)	(.037)
Athlete	.123	.115
	(.025)	(.027)
Selectivity-group dummies	No	Yes

工具变量：数量还是质量？

父母面临子女数量 X 和质量（Y，如受教育程度）之间的权衡，由于很多其他因素（父母收入水平、受教育程度等）都会同时影响 X, Y，因此我们需要寻找只与 X 有关的工具变量 Z。



$Y \Leftarrow \hat{X}(Z)$



工具变量：两阶段回归（2SLS）

第一阶段（first stage）：$X_i = c + \gamma Z_i + e_i$

第二阶段（second stage）：$Y_i = \alpha + \beta \hat{X}_i + \varepsilon_i$

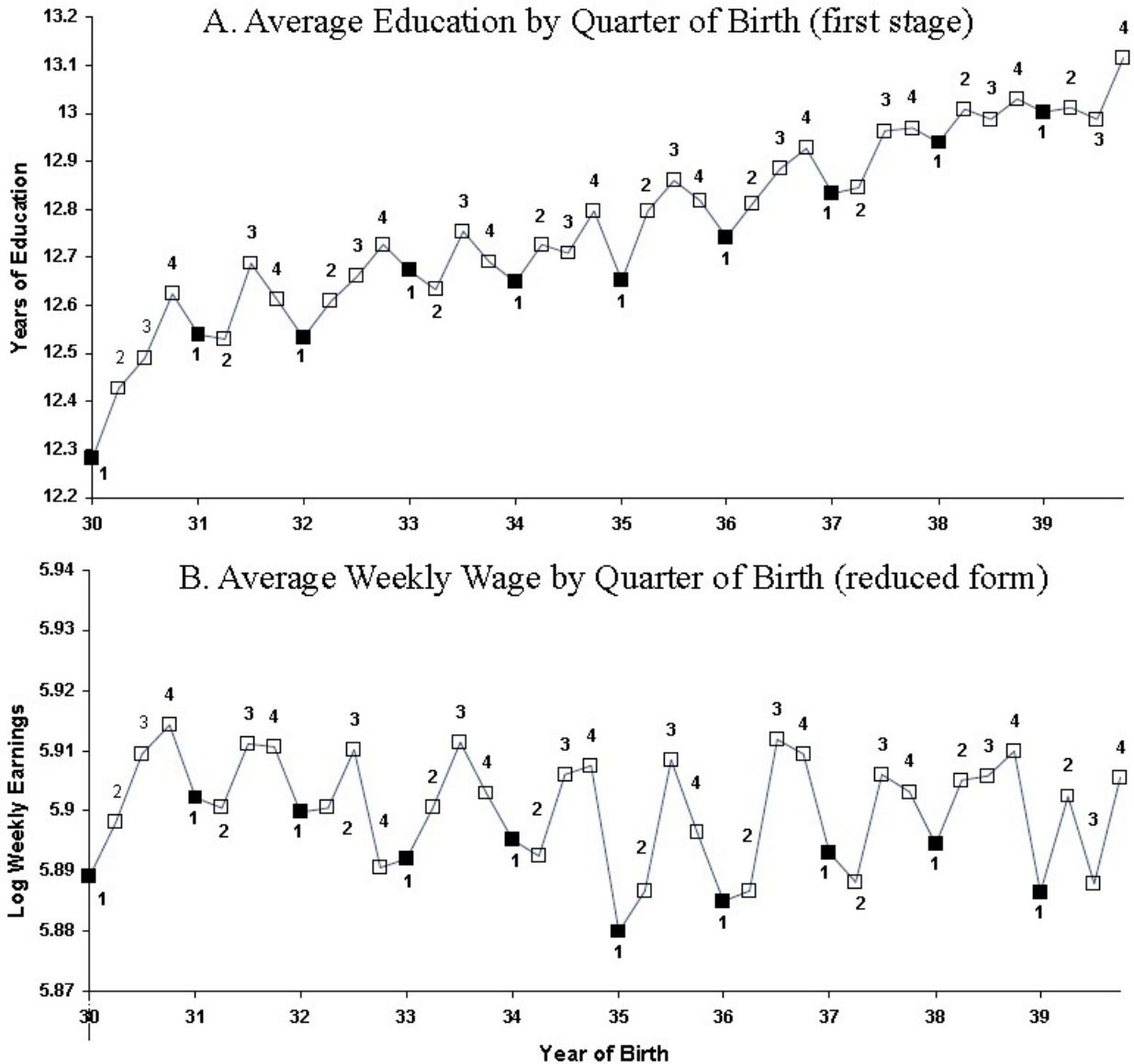
Quantity-quality first stages

	Twins instruments		Same-sex instruments		Twins and same-sex instruments
	(1)	(2)	(3)	(4)	(5)
Second-born twins	.320	.437			.449
	(.052)	(.050)			(.050)
Same-sex sibships			.079	.073	.076
			(.012)	(.010)	(.010)
Male		−.018		−.020	−.020
		(.010)		(.010)	(.010)
Controls	No	Yes	No	Yes	Yes

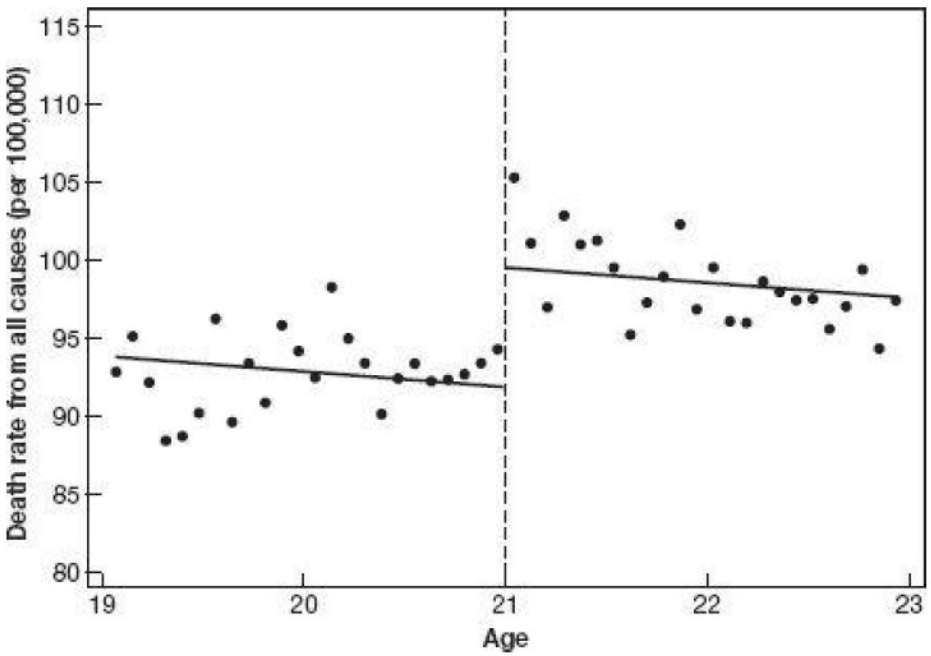
OLS and 2SLS estimates of the quantity-quality trade-off

		2SLS estimates		
Dependent variable	OLS estimates (1)	Twins instruments (2)	Same-sex instruments (3)	Twins and same-sex instruments (4)
Years of schooling	−.145	.174	.318	.237
	(.005)	(.166)	(.210)	(.128)
High school graduate	−.029	.030	.001	.017
	(.001)	(.028)	(.033)	(.021)
Some college (for age ≥ 24)	−.023	.017	.078	.048
	(.001)	(.052)	(.054)	(.037)
College graduate (for age ≥ 24)	−.015	−.021	.125	.052
	(.001)	(.045)	(.053)	(.032)

工具变量：工资、教育和出生时间



断点回归：危险的21岁

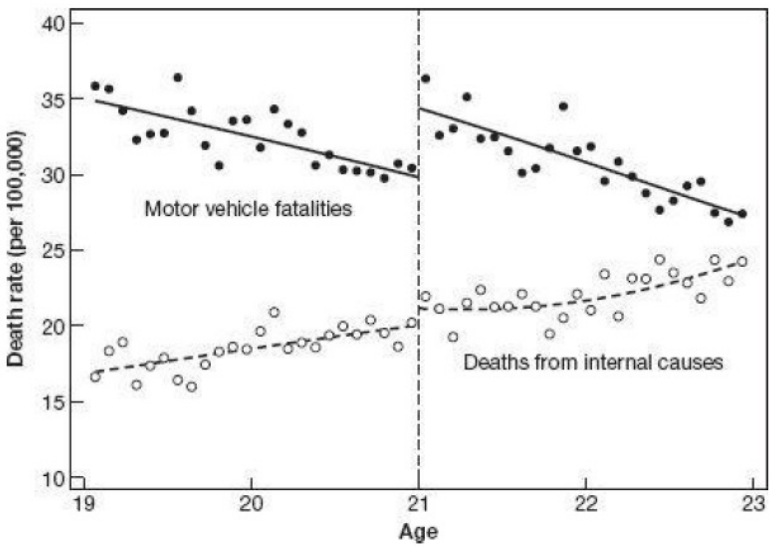


KATY: Is this really what you’ re gonna do for the rest of your life?
BOON: What do you mean?
KATY: I mean hanging around with a bunch of animals getting drunk every weekend.
BOON: No! After I graduate, I’ m gonna get drunk every night.

断点回归：哑变量的作用

记 M_a 为年龄 a 的死亡率，$D_a = 0, 1$ 为是否达到21岁的哑变量，回归以下方程：

$$M_a = \alpha + \beta D_a + \gamma a + \varepsilon_a$$



Dependent variable	Ages 19–22 (1)	Ages 19–22 (2)	Ages 20–21 (3)	Ages 20–21 (4)
All deaths	7.66 (1.51)	9.55 (1.83)	9.75 (2.06)	9.61 (2.29)
Motor vehicle accidents	4.53 (.72)	4.66 (1.09)	4.76 (1.08)	5.89 (1.33)
Suicide	1.79 (.50)	1.81 (.78)	1.72 (.73)	1.30 (1.14)
Homicide	.10 (.45)	.20 (.50)	.16 (.59)	−.45 (.93)
Other external causes	.84 (.42)	1.80 (.56)	1.41 (.59)	1.63 (.75)
All internal causes	.39 (.54)	1.07 (.80)	1.69 (.74)	1.25 (1.01)
Alcohol-related causes	.44 (.21)	.80 (.32)	.74 (.33)	1.03 (.41)
Controls	age	age, age^2, interacted with over-21	age	age, age^2, interacted with over-21
Sample size	48	48	24	24

断点回归：内贾德作弊了吗？

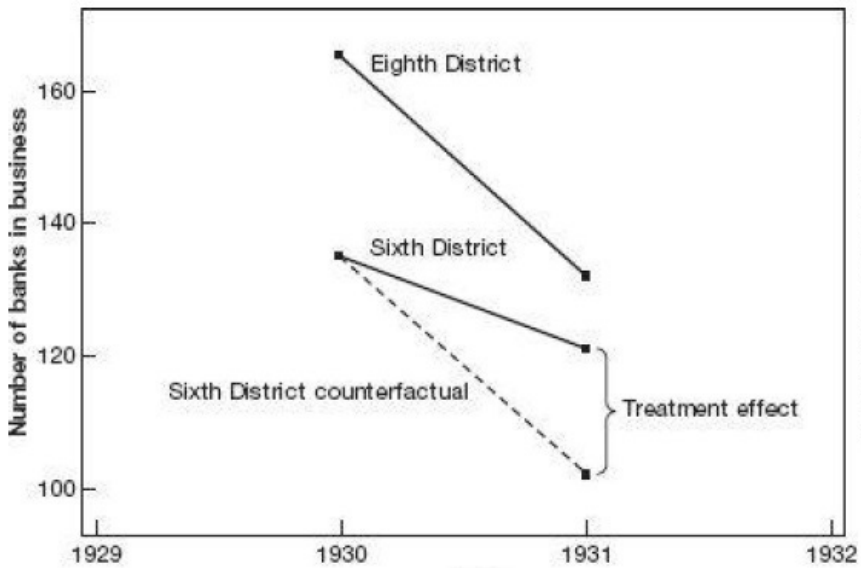
- 2009年，时任伊朗总统的M.艾哈迈迪内贾德在总统选举中以较大优势获胜。很多人指责有人暗中操控选票。
- 哥伦比亚大学的两名研究生B.比伯和A.斯卡想出了一个好办法：
  - 4名主要候选人在29个省的得票数，共116个样本；
  - 如果票数没有造假，那么这些数字的末位数应该是随机的，也就是0-9每个数字出现的概率为10%；
  - 实际上，7出现的次数过多，几乎是正常概率的两倍。

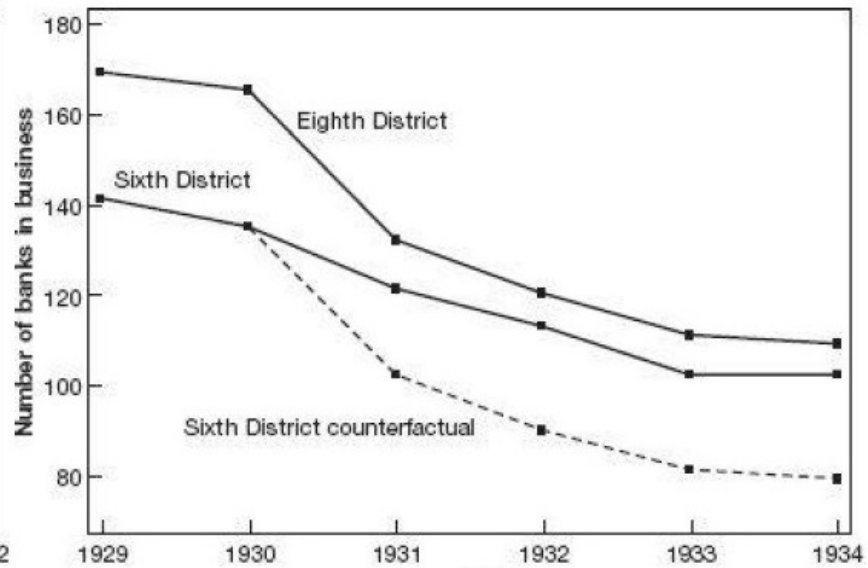
双差分：大萧条时央行应该救银行吗？

Wholesale firm failures and sales in 1929 and 1933

	1929	1933	Difference (1933–1929)
Panel A. Number of wholesale firms			
Sixth Federal Reserve District (Atlanta)	783	641	−142
Eighth Federal Reserve District (St. Louis)	930	607	−323
Difference (Sixth–Eighth)	−147	34	181
Panel B. Net wholesale sales ($ million)			
Sixth District Federal Reserve (Atlanta)	141	60	−81
Eighth District Federal Reserve (St. Louis)	245	83	−162
Difference (Sixth–Eighth)	−104	−23	81

双差分：识别救助效果





$$Y_{it} = 167 - \underset{(8.8)}{29}\, D_i - \underset{(7.6)}{49}\, POST_t + \underset{(10.7)}{20.5}\,(D_i \times POST_t) + \varepsilon_{it}$$

双差分：水污染的后果

下图为John Snow(1855)采集的伦敦各区1849和1854年霍乱死亡人口，及供水公司的数据。1852年之前两家公司的水源均有污染，之后Lambeth公司找到了清洁水源，Southwark and Vauxhall公司没有变化。

Sub-Districts.	Deaths from Cholera in 1849.	Deaths from Cholera in 1854.	Water Supply.
St. Saviour, Southwark	283	371	Southwark & Vauxhall Company only.
St. Olave	157	161	
St. John, Horsleydown	192	148	
St. James, Bermondsey	249	362	
St. Mary Magdalen	259	244	
Leather Market	226	237	
Rotherhithe*	352	282	
Wandsworth	97	59	
Battersea	111	171	
Putney	8	9	
Camberwell	235	240	
Peckham	92	174	
Christchurch, Southwar	256	113	
Kent Road	267	174	
Borough Road	312	270	
London Road	257	93	Lambeth Company, and Southwark and Vauxhall Compy.
Trinity, Newington	318	210	
St. Peter, Walworth	446	388	
St. Mary, Newington	143	92	
Waterloo Road (1st)	193	58	
Waterloo Road (2nd)	243	117	
Lambeth Church (1st)	215	49	
Lambeth Church (2nd)	544	193	
Kennington (1st)	187	303	
Kennington (2nd)	153	142	
Brixton	81	48	
Clapham	114	165	
St. George, Camberwell	176	132	
Norwood	2	10	Lambeth Company only.
Streatham	154	15	
Dulwich	1	—	
Sydenham	5	12	
First 12 sub-districts	2261	2458	Southwk. & Vauxhall.
Next 16 sub-districts	3905	2547	Both Companies.
Last 4 sub-districts	162	37	Lambeth Company.

请问清洁水源可能使因霍乱而死亡的人口下降多少？你能先大致估算，然后用计量的方法进行检验吗？

计量回归模型分类

- 根据数据类型分类：

	x 连续	x 离散
y 连续	线性与非线性回归	方差分析
y 离散	广义线性模型	

- 根据数据性质分类：
  - 截面数据（cross-section）：如上
  - 面板数据（panel）：固定/随机效应模型（Fixed/Random effect model）
  - 时序数据（time series）

古典线性回归模型的假定

1. 自变量 y 与因变量 $X=(x_1,x_2...x_K)$ 之间存在线性关系：

$$y=X\beta+\varepsilon$$

2. 给定数据样本，$N\times K$ 维矩阵 X 满秩：

$$rank\,(X)=K<N,\quad 或\ X^TX可逆$$

3. 随机误差与 X 相互独立且期望为零：

$$E\,(\varepsilon|X)=E\,(\varepsilon)=0,\quad 或\ E\left(X^T\varepsilon\right)=0$$

4. 随机误差之间相互独立（$E\,(\varepsilon_i\varepsilon_j)=0,\ i\neq j$），且服从同样的正态分布（$\varepsilon_i\sim N\left(0,\sigma^2\right),\ i=1,2...N$），用矩阵形式表示如下（其中 I_N 为 N 维单位方阵）：

$$\varepsilon\sim N\left(0,\sigma^2\mathrm{I}_N\right),\quad 或\ E\left(\varepsilon\varepsilon^T\right)=\sigma^2\mathrm{I}_N$$

普通最小二乘（OLS）估计

$$\begin{aligned}\hat{\beta}^{OLS} &= \arg\min_{\beta} \varepsilon^T \varepsilon \\ &= \arg\min_{\beta} \|y - X\beta\|\end{aligned}$$

- $\hat{\beta}^{OLS} = \left(X^T X\right)^{-1} X^T y$
- $Var\left(\hat{\beta}^{OLS}\right) = \sigma^2 \left(X^T X\right)^{-1}$
- $\hat{\sigma}^2 = s^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{N-K}$

OLS估计量的性质

- 无偏性

$$E\left(\hat{\beta}^{OLS}\right)=\beta,\quad E\left(\hat{\sigma}^2\right)=\sigma^2$$

- 有效性

$\hat{\beta}^{OLS}$ 是最佳线性无偏估计量（BLUE），$Var\left(\hat{\beta}^{OLS}\right)$ 在所有无偏估计量的方差中最小（Gauss-Markov定理）。

- 一致性 $(N\to\infty)$

$$\hat{\beta}^{OLS}\xrightarrow{p}\beta,\quad \hat{\sigma}^2\xrightarrow{p}\sigma^2$$

- 渐进分布 $(N\to\infty)$

$$\hat{\beta}^{OLS}\sim N\left(\beta, s^2\left(X^TX\right)^{-1}\right)$$

其中 $\hat{\beta}_k^{OLS}\sim N\left(\beta_k,\ \frac{s^2}{\sum_{i=1}^N x_{ik}^2}\right)$

存在内生性问题：IV估计

- 古典线性回归模型的假定 3 不成立，则解释变量存在内生性问题，将导致 $\hat{\beta}$ 有偏。

$$E\left(X^T\varepsilon\right) \neq 0$$

- 假定存在内生变量 X 的工具变量（instrumental variable, IV）Z，满足：
  - 外生性条件：$E\left(Z^T\varepsilon\right) = 0$
  - 相关性条件：$Z^TX = \Sigma_{ZX}$ 可逆
  - 满秩条件：$Z^TZ = \Sigma_{ZZ}$ 可逆
- 工具变量（IV）估计：

$$\hat{\beta}^{IV} = \left(Z^TX\right)^{-1}Z^Ty$$
$$Var\left(\hat{\beta}^{IV}\right) = \sigma^2\left(\Sigma_{ZX}^T\Sigma_{ZZ}^{-1}\Sigma_{ZX}\right)^{-1}$$
$$\hat{\sigma}^2 = s^2 = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{N-K}$$

存在异方差和残差自相关：GLS估计

- 古典线性回归模型的假定 4 不成立，存在异方差和残差自相关问题，将影响参数估计的有效性，使置信区间估计和统计检验有误：

$$\varepsilon \sim N\left(0, \sigma^2 \Omega\right), \quad \Omega \neq \mathrm{I_N}$$

- 广义最小二乘（Generalized least squares, GLS）估计：

$$\begin{aligned}
\hat{\beta}^{GLS} &= \left(X^T \Omega^{-1} X\right)^{-1} X^T \Omega^{-1} y \\
Var\left(\hat{\beta}^{GLS}\right) &= \sigma^2 \left(X^T \Omega^{-1} X\right)^{-1} \\
\hat{\sigma}^2 &= s^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{N-K}
\end{aligned}$$

由于我们不可能根据 N 个样本估计出对称矩阵 Ω 中的 $N(N+1)/2$ 个元素，因此在实际应用中，一般先做OLS估计，再根据方差检验结果对 Ω 的具体形式做简化假定（如只考虑异方差、自相关，或直接假定 ω_{ij} 的函数形式），最后用简化的 $\hat{\Omega}$ 作为 Ω 的一致估计，得到FGLS（Feasible GLS）估计量。