

政治经济学前沿方法论与量化分析

第四讲 多元线性回归

上课地点： 善斋306C
上课时间：周二第六大节

龙治铭
善斋307C
zhiminglong@tsinghua.edu.cn



清华大学
Tsinghua University

目录

CONTENTS



一

多元线性回归的理论基础



二

非理想情况的处理办法



三

多元线性回归在Stata中的实现



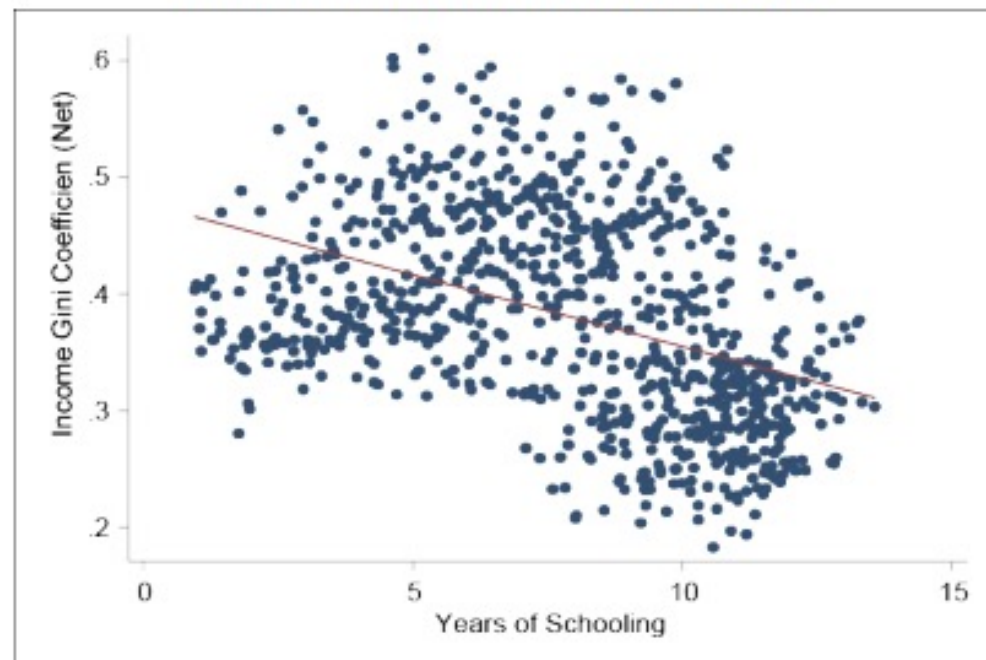
四

参考文献

1

多元线性回归的理论基础

Figure 7: Education Attainment and Income Gini Coefficients across Countries, Five-Year Intervals from 1980 to 2015



读书可否改变命运？Lee and Lee (2018)：一国平均受教育程度越高，经济不平等程度越低。

※毛主席对苏联《政治经济学教科书》的批评：这本书的写法很不好，总是从概念入手。研究问题，要从人们看得见、摸得到的现象出发，来研究隐藏在现象后面的本质，从而揭露客观事物的本质的矛盾。《资本论》对资本主义经济的分析，就是用这种方法，总是从现象出发，找出本质，然后又用本质解释现象，因此，能够提纲挈领。教科书对问题不是从分析入手，总是从规律、原则、定义出发，这是马克思主义从来反对的方法。原理、原则是结果，这是要进行分析，经过研究才能得出的。人的认识总是先接触现象，通过现象找出原理、原则来。而教科书与此相反，它所用的方法，不是分析法，而是演绎法。形式逻辑说，人都要死，张三是人，所以张三要死。这里，人都要死是大前提。教科书对每个问题总是先下定义，然后把这个定义作为大前提，来进行演绎，证明他们所要说的道理。他们不懂得，大前提也应当是研究的结果，必须经过具体分析，才能够证明是正确的。

教科书的写法，不是高屋建瓴，势如破竹，没有说服力，没有吸引力，读起来没有兴趣，一看就可以知道是一些只写文章、没有实际经验的书生写的。这本书说的是书生的话，不是革命家的话。他们做实际工作的人没有概括能力，不善于运用概念、逻辑这一套东西；而做理论工作的人又没有实际经验，不懂得经济实践。两种人，两方面——理论和实践没有结合起来。同时作者们没有辩证法。没有哲学家头脑的作家，要写出好的经济学来是不可能的。马克思能够写出《资本论》，列宁能够写出《帝国主义论》，因为他们同时是哲学家，有哲学家的头脑，有辩证法这个武器。

※回归分析(Regression Analysis)其实是一种归纳法，观察现象，从现象中找出统计规律，然后又用规律解释现象。

※散点图：形象、直观，便于发现可能的统计规律。

例：著名的菲利普曲线（Phillips, 1968），通货膨胀和失业率之间似乎存在负相关性。

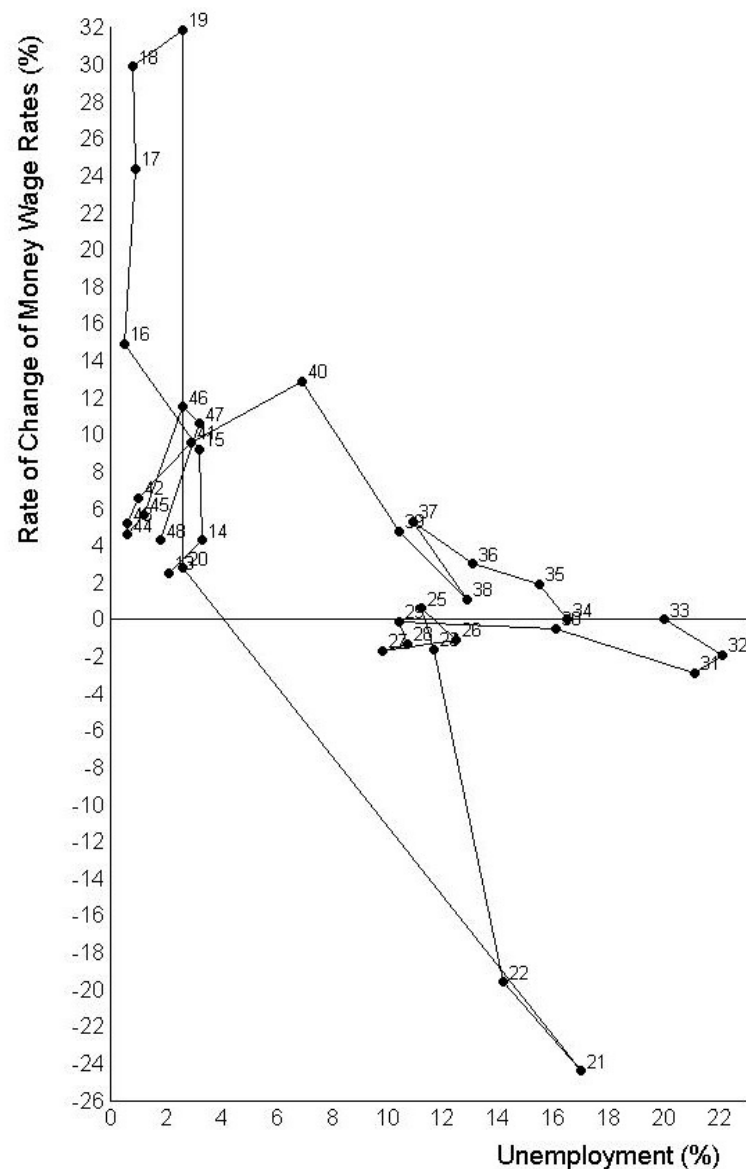
围绕着这一条曲线所暗示的统计规律，经济学家提出了不同的经济理论来解释这一现象，主导了二战以后的宏观经济研究方向（括号内为获得诺贝尔奖年份）：

支持方：Samuelson (1970)、Solow (1987)、Mundell (1999)、Prescott (2004)、Sims (2011)

反对方：Hayek (1974)、Friedman (1976)、Lucas (1995)、Phelps (2006)、Sargent (2011)

※写论文是相反的演绎法：先提出一个经济理论模型，根据理论模型提出回归模型，再通过实证分析检验模型是否成立。

理论模型的提出需要归纳法以及符合现实。



※应变变量 y ，自变量 x_1, x_2, \dots, x_{k-1} 外生且不共线，随机扰动项 ε 服从高斯分布，即 $\varepsilon \sim N(0, \sigma^2)$

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_{k-1} x_{k-1} + \varepsilon$$

样本容量为 T ，其中 $\alpha_0, \alpha_1, \dots, \alpha_{k-1}$ 为参数

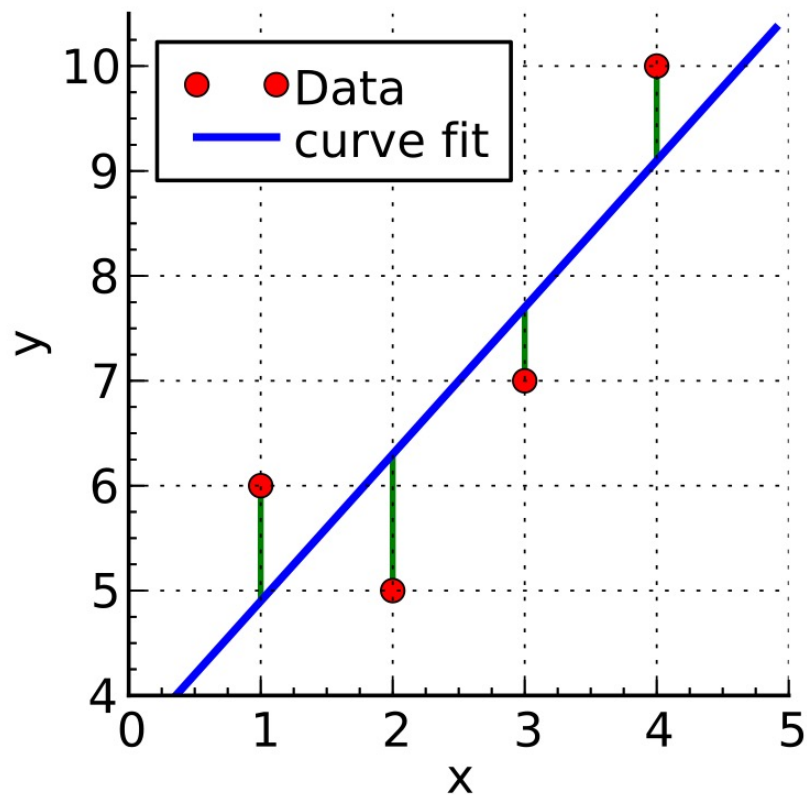
※问题：有样本容量为 T 的 $y, x_1, x_2, \dots, x_{k-1}$ 的观测值，估计 k 个参数 $\alpha_0, \alpha_1, \dots, \alpha_{k-1}$ 。估计量用 $\hat{\alpha}_0, \hat{\alpha}_1, \dots, \hat{\alpha}_{k-1}$ 表示。如何实现？

※最小二乘法（OLS）：样本空间里每一个点到“直线”（ $k-1$ 维超平面）距离最短（原问题被转化为求最小值的问题）：

$$y = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2 + \dots + \hat{\alpha}_{k-1} x_{k-1} + \hat{\varepsilon}$$

$\hat{\varepsilon}$ 称为残差值，是随机扰动项 ε 的估计量

※最大似然估计（MLE）：既然已经取到了给定的样本数据，那么表明取到这一样的概率比较大。如果随机扰动项的概率分布已知，例如高斯分布，对于给定的样本数据，使得随机扰动项概率最大（似然函数，证明从略）的参数是一组比较可靠的参数（原问题被转化为求最大值的问题）。



严格的讲，点到直线的距离是垂线段的长度，OLS是竖直方向距离平方和最短。

※将多元线性回归方程改写为矩阵形式更加简洁：

应变量 y 的 T 个观测值写为： $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_T \end{pmatrix}$, x_1, x_2, \dots, x_{k-1} 连同常数项写为矩阵 $\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{k-1,1} \\ 1 & x_{1,2} & x_{k-1,2} \\ \dots & \dots & \dots \\ 1 & x_{1,T} & x_{k-1,T} \end{pmatrix}$

一般约定用黑体的小写英文字母表示向量，黑体的大写英文字母表示矩阵。

回归模型则变为： $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

最小二乘法估计量 $\underset{(k \times 1)}{\hat{\boldsymbol{\beta}}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \dots \\ \hat{\alpha}_{k-1} \end{pmatrix}$

※高斯马尔科夫定理：在理想条件下，OLS estimator is BLUE estimator.

※理想条件：1) $\text{Rang}(\mathbf{X}'\mathbf{X})=k$ 矩阵 $\mathbf{X}'\mathbf{X}$ 满秩，变量之间不共线

2) $\text{Cov}(\mathbf{x}, \boldsymbol{\epsilon})=0$ 自变量是外生的

3) $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2 \mathbf{I}_T$ 其中 \mathbf{I}_T 为 T 维的单位矩阵 $\mathbf{I}_T = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix}$ 同方差性和非自相关性

4) $\boldsymbol{\epsilon} \sim \mathbf{N}(0, \sigma^2)$ 随机扰动项服从正态分布。正态性

※ $\text{Rang}(\mathbf{X}'\mathbf{X})=k$ 满秩及其意义

※ 不满秩的后果：多重共线性 (Multicollinearity)

- 1) 如果 $\text{Rang}(\mathbf{X}'\mathbf{X}) < k$, $(\mathbf{X}'\mathbf{X})^{-1}$ 不存在, OLS 估计量不唯一 (有无穷多个)
- 2) $\text{Rang}(\mathbf{X}'\mathbf{X}) < k$ 是极端情况, 大多数情况是尽管满秩, 但矩阵中有两列线性相关程度很高 (几乎是同一个变量或者包含大量重复的信息), 此时回归结果极为不稳定, 某一个变量一个观测值的极小变化都会引起回归结果的剧烈变化。

※ 不满秩的原因：

- 1) 过度拟合 (over fitting)。事物是广泛联系的, 自变量越多, 拟合结果当然越好。自变量数量过多, 样本容量过小 (极端情况: $T < k$)
- 2) 自变量之间存在高度的线性相关性, 例如, 若在解释股价的时候使用了成交量、总金额、总市值、换手率和流通股本等变量, 换手率 = 总金额 / 总市值, 多个变量之间相互依存, 包含了大量的重复信息。
- 3) 哑变量 (Dummy variable) 陷阱。哑变量 $D = \begin{cases} 1, & \text{满足某种条件} \\ 0, & \text{otherwise} \end{cases}$ 哑变量本身接近于常数, 多个哑变量和常数共同构成了高度线性相关的组合。一般来说, 哑变量超过三个, 多重共线性问题会变得非常严重!

理想条件的含义2：外生性 (exogenous)

※ $\text{Cov}(x, \epsilon) \neq 0$ 或者 $E(\epsilon|X) \neq 0$ 内生性问题 (Endogeneity)

※ 内生性问题的后果：OLS 估计量是有偏估计 (MLE不唯一)

存在内生性问题，回归参数是有偏估计，结论不可靠，经济学/社会科学最为严重的问题。内生性问题是统计学和计量经济学的区别所在。

广义的内生性问题：动态模型中的识别问题 (Identification problem, 时间序列部分再详细介绍)

※ 引起内生性问题的原因：

1) 遗漏变量。

例子：读书是否能够改变命运？大量的实证研究表明，一个人的收入和所接受教育的年限有正相关性，但是不是只用教育水平就可以解释收入差距呢？性别、种族的歧视，家庭背景等都是遗漏变量。

2) 测量误差。

例子：在新古典主义经济学中，产出由劳动力投入、资本投入和技术进步决定，技术进步如何测量？总是有误差的。R&D可能和残差值相关。

3) 互为因果 (feedback effect或Simultaneity) 。

巴罗 (1991) 认为，教育水平 (先用入学率，后用受教育年限代替) 决定了经济增长的差异。那么是教育决定了经济增长的差异呢还是反过来因为经济增长了，我们有更多的钱办教育？

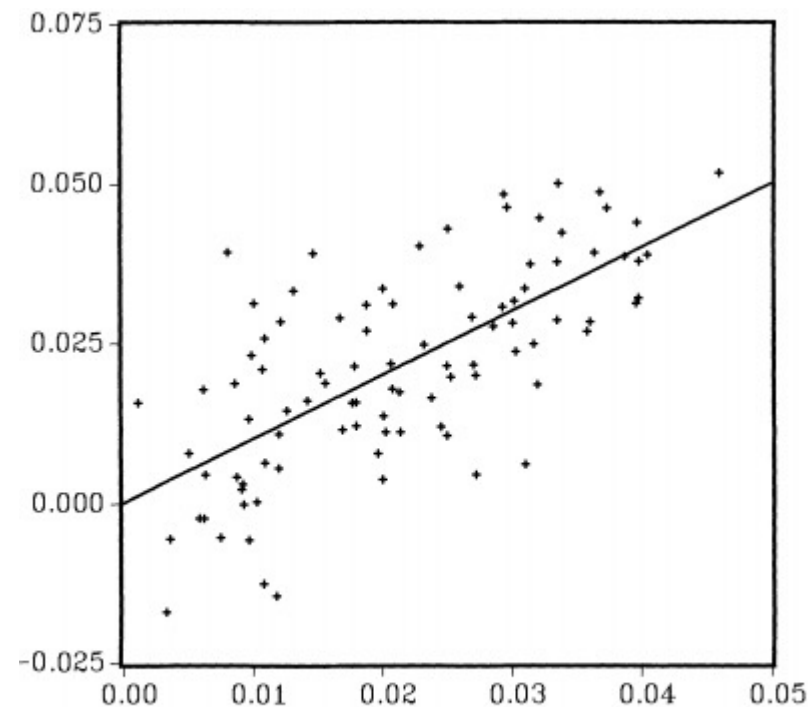


FIGURE III
Partial Association Between per Capita Growth and School-Enrollment Variables
(from regression 1 of Table I)

理想条件的含义3：从外生性到识别问题 (Identification)

※ 上面几节我们用了股市的价格和成交额作为描述性统计的例子，似乎价格和成交额之间有着很强的线性相关性。可不可以做线性回归分析呢？

※ 理论模型

$$\text{Demand : } Q_t^d = \alpha + \beta P_t + \epsilon_t^d$$

$$\text{Supply : } Q_t^s = \gamma + \delta P_t + \epsilon_t^s$$

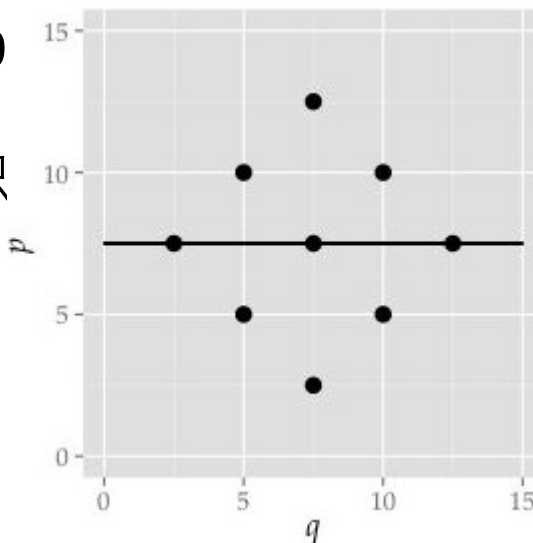
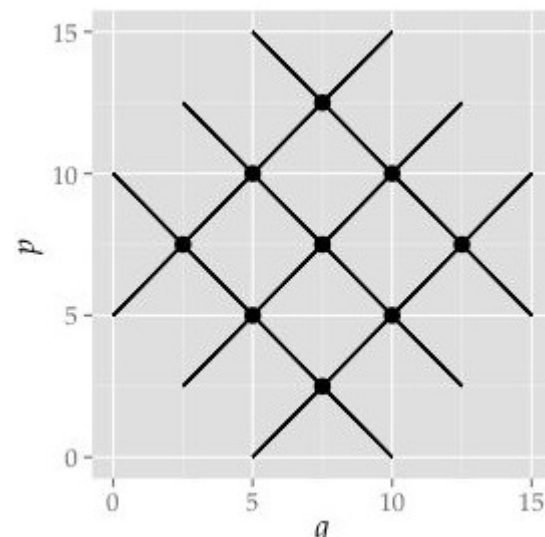
$$\text{Market clears : } Q_t^d = Q_t^s = Q_t$$

$$P_t = \frac{\gamma - \alpha}{\beta - \delta} + \frac{\epsilon_t^s - \epsilon_t^d}{\beta - \delta}$$

我们发现，均衡时的价格 P_t 与两个方程的随机扰动项均相关，存在内生性问题。使用OLS得到的参数（2个），我们不知道得到的究竟是需求还是供给函数（4个参数）。

※ 识别过程：如果我们相信，价格为0时，供给也为0，那么可以限定 $\gamma = 0$ （排除性限制条件exclusion restriction, Blanchard, 1993）。价格达到收入约束水平时，需求为0： $\delta = f(R)$ 。通过这样的办法，使得模型可以被识别。常用的识别办法是让参数矩阵变为三角矩阵。

※ 识别问题需要做出一定的假设，然而，很多经济学论文的假设通常却是不符合现实的，仅仅是为了识别而识别。（见我的博士论文第九章）。识别问题是统计学以外的知识（Romer, 2015）



理想条件的含义4：白噪音

※ $E(\epsilon\epsilon') = \sigma^2 \mathbf{I}_T$ 白噪音（White noise）的意义：当随机扰动项是白噪音的时候，
OLS估计量是一致无偏估计量（证明略）

残差值是白噪音，表明没有其他信息可以再从白噪音中分离出来，
模型得到了很好的拟合。

※ 不是WN的形式：

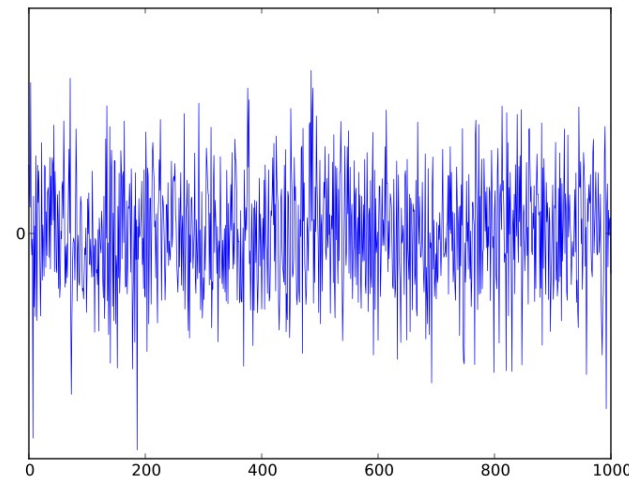
- 1) 主对角线上元素不相等-异方差性（Heteroscedasticity）
- 2) 除主对角线上以外元素不为0-自相关性（autocorrelation）

※ 异方差性（Heteroscedasticity）的后果：

OLS估计量仍然是无偏和渐进一致的，但不再是最小方差，（参数的）方差估计有误，因此t值不可信。

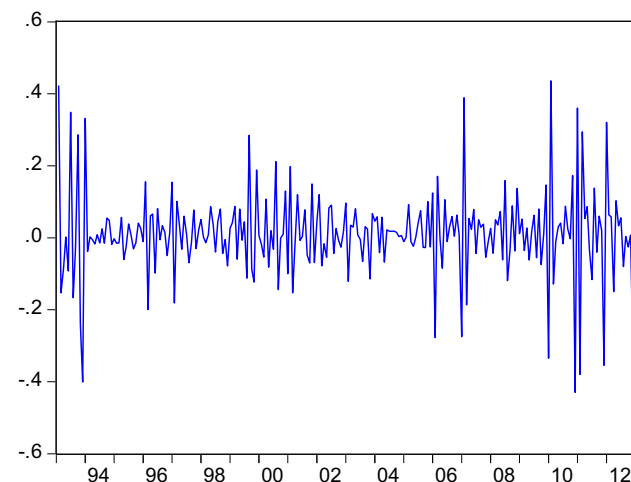
※ 自相关性（autocorrelation）

OLS估计量仍然是无偏和渐进一致的，但不再是最小方差，（参数的）方差估计有误，因此t值不可信。



一个高斯白噪音

GEXPENC



理想条件的含义5：正态性

※ $\epsilon \sim N(0, \sigma^2)$ 随机扰动项服从正态分布

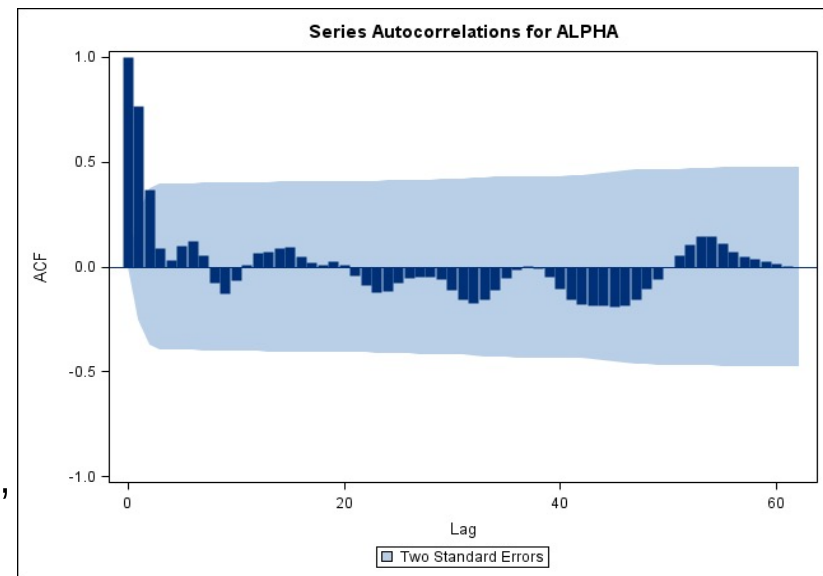
白噪音有许多种，高斯白噪音只是其中一种，表明无穷多相互独立的其他因素叠加。

※ 随机扰动项不是白噪音的可能原因有：











- 1) 时间序列不平稳，存在高阶矩（如经济环境剧烈变化时，波动增强；经济平稳运行时，波动稳定；存在异方差性）
- 2) 某些经济政策具有滞后效应（财政支出、货币政策不会立竿见影，有一定的滞后期，存在自相关性）
- 3) 某些经济变量存在惯性（由俭入奢易，由奢入俭难，消费不太可能一天之内就降下去）

※ 随机扰动项不是高斯分布的后果：OLS估计量不是最优（方差不是最小）

※ 我们可以看到，违背正态性的后果相对不太严重的，因此若残差值是白噪音表明回归模型已经较好的建模，若回归残差值是高斯分布，则是理想的建模。不能达到理想建模，可以退而求其次，至少保证残差值是白噪音，



摘选自我的博士论文第三章：投资占资本比重的自相关函数

Correlogram of ACTIVEPOPUGROWTH						
Date: 05/28/13 Time: 19:28						
Sample: 1952 2012						
Included observations: 59						
Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob	
		1	0.090	0.090	0.5053	0.477
		2	0.068	0.061	0.8012	0.670
		3	-0.008	-0.019	0.8048	0.848
		4	0.012	0.010	0.8140	0.937
		5	0.049	0.050	0.9744	0.965

劳动力人口增长率的自相关函数：摘自我的硕士论文

2

非理想情况的处理办法

理论是灰色
的，而生命
之树长青

——歌德《浮士德》

※经验判断：

- 1) 多重共线性是由变量高度自相关引起的，因此加入或删除一个变量对于回归结果影响很大，如果出现这种状况，可以判定可能存在Multicollinearity
- 2) 变量高度自相关是因为变量接近或包含重复信息，若自变量是接近的变量，则先验的可以认为可能存在Multicollinearity
- 3) 相同或接近的变量在回归分析中，一般不会两个都显著，因此若变量（t检验）不显著，而方程却具有整体显著性（F检验），则可以判定可能存在Multicollinearity
- 4) 同理，若一个变量在简单线性回归中是显著的，而在多元线性回归中不显著，则可以判定可能存在Multicollinearity

※定量判断：

- 1) 计算方差膨胀因子（Variance inflation factor, VIF）
判断标准：mean VIF > 6 或 ind VIF > 10 认为存在Multicollinearity
- 2) Farrar–Glauber test（1967）批评较多，较少使用
- 3) Condition number test: 条件数（CN）大于30，则存在较为严重的Multicollinearity

※多重共线性的判断相对较为主观

※具体情况具体分析

- 1) 删除某个变量。因为多重共线性通常由重复或者相近的变量引起，在不引起模型设定基础性改变时，可以删除相似变量中的一个或者若干个。可能的危险：内生性问题。
- 2) 删除常数项或哑变量。太多的哑变量必然会导致多重共线性。一定要用两个哑变量时，最好删掉常数项。危险：R²可能大幅度降低甚至为负。
- 3) 扩大样本容量。最理想的方法，使得T远大于k。
- 4) 对变量进行转换，例如谨慎同时使用换手率和交易额、总市值三个相关性较高的变量在同一线性回归中。
- 5) 脊回归 (Ridge regression) $\hat{\beta}_{\text{Ridge}} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}'\mathbf{Y}$ 多重共线性的的问题在于矩阵 $\mathbf{X}'\mathbf{X}$ 不是满秩因此不可逆（或者接近奇异矩阵），在主对角线上人为加上一个常数，可以解决这个问题。 λ 的选择：使用扫描的办法选择最优值（可以根据信息准则或R²判断）。脊回归是贝叶斯方法的特殊形式（后面细讲）。
- 6) 主成分回归（principal component regression）或偏最小二乘回归（partial least squares regression）（不展开讲）
- 7) 压缩的哑变量（可以同时处理残差的自相关性，见我的博士论文第五章，7-steps-Fisher检验）

※经验判断：三种引起内生性的原因：遗漏变量、测量误差和互为因果

※统计检验：Hausman-Wu test

思想：存在内生性问题的估计量是有偏的估计，假设得到的估计量为 b_1 （有效的），已经有另外一个一致估计量 b_0 （方差更大，效率低），将 b_1 与 b_0 作比较，若 b_1 收敛于 b_0 ，则 b_1 也是一致估计量，反之则存在内生性。

Hausman-Wu 统计量 H 渐进服从卡方分布（证明从略）

Hausman-Wu 检验的本质也是工具变量法

其他检验：Wald block exogenous test (动态模型)

※工具变量：工具变量 z 与内生性变量 x 相关，但与随机扰动项 ϵ 不相关。

※内生性问题的解决办法：具体问题具体分析

遗漏变量：添加遗漏变量

测量误差：找出可能的测量误差，减少误差。

互为因果：寻找工具变量（Instrumental variables）。广义的讲，工具变量法适用于所有的内生性问题的处理，但工具变量的寻找非常困难。

一个好的IV=一篇AER

※Two-stage least squares (2LS, Heckman, 1997) :假定变量 x 具有内生性, 现找到一个工具变量 z ,与 x 相关且严格外生, 与原模型随机扰动项 ϵ 不相关。那么我们可以用两步最小二乘法得到无偏一致估计量 $\hat{\beta}$ 。
原模型：

$$Y = X\beta + \epsilon$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

第一步：

$$X = Z\delta + \text{errors}$$

$$\hat{\delta} = (Z^T Z)^{-1} Z^T X,$$

得到 X 的预测值：

$$\hat{X} = Z\hat{\delta}$$

带入原模型, 用 \hat{X} 替换 X ：

$$Y = \hat{X}\beta + \text{noise}$$

得到2SLS estimator：

$$\beta_{2SLS} = (X^T P_Z X)^{-1} X^T P_Z Y$$

其中：

$$P_Z = Z(Z'Z)^{-1}Z'$$

※当工具变量个数大于内生性变量个数时, 称为“过度识别” (over-identification), 过度识别是我们想要的, 这个时候2SLS 就是广义矩估计。

※ 面板回归 (Panel regression) : 过去可以影响未来, 然而未来无法影响过去, 因此, 某个变量的滞后阶可能与当期的误差有关, 与未来这个变量的取值有关, 但是与未来的误差无关, 是工具变量的一个很好的选择。(后面详细介绍面板回归)。

※ 双重差分法 (Difference in difference) : 学过高中生物的同学知道, 要研究某个变量的影响, 需要设置实验组和对照组, 保证其他条件一致, 控制某个变量变化, 考察它对结果的影响。DID的思想类似, 出现了一次外部冲击, 这次冲击影响了一部分样本, 对另一部分样本则无影响, 将受到冲击的部分设为实验组, 未受冲击的设为对照组, 冲击前后两组结果 (已经差分一次) 再差分之后就是冲击的真实影响。Card and Krueger (1994) 最低工资法案:

New Jersey	Pennsylvania	Difference	
February	20.44	23.33	-2.89
November	21.03	21.17	-0.14
Change	0.59	-2.16	2.75

房地产调控政策(限价)有效吗?

	2009	2011	Difference
广州(限价)	16,000	20,000	4,000 ↑
佛山(不限价)	12,000	17,000	5,000 ↑
Difference			-1,000

$$y = \beta_0 + \beta_1 T + \beta_2 S + \beta_3 (T \cdot S) + \varepsilon$$

β_3 就是冲击的真实影响。

借鉴自然学科里的实验设计的方法：

- ※ 断点回归设计（RDD）：Pinotti, Paolo. "Clicking on heaven's door: The effect of immigrant legalization on crime." *American Economic Review* 107.1 (2017): 138-68.
- ※ 倾向得分匹配分析（PSM）：logit模型中再详细介绍
- ※ 合成控制法（SMC）：Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program." *Journal of the American statistical Association* 105.490 (2010): 493-505.
- ※ 安慰剂检验（Placebo test）：Zwick, Eric, and James Mahon. "Tax policy and heterogeneous investment behavior." *American Economic Review* 107.1 (2017): 217-48.
- ※ 学习前沿量化方法的好办法是看AER的论文，方法、数据、代码都是公开的，可以通过复刻学习掌握这种技术。
- ※ 提出新的量化方法的好办法是看Econometrica（以及统计四大天王：Annals of Statistics、Journal of American the Statistical Association、Journal of the Royal Statistical Society、Biometrika）的论文

※ 异方差性的检验：根据引起异方差性的可能原因，如残差的异方差性可能来自自变量或者其平方，设定特定的模型，检验假定模型的整体显著性。

※ White test：Lagrange multiplier (LM) test 统计量服从卡方分布（证明从略）

※ Breusch-Pagan/ Cook-Weisberg test：统计量服从卡方分布（证明从略）

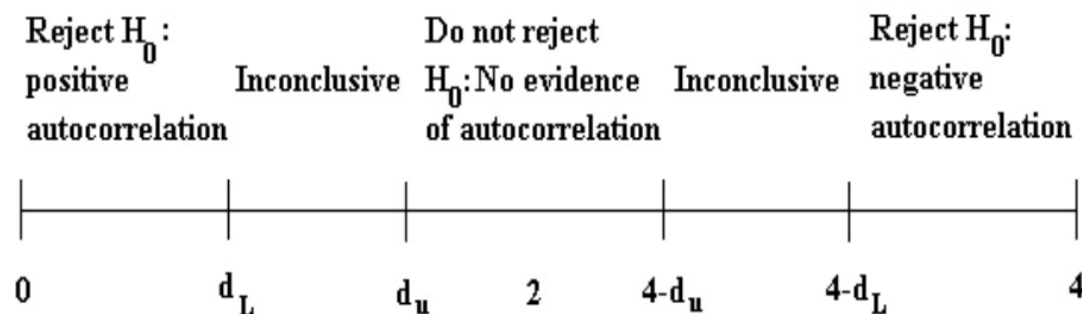
※ 异方差性的解决办法：异方差性的存在，使得参数的方差估计量不再一致，White 提出了一个Robust correction, 使用方差的一致估计量代替参数方差（具体公式从略）。修改只改变参数的方法，不改变参数的值，因此t值有所变化（变小），变量显著性降低。Newey-West correction: 适用于既有自相关又有异方差的情况。

※ 自相关性的检验：思想类似于异方差性，根据引起异方差性的可能原因，设定特定的模型，检验假定模型的整体显著性。

※ Durbin-Watson test:局限性：只能检测一阶自相关（统计量取值范围0-4）

※ Breusch-Godfrey test（P阶）：LM test,服从卡方分布

※ Ljung-Box test（特定阶数）.服从卡方分布



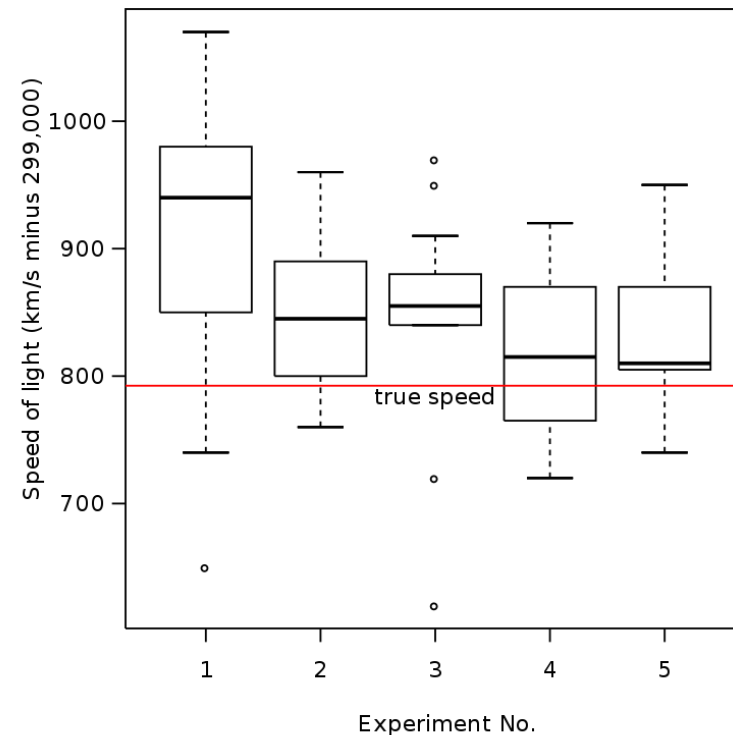
- ※检验：见上一节的检验方法
- ※处理：不满足正态性，结果只是不是最优估计量，大多数情况下无需处理

※ 检验：

- 1) 画图：线箱图 (Box plot)
- 2) Hadi 距离等
- 3) ARMA-12 outlier test

※ 处理办法：具体情况具体分析

- 1) 删除 (Chow, 1993, 2001, QJE和JDC, 删除了三年自然灾害时期)
- 2) 根据离群值的重要程度, 看是否需要设计断点回归、Changes of regimes models 等



- ※ R² 模型的解释力

$R^2 = 1 - ESS/TSS$ 取值范围[0,1] R^2 越大越好

注意：R²在特殊情况下可能为负（删除常数项的回归）

- ※ Fisher 检验：模型整体的显著性

- ※ 信息准则：AIC,SIC,HQ 信息准则越小越好

- ※ 其他：loglikelihood 越大越好

- ※ 参数的稳定性：邹至庄检验Chow test

3

多元线性回归在Stata中的实现



※宏观数据通常是时间序列，涉及时间序列的建模的问题
(下一讲调整为时间序列，需要先理解时间序列，才能再讲面板回归)

※国家统计局微观数据库（需要申请）

<http://microdata.stats.gov.cn/>

※美国数据库（FRED）：<https://fred.stlouisfed.org/>

※AER论文附录数据：<https://www.aeaweb.org/journal>
(通常不是单一方法,不适合侧重某一方法)

※stata自带数据：

输入命令：`help dta_examples`

在lifeexp 这一栏点describe, 可以看到对数据的描述
点use, 导入该数据库

※问题：人均预期寿命与人口增长率、人均GDP、
安全饮用水等是否有关系

The datasets listed here are installed with Stata. You can also see a complete list of the datasets used in the Stata documentation that are available via the Internet. Each manual title is listed as a link that will take you to the list of datasets for that manual.

<code>auto.dta</code>	<code>use</code>	<code>describe</code>
<code>auto2.dta</code>	<code>use</code>	<code>describe</code>
<code>autornrd.dta</code>	<code>use</code>	<code>describe</code>
<code>bplong.dta</code>	<code>use</code>	<code>describe</code>
<code>bpwide.dta</code>	<code>use</code>	<code>describe</code>
<code>cancer.dta</code>	<code>use</code>	<code>describe</code>
<code>census.dta</code>	<code>use</code>	<code>describe</code>
<code>citytemp.dta</code>	<code>use</code>	<code>describe</code>
<code>citytemp4.dta</code>	<code>use</code>	<code>describe</code>
<code>educ99gdp.dta</code>	<code>use</code>	<code>describe</code>
<code>gnp96.dta</code>	<code>use</code>	<code>describe</code>
<code>lifeexp.dta</code>	<code>use</code>	<code>describe</code>
<code>network1.dta</code>	<code>use</code>	<code>describe</code>
<code>network1a.dta</code>	<code>use</code>	<code>describe</code>
<code>nlsw88.dta</code>	<code>use</code>	<code>describe</code>
<code>nlswide1.dta</code>	<code>use</code>	<code>describe</code>
<code>pop2000.dta</code>	<code>use</code>	<code>describe</code>
<code>sandstone.dta</code>	<code>use</code>	<code>describe</code>
<code>sp500.dta</code>	<code>use</code>	<code>describe</code>
<code>surface.dta</code>	<code>use</code>	<code>describe</code>
<code>tsline1.dta</code>	<code>use</code>	<code>describe</code>
<code>tsline2.dta</code>	<code>use</code>	<code>describe</code>
<code>uslifeexp.dta</code>	<code>use</code>	<code>describe</code>
<code>uslifeexp2.dta</code>	<code>use</code>	<code>describe</code>
<code>voter.dta</code>	<code>use</code>	<code>describe</code>
<code>xtline1.dta</code>	<code>use</code>	<code>describe</code>

※人均预期寿命（life expectancy）：反映了人们的生活水平，医疗水平越高、经济文化程度越发达、人均预期寿命越高。

※数据库中的变量：

Region: 分为欧洲/亚洲、北美、南美

Country: 国家名

Popgrowth: 年均增长率

Lexp: 出生时预期寿命

Gnppc: 人均国民收入

Safewater:

※假设：

假设1：经济越发达，人均预期寿命越高

假设2：医疗水平越高，婴儿死亡率越低，人口增长率越高（一定时间内），人口预期寿命越长

假设3：饮用水越安全，人们生活水平越健康，人均预期寿命越高

※思考：在这里，region应不应该作为解释变量？

```
. describe
```

```
Contains data from http://www.stata-press.com/data/r15/lifeexp.dta
```

```
obs:           68           Life expectancy, 1998
vars:           6           26 Mar 2016 09:40
size:          2,652        (_dta has notes)
```

variable name	storage type	display format	value label	variable label
region	byte	%12.0g	region	Region
country	str28	%28s		Country
popgrowth	float	%9.0g		* Avg. annual % growth
lexp	byte	%9.0g		* Life expectancy at birth
gnppc	float	%9.0g		* GNP per capita
safewater	byte	%9.0g		*
* indicated variables have notes				

```
Sorted by:
```

描述性统计

※按组分的均值图：grmeanby region, summarize(lexp)
欧洲最高，北美其次，南美较低

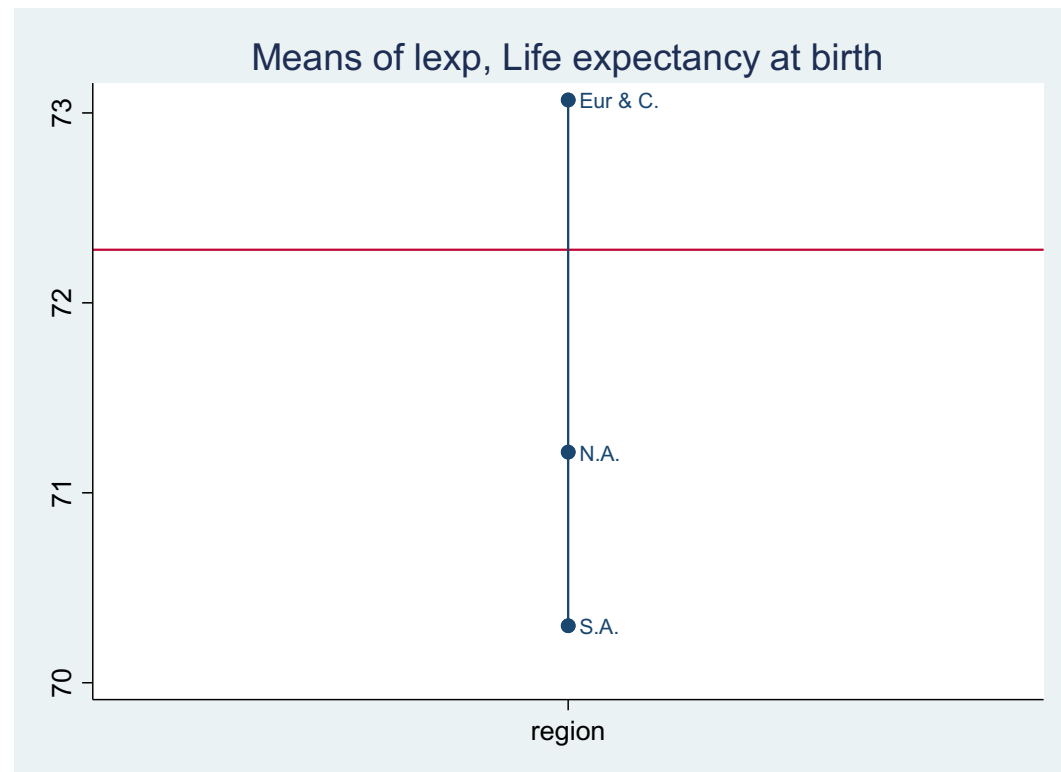
※按组分的统计性描述：by region, sort : summarize
我们可以看到，质量型变量的数字特征没有任何意义
质量型变量的描述型统计在逻辑回归中介绍

※通过协方差矩阵探索变量之间可能的关系
correlate lexp popgrowth gnppc safewater
注：可以先点lexp以便探索

```
. correlate lexp popgrowth gnppc safewater
(obs=37)
```

	lexp	popgro~h	gnppc	safewa~r
lexp	1.0000			
popgrowth	-0.4750	1.0000		
gnppc	0.6948	-0.4902	1.0000	
safewater	0.8177	-0.3935	0.7063	1.0000

假设1和3可能成立，假设2可能是非线性关系
人均国民收入和安全饮用水之间相关性很高，可能存在多重共线性，



```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
region	68	1.5	.7431277	1	3
country	0				
popgrowth	68	.9720588	.9311918	-.5	3
lexp	68	72.27941	4.715315	54	79
gnppc	63	8674.857	10634.68	370	39980
safewater	40	76.1	17.89112	28	100

假设1

※散点图：twoway (scatter lexp gnppc)

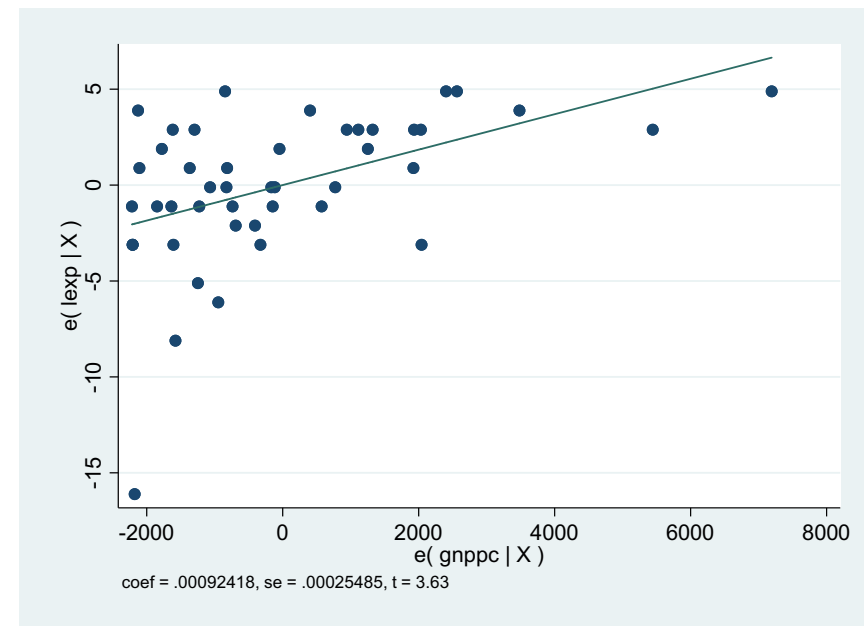
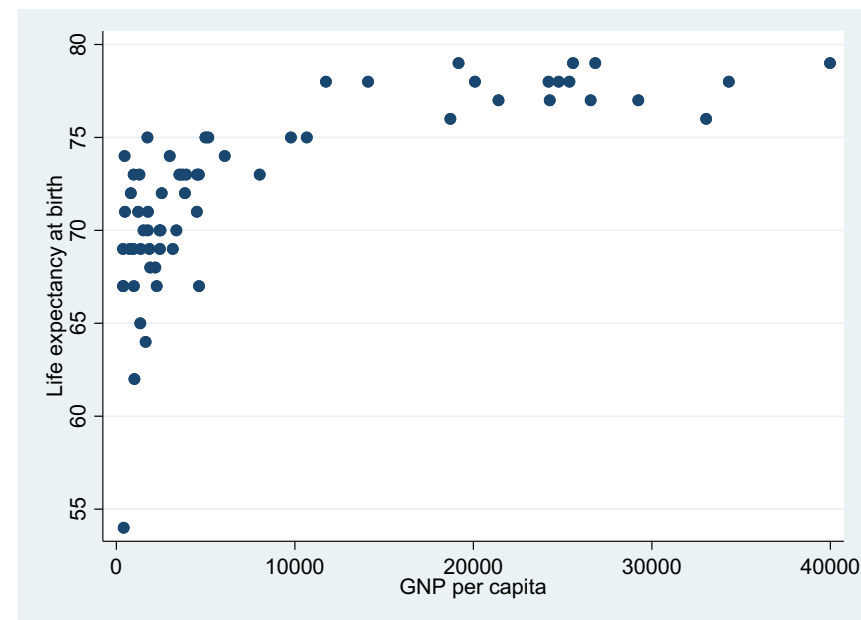
直观印象：人均收入低于10000时，人均预期寿命随收入增长较快
原因：人的寿命增加受到自然属性的限制，收入只是社会属性。

※简单线性回归的验证：regress lexp gnppc if gnppc<10000
avplots

Source	SS	df	MS	Number of obs	=	45
Model	158.424815	1	158.424815	F(1, 43)	=	13.15
Residual	518.01963	43	12.0469681	Prob > F	=	0.0008
Total	676.444444	44	15.3737374	R-squared	=	0.2342
				Adj R-squared	=	0.2164
				Root MSE	=	3.4709

lexp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
gnppc	.0009242	.0002548	3.63	0.001	.0004102 .0014381
_cons	67.72003	.8381317	80.80	0.000	66.02977 69.41028

R2较低，模型解释力较差，单一经济因素不足以解释人均预期寿命的差异



假设2

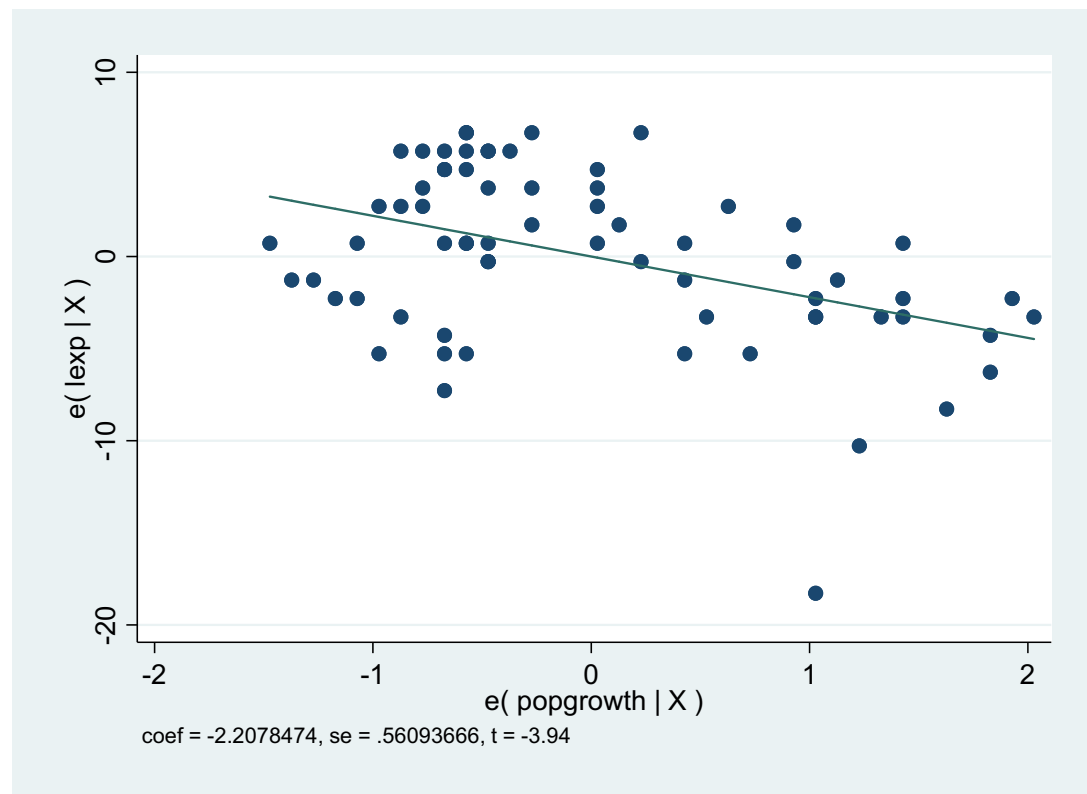
※散点图：regress lexp popgrowth
avplots

```
. regress lexp popgrowth
```

Source	SS	df	MS	Number of obs	=	68
Model	283.198638	1	283.198638	F(1, 66)	=	15.49
Residual	1206.49254	66	18.28019	Prob > F	=	0.0002
				R-squared	=	0.1901
				Adj R-squared	=	0.1778
Total	1489.69118	67	22.2341967	Root MSE	=	4.2755

lexp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
popgrowth	-2.207847	.5609367	-3.94	0.000	-3.327794	-1.087901
_cons	74.42557	.7524218	98.91	0.000	72.92331	75.92783

```
. avplots
```



散点图显示，人口增长率越高，人均预期寿命越低。这可能是因为我们已经处于后工业化时代，富裕的国家生育率较低，因此人口增长率低，如Becker的解释。

同样的，单一的人口因素也无法充分解释预期寿命

一个可能的马克思主义角度的解释：低收入国家资本有机构成更低，对劳动力的需求更高。

假设3

※散点图：regress lexp safewater
avplots

```
. sktest e0
```

Skewness/Kurtosis tests for Normality						
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2 (2)	Prob>chi2	joint
e0	40	0.8002	0.3326	1.05	0.5907	

```
. swilk e0
```

Shapiro-Wilk W test for normal data					
Variable	Obs	W	V	z	Prob>z
e0	40	0.95182	1.905	1.356	0.08758

```
. sfrancia e0
```

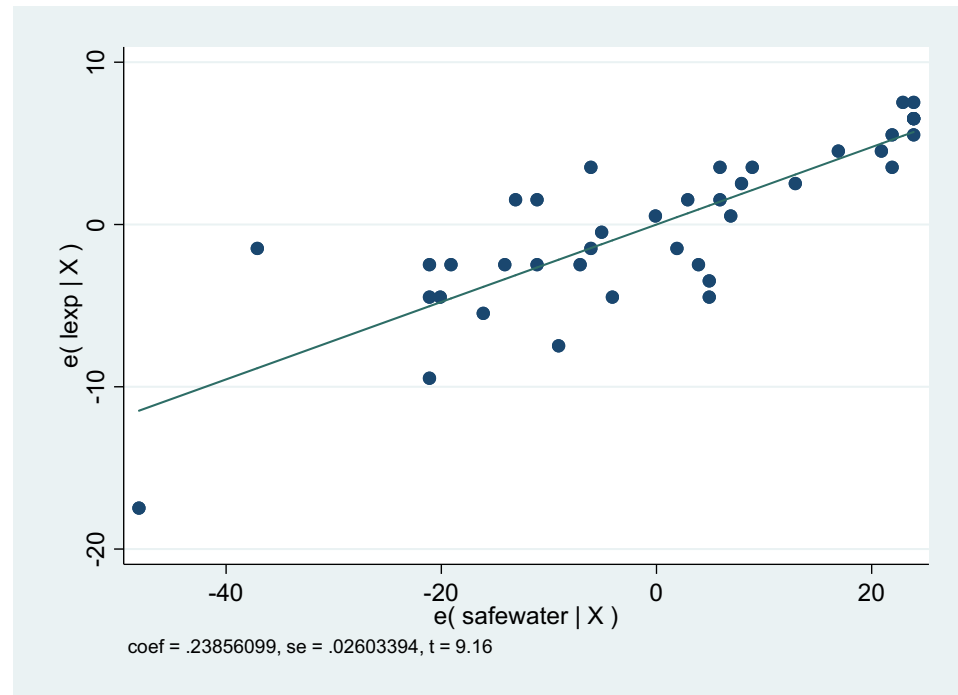
Shapiro-Francia W' test for normal data					
Variable	Obs	W'	V'	z	Prob>z
e0	40	0.94926	2.224	1.490	0.06816

※简单线性回归效果如何？

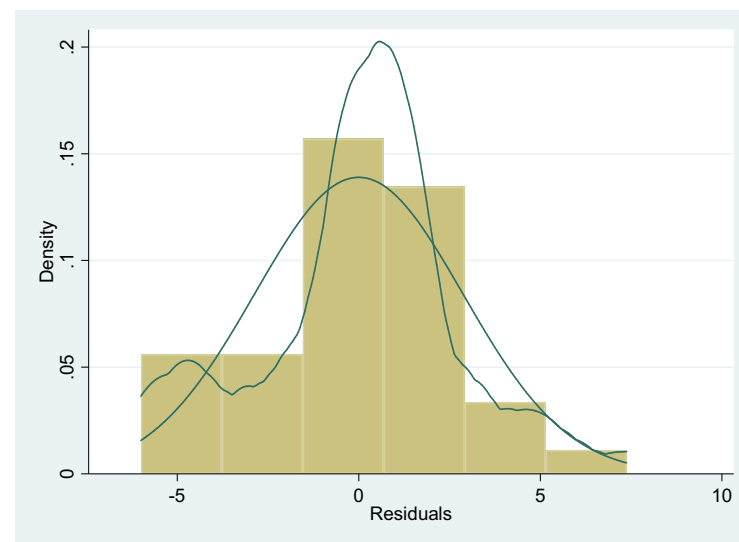
保存残差值：predict e0, residual

检验正态性：相当好

饮用水安全可能是非常重要的因素



这可能是一



※多重共线性：前面的分析表明，人均国民收入和安全饮用水之间相关性很高，可能存在多重共线性

计算方差膨胀因子：estat vif

方差膨胀因子表明多重共线性不严重，这个问题可以在一定程度上容忍。

```
. estat vif
```

Variable	VIF	1/VIF
gnppc	2.23	0.447849
safewater	2.01	0.498202
popgrowth	1.32	0.755252
Mean VIF	1.85	

※ regress lexp popgrowth gnppc safewater

初步结果的若干评论：

- 1) R^2 比单独的简单线性回归都高，这表明三个变量的确都增强了模型的解释力
- 2) 三个变量参数的正负值与描述性统计一致符合预期
- 3) 人均GDP和人口增长率变得不显著，但在简单线性回归中它们都是显著的，这表明存在多重共线性，与VIF结论不一致。

```
. regress lexp popgrowth gnppc safewater
```

Source	SS	df	MS	Number of obs	=	37
Model	682.388693	3	227.462898	F(3, 33)	=	26.97
Residual	278.31401	33	8.43375787	Prob > F	=	0.0000
				R-squared	=	0.7103
				Adj R-squared	=	0.6840
Total	960.702703	36	26.6861862	Root MSE	=	2.9041

lexp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
popgrowth	-.7639202	.6008591	-1.27	0.212	-1.986377	.4585369
gnppc	.0000883	.0000704	1.26	0.218	-.0000548	.0002315
safewater	.184967	.0383851	4.82	0.000	.106872	.263062
_cons	57.80826	2.880981	20.07	0.000	51.94686	63.66966

※解决办法：

删去常数项（结果变化很大，表明多重共线性严重且删除常数项办法不好）

删去人均收入？

lexp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
popgrowth	4.362882	1.946607	2.24	0.032	.4069011	8.318864
gnppc	-.0004314	.0002342	-1.84	0.074	-.0009074	.0000445
safewater	.8869286	.0565468	15.68	0.000	.7720117	1.001845

多元线性回归2

※组间差异过大，放在一起可能不太好，分组回归。

```
by region, sort : regress lexp  
popgrowth gnppc safewater
```

※结果：组间显著性差异较大。
对于欧洲和亚洲国家（样本中主要是欧洲）来说：人口增长率不显著，这可能是因为欧洲人口增长率都较低。
北美：人均收入不显著，可能因为北美国家收入都较高
南美：人口增长率不显著，可能因为人口增长率都较高；安全饮用水不显著，可能因为数据缺失。

※结论：分组回归结果更好，但各组回归样本容量太小，降低了统计可信性。

解决办法：寻找更大样本的数据

lexp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
popgrowth	-.6322132	.6186807	-1.02	0.327	-1.980203	.7157762
gnppc	.0001658	.000062	2.67	0.020	.0000307	.0003009
safewater	.1095346	.0448209	2.44	0.031	.0118782	.207191
_cons	62.61165	3.077808	20.34	0.000	55.90568	69.31762

lexp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
popgrowth	-3.246141	1.517071	-2.14	0.070	-6.833444	.3411617
gnppc	-.0001516	.0002505	-0.61	0.564	-.0007439	.0004407
safewater	.3098615	.0651519	4.76	0.002	.1558019	.4639212
_cons	54.46161	5.103504	10.67	0.000	42.39374	66.52948

lexp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
popgrowth	5.005481	3.0607	1.64	0.153	-2.483782	12.49474
gnppc	.0016928	.0006884	2.46	0.049	8.34e-06	.0033773
safewater	.1631062	.0860207	1.90	0.107	-.0473789	.3735913
_cons	42.85596	12.3444	3.47	0.013	12.6503	73.06162

多元线性回归3

※人均收入与其他变量相关性都较高，可以删去，而前面分析表明人均收入和预期寿命之间的关系仅在一定范围内才具有较强的线性相关性

※因此，要么我们使用非线性模型：NLS，要么删除人均收入，要么加入一个哑变量D（收入低于10000为1，高于10000为0），新的回归变量为D*gnppc。
第三种做法多重共线性仍然存在

创建dummy:

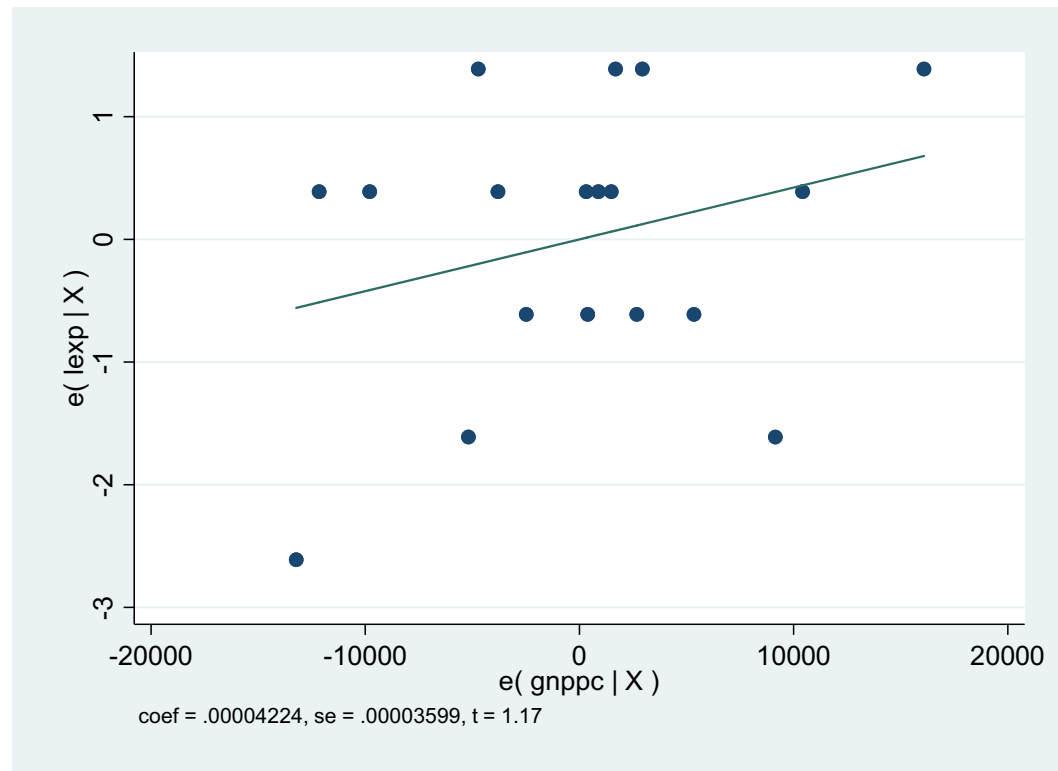
generate d = 0 /*第一步：所有的都等于0*/

replace d = 1 if gnppc<10000 /*第二步：人均收入小于10000的为1*/

replace d = . if missing(gnppc)/*第三步：缺失值仍然是缺失值*/

gen gnppc_d=gnppc*d)/*生成一个新的变量*/

※结果：与原回归相比，人口增长率的显著性得到了改善



	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
popgrowth	-1.007996	.5802893	-1.74	0.092	-2.188604	.1726115
gnppc_d	.0000763	.0002106	0.36	0.720	-.0003521	.0005046
safewater	.2154835	.0301238	7.15	0.000	.1541962	.2767709
_cons	56.34161	2.761789	20.40	0.000	50.72271	61.96052

多元线性回归4

※删掉人均收入可能是较好的办法

R2没有显著变化，解释力较高

所有变量均显著

残差值的检验：

predict e, residual

正态性检验：

sktest e

swilk e

sfrancia e

通过正态性检验，拟合得相当好

但异方差性检验有瑕疵：estat hettest, iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: fitted values of lexp

chi2(1) = 13.19

Prob > chi2 = 0.0003

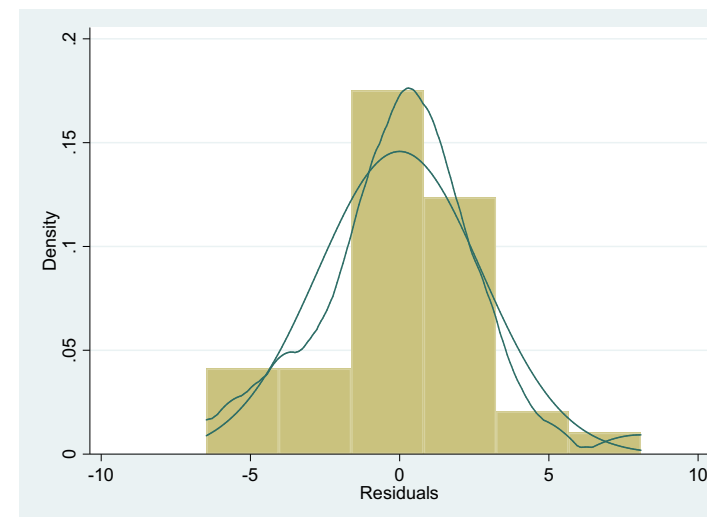
使用Robust regression.

美中不足：样本容量偏小（40*3）。

```
. regress lexp popgrowth safewater
```

Source	SS	df	MS	Number of obs	=	40
Model	739.857403	2	369.928701	F(2, 37)	=	46.86
Residual	292.117597	37	7.89507019	Prob > F	=	0.0000
				R-squared	=	0.7169
				Adj R-squared	=	0.7016
Total	1031.975	39	26.4608974	Root MSE	=	2.8098

lexp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
popgrowth	-1.035783	.5367615	-1.93	0.061	-2.123366	.0517986
safewater	.2155804	.0278255	7.75	0.000	.1592005	.2719602
_cons	56.51166	2.568017	22.01	0.000	51.30837	61.71496



※内生性问题：

测量误差：不清楚

互为因果：寿命的提升会不会引起用水更加安全和人口增长率的变化？似乎没有直接的feedback effect

遗漏变量：必然存在遗漏变量

局限于样本数据，没有好的办法设计工具变量，五一放假前复刻Acemoglu（1991），学习如何构建工具变量。

※样本偏差：

1) 全世界200多个国家和地区，样本只有68个，本身是不是cherry pick?

古巴等特例

2) 特定时间点的数据，是不是只反映了人类社会某个特定历史阶段的经济社会规律？

※样本容量：样本容量偏小，统计可靠性不高，组内回归样本容量更小。

※理论基础：纯统计模型缺乏理论基础，究竟哪些变量该放在回归，哪些变量又该排除，纯统计模型无法回答这样的问题，而统计关系又是敏感易变的（多重共线性）。建模应当给予经济学理论而非纯统计模型（Lucas, 1974）。大数据方法受到的批评。

4

参考文献

站在巨人的肩膀上

"If I have seen further
it is by standing on
the shoulders of
Giants. "

by Isaac Newton in
1675

见网络学堂附件

下节课见

马克思主义学院

龙治铭



清华大学
Tsinghua University