奥地利统计学期刊 *MMMMMM*
YYYY, 第*VV*卷, *1–*
xx.http://www.ajs.or.at/

arXiv:2301.00921v1 [stat.ME] 3 Jan 2023

计数数据的多变量广义线性混合模型

Guilherme P.Silva
巴拉那联邦大学

Henrique A.Laureano佩克诺普林西皮研究所

Ricardo R. Petterle
巴拉那联邦大学

Paulo J. R.J´
uniorParan´a 联邦大学

Wagner H.
BonatParan´a 联邦大学

摘要

计数数据的单变量回归模型已有丰富的研究文献。然而，多变量计数数据方面的研究却并非如此。因此，我们提出多变量广义线性混合模型框架，用于处理一组多变量响应，并通过服从多变量正态分布的随机效应来衡量响应之间的相关性。该模型基于包含随机截距的广义线性混合模型，估计过程与标准广义线性混合模型相同，即通过拉普拉斯近似积分消去随机效应。我们使用 R 中提供的 TMB 软件包高效地实现了该模型。我们采用了泊松分布、负二项分布（NB）和 COM-泊松分布。为评估估计量的性质，我们开展了一项模拟研究，针对每种分布考虑四种不同的样本量和三种不同的相关值。对于泊松分布和负二项分布，我们获得了无偏且相合的估计量；对于 COM-泊松分布，估计量是相合的，但存在偏差，尤其是离散度、方差和相关参数估计量。我们将这些模型应用于两个数据集。第一个数据集涉及在澳大利亚 30 个不同地点采集的样本，记录了 41 种不同蚂蚁物种各自出现的次数；这意味着需要同时估计多达 820 个方差-协方差参数和 41 个离散参数，更不用说回归参数了。第二个数据集来自澳大利亚健康调查，包含 5 个响应变量和 5190 名受访者。根据广义离散指数，这些数据集可视为过度离散。根据三项拟合优度指标——赤池信息准则、贝叶斯信息准则和最大化对数似然值，COM-Poisson模型优于另外两个竞争模型。因此，该模型估计出的参数标准误更小，显著相关系数的数量也更多。由此可见，所提出的模型能够处理多变量计数数据，无论响应变量呈欠离散、等离散还是过度离散，都能在考虑协变量影响的同时衡量变量间的任意相关性。

关键词：拉普拉斯近似；多变量模型；优化；过度离散计数数据；回归模型；Template ModelBuilder。

1. 引言

众所周知，在广义线性模型（GLM）框架下，泊松分布是处理计数数据最常用的模型（Nelder 和 Wedderburn 1972）。然而，它仅适用于等离散计数数据，即响应变量的均值等于方差的情况。作为泊松模型的替代方案，单变量计数数据的统计学文献十分丰富，既涵盖过度离散，也涵盖欠离散。例如，以均值和离散参数为参数的负二项分布（NB）II 型分布（称为方差的二次参数化）（Winkelmann 2008）、障碍模型和零膨胀模型（Zeileis、Kleiber 和 Jackman 2008），以及混合泊松回归（Winter 和 B¨urkner 2021）是用于建模过度离散计数数据的常见选择；伽马计数分布（Zeviani、Ribeiro Jr、Bonat、Shimakura 和 Muniz 2014）、Conway-Maxwell-Poisson 分布（COM-泊松分布）（Shmueli、Minka、Kadane、Borle 和 Boatwright 2005）以及基于泊松 Tweedie 分布的扩展泊松 Tweedie 分布（Bonat、Jørgensen、Kokonendji、Hinde 和 Dem´etrio 2018）（El-Shaarawi、Zhu 和 Joe 2011；Jørgensen 和 Kokonendji 2016）可以处理欠离散、等离散和过度离散这三种情况，尽管其概率质量函数（pmf）的计算依赖数值方法。

另一方面，关于多变量数据的文献很少。然而，研究人员越来越需要分析包含多个响应变量的数据集。这样做的好处是能够更好地考察响应变量之间的关系（Bonat 和 Jørgensen 2016）；同时，更多可用数据也能让数值方法更充分地发挥作用，从而估计模型参数。

Winkelmann （2008）提出并介绍了一些处理多变量数据的方法，重点在于构建多元分布。然而，这些方法也有一些实际局限性。多变量泊松-对数正态回归（MPLR）和潜变量泊松-正态回归模型仅适用于过度离散数据。多变量泊松-伽马混合模型（MPGM）和多变量负二项模型（MNBM）仅适用于具有正相关的过度离散数据。另一个例子是 Copula 框架，它可以用于构建多元分布。但对于许多响应变量之间的负相关，很难进行建模，尤其是在计数数据中（Nikoloulopoulos 和 Karlis 2009）。 Inouye、Yang、Allen 和 Ravikumar（2017）提出了三种替代方法，仅用于处理等离散和过度离散数据，这些数据通过联结函数或分布混合构建。

此外，人们还开发了在实际应用中不受限制的分布。Famoye(2015) 提出了多元广义泊松回归模型，该模型能够处理各种类型的离散程度和相关性，并采用基于最大似然（ML）范式的方法进行估计。类似地，Mu˜noz-Pichardo, Pino-Mej´ıas, Garc´ıa-Heras, Ruiz-Mu˜noz, and Gonz´alez-Regalado (2021) 提出了多元条件泊松回归模型，其中响应变量之间的关系通过线性预测子中的一个系数来衡量。反过来，响应变量之间的依赖关系以其他响应变量为条件。

基于估计函数的灵活建模框架——多变量协方差广义线性模型（MCGLM）也能拟合此类数据（Bonat2016）。该方法仅采用二阶矩假设，并基于准最大似然法估计参数（Wedderburn1974）。它采用与广义估计方程（GEE）类似的方法处理相关数据（Liang and Zeger 1986），从而能够同时处理多变量响应和相关数据。我们还可以提及基于贝叶斯推断的方法，例如通过 `MCMCglmm` 软件包实现的 MCMC广义线性混合模型（Hadfield 2010），以及使用 Stan 的贝叶斯回归模型——`brms` 软件包（B ¨Urkner 2018），但代价是计算时间更长。本文将不讨论贝叶斯模型。

另一种方法是使用层次广义线性模型这一类模型，来建模同一个体内响应变量之间的相关性（Lee and Nelder1996）。这一类模型可以建模

相关变量或个体之间的相关性，可通过随机效应（一个未观测变量）来建模，该随机效应可以服从任意分布。当随机效应服从高斯分布时，我们得到广义线性混合模型（GLMM）。不过，GLMM 通常用于建模样本单元之间的相关性，而非响应变量之间的相关性。此类方法已在成熟的 R 软件包中实现，例如 R 核心团队 2020 所开发的软件包 glmmTMB （ Brooks, Kristensen, vanBenthem, Magnusson, Berg, Nielsen, Skaug, Maechler, andBolker2017）、lme4（Bates, M¨achler, Bolker, and Walker2015）和 nlme（Pinheiro, Bates, DebRoy,Sarkar,andR Core Team 2017）中。

本文提出在广义线性混合模型的框架下对多变量计数数据进行建模，以适应响应变量之间的相关性。参数估计值通过最大似然法获得（Aldrich 等 1997）。我们使用服从多变量正态分布的随机效应来适应响应变量之间的相关性。估计过程与广义线性混合模型中的过程类似，并通过 R （R 核心团队 2020）借助 TMB 软件包（Kristensen、Nielsen、Berg、Skaug 和 Bell 2016）实现。

本文将重点讨论多变量过度离散数据，而 da Silva、Laureano、Petterle、J´unior 和 Bonat （2022）提出了类似的方法，重点关注欠离散数据。这种针对多变量计数数据的建模方法针对响应变量的三种不同分布进行了评估：泊松分布、负二项分布和采用均值参数化的 COM-泊松分布（Huang 2017）。尽管 COM-泊松分布不属于指数族，但我们仍将其视为广义线性混合模型框架，因为无论所使用的分布如何，估计过程都保持不变。

本文共分为六节，包括本引言。第2节介绍用于展示模型应用的数据集。第3节提出多变量广义线性混合模型（MGLMM）。第4节展示用于评估估计量性质的模拟研究结果。第5节呈现将模型应用于第2节所述数据集的结果。最后，第6节讨论本文的主要贡献及未来工作。

2. 数据集

2.1. 数据集 I：澳大利亚健康调查

AHS 是澳大利亚规模最大的健康问题调查。本文使用的数据可通过`mcglm` 软件包（Bonat 2018）中的`ahs`对象获取。本研究的主要目的是探究，更多的医疗保健服务可及性以及性别、年龄和收入等人口统计学协变量，是否与患者使用医疗服务的次数有关。

该数据包含 5190 名受访者和 15 个变量，其中 10 个被视为协变量，5 个则作为横断面研究中的计数型响应变量，研究对象为 18 岁以上的个体。这五个响应变量分别为 Ndoc（与医生或专科医生的就诊次数）、Nndoc（与医疗专业人员的就诊次数）、Nadm（过去12个月内入住医院、精神病院、护理院或康复疗养院的次数）、Nhosp（最近一次住院期间的住院夜数）和 Nmed（过去两天使用的处方药和非处方药总数）。补充材料中的表 S1 列出了各协变量的说明。

图1 展示了各响应变量的条形图。图中显示，所有变量都存在一定程度的过度离散，这可能是由于分布右侧重尾且可能存在一些零值过多所致。需要指出的是，这些解释仅适用于边际层面，因为它们未考虑协变量的影响。表1 给出了5个响应变量的一些描述性统计。可以看出，这5个响应变量之间存在正相关（多数情况下相关性较弱）；鉴于它们是在同一个体上测量的，这种相关性可以接受。Nadm 与 Nhosp 之间除外，其相关性为0.996。此外，均值与方差的关系

显示这些变量均存在过度离散，其中 Nhosp 的过度离散程度最高。其特征是 Fisher 离散指数（离散指数 DI）大于1（Fisher1934）。此外，广义离散指数（GDI）（Kokonendji 和 Puig 2018）也将该数据集判定为过度离散，因为其值大于1，且其95%置信区间不包含零。不过，这些结果仍需通过模型拟合加以确认。

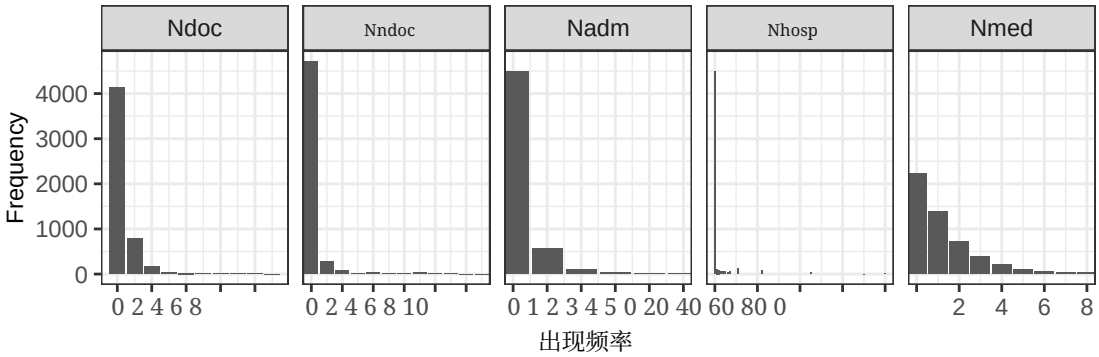


图1：澳大利亚健康调查（AHS）数据中各响应变量计数的条形图。Ndoc（与医生或专科医生的就诊次数）、Nndoc（与医疗专业人员的就诊次数）、Nadm（过去12个月内在医院、精神病院、护理院或康复疗养院的住院次数）、Nhosp（最近一次住院期间的住院夜数）和 Nmed（过去两天使用的处方药和非处方药总数）。

表1：澳大利亚健康调查（AHS）响应变量的 Spearman 相关性、均值、方差和离散指数（DI）。广义离散指数（GDI）及其标准误（SE）为 17.944（1.99）。Ndoc（与医生或专科医生的就诊次数）、Nndoc（与医疗专业人员的就诊次数）、Nadm（过去12个月内在医院、精神病院、护理院或康复疗养院的住院次数）、Nhosp（最近一次住院期间的住院夜数）和 Nmed（过去两天使用的处方药和非处方药总数）。

	斯皮尔曼相关 ρ				Mean	方差	DI
	Nndoc	Nadm	Nhosp	Nmed			
Ndoc	0.11	0.158	0.160	0.290	0.302	0.637	2.111
Nndoc		0.107	0.113	0.165	0.215	0.932	4.341
Nadm			0.996	0.135	0.174	0.258	1.484
Nhosp				0.139	1.334	37.455	28.083
Nmed					1.218	2.423	1.989

2.2. 数据集 II：蚂蚁物种丰度

这些数据来自一项在澳大利亚东南部开展的研究，研究者包括 Gibb、Stok- losa、Warton、Brown、Andrew 和 Cunningham (2015)，这些数据可通过 R 核心团队 2020 提供的 R 软件中的 `mvabund`软件包（Wang、Naumann、Eddelbuettel、Wilshire 和 Warton 2020）获取。这项研究在澳大利亚东南部的 30 个地点开展，时间为 2007 年 11 月至 2008 年 4 月；研究人员统计了为期 18 天的调查期间落入陷阱诱捕器的 41 种不同蚂蚁物种的数量。该数据集主要用于考察环境变量与不同蚂蚁物种出现频率及蚂蚁物种共同出现之间的关系。所考虑的响应变量是 41 个物种各自在 30 个地点的出现情况，而协变量包括

5 个环境变量，其完整说明见补充材料表 S2。响应变量的名称以索引编号（1,...,41）开头，后接各物种的缩写名称。

图2 展示了 ANT数据中各响应变量的条形图。这41种不同的物种表现出很大的变异性。有些物种仅在一个地点被观察到一次，例如 多刺蚁属 和 火蚁属，而 伊瑞多蚁属 和 切叶蚁属 则在多个地点被观察到20次。

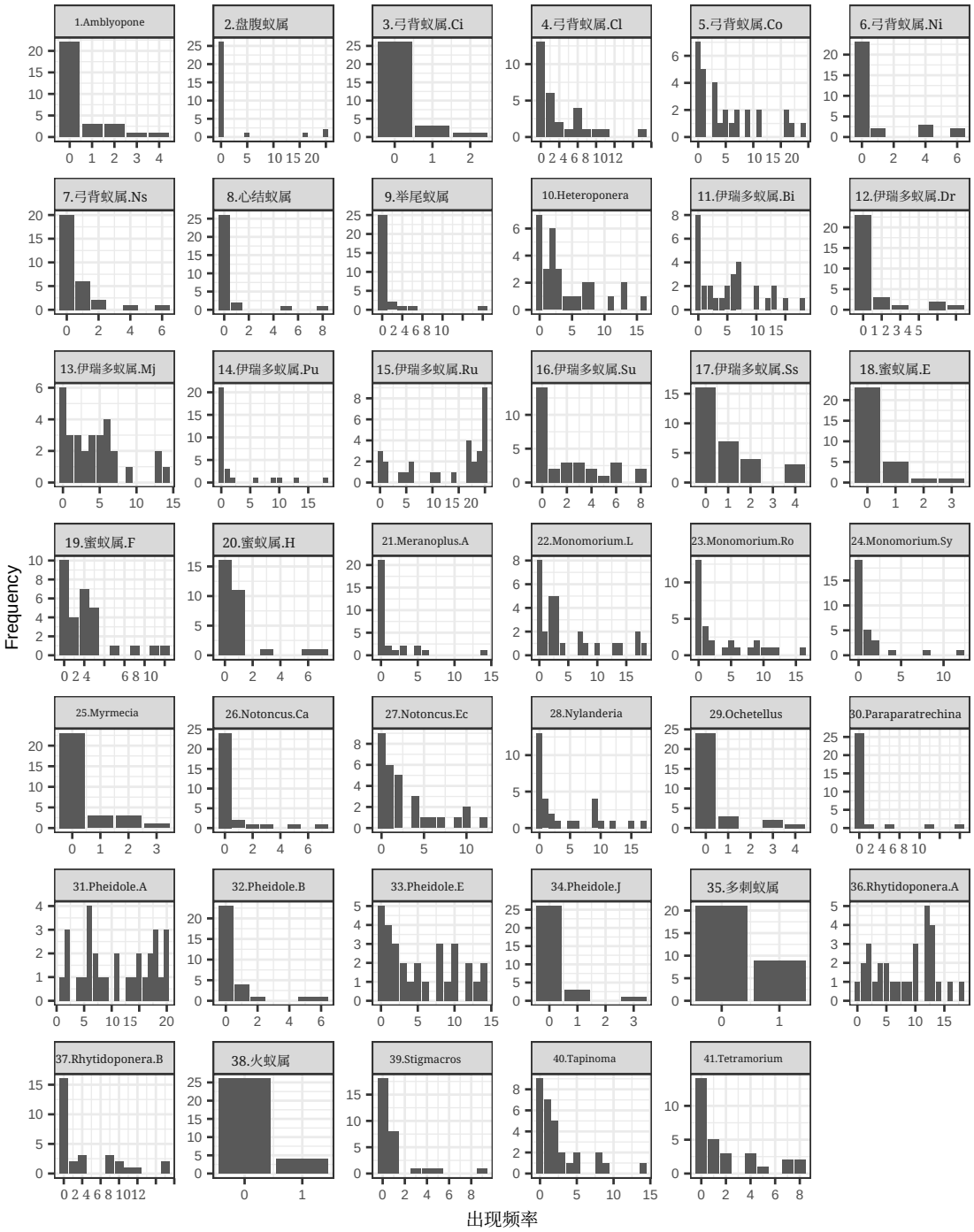


图2：ANT数据中各蚂蚁属出现频率的条形图。

表2 列出了每个响应变量的均值、方差和离散指数，以及该数据集的广义离散指数。除多刺蚁属和火蚁属外，几乎所有变量的方差都大于均值；这一点可由离散指数>1来描述。根据广义离散指数 $= 11$，543以及95%置信区间的下限，该数据集被归类为过度离散，且下限

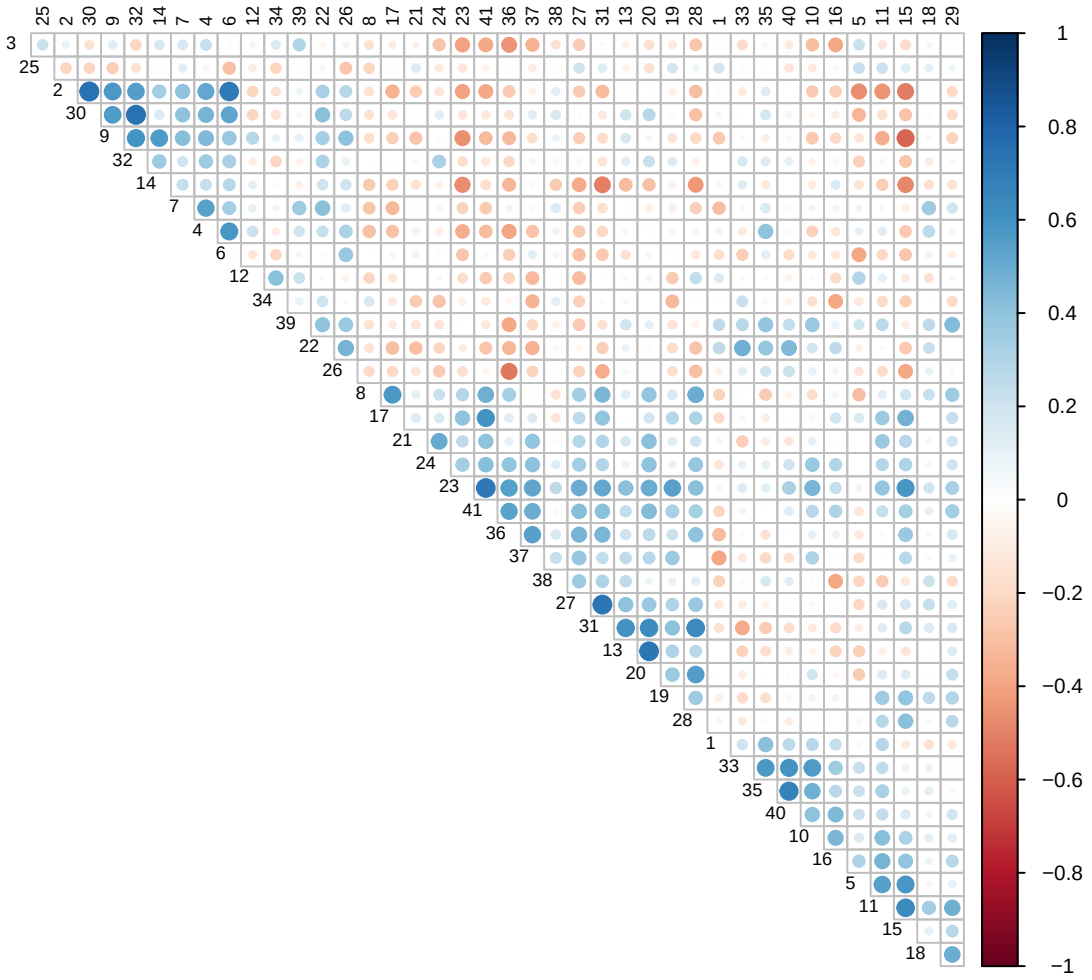
大于1。

表2：ANT数据中各属的均值、方差和离散指数（DI）。广义离散指数（GDI）及其标准误（SE）为 11.54（.92）

响应	Mean	方差	DI	响应	Mean	方差	DI
1.Amblyopone	0.533	1.085	2.034	21.Meranoplus.A	1.333	8.713	6.534
2.盘腹蚁属	2.033	32.999	16.229	22.Monomorium.L	4.733	32.409	6.847
3.弓背蚁属.Ci	0.167	0.213	1.276	23.Monomorium.Ro	3.433	20.944	6.100
4.弓背蚁属.Cl	1.933	7.099	3.672	24.Monomorium.Sy	1.167	6.902	5.916
5.弓背蚁属.Co	5.300	33.252	6.274	25.Myrmecia	0.400	0.662	1.655
6.弓背蚁属.Ni	0.867	3.430	3.958	26.Notoncus.Ca	0.633	2.654	4.191
7.弓背蚁属.Ns	0.667	1.816	2.724	27.Notoncus.Ec	2.900	12.300	4.241
8.心结蚁属	0.500	2.879	5.759	28.Nylanderia	3.733	25.720	6.889
9.举尾蚁属	0.567	3.633	6.412	29.Ochetellus	0.433	1.082	2.496
10.Heteroponera	4.067	19.857	4.883	30.Paraparatrechina	0.700	4.976	7.108
11.伊瑞多蚁属.Bi	5.333	26.437	4.957	31.Pheidole.A	11.100	40.300	3.631
12.伊瑞多蚁属.Dr	0.600	1.834	3.057	32.Pheidole.B	0.567	2.047	3.613
13.伊瑞多蚁属.Mj	4.300	15.803	3.675	33.Pheidole.E	5.467	22.809	4.172
14.伊瑞多蚁属.Pu	2.033	20.447	10.056	34.Pheidole.J	0.200	0.372	1.862
15.伊瑞多蚁属.Ru	13.300	59.045	4.439	35.多刺蚁属	0.300	0.217	0.724
16.伊瑞多蚁属.Su	2.133	6.809	3.192	36.Rhytidoponera.A	8.300	25.528	3.076
17.伊瑞多蚁属.Ss	0.900	1.610	1.789	37.Rhytidoponera.B	2.400	11.834	4.931
18.蜜蚁属.E	0.333	0.506	1.517	38.火蚁属	0.133	0.120	0.897
19.蜜蚁属.F	2.133	6.740	3.159	39.Stigmacros	0.967	3.826	3.958
20.蜜蚁属.H	0.900	2.783	3.092	40.Tapinoma	2.533	11.154	4.403
				41.Tetramorium	1.933	7.030	3.636

图3 通过相关图中的斯皮尔曼相关，探究不同物种的出现频率之间的关系。边际相关范围从 -.58 到 .74，均匀分布于相关参数的所有可能取值之间 $\rho = (-1,1)$.

图3：使用斯皮尔曼相关分析 ANT 物种出现频率的相关图



3. 计数数据的多元广义线性混合模型（MGLMM）

本节沿用 da Silva 等人(2022) 中相同的描述，并较为简洁地完整介绍 MGLMM 模型。尽管解释内容相同，但本文将探讨过度离散计数模型。总之，我们将标准广义线性混合模型（Breslow and Clayton 1993）扩展至多个响应变量。

令 Y_{ij} 为 $i = 1, \ldots, n$ 个独立个体和 $r = 1, \ldots, k$ 个响应变量对应的随机变量。此外，设 x^{irj} 表示每个个体 i 和响应变量 r 的第 j 个协变量的取值，其中共有 k 个已知协变量。首先，我们指定一个仅含随机截距的标准广义线性混合模型联合模型：

$$\mathrm{Y}_{ir} \mid \mathrm{b}_{ir} \sim f(\mu_{ir}; \phi_{ir}),$$

其中，给定随机效应时，Y_{ir} 的条件分布服从同一分布 f，但允许均值和离散度不同。接下来，线性预测子：

$$\mathrm{g}_r(\mu_{ir}) = x_{irj}^{\top}\boldsymbol{\beta}_r + \mathrm{b}_{ir},$$

其中 $\mathrm{g}_r(.)$ 是适当的连接函数，$\boldsymbol{\beta}_r$ 是参数估计值向量，b_{ir} 是受试者 i 和响应变量 r的随机截距值。最后，随机效应分布指定如下：

$$\begin{pmatrix} \mathrm{b}_{i1} \\ \mathrm{b}_{i2} \\ \vdots \\ \mathrm{b}_{ir} \end{pmatrix} \sim \mathrm{NM}\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \underset{r\times r}{\boldsymbol{\Sigma}} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1r}\sigma_1\sigma_r \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \dots & \rho_{2r}\sigma_2\sigma_r \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{r1}\sigma_r\sigma_1 & \rho_{r2}\sigma_r\sigma_2 & \dots & \sigma_r^2 \end{bmatrix} \right),$$

其中，每个随机效应的均值为 0，方差为 σ^2，且任意一对随机效应之间的相关性为 $\rho_{rr'}$($r \neq r'$)。该框架具有通用性，可应用于任意分布f 和连接函数 $g_r(.)$。不过，由于我们处理的是计数数据，因此仅考虑泊松分布、II型负二项分布和 COM-Poisson分布，以及对数链接函数。也可以将该模型扩展为针对每个响应变量 (f_r) 使用不同的分布，但本文不讨论这一点。

最大似然是所采用的估计程序，其详细描述见da Silva 等人(2022)。我们通过数值积分结合拉普拉斯近似，对随机效应进行了积分消去（Tierney和Kadane 1986）。我们高效地实现了牛顿-拉夫森法，以执行内部优化。我们使用一阶导数方法优化边际似然，例如 `BFGS` 和 `OPTIM` 例程，这些例程可在 `R` 软件中使用。联合似然和边际似然的导数通过自动微分获得（Baydin、Pearlmutter、Radul和Siskind 2017）。这一过程通过 `TMB` 软件包高效实现，该软件包提供 `C++` 模板，用户须提供目标函数，在本例中即对数似然函数。

该模型中随机变量的变异性由两个参数衡量：随机效应的方差和概率质量函数（pmf）的离散度。该模型的一个关键特点是具有很强的灵活性，能够估计这两种方差，且不存在可辨识性问题。

用于生成本文结果的代码可在 https://github.com/guilhermeparreira/papers/tree/master/AJS_多变量广义线性混合模型_过度离散_计数以及 http://www.leg.ufpr.br/doku.php/publications:papercompanions。

4. 模拟研究

本节中，我们通过模拟研究评估最大似然估计量的性质（偏差和一致性）。我们考虑了一个针对计数数据的二元回归模型。我们设计了 12 种模拟情景，包含四种不同的样本量：100、250、500 和 1000，以及三种随机效应之间不同的相关性，$\rho = -0.5$，0，0。5。对于回归结构，我们仅为每个响应考虑了一个截距，其中 β_{01} =log(7) 和 β_{02} =log(1。5)。随机效应的方差为 $\sigma_1^2 = .3$ 和 $\sigma_2^2 = .15$。负二项分布和 COM-泊松分布的离散参数分别等于 $\phi = 1$ 和 $\nu = .7$，这会产生轻微过度离散。

对于每种设计，我们分别为泊松分布、COM-泊松分布和负二项分布生成了 150、200 和 300 个数据集。我们的初衷是为每种分布生成 100 个数据集。然而，由于并非每次重复都能得到参数估计值的标准误（SE），因此有必要根据每种分布中标准误计算失败的次数，按比例增加生成的数据集数量，以确保至少获得 100 个有效估计。以下三个小节分别介绍泊松情景、负二项分布情景和 COM-泊松情景的结果。

4.1. 泊松情景

图4 展示了双变量泊松回归模型在不同样本量和模拟情景下，基于平均标准误计算的平均偏差及其置信区间。结果表明，在所有模拟情景中，随着样本量增加，平均偏差和标准误均趋近于 0。这说明最大似然估计量在大样本下具有一致性和无偏性。

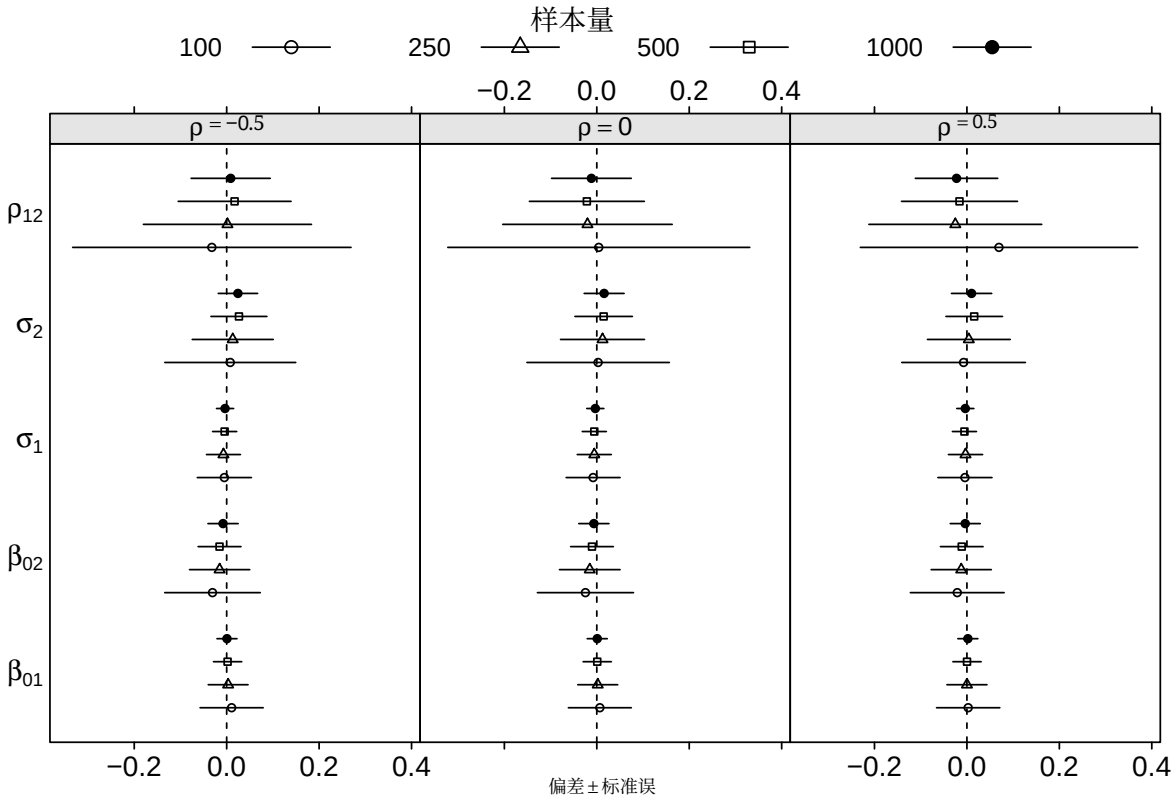


图4：基于双变量泊松回归模型的四种不同样本量和三个相关系数，平均偏差及其95%置信区间（根据平均标准误（标准误）计算）。各参数的真实值分别为 β_{01} =log(7)、 β_{02} = log(1.5)、$\sigma_1^2 = .3$ 和 $\sigma_2^2 = .15$。

图 5 展示了双变量泊松回归模型在不同样本量和模拟情景下各参数的覆盖率。总体而言，所有经验覆盖率均接近 95% 的名义水平，大约介于 90% 和 98% 之间。具体而言，回归参数 $\boldsymbol{\beta}_{0r}$ 的覆盖率略高于名义水平。对于随机效应方差 $\boldsymbol{\sigma}_r$，覆盖率略低于名义水平。最后，对于随机效应之间的相关性 ρ，当$\rho = .5$时，覆盖率略低；当 $\rho = \{0, -0.5\}$ 时，与 95% 的名义水平相比，覆盖率略高。

即使对于泊松分布这种最简单的估计情形（因为不存在离散参数），1800 次（$3\times4\times150$）模拟中仍有 13 次未能返回某些参数估计值的标准误，或因标准误过大而产生极端值；这些结果未纳入分析。我们使用 `PORT`算法估计模型，因为在大多数情况下，它比 `BFGS` 更稳定、速度更快。Kristensen 等人也（2016）也有类似发现：他们考察了 9 个不同模型设定下的研究案例，涵盖从线性回归到多元随机波动率模型；结果表明，`PORT`的表现优于 `BFGS`。

4.2. 负二项情景

图 6 表明，在所有模拟情景中，随着样本量增加，平均偏差和标准误均趋近于 0。这说明最大似然估计量在大样本下具有一致性和无偏性。随机效应的方差 σ_1 和 σ_2 略有低估。

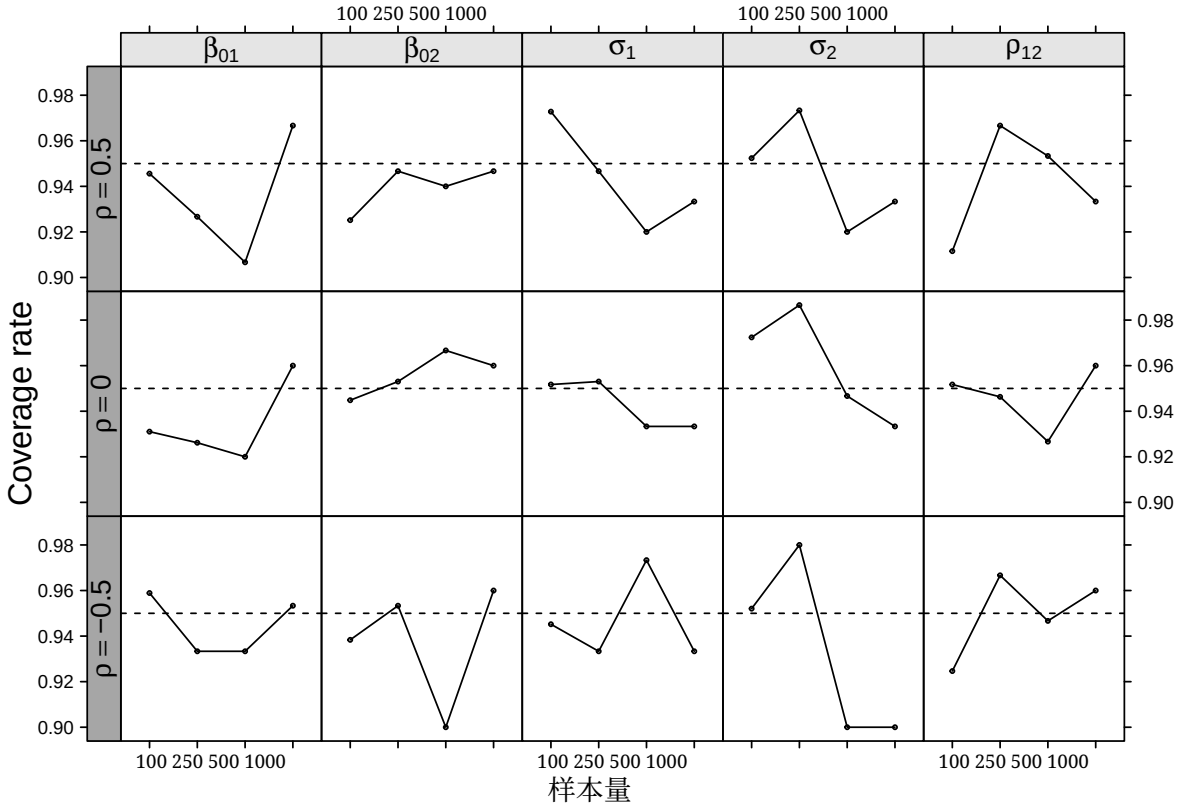


图 5：泊松回归模型中各参数（按列）的覆盖率随样本量和相关系数变化的情况。各参数的真实值为 β_{01} =log(7)、 $\beta_{02} = \log(1.5)$、 $\sigma_1^2 = .3$ 和 $\sigma_2^2 = .15$。

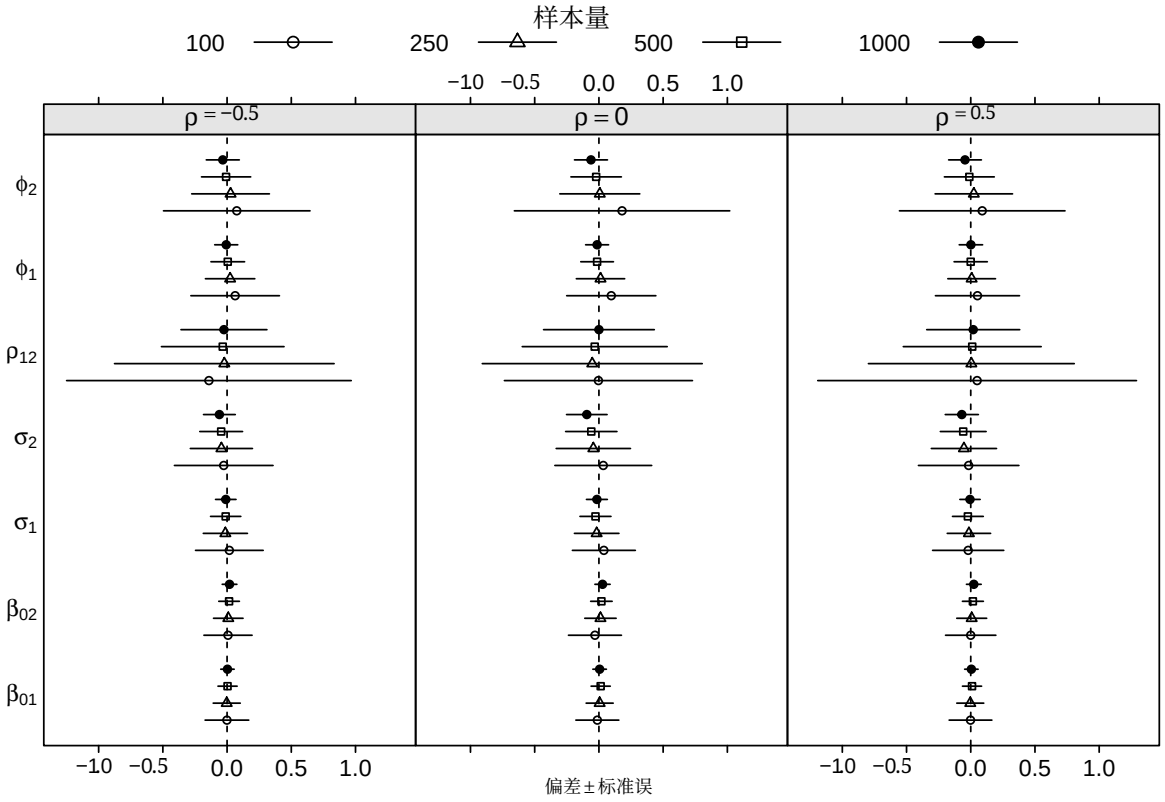


图 6：双变量负二项（NB）回归模型在四种不同样本量和三个相关系数下的平均偏差及其基于平均标准误（SE）的 95%置信区间。各参数的真实值分别为 $\beta_{01} = \log(7)$、$\beta_{02} = \log(1.5)$、$\phi = 1$、$\sigma_1^2 = .3$ 和 $\sigma_2^2 = .15$。

总体而言，图7 显示，经验覆盖率接近 95% 的名义水平。具体来说，回归参数 $\boldsymbol{\beta}_{0r}$、随机效应方差 $\boldsymbol{\sigma}_r$ 和随机效应之间的相关性 ρ 的覆盖率略高于名义水平。相反，当样本量等于 100 时，ρ 的估计覆盖率接近 80%，而 $\rho = 0$。

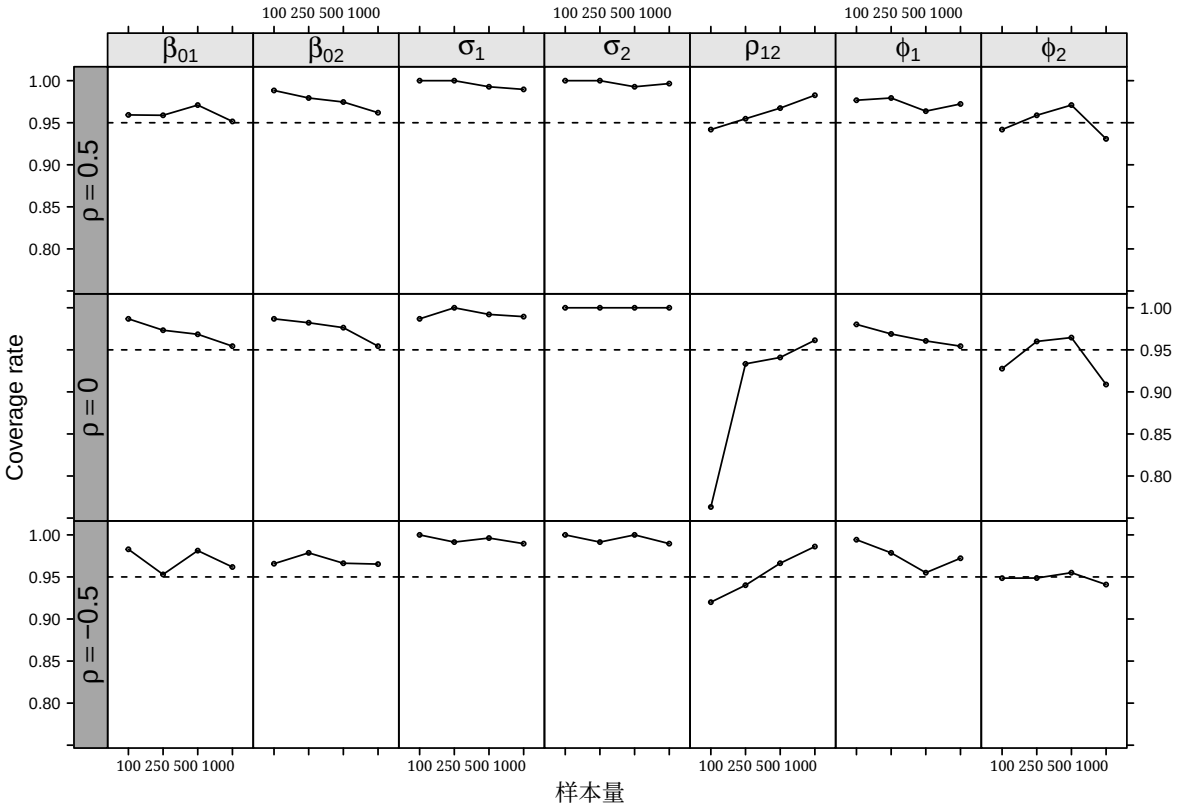


图7：双变量负二项（NB）回归模型中，各参数（按列排列）的覆盖率随样本量和相关系数的变化。各参数的真实值分别为 $\beta_{01} = \log(7)$、 $\beta_{02} = \log(1.5)$、 $\phi = 1$、 $\sigma_1^2 = .3$ 和 $\sigma_2^2 = .15$。

与泊松回归模型相比，双变量 NB 回归模型的估计问题更为严重，在 3600 次模拟中有 740 次（$3 \times 4 \times 300$）（20,6%）出现了估计问题。此外，还需要从结果中剔除离散参数 ϕ 大于 5 的迭代（我们的模拟值为 1）。

4.3. COM-泊松情景

图 8 显示，在所有模拟情景中，随着样本量增加，标准误趋于 0，但偏差并非如此。这表明最大似然估计量不具有一致性。此外，几乎所有估计值都有偏差。尽管 ν_2 被高估， ν_1 则被低估（且偏差不会随样本量增加而减小）。相关参数 ρ表现出典型的特征：当数据是在 $\rho = 0$ 条件下模拟时，几乎没有偏差；在$\rho = .5$ 条件下，偏差为负；在 $\rho = -.5$ 条件下，偏差为正；模型会使相关参数趋向于零。随机效应的标准差 σ_2 被高估，而 σ_1 则略微被低估。对于回归参数， β_{02} 略微被低估，而 β_{01} 则显示偏差可以忽略不计。

参数之间一种可能存在的有趣关系是，当 ν 被高估时（使模型的离散度降低： $\nu > 1$），σ 也被高估（使模型的方差增加）。相反，当 ν 被低估时（使模型的离散度进一步增加: $\nu < 1$）， σ 也被低估（使模型的方差降低）。似乎当 ν 使模型的离散度增加时， σ 会使方差降低，反之亦然。这可能与这些参数之间存在非正交性有关。

总的来说，图 9 显示经验覆盖率与名义水平 f

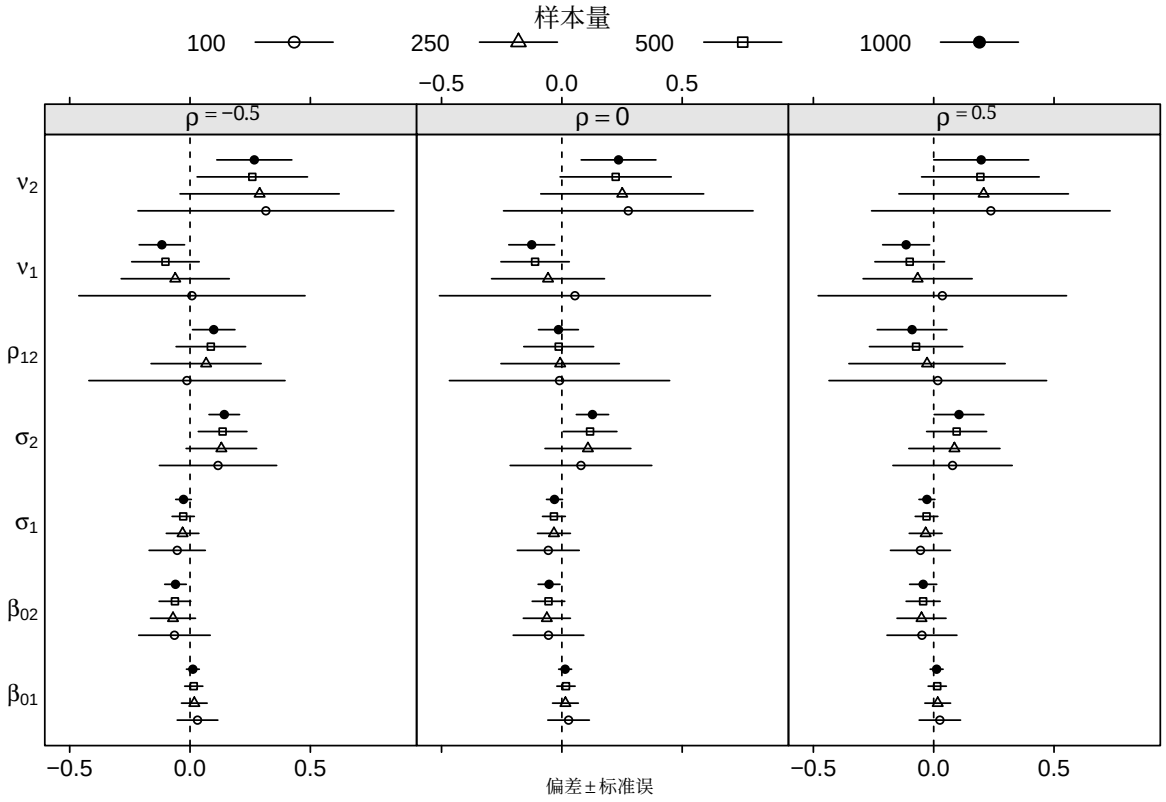


图 8：双变量 COM-Poisson 回归模型在四种不同样本量和三种相关系数下的平均偏差及其基于平均标准误（SE）的 95%置信区间。各参数的真实值分别为 $\beta_{01} = \log(7)$、$\beta_{02} = \log(1.5)$、$\nu = .7$、$\sigma_1^2 = .3$ 和 $\sigma_2^2 = .15$。

95% 不接近，这与 图 8中呈现的结果一致，即使样本量更大，偏差也没有减小。具体而言，σ_2 的覆盖率在所有参数中最差（覆盖率随样本量增加而降低），其次是 ν_2。这并不令人意外，因为这些参数在图 8中的偏差最大。相比之下，β_{01}、σ_1 和 ρ（当数据由 $\rho = 0$生成时）的覆盖率接近 95% 水平。

与泊松回归模型相比，COM-泊松分布回归模型的估计问题更严重；与负二项分布模型相比则较轻，共有 245 个（10.2%）模型属于此类，占 2400 个（$3 \times 4 \times 200$）中的一部分。除采用泊松分布模型的规则对极端值进行分类外，还必须剔除离散参数 ν 大于 4 且 ρ 估计值的标准误大于 2 的模型。

5. 数据分析

本节介绍第 2 节中所呈现的两个数据集的数据分析。我们通过以下方式谨慎地选取了第 3 节中模型的初始值。首先，对于每个数据集，我们都使用`MCGLM` 软件包（Bonat2016）通过准似然拟合泊松分布模型，以获得回归参数和方差参数的初始参数估计值（基于残差的方差）。由于方法不同，我们将相关参数设为 0。这些是泊松分布模型的初始值；泊松分布模型的最终估计值成为负二项分布模型的初始值，而负二项分布模型的初始估计值则成为 COM-泊松分布模型的初始值。负二项分布模型的离散参数初始值设为 $\phi = 1$，即轻微

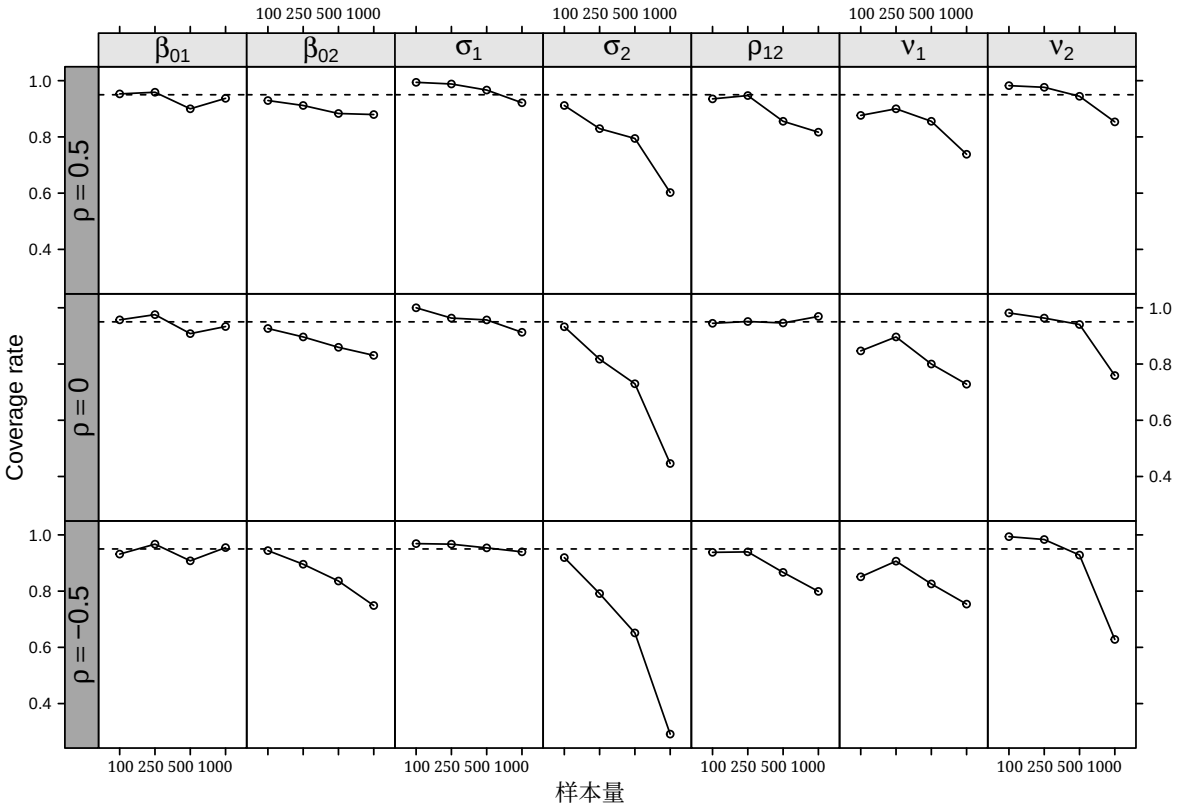


图 9：双变量 COM-Poisson 回归模型中各参数（按列排列）在不同样本量和相关系数下的覆盖率。各参数的真实值分别为 $\beta_{01} = \log(7)$、$\beta_{02} = \log(1.5)$、$\nu = .7$、$\sigma_1^2 = .3$ 和 $\sigma_2^2 = .15$。

过度离散；而 COM-Poisson 模型的离散参数初始值设为 $\nu = 1$，即等离散。

估计过程首先从 AHS数据中抽取一个容量为 300 的简单随机样本（SRS，随机种子为 2390）。利用该样本得到的参数作为初始值，将模型拟合到完整数据集。对于 ANT数据，由于数据集规模较小（30），我们没有采用这一步。随后，我们对同一模型进行多次优化，并依次使用不同的优化器。第一轮使用 PORT程序，其最终估计值作为初始值，用 BFGS 优化同一模型；接着将 BFGS 的最终估计值作为初始值，再使用 PORT 优化；最后又用 PORT 进行一轮优化。之所以采用这一策略，是因为这些模型较为复杂，难以估计，而这一过程得到了合理的结果。

除了第 3 节中介绍的完整模型外，我们还尝试估计四种更简单的模型，以检验模型是否过度指定。我们通过固定 $\boldsymbol{\Sigma}$ 或离散参数的取值来简化模型。第一种简化旨在使响应变量相互独立，即将相关参数设为 0。第二种简化中，我们固定$\phi = 1$ 和 $\nu = 1.5$，分别使负二项分布模型和 COM-Poisson 模型呈现轻微过度离散和欠离散。在第三种简化模型中，我们将方差设为 1。最后一种策略仅使用一个方差参数来估计所有响应变量。最后三种简化模型旨在回答能否在同一模型中同时估计离散度参数和方差参数。我们只对负二项分布模型和 COM-Poisson 模型进行了检验，因为它们包含离散度参数。

报告的结果仅基于完整数据，并采用了各分布的最佳模型版本；这些模型均已剔除无法估计标准误的协变量。随后，我们给出了最大化的对数似然值、赤池信息准则和贝叶斯信息准则（AIC 和 BIC），

以及估计参数的数量（np）。

5.1. AHS数据

表 3 展示了不同分布和设定下 AHS数据的拟合优度指标。

表 3：不同分布和设定下 AHS数据的拟合优度指标。估计参数数量（np）、赤池信息准则和贝叶斯信息准则（AIC 和 BIC），以及对数似然值（Loglik）。

模型	np	AIC	BIC	对数似然
泊松分布	70	33814	34272	-16837
NB	75	33826	34318	-16838
COM-泊松分布	75	33106	33598	-16478

在所有分布及其各自的简化版本中，COM-泊松完整模型表现最佳。

图10 展示了按结果和最终模型划分的回归参数估计值及95%置信区间。可以看出，所有模型的置信区间幅度几乎相同，但对于Ndoc响应变量的大多数估计值，COM-泊松模型的区间略小于其他模型；对于其他响应变量，差异很小。泊松模型与负二项分布模型的点估计值非常接近，而COM-泊松模型在一些估计值上有所不同，例如Ndoc的freepoor、年龄、性别、疾病和有限慢性病。这可能与图8中发现的偏差有关。图 8。

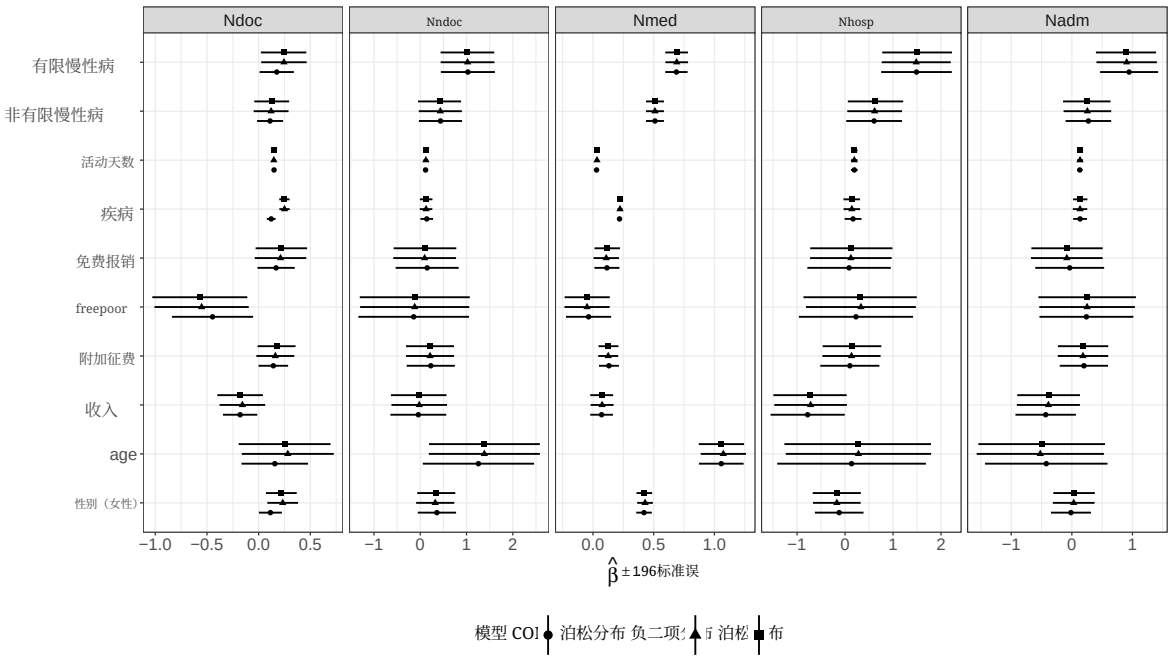


图10：澳大利亚健康调查（AHS）数据中按结果和最终模型划分的回归参数估计值及95%置信区间。

表4 展示了各模型和结果的离散度估计值。尽管根据表1中给出的广义离散指数（GDI） $= 17.94$ ，我们可以认为这些数据略微过度离散，但我们看到 ϕ 趋于无穷大（表明负二项分布（NB）模型为等离散模型）

分布），而且 ν 对所有响应变量均大于1，表明存在欠离散。预期结果应为 ϕ 和 ν 趋近于零，从而导致两种分布都出现过度离散。然而，如第5.2小节所述，这种现象可能是由随机效应的方差引起的。对于COM-泊松分布模型，唯一表明存在过度离散的 ν 估计值对应Nmed，其样本离散指数为1.99（轻微过度离散）。

表4：澳大利亚健康调查（AHS）数据中各模型和结果的参数估计值及标准误（SE）的离散度

	负二项分布(ϕ)		COM-泊松分布(ν)	
结果	估计量	SE	估计量	SE
Ndoc	1.7e+04	1.8e+05	9.160	0.432
Nndoc	5.9e+03	6.0e+04	2.837	0.336
Nmed	3.6e+03	2.4e+04	0.674	0.032
Nhosp	8.5e+03	1.2e+05	5.110	0.615
Nadm	8.5e+03	8.1e+04	6.665	0.431

我们在（1）中给出了 COM-Poisson模型的相关系数估计值，并在括号内列出其标准误。在返回的 10 个相关系数中，COM-Poisson模型有 6 个相关系数显著。矩阵中的星号表示系数在 5% 水平上显著。尽管我们并不特别关注对每个相关系数估计值进行解读，但了解其中哪些显著仍然很重要，因为显著性可能与较小的标准误有关。

我们可以看到 Nadm 与 Nhosp 之间几乎完美且显著的相关性，以及 Nmed 与 Ndoc 之间的强相关性。我们在 表1 中也观察到了这一模式。

$$\text{COM-Poisson}_{\rho} = \begin{array}{c} \begin{matrix} \text{Ndoc} & \text{Nndoc} & \text{Nmed} & \text{Nhosp} & \text{Nadm} \end{matrix} \\ \begin{bmatrix} 1 & -0.07(0.03)^{*} & 0.78(0.09)^{*} & -0.02(0.02) & -0.02(0.02) \\ & 1 & 0.34(0.14)^{*} & -0.04(0.03) & -0.05(0.03) \\ & & 1 & 0.25(0.09)^{*} & 0.25(0.09)^{*} \\ & & & 1 & 0。99(<.01)^{*} \\ & & & & 1 \end{bmatrix} \end{array} \begin{matrix} \text{Ndoc} \\ \text{Nndoc} \\ \text{Nmed} \\ \text{Nhosp} \\ \text{Nadm} \end{matrix} \quad (1)$$

5.2. ANT数据

我们在表 5中列出了每种分布的最佳模型。由于按照采用的准则，这两个模型在中间模型中等效，因此我们在此纳入了两个 COM-Poisson分布模型。

表 5：ANT数据中各分布最佳参数化模型的拟合优度指标。估计参数的数量（np）、赤池信息准则和贝叶斯信息准则（AIC 和 BIC），以及对数似然值（Loglik）。

模型	np	AIC	BIC	对数似然
泊松分布	1057	5085.08	6566.15	-1485.54
NB	1098	5166.98	6705.50	-1485.49
COM-泊松分布	1098	4965.87	6504.38	-1384.93
固定离散度康普松分布	1057	5016.70	6497.77	-1451.35

从表 5 可以看出，最佳模型来自 COM-Poisson分布；泊松模型和负二项分布模型的对数似然结果相近。COM-Poisson模型的对数似然和贝叶斯信息准则更高，赤池信息准则则低于固定离散度模型。对这两个 COM-Poisson模型进行对数似然比检验（似然比检验）得到 $p < .00001$ $(\chi_{41} = 132.84)$，这为完全指定模型提供了支持。因此，我们将列出每种分布的完整模型估计值。

图11 展示了前12个响应变量按结果和最终模型划分的回归参数估计值及其95%置信区间。我们在补充材料的图S1、S2和S3中展示了其余响应变量的相同图形，因为我们发现所有响应变量都呈现出相似的模式。首先可以看到，并非所有响应变量都具有相同的线性预测子（第二、第三和第九个响应变量中未呈现协变量“野生哺乳动物粪便”）。其次，有些回归估计值的标准误仍未由模型返回，这些估计值以空心圆表示。需要区分模型返回了标准误的情况和标准误极小的情况（以实心圆表示，例如COM-泊松分布中响应变量1–8和10–12对应的裸地）。

总体而言，COM-Poisson模型的置信区间比其他模型更窄，而所有模型的点估计值都较为接近。与其他协变量相比，野生哺乳动物粪便协变量的置信区间最大。仍有一些回归系数非常接近零，这表明采用更好的变量选择方法或许能改善模型拟合。

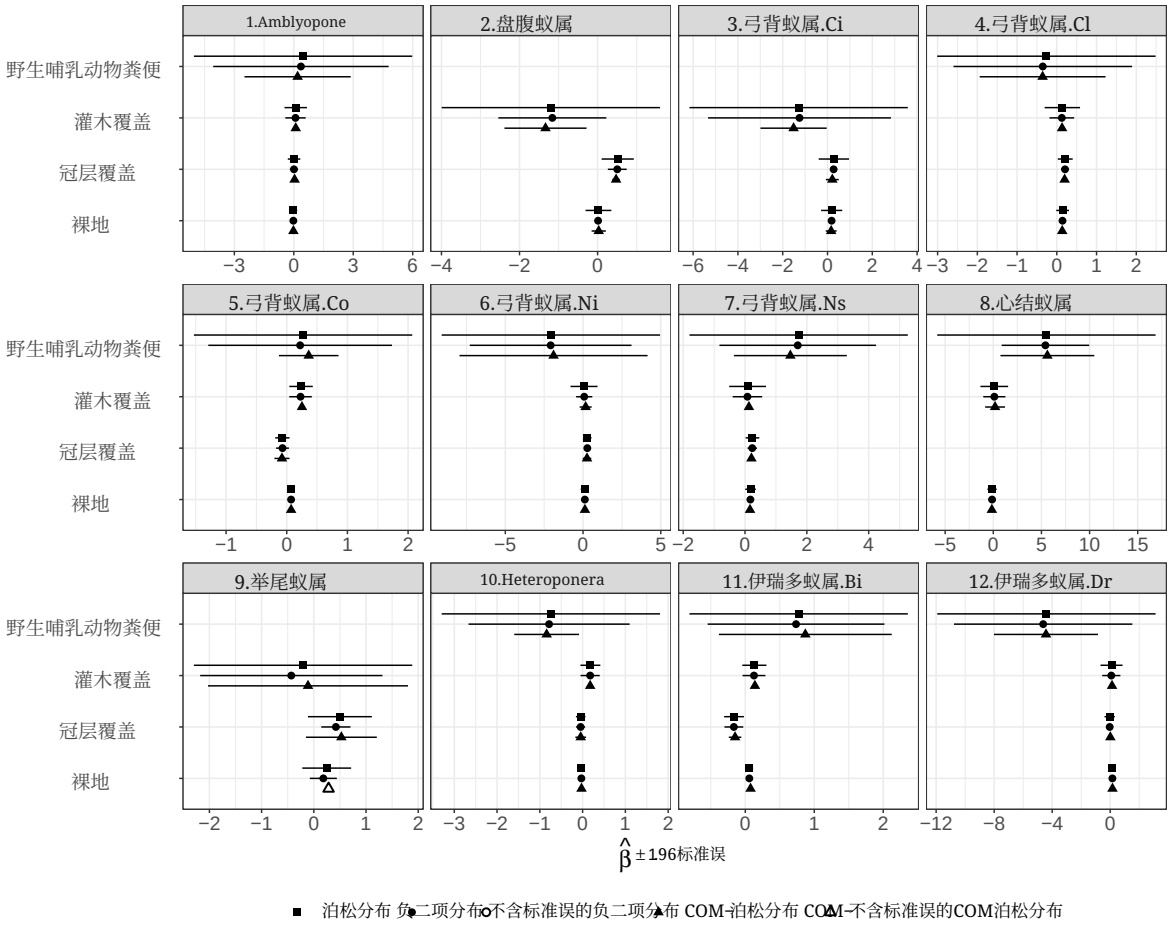
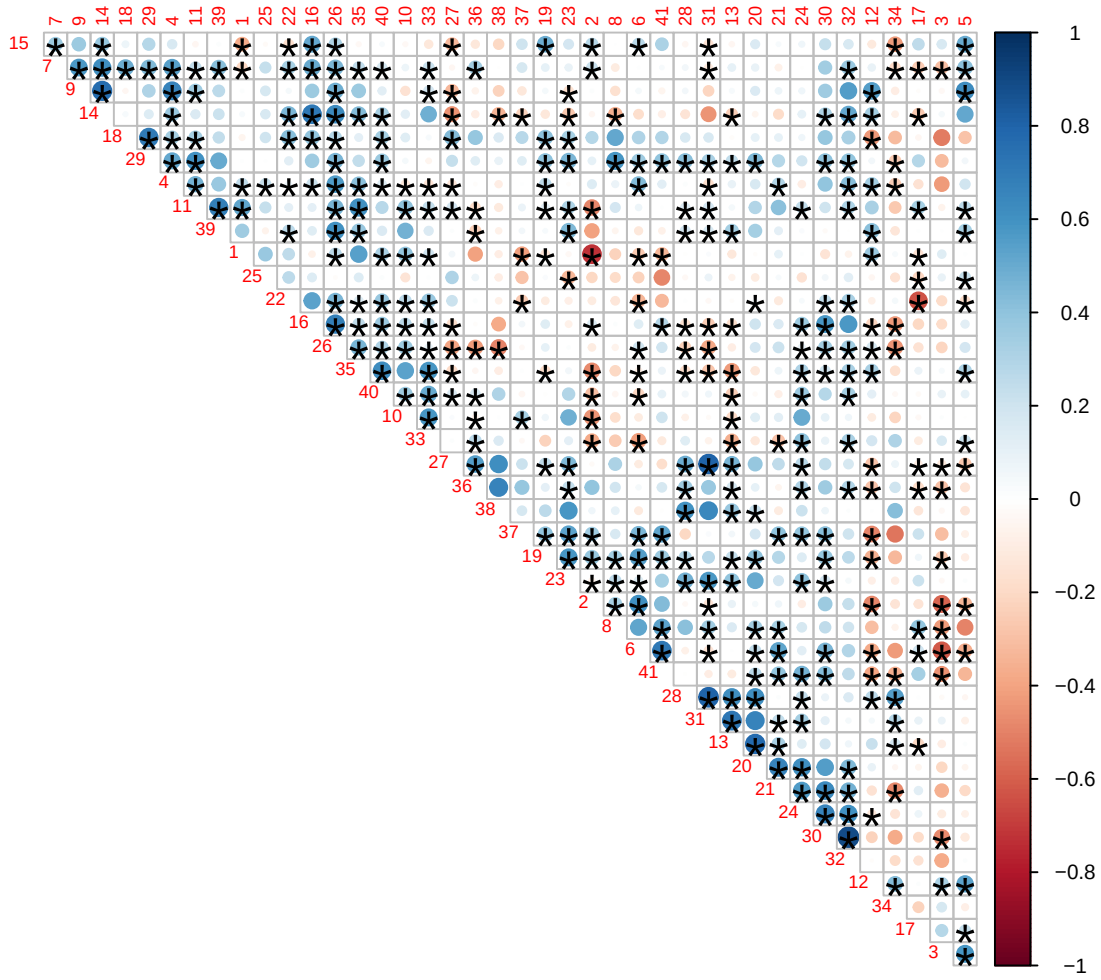


图11：ANT数据中各结果和最终模型的回归参数估计值及95%置信区间。

相关　COM-Poisson模型的相关系数见图12。在这些　e

返回的820个相关系数中，COM-Poisson模型有333个显著相关系数，负二项分布有278个，泊松分布有83个。这表明，COM-Poisson模型的相关系数比其他模型的相应系数标准误更小。图中的星号表示在5%水平上显著的系数。图中呈现的相关模式与边际相关中发现的模式有所不同。这表明，在计算相关系数时，考虑线性预测子的影响十分重要。例如，响应变量1与2、响应变量2与3之间的样本相关几乎为零；而这两对响应变量的随机效应之间，COM-Poisson模型的相关性为负，且在5%水平上显著；负二项分布模型中只有一对显著，泊松分布模型中则都不显著，这可能是由于标准误较大。当$\hat{\rho}$在Y1与Y2的随机效应之间等于-.7、在Y2与Y3的随机效应之间等于-.54时，

图12：基于 COM-Poisson模型的蚂蚁物种出现频率相关图。星号表示在 5% 显著性水平上相关性显著



6. 讨论

本文旨在提出一种适用于计数数据的多变量广义线性混合模型。该模型的一大优势是能够处理多个结局。我们基于带随机截距的广义线性混合模型构建框架，并建立了一个联合模型，其中随机效应服从多元正态分布。因此，我们可以测量由多变量响应变量集所引出的随机效应的方差和相关性。计数变量所使用的分布包括泊松分布、负二项分布和 COM-泊松分布。估计

过程与广义线性混合模型相同。我们使用 TMB包实现该模型，因为它提供了最先进的 C++ 库，可处理自动微分、CppAD (BellBM2005）、线性代数、Eigen C++(Guennebaud, Jacob 等 2010）以及并行计算，BLAS(Blackford, Petitet, Pozo, Remington, Whaley, Demmel, Dongarra, Duff, Hammarling, Henry 等 2002）等。

为了评估极大似然估计量的性质，我们针对每种分布开展了模拟研究，考察了相关参数的三个不同取值和四种不同的样本量。我们通过平均偏差、基于平均标准误的置信区间以及名义水平为 95% 的覆盖率来评估这些估计量。对于泊松分布，在大样本下，我们获得了无偏且相合的估计量，偏差区间至多介于 (-.2;.2) 之间，但样本量等于 100 且参数取值为 ρ时除外。在所考察的所有情形中，覆盖率均接近 95%，介于 90% 和 98% 之间。与泊松分布相比，负二项分布表现出更大的变异性，尤其是因为负二项分布中的离散参数 ϕ 是对过度离散进行建模所必需的。对于负二项分布，在大样本下，我们同样获得了无偏且相合的估计量，且大多数情况下偏差区间介于 (-.5;.5) 之间。具体而言，我们发现相关参数和 ϕ_2 参数估计值的置信区间更宽，而 σ_2 和 ν_2 参数估计值的偏差较小。大多数情况下，覆盖率等于或高于 95%；当$\rho = 0$ 且样本量等于 100 时，相关参数的覆盖率低于 80%。

对于 COM-Poisson模型，参数估计量既不一致，也有偏差。回归参数在 β_{02} 时被低估，在 β_{01}时则没有偏差。相关参数 ρ 始终被估计得趋近于零：当 $\rho = -.5$ 时被高估，$\rho = +.5$ 时被低估，而在 $\rho = 0$ 时没有偏差。随机效应 σ_2 的标准差被高估，而 σ_1 则被低估。这种现象可能与离散参数 ν有关：当从 σ_2 中捕获更多方差时，从 ν_2中捕获的离散度就更少；另一方面，当从 σ_1中捕获的方差更少时，从 ν_1中捕获的离散度就更多。

由于大多数参数都有偏差，因此所有情景下的覆盖率都没有接近 95%。对于σ_2、 β_{02} 和 ν_r ，覆盖率介于 70% 至 90% 之间；对于 ρ ，在 3 种情景中的 2 种（$\rho = \{-0.5, 0.5\}$）下，覆盖率接近 80%。其他参数 β_{01}、 σ_1 和 ρ在 $\rho = 0$ 时的覆盖率接近 95%。

之后，使用各模型及其变体对这两个数据集进行了分析：无相关性、固定离散度、固定方差，以及所有随机效应具有共同方差。第一个数据集来自 AHS 调查，包含五个随机变量和 5190 名参与者。根据所采用的拟合度量，COM-Poisson模型的拟合效果最佳。COM-泊松分布、负二项分布和泊松模型对 β 和 ρ 的估计标准误相近。ν 参数捕捉到了 5 个欠离散响应变量中的 4 个以及一个过度离散响应变量，随后是一个较小的 σ。与其他模型相比，COM-Poisson模型得出了更多显著的相关值。

第二个数据集是 ANT，其中包含 41 个响应变量，用于计数澳大利亚 30 个地点的蚂蚁物种数量。根据广义离散指数（GDI），该多变量响应可视为过度离散。就赤池信息准则和对数似然而言，COM-Poisson模型也是最佳模型；贝叶斯信息准则最佳的模型则是固定离散度的 COM-Poisson 模型。在几乎所有比较中，COM-Poisson 模型的标准误都小于其他模型。对于所有响应变量， ν 参数均大于 1，表明存在欠离散。

因此，我们建议对计数数据使用多元广义线性混合模型框架。具体而言，在两个真实数据集中，COM-Poisson模型取得了最佳结果。它的主要优势在于能够同时对所有响应变量进行建模，并测量随机效应之间的相关性。

该模型的估计过程计算量很大，实现该模型是一项计算挑战。例如，估计 COM-Poisson 模型

用于 AHS数据的模型颇为繁琐。在配备 AMD Opteron 6136 处理器、92 GB 内存的 Debian GNU/LINUX 8（jessie）服务器上使用两个线程进行拟合，耗时 5 天。改进计算实现，甚至采用其他框架，都是未来工作的方向。在此背景下，改进 COM-Poisson 模型的估计过程需求迫切；同时，也有必要了解 COM-Poisson分布能否取代用于过度离散数据集的负二项分布。

参考文献

Aldrich J 等（1997）。“RA Fisher 与最大似然法的形成（1912—1922）。”《Statistical Science》，12(3)，162–176。Bates D，M¨achler M，Bolker B，Walker S（2015）。“使用 lme4 拟合线性混合效应模型。”《统计软件杂志》，67(1)，1–48。doi:10.18637/jss.v067.i01。Baydin AG，Pearlmutter BA，Radul AA，Siskind JM（2017）。“机器学习中的自动微分：综述。”《机器学习研究杂志》，18(1)，5595–5637。Bell BM（2005）。CppAD：用于 C++ 算法微分的软件包。网址：http://www.coin-or.org/CppAD。Blackford LS，Petitet A，Pozo R，Remington K，Whaley RC，Demmel J，Dongarra J，Duff I，Hammarling S，Henry G 等（2002）。“一套更新的基础线性代数子程序（BLAS）。”《ACM数学软件汇刊》，28(2)，135–151。Bonat WH（2016）。mcglm：多元协方差广义线性模型。R 软件包版本 0.3.0，网址：https://CRAN.R-project.org/package=mcglm。Bonat WH（2018）。“R 中的多响应变量回归模型：mcglm 软件包。”《统计软件杂志》，84(4)，1–30。doi:10.18637/jss.v084.i04。Bonat WH，Jørgensen B（2016）。“多元协方差广义线性模型。”《皇家统计学会杂志：C辑（应用统计）》，第 649–675 页。Bonat WH，Jørgensen B，Kokonendji CC，Hinde J，Dem´etrio CG（2018）。“扩展泊松–Tweedie 分布：计数数据的性质与回归模型。”《统计建模》，18(1)，24–49。Breslow NE，Clayton DG（1993）。“广义线性混合模型中的近似推断。”美国统计学会杂志，88(421)，9–25。Brooks ME，Kristensen K，van Benthem KJ，Magnusson A，Berg CW，Nielsen A，Skaug HJ，Maechler M，Bolker BM（2017）。“glmmTMB 在零膨胀广义线性混合建模软件包中兼顾速度与灵活性。”《R Journal》，9(2)，378–400。网址：https://journal.r-project.org/archive/2017/RJ-2017-066/index.html.¨BUrkner PC（2018）。“使用 R 软件包 brms 进行高级贝叶斯多层建模。”《R Journal》，10(1)，395–411。doi:10.32614/RJ-2018-017。da Silva GP，Laureano HA，Petterle RR，J´unior PJR，Bonat WH（2022）。“欠离散计数数据的多变量广义线性混合模型。”doi:10.48550/ARXIV. 2205.10486。网址：https://arxiv.org/abs/2205.10486。El-Shaarawi AH，Zhu R，Joe H（2011）。“使用泊松–Tweedie 族对物种丰度建模。”《Environmetrics》，22(2)，152–164。Famoye F（2015）。“多元广义泊松回归模型。”《统计通讯：理论与方法》，44(3)，497–511。ISSN 0361-0926。doi:10.1080/03610926.2012.743565。

Fisher RA（1934）。“确定频数的方法对频数估计的影响。”《优生学年鉴》，6(1)，13–25。Gibb H，Stoklosa J，Warton DI，Brown A，Andrew NR，Cunningham S（2015）。“形态能否预测蚂蚁物种及其群落的营养级和栖息地利用？”《生态学》，177(2)，519–531。Guennebaud G，Jacob B 等（2010）。“Eigen v3。”http://eigen.tuxfamily.org。Hadfield JD（2010）。“多响应广义线性混合模型的 MCMC 方法：MCMCglmm R 软件包。”《统计软件杂志》，33(2)，1–22。网址：http://www.jstatsoft.org/v33/i02/。Huang A（2017）。“离散计数数据的均值参数化 Conway–Maxwell–Poisson 回归模型。”《统计建模》，17(6)，359–380。Inouye DI，Yang E，Allen GI，Ravikumar P（2017）。“源自泊松分布的计数数据多元分布综述。”《Wiley 跨学科评论：计算统计学》，9(3)，e1398。Jørgensen B，Kokonendji CC（2016）。“离散离散度模型及其 Tweedie 渐近性质。”《AStA 统计分析进展》，100(1)，43–78。Kokonendji CC，Puig P（2018）。“多元计数分布的 Fisher 离散指数：综述与新提案。”《多元分析杂志》，165，180–193。Kristensen K，Nielsen A，Berg CW，Skaug H，Bell BM（2016）。“TMB：自动微分与拉普拉斯近似。”《统计软件杂志》，70(5)，1–21。doi：10.18637/jss.v070.i05。Lee Y，Nelder JA（1996）。“层级广义线性模型。”《皇家统计学会杂志：B辑（方法论）》，58(4)，619–678。ISSN 00359246。网址：http://www.jstor.org/stable/2346105。Liang KY，Zeger SL（1986）。“使用广义线性模型进行纵向数据分析。”Biometrika，73(1)，13–22。Mu˜noz-Pichardo JM，Pino-Mej´ıas R，Garc´ıa-Heras J，Ruiz-Mu˜noz F，Gonz´alez-Regalado ML（2021）。“计数数据的多元泊松回归模型。”《应用统计》，0(0)，1–17。doi:10.1080/02664763.2021.1877637。https://doi.org/10.1080/02664763.2021.1877637，网址：https://doi.org/10.1080/02664763.2021.1877637。Nelder JA，Wedderburn RW（1972）。“广义线性模型。”《英国皇家统计学会杂志：A辑（总论）》，135(3)，370–384。Nikoloulopoulos AK，Karlis D（2009）。“使用联结函数对多元计数数据建模。”《统计通讯：模拟与计算》，39(1)，172–187。Pinheiro J，Bates D，DebRoy S，Sarkar D，R 核心团队（2017）。nlme：线性与非线性混合效应模型。R 软件包版本 3.1-131，网址：https://CRAN.R-project.org/ package=nlme。R 核心团队（2020）。R：一种用于统计计算的语言和环境。R 统计计算基金会，奥地利维也纳。网址：https://www.R-project.org/。Shmueli G，Minka TP，Kadane JB，Borle S，Boatwright P（2005）。“一种适用于离散数据拟合的有用分布：Conway–Maxwell–Poisson 分布的复兴。”《皇家统计学会杂志：C辑（应用统计）》，54(1)，127–142。

Tierney L, Kadane JB（1986）。“后验矩与边际密度的精确近似。”美国统计学会杂志，**81**(393), 82– 86。doi:10.1080/01621459.1986.10478240。https://www.tandfonline.com/doi/pdf/10.1080/01621459.1986.10478240，网址 https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478240。

Wang Y, Naumann U, Eddelbuettel D, Wilshire J, Warton D（2020）。*mvabund*：多元丰度数据分析的统计方法。R 软件包版本 4.1.3，网址 https://CRAN.R-project.org/package=mvabund。

Wedderburn RW（1974）。“准似然函数、广义线性模型与高斯—牛顿法。”*Biometrika*, **61**(3), 439–447。

Winkelmann R (2008). 计数数据的计量经济学分析. 施普林格科学与商业媒体。

Winter B, B¨urkner PC (2021). “面向语言学家的泊松回归：使用 brms 进行计数数据建模的教程式介绍。”《语言与语言学指南》, **15**(11), e12439.

Zeileis A, Kleiber C, Jackman S (2008).“R 中的计数数据回归模型。”《统计软件杂志》, **27**(8), 1–25.

Zeviani WM, Ribeiro Jr PJ, Bonat WH, Shimakura SE, Muniz JA (2014). “伽马计数分布在实验性欠离散数据分析中的应用。”《应用统计学杂志》, **41**(12), 2616–2626.

所属机构：
Guilherme P. Silva 统计与地理信息实验室 巴拉´
纳联邦大学 CEP-81530-015 库里蒂巴（PR），
巴西 电子邮件：
guilhermeparreira.silva@gmail.com网址：
https://github.com/guilhermeparreira