



Notes of Econometrics

计量经济学笔记

作者：董坤霖

组织：金融工程系，金融学院，对外经济贸易大学

时间：February 14, 2022

版本：1.0

邮箱：dkl_0707@163.com



该吃吃 该喝喝
遇事儿别往心里搁

前言

计量经济学是定量金融的入门课，也是必修课里较为硬核的一门课，无论你以后读的是金工、经济还是什么公司金融，之后的学习过程中都会用到计量经济学。当然，由于篇幅限制，本人在这里仅仅介绍最为基础的回归方法，对于后续更深的计量知识本人不做过多的展开。

计量经济学这门课，重在理解、掌握和整理，不在刷题。在一年的讲解过程中，本人见过许多同学相信所谓的“题库”，在没有对知识点进行理解的情况下直接大量刷题，结果并不是很理想。需要说明的是，这门课的考试框架虽然每年都类似，但是具体内容大不相同，不会重复出题，也不会特别设置偏题、怪题。一般来说，只要你认真整理、及时巩固并复习，完全是可以考到自己想要的分数的。当然，由于不同人对于同一门课具有不同的理解思路，本人在这里仅仅只是给出了一种计量学习框架。如果你想要完全学懂这门课，请结合自身的思维方式对本笔记进行使用。

对于这本笔记，本人花了相当长时间对自己的讲课手稿进行不断的整理编撰，这过程中耗费了大量心血。其中在某些教材没有讲清楚的地方加了一些补充，同时也进行了一些微观计量经济学、金融时间序列分析等方面知识的部分拓展帮助理解。当然，由于这门课只是个定量课程的入门课，我并没有加入一些难度过大的如矩阵代数的内容，所有拓展的内容都会在前面打*。也希望大家不要随随便便去高价卖这个笔记，这个笔记只是想帮大家学懂这门课，而不是作为一种资产进行高价出售。

另外，这个笔记纯粹只是辅助大家理解，并不如某些大牛所写的教材，它缺少实证分析的内容，也没有大量习题可供练习，仅仅只是本人根据自身一年来的教学经历和在科研项目、实习经历、后续课程等中对计量经济学框架的简单梳理，目的只是帮助大家学好这门课，考一个自己满意的分数，想要在计量领域上有更多发展还得自己多研究、多读论文、多与周围人探讨。

由于学识有限，其中难免有不准确的理解和表达，若同学们发现笔记中的遗漏不足之处，欢迎通过邮箱联系本人进行更正。另外，Ethan Deng分享的LaTex模板(GitHub地址:<https://github.com/ElegantLaTeX/>)在本人的笔记排版上提供了极大的便利，对此本人表示感谢！

Kunlin Dong

February 14, 2022

目录

1	计量经济学课程框架	1
2	数学基础	3
2.1	概率论知识	3
2.1.1	概率框架构建	3
2.1.2	数字特征	5
2.1.2.1	一阶矩：期望	5
2.1.2.2	二阶矩：方差和协方差	6
2.1.2.3	*三阶矩：偏度	7
2.1.2.4	*四阶矩：峰度	8
2.1.3	*高阶矩和低阶矩的进一步说明	8
2.1.4	几个分布	8
2.1.4.1	Bernoulli分布	9
2.1.4.2	正态分布及其衍生分布	9
2.1.5	大数定律和中心极限定理	11
2.2	统计学知识	11
2.2.1	几个概念	11
2.2.2	估计量评价标准	12
2.2.3	常见统计量及常见的抽样分布	13
2.2.4	参数估计方法	14
2.2.5	置信区间和假设检验	14
2.2.5.1	置信区间	14
2.2.6	假设检验	15
3	一元线性回归	17
3.1	一元线性回归模型概述	17
3.2	OLS估计	17
3.3	拟合效果评价	18
3.4	计量效果评价	19
3.4.1	OLS假设	19
3.4.2	抽样分布	22
3.4.3	置信区间和假设检验	23
3.4.4	LSA1假设增强	23
第3章练习	26
4	多元线性回归模型	28
4.1	遗漏变量偏差	28
4.2	多元线性回归模型表达式	29
4.3	模型估计	29
4.4	拟合效果评价	30
4.5	计量效果评价	30
4.5.1	模型假设	30

4.5.2 抽样分布和置信区间	31
4.5.3 假设检验	31
4.5.3.1 单个约束检验	31
4.5.3.2 多个约束检验	32
4.5.4 LSA1假设放松	34
第 4 章 练习	35
5 非线性回归模型	37
5.1 因果效应非常数	37
5.2 多项式回归	37
5.3 对数回归模型	38
5.3.1 线性对数模型	38
5.3.2 对数线性模型	38
5.3.3 双对数模型	39
5.4 交互项回归	39
5.4.1 X_i 和 W_i 是二值变量	39
5.4.2 X_i 是连续变量, W_i 是二值变量	40
5.4.3 X_i 和 W_i 是连续变量	40
第 5 章 练习	41
6 线性回归常见问题	42
6.1 遗漏变量偏差	42
6.2 回归函数非线性	42
6.3 测量误差	42
6.3.1 传统测量误差	43
6.3.2 最佳猜测测量误差	43
6.4 缺失数据和样本选择	44
6.4.1 数据缺失完全随机	44
6.4.2 数据缺失与X相关	45
6.4.3 数据缺失与Y相关	45
6.5 双向因果	46
第 6 章 练习	46
7 面板数据回归	47
7.1 面板数据	47
7.2 不可测的遗漏变量偏差	47
8 二值因变量回归	48
9 工具变量回归	49
10 ElegantL^AT_EX 系列模板介绍	50
10.1 ElegantBook 更新说明	50
10.2 模板安装与更新	50
10.2.1 在线使用模板	50
10.2.2 本地免安装使用	51
10.2.3 发行版安装使用	51

10.2.4 更新问题	51
10.2.5 其他发行版本	51
10.3 关于提交	51
11 ElegantBook 设置说明	52
11.1 语言模式	52
11.2 设备选项	52
11.3 颜色主题	52
11.4 封面	53
11.4.1 封面个性化	53
11.4.2 封面图	53
11.4.3 徽标	54
11.4.4 自定义封面	54
11.5 章标题	54
11.6 数学环境简介	54
11.6.1 定理类环境的使用	55
11.6.2 其他环境的使用	55
11.7 列表环境	55
11.8 参考文献	56
11.9 添加序章	56
11.10 目录选项与深度	56
11.11 章节摘要	57
11.12 章后习题	57
第 11 章 练习	57
11.13 旁注	58
12 字体选项	59
12.1 数学字体选项	59
12.2 使用 newtx 系列字体	59
12.2.1 连字符	59
12.2.2 宏包冲突	59
12.3 中文字体选项	60
12.3.1 方正字体选项	60
12.3.2 其他中文字体	60
13 ElegantBook 写作示例	62
13.1 Lebesgue 积分	62
13.1.1 积分的定义	62
第 13 章 练习	64
14 常见问题集	65
15 版本更新历史	66
A 基本数学工具	69
A.1 求和算子与描述统计量	69

B Elegant^{LT}E_X 系列模板介绍	70
B.1 ElegantBook 更新说明	70
B.2 模板安装与更新	70
B.2.1 在线使用模板	70
B.2.2 本地免安装使用	71
B.2.3 发行版安装使用	71
B.2.4 更新问题	71
B.2.5 其他发行版本	71
B.3 关于提交	71
C ElegantBook 设置说明	72
C.1 语言模式	72
C.2 设备选项	72
C.3 颜色主题	72
C.4 封面	73
C.4.1 封面个性化	73
C.4.2 封面图	73
C.4.3 徽标	74
C.4.4 自定义封面	74
C.5 章标题	74
C.6 数学环境简介	74
C.6.1 定理类环境的使用	75
C.6.2 其他环境的使用	75
C.7 列表环境	75
C.8 参考文献	76
C.9 添加序章	76
C.10 目录选项与深度	76
C.11 章节摘要	77
C.12 章后习题	77
第 C 章 练习	77
C.13 旁注	78
D 字体选项	79
D.1 数学字体选项	79
D.2 使用 newtx 系列字体	79
D.2.1 连字符	79
D.2.2 宏包冲突	79
D.3 中文字体选项	80
D.3.1 方正字体选项	80
D.3.2 其他中文字体	80
E ElegantBook 写作示例	82
E.1 Lebesgue 积分	82
E.1.1 积分的定义	82
第 E 章 练习	84

F 常见问题集	85
G 版本更新历史	86
A 基本数学工具	89
A.1 求和算子与描述统计量	89

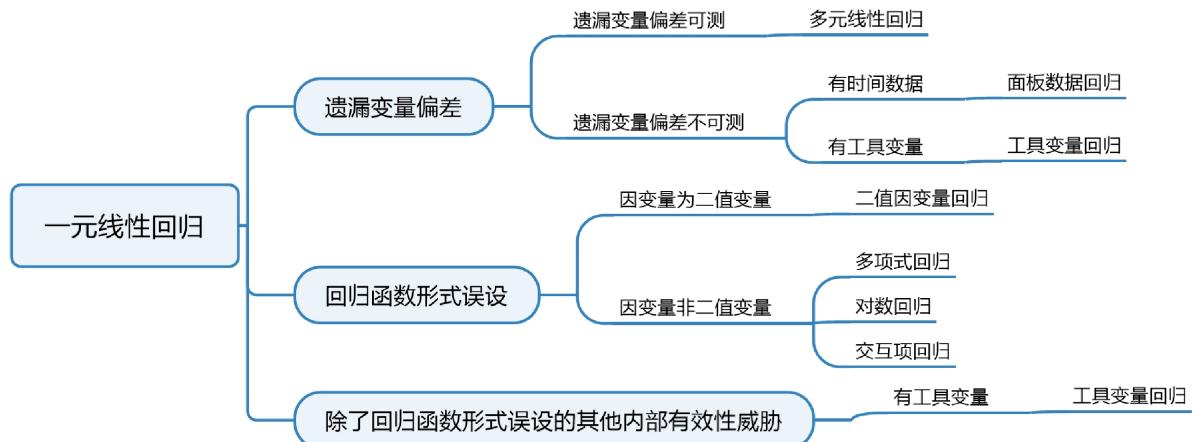
第1章 计量经济学课程框架

在学任何一门定量课程之前，我们必须要知道这门课的学习目的，以及课程学习框架。你学完了这门课，必须要知道这门课到底在干啥，他解决了哪些问题，他没解决的问题又有哪些。无论你后续想走的是学术还是业界，明白这些概念往往会对你的发展有非常大的帮助。

那么，作为定量课程的入门课，首先我们有了第一个问题：计量经济学到底在干啥？这门课跟其他课程不同，经济学只是他的研究背景，比如劳动力的增多对老板工资提高多少，招股说明书的异质性会对IPO定价的作用，漂亮国的军费开支增多对本国的GDP的效应等等，而真正的侧重点在于前面的“计量”两个字。那么，计量经济学到底在计量啥呢？在经济学里面，我们关心的是在一定条件下，一个变量变动对另一个变量的影响，这种影响我们称之为因果效应。计量经济学干的事情就是以量化的方式对这种因果效应进行呈现。而我们量化的方式，往往是利用回归方程进行研究。从这个角度上来说，计量经济学也可以被称作是回归分析导论。

根据以上论述，我们下一步要确立的，应该是回归方程形式。首先对于一个因变量 Y_i ，我们将影响因素分为 X_i 和除了 X_i 以外影响 Y_i 的因素，我们称之为 u_i 。而后我们建立回归方程： $Y_i = f(X_i) + u_i$ ，其中 X_i 是解释变量， Y_i 是因变量， f 是回归函数， u_i 是随机误差项。在最简单的情景下，我们确定回归函数 f 为线性函数，即 $Y_i = \beta_0 + \beta_1 X_i + u_i$ ，这就是我们的一元线性回归模型。

当然，因为这种模型有很多问题，比如可测情况下的遗漏变量偏差，这时候需要我们加入遗漏变量，变成多元线性回归模型，如果是不可测的情况，我们往往会采用面板数据回归的方式对遗漏变量进行“降维打击”，或者“请求辅助支援”利用工具变量进行回归。如果是因果效应不是常数的问题，或者因变量是个二值变量，我们可能得改改 f ，前者往往改成多项式/对数/交互项模型，后者往往改成Probit/Logit模型。如果还有其它问题，比如双向因果这种最恶心的问题，我们剩下的唯一办法基本上是利用工具变量进行回归。这就是计量经济学的宏观上的整体框架，见图 1.1。本质上无论再怎么改，初级计量始终都围绕着一元线性回归模型，所以各位在学习前面的内容时一定要好好掌握。



那么，如果是从微观角度上看，计量经济学该怎么学习某一个回归模型呢？一般而言，我们是按照提出问题-分析问题-解决问题的方式进行研究。分为以下步骤。

1. 模型的提出背景是什么？任何一个模型都是为了解决实际应用问题而服务的，这是定量分析课程的特点，我们的模型要么是解决上一个模型遗留的问题，要么是发现了新的问题去解决它。
2. 模型的表达式长啥样？根据这个表达式，怎么解释回归模型系数，怎么进行系数估计？
3. 如何评价这个模型好坏？评价一般分为两种：拟合效果评价和计量效果评价。

4. 对于拟合效果评价，主要是看 R^2 或者调整 R^2 , SER或者RMSE, 这其实是看看模型是否能够很好的拟合数据。
5. 对于计量效果评价，主要是看你的研究的问题条件是否符合你的模型假设。如果符合，你去算回归系数的抽样分布，进而去算置信区间，假设检验等等。

当然，如果你发现这个模型还可以在加强一些条件，那可以加上去看看是否能让回归系数的抽样分布变的更简洁，比如一元线性回归加上了误差项同方差可以大大简化回归系数的抽样分布方差；但如果说你觉得这个条件太苛刻，那你可以进行放松，比如多元线性回归中，通过加入了控制变量将条件均值为0这个假设放松成了条件均值独立。

6. 如果你发现你无论怎么改，这个模型还是有问题，那么最好的办法就是换个模型看看能不能解决。

上述过程可被总结为图 1.2。

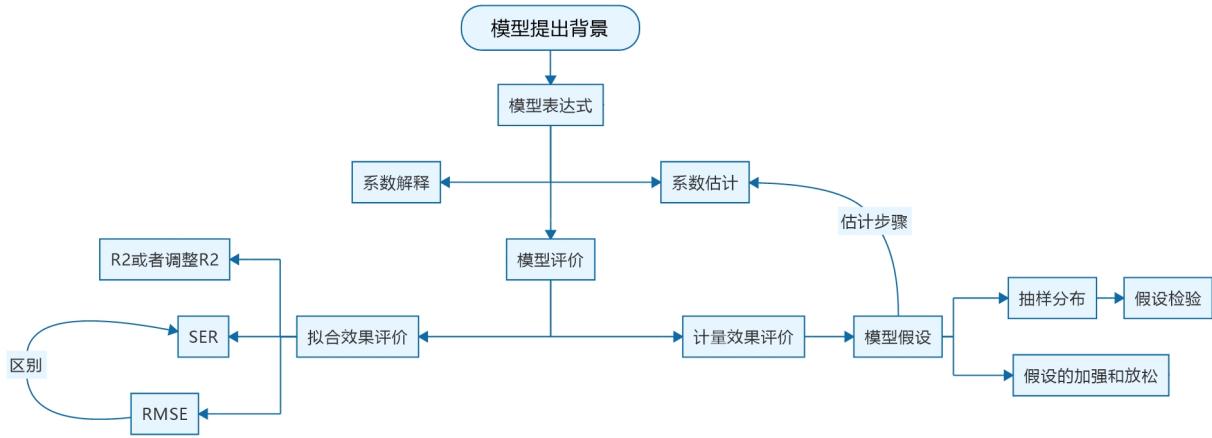


图 1.2: 计量经济学微观框架

**以上就是计量经济学的全部框架。计量经济学本质上只讲了最为基础的计量方法，如果你想更进一步学习这个计量内容或者想走学术方向，可以选择微观经济学这门课程进行更进一步学习（当然这门课的难度个人感觉略大）。如果你对截面数据上的因果效应的计量不感兴趣，而是对时间序列上的预测更为感兴趣，可以选择金融时间序列分析这门课进行学习。再或者，你对于计量中的回归效果不满意，可以选择机器学习方向进行更进一步的学习。总而言之，只要是走定量方向，无论你未来是走学术还是走业界，计量经济学大概率是你需要着重花时间学习的一门课。

第2章 数学基础

初级计量这门课的数学基础主要分为以下部分：一、概率论知识。二、统计学知识。其中概率论部分重点掌握随机变量的数字特征、正态分布及其性质、大数定律和中心极限定理。统计学部分重点掌握极大似然估计和各种假设检验。

2.1 概率论知识

世间万物本身具有随机性，尤其是经济金融领域，我们无法像自然学科那般建立完美的对照实验对问题进行研究。换言之，我们需要一个理论去测度这种随机性。而概率论的提出就是为此服务。概率论主要是在理论上对事件发生的“可能性”提供了概率这个数学工具进行测度，并以此为依据对发生的“可能性”的特征进行了描述。

不过，对于单单的计量本身不需要掌握过深的概率论知识，我们仅需要建立起最为基础的概率框架，不需要掌握其中过深的理论证明。

2.1.1 概率框架构建

首先，我们定义一个样本空间 Ω ，其中的每个元素 $A \in \Omega$ 被称作样本点或者基本事件，在同一个问题的研究框架下，通常只是针对一个样本空间。而后，我们定义概率如下：

定义 2.1 (概率)

设 Ω 为样本空间， $A \in \Omega$ 为样本点或者基本事件，对于每个事件我们定义函数 $P : \Omega \rightarrow [0, 1]$ 。若满足以下条件：

- $P(A) \in [0, 1]$
- $P(\Omega) = 1$
- 对于事件 A_1, A_2, \dots ，始终有 $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

则我们称 $P(A)$ 为事件A的概率。



由于实际中我们更经常遇到的是已知某一事件的发生，想要求另一事件的概率。为此我们定义了条件概率。

定义 2.2 (条件概率)

设 A, B 为两个事件且有 $P(A) > 0$ ，则我们将 $\frac{P(AB)}{P(A)}$ 称为已知事件A发生的情况下，事件B发生的概率，即

$$P(B|A) = \frac{P(AB)}{P(A)} \quad (2.1)$$



由条件概率我们也可以引申出事件之间相互独立的定义。

定义 2.3 (事件的独立性)

对于任意事件 $A, B \in \Omega$ ，若有

$$P(AB) = P(A)P(B) \quad (2.2)$$

则称 A, B 相互独立。当然，在 $P(A) > 0$ 的条件下，根据条件概率定义，我们亦可以写成如下形式：

$$P(B|A) = P(B) \quad (2.3)$$



需要注意的是，独立意味着不相关，但不相关不能推出着独立。

另外，在实际中我们也常常会把一个复杂事件拆成若干个不相容的简单事件之和，为此有全概率公式。

定理 2.1 (全概率公式)

设样本空间为 Ω ，事件 A_1, A_2, \dots, A_n 互不相容， $P(A_i) > 0, i = 1, 2, \dots, n$ ，且 $\Omega = \bigcup_{i=1}^n A_i$ ，对于任意事件 $B \in \Omega$ ，有

$$P(B) = \sum_{i=1}^n P(A_i)P(B|A_i) \quad (2.4)$$



条件概率的引入和全概率公式的提出表明，对于一个结果事件B的发生概率，我们可以将其写成是在一系列已知原因事件 A_i 发生下结果事件B发生的概率乘以原因事件 A_i 的概率乘积之和。即可以“由因推果”。

但同时我们也注意到，由于 $P(A_iB) = P(A_i|B)P(B) = P(B|A_i)P(A_i)$ ，即 $P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$ 。这表明，已知结果B发生，我们想要求这个结果是由原因 A_i 引起的概率（果→因），可以先求已知原因 A_i 发生，结果B发生的概率（因→果）和在不知道原因 A_i 是否发生的情况下，结果B发生的概率（即单纯的“果”）。这就是贝叶斯公式(Bayes' theorem)。

定理 2.2 (贝叶斯公式)

设样本空间为 Ω ，事件 $B \in \Omega$ ，事件 A_1, A_2, \dots, A_n 互不相容，且 $\Omega = \bigcup_{i=1}^n A_i$ 。若 $P(B) > 0$ 且 $P(A_i) > 0, i = 1, 2, \dots, n$ ，则有

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)}, i = 1, 2, \dots \quad (2.5)$$



*这里，我们称 $P(A_i)$ 为先验概率(prior probability)，即根据以往的经验分布和分析，随机实验前就能计算出的某一原因的概率； $P(B|A_i)$ 为后验概率(posterior probability)，即已经知道结果B发生了，想要考察是由事件 A_i 引起的概率。

以上为最基础的事件概率的定义和相关定理。但由于样本空间千变万化，有数值型，也有文字描述型。对于文字描述型，我们不太好去刻画事件发生的概率大小，比如，你想要去考察你今天出门遇到堵车的概率，这时候单单用文字描述是不方便建立模型进行研究的，需要对这个事件进行抽象来建模。为此我们进一步定义了随机变量。随机变量可以把它简单理解成是一个随机实验结果的集合 \mathcal{F} 到实数域 \mathbb{R} 的实值单值函数，或者更简单的对随机结果的数值概括。

定义 2.4 (随机变量)

设随机实验的样本空间为 Ω ， $\omega \in \Omega$ 为其中的样本点或者基本事件，某个实验所有可能的结果为事件集 \mathcal{F} 。对于每个事件我们定义实值单值函数 $X = X(\omega)$ ，且对 $\forall x \in \mathbb{R}$ ，有 $\{\omega : X(\omega) < x\} \in \mathcal{F}$ 。则我们将X称为随机变量。



当然，这里的“事件集”定义较为模糊，但是它在测度论中对其有更为严格的规定，要求这里所谓的“事件集”是一个 σ 代数。由于计量中不会探讨这么复杂的问题，我们对此不再过多赘述。

有了基本的概率框架后，我们定义了分布函数 $F(x)$ ，并将随机变量分为了离散型变量和连续型变量。离散型随机变量是仅取离散值，如 $0, 1, 2, \dots$ ；连续型随机变量为取一系列连续的可能的值，如 $[0, 1], [0, \infty)$ 等，并且对于连续型随机变量，存在非负函数 $f(x)$ 使得 $\forall x \in \mathbb{R}$ ，有 $F(x) = \int_{-\infty}^x f(t)dt$ ，这里的 $f(x)$ 即是概率密度函数。

另外，类似于事件，随机变量有时也有相互独立的关系。

定义 2.5 (随机变量的独立性)

对于n个随机变量 $\{\xi_i\}_{i=1}^n$, 对 $\forall x_i \in \mathbb{R}, i = 1, 2, \dots, n$, 有

$$P\{\xi_1 < x_1, \dots, \xi_n < x_n\} = P\{\xi_1 < x_1\} \cdots P\{\xi_n < x_n\} \quad (2.6)$$

则我们称 $\xi_1, \xi_2, \dots, \xi_n$ 相互独立。



当然我们也可以类似的扩展成多维随机变量（或者说随机向量）的定义。这里不再展开。但是我们更需要知道的是多维随机变量的独立性性质。

定理 2.3 (随机向量的分量独立性)

若有m维随机变量 $X = (X_1, X_2, \dots, X_m)$ 和n维随机变量 $Y = (Y_1, Y_2, \dots, Y_n)$ 相互独立，则

- 对于 $\forall i = 1, 2, \dots, m$ 和 $\forall j = 1, 2, \dots, n$, X_i, Y_j 相互独立
- 对于连续函数 h, g , $h(X_1, X_2, \dots, X_m)$ 和 $g(Y_1, Y_2, \dots, Y_n)$ 相互独立



2.1.2 数字特征

在构建好概率框架后，我们其实就能根据分布函数来研究出随机变量的性质了。但实际上，很多随机变量的分布函数并不能通过简单的观察概率密度图就能知道他的分布函数长啥样，并且我们有时也不需要知道他的分布函数，只需要知道它的一些特征就够了。随机变量的数字特征，主要是一阶矩（期望）、二阶矩（方差和协方差）、三阶矩（偏度）、四阶矩（峰度）。我们依次介绍如下。

2.1.2.1 一阶矩：期望

首先对于一阶矩，即期望，我们定义如下：

定义 2.6 (期望)

对于离散型随机变量X, 假设他的分布律为 $P(X = x_k) = p_k, k = 1, 2, 3, \dots$, 若级数 $\sum_{k=1}^{\infty} |x_k| p_k < \infty$, 则我们称此级数为随机变量X的期望, 记为 $E(X)$, 即

$$E(X) = \sum_{k=1}^{\infty} x_k p_k \quad (2.7)$$

若级数 $\sum_{k=1}^{\infty} |x_k| p_k < \infty$ 不存在, 则我们称X的期望不存在。类似的, 对于连续型随机变量X, 设它的概率密度函数为 $f(x)$, 若积分 $\int_{-\infty}^{\infty} |x| f(x) dx < \infty$, 则我们称此积分为随机变量X的期望, 即

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \quad (2.8)$$

同样的, 若积分 $\int_{-\infty}^{\infty} |x| f(x) dx < \infty$ 不存在, 则我们称X的期望不存在。



对于随机变量的期望, 我们可以简单理解成是概率意义上的加权平均, 为此我们有时在不至于混淆的情况下, 把期望称为是均值。一般而言, 若是我们知道一个随机变量的分布, 我们对随机变量的预测即是期望。

定理 2.4 (函数期望)

设X为随机变量, $g(X)$ 为随机变量的函数。若X为离散型, 分布律 $P(X = x_k) = p_k$, 有

$$E[g(X)] = \sum_{k=1}^{\infty} g(x_k) p_k \quad (2.9)$$

若X为连续型，密度函数为 $f(X)$ ，有

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx \quad (2.10)$$

另外，条件期望也需要知道定义。

定义 2.7 (条件期望)

对于离散型随机变量Y，设在给定随机变量X下，已知条件分布律为 $P(Y = y_k|X) = p_k, k = 1, 2, 3, \dots$ ，若级数 $\sum_{k=1}^{\infty} |y_k|p_k < \infty$ ，则我们称此级数为随机变量X的条件期望，记为 $E(Y|X)$ ，即

$$E(Y|X) = \sum_{k=1}^{\infty} y_k P(Y = y_k|X) \quad (2.11)$$

若级数 $\sum_{k=1}^{\infty} |y_k|p_k < \infty$ 不存在，则我们称X的条件期望不存在。类似的，对于连续型随机变量Y，设它的概率密度函数为 $f(y|x)$ ，若积分 $\int_{-\infty}^{\infty} |y|f(y|x)dy < \infty$ ，则我们称此积分为随机变量X的期望，即

$$E(Y|X) = \int_{-\infty}^{\infty} yf(y|x)dy \quad (2.12)$$

同样的，若积分 $\int_{-\infty}^{\infty} |y|f(y|x)dy < \infty$ 不存在，则我们称X的条件期望不存在。

对于条件期望，你可以把它理解成是关于X的函数。即 $E(Y|X) = g(X)$ 。

同时，对于期望，我们有以下性质：

1. a, b, c 为常数， X, Y 为随机变量，则 $E(aX + bY + c) = aE(X) + bE(Y) + c$
2. 若随机变量 X, Y 相互独立，则 $E(XY) = E(X)E(Y), E(Y) = E(Y|X)$
3. 条件均值迭代法则：对随机变量X，Y 有 $E(Y) = E[E(Y|X)]$ 。
4. 若有随机变量 X_1, X_2, Y 且 X_2, Y 相互独立，则 $E(Y|X_1) = E(Y|X_1, X_2)$
5. 若一系列随机向量 $(X_i, Y_i)_{i=1}^n$ 相互独立，则有 $E(Y_i|X_1, X_2, \dots, X_n) = E(Y_i|X_i)$
6. (Cauchy-Schwarz Inequality): $[E(XY)]^2 \leq E(X^2)E(Y^2)$ ，如果令 $Y=1$ ，则有 $[E(X)]^2 \leq E(X^2)$

注意，性质1在任意情况下都能成立，性质2仅在独立情况下成立。性质3极为常用，在证明系数估计无偏性时有很大作用。性质4也比较常用，理解这条其实只要回顾一下独立随机向量的分量独立性：对于 $\forall j \neq i$ ， (X_i, Y_i) 和 (X_j, Y_j) 相互独立，则 Y_i, X_j 相互独立，即 $E(Y_i|X_1, X_2, \dots, X_n) = E(Y_i|X_i)$

2.1.2.2 二阶矩：方差和协方差

二阶矩主要是方差和协方差。我们分别定义如下：

定义 2.8 (方差和协方差)

设X为随机变量，若期望 $E[(X - E(X))^2]$ 存在，则称该期望为X的方差，记为 $Var(X)$ 或者 σ_X^2 。即

$$Var(X) = E[(X - E(X))^2] \quad (2.13)$$

其中， $\sigma_X = \sqrt{Var(X)}$ 称作是X的标准差。

若Y也为随机变量，且期望 $E[(X - E(X))(Y - E(Y))]$ 存在，则称该期望为X和Y的协方差，记为 $Cov(X, Y)$ 。即

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] \quad (2.14)$$

可以看到，方差其实是协方差的一种，即 $Var(X) = Cov(X, X)$ 。

另外还有条件方差定义如下：

定义 2.9 (条件方差)

对于离散型随机变量Y，设在给定随机变量X下，条件期望 $E[Y - E(Y|X)]^2$ 存在，则我们称此条件期望为随机变量X的条件方差，记为 $Var(Y|X)$ ，即

$$Var(Y|X) = E\{[Y - E(Y|X)]^2|X\} \quad (2.15)$$

为了反映X, Y之间的线性相关性，我们定义了相关系数如下：

定义 2.10 (相关系数)

设X, Y为随机变量，我们定义相关系数为

$$corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \quad (2.16)$$

其中， $corr(X, Y)$ 有时也记做 $\rho_{X,Y}$ 。

对于方差、协方差和相关系数，我们有以下性质：

1. a, b, c 为常数， X, Y, Z 为随机变量， $Var(aX + bY) = a^2 \sigma_X^2 + b^2 \sigma_Y^2 + 2\rho_{X,Y}\sigma_X\sigma_Y$, $Cov(aX + bY + c, Z) = aCov(X, Z) + bCov(Y, Z)$
2. X, Y 是随机变量，则 $E(XY) = Cov(X, Y) + E(X)E(Y)$, $E(X^2) = Var(X) + [E(X)]^2$
3. 若随机变量X, Y相互独立，则 $Var(X + Y) = Var(X) + Var(Y)$
4. 若 $E(Y|X) = 0$ ，则 $Cov(X, Y) = corr(X, Y) = 0$
5. 条件方差迭代法则：对于随机变量X, Y，在任意情况下都有： $Var(Y) = Var[E(Y|X)] + E(Var(Y|X))$
6. * $Var(X) = 0$ 等价于 $P(X = E(X)) = 1$

其中，性质1-3在为条件均值或者方差时也成立。

对于性质4，反之不一定成立，比如假设X服从正态分布， $Y = X^2$, $Cov(X, Y) = E(XY) - E(X)E(Y) = E(X^3) - E(X)E(X^2) = 0 - 0 * 1 = 0$ ，但是 $E(Y|X) = X^2 \neq 0$ 。

对于性质5，这条定理主要是在后面的OLS同方差推导会用到，别的用的不多。

性质6了解即可。

对于相关系数，我们有如下说明：

1. $-1 \leq \rho_{X,Y} \leq 1$
2. 若 $\rho_{X,Y} > 0$ ，则称X, Y为正相关；若 $\rho_{X,Y} < 0$ ，则称X, Y负相关；若 $\rho_{X,Y} = 0$ ，则称X, Y不相关
3. 若 $\rho_{X,Y} = 1$ ，则称X, Y完全正相关；若 $\rho_{X,Y} = -1$ ，则称X, Y完全负相关；
4. 相关系数对线性变换不受影响，即对于任意常数a, b和随机变量X, Y有 $corr(aX + c, bY + d) = corr(X, Y)$
5. 独立相关系数必为0，相关系数不为0一定不独立。

2.1.2.3 *三阶矩：偏度

偏度是对数据在分布图上倾斜程度的度量，是统计数据非对称分布的特征。我们对偏度定义如下：

定义 2.11 (偏度)

设X为随机变量，若期望 $E[(X - E(X))^3]$ 存在，则称该期望除以标准差的三次方为偏度，即

$$Skewness(X) = \frac{E[X - E(X)]^3}{\sigma_X^3} \quad (2.17)$$

当偏度大于0时，我们称数据为正偏或者右偏。当偏度小于0时，我们称数据为负偏或者左偏。当偏度等于0时，数据呈对称分布。

另外，关于左偏和右偏的区分，其实主要看的是左右尾巴厚度的大小比较。如图2.1 和图2.2，当数据为右偏时，数据右边尾巴往往较厚，数据为左偏时，数据左边尾巴往往较厚。

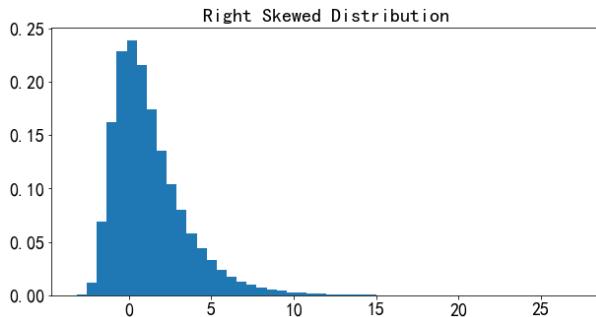


图 2.1: 右偏分布

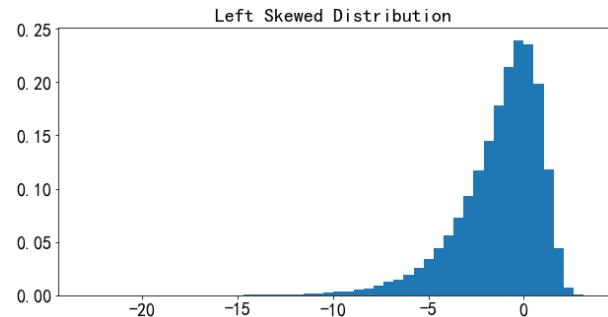


图 2.2: 左偏分布

2.1.2.4 *四阶矩: 峰度

峰度为对数据尾部大小的衡量。定义如下

定义 2.12 (峰度)

设 X 为随机变量, 若期望 $E[(X - E(X))^4]$ 存在, 则称该期望除以标准差的四次方为偏度, 即

$$Kurtosis(X) = \frac{E[X - E(X)]^4}{\sigma_X^4} \quad (2.18)$$



一般而言, 我们计算峰度往往将数据与正态分布进行比较。若峰度大于三, 则我们称数据为厚尾分布(fat-tailed), 若峰度小于三, 我们称数据为细尾分布(thin-tailed)。我们所熟知的t分布就是一种厚尾分布。

对于经济金融数据, 很多数据都会出现厚尾特征。厚尾意味着大幅偏离均值的极端值较多, 由于计量的目的是量化一般情况下的因果效应而非预测, 极端数据并不能给我们提供帮助, 有时反倒有害(但时间序列预测则大不相同)。后面对此会有更进一步的说明。

2.1.3 *高阶矩和低阶矩的进一步说明

事实上, 若随机变量的高阶矩存在, 那么低阶矩一定存在。反之则不一定成立。对于前者, 可以考虑Hölder不等式: 对 $\forall p, q > 1, f \in L^p(\Omega), g \in L^q(\Omega)$, 则有

$$\int_{\Omega} f(x)g(x)dx \leq (\int_{\Omega} |f(x)|^p dx)^{\frac{1}{p}} (\int_{\Omega} |g(x)|^q dx)^{\frac{1}{q}} \quad (2.19)$$

对 $\forall a \geq 1, p > 1, q > 1$, 有

$$E(X^a) = \int_{\Omega} x^a * 1dF(x) \leq (\int_{\Omega} x^{ap} * dF(x))^{\frac{1}{p}} (\int_{\Omega} 1^q * dF(x))^{\frac{1}{q}} = (\int_{\Omega} x^{ap} * dF(x))^{\frac{1}{p}} \quad (2.20)$$

令 $b = ap > 1$, 则有

$$E(X^a) \leq (\int_{\Omega} x^{ap} * dF(x))^{\frac{1}{p}} = [E(X^b)]^{\frac{a}{b}} \quad (2.21)$$

显然, 当 $E(X^b)$ 存在时, $E(X^a)$ 也存在, 证毕。

2.1.4 几个分布

在初级计量当中有两大基础分布, 一是离散型的Bernoulli分布, 二是连续型的正态分布。Bernoulli分布较为简单, 其也被称作是两点分布或者0-1分布。

2.1.4.1 Bernoulli分布

定义 2.13 (Bernoulli分布)

设随机变量X仅取0或者1，且对于 $0 < p < 1$ ，有

$$P(X = k) = \begin{cases} p & k=0 \\ 1-p & k=1 \end{cases} \quad (2.22)$$

或写成简单粗暴的形式： $P(X = k) = p^k(1-p)^{1-k}$, $k = 0, 1, 0 < p < 1$. 则称随机变量X服从Bernoulli分布。



注意，这里不是所谓的二项分布。二项分布是n重Bernoulli实验的分布，即Bernoulli分布的衍生。初级计量里没有涉及到二项分布及其衍生，故仅对Bernoulli分布介绍如上。

2.1.4.2 正态分布及其衍生分布

正态分布是计量中出现最多的一个分布，所以必须对他的定义和性质均有所了解。

定义 2.14 (正态分布)

设随机变量X服从均值为 μ ，方差为 σ^2 的概率分布，且其概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (2.23)$$

则称X服从一维正态分布，记为 $X \sim N(\mu, \sigma^2)$ 。



当然，正态分布也可以拓展成多维形式。在计量中，正态分布的概率密度函数可能并不是特别重要。重要的是他的相关性质，列举如下：

- 若n个随机变量服从多维正态分布，则其边缘分布均为一维正态分布
- 若n个随机变量服从多维正态分布，则他们分量的线性组合依旧为正态分布
- 若n个随机变量服从多维正态分布，且其协方差矩阵为对角阵（即两两随机变量的协方差为0）。则n个随机变量相互独立。反之也成立。
- 若n个随机变量均服从正态分布，且相互独立。他们服从联合正态分布
- 标准正态分布的k阶矩公式：

$$E(X^k) = \begin{cases} 0 & k \text{ 为奇数} \\ (k-1)(k-3)\cdots 1 & k \text{ 为偶数} \end{cases} \quad (2.24)$$

上述定理可被概述为如下导图：对于第四条，仅需掌握前四阶矩即可。



图 2.3: 正态分布性质

正态分布还需要掌握他的衍生分布的相关性质，他的衍生分布为 χ 分布， t 分布， F 分布。各个分布的密度

函数无需掌握，但要知道他们如何从正态分布衍生而来。

定义 2.15 (卡方分布)

设随机变量 X_1, X_2, \dots, X_n 相互独立，且均服从标准正态分布 $N(0, 1)$ ，则我们称随机变量 $\chi^2 = \sum_{i=1}^n X_i^2$ 服从自由度为 n 的卡方分布，记为 $\chi^2 \sim \chi^2(n)$



卡方分布需要知道以下性质：

- 可加性： $\chi_1^2 \sim \chi^2(n_1), \chi_2^2 \sim \chi^2(n_2)$ ，则有 $\chi^2 = \chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$.
- 期望和方差： $\chi^2 \sim \chi^2(n)$ ，则有 $E(\chi^2) = n, Var(\chi^2) = 2n$

第二条可以作为期望方差练习留给各位证明。

卡方分布的概率密度函数图像如图2.4。

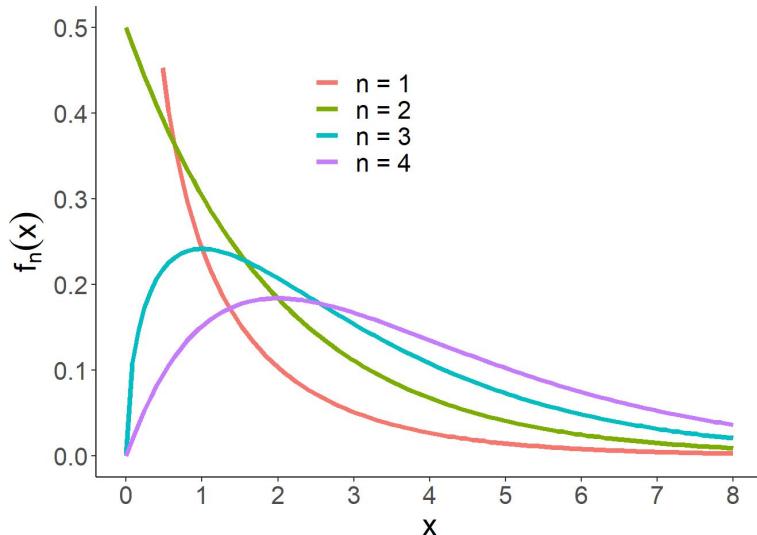


图 2.4: 卡方分布密度函数

接着是t分布。t分布主要是小样本分布。

定义 2.16 (t分布)

设随机变量 X, Y 相互独立，且 $X \sim N(0, 1)$, $Y \sim \chi^2(n)$ 。则我们称随机变量 $t = \frac{X}{\sqrt{Y/n}}$ 服从自由度为 n 的t分布，记为 $t \sim t(n)$



t分布需要知道他的正态近似：当 n 趋近于无穷时，t分布将为正态分布。证明是利用大数定律：由于随机变量 $X_1^2, X_2^2, \dots, X_n^2$ 独立同分布，且具有相同的期望 0，则当 n 很大时，他们的均值依概率收敛到 0。为此，当样本数据量够大时，我们常常用正态分布作为 t 分布的代替。

另外，厚尾的特征也可以从图2.5看出。当自由度较小时，t分布的两边的尾部明显要大于正态分布（即 $n \rightarrow \infty$ 时的t分布）。

最后是F分布。这个分布仅需大概掌握定义及其近似即可。

定义 2.17 (F分布)

设随机变量 U, V 相互独立，且 $U \sim \chi^2(m)$, $V \sim \chi^2(n)$ 。则我们称随机变量 $F = \frac{U/m}{V/n}$ 服从第一自由度为 m ，第二自由度为 n 的F分布，记为 $F \sim F(m, n)$



F分布的卡方近似：在 $n \rightarrow \infty$ 的情况下， $V/n \rightarrow 1$ 则有 $mF \rightarrow \chi^2(m)$

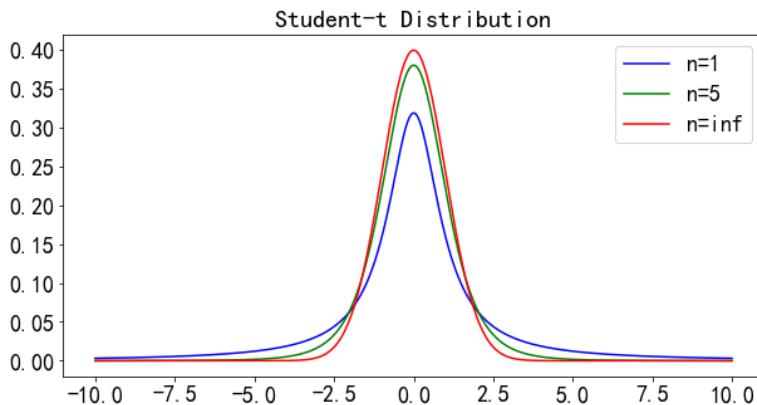


图 2.5: t 分布密度函数

2.1.5 大数定律和中心极限定理

概率论中的大数定律和中心极限定理有好多个，这里计量仅需掌握Khinchin大数定律和独立同分布的中心极限定理。

计量中有两个定理会较为常用：大数定律和中心极限定理。前者讲述的是独立同分布的随机变量的样本均值会随样本量增大依概率收敛到总体均值（即期望），后者讲述的是对于任意独立同分布的随机变量的样本均值会随着样本量增大分布收敛到正态分布。由于样本量无限时，方差为0，此时有 $P(X = \mu) = 1$ ，为此在记忆时可以看成前者近似是后者的一个推论。

定理 2.5 (大数定律)

设随机变量 X_1, X_2, \dots, X_n 独立同分布，且具有相同数学期望 $E(X_i) = \mu$ 。则对 $\forall \epsilon > 0$ 有

$$\lim_{n \rightarrow \infty} P\left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| < \epsilon \right\} = 1 \quad (2.25)$$



事实上，你会注意到辛钦大数定律甚至不需要方差有限的条件。证明不需要掌握，感兴趣可自行查阅概率论教材。

定理 2.6 (中心极限定理)

设随机变量 X_1, X_2, \dots, X_n 独立同分布，且具有相同数学期望 $E(X_i) = \mu$ ，方差 $Var(X_i) = \sigma^2 > 0$ ，则有

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} N(0, 1) \quad (2.26)$$



两个定理都比较常用。前者一般来说是在基础的一致性证明中出现比较多，比如方差、协方差这些基础统计量的一致性。后者相对更为常用点，主要是需要证明系数的渐进正态性会用。

2.2 统计学知识

如果说前面的概率论测度了事件发生的可能性，那么统计学则是将理论模型和真实数据联系起来，度量理论模型的准确性。为此从这个角度上看，统计学比概率论还更为重要。

2.2.1 几个概念

首先是需要知道一些统计相关的概念。

总体和样本 这是计量中最为基础也是最重要的概念。总体是包含所有研究的个体的集合，由于我们不能观测到全部的个体，只能抽取出来一部分。从总体中抽取出来的就是我们所观察到的样本。比如，你想度量今天食堂的饭菜总体质量，那你肯定是挑几个窗口都去吃一口看看质量如何，而不是把食堂今天提供的所有饭菜都吃光。那么这时候，食堂提供的全部的饭菜就是总体，你吃的那几口就是样本。我们是通过样本的结果外推到总体进而得出结论。为此，由于“抽样”这个事件本身具有随机性，所以我们所抽取的样本是个随机变量。

简单随机样本 简单随机样本指的是通过简单随机抽样得到的样本，这个样本必须满足独立同分布(i.i.d, independent identical distribution)的性质。还是用上面的那个例子，对于一个简单随机抽样的吃法，你不能吃了窗口A后，觉得难吃想早点结束抽样，直接继续吃旁边的窗口B，即样本之间会具有相关性，这是不允许的。你必须是通过电脑模拟或者掷骰子的方式去找到你所抽取的样本。初级计量主要是截面回归，一般都为简单随机样本。

统计量 由于我们观测到的样本实质上是个n维向量，换言之，你有2000个样本，那么你所观测到的就是一个2000维的样本向量，但事实上直接去看这2000维样本向量看不出啥东西来，我们必须利用数学工具去找出这2000维样本向量的特征，这个数学工具就是统计量：设 X_1, X_2, \dots, X_n 为来自总体X的样本， $g(X_1, X_2, \dots, X_n)$ 为样本的函数且不含有任何未知参数，则 $g(X_1, X_2, \dots, X_n)$ 是统计量。

估计值和估计量 设我们有样本 X_1, X_2, \dots, X_n ，样本值为 x_1, x_2, \dots, x_n ，统计量 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 为估计量，他的观测值 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 为估计值。换言之，估计量是个随机变量，估计值是个数。

2.2.2 估计量评价标准

样本只有一个，但由于函数的任意性，估计量可以有无数个。为此我们必须有估计量的评价标准，以挑选出我们认为最好的估计量。这些标准分别是无偏性、一致性、有效性。

假设我们的待估计的参数为 θ :

无偏性 设估计量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 的数学期望存在，且 $\forall \theta \in \Theta$ 均有

$$E(\hat{\theta}) = \theta \quad (2.27)$$

则我们称该估计量具有无偏性。

一致性 设参数 θ 的估计量是 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ ，若 $\forall \theta \in \Theta$ 有

$$\lim_{n \rightarrow \infty} P\{|\hat{\theta} - \theta| < \epsilon\} = 1 \quad (2.28)$$

或者写成 $\hat{\theta} \xrightarrow{P} \theta$ ，则我们称 $\hat{\theta}$ 具有一致性。

有效性 设参数 θ 的无偏估计量分别为 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ ，若 $\forall \theta \in \Theta$ 有

$$Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2) \quad (2.29)$$

则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效。

另外关于一致性有如下说明：

- 我们唯一学过的“一致收敛”只在大数定律里出现过。换言之在我们的知识范围内根本上只有样本均值是一致收敛的。
- 连续函数的概率收敛定理：若有 $X_n \xrightarrow{P} a, Y_n \xrightarrow{P} b$ ，又有函数 $g(x, y)$ 在 $g(a, b)$ 处连续，则我们有 $g(X_n, Y_n) \xrightarrow{P} g(a, b)$ 。该条定理极容易遗漏，初级计量中几乎所有估计量的一致性证明都会用到该定理。

现在来考虑一个例子。

例题 2.1 设从总体 X 抽出的简单随机样本为 X_1, X_2, \dots, X_n ($n > 1$)，待估参数为总体均值 μ_X ，试问估计量 \bar{X} 和估计量 X_1 的无偏性、一致性、有效性？

解

无偏性 $E(\bar{X}) = \frac{1}{n}E(\sum_{i=1}^n X_i) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{1}{n}\sum_{i=1}^n \mu_X = \mu_X$ 。注意，样本为简单随机样本， X_1, X_2, \dots, X_n 独立同分布，为此他们都具有相同的均值。根据定义这两个估计量都具有无偏性。

一致性 一致性的要求是：随着样本量的增大，估计量依概率收敛到对应的参数。对于样本均值，由大数定律他一定能收敛到总体均值， \bar{X} 一致性条件满足。但对于 X_1 ，无论样本量多大，他始终还是 X_1 ，这玩意一直是个随机变量，压根不能依概率收敛，更别提收敛到待估参数 μ_Y 。所以 \bar{X} 具有一致性， X_1 不具有一致性。

有效性 $Var(\bar{X}) = \frac{1}{n^2}Var(\sum_{i=1}^n X_i)$ ，由于 X_1, X_2, \dots, X_n 独立同分布，则 $Var(\bar{X}) = \frac{1}{n^2} * nVar(X) = \frac{1}{n}Var(X)$ ，而 $Var(X_1) = Var(X)$ ，显然有 $Var(\bar{X}) < Var(X_1)$ ， \bar{X} 比 X_1 有效

2.2.3 常见统计量及常见的抽样分布

先说常见统计量。对于样本 X_1, X_2, \dots, X_n 有

- 样本均值 $\bar{X} = \sum_{i=1}^n X_i$
- 样本方差 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
- 样本标准差 $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$
- 样本协方差 $S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$

其中，-1减去的是均值 \bar{X} 自由度。对于不同的情况，标准差需要减去的自由度不同。

*这里，自由度更准确的理解是“自由维度”，即取值不受约束的变量个数，或者通俗点是变量个数减去约束等式个数。比如卡方分布，他有 n 个 X_i ，可变化的变量有 n 个，所以自由度是 n 。对于样本方差，因为他考虑的是离均值的变化情况，计算方差时候均值必须给出，这里就蕴含了 $\sum_{i=1}^n X_i = \bar{X}$ 这个等式，所以自由度为 $n-1$ 。当然你甚至还能从啥二次型的秩的角度去看自由度的定义，不过那个不好理解这里不再展开。

可以证明，上述统计量分别一致收敛到总体均值、方差、标准差、协方差。各位可以留作一致性证明的习题作为练习。

另外，还有个统计量叫做标准误。他其实是均值的标准差。由于在有相同期望，方差有限的*i.i.d*样本中，均值 $\bar{X} \stackrel{d}{\sim} N(\mu_X, \frac{\sigma_X^2}{n})$ 。则我们将 $SE(\bar{X}) = \frac{\sigma_X}{\sqrt{n}}$ 的估计量定义为标准误。

现在来说说抽样分布。初级计量除了以上介绍的三大正态衍生分布都是来源于正态总体的抽样分布—— χ^2 分布，t分布，F分布之外，还需要掌握以下从正态分布衍生而来的抽样分布。

- 单样本：设样本 X_1, X_2, \dots, X_n 来源于 $N(\mu, \sigma^2)$ ，则对均值 \bar{X} 有 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ ，和 $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$
- 双样本：设样本 X_1, X_2, \dots, X_{n_1} 来源于 $N(\mu_1, \sigma_1^2)$ ， Y_1, Y_2, \dots, Y_{n_2} 来源于 $N(\mu_2, \sigma_2^2)$ ，则对均值 \bar{X}, \bar{Y} 之差有： $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$ ，和 $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(n_1+n_2-2)$ 。若有 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ，则上述式子可被简写为 $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1+n_2-2)$ ，其中 $S_w = \frac{(n_1-1)S_1^2 + (n_2-2)S_2^2}{n_1+n_2-2}$

其中，将总体标准差换成样本标准差后，从正态分布到t分布的原因是正态分布的标准差 S 满足 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ ，且有 \bar{X} 和 S 相互独立， $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} / \sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

2.2.4 参数估计方法

介绍完最为基础的概念。我们来说说如何估计参数。初级计量中对于回归方程有两种参数估计方式:最小二乘法 (Ordinary Least Squares, OLS) 和极大似然估计 (Maximum Likelihood Estimate, MLE)。

对于OLS, 其主要是针对线性回归方程进行参数估计。比如我们有 $Y_i = \beta_0 + \beta_1 X_i + u_i$ 这个回归方程, 所要做的便是通过调整参数 (b_0, b_1) 最小化残差平方和 $Q(b_0, b_1) = \sum_{i=1}^n \hat{u}_i^2$, 最小化后的得到的参数 (b_0, b_1) 即是是我们估计出来的 $(\hat{\beta}_0, \hat{\beta}_1)$ 。由于初级计量大多数都是线性方程 (不是线性也能转化成线性的), OLS是最为常用的估计方法。

对于MLE, 其思想主要是去最大化当前观测到的样本值出现的概率。分为以下步骤:

1. 写出似然函数 $L(\theta; X_1, X_2, \dots, X_n)$, 这个函数是关于随机变量 X_1, X_2, \dots, X_n 的联合分布, 但同时本质上是一个关于参数 θ 的函数。对于独立同分布的情况, 若为离散型, 我们可以写成 $L(\theta) = L(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta) = \prod_{i=1}^n P(X_i = x_i; \theta)$, 对于连续型, 假设概率密度函数为 $f(x; \theta)$, 则 $L(\theta) = \prod_{i=1}^n f(x_i; \theta)$
2. 由于概率值取值范围为 $[0, 1]$, 若样本数量较大时, 似然函数可能会非常接近于0, 为了方便求解我们通常取对数。即计算 $\ln L(\theta) = \sum_{i=1}^n \ln P(X_i = x_i; \theta)$ 或者 $\ln L(\theta) = \sum_{i=1}^n \ln f(x_i; \theta)$
3. 接着最大化对数似然函数 $\ln L(\theta)$, 即 $\max_{\theta} \ln L(\theta)$ 。一般来说, 我们是采取让其各分量的偏导等于0 (即 $\frac{\partial \ln L(\theta)}{\partial \theta} = 0$) 进行计算。但实际当中这个偏导可能不好求, 往往会采取梯度下降法或者牛顿法等数值方法去求解。求解得到的 $\hat{\theta}$ 即为所求。初级计量中一般要求写到偏导等于0即可, 其余不做要求。

2.2.5 置信区间和假设检验

2.2.5.1 置信区间

前面的参数估计主要是点估计, 他给的是未知参数的估计值, 但是没有给相应的精度。为此有了置信区间。

定义 2.18 (置信区间)

设总体X的分布函数为 $F(x; \theta)$, 待估参数为 $\theta \in \Theta$, $\forall \alpha \in (0, 1)$, 若有两个统计量 $\hat{\theta}_L = \hat{\theta}_L(X_1, X_2, \dots, X_n)$ 和 $\hat{\theta}_U = \hat{\theta}_U(X_1, X_2, \dots, X_n)$ (其中 $\hat{\theta}_L < \hat{\theta}_U$), $P(\hat{\theta}_L < \theta < \hat{\theta}_U) \geq 1 - \alpha$, 则我们称随机区间 $[\hat{\theta}_L, \hat{\theta}_U]$ 为置信水平 $1 - \alpha$ 的置信区间。



需要注意的是, 由于这里有两个参数 $\hat{\theta}_L$ 和 $\hat{\theta}_U$, 给定了置信水平只是控住了 $\hat{\theta}_L$ 和 $\hat{\theta}_U$ 之间在概率上的"相对距离", 故置信区间本质上可以有很多个, 我们只是选了区间长度最短的那个。

构造置信区间的方法为枢轴量法:

1. 构造样本 X_1, X_2, \dots, X_n 的函数 $W = W(X_1, X_2, \dots, X_n; \theta)$, 其中W的分布已知, 且该分布不依赖于未知参数 θ
 2. 适当选取两个常数a, 使得对于给定的置信水平 $1 - \alpha$ 有 $P(a \leq W(X_1, X_2, \dots, X_n; \theta) \leq b) = 1 - \alpha$
 3. 经过对上述不等式等价变换得到 $P(\hat{\theta}_L < \theta < \hat{\theta}_U) \geq 1 - \alpha$, 其中 $[\hat{\theta}_L, \hat{\theta}_U]$ 即为置信度 $1 - \alpha$ 置信区间
- 现在看个例子。

例题 2.2 设总体X已知, 且有 $X \sim N(\mu, \sigma^2)$, 其中参数 μ 是未知参数, σ^2 为已知参数。设 X_1, X_2, \dots, X_n 为样本, 试求置信度为 $1 - \alpha$ 的置信区间。

解 由题知, 统计量 \bar{X} 为 μ 的无偏一致估计, 且由中心极限定理: $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, 为此我们有

$$P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}\right) = 1 - \alpha$$

经过等价变换有

$$P\left(\bar{X} - \frac{\sigma}{\sqrt{n}}z_{\alpha/2} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right) = 1 - \alpha$$

此时获得置信区间($\bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$)

2.2.6 假设检验

前面简单介绍了估计量的相关概念，什么样的估计量是个好的估计量，以及估计量的精度有多大。那么现在就有了个问题：假设你对估计量有个事先的预估值，那么你想检验你的假设是否正确，应该怎么做呢？这就是假设检验的提出。假设检验本身的思想是“小概率事件不太可能发生”，当我们在原假设下通过样本计算的统计量比较极端（原假设成立下，本不太可能发生的小概率事件发生了），那么我们可以拒绝相应的原假设。

讲假设检验前，首先明白几个概念。

原假设和备择假设 原假设 H_0 一般来说是某一个观点、某个社会常识、大概率发生的事件（也就是说，肯定句并不一定是原假设）。比如，把学校里所有大肚子的人抓起来去排队检查是否怀孕，但事实上怀孕本身概率就不大，更多是吃太肥了才导致看上去跟怀孕一样（图 2.6），为此原假设就是：这个人没有怀孕。备择假设 H_1 一般来说，是原假设的反面。

在初级计量中，一般原假设都是系数为0或者等于某个数，所以这块内容如果实在理解不了，硬背“系数为0或者等于某个数就是原假设”其实也行。但如果你后续课程还需要用到计量的话，最好还是尝试在学习过程中理解这点。



图 2.6: 原假设的确立

两类错误 一般而言，对于我们所做的假设会犯两种错误——第I类错误：假设是对的，但是你却拒绝了它（弃真）。第II类错误：假设是错的，但是你却接受了它（取伪）。理论研究表明，当样本容量增大时两类错误都会有所减小，但是在样本容量固定的情况下，当你减小其中一类犯错误的概率，另一类犯错误的概率也会增大。根据前面所述，原假设一般被确立为某个观点或者某个比较可能发生的情况，考虑到犯错误的成本，我们往往会减小第一类错误的概率。

同样，还是考虑上文的例子，如果说在相同样本容量下，你只控制了第二类错误减小而使得第一类错误增大，那么你将检验出一大帮胖子怀孕，这个代价（尤其是男人怀孕）明显比没有检查出怀孕的大得多。具体可以看图2.7。

当然并不是说第二类错误就不重要，只是一般来说，第一类错误的代价比较高，所以我们才会选择性“放弃”控制第二类错误。

显著性水平 显著性水平 α 是原假设为真下你可能犯错误的概率。初级计量中一般来说都取0.05（根据要求可能会取0.01）。结合前面置信区间的概念，你可以把它理解为掉出置信区间的概率，换言之，对于一个概率密度函

数下的面积，往往由显著性水平 α 和置信水平 $1 - \alpha$ 构成。



图 2.7: 第一类错误的成本和第二类错误的成本

假设检验的步骤如下。

1. 确立原假设 H_0 和备择假设 H_1 。这步非常关键，假设确立错了后面都错了。
2. 选择统计量确定拒绝域。拒绝域是统计量落入此区域后就拒绝原假设的集合，通俗来说拒绝域就是统计量非常极端的集合。比如，考虑 $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$ ，那么所谓的“极端集合”自然是在两个尾部，即在给定小概率 α 下， t 落入该区域的概率应该是 α ，即 $|t| = |\frac{\bar{X}-\mu}{S/\sqrt{n}}| > t_{\alpha/2}$ 。此时，结合假设检验的思想“小概率事件不太可能发生”，这个小概率 α 就是上文所说的显著性水平。
3. 给定显著性水平 α 做判断：一般来说有两种方式：

临界值法：临界值即是恰好拒绝原假设的统计值，将你计算的统计量和临界值进行比较进行判断。

p值法：p值是原假设为真的条件下，比所得到的样本观察结果更极端结果出现的概率。当 $p < \alpha$ ，说明原假设为真的条件下，所可能出现结果比样本观察结果还要极端的概率小于你预设的所能容忍的拒绝 H_0 犯错误的概率，这时候你可以放心拒绝原假设。

但是要注意的是，如果你不能拒绝原假设，那么你就老老实实写不能拒绝原假设，不要写接受原假设。因为你接受了原假设隐含了第二类错误，即“取伪”的错误。

目前在初级计量中，出现的检验有三种：t检验，F检验，J检验。t检验主要是为了检验回归方程里单个系数是否为0，F检验主要针对多个系数是否联合为0，J检验主要是针对工具变量回归中工具变量是否和误差项完全不相关的检验。对于最后一个检验，各位死记就行（原理比较复杂），但对于前两个要理解掌握。

下面我们看个例子。

例题 2.3 已知统计量 $X \sim N(\mu, \sigma^2)$ ，我们现在通过现有的样本计算出了 \bar{X} 和 X 的标准误 $\frac{S}{\sqrt{n}}$ ，试问如何检验 $\mu = 0$ ？
解

1. 原假设 $H_0 : \mu = 0$ ，备择假设 $H_1 : \mu \neq 0$
2. 我们知道了 \bar{X} 和标准误 $\frac{S}{\sqrt{n}}$ ，那么联想到的便是t统计量： $t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ ，由于t统计量服从t分布，那么拒绝域形式为自然就是为 $|t| = |\frac{\bar{X}}{S/\sqrt{n}}| > k$
3. 现在进行判断：临界值法：给定显著性水平 α ，恰好拒绝原假设的临界值为 $t_{\alpha/2}$ ，则当 $|t| > t_{\alpha/2}$ 时候拒绝原假设。p值法：p值是原假设为真的条件下，比所得到的样本观察结果更极端结果出现的概率，所以这里 $p = P\{t > \frac{\bar{X}-\mu}{S/\sqrt{n}}\}$ ，若 $p < \alpha$ 则拒绝原假设。

后续我们还会在对计量中的假设检验做进一步说明。

第3章 一元线性回归

3.1 一元线性回归模型概述

前面提过，计量经济学的目的是研究一个变量 X_i 的变化对另一个变量 Y_i 的影响，即 $Y_i = f(X_i) + u_i$ ，那么，在最为简单的情况下，我们假设函数 f 为线性函数，即

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i = 1, 2, \dots, n \quad (3.1)$$

如果是对上述总体回归模型进行估计，我们写成以下形式：

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \text{式中 } Y_i = \hat{Y}_i + \hat{u}_i \quad (3.2)$$

符号说明和系数解释如下。

- $(X_i, Y_i)_{i=1}^n$ 为样本，其中 X_i 为解释变量，又称为自变量、回归变量； Y_i 为被解释变量，又称为因变量
- u_i 为随机误差项，代表了除了 X_i 外能够影响 Y_i 的因素
- β_0 为总体回归方程的截距项
- β_1 为总体回归方程的斜率
- $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$ 被称作总体回归线，这里隐含了假设 $E(u_i|X_i) = 0$ （后续会对此进行说明）。并且通过这条总体回归线也可以得到 β_0 和 β_1 的含义：

如果 X_i 是非二值变量， β_1 是 X_i 变动一单位时平均对 Y_i 的影响；如果 X_i 是二值变量即 $X_i \in \{0, 1\}$ ，那么 β_1 系数解释为 $X_i = 1$ 和 $X_i = 0$ 时的 $E(Y_i|X_i)$ 之差

β_0 一般来说没有经济意义，但从数学意义上它代表了总体回归线的高度。如果 X_i 是二值变量， β_0 系数解释为 $X_i = 0$ 时候的均值。

- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ 是样本回归线。 \hat{Y}_i 是通过估计的回归方程和 X_i 估计出来的预测值， $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是通过回归方程估计出来的截距项和斜率， \hat{u}_i 是残差

着重要说明的是：

1. 总体回归线是我们所研究全部个体的回归线，但是由于我们拿不到全部的个体，只能通过抽取其中的一部分样本去估计出样本回归线，有时因为数据不符合模型假设可能会导致样本回归线与总体回归线差异过大。
2. 注意残差和随机误差项的符号，前者带帽，后者不带帽。残差项别写错，是 $\hat{u}_i = Y_i - \hat{Y}_i$
3. 总体回归线他的参数 β_0 和 β_1 是个常数，他代表了所研究的全部个体，但是 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是通过样本 $(X_i, Y_i)_{i=1}^n$ 估计出来的，而我们抽取到的样本是个随机变量，所以 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 也是随机变量，而不是完全确定的数。
4. 注意 β_1 系数解释，他刻画的是给定 X_i 情况下， X_i 所变化一单位， Y_i 的条件均值的变化，而不是单单 Y_i 的变化。因为所拿到的只有样本 $(X_i, Y_i)_{i=1}^n$ ，我们无法预测也无法计算得到随机误差项。

3.2 OLS估计

讲完了一元线性回归模型的表达式和系数解释，现在来看看模型的估计。一元线性回归模型一般采取最小二乘法（Ordinary Least Squares, OLS）进行估计，目标是使得拟合样本回归线后得到的残差平方和最小化，换言之即是：令 $Q(b_0, b_1) = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$

$$\min_{b_0, b_1} Q(b_0, b_1) = \min_{b_0, b_1} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2 \quad (3.3)$$

为了求最小值，对 $Q(b_0, b_1)$ 求偏导并令其等于0。即

$$\begin{cases} \frac{\partial Q}{\partial b_1} = -2 \sum_{i=1}^n (Y_i - b_0 - b_1 X_i) X_i = 0 \\ \frac{\partial Q}{\partial b_0} = -2 \sum_{i=1}^n (Y_i - b_0 - b_1 X_i) = 0 \end{cases} \quad (3.4)$$

化简得

$$\begin{cases} \sum_{i=1}^n X_i Y_i - b_0 \sum_{i=1}^n X_i - b_1 \sum_{i=1}^n X_i^2 = 0 \\ \sum_{i=1}^n Y_i - nb_0 - b_1 \sum_{i=1}^n X_i = 0 \end{cases} \quad (3.5)$$

经过整理得到

$$\begin{cases} \hat{\beta}_1 \equiv b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \hat{\beta}_0 \equiv b_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{cases} \quad (3.6)$$

以上便是系数估计的全部步骤，其中式子3.3、3.4和3.6是考试中的给分点。另外，注意到 $\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ 通过将 $\hat{\beta}_1$ 和 $\hat{\beta}_0$ 回代3.3得到正规方程：

$$\begin{cases} \sum_{i=1}^n X_i \hat{u}_i = 0 \\ \sum_{i=1}^n \hat{u}_i = 0 \end{cases} \quad (3.7)$$

这可以推出来以下结论：

1. $S_{X_i \hat{u}_i} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(\hat{u}_i - \bar{\hat{u}}_i) = 0$, 即解释变量和残差的样本协方差为0
2. $\bar{\hat{u}}_i = 0$, 但要注意的是, 这不意味着 $\bar{u}_i = 0$
3. Y_i 的样本观测值均值等于预测值均值, 且拟合直线必过 (\bar{X}, \bar{Y}) 。因为 $\bar{Y} = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i + \hat{u}_i) = \bar{\hat{Y}} = \frac{1}{n} \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$

3.3 拟合效果评价

拟合效果评价有三个指标：回归标准误差、均方根误差、 R^2 。

其中，回归标准误差和均方根误差需要掌握他们表达式的区别。回归标准误差(Standard Error of Regression, SER)是随机误差项 u_i 的标准差 $\sigma_{u_i}^2$ 的估计量，均方根误差(Root of Mean Squared Errors, RMSE)是残差平方和的均值的平方根，即

$$\begin{cases} SER = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (\hat{u}_i - \bar{\hat{u}}_i)^2} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2} \\ RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2} \end{cases} \quad (3.8)$$

其中，SER的n-2为经过调整的自由度，即n减去3.3中等式数量2。RMSE不需要经过自由度调整，因为他取的是残差平方和的均值。

R^2 最重要的是它的定义。首先我们定义以下平方和：

- 被解释平方和(Explained Sum of Squares, ESS): $ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$

- 总平方和(Total Sum of Squares, TSS): $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$

- 残差平方和(Sum of Squared Residual, SSR): $TSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$

以上平方和，注意不要写成SST、SSE、RSS等形式，别的教材可能确实会这么写，但我们还是按照我们教材的符号规定。

另外我们注意到， $TSS=ESS+SSR$ （留作习题）。我们定义

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \quad (3.9)$$

R^2 为此可以被解释成是被解释差异部分占总体差异部分的比例。 R^2 越大，说明回归拟合效果越好。

R^2 需要知道以下性质：

1. $0 \leq R^2 \leq 1$ ，这是因为ESS、TSS、SSR均大于0。另外，当 $R^2 = 0$ 时，我们认为 X_i 完全无法解释 Y_i ，当 $R^2 = 1$ 时，我们认为 X_i 完全能够解释 Y_i 的变动
2. 在 X_i 非常数的情况下，始终有 $R^2 = 0 \Leftrightarrow \beta_1 = 0$
3. $R^2 = [\text{corr}(Y_i, \hat{Y}_i)]^2$ （留作习题）。从这个角度也能看出， R^2 越接近于1， Y_i, \hat{Y}_i 相关性越强，回归效果越好。该式在多元条件下也成立。

另外，需要注意的是，计量更重要的是他对因果效应的计量准不准， R^2 由于代表的仅仅是他的预测效果，所以在计量意义上它无关紧要，为此所谓“ R^2 代表了计量模型的好坏”这句话不对。只是在后续的模型比较中，如果两个模型在计量效果上都类似，那么我们可能更倾向于采取 R^2 高，即回归效果更好的模型。

**拟合效果评价并不是计量care的内容，大多数时候你看到的计量论文中 R^2 都比较低，一般都不会到0.6以上。但是，对于后续时间序列，或是机器学习等课程而言，他们并不太关心模型估计系数是不是真的和总体回归模型的那个系数接近，而更关心的是我能不能拿这个模型去做预测。在这种情况下，RMSE和 R^2 都极其重要。

3.4 计量效果评价

3.4.1 OLS假设

一个模型计量的好不好，重要的是数据是否能够很好的符合模型假设。为此我们首先了解一元线性回归模型的三条最小二乘假设（Least Squares Assumptions, LSA）。

1. LSA1: $E(u_i|X_i) = 0$ ，即给定 X_i 的情况下， u_i 的条件均值为0。
2. LSA2: (X_i, Y_i) i.i.d，即 (X_i, Y_i) 独立同分布
3. LSA3: $E(X_i^4) < \infty, E(Y_i^4) < \infty$ ，即 X_i, Y_i 具有有限四阶矩（或者有限峰度），也可以说数据没有较大异常值

LSA1 这条假设是OLS中最重要的数据关系假设，对随机误差项的条件分布在一阶矩上进行限制。他和LSA2共同确保了 $\hat{\beta}_1$ 的无偏性，以及单独的LSA1能确定一致性。我们证明如下：

练习 3.1 在最小二乘框架下，我们有 $E(\hat{\beta}_1) = \beta_1$

证明 由 $Y_i = \beta_0 + \beta_1 X_i + u_i$ 知， $\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u}$ 两式相减有：

$$Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + (u_i - \bar{u}) \quad (3.10)$$

由3.6知,

$$\begin{aligned}
 \hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
 &= \frac{\sum_{i=1}^n (X_i - \bar{X})[\beta_1(X_i - \bar{X}) + (u_i - \bar{u})]}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
 &= \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^n (X_i - \bar{X})^2}
 \end{aligned} \tag{3.11}$$

注意到 $\sum_{i=1}^n (X_i - \bar{X})\bar{u} = \bar{u}(\sum_{i=1}^n X_i - n\bar{X}) = 0$, 有

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \tag{3.12}$$

这时候没法再化简了, 直接取期望, 并利用条件均值迭代公式:

$$\begin{aligned}
 E(\hat{\beta}_1) &= \beta_1 + E\left(\frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \\
 &= \beta_1 + E\left[E\left(\frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})^2} | X_1, X_2, \dots, X_n\right)\right] \\
 &= \beta_1 + E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})E(u_i | X_1, X_2, \dots, X_n)}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]
 \end{aligned} \tag{3.13}$$

利用LSA2, $(X_i, u_i) i.i.d$ (后面会对此进行说明), 则有: $E(u_i | X_1, X_2, \dots, X_n) = E(u_i | X_i) = 0$

故以上式最终化简为

$$E(\hat{\beta}_1) = \beta_1 + E[0] = \beta_1 \tag{3.14}$$

上述证明过程中,

- 注意3.10, 这里的 \bar{u} 是随机误差项的均值, 而不是残差项的均值。
- 在写条件均值迭代的时候, 一定要将 X_1, X_2, \dots, X_n 都囊括进去。因为你给了 X_i 是没办法 $E[(X_1 - \bar{X})u_1 | X_i]$ 的。当然, 还有一致性。

练习 3.2 $\hat{\beta}_1$ 具有一致性。

证明

$$\begin{aligned}
 \hat{\beta}_1 &= \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \\
 &= \frac{\hat{Cov}(X_i, Y_i)}{\hat{Var}(X_i)}
 \end{aligned} \tag{3.15}$$

这意味着 $\hat{\beta}_1 \xrightarrow{P} \frac{Cov(X_i, Y_i)}{Var(X_i)}$, 此时

$$\begin{aligned}\hat{\beta}_1 &= \frac{Cov(X_i, \beta_0 + \beta_1 X_i + u_i)}{Var(X_i)} \\ &= \beta_1 + \frac{Cov(X_i, u_i)}{Var(X_i)} \\ &= \beta_1\end{aligned}\quad (3.16)$$

上述证明中, 利用了这条假设的推论: $Cov(X_i, u_i) = corr(X_i, u_i) = 0$, 即 X_i, u_i 必须完全不相关。这个推论意味着 X_i 和除了 X_i 外, 其他能够影响 Y_i 的因素 u_i 必须互不影响 (但是 $corr(X_i, u_i) = 0$ 不能推出 $E(u_i|X_i) = 0$, 比如 $X_i \sim N(0, 1)$ 但是 $u_i = X_i^2$)。另外, 若 $corr(X_i, u_i) \neq 0$, 则由逆否命题知 $E(u_i|X_i) \neq 0$, 即违背了 LSA1。

LSA2 这条假设主要是规定了你数据的抽样方式, 必须是一个简单随机抽样, 样本之间不能有相关性。这个主要是在截面数据中比较容易满足 (当然并不是说截面数据一定 i.i.d., 比如地区之间如果有相关性就得用下空间计量经济学), 但在包含时间维度的数据中——时间序列数据和面板数据中很难满足。比如股价数据从统计上基本满足随机游走过程 $\ln Y_t = \ln Y_{t-1} + \epsilon_t$, ϵ_t 为收益率序列, 你可以看成是单纯的独立同分布的序列。这时候 Y_t 和前一日数据 Y_{t-1} 具有极强的相关性, 你不能拿这种数据不能简简单单用 OLS 做。另外, 由于总体回归方程 $Y_i = \beta_0 + \beta_1 X_i + u_i$ 中, β_0 和 β_1 是常数, 确定了 X_i, Y_i, u_i 中的两个就可以推出另外一个, 所以该假设也可以写成是: $(X_i, u_i) \text{i.i.d.}$

注 ***在金融时间序列当中, 如果 X_i, Y_i 自相关性很强, 那么由于某些时候数据会出现趋同的特征, 这时候的回归一般没什么意义, 只是建立起了虚假的回归关系 (如果可以协整另说)。如图 3.1 所示, 拿两个随机游走序列进行回归, R^2 和 t 统计量并不是很稳定, 有时候可能 R^2 会高达 0.95, t 值高达 60 以上。但是这时候用于检验是否会出现伪回归的 DW 统计量非常低, 基本不超过 0.6, 说明残差序列也存在序列自相关。这时候我们称这种回归“伪回归”。

LSA3 这条假设是技术性假设, 主要是使得 $\hat{\beta}_1$ 方差有限, 并且满足正态分布。而当违反该假设即数据中出现少量较大异常值时, 如图 3.2, 回归出来的斜率往往变化较大。考虑到我们是为了计量因果效应, 一般在回归前都会先去除异常值。另外, 同样因为总体回归方程中的 X_i, Y_i, u_i 有“知二推三”, 所以该假设也可以写成是: $E(X_i^4) < \infty, E(u_i^4) < \infty$ 。

***当你发现你的数据有较多异常值, 但你又不想删去他们时, 可以选择取对数形式降低较大极端值影响, 或者利用中位数回归。中位数回归是为了最小化残差的绝对值和, 即 $\min_{b_1, b_0} |Y_i - b_0 - b_1 X_i|$, 他可以比较好的容忍异常值, 但由于绝对值的存在不能求导并不如 OLS 一样好求解。所以一般而言, 如果能用 OLS 还是用 OLS 回归。

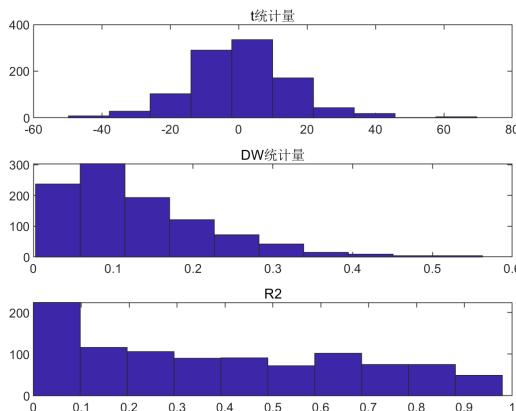


图 3.1: 伪回归

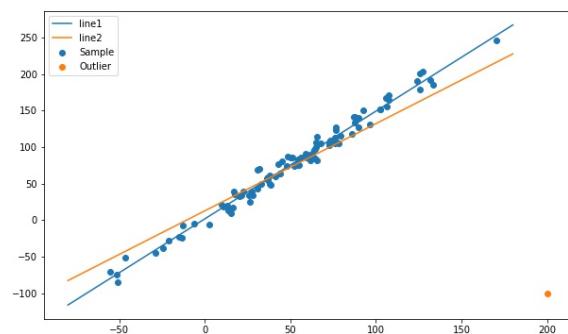


图 3.2: 异常值对回归结果的影响

我们将上述内容整理成以下表格。

	LSA1	LSA2	LSA3
内容	$E(u_i X_i) = 0$	$(X_i, Y_i) \text{i.i.d}$	$E(X_i^4) < \infty, E(Y_i^4) < \infty$
等价形式	/	$(X_i, u_i) \text{i.i.d}$	$E(X_i^4) < \infty, E(u_i^4) < \infty$
推论	$\text{corr}(X_i, u_i) = 0$	/	/
作用	确保 $\hat{\beta}_1$ 的无偏性和一致性	确定样本抽样方式	保证异常值不会干扰到回归结果
违反后果	$\hat{\beta}_1$ 不再具有无偏性	回归可能不具有意义，即“伪回归”	估计出来的参数会有很大变化

3.4.2 抽样分布

了解了一元线性回归的三个假设，现在我们来看看 $\hat{\beta}_1$ 的抽样分布。 $\hat{\beta}_0$ 的抽样分布不要求掌握，因为我们更关心的是量化因果效应而非计算回归线的高度。

我们现在已知 3.6，那么，和证明无偏性类似，有：

$$\begin{aligned}
\hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \beta_1 + \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})u_i}{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})u_i}{S_X^2} \left(\frac{n}{n-1} \right)
\end{aligned} \tag{3.17}$$

在大样本下，我们知道

$$\bar{X} \xrightarrow{P} \mu_X, S_X^2 \xrightarrow{P} \sigma_X^2 \tag{3.18}$$

所以有

$$\hat{\beta}_1 \xrightarrow{P} \beta_1 + \frac{1}{n} \sum_{i=1}^n \frac{(X_i - \mu_X)u_i}{\sigma_X^2} \tag{3.19}$$

令 $v_i = (X_i - \mu_X)u_i$ ，由 LSA2， $\{v_i\}_{i=1}^n$ 独立同分布。

由 LSA1: $E(v_i) = E[(X_i - \mu_X)u_i] = E\{E[(X_i - \mu_X)u_i|X_i]\} = E\{(X_i - \mu_X)E[u_i|X_i]\} = 0$.

由 LSA3, $Var(v_i) \leq E[((X_i - \mu_X)u_i)^2] \leq \sqrt{E[(X_i - \mu_X)^4]E(u_i^4)} < \infty$

由于 v_i 具有期望和有限方差，且独立同分布，故由中心极限定理有：

$$\bar{v} \xrightarrow{d} N(0, \frac{Var(v_i)}{n}) \tag{3.20}$$

所以

$$\hat{\beta}_1 \xrightarrow{d} N(\beta_1, \frac{\sigma_v^2}{n\sigma_X^4}) \tag{3.21}$$

或者写成

$$\hat{\beta}_1 \xrightarrow{d} N(\beta_1, \frac{Var[(X_i - \mu_X)u_i]}{n\sigma_X^4}) \tag{3.22}$$

根据上述式子，我们有以下结论：当其他条件不变的前提下

- X_i 的方差增大， $Var(\hat{\beta}_1)$ 越小（该条在后面的同方差看的更为明显）
- u_i 的方差越大， $Var(\hat{\beta}_1)$ 越大
- 样本量 n 越大， $Var(\hat{\beta}_1)$ 越小

可以看到在抽样分布的证明过程当中，LSA1-3 都有用到，其中，LSA3 只是一条保证总体方差有限的技术性假设。对于方差项，我们后面加上同方差假设后还能进行进一步的化简。

3.4.3 置信区间和假设检验

根据以上对抽样分布的分析，我们得出了 $\hat{\beta}_1$ 的渐进分布，现在，我们可以利用这个分布进行置信区间的构造和假设检验。

对于置信区间，前面我们提过，置信区间是在给定置信度 $1-\alpha$ 的条件下，真值 β_1 落入该区间的概率。那么，依照枢轴量法和大样本下 $\hat{\beta}_1$ 的抽样分布，我们有

$$P\left(\left|\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}\right| > t_{\alpha/2}\right) = 1 - \alpha \quad (3.23)$$

其中 $SE(\hat{\beta}_1)$ 是 $\hat{\beta}_1$ 的标准误，即 $\hat{\beta}_1$ 标准差的估计量。根据大样本下 $\hat{\beta}_1$ 的分布的方差

$$Var(\hat{\beta}_1) = \frac{Var[(X_i - \mu_X)u_i]}{n(\sigma_X^2)^2} \quad (3.24)$$

我们估计

$$SE(\hat{\beta}_1) = \sqrt{\frac{\frac{1}{n-2} \sum_{i=1}^n ((X_i - \bar{X})u_i)^2}{\frac{n}{n} \left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right)^2}} \quad (3.25)$$

式中， $\frac{1}{n-2} \sum_{i=1}^n ((X_i - \bar{X})u_i)^2$ 是 $Var[(X_i - \mu_X)u_i]$ 的估计量， $n-2$ 是 v_i 的自由度，这里因为OLS估计中两个偏导等于0的等式，自由度要减去2。 $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 是对 σ_X^2 的估计。

当然，由于标准误的具体计算比较繁琐，我们一般都会借助计算机软件求出。

式 3.25 等价于

$$P(\hat{\beta}_1 - t_{\alpha/2} * SE(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + t_{\alpha/2} * SE(\hat{\beta}_1)) = 1 - \alpha \quad (3.26)$$

取 $\alpha = 0.05$, $t_{\alpha/2} = 1.96$, β_1 在95%置信水平下的置信区间为

$$[\hat{\beta}_1 \pm 1.96 * SE(\hat{\beta}_1)] \quad (3.27)$$

现在来看假设检验。步骤如下：

1. 一般而言，我们比较关心的是是否有因果效应，那么便可确立 $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$
2. 由于 $\hat{\beta}_1 \stackrel{d}{\sim} N(\beta_1, \frac{Var[(X_i - \mu_X)u_i]}{n\sigma_X^4})$ ，此时 β_1 未知，方差未知，那么可以确定统计量为t统计量： $t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$ ，拒绝域形式为 $P(|t| > t_{\alpha/2}) = 1 - \alpha$
3. 计算出来t统计量 t_{actual} 后，有两种方式：

临界值法：在显著性水平 $\alpha = 0.05$ 下，临界值为 $t_0 = 1.96$ 。故只要让 $|t| > t_0$ 即可拒绝 H_0

p值法：p值是 H_0 为真的情况下比样本观察结果更为极端的概率。所以 $p = P(|t| > |t_{actual}|)$ ，若 $p < 0.05$ ，则说明拒绝原假设的概率仍然在容忍范围内，那么我们可以放心拒绝原假设。

例题 3.1 已知一元线性回归方程 $price_i = \beta_0 + \beta_1 mpg_i + u_i$, stata回归结果如图 3.3, 试写出 β_1 的置信区间，并检验 $H_0 : \beta_1 = 0$

解 置信度为95%的置信区间为 $[-238.8943 - 1.96 * 57.47791, -238.8943 + 1.96 * 57.47791] = [-351.5510, -126.2376]$ （之所以和表格中的不太一样，是因为 $t_{\alpha/2}$ 精度问题，考试时会有精度要求）

相应的可以计算t统计量， $t = -238.8943 / 57.47791 = -4.1563$ ，显然有 $|t| > 1.96$ ，在5%的显著性水平下拒绝 H_0 。

3.4.4 LSA1假设增强

对于LSA1，我们规定了他的条件分布一阶矩 $E(u_i | X_i) = 0$ ，和LSA2推出了 $\hat{\beta}_1$ 是一致无偏估计量，并加上LSA3推出了 $\hat{\beta}_1$ 的渐进分布。但我们注意到这个渐进分布的方差有点儿繁琐，为此我们思考能否通过加强条件分布的要求（即LSA1），进而去简化这个方差，并看看增强后的 $\hat{\beta}_1$ 是否还有更好的性质。

Linear regression	Number of obs	=	74
	F(1, 72)	=	17.28
	Prob > F	=	0.0001
	R-squared	=	0.2196
	Root MSE	=	2623.7

price	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
mpg	-238.8943	57.47701	-4.16	0.000	-353.4727 -124.316
_cons	11253.06	1376.393	8.18	0.000	8509.272 13996.85

图 3.3: 例题 3.1 回归结果

第一步加强自然是加强条件分布的二阶矩。令条件分布的方差为常数（即同方差条件），即有

$$Var(u_i|X_i) = Constant \quad (3.28)$$

有了这个条件后，尝试去化简 $\hat{\beta}_1$ 的方差。由于

$$Var(\hat{\beta}_1) = \frac{Var((X_i - \mu_X)u_i)}{n\sigma_X^4} \quad (3.29)$$

分子是个无条件方差，包含了 u_i 项，我们的条件是给定了条件方差，为此不妨联想到条件方差迭代公式：

$$\begin{aligned} Var((X_i - \mu_X)u_i) &= Var(E((X_i - \mu_X)u_i|X_i)) + E(Var(X_i - \mu_X)u_i|X_i) \\ &= Var((X_i - \mu_X)E(u_i|X_i)) + E((X_i - \mu_X)^2 Var(u_i|X_i)) \\ &= 0 + E((X_i - \mu_X)^2) Var(u_i|X_i) \\ &= Var(X_i) Var(u_i|X_i) \end{aligned} \quad (3.30)$$

又由于

$$Var(u_i) = Var(E(u_i|X_i)) + E(Var(u_i|X_i)) = E(Var(u_i|X_i)) = Var(u_i|X_i) \quad (3.31)$$

所以有分子化简如下：

$$Var((X_i - \mu_X)u_i) = \sigma_X^2 \sigma_u^2 \quad (3.32)$$

为此有

$$Var(\hat{\beta}_1) = \frac{\sigma_u^2}{n\sigma_X^2} \quad (3.33)$$

那么，对 u_i 的条件二阶矩加以限制之后，是否还有更进一步的结论？这就是我们的Gauss-Markov定理。

定理 3.1 (Gauss-Markov定理)

在给定LSA1-3的条件下，若有同方差条件 $Var(u_i|X_i) = Constant$ ，则在给定 X_1, X_2, \dots, X_n 下， $\hat{\beta}_1$ 为最佳线性无偏估计量 (Best Linear Unbiased Estimator, BLUE)，即 $\hat{\beta}_1$ 是所有线性无偏估计量中方差最小的那一个。

注 该定理中，所谓线性估计量是指表示为 Y_1, Y_2, \dots, Y_n 的线性函数，而我们的OLS估计量恰好就能表示成 $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum_{i=1}^n \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} Y_i = \sum_{i=1}^n a_i Y_i$ (其中 $a_i = \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}$)。

证明 首先该估计量是线性无偏的，且有

$$\hat{\beta}_1 = \sum_{i=1}^n a_i Y_i \quad (3.34)$$

式中, $a_i = \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}$ 现在我们需要证明其“最佳”, 即在所有线性无偏估计量中方差最小。设线性无偏估计量 $\tilde{\beta}_1 = \sum_{i=1}^n \lambda_i Y_i$, 则有

$$\begin{aligned} E(\tilde{\beta}_1 | X_1, X_2, \dots, X_n) &= \sum_{i=1}^n \lambda_i (\beta_0 + \beta_1 X_i + E(u_i | X_i)) \\ &= (\sum_{i=1}^n \lambda_i) \beta_0 + (\sum_{i=1}^n \lambda_i X_i) \beta_1 \\ &= \beta_1 \end{aligned} \quad (3.35)$$

则有

$$\begin{cases} \sum_{i=1}^n \lambda_i = 0 \\ \sum_{i=1}^n \lambda_i X_i = 1 \end{cases} \quad (3.36)$$

设 $\lambda_i = a_i + d_i$, 由于 $\hat{\beta}_1$ 是 $\tilde{\beta}_1$ 取值的一部分, 那么有

$$\begin{cases} \sum_{i=1}^n a_i = 0 \\ \sum_{i=1}^n a_i X_i = 1 \end{cases} \quad (3.37)$$

即

$$\begin{cases} \sum_{i=1}^n d_i = 0 \\ \sum_{i=1}^n X_i d_i = 1 \end{cases} \quad (3.38)$$

由此,

$$\sum_{i=1}^n (X_i - \bar{X}) d_i = 0 \quad (3.39)$$

我们可以得出

$$\begin{aligned} Var(\tilde{\beta}_1 | X_1, X_2, \dots, X_n) &= \sum_{i=1}^n \lambda_i^2 Var(\beta_0 + \beta_1 X_i + u_i | X_i) \\ &= \sum_{i=1}^n \lambda_i^2 \sigma_u^2 \\ &= \sum_{i=1}^n (a_i + d_i)^2 \sigma_u^2 \\ &= Var(\hat{\beta}_1 | X_1, X_2, \dots, X_n) + \sum_{i=1}^n (2a_i d_i + d_i^2) \sigma_u^2 \\ &= Var(\hat{\beta}_1 | X_1, X_2, \dots, X_n) + \sigma_u^2 (2 \sum_{i=1}^n a_i d_i + \sum_{i=1}^n d_i^2) \end{aligned} \quad (3.40)$$

而

$$\begin{aligned} \sum_{i=1}^n a_i d_i &= \frac{\sum_{i=1}^n (X_i - \bar{X}) d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= 0 \end{aligned} \quad (3.41)$$

所以有

$$Var(\tilde{\beta}_1|X_1, X_2, \dots, X_n) = Var(\hat{\beta}_1|X_1, X_2, \dots, X_n) + \sigma_u^2 \left(\sum_{i=1}^n d_i^2 \right) \geq Var(\hat{\beta}_1|X_1, X_2, \dots, X_n) \quad (3.42)$$

当 $d_i = 0$ 即 $\tilde{\beta}_1 = \hat{\beta}_1$ 时，等号成立。所以 $\hat{\beta}_1$ 是最佳线性无偏估计量。

同方差的条件的加强使得标准误形式变的非常简洁。但是同方差下的标准误公式只能在同方差情况下使用，不能在异方差情况下使用。相反，我们所谓异方差下的标准误其实是仅在LSA1-3下得出的标准误，他没有同方差这个条件的加强，是个稳健的标准误（为此在stata中异方差下稳健标准误的回归命令为reg Y X, robust），无论是在同方差形式还是异方差情况下都能使用。

但如果对 u_i 的条件分布进一步加强，我们还能得到以下结论。

定理 3.2

在给定LSA1-3和同方差条件 $Var(u_i|X_i) = Constant$ ，且在给定 X_1, X_2, \dots, X_n 下， u_i 服从条件正态分布，那么 $\hat{\beta}_1$ 为最佳一致估计量，即 $\hat{\beta}_1$ 是所有一致估计量中方差最小的那个。



这个证明比较复杂，我们在此忽略。

同样的，我们还能得到， $\hat{\beta}_1$ 将服从精确的正态分布。这是因为

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) u_i \quad (3.43)$$

在已知 X_1, X_2, \dots, X_n 的情况下，若 u_i 服从正态分布，那么整体的 $\hat{\beta}_1$ 也服从精确正态分布。为此 $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = (\beta_0 + \beta_1 X_i + u_i) - \hat{\beta}_1 \bar{X}$ 服从精确正态分布。相应的两个对应的t统计量也服从精确的自由度为 $n-2$ 的t分布。

第3章 练习

1. 尝试写出无截距项的一元线性模型的系数估计过程。
2. ESS、TSS和SSR的关系：请证明在一元线性回归模型 $Y_i = \beta_0 + \beta_1$ 的OLS估计下，有 $TSS = ESS + SSR$ ，但是去掉截距项 β_0 的OLS估计下，该式子不一定成立。
3. 证明：一元线性回归模型中， $R^2 = [corr(Y_i, \hat{Y}_i)]^2$
4. 证明： $E(\hat{\beta}_0) = \beta_0$
5. 判断下面命题的正确性：
 - A. 同方差下的标准误在异方差下可以使用。
 - B. 异方差下的标准误在异方差下可以使用。
 - C. 同方差下的标准误在同方差下可以使用。
 - D. 异方差下的标准误在同方差下可以使用。
6. 对于 $Y_i = \beta_0 + \beta_1 X_i + u_i$ ，若 β_1 变成原来的两倍，则 $SE(\hat{\beta}_1)$ 变化是原来的多少？
7. (二值变量下的一元线性回归) 考虑

$$Y_i = \beta_0 + \beta_1 X_i + u_i, X_i \in \{0, 1\}$$

那么显然有

$$E(Y_i|X_i = 1) = \beta_0 + \beta_1$$

$$E(Y_i|X_i = 0) = \beta_0$$

现在我们要估计 β_1 ，用另一种角度去思考：样本均值是对总体均值的无偏一致估计，那么我们尝试用在不同 X_i 下 Y_i 取值的均值来作为条件均值的估计——即 $X_i = 0$ 或者 1 时，样本 Y_i 的均值分别记为 \bar{Y}_1 和 \bar{Y}_0 ，则有 $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$

(1) 请利用式子 3.6 说明两个 $\hat{\beta}_1$ 的估计结果是完全一样的。

(2) 分别在同方差和异方差条件下 $\hat{\beta}_1$ 的标准误。并写出检验 $\beta_1 = 0$ 的步骤。

8. 某人做好一个一元线性回归模型 $Y_i = \beta_0 + \beta_1 X_i + u_i$ 后，将残差计算了出来，并求得了残差和 X_i 的样本相关系数，结果等于0，由此得出：模型满足LSA1，试问这过程中有什么问题？

第4章 多元线性回归模型

前面我们讲完了最为基础的一元线性回归模型，了解了计量的基本建模流程。但毕竟是最为基础的回归模型，有的条件，比如LSA1： $E(u_i|X_i) = 0$ 要求影响 Y_i 的其他因素 u_i 和 X_i 一定要不相关。而这个条件几乎在所有研究背景都很难满足，不满足的情况之一就是遗漏变量偏差。

4.1 遗漏变量偏差

首先我们来看看，何为遗漏变量

定义 4.1 (遗漏变量)

对于回归模型 $Y_i = f(X_i) + u_i$ ，若变量 Z_i 能影响 Y_i 但没有包含在被解释变量 X_i 中（即包含在随机误差项 u_i 中），我们称 Z_i 叫做遗漏变量。



从这可以看出，不是所有遗漏变量都会影响我们对因果效应的计量。一般来说，只要你没有动到LSA1： $E(u_i|X_i) = 0$ ， u_i 虽然是 X_i 的函数但是概率意义上平均为0，那么就没有任何关系。比如 Y_i 的统计误差是完全随机的，和任何变量没有关系，那么这个因素便不会影响LSA1。但是，如果说你的遗漏变量和 X_i 相关，那么便会影响到 β_1 的估计，这就是遗漏变量偏差。

定义 4.2 (遗漏变量偏差)

对于回归模型 $Y_i = f(X_i) + u_i$ ，若变量 Z_i 满足以下条件：

- Z_i 是个遗漏变量，即能影响 Y_i 但没有包含在被解释变量 X_i 中
- Z_i 与 X_i 相关，即 $corr(Z_i, X_i) \neq 0$

则我们称 Z_i 导致的偏差是遗漏变量偏差。



那么，我们不妨来看看遗漏变量偏差对 $\hat{\beta}_1$ 的影响。我们考虑两个方程：

现实中真实总体的回归方程，不含有任何的模型偏误(长回归方程)

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i \quad (4.1)$$

你提出的回归方程(短回归方程)

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (4.2)$$

由于我们估计的

$$\hat{\beta}_1 = \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \xrightarrow{P} \frac{Cov(X_i, Y_i)}{Var(X_i)} \quad (4.3)$$

而

$$\begin{aligned} Cov(X_i, Y_i) &= Cov(X_i, \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i) \\ &= \beta_1 Var(X_i) + \beta_2 Cov(X_i, Z_i) \end{aligned} \quad (4.4)$$

式中， $Cov(X_i, e_i) = 0$ 是因为长回归方程代表的是真实模型，不含有任何模型偏误，即LSA1成立。所以有

$$\hat{\beta}_1 \xrightarrow{P} \beta_1 + \beta_2 \rho_{X,Z} \frac{\sigma_Z}{\sigma_X} \quad (4.5)$$

这里我们也可以看到遗漏变量偏差的两个条件。一、他必须是个遗漏变量，是 Y_i 的决定因素，但是没有包含在回归模型当中，否则他就是个无关变量，即回归系数 $\beta_2 = 0$ 。二、他必须和变量 X_i 有相关性，也就是 $\rho_{X,Z}$ 不为0。

式 4.5 告诉我们一件很严重的事情：如果有遗漏变量偏差， $\hat{\beta}_1$ 和 β_1 之间不是乘上某个数的关系，而是加减

的关系。这就意味着你估计的 $\hat{\beta}_1$ 方向和 β_1 的方向可能会截然不同，比如你认为 β_1 应该是正的，但是 $\hat{\beta}_1$ 不显著，甚至是显著为负的，即无论你怎么调你的抽样数量，你的估计有可能偏大也可能偏小。

另外，式 4.5 也告诉我们， $\hat{\beta}_1$ 估计是偏大还是偏小，取决于 β_1 和 $\rho_{X,Z}$ 是否同号。如果是同号，那么你的估计将会偏大，如果是异号，那么你的估计会偏小。

为此，如果说你知道这个遗漏变量是啥且你能搞到这个变量的数据，那么修正这个模型的最好办法就是直接把这个变量加进去，即用多元线性回归。

4.2 多元线性回归模型表达式

一般来说，多元线性回归模型一般形式如下：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} + u_i \quad (4.6)$$

如果写成矩阵形式，令 $Y = [Y_1, Y_2, \dots, Y_n]^T$, $\beta = [\beta_0, \beta_1, \beta_2, \dots, \beta_n]^T$, $X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{21} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{kn} \end{bmatrix}$

$$u = (u_1, u_2, \dots, u_n),$$

那么模型表达式如下：

$$Y = X\beta + u \quad (4.7)$$

对于式 4.6 有：

$$E(Y_i | X_{1i}, X_{2i}, \dots, X_{ni}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} \quad (4.8)$$

那么，当控制住 X_{2i}, \dots, X_{ni} 不变时，有

$$E(Y_i | X_{1i} = x + 1, X_{2i}, \dots, X_{ni}) - E(Y_i | X_{1i} = x, X_{2i}, \dots, X_{ni}) = \beta_1 \quad (4.9)$$

所以 β_1 的系数解释为：控制住 X_{2i}, \dots, X_{ni} 不变时， X_{1i} 变化一单位时平均对 Y_i 的影响。

4.3 模型估计

模型估计与一元线性回归完全相同，还是最小化残差平方和，即有

$$\min_{b_0, b_1, b_2, \dots, b_n} \sum_{i=1}^n (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i} - \cdots - b_n X_{ni})^2 \quad (4.10)$$

或者写成矩阵形式更方便求解：

$$\min_b (Y - Xb)^T (Y - Xb) \quad (4.11)$$

***如果写成矩阵形式，若你了解矩阵代数过程是非常简单的。采取分母布局直接对向量 b 求导有

$$-2 * (Y - Xb)^T * X = 0 \quad (4.12)$$

解得

$$\beta \equiv b = (X^T X)^{-1} X^T Y \quad (4.13)$$

4.4 拟合效果评价

一元线性回归类似，还是SER、RMSE、 R^2 ，只是SER进行了自由度调整不同：

$$\begin{cases} SER = \sqrt{\frac{1}{n-k-1} \sum_{i=1}^n (\hat{u}_i - \bar{\hat{u}})^2} = \sqrt{\frac{1}{n-k-1} \sum_{i=1}^n \hat{u}_i^2} \\ RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2} \end{cases} \quad (4.14)$$

R^2 和原来完全相同，定义同样为 $R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$ ，也有 $R^2 = [corr(Y_i, \hat{Y}_i)]^2$

但这样会产生一个问题。由于多元线性回归当中，变量数并没有收到限制，这就会导致一个问题：为了追求更高的 R^2 ，我们往往会加入越多的变量。为了理解这点，考虑

$$Q(b_0, b_1, b_2, \dots, b_n, b_{n+1}) = \sum_{i=1}^n (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i} - \dots - b_n X_{ni} - b_{n+1} X_{(n+1)i})^2 \quad (4.15)$$

在不加入这个变量进行回归的前提下，我们其实是在干以下最优化问题：

$$\min_{b_0, b_1, b_2, \dots, b_n} Q(b_0, b_1, b_2, \dots, b_n, 0) \quad (4.16)$$

而我们有 $\min_{b_0, b_1, b_2, \dots, b_n, b_{n+1}} Q(b_0, b_1, b_2, \dots, b_n, b_{n+1}) \leq Q(b_0, b_1, b_2, \dots, b_n, 0)$ 恒成立，对不等号右边取最小值，有

$$\min_{b_0, b_1, b_2, \dots, b_n, b_{n+1}} Q(b_0, b_1, b_2, \dots, b_n, b_{n+1}) \leq \min_{b_0, b_1, b_2, \dots, b_n} Q(b_0, b_1, b_2, \dots, b_n, 0) \quad (4.17)$$

这也就表明了，增加了变量个数，本质上是减少了最优化问题中的约束条件，回归后的残差平方和SSR会进一步的减小，即增大 R^2 。为此我们引入调整 R^2 ，避免因为变量的个数增大而导致评价指标减小：

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{ESS/(n-1)} = 1 - \frac{s_{\hat{u}}}{s_Y} \quad (4.18)$$

可以看得出来， \bar{R}^2 大小也能评价拟合效果如何。由于ESS和SSR都大于0， \bar{R}^2 一定小于1；但是 \bar{R}^2 不一定大于0，原因是因为当回归效果比较不好即SSR非常大时，由于 $\frac{n-1}{n-k-1} > 1$ ， $\frac{SSR}{ESS} * \frac{n-1}{n-k-1}$ 可能会大于1。

4.5 计量效果评价

4.5.1 模型假设

多元线性回归模型假设和一元类似，但由于我们在矩阵求解过程中，引入了 $(X^T X)^{-1}$ ，这也表明我们要求矩阵X一定是个列满秩矩阵，但同时为了保证求解的精度，他也不能是病态矩阵(X接近列不满秩的状态)。即：多元线性回归模型中，一个变量不能是且不能接近是一个变量的线性组合。

故假设如下：

1. LSA1: $E(u_i | X_1, X_2, \dots, X_k) = 0$ ，即给定 X_1, X_2, \dots, X_k 的情况下， u_i 的条件均值为0。
2. LSA2: $(X_1, X_2, \dots, X_k, Y_i) i.i.d.$ ，即 (X_i, Y_i) 独立同分布
3. LSA3: $E(X_j^4) < \infty$ ($j = 1, 2, \dots, k$)， $E(Y_i^4) < \infty$ ，即样本中每个变量都具有有限四阶矩（或者有限峰度），也可以说数据没有较大异常值
4. 不能有较大的多重共线性

这里的多重共线性指代的便是上文说的“多元线性回归模型中，一个变量不能是且不能接近是一个变量的线性组合”。注意这里的变量也包含常数项1。

首先我们来说说完全多重共线性。完全多重共线性一般来说有两种方式，要么是你手贱多加了一个完全一样的变量，要么是“虚拟变量陷阱”。虚拟变量陷阱指的是：一个定性的变量有着m个结果，但你引入了m个虚拟变量并做了带截距项的回归，这时候回归系数是求不出来的。比如，性别一般来说有两种取值（排除异常情

况)：男性或者女性。那么假如说你引入

$$D_{1i} = \begin{cases} 1, & \text{性别为男} \\ 0, & \text{性别为女} \end{cases} \quad (4.19)$$

和

$$D_{2i} = \begin{cases} 1, & \text{性别为女} \\ 0, & \text{性别为男} \end{cases} \quad (4.20)$$

并做回归 $Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$, 那么由于 $1 = D_{1i} + D_{2i}$ 恒成立, 这时候矩阵X列不满秩, 无法求解系数, 即落入“虚拟变量陷阱”。

对于较大的多重共线性, 往往是由于变量之间具有高度相似性所导致的。比如, 考虑你贸食堂收入受那些因素影响, 你加入了学生生活费、食堂饭菜价格、员工月收入等等。但是, 就大部分人情况来看, 如果你点外卖的费用和次数都大致固定的话, 食堂饭菜价格大致和生活费是大致呈正比的(至少俺周围人都这样, 点外卖全看美团有没有优惠券), 那么这时候可能就会有较大共线性问题。此外, 有时候你加的变量太多, 变量与变量之间可能也确实会存在着你还不是很清楚的共线性关系(在机器学习里尤为常见, 解决方式通常用L1的LASSO回归或者L2的岭回归)。这时候我们检测共线性的方法是用方差膨胀因子(VIF, Variance Inflation Factor)。

$$VIF_i = \frac{1}{1 - R_i^2} \quad (4.21)$$

其中 R_i^2 是你拿一个变量 X_i 对其余变量做回归的 R^2 。当 $0 < VIF < 10$ 时(有的是取5), 说明不存在较大的多重共线性; 当 $10 \leq VIF < 100$ 时说明有较大多重共线性; 当 $VIF \geq 100$ 时有严重多重共线性。

4.5.2 抽样分布和置信区间

抽样分布在给定以上假设还是满足渐进正态的。这点我们不再提及。置信区间也是同样构造方式。

4.5.3 假设检验

多元线性回归的假设检验比较多样, 分为以下几种: 一、单个约束(一个等号)检验, 要么检验单个系数是否为0, 要么检验系数之间是否满足一定的关系。二、多个约束(多个等号)检验, 一般都是检验多个系数是否联合为0。

4.5.3.1 单个约束检验

单个约束的检验本质上还是构造t统计量——计算t统计量=系数/标准误, 比较他的绝对值是否大于临界值1.96, 或者对应的p值是否小于0.05。对于单个系数是否为0的和一元线性回归一毛一样, 具体的检验步骤我们不再赘述, 我们主要想说的是多个系数之间单个约束的检验, 考试有时会作为稍难的题进行考察(但不是每年都考)。

两个系数相等 这种是最简单的。考虑

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad (4.22)$$

现在考虑 $H_0 : \beta_1 = \beta_2$ 这个约束。如果是stata直接“test $X_1 = X_2$ ”就行, 但实际上这个假设等价于 $H_0 : \beta_1 - \beta_2 = 0$, 那么我们可以在回归方程中凑出 $(\beta_1 - \beta_2)$ 这个系数, 然后按上述逻辑去检验他是否为0。

我们变换如下: 令 $\alpha = \beta_1 - \beta_2$, $W_i = X_{1i} + X_{2i}$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \\ &= \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + \beta_3 X_{3i} + u_i \\ &= \beta_0 + \alpha X_{1i} + \beta_2 W_i + \beta_3 X_{3i} + u_i \end{aligned} \quad (4.23)$$

此时我们只需要检验系数 $\alpha = 0$ 即可。

系数线性组合为0 这种稍微复杂点，但也是类似的道理。同样还是凑。考虑 $H_0 : \beta_1 + 2\beta_2 + 3\beta_3 = 0$ ，令 $\alpha = \beta_1 + 2\beta_2 + 3\beta_3$, $W_{2i} = X_{2i} - 2X_{1i}$, $W_{3i} = X_{3i} - 3X_{1i}$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \\ &= \beta_0 + (\beta_1 + 2\beta_2 + 3\beta_3) X_{1i} + \beta_2 (X_{2i} - 2X_{1i}) + \beta_3 (X_{3i} - 3X_{1i}) + u_i \\ &= \beta_0 + \alpha X_{1i} + \beta_2 W_{2i} + \beta_3 W_{3i} + u_i \end{aligned} \quad (4.24)$$

系数线性组合为非零常数 这种是最复杂的。凑也是凑，但是我们要稍微转变下思路。设 $H_0 : \beta_1 + 2\beta_2 + 3\beta_3 = 4$ 如果只是变换等号右边，我们会发现

$$Y_i = \beta_0 + (\beta_1 + 2\beta_2 + 3\beta_3 - 4) X_{1i} + \beta_2 (X_{2i} - 2X_{1i}) + \beta_3 (X_{3i} - 3X_{1i}) + 4X_{1i} + u_i \quad (4.25)$$

可以看到，由于 $4X_{1i}$ 非常碍事，你变换了回归方程，但是没有完全变换，咋办？注意到 $4X_{1i}$ 在给定样本还没回归前就已经是个数而不是个需要通过模型整出来的东西，直接扔到等式左边，即

$$Y_i - 4X_{1i} = \beta_0 + (\beta_1 + 2\beta_2 + 3\beta_3 - 4) X_{1i} + \beta_2 (X_{2i} - 2X_{1i}) + \beta_3 (X_{3i} - 3X_{1i}) + u_i \quad (4.26)$$

令 $\alpha = \beta_1 + 2\beta_2 + 3\beta_3$, $Z_i = Y_i - 4X_{1i}$, $W_{2i} = X_{2i} - 2X_{1i}$, $W_{3i} = X_{3i} - 3X_{1i}$, 有

$$Z_i = \beta_0 + \alpha X_{1i} + \beta_2 W_{2i} + \beta_3 W_{3i} + u_i \quad (4.27)$$

同样做回归检验 $H_0 : \alpha = 0$ 就行。

4.5.3.2 多个约束检验

如果说单个约束检验使用的是t检验，那么多个约束的检验应该使用的是F检验而不是t检验。

首先，为什么对于多个约束的检验，不能将其拆分成一个个的约束进行检验？考虑一个比较简单的形式，对于二元变量回归（注意，二元是两个元素的回归，二值是一个变量但仅取0或1的回归）：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \quad (4.28)$$

原假设 $H_0 : \beta_1 = \beta_2 = 0$ 。现在考虑拆分成单个约束的检验，即将 H_0 拆分为 $\beta_1 = 0$ 和 $\beta_2 = 0$ ，并分别计算t统计量=回归系数/标准误，拿他的绝对值和1.96比看是否能拒绝。只要拒绝一个，那么我们就能拒绝 H_0 。

但这样做有个问题：我们的假设是 $H_0 : \beta_1 = \beta_2 = 0$ ，对应的一次一个检验是 $H_{01} : \beta_1 = 0$ 和 $H_{02} : \beta_2 = 0$ 。假设 t_1 和 t_2 相互独立，我们要求的显著性水平是5%，那么“一次一个”能够恰好拒绝原假设的概率为：

$$\begin{aligned} P(|t_1| > 1.96 \text{ 或 } |t_2| > 1.96 | H_0 \text{ 为真}) &= 1 - P(|t_1| \leq 1.96 \text{ 且 } |t_2| \leq 1.96 | H_0 \text{ 为真}) \\ &= 1 - P(|t_1| \leq 1.96 | H_0 \text{ 为真}) P(|t_2| \leq 1.96 | H_0 \text{ 为真}) \\ &= 1 - (1 - 0.05)^2 \\ &= 9.75\% > 5\% \end{aligned} \quad (4.29)$$

显然，如果“一次一个”在每一次采取的显著性水平都为5%时，很有可能会导致最终合起来的结果超过了显著性水平下所能容忍的犯错概率。为此一种做法是去缩减决策时每一步的显著性水平，但那样不现实，在约束个数增多时无论哪个都拒绝不了。更经常的我们会用F检验。在二元回归的情况下，F统计量

$$F = \frac{1}{2} \left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2} \right) \quad (4.30)$$

在大样本下， $F \sim F(q, \infty) = \chi^2(q)/q$, q 为约束个数，这里 $q = 2$ 。所以，当 $F > F_\alpha(2, \infty)$ ，在显著性水平 α 下拒绝原假设。

***这里，之所以是大于，是因为F统计量构造了平方项，比如检验 $\beta_1 = 0$ ，他会对相应的估计做一个平方，即看 $\hat{\beta}_1^2$ 是否足够远离0，至于什么叫做足够，那得具体看看F统计量如何构造。如果实在理解不了，你就粗略记住初级计量中所有涉及到F统计量 χ^2 统计量的检验都得让这些统计量大于某个临界值。

*基于上述F统计量，我们可以构建二元回归下两个系数的置信集。由于 $F = \frac{1}{2} \left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2} \right)$ ，在95%的置

信度下有 $F \leq F_{5\%}(2, \infty) = 3.00$, 即有

$$\frac{1}{2(1 - \hat{\rho}_{t_1, t_2}^2)} \left[\left(\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \right)^2 + \left(\frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \right)^2 - 2\hat{\rho}_{t_1, t_2}^2 \left(\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \right) \left(\frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \right) \right] \leq 3 \quad (4.31)$$

显然, 该置信集是一个以 $(\hat{\beta}_1, \hat{\beta}_2)$ 为中心的椭圆。

那么, 说完了为什么要用F检验, 我们下面来看看联合检验下的F统计量如何构造。

只有一个系数 对于单个系数的F统计量, 由于大样本下有 $t^2 = \chi^2(1) = F(1, \infty)$, 所以 $F = t^2$, 且 $F \sim F(1, \infty) = \chi^2(1)$

二元回归中两个系数的检验 和前文所叙相同, 大样本下 $F = \frac{1}{2} \left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1, t_2} t_1 t_2}{1 - \hat{\rho}_{t_1, t_2}^2} \right) \sim F(2, \infty)$

多个系数下, 异方差下的检验 ***如果是异方差的情况, 设约束为 $R\beta = r$, 因为 $\hat{\beta} = (X^T X)^{-1} X^T Y = \beta + (X^T X)^{-1} X^T U$, 在已知X的情况下, 由中心极限定理, $\hat{\beta} \sim N(\beta, \Sigma_{\hat{\beta}})$, 所以在原假设下有 $R\hat{\beta} - r \sim N(0, R\Sigma_{\hat{\beta}}R^T)$, 所以大样本下有: $F = (R\hat{\beta} - r)^T (R\Sigma_{\hat{\beta}}R^T)^{-1} (R\hat{\beta} - r) / q \sim \chi^2(q) / q = F(q, \infty)$ 。

不过, 由于异方差下F统计量构造式子比较复杂, 推导过程比较考察线性代数的功底, 初级计量不会涉及。我们在此不进行进一步的论述。真正使用的时候, 我们往往会借助stata的test命令进行计算。

多个系数下, 同方差下的检验 这个是初级计量的重中之重。推导是由上述异方差情况下的式子加上同方差化简而来。F统计量构造为

$$F = \frac{(SSR_r - SSR_u) / q}{SSR_u / (n - k_u - 1)} \quad (4.32)$$

符号说明如下:

- SSR_u 指的是不受 H_0 约束限制回归下的残差平方和
- SSR_r 指的是 H_0 成立下回归的残差平方和
- n 是样本量
- k_u 是不受 H_0 约束限制回归的变量个数
- q 是 H_0 的约束个数, 你可以简单看成是 H_0 中等号的个数。

如果你把 $R^2 = 1 - \frac{SSR}{TSS}$ 代入, 那么上述式子可以写成

$$F = \frac{(R_u^2 - R_r^2) / q}{(1 - R_u^2) / (n - k_u - 1)} \quad (4.33)$$

式中,

- R_u^2 指的是不受 H_0 约束限制回归下的 R^2
- R_r^2 指的是 H_0 成立下回归的 R^2
- n 是样本量
- k_u 是不受 H_0 约束限制回归的变量个数
- q 是 H_0 的约束个数, 你可以简单看成是 H_0 中等号的个数。

如果给定X下误差项U服从正态分布, 那么上述统计量精确服从 $F(q, n - k_u - 1)$ 的分布。但如果没给, 大样本下, 他服从 $F(q, \infty) = \chi^2(q) / q$ 分布。

综上不同情况的F统计量构建如下:

$$F = \begin{cases} t^2 & \text{单个系数等于0} \\ \frac{1}{2} \left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1, t_2} t_1 t_2}{1 - \hat{\rho}_{t_1, t_2}^2} \right) & \text{二元回归中两个系数联合为0} \\ \frac{(R_u^2 - R_r^2) / q}{(1 - R_u^2) / (n - k_u - 1)} & \text{同方差下多个系数} \\ (R\hat{\beta} - r)^T (R\Sigma_{\hat{\beta}}R^T)^{-1} (R\hat{\beta} - r) / q & \text{异方差下多个系数, 一般而言用stata中的test命令} \end{cases} \quad (4.34)$$

例题 4.1 利用stata中内置的数据集nlsw88, 分析妇女工资受那些因素所影响。方程为 $wage_i = \beta_0 + \beta_1 age_i + \beta_2 south_i + \beta_3 c_city_i + \beta_4 never_married_i + u_i$, 回归结果如图4.1, 请你在同方差条件下检验 $H_0 : \beta_2 = \beta_3 = \beta_4$ 。

VARIABLES	(1) wage	(2) wage
age	-0.058 (-1.48)	-0.068* (-1.71)
south	-1.539*** (-6.52)	
c_city	0.605** (2.24)	
never_married	0.599 (1.40)	
Constant	10.461*** (6.60)	10.430*** (6.59)
Observations	2,246	2,246
R-squared	0.022	0.001

Robust t-statistics in parentheses

*** p<0.01, ** p<0.05, * p<0.1

图 4.1: 例题 4.1 回归结果

解 不受 H_0 约束下的回归为(1), 受 H_0 约束下的回归为(2)。所以 $R_u^2 = 0.022$, $R_r^2 = 0.001$, $n = 2246$, $k_u = 4$, $q = 3$ 。所以 $F = \frac{(0.022-0.001)/3}{(1-0.022)/(2246-4-1)} = 16.040$, 由于 $F > \chi_{5\%}^2(3)/3 = 2.60$, 故在 5% 的显著性水平下拒绝原假设。

4.5.4 LSA1 假设放松

经过以上论述, 我们明白了多元线性回归建立的基本流程, 但是回归模型假设, 我们有个问题: 是不是我们加入的每个变量都要保证误差项的条件均值 $E(u_i|X_{1i}, X_{2i}, \dots, X_{ni}) = 0$? 因为我们引入多元线性回归模型, 主要目的是尽可能去除遗漏变量偏差的影响, 进而估计我们关心的系数 β_1 , 至于 $\beta_2, \beta_3, \dots, \beta_n$ 我们都不 care。那么, 基于这个背景下我们能不能对 LSA1 进行放松? 答案是可以的。

如果我们将条件均值为 0, 改为条件均值独立, 也就是对于多元线性回归

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i \quad (4.35)$$

有

$$E(u_i|X_i, W_i) = E(u_i|W_i) \quad (4.36)$$

式中的 W_i 叫做控制变量。意思是使得整体随机误差项的随机误差项的条件均值最终可以由控制变量所表述, 即仅为控制变量的函数。这也就是为什么称 W_i 为“控制变量”的原因。

现在我们来看看, 假设放松后会有什么结论。

首先来看 β_1 是否具有因果效应解释。

$$\begin{aligned} & E(Y_i|X_i = x + \Delta x, W_i) - E(Y_i|X_i = x, W_i) \\ &= (\beta_0 + \beta_1 X_i + \beta_2 W_i + E(u_i|X_i = x + \Delta x, W_i)) - (\beta_0 + \beta_1 X_i + \beta_2 W_i + E(u_i|X_i = x, W_i)) \\ &= \beta_1 \Delta x + (E(u_i|X_i = x + \Delta x, W_i) - E(u_i|X_i = x, W_i)) \\ &= \beta_1 \Delta x + (E(u_i|W_i) - E(u_i|W_i)) \\ &= \beta_1 \Delta x \end{aligned} \quad (4.37)$$

显然是有的。

其次再来看无偏性。仅考虑条件均值为线性函数的情况, 即

$$E(u_i|X_i, W_i) = E(u_i|W_i) = \gamma_0 + \gamma_1 W_i \quad (4.38)$$

令 $v_i = u_i - E(u_i|X_i, W_i)$, 则

$$E(v_i|X_i, W_i) = E(u_i|X_i, W_i) - E[E(u_i|X_i, W_i)] = E(u_i|X_i, W_i) - E(u_i|X_i, W_i) = 0 \quad (4.39)$$

令 $\delta_0 = \beta_0 + \gamma_0, \delta_2 = \beta_2 + \gamma_1$, 原方程改写为:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 W_i + \gamma_0 + \gamma_1 W_i + v_i \\ &= (\beta_0 + \gamma_0) + \beta_1 X_i + (\beta_2 + \gamma_1) W_i + v_i \\ &= \delta_0 + \beta_1 X_i + \delta_2 W_i + v_i \end{aligned} \quad (4.40)$$

而该式中 $E(v_i|X_i, W_i) = 0$, 这表明放松LSA1后实际估计的是式4.40, 由于 X_i 前面的系数不变, 故有 $E(\hat{\beta}_1) = \beta_1$ 。

但是由于 $\delta_2 = \beta_2 + \gamma_1 \neq \beta_2$, $E(\hat{\beta}_2) = \delta_2 = \beta_2 + \gamma_1 \neq \beta_2$ 。

故条件均值独立有如下结论:

1. β_1 估计具有因果效应
2. $\hat{\beta}_1$ 无偏
3. $\hat{\beta}_2$ 有偏

结论 以上便是多元线性回归的基本内容, 他解决了一元线性回归最大的问题: 可测的遗漏变量偏差。但除此之外, 一元线性回归还可能会存在其他模型设定形式上的问题。我们后面一一阐述。

第4章 练习

1. 请比较一元和多元线性回归中, SER、 R^2 、调整 R^2 的区别, 并说明为啥一元线性回归没人用调整 R^2
2. 请写出联合假设下, 同方差的F统计量, 并解释为啥F统计量越大, 说明越能拒绝原假设。
3. 线性回归中, 如果 R^2 和调整 R^2 都很低, 那么是不是说明回归效果不好呢?
4. stata自带的nlsw88数据集是一个关于妇女工资影响因素的数据集, wage代表的是妇女工资, age是妇女岁数, ttl_exp是妇女的工作经验(已量化成连续型数据)。回归结果如图4.2, 回答以下问题:
(1)根据回归结果(1), 写出age的系数解释、置信区间、p值, 并检验系数是否显著。 (2)为什么加入了妇女

VARIABLES	(1) wage	(2) wage
age	-0.132*** (-3.46)	-0.068* (-1.71)
ttl_exp	0.342*** (15.16)	
Constant	8.649*** (5.64)	10.430*** (6.59)
Observations	2,246	2,246
R-squared	0.075	0.001

Robust t-statistics in parentheses

*** p<0.01, ** p<0.05, * p<0.1

图4.2: nlsw88回归结果

工作经验ttl_exp后, 回归系数变小了? 并分析变小的原因。

(3)现在有定性变量race, 取值为black,white,other。创建虚拟变量

$$D_{1i} = \begin{cases} 1, \text{race是black} \\ 0, \text{race不是black} \end{cases}$$

$$D_{2i} = \begin{cases} 1, \text{race是white} \\ 0, \text{race不是white} \end{cases}$$

和

$$D_{3i} = \begin{cases} 1, \text{race是other} \\ 0, \text{race不是other} \end{cases}$$

， 并将三个虚拟变量全部引入回归模型(1)， 试问有什么后果？

(4)试问解决了(3)的问题后， 还会存在哪些遗漏变量？

第5章 非线性回归模型

本章我们根据前面的一元线性回归进行拓展。为的是解决一元线性回归中因果效应非常数的问题。

5.1 因果效应非常数

在经济数据当中，因果效应非常数——即X变化一单位，Y变动是某些变量的函数的情况非常常见。比如，考虑微观经济学中的Cobb-Douglas生产函数模型: $Y = AL^\alpha K^\beta \epsilon$ ，Y是GDP，A为技术水平，L是劳动力，K是资本， ϵ 是扰动项。让L变化一个单位，Y变化的均值显然并不是个常数。再比如实际中的例子：Boston房价数据集中，给人口处于较低地位的百分比lstat和波士顿郊区房价中位数mdev画个散点图（图 5.1），从图上看明显不是线性关系。所以我们有必要对原来的一元线性回归模型进行一定程度上的拓展。

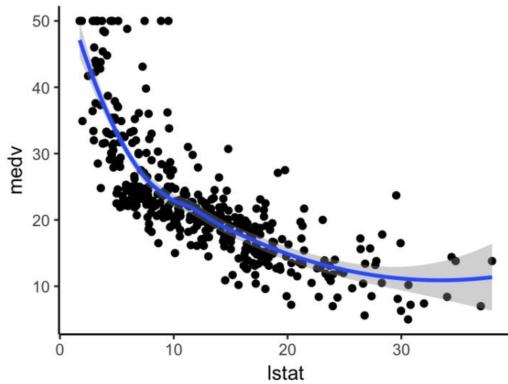


图 5.1: lstat 和 mdev 散点图

5.2 多项式回归

首先第一个想法：对于任意连续函数，我们都可以用多项式进行逼近(Weierstrass第一逼近定理)，那么我们不妨就利用这点对原来一元线性回归模型进行修正，写成 $Y_i = P(X_i) + u_i$ 的形式。即：

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_r X_i^r + u_i \quad (5.1)$$

这样就能解决因果效应非常数的问题。但是有个明显的缺点，因果效应不好解释。考虑阶数 $r=2$ ，比如我们让 X_i 变化一个单位， Y_i 的均值变化为

$$\begin{aligned} & E(Y_i|X_i = x + \Delta x) - E(Y_i|X_i = x) \\ &= (\beta_0 + \beta_1(x + \Delta x) + \beta_2(x + \Delta x)^2 + E(u_i|X_i = x + \Delta x)) - (\beta_0 + \beta_1x + \beta_2x^2 + E(u_i|X_i = x)) \\ &= \beta_1\Delta x + \beta_2(2x\Delta x + (\Delta x)^2) \end{aligned} \quad (5.2)$$

可以看到非常难解释 β_1 和 β_2 。这也就是计量论文中几乎没见到用多项式的原因。（但是作为拟合和预测还是常用的）。

对于该模型的估计，我们还是用OLS。实践中，我们会利用原本的 X_i 产生 X_i 的高次项 $X_i^2, X_i^3, \dots, X_i^r$ ，然后将他们统一放到模型中进行OLS回归。

模型假设与原来相同。如果你觉得有必要加入控制变量那就加进去。（**但是到了多重共线性，数值上，多项回归式可能会产生较大多重共线性。比如 X_i 在1附近的时候， $X_i \approx X_i^2 \approx \dots \approx X_i^r$ ，这时候用多项式回归就有比较大的问题。）

对于假设检验，根据原假设不同，我们分为以下几种类型。考虑 $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3$ ：

1. H_0 : 回归函数为多项式， H_1 : 回归函数为线性函数。此时等价于检验 $H_0 : \beta_2 = \beta_3 = 0$

2. H_0 : 回归函数为三次多项式, H_1 : 回归函数为二次多项式。此时等价于检验 $H_0 : \beta_3 = 0$
 3. $H_0 : X_i$ 对 Y_i 有因果效应, $H_1 : X_i$ 对 Y_i 没有因果效应。此时所有带 X_i 项的系数都要检验, 即检验 $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

5.3 对数回归模型

针对多项式回归因果效应不好解释的问题, 对数回归模型成为一个更好的选择。他有三种形式:

1. 线性对数模型 $Y_i = \beta_0 + \beta_1 \ln X_i + u_i$
2. 对数线性模型 $\ln Y_i = \beta_0 + \beta_1 X_i + u_i$
3. 双对数模型 $\ln Y_i = \beta_0 + \beta_1 \ln X_i + u_i$

如果你觉得容易混淆, 可以按以下方式: 前面两个字对应的是 Y_i , 后面两个字对应的是 X_i 。比如, 线性对数模型中, "线性" 对应 Y_i , "对数" 对应 X_i , 所以总的回归模型为

$$Y_i = \beta_0 + \ln X_i + u_i$$

对数回归模型重在掌握他的因果效应解释。但其实都大同小异。首先对 $\ln(1+a)$ 在 $a=0$ 处进行 Taylor 展开有

$$\ln(1+a) = 0 + a + o(a) \quad (5.3)$$

当 a 很小时, 有近似

$$\ln(1+a) \approx a \quad (5.4)$$

利用这点, 我们考虑当 $\frac{\Delta x}{x}$ 很小时,

$$\ln(x + \Delta x) - \ln(x) = \ln\left(1 + \frac{\Delta x}{x}\right) \approx \frac{\Delta x}{x} \quad (5.5)$$

5.3.1 线性对数模型

对于线性对数模型下的总体回归函数,

$$Y = \beta_0 + \beta_1 \ln X \quad (5.6)$$

给 X 一个 ΔX 的变动, Y 变动

$$\begin{aligned} \Delta Y &= \beta_1 (\ln(X + \Delta X) - \ln(X)) \\ &= \beta_1 \frac{\Delta X}{X} \end{aligned} \quad (5.7)$$

所以系数解释为: X_i 变化 1% 的时候, Y_i 变化 $0.01\beta_1$

5.3.2 对数线性模型

对于一般的对数线性函数:

$$\ln(Y) = \beta_0 + \beta_1 X \quad (5.8)$$

给 X 一个 ΔX 的变动,

$$\begin{aligned} \ln(Y + \Delta Y) - \ln(Y) &= \beta_1 (X + \Delta X) - X \\ &= \beta_1 \Delta X \end{aligned} \quad (5.9)$$

而有

$$\frac{\Delta Y}{Y} = \ln(Y + \Delta Y) - \ln(Y) = \beta_1 \Delta X \quad (5.10)$$

所以系数解释为: X_i 变化 1 单位的时候, Y_i 变化 $100\beta_1\%$

5.3.3 双对数模型

对双对数模型下的总体回归函数：

$$\ln(Y) = \beta_0 + \beta_1 X \quad (5.11)$$

给 X 一个 ΔX 的变动，

$$\begin{aligned} \ln(Y + \Delta Y) - \ln(Y) &= \beta_1(\ln(X + \Delta X) - \ln(X)) \\ &= \beta_1 \frac{\Delta X}{X} \end{aligned} \quad (5.12)$$

又有

$$\frac{\Delta Y}{Y} = \ln(Y + \Delta Y) - \ln(Y) = \beta_1 \frac{\Delta X}{X} \quad (5.13)$$

所以系数解释为： X_i 变化1%的时候， Y_i 变化 $100\beta_1\%$

或者也可以利用微观经济学中弹性的概念， β_1 为 X_i 对 Y_i 的弹性。

模型估计也是把 X_i 或者 Y_i 转化成对数形式进行线性回归。剩下的模型假设等等与一元线性回归相同。

但涉及到拟合效果评价，值得注意的是所有评价指标，比如 R^2 ，一元线性回归只能和线性对数模型进行比较，对数线性模型只能和线性对数模型进行比较。这是因为我们回归时的因变量不同，前者是单纯的 Y_i ，后者是 $Z_i = \ln(Y_i)$ ，两个因变量不同没法进行拟合效果的比较。

5.4 交互项回归

有的时候，因果效应不是关于 X_i 的函数，他有可能是关于别的变量的函数，比如如下式子：

$$\begin{aligned} Y_i &= \beta_0 + (\beta_1 + \beta_3 W_i)X_i + \beta_2 W_i + u_i \\ &= \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3(X_i * W_i) + u_i \end{aligned} \quad (5.14)$$

显然当你控制住 W_i 不变时，他确实是个常数，但要是 W_i 本身不是常数时，此时因果效应也不是常数。这时候我们称 $X_i * W_i$ 为交互项。

要理解这点，我们依次来看看以下情形。

5.4.1 X_i 和 W_i 是二值变量

当 X_i 和 W_i 是二值变量时，对于 Y_i 的条件期望取值有以下四种情况。

	$W_i = 1$	$W_i = 0$
$X_i = 1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_1$
$X_i = 0$	$\beta_0 + \beta_2$	β_0

此时，如果我们要解释交互项前面的系数 β_3 ，由上表：

$$E(Y_i|X_i = 1, W_i = 1) - E(Y_i|X_i = 1, W_i = 0) = \beta_2 + \beta_3 \quad (5.15)$$

和

$$E(Y_i|X_i = 0, W_i = 1) - E(Y_i|X_i = 0, W_i = 0) = \beta_2 \quad (5.16)$$

由于 $E(Y_i|X_i = x, W_i = 1) - E(Y_i|X_i = x, W_i = 0)$ 可以看做是在不同 X_i 取值下， W_i 对 Y_i 均值的效应。为此， β_3 的系数解释为： $X_i = 1$ 和 $X_i = 0$ 时， W_i 对 Y_i 均值的效应之差。

5.4.2 X_i 是连续变量, W_i 是二值变量

当 X_i 是连续变量, W_i 是二值变量时, 在 W_i 不同取值下 Y_i 对 X_i 进行回归, 那么我们有如下式子:

$$\begin{cases} Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_i + u_i & W_i = 1 \\ Y_i = \beta_0 + \beta_1 X_i + u_i & W_i = 0 \end{cases} \quad (5.17)$$

分以下几种情况, 如图 5.2 :

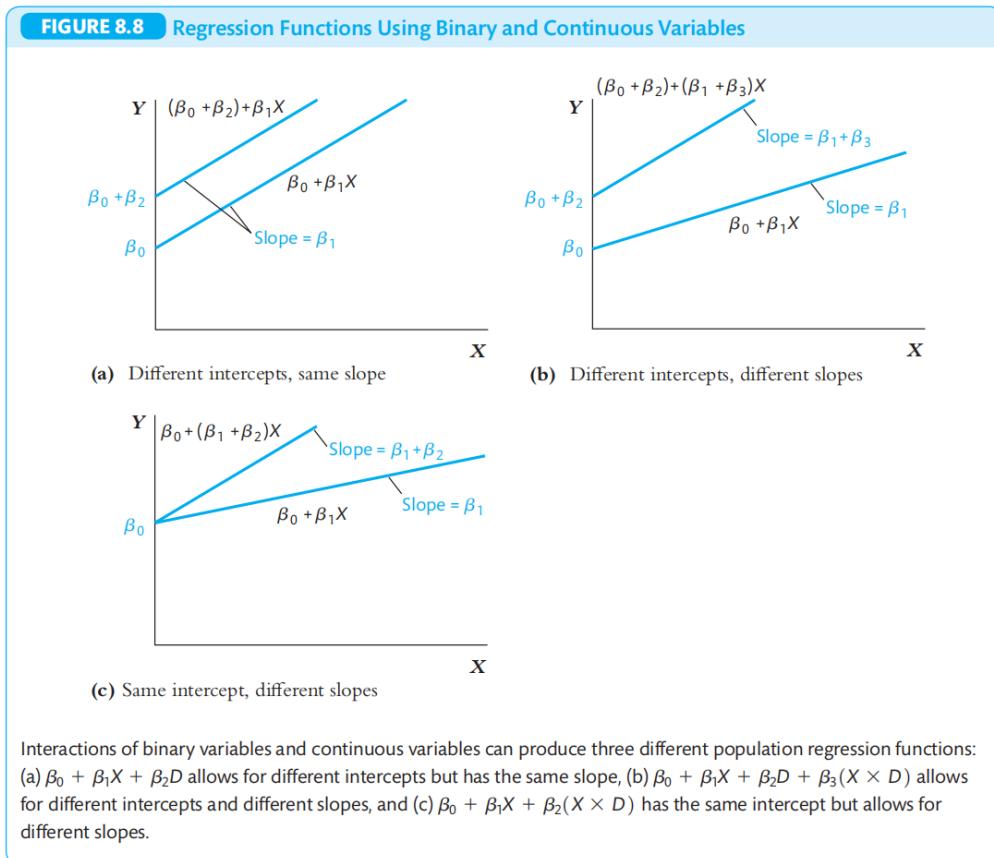


图 5.2: 连续变量和二值变量

斜率相同, 截距相同 由于两条直线完全重合, 此时, $\beta_2 = \beta_3 = 0$, 这里并没有在图中显示出。

斜率相同, 截距不同 如图(a), 由于斜率相同但截距不同, 我们有 $\beta_3 = 0$ 且 $\beta_2 \neq 0$

斜率不同, 截距相同 如图(c), 此时截距相同斜率不同, 我们有 $\beta_3 = 0$ 且 $\beta_2 \neq 0$

斜率不同, 截距不同 如图(b), 这种情况下, 我们有 $\beta_2 \neq 0$ 且 $\beta_3 \neq 0$

对于因果效应解释, 由于

$$E(Y_i|X_i = x, W_i = 1) - E(Y_i|X_i = x, W_i = 0) = \beta_2 + \beta_3 x \quad (5.18)$$

同样上式可以看做是不同 X_i 取值下, W_i 对 Y_i 均值的效应。为此, β_3 的系数解释为: X_i 增加一单位时, W_i 对 Y_i 均值的效应的增量。

5.4.3 X_i 和 W_i 是连续变量

通过对以上的探讨, 我们发现对于交互项因果效应的解释无非是一个“增量”的概念, 我们推广到连续型

变量:

$$E(Y_i|X_i = x, W_i = w + 1) - E(Y_i|X_i = x, W_i = w) = \beta_2 + \beta_3 x \quad (5.19)$$

为此有因果效应解释还是: X_i 增加一单位时, W_i 对 Y_i 均值的效应的增量。

后续的假设检验与多元线性回归是相同的方式。比较特殊的是以下两种情况, 注意区分。

1. $H_0: X_i$ 对 Y_i 的因果效应是常数。这种情况下只要检验交互项系数 $\beta_3 = 0$ 即可。

2. $H_0: W_i$ 对 Y_i 有影响。这种情况下, 需要检验所有带 W_i 的项, 即 $H_0: \beta_2 = \beta_3 = 0$ 。

**另外, 值得注意的是, 这种非线性回归也可能存在一定的多重共线性。因为 $X_i * W_i$ 这个乘积在 W_i 集中密集在某个数的时候可能正比于 X_i 。所以对于非线性回归还是要谨慎使用。

第5章 练习

1. 梳理下对数回归模型的因果效应解释。

2. 考虑回归 $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 (X_i * W_i) + u_i$ 。假设我们对于不同变量下的 Y_i 对应的均值如下: 请

	$W_i = 1$	$W_i = 0$
$X_i = 1$	664.1	666.2
$X_i = 0$	645.9	640.5

你根据以上表格估计模型系数? (提示: 条件期望可以用在给定条件下平均值作为相应估计)

3. 假设我们在不同二值变量 D_i 的取值下, 我们有

$$\begin{cases} Y_i = \beta_{01} + \beta_{11} X_i + u_i & D_i = 1 \\ Y_i = \beta_{02} + \beta_{12} X_i + u_i & D_i = 0 \end{cases} \quad (5.20)$$

现在你想看看两个回归方程的斜率是否相同, 本质上是在做什么检验? 并写出检验过程。

4. * (DID, 双重差分法) 我们考虑以下实验:

某地方政府正在考虑修建地铁路对本地GDP的影响。现在该政府拿到往年每次修建本地铁路前后的GDP数据集如下:

- T_i : 时间
- W_i : 此时是否已经修建铁路。如果是取1, 否则取0.
- Y_i : 此时的本地GDP

简便起见, 假设除了修建铁路这个因素之外的其他条件不变情况下, GDP大致呈未知常数稳定增长。那么我们考虑回归如下:

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2 T_i + \beta_3 (W_i * T_i) + u_i$$

回答以下问题:

- (1)为什么不能直接拿 Y_i 对 W_i 进行回归?
- (2)请你解释交互项系数。并写出假设检验步骤。
- (3)请你检验修建铁路是否对GDP有显著影响?

第6章 线性回归常见问题

首先需要明白的一点是，我们做计量经济学的回归，所需要考虑的是因果效应计量得准不准确，换言之，你这个回归有不有效。那么，对于这个有效性，我们分为两种有效性。

1. 内部有效性：对于所研究总体，通过回归做出的因果推断结论都是正确的。我们称这种情况为内部有效性。
2. 外部有效性：你做出的结论不光对研究总体有效，还能外推到其他研究总体上，这种情况称之为外部有效。

举个例子：考虑地方城投债的增发对地方GDP有什么促进作用。中国有34个省级行政区。如果你考虑浙江地区的城投债增发，那么如果整个过程没有任何问题，那么这时候内部有效性就达到了。然而，如果你想外延到整个中国，比如贵州这些地区，地方政府不容易违约，债券是发出来了但是没人敢买，那么可能效果就没浙江那么明显。这时候就说明外部有效性没有达到。但是，如果你考虑的是相似的比如福建、上海这种同样是东部沿海城市，由于情况都类似，那么外部有效性就比较容易达到。

一般而言，我们更多关心的是内部有效性问题。因为有些时候内部有效性不好达到更别提外部有效性了。在我们的初级计量当中，线性回归一般来说有五大内部有效性问题：

1. 遗漏变量偏差
2. 回归函数非线性
3. 测量误差
4. 缺失数据和样本选择
5. 双向因果

我们依次介绍如下。

6.1 遗漏变量偏差

该问题我们在前面探讨过。无非是两个条件：一、 W_i 是个遗漏变量，对 Y_i 有影响但是没包括在模型里。二、 W_i 和 X_i 具有相关性。如果这两个条件成立，那么就会导致遗漏变量偏差， β_1 的估计是有偏且不一致的（方向也可能是反的）。

遗漏变量偏差分两种，一种是可测的遗漏变量偏差，即你遗漏的这个变量能拿到确确实实的定量数据，解决比较简单，把遗漏变量丢进去完事。另一种是不可测的遗漏变量偏差，这种情况下你只能通过后续的面板数据回归或者工具变量回归解决。

6.2 回归函数非线性

这个问题在上节部分有讨论过，如果被解释变量他是个连续型变量，那么这里的核心问题是因果效应非函数，解决方式有三个：多项式、对数、交互项。如果是个二值变量，那么可以考虑后面学的Logit/Probit回归。

6.3 测量误差

什么是测量误差？回忆一下你初中学的，如果你用的尺子刻度不准，那么实际的长度和你测量的长度就会有偏差。经济学中的测量误差也是如此。比如，你在问卷上发布了个调查问卷，但是有可能填问卷的人小学语文没学好给出了错误答案，再或者是问卷上只能填写整数值而只好给出了近似的数值，这都属于测量误差。

首先我们来看看解释变量 X_i 如果有测量误差会有什么影响。假设真实的数据为 X_i ，你收集到的数据为 \widetilde{X}_i 。那么测量误差

$$v_i = \widetilde{X}_i - X_i \quad (6.1)$$

为了探讨这个测量误差问题，我们假设LSA1： $E(u_i|X_i) = 0$ 成立。即有

$$Cov(u_i, X_i) = 0 \quad (6.2)$$

此时 X_i 的测量误差有两种特殊形式：传统测量误差和最佳猜测测量误差。

6.3.1 传统测量误差

传统测量误差顾名思义，就是你的测量误差非常传统，跟你初中物理学的一毛一样，单纯是量不准导致的纯纯的噪声。不妨假设它和其他能够影响 Y_i 的其他因素 u_i 一般没半毛钱关系，即

$$Cov(v_i, X_i) = 0 \quad (6.3)$$

和

$$Cov(v_i, u_i) = 0 \quad (6.4)$$

或者通过 $Cov(X_i, u_i) = 0$ ，也可以写成 $Cov(\tilde{X}_i, u_i) = 0$

由于我们实际回归的是 \tilde{X}_i ，即

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ &= \beta_0 + \beta_1 (X_i + v_i) + (u_i - \beta_1 v_i) \\ &= \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i \end{aligned} \quad (6.5)$$

式中， $\tilde{u}_i = u_i - \beta_1 v_i$

而

$$\tilde{\beta}_1 = \beta_1 + \frac{\frac{1}{n-1} \sum_{i=1}^n (\tilde{X}_i - \bar{\tilde{X}})(\tilde{u}_i - \bar{\tilde{u}})}{\frac{1}{n-1} \sum_{i=1}^n (\tilde{X}_i - \bar{\tilde{X}})^2} \xrightarrow{P} \beta_1 + \frac{Cov(\tilde{X}_i, \tilde{u}_i)}{Var(\tilde{X}_i)} \quad (6.6)$$

分子：

$$\begin{aligned} Cov(\tilde{X}_i, \tilde{u}_i) &= Cov(X_i + v_i, u_i - \beta_1 v_i) \\ &= Cov(X_i, u_i) + Cov(v_i, u_i) - \beta_1 Cov(X_i, v_i) - \beta_1 Var(v_i) \\ &= -\beta_1 \sigma_v^2 \end{aligned} \quad (6.7)$$

分母：

$$\begin{aligned} Var(\tilde{X}_i) &= Var(X_i + v_i) \\ &= \sigma_X^2 + \sigma_v^2 \end{aligned} \quad (6.8)$$

总的有

$$\tilde{\beta}_1 \xrightarrow{P} \beta_1 - \left(\frac{\sigma_v^2}{\sigma_X^2 + \sigma_v^2} \right) \beta_1 = \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_v^2} \right) \beta_1 \quad (6.9)$$

所以，可以看出，传统测量误差越大，说明回归系数整体越靠近0（注意不是低估，如果是负的就是高估）。并且，测量误差方差趋近于无穷时 $\tilde{\beta}_1 = 0$ 。（也可以这么理解：回归的 \tilde{X}_i 此时接近于一个超大的噪声，相当于你随便生成个随机数，拿 Y_i 对它进行回归，自然是解释不了 Y_i 啥东西的，此时回归系数必然为0）

对于被解释变量 Y_i ，传统测量误差由于相当于噪声，可以被随机误差项 u_i 吸收，有测量误差也不会影响系数的一致性和无偏性。但是 $\tilde{\beta}_1$ 的方差会比原来大，如果这个测量误差方差特别大，那么如果样本不够大有可能影响具体的假设检验（LSA3不成立）。

6.3.2 最佳猜测测量误差

这种测量误差比较特殊，他要求被调查者可以不记得 X_i ，但是要求你得根据你的背景给一个 X_i 一个尽可能的猜测，即 $\tilde{X}_i = E(X_i|\Omega_i)$ 。

也就是说，对于测量误差

$$\omega_i = \widetilde{X}_i - X_i = E(X_i|\Omega_i) - X_i \quad (6.10)$$

有

$$E(\omega_i|\Omega_i) = E(X_i|\Omega) - E(X_i|\Omega) = 0 \quad (6.11)$$

即你在给定所有信息下，测量误差的期望理应等于0.

假设 $E(u_i|\Omega_i) = 0$ 。此时这个测量误差有

$$\begin{aligned} Cov(\widetilde{X}_i, \widetilde{u}_i) &= Cov(\widetilde{X}_i, u_i - \beta_1 \omega_i) \\ &= Cov(\widetilde{X}_i, u_i) - \beta_1 Cov(\widetilde{X}_i, \omega_i) \end{aligned} \quad (6.12)$$

对于前者，由于

$$\begin{aligned} Cov(\widetilde{X}_i, u_i) &= E(\widetilde{X}_i u_i) - E(\widetilde{X}_i)E(u_i) \\ &= E[E(\widetilde{X}_i u_i|\Omega_i)] - E(\widetilde{X}_i)E(E(u_i|\Omega_i)) \\ &= E[\widetilde{X}_i E(u_i|\Omega_i)] - 0 \\ &= 0 \end{aligned} \quad (6.13)$$

对于后者，由于

$$\begin{aligned} Cov(\widetilde{X}_i, \omega_i) &= E[E(\widetilde{X}_i \omega_i|\Omega_i)] - E(\widetilde{X}_i)E(\omega_i) \\ &= E[\widetilde{X}_i E(\omega_i|\Omega_i)] - E(\widetilde{X}_i)E(E(\omega_i|\Omega_i)) \\ &= 0 - 0 \\ &= 0 \end{aligned} \quad (6.14)$$

所以有

$$Cov(\widetilde{X}_i, \widetilde{u}_i) = 0 \quad (6.15)$$

此时，通过上式可以证明 $\hat{\beta}_1$ 是一致的。无偏性也可以证明（但这个可能需要点信息流下条件期望的性质，不再展开）。

对于传统测量误差，我们一般都用工具变量进行解决。最佳猜测的测量误差我们其实并没有太大所谓只要能保持一致性无偏性就行，但由于这种情况下误差项的方差比较大，我们要尽可能保证样本量够多。

注 最佳猜测下的测量误差的证明其实相对复杂，因为这还涉及到怎么去整这个“信息”。如果你不想掌握证明，跳过直接掌握结论即可。但前面传统测量误差的证明你得理解，这还在我们的知识范围内。

6.4 缺失数据和样本选择

由于我们回归的时候必须要有全部变量的数据，所以哪怕你只缺失了一个变量的数据回归都得把这个数据点剔除。换言之，缺失数据本质上是在缺失样本，我们分三种情况，

1. 数据缺失完全随机
2. 数据缺失与X相关
3. 数据缺失与Y相关

6.4.1 数据缺失完全随机

这种情况是最没有影响的。由于你本身样本抽取是随机的，数据缺失也是随机的，结果也就是少抽了一点样本。如果你有1000份问卷，你家养的狗（尤其哈士奇、柴犬等大型犬）拆家过程中把问卷吃了500份，那么不要慌，问题不大，家没了，问卷还在，狗的行为不会引起任何偏差。



图 6.1: 狗的行为不会引起任何偏差

6.4.2 数据缺失与X相关

如果你养的狗比较聪明，他会根据X的变量大小来吃问卷。那么，因为我们研究的是X的变动对Y的影响，如果只考虑内部有效性，你最多也就是样本量小了，研究范围小了。狗不仅不会引起任何偏差，还帮你减轻了工作负担（善解人意的狗）。

6.4.3 数据缺失与Y相关

如果你养的狗是根据Y的大小去吃问卷，那么这时候建议你要么把狗送人，要么好好教育你的狗，要么把狗炖了下饭。因为这时候可能涉及到样本选择偏差——用于分析的数据不再是从总体中简单随机抽样而得，他经过了你家狗的二次加工，不再代表原来的研究总体。

我们举个简单的例子。假设你找到一系列的特定环境下人狗发生冲突的新闻，但是如果是非常猎奇的新闻媒体，只会报道人咬狗的事件，不会报道狗咬人的事件，那么无论你的环境怎么变化，结果都是人咬狗，而不是狗咬人。这种情况下你估计的回归系数基本上是有偏的，他不可能代表原来总体的回归系数。



图 6.2: 人咬狗事件

至于怎么解决，一般来说，如果你能找到是啥原因导致的 Y_i 减少并且能把他数据找到，那么直接用多元线性回归。但是一般来说我们都找不到，我们更多的会用工具变量解决。

6.5 双向因果

双向因果顾名思义，你可能会发现X能影响Y，但实际上Y也能影响X。举个例子，我们想估计哈士奇幼崽的需求弹性。价格上升，购买哈士奇的需求自然下降。但需求下降时，供给大于需求，厂商自然会降价。这样，价格上升，需求下降，但需求减少，价格也会下降，但形成双向因果。

那么，双向因果有哪些影响呢？我们看以下回归方程：

$$\ln Q_i = \beta_0 + \beta_1 \ln P_i + u_i \quad (6.16)$$

式中， β_1 可以解释为需求弹性，这由上一章的内容可以得到。

那么，假设我们有一个外部冲击，比如市场上多出一群人嫌钱太多想养哈士奇，这个冲击完全来源于外部，和 $\ln P_i$ 没关系，那么我们只能看成是 u_i 增加导致的 $\ln Q_i$ 增加，但是，由于此时需求大于供给，价格上升即 $\ln P_i$ 增加，所以此时我们会发现： $\text{corr}(\ln P_i, u_i) \neq 0$ ，LSA1不成立。此时 $\hat{\beta}_1$ 估计是有偏且不一致的。

和遗漏变量偏差不同，双向因果不能通过多元线性回归解决，只能通过工具变量分离出和 u_i 不相关的部分进行回归。后续工具变量这块我们还会进一步的讨论。

第6章 练习

1. 请你梳理本节五大问题，并在学完全部课程后看看这几个问题该怎么解决。

第7章 面板数据回归

后面三章是线性回归的拓展模型。他们针对不同的情况能够解决不同的问题。

7.1 面板数据

什么是面板数据？说白了，就是不仅包含个体维度上的数据，还包括时间维度上的数据。我们一般将它记做 X_{it} 。其中， i 是个体维度， t 是时间维度，如图，以国家统计局的城镇人口数据为例，往行的方向是每一年即时间维度数据，往列的方向是每一个地区即个体维度数据。如果你单单挑选某一年，那么你获得的就是截面数据；如果你单单截取某一个地区，那么你获得的是时间序列数据。

地区	2019年	2018年	2017年	2016年	2015年
北京市	1865	1863	1878	1880	1877
天津市	1304	1297	1291	1295	1278
河北省	4374	4264	4136	3983	3811
山西省	2221	2172	2123	2070	2016
内蒙古自治区	1609	1589	1568	1542	1514
辽宁省	2964	2968	2949	2949	2952
吉林省	1568	1556	1539	1530	1523
黑龙江省	2284	2268	2250	2249	2241
上海市	2144	2136	2121	2127	2116
江苏省	5698	5604	5521	5417	5306
浙江省	4095	3953	3847	3745	3645
安徽省	3553	3459	3346	3221	3103
福建省	2642	2594	2534	2464	2403
江西省	2679	2604	2524	2438	2357
山东省	6194	6147	6062	5871	5614
河南省	5129	4967	4795	4623	4441
湖北省	3615	3568	3500	3419	3327
湖南省	3959	3865	3747	3599	3452

图 7.1: 中国城镇人口数据

7.2 不可测的遗漏变量偏差

第8章 二值因变量回归

Copyright to Kunlin Dong @SBF UBE

第9章 工具变量回归

Copyright to Kunlin Dong @SBF UBE

第 10 章 ElegantL^AT_EX 系列模板介绍

ElegantL^AT_EX 项目组致力于打造一系列美观、优雅、简便的模板方便用户使用。目前由 **ElegantNote**, **ElegantBook**, **ElegantPaper** 组成，分别用于排版笔记，书籍和工作论文。强烈推荐使用最新正式版本！本文将介绍本模板的一些设置内容以及基本使用方法。如果您有其他问题，建议或者意见，欢迎在 GitHub 上给我们提交 issues 或者邮件联系我们。

我们的联系方式如下，建议加入用户 QQ 群提问，这样能更快获得准确的反馈，加群时请备注 L^AT_EX 或者 ElegantL^AT_EX 相关内容。

- 官网: <https://elegantlatex.org/>
- GitHub 网址: <https://github.com/ElegantLaTeX/>
- CTAN 地址: <https://ctan.org/pkg/elegantbook>
- 下载地址: 正式发行版, 最新版
- 微博: ElegantL^AT_EX
- 微信公众号: ElegantL^AT_EX
- 用户 QQ 群: 692108391
- 邮件: elegantlatex2e@gmail.com

10.1 ElegantBook 更新说明

此次为 4.x 第一个版本，在 3.x 基础上，主要更新了定理以及参考文献的支持方式，具体内容有：

1. **重要改动:** 由原先的 Bib^LT_EX 改为 biblatex 编译方式（后端为 biber），请注意两者之间的差异；
2. **重要改进:** 修改对于定理写法兼容方式，提高数学公式代码的兼容性；
3. 页面设置改动，默认页面更宽；方便书写和阅读；
4. 支持目录文字以及页码跳转；
5. 不再维护 pdfL^AT_EX 中文支持方式，请务必使用 X_EL^AT_EX 编译中文文稿。
6. 增加多语言选项，法语 lang=fr、德语 lang=de、荷兰语 lang=nl、匈牙利语 lang=hu、西班牙语 lang=es、蒙古语 lang=mn 等。

 **笔记** 如果你使用旧版本切换到新版本时，遇到问题时，请核对文档中是否有 pageanchor 字样。如果有，请删除文档中的 \hypersetup{pageanchor=true}，并且在 \maketitle 和 \tableofcontents 之间添加 \frontmatter。2.x 版本的用户请仔细查看跨版本转换。

10.2 模板安装与更新

你可以通过免安装的方式使用本模板，包括在线使用和本地（文件夹内）使用两种方式，也可以通过 T_EX 发行版安装使用。

10.2.1 在线使用模板

我们把三套模板全部上传到 Overleaf 上了，网络便利的用户可以直接通过 Overleaf 在线使用我们的模板。使用 Overleaf 的好处是无需安装 T_EX Live 2020，可以随时随地访问自己的文件。查找模板，请在 Overleaf 模板库里面搜索 elegantlatex 即可，你也可以直接访问 [搜索结果](#)。选择适当的模板之后，将其 Open as Template，即可把模板存到自己账户下，然后可以自由编辑以及与别人一起协作。更多关于 Overleaf 的介绍和使用，请参考 Overleaf 的[官方文档](#)。

注 Overleaf 上，中文需要使用 X_EL^AT_EX 进行编译，英文建议使用 pdfL^AT_EX 编译。

10.2.2 本地免安装使用

免安装使用方法如下，从 GitHub 或者 CTAN 下载最新版，严格意义上只需要类文件 `elegantbook.cls`。然后将模板文件放在你的工作目录下即可使用。这样使用的好处是，无需安装，简便；缺点是，当模板更新之后，你需要手动替换 `cls` 文件。

10.2.3 发行版安装使用

本模板测试环境为 Win10 和 TeX Live 2021，如果你刚安装 TeX Live 2021 用户，安装后建议升级全部宏包，升级方法：使用 cmd 运行 `tlmgr update --all`，如果 `tlmgr` 需要更新，请使用 cmd 运行 `tlmgr update --self`，如果更新过程中出现了中断，请改用 `tlmgr update --self --all --reinstall-forcibly-removed` 更新。

10.2.4 更新问题

如果使用 `tlshell` 无法更新模板，请使用命令行全部更新全部宏包或者使用免安装的方法使用本模板。

通过命令行（管理员权限）输入下面的命令对 `tlmgr` 自身和全部宏包进行更新。

```
tlmgr update --self
tlmgr update --all
```

更多的内容请参考 [How do I update my TeX distribution?](#)

10.2.5 其他发行版本

由于宏包版本问题，本模板不支持 CTeX 套装，请务必安装 TeX Live。更多关于 TeX Live 的安装使用以及 CTeX 与 TeX Live 的兼容、系统路径问题，请参考官方文档以及 [一份简短的关于安装 L^AT_EX 安装的介绍](#)。

10.3 关于提交

出于某些因素的考虑，ElegantL^AT_EX 项目自 2019 年 5 月 20 日开始，不再接受任何非作者预约性质的提交（pull request）！如果你想改进模板，你可以给我们提交 issues，或者可以在遵循协议（LPPL-1.3c）的情况下，克隆到自己仓库下进行修改。

第 11 章 ElegantBook 设置说明

本模板基于基础的 book 文类，所以 book 的选项对于本模板也是有效的（纸张无效，因为模板有设备选项）。默认编码为 UTF-8，推荐使用 TeX Live 编译。本文编写环境为 Win10 (64bit) + TeX Live 2021，英文支持 pdfLaTeX，中文仅支持 XeLaTeX 编译。

11.1 语言模式

本模板内含两套基础语言环境 `lang=cn`、`lang=en`。改变语言环境会改变图表标题的引导词（图，表），文章结构词（比如目录，参考文献等），以及定理环境中的引导词（比如定理，引理等）。不同语言模式的启用如下：

```
\documentclass[en]{elegantbook}  
\documentclass[lang=cn]{elegantbook}
```

除模板自带的两套语言设定之外，由网友提供的其他语言环境设置如下：

- 由 VincentMVV 提供的意大利语翻译 `lang=it`，相关讨论见 [Italian translation](#)；
- 由 abfek66 提供的法语翻译 `lang=fr`，相关讨论见 [Italian translation](#)；
- 由 inktvis75 提供的荷兰语翻译 `lang=nl`，相关讨论见 [Dutch Translation](#)；
- 由 palkotamas 提供的匈牙利语翻译 `lang=hu`，相关讨论见 [Hungarian translation](#)；
- 由 Lisa 提供的德语翻译 `lang=de`，相关讨论见 [Deutsch translation](#)；
- 由 Gustavo A. Corradi 提供的西班牙语的翻译 `lang=es`，相关讨论见 [Spanish translation](#)；
- 由 Altantssooj 提供的蒙古语的翻译 `lang=mn`，相关讨论见 [Mongolian translation](#)。

注 以上各个语言的设定均为网友设定，我们未对上述翻译进行过校对，如果有问题，请在对应的 issue 下评论。并且，只有中文环境 (`lang=cn`) 才可以输入中文。

11.2 设备选项

最早我们在 ElegantNote 模板中加入了设备选项 (`device`)，后来，我们觉得这个设备选项的设置可以应用到 ElegantBook 中¹，而且 Book 一般内容比较多，如果在 iPad 上看无需切边，放大，那用户的阅读体验将会得到巨大提升。你可以使用下面的选项将版面设置为 iPad 设备模式²

```
\documentclass[pad]{elegantbook} %or  
\documentclass[device=pad]{elegantbook}
```

11.3 颜色主题

本模板内置 5 组颜色主题，分别为 `green`³、`cyan`、`blue`（默认）、`gray`、`black`。另外还有一个自定义的选项 `nocolor`。调用颜色主题 `green` 的方法为

```
\documentclass[green]{elegantbook} %or  
\documentclass[color=green]{elegantbook}
```

¹不过因为 ElegantBook 模板封面图片的存在，在修改页面设计时，需要对图片进行裁剪。

²默认为 normal 模式，也即 A4 纸张大小。

³为原先默认主题。

表 11.1: ElegantBook 模板中的颜色主题

	green	cyan	blue	gray	black	主要使用的环境
structure						chapter section subsection
main						definition exercise problem
second						theorem lemma corollary
third						proposition

如果需要自定义颜色的话请选择 `nocolor` 选项或者使用 `color=none`, 然后在导言区定义 `structurecolor`、`main`、`second`、`third` 颜色, 具体方法如下:

```
\definecolor{structurecolor}{RGB}{0,0,0}
\definecolor{main}{RGB}{70,70,70}
\definecolor{second}{RGB}{115,45,2}
\definecolor{third}{RGB}{0,80,80}
```

11.4 封面

11.4.1 封面个性化

从 3.10 版本开始, 封面更加弹性化, 用户可以自行选择输出的内容, 包括 `\title` 在内的所有封面元素都可为空。目前封面的元素有

表 11.2: 封面元素信息

信息	命令	信息	命令	信息	命令
标题	<code>\title</code>	副标题	<code>\subtitle</code>	作者	<code>\author</code>
机构	<code>\institute</code>	日期	<code>\date</code>	版本	<code>\version</code>
箴言	<code>\extrainfo</code>	封面图	<code>\cover</code>	徽标	<code>\logo</code>

另外, 额外增加一个 `\bioinfo` 命令, 有两个选项, 分别是信息标题以及信息内容。比如需要显示User Name: 111520, 则可以使用

```
\bioinfo[User Name]{111520}
```

封面中间位置的色块的颜色可以使用下面命令进行修改:

```
\definecolor{customcolor}{RGB}{32,178,170}
\colorlet{coverlinecolor}{customcolor}
```

11.4.2 封面图

本模板使用的封面图片来源于 pixabay.com⁴, 图片完全免费, 可用于任何场景。封面图片的尺寸为 1280 × 1024, 更换图片的时候请严格按照封面图片尺寸进行裁剪。推荐一个免费的在线图片裁剪网站 fotor.com。用户

⁴感谢 ChinaTeX 提供免费图源网站, 另外还推荐 pexels.com。

QQ 群内有一些合适尺寸的封面，欢迎取用。

11.4.3 徽标

本文用到的 Logo 比例为 1:1，也即正方形图片，在更换图片的时候请选择合适的图片进行替换。

11.4.4 自定义封面

另外，如果使用自定义的封面，比如 Adobe illustrator 或者其他软件制作的 A4 PDF 文档，请把 `\maketitle` 注释掉，然后借助 `pdfpages` 宏包将自制封面插入即可。如果使用 `titlepage` 环境，也是类似。如果需要 2.x 版本的封面，请参考 `etitlepage`。

11.5 章标题

本模板内置 2 套章标题显示风格，包含 `hang`（默认）与 `display` 两种风格，区别在于章标题单行显示（`hang`）与双行显示（`display`），本说明使用了 `hang`。调用方式为

```
\documentclass[hang]{elegantbook} %or
\documentclass[titlestyle=hang]{elegantbook}
```

在章标题内，章节编号默认是以数字显示，也即第 1 章，第 2 章等等，如果想要把数字改为中文，可以使用

```
\documentclass[chinese]{elegantbook} %or
\documentclass[scheme=chinese]{elegantbook}
```

11.6 数学环境简介

在我们这个模板中，我们定义了两种不同的定理模式 `mode`，包括简单模式（`simple`）和炫彩模式（`fancy`），默认为 `fancy` 模式，不同模式的选择为

```
\documentclass[simple]{elegantbook} %or
\documentclass[mode=simple]{elegantbook}
```

本模板定义了四大类环境

- 定理类环境，包含标题和内容两部分，全部定理类环境的编号均以章节编号。根据格式的不同分为 3 种
 - `definition` 环境，颜色为 `main`；
 - `theorem`、`lemma`、`corollary` 环境，颜色为 `second`；
 - `proposition` 环境，颜色为 `third`。
- 示例类环境，有 `example`、`problem`、`exercise` 环境（对应于例、例题、练习），自动编号，编号以章节为单位，其中 `exercise` 有提示符。
- 提示类环境，有 `note` 环境，特点是：无编号，有引导符。
- 结论类环境，有 `conclusion`、`assumption`、`property`、`remark`、`solution` 环境⁵，三者均以粗体的引导词为开头，和普通段落格式一致。

⁵本模板还添加了一个 `result` 选项，用于隐藏 `solution` 和 `proof` 环境，默认为显示（`result=answer`），隐藏使用 `result=noanswer`。

11.6.1 定理类环境的使用

由于本模板使用了 `tcolorbox` 宏包来定制定理类环境，所以和普通的定理环境的使用有些许区别，定理的使用方法如下：

```
\begin{theorem}{theorem name}{label}
The content of theorem.
\end{theorem}
```

第一个必选项 `theorem name` 是定理的名字，第二个必选项 `label` 是交叉引用时所用到的标签，交叉引用的方法为 `\ref{thm:label}`。请注意，交叉引用时必须加上前缀 `thm:`。

在用户多次反馈下，4.x 之后，引入了原生定理的支持方式，也就是使用可选项方式：

```
\begin{theorem}[theorem name] \label{thm:theorem-label}
The content of theorem.
\end{theorem}
% or
\begin{theorem} \label{thm:theorem-without-name}
The content of theorem without name.
\end{theorem}
```

其他相同用法的定理类环境有：

表 11.3: 定理类环境

环境名	标签名	前缀	交叉引用
definition	label	def	<code>\ref{def:label}</code>
theorem	label	thm	<code>\ref{thm:label}</code>
lemma	label	lem	<code>\ref{lem:label}</code>
corollary	label	cor	<code>\ref{cor:label}</code>
proposition	label	pro	<code>\ref{pro:label}</code>

11.6.2 其他环境的使用

其他三种环境没有选项，可以直接使用，比如 `example` 环境的使用方法与效果：

```
\begin{example}
This is the content of example environment.
\end{example}
```

这几个都是同一类环境，区别在于

- 示例环境（`example`）、练习（`exercise`）与例题（`problem`）章节自动编号；
- 注意（`note`），练习（`exercise`）环境有提醒引导符；
- 结论（`conclusion`）等环境都是普通段落环境，引导词加粗。

11.7 列表环境

本模板借助于 `tikz` 定制了 `itemize` 和 `enumerate` 环境，其中 `itemize` 环境修改了 3 层嵌套，而 `enumerate` 环境修改了 4 层嵌套（仅改变颜色）。示例如下

- first item of nesti;
 - second item of nesti;
 - first item of nestii;
 - second item of nestii;
 - first item of nestiii;
 - second item of nestiii.
1. first item of nesti;
 2. second item of nesti;
 - (a). first item of nestii;
 - (b). second item of nestii;
 - I. first item of nestiii;
 - II. second item of nestiii.

11.8 参考文献

此模板使用了 `biber` 来生成参考文献，也即使用 `biblatex` 宏包，在中文示例中，使用了 `gbt7714` 宏包。参考文献示例：[cn1, en2, en3] 使用了中国一个大型的 P2P 平台（人人贷）的数据来检验男性投资者和女性投资者在投资表现上是否有显著差异。

你可以在谷歌学术，Mendeley，Endnote 中获得文献条目（bib item），然后把它们添加到 `reference.bib` 中。在文中引用的时候，引用它们的键值（bib key）即可。注意需要在编译的过程中添加 `biber` 编译。

为了方便文献样式修改，模板引入了 `citestyle` 和 `citestyle` 选项，默认均为数字格式（numeric），如果需要设置为国标 GB7714-2015，需要使用：

```
\documentclass[citestyle=gb7714-2015, bibstyle=gb7714-2015]{elegantbook}
```

如果需要添加排序方式，可以在导言区加入

```
\ExecuteBibliographyOptions{sorting=ynt}
```

启用国标之后，可以加入 `sorting=gb7714-2015`。

11.9 添加序章

如果你想在第一章前面添序章，不改变原本章节序号，可以在第一章内容前面使用

```
\chapter*{Introduction}
\markboth{Introduction}{Introduction}
The content of introduction.
```

11.10 目录选项与深度

本模板添加了一个目录选项 `toc`，可以设置目录为单栏（`onecol`）和双栏（`twocol`）显示，比如双栏显示可以使用

```
\documentclass[twocol]{elegantbook}
\documentclass[toc=twocol]{elegantbook}
```

默认本模板目录深度为 1，你可以在导言区使用

```
\setcounter{tocdepth}{2}
```

将其修改为 2 级目录（章与节）显示。

11.11 章节摘要

模板新增了一个章节摘要环境（introduction），使用示例

```
\begin{introduction}
    \item Definition of Theorem
    \item Ask for help
    \item Optimization Problem
    \item Property of Cauchy Series
    \item Angle of Corner
\end{introduction}
```

效果如下：

内容提要

- | | |
|--|--|
| <input type="checkbox"/> Definition of Theorem
<input type="checkbox"/> Ask for help
<input type="checkbox"/> Optimization Problem | <input type="checkbox"/> Property of Cauchy Series
<input type="checkbox"/> Angle of Corner |
|--|--|

环境的标题文字可以通过这个环境的可选参数进行修改，修改方法为：

```
\begin{introduction}[Brief Introduction]
    ...
\end{introduction}
```

11.12 章后习题

前面我们介绍了例题和练习两个环境，这里我们再加一个，章后习题（problemset）环境，用于在每一章结尾，显示本章的练习。使用方法如下

```
\begin{problemset}
    \item exercise 1
    \item exercise 2
    \item exercise 3
\end{problemset}
```

效果如下：

第 11 章 练习

1. exercise 1
2. exercise 2
3. exercise 3
4. 测试数学公式

$$a^2 + b^2 = c_{2_i}(1, 2)[1, 23] \quad (11.1)$$

注 如果你想把 problemset 环境的标题改为其他文字，你可以类似于 introduction 环境修改 problemset 的可选参数。另外，目前这个环境会自动出现在目录中，但是不会出现在页眉页脚信息中（待解决）。

解 如果你想把 problemset 环境的标题改为其他文字，你可以类似于 introduction 环境修改 problemset 的可选参数。另外，目前这个环境会自动出现在目录中，但是不会出现在页眉页脚信息中（待解决）。

11.13 旁注

在 3.08 版本中，我们引入了旁注设置选项 `marginpar=margintrue` 以及测试命令 `\elegantpar`，但是由此带来一堆问题。我们决定在 3.09 版本中将其删除，并且，在旁注命令得到大幅度优化之前，不会将此命令再次引入书籍模板中。对此造成各位用户的不方便，非常抱歉！不过我们保留了 `marginpar` 这个选项，你可以使用 `marginpar=margintrue` 获得保留右侧旁注的版面设计。然后使用系统自带的 `\marginpar` 或者 `marginnote` 宏包的 `\marginnote` 命令。

注 在使用旁注的时候，需要注意的是，文本和公式可以直接在旁注中使用。

```
% text
\marginpar{margin paragraph text}

% equation
\marginpar{
  \begin{equation}
    a^2 + b^2 = c^2
  \end{equation}
}
```

但是浮动体（表格、图片）需要注意，不能用浮动体环境，需要使用直接插图命令或者表格命令环境。然后使用 `\captionof` 为其设置标题。为了得到居中的图表，可以使用 `\centerline` 命令或者 `center` 环境。更多详情请参考：[Caption of Figure in Marginpar](#)。

```
% graph with centerline command
\marginpar{
  \centerline{
    \includegraphics[width=0.2\textwidth]{logo.png}
  }
  \captionof{figure}{your figure caption}
}

% graph with center environment
\marginpar{
  \begin{center}
    \includegraphics[width=0.2\textwidth]{logo.png}
    \captionof{figure}{your figure caption}
  \end{center}
}
```

第 12 章 字体选项

字体选项独立成章的原因是，我们希望本模板的用户关心模板使用的字体，知晓自己使用的字体以及遇到字体相关的问题能更加便捷地找到答案。

重要提示：从 3.10 版本更新之后，沿用至今的 newtx 系列字体被重新更改为 cm 字体。并且新增中文字体（chinesefont）选项。

12.1 数学字体选项

本模板定义了一个数学字体选项（math），可选项有三个：

1. `math=cm`（默认），使用 L^AT_EX 默认数学字体（推荐，无需声明）；
2. `math=newtx`，使用 `newtxmath` 设置数学字体（潜在问题比较多）。
3. `math=mtpro2`，使用 `mtpro2` 宏包设置数学字体，要求用户已经成功安装此宏包。

12.2 使用 newtx 系列字体

如果需要使用原先版本的 newtx 系列字体，可以通过显示声明数学字体：

```
\documentclass[math=newtx]{elegantbook}
```

12.2.1 连字符

如果使用 newtx 系列字体宏包，需要注意下连字符的问题。

$$\int_{R^q} f(x, y) dy \text{off} \quad (12.1)$$

的代码为

```
\begin{equation}
\int_{R^q} f(x, y) dy \text{of } \kern0pt f
\end{equation}
```

12.2.2 宏包冲突

另外在 3.08 版本中，有用户反馈模板在和 `yhmath` 以及 `esvect` 等宏包搭配使用的时候会出现报错：

```
LaTeX Error:
Too many symbol fonts declared.
```

原因是在使用 `newtxmath` 宏包时，重新定义了数学字体用于大型操作符，达到了 **最多 16 个数学字体** 的上限，在调用其他宏包的时候，无法新增数学字体。为了减少调用非常用宏包，在此给出如何调用 `yhmath` 以及 `esvect` 宏包的方法。

请在 `elegantbook.cls` 内搜索 `yhmath` 或者 `esvect`，将你所需要的宏包加载语句取消注释即可。

```
%%% use yhmath pkg, uncomment following code
% \let\oldwidering\widering
% \let\widering\undefined
% \RequirePackage{yhmath}
% \let\widering\oldwidering
```

```
%%% use esvect pkg, uncomment following code
% \RequirePackage{esvect}
```

12.3 中文字体选项

模板从 3.10 版本提供中文字体选项 `chinesefont`, 可选项有

1. `ctexfont`: 默认选项, 使用 `ctex` 宏包根据系统自行选择字体, 可能存在字体缺失的问题, 更多内容参考 [ctex 宏包官方文档](#)¹。
2. `founder`: 方正字体选项, 调用 `ctex` 宏包并且使用 `fontset=none` 选项, 然后设置字体为方正四款免费字体, 方正字体下载注意事项见后文。
3. `nofont`: 调用 `ctex` 宏包并且使用 `fontset=none` 选项, 不设定中文字体, 用户可以自行设置中文字体, 具体见后文。

注 使用 `founder` 选项或者 `nofont` 时, 必须使用 `XeLaTeX` 进行编译。

12.3.1 方正字体选项

由于使用 `ctex` 宏包默认调用系统已有的字体, 部分系统字体缺失严重, 因此, 用户希望能够使用其它字体, 我们推荐使用方正字体。方正的方正书宋、方正黑体、方正楷体、方正仿宋四款字体均可免费试用, 且可用于商业用途。用户可以自行从[方正字体官网](#)下载此四款字体, 在下载的时候请**务必**注意选择 GBK 字符集, 也可以使用 [ETEX 工作室](#)提供的[方正字体](#), 提取码为: `njy9` 进行安装。安装时, Win 10 用户请右键选择为全部用户安装, 否则会找不到字体。

字体名称	编码	单价	实付价	交易状态	操作
订单号: C20200204164821OW1F					2020-02-04 16:48:21
方正仿宋_GBK • 免费商用	简繁扩展(GBK)	¥ 0.00			
方正黑体_GBK • 免费商用	简繁扩展(GBK)	¥ 0.00		免费 已完成	下载字体
方正书宋_GBK • 免费商用	简繁扩展(GBK)	¥ 0.00			
方正楷体_GBK • 免费商用	简繁扩展(GBK)	¥ 0.00			

12.3.2 其他中文字体

如果你想完全自定义字体², 你可以选择 `chinesefont=nofont`, 然后在导言区设置

```
\setCJKmainfont [BoldFont={FZHei-B01}, ItalicFont={FZKai-Z03}]{FZShuSong-Z01}
\setCJKsansfont [BoldFont={FZHei-B01}, ItalicFont={FZHei-B01}]{FZHei-B01}
\setCJKmonofont [BoldFont={FZHei-B01}, ItalicFont={FZHei-B01}]{FZFangSong-Z02}
```

¹可以使用命令提示符, 输入 `texdoc ctex` 调出本地 `ctex` 宏包文档

²这里仍然以方正字体为例。

```
\setCJKfamilyfont{zhsong}{FZShuSong-Z01}
\setCJKfamilyfont{zhhei}{FZHei-B01}
\setCJKfamilyfont{zhkai}{FZKai-Z03}
\setCJKfamilyfont{zhfs}{FZFangSong-Z02}
\newcommand*\songti{\CJKfamily{zhsong}}
\newcommand*\heiti{\CJKfamily{zhhei}}
\newcommand*\kaishu{\CJKfamily{zhkai}}
\newcommand*\fangsong{\CJKfamily{zhfs}}
```

第 13 章 ElegantBook 写作示例

内容提要

- 定义 ??
- 柯西列性质 E.1.1
- Fubini 定理 E.1.1
- 韦达定理
- 最优化原理 E.1.1

13.1 Lebesgue 积分

在前面各章做了必要的准备后，本章开始介绍新的积分。在 Lebesgue 测度理论的基础上建立了 Lebesgue 积分，其被积函数和积分域更一般，可以对有界函数和无界函数统一处理。正是由于 Lebesgue 积分的这些特点，使得 Lebesgue 积分比 Riemann 积分具有在更一般条件下的极限定理和累次积分交换积分顺序的定理，这使得 Lebesgue 积分不仅在理论上更完善，而且在计算上更灵活有效。

Lebesgue 积分有几种不同的定义方式。我们将采用逐步定义非负简单函数，非负可测函数和一般可测函数积分的方式。

由于现代数学的许多分支如概率论、泛函分析、调和分析等常常用到一般空间上的测度与积分理论，在本章最后一节将介绍一般的测度空间上的积分。

13.1.1 积分的定义

我们将通过三个步骤定义可测函数的积分。首先定义非负简单函数的积分。以下设 E 是 \mathcal{R}^n 中的可测集。

定义 13.1 (可积性)

设 $f(x) = \sum_{i=1}^k a_i \chi_{A_i}(x)$ 是 E 上的非负简单函数，其中 $\{A_1, A_2, \dots, A_k\}$ 是 E 上的一个可测分割， a_1, a_2, \dots, a_k 是非负实数。定义 f 在 E 上的积分为 $\int_a^b f(x) dx$

$$\int_E f dx = \sum_{i=1}^k a_i m(A_i) \pi \alpha \beta \sigma \gamma \nu \xi \epsilon \varepsilon. \oint_a^b \oint_a^b \prod_{i=1}^n$$
 (13.1)

一般情况下 $0 \leq \int_E f dx \leq \infty$ 。若 $\int_E f dx < \infty$ ，则称 f 在 E 上可积。



一个自然的问题是，Lebesgue 积分与我们所熟悉的 Riemann 积分有什么联系和区别？在 4.4 在我们将详细讨论 Riemann 积分与 Lebesgue 积分的关系。这里只看一个简单的例子。设 $D(x)$ 是区间 $[0, 1]$ 上的 Dirichlet 函数。即 $D(x) = \chi_{Q_0}(x)$ ，其中 Q_0 表示 $[0, 1]$ 中的有理数的全体。根据非负简单函数积分的定义， $D(x)$ 在 $[0, 1]$ 上的 Lebesgue 积分为

$$\int_0^1 D(x) dx = \int_0^1 \chi_{Q_0}(x) dx = m(Q_0) = 0$$
 (13.2)

即 $D(x)$ 在 $[0, 1]$ 上是 Lebesgue 可积的并且积分值为零。但 $D(x)$ 在 $[0, 1]$ 上不是 Riemann 可积的。

有界变差函数是与单调函数有密切联系的一类函数。有界变差函数可以表示为两个单调递增函数之差。与单调函数一样，有界变差函数几乎处处可导。与单调函数不同，有界变差函数类对线性运算是封闭的，它们构成一线空间。练习题 E.1 是一个性质的证明。

练习 13.1 设 $f \notin L(\mathcal{R}^1)$ ， g 是 \mathcal{R}^1 上的有界可测函数。证明函数

$$I(t) = \int_{\mathcal{R}^1} f(x+t) g(x) dx \quad t \in \mathcal{R}^1$$
 (13.3)

是 \mathcal{R}^1 上的连续函数。

解 即 $D(x)$ 在 $[0, 1]$ 上是 Lebesgue 可积的并且积分值为零。但 $D(x)$ 在 $[0, 1]$ 上不是 Riemann 可积的。

证明 即 $D(x)$ 在 $[0, 1]$ 上是 Lebesgue 可积的并且积分值为零。但 $D(x)$ 在 $[0, 1]$ 上不是 Riemann 可积的。

定理 13.1 (Fubini 定理)

(1) 若 $f(x, y)$ 是 $\mathcal{R}^p \times \mathcal{R}^q$ 上的非负可测函数，则对几乎处处的 $x \in \mathcal{R}^p$, $f(x, y)$ 作为 y 的函数是 \mathcal{R}^q 上的非负可测函数， $g(x) = \int_{\mathcal{R}^q} f(x, y) dy$ 是 \mathcal{R}^p 上的非负可测函数。并且

$$\int_{\mathcal{R}^p \times \mathcal{R}^q} f(x, y) dx dy = \int_{\mathcal{R}^p} \left(\int_{\mathcal{R}^q} f(x, y) dy \right) dx. \quad (13.4)$$

(2) 若 $f(x, y)$ 是 $\mathcal{R}^p \times \mathcal{R}^q$ 上的可积函数，则对几乎处处的 $x \in \mathcal{R}^p$, $f(x, y)$ 作为 y 的函数是 \mathcal{R}^q 上的可积函数，并且 $g(x) = \int_{\mathcal{R}^q} f(x, y) dy$ 是 \mathcal{R}^p 上的可积函数。而且 E.4 成立。



E.1.1

笔记 在本模板中，引理 (lemma)，推论 (corollary) 的样式和定理 E.1.1 的样式一致，包括颜色，仅仅只有计数器的设置不一样。

我们说一个实变或者复变量的实值或者复值函数是在区间上平方可积的，如果其绝对值的平方在该区间上的积分是有限的。所有在勒贝格积分意义下平方可积的可测函数构成一个希尔伯特空间，也就是所谓的 L^2 空间，几乎处处相等的函数归为同一等价类。形式上， L^2 是平方可积函数的空间和几乎处处为 0 的函数空间的商空间。

命题 13.1 (最优化原理)

如果 u^* 在 $[s, T]$ 上为最优解，则 u^* 在 $[s, T]$ 任意子区间都是最优解，假设区间为 $[t_0, t_1]$ 的最优解为 u^* ，则 $u(t_0) = u^*(t_0)$ ，即初始条件必须还是在 u^* 上。



我们知道最小二乘法可以用来处理一组数据，可以从一组测定的数据中寻求变量之间的依赖关系，这种函数关系称为经验公式。本课题将介绍最小二乘法的精确定义及如何寻求点与点之间近似成线性关系时的经验公式。假定实验测得变量之间的 n 个数据，则在平面上，可以得到 n 个点，这种图形称为“散点图”，从图中可以粗略看出这些点大致散落在某直线近旁，我们认为其近似为一线性函数，下面介绍求解步骤。

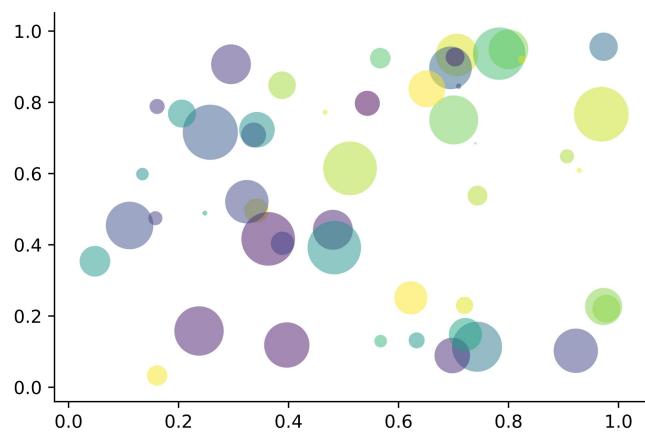


图 13.1: 散点图示例 $\hat{y} = a + bx$

以最简单的一元线性模型来解释最小二乘法。什么是一元线性模型呢？监督学习中，如果预测的变量是离散的，我们称其为分类（如决策树，支持向量机等），如果预测的变量是连续的，我们称其为回归。回归分析中，如果只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线

性回归分析。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。对于二维空间线性是一条直线；对于三维空间线性是一个平面，对于多维空间线性是一个超平面。

性质 柯西列的性质

1. $\{x_k\}$ 是柯西列，则其子列 $\{x_k^i\}$ 也是柯西列。
2. $x_k \in \mathcal{R}^n$, $\rho(x, y)$ 是欧几里得空间，则柯西列收敛， (\mathcal{R}^n, ρ) 空间是完备的。

结论 回归分析 (regression analysis) 是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。运用十分广泛，回归分析按照涉及的变量的多少，分为一元回归和多元回归分析；按照因变量的多少，可分为简单回归分析和多重回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。

第 13 章 练习

1. 设 A 为数域 K 上的 n 级矩阵。证明：如果 K^n 中任意非零列向量都是 A 的特征向量，则 A 一定是数量矩阵。
2. 证明：不为零矩阵的幂零矩阵不能对角化。
3. 设 $A = (a_{ij})$ 是数域 K 上的一个 n 级上三角矩阵，证明：如果 $a_{11} = a_{22} = \dots = a_{nn}$ ，并且至少有一个 $a_{kl} \neq 0 (k < l)$ ，则 A 一定不能对角化。

第 14 章 常见问题集

我们根据用户社区反馈整理了下面一些常见的问题，用户在遇到问题时，应当首先查阅本手册和本部分的常见问题。

1. 有没有办法章节用“第一章，第一节，（一）”这种？

见前文介绍，可以使用 `scheme=chinese` 设置。

2. 大佬，我想把正文字体改为亮色，背景色改为黑灰色。

页面颜色可以使用 `\pagecolor` 命令设置，文本命令可以参考[这里](#)进行设置。

3. Package `ctex` Error: CTeX fontset ‘Mac’ is unavailable.

在 Mac 系统下，中文编译请使用 Xe^LAT_EX。

4. ! LaTeX Error: Unknown option ‘`scheme=plain`’ for package ‘`ctex`’.

你用的 CTeX 套装吧？这个里面的 `ctex` 宏包已经是 10 年前的了，与本模板使用的 `ctex` 宏集有很大区别。不建议 CTeX 套装了，请卸载并安装 TeX Live 2021。

5. 我该使用什么版本？

请务必使用[最新正式发行版](#)，发行版间不定期可能会有更新（修复 bug 或者改进之类），如果你在使用过程中没有遇到问题，不需要每次更新[最新版](#)，但是在发行版更新之后，请尽可能使用最新版（发行版）！最新发行版可以在 GitHub 或者 TeX Live 2021 内获取。

6. 我该使用什么编辑器？

你可以使用 TeX Live 2021 自带的编辑器 TeXworks 或者使用 TeXstudio，TeXworks 的自动补全，你可以参考我们的总结 [TeXworks 自动补全](#)。推荐使用 TeX Live 2021 + TeXstudio。我自己用 VS Code 和 Sublime Text，相关的配置说明，请参考 [LaTeX 编译环境配置：Visual Studio Code 配置简介](#) 和 [Sublime Text 搭建 LaTeX 编写环境](#)。

7. 您好，我们想用您的 ElegantBook 模板写一本书。关于机器学习的教材，希望获得您的授权，谢谢您的宝贵时间。

模板的使用修改都是自由的，你们声明模板来源以及模板地址（GitHub 地址）即可，其他未尽事宜按照开源协议 LPPL-1.3c。做好之后，如果方便的话，可以给我们一个链接，我把你们的教材放在 ElegantLaTeX 用户作品集里。

8. 请问交叉引用是什么？

本群和本模板适合有一定 LaTeX 基础的用户使用，新手请先学习 LaTeX 的基础，理解各种概念，否则你将寸步难行。

9. 定义等环境中无法使用加粗命令么？

是这样的，默认中文并没有加粗命令，如果你想在定义等环境中使用加粗命令，请使用 `\heiti` 等字体命令，而不要使用 `\textbf`。或者，你可以将 `\textbf` 重新定义为 `\heiti`。英文模式不存在这个问题。

10. 代码高亮环境能用其他语言吗？

可以的，ElegantBook 模板用的是 `listings` 宏包，你可以在环境（`lstlisting`）之后加上语言（比如 Python 使用 `language=Python` 选项），全局语言修改请使用 `lset` 命令，更多信息请参考宏包文档。

11. 群主，什么时候出 Beamer 的模板（主题），ElegantSlide 或者 ElegantBeamer？

由于 Beamer 中有一个很优秀的主题 Metropolis。后续确定不会再出任何主题/模板，请大家根据需要修改已有主题。

第 15 章 版本更新历史

根据用户的反馈，我们不断修正和完善模板。截止到此次更新，ElegantBook 模板在 GitHub 上有将近 100 次提交，正式发行版本（release）有 17 次。由于 3.00 之前版本与现在版本差异非常大，在此不列出 3.00 之前的更新内容。

2021/05/02 更新：版本 4.1 正式发布。

- ① **重要改动：**由原先的 `BIBTEX` 改为 `biblatex` 编译方式（后端为 `biber`），请注意两者之间的差异；
- ② **重要改进：**修改对于定理写法兼容方式，提高数学公式代码的兼容性；
- ③ 页面设置改动，默认页面更宽；方便书写和阅读；
- ④ 支持目录文字以及页码跳转；
- ⑤ 不再维护 `pdflATEX` 中文支持方式，请务必使用 `XeLATEX` 编译中文文稿。
- ⑥ 增加多个语言选项，法语 `lang=fr`、荷兰语 `lang=nl`、匈牙利语 `lang=hu`、西班牙语 `lang=es`、蒙古语 `lang=mn` 等。

2020/04/12 更新：版本 3.11 正式发布，**此版本为 3.x 最后版本。**

- ① **重要修正：**修复因为 `gbt7714` 宏包更新导致的 `natbib` option clash 错误；
- ② 由于 `pgfornament` 宏包未被 TeX Live 2020 收录，因此删除 base 相关的内容；
- ③ 修复部分环境的空格问题；
- ④ 增加了意大利语言选项 `lang=it`。

2020/02/10 更新：版本 3.10 正式发布

- ① 增加数学字体选项 `math`，可选项为 `newtx` 和 `cm`。
重要提示：原先通过 `newtxmath` 宏包设置的数学字体改为 LATEX 默认数学字体，如果需要保持原来的字体，需要显式声明数学字体 (`math=newtx`)；
- ② 新增中文字体选项 `chinesefont`，可选项为 `ctexfont`、`founder` 和 `nofont`。
- ③ 将封面作者信息设置为可选，并且增加自定义信息命令 `\bioinfo`；
- ④ 在说明文档中增加版本历史，新增 `\datechange` 命令和 `change` 环境；
- ⑤ 增加汉化章节选项 `scheme`，可选项为汉化 `chinese`；
- ⑥ 由于 `\lvert` 问题已经修复，重新调整 `ctex` 宏包和 `amsmath` 宏包位置。
- ⑦ 修改页眉设置，去除了 `\lastpage` 以避免 page anchor 问题，加入 `\frontmatter`。
- ⑧ 修改参考文献选项 `cite`，可选项为数字 `numbers`、作者-年份 `authoryear` 以及上标 `super`。
- ⑨ 新增参考文献样式选项 `bibstyle`，并将英文模式下参考文献样式 `apalike` 设置为默认值，中文仍然使用 `gbt7714` 宏包设置。

2019/08/18 更新：版本 3.09 正式发布

- ① `\elegantpar` 存在 bug，删除 `\elegantpar` 命令，建议用户改用 `\marginnote` 和 `\marginpar` 旁注命令。
- ② 积分操作符统一更改为 `esint` 宏包设置；
- ③ 新增目录选项 `toc`，可选项为单栏 `onecol` 和双栏 `twocol`；
- ④ 手动增加参考文献选项 `cite`，可选项为上标形式 `super`；
- ⑤ 修正章节习题（`problemset`）环境。

2019/05/28 更新：版本 3.08 正式发布

- ① 修复 `\part` 命令。

-
- ② 引入 Note 模板中的 `pad` 选项 `device=pad`。
 - ③ 数学字体加入 `mtpro2` 可选项 `math=mtpro2`, 使用免费的 `lite` 子集。
 - ④ 将参考文献默认显示方式 `authoyear` 改为 `numbers`。
 - ⑤ 引入旁注命令 `\marginpar` (测试)。
 - ⑥ 新增章节摘要环境 `introduction`。
 - ⑦ 新增章节习题环境 `problemset`。
 - ⑧ 将 `\equation` 重命名为 `\extrainfo`。
 - ⑨ 完善说明文档, 增加致谢部分。

2019/04/15 更新: 版本 3.07 正式发布

- ① 删除中英文自定义字体总设置。
- ② 新增颜色主题, 并将原绿色默认主题设置为蓝色 `color=blue`。
- ③ 引入隐藏装饰图案选项 `base`, 可选项有显示 `show` 和隐藏 `hide`。
- ④ 新增定理模式 `mode`, 可选项有简单模式 `simple` 和炫彩模式 `fancy`。
- ⑤ 新增隐藏证明、答案等环境的选项 `result=noanswer`。

2019/02/25 更新: 版本 3.06 正式发布

- ① 删除水印。
- ② 新封面, 新装饰图案。
- ③ 添加引言使用说明。
- ④ 修复双面 `twoside`。
- ⑤ 美化列表环境。
- ⑥ 增加 `\subsubsection` 的设置。
- ⑦ 将模板拆分成中英文语言模式。
- ⑧ 使用 `lstlisting` 添加代码高亮。
- ⑨ 增加定理类环境使用说明。

2019/01/22 更新: 版本 3.05 正式发布

- ① 添加 `xeCJK` 宏包中文支持方案。
- ② 修复模板之前对 `TikZ` 单位的改动。
- ③ 更新 logo 图。

2019/01/15 更新: 版本 3.04 正式发布

- ① 格式化模板代码。
- ② 增加 `\equation` 命令。
- ③ 修改 `\date`。

2019/01/08 更新: 版本 3.03 正式发布

- ① 修复附录章节显示问题。
- ② 小幅优化封面代码。

2018/12/31 更新: 版本 3.02 正式发布

- ① 修复名字系列命令自定义格式时出现的空格问题, 比如 `\listfigurename`。
- ② 英文定理类名字改为中文名。
- ③ 英文结构名改为中文。

2018/12/16 更新：版本 3.01 正式发布

- ① 调整 `ctex` 宏包。
 - ② 说明文档增加更新内容。
-

2018/12/06 更新：版本 3.00 正式发布

- ① 删除 `mathpazo` 数学字体选项。
- ② 添加邮箱命令 `\mailto`。
- ③ 修改英文字体为 `newtx` 系列，另外大型操作符号维持 `cm` 字体。
- ④ 中文字体改用 `ctex` 宏包自动设置。
- ⑤ 删除 `xeCJK` 字体设置，原因是不同系统字体不方便统一。
- ⑥ 定理换用 `tcolorbox` 宏包定义，并基本维持原有的定理样式，优化显示效果，支持跨页；定理类名字重命名，如 `etheorem` 改为 `theorem` 等等。
- ⑦ 删去自定义的缩进命令 `\Indent`。
- ⑧ 添加参考文献宏包 `natbib`。
- ⑨ 颜色名字重命名。

附录 A 基本数学工具

本附录包括了计量经济学中用到的一些基本数学，我们扼要论述了求和算子的各种性质，研究了线性和某些非线性方程的性质，并复习了比例和百分数。我们还介绍了一些在应用计量经济学中常见的特殊函数，包括二次函数和自然对数，前 4 节只要求基本的代数技巧，第 5 节则对微分学进行了简要回顾；虽然要理解本书的大部分内容，微积分并非必需，但在一些章末附录和第 3 篇某些高深专题中，我们还是用到了微积分。

A.1 求和算子与描述统计量

求和算子 是用以表达多个数求和运算的一个缩略符号，它在统计学和计量经济学分析中扮演着重要作用。如果 $\{x_i : i = 1, 2, \dots, n\}$ 表示 n 个数的一个序列，那么我们就把这 n 个数的和写为：

$$\sum_{i=1}^n x_i \equiv x_1 + x_2 + \dots + x_n \quad (\text{A.1})$$

附录 B ElegantL^AT_EX 系列模板介绍

ElegantL^AT_EX 项目组致力于打造一系列美观、优雅、简便的模板方便用户使用。目前由 **ElegantNote**, **ElegantBook**, **ElegantPaper** 组成，分别用于排版笔记，书籍和工作论文。强烈推荐使用最新正式版本！本文将介绍本模板的一些设置内容以及基本使用方法。如果您有其他问题，建议或者意见，欢迎在 GitHub 上给我们提交 issues 或者邮件联系我们。

我们的联系方式如下，建议加入用户 QQ 群提问，这样能更快获得准确的反馈，加群时请备注 L^AT_EX 或者 ElegantL^AT_EX 相关内容。

- 官网: <https://elegantlatex.org/>
- GitHub 网址: <https://github.com/ElegantLaTeX/>
- CTAN 地址: <https://ctan.org/pkg/elegantbook>
- 下载地址: 正式发行版, 最新版
- 微博: ElegantL^AT_EX
- 微信公众号: ElegantL^AT_EX
- 用户 QQ 群: 692108391
- 邮件: elegantlatex2e@gmail.com

B.1 ElegantBook 更新说明

此次为 4.x 第一个版本，在 3.x 基础上，主要更新了定理以及参考文献的支持方式，具体内容有：

1. **重要改动:** 由原先的 Bib^LT_EX 改为 biblatex 编译方式（后端为 biber），请注意两者之间的差异；
2. **重要改进:** 修改对于定理写法兼容方式，提高数学公式代码的兼容性；
3. 页面设置改动，默认页面更宽；方便书写和阅读；
4. 支持目录文字以及页码跳转；
5. 不再维护 pdfL^AT_EX 中文支持方式，请务必使用 X_EL^AT_EX 编译中文文稿。
6. 增加多语言选项，法语 `lang=fr`、德语 `lang=de`、荷兰语 `lang=nl`、匈牙利语 `lang=hu`、西班牙语 `lang=es`、蒙古语 `lang=mn` 等。

 **笔记** 如果你使用旧版本切换到新版本时，遇到问题时，请核对文档中是否有 `pageanchor` 字样。如果有，请删除文档中的 `\hypersetup{pageanchor=true}`，并且在 `\maketitle` 和 `\tableofcontents` 之间添加 `\frontmatter`。
2.x 版本的用户请仔细查看跨版本转换。

B.2 模板安装与更新

你可以通过免安装的方式使用本模板，包括在线使用和本地（文件夹内）使用两种方式，也可以通过 T_EX 发行版安装使用。

B.2.1 在线使用模板

我们把三套模板全部上传到 Overleaf 上了，网络便利的用户可以直接通过 Overleaf 在线使用我们的模板。使用 Overleaf 的好处是无需安装 T_EX Live 2020，可以随时随地访问自己的文件。查找模板，请在 Overleaf 模板库里面搜索 `elegantlatex` 即可，你也可以直接访问 [搜索结果](#)。选择适当的模板之后，将其 `Open as Template`，即可把模板存到自己账户下，然后可以自由编辑以及与别人一起协作。更多关于 Overleaf 的介绍和使用，请参考 Overleaf 的[官方文档](#)。

注 Overleaf 上，中文需要使用 X_EL^AT_EX 进行编译，英文建议使用 pdfL^AT_EX 编译。

B.2.2 本地免安装使用

免安装使用方法如下，从 GitHub 或者 CTAN 下载最新版，严格意义上只需要类文件 `elegantbook.cls`。然后将模板文件放在你的工作目录下即可使用。这样使用的好处是，无需安装，简便；缺点是，当模板更新之后，你需要手动替换 `cls` 文件。

B.2.3 发行版安装使用

本模板测试环境为 Win10 和 TeX Live 2021，如果你刚安装 TeX Live 2021 用户，安装后建议升级全部宏包，升级方法：使用 cmd 运行 `tlmgr update --all`，如果 `tlmgr` 需要更新，请使用 cmd 运行 `tlmgr update --self`，如果更新过程中出现了中断，请改用 `tlmgr update --self --all --reinstall-forcibly-removed` 更新。

B.2.4 更新问题

如果使用 `tlshell` 无法更新模板，请使用命令行全部更新全部宏包或者使用免安装的方法使用本模板。

通过命令行（管理员权限）输入下面的命令对 `tlmgr` 自身和全部宏包进行更新。

```
tlmgr update --self
tlmgr update --all
```

更多的内容请参考 [How do I update my TeX distribution?](#)

B.2.5 其他发行版本

由于宏包版本问题，本模板不支持 CTeX 套装，请务必安装 TeX Live。更多关于 TeX Live 的安装使用以及 CTeX 与 TeX Live 的兼容、系统路径问题，请参考官方文档以及 [一份简短的关于安装 L^AT_EX 安装的介绍](#)。

B.3 关于提交

出于某些因素的考虑，ElegantL^AT_EX 项目自 2019 年 5 月 20 日开始，不再接受任何非作者预约性质的提交（pull request）！如果你想改进模板，你可以给我们提交 issues，或者可以在遵循协议（LPPL-1.3c）的情况下，克隆到自己仓库下进行修改。

附录 C ElegantBook 设置说明

本模板基于基础的 book 文类，所以 book 的选项对于本模板也是有效的（纸张无效，因为模板有设备选项）。默认编码为 UTF-8，推荐使用 TeX Live 编译。本文编写环境为 Win10 (64bit) + TeX Live 2021，英文支持 pdfLaTeX，中文仅支持 XeLaTeX 编译。

C.1 语言模式

本模板内含两套基础语言环境 `lang=cn`、`lang=en`。改变语言环境会改变图表标题的引导词（图，表），文章结构词（比如目录，参考文献等），以及定理环境中的引导词（比如定理，引理等）。不同语言模式的启用如下：

```
\documentclass[en]{elegantbook}
\documentclass[lang=cn]{elegantbook}
```

除模板自带的两套语言设定之外，由网友提供的其他语言环境设置如下：

- 由 VincentMVV 提供的意大利语翻译 `lang=it`，相关讨论见 [Italian translation](#)；
- 由 abfek66 提供的法语翻译 `lang=fr`，相关讨论见 [Italian translation](#)；
- 由 inktvis75 提供的荷兰语翻译 `lang=nl`，相关讨论见 [Dutch Translation](#)；
- 由 palkotamas 提供的匈牙利语翻译 `lang=hu`，相关讨论见 [Hungarian translation](#)；
- 由 Lisa 提供的德语翻译 `lang=de`，相关讨论见 [Deutsch translation](#)；
- 由 Gustavo A. Corradi 提供的西班牙语的翻译 `lang=es`，相关讨论见 [Spanish translation](#)；
- 由 Altantssooj 提供的蒙古语的翻译 `lang=mn`，相关讨论见 [Mongolian translation](#)。

注 以上各个语言的设定均为网友设定，我们未对上述翻译进行过校对，如果有问题，请在对应的 issue 下评论。并且，只有中文环境 (`lang=cn`) 才可以输入中文。

C.2 设备选项

最早我们在 ElegantNote 模板中加入了设备选项 (`device`)，后来，我们认为这个设备选项的设置可以应用到 ElegantBook 中¹，而且 Book 一般内容比较多，如果在 iPad 上看无需切边，放大，那用户的阅读体验将会得到巨大提升。你可以使用下面的选项将版面设置为 iPad 设备模式²

```
\documentclass[pad]{elegantbook} %or
\documentclass[device=pad]{elegantbook}
```

C.3 颜色主题

本模板内置 5 组颜色主题，分别为 `green`³、`cyan`、`blue`（默认）、`gray`、`black`。另外还有一个自定义的选项 `nocolor`。调用颜色主题 `green` 的方法为

```
\documentclass[green]{elegantbook} %or
\documentclass[color=green]{elegantbook}
```

¹不过因为 ElegantBook 模板封面图片的存在，在修改页面设计时，需要对图片进行裁剪。

²默认为 normal 模式，也即 A4 纸张大小。

³为原先默认主题。

表 C.1: ElegantBook 模板中的颜色主题

	green	cyan	blue	gray	black	主要使用的环境
structure						chapter section subsection
main						definition exercise problem
second						theorem lemma corollary
third						proposition

如果需要自定义颜色的话请选择 `nocolor` 选项或者使用 `color=none`, 然后在导言区定义 `structurecolor`、`main`、`second`、`third` 颜色, 具体方法如下:

```
\definecolor{structurecolor}{RGB}{0,0,0}
\definecolor{main}{RGB}{70,70,70}
\definecolor{second}{RGB}{115,45,2}
\definecolor{third}{RGB}{0,80,80}
```

C.4 封面

C.4.1 封面个性化

从 3.10 版本开始, 封面更加弹性化, 用户可以自行选择输出的内容, 包括 `\title` 在内的所有封面元素都可为空。目前封面的元素有

表 C.2: 封面元素信息

信息	命令	信息	命令	信息	命令
标题	<code>\title</code>	副标题	<code>\subtitle</code>	作者	<code>\author</code>
机构	<code>\institute</code>	日期	<code>\date</code>	版本	<code>\version</code>
箴言	<code>\extrainfo</code>	封面图	<code>\cover</code>	徽标	<code>\logo</code>

另外, 额外增加一个 `\bioinfo` 命令, 有两个选项, 分别是信息标题以及信息内容。比如需要显示User Name: 111520, 则可以使用

```
\bioinfo{User Name}{111520}
```

封面中间位置的色块的颜色可以使用下面命令进行修改:

```
\definecolor{customcolor}{RGB}{32,178,170}
\colorlet{coverlinecolor}{customcolor}
```

C.4.2 封面图

本模板使用的封面图片来源于 pixabay.com⁴, 图片完全免费, 可用于任何场景。封面图片的尺寸为 1280 × 1024, 更换图片的时候请严格按照封面图片尺寸进行裁剪。推荐一个免费的在线图片裁剪网站 fotor.com。用户

⁴感谢 ChinaTeX 提供免费图源网站, 另外还推荐 pexels.com。

QQ 群内有一些合适尺寸的封面，欢迎取用。

C.4.3 徽标

本文用到的 Logo 比例为 1:1，也即正方形图片，在更换图片的时候请选择合适的图片进行替换。

C.4.4 自定义封面

另外，如果使用自定义的封面，比如 Adobe illustrator 或者其他软件制作的 A4 PDF 文档，请把 `\maketitle` 注释掉，然后借助 `pdfpages` 宏包将自制封面插入即可。如果使用 `titlepage` 环境，也是类似。如果需要 2.x 版本的封面，请参考 `etitlepage`。

C.5 章标题

本模板内置 2 套章标题显示风格，包含 `hang`（默认）与 `display` 两种风格，区别在于章标题单行显示（`hang`）与双行显示（`display`），本说明使用了 `hang`。调用方式为

```
\documentclass[hang]{elegantbook} %or
\documentclass[titlestyle=hang]{elegantbook}
```

在章标题内，章节编号默认是以数字显示，也即第 1 章，第 2 章等等，如果想要把数字改为中文，可以使用

```
\documentclass[chinese]{elegantbook} %or
\documentclass[scheme=chinese]{elegantbook}
```

C.6 数学环境简介

在我们这个模板中，我们定义了两种不同的定理模式 `mode`，包括简单模式（`simple`）和炫彩模式（`fancy`），默认为 `fancy` 模式，不同模式的选择为

```
\documentclass[simple]{elegantbook} %or
\documentclass[mode=simple]{elegantbook}
```

本模板定义了四大类环境

- 定理类环境，包含标题和内容两部分，全部定理类环境的编号均以章节编号。根据格式的不同分为 3 种
 - `definition` 环境，颜色为 `main`；
 - `theorem`、`lemma`、`corollary` 环境，颜色为 `second`；
 - `proposition` 环境，颜色为 `third`。
- 示例类环境，有 `example`、`problem`、`exercise` 环境（对应于例、例题、练习），自动编号，编号以章节为单位，其中 `exercise` 有提示符。
- 提示类环境，有 `note` 环境，特点是：无编号，有引导符。
- 结论类环境，有 `conclusion`、`assumption`、`property`、`remark`、`solution` 环境⁵，三者均以粗体的引导词为开头，和普通段落格式一致。

⁵本模板还添加了一个 `result` 选项，用于隐藏 `solution` 和 `proof` 环境，默认为显示（`result=answer`），隐藏使用 `result=noanswer`。

C.6.1 定理类环境的使用

由于本模板使用了 `tcolorbox` 宏包来定制定理类环境，所以和普通的定理环境的使用有些许区别，定理的使用方法如下：

```
\begin{theorem}{theorem name}{label}
The content of theorem.
\end{theorem}
```

第一个必选项 `theorem name` 是定理的名字，第二个必选项 `label` 是交叉引用时所用到的标签，交叉引用的方法为 `\ref{thm:label}`。请注意，交叉引用时必须加上前缀 `thm:`。

在用户多次反馈下，4.x 之后，引入了原生定理的支持方式，也就是使用可选项方式：

```
\begin{theorem}[theorem name] \label{thm:theorem-label}
The content of theorem.
\end{theorem}
% or
\begin{theorem} \label{thm:theorem-without-name}
The content of theorem without name.
\end{theorem}
```

其他相同用法的定理类环境有：

表 C.3: 定理类环境

环境名	标签名	前缀	交叉引用
definition	label	def	<code>\ref{def:label}</code>
theorem	label	thm	<code>\ref{thm:label}</code>
lemma	label	lem	<code>\ref{lem:label}</code>
corollary	label	cor	<code>\ref{cor:label}</code>
proposition	label	pro	<code>\ref{pro:label}</code>

C.6.2 其他环境的使用

其他三种环境没有选项，可以直接使用，比如 `example` 环境的使用方法与效果：

```
\begin{example}
This is the content of example environment.
\end{example}
```

这几个都是同一类环境，区别在于

- 示例环境（`example`）、练习（`exercise`）与例题（`problem`）章节自动编号；
- 注意（`note`），练习（`exercise`）环境有提醒引导符；
- 结论（`conclusion`）等环境都是普通段落环境，引导词加粗。

C.7 列表环境

本模板借助于 `tikz` 定制了 `itemize` 和 `enumerate` 环境，其中 `itemize` 环境修改了 3 层嵌套，而 `enumerate` 环境修改了 4 层嵌套（仅改变颜色）。示例如下

- first item of nesti;
 - second item of nesti;
 - first item of nestii;
 - second item of nestii;
 - first item of nestiii;
 - second item of nestiii.
1. first item of nesti;
 2. second item of nesti;
 - (a). first item of nestii;
 - (b). second item of nestii;
 - I. first item of nestiii;
 - II. second item of nestiii.

C.8 参考文献

此模板使用了 biber 来生成参考文献，也即使用 biblatex 宏包，在中文示例中，使用了 gbt7714 宏包。参考文献示例：[cn1, en2, en3] 使用了中国一个大型的 P2P 平台（人人贷）的数据来检验男性投资者和女性投资者在投资表现上是否有显著差异。

你可以在谷歌学术，Mendeley，Endnote 中获得文献条目（bib item），然后把它们添加到 reference.bib 中。在文中引用的时候，引用它们的键值（bib key）即可。注意需要在编译的过程中添加 biber 编译。

为了方便文献样式修改，模板引入了 `bibstyle` 和 `citestyle` 选项，默认均为数字格式（numeric），如果需要设置为国标 GB7714-2015，需要使用：

```
\documentclass[citestyle=gb7714-2015, bibstyle=gb7714-2015]{elegantbook}
```

如果需要添加排序方式，可以在导言区加入

```
\ExecuteBibliographyOptions{sorting=ynt}
```

启用国标之后，可以加入 `sorting=gb7714-2015`。

C.9 添加序章

如果你想在第一章前面添序章，不改变原本章节序号，可以在第一章内容前面使用

```
\chapter*{Introduction}
\markboth{Introduction}{Introduction}
The content of introduction.
```

C.10 目录选项与深度

本模板添加了一个目录选项 `toc`，可以设置目录为单栏（`onecol`）和双栏（`twocol`）显示，比如双栏显示可以使用

```
\documentclass[twocol]{elegantbook}
\documentclass[toc=twocol]{elegantbook}
```

默认本模板目录深度为 1，你可以在导言区使用

```
\setcounter{tocdepth}{2}
```

将其修改为 2 级目录（章与节）显示。

C.11 章节摘要

模板新增了一个章节摘要环境（introduction），使用示例

```
\begin{introduction}
    \item Definition of Theorem
    \item Ask for help
    \item Optimization Problem
    \item Property of Cauchy Series
    \item Angle of Corner
\end{introduction}
```

效果如下：

内容提要

- | | |
|--|--|
| <input type="checkbox"/> Definition of Theorem
<input type="checkbox"/> Ask for help
<input type="checkbox"/> Optimization Problem | <input type="checkbox"/> Property of Cauchy Series
<input type="checkbox"/> Angle of Corner |
|--|--|

环境的标题文字可以通过这个环境的可选参数进行修改，修改方法为：

```
\begin{introduction}[Brief Introduction]
    ...
\end{introduction}
```

C.12 章后习题

前面我们介绍了例题和练习两个环境，这里我们再加一个，章后习题（problemset）环境，用于在每一章结尾，显示本章的练习。使用方法如下

```
\begin{problemset}
    \item exercise 1
    \item exercise 2
    \item exercise 3
\end{problemset}
```

效果如下：

第 C 章 练习

1. exercise 1
2. exercise 2
3. exercise 3
4. 测试数学公式

$$a^2 + b^2 = c_{2_i}(1, 2)[1, 23] \quad (\text{C.1})$$

注 如果你想把 problemset 环境的标题改为其他文字，你可以类似于 introduction 环境修改 problemset 的可选参数。另外，目前这个环境会自动出现在目录中，但是不会出现在页眉页脚信息中（待解决）。

解 如果你想把 problemset 环境的标题改为其他文字，你可以类似于 introduction 环境修改 problemset 的可选参数。另外，目前这个环境会自动出现在目录中，但是不会出现在页眉页脚信息中（待解决）。

C.13 旁注

在 3.08 版本中，我们引入了旁注设置选项 `marginpar=margintrue` 以及测试命令 `\elegantpar`，但是由此带来一堆问题。我们决定在 3.09 版本中将其删除，并且，在旁注命令得到大幅度优化之前，不会将此命令再次引入书籍模板中。对此造成各位用户的不方便，非常抱歉！不过我们保留了 `marginpar` 这个选项，你可以使用 `marginpar=margintrue` 获得保留右侧旁注的版面设计。然后使用系统自带的 `\marginpar` 或者 `marginnote` 宏包的 `\marginnote` 命令。

注 在使用旁注的时候，需要注意的是，文本和公式可以直接在旁注中使用。

```
% text
\marginpar{margin paragraph text}

% equation
\marginpar{
\begin{equation}
a^2 + b^2 = c^2
\end{equation}
}
```

但是浮动体（表格、图片）需要注意，不能用浮动体环境，需要使用直接插图命令或者表格命令环境。然后使用 `\captionof` 为其设置标题。为了得到居中的图表，可以使用 `\centerline` 命令或者 `center` 环境。更多详情请参考：[Caption of Figure in Marginpar](#)。

```
% graph with centerline command
\marginpar{
\centerline{
\includegraphics[width=0.2\textwidth]{logo.png}
}
\captionof{figure}{your figure caption}
}

% graph with center environment
\marginpar{
\begin{center}
\includegraphics[width=0.2\textwidth]{logo.png}
\captionof{figure}{your figure caption}
\end{center}
}
```

附录 D 字体选项

字体选项独立成章的原因是，我们希望本模板的用户关心模板使用的字体，知晓自己使用的字体以及遇到字体相关的问题能更加便捷地找到答案。

重要提示：从 3.10 版本更新之后，沿用至今的 newtx 系列字体被重新更改为 cm 字体。并且新增中文字体 (`chinesefont`) 选项。

D.1 数学字体选项

本模板定义了一个数学字体选项 (`math`)，可选项有三个：

1. `math=cm` (默认)，使用 L^AT_EX 默认数学字体 (推荐，无需声明)；
2. `math=newtx`，使用 `newtxmath` 设置数学字体 (潜在问题比较多)。
3. `math=mtpro2`，使用 `mtpro2` 宏包设置数学字体，要求用户已经成功安装此宏包。

D.2 使用 newtx 系列字体

如果需要使用原先版本的 newtx 系列字体，可以通过显示声明数学字体：

```
\documentclass[math=newtx]{elegantbook}
```

D.2.1 连字符

如果使用 newtx 系列字体宏包，需要注意下连字符的问题。

$$\int_{R^q} f(x, y) dy \text{off} \quad (\text{D.1})$$

的代码为

```
\begin{equation}
\int_{R^q} f(x, y) dy \text{of } \kern0pt f
\end{equation}
```

D.2.2 宏包冲突

另外在 3.08 版本中，有用户反馈模板在和 `yhmath` 以及 `esvect` 等宏包搭配使用的时候会出现报错：

```
LaTeX Error:  
Too many symbol fonts declared.
```

原因是在使用 `newtxmath` 宏包时，重新定义了数学字体用于大型操作符，达到了 **最多 16 个数学字体** 的上限，在调用其他宏包的时候，无法新增数学字体。为了减少调用非常用宏包，在此给出如何调用 `yhmath` 以及 `esvect` 宏包的方法。

请在 `elegantbook.cls` 内搜索 `yhmath` 或者 `esvect`，将你所需要的宏包加载语句取消注释即可。

```
%%% use yhmath pkg, uncomment following code
% \let\oldwidering\widering
% \let\widering\undefined
% \RequirePackage{yhmath}
% \let\widering\oldwidering
```

```
%%% use esvect pkg, uncomment following code
% \RequirePackage{esvect}
```

D.3 中文字体选项

模板从 3.10 版本提供中文字体选项 `chinesefont`, 可选项有

1. `ctexfont`: 默认选项, 使用 `ctex` 宏包根据系统自行选择字体, 可能存在字体缺失的问题, 更多内容参考 [ctex 宏包官方文档](#)¹。
2. `founder`: 方正字体选项, 调用 `ctex` 宏包并且使用 `fontset=none` 选项, 然后设置字体为方正四款免费字体, 方正字体下载注意事项见后文。
3. `nofont`: 调用 `ctex` 宏包并且使用 `fontset=none` 选项, 不设定中文字体, 用户可以自行设置中文字体, 具体见后文。

注 使用 `founder` 选项或者 `nofont` 时, 必须使用 `XeLaTeX` 进行编译。

D.3.1 方正字体选项

由于使用 `ctex` 宏包默认调用系统已有的字体, 部分系统字体缺失严重, 因此, 用户希望能够使用其它字体, 我们推荐使用方正字体。方正的方正书宋、方正黑体、方正楷体、方正仿宋四款字体均可免费试用, 且可用于商业用途。用户可以自行从[方正字体官网](#)下载此四款字体, 在下载的时候请务必注意选择 GBK 字符集, 也可以使用 [ETEX 工作室](#)提供的[方正字体](#), 提取码为: `njy9` 进行安装。安装时, Win 10 用户请右键选择为全部用户安装, 否则会找不到字体。

字体名称	编码	单价	实付价	交易状态	操作
订单号: C20200204164821OW1F					2020-02-04 16:48:21
方正仿宋_GBK • 免费商用	简繁扩展(GBK)	¥ 0.00			
方正黑体_GBK • 免费商用	简繁扩展(GBK)	¥ 0.00		免费 已完成	下载字体
方正书宋_GBK • 免费商用	简繁扩展(GBK)	¥ 0.00			
方正楷体_GBK • 免费商用	简繁扩展(GBK)	¥ 0.00			

D.3.2 其他中文字体

如果你想完全自定义字体², 你可以选择 `chinesefont=nofont`, 然后在导言区设置

```
\setCJKmainfont [BoldFont={FZHei-B01},ItalicFont={FZKai-Z03}]{FZShuSong-Z01}
\setCJKsansfont [BoldFont={FZHei-B01},ItalicFont={FZHei-B01}]{FZHei-B01}
\setCJKmonofont [BoldFont={FZHei-B01},ItalicFont={FZHei-B01}]{FZFangSong-Z02}
```

¹可以使用命令提示符, 输入 `texdoc ctex` 调出本地 `ctex` 宏包文档

²这里仍然以方正字体为例。

```
\setCJKfamilyfont{zhsong}{FZShuSong-Z01}
\setCJKfamilyfont{zhhei}{FZHei-B01}
\setCJKfamilyfont{zhkai}{FZKai-Z03}
\setCJKfamilyfont{zhfs}{FZFangSong-Z02}
\newcommand*\songti{\CJKfamily{zhsong}}
\newcommand*\heiti{\CJKfamily{zhhei}}
\newcommand*\kaishu{\CJKfamily{zhkai}}
\newcommand*\fangsong{\CJKfamily{zhfs}}
```

附录 E ElegantBook 写作示例

内容提要

- 定义 ??
- Fubini 定理 E.1.1
- 最优化原理 E.1.1
- 柯西列性质 E.1.1
- 韦达定理

E.1 Lebesgue 积分

在前面各章做了必要的准备后，本章开始介绍新的积分。在 Lebesgue 测度理论的基础上建立了 Lebesgue 积分，其被积函数和积分域更一般，可以对有界函数和无界函数统一处理。正是由于 Lebesgue 积分的这些特点，使得 Lebesgue 积分比 Riemann 积分具有在更一般条件下的极限定理和累次积分交换积分顺序的定理，这使得 Lebesgue 积分不仅在理论上更完善，而且在计算上更灵活有效。

Lebesgue 积分有几种不同的定义方式。我们将采用逐步定义非负简单函数，非负可测函数和一般可测函数积分的方式。

由于现代数学的许多分支如概率论、泛函分析、调和分析等常常用到一般空间上的测度与积分理论，在本章最后一节将介绍一般的测度空间上的积分。

E.1.1 积分的定义

我们将通过三个步骤定义可测函数的积分。首先定义非负简单函数的积分。以下设 E 是 \mathcal{R}^n 中的可测集。

定义 E.1 (可积性)

设 $f(x) = \sum_{i=1}^k a_i \chi_{A_i}(x)$ 是 E 上的非负简单函数，其中 $\{A_1, A_2, \dots, A_k\}$ 是 E 上的一个可测分割， a_1, a_2, \dots, a_k 是非负实数。定义 f 在 E 上的积分为 $\int_a^b f(x) dx$

$$\int_E f dx = \sum_{i=1}^k a_i m(A_i) \pi \alpha \beta \sigma \gamma \nu \xi \epsilon \varepsilon. \oint_a^b \oint_a^b \prod_{i=1}^n$$
 (E.1)

一般情况下 $0 \leq \int_E f dx \leq \infty$ 。若 $\int_E f dx < \infty$ ，则称 f 在 E 上可积。



一个自然的问题是，Lebesgue 积分与我们所熟悉的 Riemann 积分有什么联系和区别？在 4.4 在我们将详细讨论 Riemann 积分与 Lebesgue 积分的关系。这里只看一个简单的例子。设 $D(x)$ 是区间 $[0, 1]$ 上的 Dirichlet 函数。即 $D(x) = \chi_{Q_0}(x)$ ，其中 Q_0 表示 $[0, 1]$ 中的有理数的全体。根据非负简单函数积分的定义， $D(x)$ 在 $[0, 1]$ 上的 Lebesgue 积分为

$$\int_0^1 D(x) dx = \int_0^1 \chi_{Q_0}(x) dx = m(Q_0) = 0$$
 (E.2)

即 $D(x)$ 在 $[0, 1]$ 上是 Lebesgue 可积的并且积分值为零。但 $D(x)$ 在 $[0, 1]$ 上不是 Riemann 可积的。

有界变差函数是与单调函数有密切联系的一类函数。有界变差函数可以表示为两个单调递增函数之差。与单调函数一样，有界变差函数几乎处处可导。与单调函数不同，有界变差函数类对线性运算是封闭的，它们构成一线空间。练习题 E.1 是一个性质的证明。

练习 E.1 设 $f \notin L(\mathcal{R}^1)$, g 是 \mathcal{R}^1 上的有界可测函数。证明函数

$$I(t) = \int_{\mathcal{R}^1} f(x+t) g(x) dx \quad t \in \mathcal{R}^1$$
 (E.3)

是 \mathcal{R}^1 上的连续函数。

解 即 $D(x)$ 在 $[0, 1]$ 上是 Lebesgue 可积的并且积分值为零。但 $D(x)$ 在 $[0, 1]$ 上不是 Riemann 可积的。

证明 即 $D(x)$ 在 $[0, 1]$ 上是 Lebesgue 可积的并且积分值为零。但 $D(x)$ 在 $[0, 1]$ 上不是 Riemann 可积的。

定理 E.1 (Fubini 定理)

(1) 若 $f(x, y)$ 是 $\mathcal{R}^p \times \mathcal{R}^q$ 上的非负可测函数，则对几乎处处的 $x \in \mathcal{R}^p$, $f(x, y)$ 作为 y 的函数是 \mathcal{R}^q 上的非负可测函数， $g(x) = \int_{\mathcal{R}^q} f(x, y) dy$ 是 \mathcal{R}^p 上的非负可测函数。并且

$$\int_{\mathcal{R}^p \times \mathcal{R}^q} f(x, y) dx dy = \int_{\mathcal{R}^p} \left(\int_{\mathcal{R}^q} f(x, y) dy \right) dx. \quad (\text{E.4})$$

(2) 若 $f(x, y)$ 是 $\mathcal{R}^p \times \mathcal{R}^q$ 上的可积函数，则对几乎处处的 $x \in \mathcal{R}^p$, $f(x, y)$ 作为 y 的函数是 \mathcal{R}^q 上的可积函数，并且 $g(x) = \int_{\mathcal{R}^q} f(x, y) dy$ 是 \mathcal{R}^p 上的可积函数。而且 E.4 成立。



E.1.1

笔记 在本模板中，引理 (lemma), 推论 (corollary) 的样式和定理 E.1.1 的样式一致，包括颜色，仅仅只有计数器的设置不一样。

我们说一个实变或者复变量的实值或者复值函数是在区间上平方可积的，如果其绝对值的平方在该区间上的积分是有限的。所有在勒贝格积分意义下平方可积的可测函数构成一个希尔伯特空间，也就是所谓的 L^2 空间，几乎处处相等的函数归为同一等价类。形式上， L^2 是平方可积函数的空间和几乎处处为 0 的函数空间的商空间。

命题 E.1 (最优性原理)

如果 u^* 在 $[s, T]$ 上为最优解，则 u^* 在 $[s, T]$ 任意子区间都是最优解，假设区间为 $[t_0, t_1]$ 的最优解为 u^* ，则 $u(t_0) = u^*(t_0)$ ，即初始条件必须还是在 u^* 上。



我们知道最小二乘法可以用来处理一组数据，可以从一组测定的数据中寻求变量之间的依赖关系，这种函数关系称为经验公式。本课题将介绍最小二乘法的精确定义及如何寻求点与点之间近似成线性关系时的经验公式。假定实验测得变量之间的 n 个数据，则在平面上，可以得到 n 个点，这种图形称为“散点图”，从图中可以粗略看出这些点大致散落在某直线近旁，我们认为其近似为一线性函数，下面介绍求解步骤。

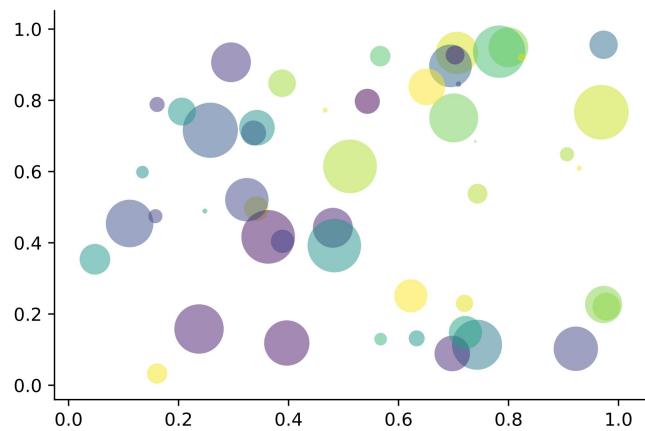


图 E.1: 散点图示例 $\hat{y} = a + bx$

以最简单的一元线性模型来解释最小二乘法。什么是一元线性模型呢？监督学习中，如果预测的变量是离散的，我们称其为分类（如决策树，支持向量机等），如果预测的变量是连续的，我们称其为回归。回归分析中，如果只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线

性回归分析。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。对于二维空间线性是一条直线；对于三维空间线性是一个平面，对于多维空间线性是一个超平面。

性质 柯西列的性质

1. $\{x_k\}$ 是柯西列，则其子列 $\{x_k^i\}$ 也是柯西列。
2. $x_k \in \mathcal{R}^n$, $\rho(x, y)$ 是欧几里得空间，则柯西列收敛， (\mathcal{R}^n, ρ) 空间是完备的。

结论 回归分析 (regression analysis) 是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。运用十分广泛，回归分析按照涉及的变量的多少，分为一元回归和多元回归分析；按照因变量的多少，可分为简单回归分析和多重回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。

第 E 章 练习

1. 设 A 为数域 K 上的 n 级矩阵。证明：如果 K^n 中任意非零列向量都是 A 的特征向量，则 A 一定是数量矩阵。
2. 证明：不为零矩阵的幂零矩阵不能对角化。
3. 设 $A = (a_{ij})$ 是数域 K 上的一个 n 级上三角矩阵，证明：如果 $a_{11} = a_{22} = \dots = a_{nn}$ ，并且至少有一个 $a_{kl} \neq 0 (k < l)$ ，则 A 一定不能对角化。

附录 F 常见问题集

我们根据用户社区反馈整理了下面一些常见的问题，用户在遇到问题时，应当首先查阅本手册和本部分的常见问题。

1. 有没有办法章节用“第一章，第一节，（一）”这种？

见前文介绍，可以使用 `scheme=chinese` 设置。

2. 大佬，我想把正文字体改为亮色，背景色改为黑灰色。

页面颜色可以使用 `\pagecolor` 命令设置，文本命令可以参考[这里](#)进行设置。

3. Package `ctex` Error: CTeX fontset ‘Mac’ is unavailable.

在 Mac 系统下，中文编译请使用 Xe^LAT_EX。

4. ! LaTeX Error: Unknown option ‘`scheme=plain`’ for package ‘`ctex`’.

你用的 CTeX 套装吧？这个里面的 `ctex` 宏包已经是 10 年前的了，与本模板使用的 `ctex` 宏集有很大区别。不建议 CTeX 套装了，请卸载并安装 TeX Live 2021。

5. 我该使用什么版本？

请务必使用[最新正式发行版](#)，发行版间不定期可能会有更新（修复 bug 或者改进之类），如果你在使用过程中没有遇到问题，不需要每次更新[最新版](#)，但是在发行版更新之后，请尽可能使用最新版（发行版）！最新发行版可以在 GitHub 或者 TeX Live 2021 内获取。

6. 我该使用什么编辑器？

你可以使用 TeX Live 2021 自带的编辑器 TeXworks 或者使用 TeXstudio，TeXworks 的自动补全，你可以参考我们的总结 [TeXworks 自动补全](#)。推荐使用 TeX Live 2021 + TeXstudio。我自己用 VS Code 和 Sublime Text，相关的配置说明，请参考 [LaTeX 编译环境配置：Visual Studio Code 配置简介](#) 和 [Sublime Text 搭建 LaTeX 编写环境](#)。

7. 您好，我们想用您的 ElegantBook 模板写一本书。关于机器学习的教材，希望获得您的授权，谢谢您的宝贵时间。

模板的使用修改都是自由的，你们声明模板来源以及模板地址（GitHub 地址）即可，其他未尽事宜按照开源协议 LPPL-1.3c。做好之后，如果方便的话，可以给我们一个链接，我把你们的教材放在 ElegantLaTeX 用户作品集里。

8. 请问交叉引用是什么？

本群和本模板适合有一定 LaTeX 基础的用户使用，新手请先学习 LaTeX 的基础，理解各种概念，否则你将寸步难行。

9. 定义等环境中无法使用加粗命令么？

是这样的，默认中文并没有加粗命令，如果你想在定义等环境中使用加粗命令，请使用 `\heiti` 等字体命令，而不要使用 `\textbf`。或者，你可以将 `\textbf` 重新定义为 `\heiti`。英文模式不存在这个问题。

10. 代码高亮环境能用其他语言吗？

可以的，ElegantBook 模板用的是 `listings` 宏包，你可以在环境（`lstlisting`）之后加上语言（比如 Python 使用 `language=Python` 选项），全局语言修改请使用 `lset` 命令，更多信息请参考宏包文档。

11. 群主，什么时候出 Beamer 的模板（主题），ElegantSlide 或者 ElegantBeamer？

由于 Beamer 中有一个很优秀的主题 Metropolis。后续确定不会再出任何主题/模板，请大家根据需要修改已有主题。

附录 G 版本更新历史

根据用户的反馈，我们不断修正和完善模板。截止到此次更新，ElegantBook 模板在 GitHub 上有将近 100 次提交，正式发行版本（release）有 17 次。由于 3.00 之前版本与现在版本差异非常大，在此不列出 3.00 之前的更新内容。

2021/05/02 更新：版本 4.1 正式发布。

- ① 重要改动：由原先的 `BIBTEX` 改为 `biblatex` 编译方式（后端为 `biber`），请注意两者之间的差异；
- ② 重要改进：修改对于定理写法兼容方式，提高数学公式代码的兼容性；
- ③ 页面设置改动，默认页面更宽；方便书写和阅读；
- ④ 支持目录文字以及页码跳转；
- ⑤ 不再维护 `pdflATEX` 中文支持方式，请务必使用 `XeLATEX` 编译中文文稿。
- ⑥ 增加多个语言选项，法语 `lang=fr`、荷兰语 `lang=nl`、匈牙利语 `lang=hu`、西班牙语 `lang=es`、蒙古语 `lang=mn` 等。

2020/04/12 更新：版本 3.11 正式发布，**此版本为 3.x 最后版本。**

- ① 重要修正：修复因为 `gbt7714` 宏包更新导致的 `natbib` option clash 错误；
- ② 由于 `pgfornament` 宏包未被 TeX Live 2020 收录，因此删除 base 相关的内容；
- ③ 修复部分环境的空格问题；
- ④ 增加了意大利语言选项 `lang=it`。

2020/02/10 更新：版本 3.10 正式发布

- ① 增加数学字体选项 `math`，可选项为 `newtx` 和 `cm`。
重要提示：原先通过 `newtxmath` 宏包设置的数学字体改为 LATEX 默认数学字体，如果需要保持原来的字体，需要显式声明数学字体 (`math=newtx`)；
- ② 新增中文字体选项 `chinesefont`，可选项为 `ctexfont`、`founder` 和 `nofont`。
- ③ 将封面作者信息设置为可选，并且增加自定义信息命令 `\bioinfo`；
- ④ 在说明文档中增加版本历史，新增 `\datechange` 命令和 `change` 环境；
- ⑤ 增加汉化章节选项 `scheme`，可选项为汉化 `chinese`；
- ⑥ 由于 `\lvert` 问题已经修复，重新调整 `ctex` 宏包和 `amsmath` 宏包位置。
- ⑦ 修改页眉设置，去除了 `\lastpage` 以避免 page anchor 问题，加入 `\frontmatter`。
- ⑧ 修改参考文献选项 `cite`，可选项为数字 `numbers`、作者-年份 `authoryear` 以及上标 `super`。
- ⑨ 新增参考文献样式选项 `bibstyle`，并将英文模式下参考文献样式 `apalike` 设置为默认值，中文仍然使用 `gbt7714` 宏包设置。

2019/08/18 更新：版本 3.09 正式发布

- ① `\elegantpar` 存在 bug，删除 `\elegantpar` 命令，建议用户改用 `\marginnote` 和 `\marginpar` 旁注命令。
- ② 积分操作符统一更改为 `esint` 宏包设置；
- ③ 新增目录选项 `toc`，可选项为单栏 `onecol` 和双栏 `twocol`；
- ④ 手动增加参考文献选项 `cite`，可选项为上标形式 `super`；
- ⑤ 修正章节习题（`problemset`）环境。

2019/05/28 更新：版本 3.08 正式发布

- ① 修复 `\part` 命令。

-
- ② 引入 Note 模板中的 `pad` 选项 `device=pad`。
 - ③ 数学字体加入 `mtpro2` 可选项 `math=mtpro2`, 使用免费的 `lite` 子集。
 - ④ 将参考文献默认显示方式 `authoyear` 改为 `numbers`。
 - ⑤ 引入旁注命令 `\marginpar` (测试)。
 - ⑥ 新增章节摘要环境 `introduction`。
 - ⑦ 新增章节习题环境 `problemset`。
 - ⑧ 将 `\equation` 重命名为 `\extrainfo`。
 - ⑨ 完善说明文档, 增加致谢部分。

2019/04/15 更新: 版本 3.07 正式发布

- ① 删除中英文自定义字体总设置。
- ② 新增颜色主题, 并将原绿色默认主题设置为蓝色 `color=blue`。
- ③ 引入隐藏装饰图案选项 `base`, 可选项有显示 `show` 和隐藏 `hide`。
- ④ 新增定理模式 `mode`, 可选项有简单模式 `simple` 和炫彩模式 `fancy`。
- ⑤ 新增隐藏证明、答案等环境的选项 `result=noanswer`。

2019/02/25 更新: 版本 3.06 正式发布

- ① 删除水印。
- ② 新封面, 新装饰图案。
- ③ 添加引言使用说明。
- ④ 修复双面 `twoside`。
- ⑤ 美化列表环境。
- ⑥ 增加 `\subsubsection` 的设置。
- ⑦ 将模板拆分成中英文语言模式。
- ⑧ 使用 `lstlisting` 添加代码高亮。
- ⑨ 增加定理类环境使用说明。

2019/01/22 更新: 版本 3.05 正式发布

- ① 添加 `xeCJK` 宏包中文支持方案。
- ② 修复模板之前对 `TikZ` 单位的改动。
- ③ 更新 logo 图。

2019/01/15 更新: 版本 3.04 正式发布

- ① 格式化模板代码。
- ② 增加 `\equation` 命令。
- ③ 修改 `\date`。

2019/01/08 更新: 版本 3.03 正式发布

- ① 修复附录章节显示问题。
- ② 小幅优化封面代码。

2018/12/31 更新: 版本 3.02 正式发布

- ① 修复名字系列命令自定义格式时出现的空格问题, 比如 `\listfigurename`。
- ② 英文定理类名字改为中文名。
- ③ 英文结构名改为中文。

2018/12/16 更新：版本 3.01 正式发布

- ① 调整 `ctex` 宏包。
 - ② 说明文档增加更新内容。
-

2018/12/06 更新：版本 3.00 正式发布

- ① 删除 `mathpazo` 数学字体选项。
- ② 添加邮箱命令 `\mailto`。
- ③ 修改英文字体为 `newtx` 系列，另外大型操作符号维持 `cm` 字体。
- ④ 中文字体改用 `ctex` 宏包自动设置。
- ⑤ 删除 `xeCJK` 字体设置，原因是不同系统字体不方便统一。
- ⑥ 定理换用 `tcolorbox` 宏包定义，并基本维持原有的定理样式，优化显示效果，支持跨页；定理类名字重命名，如 `etheorem` 改为 `theorem` 等等。
- ⑦ 删去自定义的缩进命令 `\Indent`。
- ⑧ 添加参考文献宏包 `natbib`。
- ⑨ 颜色名字重命名。

附录 A 基本数学工具

本附录包括了计量经济学中用到的一些基本数学，我们扼要论述了求和算子的各种性质，研究了线性和某些非线性方程的性质，并复习了比例和百分数。我们还介绍了一些在应用计量经济学中常见的特殊函数，包括二次函数和自然对数，前 4 节只要求基本的代数技巧，第 5 节则对微分学进行了简要回顾；虽然要理解本书的大部分内容，微积分并非必需，但在一些章末附录和第 3 篇某些高深专题中，我们还是用到了微积分。

A.1 求和算子与描述统计量

求和算子 是用以表达多个数求和运算的一个缩略符号，它在统计学和计量经济学分析中扮演着重要作用。如果 $\{x_i : i = 1, 2, \dots, n\}$ 表示 n 个数的一个序列，那么我们就把这 n 个数的和写为：

$$\sum_{i=1}^n x_i \equiv x_1 + x_2 + \dots + x_n \quad (\text{A.1})$$