国际关系研究的定量方法：定义、规则与操作*

庞　珣

【内容提要】　随着国际关系数据的长期积累和中层理论及微观理论在国际关系研究中占据主导地位，定量方法已经成为国际关系学中最重要的研究方法之一，也是国际关系学者工具箱中的必备，正确理解和使用定量方法对促进学科的发展具有重要意义。有的国际关系研究者对定量方法存在一些误解，错误和不当使用也屡见不鲜。有鉴于此，作者具有针对性地对定量方法的概念、原则和操作进行基本介绍和重点澄清。定量方法具有透明度高、可重演性和明确程序化的优势，但不适用于探索规范性理论或寻找历史的真相。定量方法的使用者不但要清楚它能服务于什么目的，更要清楚定量方法功能的边界。定量方法在各个操作步骤上有具体的要求和原则，研究者需要严格遵循、诚实报告。只有正确而谨慎地使用定量方法，才能够达到推进学术的目的，也能够真正维护定量方法的声誉和前景。

【关键词】　定量方法；实证主义；假设检验；科学方法；证伪

【作者简介】　庞珣，清华大学国际关系学系副教授。（北京　邮编：100084）

【中图分类号】　D815　C3　【文献标识码】　A　【文章编号】　1006－9550(2014)01－0005－21

* 本文受到王雪莲教育基金资助。感谢《世界经济与政治》杂志匿名审稿人提出的修改意见，文责概由笔者自负。

定量方法(quantitative methods)是实证主义(positivism)研究的主要方法之一。随着社会数据的长期积累和系统搜集,定量方法在社会科学研究中已变得日益普遍和重要。这一趋势在国际关系学中也表现明显:早期的国际关系研究均以历史和思辨方法为主,但如今定量方法已经跻身于最重要的研究方法和路径之列。不过,自定量方法被引入国际关系研究以来,它一直是方法论争论的焦点,拥护者视之为学科科学化之必然与必须,反对者则怨责定量方法降低了学科的深度与广度。① 产生方法论之争的根源,虽然可以追溯到本体论与认识论的根本差异与难以兼容,但亦可部分地归咎于对定量方法的种种误解。

有鉴于此,本文旨在针对一些常见的偏差理解和不当运用,结合实际研究的例子,对定量方法是什么、不是什么以及使用的步骤和原则进行澄清和梳理。学界对定量方法的理解并无绝对定论,本文尽量采用普遍认同的观点,但在论述中也不可避免地受到笔者自身的方法论训练和研究经验影响。

一 定量方法在国际关系学研究中的应用现状

很多人认为国际关系学定量方法具有"盎格鲁-撒克逊"特色,尤其是美国国际关系研究的特殊传统。但是,"国际关系教学、研究和政策调查项目(TRIP)"在欧美亚非 20 个国家和地区对从事国际关系和对外政策分析的学者进行问卷调查,发现美国学者将定量研究作为其最主要研究方法的比例其实远低于欧洲的爱尔兰、挪威和亚洲的中国香港。美国将定量方法作为第二研究方法的比例也只处于受调查国家的中等水平。②

TRIP 调查显示,研究者将定量方法作为第一方法的整体比例不高(约 20%),定量方法的使用也没有主导美国的国关研究。但与这一现象相对照的是,国际关系领域的顶级国际期刊却呈现出定量研究所占比例逐年上升和占主导地位的趋势。根据刊物的影响因子排名,12 种顶级国际关系以及相关刊物③(只选取其中刊发的国关研究成果)上发表使用定量方法的文章的比例,呈现了非常显著的增长趋势。

---

① David Collier and Henry E. Brady, *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, Lanham: Rowman & Little Field Publishers, 2004; Gary King, Robert O. Keohane and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton: Princeton University Press, 1994.

② Daniel Maliniak, Susan Peterson and Michael J. Tierney, *TRIP around the World: Teaching, Research, and Policy Views of International Relations Faculty in 20 Countries*, The Institute for the Theory and Practice of International Relations, 2012.

③ 具体包括 *American Political Science Review* (*APSR*), *American Journal of Political Science* (*AJPS*), *British Journal of Political Science* (*BJPS*), *European Journal of International Relations* (*EJIR*), *International Organization* (*IO*), *International Security* (*IS*), *International Studies Quarterly* (*ISQ*), *Journal of Conflict Resolution* (*JCR*), *Journal of Peace Research* (*JPR*), *Journal of Politics* (*JOP*), *Security Studies* (*SS*), *and World Politics* (*WP*)。

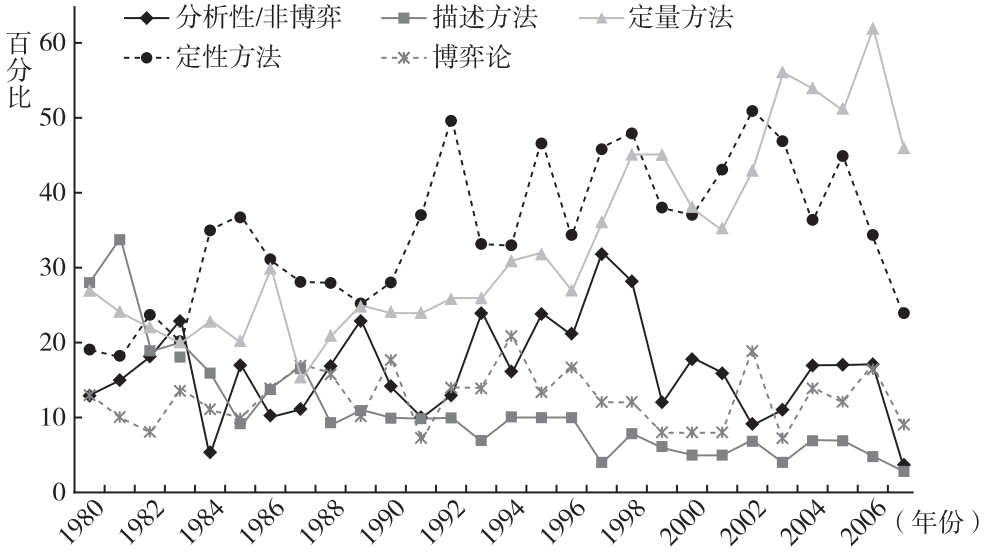

图 1　12 种期刊发表国际关系文章使用方法百分比

资料来源：Daniel Maliniak, Susan Peterson and Michael J. Tierney, "International Relations in the US Academy," *International Studies Quarterly*, Vol. 55, No. 2, 2011, pp. 437–464。

从图 1 可见，在这些刊物上发表的论文在过去 30 年中使用定量方法的比例总体有大幅增长，在近 10 年内取代定性方法成为发表论文最主要的研究方法，占到了 50% 以上。[①] 这表明，由美国学术界主导的"顶级"学术刊物，的确呈现了注重定量研究的特点。

所以，无论是从 TRIP 的学科调查结果或者基于国际关系学者的观察和感受，定量方法的重要性和广泛性不容忽视。学习和使用定量方法已成为国关学者职业训练的必备内容。无论国关学者是否决定在其研究中使用定量方法，定量方法无疑都需要被放入研究者的工具箱中，以备不时之需。同时，在中国学术"走出去"的大背景下，国际关系学者也需要在国际学术刊物上发表研究成果，正确使用定量方法将成为中国学者在国际上展示研究和发挥影响的重要条件。

二　常见的错误理解和操作

尽管定量方法在国际关系研究中的使用越来越广泛，尤其是在高水平发表中占主

① Daniel Maliniak, Susan Peterson and Michael J. Tierney, "International Relations in the US Academy," *International Studies Quarterly*, Vol. 55, 2011, pp. 437–464.

导地位,但很多人对定量研究的理解和使用都存在着一些偏差和误区。这些偏差和误区,有的来自定量方法批评者,有的来自倡导者。对定量方法的错误理解和使用可能对定量研究本身的发展、甚至对国际关系实证研究的发展造成严重的负面影响,因此有必要予以特别指出和分析。

(一)定量研究不需要国际关系理论支撑

有些定量研究的批评者认为,定量研究从数据和统计模型出发,得到实证分析结果,不需要运用国际关系理论进行分析。定量研究者只需掌握数据分析工具和统计知识及软件的使用,无须有扎实的国关理论功底和深入的理论思考。

恰恰相反,国际关系定量研究的前提是清晰的国关理论。统计建模要求研究者对与研究问题相关的理论有全面的掌握和深入分析,以提出简洁的假设并控制替代解释的干扰。得到统计分析结果后,研究者还必须对实证结果进行理论上的解读和评估。因此,定量研究不能没有理论,相反,若没有理论的话,则不需要、也不可能进行定量研究。

当然,定量研究不适合复杂的或模糊的理论,不适合大理论(也即范式意义上的理论),也不适合思辨性质的理论。定量方法本身也不发展理论,它只检验理论,但在使用定量方法的研究中,研究者通常要根据理论、逻辑和观察先发展理论,再用定量方法来检验理论。

(二)定量分析过度简化以至于远离现实、失去意义

定量方法的批评者经常对定量方法的高度简化表示不满,认为高度简化的统计模型距离现实太远,以至于其可信性和实用性让人怀疑,甚至认为定量研究者是一些对政治现实毫无感觉和兴趣的人。

但模型对现实进行简化,并非使用统计方法的结果,而是国际关系或政治学理论(如果我们还需要理论的话)本身就是对现实的高度简化,简化到能够让研究者在纷繁复杂的现实中把握并提取其感兴趣的关系,进行分离研究。当高度简化的理论形成之后,定量方法的统计建模只是对理论进行数学或统计表达而已。在这一步骤中,研究者根据理论假定和数据类型选择统计模型,而当理论被表达为统计模型后,简化都具有高度透明性,并且,谨慎的定量研究者也将会对这些假定进行系统检查。

(三)国际关系数据质量低,基于这些数据的定量研究不可靠

有一些学者对定量方法和统计模型本身并不反对和怀疑,但是基于对国际关系数据的检查,发现国际关系学的数据本身的问题很多,从而令这些学者对定量研究表示担忧和对定量分析结果表示质疑。

诚然,国际关系领域中的很多数据都涉及宏观加总数据和欠发达国家的数据,数

据缺失、测量误差以及概念化困难等问题极为常见。但是,在整个社会科学领域,完美的数据几乎没有,任何学科(包括自然科学)的数据都存在这样或那样的棘手问题。统计学中大量的理论与工具的诞生和发展正是为了处理各种数据问题,而且,统计方法和模型的发展往往是由数据中的棘手问题所驱动的。国际关系中的数据问题多而复杂,恰恰意味着国际关系学者应该学习并掌握更多的统计知识和工具。同时也意味着,国际关系学者可能对统计学本身的发展和为别的学科定量分析在方法上做出贡献。而且,在定量方法中,数据和统计模型的运用只是为了证伪理论和假设,而证伪在实际研究中具有暂时性。一个没有被现阶段数据和统计技术所证伪的假设,在将来有了新数据、新方法之后还要接受持续的检验。因此,本着"证伪"的目的,数据的不完美并不是避免使用定量方法的理由。

(四)定量方法是最科学的研究方法

这一误解来自定量方法的热情鼓吹者。他们将提倡科学方法等同于提倡使用定量方法,甚至将定量方法的使用看成是判断一个研究是否为"科学研究"的标志。但是,科学研究本身是由一套程序来定义的,而这套程序与定量方法相关的部分只有实证检验这一个环节;即使是在实证检验中,定量方法也只是理论检验的方法之一。定性方法,如比较案例研究、过程追踪、田野调查、访谈、档案分析等,都可以成为理论检验的方法,也是科学研究中的常用方法。而何种检验方法更好,是由研究的具体问题、理论和实证信息(数据的形式)情况所决定的。定量方法由于具有高度的透明性、可重复性和系统性享有很大优势,但这并不意味着定量方法在具体的研究中总是最好的选择。

(五)定量方法就是从数据分析中得出解释

这种误解认为定量方法就是用数据说话,在数据分析的基础上得出对问题的解释。持有这种认识的不仅是定量方法初学者,一些较为有经验的定量研究者也经常流露出这种数据挖掘(data-mining)的认识。

这一误解在大数据时代更有其肥沃的土壤。在当今信息爆炸的时代,使用统计方法和计算机软件对海量数据进行信息提取和分析,被证明具有很高的预测准确性和实用价值。一些政治科学家也在从事数据挖掘和机器学习的研究。① 但是,即使是基于海量数据的数据挖掘,也只是对相关关系进行发现,而不能发现因果。因果关系在认识论上有特定的要求,本质上是演绎性质而非归纳得出,而数据挖掘从本质上是归纳性质的研究。② 一

① 例如,Justin Grimmer and Brandon Steward, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis*, Vol. 21, No. 3, 2013, pp. 267-297。

② Stephen L. Morgan, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, Cambridge: Cambridge University Press, 2007.

些定量方法的研究者不从演绎的逻辑和理论推理出发得到理论,却从数据和模型出发,基于实证分析的结果来得出某种解释,这是对定量研究方法的错误使用。避免从数据和统计结果出发来建立理论或发展解释,严格遵守定量方法是一种理论检验而非理论发展(发现)的方法,这是一条必须遵守的重要原则。

(六)统计技术越高级、统计模型越复杂,定量方法就越可靠

定量研究者中不乏一些技术论者。这些人简单而武断地凭借统计技术的高低来判断定量研究的质量。定量方法作为工具,要为研究目的服务。统计模型的选择和统计技术的运用,由研究目的和数据形态决定。方法的优劣不可以从方法本身来判定,而要以它是否服务于目的来评估。当研究者具有简明的理论假设和高质量的数据,可能简单线性模型和多元回归就是最好的方法;而研究者如果在理论部分和数据搜集上漫不经心,期待用复杂的统计模型和技术来做出高质量的研究则几乎是不可能的。在定量研究中,若简单模型够用则尽量使用简单模型。研究者不应为了要造成深奥的印象或炫耀统计技术而舍近求远,选用复杂的模型和技术。

三 实证研究中的定量方法

人们通常的印象是,定量研究就是数据处理和统计软件运用,在学术论文或专著中表现为使用大量数学公式、统计图表和统计术语的运用等。数据、软件、数学公式、图表、统计术语等确为定量方法的表征,但这些外在的表现并不是定量方法的定义特征。

(一)什么是定量方法

定量方法是对经验数据(包括实验性数据和观察数据)进行统计推论、从而对理论假设进行检验的过程。定量方法服务于实证主义研究。实证主义研究承认客体与主体的分离,以分析为主,旨在对客观世界进行发现。[①] 实证主义从本体论和认识论上都区别于社会科学研究中的"阐释"或"批判"传统。作为实证研究的一种,定量研究的方法和思想不适用于阐释性和批判性研究,只有采取实证主义的本体论和认识论,才可能、也才有必要遵循定量的研究路径和方法。但是,定量研究只是类型众多的实证主义研究的一种,只要是承认研究的目的是为了认识客观世界现象而非在规范意义上对现象进行阐释和批判的,都是实证主义研究,而无论研究是采用了定性、定量还是历史的方法。[②]

---

① Steve Smith, Ken Booth and Marysia Zalewski, eds., *International Theory: Positivism and Beyond*, Cambridge: Cambridge University Press, 1994.

② Willnat Brains, et al., *Empirical Political Analysis*, New York: Person Education, Inc., 2005, p. 82.

要正确认识定量方法在科学研究中的作用,有必要对科学研究做进一步说明。演绎和证伪是科学方法的两大决定性特点。理论并非完全来源于对现实观察的归纳。基于再多事件的观察也不可能上升为普遍的理论,理论的得来必然是演绎的结果,而理论的源泉究竟是现实世界还是某种神秘的灵感,科学方法论倾向于不予讨论。科学研究理论的演绎性质决定,定量方法不用于理论创造。这一点听上去也许让定量方法的推崇者感到沮丧。不过,科学研究是一个循环的过程,当定量方法发现了与理论相悖的事实时(或称为证伪),研究者必须重新思考并修正理论。但理论的修正仍要经历一个重新演绎的过程,而不能够基于定量分析的实证结果,也即不能基于对现实的归纳。①

定量方法是科学研究中进行假设检验的方法之一。科学方法主要由一套明确的程序来定义,主要是根据这套程序来发展解释性和演绎性的理论假设并用实证证据来检验理论假设。② 定量方法服务于科学研究,但科学研究不一定只使用定量方法,因为定性方法也可用于假设检验。定量方法只对理论进行检验,这是科学理论演绎特性的要求。这一点严格区分了使用定量方法的科学研究和进行数据挖掘的统计研究。在社会科学尤其是国际关系学中,数据挖掘被认为是非理论性的,也是非科学的方法。尽管大数据时代的到来让一些学者开始从数据出发,发现重要的相关关系,以便更好地进行预测,但因果解释仍然是社会科学包括国际关系学的真正追求。在科学哲学和方法论的意义上,关于证明与证伪以及证伪的判定等存在着长期的争论。③ 并非所有人都认同卡尔·波普尔(Karl Popper)的证伪思想,也并非所有用于科学研究的方法都意在证伪。但是,定量方法却严格地用于证伪,这在很大程度上是由于定量方法目前使用的统计检验思想和原理本就是严格建立在演绎认识论基础上的证伪方法。④

具体来说,定量方法开始于待检验的假设,该假设是关于概念之间清晰的因果关系,而这种关系通常是或然性(probabilistic)而非决定性(deterministic)的。对假设中的概念进行概念化后,定量方法要求把"概念"翻译为"变量",从而将理论问题转化为统计问题。概念转化为变量后,即可以进行测量,测量要求系统化和标准化。确定了

---

① Karl Popper, *The Logic of Scientific Discovery*, Florence: Routledge, 2002; Thomas S. Kuhn, *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press, 1996.

② 阎学通、孙学峰:《国际关系研究实用方法》,北京:人民出版社 2001 年版。

③ Imre Lakatos and Alan Musgrave, eds., *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press, 1970; Imre Lakatos, *The Methodology of Scientific Research Programmes: Philosophical Papers*, Vol. 1, Cambridge: Cambridge University Press, 1978.

④ Erich L. Lehmann, "The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?" *Journal of the American Statistical Association*, Vol. 88, No. 424, 1993, pp. 1242-1249; Johannes Lenhard, "Models and Statistical Inference: The Controversy between Fisher and Neyman-Pearson," *British Journal for the Philosophy of Science*, Vol. 57, No. 1, 2006, pp. 69-91.

测量方法之后,研究者对现实进行观察和信息收集,形成数据,这些数据被称做变量的体现(realization)。变量可以是定量的变量或定性的变量,但在定量研究中,一概采取数值形式。定性变量的取值没有数值上的意义,只有类别或排序意义。数据的类型由变量类型决定,而数值则由明确的测量系统生成。研究者进而对数据进行统计分析,无论是复杂或简单的分析,其目的和最终结果都是对假设进行检验,并根据事先确定的标准,做出是否拒绝(reject)假设的决定。整个定量研究程序,要求概念定义-测量-数据生成-数据分析-假设检验等各个环节保持高度的透明,并以其他研究者能够重演(replicate)这一程序并取得相同结果为要求——即无论什么样的研究者,如果遵从该研究给定的定义、测量方法、数据生成和数据分析方法,能够得到相同的结果。

(二)什么不是定量方法

定量方法必须建立在数值数据的基础上,但不是所有对数值数据进行分析的研究都是科学研究意义上的定量方法。定量方法必须对具有一定普遍性和演绎性的理论进行实证检验。一些研究尽管进行了大量的数据分析并运用了复杂的统计模型,却由于缺乏对演绎性理论的追求而不能够称为定量研究。常被误认为是定量研究、但实际上却不是的主要有以下两类:

第一类是仅止于数值描述或统计描述的实证研究。这类研究通常可见于各种报告,如《中国对外援助白皮书》、《世界银行发展报告》等,或是案例分析、过程追踪或历史回顾中的大量数据描述,例如,彼得·克拉格伦德(Peter Kragelund)关于几类非传统援助国的介绍和比较中,运用了大量的援助分布数据,文中也有大量图表对数据进行归类和分析,但这一研究属于描述性研究而非定量研究。① 这类研究运用了大量的数据,有些数据甚至是原始数据,而且也对数据进行了一定的统计分析,如平均值、标准方差、增长率等,也有大量的图表报告,但是,数据描述和分析在这些研究中作为对事实的陈述和分析,不含统计推论,因此不是定量方法。当然,在定量研究中,往往要给出统计描述,但统计描述本身并非出示任何实证证据,而在于增加数据特征的透明度和便于其他研究者评判统计模型的选择是否恰当。定量研究不止步于统计描述,而必须进一步进行统计推论。统计推论即是从样本到总体的推论,从可观察到不可观察的推论,从已知到未知的推论,而推论的目的是要进行假设检验。

第二类大量使用数值数据和复杂统计方法、但却并非定量研究方法的研究是计算机模拟。这类研究使用计算机生成的数据进行分析,由于数据不是实证数据,并非来自于

① Peter Kragelund, "The Return of Non-DAC Donors to Africa: New Prospects for African Development?" *Development Policy Review*, Vol. 26, No. 5, 2008, pp. 555-584.

现实世界,因此这类研究不是实证研究,使用的统计方法就不能够被称为定量方法。进行模拟的主要目的在于发现理论中的逻辑谬误或统计模型中的技术问题。因为数据的生成过程完全由研究者设计并控制,研究者确知在这些数据生成过程中的各种参数值,运用研究者发展的理论模型或统计模型去推论这些参数值,可以检验这些理论和模型在逻辑或技术上的谬误。但定量研究要求运用实证数据,即是研究者无法确知的实际过程,来检验理论,这种检验并非检验理论或模型自身内部的严密性和技术误差,而是检验理论与现实是否相矛盾。计算机模拟在研究中具有重要性,因为一个内部具有逻辑矛盾的理论,在进行实证检验前就可以被淘汰或修正,或者一个具有技术问题的统计模型会造成检验偏差,这些都可以在模拟研究中得到排除,从而使得定量分析的假设检验更少犯错。这类方法在统计学、工程学、经济学及其他学科(包括政治学)已经得到广泛运用,但其目的是为了提高定量研究的质量,本身并非定量研究。笔者于 2010 年发表于《政治分析》(Political Analysis)杂志上的文章发展了分析长面板数据的非线性动态模型,用大量的计算机生成的数据进行蒙特卡洛试验,目的在于检验新模型和算法的可靠性及其相对于传统模型的优势,尽管这样的研究使用了复杂的统计知识和大量的数据,但仍然不能算是实证研究;其中使用的方法可以称为统计方法,但却不是"定量方法"。① 大量的统计学、计量经济学和政治学方法的论文都属于此类,不是严格意义上的定量研究。

四 定量方法的实施步骤

对研究方法的理论探讨虽然有其重要性,但方法只有在使用中才有其价值,也只有在操作中人们才能够真正理解其背后的本体论和认识论基础。本文这一部分对定量方法的使用步骤进行逐一讨论,希望在讨论中进一步对上文中提到的一些常见误解进行澄清,同时也为初学者提供一个定量方法使用的简单指南。

定量方法有一套相对确定的程序,每一个步骤都需要遵循一些具体的原则。需要强调的是,没有任何理由相信,定量研究比定性研究甚至描述性、阐释性研究更接近真理;也没有任何理由认为,定量研究比其他研究方法更可能排除谬误。定量研究方法的真正优势在于它透明程度高、较容易为研究同行重演并评判,从而有助于凝聚学术共同体的智慧和努力,以推进科学发展的进程。对定量方法在使用过程中所要遵循的

① Xun Pang, "Modeling Heterogeneity and Serial Correlation in Binary Time-Series Cross-Sectional Data: A Bayesian Multilevel Model with AR(p) Errors and Non-Nested Clusterings," *Political Analysis*, Vol. 18, No. 4, 2010, pp. 470-498.

原则和准则,学界具有高度的共识,即对程序公开与信息透明的要求。

下文在讨论定量研究方法时将比较多地运用到两个国际关系定量研究的实例。其中一个是国际安全方面的研究,弗吉尼娅·福特纳(Virginia Page Fortna)发表在《国际研究季刊》上的《国际维和行动是否有助于维护内战后的和平状态》(简称《维和与内战》)。① 另外一个实例是海伦·米尔纳(Helen V. Milner)和久保田惠子(Keiko Kubota)发表在《国际组织》上的一篇国际政治经济学的文章《为什么走向贸易开放?发展中国家民主化和贸易政策转变》(在下文中用《民主化与贸易开放》代指),主要分析民主化对贸易政策开放程度的影响。② 选取这两个例子的主要原因是它们发表的刊物为国际关系学顶级刊物,并且两篇文章分属国际安全和国际政治经济学两个领域;同时,它们有较高的引用率,是很多国际关系学研究生课程的必读材料。需要指出的是,在政治学方法论上,这两个例子的研究并非完美,在本文中不是作为范例,而是作为实例来帮助说明定量研究的步骤。

(一)第一步:确定研究问题是否适用定量方法

在此必须再次强调,定量方法适用的研究问题非常有限,并具有特定的要求。许多极具重要性的研究问题并不能使用定量方法来研究。定量方法不适合一切规范性问题,比如解答"什么是公正的国际政治经济秩序"、"人权高于主权还是主权高于人权"、"发达国家是否负有在经济上援助不发达国家的义务"等问题。此外,由于定量方法主要用于对具有一定普遍性的、在一定程度上重复发生的现象的解释性研究,因此如果研究问题是还原一个特定事件本身或对特定现象进行个别解释,则无法使用定量研究。例如,古巴导弹危机的决策过程是怎样的?什么因素导致了冷战结束?等等。这些研究问题要求对具体的事件(现象)进行还原或做出具体的解释。它们的研究结果也许能够帮助解释相似事件,但研究本身并不追求普遍性。此类问题最常见的(也许是最合适的)研究方法是历史研究的方法,对史料进行鉴别分析和综合归纳。

第一,定量方法只适用于特定的研究问题。这类问题必须是关于具有一定抽象性的、可测量的概念之间的关系。波普尔所定义的"概念"必须是具有绝对普遍性的概念。但是在实际的国际关系研究中,定量方法往往只要求具有一定抽象性的概念——一个概念是多个现象的抽象,但在范围上可以有时空限定。如战争是一个具有普遍性的概念,显然符合定量研究的要求。但是,中美贸易不是一个普遍性的概念,因为它受

① Virginia Page Fortna, "Does Peacekeeping Keep Peace? International Intervention and the Duration of Peace after Civil War," *International Studies Quarterly*, Vol. 48, No. 2, 2004, pp. 269-292.

② Helen V. Milner and Keiko Kubota, "Why the Move to Free Trade? Democracy and Trade Policy in the Developing Countries," *International Organization*, Vol. 59, No. 1, 2005, pp. 107-143.

到“中美”的限定,不符合波普尔对概念的定义。但在实际研究中,中美贸易却具有一定的抽象性,它可以包含自中美有贸易关系以来每年的相互贸易活动,定量研究可以用于包含这类概念的研究问题。

第二,定量方法要求概念具有可测量性。一些研究问题涉及难以测量的概念,进行定量研究前,必须将这些概念转化为可测量的概念(或者研究者必须找到合适的代理概念)。例如,研究问题是文化传统如何影响国际合作。文化传统包含太多的意义,是一个难以测量的概念,如果可以用其他可测量的概念作为代理概念,定量方法才能用于研究此问题。

第三,定量方法要求研究的概念间关系具有清晰性、方向性和明确性。关系可以是线性、非线性、独立关系或条件关系等,可以具有或然性,但必须清晰而明确。线性要具有明确的方向,非线性要具有变化特征,条件关系要明确条件是什么以及条件的影响方向等。如果概念之间的关系错综复杂、枝节横生,以至于从因到果的路径无法给出一个可以明确陈述的简洁关系,这样的研究就不适合用定量研究。

(二)第二步:将理论形成待检验的假设

研究中的核心理论往往包括不止一环的因果链条,定量方法可以用于检验所有的因果环节,也可以只检验两端概念之间的关系。比如,理论“概念 X 导致概念 Y 是通过对概念 Z_1 和概念 Z_2 的顺序影响”,即因果链条为 $X \rightarrow Z_1 \rightarrow Z_2 \rightarrow Y$。使用定量方法进行假设检验时,本着证伪的原则,待检验的假设可以只为 $X \rightarrow Y$,中间环节可以不加检验。但是,如果需要证伪理论的整个逻辑链,就要将每一个链条形成一个假设,即待检验的假设有三个:$X \rightarrow Z_1$, $Z_1 \rightarrow Z_2$, $Z_2 \rightarrow Y$。定性研究注重整个因果过程研究(如过程追踪方法),但大部分定量研究只是对两端概念之间的关系进行检验,因为定量研究的主要目的不是“证明”而是“证伪”。如在《维和与内战》一文中,待检验的核心假设是“维和行动延长了内战后的和平状态”;而《民主化与贸易开放》的待检验核心假设为“发展中国家政治民主化带来更为开放的贸易政策”。这两个假设背后的理论都有不止一环的因果链条。例如,在“发展中国家贸易开放将有利于劳动力要素所有者”这一前提下,《民主化与贸易开放》一文的理论逻辑链条为“政治转型(X)导致劳动要素禀赋所有者政治影响力上升(Z_1),导致更开放的贸易政策(Y)”。如果要“证明”的话,X 到 Z_1 和 Z_1 到 Y 都必须要检验。但证伪则可以只检验 $X \rightarrow Y$。

确定了待检验的假设之后,定量方法的使用者需要将理论假设用统计假设的形式表达出来。虽然在论文写作中这一部分通常被省略,但待检验的假设要求能够表达成以下数学形式:

$$E(Y)=f(X)$$

其中,$E(Y)$ 是待解释现象的期望值。在定量研究中,一般只对现象的普遍特征 $E(Y)$ 进行解释,而不解释具体的一个现象(表示为观察到的 y_i),这也是为什么定量研究必须是有一定程度普遍性的研究——对一类现象的共有特征而非一个现象的特殊部分进行研究。公式中,$f(X)$ 是 X 的函数,而方程的性状就是待解释的现象 Y 和解释现象 X 之间的关系。研究者的理论是对这种关系是什么($f(X)$ 的性状)以及为什么的研究,而待检验的假设就是用观察到的具体现象 $y_1,y_2,\cdots,y_n$ 及相应的 $x_1,x_2,\cdots,x_n$ 来检验 Y 和 X 之间的关系是否为 $f(X)$。因此,函数 $f(X)$ 是待检验的假设。

这里,我们又回到定量研究能够胜任的理论必须十分清晰并高度简化这一点上。待检验的假设是 Y 和 X 之间的关系 $f(X)$。要使用统计方法对这一假设进行检验,方程 $f(X)$ 的性状不能过于复杂,否则将无法进行统计上的假设检验。使用定量方法的研究,通常情况是对 X 的变化带来 $E(Y)$ 的增减进行理论、逻辑和实证研究,在数学上面表达为,函数是 X 的增函数还是减函数,也即是其一阶导数是正还是负:

$$\frac{df(X)}{dX}>0 \text{ 或 } \frac{df(X)}{dX}<0$$

最简单的假设检验就是假设 $E(Y)$ 和 X 之间具有线性关系 $E(Y)=f(X)=\beta X$,其中 β 是未知参数。这样,检验 X 和是否对 Y 有影响就是检验假设 $\beta=0$,X 是否对 Y 有正面或负面的影响,即为检验假设 $\beta>0$ 或 $\beta<0$。当然,定量方法可以检验一些比线性关系更为复杂而动态的理论假设,但是研究者必须控制概念的数量以及尽量简化概念之间的关系。在《维和与内战》和《民主化与贸易开放》两个研究中,待检验的假设都转化为对 X 的系数是否为正的检验。①

(三)第三步:变量测量、样本抽样和数据生成

数据(data)在英文中通常是信息的同义词,或指信息的载体。数据本身并不一定要采取数值(numeric)形式,而且以数值形式表现出来的数据也许并没有数量意义。但在定量研究中,数据专指数值数据,是变量的具体取值。这就是说,数据的背后是变量,变量的背后是概念。数据信息不在于数值本身,而在于通过可观察到的数据,研究者可以对感兴趣的变量或概念进行实证认识。因此,数据取样的质量在于它是否反映了变量的分布特征。由于变量的分布特征对研究者来说通常未知,比如待解释现象(因变量)的均值 $E(Y)$ 是变量分布的重大特征,但它却是未知的,它的取值及变化原

① 两篇文章中使用的都是非线性模型。非线性模型中 X 对 Y 的影响往往不仅仅取决于其参数,但是由于对其他因素进行了控制,关于变量之间关系的假设检验仍然可以简化为对系数的检验。

因正是研究的目的和任务。如果我们得到的数据 $y_1, y_2, \cdots, y_n$ 能够反映出变量 Y 的分布,那我们就能够从可知到未知进行推论。怎么才能知道一组具体的数据 $y_1, y_2, \cdots, y_n$ 是否反映了变量 Y 的分布特征?随机抽样和样本大小是核心。随机抽样是保证数据对变量分布具有良好代表性的黄金法则,而样本越大这种代表性越准确。随机样本和样本规模直接影响到样本数据质量及之后的假设检验。①

测量误差也是影响数据质量的重要因素。测量首先是要基于概念的定义,概念化是测量的前提。概念化的好坏直接决定了测量的质量。测量本身可能具有误差。测量的误差及其处理是个重要而棘手的课题,研究者必须仔细考察数据是否有测量误差,有什么类型的测量误差,从而采用适当的统计方法加以处理和纠正,或者进行重新测量。②

在任何情况下,生成过程和标准不明确、不透明的数据都不能使用。由于数据的来源、测量和抽样对同行评判一个定量研究极其重要,定量研究的研究者要求尽可能提供其使用的数据信息。惯例上,由于篇幅的限制,研究者在研究论文中简要报告变量定义、测量单位、数据分布信息及数据来源,更为详细的信息,包括使用的具体数据及其编码手册、数据来源等信息在别处公开,供其他研究者检查和使用。大部分研究者会在自己的网页上将这些信息长期公开。例如,从《维和与内战》和《民主化与贸易开放》的作者网页上都可以得到编码手册和原始数据。③

(四)第四步:数据处理和描述性统计分析

定量研究者获取数据样本后将对数据进行整理和处理。研究者需要意识到,数据的生成过程可能发生各种各样的测量和编码问题,有些由编码者个人疏忽造成,有些是系统性的测量误差,有些是数据编码的习惯不适合研究者所使用的统计软件。缺乏数据处理经验的定量研究者常常将他人生成的数据直接用于定量分析和假设检验中,得到错误的或偏差的结论却难以自知。

这一步的每一种工作都涉及特定的统计知识和技术,一本好的数据分析手册或教材,必然包含如何对数据进行前期处理。④ 囿于篇幅,本文不能详细阐述和介绍,在此只列出下面一个大致的工作清单,并加以简要说明。

---

① 关于随机抽样的入门介绍,可参见 Alan Agresti and Barbara Finlay, *Statistical Methods for the Social Sciences*, Upper Saddle River: Prentice Hall, 1999, chapter 2。关于取样的初步介绍,可参见 Earl Babbie, *The Practice of Social Research*, Belmont: Wadsworth, 2010。

② 关于测量误差的种类及其影响,可参见 Jeffrey M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press, 2001, chapter 4.4。

③ http://www.columbia.edu/vpf4/research.htm,登录时间:2013 年 12 月 18 日; http://www.princeton.edu/hmilner/published.html,登录时间:2013 年 12 月 18 日。

④ Julian J. Faraway, *Linear Models with R*, Boca Raton: Chapman & Hall/CRC, 2005.

第一,将数据储存为计算机可以正确识别的形式。初学者往往忽视以数值形式出现的数据有定量数据(quantitative data)或种类数据(categorical data)两类,种类数据又根据种类是否有某种序列(如程度、级别等)关系而有所区分等。如不进行性质上的处理,在很多情况下,计算机将一律把这些数据看成定量数据,给后来的分析造成很大偏差。

第二,对数据编码进行调整。如缺失数据在一些数据库中被记为"999"、"99"、"9"或"0",这是供一些较早的统计软件识别的编码法,而使用不同的统计软件,缺失数据的机器识别不一样,研究者要根据自己使用的软件改变这些编码,否则缺失数据代码"999"将被一些软件读为数值九百九十九。比如,使用 R 软件的研究者,需要将缺失数据编码一律置换成"NA",计算机才知道这些是缺失数据。除了缺失数据,还有其他一些编码问题,研究者需要根据数据生成者提供的手册一一仔细检查。

第三,缺失数据的估计和补全。为了尽量减少数据信息损失和缺失数据处理不当可能带来的选择偏差,研究者应该使用各种"多重填补(imputation)"或"数据增广(data augmentation)"的方法补齐不同数据结构中的缺失数据。①

第四,对单个变量的数据分布情况进行分析,以对样本的质量和数据问题进行评估。在实践经验和统计知识的基础上,研究者一般对将要使用的变量有一个基本的认识,如大致知道其分布的范围和形状等。例如,对数据的分布检查发现,取值范围只能是正实数的变量(如国内生产总值)出现为负的数据,或者代表比例的变量获得了大于 1 的取值等,研究者必须对数据的测量和编码进行再检查。数据初步检查的部分结果——尤其是变量的数据分布情况——在定量研究中一般会在正文中报告,如《民主化与贸易开放》在第 121 页上的描述性统计特征表,包括变量实证分布的均值、标准差、极大值、极小值等。②

第五,变量间相关性的简单分析。在模型估计和统计检验前,定量研究要对变量之间简单的相关关系进行一个初步检测。我们希望所选择的解释变量与因变量之间具有较强的相关关系,但却不希望因变量之间的相关性太强。因变量之间的高度相关性实际上是两组数据所包含的信息雷同,对因变量的解释重合范围大,在统计上则会

---

① Donald B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley-Interscience, 2004; Joseph L. Schafer and Recai M. Yucel, "Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values," *Journal of Computational & Graphical Statistics*, Vol. 11, No. 2, 2002, pp. 437-457; Xiaowei Yang, "Markov Transition Models for Binary Repeated Measures with Ignorable and Nonignorable Missing Values," *Statistical Methods in Medical Research*, Vol. 16, No. 4, 2007, pp. 347-364; James Honaker, et al., "What to Do about Missing Values in Time Series Cross-Section Data," *American Journal of Political Science*, Vol. 54, No. 3, 2010, pp. 561-581; Gary King, et al., "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review*, Vol. 95, No. 1, 2001, pp. 49-69.

② Helen V. Milner and Keiko Kubota, "Why the Move to Free Trade? Democracy and Trade Policy in the Developing Countries," p. 121.

造成较严重的技术问题。《维和与内战》一文中的表 1 到表 3 都是对维和行动与战后和平的持续时间进行简单的相关性分析。①

(五)第五步:选择或建立统计模型

研究者根据待检验的理论假设、与研究问题相关的理论以及变量特征建立恰当的统计模型。笔者常常被问到一个问题:"我们这个研究的定量分析部分要求不高,能不能拿个现成的统计模型来分析?"此类问题是出于对统计建模的不了解。针对任何一个具体的研究都没有所谓"现成"的统计模型。定量研究中的统计模型实际上只是包含待检验的理论和相关替代理论的统计(数学)表达。没有对研究现象的深入思考,没有对相关理论及其关系的全面分析,研究者无法进行高质量的统计建模。一些定量研究的初学者不懂得统计建模需要理论支撑,人云亦云或照猫画虎地使用模型,模型质量低下,导致假设检验得出错误或偏差的结论。统计学家和计量经济学家只建立普遍的、抽象的模型,其主要任务也不在于"建模",而在于发展模型估计的方法和对新估计量的性质进行探讨。对具体问题的定量研究和统计建模,要求研究者对研究问题进行深入参考和全面了解,因为任何一个具体的模型都是一组具体的相关理论的数学体现。

《维和与内战》运用了生存模型(survival model),而《民主化与贸易开放》以关税水平测量贸易开放程度时运用了面板线性模型,在用高盛集团的开放指标(开放与不开放)测量开放程度时却使用了条件 Logit 模型。这两个研究的模型都包含了十几个解释变量。那么研究中的模型是如何建立起来的呢?下面对建模的基本环节和任务进行一个简单的说明。

抽象的统计模型由三部分构成:变量、解释变量的函数和残差项(error term),也即

$$Y = f(X) + \varepsilon$$

这个抽象的模型在具体的研究中通过完成以下三个任务而得以具体建立:

第一,变量及变量选择。在上面的抽象模型中,Y 称为因变量。在建模过程中,Y 的选择比较明确,就是待解释的现象。X 称为解释变量,通常包括多个变量。变量选择问题主要是选择 X 的问题。显然,待检验的核心假设中包含的解释变量应该被包括在 X 中,因为定量研究的目的就是要检验它们和 Y 之间的关系。变量选择的难点主要在于,除这些待检验核心假设中的变量外,还有什么样的变量应该被包含在模型中?

那些进入模型的其他变量通常又被称为"控制变量"。控制变量的概念来自实验性科学研究。在实验室条件下,研究者可以控制一些条件,以期精确观察某一条件的变化

① Virginia Page Fortna, "Does Peacekeeping Keep Peace? International Intervention and the Duration of Peace after Civil War," pp. 272-274.

如何影响到现象的变化。为什么要控制这些因素呢?“控制”的意思就是让这些可能影响结果的因素“不变”,而不变的因素是不可以用来解释变化的,于是就排除了现象变化的他因。但是,在非实验室条件下,我们无法控制一些因素的变化。把这些可能的影响因素包括在统计模型中,可以达到类似的“控制”作用——至少研究者可以观察到这些因素如何变化并将其变化的影响进行分离,从而得到待检验假设中的变量对因变量的影响。因此,选择控制变量的两大原则是:(1)这些变量可能影响到因变量的变化;(2)这些变量和核心解释变量相关。一个变量如果同时符合这两个原则,就需要进入模型成为控制变量。第一个原则比较容易理解:如果一个因素与结果无关,它的变化也就不会影响到待检验的关系。第二个原则意味着,如果一些因素影响到了结果,但这些影响平行于我们要集中考察的影响机制,那么不考虑这些因素也不会影响我们对核心变量的变化如何影响到结果的判断,因此模型没有必要包含这些因素。具体如何确定哪些变量需要控制,除了这两大普遍原则外,还有以著名后门准则为代表的因果图形分析法和一些统计方法。①

《维和与内战》和《民主化与贸易开放》两篇文章都用了大量的篇幅来讨论相关替代性解释,从而确定控制变量的选择。《民主化与贸易开放》详细阐述了国际政治经济学文献中关于贸易开放的国际因素、决策者因素和时机因素(危机等),这些是与政治转型相联系而又不同的解释。这些替代性解释在其建模中都体现在了控制变量的选择上。《维和与内战》更是在理论和逻辑上深入讨论了对内战后和平持续情况的其他解释以及这些解释性因素和维和之间的相关性,如内战的烈度、持续时间和内战性质等因素,既影响到战后和平的维持,也影响到国际是否对内战进行干预的维和行动。研究者在建模中充分控制了体现这些相关替代性解释的变量。

第二,对因变量和自变量之间的关系进行大致假定。这在统计上就是对上文中提到的抽象模型中函数 $f(X)$ 的确定(假定)。关于函数 $f(X)$ 的选择,部分取决于因变量的分布特征。比如,如果 Y 只能有 1 和 0 两种取值,$f(X)$ 则不能是线性函数。在《民主化与贸易开放》一文中,高盛集团的开放指标是两分变量(dichotomous variable),因此作者采用了相应的非线性函数假设,而在《维和与内战》一文中,因变量是战后和平状态的持续时间,一般这种因变量性质意味着生存模型的函数假设比较合理。

当然 $f(X)$ 的形状可以根据实际的理论而复杂多变,模型建立时也可以找到适当的函数来表达这种复杂多变的关系(如非参数模型)。但是,定量研究者们一般尽量避免使用复杂的函数。过于复杂的函数假设,不但会带来下一步模型或参数估计中的技术困难问题,更重要的是,复杂的函数形状意味着复杂的关系,使得对实证结果分析

① Judea Pearl, *Causality: Models, Reasoning and Inference*, Cambridge: Cambridge University Press, 2009.

解读也变得困难。

第三,划定研究的解释范围。任何统计模型都要包括一个残差项 ε。这是一个随机变量,包含了所有对结果具有影响、但模型不加以解释或控制的因素。这些因素往往是无法观察或无法测量的,或者这些因素与模型要考察的那些影响因素 X 无关。这一项划定了研究的范围——在社会科学研究中,任何研究都不可能穷尽所有的影响因素,而特定的研究只是集中在对为数很少的因素进行分析和控制。有了这一随机项,我们的理论才能够对现实进行大量简化,它允许我们把许多因素排除在研究的解释范围之外。同时,这一随机项也决定了研究发展出来的理论都是或然理论——理论只对现象的平均趋势进行解释,而无法(也不追求)对一时一地的具体现象进行精确解释,因为任何个体的解释和预测必须要加入残差项 ε,而这一项是未知的、随机的。

(六)第六步:参数估计和假设检验结果

统计模型将研究者发展的理论连同其他相关的可能解释进行一定程度简化后,以数学的形式表现出来。但是模型本身对理论持中立的态度,将研究者认为的因果关系全都表现为未知的参数。例如下面这个简单的多元线性模型:

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

这个模型只表明,因素 $X_1, X_2, \cdots, X_k$ 对 Y 有线性关系,但由于 $\beta_1, \beta_2, \cdots, \beta_k$ 全是未知参数,因此模型对这些因素是否有影响以及有什么影响持开放性态度。一个变量被包含在模型中,也不意味着它一定对因变量 Y 有影响——如果参数 β_1 实际等于 0,那么 X_1 对 Y 没有影响。同时,所有参数 β 的符号也未知—— X 对 Y 究竟有正向或负向的影响,模型不进行假定。另外,模型也不确定参数的大小,即不回答 X 对 Y 影响的大小。但是,只有知道了这些参数的大小和正负,理论或假设才最终得到了表达。而要得到关于这些参数的知识,就需要分析数据里蕴藏的信息,这就是参数估计——运用数据信息,对模型中的未知参数进行估计,从而获得关于因变量和自变量关系的信息。

通过一定的统计估计方法,我们可以得到参数的估计值 $\hat{\beta}$。这个值是否就是定量分析所得到的 X 和 Y 之间的关系呢?回答是否定的。例如,参数估计值 $\hat{\beta}_1 = 3$,即使是这个值"统计显著",也不表明在其他条件不变的情况下,X_1 增加 1 个单位导致 Y 增加 3 个单位,甚至不表明 X_1 和 Y 之间具有因果关系。这听上去有些违背人们的通常认识,但却是对模型估计的正确认识。定量分析没有"证明"功能,得到的参数估计值也就不是因变量和自变量之间的关系。参数估计量(包括估计值和标准差)只能用来证伪核心假设中两者的关系,也即我们所说的定量假设检验。

研究者的核心假设往往是关于什么因素(X)的上升导致 Y 的上升(下降)。在线性

参数模型中,这一假设就可以简单地表示为$\beta > 0$或$\beta < 0$,这也是定量研究者在整个统计分析中要试图证伪的假设。而参数估计值$\hat{\beta}$及其标准差$se(\hat{\beta})$正是用于对关于参数β的假设的检验。当然,研究者都希望自己的核心假设能够经得起实证检验,即"没有被证伪",但即使"未被证伪"也不等于"被证明"。目前得到最广泛运用的统计假设检验的方法是元假设显著检验或基于p值的检验。元假设显著检验方法需要设定"元假设"(H_0)和"替代假设"(H_1),两者是非此即彼的关系(补集),也就是拒绝H_0就等于接受H_1,反之亦然。而基于p值的检验不需要设定替代假设,因此如果用p值检验,则一般不用"接受假设"这样的说法,只有"拒绝"与"不能拒绝"的区别。基于数学上的考虑,在两种假设检验中,元假设一般为研究者试图要"证伪"的假设,也即是与研究者的核心假设相反的假设。如研究者的核心假设是X对Y有作用,即$\beta \neq 0$,在假设检验中,元假设就为$H_0:\beta = 0$。研究者私心里期待被拒绝的是元假设。如果这一假设被证伪,在元假设显著检验里,研究者就可以接受与元假设具有非此即彼关系的替代假设$H_1:\beta \neq 0$,这也即是研究者的核心假设。在p值检验中,没有替代假设,但对元假设的拒绝就表明研究者的核心假设$\beta \neq 0$可以不被拒绝。

这里需要强调的是,统计中的假设检验并非是对"真理"和"谬误"的检验和区分,而是一种帮助研究者做出"拒绝某假设"这一决定的决策依据和理论。这一点在统计导论课中往往让学生感到吃惊和不习惯,而还有一些讲授统计课的教师不向学生澄清这一点,导致学生在之后的学习中一直持有错误的认识,即统计检验是对真假的检验。假设检验中的一些术语,如"显著程度"、"置信区间"以及p值,完全不代表对元假设或替代性假设为"真"或为"假"的概率或程度进行判断。从科学哲学和认识论上讲,一个假设只可能是"真"或是"假",谈论它们多大程度上或概率上是真是假是错误的。[①] 因此,术语"在1%的程度上统计显著"、"p值为0.001"或"99%置信区间",只表示一个意思:研究者如果拒绝一个假设,而这个假设实际为真,他犯这样的错误的概率有多大。这些指标只是决策论中的指标,而非对真理和谬误的评估,因为作为研究者,我们可能永远不知道真理在何处,也不知道我们离真理有多远。我们所能做的只是降低犯错误的概率。假设检验实际上只是以一种系统决策的过程,让研究者大致了解自己犯错的概率。当我们觉得拒绝假设$\beta = 0$,我们犯错(拒绝了一个真假设)的概率小于0.01时,我们做出拒绝元假设和接受替代$\beta \neq 0$决定的信心比较大,这也是为什么"置

① James O. Berger and Thomas Sellke, "Testing a Point Null Hypothesis: The Irreconciliability of P Values and Evidence," *Journal of the American Statistical Association*, Vol. 82, No. 397, 1987, pp. 112-122; Eduard Brandstatter, "Confidence Intervals as an Alternative to Significance Testing," *Methods of Psychological Research Online*, Vol. 4, 1999, pp. 32-46.

信区间”在英文里就是“confidence interval”。这里“confidence”是对所做决定的“自信”程度而非假设本身为真为假的程度。

在《维和与内战》一文中,核心假设是对维和(变量“peacekeeping”)的 β 是否为正的检验。文章对不同的模型进行了估计,也对不同类型的维和行动进行了分别检验。例如,表 7(第 284 页)中,我们看到对所有的维和行动进行分析,“维和”的参数估计值 $\hat{\beta}$ 是 0.32,标准方差 $se(\hat{\beta})$ 是 0.18。① 这里,估计值 0.32 除了是个正数之外,单独看没有任何意义。当我们将标准方差和参数估计值结合起来,根据一定的分布假设,就可以得到假设检验的结果,即拒绝维和的参数 $\beta = 0$(也即维和对战后和平的维持无影响)可能犯错误的几率小于 0.1。由于参数估计值是正数,如果不为 0 则为正,因此研究者可以得出结论说,根据这个模型和数据,文章的核心假设没有被证伪,我们可以比较有信心地暂时接受这样一种理论,那就是维和行动可以帮助维持内战后和平。

(七)第七步:模型假定的重新检查

有的定量研究者在建模、参数估计和假设检验之后,即结束了定量分析过程。而实际上,定量分析至此离结束尚早。理论对现实加以简化、统计模型对理论加以简化,都是建立在一系列前提假定的基础上。没有前提假定,就不可能有简化。但是,这些模型的前提假定是否合理以及这些假设是否对检验结论有重大影响,谨慎的科学研究者需要进一步检查,这通常被称为“鲁棒性”检查(robustness checking)。鲁棒性检查主要包括、但不限于以下部分:

第一,控制变量的不同选择。由于解释变量之间的相关性,不同变量放入或拿出模型,可能会影响到核心参数的估计值及其标准差的变化,从而影响到假设检验的结果。一些学者习惯于估计多个具有不同变量选择的模型,从而展示其核心假设检验结果的稳定性。在《民主化与贸易开放》(几乎全部的回归结果报告表)和《维和与内战》(表 9)两篇文章中都采用把一些控制变量拿出或放进模型中的办法来显示结果的稳定性。笔者认为,从建模的角度看,这种任意放入或拿出变量的方法并不可取,因为在前面选择控制变量时,我们有理论要求必须要控制这些相关替代性解释,否则将会影响到统计结果。如果为了检查鲁棒性而拿出这些变量,实际上是用错误的建模来检验参数估计值的鲁棒性。不幸的是,虽然这种做法并不可取,但似乎已经约定俗成,尤其在国际关系研究中非常普遍。

第二,使用变量的不同测量方式。同一个概念,在将其量化的过程中,可以使用不

① Virginia Page Fortna, “Does Peacekeeping Keep Peace? International Intervention and the Duration of Peace after Civil War,” p. 284.

同性质的变量,例如“贸易开放”这一概念,可以转化为“哑变量”(即开放与封闭),但也可以是一个连续变量(开放程度的连续变化)。同时,同一个变量可能有不同的测量体系或标准,如按照死亡人数的最低标准来定义武装冲突为战争或非战争,高于1000 人为战争,否则为非战争。在测量中,不同的战争伤亡人数门槛值可以生成不同的战争数据。谨慎的研究者在进行鲁棒性检查的时候,会检查不同的变量定义和测量方式是否影响假设检验的结果。《民主化与贸易开放》一文中,作者用了两种非常不同的关于贸易政策开放的测量,运用不同的测量方式,仍然得出了政体影响显著的结果,于是我们认为这样的分析相对于单一的测量方式要更可靠。

第三,对统计上的假定进行残差检验。这一检验主要是用模型估计返回的残差值,对模型在统计上的假定进行检验,看是否有理由怀疑模型假定的变量间关系形式与数据间关系相差过大,也即拟合优度(goodness-of-fit),是模型对数据的符合程度的一个反向检查。这些检查往往建立在估计结果得到的残差的分析上。一般情况下,限于文章的篇幅,这些检验不在正文中一一报告,但却是研究者必须要进行的环节。《民主化与贸易开放》和《维和与内战》在文中都没有报告这一部分,但我们倾向于相信他们做过这些检测。当然,在对一些学术研究进行重演时,我们经常在残差检查中发现一些严肃的问题,这意味着作者可能没有进行过此类检查。

第四,统计模型类型的不同选择。有一些统计假定是否影响到了统计检验的结果,还可以通过对不同类型的模型进行参数估计来观察。在建模和统计分析中,对同一问题往往有不同类型的统计模型可供选择。研究者根据模型的不同假定和要求,选取他认为最佳的模型进行统计检验。但在鲁棒性检查中,研究者会将大致适合该问题的统计模型进行估计,如果基于不同类型的模型而进行的假设检验的结果大致相同,则说明结果具有鲁棒性。在我们选择的这两篇实例中都没有进行这方面的检查。比如《民主化与贸易开放》一文中,面板分析可以有多种不同的模型类型选择,但作者没有检查是否其他类型的模型会得出不一样的结果。在米尔纳与笔者等人合著的《国际体系与国内政治:从国际关系中的复杂互动关系到实证模型》一文中,我们对该研究的核心假设以不同的模型类型进行了鲁棒性检查。①

总之,研究者对其定量研究中做出“拒绝元假设”(也即是研究者关心的核心假设经受住了实证检验,可以暂时保留)决定的“信心”程度或这一决定的可靠程度,不仅仅要

① Stephen Chaudoin, Helen Milner and Xun Pang, “International Systems and Domestic Politics: Linking Complex Interactions with Empirical Models in International Relations,” Conditionally accepted(有条件接受) by *International Organization*。

看基于主要模型的假设检验中的“统计显著程度”或“置信区间”,还要看基于不同变量类型、测度标准和模型假设,这一假设检验的结果是否具有稳定性。稳定性越高,研究者越有信心拒绝“元假设”,同时对自己发展的理论或假设越有信心。研究者应该乐于穷尽所有的可能检查,让其理论和假设经受各种可能的检验,来提高其理论的可信度。

(八)第八步:结果分析和解读

最后,研究者在让自己的假设经受了各种检验后,还必须对各种统计结果进行解读。对实证结果进行分析和解读,在定量分析中不仅不可缺少,而且非常重要。需要注意的是,这一步骤所要运用的不仅是研究者的统计知识和对测量与数据所掌握的信息,更要依赖研究者对其研究问题领域的专业知识。如果说在鲁棒性检验中,研究者的主要任务是检查检验结果是否在统计上可靠,那么在这一部分,研究者的主要任务是对统计结果进行实质性解读,看检验结果(包括控制变量影响的检验结果)是否在理论上或实质性意义上能够讲得通。如果出现理论上、逻辑上和常识上都难以理解的统计结果,研究者需要进一步怀疑自己的统计工具和数据及其测量的可靠性,并对此做出解释。

五 结论

随着定量方法在国际关系研究中的普遍运用,对定量方法的正确理解和使用对促进学科的发展具有重要意义。有的国际关系研究者对定量方法存在一些误解,错误和不当使用也不鲜见。我们需要明确的是,定量方法只是科学研究方法之一,而科学研究也只是学术研究中的一类研究形式和路径。定量方法本身并没有任何“魔法”,能够让研究更接近“真理”,相反,和其他任何方法一样,偏差的理解和错误的操作,都可能致使定量方法产生错误的研究结果。各种研究方法各有优势,也各有局限,使用者不但要清楚它能服务于什么目的,更要清楚它无能为力或不适合的使用领域是什么。定量方法具有透明度高、可重演性和明确程序化的优势,但不适用于探索规范性理论或寻找历史的真相。定量方法的推崇者和实践者务必避免“手持一把榔头,便无处不是钉子”。定量方法在各个操作步骤上有具体的要求和原则,研究者需要严格遵循、诚实报告。只有正确而谨慎地使用定量方法,才能够达到推进学术的目的,也能够真正维护定量方法的声誉和前景。

[收稿日期:2013 - 12 - 03]
[修回日期:2013 - 12 - 21]
[责任编辑:主父笑飞]