定量研究方法  
**Quantitative Research Methods**

**授课教师**彭泽宇 青年副研究员（[zpeng@fudan.edu.cn](mailto:zpeng@fudan.edu.cn)）

**助教**蒋屹阳、周督竣、肖伟林、张睿明、林佳怡

**目录**

[第一讲 绪论：定量研究方法在社科研究中的应用 4](#_Toc185604779)

[一、观察性研究 4](#_Toc185604780)

[（一）观察性研究的案例 4](#_Toc185604781)

[（二）观察性研究的局限性 5](#_Toc185604782)

[二、随机对照试验 5](#_Toc185604783)

[（一）随机对照试验的案例 5](#_Toc185604784)

[（二）随机对照试验的局限性 5](#_Toc185604785)

[第二讲 R语言 5](#_Toc185604786)

[第三讲 概率与条件概率 6](#_Toc185604787)

[一、概率的基本概念 6](#_Toc185604788)

[（一）概率的定义 6](#_Toc185604789)

[（二）概率模型的基本概念 6](#_Toc185604790)

[（三）概率公理 6](#_Toc185604791)

[（四）补集与全概率公式 7](#_Toc185604792)

[二、条件概率 7](#_Toc185604793)

[三、贝叶斯法则 7](#_Toc185604794)

[四、独立性 7](#_Toc185604795)

[第四讲 随机变量和概率分布 8](#_Toc185604796)

[一、随机变量 8](#_Toc185604797)

[（一）随机变量的概念 8](#_Toc185604798)

[（二）随机变量的种类 8](#_Toc185604799)

[二、分布 8](#_Toc185604800)

[（一）伯努利分布 8](#_Toc185604801)

[（二）均匀分布 9](#_Toc185604802)

[（三）二项分布 9](#_Toc185604803)

[（四）正态分布 10](#_Toc185604804)

[三、期望和方差 11](#_Toc185604805)

[（一）期望 11](#_Toc185604806)

[（二）方差 12](#_Toc185604807)

[（三）协方差 12](#_Toc185604808)

[第五讲 点估计 13](#_Toc185604809)

[一、估计 13](#_Toc185604810)

[（一）估计及其相关概念 13](#_Toc185604811)

[（二）估计量的性质 13](#_Toc185604812)

[二、适用于任何样本容量的有限样本特征 13](#_Toc185604813)

[（一）无偏性 13](#_Toc185604814)

[（二）相对有效性 14](#_Toc185604815)

[（三）平均数平方误差 14](#_Toc185604816)

[三、适用于大样本容量的特征 14](#_Toc185604817)

[（一）一致性 14](#_Toc185604818)

[（二）渐进正态性 14](#_Toc185604819)

[第六讲 区间估计 15](#_Toc185604820)

[一、区间估计的相关概念 15](#_Toc185604821)

[（一）区间估计的概念 15](#_Toc185604822)

[（二）标准误 15](#_Toc185604823)

[二、置信区间 16](#_Toc185604824)

[（一）置信区间：*σ*2已知 16](#_Toc185604825)

[（二）置信区间：*σ*2未知 16](#_Toc185604826)

[三、构建置信区间的流程 17](#_Toc185604827)

[第七讲 假设检验 18](#_Toc185604828)

[一、单样本假设检验 18](#_Toc185604829)

[（一）假设 18](#_Toc185604830)

[（二）检验水平*α* 18](#_Toc185604831)

[（三）检验统计量 19](#_Toc185604832)

[（四）假设检验中的*p*值 20](#_Toc185604833)

[（五）置信区间与假设检验 20](#_Toc185604834)

[（六）实际显著性与统计显著性 20](#_Toc185604835)

[二、双样本检验 21](#_Toc185604836)

[三、假设检验的流程 21](#_Toc185604837)

[第八讲 简单线性回归 21](#_Toc185604838)

[一、相关性 22](#_Toc185604839)

[（一）相关性的概念 22](#_Toc185604840)

[（二）相关性的性质 22](#_Toc185604841)

[二、最小二乘法 23](#_Toc185604842)

[（一）简单线性模型 23](#_Toc185604843)

[（二）最小二乘法与相关系数 23](#_Toc185604844)

[第九讲 最小二乘法的性质与应用 24](#_Toc185604845)

[一、回归线与均值 24](#_Toc185604846)

[（一）回归线与样本平均值 24](#_Toc185604847)

[（二）回归线与残差的平均值 24](#_Toc185604848)

[（三）均方根误差 24](#_Toc185604849)

[二、模型拟合度：总平方和、解释平方和与残差平方和 25](#_Toc185604850)

[（一）总平方和、解释平方和与残差平方和的定义及性质 25](#_Toc185604851)

[（二）决定系数*R*2 25](#_Toc185604852)

[三、用于预测的回归模型 25](#_Toc185604853)

[第十讲 最小二乘法的统计性质与推断 26](#_Toc185604854)

[一、最小二乘法的统计性质 26](#_Toc185604855)

[（一）总体回归模型与估计回归模型 26](#_Toc185604856)

[（二）最小二乘法的六大假设 26](#_Toc185604857)

[（三）最小二乘法的无偏性 28](#_Toc185604858)

[（四）最小二乘法的方差 29](#_Toc185604859)

[二、斜率系数的估计与检验 29](#_Toc185604860)

[（一）斜率系数的假设检验 29](#_Toc185604861)

[（二）斜率系数的区间估计 30](#_Toc185604862)

[三、小结 30](#_Toc185604863)

[第十一讲 含有两个自变量的回归模型，遗漏变量偏误及共线性 31](#_Toc185604864)

[一、多元回归模型的动机与意义 31](#_Toc185604865)

[二、含有一个二元协变量的多元回归 31](#_Toc185604866)

[（一）对回归斜率的解释 31](#_Toc185604867)

[（二）二元协变量（covariate） 31](#_Toc185604868)

[三、含有一个连续协变量的多元回归 32](#_Toc185604869)

[四、多元回归的机制 32](#_Toc185604870)

[（一）预测值与残差 32](#_Toc185604871)

[（二）二元（多元）回归中的最小二乘法 33](#_Toc185604872)

[五、多元回归的模型拟合 34](#_Toc185604873)

[六、多元回归假设 34](#_Toc185604874)

[七、遗漏变量偏误 35](#_Toc185604875)

[（一）遗漏变量偏误的概念与案例 35](#_Toc185604876)

[（二）加入无关变量的情况 36](#_Toc185604877)

[八、多元回归的方差和多重共线性问题 37](#_Toc185604878)

[（一）多元回归的方差 37](#_Toc185604879)

[（二）多重共线性问题 37](#_Toc185604880)

[第十二讲 含有定性信息的多元回归分析 38](#_Toc185604881)

[一、单类别虚拟自变量 38](#_Toc185604882)

[（一）虚拟变量的概念 38](#_Toc185604883)

[（二）虚拟变量的应用 38](#_Toc185604884)

[二、多类别虚拟自变量 39](#_Toc185604885)

[三、交互项与交互效应 40](#_Toc185604886)

[（一）含有两个虚拟变量的交互效应 40](#_Toc185604887)

[（二）含有一个虚拟变量和一个连续变量的交互效应 40](#_Toc185604888)

[（三）含有两个连续变量的交互效应 41](#_Toc185604889)

[四、二值因变量：线性概率模型 41](#_Toc185604890)

[（一）线性概率模型的概念 41](#_Toc185604891)

[（二）线性概率模型的案例 42](#_Toc185604892)

[（三）线性概率模型的局限性与预测 42](#_Toc185604893)

第一讲 绪论：定量研究方法在社科研究中的应用

2024.9.6

社会科学量化研究通常分为观察性研究（observational studies）和随机对照试验（randomized controlled trials, RCTs）。

* **观察性研究：**研究人员无法在现实世界中随机分配干预措施，只能观察和收集自然发生的事件中的数据；
* **随机对照试验：**研究人员随机分配干预给研究对象（观察值，observations）。

一、观察性研究

（一）观察性研究的案例

1. 案例背景：无差别炮击是否会引起更强烈的反抗

无差别炮击、轰炸是军事冲突双方常见的策略，同时也是国际安全、军事理论研究中常见的话题。由此催生了一个社会科学问题：在内战中政府军无差别炮击是否会引起民众更强烈的反抗？

这个问题的潜在答案有两个：会引起强烈反抗；或，不会引起强烈反抗，并有效遏制镇压了反抗势力。

我们首先运用推理，提出导向上述答案的可能原因：

* **会引起强烈反抗**
  + 加剧对政府军的仇恨
  + 向叛军/抵抗组织寻求保护
* **有效遏制镇压反抗势力**
  + 减少叛军可获得的资源
  + 威慑恐吓当地民众
  + 打击叛军的后勤网络

我们可以运用定量研究方法验证上述假说。当然，对于这种社会科学问题（尤其是与国际经济军事相关的问题），观察性研究的数据采集难度是很大的。

2. 案例研究：自变量、因变量与干扰因素

杰森·力欧（Jason Lyall, 2009）收集了2000年至2005年俄罗斯联邦军队对车臣村庄实施炮击的数据。数据中存在自变量和因变量之分：

* **自变量（解释变量、预测变量）：**相对稳定且能够决定其他变量的数值的变量；
* **因变量（被解释变量、响应变量）：**数值取决于自变量或其他变量的变量。

我们需要以实证方式验证自变量与因变量之间的关系。在该案例中，自变量是该村是否受到过俄军的无差别炮击；因变量是炮击前后俄军和亲政府武装在该村遭受车臣反政府武装袭击的次数。

当然，存在影响自变量的干扰因素，如人口、贫困率、地形高度（海拔）、是否有俄军进驻、控制该村的车臣领导人派系等。干扰因素又称控制变量。

（二）观察性研究的局限性

观察性研究可能存在选择性偏误（selection bias），其内在效度（internal validity）可能会受到影响，难以证实因果关系。例如，在上面的例子中，研究者可能没有观测到一个同时影响了俄军炮击某一村庄的决定以及当地袭击俄军及亲俄武装的频率的干扰因素。

因此，在随机对照试验当中，研究者可以将干预变量进行随机分配，从而解决选择性偏误问题。

二、随机对照试验

（一）随机对照试验的案例

1. 案例背景：劳动力市场是否存在种族歧视

种族歧视是以美国为代表的西方国家中普遍存在的社会问题。西方国家的劳动力市场是否真的存在种族歧视是社会科学界常年争论的学术话题——例如，不同种族之间的失业率差异是否是由于其他因素（比如教育程度差异）造成的？

在实验中，研究者很难操纵或改变一个人的种族。因此，一个重要的问题就是：如何设计一个随机试验来检视雇主是否会区别对待不同种族的求职者？

2. 案例研究：随机对照试验的设计

经济学家玛丽安娜·伯特兰德（Mariannne Bertrand）和赛德希尔·穆来纳森（Sendhil Muillainathan）进行了以下试验来考察美国劳动力市场的种族歧视问题：研究人员向在报纸上登广告的潜在雇主发去虚构的求职简历；只改变求职者的名字，而将其他信息保持不变。一些求职者用了非常典型的非洲裔美国人的名字（如Lakisha），而其他求职者用了典型的白人名字（如Emily）；比较这两个群体之间的电话回复率。

（二）随机对照试验的局限性

随机对照试验的内在效度普遍高于观察性研究，通常被认为是建立许多科学学科的因果关系的黄金标准。然而，出于道德和操作的种种因素，在很多情况下研究者无法在现实世界随机分配干预措施。例如，在上文提及的“无差别炮击”案例中，社会科学家就不可能将每次炮击的目标随机分配；在本案例中，寄出大量虚构的求职简历也可能会扰乱用人单位人事部门的正常运转。

随机对照试验可能缺乏“外在效度（external validity）”。如在本案例中，研究对象是否包含一个国家/社会中所有族群？在美国，简历上一般不贴照片；那么这个实验是否符合其他国家的社会文化背景？

第二讲 R语言

2024.9.13 / 2024.9.20 / 2024.9.27

见“R语言语法速查”。

第三讲 概率与条件概率

2024.9.27 / 2024.9.29

一、概率的基本概念

概率论是解决社会科学研究中不确定性的重要工具。社会科学研究一般由两步构成：

* 构建理论：在所学的基础上建立一个概率模型；
* 运用数据和其他经验性证据证明概率模型从而支持你的理论。

（一）概率的定义

**概率**是一系列在现实世界中用来测量随机性并对随机性建模的数学工具。

频率统计学派（frequentist）认为，概率是对应频率的极限，是事件发生次数与同样条件下重复试验次数之比，即当在相同条件下反复进行的实验次数接近无穷次是某事件发生频率的极限。

贝叶斯统计学派（Bayesian）认为，概率是用来测量一个人认为某件事发生可能的主观信念。

（二）概率模型的基本概念

**试验（experiment）**是一个或一组产生与某一问题相关的随机事件的行动。

**样本空间（sample space）**是试验所有可能的一组结果，通常用Ω表示。

**事件（event）**是样本空间的一个子集。

（三）概率公理

第一，任一事件*A*的概率非负，即：

第二，样本空间中所有结果中任一发生的概率是1，即：

第三，概率遵循加法法则，即：

对于任意一给定的事件*A*和*B*：

如果事件*A*和*B*是互斥的：

（四）补集与全概率公式

**补集**是所有不属于给定的事件的结果的集合，通常用上标*c*表示。因为*A*和*Ac*是互斥的，且它们一起构成了整个样本空间，故根据加法法则，对于任一给定的事件*A*，有：

同时，我们可以使用韦恩图得出全概率公式：

二、条件概率

事件*A*的**条件概率**指的是事件*A*随着事件*B*的发生而发生，用*P*(*A*|*B*)表示：

在这一公式中，*P*(*A和B*)是两个事件都发生的概率即**联合概率（joint probability）**，其中，*P*(*B*)是事件*B*的**边际概率**。整理公式，我们可以得到乘法法则（multiplication rule）：

综合之前提到的全概率公式，可得出全概率公式的另一种表达方式：

三、贝叶斯法则

贝叶斯法则如下：

其中，*P*(*A*)被称为先验概率（prior probability），反映了一个人关于事件*A*发生可能性的初始信念。在观测到事件*B*代表的数据后，我们更新了自己的信念并得到*P*(*A*|*B*)，*P*(*A*|*B*)被称为后验概率（posterior probability）。

贝叶斯法则的公式是基于条件概率公式推导出的，故可看作是条件概率公式的一种转写。

四、独立性

在直觉上，两个事件的**独立性（independence）**指的是一个事件的信息不会给我们更多关于另一个事件是否发生的信息。根据乘法法则，当且仅当事件*A*和*B*的联合概率等于边际概率的乘积，那么*A*和*B*互相独立的，即：

给定事件*C*，*A*和*B*的**条件独立**意味着给定*C*时*A*和*B*的联合概率等于两个条件概率的乘积：

同时，这也意味着：

第四讲 随机变量和概率分布

2024.9.29 / 2024.10.11 / 2024.10.18

一、随机变量

（一）随机变量的概念

随机变量*X*指的是将样本空间中的每个事件赋予一个实数的函数（即赋予每一个事件一个数）。

随机变量的数值必须表示互斥（mutually exclusive），不同数值不能表示同一事件（如事件*A*不能同时用1和2来表示）。

随机变量的数值必须是完备的（exhaustive），全部事件应该有一些数值来表示。

（二）随机变量的种类

* **离散随机变量：**数值为有限个，如家庭成员个数、态度（支持/反对）等。其概率分布函数包括概率质量函数（probability mass function, PMF）、累积分布函数（cumulative distribution function, CDF）。
* **连续随机变量：**在某个实线区间内有无限个取值，如长度、GDP等。其概率分布函数包括概率密度函数（probability density function, PDF）、累积分布函数。

二、分布

（一）伯努利分布

1. 伯努利随机变量的定义

我们先定义一个二元随机变量（binary random variable），该二元随机变量的数值为两个不同的数。我们将之称为**伯努利随机变量（Benoulli random variable）**。我们一般认为，事件*X*=1表示成功，*X*=0表示失败；我们用*p*表示成功的概率。

伯努利随机变量可以用**概率质量函数（PMF）**表示。一个随机变量*X*的PMF的定义是：在随机变量取某一值*x*时，*f*(*x*)=*P*(*X*=*x*)的概率。以伯努利随机变量为例，PMF在*x*=1（成功）时取*p*值，在*x*=0（失败）时取值为1−*p*，除此之外函数*f*(*x*)的值对应其他任何*x*值都是0。

离散随机变量的**累积分布函数（CDF）***F*(*x*)表示随机变量*X*取值小于或等于某一特定值*x*时的累积概率，即*F*(*x*)=*P*(*X*≤*x*)。因此CDF表示PMF *f*(*x*)对于所有*x*值的和。任何离散随机变量PMF *f*(*x*)和CDF *F*(*x*)之间的关系可以写为：

其中，*k*表示所有小于或等于*x*的随机变量*X*的值。

2. 伯努利随机变量的概率分布函数

一个成功概率为*p*的伯努利随机变量的概率质量函数（PMF）为：

其中*f*(1)和*f*(0)分别代表成功和失败的概率。

一个成功概率为*p*的伯努利随机变量的累积分布函数（CDF）为：

若*x*≥1，函数值为1-*p*+*p*=1。

（二）均匀分布

1. 均匀随机变量的定义

**均匀随机变量（uniform random variable）**是连续随机变量的一种简单类型。均匀随机变量在给定区间[*a*, *b*]内取一值的可能性相同。

均匀随机变量可用概率密度函数（PDF）或者累积分布函数（CDF）来表示。

概率密度函数的数值非负且可以大于1，但是PDF下方的区域面积必须等于1。需要注意的是，与概率质量函数（PMF）不同，PMF的数值不能大于1，而PDF的取值可以大于1。**均匀分布的PDF**是一条由定义的水平线，因为PDF下方区域的面积等于1。

**均匀分布的CDF** *F*(*x*)表示随机变量*X*取值小于或等于给定值*x*的概率，即*P*(*X*≤*x*)。如果*a*≤*x*< *b*，PDF下方与*x*对应的面积就是CDF的值。数理上，我们可以用积分来表示这个概念：

2. 均匀随机变量的概率分布函数

在区间[*a*, *b*]的均匀随机变量的概率密度函数为：

在区间[*a*, *b*]的均匀随机变量的累积分布函数为：

（三）二项分布

1. 二项随机变量的定义

**二项分布（binomial distribution）**是对伯努利分布的一般化。二项随机变量*X*记录了*n*次独立且相同的、成功概率为*p*的试验中的成功次数。

一个**二项随机变量**是*n*个独立并且相同分布（independently and identically distributed, 或者简称为i.i.d.）的伯努利随机变量之和。在二项分布中，*X*可以取0到*n*的整数。

2. 二项随机变量的概率分布函数

试验次数*n*，成功率为*p*的二项随机变量的**概率质量函数**公式为：

其中，

即*n*选*x*。

对于*x*=0, 1, ..., *n*, **累积分布函数**可以写为：

（四）正态分布

1. 正态随机变量的定义

一个**正态随机变量（normal random variable）**可以取实线(−∞, ∞)上任意一点。一个正态分布有两个重要参数：平均数*µ*，标准差*σ*。一个正态随机变量*X*可以写作*X*∼*N*(*µ*, *σ*2)，其中*σ*2表示方差。

2. 正态随机变量的分布函数

一个正态随机变量的**概率密度函数（PDF）**为：

对于实线上任意一个*x*，正态随机变量的**累积分布函数（CDF）**为：

CDF代表PDF下方从-∞到*x*所形成的面积。

正态分布PDF形状为钟型，中心为平均数，分布范围由标准差大小决定。正态分布PDF下的面积为1。不同的平均数只移动PDF和CDF，但不改变形状。标准差增加可造成PDF更平坦，CDF增长趋势更平缓。

3. z分数

假设随机变量*X*∼*N*(*µ*, *σ*2)。令*c*为任一常数，下列性质成立：

* 定义为*Z*=*X*+*c*的随机变量服从*Z*∼*N*(*µ*+*c*, *σ*2)的正态分布；
* 定义为*Z*=*cX*的随机变量服从*Z*∼*N*(*cµ*, (*cσ*)2)的正态分布。

根据以上性质，我们可以推出：

我们称为正态随机变量的***z*分数**，即：

*z*分数也可以理解为随机变量*X*的某一个数值与平均值*µ*之间有多少个标准差*σ*的距离。如果数据按照正态分布，则约2/3的数值在平均值的一个标准差之内，约95%的数值在平均值的两个标准差之内。

计算正态随机变量位于平均值*k* (*k* > 0)个标准差之间的概率可用：

其中，*P*(-*k* ≤ *Z* ≤ *k*)也可以被转写成CDF在-*k*和*k*取值时的差，即：

三、期望和方差

（一）期望

1. 期望的基本概念

**期望（expectation）**代表一个概率分布下的平均值，也称为总体平均值（population mean）。随机变量的期望在给定的概率模型中是固定的。我们通常用*E*(*X*)来表示随机变量*X*的期望。

用*E*(*X*)表示的随机变量的期望值定义为：

其中，*f*(*x*)是离散变量*X*的概率质量函数（PMF）或连续变量*X*的概率密度函数（PDF）。

2. 期望的性质

令*X*和*Y*为随机变量，*a*和*b*是任意常数。期望具有如下线性性质：

如果*X*和*Y*独立，则有：

应注意，*E*(*XY*)=*E*(*X*)*E*(*Y*)是*X*和*Y*独立的必要非充分条件。

（二）方差

1. 方差的基本概念

随机变量的标准差（standard deviation）的平方即**方差（variance）**。我们通常用*V*(*X*)来表示随机变量*X*的方差。

随机变量*X*的方差被定义为：

上式可被转写为：

2. 方差的性质

令*X*和*Y*为随机变量，*a*和*b*是任意常数。方差具有如下性质：

如果*X*和*Y*独立，则有：

（三）协方差

协方差用于度量两个随机变量*X*和*Y*的相关变化。我们通常用Cov(*X*, *Y*)来表示随机变量*X*和*Y*的协方差。

随机变量*X*和*Y*的协方差被定义为：

若*X*=*Y*，则Cov(*X*, *Y*)=*V*(*X*)。

如果*X*和*Y*独立，则有：

应注意，Cov(*X*, *Y*)=0是*X*和*Y*独立的必要非充分条件。

第五讲 点估计

2024.10.18 / 2024.10.25

一、估计

（一）估计及其相关概念

**统计推断**是利用来自某总体的一个样本从而对总体有所了解。统计推断的方法大致可归类为**估计（estimation）**和假设检验（hypothesis testing）；估计又分为点估计（point estimate） 和区间估计（interval estimate）。

与估计相关的概念有估计目标、估计量与估计值。

**估计目标（estimand）**是我们需要估计的总体的参数（parameter）。通常用希腊字母表示（如*μ*、*θ*）。

**估计量（estimator）**是我们估算估计目标的方法。通常用希腊字母加hat表示（如）。

**估计值（estimate）**是我们运用估计量从所得样本中进行估算所得的具体数值。

（二）估计量的性质

估计量是随机变量，其随机性来源于重复抽样。由重复随机抽样而得来的估计量的分布常被称为抽样分布（sampling distribution）。估计量的性质指的是其抽样分布的特征。

适用于任何样本容量的有限样本特征有：

* **无偏性（unbiasedness）：**估计量的抽样分布是否围绕在估计目标参数附近（）；
* **相对有效性（efficiency）：**抽样分布的方差是否较小（）。

适用于大样本容量的特征有：

* **一致性（consistency）：**当我们的样本容量（*n*）足够大的时候，估计量的抽样分布是否趋近于估计目标；
* **渐进正态性（asymptotic normality）：**当我们的样本容量（*n*）足够大的时候，估计量的抽样样本是否呈正态分布。

二、适用于任何样本容量的有限样本特征

（一）无偏性

**偏误（bias）**指的是估计量的期望与估计目标的期望值的差。

我们称一个估计量的**无偏的**，如果

如果我们能够从总体中抽取关于*X*的无限多个样本，并且每次都计算一个估计值，那么将所有随机样本的这些估计值平均起来，我们便得到*θ*。

（二）相对有效性

相对有效性是比较不同无偏估计量的方法之一。在多个无偏误的估计量当中，我们通常选取方差最小、相对有效性最高的那一个。

**相对有效性**就是：比较和，如果的方差小于的方差，那么更有效。

（三）平均数平方误差

一般而言，我们倾向选择无偏并且相对有效的估计量。如果估计量都存在偏误，我们该如何选取合适的估计方案？此时就需要引入平均数平方误差，即均方误（mean-squared error, MSE）：

可见，均方误实际上是估计量的方差与偏误平方之和。

三、适用于大样本容量的特征

（一）一致性

如果一个估计量的抽样分布随着样本容量*n*的增加越来越集中于估计目标*θ*，那么就是**一致的**。

即

一致性体现在估计量的期望值越来越接近估计目标以及估计量的方差逐渐趋近于0。一致的估计量不一定无偏，无偏误的估计量不一定具有一致性。

（二）渐进正态性

1. 大数定理

假设我们得到一个有*n*个观测值，独立并且相同分布的随机变量的样本*X*1, *X*2, ..., *Xn*。大数定理指在样本容量足够大的情况下：

其中，*μ*可以用*E*(*X*)来表示。

2. 中心极限定理

**中心极限定理（central limit theorem, CLT）**是描述渐进正态性的一个强有力的理论。中心极限定理指出，样本平均数的分布随着样本数的增加接近正态分布。这一结论可以广泛应用于各种分布中。

中心极限定理的内容是：假设我们有独立并且相同分布（i.i.d）的随机变量的样本*X*1, *X*2, ..., *Xn*，形成一个均值为*μ*，方差为*σ*2的概率分布。样本平均数用表示，那么中心极限定理为：

其中，*μ*可以用*E*(*X*)来表示，*σ*2可以用*V*(*X*)来表示。

这一公式的含义是：随着样本规模的增加，分布样本平均数的*z*分数收敛成标准正态分布，即*N*(0,1)。如果样本容量*n*足够大，我们可以用样本标准差*Sn*来代替*σ*。

在样本容量足够大的情况下，我们可以通过正态分布的性质推出中心极限定理的另一种写法：

这一公式的含义是：随着样本规模的增加，样本平均数的分布收敛成正态分布，即。

第六讲 区间估计

2024.10.25

一、区间估计的相关概念

（一）区间估计的概念

运用估计量估计我们只能得到一个标量数值*θ*，也就是点估计。由于我们运用了重复多次随机抽样的方法来对总体参数进行估计，我们同时也关注每次估计的不确定性（可以理解为每次抽样的计算出的估计值的波动范围）。

一个区间的估计量的表示方法是：

区间代表了估计目标可能被包含的区域。

（二）标准误

**标准误（standard error）**是推断统计中的重要概念。标准误用于刻画随机抽样所得的样本平均值的离散程度。对于样本平均数，标准误的一般化公式如下：

假设我们有独立且相同分布的随机变量样本{*X*1, *X*2, ..., *Xn*}，并且已知总体方差*σ*2的时候，样本平均数的标准误为：

我们同样可以通过样本方差来估计总体方差*σ*2：

二、置信区间

我们回顾中心极限定理：

我们可以根据上述结论构建出另一个对不确定性的测量指标，即**置信区间（confidence interval）**。置信区间给出了可能包括估计目标真实值的数值范围。

为计算置信区间，研究者们制定了**置信水平/置信度（confidence level）**，这决定了在多大程度上确信区间内包括估计目标的真实值。研究者通常选取95%、90%和99%作为置信水平（其中95%置信水平最为常见）。置信水平通常被写为(1−*α*)×100%（*α*=0.05对应95%的置信水平）。

（一）置信区间：*σ*2已知

假设我们有独立并且相同分布的随机变量的样本*X*1, *X*2, ..., *Xn*，并且来自于总体平均值为*µ*，方差为*σ*2的正态分布中，总体方差已知的情况下，对于样本平均数的(1−*α*)×100%置信区间为：

其中为标准误，为**临界值（critical value）**。

临界值等于标准正态分布N(0, 1)的分位数。最大临界值和最小临界值围成的面积为置信水平(1−*α*)×100%。置信水平越高，最大临界值越高，在其他条件不变的情况下，所得置信区间的范围越大。

*n*%置信区间的含义是：在重复数据产生的过程中，有*n*%的数据的区间内包含估计目标。

（二）置信区间：*σ*2未知

1. 学生t分布

如果总体方差*σ*2未知，我们可以用样本方差*Sn*2来估计*σ*2。是*σ*2无偏且具有一致性的估计量。

假设我们有独立且相同分布的随机变量样本{*X*1, *X*2, ..., *Xn*}，并且总体方差*σ*2未知的时候，样本平均数的标准误是：

如果*σ*2未知，我们则需要运用**学生t分布（Student’s t-distribution）**来构建置信区间。

假设我们有独立并且相同分布的随机变量的样本*X*1, *X*2, ..., *Xn*，并且来自于总体平均值为*µ*，方差为*σ*2的正态分布中。那么样本平均数和标准误的t统计量服从自由度为*n*-1的学生t分布：

t分布与正态分布相似，但是尾部更大。随着自由度（df）的增加，t分布接近正态分布。一般而言，当自由度大于100的时候，t分布和正态分布是没有显著区别的。自由度为观测值的个数（样本容量）减去待估参数的个数；在估计*µ*的例子中，待估参数为1。

2. *σ*2未知时的置信区间公式

因此，在总体方差未知的情况下，对于样本平均数的(1−*α*)×100%的大样本置信区间为：

其中为标准误，为临界值。

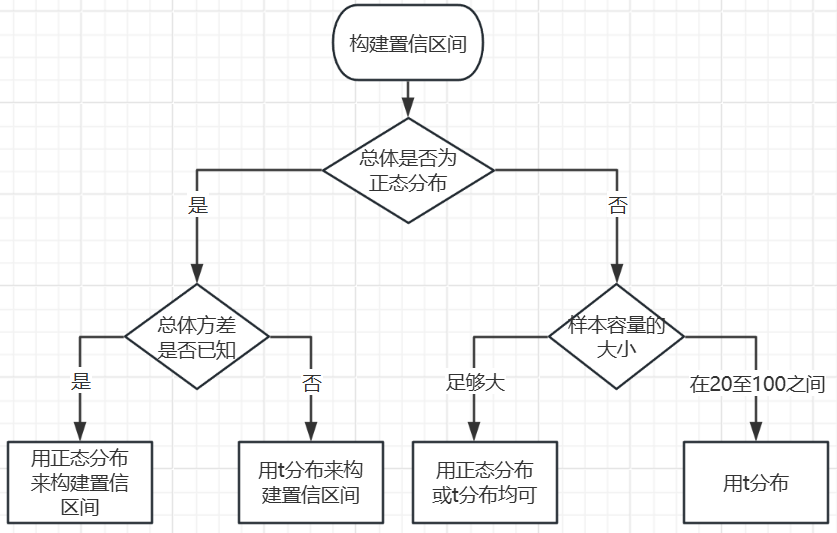
在相同置信水平的情况下，t分布的临界值要稍大于正态分布的临界值。

3. t分布与中心极限定理

随着自由度（df）的增加，t分布接近正态分布。随着*n*→∞，则：

这意味着当样本容量足够大的时候我们不需要假设总体服从正态分布。

三、构建置信区间的流程



第七讲 假设检验

2024.10.25 / 2024.11.1

在实证研究中，我们通常对研究问题有一个明确的肯定或者否定的答复；而运用样本数据来回答这类问题的方法就是假设检验。

一、单样本假设检验

为了理解假设检验的含义，我们举一个美国政治的例子。

（一）假设

假设美国前总统特朗普声称拜登在2020年总统大选中舞弊，只有45%的美国选民支持拜登；但我们认为拜登的支持率很可能超过45%。

我们把这个最终可能被否定的假设，即“2020年总统大选中只有45%美国选民支持拜登”称为**零假设（null hypothesis）**。特朗普认为的45%支持率通常被称为预先指定值*μ*0，设拜登在2020年的支持率为*μ*，则零假设可以写成：

零假设就像正在受审的被告人，要认定被告人有罪，就必须拿出强有力的数据/证据来反对*H*0。我们不认为拜登在2020年大选期间支持率只有45%，在这种情况下我们会提出**备选假设（alternative hypothesis）**：

（二）检验水平*α*

提出备选假设之后，我们运用美国大选研究的样本（n = 1000），发现这组随机样本中有58.3%的选民支持拜登。那么我们观测到的数据能否推翻（拒绝）特朗普的观点？58.3%和45%是有差距的，但是这种差距是因为抽样误差造成的，还是因为*H*0与真实的支持率有差异造成的？要回答这个问题，我们要选取一个**检验水平*α*（level of test α）**和一个**检验统计量**。

在假设检验中我们会遇到四种情况：

|  |  |  |
| --- | --- | --- |
|  | **保持*H*0** | **拒绝*H*0** |
| ***H*0是正确的** | 正确 | I类错误 |
| ***H*0是错误的** | II类错误 | 正确 |

其中包括了两类错误：

* **I类错误（弃真）：**拒绝了正确的*H*0；
* **II类错误（取伪）：**未能拒绝错误的*H*0。

研究者更关注如何控制第一类错误，在做检验的时候会在一开始就设定一个*α*值，即检验水平*α*（通常也被称为显著性水平）。*α*值代表着我们对犯第一类错误的容忍程度，常用的*α*值有0.10、0.05和0.01。在检测拜登支持率的这个例子中，我们将*α*值设为0.05。

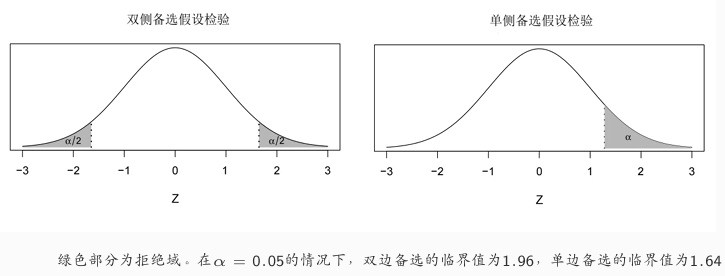
（三）检验统计量

**检验统计量**是观测数据中的某种函数。拜登在抽样民调中的支持率同时也是样本的平均值是，我们选择该样本的平均值的z分数作为检验统计量。根据中心极限定理，z分数与0.45之间差距的z分数是：

其中，我们可以用样本方差来估计。由于样本容量*N*足够大（n=1000），我们可以运用中心极限定理来近似零假设下z分数的分布：

接下来我们可以用**R**计算出，将其代入可得：

再将*Zn*的绝对值与标准正态分布中的临界值*z*作比较。当备选假设为*µ*>*µ*0或者*µ*<*µ*0时，我们选择与显著值*α*对应的单边备选临界值；当备选假设为*µ*≠*µ*0时，我们选择与显著值*α*对应的双边备选临界。



我们先看单边备选临界值的情况。由于我们的备选假设是*µ*>0.45，所以我们选择单边备选临界值*z*=1.64。可见：

检验统计量z分数的绝对值大于临界值，所以我们拒绝拜登的支持率只有45%这个零假设。这意味着58.3%和45%之间的差距并不仅仅是因为抽样的误差而造成的，而是拜登的实际支持率与预先指定值*µ*0之间的差异造成的。

再看双边备选临界值的情况。如果我们的备选假设是*µ*≠0.45，即拜登的支持度不是45%，那我们选择双边备选临界值*z*=1.96。可见：

检验统计量z分数的绝对值大于临界值，所以我们同样拒绝拜登的支持率只有45%这个零假设。

（四）假设检验中的*p*值

除通过比较统计量z分数的绝对值和临界值*z*之外，研究者还会通过检视*p*值（p-value）的大小来决定是否拒绝零假设。***p*值**指的是在零假设正确的情况下，至少观测到一次检验统计量的概率；*p*值越小，拒绝零假设的证据越强。

如果备选假设为单边（即*H*1:*µ*>0.45）：

如果备选假设为双边（即*H*1:*µ*≠0.45）：

如果我们选取*α*=0.05作为显著性，那么在上述的*p*值检验中，当*p*值<0.05时，我们拒绝*H*0；当*p*值≥0.05时，我们不拒绝*H*0。可见，上述两个*p*值都远低于0.05，所以我们拒绝*H*0。

（五）置信区间与假设检验

置信区间和双边假设检验存在一对一的关系。如果在上一个例子中选择双边备选假设，研究者可以通过观察的置信区间是否包括0.45来判断是否拒绝*H*0。如果置信区间包括0.45，则不拒绝*H*0；如果置信区间不包括0.45，则拒绝*H*0。

（六）实际显著性与统计显著性

在拜登支持率的例子中，我们拒绝*H*0。这意味着样本中拜登的支持度和特朗普所声称的支持率之间的差距是有**统计显著性**的。而这种统计显著性体现在检验统计量的z分数很大但是当我们回顾公式：

我们会发现z分数大的一个原因可能是我们的样本容量（n =1000）非常大或者方差*σ*2很小。但是13.3%的差距是否足够大，即是否具有**实际显著性**，我们需要结合前人的结论再加以讨论。

二、双样本检验

我们可以将单样本假设检验的框架和流程推广到双样本检验当中。例如我们想知道拜登在2020年的支持率与希拉里·克林顿在2016年的支持率是否一致。

我们设拜登在2020年的支持情况为独立且相同的随机变量，并且来自于平均值为*µa*，方差为*σa*2的总体概率分布，希拉里在2016年的支持情况为独立且相同的随机变量，并且来自于平均值为*µb*，方差为*σb*2的总体概率分布。两个样本抽取过程是相互独立的；两个随机变量的样本容量足够大（*na*=1000；*nb* =1100）。当样本容量足够大的时候，我们可以用样本平均和估计*µa*和*µb*，用样本方差*Sna*2和*Snb*2来估计*σa*2和*σb*2。

我们将两个总体具有相同平均值作为零假设：*H*0 : *µa*=*µb*，即*H*0 : *µa*−*µb*=0。我们选择双边备选假设：*H*1 :*µa*≠*µb*（拜登和希拉里的支持率不一致）。我们将检验水平*α*设为0.05。下面，我们计算检验统计量：

代入数据，计算出统计量*Zn*=2.90，与临界值1.96比较。由于*Zn*=2.90>1.96，所以我们拒绝*H*0。

三、假设检验的流程

1. 设置零假设*H*0；
2. 设置备选假设*H*1并确定采取单边检验还是双边检验；
3. 设定检验水平*α*；
4. 选取检验统计量*Zn*。如果*n*足够大，*Zn*∼*N*(0,1)；
5. 计算统计量*Zn*；
6. 根据检验水平选取临界值*z*；
   1. 检验水平*α*=0.05；
   2. 临界值*z*——单边：1.64；双边：1.96。
7. 比较*Zn*与*z*；
   1. 如果统计量|*Zn*|>*z*，拒绝*H*0；
   2. 如果统计量|*Zn*|≤*z*，不拒绝（保持）*H*0。
8. 同时我们也可以计算*p*值，再比较*p*值与*α*值。
   1. *p*值<*α*值，拒绝*H*0；
   2. *p*值≥*α*值，不拒绝*H*0。

第八讲 简单线性回归

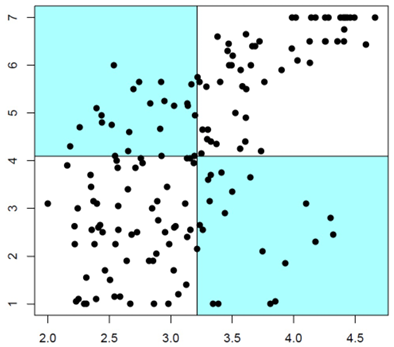
2024.11.10

一、相关性

（一）相关性的概念

我们回顾协方差这一概念——协方差用于度量两个随机变量*X*和*Y*的相关变化。随机变量*X*和*Y*的**样本协方差**可以写成数学表达公式可以写成：

当Cov(*X*, *Y*)为正时，整体数据云的趋势向上；为负时，则向下。如图：



协方差的一个标准化表达方式为相关系数。**相关系数**衡量两个变量相互关联的程度，样本相关系数的数学表达公式可以写成：

相关系数是协方差的一种标准化的表达方式：

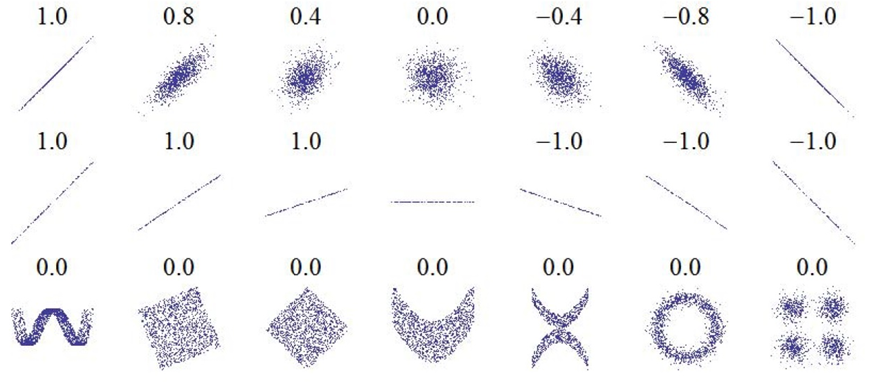
其中，*Sx*、*Sy*分别是变量*X*和*Y*的标准差。

（二）相关性的性质

下面列举了相关性的一些性质：

1. 相关系数是一种刻画两个随机变量**线性**关系的指标；
2. 相关系数为正（负）时，一个观测值（observation）的*X*值大于（小于）其样本平均数，那么这个观测值所对应的*Y*值很可能也大于其样本平均值；
3. 相关系数为正（负），散点图中的数据云呈上升（下降）趋势；
4. 相关系数的绝对值大，则数据点紧紧环绕在直线周围；
5. 相关系数的范围是[−1,1]，1（-1）为完全正（负）相关；
6. *X*和*Y*相关并不意味着二者具有因果关系。

相关系数与数据云的趋势如下图所示：



二、最小二乘法

（一）简单线性模型

1. 线性模型

我们还可以运用**线性模型**来表示两个变量之间的线性关系：

其中，*Y*为结果变量（响应变量、因变量），如投票率、政策效应、战争等；*X*为预测因子（解释变量、自变量），如政治制度、选民特征、政策干预等。

截距*α*表示*X*为0的时候*Y*的平均值。斜率*β*度量当*X*增加一个单位时*Y*平均增加的量。研究者允许观测值偏离完美线性关系，所以*ϵ*代表误差（扰动）项。

2. 回归模型

以上等式中的*α*和*β*对于研究者来说是未知的，研究者需要通过数据去估计斜率和截距。研究者通常用**估计模型（拟合模型）**来估计*α*和*β*，这一类模型也叫做**回归模型**：

其中，和分别代表了*α*和*β*的估计值。我们把观测到的*X*=*x*值代入回归模型，就能够计算出，即结果变量的预测值（或拟合值）。

观测结果值*Y*与预测值之间的差叫做**残差（residual，预测误差）**：

（二）最小二乘法与相关系数

为了画出一条“最佳”的回归线来描述*X*和*Y*之间的线性关系，选取“最佳”的和，我们就需要最小化**残差平方和（sum squared residual, SSR）**。

经过推导，我们可以得出最小二乘估计的和：

根据协方差和样本方差的公式，我们发现：

根据相关系数的公式，我们发现：

其中，*ρXY*是*X*与*Y*的相关系数。

由此可见，若数值大，则原因可能是：

* *X*和*Y*之间的相关性强；
* *Y*的标准差/方差大；
* *X*的标准差/方差小。

第九讲 最小二乘法的性质与应用

2024.11.15

一、回归线与均值

在上一讲中，我们已经知道，截距和斜率参数的最小二乘估计和由下式给出：

（一）回归线与样本平均值

回顾回归线的公式：

将以及代入上式，可得，这就意味着回归线总是通过数据的中心点。

（二）回归线与残差的平均值

回顾残差之和的公式：

将其除以样本容量*n*，可得残差平均值为0（过程略）。

（三）均方根误差

由于残差的平均值为0，故需要一个新的指标来表示回归预测误差的平均程度。我们引入**均方根误差（root mean squared error, RMSE）**，它表示回归预测误差的平均幅度。

二、模型拟合度：总平方和、解释平方和与残差平方和

（一）总平方和、解释平方和与残差平方和的定义及性质

**总平方和（total sum of squares, TSS）**的公式是：

总平方和可被拆解为两部分：**解释平方和（explained sum of squares, ESS）**与**残差平方和（sum of squares residual, SSR）**。

三个平方和之间存在关系。在一个回归模型中，解释平方和是可被回归模型解释的部分，残差平方和则是没有被回归模型预测到的部分。因此：

（二）决定系数*R*2

我们引入**决定系数*R*2**，其为预测值*X*可以解释的总平方和（TSS）的比例，表明线性模型与数据的拟合程度。

*R*2的性质包括：

* 0≤*R*2≤1；
* 当*R*2=1时，所有的数据点都在回归线上；
* 当*R*2=0时，*X*和*Y*之间没有相关性。

应注意，*R*2越大，数据的拟合程度**不一定**越好。

三、用于预测的回归模型

回归模型刻画了作为*X*的函数*Y*如何随着预测变量*X*变化而变化。我们也可以用条件平均值和条件期望函数来表示回归模型。*E*[*Y*|*X*]描述了在所有*X*可以取到的值的条件下*Y*的平均值是如何变化的。我们可以用观察到的样本来拟合出*E*[*Y*|*X*]，从而运用拟合模型来进行预测推断。

回归模型可以用作预测新的观察值。假设我们根据样本数据拟合出了回归线，我们可以将一个新的*X*值*xnew*代入回归模型算出*Y*的期望值：

如果代入两个不同的*X*值*xnew*,1、*xnew*,2，我们可以算出两个*Y*值的期望差异：

可见，可以被解释成当*X*改变一个单位时两个数据点的*Y*值的平均/期望差异。在实际问题中，我们可以用这样的语句表述：

* 若*X*增加（减少）*n*个单位，则*Y*的预期变化为；
* 若*X*增加（减少）*n*个单位与*Y*预期变化的相关。

但不能表述为“*X*增加（减少）*n*个单位导致*Y*的变化为”，因为没有零条件均值（zero conditional mean）等其他条件，我们不能够从回归模型中推出因果关系。

第十讲 最小二乘法的统计性质与推断

2024.11.29

一、最小二乘法的统计性质

（一）总体回归模型与估计回归模型

在现实中，我们往往不知道在总体中*X*和*Y*之间的关系。因此，我们要用样本估计总体，并且运用无偏性等概念来评估最小二乘法回归模型，对斜率参数进行假设检验。

总体回归模型可以写成：

其中，*Y*、*X*分别是因变量和自变量，*α*、*β*分别是截距和斜率，*ϵ*是误差项（扰动项，可以看作是一个*X*以外的所有未知并能够影响*Y*的因素）。

估计模型可以写成：

且

其中，、分别是截距和斜率的估计值，是预测值，是残差（没有被*X*解释的*Y*的部分，可以看作是对*ϵ*误差项的估计）。

（二）最小二乘法的六大假设

最小二乘法具有如下假设：

1. **线性于参数：**总体回归模型中的参数是线性的；
2. **随机抽样：**观测到的数据是从总体中通过随机抽样得到的；
3. **自变量的样本波动性：**自变量*X*必须有波动（变化）；
4. **零条件均值：**给定自变量*X*的任何值，误差项的期望值为零；
5. **同方差性：**给定自变量的任何值，误差都具有相同的方差；
6. **正态性：**误差项和自变量相互独立且服从正态分布。

满足上述假设中的一个或多个时，就可以作出不同层次的统计推断：

* **描述性统计：**假设3；
* **无偏性和一致性：**假设1~4；
* **高斯马科夫定理（BLUE）大样本推断：**假设1~5；
* **经典线性模型（BUE）小样本推断：**假设1~6。

下面我们主要讨论假设1~5。

1. 线性于参数

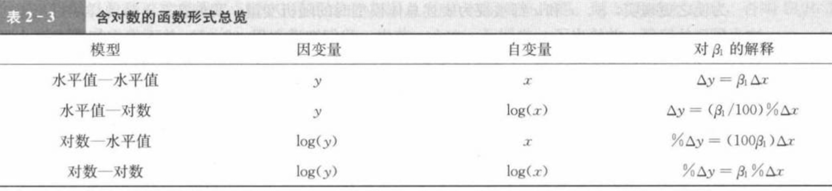
线性于参数指的是，在总体模型中，参数是线性的并且因变量*Y*与自变量*X*和误差*ϵ*的关系如下：

注意在这条假设下变量可以是非线性的，这意味着我们可以根据散点图中数据云的形状加入非线性因素。下面我们就来讨论这一点。

在线性回归模型中，我们可以将非线性因素加入到变量中。我们可以通过对变量进行合理转换来加入非线性因素。当数据向右偏斜（数据点向左扎堆）的时候，我们可以将变量进行对数（log）转换。对数转换仅适用于连续变量，不适用于离散变量。

当我们对变量进行对数转换的时候，对斜率*β*的解释会有所变化：

* log(*Y*)关于*X*的回归→百分之100 *β*的*Y*增量与一个单位的*X*的增量有关；
* *Y*关于log(*X*)的回归→*Y*的增量(*β*)与百分之一的*X*的增量有关；
* log(*Y*)关于log(*X*)的回归→百分之*β*的*Y*的增量与百分之一的*X*的增量有关。



2. 随机抽样

随意抽样指的是样本为服从总体模型方程的随机样本{(*Xi*, *Yi*):(*i* =1, 2, ..., *n*)}。

满足该假设的一个例子是通过随机抽样产生的截面问卷调查数据。可能违反该假设的例子包括时间序列数据（可能产生序列相关）以及不是通过随机抽样而获得的数据。

3. 自变量的样本波动性

自变量的样本波动性指的是*X*的样本*X*1, *X*2, ..., *Xn*不是完全相同的数值。注意到，在*X*1, *X*2, ..., *Xn*完全相同的情况下：

此时分母为0，则无法被识别。

4. 零条件均值

零条件均值指的是，给定自变量*X*的任何值，误差项的期望值为零，即

在随机抽样的条件下，该假设可以写成：对所有*X*1, *X*2, ..., *Xn*，都有*E*(*ϵi*|*Xi*)=0；这也意味着Cov(*X*, *ϵ*)=0。

为了更好地理解这一假设，我们以薪水与受教育程度之间关系为例：

注意到*ϵ*代表受教育程度以外的所有未知并能够影响薪水的因素。假设其中的能力因素与受教育程度产生关联，即Cov(教育, 能力)≠0。由此，我们可以得出*E*(能力|教育=6)≠*E*(能力|教育=8)≠*E*(能力|教育=10)，而这违反“对所有*X*1, *X*2, ..., *Xn*，都有*E*(*ϵi*|*Xi*)=0”这一假定。

5. 同方差性（homoskedasticity）

我们下面开始讨论如何计算估计量的方差，即*V*(*β*)。我们回顾公式：

我们知道计算的方差可能与误差*ϵ*有关。在假设1~4下，*V*(*β*)的方差能够计算出来，不过为了简化方差计算，我们要增加一个关于误差*ϵ*和自变量*X*的假设，即同方差假设。

同方差性指的是，给定自变量的任何值，误差的方差为常数且不随自变量变化而改变，即

由于误差的方差本身不取决于自变量，所以

（三）最小二乘法的无偏性

注意到，如果满足上述假设1~4，则可以得出最小二乘法是一种无偏的估计量，即

对这一点的证明可以经由迭代期望定律得出，此处略。**迭代期望定律（law of iterated expectation）**表明，对于任意两个随机变量*X*和*Y*，以下等式成立：

在给定*X*的情况下，内部期望值在*Y*上取平均值，产生关于*X*的函数，而外部期望值在*X*上将得到的条件期望函数取平均值。

最小二乘法的无偏性的一个应用是二分自变量。假设我们有自变量为二分变量（*X*=1或*X*=0）的回归模型，并且满足最小二乘假设1~4，即首先有

在这种情况下，我们用最小二乘法估计的是总体参数*β*的无偏估计量。同时也表示*X*对*Y*的**因果效应**。

如果只满足假设1~3，即无法满足零条件均值时，则偏误就会产生，是一个有偏的估计量。

（四）最小二乘法的方差

假设1~5被称为高斯-马尔科夫定理（Gauss-Markov assumptions）。一个违反假设5的情况是异方差，即

在假设1~5成立的情况下，有

其中，。

在实际应用中，绝大多数情况下误差*ϵ*都是未知的。在这种情况下我们可以用样本中的残差的来估计误差*ϵ*的方差。用残差估计误差的方差的表达式可以写成

其中*n*-2为一元回归模型的自由度。

在假设1~5成立的情况下，我们可以用残差估计误差的方差，同时估计的方差：

同时通过开根号我们也可以获得的标准误：

二、斜率系数的估计与检验

（一）斜率系数的假设检验

在得出的标准误之后，我们可以对其进行假设检验。研究者通常将零假设和备选假设分别设置为：

设定检验水平*α*。

推荐使用*t*统计量，由自由度为*n*-2的学生t分布来计算：

根据检验水平和自由度选取相应的临界值*t*：

* 当|*T*|>*t*时，拒绝*H*0；
* 当|*T*|≤*t*时，不拒绝*H*0。

同时我们也可以计算*p*值，再比较*p*值与检验水平。

* *p*值<*α*值，拒绝*H*0；
* *p*值≥*α*值，不拒绝*H*0。

（二）斜率系数的区间估计

在得出的标准误之后，我们也可以构建斜率系数的置信区间：

其中为标准误，为临界值。若置信区间不包含*β*0（也就是0），则拒绝*H*0。

三、小结

总而言之，满足不同的假设，我们可以进行相应的统计推断：

* **描述性推断**
  + 简单描述样本中*X*与*Y* 之间的关系；
  + 不对总体参数进行估计；
  + 只需满足假设3（自变量的样本波动）。
* **预测推断**
  + 考察问题：观测到一个新的自变量数值，因变量的预期值是多少；
  + 需满足假设2（随机抽样）和假设3；最好能满足假设1（线性于参数）。
* **因果推断**
  + 对反事实（couterfactual）的推断
    - 考察问题例如：在其他条件不变的情况下，如果我们不增加研发投入，公司的销售额会不会停止增长？如果工会密度不再增长，该国还会提高社会支出吗？
  + 需满足假设1~4（零条件均值）；
  + 特别需要考虑的是自变量*X*会与误差*ϵ*中的哪些因素产生关联。
* **因果推断+假设检验**
  + 需满足假设1~4；最好能满足假设5（同方差性）；
  + 如果出现异方差的情况，也有多种方法能够计算出标准误。这些方法被称为异方差稳健标准误（heteroskedasticity-robust standard errors）。

第十一讲 含有两个自变量的回归模型，遗漏变量偏误及共线性

2024.12.6 / 2024.12.13

一、多元回归模型的动机与意义

多元回归模型让回归模型中包含更多变量从而总结归纳出更多描述性的信息，增强拟合度以及提高模型的预测能力，控制混淆变量（confounding variable）从而进行因果推断。

多元回归模型加入了更多的非线性因素，如

多元回归模型也加入了交互效应，如

二、含有一个二元协变量的多元回归

为了理解二元回归，我们举社会支出与工会密度为例。我们感兴趣的变量有：

* *Y*：社会支出占GDP的比例；
* *X*1：工会密度（工会会员占劳动人口的百分比）；
* *X*2：妇女劳动参与度（1=高参与度；0=中等或低参与度）或者失业率（连续变量）。

一元回归解释的问题是：一个国家的工会密度能够很好地预测这个国家的社会支出情况吗？

多元回归解释的问题是：当我们将妇女劳动参与程度或是失业率作为控制变量，工会密度还能够很好地预测国家的社会支出情况吗？

（一）对回归斜率的解释

1. 对一元回归斜率的解释

我们回顾对一元回归斜率的解释。我们只做社会支出关于工会密度的回归：

对斜率的解释是：工会密度增长一个单位（百分点）与社会支出增长0.0971个百分点相关。

2. 对二元回归斜率的解释

我们可以让回归模型包含更多信息，例如包含妇女劳动参与程度这个二元变量，对此就可能产生两种解释：

* 妇女劳动参与程度高的国家社会支出可能也高；
* 妇女劳动参与程度较低的国家社会支出可能较低。

（二）二元协变量（covariate）

1. 加入二元协变量

我们将一个协变量加入到我们的预测公式：

以上公式表示在函数形式为线性的假设下，我们希望利用变量*X*1和*X*2所包含的信息去预测因变量*Y*。

注意，在上式中，我们把自变量写成*Xij*形式：*i*=1, ..., *n*表示第*i*个观测值；*j*=1, ..., *p*表示第*j*个自变量。为简洁，我们通常简写为*Xj*。*X*1以外的自变量叫协变量。

同时我们还可以用变量的具体名称来写出预测公式：

2. 解释二元协变量

下面我们尝试解释二元协变量。如果我们将*X*1（工会密度）保持不变，将*X*2=1代入回归模型，则：

将*X*2=0代入回归模型，则：

我们得到了两条斜率相同但是截距不同的预测线。截距差表示的是在工会密度保持不变的情况下妇女劳动参与率高的国家和妇女劳动参与率低的国家之间在社会支出方面的预计差异。这意味着，在工会密度保持不变的情况下妇女劳动参与率高的国家比妇女劳动参与率低的国家之间在社会支出方面预计平均高出个百分点。

三、含有一个连续协变量的多元回归

我们将一个连续协变量加入到我们的预测公式：

* *Y*：社会支出占GDP的比例；
* *X*1：工会密度（工会会员占劳动人口的百分比）；
* *X*2：失业率（连续变量）。

则回归公式可以写成：

这可被解释为：在相同失业率的情况下，工会密度增长一个单位（百分点）与社会支出增长个百分点相关；在工会密度保持不变的情况下，失业率增长一个单位（百分点）与社会支出增长个百分点相关。

我们也可以同时代入数值*X*1和*X*2来做更复杂的预测。比如说我们想知道荷兰和瑞典在2004年社会支出的预测差距，我们将荷兰与瑞典对应的工会密度与失业率代入，分别得到预测结果。

四、多元回归的机制

（一）预测值与残差

在多元回归中，预测值和残差的定义和简单线性回归类似。预测（拟合）值是：

其中*i*=1, ..., *n*。

残差是：

其中*i*=1, ..., *n*。有两个自变量回归模型在三维空间中呈现出截面的形状，观测结果与回归截面的距离表示残差。

（二）二元（多元）回归中的最小二乘法

1. 使用最小二乘法最小化残差平方和

在简单线性回归模型中，我们运用了最小二乘法，选择来最小化残差平方和。在二元回归模型中（两个自变量*Xi*和*Zi*），我们用类似的方法，选取使得残差平方和最小化。

推导过程略。最后我们用最小二乘法估计出的可以写成

其中，*SX*2和*SZ*2为样本方差（即*V*(*X*)和*V*(*Z*)）。识别必须满足

违反这一条件的情况有：

* *X*或*Z*是常数；
* 任意一个自变量是另外一个自变量的线性函数。

2. 二元（多元）回归分析的性质

根据上述推导，二元（多元）回归分析有如下性质：

第一，残差的样本平均值为零，即

第二，观测结果的平均值等于预测结果的平均值，即

第三，一定在回归截面上。

3. 对多元回归“排除其他变量影响”的解释

假设我们得到回归公式，则最小二乘估计量的另一种表达形式是

其中，是利用样本将*X*对*Z*进行回归而得到的残差，即

以下两个回归公式可以得出相同的：

残差是*X*与*Z*不相关的部分，即是排除*Z*对*X*的影响之后能够解释*X*的部分。于是，代表在排除*Z*对*X*的影响之后，*Y*和*X*之间的样本关系。

五、多元回归的模型拟合

我们首先回顾一元回归的相关概念——在多元回归中，这些概念仍适用：

第一，总平方和、解释平方和、残差平方和。

第二，决定系数*R*2。

增加自变量的个数，*R*2会增大。当存在多个自变量时，我们会根据自由度*n*-*p*-1（*p*=自变量个数）来计算经调整后的*R*2（adjusted *R*2）。

六、多元回归假设

多元回归假设与一元回归假设类似：

1. **线性于参数：**总体回归模型中的参数是线性的；
2. **随机抽样：**观测到的数据是从总体中通过随机抽样得到的；
3. **不存在完全共线性：**没有一个自变量是常数，自变量之间也不存在严格的线性关系；
4. **零条件均值：**给定自变量的任何值，误差项的期望值为零；
5. **同方差性：**给定自变量的任何值，误差都具有相同的方差；
6. **正态性：**误差项和自变量相互独立且服从正态分布。

与一元回归类似地，满足一定的假设后，就可以进行不同种类的统计推断。

在上述假设中，假设3“不存在完全共线性”与一元回归不同，下面对此进行解释。不存在完全共线性意味着在样本中没有一个自变量是常数，自变量之间也不存在严格的线性关系。符合这一假设的例子有：

违反这一假设的例子有：

七、遗漏变量偏误

（一）遗漏变量偏误的概念与案例

1. 设定不足模型问题

在满足假设1至假设4的情况下，下列对总体参数*βj*的任意值都成立：

证明略。

假设我们的总体回归模型满足OLS假设1至4：

以及样本回归模型：

假设我们关注*X*1对*Y*的偏效应*β*1，由于数据不足无法测量*X*2，我们得出一个只包含*X*1的设定不足（underspecified）模型：

则此处存在一个问题：还是的无偏估计量吗？为了回答这个问题，我们引入两个案例。

2. 案例分析

考虑以下总体回归模型：

假设我们无法得到教育水平的数据，模型中只包含看新闻的频次：

则此时与的大小关系是什么？实际上，教育水平和看新闻频次和普通话水平呈正相关，高估了看新闻对普通话水平的效应，因此。

再考虑以下总体回归模型：

假设我们无法得到当地政府合法性的数据，模型中只包含看联合国是否派驻维和部队这一自变量：

则此时与的大小关系又是什么？实际上，是否派驻维和部队与当地政府的合法性呈负相关，当地政府的合法性和当地的社会秩序呈正相关，低估了维和部队维持当地社会秩序的作用，因此。

3. 遗漏变量偏误的概念

由上述案例，我们总结出一个一般性情况。假设我们的总体回归模型满足OLS假设1至4：

则只包含*X*1的设定不足模型是：

第十讲中，我们提到了一元回归和二元回归斜率系数之间的关系：

其中

* 是*X*2对*X*1做回归所得的斜率。；
* 表示的是在控制*X*1的前提下，*Y*与*X*2之间的关系；
* 或。

我们还可以得出遗漏变量偏误的表达方式（过程略）：

由此，我们可以看出偏误的方向取决于和的符号。任何没有被加入到模型中的，并且同时与自变量*X*以及因变量*Y*产生关联的变量叫做**遗漏变量（omitted variable）**或**混淆变量（confounding variable）**。下表总结了遗漏变量偏误的方向：

|  |  |  |
| --- | --- | --- |
|  | **Cor(*x*1, *x*2)>0** | **Cor(*x*1, *x*2)<0** |
| ***β*2>0** | 偏误为正 | 偏误为负 |
| ***β*2<0** | 偏误为负 | 偏误为正 |

其中偏误为正代表高估了*X*1对*Y*的影响；偏误为负代表低估了*X*1对*Y*的影响。如果回归模型未能把重要的混淆变量包括在内，那么模型对斜率参数的估计就是有偏且不一致的。

（二）加入无关变量的情况

假设我们的总体回归模型满足OLS假设1至4：

我们多加入了一个*X*2：

并且*X*2和*Y*不相关，即。我们知道，如果，那么，仍然是一个无偏的估计量。因此，在回归模型加入一个无关变量不会导致偏误，但会使斜率参数估计量的方差增大。

八、多元回归的方差和多重共线性问题

（一）多元回归的方差

首先我们回顾最小二乘法的第五个假设：给定自变量的任何值，误差都具有相同的方差，即

在满足假设1至5的前提下,以自变量的样本为条件，估计量的方差为

其中，为变量*Xj*中的总平方和（样本总波动），*Rj*2则是将*Xj*对截距项以及所有其他自变量进行回归所得到的*R*2。

下面我们从计算标准误的角度推导上式。如同一元回归，我们可以用样本回归中得到的残差来估计误差的方差*σϵ*2：

通常被叫做回归模型的残差标准误。*n*−*p*−1代表自由度。其中*n*为样本容量，*p*为自变量个数。

我们现在可以估计出斜率参数的标准误：

将此式两边平方，即可得到斜率系数方差：

我们可以看到方差由三个部分组成：

* *σϵ*2是误差的方差。误差的方差越大，的方差越大；
* TSS*j*为总平方和，描述的是自变量*Xj*的总样本波动。当样本容量增大或自变量*Xj*的方差增大的时候，的方差会减小；
* *Rj*2描述了自变量间的线性关系，*Rj*2越大，的方差越大。

（二）多重共线性问题

1. 多重共线性的概念

多重共线性指的是两个或多个自变量之间的相关性过强而导致*Rj*2增大，OLS斜率估计值的方差过大的情况。完全共线性违反了假设3，同时在自变量完全共线的情况下*Rj*2=1。

注意到斜率系数方差分子中的(1- *Rj*2)。当*Rj*2接近于1的时候，我们仍可以用最小二乘法估计*βj*，但是斜率估计值的方差和标准误会非常大，导致与0没有显著差别。

2. 如何发现多重共线性问题

多重共线性不会导致偏误，但是较严重的共线性会使估计量变得不精确。在拟合回归模型之前，可以计算Cor(*X*1, *X*2)，如果二者的相关性系数非常高，则可能有共线性的危险。

拟合回归模型之后，可以考察回归模型是否存在以下问题：

* 加入或删去一个协变量之后其他变量上的系数有很大的改变；
* 回归结果不显著但是*Rj*2仍然很高；
* 回归系数上的符号和自己期望的符号相反。

3. 如何解决多重共线性问题

我们可以通过扩大样本容量、删去或者合并共线性较强的自变量、考虑加入非线性元素等方法解决多重共线性问题。

第十二讲 含有定性信息的多元回归分析

2024.12.13 / 2024.12.20

一、单类别虚拟自变量

（一）虚拟变量的概念

**虚拟变量（dummy variable, 二值/二元变量binary variable，指示变量indicator variable）**通常以二值的形式出现，例如一个人是男还是女，一名学生是否骑车上学，一个人是否落户上海等。

在回归分析中，我们通常运用虚拟变量来包含定性信息，从而比较样本中不同组别和类别的差异，例如不同区域间的差异。回归模型中的虚拟变量能让我们获取每个组别中结果变量的条件均值，例如：不同地域居民的平均收入是否存在差距？本科学历及以上的人的平均工资是否高于没有本科学历的人？

虚拟变量也可以作为**交互项（interaction term）**来检验效应的异质性，例如：一项政策干预的效果是否在老年人群体中更大？

（二）虚拟变量的应用

为了考察虚拟变量的应用，我们关注一个案例——党员和非党员在性别态度上的差异（例如是否同意“男人以事业为重，女人以家庭为重”）。

受访者*i*的政治面貌可分为党员和非党员（包括民主党派、团员、无党派人士和群众），即

我们可以用一个虚拟变量来表示受访者的政治面貌：

因变量*Y*为是否同意“男人以事业为重，女人以家庭为重”这个观点（1=完全不同意；5=完全同意）。我们将因变量*Y*对自变量*X*进行回归：

预测回归公式为。预测回归公式代表了党员和非党员之间在性别态度方面期望平均值的差异，即

期望均差为

当我们加入更多的协变量的时候，虚拟变量上面的系数代表着该虚拟自变量对因变量的偏效应。

二、多类别虚拟自变量

为了考察多类别虚拟自变量的应用，我们同样关注一个案例——中国不同地区之间城市居民幸福感是否存在差异。中国有七大地理区域（*m*=7）：

我们通常需要将*m*-1个，即6个虚拟变量加入到回归模型中进行分析，每一个虚拟变量对应一个类别。将全部类别做成虚拟变量加入到回归公式会导致完全共线性，是所谓虚拟变量陷阱（dummy variable trap）。有意空缺的类别叫做基准组（base/bench-mark group）。

我们设立的6个虚拟变量如下表：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **类别** | ***D*1** | ***D*2** | ***D*3** | ***D*4** | ***D*5** | ***D*6** |
| **华东** | 1 | 0 | 0 | 0 | 0 | 0 |
| **华北** | 0 | 1 | 0 | 0 | 0 | 0 |
| **东北** | 0 | 0 | 1 | 0 | 0 | 0 |
| **华中** | 0 | 0 | 0 | 1 | 0 | 0 |
| **华南** | 0 | 0 | 0 | 0 | 1 | 0 |
| **西南** | 0 | 0 | 0 | 0 | 0 | 1 |
| **西北** | 0 | 0 | 0 | 0 | 0 | 0 |

其中，西北是基准组。因变量为居民的主观幸福感（1=非常不幸福；5=非常幸福），我们得到回归模型

三、交互项与交互效应

在回归分析中，我们可以用交互项进行异质性检验（例如：教育的回报〔*X*1〕效应是否取决于性别〔*X*2〕？）；同时，交互项也可以使我们感兴趣的自变量上的系数随着不同的组别而改变，从而进一步比较不同组别之间的差异。

我们可以将以下变量组合作为交互项进行分析：两个或多个虚拟变量；虚拟变量和连续变量；两个或多个连续变量。

（一）含有两个虚拟变量的交互效应

为了理解含有两个虚拟变量的交互效应，我们延续上文“党员和非党员在性别态度上的差异”的案例。现在我们有两个虚拟变量和一个交互项：

我们仍然将性别态度作为因变量*Y*对自变量*X*进行回归：

通过R，我们可以求得约为-0.30051。交互项上的系数代表了党员效应在男、女不同组别中的差异，因此这一系数的实际意义是：党员对性别态度的效应在女性组中比男性组中大。

由上述案例，我们得出一般结论。一个具有交互项的线性回归模型示例是：

该模型假设*X*1的影响线性取决于*X*2。也就是说，当我们将*X*2增加一个单位时，与*X*1增加一个单位相关的平均结果的变化上升了*β*3。

需要注意的是，该模型中不能将*β*1或*β*2不能直接解释成偏效应。应求出偏导数，即：

（二）含有一个虚拟变量和一个连续变量的交互效应

为了理解含有一个虚拟变量和一个连续变量的交互效应，我们现在关注受教育程度（是否拥有本科及以上学历）对性别态度的效应是否根据年龄的变化而变化。我们有一个虚拟变量和一个连续变量组成的交互项：

我们仍然将性别态度作为因变量*Y*对自变量*X*进行回归：

其中表示的是年龄相差一年的两组之间是否拥有本科学历对性别态度的平均效应的估计差异。

通过R，我们可以求得约为-0.004882，即年龄相差一年的两组受访者之间本科学历对性别态度的平均效应差异为-0.004882；这意味着，本科学历对性别态度的效应在较年长的受访者中更大。

（三）含有两个连续变量的交互效应

为了理解含有两个连续变量的交互效应，我们现在关注受教育程度（受教育年数）对受访者幸福感的效应是否根据年龄的变化而变化。我们有两个连续变量组成的交互项：

我们以幸福感作为因变量*Y*对自变量*X*进行回归：

其中表示的是年龄相差一年的两组之间教育年数增长一年对幸福感的平均效应的估计差异。

通过R，我们可以求得约为-0.000579，即年龄相差一年的两组受访者之间教育年数增长一年对幸福感的平均效应的差异为-0.000579；这意味着受教育程度对幸福感的效应在较年长的受访者中更小。

四、二值因变量：线性概率模型

（一）线性概率模型的概念

在社会科学研究中我们也会遇到因变量为二值变量的情况，即

二值因变量的例子包括是否在选举中投票、是否接种疫苗、两国是否为军事同盟、两国是否处于战争状态等。为解释一个事件的二值结果，我们可以用以下回归模型：

当*Y*是一个二值变量时：

成功（*Y*=1）的概率等于*Y*的期望值，所以我们得到

其中，*P*(*Y*=1|*X*1 … *Xp*)被称为响应概率（response probability）；由于其为*Xj*的一个线性函数，所以此模型又被称为**线性概率模型（linear probability model, LPM）**。其中，*βj*表示的是*Xj*变化之后，成功概率相应的变化（不是“在其他条件不变的情况下，*Xj*提高一个单位，*Y*的变化量”）。

我们也可以用样本拟合出以下回归模型：

其中，表示在每个*Xj*等于0时预期成功的平均概率；斜率度量的是当*Xj*提高一个单位时，*P*(*Y*=1|*X*1 … *Xp*)的预期变化。

（二）线性概率模型的案例

下面，我们举出一个线性概率模型的案例。我们关注影响城市居民接种新冠疫苗的因素，即：

此模型包含下述自变量：

* Age：受访者年龄；
* Female：受访者性别（1=女性；0=男性）；
* Party\_Member：是否为党员；
* Education：接受教育年数。

这意味着：保持方程中其他因素不变，每年长一岁，预计接种新冠疫苗的概率就会改变；保持方程中其他因素不变，女性平均接种新冠疫苗的概率与男性平均接种新冠疫苗的概率的差值为；其他系数以此类推。

（三）线性概率模型的局限性与预测

需要注意的是，线性概率模型具有一些局限性。例如，将某些特定的组合数值代入预测公式后，可能会得到大于1或小于0的预测值；此外，因变量可能与自变量呈非线性的关系。

尽管线性概率模型的预测值有时会超出[0, 1]，我们仍可以利用一些方法来预测二值[0, 1]结果。例如，若为回归模型的拟合值，且个别拟合值可能在[0, 1]范围之外，那么我们可以定义一个新的预测值：

可见，的取值不是0就是1。我们再将和观测值*Yi*进行比较，可以得出正确预测*Yi*=1和*Yi*=0的频率以及全部预测正确的百分比，是为全部正确预测百分比（percent correctly predicted）。