计量经济学

崔文昊

经济管理学院
北京航空航天大学

课程简要说明

课程名称: 计量经济学

- 学分/学时: 3 学分/48 学时
- 课程性质: 教育部规定核心课程

主讲教师: 崔文昊

- 10676@buaa.edu.cn
- 新主楼 A1018

教材及参考书

教材： **以课堂讲义为主**（《计量经济学（第 5 版)》用于预复习，注意与课堂讲义内容不完全重合)

主要参考教材：

- 《计量经济学（第 5 版)》，李子奈、潘文卿，高等教育出版社，2020
- 《计量经济学学习指南与练习》，潘文卿、李子奈，高等教育出版社，2021
- Econometrics by Badi H. Baltagi
- Econometrics by Fumio Hayashi

课程内容

- 绪论
- 一元回归模型
- 多元回归模型
- 大样本理论
- 非线性模型
- 多元线性模型拓展：异方差性，内生性
- 时间序列
- 面板数据
- 多元线性模型拓展：序列相关性
- 联立方程计量经济模型 *

成绩

课程成绩

- 课堂表现：15 分
- 平时作业：25 分
- 期末考核：60 分

绪论

计量经济学 (Econometrics)

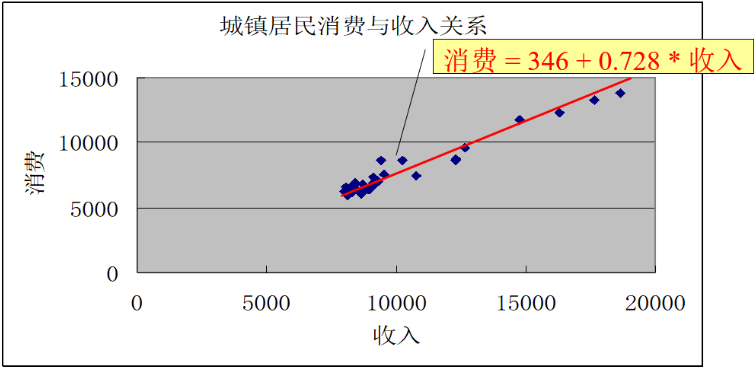
- 经济学的一个分支学科
- 以揭示经济活动中客观存在的数量关系为内容
- 在经济学科中居于最重要的地位
- 把经济理论、数学和统计推断作为工具，应用于经济现象分析

计量经济学的定义

定义：经济理论、统计学和数学的结合

- 弗里希："经验表明，统计学、经济理论和数学这三者对于真正了解现代经济生活的数量关系来说，都是必要的，但本身并非是充分条件。三者结合起来，就是力量，这种结合便构成了计量经济学。"
- 萨缪尔森："计量经济学可以定义为实际经济现象的数量分析。这种分析基于理论与观测的并行发展，而理论与观测，又是通过适当的推断方法得以联系。"
- 戈登伯格："计量经济学可以定义为这样的社会科学：它把经济理论、数学和统计推断作为工具，应用于经济现象分析。"

简单的例子



- 数据来自统计学，线性关系来自经济学，数量公式来自数学
- 三者结合，建立了计量经济学模型

计量经济学体系

初、中、高级计量经济学

- 初级以计量经济学的数理统计学基础知识和经典的线性单方程模型理论与方法为主要内容
- 中级以用矩阵描述的经典线性单方程模型理论与方法、经典线性联立方程模型理论与方法，以及传统的应用模型为主要内容
- 高级以非经典的、现代的计量经济学模型理论、方法与应用为主要内容

本课程定位于中级水平上，适当引入高级的内容

计量经济学体系

理论计量经济学和应用计量经济学

- 理论计量经济学是以介绍、研究计量经济学的理论与方法为主要内容，侧重于理论与方法的数学证明与推导，与数理统计联系极为密切。除介绍计量经济模型的数学理论基础、普遍应用的计量经济模型的参数估计方法与检验方法外，还研究特殊模型的估计方法与检验方法
- 应用计量经济学则以建立与应用计量经济学模型为主要内容，强调应用模型的经济学和经济统计学基础，侧重于建立与应用模型过程中实际问题的处理

本课程是二者的结合，偏侧重于理论

经典计量经济学和非经典计量经济学

经典计量经济学

- 一般指 20 世纪 70 年代以前发展并广泛应用的计量经济学
- 包括: 单方程模型 (Single Equation Model)，联立方程模型 (Simultaneous Equations Model)
- 以线性模型为主要形式

非经典计量经济学

- 一般指 20 世纪 70 年代末以来发展的计量经济学理论、方法及应用模型，也称为现代计量经济学
- 包括: 微观计量经济学 (Microeconometrics)，非参数计量经济学 (Nonparametric Econometrics), 时间序列计量经济学 (Time-Series Econometrics) 等

计量经济学体系

本课程以经典计量经济学为主

- 适当引入一些简单的、应用较多的现代计量经济学理论方法
- 从理论方法角度，经典计量经济学理论方法是非经典计量经济学理论方法的基础
- 从应用角度，经典计量经济学模型仍然是目前应用最为普遍的计量经济学模型

建立计量经济学模型的步骤和要点

理论模型的建立 (4 项任务)

$$y_i = g(X_i, \beta) + u_i, \quad i = 1, 2, \ldots, N$$

- 确定模型包含的变量 (y, X)
  - y 为被解释变量 (explained or dependent variable)
  - X 为解释变量 (explanatory or independent variables)
  - 不遗漏显著的变量
  - 考虑数据的可得性
  - 考虑入选变量之间的关系 (不能完全线性相关)
- 确定模型的数学形式 (依据经济学理论)
  - 利用经济行为理论和数理经济学的成果
  - 根据样本数据作出的变量关系图
- 确定随机扰动项 u 的概率分布特性
- 拟定模型中待估计参数 β 的理论期望值区间 (比如供给曲线斜率为正，需求曲线斜率为负)

样本数据的收集

几类常用的样本数据

- 截面数据 (Cross-sectional Data)
- 时间序列数据 (Time-series Data)
- 面板数据 (Panel Data)

截面数据 (Cross-sectional Data)

- 本课程主要基于截面数据
- 在同一时间（时期或时点）截面上反映一个总体的一批（或全部）个体的同一特征变量的观测值
- 比如某一特定年份的普查数据
- 也称横截面数据、静态数据

时间序列数据 (Time-series Data)

- 本课程只对时间序列数据进行简单的介绍
- 是一批按照时间先后排列的统计数据
- 比如国内生产毛额 (GDP)、消费者物价指数 (CPI)、加权股价指数、利率、汇率等等都是时间序列
- 时间序列的时间间隔可以是分秒 (如高频金融数据)，可以是日、周、月、季度、年、甚至更大的时间单位

面板数据 (Panel Data)

- 本课程只对面板数据数据进行简单的介绍 (处理方法和截面数据以及时间序列数据类似)
- 指在时间序列上取多个截面，在这些截面上同时选取样本观测值所构成的样本数据
- 反映了空间和时间两个维度的经验信息
- 例如我国 31 个省市自治区 2000-2021 年的地区国内生产总值

建立计量经济学模型的步骤和要点

数据质量

- 完整性 (遗失数据)
- 准确性 (测量误差并且符合模型所需要)
- 可比性 (不同时点价格与统计口径)
- 一致性 (样本与母体, 比如不能用企业数据替代行业数据)

建立计量经济学模型的步骤和要点

模型参数的估计

- 模型参数估计方法
  - 最小二乘法 (Least Squares Estimation)
  - 最大似然估计 (Maximum Likelihood Estimation)
  - 矩估计 (Method of Moments)
- 关于应用软件的使用
  - 课堂教学结合 Matlab
  - https://s.buaa.edu.cn/
  - STATA, SAS, SPSS,R, Matlab, Python: 能够熟练使用一种

建立计量经济学模型的步骤和要点

模型的检验

- 经济意义检验
  - 根据经济学理论拟定的符号、大小、关系，对参数估计结果的可靠性进行判断
  - 例如价格上升，需求下降 (对应的系数为负)
- 统计检验
  - 拟合优度检验
  - 显著性检验
- 计量经济学检验
  - 异方差检验
  - 序列相关检验
- 稳健性检验

建立计量经济学模型的步骤和要点

计量经济学模型成功的三要素

- 理论 (建模)
- 数据 (采样)
- 方法 (估计 + 检验)

计量经济学模型的应用

1. 结构分析

- 是对经济现象中“变量之间相互关系”的研究
- 计量经济学模型的功能，是揭示经济现象中变量之间的相互关系
- 前提是模型设定和统计推断都是正确的
- 主要关注点是边际效应 $\frac{\partial E[y]}{\partial X_i}$
- 是其他功能的基础

计量经济学模型的应用

2. 经济预测

- 计量经济学模型，是以模拟历史、从已发生的经济活动中找出变化规律为主要技术手段
- 经济预测，不应该成为计量经济学模型的主要应用领域
- 对于非稳定发展的经济过程，对缺乏规范行为理论的经济活动，计量经济学模型预测功能失效

计量经济学模型的应用

3. 政策评价

- 经济政策不能实验，计量经济学模型“经济政策实验室”的功能所产生的效用是巨大的
- 政策评价，应该成为计量经济学模型的主要应用领域

计量经济学模型的应用

4. 理论检验与发展

- 实践，是检验真理的唯一标准
- 计量经济学模型，提供了一种检验经济理论的好方法
- 对理论假设的检验，可发现和发展理论

统计基础复习

估计方法

- 矩估计
- 最大似然估计

估计值的性质

- 无偏性
- 有效性
- 一致性
- 充分性

练习题 1

我们考虑从一个指数分布 $(X \sim exp(\theta), f(X) = \frac{1}{\theta}e^{-\frac{X}{\theta}})$ 之中进行独立同分布的抽样，并将观测值记作 X_i,

- $\bar{X}$ 是否是 θ 的充分统计量
- 计算矩估计和最大似然估计量 $\hat{\theta}$
- 计算 $\hat{\theta}$ 的期望和方差
- $\hat{\theta}$ 是否具有有效性

假设检验

- H_0 vs H_1
- 假设检验的两类错误
- Neyman-Pearson Lemma
- 置信区间

练习题 2

我们考虑从一个指数分布 ($X \sim exp(\theta), f(X) = \frac{1}{\theta} e^{-\frac{X}{\theta}}$) 之中进行独立同分布的抽样，并将观测值记作 X_i,

- 基于 Neyman-Pearson Lemma 计算 uniformly most powerful critical region 给定 $\alpha \leq 0.05$, 用来检验 $H_0 : \theta = 2$ vs $H_1 : \theta = 1$， 其中 $N = 10$
- 假设我们对于估计 $\beta = \theta^2$ 感兴趣 ($Var(X_i) = \beta$), 我们可以采用矩估计得到 $\hat{\beta} = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2$。$\hat{\beta}$ 是否具有有效性

一元线性回归模型 (Simple Linear Regression Model)

变量间的关系

- 确定性关系或函数关系:
  - 研究的是确定性现象/非随机变量间的关系
  - 比如，圆面积 $= f(\pi, r) = \pi r^2$
- 统计依赖或相关关系:
  - 研究的是非确定性现象/随机变量间的关系。
  - 比如，农作物产量 $= f($气温，降雨量，阳光，施肥量，$\ldots)$
  - 主要是通过相关分析 (correlation analysis) 或回归分析 (regression analysis) 来考察

一元线性回归模型 (Simple Linear Regression Model)

回归分析，仅对存在因果关系而言

- 不存在线性相关，并不意味着不相关
- 存在相关关系，并不一定存在因果关系
  - 相关分析对称地对待任何（两个）变量，两个变量都被看作是随机的
  - 回归分析对变量的处理方法存在不对称性，即区分应变量 (被解释变量 y，果) 和自变量 (解释变量 X，因)，前者是随机变量，后者不一定是

一元线性回归模型 (Simple Linear Regression Model)

回归分析的基本概念

- 回归分析 (regression analysis)，是研究一个变量关于另一个（些）变量的具体依赖关系的计算方法和理论
- 其目的在于通过后者的已知或设定值，去估计和（或）预测前者的（总体）均值
- 两类变量
  - y: 被解释变量或应变量
  - X: 解释变量或自变量

回归分析的主要内容

- 根据样本观察值对经济计量模型参数进行估计，求得回归方程
- 对回归方程、参数估计值进行显著性检验
- 利用回归方程进行分析、评价及预测

一元线性回归模型 (Simple Linear Regression Model)

总体回归函数

- 总体回归函数 $E[y_i|X_i] = f(X_i)$
- 说明被解释变量 y 的平均状态（总体条件期望）、随解释变量 X 变化的规律
- 可是线性或非线性的 (在本节我们仅考虑最简单的情形 $E[y_i|X_i] = \alpha + \beta X_i$)
- α,β 是未知参数，称为回归系数 (regression coefficients)

随机扰动项

- 观察值偏离它的期望值的“离差”，是一个不可观测的随机变量
- $u_i = y_i - E[y_i|X_i]$
- 表明被解释变量除了受解释变量的系统性影响外，还受其他因素的随机性影响

一元线性回归模型

$$y_i = \alpha + \beta X_i + u_i, \quad i = 1, 2, \ldots, N$$

u_i 主要包括下列因素

- 在解释变量中被忽略因素的影响
- 变量观测值的观测误差的影响
- 模型关系的设定误差的影响
- 其它随机因素的影响

一元线性回归模型 (Simple Linear Regression Model)

示例 2.1.1

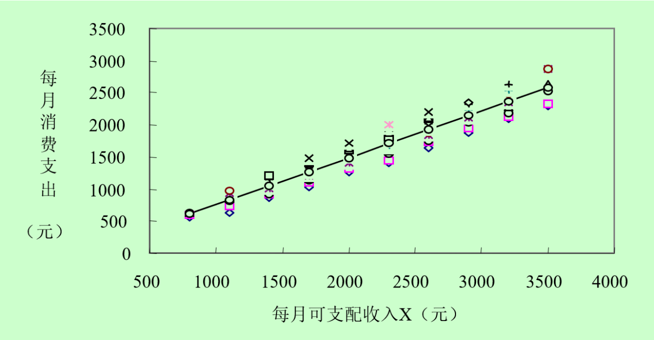
- 一个假想的社区有 99 户家庭组成，欲研究该社区每月家庭消费支出 Y 与每月家庭可支配收入 X 的关系。即如果知道了家庭的月收入，能否预测该社区家庭的平均月消费支出水平。
- 为达到此目的，将该 99 户家庭划分为组内收入差不多的 10 组，以分析每一收入组的家庭消费支出

一元线性回归模型 (Simple Linear Regression Model)

表 2.1.1　某社区家庭每月收入与消费支出统计表

	每月家庭可支配收入X（元）									
	800	1100	1400	1700	2000	2300	2600	2900	3200	3500
每月家庭消费支出Y（元）	561	638	869	1023	1254	1408	1650	1969	2090	2299
	594	748	913	1100	1309	1452	1738	1991	2134	2321
	627	814	924	1144	1364	1551	1749	2046	2178	2530
	638	847	979	1155	1397	1595	1804	2068	2266	2629
		935	1012	1210	1408	1650	1848	2101	2354	2860
		968	1045	1243	1474	1672	1881	2189	2486	2871
			1078	1254	1496	1683	1925	2233	2552	
			1122	1298	1496	1716	1969	2244	2585	
			1155	1331	1562	1749	2013	2299	2640	
			1188	1364	1573	1771	2035	2310		
			1210	1408	1606	1804	2101			
				1430	1650	1870	2112			
				1485	1716	1947	2200			
						2002				
共计	2420	4950	11495	16445	19305	23870	25025	21450	21285	15510

一元线性回归模型 (Simple Linear Regression Model)



- 随着收入的增加，消费“平均地说”也在增加，且 y 的条件均值、均落在一根正斜率的直线上
- 将居民消费支出看成是其可支配收入的线性函数 $E[y_i|X_i] = \alpha + \beta X_i$

一元线性回归模型的参数估计

- 估计方法
  - 参数的普通最小二乘估计
  - 参数估计的最大似然法
  - 参数估计的矩估计
- 估计量的性质
  - 无偏性
  - 有效性
  - 高斯-马尔可夫定理 (Gauss-Markov Theorem)，最佳线性无偏估计量 (best liner unbiased estimator, BLUE)

多元线性回归模型 (Multiple Linear Regression Model)

多元线性回归模型

- 总体回归函数 $E[y_i|X_i] = \alpha + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$
- $u_i = y_i - E[y_i|X_i]$
- $y_i = \alpha + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i,\ i = 1, 2, \ldots, N$
  - K 为解释变量的数目 (注意：本课程之中默认包含截距项)
  - β_j 称为回归参数 (regression coefficient)

总体回归函数

- 描述在给定解释变量 X_{ji} 条件下被解释变量 y_i 的条件均值
- β_j 表示在其他解释变量保持不变的情况下,X_{ji} 每变化 1 个单位时，y_i 的均值 $E[y_i]$ 的变化
- β_j 给出了 X_{ji} 的单位变化对 y_i 均值的 “净”(不含其他变量) 影响

多元线性回归模型 (Multiple Linear Regression Model)

模型的基本假设

- 模型设定正确假设 (线性模型假设)
- 严格外生性假设 ($E[u_i|X_{2i}, X_{3i}, \ldots, X_{Ki}] = 0$)
- 无完全共线性假设
- 同方差假设 ($var(u) = \sigma^2 I_N$)
- 正态分布假设 $u \sim N(0, \sigma^2 I_N)$

多元线性回归模型的参数估计

- 估计方法
  - 参数的普通最小二乘估计
  - 参数估计的最大似然法
  - 参数估计的矩估计
- 估计量的性质
  - 无偏性
  - 有效性
  - 高斯-马尔可夫定理 (Gauss-Markov Theorem), 最佳线性无偏估计量 (best liner unbiased estimator, BLUE)

多元线性回归模型 (Multiple Linear Regression Model)

示例

- 地区城镇居民消费模型
  - y_i: 地区城镇居民人均消费
  - X_{2i}: 地区城镇居民人均可支配收入
  - X_{3i}: 前一年地区城镇居民人均消费
  - 样本：2006 年，31 个地区

多元线性回归模型 (Multiple Linear Regression Model)

地区	2006年消费支出 Y	2006年可支配收入 X_1	2005年消费支出 X_2	地区	2006年消费支出 Y	2006年可支配收入 X_1	2005年消费支出 X_2
北　京	14825.4	19977.5	13244.2	湖　北	7397.3	9802.7	6736.6
天　津	10548.1	14283.1	9653.3	湖　南	8169.3	10504.7	7505.0
河　北	7343.5	10304.6	6699.7	广　东	12432.2	16015.6	11809.9
山　西	7170.9	10027.7	6342.6	广　西	6792.0	9898.8	7032.8
内蒙古	7666.6	10358.0	6928.6	海　南	7126.8	9395.1	5928.8
辽　宁	7987.5	10369.6	7369.3	重　庆	9398.7	11569.7	8623.3
吉　林	7352.6	9775.1	6794.7	四　川	7524.8	9350.1	6891.3
黑龙江	6655.4	9182.3	6178.0	贵　州	6848.4	9116.6	6159.3
上　海	14761.8	20667.9	13773.4	云　南	7379.8	10069.9	6996.9
江　苏	9628.6	14084.3	8621.8	西　藏	6192.6	8941.1	8617.1
浙　江	13348.5	18265.1	12253.7	陕　西	7553.3	9267.7	6656.5
安　徽	7294.7	9771.1	6367.7	甘　肃	6974.2	8920.6	6529.2
福　建	9807.7	13753.3	8794.4	青　海	6530.1	9000.4	6245.3
江　西	6645.5	9551.1	6109.4	宁　夏	7205.6	9177.3	6404.3
山　东	8468.4	12192.2	7457.3	新　疆	6730.0	8871.3	6207.5
河　南	6685.2	9810.3	6038.0				

多元线性回归模型 (Multiple Linear Regression Model)

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

$$\hat{y}_i = 143.3265 + 0.555644 * X_{2i} + 0.250085 * X_{3i}$$

$$\hat{\alpha} = 143.3265$$

$$\hat{\beta}_2 = 0.555644$$

$$\hat{\beta}_3 = 0.250085$$

大样本理论 (Asymptotic Theory)

大样本理论

- 大数定律
  - 切比雪夫不等式
  - 估计量的一致性
- 中心极限定理
  - 矩母函数
  - 估计量的渐进分布

假设检验

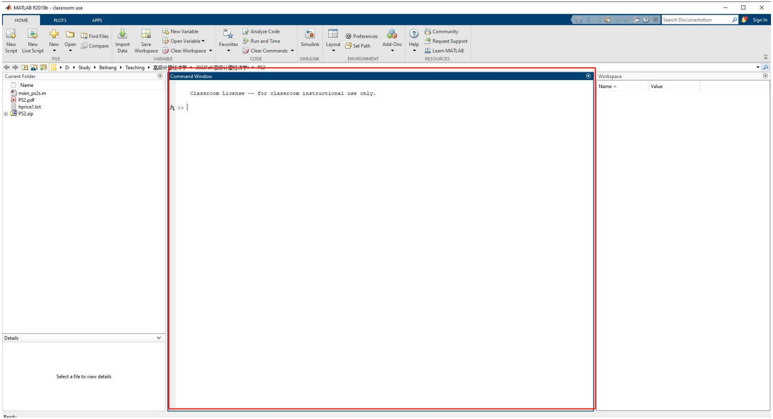
- 有限样本假设检验 (需要假设 $u_i \sim N(0, \sigma^2)$)
  - t 检验
  - F 检验
- 大样本假设检验 (不需要假设 $u_i \sim N(0, \sigma^2)$)
  - t 检验
  - $Wald$ 检验

Matlab 基础

下载地址： https://s.buaa.edu.cn

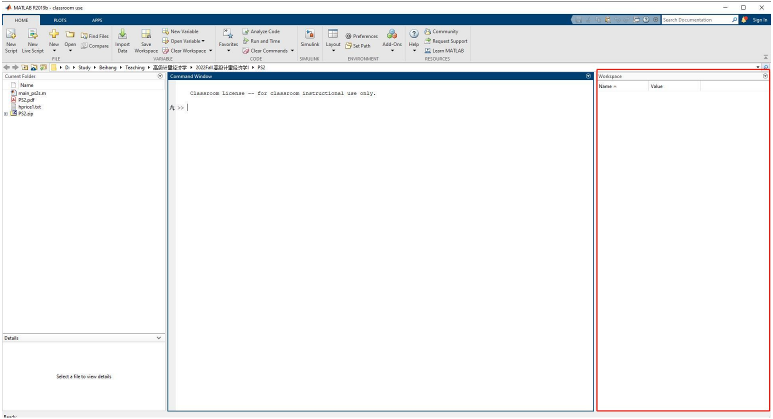
按照步骤进行安装

Matlab 基础



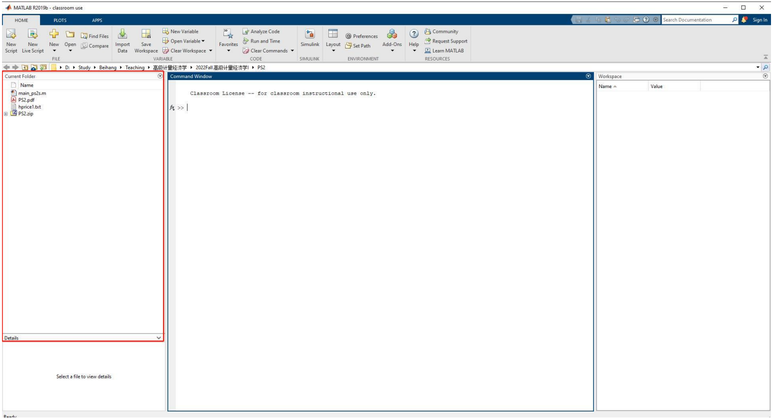
命令窗口：输入命令的地方

Matlab 基础



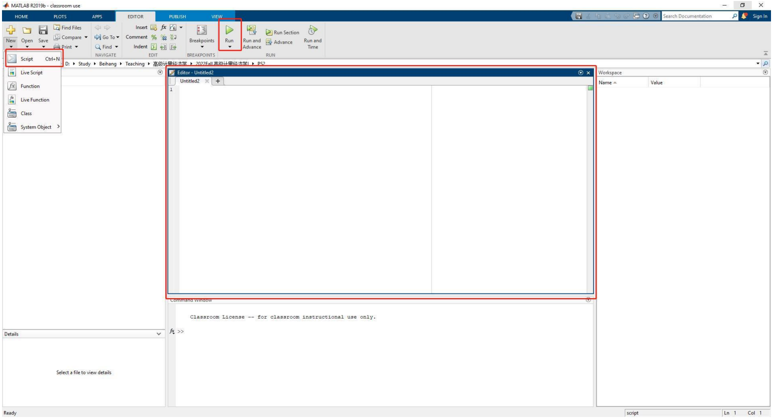
工作空间：查看结果的地方

Matlab 基础



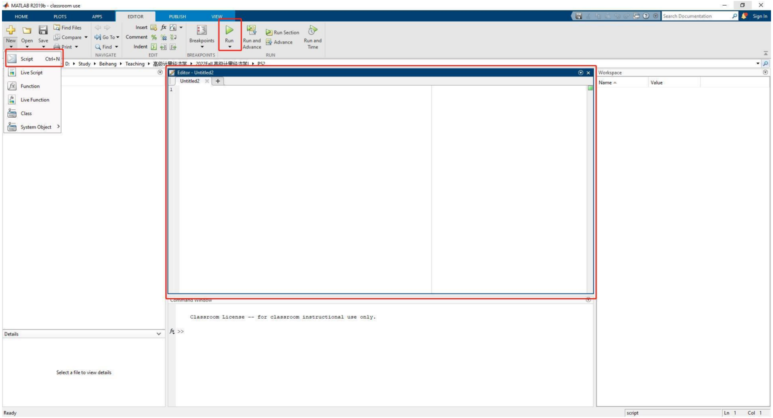
文件路径：当需要读取数据的时候，需要确保路径准确

Matlab 基础



在使用 Matlab 时，为了易于保存和编辑，一般会新建一个 script，然后在 script 之中输入代码，输入完成后点击 run 即可运行代码

Matlab 基础



在使用 Matlab 时，为了易于保存和编辑，一般会新建一个 script，然后在 script 之中输入代码，输入完成后点击 run 即可运行代码

Matlab 基础

Editor - D:\Study\Beihang\Teaching\本科计量经济学\2022 Fall\Slides\lab1\main_lab1_short.m

main_lab1_short.m

```
% A line starting with a % is a comment

% clear will erase all the variables that Matlab currently hold in memory
clear;

% This would close everything but Matlab itself: opened files, graphic windows, ...
close all;

```

- % 表示注释行
- clear 命令会删除 Matlab 之中存储的所有数据
- close all 命令会关闭 Matlab 之中打开的额外的窗口：比如绘制的图表

Matlab 基础

```
% We can create variables
x = 4;

% Contrary to SAS, a command does not need to be terminated by a ;
% If we do not put a ; at the end, the results are displayed in
% the output window
y = 5
```

- 当一个命令以; 作为结尾，表示不在命令窗口输出结果

```
Command Window
>> main_lab1_short

y =

     5

>>
```

Matlab 基础

```
% We can create vectors. The [ and ] signs are telling Matlab that what's
% in between are elements of a vector (or matrix).
% A space is treated as a comma: it tells Matlab to stack the numbers in a row
% A ; tells Matlab to stack the number in a column
a = [1 2 3 4]
b = [5, 6, 7, 8]
c = [9; 10; 11; 12]
```

- 用 [ ] 可以创建向量和矩阵
- 行间的元素用空格或者逗号隔开
- 列间的元素用分号隔开

Command Window

```
a =

     1     2     3     4


b =

     5     6     7     8


c =

     9
    10
    11
    12
```

Matlab 基础

```
% We can stack vectors to create a matrix
d = [a; b]

% We can create a matrix directly
e = [1, 2, 3, 4; 5, 6, 7, 8]
```

- 我们通过合并向量和矩阵来创建新的矩阵
- 我们也可以直接创建新的矩阵

```
d =

     1     2     3     4
     5     6     7     8


e =

     1     2     3     4
     5     6     7     8
```

Matlab 基础

```
% you can create matrices of zeros, ones or an identity matrix
e = zeros(4,5)
e = ones(3,2)
e = eye(3)
```

```
e =

     0     0     0     0     0
     0     0     0     0     0
     0     0     0     0     0
     0     0     0     0     0
```

```
e =

     1     1
     1     1
     1     1
```

```
e =

     1     0     0
     0     1     0
     0     0     1
```

Matlab 基础

```
% We can add, subtract, multiply and divide numbers
f = 1+2
f = 1-2
f = 1*2
f = 1/2

% We can multiply matrices and vectors
f = d*c

% We can add and subtract matrices and vectors
f = a-b
```

```
f =

     3

f =

    -1

f =

     2

f =

    0.5000

f =

   110
   278

f =

    -4     -4     -4     -4
```

Matlab 之中可以直接进行 (矩阵) 四则运算

Matlab 基础

```
% We can invert a matrix
g = [10, 3; 3, 5]
h = inv(g)
g*h
```

Matlab 可以用 inv 求逆矩阵

```
Command Window
g =

    10     3
     3     5


h =

    0.1220   -0.0732
   -0.0732    0.2439


ans =

    1.0000   -0.0000
   -0.0000    1.0000
```

Matlab 基础

```
% We can use ... to put one statement over multiple lines
i = [1, 2, 3, ...
     4, 5, 6]

% We can transpose vectors and matrices with '
i = i'
```

```
i =

     1     2     3     4     5     6


i =

     1
     2
     3
     4
     5
     6
```

- 当命令过长时，可以用... 进行换行
- 单引号' 表示矩阵的转置

Matlab 基础

```
% On vectors and matrices we can do element-by-element operations
% (multiplications and divisions) using the dot operators
f = a .* b
f = a ./ b
```

乘除前面加一点（点乘和点除）表示矩阵和向量元素对元素的乘除运算

```
f =

     5    12    21    32


f =

    0.2000    0.3333    0.4286    0.5000
```

Matlab 基础

```
% We can read a data file in text format (columns separated by spaces) using
% the load function
j = load('baseball.csv')
```

load 函数可以用于读取数据 (txt, csv, ...)

```
j =

   194663079          97        2006
   120099824          86        2006
   103472000          89        2006
   102750667          90        2006
   101084963          97        2006
    98447187          88        2006
    94424499          66        2006
    92551503          82        2006
    90156876          79        2006
    90056419          76        2006
    88891371          83        2006
    88273333          85        2006
    87959833          78        2006
    82612866          95        2006
    72585582          70        2006
    71915000          87        2006
    69896141          88        2006
    68228662          80        2006
```

Matlab 基础

```
% I can select the subset of a matrix. For example, rows 4 to 6 and columns
% 2 and 3
k = j(4:6,2:3)
```

可以快捷的选取已有变量的特定行列来创建新的变量

```
k =

    90        2006
    97        2006
    88        2006
```

Matlab 基础

```
% Now I want to regress the number of wins on an intercept and the payroll
% (measured in millions of dollars)
y = j(:,2);

N = length(y); % Matlab is case-sensitive. You can have a variable named 'N'
               % and another variable named 'n'
x = zeros(N,2);
x(:,1) = ones(N,1);
x(:,2) = j(:,1)/1000000;

betaHat = inv(x'*x)*x'*y
```

- 我们做一个简单的线性回归 $Y_i = \alpha + \beta X_i + u_i$
- Y_i 表示获胜场次，X_i 表示球队工资支出
- Matlab 区分大小写

```
betaHat =

   68.4069
    0.1638
```

Matlab 基础

计算 $\hat{u}_i$, $\hat{\sigma}^2$, $Var[\hat{\beta}|X_i]$, t 统计量, p 值

```
% I can compute the residuals
uHat = y - x*betaHat;

% I can estimate the variance of the error term
sigma2Hat = sum(uHat.^2)/(N-2);

% I can compute the variance matrix of betaHat
varBetaHat = sigma2Hat*inv(x'*x)

% The standard errors are
stdBetaHat = sqrt(diag(varBetaHat))

% I can compute the p-value for the hypothesis that the payroll coefficient
% is equal to zero (2-sided)
t = (betaHat(2)-0)/stdBetaHat(2)
pValue = 2*(1-normcdf(abs(t)))
```

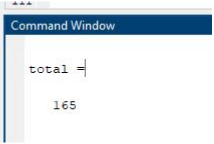
Command Window

```
varBetaHat =

   18.2759   -0.2015
   -0.2015    0.0026


stdBetaHat =

    4.2750
    0.0509


t =

    3.2209


pValue =

    0.0013

>>
```

Matlab 基础

```
% We can do loops using the "for" statement. Quick example: adding the
% numbers from 10 to 20

total = 0;
for i=10:1:20
    i
    total = total+i
end
total
```

- Matlab 之中可以用 for 循环来进行迭代计算
- 比如计算 $10+11+\cdots+19+20$
- 除了 for 循环，常用的语句还有 while 循环，if 条件
- 自己在命令窗口输入 help while 和 help if 来学习具体的用法



非线性模型估计方法

- 离散选择模型
  - 在经典计量经济学模型中，被解释变量 y 通常被假定为连续变量
  - 1962 年，Warner 首次将应用于经济研究领域，用以研究公共交通工具和私人交通工具的选择问题
  - 70、80 年代，离散选择模型被普遍应用于经济布局、企业定点、交通问题、就业问题、购买决策等经济决策领域的研究
  - 模型的估计方法，主要发展于 80 年代初期

- 以二元选择模型 (Binary Choice Model) 为例
  - 对于单个方案的取舍：例如，购买者对某种商品的购买决策问题，求职者对某种职业的选择问题，投票人对某候选人的投票决策，银行对某客户的贷款决策
  - 对于两个方案的选择：例如，两种出行方式的选择，两种商品的选择

二元选择模型 (Binary Choice Model)

- y_i 为观测值为 1 和 0 的决策被解释变量
- X_{ji} 为解释变量，包括选择对象的属性和选择主体的属性
  - 比如你是银行信用卡中心的员工，需要决定是否给申请人发放信用卡
  - $y_i = 1$ 表示发放，$y_i = 0$ 表示不发放
  - X_{ji} 往往包含婚姻状况，收入，年龄，性别，受教育年限等变量

非线性模型 (Nonlinear Model)

模型估计：以 Probit 模型为例

- 非线性模型的估计方法
- 估计量的一致性
- 估计量的渐近分布
- 估计量的经济学含义，边际效用 (marginal effect)

非线性假设检验

- Delta Method

其他非线性模型

- Logit 模型
- Tobit 模型

线性模型拓展：异方差性 (Heteroskedasticity)

异方差的概念

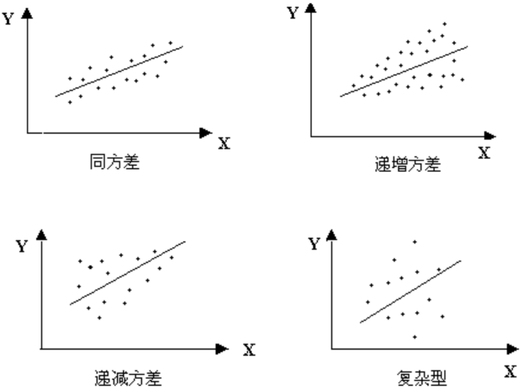
- 对不同的样本点 i，随机误差项的方差不再是常数，而互不相同，则认为出现了异方差性
  - 比如，我们考虑每月家庭消费支出 y_i 与每月家庭可支配收入 X_i 的关系
  - 很显然，每月家庭可支配收入 X_i 越高，其对应的方差也会越大，也就是 u_i 的方差会随着 X_i 增大而增大

- 异方差的类型
  - 同方差：$E[u_i^2] = \sigma^2$ 为常数
  - 异方差：$E[u_i^2] = \sigma_i^2 = f(X_i)$ 与解释变量观测值 X_i 有关

线性模型拓展： 异方差性 (Heteroskedasticity)

异方差的类型

- 单调递增型： σ_i^2 随 X_i 的增大而增大
- 单调递减型： σ_i^2 随 X_i 的增大而减小
- 复杂型： σ_i^2 与 X_i 的变化呈复杂形式



线性模型拓展：异方差性 (Heteroskedasticity)

示例 1

- 研究居民家庭的储蓄行为 $y_i = \alpha + \beta_2 X_{2i} + u_i$
- y_i: 第 i 个家庭的储蓄额
- X_{2i}: 第 i 个家庭的可支配收入

- 高收入家庭：储蓄的差异较大
- 低收入家庭：储蓄则更有规律性，差异较小
- u_i 的方差呈现单调递增型变化

线性模型拓展：异方差性 (Heteroskedasticity)

示例 2

- 以绝对收入假设为理论假设，建立居民消费函数
- 将居民按照收入等距离分成 n 组，取组平均数为样本观测值
  $y_i = \alpha + \beta_2 X_{2i} + u_i$
- y_i: 第 i 组观测值的消费
- X_{2i}: 第 i 组观测值的收入

- 一般情况下，中等收入组人数多，两端收入组人数少
- 人数多的组平均数的误差小，人数少的组平均数的误差大
- u_i 的方差随着解释变量观测值的不同而不同，且呈 U 形

线性模型拓展：异方差性 (Heteroskedasticity)

示例 3

- 某一行业的企业为样本建立企业生产函数模型
  $y_i = A_i^{\beta_1} K_i^{\beta_2} L_i^{\beta_3} e^{u_i}$
- y_i: 企业 i 的产出量
- A_i: 企业 i 的技术，K_i: 企业 i 的资本，L_i: 企业 i 的劳动

- 每个企业所处的外部环境对产出量的影响，被包含在随机误差项中
- 对不同的企业，它们对产出量的影响程度不同，造成了随机误差项的异方差性
- 随机误差项的方差并不随某一个解释变量观测值的变化而呈规律性变化，呈现复杂型

异方差性的解决办法

- 异方差性的后果
  - 不具有有效性 (不再是 BLUE)
  - 变量的显著性检验失去意义
- 异方差性的解决办法
  - 加权最小二乘法 (Weighted Least Squares)
  - 异方差稳健标准误法
- 异方差性的检验
  - 图示法
  - 怀特 (White) 检验

线性模型拓展：内生性 (Endogeneity)

内生性的概念

- $E[X_i u_i] \neq 0$
- $\hat{\beta}$ 是有偏的、不一致的

- 内生性的解决办法
  - 工具变量 (Instrument variables)：模型估计过程中被作为工具使用，以替代模型中与随机误差项相关的随机解释变量
  - 与所替代的内生性变量高度相关
  - 与随机误差项不相关
  - 两阶段最小二乘法

- 内生性检验
  - 弱工具变量检验
  - 过度识别检验

时间序列分析 (Time Series Analysis)

经典计量经济模型常用到的数据有

- 截面数据 (cross-sectional data)
- 时间序列数据 (time-series data)
- 面板数据 (panel data)

时间序列数据，是最常见也是最常用到的数据

- 是一批按照时间先后排列的统计数据

时间序列分析 (Time Series Analysis)

经典时间序列分析模型

- 平稳时间序列模型
  - Moving average (MA) 模型
  - Autoregression (AR) 模型
  - ARMA 模型
- 非平稳时间序列模型
  - 单位根检验
  - 协整检验 (本课程暂不涉及)

平稳性

- 经典回归分析暗含着一个重要假设：数据是平稳的
- 当数据非平稳时
  - 大样本下的统计推断基础 (中心极限定理和大数定律) 被破坏
  - 往往导致出现 “虚假回归”(Spurious Regression)
    - 表现为两个本来没有任何因果关系的变量，却有很高的相关性 (如都含有时间趋势项 t)

平稳性的定义

- 假定时间序列 $\{X_t\}$ 满足下列条件
  - 均值 $E(X_t) = \mu$, 是与时间 t 无关的常数
  - 方差 $Var(X_t) = \sigma^2$, 也是与时间 t 无关的常数
  - 协方差 $Cov(X_t, X_{t-h}) = \gamma_h$, 只与时期间隔 h 有关、而与时间 t 无关的常数
- 则称该随机时间序列是平稳的 (stationary), 也称作弱平稳

平稳性的随机过程示例

- 白噪声（white noise）过程
  - X_t 服从某一特定的分布
  - 并且相对于 t 独立
  - 且满足 $E(X_t) = 0, Var(X_t) = \sigma^2$
- 则容易得出
  - 均值 $E(X_t) = 0$，与时间 t 无关
  - 方差 $Var(X_t) = \sigma^2$，与时间 t 无关
  - 协方差 $Cov(X_t, X_{t-h}) = 0$，与 h、t 无关

非平稳性的随机过程示例

- 随机游走 (random walk) 过程
  - $X_t = X_{t-1} + u_t$，其初始值 $X_0 = 0$
  - u_t 服从某一特定的分布
  - 并且相对于 t 独立
  - 且满足 $E(u_t) = 0, Var(u_t) = \sigma^2$
- 则容易得出
  - 均值 $E(X_t) = 0$，与时间 t 无关
  - 方差 $Var(X_t) = t\sigma^2$，与时间 t 有关
  - 协方差 $Cov(X_t, X_{t-h}) = (t-h)\sigma^2$，与 h、t 有关

差分平稳示例

- 随机游走 (random walk) 差分过程
  - $X_t = X_{t-1} + u_t$，其初始值 $X_0 = 0$
  - u_t 服从某一特定的分布
  - 并且相对于 t 独立
  - 且满足 $E(u_t) = 0, Var(u_t) = \sigma^2$
- 则容易得出 $\Delta X_t = X_t - X_{t-1} = u_t$ (一阶差分 (first difference))
  - u_t 是一个随机游走过程
  - 故而是平稳的
  - 如果一个时间序列是非平稳的，它常常可通过取差分的方法而形成平稳序列

时间序列分析

- 时间序列模型
  - MA 模型
  - AR 模型
- 单位根检验
  - Dickey-Fuller 检验
  - Augmented Dickey-Fuller 检验

面板数据

- 指在时间序列上取多个截面，在这些截面上同时选取样本观测值所构成的样本数据
  - 经济分析时，经常会遇到时间序列和横截面两者相结合的数据
    - 企业投资需求分析中，会遇到“多个企业的若干指标的不同月度或季度时间序列”
    - 在城镇居民消费分析中，会遇到“不同省市地区的反映居民消费和居民收入的年度时间序列”
- 将前述的企业或地区等统称为个体，这种具有“三维（指标 X，个体 i，时间 t）信息的数据结构”称为面板数据（panel data）

面板数据模型

- 在之前的课程制作，在分析时只利用了面板数据中的某些二维数据信息 (时间序列建模或横截面数据建模)
- 实际经济分析中，仅利用二维信息的模型在很多时候往往不能满足分析问题的需要
- 例如，在生产函数分析中 $y = A^{\beta_1} K^{\beta_2} L^{\beta_3} e^u$
  - 利用横截面数据，只能对规模经济进行分析
  - 利用混有规模经济和技术革新信息的时间序列数据，只有在假设规模收益不变的条件下、才能实现技术革新的分析
  - 利用面板数据，可同时分析企业的规模经济和技术革新，可实现规模经济和技术革新的综合分析

面板数据模型

- 面板数据含有 (指标 X, 个体 i, 时间 t) 三维信息
- 利用面板数据模型可以构造和检验比以往单独使用横截面数据或时间序列数据更为真实的行为方程，可进行更加深入的分析
- 已经成为近年来计量经济学理论方法的重要发展之一

- 固定效应

线性模型拓展：序列相关性 (Serial Correlation)

序列相关性的概念

- 在我们已经学习过的模型之中，均假设 u_i 之间不存在相关性 (No Autocorrelation)
- 以截面数据为样本时，如果模型随机项之间存在相关性, 称为 Spatial Autocorrelation
- 以时序数据为样本时，如果模型随机项之间存在相关性，称为 Serial Autocorrelation
- 现实之中，一般统称为序列相关性 (Serial Correlation or Autocorrelation)
- 统计学上表现为 $cov(u_i, u_j) = E[u_i u_j] \neq 0$

序列相关性的原因

- 没有包含在解释变量中的、被解释经济变量 “固有惯性”(比如本期 GDO 取决于过去数期的 GDP)
- 模型设定偏误
  - 函数形式有误
  - 遗漏了重要的解释变量
- 数据经过预处理 (如插值补齐)，数据误差等因素

- 时序数据作为样本时，一般重点考虑序列相关性
- 截面数据作为样本时，一般较少考虑序列相关性

序列相关性的解决办法

- 序列相关性的后果
  - 参数估计量非有效
  - 变量的显著性检验失去意义
  - 与异方差性引起的后果相同
- 序列相关性的解决办法
  - 广义最小二乘法 (广义差分法)
- 序列相关性的检验

期末考试

期末考试题型

- 选择题 (约 12 道)
- 判断题 (约 5 道)
- 论述题 (约 1 道)
- 计算、推断、证明题 (约 3 道)

- 难度与往年期末考试相当 (建议参考去年期末进行复习)，简单于平时作业
- 具体时间学校统一安排，时间 120 分钟
- 目前预计是线上考试

期末考试

题型举例

半对数模型 $Y_i = \beta_0 + \beta_1 LnX + \mu_i$ 中，参数 β_1 的含义是（ D ）

A. Y 关于 X 的弹性

B. X 的绝对量变动，引起 Y 的绝对量变动

C. Y 关于 X 的边际变动

D. X 的相对变动，引起 Y 的期望值绝对量变动

期末考试

回归分析中使用的距离是点到直线的垂直坐标距离，最小二乘准则是指（ D ）

A. 使$\left|\sum_{t=1}^{n}(Y_t-\hat{Y}_t)\right|$达到最小值

B. 使$\min\left|Y_i-\hat{Y}_i\right|$达到最小值

C. 使$\max\left|Y_t-\hat{Y}_t\right|$达到最小值

D. 使$\sum_{t=1}^{n}(Y_t-\hat{Y}_t)^2$达到最小值

期末考试

1、经典线性回归模型（CLRM）中的干扰项不服从正态分布的，OLS 估计量将有偏的。

错

即使经典线性回归模型（CLRM）中的干扰项不服从正态分布的，OLS 估计量仍然是无偏的。

因为 $E(\hat{\beta}_2) = E(\beta_2 + \sum K_i \mu_i) = \beta_2$，该表达式成立与否与正态性无关。

期末考试

1、（12 分）某公司想决定在何处建造一个新的百货店，对已有的 30 个百货店的销售额作为其所处地理位置特征的函数进行回归分析，并且用该回归方程作为新百货店的不同位置的可能销售额，估计得出（括号内为估计的标准差）

$$\hat{Y}_t = 30 + 0.1 \times X_{1t} + 0.01 \times X_{2t} + 10.0 \times X_{3t} + 3.0 \times X_{4t}$$

$$(0.02) \qquad (0.01) \qquad (6.8) \qquad (1.0)$$

其中：Y_t ＝第 i 个百货店的日均销售额（美元）；

X_{1t} ＝第 i 个百货店前每小时通过的汽车数量（辆）；

X_{2t} ＝第 i 个百货店所处区域内的人均收入（美元）；

X_{3t} ＝第 i 个百货店内所有的桌子数量；

X_{4t} ＝第 i 个百货店所处地区竞争店面的数量。

请回答以下问题：

（1） 说出本方程中系数 0.1 和 0.01 的经济含义。

（2） 各个变量前参数估计的符号是否与经济理论期望的符号一致？

（3） 在 α ＝0.05 的显著性水平下检验变量 X_{1t} 、 X_{3t} 的显著性。

Thank you