

社会网络规模的影响因素： 不同估计方法的比较*

黄荣贵 桂勇

摘要：在对社会网规模的估算中，OLS方法的适当性和实际效果目前还缺乏较为系统的考察。本文在对现有统计学文献进行梳理的基础上，初步探讨了在估算社会网规模时OLS方法可能存在的问题及可供选择的替代方案。通过对CGSS 2003数据集的统计分析及计算机模拟分析，本文在经验层面发现，OLS估计结果确实与替代方案（特别是负二项回归模型）的估计结果存在明显的差异。最后，本文建议研究者采用可供替代的建模方式研究社会网络规模的影响因素。

关键词：网络规模 OLS 负二项回归

在对社会网规模的估算中，现有文献往往采用OLS方法（例如，张文宏，2005；边燕杰，2004；边燕杰、李煜，2000；Hampton, 2003；La Due Lake & Huckfeldt, 1998；Pinkster & Volker, 2009）。不过，采用OLS对社会网规模进行估算有几个值得特别注意的地方：第一，不同OLS建模策略对结果的影响可能是不一样的。例如，直接对网络规模进行OLS回归与对因变量进行对数变换就可能带来不同结果。然而，研究者很少比较不同OLS建模策略对结果的影响。第二，实际研究中常常对因变量进行对数变换，但是并没有明确对数变换过程中如何处理0取值的问题。如果对“因变量+正数”进行对数变换，研究者似乎很少对该正数的选择做出说明。第三，在某些特定条件下，采用OLS对社会网规模进行估算可能直接影响研究结论，特别是在社会网规模较小、总体均值不大但是方差比较大，或者自变量影响效应较低的情况下更是如此。本文的目的，就是对不同方法估算社会网规模进行较为详细的比较，以便对相应研究的建模策略提供一个参考。

* 本文曾在“中国社会学会社会网与社会资本研究专业委员会成立大会暨第五届社会网与关系管理学术研讨会”上宣读。作者感谢国家社会科学基金项目（05CSH17）与教育部人文社会科学研究规划基金项目（05JA840004）的支持，感谢匿名评审对本文提出的宝贵意见和建议。

一、OLS 估计可能存在的问题及可供选择的替代模型

社会网络规模是一个计数型变量，是取值为非负数的变量。并且，其分布往往呈现明显的偏态性。在经验研究中，研究者往往将社会网络规模看作是服从正态分布的变量，并使用 OLS 方法研究其影响因素。正如统计学文献指出，当解释变量是计数变量而使用 OLS 进行估计，将可能导致非有效的、不一致甚至是有偏的估计量(Long, 1997: 217; King, 1988)。在使用 OLS 方法估算社会网的规模时存在如下问题：首先，OLS 模型的预测值可能是负数，对于网络规模来说，这是无意义的。可以说，使用 OLS 研究网络规模的影响因素，是对统计模型的错误设定。模型的错误设定将导致估计量的一致性(DeMaris, 2004: 352)。其次，对于计数型因变量而言，OLS 模型的残差项不再是同方差的（即存在异方差）。异方差的存在，意味着标准误及相应的显著度不再是准确的估计(DeMaris, 2004: 353)。针对异方差问题，一种常见的做法是对因变量 Y 进行对数变换。如果这么做，研究者就面临一个挑战：如何处理取值为 0 的个案？如果分布的期望较大，分布较接近正态分布，0 取值的个案所占的比例也较少，一种可能的处理办法是将 0 取值剔除，仅考察取值大于 0 的个案。但是，当 0 取值的个案所占的比例很大的时候，这种做法显然不是好的选择。如果将 0 纳入分析，并且希望进行对数变换，一种可能是对 $Y_i + n$ 进行对数变换 (n 是一个常数，常用的取值有 0.1、0.5 和 1 等)，然后对 $\log(Y_i + n)$ 进行 OLS 估计。在社会学研究中，研究目的往往是考察自变量对因变量 Y 的影响。为此，即使在模型估计中对因变量 Y 进行对数变换，在阐释模型结果时，研究者依然需要计算自变量与因变量 Y 的条件期望两者的关系。然而，对第二种对数变换方式而言，该过程似乎并不容易 (Wooldridge, 2002: 645)。在这个意义上说，更加合适的做法是直接对 Y 的条件期望进行建模。上文从统计理论上指出了使用 OLS 估计可能存在的问题。此外，经验研究也指出，使用 OLS 估计对结果可能具有不可忽视的影响，比如，模型系数的显著性可能出现逆转。换言之，在 OLS 中不显著的自变量可能在计数变量模型中变为显著 (Krain, 1998)。

从现有文献来看,针对计数变量进行分析的常用统计模型包括泊松回归(Poisson regression)、负二项回归(Negative binomial regression)和零膨胀回归模型(Zero-inflated model)等(Long, 1997: 221-247)^①。本文考察以上统计模型,理由如下:首先,以上统计模型是处理计数型变量的常见模型(例如 Cameron & Trivedi, 1998; Winkelmann, 2008)。其次,常用统计软件提供以上模型的估计程序(比如 StataCorp, 2009; Zeileis, Kleiber & Jackman, 2008)。在这个意义上说,以上模型具有较强的实用性。接下来,作者将简要介绍泊松回归、负二项回归以及零膨胀回归模型。^②

(一)泊松回归

在泊松回归中,因变量 Y 假定服从泊松分布。其模型为 $E(Y_i | X_i) = \exp(X_i \beta)$ 。该模型保证了因变量的取值为非负数。从泊松分布的特性来看,当分布的均值增加时,0 出现的概率也随之下降,分布也越来越接近正态分布。这意味着,当网络规模较大时,使用 OLS 估计的问题相对较小。然而,当网络规模的均值较小时,0 出现的概率比较大,其分布的偏态性也更加明显,这时候使用 OLS 建模的问题更大一些。

在泊松回归中,因变量的条件方差应该等于条件期望。在实际应用中,因变量的条件方差往往大于条件期望,这种现象被称为过分离散(over-dispersion)。过分离散虽然不影响泊松回归系数的一致性,但是标准误将会被低估,本来不显著的变量可能会变成显著。过分离散状态的出现,可能的原因包括遗漏解释变量,或者所研究的对象存在社会感染(contagion),即彼此之间并不独立。例如,就朋友网的规模而言,即使个人其他特征不变,认识朋友这个过程本身可能影响到下一阶段朋友网建立的速度。因为过分离散现象不影响泊松回归模型系数的一致性,因此一种处理方案是对标准误进行调整,比如使用 Quasi-ML 进行估计,该模型被称为 Quasi-Poisson 回归模型。

(二)负二项回归

^① 虽然 Tobit(杜宾)模型可以解决预测变量为负的问题,但该模型的主要目的是处理截尾数据(censored data,也有学者翻译为删失数据)、样本选择偏误和角点解(corner solution)——即变量正值连续分布,但以正概率取零值的数据结构。本文不考察 Tobit 模型。

^② 本文并不会进行详细的数学推导,感兴趣的读者可参考 Long, 1997; Cameron & Trivedi, 1998。

负二项回归是一种较为常用的处理过分离散现象的模型。负二项回归的模型为： $E(Y_i | X_i) = \exp(X_i \beta) * \delta_i$ ，其中 δ_i 服从以 ν_i 为参数的伽玛（Gamma）分布。为了使模型能够识别(identification)，需要增加两个限定条件： $E(\delta_i) = 1, \nu_i = \alpha^{-1}$ (for $\alpha > 0$)。经数学推导可知， Y_i 的条件期望与泊松回归中的条件期望一致。 Y_i 条件方差等于 $\text{Var}(Y_i | X_i) = \mu_i + \alpha \mu_i^2$ ，其中 μ_i 是条件期望。

(三)零膨胀回归模型

在计数型变量建模的过程中，还经常遇到一种情况，那就是0取值所占的比例远高于模型所预测的比例。对于“过度”出现的0取值，可能具有不同的生成机制。一种可能是，对于某个群体而言，某种类型的社会网是不可能存在的，笔者称之为“结构性的0网络规模”。例如，对于长年在外工作的农民工而言，其拜年网的规模可能属于“结构性的0网络规模”；对于部分独居老人而言，朋友网的规模也可能属于“结构性的0网络规模”。另一种可能是，某个群体的社会网络规模可以是非0的，实际观察值为0是因为随机性所造成。因为在0网络规模的产生机制可能存在不同的类型，研究者在数据分析时应该直接对可能存在的不同机制分别进行建模。对于这种情况，研究者可以使用零膨胀回归模型进行建模。该模型具有两个部分：对结构性0取值的产生机制使用logit回归进行建模；对泊松过程或负二项过程生成的部分使用泊松回归或负二项回归进行建模。如果观察到的数据中取值为0的比例比以上模型所预测要高，则零膨胀回归模型是一种可供选择的模型。

由于现有文献对“结构性0网络规模”的理论探讨较少，在设定零模型的影响因素时缺乏理论指引，也难以使用社会网络相关的理论评估该模型所得结论的适当性。因此，在下一节的经验比较中，本文仅仅拟合零膨胀模型（见附录），但不对其进行详细的考察和评估。

(四)模型选择

数学上可以证明，当负二项回归的参数 α 变为0时，负二项回归的条件方差等价于泊松回归的条件方差，负二项回归将退化为泊松回归。在这个意义上说，泊松回归和负二项回归是嵌套（nested）模型，可以使用似然比检验(likelihood test)来比较两者的拟合程度(Cameron & Trivedi, 1998)。泊松回归、负二项回归与零膨胀回归模型不是嵌套

模型，不能使用似然比检验来比较模型的拟合程度，而应该用 Vuong 检验来比较非嵌套模型的拟合程度(Vuong, 1989)。值得注意的是，不管使用哪种模型，模型和数据的拟合程度仅仅是模型选择的一个标准。在经验研究中，还需要结合理论知识对模型的合理性进行综合选择。接下来，笔者以 CGSS 2003 的数据为例，比较不同模型在社会网规模研究中的不同效果。

从现有社会网络研究的文献可知，大部分研究者对网络规模取对数，然后使用 OLS 进行估计，但是并没有指出如何处理 0 取值的处理方式（例如，张文宏，2005；边燕杰，2004；边燕杰、李煜，2000）。本文的第二部分比较不同模型的时候，仅针对以网络规模对数为因变量的 OLS 模型和计数变量模型。

二、不同模型估计社会网规模的不同效果： 以 CGSS 2003 调查数据为例

（一）数据来源与变量测量

本研究使用的数据是中国人民大学社会学系于 2003 年所收集的中国综合社会调查数据 (CGSS 2003)。抽样方案为四阶段不等概率抽样：第一阶段以区、县为初级抽样单位。抽样框来自行政规划资料，共有 2801 个区、县单位；第二阶段以街道、乡镇为二级抽样单位；第三阶段以居民委员会、村民委员会为三级抽样单位；第四阶段以家庭住户为单位，在每户中确定 1 人为最终单位^①。笔者所使用的数据集共有 5279 个被观察对象^②，其中男性 2687 人，占 50.9%；已婚者有 4831 人，占 91.9%；受过高等教育者有 1070 人，占 20.3%。样本的平均年龄 44 岁，年收入平均 10370 元，中位值 8000 元，标准差为 12900.67。

基于社会资本研究，本小节将具体探讨社会网规模的影响因素。学界对社会资本的定义有不同的理解（张文宏，2007）。本文所指社会资本是个人的社会网络联系，个人社会网络联系越多，社会资本存量也越大。即，网络规模是因变量。主要的解释变量是被访者的阶层。

^① 关于数据收集的更加详细的资料，请参见其网站 www.chinagss.org

^② 考虑到下文统计模型将引入职业作为主要解释变量，无职业信息的观察对象在数据预处理时直接删除。5297 是具有职业信息的个案数。

此外，作者结合社会资本的研究文献选取控制变量。变量的测量如下：

1. 因变量：网络规模

在本研究中，因变量是讨论网和拜年网的网络规模。讨论网规模的测量指标是：“在过去半年内，与您和讨论过对您来说是重要的问题的人的个数”。讨论网络平均规模是 5.25 人，其中 4.86% 的被访者的讨论网规模为 0。拜年网的网络规模测量指标是：“在今年春节期间，以各种方式互相拜年、交往的亲属、亲密朋友和其他人共有多少人？”拜年网的平均网络规模是 28.24，其中 4.71% 的被访者的网络规模为 0（详见表 1）。

2. 自变量：社会网络的影响因素

根据现有文献可知，阶层(张文宏, 2005; 边燕杰, 2004; 边燕杰、李煜, 2000)是影响社会网络的重要因素。此外，工作状态，特别是是否全职工作可能会直接影响个人的社会网规模。在本文中，阶层的一个测量指标是被访者最近的职业；在某种程度上，该指标同时反映了被访者的阶层和工作状态。具体而言，使用的职业分类沿用 CGSS 2003 所提供的职业分类标准，包括：下岗，家务劳动者，退休、农、林、牧、渔、水利业生产人员，办事人员和有关人员，专业技术人员，个体户，机关、党群组织、企事业单位负责人。

个人所获取的社会资本总量存在着一定的代际传递性，依赖于家庭所拥有的社会资本总量。父母的教育，尤其是母亲的教育水平与孩子的社会资本具有密切相关 (Halpern, 2005)。本研究将引入父母的教育水平作为解释变量（父母是否受过高中或者以上的教育）。普特南指出，看电视可能对社会资本具有消极的影响，而一起打保龄球则有助于增进社会资本 (Putnam, 2000)。后续研究指出，看电视和社会流动性对社会资本具有交互效应，其影响并不必然是负面的（比如，Kang & Kwak, 2003）。因此，被访者的空余时间消耗在看电视的程度将是其中一个解释变量。对于中国人而言，一起打牌或者麻将对于社会资本的影响可能与一起打保龄相似，因为两者都是集体性的休闲活动。基于这一认识，笔者将周末是否经常打牌或者麻将作为自变量。

基于对现有文献的回顾，本研究引入的其他变量包括户口所在地（边燕杰, 2004），性别、婚姻状态和年龄（张文宏, 2005），教育程度和收入（边燕杰、李煜, 2000）。户口所在地的测量为“被访者户口是否属于本县市”。本研究对年龄变量进行转换，模型中的年龄指标 = (实际年龄 - 样本年龄平均值) / 10。教育程度变量方面，被访者被分为

受过高等教育与未受过高等教育两类。为了减少收入极值的影响，根据个人年收入，本文将被访者分为高收入群体和低收入群体，收入大于或者等于平均值者属于高收入群体，否则属于低收入群体。

表 1 网络规模的描述统计

	讨论网	拜年网
均值	5.25	28.24
方差	270.89	821.40
0 网络规模的比例	4.86%	4.71%
个案数	4852	4779

注：1) 均值和方差使用人权重加权^①。2) 由于变量包含缺失值，表中汇报的有效个案数小于前文所提及的个案总数（5297）。^②

（二）网络规模的影响因素：不同估计方法之比较

从表 1 可知，网络规模的方差远大于其均值，因此存在过分离散现象。在模型拟合过程中，统计检验也证实过分离散现象的存在。因此在这里不汇报泊松回归的结果，而是汇报 Quasi-Poisson 回归的结果^③。

1. 讨论网

比较不同建模方法对讨论网网络规模的影响因素的估计（见下表 2），可以发现：

（1）即使同样采用 OLS 进行估计，把 0 网络规模的样本放入模型和把 0 网络规模的样本剔除出模型将直接影响最后的结论。比较模型 1 和模型 2 可知，模型 2 中关于阶层的几个变量（家务劳动者、专业技术人员、党政机关企事业负责人）不再显著。

（2）总体而言，Quasi-Poisson 模型各个系数的显著性介乎于 OLS 模型和负二项回归模型之间，唯有看电视一项例外。换言之，Quasi-Poisson 模型部分系数的显著性与 OLS 相一致，部分回归系数的显著性与负二项回归模型一致。

^① CGSS 2003 数据集提供不同的加权重，人权重是其中一种加权方式。下同。

^② 由于缺失值的存在，下文统计模型的有效个案数均小于个案总数（5297）。本文并没有对变量的极值进行特别处理。

^③ 基于上文所提及的原因，本文不对“零膨胀模型”进行深入的讨论。但是，本文拟合了零膨胀负二项回归模型，详细结果见附表 1。从社会网络的代际传递的理论来看，家庭背景可能是影响 0 网络规模的一个结构性因素，因此在拟合零膨胀负二项回归模型时，0 模型分部引入了父母亲的教育程度。此外，是否是本县市人也作为一个结构性的制约因素引入 0 模型分部。

(3) 将模型 4 与其他 3 个模型相比, 并且结合现有的研究结果, 似乎模型 4 的估计结果与理论预期更为一致: 第一, 本县市人的系数为正, 并且显著, 这意味着讨论网规模与社会流动性之间存在着反向的关系。第二, 年龄系数不显著, 这可能是没有细分讨论网的子类型导致的。现有研究指出, 不同年龄的群体的社交模式具有很大的差异, 年轻群体在朋友网络方面占有优势, 但是老年人群体在邻里网络和志愿活动方面可能更有优势 (Halpern, 2005); 因此, 没有细分讨论网的子类型可能掩盖其中的差异, 导致年龄系数不显著。第三, 在模型 4 中, 家务劳动者都不显著, 并且这一结果与理论预期相一致。如果讨论网是通过市场联系和组织联系而展开 (边燕杰, 2004), 我们没有理由认为家务劳动者的讨论网规模更大。第四、其他阶层变量的影响方面, 似乎模型 1 和模型 4 的结果比较一致, 模型 2 和模型 3 的结果比较一致; 然而, 总的来说, 似乎 OLS 估计低估了个体户阶层讨论网的规模。个体户基层具有相当强的市场联系, 讨论网的规模较大与常识相符, 虽然这种网络联系是否能转换为拜年网值得进一步考察。第四, 负二项回归中, 看电视对于讨论网规模具有促进作用。这一发现与模型 1、模型 2 是一致的。该发现与现有研究结论——看电视会降低人际间的联系——相左。这说明看电视对社会资本没有普遍的影响。

与现有理论预测不同的是, 周末与朋友打牌或麻将、父母的教育程度对讨论网规模没有影响。一种可能是周末与朋友打牌或麻将主要以休闲、娱乐为目的, 这种社会联系无法转换为以社会讨论为基础的社会资本。父母教育程度不显著意味着当代中国存在着较大的代际流动。最后, 四个模型一致指出, 低收入者的讨论网规模比较小。

表 2 讨论网规模的阶层差异

自变量	模型 1		模型 2		模型 3		模型 4	
	OLS		OLS		Quasi Poisson		Negative Binomial	
	Log(网络规模)		Log(网络规模+01)		网络规模		网络规模	
	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.
截距	0.93**	0.10	0.96**	0.13	1.34**	0.34	1.31**	0.13
本县市人	0.03	0.06	-0.06	0.08	0.25	0.21	0.20*	0.08
年龄	-0.06**	0.02	-0.10**	0.02	0.01	0.05	-0.02	0.02

社会学研究

男性	0.01	0.03	-0.03	0.04	0.13	0.09	0.10**	0.04
高教育程度	0.07+	0.04	0.07	0.05	-0.07	0.13	-0.04	0.05
低收入群体	-0.07*	0.03	-0.11**	0.04	-0.30**	0.09	-0.24**	0.03
已婚	0.03	0.06	0.08	0.07	-0.02	0.18	0.01	0.07
下岗	0.03	0.06	-0.03	0.07	0.04	0.19	0.02	0.07
家务劳动者	0.22*	0.10	0.11	0.12	0.16	0.34	0.13	0.12
退休	0.08	0.06	0.11	0.07	-0.07	0.18	-0.01	0.07
农、林、牧、 渔、水利业 生产人员	0.00	0.12	-0.14	0.15	-0.02	0.40	-0.01	0.14
办事人员和 有关人员	0.06	0.05	0.02	0.06	0.05	0.15	0.04	0.06
专业技术人 员	0.12*	0.06	0.12+	0.07	0.19	0.17	0.18**	0.07
个体户	0.10+	0.05	0.08	0.07	0.59**	0.15	0.52**	0.06
机关、党群 组织、企事 业单位负责 人	0.16**	0.06	0.14+	0.08	0.18	0.18	0.23**	0.07
常看电视	0.06**	0.02	0.04*	0.02	0.06	0.05	0.06**	0.02
常玩牌或打 麻将	-0.00	0.01	-0.00	0.02	-0.01	0.05	-0.01	0.02
母亲教育程 度(>=高中)	-0.02	0.06	0.00	0.07	-0.01	0.18	0.00	0.07
父亲教育程 度(>=高中)	0.04	0.04	0.08	0.05	0.01	0.13	0.04	0.05
Theta	--		--		--		1.17	0.03
个案数	3524		3684		3684		3684	
AIC	9152.33		11491.59		---		20706.99	

注：1) 以人权重加权。2) 模型 1 仅引入网络规模大于 0 的个案。3) 阶层变量以“生产、运输设备操作人员及有关人员”为基准项。4) 由于不同模型的因变量不同，AIC 统计量不具有可比性。因此，研究者无法直接根据 AIC 进行模型选择。下同。

+ p<.10 * p<.05 ** p<.01

2. 拜年网

对拜年网网络规模的统计分析发现（详见下表 3），不同建模方法的结果依然存在差异，但是彼此之间的差异较讨论网的差异小。表 2 中不同模型比较发现有 27 项显著的差异（以 0.05 作为显著的分界点），但是表 3 仅有 13 项差异。这可能与拜年网的平均规模比较大（拜年网网络规模的均值为 28.24）有一定的关系。就表 3 各模型之间的差异而言，我们可以发现：

（1）即使同样采用 OLS 估计，是否将 0 规模样本包括进来对最终结果有较为显著的影响。

（2）Quasi-Poisson 回归模型各个系数的显著性几乎与负二项回归模型一致，仅有婚姻状态例外。然而，虽然负二项回归模型中的已婚群体的系数不显著，但是其显著度已经非常接近 0.05（实际显著度为 0.054）。

（3）4 个模型的主要差异集中于阶层变量：仅有家务劳动者、专业技术人员两个阶层变量的统计结果在 4 个模型是一致的。党政机关、企事业单位等阶层变量仅在模型 1 显著，办事人员及有关人员阶层变量仅在模型 2 是不显著的。换言之，模型 3 和模型 4 的结论比较一致，并且部分得到 OLS 模型的支持。如果我们接受模型 3 和模型 4 的结论，则意味着党政机关、企事业单位负责人的拜年网规模并不比工人阶级的规模大，该发现与现有的研究并不一致（边燕杰，2004；边燕杰、李煜，2000；张文宏，2005）。一种可能是，在市场化进程中，传统的权力关系对社会网的影响开始消减。另一种可能是，拜年网的生成可能受协调组合（assortative mixing）机制的影响^①。即，高权者给高权者拜年，而低权者给低权者拜年（关于拜年网的阶层分割与渗透的详细讨论，参见边燕杰等,2005）^②。因此权力可能和拜年网规模无直接关系。对于以上猜测，需要研究者进一步考证。

值得一提的是，看电视对于拜年网规模似乎没有影响；与之相对，常和朋友玩牌或麻将的人的拜年网规模更大。一种可能是，一起玩牌或者麻将的朋友本身就是属于核心社会网成员，因此彼此之间拜年的可能性比较大。父母教育程度对拜年网规模没有影响。

^① 关于社会网络生成机制的简要评述，可参考 Goodreau et al.,2009:105-106。

^② 作者感谢匿名评审指出这一解释。

3. 讨论网与拜年网的比较

对表 2 和表 3 的比较分析显示，似乎讨论网和拜年网是两种性质不同的社会网络，各自具有不同的影响因素。以负二项回归模型为例，对讨论网和拜年网进行比较（见下表 4）。结果显示，共有 13 个变量对讨论网或者拜年网有影响，这 13 个变量中仅有 3 个变量的影响是一致的（低收入群体、办事人员和有关人员、专业技术人员）。其中，有几个变量值得进行简要的讨论：第一，家务劳动者仅对拜年网规模具有正向的影响，这可能反映了中国传统的家庭分工模式。第二，个体户阶层对讨论网规模有影响，但对拜年网规模没有影响。第三，看电视仅对讨论网规模有影响，这可能因为看电视是获得信息的渠道，而获取信息有助于朋友之间进行讨论。相比之下，玩牌或麻将可能发生在熟人和朋友之间。一直可能的解释是，熟人和朋友网的规模会影响被访者是否玩牌或者玩麻将，而不是相反。最后，虽然没有在表 4 中反映出来，然而，表 2 和表 3 中大部分模型都表明，父母的教育程度似乎与个人的网络规模没有太多的关系。这可能与当前中国社会流动较大有关。

此外，根据对数变换模型^①和负二项回归模型对网络规模进行预测，然后计算实际网络规模与预测规模的相关系数，结果如下：（1）对讨论网而言，对数变换模型所得相关系数为 0.04，负二项回归模型所得相关系数为 0.07。（2）对拜年网而言，对数变换模型所得相关系数为 0.23，负二项回归模型所得相关系数为 0.24。比较可知，负二项回归模型预测效果优于对数变化模型。并且，当网络规模较小时（比如，讨论网）该优势更明显。

表 3 拜年网网络规模的阶层差异

模型类型 因变量	模型 1 OLS Log(规模网络)		模型 2 OLS Log(网络规模+01)		模型 3 Quasi Poisson 网络规模		模型 4 Negative Binomial 网络规模	
	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.
截距	2.74**	0.12	2.69**	0.17	3.25**	0.12	3.19**	0.11
本县市人	0.00	0.07	-0.08	0.10	-0.03	0.07	-0.02	0.07
年龄	-0.11**	0.02	-0.13**	0.03	-0.09**	0.02	-0.08**	0.02

^① 表 2 和表 3 的模型 2。

男性	0.04	0.03	-0.09+	0.05	0.00	0.03	0.01	0.03
高教育程度	0.14**	0.05	0.21**	0.07	0.14**	0.05	0.14**	0.04
低收入群体	-0.18**	0.03	-0.26**	0.05	-0.23**	0.03	-0.21**	0.03
已婚	0.17**	0.06	0.10	0.10	0.14*	0.06	0.12+	0.06
下岗	-0.02	0.06	-0.17+	0.09	-0.02	0.07	-0.03	0.06
家务劳动者	0.36**	0.11	0.32*	0.16	0.39**	0.11	0.38**	0.11
退休	0.02	0.06	-0.11	0.09	-0.04	0.07	-0.03	0.06
农、林、牧、渔、水利业生产人员	-0.04	0.14	-0.17	0.21	-0.02	0.16	-0.03	0.14
办事人员和有关人员	0.12*	0.05	0.07	0.07	0.13*	0.05	0.12*	0.05
专业技术人员	0.18**	0.06	0.20*	0.09	0.13*	0.06	0.15*	0.06
个体户	0.04	0.06	0.05	0.09	0.09	0.06	0.07	0.06
机关、党群组织、企事业单位负责人	0.17*	0.07	0.06	0.10	0.06	0.07	0.08	0.07
常看电视	0.00	0.02	0.02	0.02	-0.01	0.02	0.01	0.02
常玩牌或打麻将	0.07**	0.02	0.12**	0.02	0.06**	0.02	0.06**	0.02
母亲教育程度 (>=高中)	0.02	0.06	-0.00	0.09	-0.02	0.06	-0.03	0.06
父亲教育程度 (>=高中)	0.06	0.05	0.14*	0.07	0.05	0.05	0.06	0.05
Theta	--		--		--		1.27	0.03
个案数	3463		3610		3610		3610	
AIC	9632.57		13018.84		--		32482.42	

表4 讨论网和拜年网影响因素的比较：以负二项回归模型为例

显著的影响因素	讨论网	拜年网
本县市人	+	ns
年龄	ns	-
男性	+	ns

高教育程度	ns	+
低收入群体	-	-
已婚	ns	+
家务劳动者	ns	+
办事人员和有关人员	+	+
专业技术人员	+	+
个体户	+	ns
机关、党群组织和企事业 单位负责人	+	ns
常看电视	+	ns
常玩牌或麻将	ns	+
影响一致的变量（总变量 数）	3(13)	

注：+：正向显著；-：负向显著；ns：不显著

三、对负二项分布变量建模的比较：一个模拟的结果

从上文的描述分析可知，社会网规模的方差远大于其均值；从不同模型的比较结果来看，负二项分布回归模型的结果也比较符合常识和理论预期。这些迹象似乎表明，社会网规模的分布比较接近负二项分布。如果这样，那接下来的问题是：如果网络规模确实服从负二项分布，使用 OLS 回归和负二项分布回归模型估计结果是否有差异；如果有，究竟差异与什么因素有关？从理论上说，这一方面可能与变量的均值有关；另一方面，可能与自变量对因变量的影响效应大小有关。为了回答该问题，笔者通过统计模拟进行分析。尽管统计模拟无法在统计学意义上证明这种差异是否存在，但模拟结果能提供相关的经验证据，具有重要的参考价值。设计统计模拟方案的时候主要考虑了以下因素：1、样本量。不失一般性，本文仅引入一个自变量。考虑到模型参数比较少，为了减低由大样本引起的虚假显著，样本量应该中等偏小。在具体模拟中，样本量为 200。2、自变量分布。为了使模拟情景与实际研究更加一致，作者并没有采用随机数作为自变量，而是在 CGSS 2003 数据集中选取“常看电视”这一变量，并从该变量中随机

抽取了 200 个观察^①。3、自变量对因变量影响效应的大小(下图中的 size of effect)和模型的截距(下图中的 a)。参照上文各回归系数的大小, 选取了 -0.15 到 0.15 之间的影响效应进行模拟。截距的取值是 1.5, 2, 2.5, 3。4、参考上文负二项回归模型中的 Theta 参数大小, 模拟中对 0.75, 1, 1.25 和 1.5 的情形进行模拟。

对每一种模拟设定, 笔者采用统计软件 R (R Development Core Team, 2010)生成因变量, 并采用四种不同的建模策略进行分析, 考察自变量系数是否显著并且和实际影响效应的方向一致。四种不同的建模策略分别是: 负二项回归模型(negative binomial regression)、对因变量直接进行 OLS 回归分析(OLS:y)、对 $\log(\text{因变量}+0.1)$ 进行 OLS 回归分析(OLS: $\log(y+0.1)$)、仅对大于 0 的因变量取对数, 并对变换后的变量进行 OLS 回归分析(OLS: $\log(y|y>0)$)。对每个模拟设定重复 500 次, 计算出“自变量系数显著并且和实际影响效应方向一致”的模型所占的比例。

^① 正如匿名评审指出的, 模拟设计并没有考虑未观察变量对模型估计的影响。然而, 本模拟的主要目的是, 在假定数据生成过程为负二项分布的基础上比较 OLS 回归和负二项回归模型的估计结果。这意味着, 负二项回归模型的误设(misspecification)对估计结果的影响不是本模拟的关注点。因此, 选择“看电视”为自变量并不失一般性。当然, 若考察模型误设对负二项回归模型估计的影响, 可以在数据生成过程中引入未观察变量。这可在将来的研究中进一步探索。

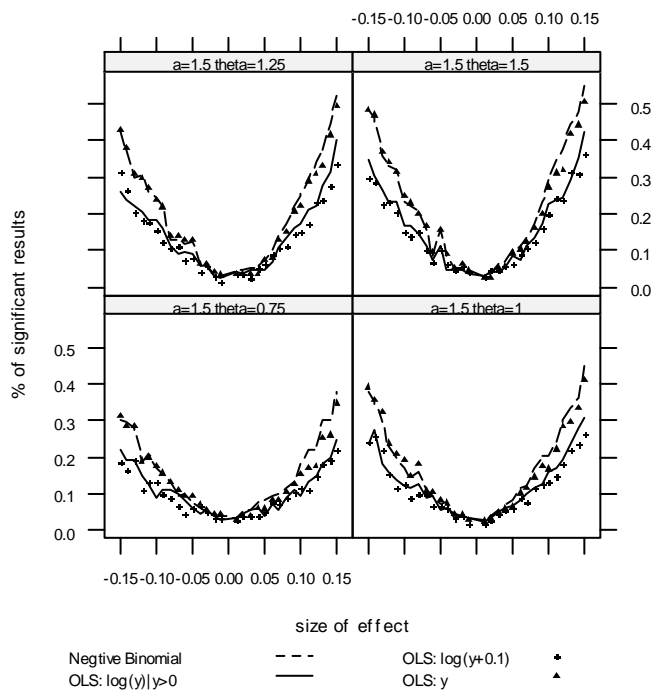


图 1 对负二项分布变量建模方法比较的模拟分析结果(I)

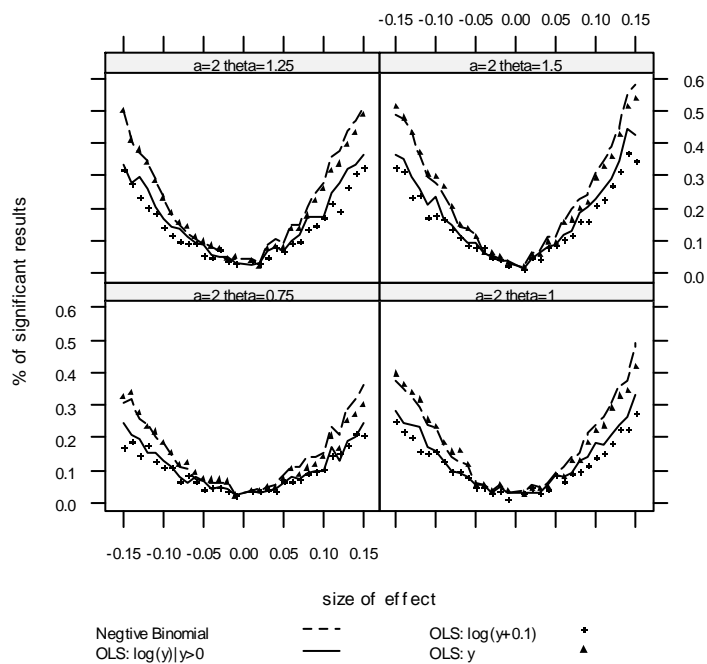


图 2 对负二项分布变量建模方法比较的模拟分析结果(II)

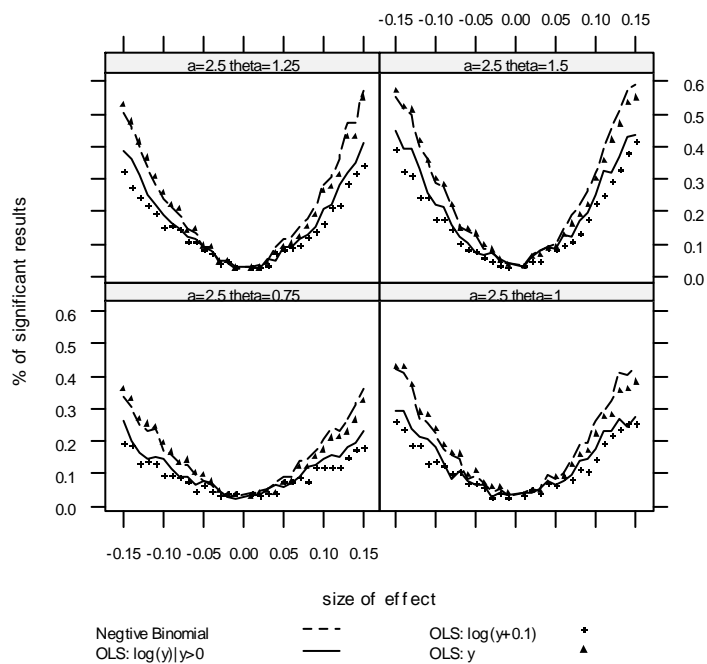


图3 对负二项分布变量建模方法比较的模拟分析结果(III)

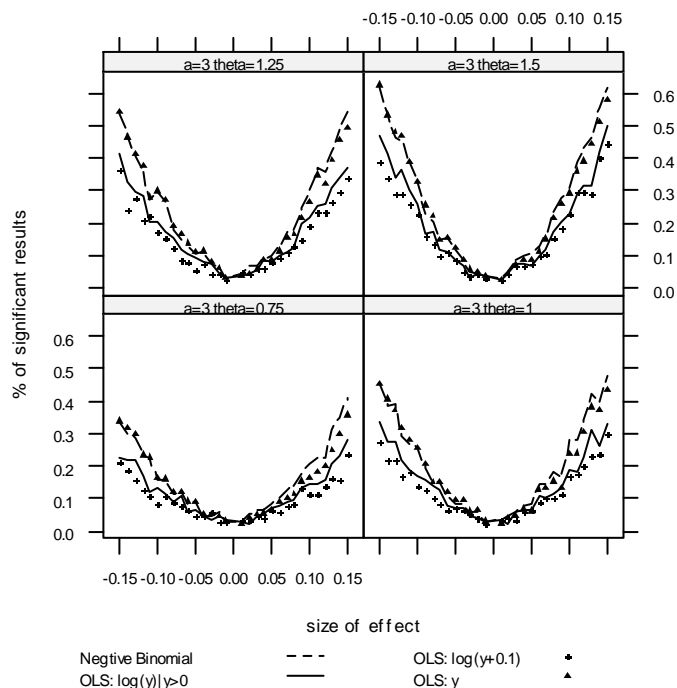


图 4 对负二项分布变量建模方法比较的模拟分析结果(III)

对模拟结果（图 1-4）进行比较，有如下发现：

1. 若以正确估计出正确方向的显著系数为标准，对于同一个摹拟情景，负二项回归模型的效果最好。负二项回归模型的效果略优于对因变量直接进行 OLS 回归分析（若还综合考虑模型系数大小，负二项回归模型更是优于 OLS 回归模型）。对 $\log(\text{因变量}+0.1)$ 进行 OLS 回归分析，仅对大于 0 的因变量取对数、并对变换后的变量进行 OLS 回归分析这两种方法的效果远弱于其他两种方法，其中对 $\log(\text{因变量}+0.1)$ 进行 OLS 回归分析的效果最差。总的来说，若自变量的实际效应不为 0，但 OLS 回归估计得到显著结果的可能性远低于负二项回归模型。值得一提的是，虽然取对数的目的是使得因变量分布比较对称，从而更加接近 OLS 回归模型的假设，并且该方法被广为使用，但是出乎意料的是，这种方法似乎是最不合适的方法，至少模拟分析的结果是这样。

2. 对于同一建模方法，自变量的影响效应的绝对值越大，模型得到显著性结果的比例越高。

3. 对于同一个均值的负二项分布，方差越小(即 θ 越大)，模型得到显著性结果的比例越高。

4. 比较图 1-4 对应列可知，对于相同的 θ 值， a 值的增加意味着分布均值增加；而分布均值越大，模型得到显著性结果的比例越高。为了考察该发现是否与自变量选择有关，笔者选择年龄作为自变量，重新模拟，其发现与之基本一致。从以上分析可知，当社会网络规模均值较小，方差比较大，并且自变量对网络规模影响效应较小的时候，使用 OLS 方法估计网络规模的影响因素，可能会误将实际有影响的因素看作没有影响，即低估了变量的显著性^①。

五、数据、变量和测量

现有的统计理论指出，使用 OLS 对计数型变量进行建模，其估计是不一致、非有效、有偏的。对 CGSS 2003 数据中的社会网规模不同估算方法效果的分析表明，OLS 估计结果与 Quasi-Poisson 回归、负二项回归结果存在明显的差异。不同估计方法之间的差异在讨论网的分析中更为明显，这可能与讨论网的平均网络规模较小有关。即使研究者仅仅关注自变量的影响方向以及模型显著性检验是否正确，而不关心影响效应的强度，OLS 估计方法依然不是一个最优的选择。模拟分析的结果显示，对于服从负二项分布的计数变量而言，采用 OLS 估计方法会高估参数的标准误。换言之，如果研究者使用 OLS 估计方法，将实际上显著的自变量误判为不显著的可能性远远高于负二项回归模型。当负二项分布的计数型变量的方差较大、均值较小、自变量的影响效应较小的时候，误判的可能性较大。这对我们的实际研究工作有着一定的指导意义。包括笔者在内的一些研究者在研究影响社会网规模的因素时，常常发现引入模型的各自变量很难取得显著的效果，可能的原因，就在于研究者没有采用合适的建模方法。实际上，研究者常常对网络规模进行对数变化，然后采用 OLS 进行分析；而模拟方

^① 在实际分析中，研究者可能发现，OLS 模型中显著的系数在负二项回归模型中反而不显著，这有两种可能：一是由随机性所引致。这并不与模拟结果矛盾，因为模拟结果是对多数模拟的平均。二是可能因变量并不是严格服从负二项分布。

法的结果恰恰显示，这是效果较差的建模方式。基于以上发现，本文认为今后关于社会网规模影响因素的研究应该尽量采用针对计数型变量的建模方法。

此外，本研究对于定量研究中的建模方法的选择具有一定的启发意义。正如博克斯和诺曼(Box & Norman,1987:424)指出，“所有的统计模型都是错误的，但是一些是有用的。”任何一个统计模型，都有相应的假定，而这些假定与数据的实际生成过程的吻合程度是有差异的。只有当模型的假定与实际数据较为吻合，其结果才比较可靠。在这个意义上说，研究者应该对建模方法的选择进行讨论，并对模型的研究假定进行检验，而不是机械地选择。就社会网规模的影响因素的相关研究而言，学者对选择何种模型估计网络规模存在不同的看法；然而，不管选择何种统计模型，给出必要的论证依然是一种好的做法。比如，范德波尔（van der Poel，1993：第5章）使用方差分析来比较网络规模的社会-结构差异时，他明确指出了使用方差分析的两个原因：一、在他的研究中，网络规模比较接近正态分布。二、方差分析更加简单。在其他章节，范德波尔主要使用泊松回归模型分析网络规模的影响因素，因为泊松回归模型在统计学意义上更加合适(见其第67页的脚注2以及关于方法论的附录B)。值得注意的是，并非所有社会网络规模的测量指标都接近正态分布。如果网络规模测量指标明显偏离正态分布，研究者应尽量采用针对计数型变量的建模方法。如果研究者对计数型变量模型依然持有怀疑的态度，本文建议研究者同时采用 OLS 和计数型变量模型进行分析，结合对模型假定的检验、模型的拟合程度和理论指引对模型进行讨论并做出最后的选择。

自变量	Model 1 讨论网		Model 2 拜年网	
	Coefficient	S.E.	Coefficient	S.E.
计数模型分部				
截距	1.31**	0.13	3.20**	0.11
本县市人	0.20*	0.08	-0.01	0.06
年龄	-0.02	0.02	-0.08**	0.02
男性	0.10**	0.04	0.03	0.03
高教育程度	-0.04	0.05	0.13**	0.04

低收入群体	-0.24**	0.03	-0.20**	0.03
已婚	0.01	0.07	0.13*	0.06
下岗	0.02	0.07	-0.01	0.06
家务劳动者	0.13	0.12	0.39**	0.10
退休	-0.01	0.07	-0.01	0.06
农、林、牧、渔、水利业生产人员	-0.01	0.14	-0.01	0.13
办事人员和有关人员	0.04	0.06	0.12**	0.05
专业技术人员	0.18**	0.07	0.15**	0.06
个体户	0.52**	0.06	0.07	0.05
机关、党群组织、企事业单位负责人	0.23**	0.07	0.10	0.06
看电视	0.06**	0.02	0.00	0.02
玩牌或打麻将	-0.01	0.02	0.05**	0.01
母亲教育程度	0.00	0.07	-0.03	0.06
父亲教育程度	0.04	0.05	0.04	0.04
Log(theta)	0.17**	0.03	0.43**	0.03
截距	-22.98	7931.28	-4.21**	0.89
母亲教育程度	-0.70	5458.18	-0.86	1.25
父亲教育程度	-0.78	3596.46	-1.40+	0.75
本县市人	1.59	7987.43	0.97	0.89
个案数	3684.00		3610.00	
Vuong 检验(零膨胀负二项回归 vs. 负二项回归)	支持负二项回归 p-value 0.4666819		支持零膨胀负二项回归 p-value 4.65622e-06	
与负二项回归比较	--		婚姻状态变为显著	

注：1) 以人权重加权。 2) 模型 1 仅引入网络规模大于 0 的个案。3) 阶层变量以“生产、运输设备操作人员及有关人员”为基准项。

+ p<.10 * p<.05 ** p<.01

参考文献：

- 边燕杰, 2004, 《城市居民社会资本的来源及作用：网络观点与调查发现》, 《中国社会科学》第 3 期。
- 边燕杰、李煜, 2000, 《中国城市家庭的社会网络资本》, 《清华社会学评论》(特辑 2)。
- 边燕杰、Ronald Breiger、Deborah Davis、Joseph Galaskiewicz, 2005, 《中国城市的职业、阶层和

- 关系网》，《开放时代》第4期。
- 张文宏, 2005, 《城市居民社会网络资本的阶层差异》，《社会学研究》第4期。
- , 2007, 《中国的社会资本研究: 概念、操作化测量和经验研究》，《江苏社会科学》第3期。
- Box, George E.P. & Draper Norman R. 1987, *Empirical Model-building and Response Surfaces*. New York: Wiley.
- Cameron, A. C. & P. K.Trivedi 1998, *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.
- DeMaris, A. 2004, *Regression with Social Data: Modelling Continuous and Limited Response Variables*. Hoboken: John Wiley & Sons.
- Halpern, David 2005, *Social Capital*. Cambridge: Polity Press.
- Hampton, K. N. 2003, "Grieving for a Lost Network: Collective Action in a Wired Suburb." *Information Society* 19 (5).
- Goodreau, Steven, James A. Kitts & Martina Morris 2009, "Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks." *Demography* 46 (1).
- Kang, Naewon & Nojin Kwak 2003, "A Multilevel Approach to Civic Participation: Individual Length of Residence, Neighborhood Residential Stability, and Their Interactive Effects With Media Use." *Communication Research* 30 (1).
- King, Gary 1988, "Statistical Models for Political Science Event Counts: Bias in Conventional Procedures and Evidence for the Exponential Poisson Regression Model." *American Journal of Political Science* 32 (3).
- Krain, Matthew 1998, "Contemporary Democracies Revisited: Democracy, Political Violence, and Event Count Models." *Comparative Political Studies* 31 (2).
- La Due Lake, R. & R. Huckfeldt 1998, "Social Capital, Social Networks, and Political Participation." *Political Psychology* 19 (3).
- Long, J. Scott 1997, *Regression models for categorical and limited dependent variables*. Thousand Oaks : Sage Publications.
- Pinkster, F. M. & B.Volker 2009, "Local Social Networks and Social Resources in Two Dutch Neighbourhoods." *Housing Studies* 24 (2).
- Putnam, R. D. 2000, *Bowling alone: The collapse and revival of American community*. New York: Simon & Schuster.
- R Development Core Team 2010, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

StataCorp. 2009, *Stata Statistical Software: Release 11*. College Station, TX: StataCorp LP.
www.stata.com.

Van der Poel, Mart. 1993, *Personal Networks: A Rational-choice Explanation of Their Size and Composition*. Lisse [Netherlands]: Swets & Zeitlinger.

Vuong, Q.H. 1989, "Likelihood ratio tests for model selection and non-nested hypotheses."
Econometrica. 57.

Winkelmann, R. 2008. *Econometric Analysis of Count Data*. New York: Springer.

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

Zeileis, A.; Kleiber, C. & S. Jackman 2008, "Regression Models for Count Data in R." *Journal of Statistical Software* 27 (8). Retrived from <http://www.jstatsoft.org/v27/i08>, 28 March 2010

作者单位：香港城市大学公共与社会行政系（黄荣贵）

复旦大学社会学系（桂 勇）

责任编辑：闻 翔

文章来源：《社会学研究》2010年第4期

中国社会学网 www.sociology.cass.cn