清华大学文科出版基金
QINGHUADAXUEWENKECHUBANJIJIN

国家自然科学基金（项目号71874094）

实验与准实验设计在语言教育研究中的应用

Application of Experimental and Quasi-Experimental Designs in Language Education Research

郭　茜　冯瑞玲　著

清華大學出版社
北京

图书在版编目 (CIP) 数据

实验与准实验设计在语言教育研究中的应用 / 郭茜, 冯瑞玲著.—北京：清华大学出版社，2022.8

ISBN 978-7-302-61619-1

Ⅰ. ①实… Ⅱ. ①郭… ②冯… Ⅲ. ①语言教学－实验研究 Ⅳ. ①H09-33

中国版本图书馆 CIP 数据核字（2022）第 147723 号

责任编辑：商成果
封面设计：常雪影
责任校对：欧　洋
责任印制：朱雨萌

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-83470000　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市天利华印刷装订有限公司
经　　销：全国新华书店
开　　本：170mm × 240mm　　**印　张**：13　　**字　数**：211 千字
版　　次：2022 年 8 月第 1 版　　**印　次**：2022 年 8 月第 1 次印刷
定　　价：69.00 元

产品编号：092145-01

序　言

《实验与准实验设计在语言教育研究中的应用》一书即将出版，受作者之邀为本书作序，我由衷地为她们感到高兴。

郭茜教授是清华大学外国语言文学系应用语言学和语言教育方向的负责人，专业背景深厚，学术功底扎实，研究成果在语言教育领域具有前沿性和引领性，是一位学术水平高、研究能力强、教学成效突出并为清华大学外国语言文学学科建设发挥了重要骨干作用的学者。我与郭茜教授自她在清华大学攻读硕士学位时就相识。郭茜教授多年来一直从事语言教学工作，在语言教育研究领域也深耕多年，先后主持国家社会科学基金、国家自然科学基金项目以及多项校级研究项目，研究成果在语言教育领域具有很强的理论和应用价值。郭茜教授在哈佛大学攻读博士学位期间，受到了良好的量化科研方法训练，回国工作之后的科研成果也多基于实验和准实验研究设计，为本书的成文打下了坚实的基础。此外，郭茜教授开设了语言教育研究设计方面的研究生课程，在授课过程中积累了丰富的经验，了解语言教育领域的研究生在研究设计和数据分析过程中的常见问题和实际困难。这有助于她在著书过程中充分考虑读者的需求，使本书的内容设计和问题探讨更具针对性。

第二位作者冯瑞玲也具有多年的语言教学经验，她对语言教学和第二语言习得研究的热爱让我印象深刻。在成为清华大学外文系郭茜教授的博士研究生之前，她已经开始英语教学的实验与准实验研究，并主持和参与了省部级和国家级相关研究课题，发表了多篇学术研究论文和多部译著。她在博士一年级和二年级时修读了清华大学外文系、教研院和社科学院的多门研究设计和量化分析课程，并将所学运用于研究的开展和成果的撰写，理论学习结合实际研究的经历帮助她更全面、深刻地理解和应用量化统计分析方法。

《实验与准实验设计在语言教育研究中的应用》一书，与已出版的很多研究设计和数据分析类书籍相比，具有非常突出的特点和优势。首先，它从语言教育研究者的视角通俗易懂地讲述如何设计研究、分析数据及解读分析结果，

并充分与语言教育研究实践相结合。其次，它有着鲜明的跨学科视角，这与语言教育研究本身的跨学科色彩一致，相信这本书的出版可以为国内语言教育尤其是外语教学研究领域带来新鲜的研究理念，从而拓宽我们的研究领域和视野。此外，该书前后呼应，前三章专注于量化分析方法，为理解后面章节的研究设计部分打下基础，读者不仅可以学习常用的实验与准实验设计，还能学习实验研究的常用数据分析方法。本书成功入选清华大学文科出版基金资助计划，在语言教育研究领域创新性地使用随机实验与准实验研究设计，开拓了外语教育研究的新范式，既是一本高水平的学术专著，也可成为语言教育研究方法相关课程的教材。相信本书的出版对我国语言教育研究方法的理论探讨和实践创新具有重要的参考价值并将发挥积极的导向作用。

张文霞

2022 年 8 月

前　言

观测数据本身不能支持因果推断，因为相关关系不等于因果关系，不能仅基于观测数据发现两个因素间存在相关关系就认为其中一个因素导致另一个因素，继而希望通过改变前者来影响后者，这样很有可能导致决策失误。决策需要基于因果推断。公认能支持因果推断的研究设计是随机实验和各种准实验。2021 年诺贝尔经济学奖三位得主（David Card、Joshua D. Angrist 和 Guido W. Imbens）的获奖原因就是他们在准实验设计领域做出了突出贡献。国内语言类研究中常见的随机实验和准实验设计较为单一，较少使用教育、经济领域常见的准实验设计，在设计的严谨性上也存在差距，导致一些研究的质量差强人意。哈佛大学 Light et al.（1990）所著 *By Design: Planning Research on Higher Education* 一书中序言的标题就是“You can’t fix by analysis what you bungled by design”（你无法用数据分析来弥补搞砸了的设计），由此可见设计对研究的重要性。

本书第一作者在哈佛大学攻读博士学位时受过系统的定量研究训练，因为就读的教育政策定量分析专业的特殊要求，选修了近十门定量研究方法类课程。自 2012 年回国以来，在解答学生研究方法相关问题以及参加硕士生、博士生毕业论文评审和答辩过程中，发现相当一部分研究的设计不能很好地支持研究结论；有些学生做的统计分析并不能回答自己的研究问题，显示出他们对统计分析缺乏基本的认识；还有的学生虽然采用的是混合式研究方法，但从论文中能明显看出他们接受的定量研究方法训练不足，对辛辛苦苦收集到的大量数据可能只是给出均值、频率等描述统计结果，浪费了宝贵的数据。本书第二作者有十二年语言教学与研究经验，近七年来一直从事国际在线协作学习研究，在与不同国家的高校同人合作教学和科研的过程中，对语言教育研究的跨学科性质和设计严谨性需求深有感触，同时也发现国内语言教育研究领域的部分教师和学者在统计分析方法和研究设计上还有不少困惑，并且对相近学科中已经很成熟的研究方法了解甚少。两位作者希望将相邻领域的研究设计引入语言教育研究中，帮助这些同人熟悉常用的数据分析方法和研究设计，从而扩展研究视野。

基于以上原因，本书面向语言教育领域的教师和研究生，在介绍一些广泛使用的量化统计方法的基础上，讨论本领域可以使用的实验与准实验研究设计（“实验”和“准实验”连用时，前者特指“随机实验”，这是相关文献中较为常见的用法）。本书既包含对常用统计分析方法的入门级讲解，帮助不熟悉这部分内容的读者快速学习掌握基本方法，同时依托巧妙且有效地利用随机实验和各种准实验研究设计的优秀研究实例，向读者展示各类研究设计适用的情形、使用时需要注意的事项、完善研究设计的策略。我们尽可能从语言教育领域遴选研究案例，但由于实验与准实验设计在教育、经济领域运用比较广，而在语言教育领域中的使用起步较晚，所以有的准实验设计在语言教育领域很难找到相关研究，只能从更广泛的教育领域选取。譬如工具变量是教育领域很常见的一种准实验设计，但现有语言教育研究中使用的工具变量只有一两种，常见的三类工具变量都没能在语言教育研究领域中找到相关研究。此外，尽管也有语言教育研究实行随机分组实验，但随机分组过程非常规范的研究往往来自教育和经济领域，语言教育研究领域中则比较缺乏。语言教育作为交叉学科，一直从教育、经济领域借鉴研究方法，我们希望读者对经典、规范的随机实验和准实验研究设计能有所了解，以期帮助他们拓宽研究视野，提升跨学科研究能力，所以书中有时会借用一些教育类论文作为研究案例。

本书总体分为两部分。第一部分由前三章的统计基础知识组成，介绍多种回归模型；第二部分（第四章至第九章）则在阐明因果推断重要性的基础上介绍随机实验和各种准实验研究设计。因为回归分析是随机实验和准实验研究中最常用的数据分析方法，也是本书第二部分中研究实例通常采用的分析方法，所以前三章可以为不熟悉回归分析方法的读者阅读后面的章节打下基础。具体到各章，第一章介绍一个预测变量和一个结果变量的简单线性回归，并介绍预测变量为虚拟变量和类别变量时的处理方法；第二章在介绍遗漏变量偏误的基础上引入多元线性回归；第三章介绍结果变量为虚拟变量和类别变量的回归分析。这三章也会将回归分析与相关分析、方差分析、卡方检验等统计方法进行比较，以帮助熟悉后面这类统计方法的读者事半功倍地学习回归分析。在介绍了这些统计分析方法后，第四章首先厘清相关关系与因果关系的差异并引入随机实验和准实验的概念；第五章介绍随机实验；第六章介绍自然实验和双重差分法；第七章介绍断点回归；第八章介绍工具变量；最后一章介绍语言教

育研究领域中其他常见的准实验研究设计，包括在实验组和对照组不完全可比时能用来减少干预效应估算偏误的倾向得分匹配法。

第一章到第三章每章结尾处有回归分析习题，供大家检查自己的掌握情况；第四章到第九章每章结尾处有练习，引导读者深入思考该章的核心内容，分析已发表论文的研究设计并使用相应的方法设计自己的研究。除此以外，每章的最后都有“进深资源推荐”，供有兴趣了解更多相关内容的读者参考。

考虑到语言教育领域的不少学者对统计学和概率论知识掌握不多，而且本书的学习目标不是深奥的数理统计原理，因此没有加入大量复杂的数学公式。本书不要求读者有定量研究方法的基础，重点也不在推导和估算过程，而在相关统计方法的精髓，聚焦科学研究实务，写作风格对于语言教育研究领域的读者而言非常友好。相信本书能帮助语言教育领域从事或意欲从事定量研究和混合式研究的师生了解随机实验与多种准实验的优缺点、研究设计及数据分析结果的解读，也能帮助他们提升研究敏感性，在适于因果推断的研究机会出现时，能及时抓住机会，有的放矢地提出研究问题，并进而设计和实施实验或准实验研究，推动他们的研究迈上一个新台阶。

由于作者水平有限，书中难免存在疏漏之处，敬请同行批评指正！

作　者

2022 年 2 月

于清华园

目　录

第一章　简单线性回归 1
第一节　基本概念 1
第二节　简单线性回归模型和普通最小二乘法 5
第三节　简单线性回归前提假设 6
第四节　分类预测变量设置 8
第五节　回归方程检验 9
第六节　简单线性回归系数解读 12
第七节　简单线性回归与相似分析方法 14
第八节　结语 15

第二章　多元线性回归 18
第一节　基本概念 18
第二节　多元线性回归模型 20
第三节　多元线性回归注意事项 21
第四节　预测变量筛选 22
第五节　回归方程检验 25
第六节　交互项 27
第七节　多元线性回归系数解读 30
第八节　多元线性回归与多因素方差分析的异同 38
第九节　结语 39

第三章　logistic 回归 41
第一节　logistic 回归基本概念 41
第二节　logistic 回归类型 43
第三节　预测变量设置 49
第四节　logistic 回归系数解读 51

第五节　logistic 回归模型评价与适用条件 56
第六节　logistic 回归与相似检验方法的比较 57
第七节　结语 60

第四章　因果推断 64
第一节　相关关系与因果关系 64
第二节　区分相关关系和因果关系的重要性 66
第三节　基于随机实验的因果推断 68
第四节　基于准实验的因果推断 69
第五节　结语 71

第五章　随机实验 73
第一节　随机实验流程 73
第二节　常见随机实验设计 77
第三节　威胁随机实验效度的因素 85
第四节　统计功效 92
第五节　干预效应估算方法 94
第六节　样本量预估方法 96
第七节　结语 100
附录　随机抽取 101

第六章　自然实验 108
第一节　基本概念 108
第二节　双重差分法估算 109
第三节　双重差分法的关键假设及假设检验 115
第四节　双重差分法拓展 117
第五节　自然实验来源 120
第六节　自然实验数据来源和选择 121
第七节　自然实验与语言教育研究 123
第八节　结语 126

第七章　断点回归 128
第一节　基本概念 128
第二节　常见断点回归设计 131
第三节　断点回归设计的内部效度 135
第四节　干预效应估算方法 139
第五节　断点回归设计的优缺点 146
第六节　结语 147

第八章　工具变量 149
第一节　基本概念 149
第二节　工具变量估算的关键假设及假设检验 150
第三节　常见工具变量 157
第四节　工具变量与其他研究设计的结合使用 162
第五节　干预效应估算方法 165
第六节　结语 168

第九章　语言教育领域中其他常见准实验设计 170
第一节　干预前的平衡性检验 170
第二节　倾向得分匹配 175
第三节　扩充对照组的范围 177
第四节　结语 180

参考文献 181

后记 194

第一章　简单线性回归

回归分析是随机实验和准实验研究中最常用的数据分析方法，掌握常见回归分析的基础知识有助于更好地理解随机实验和准实验研究设计及其效应估算方法。为了帮助读者更好地理解本书第二部分的章节内容（随机实验和各种准实验设计），本书第一部分（前三章）主要介绍多种回归分析的基础知识，其中线性回归分析是经典统计中最常用的量化分析方法之一。

所谓线性回归，是指确定结果变量与预测变量之间为直线型关系的一种统计分析方法。线性回归分析要求结果变量必须是连续数值变量，但对预测变量的测量尺度没有要求。它探究一个或多个预测变量与结果变量之间的相关关系，帮助我们用一个变量（预测变量）的值估计或预测另一个变量（结果变量）的值。比如英语学习时长与英语成绩相关，我们可以用英语学习时长估计或预测英语成绩，这种只有一个预测变量的线性回归被称为简单线性回归，也叫一元线性回归，是线性回归最基本的形式。本章介绍简单线性回归的一些基本概念、回归模型、回归估计方法、前提假设、二分类变量和多分类变量做预测变量时的处理办法、回归方程检验、回归分析结果中的系数解读、简单线性回归与相关分析和单因素方差分析（one-way analysis of variance，one-way ANOVA）的联系与区别等内容。学习本章内容可为第二章多元线性回归和第三章 logistic 回归的学习打基础、做铺垫。

第一节　基 本 概 念

变量类型

变量类型影响回归模型的建立，所以有必要先对变量分类进行描述。按照测量尺度，变量可以分为四种。

（1）定类变量，也叫名义变量、称名变量（nominal variable），是描述事物特性的变量，将事物区分为互斥的不同类别，如季节、颜色等，在进行数据分

析前通常要将这类变量进行重新编码，用数字给每个类别进行赋值，进行量化处理。定类变量还可以进一步分为二分类变量（binary variable 或 dichotomous variable）和多分类变量（multinomial variable），二分类变量取值仅为 0 或 1 时，也称为虚拟变量或哑变量（dummy variable）。如 *female*（女性取值为 1，男性取值为 0）是虚拟变量，但性别（取值为“男”或“女”）不是虚拟变量。

（2）定序变量（ordinal variable）是描述事物等级的变量（如受教育程度、英语水平），可以排序比较大小（如中学教育水平高于小学教育水平），但变量的两个取值之间的差没有意义（例如不能将中学教育水平与小学教育水平间的差和大学教育水平与中学教育水平间的差进行比较），因此不能进行加减等算术运算。

（3）定距变量（interval variable）的两个取值之差有意义（如 1℃和 3℃间的差与 2℃和 4℃间的差大小相等），但这种变量没有绝对零点（如 0℃不代表没有温度），因此可以进行加减运算，但不能进行乘除运算。如摄氏和华氏温度、智商都是定距变量。

（4）定比变量（ratio variable）和定距变量一样，两个数值之差有意义；除此以外，它还有真正的零点（如人数为 0 表示一个人都没有），因此两个数值之间的比值也有意义（如 8 公斤是 4 公斤的两倍重）。定比变量不仅可以进行加减运算，还可以进行乘除运算。这种变量包含的数据信息最多，测量尺度最高。

这四种变量的测量尺度依次增高。测量尺度高的变量可以转换为测量尺度低的变量，这时会损失掉一些信息，反之则不行。定类变量和定序变量都是分类变量（categorical variable，也称类别变量），前者是无序分类变量（unordered categorical variable，指类别之间无程度和顺序差别的分类变量），后者是有序分类变量（ordinal categorical variable，指各类别之间有程度差别的分类变量）。定距变量和定比变量都是数值型变量（metric variable），实际研究中，多数情况下我们对定距变量和定比变量不做区分，本书中的连续型变量（continuous variable）即包括这两类变量。

按照变量在回归模型中的角色，可分为以下几种。

（1）结果变量（outcome variable），即被估计或被预测的变量，也称为因变量（dependent variable）、响应变量（response variable）、被解释变量（explained variable）、回归应变量（regressed variable）。

（2）预测变量（predictor），也称为自变量（independent variable）、解释变量（explanatory variable）、输入变量（input variable）、回归量（regressor），一般是被研究者操纵的变量。

（3）多元线性回归中的预测变量又可以根据不同的角色分为主要预测变量、控制变量/协变量、调节变量、中介变量等。这部分在本书第二章中介绍。

标准差、方差和标准误

在量化数据分析结果汇报中，我们经常看到标准差（standard deviation，SD）、方差（variance）和标准误（standard error，SE）。很多人不清楚三者的区别与联系，这里一起加以说明。

标准差常用 σ 表示，用来衡量一个样本的离散程度，常用于描述统计。标准差越大，说明这组样本数据离散程度越大，集中程度越小。比如考察两个班的英语成绩时对比它们的标准差，标准差较小的班级说明这个班学生们的成绩比较接近，大家的分数与班级平均分的差异较小，该班级学生英语成绩的离散程度较小；标准差较大的班级说明这个班学生们的成绩相差比较大，大家的分数与班级平均分差异较大，该班级学生英语成绩的离散程度较大。假设一组数据有 n 个取值，$x_1, x_2, \cdots, x_n$，其均值为 $\bar{x}$，则：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

方差（variance）常用 σ^2 来表示，是标准差的平方，用来描述一组数据与其数学期望值（即均值）的离散程度。

标准误的全称是样本均值标准误（standard error for the sample mean，SE），用来衡量样本均值与总体（population）均值的差异，即抽样误差。标准误越小，样本数据越能代表总体数据，用样本统计量（sample statistic）推断总体参数（population parameter）的可靠性就越大。标准误与标准差的区别在于：对一个总体进行多次抽样，每个样本都有自己的平均值，这些平均值的标准差即标准误，而标准差是针对某一个样本的数据。二者的数学换算关系如下（n 指样本量）：

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

p 值和显著性水平

p 值是指当零假设为真时，得到样本观测结果（如某预测变量的估算系数值）甚至更极端结果的概率。*p* 值用来判断假设检验的结果是否具有统计学意义。假设分为零假设（null hypothesis，H_0；假设干预在总体人群中无效）和备择假设（alternative hypothesis，H_1；假设干预在总体人群中有效），前者一般是研究者想通过数据分析予以推翻的假设（如两组之间的差异为 0 或两组数据的相关系数为 0），备择假设则是与前者相对立的假设（如两组之间的差异不为 0 或两组数据的相关系数不为 0）。在研究中，通常先选定显著性水平（significance level，也称 α level）。显著性水平是指某个预测变量对结果变量实际上没有效应时我们使用的样本会发现显著效应的概率。语言教育研究中通常采用 0.05 的显著性水平，以干预研究为例，这个显著性水平表示假设干预在总体人群中无效（即零假设），那么我们一次抽样只有 5%的概率会碰巧得到干预有效的结果，我们认为这个概率很低（有时候会采用更严苛的标准，如 0.01 或者 0.001），因此认为我们的样本不是出自零假设的样本，所以拒绝零假设，认为干预实际上有效。通常我们将 *p* 值和显著性水平比较，如果显著性水平设为 0.05，$p < 0.05$ 表示如果零假设成立，在一次抽样中出现当前结果以及更极端结果的概率小于 5%。*p* 值比显著性水平低得越多，我们越有信心认为零假设成立时不可能得到当前结果，也就越有信心拒绝零假设。有关显著性水平和统计显著性等概念，可以参看《大数据时代下的统计学（第 2 版）》（杨轶莘，2019）一书相关部分或是节选自该书相关内容的“量化研究方法”公众号文章《“凑巧”可以拒绝吗？统计学的重要工具——假设检验》。

置信区间

置信区间（confidence interval）是通过样本统计量所构建的总体参数的估计区间（即可能范围），表示总体参数的真实值有一定概率落在这个估计区间里。这里的“一定概率”即置信水平，等于 1 减去显著性水平 α，例如显著性水平设为 0.05，则置信水平为 95%，这也是最常见的置信水平。95%置信区间是根

据样本观测值（observations）估计出的数值区间，表示总体参数的真实值有 95% 的可能性落在这个区间内。如果置信区间内包含 0（如区间为–0.03～1.09），则不能排除总体参数为 0 的情况，即不能拒绝零假设；如果置信区间的上下限都大于 0 或都小于 0，则可以在相应的显著性水平上（如 0.05）拒绝零假设，认为总体参数显著不等于 0。

自由度

自由度（degree of freedom）指可自由变动的样本观测值的数目。在样本量 n 和样本均值已知的情况下，样本数据的总和也是确定值。如果已知 n 个观测值的总和及前（$n-1$）个观测值的总和，第 n 个观测值的取值就唯一确定，所以这组数据的自由度为（$n-1$）。比如，已知一个班 20 名学生的英语平均成绩为 76 分，我们可以随意猜测其中 19 名学生的成绩，但一旦确定了这 19 名学生的成绩，最后一名学生的成绩也就确定了，所以这组数据的自由度为 19，即样本量 20 减去 1。

第二节　简单线性回归模型和普通最小二乘法

简单线性回归分析是根据预测变量 x 和结果变量 y 的相关关系，建立 x 与 y 的线性回归方程，并对 y 进行估计或预测。由于语言教育领域研究的现象一般受多种因素影响，简单线性回归应用不多，我们在建立回归模型时一般都会添加控制变量，以期通过统计方法控制干扰因素。但了解简单线性回归模型及其估算原理和系数解读等是学习和应用多元线性回归的基础，所以本章内容非常重要。

假如我们要考察预测变量 x 与结果变量 y 之间的关系，设它们之间的回归模型为：

$$y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i \qquad (1\text{-}1)$$

其中，y 是结果变量；x 是预测变量；α_0 是常数项（constant），也称为截距（intercept）；α_1 是回归系数（coefficient），也是数学意义上的斜率（slope）；ε 是随机误差，也就是俗称的误差项（error term），表示除去预测变量对结果变量影响后的随机误差，即回归方程的预测值与真实值的差距；下标 i 是样本中

每个个体的序号。

通过样本观测值得到方程（1-1）中 α_0 与 α_1 的估计值 $\widehat{\alpha_0}$ 与 $\widehat{\alpha_1}$，那么我们就可以列出经验回归方程如下：

$$\hat{y} = \widehat{\alpha_0} + \widehat{\alpha_1} x \qquad (1\text{-}2)$$

我们也可以称方程（1-2）是理论模型的拟合直线（fitted line）或经验回归直线。简单线性回归主要是利用样本观测值估计未知参数 α_0、α_1 和方差 σ^2，对模型和回归系数做显著性检验，并对回归系数做置信区间估计。

在简单线性回归模型参数估计中，我们通常用普通最小二乘法（ordinary least square，OLS）得到截距 α_0 和斜率 α_1 的估计值 $\widehat{\alpha_0}$ 和 $\widehat{\alpha_1}$。有了 $\widehat{\alpha_0}$ 和 $\widehat{\alpha_1}$，就可以获得每个样本观测值的拟合值 $\hat{y}$，每个 $\hat{y}$ 都在拟合直线上。ε_i 是样本个体 i 的回归拟合值 $\hat{y}_i$ 与观测值 y_i 之间的差异，即残差。ε_i 的期望值（即平均值）为 0，但实际上并非每个观测值的残差都为 0。普通最小二乘法的中心思想是使得观测值与估计值之差的平方和达到最小，“二乘”指取观测值和估计值之差的平方，即残差平方，“最小”指残差的平方和（sum of squares error，SSE）最小，即 $\sum(y_i - \hat{y}_i)^2$ 最小。普通最小二乘法使得回归函数尽可能好地拟合一组观测值，使经验回归直线处于样本数据点的中心位置。为什么用残差平方和而不是残差和呢？因为残差有正数和负数之分，加总后会相互抵消得 0，无法取最小值，取平方则可以解决这个问题。

第三节　简单线性回归前提假设

线性回归是一种参数估计方法，需要设定一些前提假设。如果样本数据不满足这些假设，回归分析的结果就会出现偏差（bias），因此有必要在开始回归分析前验证一下这些假设。简单线性回归的前提假设主要有四个，分别是线性（linear）、独立性（independent）、正态性（normal）、方差齐性（equal variance），可以简单将其记为“LINE”。对这些假设的检验都可以通过统计分析软件来实现。

所谓线性，是指结果变量与预测变量之间存在线性关系，如果二者之间的关系非线性，就不能用线性回归了。线性关系可以通过绘制散点图（scatterplot）来考察结果变量随预测变量变化而变化的情况，如果二者之间很明显没有线性

关系，可以尝试对数据进行转换（如对数转换、平方根转换），或使用其他分析方法（如非线性回归）。

独立性是指结果变量的各观测值之间相互独立，不能相互影响。相应地，残差之间也应相互独立，即不存在自相关性。可以用 Durbin-Watson 检验（DW 检验）检验独立性，若残差之间不相关，说明不存在自相关性，则满足独立性假设。该统计量的值落在（0,4）内，DW = 2 意味着没有自相关性，0 < DW < 2 表明残差间存在正相关，2 < DW < 4 表明残差间存在负相关，一般 DW 值为 1～3 可以接受。独立性在语言教育研究中经常难以实现，比如对同一个群体重复测试（时间序列数据，见第六章），这种情况下的观测值之间可能会相互影响；又比如同一班级内学生的成绩可能相互影响。在数据分析时可以通过聚类调整后的标准误（cluster standard error）处理自相关问题，如统计软件 Stata 中可以使用 cluster 这一回归指令选项。它针对的问题是同一组内个体（如同一班级内学生成绩）的误差项之间存在相关性，即观测值之间不独立。该选项对标准误进行聚类调整，调整后回归系数的标准误会增加，由于通常用回归系数除以其标准误得到的商判断系数的显著性，商越大显著性越高，所以聚类调整可以避免低估标准误，因此得到的回归结果更稳健。例如在各班级内将学生随机分配到实验组和对照组，检验对实验组进行干预的效果可以使用以下指令：

regress grade treatment, cluster(class)

其中，grade 是表示成绩的结果变量；treatment 是代表分组情况的虚拟变量（实验组取值为 1，对照组取值为 0）；class 是代表各班的分类变量；逗号后的 cluster 选项针对班级内学生成绩间可能存在的相关性进行聚类调整。由于使用聚类调整标准误的结果更稳健，所以论文审稿人在观测值之间不独立时会关注回归中是否进行了聚类调整。回归中如果加了 cluster 选项，我们可以在数据分析部分主动报告，如“we clustered standard errors at the class level to account for the potential error correlation within each class”；另外，在用表格展示回归结果时也可以加上“robust standard errors clustered at class level”之类的说明。

正态性严格来讲是指残差应该满足正态分布，但我们可以直接观察数据是否为正态分布，具体可以通过绘制直方图（histogram）或 Q-Q 图（quantile-quantile plot）实现。如果残差满足正态分布，则 Q-Q 图中的散点会近似落在一条直线上；如果不满足正态分布，散点会偏离该直线。如果数据不满足正态性，置信

区间会很不稳定，可以考虑对数据进行转换（如对数转换、平方根转换、倒数转换），或采用非参数估计方法。

方差齐性意为方差相等，即对于每个预测变量取值，对应的结果变量都有相同的方差。满足这一条件的话，我们说模型具有同方差性（homoscedasticity）；若不满足，则为异方差性（heteroscedasticity）。截面数据（见第六章）容易出现异方差性（何晓群，2017）。可以通过绘制残差图来判断方差齐性，即以结果变量的回归估计值为横坐标，以残差为纵坐标作图，如果残差无明显的规律分布，表明方差齐性；如果残差范围有逐渐扩大或缩小的趋势，则可能存在方差不齐的情况。如果数据不满足方差齐性，我们可以对数据进行一些转换。一般情况下，如果异方差性不是很严重，仍然可以采用线性回归。

第四节　分类预测变量设置

前面提到，线性回归模型对预测变量的类型没有要求，分类变量和数值型变量均可纳入回归方程。分类变量若是二分类变量，一般我们会把它转换为取值 0 或 1 的虚拟变量再纳入方程；若是无序多分类变量（如季节），一般需要将其设置为多个虚拟变量再进行回归分析；有序多分类变量（如学历水平）则视情况而定，一般视为分类变量，但当定序变量的取值比较多且间隔比较均匀时，也可以按数值型变量处理。

先来看二分类变量的设置。以性别为例，这是个分类变量，有两个类别，即男和女，我们可以设置 *female* 这个虚拟变量（女性取值为 1，男性取值为 0）取代性别变量，然后把虚拟变量 *female* 纳入回归方程。当然我们也可以采用虚拟变量 *male*（男性取值为 1，女性取值为 0）。对于只有两个类别的变量，设置一个虚拟变量即可，因为一个虚拟变量（如 *female*）足以区分两个类别（如两种性别）。

再来看多分类变量的设置。多分类变量不同水平的变化（如学历水平从小学毕业提升到中学毕业和从中学毕业提升到大学本科毕业）对结果变量的影响可能不同，即预测变量每个单位的变化带来结果变量的变化不是常数，二者之间不是直线关系，因此不能将这类多分类变量直接纳入回归分析，而需要将多分类变量设置为多个虚拟变量。有序多分类变量一般也需要设置多个虚拟变量，

但研究者可以从专业知识出发进行判断，如果该预测变量取值比较多且间隔比较均匀，可以作为数值型变量处理。一般 n 个分类的预测变量需要设置 $n-1$ 个虚拟变量（Harrell，2001）。比如我们把学生的英语水平分为高、中、低三类，那么我们可以用两个虚拟变量来代替学生的英语水平这个多分类变量。

表 1-1 中，变量“高水平”取 1 代表高水平，0 代表非高水平；变量“中水平”取 1 代表中水平，0 代表非中水平；两个变量均取 0 时代表低水平。因此两个虚拟变量即可以表示三分类变量，这两个虚拟变量需要同时纳入或退出回归模型（即同进同出）。如果使用 SPSS 软件进行线性回归分析，需要先将多分类变量转换为多个虚拟变量。如果使用 Stata 软件则只需在分类变量前加“i.”（如将表示英语水平的变量“*Prof*”变成“i.*Prof*”）。

表 1-1　将多分类预测变量设置为多个虚拟变量示例

英语水平	高水平	中水平
高	1	0
中	0	1
低	0	0

分类变量纳入回归分析需要有一个参考对象，即用于做比较的类别，如表 1-1 中的参考类别（omitted category）为“低水平”。实际研究中，研究者可以根据研究需要设置参考类别，但要保证参考类别有实际意义，同时参考类别组的观测值数量不能太少。

第五节　回归方程检验

建立线性回归模型，根据样本观测值得到各个参数的普通最小二乘法估计值后，我们要对回归方程进行必要的检验，以确定模型是否合理、系数是否显著。这里主要介绍回归方程显著性检验（F 检验）和预测变量系数显著性检验（t 检验）及回归模型拟合检验。在简单线性回归中，F 检验和 t 检验是等价的。

F 检验

简单线性回归的 F 检验用来评价预测变量与结果变量间的线性关系是否显

著，即回归方程本身是否有效。由于简单线性回归中只有一个预测变量，所以回归方程显著性等价于预测变量系数的显著性。

我们先来看几个相关概念：

（1）总离差平方和（sum of squares total，SST），反映结果变量的波动大小，等于结果变量的样本观测值与其观测值均值的差值平方和，即

$$\mathrm{SST}=\sum_{i=1}^{n}(y_i-\bar{y})^2$$

（2）回归平方和（sum of squares due to regression，SSR），反映 SST 中能被经验回归方程解释的部分，等于结果变量的估计值与其观测值均值的差值平方和，即

$$\mathrm{SSR}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

（3）残差平方和（sum of squares error，SSE），反映 SST 中不能被经验回归方程解释的部分，等于结果变量观测值与其估计值的差值平方和，即

$$\mathrm{SSE}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

SST 是 SSR 和 SSE 之和，即样本总波动性 = 回归方程可解释的波动性 + 误差造成的波动性。根据这三个参数的定义可知，SSR 越大，回归方程的解释度就越高，由此构造 F 值检验统计量如下：

$$F=\frac{\dfrac{\mathrm{SSR}}{1}}{\dfrac{\mathrm{SSE}}{n-2}}$$

其中，n 是样本量。根据设定的显著性水平 α、自由度（1，$n-2$）查 F 分布表，得到相应的临界值 F_α。若 $F>F_\alpha$，则回归方程具有显著意义，回归结果显著；若 $F<F_\alpha$，则回归方程无显著意义，回归结果不显著。在使用统计软件时，我们无须手动计算 F 统计量，只需要看 F 检验结果中的 p 值是否小于设定的显著性水平 α，如表 1-2 中的 Sig.值。

表 1-2　简单线性回归 F 检验结果示例

模型		平方和	自由度	均方	F	Sig.
1	回归	150.126	1	150.126	19.258	0.000
	残差	600.250	77	7.795		
	总计	750.376	78			

t 检验

回归分析中，t 检验的目的是检验零假设 H_0（即预测变量系数等于 0），以此来检验回归系数的显著性。若不能拒绝零假设，说明预测变量与结果变量之间无显著的线性关系；若能拒绝零假设，则接受备择假设 H_1（即预测变量系数不等于 0），说明预测变量与结果变量之间有显著的线性关系。t 统计值是普通最小二乘法估算的回归系数除以系数的标准误（即系数标准差的样本估计值），服从自由度为 $n-2$ 的 t 分布。若设定显著性水平为 α，则双尾检验的临界值为 $t_{\alpha/2}$。若 $|t| \geqslant t_{\alpha/2}$，我们可以拒绝零假设，认为预测变量系数显著不为 0，即预测变量与结果变量有显著线性关系；否则，我们无法拒绝零假设，就认为预测变量与结果变量间没有线性关系。使用统计软件时，与 F 检验类似，我们只需要看预测变量系数的 p 值是否小于设定的显著性水平 α，如果小于此值，则表示预测变量与结果变量间存在显著的线性关系。

实际研究中，若不能拒绝零假设，还需要查明原因，为下一步改进研究做准备。不能拒绝零假设，有以下几种可能的原因：（1）预测变量与结果变量间无显著相关关系；（2）二者显著相关，但不是线性关系；（3）存在遗漏变量偏误（见第二章）。

回归方程拟合优度

上面的两项检验可以告诉我们整个回归方程及回归系数的显著性，但不能显示回归方程对样本数据的拟合情况。判定系数（coefficient of determination）R^2（$0 \leqslant R^2 \leqslant 1$）是检验回归方程拟合优度的统计量，也称为“决定系数”或“确定系数”。它是指在结果变量的总变化（总离差平方和）中，由回归方程解释的变化（回归平方和）所占比重。就好像检验一件衣服是否合身，我们检验通过

经验回归方程拟合出的值在多大程度上符合样本观测值（李连江，2017）。

若总离差平方和 SST 不为零（只有所有样本的结果变量都相等时 SST 才为零），可以得到判定系数：

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

其中，SSR 为回归平方和，SSE 为残差平方和，SST 为总离差平方和。R^2 值越接近 1 说明回归方程与样本值拟合越好，反之，则回归方程与样本值拟合越差。如果该系数等于 1，说明所有数据点都落在了拟合回归直线上；若等于 0，说明拟合回归直线与坐标轴横轴平行，不管预测变量 x 取何值，结果变量 y 的取值都是一个常数，即预测变量和结果变量毫无关系，完全无法解释结果变量的样本差异。R^2 值受样本量的影响，若样本量与预测变量个数接近，R^2 值容易接近 1，这时对回归方程拟合优度的解释要谨慎。此外，拟合优度并非检验模型优劣的唯一标准，在实际研究中还要基于理论对模型结构给予合理的解释（何晓群，2017）。

第六节　简单线性回归系数解读

简单线性回归系数表示预测变量每增加一个单位时结果变量的平均变动。下面我们按预测变量类型（虚拟变量、多分类变量、连续变量）介绍回归系数解读方式。

虚拟变量回归系数解读

这里以郭茜等（Guo et al.，2016）对完成句子（sentence completion）题型中是否预览选项（previewing）与答题时间（time，单位为毫秒）关系的研究为例。预测变量 *previewing* 是个虚拟变量，有预览行为时取值为 1，无预览行为时取值为 0。我们利用该研究的数据拟合了 *previewing* 与 *time* 之间的关系，得到经验回归方程如下：

$$\widehat{time} = 39387 + 8786\,previewing \qquad (1\text{-}3)$$

其中，截距项 39387 是无预览行为时的平均答题时间（这时 $previewing = 0$），

而预测变量系数 8786 则是有预览行为时的答题时间与无预览行为时答题时间的平均值差异，即与无预览行为时相比，有预览行为时平均每道题多用约 8.8 秒（Relative to no previewing, previewing behavior was associated with about 8.8 seconds more time per item）。但该研究不是实验或准实验设计，所以我们不能说预览行为使平均答题时间增加了约 8.8 秒，要尽量避免因果推断描述。当然，论文中的数据汇报要更复杂一些，比如要提供 p 值、决定系数、标准误等。这里主要是帮助大家理解系数的含义，如果要撰写研究成果，还需多了解目标期刊的要求，学习已发表论文的结果汇报方式。

多分类变量回归系数解读

郭茜等（Guo et al.，2016）的研究中还有一个四分类预测变量——英语水平，包括中低水平（*lower-intermediate*）、中高水平（*upper-intermediate*）、高水平（*advanced*）和母语水平（*native*），我们以中低水平为参考类别，设置三组虚拟变量，拟合得到答题时间与英语水平之间的经验方程如下：

$$\widehat{time} = 56438 - 3849 upper\text{-}intermediate - 15801 advanced - 36252 native \quad (1\text{-}4)$$

其中，截距项是三个虚拟变量都取值为 0 时的平均答题时间估计值，即参考类别（中低水平组）的平均答题时间估计值，为 56.4 秒。三个虚拟变量系数分别代表各虚拟变量取值为 1 的组别与参考类别组在平均答题时间估计值上的差异。可见，与中低水平组相比，中高水平组平均每题少用约 3.8 秒，高水平组平均每题少用约 15.8 秒，母语组平均每题少用约 36.3 秒（Compared with participants in the lower-intermediate proficiency group, participants in the upper proficiency group on average spent about 3.8 seconds less per item, participants in the advanced proficiency group spent about 15.8 seconds less per item, and native speakers spent about 36.3 seconds less per item）。

连续变量回归系数解读

我们用一组数据来拟合每周英语学习时长（范围 6～8 小时）和英语成绩的关系，得到经验回归方程如下：

$$\widehat{grade} = -27.89 + 13.69 hours \quad (1\text{-}5)$$

预测变量 *hours* 的系数可解读为：每周英语学习时间每增加 1 小时，英语成绩增加约 13.7 分（On average, one more English-study hour is associated with about 13.7 more points in English grade）。截距项（–27.89）是否意味着如果一个学生的每周英语学习时长为 0，他的英语成绩就会为负数呢？很显然，这不符合实际。这提醒我们，不要超出数据范围（如这里的学习时长范围是 6～8 小时）去解读系数。

第七节　简单线性回归与相似分析方法

相关分析

相关分析是研究两个变量之间相关关系的统计分析方法，不区分预测变量和结果变量。不同测量尺度的变量需要采用不同的相关分析方法。两个变量均为定距或定比变量且满足正态分布时，用皮尔逊 r 相关；两个变量均为定序变量时，用肯德尔 τ 相关或斯皮尔曼 ρ 相关；一个变量为定距或定比变量，另一个为定序变量，或两个定距或定比变量不都满足正态分布时，我们用斯皮尔曼 ρ 相关（许宏晨，2013）。

简单线性回归与相关分析

相关是回归分析的前提。只有两个变量具有显著相关关系，回归分析才有意义。所以一般在回归分析之前可以先进行相关分析。两种分析也有一些区别，具体列举如下。

（1）对变量的处理不同：简单线性回归中区别结果变量与预测变量，两者位置不可换；而相关分析中，两个变量地位一样，不区分预测变量和结果变量。

（2）分析目的不同：简单线性回归不仅可以揭示两个变量的相关关系，还可以预测结果变量如何随预测变量的变化而变化，探究两个变量之间是否为直线关系；但相关分析不关心变量间是直线还是曲线关系，只关心变量间关系的强度和方向。

单因素方差分析

单因素方差分析（one-way ANOVA）是检验各类别的结果变量平均值是否

存在显著差异的统计分析方法。它的结果变量必须是连续变量，预测变量必须是多分类变量（三个或更多分类；只有两个分类时用 *t* 检验）。如果单因素方差分析发现显著差异，可以进行事后检验（post-hoc test），以确定哪两个类别之间存在显著差异。根据预测变量各类别之间是否相关，可以分为被试内（within-subject）单因素方差分析和被试间（between-subject）单因素方差分析。比如一组参与者参加了三次测试，想对比三次测试的成绩间是否有显著差异，用被试内单因素方差分析；如果三组参与者参加了同一项测试，对比这三组的测试成绩间是否有显著差异，则用被试间单因素方差分析。

简单线性回归与单因素方差分析

简单线性回归和单因素方差分析都应用于结果变量为连续变量的情形，都可以用来分析组别之间的差异。但二者有诸多区别，具体如下。

（1）预测变量类型不同：简单线性回归的预测变量既可以是分类变量也可以是连续变量；而单因素方差分析的预测变量只能是分类变量，通常代表组别。

（2）分析目的不同：简单线性回归主要检验结果变量是否随着预测变量的变化而变化；而单因素方差分析一般是比较三组或以上的组间差异；若简单线性回归分析的预测变量是多分类变量，也可以实现各组间结果变量的比较。

（3）预测变量是多分类变量时，简单线性回归无法直接实现两两比较，需要人工选择不同的参考类别，建立多个回归模型来实现；而单因素方差分析可以通过事后检验实现两两比较，但这建立在方差分析结果显著的基础上，而回归分析无论整体方程是否显著都可以得出组间两两比较的结果（Christensen，2016）。

第八节　结　　语

本章介绍了量化统计分析中的一些基础概念及简单线性回归的原理、系数解读等重点信息，并分析了简单线性回归与语言教育研究中其他常用量化分析方法的异同。值得注意的是，语言教育所研究的现象一般受多种因素影响，实际研究中往往难以完全控制干扰变量，所以简单线性回归经常遗漏重要变量，导致误差项与预测变量相关。因此，简单线性回归的实际应用并不多，但它是

我们理解和使用多元线性回归的基础，为多元线性回归做铺垫，尤其当预测变量很多的时候，直接将所有预测变量纳入模型是不可取的，这时简单线性回归可用来初步排除一些可能与结果变量关系不大的预测变量。

练习

1. 通过分析 80 位中国英语学习者作文语料中的词汇密度，我们得到以下经验回归方程：

$$\widehat{lexiDensity} = 82.451 - 0.356K1Word$$

其中，*lexiDensity*（lexical density）是词汇密度，指实意词在文本中所占比例（百分比形式表达）；*K1Word* 是 1000 个最常见词汇在文本中所占比例（百分比形式表达）。

（i）当 $K1Word = 86.41$ 时，*lexiDensity* 取值是多少？

（ii）为什么 *K1Word* 的系数为负数？

2. 我们收集了一个接受过学术英语训练的班级的数据，并与另一个未接受该训练班级的数据比较。通过数据分析，得到以下经验回归方程：

$$\widehat{acaWord} = 7.90 + 4.91training$$

其中，*acaWord* 是学术词汇得分（即学术词汇在文本中所占比例，百分比形式表达）；变量 *training* 是个虚拟变量，接受训练的学生取值为 1，其他学生取值为 0。方程中的系数均具有统计显著性。

（i）这里的截距项有什么含义？

（ii）预测变量 *training* 的回归系数有什么含义?

（iii）我们需要把 *training* 和一个代表未受训练组的虚拟变量（如 *noTraining*）一起纳入回归模型吗？为什么？

（iv）通过这个经验回归方程我们可以得出结论“学术英语训练提高了学生的学术英语词汇使用得分”吗？为什么？

3. 在一项英语文本的语体研究中，我们用多分类变量 *Register* 作为预测变量（包含新闻报道 *news*、自传 *biography*、学术文本 *academic text* 和小说 *fiction* 四个类别），*FPpronoun*（第一人称代词得分）作为结果变量。Stata 软件汇报的简单线性回归结果如下：

```
. reg FPpronoun i.Register
```

Source	SS	df	MS	Number of obs	=	230
				F(3, 226)	=	28.66
Model	307.651866	3	102.550622	Prob > F	=	0.0000
Residual	808.650082	226	3.57809771	R-squared	=	0.2756
				Adj R-squared	=	0.2660
Total	1116.30195	229	4.874681	Root MSE	=	1.8916

FPpronoun	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Register						
biography	1.362403	.3574762	3.81	0.000	.65799	2.066815
academic text	-.9604886	.3550305	-2.71	0.007	-1.660082	-.2608954
fiction	2.074154	.4524412	4.58	0.000	1.182611	2.965696
_cons	1.441364	.2851673	5.05	0.000	.8794368	2.00329

（i）请根据上表写出回归经验方程。

（ii）请解释变量 *biography* 的系数含义。

（iii）请解释变量 *academic text* 的系数含义

进深资源推荐

[1] Matthijs Rooduijn 和 Emiel van Loon（阿姆斯特丹大学）开设的网上课程 Basic statistics（课程网址：https://www.coursera.org/learn/basic-statistics）第二、六、七周内容。

[2] Annemarie Zand Scholten 和 Emiel van Loon（阿姆斯特丹大学）开设的网上课程 Inferential statistics（课程网址：https://www.coursera.org/learn/inferential-statistics#syllabus）第二、三、五周内容，可以与上一个慕课课程结合学习。

[3] 李连江，2017. 戏说统计[M]. 北京：中国政法大学出版社：第四章.

第二章　多元线性回归

第一章介绍了简单线性回归。简单线性回归是用一个预测变量解释结果变量的变化，然而大部分语言现象背后都有众多因素起作用，难以用一个因素解释，所以语言教育研究中很少使用简单线性回归，而是用多个预测变量共同解释或预测一个结果变量，这时要用到多元线性回归（multiple linear regression）分析两个或更多预测变量与一个结果变量间的关系，其中与研究问题相关的预测变量叫主要预测变量（major predictor/key predictor）。多元线性回归与简单线性回归的区别主要在于预测变量的数量。本章我们将介绍多元线性回归的基本概念、回归模型及其注意事项、预测变量筛选、回归方程检验、交互项、回归系数解读、与多因素方差分析的异同等内容。

第一节　基 本 概 念

遗漏变量偏误

人们通常认为女生英语口语流利程度高于男生，如果我们将性别转换为虚拟变量 *female*（女生 = 1，男生 = 0）作预测变量，口语流利度作结果变量，可以构建以下简单线性回归方程：

$$fluency_i = \alpha_0 + \alpha_1 female_i + \varepsilon_i \qquad (2\text{-}1)$$

同时，日常教学经验告诉我们，英语水平与口语流利度也密切相关，所以可以用英语水平作为预测变量来解释或预测口语流利度，新构建的回归方程如下：

$$fluency_i = \beta_0 + \beta_1 proficiency_i + \varepsilon_i \qquad (2\text{-}2)$$

如此看来，口语流利度既可以被性别预测也可以被英语水平预测，如果性别又与英语水平相关，那么方程（2-1）就把英语水平对结果变量的预测效应归给了性别，方程（2-2）则把性别对结果变量的预测效应归给了英语水平。这两

个方程的普通最小二乘法估算结果都是有偏的（biased），因为它们均遗漏了重要变量，造成了遗漏变量偏误（omitted variable bias）。

遗漏变量偏误必须满足两个条件：（1）遗漏变量必须与结果变量相关，因为方程中没有包含这个变量，所以它进入了误差项，成为误差项的一部分；（2）遗漏变量必须与一个或多个预测变量相关，由于遗漏变量在误差项中，这意味着误差项与预测变量相关，违背了普通最小二乘法回归的一条基本假设，即误差项与预测变量不相关。遗漏变量偏误是非随机实验研究经常遇到的问题，因为观测数据通常无法控制一些与结果变量及预测变量相关的干扰因素，而这些干扰因素如果不纳入回归模型，就会导致预测变量与误差项相关，回归结果有偏，预测结果不准确。出现遗漏变量时，会把遗漏变量对结果变量的贡献归给模型中的预测变量，导致高估该预测变量与结果变量间的关系，即高估预测变量系数的绝对值。

遗漏重要变量会降低普通最小二乘法线性回归估计结果的准确性。语言教育研究中，结果变量除了与研究者关注的预测变量（即主要预测变量）相关外，通常还会与数个其他变量相关，仅用简单线性回归分析可能导致遗漏变量偏误。我们可以采用随机实验或自然实验解决遗漏变量偏误问题（相关内容请见本书第五章和第六章）。如果不能实施这两种设计，还可以根据理论和研究目的收集尽可能丰富的数据，在回归模型中纳入更多预测变量，以避免遗漏重要变量。不过纳入与结果变量无关的预测变量会影响判定系数 R^2。在本章的预测变量筛选一节，我们会详细讲解如何选择预测变量。

控制变量

控制变量（control variable）是多元线性回归中研究者并不关注的因素，但这些因素会干扰研究者所关注的主要预测变量与结果变量间的关系，所以也称为混淆变量，是我们在研究中要控制的因素。如果这些因素没有以实验设计的方式控制住，就需要用统计方法在数据分析中加以控制，否则有可能造成遗漏变量偏误（李连江，2017）。

调节变量

如果预测变量 x 与结果变量 y 有关系，但二者之间的关系受第三个变量 z

的影响，那么 *z* 就是 *x* 与 *y* 关系的调节变量（moderator）（Baron et al.，1986）。调节变量可以是分类变量，也可以是连续变量。我们生活中经常说的“视情况而定”中的“情况”就起着调节变量的作用。如教学方法对英语学习效果的影响可能因学生的英语水平、学习风格而异，那么英语水平、学习风格就是教学方法与英语学习效果间的调节变量。调节效应在回归模型中以预测变量与调节变量相乘的形式出现，即交互项（interaction，也称为“交叉项”，详见本章第六节）。

中介变量

中介变量（mediator）也是回归分析中的重要概念。如果预测变量 *x* 通过影响变量 *m* 来影响结果变量 *y*，那么 *m* 就是 *x* 与 *y* 关系中的中介变量（Baron et al.，1986）。如果回归分析中，我们发现预测变量 *x* 与结果变量 *y* 显著相关，*x* 与变量 *m* 显著相关，*m* 与 *y* 也显著相关，而且在含有 *x* 与 *y* 的回归模型中纳入变量 *m* 后，*x* 的标准化回归系数（数据标准化之后得到的回归系数，详见本章第七节）绝对值显著下降，则说明 *m* 起中介变量的作用。所以中介变量可以解释预测变量与结果变量之间的关系，而调节变量可以影响预测变量与结果变量间的关系。比如家庭经济社会背景（socioeconomic status，SES）与学生英语水平的关系中，性别可能是调节变量，即 SES 对不同性别学生的英语水平影响不同；而教育则可能是中介变量，即 SES 通过教育影响学生的英语水平。

第二节　多元线性回归模型

假若我们要考察 *k* 个预测变量 $x_1, x_2, \cdots, x_k$ 与结果变量 *y* 之间的关系，设它们之间的多元回归模型为：

$$y_i = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \cdots + \gamma_k x_{ki} + \varepsilon_i \qquad (2\text{-}3)$$

其中，γ_0 是常数项；$\gamma_1, \gamma_2, \cdots, \gamma_k$ 是相应预测变量的偏回归系数，表示当其他变量保持不变时，该预测变量每增加一个单位，结果变量平均变化的数值；ε 是误差项，表示除去 *k* 个预测变量对结果变量影响后的随机误差，即回归方程的预测值与真实值的差距。这里的误差项与简单线性回归的误差项一样服从相关假设。多元线性回归方程的参数也可用普通最小二乘法进行估计。对于

方程（2-3）而言，普通最小二乘法估计就是要找 $\gamma_0,\gamma_1,\cdots,\gamma_k$ 的估计值 $\widehat{\gamma_0},\widehat{\gamma_1},\cdots,\widehat{\gamma_k}$，得到经验回归方程：

$$\hat{y}=\widehat{\gamma_0}+\widehat{\gamma_1}x_1+\widehat{\gamma_2}x_2+\cdots+\widehat{\gamma_k}x_k \tag{2-4}$$

其中，$\hat{y}$ 为观测值 y 的回归拟合值，简称拟合值。使得真实 y 值与拟合值 $\hat{y}$ 之差的平方和（残差平方和，SSE）最小的 $\widehat{\gamma_0},\widehat{\gamma_1},\cdots,\widehat{\gamma_k}$，即回归参数 $\gamma_0,\gamma_1,\cdots,\gamma_k$ 的普通最小二乘法估计值。

第三节　多元线性回归注意事项

为使参数估计量具有良好的统计性质，多元线性回归需要满足一些条件，除了第一章第三节里提到的简单线性回归的线性、独立性、正态性和方差齐性假设以外，还包括：

（1）预测变量与随机误差项之间相互独立，即不存在遗漏变量偏误。

（2）自变量之间不存在多重共线性（multicolinearity）。多重共线性是指某些自变量之间存在较强的相关性，这会威胁回归模型的稳定性和偏回归系数估计的准确性，这时普通最小二乘法估算结果会出现问题。比如把英语水平和英语学习年限同时作为预测变量放入多元线性回归模型中时，可能会产生多重共线性的问题，因为这两个预测变量间通常有较强的正相关关系。

可以通过以下途径判断是否存在多重共线性：

（1）回归模型中某些预测变量之间显著相关，一般认为相关系数大于 0.7 且 p 值小于设定的显著性水平，可初步判断这些预测变量之间存在多重共线性。

（2）回归方程通过显著性检验（F 检验），但绝大多数预测变量的回归系数 t 检验却不显著。

（3）某些变量的偏回归系数的正负方向不符合常理或理论逻辑。

（4）回归估算结果中的共线性诊断统计量，即容忍度（tolerance）和方差膨胀因子（variance inflation factor，VIF）可用来判断多重共线性，这两个统计值互为倒数。当容忍度小于 0.1 或方差膨胀因子大于 10 时，显示预测变量之间存在多重共线性问题。

由于实际研究中很多因素彼此相关，所以多元回归模型中多重共线性的问题难以完全避免。常见的解决办法有两个：

（1）若预测变量之间存在相关关系，则删除其中的次要变量，保留更为重要的变量，使得方程中保留下来的预测变量间尽可能不相关，从而减少变量之间的重复信息，避免在模型拟合时出现多重共线性的问题。

（2）通过前进法或逐步回归法确定最终进入回归模型的预测变量，而非将所有可能的预测变量同时纳入回归模型。

第四节　预测变量筛选

在确定主要预测变量后，研究者都希望构建解释能力和预测效果优良的多元线性回归模型，但如本章开头所说，语言教育领域的现象通常有很多影响因素，遗漏重要变量会影响回归结果的准确性，但如果把诸多因素都纳入回归模型，不仅增加数据收集的难度，回归方程也会异常复杂，容易造成模型过度拟合（over-fitting），即包含过多变量，过分注重细节，但未能抓住主要趋势，看似对样本数据拟合很好，但推广性（即泛化，generalizability）较低。同时，预测变量过多有可能导致样本量相对于预测变量数量而言偏小，也会导致标准误变大（何晓群，2017）。如果与结果变量关系较弱的预测变量过多，负面影响会更加严重。在选择预测变量时，我们要保证预测变量与结果变量有显著的线性相关关系，这样可以保证模型的准确性；预测变量之间不能高度相关，以避免多重共线性。所以拟合多元回归模型的目标是确定几个重要但彼此不相关的预测变量（何晓群，2017），用尽可能少的预测变量尽可能多地减小残差，以取得模型解释力（调整 R^2）和简洁性的最佳平衡（陈强，2010）。预测变量筛选就是实现这一拟合目标的过程。一个变量是否纳入模型要看其显著性水平是否低于设定的剔出变量的检验水准（α level），为了防止丢失重要变量，可以适当放宽进入模型的标准，将检验水准 α 定在 0.15～0.20，样本量大时可以定为 0.05（Hosmer et al.，2000）。α 值越小，说明预测变量筛选的标准越高。但如果一个变量很重要，即使它不符合标准，也可以将其纳入模型。

下面介绍三种预测变量筛选办法：前进法（forward elimination）、后退法（backward elimination）和逐步回归法（stepwise regression）。

前进法

假设共有 m 个备选预测变量，前进法的起始回归模型只有截距项，然后将

预测变量逐个纳入回归模型，一次只加一个变量。我们首先计算结果变量与每一个预测变量的简单线性回归方程，然后将回归平方和（SSR，结果变量的估计值与其观测值均值的差值平方和）最大（即对结果变量贡献最大）且 F 检验有显著性的预测变量纳入方程。再将其他变量分别纳入当前的一元线性回归模型，这时共有 $m-1$ 个二元回归方程，计算各个二元方程中新纳入变量的偏回归平方和（即表 2-1 中的 R^2 变化量，这代表在回归模型已有其他预测变量的前提下，新加入的预测变量对回归模型的贡献），并对其进行 F 检验。预测变量筛选中的 F 检验与回归方程显著性的 F 检验（对整个回归模型的显著性进行检验）不同，它计算每个预测变量的 F 统计量和 p 值（见表 2-1 中的 F 变化量和显著性 F 变化量），统计软件会自动计算这两个值，无须研究者手动计算，这里我们不多展开。如果一个变量的偏回归平方和最大且 F 检验结果显著，则该预测变量进入回归模型，否则就不纳入模型。按照这种逻辑，在现有模型的基础上继续添加预测变量，直到没有可添加的预测变量为止。前进法的优势在于可纳入单独作用较强的变量并有效避免多重共线性，但由于纳入标准是该预测变量进入模型时的统计意义，而没有顾及后续新加入的预测变量对已纳入预测变量的影响，且预测变量一旦进入模型就不会再被删除，很可能最后的回归模型中有些预测变量的系数已经不再具有统计显著性，这是前进法的劣势。

表 2-1　前进法预测变量筛选结果示例

模型摘要（结果变量为总成绩标准分）									
模型	R	R^2	调整后 R^2	标准误	更改统计				
					R^2 变化量	F 变化量	自由度 1	自由度 2	显著性 F 变化量
1	0.602	0.362	0.358	7.9948	0.362	75.054	1	132	0.000
2	0.777	0.604	0.598	6.3211	0.242	80.157	1	131	0.000
3	0.848	0.720	0.713	5.3401	0.115	53.552	1	130	0.000
4	0.904	0.816	0.811	4.3390	0.097	67.911	1	129	0.000
5	0.938	0.879	0.874	3.5361	0.063	66.222	1	128	0.000

注：模型 1 中的预测变量为听力，模型 2 中的预测变量为听力、翻译，模型 3 中的预测变量为听力、翻译、写作，模型 4 中的预测变量为听力、翻译、写作、词汇，模型 5 中的预测变量为听力、翻译、写作、词汇、阅读。

后退法

后退法与前进法的程序相反，是先将所有备选预测变量都纳入回归模型，即建立全变量模型，然后逐步剔除系数不具有统计显著性的预测变量，首先被剔除的是方程中偏回归平方和最小且 F 检验不显著的预测变量，然后重复这一过程，直到没有多余的预测变量符合剔除标准。预测变量一旦被删除，就永久被排除在模型之外。后退法有利于纳入与其他变量一起联合作用较强的变量，它弥补了前进法的不足，但也失去了前进法的优势，无法有效避免多重共线性。

逐步回归法

逐步回归法在前进法的基础上结合了后退法，是一种双向筛选变量的方法，即每纳入一个新的预测变量，都会对其他已纳入模型的预测变量进行评估和检验，如果新预测变量的加入使得旧预测变量对结果变量的效应不再具有统计显著性，则剔除这个旧预测变量。这个过程反复进行，直到没有显著的预测变量可以纳入方程也没有不显著的预测变量可以从方程中剔除。最终回归模型中所有预测变量对结果变量的效应都具有统计显著性，而未纳入模型的预测变量对结果变量均无显著效应，从而实现最优的预测变量组合，最大程度地解释结果变量的差异（variation）。一般情况下，准入标准 p 值小于剔除标准的 p 值，即准入标准更严格。逐步回归法综合了前两种方法的优势，因此在研究实践中最为常用。

上述预测变量的选取准则基于偏回归平方和。除了这个准则，我们还可以考虑其他准则。如调整决定系数（即调整 R^2，adjusted R^2）考虑样本量和预测变量个数的影响，当回归方程有不止一个预测变量时，可以用来判断回归的拟合情况。该系数的值越大，所对应的回归方程拟合越好，本章第五节最后一段对该系数有更具体的讲解。赤池信息准则（Akaike information criteria，AIC）也可用于预测变量筛选，AIC 是日本统计学家 Akaike 在极大似然基础上提出的准则，该值越小，对应的回归模型越优越。C_p 统计量也是一种预测变量筛选的标准，该统计量越小，模型方程越好。何晓群（2017）对这三个准则有详细讲述，感兴趣的读者可以参考。

实际研究中，我们可以根据研究目的对不同变量采用不同的筛选办法。既要利用统计检验结果筛选预测变量，又不能完全按照逐步回归的结果来确定所

有预测变量，而是要结合理论和逻辑确定最终的预测变量筛选结果，还可以考虑添加预测变量之间的交互项（详见本章第六节）。回归模型的建立是个繁琐复杂的探索过程，需要不断尝试、检验、修正、再检验直到最终确定模型。

研究实例 2.1：逐步多元线性回归

MacIntyre et al.（2020）研究了语言教师在新冠肺炎疫情期间线上教学过程中的应对策略及这些策略与教师的压力、幸福感和消极情绪之间的关系。他们对五个积极型结果变量和五个消极型结果变量分别用 14 种应对策略进行逐步多元线性回归，并对最终的回归估算结果进行解读。研究者指出，逐步回归中只保留预测效应显著的预测变量，这一过程中需要注意的是每一个预测变量的系数都要结合其他预测变量的效应一起解读（即查看多元回归模型中预测变量的偏回归系数），而不能只看某个预测变量自身与结果变量的关系（即不能建立简单线性回归模型来考察某个预测变量的效应），因为每次删除或添加一个预测变量都可能引起模型中其他预测变量偏回归系数的变化。

第五节　回归方程检验

多元线性回归模型与简单线性回归模型一样，在得到参数的普通最小二乘法估计值后，也需要进行必要的检验，以确定模型是否合理。这一节主要介绍回归方程显著性检验（*F* 检验）、偏回归系数显著性检验（*t* 检验）及回归方程拟合优度检验。

回归方程显著性检验

回归方程的显著性检验用于评价预测变量与结果变量间的线性关系是否显著，即回归方程本身是否有效。回归方程的显著性检验通常采用 *F* 检验，多元线性回归方程中只要有一个预测变量与结果变量的线性关系显著，*F* 检验就会得到显著结果。*F* 统计量的计算公式为：

$$F=\frac{\dfrac{\sum(\hat{y}-\bar{y})^2}{k}}{\dfrac{\sum(\hat{y}-\bar{y})^2}{n-k-1}}=\frac{\dfrac{\mathrm{SSR}}{K}}{\dfrac{\mathrm{SSE}}{n-k-1}}$$

其中，n 为样本数量，k 为预测变量数量。根据设定的显著性水平 α、自由度（k，$n-k-1$）查 F 分布表，得到相应的临界值 F_α。若 $F > F_\alpha$，则回归方程具有显著意义，回归效果显著；若 $F < F_\alpha$，则回归方程无显著意义，回归效果不显著。统计软件可以提供 F 检验的 F 值和 p 值，只需要看 F 检验结果中的 p 值是否小于设定的显著性水平 α（如 0.05）即可：如果 p 值小于 α，则表示回归方程具有显著意义。

偏回归系数显著性检验

简单线性回归中，方程显著性检验（F 检验）和回归系数显著性检验（t 检验）是等价的。但多元线性回归中，F 检验得到显著结果并不意味着每个预测变量与结果变量间都有显著的线性关系，需要对每个预测变量的偏回归系数分别进行显著性检验。如果某个预测变量没有通过检验，说明这个预测变量与结果变量的关系不显著。在回归方程显著的基础上对每个预测变量进行显著性检验时，我们假设预测变量 x_i 的系数为 0（H_0），如果能拒绝这个零假设，则 x_i 的系数显著不为 0，即该预测变量与结果变量有显著线性关系；如果不能拒绝零假设，则说明该预测变量与结果变量的线性关系不显著，可以考虑从回归方程中删除这个预测变量。偏回归系数显著性检验可以通过 t 检验完成，一般统计软件也会提供估算结果，根据相应的 p 值是否小于显著性水平 α 即可判断该预测变量与结果变量间是否具有显著的线性关系。t 检验结果也可用于预测变量筛选过程，比如在决定纳入或删除预测变量时，可以参考 t 值的绝对值大小，绝对值最大的可以先纳入回归方程，绝对值最小的可以先从回归方程中删除（何晓群，2017）。

回归方程拟合优度检验

在多元线性回归模型中，可以用决定系数 R^2（$0 \leqslant R^2 \leqslant 1$）来衡量回归模型的拟合优度。第一章对这个概念已有基本介绍。R^2 值受样本量影响，同时也受预测变量个数影响，一般而言，随着预测变量个数增加，即使这些预测变量与结果变量间没有显著关系，预测误差也会变小，SSE 随之减小，相应地 R^2 值会增大。因此要对比几个多元回归模型的拟合优度，R^2 值可能并不合适。这时要对 R^2 进行调整，将 SSR 与 SST 分别除以各自的自由度，以抵消预测变量个数

增加对 R^2 值估算的影响。调整 R^2 值也叫调整决定系数或自由度调整复决定系数，它总是小于 R^2 值。

第六节　交　互　项

在回归模型中加入交互项是一种常见的数据处理方式，可以极大地拓展回归估计对变量之间依赖关系的解释，这也是多元线性回归相对于简单线性回归而言的又一大优势。交互效应是指一个预测变量与结果变量的关系受到另一个变量的影响，即一个预测变量与结果变量的关系会因另一个预测变量的水平不同而有所不同。有无交互作用的初步判断主要来自研究者的专业知识，比如教学方法的效果因学生的认知水平不同而不同，这时教学方法与认知水平间可能存在交互效应。在多个预测变量进入回归模型时，我们有必要考察变量间的交互效应。可通过在回归方程中加入预测变量的乘积项来检验交互效应。如果两个变量之间存在显著的交互效应，其主效应的研究意义会让步于交互效应（吴诗玉，2019）。此处仅讨论两个变量之间的交互项，即二重交互。三重交互项的解读非常复杂，且实际研究中应用较少，我们不做讨论。最常见的交互项分析方法是分层回归，即先建立一个只包含主效应但没有交互项的回归模型，估算出 R^2 值，再建立带有交互项的回归模型并估算 R^2 值。如果第二个 R^2 值显著更大，则说明交互效应显著。分层回归的更多细节可以参考 Cohen et al. 2003 年出版的 *applied multiple regression/correlation analysis for the behavioral science* 一书。

调节效应和交互效应在回归模型中都是以交互项的形式出现，从统计分析的角度看，它们没有区别。在讨论交互效应时不考虑哪个变量是主要预测变量、哪个是调节变量，即交互项里的变量作用是对称的，哪个变量做调节变量都可以；而讨论调节效应时则必须区分主要预测变量和控制变量，二者的作用不能互换（温忠麟等，2005）。

依据变量类型的不同，交互项可以分为三种：分类变量与分类变量交互；连续变量与分类变量交互；连续变量与连续变量交互（Jaccard，2003）。这三种交互项没有本质区别，只是结果解读上稍有差异。

分类变量与分类变量的交互项

虚拟变量可以直接纳入线性回归模型，分类预测变量则需要先设置为虚拟变量才能进行回归分析。分类变量与分类变量的交互项在回归模型中其实表现为虚拟变量与虚拟变量之间的交互项。比如要研究影子练习（shadowing practice，指几乎同步地跟着听力材料进行复述）对复述流利度的影响，结果变量为复述流利度分数，表达为单位时长内的音节数量，是一个连续变量；预测变量影子练习（*shad*）是一个虚拟变量，影子练习组取值为 1，无文本听力练习组取值为 0；调节变量英语高水平（*highProf*）也是虚拟变量，英语水平高的学生取值为 1，其他学生取值为 0。普通最小二乘法线性回归模型如下：

$$fluency_i = \lambda_0 + \lambda_1 shad_i + \lambda_2 highProf_i + \lambda_3 shad_i \times highProf_i + \varepsilon_i \qquad (2\text{-}5)$$

其中，交互项系数 λ_3 表示高英语水平学生中影子练习组与无文本听力组的复述流利度差异值与低英语水平组该差异值之差。若 λ_3 为正数且显著，说明相比低水平学生，影子练习对高水平学生的复述流利度影响更大；如果系数 λ_3 为负数且显著，则说明相比低水平学生，影子练习对高水平学生的复述流利度的正面影响较小或者负面影响较大（有关交互项系数的解读，具体参见本章第七节）。

上一段的交互项系数解读将影子练习视为主要预测变量，将英语水平视为调节变量。如果这两个变量都是我们感兴趣的预测变量，也可以将英语水平视为主要预测变量，将影子练习视为调节变量。研究实践中我们需要选定主要预测变量（即感兴趣的预测变量）和调节变量来解释交互项系数，以便于理解和汇报。

如果预测变量影子练习包括三个类别，分别为有文本影子练习、无文本影子练习、无文本听力练习，以无文本听力练习为参考类别，则回归分析中可以加入表示有文本影子练习和无文本影子练习的变量（*tShad* 和 *ntShad*）以及它们各自与调节变量（*highProf*）的交互项。普通最小二乘法线性回归模型如下：

$$\begin{aligned} fluency_i = {} & \pi_0 + \pi_1 tShad_i + \pi_2 ntShad_i + \pi_3 highProf_i + \pi_4 tShad_i \times highProf_i + \\ & \pi_5 ntShad_i \times highProf_i + \varepsilon_i \end{aligned} \qquad (2\text{-}6)$$

交互项系数 π_4 和 π_5 的含义可参照回归模型（2-5）的交互项系数含义。

连续变量与分类变量的交互项

连续变量与分类变量组成的交互项在实际研究中较为常见。依然以影子练习研究为例，依然使用影子练习的虚拟变量（影子练习组取值为 1，无文本听力练习组取值为 0）；英语水平则用大学英语四级考试成绩作为衡量指标（Guo et al.，2014；金艳等，2013）。普通最小二乘法回归模型如下：

$$fluency_i = \theta_0 + \theta_1 shad_i + \theta_2 CET_i + \theta_3 shad_i \times CET_i + \varepsilon_i \qquad (2\text{-}7)$$

大学英语四级考试成绩是连续变量，所以这里的交互项是一个分类变量与一个连续变量的乘积。以变量影子练习为主要预测变量，以英语成绩为调节变量，这时交互项系数 θ_3 表示影子练习与复述流利度的关系如何随英语成绩不同而有所不同。为了使系数 θ_1 有意义，需要对预测变量英语成绩进行中心化（centering）处理后再进行多元线性回归估算。中心化的原因和具体方法详见本章第七节。

连续变量与连续变量的交互项

交互项中两个变量都是连续变量的情况也比较常见，同样考察预测变量与结果变量的关系是否依赖于另一个预测变量的取值变化。例如要考察语言焦虑和英语水平与复述流利度间的关系，可以在回归模型中纳入两个预测变量的交互项，建立回归模型如下：

$$fluency_i = \pi_0 + \pi_1 anx_i + \pi_2 CET_i + \pi_3 anx_i \times CET_i + \varepsilon_i \qquad (2\text{-}8)$$

其中，*anx* 是语言焦虑，以语言焦虑分数为指标；*CET* 指大学英语四级考试成绩，是英语水平的测量指标。两个预测变量都是连续变量，但若 0 不在数据范围内的话，需要先对它们进行中心化处理再进行回归分析，这样偏回归系数 π_1 和 π_2 才有实质性意义。若交互项系数 π_3 为正且显著，说明英语水平越高时，语言焦虑对复述流利度的正面影响越大或负面影响越小；反之，则正面影响越小或负面影响越大。解读时需要关注焦虑影响的方向。

在回归模型中构造交互项时必须同时纳入交互项和组成交互项的所有变量（如上述模型中的 *anx* 和 *CET*）。模型纳入交互项后，对预测变量系数的解读不

同于无交互项模型中预测变量系数的解读，在本章第七节结果解读部分我们会详细介绍这个区别。

研究实例 2.2：有交互项的多元线性回归

郭茜等（Guo et al.，2016）研究完成句子题型中预览选项（previewing）与答题时间的关系时考察了英语水平的调节作用。主要预测变量 *previewing* 是个虚拟变量，英语水平是个四分类变量，所以设置了三个虚拟变量加入回归分析，相应地交互项也有三个。结果显示英语水平对预测变量与结果变量的关系有调节作用：对于中低水平的学生（研究参与者中英语水平最低的一组）来说，预览行为能节省平均答题时间；但对英语水平更高的学生而言，预览行为会增加平均答题时间。研究者用柱形图展示了交互效应。

第七节　多元线性回归系数解读

本节介绍多元线性回归结果的解读办法，包括未纳入交互项的自变量偏回归系数解读、纳入交互项的自变量偏回归系数解读以及交互项系数解读，其中交互项系数的解读又根据组成交互项的变量类型分别进行介绍。回归分析结果本身只是说明某个预测变量与结果变量之间的相关关系，并不能就此得出因果推断。因果关系需要通过实验或准实验设计来确定。此外，结果的解读和汇报可以借助图和表完成，但不同期刊可能风格有所不同，所以本小节聚焦多元线性回归结果的解读，不附相应的表格。撰写论文时可以参考目标期刊的已发表论文，按照期刊格式要求汇报回归结果。

未纳入交互项的自变量偏回归系数解读

先来看分类变量的系数解读。仍以影子练习研究为例，结果变量是复述流利度，以得分形式表达；主要预测变量为影子练习，包括三个类别，分别为有文本影子练习、无文本影子练习、无文本听力练习；控制变量是虚拟变量英语高水平，高水平学习者取值为 1，其他学习者取值为 0。不考虑调节效应时建立回归模型如下：

$$fluency_i = \eta_0 + \eta_1 tShad_i + \eta_2 ntShad_i + \eta_3 highProf_i + \varepsilon_i \qquad (2\text{-}9)$$

其中，预测变量影子练习在模型中由两个虚拟变量表示，*tShad* 指有文本影子练习，*ntShad* 指无文本影子练习，参考类别为无文本听力练习；*highProf* 是代表英语高水平的虚拟变量。假若有文本影子练习这个虚拟变量的系数估算为 1.36（且具有统计显著性，下同），它的含义为“控制英语水平的情况下/英语水平一致的情况下，有文本影子练习组的复述流利度平均分比无文本听力练习组高 1.36 分”（Controlling for the English proficiency/Holding the English proficiency constant, the estimated average retelling fluency score of the group of shadowing with text is 1.36 points higher than that of the group of listening without text）。如果一个预测变量本身就是虚拟变量（如英语高水平），则可以直接纳入回归模型，其偏回归系数意为该虚拟变量取值为 1 的组与取值为 0 的组（如英语高水平的学习者和其他学习者）在结果变量上的平均值之差。

接下来看连续变量的系数解读。仍以影子练习研究为例，要研究英语水平对复述流利度的影响，简单地将学生分为高水平组和低水平组未免有点粗糙，可以用大学英语四级考试成绩作为英语水平的衡量指标（Guo et al.，2014；金艳等，2013），那么英语水平就变成了连续变量。回归模型如下：

$$fluency_i = \theta_0 + \theta_1 shad_i + \theta_2 CET_i + \varepsilon_i \qquad (2\text{-}10)$$

假设英语成绩的偏回归系数 θ_2 的估计值为 0.16，它表示“其他变量保持不变的情况下，大学英语四级考试成绩每增加 1 分，学生的复述流利度平均分增加 0.16 分”（Holding the other variables fixed, one more point in students' CET-4 score is associated with 0.16 more point in students' retelling fluency score）。由于大学英语四级考试满分为 710 分，跨度较大，通常样本的成绩标准差也较大，可以考虑汇报四级考试成绩每增加 10 分对应的复述流利度分数变化，即“其他变量保持不变的情况下，大学英语四级考试成绩每增加 10 分，学生的复述流利度分数增加 1.6 分”（Holding the other variables fixed, a 10-point increase in CET-4 is associated with an increase of 1.6 points in students' retelling fluency score）。

多元回归分析不能像方差分析一样直接进行事后两两比较，若想知道有文本影子练习组与无文本影子练习组在复述流利度上的区别，需要另建立一个多元回归模型，在将影子练习这个三分类变量纳入回归分析时选择有文本影子练

习或无文本影子练习作为参考类别，这样就可以得到这两个类别之间的差异了。这个操作在统计软件中很容易实现。

纳入交互项的自变量偏回归系数解读

若两个预测变量组成交互项，此时这两个预测变量各自的偏回归系数就不是主要考虑因素了，其偏回归系数的解读也与上文中的解读方式有所不同，代表的是交互项中另一预测变量取值为 0 时该预测变量对结果变量的效应。下面我们依照组成交互项的预测变量类型的不同，对三种情况下预测变量偏回归系数的解读分别加以说明。

第一，我们看两个分类变量组成交互项时如何解读相应预测变量本身的偏回归系数。这里以前文中的回归模型（2-5）为例：

$$fluency_i = \lambda_0 + \lambda_1 shad_i + \lambda_2 highProf_i + \lambda_3 shad_i \times highProf_i + \varepsilon_i$$

该模型中两个分类变量均为虚拟变量：接受影子训练的学生变量 *shad* 取值为 1，其他学生（无文本听力练习）*shad* 取值为 0；高英语水平组的学生 *highProf* 变量取值为 1，其他学生 *highProf* 取值为 0。两个预测变量四种取值组合情况下的结果变量取值估算如下：

当 $shad = 1, highProf = 1$ 时，$\widehat{fluency} = \hat{\lambda}_0 + \hat{\lambda}_1 + \hat{\lambda}_2 + \hat{\lambda}_3$ （2-11）

当 $shad = 1, highProf = 0$ 时，$\widehat{fluency} = \hat{\lambda}_0 + \hat{\lambda}_1$ （2-12）

当 $shad = 0, highProf = 1$ 时，$\widehat{fluency} = \hat{\lambda}_0 + \hat{\lambda}_2$ （2-13）

当 $shad = 0, highProf = 0$ 时，$\widehat{fluency} = \hat{\lambda}_0$ （2-14）

由式（2-12）和式（2-14）做差可知，λ_1 是英语水平取值为 0 时（即低英语水平组），影子练习组与非影子练习组的复述流利度分数差异。由式（2-13）和式（2-14）做差可知，λ_2 为影子练习取值为 0 时（即无文本听力练习组），高英语水平组与低英语水平组的复述流利度分数差异。假如 λ_1 的估计值为 0.89，可以解读为“对于低英语水平的学生而言，影子练习组的复述流利度平均分比非影子练习组高 0.89 分”（For students with lower English proficiency, the estimated average retelling fluency score is 0.89 point higher for those with the shadowing practice than for those without the shadowing practice）。系数 λ_2 的估算结果解读同理，不再赘述。

第二，我们看连续变量与分类变量组成交互项时如何解读相应预测变量的偏回归系数。回归模型沿用模型（2-7）：

$$fluency_i = \theta_0 + \theta_1 shad_i + \theta_2 CET_i + \theta_3 shad_i \times CET_i + \varepsilon_i$$

估算过程可参考上一种情况。假设θ_1的估计值为 0.21，可以解读为“当学生的大学英语四级考试成绩为 0 时，影子练习组的复述流利度平均分比非影子练习组高 0.21 分”（For students whose CET-4 score is 0, the estimated average retelling fluency score of those with the shadowing practice is 0.21 point higher than that of students without the shadowing practice）。这么解读逻辑上没问题，但 0 不在四级考试成绩的数据范围内，因此上面这种解读没有实质意义。

解决这个问题的方法之一是用总平均中心化（简称“总平减”，grand mean centering）的方式对该变量进行中心化处理，即用这个变量的各观测值减去其均值，这样新变量的 0 值其实是旧变量的均值，就有意义了（Jaccard，2003）。中心化处理不会改变整个回归模型的显著性检验结果，也不会改变交互项的系数及其显著性检验结果，但会改变偏回归系数的估计值。假设收集的大学英语四级考试成绩数据的均值为 400 分，那么可以将每个学生的考试成绩都减去 400，生成新的变量 *CETc*，*CETc* 是一个均值为 0 的连续型变量。用 *CETc* 代替模型（2-7）中的 *CET*，得到新的回归模型：

$$fluency_i = \theta_0 + \theta_1 shad_i + \theta_2 CETc_i + \theta_3 shad_i \times CETc_i + \varepsilon_i \qquad (2\text{-}15)$$

假设这时系数θ_1估计值为 0.62，可以解读为“对于大学英语四级考试成绩为 400 分（英语水平中等）的学生而言，影子练习组的复述流利度平均分比非影子练习组高 0.62 分”（For students whose CET-4 score is 400, the estimated average retelling fluency score of those with the shadowing practice is 0.62 point higher than that of students without the shadowing practice）。

第三，当连续变量与连续变量组成交互项时，相应预测变量本身的偏回归系数如何解读呢？比如模型（2-8）中有语言焦虑得分与大学英语四级考试成绩的交互项，这时可能需要对两个变量都进行中心化处理再进行回归估算，其偏回归系数的解读方式与第二种情况基本一致，在此不再赘述。

交互项系数解读

前面讨论的是各个预测变量的偏回归系数解读，在纳入交互项的模型中，

交互项系数的解读更为重要。这里仍以影子练习研究为例来探讨不同类型变量所组成交互项的系数解读方法。

第一，我们看分类变量与分类变量组成的交互项系数。以两个虚拟变量的交互项为例，仍然分析回归模型（2-5）：

$$fluency_i = \lambda_0 + \lambda_1 shad_i + \lambda_2 highProf_i + \lambda_3 shad_i \times highProf_i + \varepsilon_i$$

模型中两个预测变量均为虚拟变量：接受影子训练的学生变量 *shad* 取值为 1，其他学生（无文本听力练习）*shad* 取值为 0；高英语水平组的学生 *highProf* 变量取值为 1，其他学生 *highProf* 变量取值为 0。

先来计算影子练习组的复述流利度，即将 *shad*=1 代入方程（2-5）：

$$\widehat{fluency} = \hat{\lambda}_0 + \hat{\lambda}_1 + \hat{\lambda}_2 highProf + \hat{\lambda}_3 highProf \quad (2\text{-}16)$$

接下来计算非影子练习组的复述流利度，即将 *shad*=0 代入方程（2-5）：

$$\widehat{fluency} = \hat{\lambda}_0 + \hat{\lambda}_2 highProf \quad (2\text{-}17)$$

两组在复述流利度上的差异为

$$\triangle \widehat{fluency} = \hat{\lambda}_1 + \hat{\lambda}_3 highProf \quad (2\text{-}18)$$

当 *highProf* = 1 时，$\triangle_1 \widehat{fluency} = \hat{\lambda}_1 + \hat{\lambda}_3$ （2-19）

当 *highProf* = 0 时，$\triangle_0 \widehat{fluency} = \hat{\lambda}_1$ （2-20）

$$\triangle_1 fluency - \triangle_0 \widehat{fluency} = \hat{\lambda}_3 \quad (2\text{-}21)$$

由此可知，影子练习对复述流利度的影响随着英语水平的不同而不同，若以影子练习为主要预测变量，可以说英语水平调节了影子练习与复述流利度间的关系，即交互项表示的是某个预测变量与结果变量间的关系随另一个预测变量取值的不同而有所不同。由于交互项是乘积形式，所以也可以说模型（2-5）中的交互项意味着影子练习调节了英语水平与复述流利度间的关系。实际研究一般有主要预测变量，所以不需要从两个角度分别对交互项系数进行解读。比如影子练习研究中，主要预测变量是影子练习，那么英语水平可视为调节变量。假若交互项系数 λ_3 的回归估算结果为 0.85，这个系数可以解读为“高英语水平组的影子练习组学生与非影子练习组学生在复述流利度分数上的差异比低英语水平组高 0.85 分”（The retelling fluency score difference between students with the

shadowing practice and those without the shadowing practice is 0.85 point higher for students of high English proficiency than for their peers）。

交互项的两个预测变量都是分类变量时，可以考虑用柱状图等形式显示两个预测变量与结果变量间的关系（如图 2-1 所示），使读者有比较直观的认识。

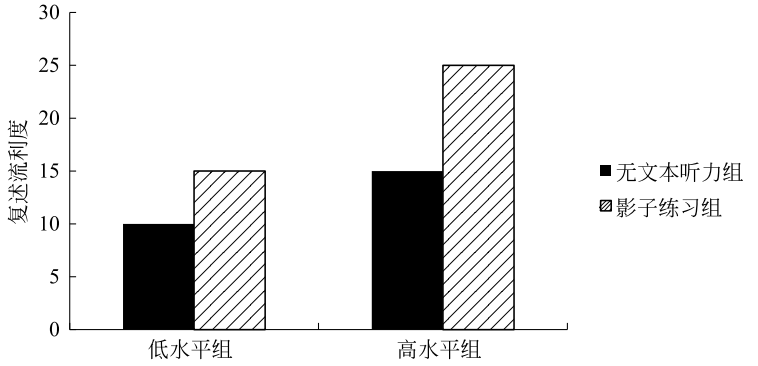

图 2-1　虚拟变量与虚拟变量交互效应示意图

第二，我们来看连续变量与分类变量组成的交互项系数的解读。仍然以影子练习为例，若我们用大学英语四级考试成绩衡量英语水平，可以得到模型（2-7）：

$$fluency_i = \theta_0 + \theta_1 shad_i + \theta_2 CET_i + \theta_3 shad_i \times CET_i + \varepsilon_i$$

模型中有虚拟变量与连续变量的交互项。我们关注的变量是影子练习，如果交互项系数 θ_3 的估计值为 0.11，则该系数可以解读为“大学英语四级考试成绩每增加 1 分，影子练习组与无文本听力练习组学生的复述流利度分数差异增加 0.11 分/影子练习组的复述流利度分数比非影子练习组多提高 0.11 分”（Corresponding to one more point in CET-4 score, the retelling fluency score difference between students with the shadowing practice and those without the shadowing practice is 0.11 point higher / One more point in the CET-4 score is associated with 0.11 more point in the increase of the average retelling fluency score for students with shadowing practice than for those without shadowing practice）。

分类变量和连续变量组成交互项时，同样可以用图形展示预测变量与结果变量间的关系，这时一般会以连续变量为横轴、结果变量为纵轴，对分类变量的不同类别分别拟合直线（如图 2-2 所示）。这些直线斜率的不同可以直观地反映交互效应（Harrell，2001）。

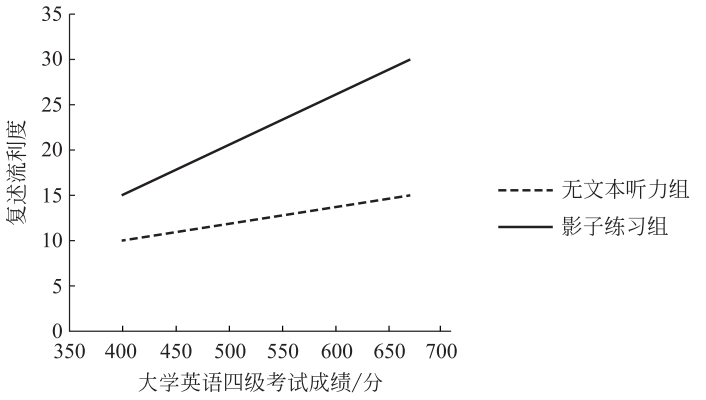

图 2-2　虚拟变量与连续变量交互效应示意图

第三，连续变量与连续变量的交互项系数该如何解读呢？仍以回归模型（2-8）为例：

$$fluency_i = \pi_0 + \pi_1 anx_i + \pi_2 CET_i + \pi_3 anx_i \times CET_i + \varepsilon_i$$

若以语言焦虑为主要预测变量，英语水平为调节变量，交互项系数π_3是指英语四级考试成绩每提高 1 分，语言焦虑与复述流利度分数间关系的变化（即语言焦虑的斜率变化）。若π_3估计值为– 0.17，它可以解读为“大学英语四级考试成绩每增加 1 分，语言焦虑得分每增加 1 分所对应的复述流利度分数变化会减少 0.17 分”（Given one more point in the CET-4 score, the change in the retelling fluency score associated with one more point in language anxiety decreases by 0.17 point）。这表明英语成绩越好的学生复述流利度得分受语言焦虑的影响越小。

组成交互项的预测变量都是连续变量时，如果也想用图形显示两个预测变量的交互关系，可以以主要预测变量为横轴、结果变量为纵轴，取调节变量的高、中、低三个值，分别为其拟合显示主要预测变量与结果变量关系的直线（如图 2-3 所示）。同样可以从直线斜率的不同看出交互效应。

需要说明的是，当模型中还有其他预测变量时，我们依然要加上“其他变量保持不变的情况下”这样的语句。为了解读方便，我们的模型中没有添加更多的预测变量，所以交互项系数的解读也省略了这部分。

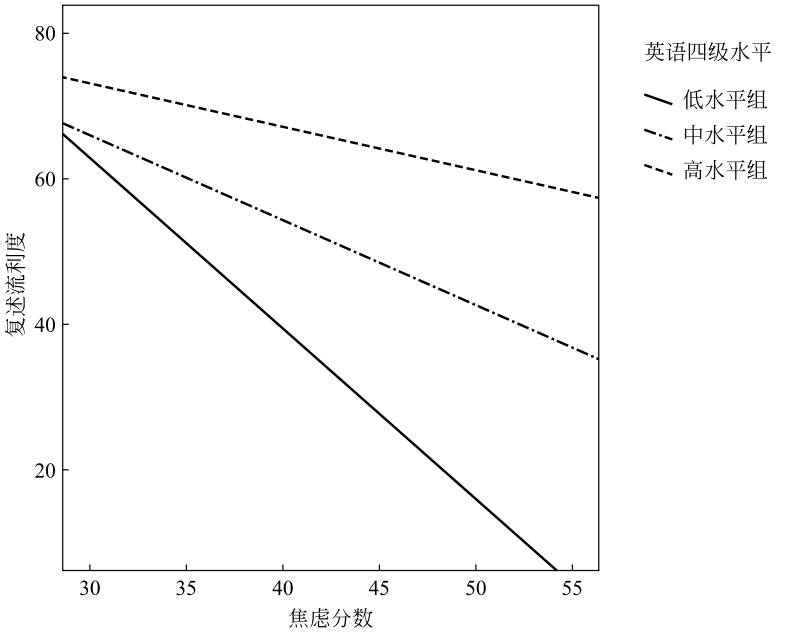

图 2-3　连续变量与连续变量交互效应示意图

标准化回归系数解读

由于预测变量的测量单位各不相同，我们无法根据偏回归系数的绝对值大小比较各预测变量对结果变量的影响大小，可以考虑将结果变量和预测变量都标准化（standardization）。最常用的标准化方法是标准差标准化，以学生成绩为例，标准差标准化是指将各学生的成绩减去所有学生成绩的平均值，然后除以所有学生成绩的标准差，这样得到的标准化得分平均值为 0，标准差为 1。标准化处理可以去除数据的不同单位，将其转化为无量纲的纯数值，这样不同单位或量级的数据能够进行比较。数据标准化之后得到的回归系数为标准化回归系数，可用来比较不同预测变量对结果变量作用的大小。该系数的绝对值越大，说明这个预测变量对结果变量的预测作用越大。统计分析软件中均提供线性回归分析结果的标准化回归系数（如表 2-2 中的标准化系数 Beta），无须我们专门对数据进行标准化处理。

表 2-2　多元回归参数估计示例

模型	结果变量：总分（标准分）					
	预测变量	未标准化回归系数		标准化回归系数	*t*	显著性
		B	标准误	Beta		
1	（常量）	11.025	2.517	—	4.380	0.000
	听力	1.643	0.130	0.456	12.606	0.000
	阅读	1.188	0.151	0.285	7.856	0.000
	词汇	1.841	0.174	0.375	10.555	0.000
	写作	2.261	0.194	0.417	11.650	0.000

截距项解读

在没有中心化预测变量的回归模型中，截距项的含义很简单，即所有预测变量（包括控制变量）的取值均为 0 时结果变量的估计值。如果对预测变量进行了中心化处理，处理后预测变量的 0 值其实是原预测变量的均值，这点在解读截距项时需要注意。

第八节　多元线性回归与多因素方差分析的异同

若回归模型中不包含连续型预测变量，且预测变量数量较少，分类变量与分类变量的交互也可以考虑用多因素方差分析（two-way analysis of variance，two-way ANOVA）。多因素方差分析也是语言教育研究中常用的量化数据分析方法，适用于结果变量为连续变量且有两个或两个以上分类预测变量的找差异研究，它回答的问题是“分类变量的不同水平组合之间在结果变量上有什么差异”。

多因素方差分析与多元线性回归有何异同呢？两者的结果变量都必须是连续变量，预测变量个数都大于 1，预测变量都可以是分类变量，都可以检验预测变量的主效应和交互效应。但这两种分析方法的分析目的、预测变量选择都有所不同。

分析目的上，多元线性回归可以用于确定两个或多个预测变量与结果变量间的关系（relationship），也可以用来比较某个分类预测变量的不同类别在结果变量上的差异（difference）；多因素方差分析则主要用于分析多个类别之间的差异。此外，多因素方差分析有事后两两比较的功能，可实现成组比较；而多元线性回归则是比较参考类别与其他各类别在结果变量上的差异，无法判断其他各类别间在结果变量上是否存在显著差异，只能通过多次建立回归模型来实现成组比较。

预测变量选择上，多元线性回归对预测变量的测量尺度没有限制；而多因素方差分析的预测变量只能是分类变量。若想在多因素方差分析中纳入连续型预测变量，则需要借助协方差分析（analysis of covariance，ANCOVA）。协方差分析是方差分析与线性回归的结合，它的结果变量是连续变量，预测变量则既有分类变量又有连续变量，不过协方差分析中连续变量只能用作控制变量，分类变量才能用作主要预测变量。因此，协方差分析可以实现多元线性回归分析的部分功能，但不如后者使用广泛。

第九节　结　　语

我们生活的世界纷繁复杂，语言教育研究亦是如此。由于语言教育多为现场研究，大多数研究情景下研究者很难执行严格的随机实验设计，这意味着我们需要在统计分析中添加与结果变量和预测变量均相关的变量，以尽量避免遗漏变量，实现预测变量对结果变量的无偏（unbiased）估计。由于语言教育领域的核心概念多以连续型变量为测量指标，因此多元线性回归堪称应用范围最广的量化分析方法之一。

练习

1. 用 Biber（1988）的多维度分析法研究一组中国英语学习者的学术英语写作语体特点，得到下面两组经验回归方程。

$$\widehat{D5} = -0.74 + 2.55\,passive$$

$$\widehat{D5} = -2.51 + 1.93\,passive + 5.02\,conjunct$$

其中，变量 *D5* 指文本的抽象性得分（分数越高，抽象性越高），变量 *passive* 是被动结构的标准分，变量 *conjunct* 是连词结构的标准分。这三个变量的得分均由 Multidimensional Analysis Tagger 软件（Nini，2015）分析得出。

（i）请解读第一个方程中 *passive* 的系数。

（ii）第二个方程中加入 *conjunct* 后，*passive* 的系数发生了什么变化，为什么？

（iii）请解读第二个方程中 *passive* 的系数。

2. 下面的经验回归方程改编自 Dynarski（2000）关于美国佐治亚州 HOPE 奖学金项目对大学入学率影响的研究。

$$\widehat{colAttend} = 0.415 + 0.115Georgia - 0.001After + 0.078Georgia \times After$$

其中，变量 *colAttend* 是大学入学率；*Georgia* 是虚拟变量，佐治亚州取值为 1，其他州取值为 0；*After* 也是虚拟变量，奖学金项目实施后取值为 1，实施前取值为 0。

（i）佐治亚州和其他州在 HOPE 奖学金项目实施前的大学入学率差异是多少？

（ii）其他州在 HOPE 奖学金项目实施前与实施后的大学入学率差异是多少？

（iii）佐治亚州在 HOPE 奖学金项目实施前与实施后的大学入学率差异是多少？

（iv）变量 *Georgia* 对预测变量 *After* 与结果变量 *colAttend* 间的关系有何调节作用？

进深资源推荐

[1] Cohen D, Cohen P, West S G, et al., 2003. Applied multiple regression/correlation analysis for the behavioral science[M]. 3rd edition. Mahwah, NJ: Lawrence Erlbaum Associates: Chapters 3-5, Chapters 7-9.

[2] Jaccard J, 2003. Interaction effects in multiple regression[M]. 2nd edition. Thousand Oaks, CA: Sage.

第三章　logistic 回归

语言教育领域的统计分析书籍较多关注数值连续的变量（如测试成绩、错误数量、反馈次数、修改次数等），而对类别选择类的结果变量关注较少。但分类变量作为结果变量在语言教育研究中也十分常见，如修正性反馈的质量（正确/错误）、错误修改的质量（已修改/部分修改/未修改）、对反馈的接纳（接纳/不接纳）、答案的正确性（正确/部分正确/错误）等。结果变量为分类变量时，线性回归会受到限制。本章介绍结果变量为分类变量（包括二分类变量和多分类变量）时的回归分析方法，主要聚焦 logistic 回归分析方法。很多量化统计书里用 logistic 回归通指 logistic 回归和 logit 回归，它们其实是同一种回归类型，适用条件也一样，但系数解读有所不同。我们遵循这一通指惯例，但会分别详细介绍 logistic 回归和 logit 回归的系数解读方法。本章通过研究案例重点介绍 logistic 回归分析方法，包括其基本概念、模型类型、预测变量设置、回归系数解读、模型评价与适用条件，同时也涉及 probit 回归、卡方检验等相似检验方法。

第一节　logistic 回归基本概念

线性概率模型

线性回归的一个重要前提是结果变量为连续变量。当结果变量为分类变量时，如果我们还用线性回归来估计结果变量，就要从概率（y 取值为 1 的比率）的角度来解释结果变量的估计值，人们通常称这种线性模型为线性概率模型（linear probability model，LPM）。比如要考察英语六级成绩与就业状态的关系，就业状态是一个二分类变量，其概率最小为 0，最大为 1。统计模型可写为：$employ_i = \alpha_0 + \alpha_1 eng_i + \cdots + \alpha_k x_k + \varepsilon_k$，其中，$employ$ 是个虚拟变量，就业状态取值为 1，未就业状态取值为 0；eng 是个连续变量，指英语六级成绩。系数 α_1 可解释为其他因素保持不变的情况下，英语六级成绩每增加 1 分，就业的概率增加 α_1。使用线性概率模型估计分类变量会产生以下几个问题：（1）残差的方差

随结果变量取值的变化而变，仍以英语六级成绩与就业的关系为例，501 分和 500 分之间的就业差异与 601 分和 600 分之间的就业差异可能有所不同，呈现异方差性（heteroscedasticity），违反线性回归模型的方差齐性假设；（2）线性概率模型估计出的概率可能超出[0,1]的区间；（3）以分类变量为结果变量的模型中，预测变量与发生概率之间是非线性关系，因此线性概率模型的回归系数和常数项都不固定，而是随自变量的取值而变化，线性概率模型无法拟合曲线关系（Wooldridge，2002；王济川等，2001），李连江（2017）称此为“直线的危机”。如果用 S 形曲线来拟合概率从极小到极大的变化，并将概率转换为发生率的自然对数，使得曲线的一端无限接近 0，另一端无限接近 1，这样概率可以作为回归分析的结果变量，同时也避免了估算的结果值小于 0 或大于 1 的情况。logistic 回归是这类分析方法中常见的一种。

发生率

logistic 回归模型将概率转变为发生率（odds，也叫发生比）。发生率是一件事情发生的概率除以不发生的概率，即 $\text{odds} = P/(1 - P)$，发生率永远为正数。用发生率测量概率，可以避免结果变量出现负值。但发生率的变化与概率的变化不对称，容易让人误解，如发生的概率由 0.98 上升到 0.99，发生率会由 49（0.98/0.02）上升到 99（0.99/0.01）。

发生率的自然对数转换

logit 回归中对发生率做自然对数转换，即转换为以 e 为底的对数，取值可以是从负无穷到正无穷的所有实数，预测变量取任何值都可以保证结果变量有意义，从而将参数间的曲线关系转换为线性关系（变量间的关系仍为非线性），因而具有很多线性回归的特点，且发生率自然对数的变化与概率的变化比较一致。发生率自然对数的正负以概率为 0.5 为界限，概率为 0.5 时，发生率为 1，其自然对数为 0；概率大于 0.5 时，发生率大于 1，发生率的自然对数为正数，且概率越大，发生率的自然对数也越大；概率小于 0.5 时，发生率小于 1，发生率的自然对数为负数，且概率越小，发生率的自然对数的绝对值越小。

发生率之比

两个发生率之比（odds ratio，OR）也称为“优势比”或“比数比”。比如

$odds_1 = P_1/(1-P_1)$，$odds_2 = P_2/(1-P_2)$，$odds_1 : odds_2$ 即二者之间的优势比。优势比有如下特性：

$$P_1 < P_2，\ P_1/(1-P_1) < P_2/(1-P_2)，\quad odds_1 : odds_2 < 1$$
$$P_1 = P_2，\ P_1/(1-P_1) = P_2/(1-P_2)，\quad odds_1 : odds_2 = 1$$
$$P_1 > P_2，\ P_1/(1-P_1) > P_2/(1-P_2)，\quad odds_1 : odds_2 > 1$$

概率、发生率、发生率的自然对数转换和优势比其实是同一事物的不同表达方式，其中概率最为直观也最容易理解。了解这几个基本概念有助于理解 logistic 回归分析的系数含义。

最大似然估计

前两章提到的线性回归用普通最小二乘法来估计总体参数，其目标是将结果变量的估计值与观测值之间的误差平方和降到最小。而 logistic 回归一般采用最大似然估计法（maximum likelihood estimate，MLE）进行参数估计。似然（likelihood）是指过去发生的概率，最大似然估计法的目标是构建与观测数据最接近的模型，“以最大概率再现样本观测数据”（王济川等，2001），即模型可对观测数据所代表的现实事件在过去发生的概率做出最准确的预测（李连江，2017）。最大似然估计法以–2loglikelihood（似然性自然对数的负二倍，简称负二倍或–2LL）为指标。似然即概率，取值范围是[0,1]，其自然对数（即对数似然值，log likelihood，LL）取值范围是(–∞,0]，乘以（–2）后得到的–2LL 指标为正数（李连江，2017）。–2LL 用来衡量模型的估计值与样本观测值之间的差距，该数值越小，说明模型拟合得越好。logistic 回归模型的目标就是减小–2LL。最大似然估计法会不断估计模型、修正模型、再估计新模型，一直到不能再显著减小–2LL，就达到了最大似然估计的目标。这一重复过程被称为迭代（iteration）。统计分析软件能自动完成迭代过程，大家只需明白这一原理即可。

第二节　logistic 回归类型

按照结果变量类别的数量，logistic 回归可分为二分类（binary）logistic 回归和多分类 logistic 回归，后者又可以根据结果变量是否为有序多分类变量分为无序多分类（multinominal）logistic 回归和有序多分类（ordinal）logistic 回归（表 3-1）。

表 3-1　logistic 回归分类

预测变量	结果变量	数据分析方法
连续/分类	二分类	二分类 logistic 回归
连续/分类	无序多分类	无序多分类 logistic 回归
连续/分类	有序多分类	有序多分类 logistic 回归

本章第一节提到，线性回归模型不能直接用来分析结果变量为分类变量的情况，logistic 回归将发生率 $P/(1-P)$转化为发生率的自然对数 $\log[P/(1-P)]$。以一元 logistic 回归为例，回归模型为：

$$\log\left[P_i/(1-P_i)\right]=\alpha_0+\alpha_1 x_i+\varepsilon_i \tag{3-1}$$

二分类 logistic 回归

二分类 logistic 回归（binary logistic regression）是最常见的 logistic 回归类型，指结果变量为二分类变量的 logistic 回归。我们以它为例来探讨 logistic 回归的估计过程。譬如要考察某校男生和女生在英语四级考试通过率上是否有差异，上面 logistic 回归模型中的预测变量 x 设为代表男性的虚拟变量（男生 = 1，女生 = 0），结果变量为表示通过英语四级考试的虚拟变量（通过 = 1，未通过 = 0）。使用 logistic 回归模型（3-1）可以得到：

$$\log \text{odds}_{男}=\log[P_{男}/(1-P_{男})]=\alpha_0+\alpha_1 \tag{3-2}$$

$$\log \text{odds}_{女}=\log[P_{女}/(1-P_{女})]=\alpha_0 \tag{3-3}$$

$$\log \text{odds}_{男}-\log \text{odds}_{女}=\alpha_1 \tag{3-4}$$

由式（3-4）可知系数 α_1 即我们感兴趣的男生和女生在英语四级通过率上的差异，不过这里是以 log odds 的形式呈现，即男生通过四级的 log odds 与女生通过英语四级的 log odds 相差 α_1。log odds 理解起来比较困难，我们可以把它转化为 odds ratio 的形式。

$$\log \text{odds}_{男}-\log \text{odds}_{女}=\log\left(\text{odds}_{男}/\text{odds}_{女}\right)=\alpha_1 \tag{3-5}$$

$$\text{odds}_{男}/\text{odds}_{女}=\mathrm{e}^{\alpha_1} \tag{3-6}$$

统计软件 SPSS 的 logistic 回归分析结果里提供 log odds ratio 和 odds ratio 两个参数。如表 3-2 中 B 列即 log odds ratio，而 Exp (B)为 e^{B}，即 B 的自然指数

运算结果。统计软件 Stata 则可以用不同指令计算 log odds ratio 与 odds ratio：使用 logit 指令的回归模型估算 log odds ratio 值，使用 logistic 指令的回归模型估算 odds ratio 值。

表 3-2　logistic 回归参数估计示例

		B	S.E.	Wals	Df	Sig.	Exp (*B*)
步骤 1	Mistake	–0.058	0.738	0.006	1	0.937	0.943
	Ease	2.534	1.051	5.811	1	0.016	12.608
	常量	–8.181	3.412	5.750	1	0.016	0.000

研究实例 3.1：二分类 logistic 回归

Nishioka 与 Durrani（2019）通过二分类 logistic 回归考察了马拉维青少年的语言资本与学习成果间的关系。结果变量学历是个二分类变量（小学或更高学历 = 1，无学历 = 0），预测变量包括语言资本（既不讲奇切瓦语也不讲英语 = 0，只讲奇切瓦语 = 1，既讲奇切瓦语也讲英语 = 2）、家庭社会经济背景（SES，包括成年人的平均家庭消费的五分位值、父母受教育情况），控制变量包括年龄、性别、城乡、家庭孩子数量、家庭成年人数量。研究者构建了两个 logistic 回归模型：第一个模型考察 SES 作为预测变量与孩子学历间的关系；第二个模型纳入语言资本，考察该变量是否为 SES 与学历关系的中介变量，判断标准是 SES 的标准化系数绝对值是否显著降低。结果发现 SES 和语言资本均与孩子学历有显著正相关，语言资本是 SES 与孩子学历之间关系的中介变量。缺乏语言资本对低 SES 家庭的孩子和非奇切瓦语区的孩子负面影响尤其大。研究者进行了多重共线性检查，发现容忍度均大于 0.2，排除了预测变量之间共线性的问题。

可能有读者困惑，该研究中的结果变量学历为什么只有两个类别，如果再多出中学、大学等学历水平，可能可以考察更多细节。作者在文中说明了马拉维的教育困境，全国数据显示只有一半多点的孩子可以通过小学毕业标准，中学的毛入学率和净入学率都非常低，所以青少年的学历用一个二分类变量足以说明情况。由此可见，我们在确定变量测量水平时不仅要考虑变量本身的特点，还要结合客观实际和研究需求。

无序多分类 logistic 回归

无序多分类 logistic 回归（multinominal logistic regression）是二分类 logistic 回归的延伸，其结果变量有多个类别，且各类别之间没有等级、大小的区分，如教学法这个变量中，1 代表讲授式教学法，2 代表任务型教学法，3 代表项目式教学法。这里结果变量的取值仅代表不同类别，数值大小之间没有等级顺序。应该用无序多分类 logistic 回归（Stata 中用 mlogit 指令）。如果结果变量有 n 个类别，那么无序多分类 logistic 回归其实是进行 $n-1$ 个二分类 logistic 回归运算，这一过程中会有一个类别作为参考类别，其他 $n-1$ 个类别均与对比类别进行比较，实现回归估计。例如三种教学方法中可以选择讲授式教学法为参考类别，从而得到两个 logistic 回归模型，一个比较任务型教学法与讲授式教学法，另一个比较项目式教学法与讲授式教学法。

研究实例 3.2：无序多分类 logistic 回归

Godfroid et al.（2015）使用限时语法判断测试（grammaticality judgment test）和非限时语法判断测试收集眼动实验数据，考察英语本族语者与非本族语者在视线扫描路径（scanpath）上的不同。研究者将扫描路径分为无回视（no regression）、未读完回视（unfinished reading with regression）和读完回视（finished reading with regression）。这个结果变量为无序多分类变量，所以研究者们用了无序多分类 logistic 回归进行估算，参考类别是无回视，这一类别代表最流畅的自动化加工机制（automatic processing），生成两个二分类 logistic 回归模型，结果变量分别为未读完回视与读完回视（二者表示受控加工，controlled processing）。预测变量包括语言背景（英语是否为母语）、时间压力（测试是否限时）、测试题目的语法正确性（是否正确），控制变量包括句尾区长度和总句长。回归结果中的重要参数是优势比（odds ratio）。

研究者指出，无序多分类 logistic 回归是二分类回归的延伸，本研究中的预测变量均为虚拟变量，优势比作为效应量指标，测量的是一个预测变量取值为 1 时事件的发生率相对于其参考类别（即预测变量取值为 0）的事件发生率的优势比。这里的事件发生是指两个二分类 logistic 回归模型中结果变量取值为 1，即出现未读完回视或读完回视。譬如以读完回视为结果变量的

回归模型中，预测变量时间压力的优势比可以理解为时间压力取值由 0 变成 1（即测试从非限时变成限时），读完回视行为发生的发生率之比。优势比值如果大于 1，表明时间压力大时，读完回视行为的发生率也大，即时间压力促进受控加工；优势比值如果小于 1，表明时间压力大时读完回视行为的发生率更小，说明时间压力促进自动化加工。

回归分析结果显示时间压力只抑制了非母语者的回视行为，但母语者与非母语者在非限时测试、题目正确时都表现出更多的回视行为。研究者认为这表明限时与非限时语法判断测试测量的内容不同，限时测试考察的是隐性语言知识，非限时测试考察的是显性语言知识，体现了不同程度的自动化加工与受控加工。

有序多分类 logistic 回归

当结果变量为定序变量时可以考虑用有序多分类 logistic 回归（ordinal logistic regression），其原理是将结果变量的多个分类依次分割为多个二分类 logistic 回归，拟合的模型个数等于结果变量类别数减 1。例如错误修改情况可以是有序三分类变量（正确、部分正确、不正确，分别赋值 1、2、3），用这个变量作结果变量进行有序多分类 logistic 回归，按结果变量的类别会拆分为两个二分类 logistic 回归，分别为（“1”对比“2+3”）和（“1+2”对比“3”），均为较低取值与较高取值的对比。此时求出的预测变量系数是预测变量每提高一个单位，结果变量提高一个单位的优势比（$\text{odds}_x : \text{odds}_{x+1}$）。

有序多分类 logistic 回归需要满足比例优势假设（proportional odds assumption，也称为“平行线假设”parallel lines assumption），即拆分后的几个二分类 logistic 回归的预测变量系数相等，仅常数项不等。所以有序多分类 logistic 回归结果只输出一组预测变量的系数。统计软件会提供平行线假设的检验结果，如果检验中 p 值大于 0.05，说明有序多分类 logistic 回归符合模型假设。如果假设不满足，则需要改用无序多分类 logistic 回归或其他统计分析方法。

研究实例 3.3：有序多分类 logistic 回归

Bulté 与 Roothooft（2020）考察了 9 个第二语言口语复杂度的量化测量指标与英语水平间的关系。英语水平由人工按照雅思水平测试的五个等级进行主

观评分。由于结果变量英语水平的五个等级是定序变量，研究者们采用了有序多分类 logistic 回归来估算口语复杂度指标对口语测试成绩的预测作用。回归结果发现，有三个指标可以显著预测英语口语成绩：词汇多样性、从句比例和形态复杂度指标。为避免多重共线性，研究者在拟合 logistic 回归前，先对各个预测变量之间的 Pearson 相关关系进行了估算，高度相关（相关系数大于 0.8）的复杂度指标未纳入回归模型，例如词汇多样性的一个指标 HD-D 与另两个词汇多样性指标 G 和 MLTD 均高度相关，因此研究者没有将 HD-D 纳入回归模型。

二分类还是多分类 logistic 回归

二分类 logistic 回归是最常见的 logistic 回归，但多分类的结果变量也可以进行 logistic 回归，无须将多分类结果变量重新编码成二分类变量。

研究实例 3.4：二分类还是多分类 logistic 回归

Loewen（2005）研究偶发型形式聚焦（incidental focus on form）与第二语言学习间的关系。研究所用语料是自然发生的、意义聚焦的第二语言课堂交际活动录音。研究者对形式聚焦情景的不同特征（如形式聚焦类型、语言焦点、复杂度、反应类型等）进行了二分类编码，采用前进法筛选预测变量："每次向模型中添加一个预测变量，以实现最优最简的回归模型。每次添加的预测变量都是对模型影响最大的一个，如果一个预测变量对模型的贡献不显著，则不将其纳入模型。因此，前进法回归追求用最少但最重要的预测变量去最大程度地解释结果变量。"研究者把显著性水平设为 0.15，针对三个分类结果变量（correction test score，suppliance test score，pronunciation test score），以 10 个预测变量（形式聚焦的不同特征：type，linguistic focus，source，complexity，directness，emphasis，response，timing，uptake，successful uptake）进行了三轮 logistic 回归预测变量的筛选，前两个结果变量的 logistic 回归模型都筛选出了 uptake 和 successful uptake 两个预测变量；而第三个结果变量的 logistic 回归模型中筛选出 complexity、source 和 successful uptake 三个预测变量。前进法回归特别适用于探索性研究，这时研究者们对预测变量和结果变量的关系还知之甚少。

Loewen（2005）在文中提到，“因为 logistic 回归只需要二分类的结果变量，所以把结果变量测试得分这个三分类变量（正确、部分正确、不正确）重新编码为二分类变量（正确、不正确）”。这样的转换丢失了数据信息，其实完全可以采用多分类 logistic 回归。具体是使用有序还是无序多分类 logistic 回归，可以根据数据的具体情况决定。

第三节　预测变量设置

logistic 回归对预测变量的类型没有限制，连续变量、二分类变量、多分类变量都可加入模型。连续变量可以直接加入模型，二分类变量转换为虚拟变量（取值为 0 或 1）即可纳入模型，这两类变量的回归系数比较容易解读。而多分类（有序或无序）变量的设置可参考第一章第四节。SPSS 软件中的 logistic 回归分析模块自带分类变量设置，不需要像线性回归分析那样先创建虚拟变量。Stata 软件中依然是在每个分类型预测变量前加“i.”（如“i.*Prof*”）即可。分类变量纳入回归分析需要一个参考对象，即跟哪个类别相比。最常见的比较方式是设置一个参考类别，其他类别都与参考类别相比。研究者可根据研究需要设置参考类别，但要保证参考类别有实际意义，比如“其他”这个类别不太适合。若预测变量数目过多或研究为探索性质，研究者需考虑自变量筛选的问题，这部分内容在第二章第四节有详细讲解，此处不再赘述。

下面重点介绍交互项的设置。本书第二章对多元线性回归中的交互项进行了解释，本章基于二分类 logistic 回归进一步介绍交互项设置。此处仅讨论两个变量之间的交互作用，即二重交互。三重交互项的解读非常复杂，且实际研究中应用较少，我们不做讨论，感兴趣的读者可以阅读 Jaccard（2001）对三重交互的解释及 Hosmer（2000）在一个奖学金项目研究中对三重交互效应的考察。

根据变量类型的不同，交互项可以分为三种：分类变量与分类变量交互；连续变量与分类变量交互；连续变量与连续变量交互（Jaccard，2001）。这三种交互项没有本质区别，只是结果解读上稍有差异。

分类变量与分类变量的交互项

虚拟变量可以直接纳入 logistic 回归模型，而多分类预测变量若对结果变量

无等比例影响，要先设置为虚拟变量再进行回归分析。分类变量与分类变量的交互项其实是一组虚拟变量之间的交互。譬如要研究不同英语水平的学习者对修正性反馈的接纳情况，结果变量是学生的改错正确性（正确取值为 1，不正确取值为 0），是一个二分类变量；预测变量是反馈类型，分为词汇反馈、语法反馈和内容反馈，这个变量纳入回归分析时需要设置两组虚拟变量；调节变量是英语高水平（高水平学生取值为 1，其他学生取值为 0）。若要考察两个变量之间的交互效应，则需要在回归分析中设置两个交互项。假设以词汇反馈为参考类别，logistic 回归方程如下：

$$repair_i = \beta_0 + \beta_1 gFeed_i + \beta_2 cFeed_i + \beta_3 highProf_i + \beta_4 gFeed_i \times highProf_i + \beta_5 cFeed_i \times highProf_i + \varepsilon_i \quad (3\text{-}7)$$

其中，反馈类型用两个虚拟变量表示，*gFeed* 代表语法反馈，*cFeed* 代表内容反馈；*highProf* 代表英语高水平。系数 β_4 和 β_5 为交互项系数。

连续变量与分类变量的交互项

倘若把上一个模型中的英语水平改为英语四级考试成绩，就是一个连续变量，交互项也就变成了连续变量与分类变量之间的乘积。具体到反馈类型与反馈使用效果（即修改是否正确）间的关系这一研究，模型中纳入反馈类型与英语水平的交互项，就可以考察反馈类型与改错正确率间的关系是否因英语水平而有所不同。也可以用分类变量做调节变量，连续变量为主要预测变量，研究英语水平与改错正确率间的关系是否因反馈类型不同而不同。

连续变量与连续变量的交互项

这种交互项考察的也是预测变量与结果变量间的关系是否随另一个预测变量的取值变化而变化。仍以英语水平与改错情况间的关系为例，考虑到反馈的数量也可能与改错情况相关（反馈数量可能影响学习者的信心和对每一条反馈所付出的注意力进而影响改错效果），我们在回归模型中纳入反馈数量（而不是反馈类型）这一调节变量，将英语水平与反馈数量进行交互。如果以英语水平为主要预测变量，可以研究英语水平与改错正确率间的关系是否因反馈数量的不同而有所不同。需要注意的是，英语水平与写作得到的反馈数量可能相关，所以实际研究中，研究者需要首先检验英语水平与反馈数量是否具有多重共线

性。若是，则不能将两者同时纳入回归模型；若否，则可以将两者同时纳入模型并考虑添加二者的交互项。

与多元线性回归一样，在 logistic 回归模型中构造交互项时必须同时纳入交互项和所有组成部分。纳入交互项之后，对非交互项系数的解读也不同于无交互项模型中的系数解读，这点会在系数解读部分详细介绍。

第四节　logistic 回归系数解读

这节介绍回归结果的解读办法，涉及无交互项回归模型的预测变量系数、有交互项回归模型的预测变量系数以及交互项系数的解读，其中交互项系数解读又根据组成交互项的变量类型分别进行介绍。与线性回归一样，logistic 回归本身无法得出因果推断，只有基于实验与准实验设计才能考察因果关系。若无实验与准实验设计，logistic 回归本质上还是一种相关关系。

未纳入交互项的自变量回归系数解读

假设某个虚拟变量 x 的系数是 β，由于 β 是 log odds 的形式，它表示在其他预测变量保持不变的情况下，该虚拟变量取值为 1 和取值为 0 时结果变量的 log odds 数值差别。这非常不好理解，所以研究者通常使用 β 的自然指数 e^{β}，将 log odds 差值转换为优势比形式，其含义为：在其他预测变量保持不变的情况下，该虚拟变量取值为 1 时结果变量指标事件发生率与虚拟变量取值为 0 时（参考类别）结果变量指标事件发生率的比值，即 $\text{odds}_{x=1} : \text{odds}_{x=0}$。这一推算过程可参考本章第二节式（3-2）至式（3-5）。若 e^{β} 值大于 1，表明该虚拟变量取值为 1 时结果变量取值为 1 的情况发生的可能性更大；反之，则表明该虚拟变量取值为 1 时结果变量取值为 1 的情况发生的可能性更小。比如一项研究考察性别与选择英语专业间的关系，将性别转换为虚拟变量女性（女 = 1，男 = 0），二分类 logistic 回归估算结果显示预测变量女性的 e^{β} 为 1.61（具有统计显著性，下同），即 $\text{odds}_{女} : \text{odds}_{男} = 1.61$，那么可以解读为“控制其他变量的情况下，女生选择英语专业的发生率是男生的 1.61 倍”（Controlling for other variables, the odds of choosing the English major for female students are 1.61 times the odds of that for male students）。多分类变量纳入回归分析时要设置为几组虚拟变量，这

些虚拟变量的系数含义也是这种解读方式，只是研究者要清楚设置的参考类别。

如果一个预测变量为连续变量，那它的回归系数 β 的自然指数 e^{β} 是 $\text{odds}_{x+1} : \text{odds}_{x}$，即在控制其他预测变量的情况下，该预测变量每增加一个单位，结果变量取值为 1 的发生率变为原值的 e^{β} 倍。如 Peng et al.（2002）研究阅读成绩及性别与是否进入阅读补习班之间的关系，二分类 logistic 回归估算结果显示预测变量阅读成绩的系数 β 为 – 0.0261，其自然指数为 0.9724，那么可以说在控制其他变量的情况下，阅读成绩每提高 1 分，学生进入补习班的发生率变为原来的 0.9724（Controlling for all other variables, for each point increase on the reading score, the odds of being recommended for remedial reading programs decrease from 1.0 to 0.9742）。

纳入交互项的自变量回归系数解读

若 x_1 与 x_2 组成交互项，此时这两个预测变量自身的系数解读也会变得不同，这时它们的系数是基于交互项的另一个变量取值为 0 时的效应。

首先我们看交互项由两个分类变量组成的情况。如果在男女生英语专业选择的研究中纳入另一个预测变量母亲受教育水平（转换为虚拟变量，本科及以上学历 = 1，其他情况 = 0），并设置女性与母亲受教育水平这两个变量的交互项，其中女性为主要预测变量，母亲受教育水平为调节变量。估算结果中预测变量女性的 e^{β} 仍然为 1.61，这时它的含义为：在其他变量保持不变的情况下，母亲没有本科或以上学历的学生中，女生选择英语专业的发生率是男生的 1.61 倍（Holding all other factors constant, for students whose mother does not have a bachelor’s or higher degree, the odds of choosing the English major for female students are 1.61 times the odds of that for male students）。这里是以两个虚拟变量为例，如果纳入交互项的预测变量是多分类变量，则在纳入交互项时转化成几个虚拟变量，各个虚拟变量自身的回归系数也是一样解读，只是研究者要清楚自己选的参考类别。

有的交互项由连续变量与分类变量组成。譬如将高考成绩（连续变量）作为预测变量，考察它与英语专业选择的关系，然后将女性（女 = 1，男 = 0）作为调节变量纳入模型，并构建高考成绩与女性的交互项。估算结果中高考成绩的 e^{β} 为 0.95，其含义为：在其他变量保持不变的情况下，高考成绩每增加 1 分，

男生选择英语专业的发生率下降到原来的 0.95（Controlling for all other variables, for a one-point increase in the College Entrance Examination score, the odds of choosing the English major decrease by 5% for male students）。

若交互项是由两个连续变量组成，如高考成绩和母亲受教育年限，其中母亲受教育年限是调节变量，那么这时高考成绩的回归系数则是母亲受教育年限为 0 时，高考成绩每增加 1 分带来的影响。解读方法与上文一致，此处不再赘述。

交互项回归系数解读

首先看分类变量与分类变量交互项的系数解读。仍以性别与英语专业选择研究为例，女性与母亲受教育水平这两个虚拟变量组成交互项，其中女性为主要预测变量，这时交互项回归系数的自然指数值（即 e^{β}）的含义是高学历母亲组的女生相比男生选择英语专业得到的优势比值与低学历母亲组的女生相比男生选择英语专业得到的优势比值的比率，即 $\left(\text{odds}_{女} : \text{odds}_{男}\right)_{高学历} : \left(\text{odds}_{女} : \text{odds}_{男}\right)_{低学历}$。经过估算，交互项的 e^{β} 为 0.65，我们可以解释为“在其他变量保持不变的情况下，高学历母亲组的女生相比男生选择英语专业的优势比是低学历母亲组的女生相比男生选择英语专业优势比的 0.65”（Holding all other factors fixed, the odds ratio of choosing the English major for female versus male students whose mother has a bachelor's or higher degree is 65% of the odds ratio for those whose mother has a lower educational level）。这个结果表明，高学历母亲组的男女生选择英语专业的差异小于低学历母亲组，即母亲的学历对于性别和英语专业选择间的关系起到调节作用，尽管女生选择英语专业的概率通常大于男生，但母亲拥有高学历会使得这种情况有所缓和。

接下来看连续变量与分类变量组成的交互项系数的解读。如果将高考成绩（连续变量）作为预测变量，考察它与英语专业选择间的关系，将女性（女生 = 1，男生 = 0）作为调节变量纳入模型，并构建高考成绩与女性的交互项。这时交互项系数的自然指数 e^{β} 表示女生组高考成绩每增加 1 分带来的结果变量指标事件发生率的变化与男生组高考成绩每增加 1 分带来的结果变量指标事件发生率变化的比值，即 $\left(\text{odds}_{x+1} : \text{odds}_{x}\right)_{女} : \left(\text{odds}_{x+1} : \text{odds}_{x}\right)_{男}$，其中 x 指高考成绩。交互项系数的自然指数估计值为 0.96，可以解释为：在其他预测变量保持不变

的情况下，女生高考成绩每增加 1 分选择英语专业发生率的改变是男生的 96%（Holding all other factors constant, the predicted odds change in choosing the English major given a one-point increase in the College Entrance Examination score for female students is 96% of that for male students）。这个结果表明，男生在英语专业的选择上受高考成绩影响更大。如果性别为主要预测变量，高考成绩为调节变量，那么交互项的 e^{β} 是高考成绩每增加 1 分时女生与男生选择英语专业的优势比的改变，即 $\left(\text{odds}_{女} : \text{odds}_{男}\right)_{x+1} : \left(\text{odds}_{女} : \text{odds}_{男}\right)_{x}$，其中 x 指高考成绩。交互项系数的自然指数 0.96 可解释为：在控制其他预测变量的情况下，高考成绩每提高 1 分，女生选择英语专业相对于男生选择英语专业的优势比会降低到之前的 96%（Controlling for all other variables, the predicted odds ratio of choosing the English major by female versus male students decreases by 4%, given a one-point increase in the College Entrance Examination score）。即随着高考成绩的提高，男女生选择英语专业的可能性的差距缩小。

最后看连续变量与连续变量组成的交互项系数的解读。比如要考察高考成绩和母亲受教育年限与选择英语专业间的关系，其中母亲受教育年限是调节变量，那么交互项的 e^{β} 是母亲受教育年限每增加 1 年对高考成绩每增加 1 分带来的英语专业选择优势比改变的效应，即 $\left(\text{odds}_{x+1} : \text{odds}_{x}\right)_{z+1} : \left(\text{odds}_{x+1} : \text{odds}_{x}\right)_{z}$，这里 x 指高考成绩，z 指母亲受教育年限。假如交互项的 e^{β} 为 0.65，该数值可以解释为：其他变量保持不变的情况下，母亲受教育年限每增加 1 年，高考成绩每增加 1 分带来的英语专业选择发生率的改变降低到之前的 65%（Holding all other factors fixed, the predicted change in the odds of choosing the English major corresponding to a one-point increase in the College Entrance Examination score decreases by 35% given a one-year increase in mothers' education time）。这个结果表明母亲受教育年限越高，高考成绩对英语专业选择的影响越小。

截距项解读

截距项的含义很简单，即所有预测变量（包括控制变量）的取值均为 0 时结果变量的自然对数值（log odds）。仍以性别（二分类变量）和母亲学历（二分类变量）与英语专业选择研究为例，截距项估计值α的自然指数 e^{α}表示其他预测变量取值为 0 的情况下，低学历母亲组的女生选择英语专业的发生率。

边际效应

对于中国读者而言，log odds ratio 和 odds ratio 都不如概率的概念贴近生活、易于理解（李连江，2017），边际效应可以帮助我们解释概率的百分比变化。边际效应多见于 probit 回归结果的汇报，本章第六节中将讲到的 Stata 中 dprobit 命令就是为了获得边际效应。若预测变量 x_1 未纳入交互项，那么它的回归系数表示“在其他预测变量保持不变的情况下，x_1 每增加一个单位，结果变量指标事件发生概率增加或下降的百分比”。若 x_1 与 x_2 组成交互项，x_1 的回归系数则表示“在其他预测变量保持不变的情况下，x_1 每增加一个单位，x_2 取值为 0 时结果变量指标事件发生概率变化的百分比”。如郭茜等（Guo et al.，2016）的研究发现有预览行为时的答题正确率比无预览行为时平均低 5%（Relative to no previewing, previewing was associated with a 5-percent lower probability of answering an item correctly）。将预测变量预览行为与低英语水平（虚拟变量，英语水平较低的学生取值为 1）的交互项纳入回归模型后，预览行为的回归系数为 – 0.106，可解读为“对非低水平组的学生而言，有预览行为时的答题正确率比无预览行为时平均低近 11%”（Previewing was associated with 11-percent lower response accuracy than no previewing for the non-lower-intermediate group）。

无序多分类和有序多分类 logistic 回归系数解读

无序多分类 logistic 回归模型其实是拟合成几个二分类 logistic 回归模型，所以结果解读与二分类 logistic 回归大致一样，只是需要注意结果变量的取值变成了“某件事发生 = 1，参考类别发生= 0”。这样的结果解读起来比较复杂，可以考虑报告各预测变量对结果变量各类别的边际效应。

有序多分类 logistic 回归模型虽然也是拟合成几个二分类 logistic 回归模型，但由于这些新拟合的模型估算的各变量系数相同，只有截距项不同，解读起来反而更为容易。

研究实例 3.5：无序多分类 logistic 回归（报告各预测变量的边际效应）

郭茜等（Guo et al., 2021a）研究学生使用自动书面修正性反馈（automated written corrective feedback）对修改用英语撰写的专业研究论文的有效性，其中一个研究问题是学生背景特征、反馈质量等因素与学生对写作自动评改（automated writing evaluation）系统所提修改建议采取的行动（接受、拒绝、

用其他方式修改）之间的关系。研究者在进行无序多分类 logistic 回归（Stata 中使用 mlogistic 指令）后进一步获得各预测变量的边际效应，如“所提修改建议准确时接受建议的概率增加 57%；标注准确时接受建议的概率增加 16%”（They had a 57% higher probability of accepting a suggestion when the suggestion was accurate and a 16% higher probability when the flagging was accurate）。也可以先进行无序多分类 probit 回归（Stata 中使用 mprobit 指令），再获取边际效应（Stata 中使用 mfx 指令），得到的结果完全可比。

研究实例 3.6：有序多分类 logistic 回归（结果解读）

Binder et al.（2015）运用有序多分类 logistic 回归考察实习对学生学业成绩水平的影响，结果变量学业成绩水平是一个定序变量。数据分析结果显示预测变量实习的系数为 0.349，其自然指数为 1.42，可知实习的学生获得更好学业成绩的发生率是不实习的学生的 1.42 倍。论文中汇报如下：“A positive coefficient for internship (b=0.349) indicated that undertaking an internship increased the odds of attaining a higher degree classification。”

第五节　logistic 回归模型评价与适用条件

要评价 logistic 回归模型的优劣，可以关注似然比检验（likelihood ratio test）、拟合优度检验（goodness-of-fit test）、Hosmer and Lemeshow（H-L）检验、信息测量指标（information measures）、伪决定系数（Pseudo *R*-Square）和混淆矩阵（confusion matrix）（Hosmer et al.，2000；王济川等，2001）。似然比检验的零假设是所有预测变量的回归系数为 0，如果显著性（sig.值）小于显著性水平（如 0.05），说明模型至少有一个预测变量的回归系数不为 0，模型有意义。该检验还提供 – 2LL 的统计值（参见本章第一节），可以用于比较不同模型的优劣。拟合优度检验提供的统计值包括 Pearson 卡方值和偏差（deviance）卡方值，但这两个统计值主要用于预测变量不多且为分类变量的情况；H-L 检验适用于含有连续变量或样本量较小的情况。三者均服从卡方分布，若卡方检验不显著，则表明模型拟合较好；反之，则模型拟合较差，还需要加入新的预测变量来提升模型的拟合优度。这三个指标都不如似然比检验稳健。信息测量指标也可以

用来比较不同模型的优劣，常见的指标包括 AIC 值和 SC 值，两个指标的值越小，说明模型拟合得越好。logistic 回归有三种伪决定系数：Cox & Snell R^2、Nagelkerke R^2、McFadden R^2，其意义与线性回归中的决定系数一样，都是值越大说明模型拟合得越好。Cox & Snell R^2 取值一般小于 1，但没有确定的取值范围。Nagelkerke R^2 和 McFadden R^2 的取值范围均为[0,1]。对于 logistic 回归，这三种伪决定系数一般都不会很高，研究者不必过于纠结它们的高低，根据自身研究领域的常规做法，选择相应的指标进行汇报即可。混淆矩阵提供总体正确率，代表回归模型预测正确的样本数占总样本数的比例，该值越大说明模型的预测能力越强，这是 logistic 回归模型拟合度的关键参数。伪决定系数和混淆矩阵均是评价模型预测准确性的指标。

logistic 回归的适用条件为：（1）结果变量是分类变量，可以是二分类变量、无序多分类变量或有序多分类变量；（2）预测变量之间不存在多重共线性；（3）各观测值之间相互独立。

第一个条件比较容易判断，只是定序变量到底适用于无序多分类 logistic 回归还是有序多分类 logistic 回归要通过平行线假设检验来判断。第二个条件多重共线性是个常见问题，尤其在一个指标存在多个参数的情况下。多重共线关系会增大标准误和均方误差，甚至导致回归系数的方向相反（陶然，2008）。我们可以在构建 logistic 回归模型前先进行预测变量间的相关分析，相关系数高于 0.8 意味着高度相关关系，这时需要舍弃一些变量，舍弃变量时要结合理论根据和变量间的相互关系；还可以用逐步回归的方法排除多重共线性。也可以直接用分类结果变量构建多元线性回归模型，得到容忍度（tolerance）和方差膨胀因子（variance inflation factor，VIF）数值的大小，来判断共线性是否存在，一般而言，容忍度小于 0.1 或方差膨胀因子大于 10，表明共线性存在。第三个条件在语言教育研究中经常难以满足，因为有很多参数的数据出自同一个研究参与者。Stata 软件中的 cluster 选项可以解决这一问题。

第六节　logistic 回归与相似检验方法的比较

本节重点讨论 logistic 回归与 probit 回归、卡方（χ^2）检验的异同。probit 回归通常译为“概率单位回归”，同样用于处理分类型结果变量，是与 logistic 回归类似的一种数据统计方法。logistic 回归模型中的误差项服从标准 logistic

分布，即二项分布，而 probit 模型中的误差项服从标准正态分布。它们均使用最大似然估计。probit 回归系数可解释为自变量对标准分数（*Z*-score）的作用（王济川等，2001），这不如 logistic 回归的系数解读直观、易懂，不过 probit 回归也可以计算边际效应（即预测变量每增加一个单位，结果变量指标发生概率的百分比变化），Stata 软件中的 dprobit 回归指令可直接输出边际效应结果。边际效应的解读比较直观，更符合中国读者的日常认知习惯。

probit 回归和 logistic 回归都是广义线性模型的一种，二者的函数图像非常接近。通常情况下可以互换使用。具体选择哪种回归模型主要取决于不同领域的常规做法。语言研究人员偏爱 logistic 回归，经济学家偏爱 probit 回归，语言教育研究者则两者都有可能使用。

和 logistic 回归一样，probit 回归也可以按照结果变量的类别分为三种：probit 回归、无序多分类 probit 回归和有序多分类 probit 回归（Stata 中的回归指令分别为 probit、mprobit 和 oprobit），此处不再赘述。

研究实例 3.7：probit 回归

郭茜等（Guo et al.，2016）考察完成句子测试中预览选项（previewing）与答案正确性和做题时间的关系。研究参与者包括不同水平的英语学习者和英语本族语者。研究问题包括：（1）预览选项与结果变量间的关系；（2）英语水平是否对两者间的关系起调节作用。要考察预览选项与答案正确性之间的关系，研究者构建了二分类 probit 回归模型，结果变量为答案正确性（正确 = 1，不正确 = 0），预测变量为预览行为，性别与语言水平为控制变量。为检验英语水平的调节作用，研究者在模型中加入了预览与英语水平的交互项，其中预览行为是主要预测变量，英语水平为调节变量，这是二分类变量与连续变量的交互项。研究者用了 Stata 软件中的 dprobit 命令，得到边际效应，在结果解读一节，我们已详细介绍如何解读该效应。该研究发现相对于无预览，预览会显著降低正确性；预览行为与英语水平有交互效应，高水平学生在无预览时正确率更高，但预览行为对中低水平学习者的答题正确性无显著影响。这表明预览选项并不能促进句子理解，反而有可能干扰理解过程，对英语水平高的人尤其如此。

卡方检验也经常用于考察分类结果变量与预测变量间的关系，它与 logistic 回归有何异同呢？卡方检验中的预测变量一般数量较少且只能是分类变量（连续型

数值变量无法纳入卡方检验），而且不能考察预测变量之间的交互作用。当预测变量分类较多时，每种情况的频数可能会很小（一般要求频数大于 5）。此外，预测变量为多分类变量时，卡方检验只能估计总体上的差异，无法描述预测变量效应的大小和方向，而 logistic 回归分析不仅能估计预测变量效应总体上的差异，还能对多分类变量设置虚拟变量，获得各虚拟变量水平的检验结果。当只有一个预测变量且卡方检验和一元 logistic 回归均符合条件时，选择哪种分析方法取决于研究领域的常规做法。总体而言，卡方检验在语言教学和语言习得研究中比一元 logistic 回归常见。但现实研究中一个结果变量的影响因素往往有多个，研究者通常关注不止一个预测变量，有时还需要考察交互效应，这时需要使用多元 logistic 回归。

研究实例 3.8：卡方检验与 logistic 回归

Chen 和 Eslami（2013）研究本族语者与英语学习者之间的实时在线文本聊天中的偶发型形式聚焦（incidental focus on form）对第二语言发展的影响（研究问题 1），并考察偶发型形式聚焦的哪些特征可以预测第二语言的发展（研究问题 2）。他们对英语学习者进行了即时后测与延迟后测，按“正确”“部分正确”“不正确”对学生答案的正确性进行编码。

对于第一个研究问题，研究者根据偶发型形式聚焦情景所涉及的语言点设计个性化测试，以测试时间（即时后测、延迟后测）为唯一预测变量，以答案正确性为结果变量进行卡方检验，发现即时测试中学生平均答对 70% 的题目，延迟后测中学生答对 64%的题目，两次测试间正确答案的频数无显著差异。对于第二个研究问题，研究者以答案正确性为结果变量，以偶发型形式聚焦的不同特征为预测变量，进行 logistic 回归分析，结果发现成功接纳（successful uptake）与反馈类型（type of feedback）是预测语法和词汇知识测试答案正确性的最主要因素。研究者认为 logistic 回归分析中的结果变量只能为二分类变量，所以将结果变量“答案正确性”的三种类型（正确、部分正确、错误）合并为二分类变量（正确、错误）；同时，为了易于解读预测变量系数，研究者把预测变量也均转换为二分类变量。虽然二分类预测变量的回归系数确实比多分类预测变量易于解释，在统计软件中也更易于操作，但多分类变量转换为二分类变量会失去数据信息，结果相对粗糙。

需要特别指出的是，有些研究者认为卡方检验可以考察变量间的关联性，但不能检验因果关系，而 logistic 回归可以考察数据间的因果关系。这种说法是

不正确的，因果关系只能通过实验和准实验设计（而非数据分析方法）确立，本书第五章至第九章会详细介绍各种实验与准实验设计以及实验效应估算方法。卡方检验和 logistic 回归本身均不能支持因果关系，只能检验结果变量的变化与预测变量的变化是否显著相关。即使预测变量可以预测结果变量，但单独这一点并不能表明因果关系存在（梅纳德，2016）。比如，研究实例 3.8 并非实验或准实验研究，所以 logistic 回归分析只能发现哪些解释变量与结果变量间的关系更大，而非影响更大。“影响”是描述因果关系的用词，不过实际操作中，研究人员经常用这种暗含因果关系的语言来描述相关关系。

第七节　结　　语

作为广义线性模型的一种，logistic 回归虽不及线性回归应用广泛，但它在语言教育研究领域有广阔的使用空间。本章中的 logistic 回归和 probit 回归是对同一问题的不同解答方法，无所谓优劣。在相似研究设计下，对数据分析方法的选择取决于研究目的和研究领域的常规做法。在结果汇报中，究竟选择优势比自然对数（log odds ratio）、优势比（odds ratio），还是概率，也主要取决于研究领域的常规做法。

练习

1. 在一个语言形式聚焦（focus on form）研究中，变量 *succUptake*、*elicit*、*preemptive* 和 *uptake* 均为虚拟变量。如果学生参考反馈信息，成功修正之前的语言错误或不足，并在紧接着的语言产出中使用正确的语言形式，*succUptake* 取值为 1，否则取值为 0。若教师不直接提供正确形式，而是引导学生产出正确形式，*elicit* 取值为 1；当教师针对学生的语言错误直接提供正确的语言形式，则取值为 0。当教师的语言形式相关反馈是为了回答学生的提问，即前摄型（preemptive）语言形式聚焦，*preemptive* 取值为 1；当学生的语言错误发生在前，教师的修正性反馈在后，即反应型（reactive）语言形式聚焦，则 *preemptive* 取值为 0。当学生对教师的反馈给出回应时，*uptake* 取值为 1，否则取值为 0。*elicit* 与 *uptake* 和 *succUptake* 均具有正相关关系。

（i）若以 *succUptake* 为结果变量、以 *elicit* 和 *preemptive* 为预测变量建立回

归模型，请解释 *elicit* 的回归系数含义。可以在 logit 回归、logistic 回归、probit 回归中任选一种回归形式。

（ii）将 *uptake* 纳入回归模型，这时 *elicit* 的回归系数会发生什么变化？为什么？

（iii）还有什么变量可以纳入回归模型，它（们）会带来什么效应？

2. 在一项写作自动评估（automated writing evaluation）研究中有一个虚拟变量 *response*，当学生根据自动书面修正性反馈（automated writing corrective feedback）成功修改一个语言错误时，*response* 取值为 1，否则取值为 0。变量 *undergraduate* 也是虚拟变量，当学生是本科生时，它取值为 1，否则取值为 0。变量 *onePub* 是虚拟变量，当学生只发表过一篇英语论文时，它取值为 1，其他情况取值为 0。

用 Stata 软件中的 dprobit 指令进行回归分析，得到如下结果：

```
. xi: dprobit Response Undergraduate onePub, cluster(ID)

Iteration 0:   log pseudolikelihood = −182.4234
Iteration 1:   log pseudolikelihood = −173.48068
Iteration 2:   log pseudolikelihood = −173.41896
Iteration 3:   log pseudolikelihood = −173.41895

Probit regression, reporting marginal effects                 Number of obs =      364
                                                              Wald chi2(2)  =     9.11
                                                              Prob > chi2   =   0.0105
Log pseudolikelihood = −173.41895                             Pseudo R2     =   0.0494

                                                  (Std. Err. adjusted for 36 clusters in ID)
------------------------------------------------------------------------------
             |                 Robust
Response |      dF/dx    Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
Underg~e*|   0.1025967   0.0505223     1.85   0.065   0.335165   0.003575   0.201619
  onePub*|  −0.1700775   0.0816969    −2.35   0.019   0.222527    −0.3302  −0.009955
---------+--------------------------------------------------------------------
  obs. P |   0.7994505
 pred. P |   0.8113193   (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
      z and P>|z| correspond to the test of the underlying coefficient being 0
```

Stata 中的 dprobit 命令估算出的回归系数是边际效应。

（i）请解释预测变量 *undergraduate* 的回归系数含义。为什么它是正数？

（ii）请解释变量 *onePub* 的回归系数含义。为什么它是负数？

之后在模型中加入 *undergraduate* 与 *onePub* 的交互项，得到以下结果：

```
. xi: dprobit Response Undergraduate onePub under_onePub, cluster(ID)

Iteration 0:   log pseudolikelihood = –182.4234
Iteration 1:   log pseudolikelihood = –171.46306
Iteration 2:   log pseudolikelihood = –171.28132
Iteration 3:   log pseudolikelihood = –171.28105

Probit regression, reporting marginal effects                 Number of obs =      364
                                                              Wald chi2(3)  =    19.43
                                                              Prob > chi2   =   0.0002
Log pseudolikelihood = –171.28105                             Pseudo R2     =   0.0611

                                                  (Std. Err. adjusted for 36 clusters in ID)
------------------------------------------------------------------------------
             |                 Robust
Response |      dF/dx    Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
Underg~e*|   0.1491705   0.0468866     2.74   0.006   0.335165   0.057274  0.241067
  onePub*| –0.1068448   0.0869451    –1.35   0.176   0.222527  –0.277254  0.063565
und~ePub*| –0.2865119    0.150895    –2.11   0.035   0.046703  –0.582261  0.009237
---------+--------------------------------------------------------------------
   obs. P |   0.7994505
  pred. P |   0.8158052  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
      z and P>|z| correspond to the test of the underlying coefficient being 0
```

（iii）请解释交互项回归系数的含义。为什么系数是负数?

进深资源推荐

[1] Jaccard J, 2001. Interaction effects in logistic regression[M]. Thousand Oaks, CA: Sage.

[2] Hosmer D, Lemeshow S, 2000. Applied logistic regression[M]. 2nd edition. New York, NY: Wiley.

[3] 斯科特•梅纳德，2016. 应用 logistic 回归分析[M]. 李俊秀，译. 2 版，上海：上海人民出版社.

[4] 瓦尼•布鲁雅，2018. logit 与 probit：次序模型和多类别模型[M]. 张卓尼，译. 上海：上海人民出版社.

[5] 一文读懂 logistic 回归的前世今生[EB/OL]. https://mp.weixin.qq.com/s/KV4rWT-k9ElWpmspQfv-4A.