Logistic 模型的系数比较问题及解决策略:一个综述

社会
2015·4
CJS
第 35 卷

洪岩璧

摘　要:本文介绍了 Logistic 模型中经常被忽视的系数比较问题,包括同一样本在不同模型间的系数比较和在不同样本或子群体间的模型系数比较。研究者往往会沿袭线性回归模型的系数比较方法,但这是不恰当的,因为 Logistic 模型存在未被观测到的异质性(残差变异)问题,所以模型间系数不能进行简单的直接比较。根据已有研究,本文总结了解决这一问题的五种策略,分别是"y^* 标准化"、KHB 分解、异质选择模型、平均偏效应(APE)和线性概率模型(LPM),然后利用 CGSS2006 数据,以教育递进率模型为例,比较这些解决策略的异同,最后总结这些策略的特征及适用情况。

关键词:Logistic 模型　系数比较　未观测到异质性　残差变异

DOI:10.15992/j.cnki.31-1123/c.2015.04.009

On the Coefficients Comparison between Logistic Regressions and the Solutions: A Brief Review

HONG Yanbi

Abstract: This paper introduced the coefficients comparison between Logistic regression, which includes comparison between models within sample and that between samples or subsamples. Due to the unobserved heterogeneity (residual variation) problem in Logistic models, it is inappropriate to follow the OLS coefficients comparison in a naive simple way. With the same dependent

* 作者:洪岩璧　东南大学人文学院社会学系(Author: HONG Yanbi, Department of Sociology, School of Humanities, Southeast University) E-mail: hongyb@seu.edu.cn

** 本文系江苏省"公民道德与社会风尚'2011'协同创新中心"、"道德国情调查研究基地"的研究成果之一,并得到东南大学"中央高校基本科研业务费专项资金"资助(2242014S30026)。[This paper is supported by The Co-Innovation Center of Civil Morality and Social Custom, Research Center of National Moral Survey, and "the Fundamental Research Funds for Central Universities" (2242014S30026).]

感谢中国人民大学中国调查与数据中心提供中国综合社会调查(CGSS)2006 年数据。刘精明、吴愈晓、叶华、曾迪洋和《社会》匿名评审人给本文提出宝贵建议,在此一并致谢。文责自负。

variable, the total variance of OLS regression function is always fixed, which is irrelevant to the number of independent variables. However, the total variance of Logistic regression function will change as the independent variables increase or decrease, because the variance of error in Logistic regression is assumed to be constant, equals $\pi^2/3$. Previous researchers proposed many solutions to this comparison problem. Based on the literature, this paper introduced five solutions: y^*-standardization, KHB decomposition, heterogeneous choice model, average partial effect (APE), and linear probability model (LPM). Y^*-standardization and KHB only work in comparison between models within sample, heterogeneous choice model only works in comparison between samples or subsamples, and APE and LPM work in both situations. Drawing up on CGSS 2006 data, using educational transition model as an example, the author then showed the use and the differences between the five solutions through examining the cohort differences in school transition and whether the effects of parental ISEI differ in two cohorts' school transition. The final part summarized the characteristics and contexts of the five solutions.

Keywords: Logistic models, coefficients comparison, unobserved heterogeneity, residual variation

一、导论

Logistic 回归模型在社会科学领域已得到广泛应用，如分层研究（升学、毕业、晋升、找工作等）、政治行为研究（投票、参与集体行动等）、市场营销（是否购买某品牌商品）、人口学研究（离婚、迁移、出生、死亡等）等，但很多研究在解释 Logistic 回归结果时常会遇到一些问题，其中一个常见的问题就是系数比较。系数比较问题包括两个方面，一是同一样本在模型中因变量相同而自变量不同时的系数比较，典型的是嵌套模型之间的比较；二是对不同样本（包括样本中的不同子群体或不同时间点数据）使用同一模型，然后对这些模型之间的系数进行比较。[1] 这里的系数指发生比率的对数（log-odds ratio，LnOR）或发生比率（odds ratio，OR）。[2]

1. 为了行文方便，下文把这两类比较分别简称为“模型间系数比较”和“样本间系数比较”。
2. 谢宇（2010：335－336）把“odds ratio”译为“发生比率比”，本文在此延续郭志刚（1999）一书中的简洁译法，仍把“odds”译为“发生比”，把“odds ratio”译为“发生比率”。

一般教科书或著作在介绍 Logistic 模型时存在两种路径:一是从回归的角度出发,与线性回归(OLS)进行类比;二是从列联表视角出发。这个视角虽然有助于理解和解释结果,却不利于理解回归(Hosmer and Lemeshow,2000:X),所以,很多著作在介绍 Logistic 模型时都以回归作为起点,[3]但这也会混淆线性回归和 Logistic 回归的一些特性,系数比较是其中一个重要方面。早在 1984 年,温什普和迈耶(Winship and Mare,1984)就已注意到这一问题,但直到近些年,尤其是埃里森(Allison,1999)的文章发表之后,这一问题才真正引起部分学者的重视。但对大多数 Logistic 回归的使用者来说,这个问题并未受到关注。在很多已发表的论文(包括中英文论文)中,都可以看到一些作者直接进行 Logistic 模型间和样本间的系数比较,忽略了其不同于 OLS 模型的地方。

本文对近年来就 Logistic 模型的系数比较问题的相关讨论做了一个简要综述,指出在 Logistic 模型的运用中可能存在的几个误区,并介绍已提出的一些解决之策。值得注意的是,本文所探讨的问题不仅适用于 Logistic 模型,也适用于其他 Logit 模型(如 Ordinal Logit 模型和 Multinomial Logit 模型)和 Probit 模型,但为了行文简洁,我们在此仅以 Logistic 模型为例进行讨论。本文先论述 Logistic 模型系数比较问题的来源,即未被观测到的异质性问题;然后介绍已有的解决策略,包括"y^* 标准化"、KBH 分解、异质选择模型、平均偏效应(APE)和线性概率模型(LPM)方法;最后以教育递进率模型为例,比较不同策略的效果。

二、问题根源:未被观测到的异质性

一些研究者在理应使用 Logit 模型的情况下(即因变量为定类或定序变量)却选择了线性回归模型,其中一个重要原因就是为了避免 Logit 模型之间系数难以直接比较的问题(胡安宁,2014;Wang and Xie,2015)。为什么不能像线性回归模型那样对 Logistic 模型进行系数比较?在这一部分,笔者将介绍穆德(Mood,2010)对这一问题的阐释。他认为,根本问

3. 从回归视角引介的例子包括郭志刚(1999)和谢宇(2010),有关列联表视角的讨论可参见鲍威斯、谢宇(2009:30－35)和唐启明(2012:288－289)的研究。

题在于未观测到的异质性(unobserved heterogeneity)，即由未纳入模型的忽略变量(omitted variables)所引起的因变量变异情况。[4]

在一般的线性回归模型中，误差项往往被假定为服从均值为 0，方差为某一常数的正态分布。所以线性回归的总方差是固定的，只要因变量不变，其相对于均值的总的离差平方和就不变。但是 Logistic 模型的总方差会随着自变量的加入或减少而变化。

首先，我们以潜变量的方式来看待二分变量。虽然观察到的因变量取值是 1(成功)和 0(失败)，但可以假想因变量是未被观测到一种倾向性，即连续变量 y^*。当 $y^*>0$ 时，$y=1$；当 $y^*\leqslant 0$ 时，$y=0$。[5]以潜变量 y^* 为因变量的模型如方程 1 所示，这和一般 OLS 模型是相同的，唯一的差别在于我们无法观测到因变量 y^*。在方程 1 中，y^* 的总方差由被解释的方差和未被解释方差两部分组成，但当我们用方程 2来估计这一潜变量模型时，却把未被解释部分的方差(残差方差)设置为固定值。

为什么要把残差方差设为固定值？在线性回归模型中，因为因变量 y 是可观测的，所以可以对残差方差进行估计，但在二分因变量模型中，由于因变量 y^* 是无法观测到的，所以必须对残差方差进行假定，否则方程就无法辨识(unidentified)(Long and Freese，2001：102)。[6]之所以要对残差方差进行标准化，是因为二分因变量 y_i 本身不含有尺度(标尺)信息，使方程中自变量系数 β_k 的绝对大小不可确定，但它们之间的相对大小是可以被估计的(谢宇，2010：340)。在 Logistic 回归模型中，误差项被设定为服从标准 Logistic 分布，即残差的均值为 0，方差为 $\pi^2/3$，约等于 3.29。[7]

由于未被解释的残差方差被设定为固定值，所以，只要被解释的方

4. 这一问题也被称做“残差变异”(residual variation)。对 Probit 模型中的这一问题分析，可参见伍德里奇(Wooldridge，2002：470－472)的研究。

5. 转换(transformation)视角和潜变量视角是理解分类变量的两种主要路径(鲍威斯、谢宇，2009)。

6. “可辨识”是指如果根据充分或完备的观测数据能确定方程参数的唯一解，那么方程就是可辨识的，需要注意的是，辨识问题不是统计推论问题，和抽样无关，而是模型设置问题(贝里，2012：26－27)。

7. 在 Probit 模型中，残差被设定为服从均值为 0，方差为 1 的标准正态分布。有关潜变量的线性模型与非线性概率模型之间的对应关系可参见朗和弗瑞斯(Long and Freese，2001：100－103)的研究。

差有所增长，那么因变量的总方差（及其标尺）就会相应增加。当因变量的标尺增加，b_1 也必然会增加。此时 b_1 的大小不仅反映了 x_1 的效应，也反映了模型中未被观测到的异质性程度。

$$y_i^* = \alpha + x_{1i}\beta_1 + \varepsilon_i \quad (1)$$

$$\ln\left[\frac{P}{1-P}\right] = a + x_{1i}b_1 \quad (2)$$

方程 2 左边的部分被定义为 Logit，即发生比的自然对数（log odds）。为了方便接下来进行阐释，笔者把方程 1 写成方程 3：

$$y_i^* = \alpha + x_{1i}\beta_1 + \sigma\varepsilon_i \quad (3)$$

与方程 1 相比，方程 3 把残差写作 $\sigma\varepsilon_i$，其中 ε_i 的方差固定为 $\pi^2/3 = 3.29$，而用 σ 进行调整，以使残差符合其真实的方差。由于 σ 无法被观测到，而且我们设定了 ε_i 的方差，因此 Logistic 模型方程 2 中的 b_1 估计的其实是 β_1/σ（即方程 3 左右两边都除以 σ），而非 β_1。[8] 换言之，我们对真实的系数 β_1 进行了调整，以使得残差方差等于标准 Logistic 分布的方差，即 3.29。

为了进一步说明加入自变量对 Logistic 模型估计的影响，我们考虑存在忽略变量（omitted variable）x_2 的情况。假定真实的模型如方程 4 所示：

$$y_i^* = \alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + \sigma\varepsilon_i \quad (4)$$

其中，ε_i 符合标准 Logistic 分布，方差为 3.29（也即调整系数 σ 为 1）。两个自变量 x_1 和 x_2 之间的关系如方程 5 所示：

$$x_{2i} = \gamma_0 + \gamma_1 x_{1i} + v_i \quad (5)$$

其中 γ_0 和 γ_1 是需要估计的系数，v_i 是与方程 4 中的 ε_i 不相关的误差项。如果方程 4 忽略了 x_2，就会产生两个问题。一是估计偏误问题，把方程 5 带入到方程 4 中，得到 x_1 的效应是“$\beta_1 + \beta_2\gamma_1$”，即忽略 x_2 情况下的 β_1 也包含了 x_2 的效应。[9] 二是残差方差的增加问题。在方程 4

8. 因此，穆德（Mood，2010）强调，Logistic 模型中估计出来的系数并不代表参数的真正效应。但这一问题在应用中并不严重，因为在非线性模型中，我们想要知道的往往是偏效应（partial effect），而非参数本身。所以对于需要确定解释变量效应的方向以及不同变量效应的相对大小而言，β/σ 和 β 的效果是一样的（Wooldridge，2002：470；Cramer，2007）。但在比较不同样本间的 Logit 系数时，β/σ 问题就显得至关重要。

9. 当 x_1 和 x_2 相关时，这一问题在线性回归中也同样存在。

中，$\sigma=1$，即真实的残差方差就是 3.29，那么 b_1 估计的就是 $\beta_1/\sigma=\beta_1$；但如果忽略了 x_2，真实的残差方差就变成了 $\mathrm{var}(\varepsilon)+\beta_2^2\ \mathrm{var}(v)$，这样 $\sigma=\sqrt{3.29+\beta^2\mathrm{var}(v)}/\sqrt{3.29}$。

因此，如果我们从方程 4 中排除了 x_2，那么 b_1 估计的就不是 β_1，而是如式 6 所示：

$$b_1=(\beta_1+\beta_2\gamma_1)[\sqrt{3.29}/\sqrt{3.29+\beta_2^2\mathrm{var}(v)}] \tag{6}$$

如果 x_1 和 x_2 不相关，那么式 6 就等于：

$$b_1=\beta_1[\sqrt{3.29}/\sqrt{3.29+\beta_2^2\mathrm{var}(x_2)}] \tag{7}$$

在线性回归模型中，如果忽略了与模型中与其他自变量无关的变量，就不会发生忽略变量偏误(omitted-variable bias)，即对估计不会产生影响(谢宇，2010：105)。但由上述分析可知，在 Logistic 模型中，即使添加的变量与其他自变量无关，对系数估计也会产生影响。根据式 7 可知，这一问题会低估变量的效应(系数绝对值减小)，而未被观测到的异质性大小取决于忽略变量的方差[$\mathrm{var}(x_2)$]及其对 y 的效应大小(β_2)。

由于我们的模型所能解释的因变量变异的比例通常都不高，[10] 即存在较多的忽略变量，所以，$\beta_2^2\mathrm{var}(v)$ 存在很大的增长空间，未被观测到的异质性问题始终萦绕不散。由式 7 可知，如果未观测的异质性与自变量无关，那么即使我们不知道未观测到异质性的大小，其影响方向也是可知的，即低估已有自变量的效应，因为调整系数 $\sigma>1$。由于调整系数在同一模型中是固定的，所以它不会影响自变量效应的方向，以及同一模型中自变量效应之间的相对大小(Wooldridge，2002：470)。

因此，未被观测的异质性问题导致我们无法像线性回归那样对 Logistic 嵌套模型之间的系数直接进行比较。同样，我们也无法直接比较同一 Logistic 模型在不同群体中的效应，如比较同样一些因素对升学的效应是否存在男女差异、族群差异和时期差异。直接简单比较所隐含的假设是：不同的样本或同一样本的子群体之间具有相同的未观测到异质性。这是一个较强的假定，但往往不符合实际。

毋庸置疑，模型之间和样本之间某变量效应的比较是社会学定量研究关注的一个核心，不可能因为上述原因而放弃比较。一个便捷的替代

10. 在一般线性回归中，相对应的是 R^2 往往并不高，表明模型并未包含所有解释因变量变异的解释变量。

选择是，尽量用连续性变量来替换二分类变量，但像升学、死亡之类的变量本身就是二分类的，没有连续性的变量可以替代。研究者针对这一问题已提出不少解决策略，虽然尚无定论，但都值得了解与借鉴。

三、应对策略

（一）同一样本，不同模型之间的系数比较

嵌套模型是定量研究中经常用到的分析策略，通过比较嵌套模型之间的系数，我们可以了解控制变量对核心自变量效应的影响。但在Logistic嵌套模型间的系数比较中，必须考虑标尺变化带来的影响。卡尔森等（Karlson，*et al.*，2013）指出，在比较嵌套模型的系数时，特定自变量的系数会受到其他自变量的“混杂效应”（confounding）和“标尺改变效应”（rescaling）两方面的影响。我们把方程3、方程4稍作更改（同时去掉了常数项，如方程3a、方程4a所示），分别称为“简化模型”（简称R）和“完全模型”（简称F）。[11]当 x_1 和 x_2 相关，且 x_2 对 y^* 存在独立效应时，就存在“混杂效应”。而“标尺改变效应”就是因上文所说的残差方差变化导致的问题，使得 x_1 的系数在不同模型中并不遵循同一测量标尺。在嵌套模型比较中，我们真正感兴趣的是“混杂效应”，但由于观察到的系数差异还包含了“标尺改变效应”，所以直接比较可能会导致错误结论。

$$y_i^* = x_{1i}\beta_{1R} + \sigma_R\varepsilon_i \tag{3a}$$

$$y_i^* = x_{1i}\beta_{1F} + x_{2i}\beta_2 + \sigma_F\varepsilon_i \tag{4a}$$

$$b_{1R} - b_{1F} = \frac{\beta_{1R}}{\sigma_R} - \frac{\beta_{1F}}{\sigma_F} \neq \beta_{1R} - \beta_{1F} \tag{8}$$

其中，b 是 β 的估计值。增加自变量后，残差方差一般会减小，即 $\sigma_R \geqslant \sigma_F$。所以如式8所示，一般而言，直接把两个嵌套模型系数相减得到的值会低估真正的“混杂效应”。

1.“y^* 标准化”

“y^* 标准化”（y^*-standardization）是针对上文提到的因变量的标尺不固定提出的。[12]温什普和迈耶（Winship and Mare，1984：517）建议，

11. 这里的简化模型和完全模型相当于谢宇（2010：138）所说的限制性模型和非限制性模型。

12. 在此我们采用唐启明（2012：327）“y^* 标准化”的提法，以区别于线性回归模型中的“y 标准化”。

可以把不同模型的系数估计值都根据潜在因变量的方差进行重新调整，那么系数在模型之间就可以比较了。具体而言，就是用系数除以各自模型潜在因变量的估计标准差 SD(y*)，然后进行比较。

$$SD(y*) = \sqrt{\mathrm{var}(x'\mathrm{b}) + \mathrm{var}(\varepsilon)} = \sqrt{\mathrm{var}(x'\mathrm{b}) + 3.29} \quad (9)$$

如式 9 所示(Karlson, *et al.*, 2013: 298; Long, 1997: 129)，模型潜在因变量的估计标准差 SD(y*)由两个部分组成：一是预测值的方差；二是残差的设定方差(3.29)。由于第二部分是固定的，所以模型间 SD(y*)的差异就来自于第一部分。而第一部分预测值的方差又取决于模型所包含的自变量。如上文所述，当增加自变量时，预测值方差就会增加，导致潜在因变量的方差也相应增加。因此，“y*标准化”就是通过用系数除以 y* 标准差，即 $b/SD(\mathrm{y}^*)$，来减小预测值方差增加的影响，使系数表达出自变量变化一个单位，因变量变化多少个潜变量 y* 的标准差单位(standard-deviation-unit change in y*)。[13] 需要注意的是，“y* 标准化”方法仅适用于同一样本内不同模型之间的系数比较，因为我们不知道不同样本之间的未观测到异质性是否存在差异。在 Stata 软件中，在执行 Logit 模型命令之后运行“listcoef, std help”命令，可以直接输出 y* 标准化后的模型系数(Long and Freese, 2001: 155)。

2. KHB 分解

卡尔森等(Karlson, *et al.*, 2013)提出了分解“混杂效应”和“标尺改变效应”的方法。KHB 方法的核心是得到 x_2 对 x_1 回归后的残差 $\widetilde{x}_2$，即方程 5 的误差项 v_i。然后用 $\widetilde{x}_2$ 替代 x_2 进入方程 4a，得到方程 10。可以证明在方程 3a、方程 4a 和方程 10 中，$\beta_{1R} = \beta^*_{1F}$ 和 $\sigma_F = \sigma^*_F$。[14]

$$y_i^* = x_{1i}\beta^*_{1F} + \widetilde{x}_{2i}\beta^*_2 + \sigma^*_F\varepsilon_i \quad (10)$$

这两个等式的具体数学证明可参见卡尔森等(Karlson, *et al.*, 2013: 292)的研究，笔者在此仅进行简单阐释，帮助读者理解其基本逻辑。首先，由于 $\widetilde{x}_2$ 是 x_2 对 x_1 回归后的残差，所以 $\widetilde{x}_2$ 与 x_1 不相关，那么 $\widetilde{x}_2$ 不会影响的 x_1 系数估计，因为 $\widetilde{x}_2$ 反映了 x_2 被 x_1 线性解释以外的变异，相当于谢宇(2010: 105)所说的“与其他自变量无关的忽略变

13. 标准化的另一原因在于潜在因变量本身没有量测单位，因而非标准化系数的大小没有意义(唐启明，2012: 327)。

14. 有关残差与偏回归估计的内容可参见谢宇(2010: 149－152)的研究。

量”。所以,得到 $\beta_{1R}=\beta_{1F}^{*}$ 。

其次,从路径分析角度来看,方程 4a 和方程 10 的预测值是一样的。方程 4a 代表了 x_1 对 y^{*} 的直接效应和经过 x_2 的间接效应,以及 x_2 对 y^{*} 的直接效应和经过 x_1 的间接效应。方程 10 则代表了 x_1 对 y^{*} 的全部效应,和 x_2 被 x_1 线性解释以外的那部分变异(即 $\widetilde{x}_2$)对 y^{*} 的全部效应,而 x_2 与 x_1 相关对 y^{*} 的那部分共同效应已经被包含在 x_1 对 y^{*} 的全部效应中了。所以方程 4a 和方程 10 只是以不同方式表述了 x_1 和 x_2 对 y^{*} 的影响,因而它们的预测值相同,方程 4a 和方程 10 的残差也必然相同,所以得到 $\sigma_F=\sigma_F^{*}$ 。因此,方程 4a 和方程 10 实际上反映的是同一个潜变量线性模型。

由于方程 4a 和方程 10 反映的同一个模型,所以它们不仅拥有相同的测量标尺,而且误差分布也相同。因此,β_{1F}^{*} 和 β_{1F} 的差异就是在控制标尺改变效应后的“混杂效应”。卡尔森等(Karlson, *et al.*, 2013)提出了三种测量系数变化的指标:

(1)差异测量:

$$b_{1F}^{*}-b_{1F}=\frac{\beta_{1F}^{*}}{\sigma_F^{*}}-\frac{\beta_{1F}}{\sigma_F}=\frac{\beta_{1R}}{\sigma_F}-\frac{\beta_{1F}}{\sigma_F}=\frac{(\beta_{1R}-\beta_{1F})}{\sigma_F} \tag{11a}$$

(2)比例测量:

$$\frac{b_{1F}}{b_{1F}^{*}}=\frac{\beta_{1F}/\sigma_F}{\beta_{1F}^{*}/\sigma_F^{*}}=\frac{\beta_{1F}}{\beta_{1R}} \tag{11b}$$

(3)百分比测量:

$$\frac{b_{1F}^{*}-b_{1F}}{b_{1F}^{*}}=\frac{(\beta_{1F}^{*}-\beta_{1F})/\sigma_F}{\beta_{1F}^{*}/\sigma_F^{*}}=\frac{\beta_{1R}-\beta_{1F}}{\beta_{1R}}\times 100\% \tag{11c}$$

上述三种指标并无本质差异,只是以不同形式表达“混杂效应”的大小,选择何种指标取决于研究者表述的需要。差异测量(11a)是基于完全模型的标尺来测量“混杂效应”,与一般的 Logit 系数具有相同的性质。而比例测量(11b)和百分比测量(11c)都不受标尺的影响,因为它们本质上是比率,测量的是潜在倾向性的偏效应,而非 Logit 系数。

此外,还可以测量在控制“混杂效应”后的“标尺改变效应”。如式 12 所示,$b_{1F}^{*}-b_{1F}$ 测量了“混杂效应”,而 $b_{1R}-b_{1F}^{*}=\frac{\beta_{1R}}{\sigma_R}-\frac{\beta_{1R}}{\sigma_F}$ 则测量了“标尺改变效应”。

$$b_{1R}-b_{1F}=\frac{\beta_{1R}}{\sigma_R}-\frac{\beta_{1F}}{\sigma_F}=\frac{\beta_{1R}}{\sigma_R}-\frac{\beta^*_{1F}}{\sigma^*_F}+\frac{\beta^*_{1F}}{\sigma^*_F}-\frac{\beta_{1F}}{\sigma_F}$$
$$=(b_{1R}-b^*_{1F})+(b^*_{1F}-b_{1F}) \quad (12)$$

卡尔森等(Karlson，*et al.*，2013)还发展出了统计量 Z_C，用以直接检验 Logit 系数改变量是否来自于控制“标尺改变效应”后的“混杂效应”。他们通过蒙特卡罗模拟数据分析指出，相比于“y^* 标准化”、APE 和 LPM 三种方法，KHB 方法的估计效果更好，结果更接近真实的系数差异。Stata 软件中已经有 khb 命令可供使用。[15]

（二）同一模型，不同组别或样本之间的系数比较

对于不同群体或样本之间的 Logistic 模型系数比较，以往通常使用的方法有两种。一是使用交互项，即对特定的自变量和样本指示变量(indicator variable)进行交互，如果交互项系数显著，我们就认为该自变量的效应在两个样本间存在差异。二是对不同样本分别进行模型估计，然后进行 Wald 卡方检验(需要假定不同样本的系数服从独立的抽样分布)，检验统计量如式 13 所示(Clogg，Petkova and Haritou，1995)。其中，β_{1i} 和 β_{2i} 分别代表两个样本的模型系数，$SE(\beta_{1i})$ 和 $SE(\beta_{2i})$ 分别为其标准误，该统计量服从标准正态分布。

$$\frac{\beta_{1i}-\beta_{2i}}{\sqrt{SE^2(\beta_{1i})+SE^2(\beta_{2i})}} \quad (13)$$

这两种方法都没有控制未观测到的异质性，其差别仅在于交互项检验假定其他变量对两个群体的作用是相同的。如果把所有解释变量都和指示变量进行交互，那就等同于分别在不同样本中进行模型估计，也即谢宇(2010：239)所说的“完全交互项”。但在 Logit 模型中，由于存在未观测到的异质性，模型系数并不代表真正的变量效应，因此难以在不同样本之间进行比较。这类似于在比较不同样本之间线性回归模型的标准化系数时所引发的问题(Allison，1999)。在讨论通径模型时，郭志刚(1999：157)指出，“标准化系数所反映的不仅是自变量对因变量的影响强度，而且还反映了模型中各变量的方差以及它们之间的协方差，甚至还反映了寓于误差项之内的未包括在模型中的那些变量的方

15. 有关 khb 命令的描述可参见：http://fmwww.bc.edu/RePEc/bocode/k/khb.html，以及科勒等(Kohler，Karlson and Holm，2011)的研究。

差。[16]因此,标准化系数有特定样本的性质,不能用于不同情况或不同总体之间的比较”。譬如,在分析受教育年限对收入的影响是否存在性别差异时,我们应该比较教育变量的非标准化系数,因为在男女两个样本中,教育变量的测量都是基于同一标尺(受教育年限)。如果采用标准化系数,其测量标尺在男女两个样本中就可能存在差别。

上文已述,在 Logit 模型中,我们得到的系数已经用 σ 调整过了(如方程 3 和方程 4 所示),类似于线性回归模型中的标准化系数。如果两个样本的残差调整系数 σ 不同,那么这两个样本的模型系数所依据的标尺(残差标准差)就不同,所以无法直接进行比较。换言之,两个样本存在不同的未观测到异质性,即残差变异(residual variation)问题。霍特科(Hoetker,2004)的一系列模拟实验表明,即使样本间的残差方差只存在较小差异,直接用上述两种传统方法(交互项和卡方检验)来比较 Logit 系数仍然会带来很大偏差,有可能显示出根本不存在的差异,或掩盖真实的差异,甚至与真实差异相反。

埃里森(Allison,1999)利用一群生物化学家组成的样本(人年数据),探讨了性别对晋升副教授的影响。他分别对男性学者和女性学者进行 Logistic 模型估计后发现,论文数量对男性晋升的作用大约是女性的 2 倍。埃里森认为这一比较结果是不可靠的,因为女性比男性具有更异质化的职业发展模式,未观测变量对女性晋升的影响比男性更大。女性样本中的调整系数 σ 大于男性样本($\sigma_F > \sigma_M$),根据式 7 可知,$b = \beta/\sigma$,所以,即使男女样本中各变量的真实效应 β_F 和 β_M 是相同的,我们所得到的女性样本系数 b_F 也会小于男性样本系数 b_M,可见,未观测到的异质性对女性样本中论文系数的影响要大于男性。

埃里森(Allison,1999)提出一个颇为繁琐和复杂的程序来检测样本间 Logistic 系数是否存在差异,但一些模拟实验证明该程序只适用于部分情形。因为他的方法有两个重要前提假定:一是在检验残差方差是否相同时,需要假定两个样本的系数相同;二是在检验某变量系数存在差异时,需要假定在两样本之间至少有一个系数是相同的(Williams,2009:546),但实际情况经常不符合这两个假定。对于是否存在残差变异的初步判断,研究者可以参考埃里森(Allison,1999)提出

16. 这等同于本文所说的未观测到的异质性。

的一个简便识别方法：如果一个群体的模型系数系统性和成比例的高于或低于另一个群体的系数，那就很可能就存在残差变异干扰。一般来说，当模型中加入较多的控制变量后，不同群体之间系数不可比较问题的严重性就会降低，因为未被观测到的异质性减小了（Allison，1999）。

1. 异质选择模型（heterogeneous choice model）

威廉姆斯（Williams，2009）提出可以用异质选择模型来解决群体间 Logit 系数的比较问题，并认为埃里森（Allison，1999）的模型和豪斯等（Hauser and Andrew，2006）的 Logistic 响应模型都是异质选择模型的子类型。异质选择模型也称位置标尺模型（location-scale model）。该模型不仅可以处理残差方差变异，还能处理其他来源的异方差（heteroscedasticiy）问题。

异质选择模型同时拟合两个方程，一个是选择方程（或位置方程），即传统的 Logit 模型估计；另一个是残差方差方程（或标尺方程），纳入那些会影响异方差性的变量，反映了潜在因变量在不同组别中是如何以不同标尺来测量的。异质选择模型的因变量可以是二分变量，也可以是定序变量。在因变量是二分变量的情况下，模型表达式如式 14 所示（这里使用的是转换的方式）：

$$\Pr(y_i = 1) = g\left[\frac{x_i\beta}{\exp(z_i\gamma)}\right] = g\left[\frac{x_i\beta}{\exp(\ln(\sigma_i))}\right] = g\left(\frac{x_i\beta}{\sigma_i}\right) \quad (14)$$

其中，g 代表联结方程（本文中以 Logit 为例，但也可以是 Probit、Complementary log-log、Log-log 和 Cauchit）。x 是第 i 个观测的一组值，所有的 x 是决定选择结果的解释变量。z 也是第 i 个观测的一组值，所有的 z 决定了群体之间在潜在因变量上的残差变异。z 不仅可以包括性别、族群等分类变量，也可以包括与残差方差相关的连续变量。需要注意的是，z 和 x 并不一定要包含相同的变量。[17] β 和 γ 是系数矩阵，它们分别表示 x 如何影响选择结果、z 如何影响方差（更准确地说是调整系数 σ 的自然对数）。在式 14 中，分子被称为选择方程，分母被称为方差方程。简单来说，异质选择模型就是在控制残差变异情

17. 一般来说，在实际研究中，z 所包含的变量相对较少。Stata 软件中 oglm 命令适用 stepwise 筛选功能，即只把那些显著影响残差变异的变量保留在方差方程中。

况下来估计 Logit 模型。

但异质选择模型并非灵丹妙药，如果模型设置错误（包括选择方程和方差方程），该模型仍有可能导致错误的结果。所以威廉姆斯（Williams，2009）建议研究者同时估计控制和未控制异方差性的模型，然后仔细考虑模型结果之间的差异是否由模型设置错误而引发。在 Stata 软件中可以用 oglm（Ordinal Generalized Linear Model）命令对该模型进行估计（Williams，2010）。

2. 平均偏效应（APE）

在 Logit 模型中，除了报告发生比率（odds ratio）之外，研究者也可以报告事件发生的概率（probability）预测值和自变量变化所引起的概率变化量（Petersen，1985）。如果要考察变量对结果变量发生概率的影响，首先需要对概率和 Logit 进行转换，两者关系是 Logistic 累积分布方程（CDF）：

$$F(\beta x_i) = \frac{\exp(\beta x_i)}{1+\exp(\beta x_i)}$$

其中 βx_i 是第 i 个观测的 Logit 值。Logistic CDF 的斜率就是 Logistic 概率分布方程（PDF），表达式如下：

$$f(\beta x_i) = \frac{\exp(\beta x_i)}{[1+\exp(\beta x_i)]^2}$$

CDF 给出的是 $y_i=1$ 的概率 $P(y_i=1)$，而特定取值 $P(y_i=1)$ 上的 PDF 则等于 $P(y_i=1)\times[1-P(y_i=1)]$。发生比率表示某自变量对 Logit 的影响是乘以一个恒定值，但自变量对因变量发生概率的影响却不能如此直观的表述。我们只能报告概率的变化，相关的指标包括“边际效应”（Marginal Effects，简称 MFX）、“平均边际效应”（Average Marginal Effects，简称 AME）和“平均偏效应”（Average Partial Effects，简称 APE）。“边际效应”测量的是在 x 某一特定取值附近的成功概率的变化率，所以“边际效应”会随着 x 的取值变化而变化（鲍威斯、谢宇，2009：56－57）。Logit 模型中 x_1 的“边际效应”（MFX）是：

$$f(\beta x_i)\beta_1 = \frac{\exp(\beta x_i)}{[1+\exp(\beta x_i)]^2}\beta_1 \qquad (15a)$$

其中，β_1 是变量 x_1 的发生比率对数（log odds-ratio）的估计值，βx_i 是

第 i 个观测的 Logit 值，$f(\beta x_i)$ 是 βx_i 的 Logistic 分布的概率分布方程（PDF）。计算 x_1 的“边际效应”时，需要把所有其他变量固定在某些取值上，一般选取均值（Mood，2010：75）。

未观测到的异质性会低估变量效应（即系数绝对值减小，如式 7 所示），但概率值则向 0.5 变化，导致 $P(y_i = 1) \times [1 - P(y_i = 1)]$ 的值向其最大值 0.25 变动，所以概率变动值可以部分抵消对系数的低估。因此，研究者建议使用“平均偏效应”（APE）进行模型间、样本间的系数比较，因为它几乎不受与自变量无关的未观测异质性影响（Cramer，2007）。APE 的表达式如式 15b 所示，计算的是 x_1 取特定值或特定区间内，“边际效应”的平均数。譬如，我们可以取均值左右0.01个标准差范围的个案来计算 APE。因此，APE 本质上是“边际效应”在样本中的加权平均数，与“边际效应”一样，APE 也会随着 x_1 取值的变化而变化，从而体现分布的非线性特征。

$$\frac{1}{N}\sum_{i=1}^{N} f(\beta x_i)\beta_1 = \frac{1}{N}\sum_{i=1}^{N} \frac{\exp(\beta x_i)}{[1+\exp(\beta x_i)]^2}\beta_1 \qquad (15b)$$

虽然卡尔森等（Karlson，*et al.*，2013）认为 APE 的效果不如 KHB 好，但 APE 的优势在于不仅适用于嵌套模型间的系数比较，也适用于不同样本之间的系数比较，并且其解释是基于概率的，更易为读者理解。Stata 软件中可用 margeff 命令计算 APE，默认报告的是根据 x_1 所有取值计算得到的 APE 值。[18]

3. 线性概率模型（LPM）

最后介绍的方法是线性概率模型（Linear Probability Model，简称 LPM），即运用线性回归来分析二分因变量。这一方法的争议颇多，因为 Logit 模型或 Probit 模型的合法性就是部分地建立在批判 LPM 之上的。从回归角度引介 Logistic 模型时，往往会指出，如果因变量是二分变量，不可以用线性回归进行估计，主要原因有三：首先，这违背了多元回归的假设，尤其是不同自变量的误差具有相同方差这一假设，即方差齐性（homoscedasticity）；其次，预测值通常会超过符合逻辑的概率范围（0－1），即预测值取值范围的荒谬性；第三，模型设置错误，因为变量

18. 也即克瑞默（Cramer，2007）所说的 ASE。margeff 命令也可以计算在某些固定值（比如均值）上的偏效应。

间关系不是线性的(郭志刚,1999;谢宇,2010:332)。

针对方差齐性问题,可以采用加权最小二乘法(Weighted Least Squares)或异方差稳健标准误(Heteroscedasticity-robust Standard Errors)来处理(鲍威斯、谢宇,2009:34;Mood,2010)。对于预测值超出[0,1]范围的取值荒谬性问题,谢宇(2010:338)认为是线性模型处理二分因变量存在的最大问题,温什普和迈耶(Winship and Mare,1984:514)也认为这是更青睐 Logit 模型和 Probit 模型的原因。但朗(Long,1997)指出,在非二分因变量的线性回归中,预测值超出可能范围的情况并不鲜见,因此只要小于 0 和大于 1 的预测值数量不是太多,就不是一个太严重的问题。穆德(Mood,2010)认为只有第三个才是最严重的、目前难以解决的问题。LPM 模型与 Logit 模型的最大区别在于,LPM 模型假定连续自变量的"边际效应"(Marginal Effect)为常数,而 Logit 模型意味着偏效应的大小是递减的(伍德里奇,2003:520)。如果非线性很重要,那么 LPM 模型可能就会使读者对变量关系理解有误。但其优点在于系数比较更容易,系数理解更直观,因此研究者在实际分析中需要权衡利弊。

穆德(Mood,2010)建议,在需要进行嵌套模型间系数比较时,线性概率模型不失为一个值得考虑的选择。[19]如伍德里奇(2003:519)对已婚妇女的劳动力市场参与数据的分析分别运用了 LPM 模型、Logit 模型和 Probit 模型,结果发现三个模型的结论一致,即每个模型所得到的系数符号都相同,而且具有统计显著性的变量也相同。卡尔森等(Karlson,*et al*.,2013)的模拟数据分析也表明,与简单直接比较 Logit 系数相比,LPM 模型系数比较更接近真实的差异。

四、实例:教育递进率模型

在这一部分,笔者以教育递进率模型为例,考察不同方法对估计结果的影响。教育递进率模型由迈耶(Mare,1980;1981)首次提出,但该模型不仅存在内生选择性问题,其教育转折方程的残差也不同,因而存在"标尺效应"问题,导致不同转折点方程中的系数相互之间无法直接

19. 在讨论序次因变量时,唐启明(2012:332)指出常规最小二乘法也是一种选择。在实际研究中,很多学者会用线性回归方法来分析序次因变量(胡安宁,2014;Wang and Xie,2014)。

进行比较(Holm and Jæger,2011)。本文使用 CGSS2006 数据分析同期群变量对被访者升学进入高中的影响,[20]同时控制性别、父母的职业(ISEI)和家庭藏书量。为了便于比较,我们选取了 20—39 岁的达到初中毕业水平的被访者,共 3 132 人。这些样本被划分为两个同期群,年轻的同期群(20—29 岁)1 471 人,年长的同期群(30—39 岁)1 661 人,变量描述统计见表 1。

数据分析分两个部分。第一部分通过嵌套模型来考察同期群效应是否会因为加入父母职业(ISEI)和家庭藏书量变量之后而发生变化。在嵌套模型比较中,运用 Logistic 模型简单比较、LPM(线性概率模型)简单比较、APE、"y* 标准化"和 KHB 分解五种方法来考察同期群变量系数在两个嵌套模型之间的变化幅度。第二部分考察父母职业在不同的同期群中的效应是否存在差异,我探讨了 Logistic 简单比较(包括分群体模型和交互项分析)、LPM 简单比较、APE 比较和 OGLM 比较(异质选择模型)这四种方法得出的结果是否存在差异。

表 1:变量描述统计表(N=3 132)

变量	均值(标准差)
高中升学率 (ss)	0.595 (0.491)
男性 (male)	0.472 (0.499)
年轻同期群 (coh)	0.470 (0.499)
18 岁时父亲职业 (iseif18)	33.384 (16.889)
18 岁时母亲职业 (iseim18)	23.019 (20.136)
18 岁时家庭藏书量 (bkn)	30.744 (20.136)

(一)同期群的影响

表 2 的第一部分报告了 Logistic 模型和线性概率模型的估计结果。根据 Logistic 模型,直接比较简化模型 R1 和完全模型 F1 的年轻同期群变量(coh)系数,在加入新变量之后,coh 系数降低了 12.4%[(0.573−0.502)/0.573]。运用 LPM 方法,直接比较线性简化模型 L1 和线性完全模型 L2 的 coh 系数,该系数降低了 14.7%[(0.136−0.116)/0.136]。然而,如表 2 第二部分所示,利用"y* 标准化"方法,得到模型 R1 和 F1 的潜变量估计标准差 SD(y*)分别是 1.837 和

20. CGSS2006 数据由中国人民大学社会学系与香港科技大学社会科学部联合实施,数据抽样方案详见《中国综合社会调查报告(2003−2008)》(中国人民大学中国调查与数据中心中国综合社会调查项目,2009)。

2.147，Beta/SD(y*)减小了预测值方差增加的影响，那么从模型 R1 到 F1，coh 系数下降了 25.1%［(0.312－0.234)/0.312］。

如表 2 第三部分所示，运用 KHB 分解方法，coh 系数降低了 26.5%［(0.683－0.502)/0.682］。这里简化模型的同期群变量系数变成了 0.683，比 Logistic 模型 R1 中的系数(0.573)大。这是因为 KHB 报告的系数是以完全模型 F1 的残差方差标尺来衡量的，[21]由于完全模型的残差方差增加了［"y* 标准化"中的 SD(y*)值揭示了相同的现象］，因此，若以 F1 模型的残差方差为标尺，简化模型的系数就会有所增加。根据表 2 第四部分所示，从模型 R1 到 F1，coh 系数的 APE 值下降了 16.5%［(0.127－0.106)/0.127］。可见，"y* 标准化"和 KHB 分解这两种方法得到的结果基本吻合；无论是 Logistic 模型还是线性模型，简单的直接比较都低估了系数的改变程度。APE 改变量虽然大于简单直接比较，但仍存在严重低估的问题。可能的原因在于，APE 可用于模型间比较的一个前提假定是新加入自变量与核心自变量不相关(Cramer，2007)，在本例中，即需要假定父母职业和家庭藏书量变量与同期群变量不相关，但方差分析显示父母职业和家庭藏书量两个变量在同期群之间均存在显著差异，不符合独立不相关假定。所以，笔者建议研究者在比较 Logit 模型间系数时应进行 y* 标准化或 KHB 分解处理。

在嵌套模型的系数比较中，当完全模型的系数小于简化模型系数时，直接比较一般会低估减少量(如式 8 所示)。因此，如果简单直接比较显示差异显著，"标尺改变效应"就不会影响差异显著这一结论。但是，如果完全模型系数大于简化模型系数，就需要特别注意。我们往往会认为新加入变量抑制了核心自变量的效应，但这很可能是由于标尺改变导致系数的增加，使系数变化改变了方向，所以，在这种情况下，一定要进行"y* 标准化"或 KHB 分解处理，避免得出错误的结论。

（二）父母职业地位影响在同期群间的差异

表 3 呈现了样本间系数比较的 Logistic、OGLM 和 LPM 三种估计方法。模型 F2 中，母亲职业与同期群交互效应显著，表明在年轻同期

21. 因此，针对同一个简化模型，如果采用了两个不同的完全模型，那么简化模型在两次 KHB 分解之后得到的两套系数是不同，因为所采取的衡量标尺是根据不同的完全模型而得到的。这和简单的嵌套模型比较不同，研究者需要注意。

表 2:嵌套模型的 Logistic 估计与 LPM 估计比较(N=3 132)

		Logistic 模型		概率线性模型(LPM)	
		简化模型(R1)	完全模型(F1)	简化模型(L1)	完全模型(L2)
一	年轻同期群(coh)	0.573***	0.502***	0.136***	0.116***
		(0.074)	(0.078)	(0.017)	(0.017)
二	SD(y*)	1.837	2.147		
	Beta/SD(y*)	0.312	0.234		
三	KHB-coh	0.683***	0.502***		
		(0.079)	(0.078)		
四	APE	0.127***	0.106***		
		(0.006)	(0.006)		

注:1. 显著性水平:* p<0.05,** p<0.01,*** p<0.001。括号中为标准误;
2. 简化模型 R1 和 L1 包含了性别变量,完全模型 F2 和 L2 包括性别、18 岁时父亲和母亲的职业、18 岁时家庭藏书量三个变量;
3. 本表没有报告其他控制变量的系数估计结果,有兴趣的读者可向作者索取。

群中,母亲职业对高中升学的影响更大。同样,直接比较两个分组模型的母亲职业系数,统计检验(方程 13)表明两个同期群之间也存在显著差异。[22]但在模型 O2 中,考虑了残差的未观测异质性之后,交互效应就不显著了,即母亲职业的影响在两个同期群之间没有呈现显著差异。

如表 3 所示,OGLM 不仅报告了选择方程的估计结果(等同于 Logistic 方程的估计结果),还报告了方差方程的结果,即 ln(sigma)值。模型 O1 的方差方程中的同期群系数表明,年轻同期群的残差标准差是年老同期群的 75.2%[exp(−0.285)]。但模型 O2 的方差方程中的同期群系数不显著。这说明不同的同期群具有不一样的升学模式,例如,1999 年的高校扩招影响到部分年轻同期群升入高中的机会和决策。

比较 OGLM(O2)和线性概率模型(L3)这两个完全模型,我们可以发现系数的方向和显著性都一致,也即基本结论相同。在 Logistic 模型(F2)中,18 岁时母亲职业与同期群的交互项具有显著性,但在 OGLM(O2)和 LPM(L3)中,该交互项都不具有显著性。因此,如穆德(Mood,2010)所言,在二分因变量模型中,线性概率模型有其自身的优势,是一个值得考虑的选择。如表 4 所示,根据 APE 值所得到的结论一方面与直接比较 Logistic 模型(F2)系数所得到的结论接近,即母亲职业与同期群的交互项是显著的。另一方面,APE 值的结果显示即使纳入交互项,同期群之间仍存在显著差异,与 OGLM 模型的结论接近。

22. 限于篇幅,本文并未报告分同期群的 Logistic 模型结果。感兴趣的读者可向作者索取。

表 3:组际差别的 Logistic 模型、OGLM 和 LPM 比较(N= 3 132)

	混合完全 Logistic 模型(F2)	混合 OGLM		混合完全 LPM (L3)
		简化模型(O1)	完全模型(O2)	
父亲职业	0.029***	0.028***	0.030***	0.007***
	(0.003)	(0.003)	(0.003)	(0.001)
母亲职业	0.007*	0.009***	0.007*	0.001*
	(0.003)	(0.002)	(0.003)	(0.001)
年轻同期群	0.174	0.363***	0.468*	0.108**
	(0.205)	(0.089)	(0.232)	(0.040)
父亲职业×同期群	0.005		−0.007	−0.001
	(0.006)		(0.008)	(0.001)
母亲职业×同期群	0.008*		0.003	0.001
	(0.004)		(0.005)	(0.001)
常数项/切点	−1.177***	1.143***	1.151***	0.258***
	(0.136)	(0.134)	(0.135)	(0.028)
方差:ln(sigma)		−0.285*	−0.417	
年轻同期群		(0.134)	(0.283)	
模型卡方(df)	409.9 (7)	409.7 (6)	412.1 (8)	—
BIC	3 881.5	3 873.7	3 887.3	4 149.2

注:1.显著性水平:* p<0.05,** p<0.01,*** p<0.001。括号中为标准误,模型卡方后面括号中为自由度;

2.所有的模型还包括了性别和家庭藏书量。

表 4:Logistic 模型对应的 APE

	Logistic 模型		
	年轻同期群	年老同期群	混合模型
男性	0.014	0.025	0.020**
	(0.026)	(0.024)	(0.007)
父亲职业	0.007***	0.007***	0.006***
	(0.001)	(0.001)	(0.000)
母亲职业	0.003***	0.002*	0.001***
	(0.001)	(0.001)	(0.000)
家庭藏书量	0.003***	0.002***	0.002***
	(0.001)	(0.000)	(0.000)
年轻同期群			0.038*
			(0.018)
父亲职业×同期群			0.001*
			(0.000)
母亲职业×同期群			0.002***
			(0.000)
N	1 471	1 661	3 132

注:显著性水平:* p<0.05,** p<0.01,*** p<0.001;括号中为标准误。

模型 F2 是常用的交互项检验组际差异的方法,但交互项检验并没有考虑可能存在的未观测到异质性情况,而是假定两个同期群具有相同的残差方差。由于简化模型 O1 表明不同同期群的残差方差显著不

同,但纳入交互项之后,模型 O2 的方差方程中的同期群系数变得不显著了。比较模型 O2 和模型 F2 中的系数,可以发现当考虑同期群的残差变异问题后,父母亲的职业变量都没有什么变化,最大的变化在于同期群变量和同期群与母亲职业的交互项。这是否表明加入交互项之后不存在残差变异问题了呢?这时需要考虑哪个模型最好地呈现了数据的情况,即进行模型选择。模型检验的指标都倾向于 OGLM 简化模型 O1。从模型卡方来看,模型 F2 和 O2 分别比模型 O1 多 1 个和 2 个自由度,但是卡方增加值却很小,都不显著。就 BIC 而言,模型 O1 的 BIC 值比模型 F1 和 O2 要低不少。因此,OGLM 简化模型 O1 是最好的,它表明在控制残差变异之后,母亲职业与同期群变量不存在显著的交互效应,而同期群之间存在显著差异。

所以,在实际分析中,应该遵循威廉姆斯(Williams,2009)的建议,即同时估计没有控制异方差和控制异方差的模型,比较两者的结果是否存在差异,并进行模型选择。而更重要的则是探究这种异方差产生的原因,这有利于我们加深对研究主题的理解,甚或打开一个新的解释路径。

五、小结

本文简述了 Logistic 模型的系数比较问题,并介绍了学者提出的五种解决策略。由于存在未被观测到的异质性及相应的标尺改变问题,Logistic 模型的模型间系数比较和组间系数比较不同于线性回归,不能像线性回归那样进行简单的直接差异比较。其他的 Logit 模型和 Probit 模型存在的问题和解决策略都和本文所介绍的内容基本相同,读者可参考推广。

表 5:二分因变量系数比较不同方法的估计效果

	体现非线性	模型间比较	群体/样本间比较	Stata 相关命令
发生比率(odds ratio)	是	否	否	logistic,logit
y* 标准化	是	是	否	listcoef
KHB 分解	是	是	否	khb
异质选择模型(HCM)	是	否	是	oglm
平均偏效应(APE)	是[a]	是	是	margeff
线性概率模型(LPM)	否	是	是	regress

注:1. a:需在不同的点取值才能体现非线性;

2. 本表部分内容来自穆德(Mood,2010:80)的表 6。

表 5 列举了二分因变量模型中系数比较的 6 种方法,其中发生比

率(或发生比率的自然对数)是研究者常用的,但在诸多方法中,传统的Logistic系数直接比较法带来的偏误最大。在同一样本的嵌套模型比较中,当混杂效应越大时,传统直接比较法导致的偏误就越大(Karlson,*et al*.,2013)。"y* 标准化"和KHB分解适用于模型间比较,但不适用于样本间比较。异质选择模型适用于样本间比较,但不适用于模型间比较。平均偏效应和线性概率模型可以进行模型间和群体间比较,但两者的假定都是线性模型,平均偏效应只有在不同的点取值才能体现出非线性,因此不能很好的拟合数据和反映数据的特征。上述这些方法可能使Logistic模型的系数比较变得更为繁琐,有时甚至难以进行比较,但正如霍特科(Hoetker,2004)所言,这虽然令人沮丧,但总胜过得到虚假结果。希望本文所介绍的方法对研究者有所裨益,以便于更好地理解多样化的社会群体和纷繁的社会现象。

参考文献(References)

鲍威斯、谢宇. 2009. 分类数据分析的统计方法[M]. 任强,等,译. 北京:社会科学文献出版社.

贝里,威廉·D. 2012. 非递归因果模型[M]. 洪岩璧、陈陈,译. 上海:格致出版社.

郭志刚. 1999. 社会统计分析方法[M]. 北京:中国人民大学出版社.

胡安宁. 2014. 教育能否让我们更健康[J]. 中国社会科学(5):116－130.

唐启明. 2012. 量化数据分析:通过社会研究检验想法[M]. 任强,译. 北京:社会科学文献出版社.

伍德里奇, J. M. 2003. 计量经济学导论[M]. 费剑平、林相森,译. 北京:中国人民大学出版社.

谢宇. 2010. 回归分析[M]. 北京:社会科学文献出版社.

中国人民大学中国调查与数据中心中国综合社会调查项目. 2009. 中国综合社会调查报告(2003－2008)[M]. 北京:中国社会出版社.

Allison, Paul D. 1999. "Comparing Logit and Probit Coefficients Across Groups." *Sociological Methods & Research* 28(2):186－208.

Clogg, Clifford C., Eva Petkova, and Adamantios Haritou. 1995. "Statistical Methods for Comparing Regression Coefficients Between Models." *The American Journal of Sociology* 100(5):1261－1293.

Cramer, J. S. 2007. "Robustness of Logit Analysis: Unobserved Heterogeneity and Mis-Specified Disturbances." *Oxford Bulletin of Economics and Statistics* 69(4):545－555.

Hauser, Robert M. and Megan Andrew. 2006. "Another Look at the Stratification of Educational Transitions: The Logistic Response Model with Partial Proportionality Constraints." *Sociological Methodology* 36(1):1－26.

Holm, Anders and Mads Meier Jæger. 2011. "Dealing with Selection Bias in Educational Transition Models: The Bivariate Probit Selection Model." *Research in Social Stratification and Mobility* 29(3):311－322.

Hoetker, Glenn. 2004. "Confounded Coefficients: Extending Recent Advances in the

Accurate Comparison of Logit and Probit Coefficients Across Groups." Working Paper, Oct. 22, 2004. University of Illinois at Urbana-Champaign. (www.public.asu.edu/.../research/Hoetker_confounded_wp.pdf).

Hosmer, David W. and Stanley Lemeshow. 2000. *Applied Logistic Regression*. New York: John Wiley & Sons.

Karlson, Kristian B., Anders Holm, and Richard Breen. 2013. "Comparing Regression Coefficients Between Same-sample Nested Models Using Logit and Probit: A New Method." *Sociological Methodology* 42(1):286－313.

Kohler, Ulrich, Kristian B. Karlson, and Anders Holm. 2011. "Comparing Coefficients of Nested Nonlinear Probability Models." *The Stata Journal* 11(3):420－438.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage Publications.

Long, J. Scott and Jeremy Freese. 2001. *Regression Models for Categorical Dependent Variables Using Stata*. College Station, Texas: Stata Press.

Mare, Robert D. 1980. "Social Background and School Continuation Decisions." *Journal of the American Statistical Association* 75(370):295－305.

Mare, Robert D. 1981. "Change and Stability in Educational Stratification." *American Sociological Review* 46(1):72－87.

Mood, Carina. 2010. "Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It." *European Sociological Review* 26:67－82.

Petersen, Trond. 1985. "A Comment on Presenting Results from Logit and Probit Models." *American Sociological Review* 50(1):130－131.

Wang, Jia and Yu Xie. 2015. "Feeling Good About the Iron Rice Bowl: Economic Sectors and Happiness in Post-Reform Urban China." *Social Science Research* 53:203－217.

Williams, Richard. 2009. "Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients Across Groups." *Sociological Methods & Research* 37(4):531－559.

Williams, Richard. 2010. "Fitting Heterogeneous Choice Models with Oglm." *The Stata Journal* 10(4):540－567.

Winship, Christopher and Robert D. Mare. 1984. "Regression Models with Ordinal Variables." *American Sociological Review* 49(4):512－525.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Sectional and Panel Data*. Cambridge: MIT Press.

责任编辑:张　军