Open Journal of Nature Science 自然科学, 2015, 3(3), 81-88
Published Online August 2015 in Hans. http://www.hanspub.org/journal/ojns
http://dx.doi.org/10.12677/ojns.2015.33011

Hans汉斯

The Application Research of Model Selection and Model Averaging in Meta-Analysis

Xiaoxiao Yin

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
Email: yinxiaoxiao2011@163.com

Received: Jul. 27th, 2015; accepted: Aug. 14th, 2015; published: Aug. 21st, 2015

Abstract

Model selection and model averaging have been the important issues which are researched by statistics and economic circles. This paper relies on theories and methods of Meta-analysis and takes the analysis of factors influencing legumes-rhizobium mutualism cooperative systems as an example. Then, the application results of model selection and model averaging method in meta-analysis are compared. The results show that model averaging method can be applied to meta-analysis and its performance is better than model selection.

Keywords

Meta-Analysis, Model Selection, Model Averaging

模型选择与模型平均在Meta分析中的应用研究

尹潇潇

云南财经大学统计与数学学院，云南 昆明
Email: yinxiaoxiao2011@163.com

收稿日期：2015年7月27日；录用日期：2015年8月14日；发布日期：2015年8月21日

摘 要

模型选择与模型平均一直是统计学与计量经济学界研究的重要问题，本文依托Meta分析理论和方法，以

文章引用: 尹潇潇. 模型选择与模型平均在 Meta 分析中的应用研究[J]. 自然科学, 2015, 3(3): 81-88.
http://dx.doi.org/10.12677/ojns.2015.33011

分析豆科植物-根瘤菌互利共生合作系统的影响因素为例，比较模型选择与模型平均方法在Meta分析中的应用效果，结果表明模型平均方法既可应用于Meta分析中，分析效果又优于模型选择。

关键词

Meta分析，模型选择，模型平均

1. 引言

通过统计建模解决一个实际问题时，往往我们可以建立多个统计模型，现在面临的问题是究竟采用哪个模型分析问题的效果较好，或者是怎样将所有模型提供的信息都充分利用起来，这就涉及到模型选择和模型组合的问题了。模型选择旨在从备选模型集中选择一个最优的模型，而模型组合(即模型平均)则为了充分利用所有模型提供的信息，给每个模型赋予一定的权重将它们组合起来。这样就避免选到一个分析效果差的模型，因为通过模型选择选出来的模型未必就是效果好的，只能说在所有模型中相对较好。因此，本文以分析豆科植物－根瘤菌互利共生系统的影响因素为例，分别使用这两种方法进行分析，然后比较这种方法的研究结果，结果表明模型平均方法不仅可以应用于 Meta 分析，而且其分析效果优于模型选择。

2. 研究背景

很多学者通过模型选择与模型平均方法分析解决一些实际问题，如 Jerald B. Johnson 等[1]在研究生态学与进化论的相关问题时，通过 Akaike information criterion (AIC)、Small sample unbiased AIC (AICc)、似然比检验以及 Schwarz criterion 等准则进行模型选择。尽管模型选择在一定条件下可以解决很多问题，但是它仍然有一些缺陷，如：可能会丢失一些有用的信息，由于问题分析都是基于所选择的模型，这样其他可能模型所反映的信息就会丢失；存在很大的风险，选择的模型可能与真实数据产生的过程比较接近，但有时相差甚远，而我们的目标并一定是寻找真实数据产生的过程，有时是估计参数或做预测。张新雨等[2]采用多种模型平均方法如 S-AIC、S-BIC、Jacknife 等构建组合模型对我国的粮食产量进行预测。Donald Berry 等[3]于贝叶斯模型平均(BMA)的方法分析补充维生素 E 和死亡率的关系；在 Meta 分析的研究中，LiQuefeng 等[4]使用 Lasso 变量选择的方法处理基因表达数据。当 Meta 分析中涉及到多个备选模型时，一个自然的问题是如何确定一个合适的模型并据此开展分析，该问题落入了模型选择的范畴。鉴于上述模型选择方法的一些缺陷，国内外学者提出了许多解决办法，其中模型平均方法就是一种常用的而且备受欢迎的方法。但是，Meta 分析中的模型平均问题还没有得到研究。接下来我们将模型平均方法应用于 Meta 分析中，研究豆科植物－根瘤菌互利共生系统的影响因素。

3. 研究方法介绍

3.1. 模型选择(Model Selection)

模型选择，伴随着不确定性，它是目前利用强大的计算机及其软件更详细的去探索数据提供信息的一个实践的例子。简而言之，模型选择就是从建立的众多可能模型中选择一个最适合解决已知问题的模型。模型选择的方法有很多，如 Akaike information criterion (AIC)、Schwartz’s Bayes information criterion (BIC)，AIC 和 BIC 的计算公式如(1)、(2)式：

$$AIC_i = -2\log \ell_i + l_i, \quad i = 1,2,\cdots,k \tag{1}$$

$$BIC_i = -2\log \ell_i + l_i \log n, \quad i = 1,2,\cdots,k \tag{2}$$

其中 k 为候选模型集中模型的个数；ℓ_i 和 l_i 分别为第 i 个模型的极大似然函数和未知参数的个数；而 n 为样本的容量。

此外，我们还有其他模型选择的方法，如 focused information criterion (FIC)，FIC 是通过极小化固定参数估计的均方误差(MSE)进行模型选择的，通常选择最小的 MSE 模型，这说明了 FIC 可应用于一些常见的情况下。虽然这些方法在统计学领域应用的非常普遍，但是它们也有各自的优点和缺点，一方面，通过 AIC、BIC 和 FIC 选择到一个最好的模型，我们可以利用该模型去解释数据的所有方面；另一方面，我们并不能确定所选择的模型对于一个估计好但对其他的估计效果也许会很差。因此，下面我们就介绍弥补模型选择缺陷的方法——模型平均。

3.2. 模型平均(Model Averaging)

模型平均，顾名思义，就是把来自不同模型的估计或者预测通过一定的权重平均起来，在一些文献中也称为模型组合，它一般包括组合估计和组合预测。事实上，模型选择是模型平均的特例，模型的权重取 0 或 1，所以模型平均所做的估计比较稳健。模型平均的关键在于如何选取组合的权重，常用的权重选择方法有 Smoothed AIC (S-AIC)、Smoothed BIC (S-BIC)、Mallow 准则、Jackknife 准则和 OPT 最优权重选择法等。下面我们简单的介绍下基于 S-AIC、S-BIC 以及 Mallow 准则的权重选择方法。本文主要运用 S-AIC 进行模型平均。

基于 S-AIC 和 S-BIC 准则的权重选择方法由 Buckland，Burnham 和 Augustin (1997) [5]提出，权重的计算公式为：

$$\omega_i = \frac{\exp\left(-xIC_i/2\right)}{\sum_{j=1}^{k}\exp\left(-xIC_i/2\right)},\ i=1,2,\cdots,k \tag{4}$$

其中 k 表示候选模型集中模型的个数；i 代表第 i 个模型；ω_i 是第 i 个模型的权重；xIC_i 为第 i 个模型的 AIC 或 BIC。由上可发现 S-AIC 和 S-BIC 的计算比较简单，因而是比较常用的权重选择方法。

Mallow 准则[6]最初是在 2007 年由 Hansen 提出的，他通过极小化 Mallow' Cp 准则构建了最小二乘模型平均估计。具体过程为：

假设候选模型集中共有 M 个近似模型，被解释变量的实际值序列记为 $Y=\left(y_1,y_2,\cdots,y_n\right)'$，预测值序列为 $\mu=\left(\mu_1,\mu_2,\cdots,\mu_n\right)'$，$\omega=\left(\omega_1,\omega_2,\cdots,\omega_M\right)'$ 为 M 个模型的权重集合，且满足 $H_n=\left\{\omega\in[0,1]^M:\sum_{m=1}^{M}\omega_m=1\right\}$。则 μ 的模型平均估计为

$$\hat{\mu}(\omega)=\sum_{m=1}^{M}\omega_m P_{(m)}Y\equiv P(\omega)Y \tag{5}$$

其中 $P_{(m)}=X_{(m)}\left(X'_{(m)}X_{(m)}\right)^{-}X'_{(m)}$，$X_{(m)}$ 是一个 $n\times k_m$ 矩阵，其元素为 $x_{ij(m)}$（$x_{ij(m)}$ 为第 m 个模型的解释变量的取值序列），k_m 为第 m 个模型含有的回归变量个数。

模型平均的 Mallow 准则为

$$C_n(\omega)=\left(Y-\hat{\mu}(\omega)\right)'\left(Y-\hat{\mu}(\omega)\right)+2\sigma^2 trP(\omega) \tag{6}$$

其中 $\sigma^2=E\left(e_i^2\middle|x_i\right)$，$x_i$ 为自变量的第 i 个取值，e_i^2 为实际值与估计值误差的平方。

通过极小化(6)式就可以得到各个模型的权重，即 $\hat{\omega}=\arg\min_{\omega\in H_n}\left(C_n(\omega)\right)$。将 $\hat{\omega}$ 带入(5)式就可以得到观测值的 Mallow Model Averaging (MMA)估计。

鉴于 S-AIC 的计算依靠 AIC，因此，本文选择 AIC 进行模型选择。模型平均与模型选择相比，模型

平均避免选到一个较差的模型；模型平均往往是将多个模型赋予权重组合起来，不会丢掉任何模型和遗失任何有用的信息，这样就可以充分利用所有的信息分析、解决问题；模型平均法并没有将建立的模型当作数据产生的真实过程，这样就可以保证估计或预测的准确性。模型平均方法经常被用于经济领域或其它领域做预测。

3.3. Meta-Analysis

针对同一个问题，可能会有很多不同的研究结果，这就需要我们采用一定的方法整合分析所有研究结果，最终得到一个相对统一且被大家认可的结论，即 Meta 分析(Meta-analysis)。Meta 分析是汇总多项原始研究的结果并分析评价其合并效应量的一系列过程，它依靠搜集已发表或未发表的具有某一可比性的文献，应用特定的统计学方法进行合并分析与综合评价[7]。Meta-analysis 又称为“荟萃分析”，是对具备特定条件的、同课题的诸多研究结果进行综合的一类统计方法。Meta-analysis 最重要的是文献资料的筛选，因为所选文献直接决定了我们提取的数据情况。Meta 分析中，Meta 回归模型通常包含以下两类：

第一，固定效应模型

$Y_i \sim N\left(\theta_i, \sigma_i^2\right)$，$i = 1, 2, \cdots, k$，其中 Y_i 为第 i 个研究的统计量，期望 $\theta_i = \theta + \beta x_i$。

第二，随机效应模型

若 $Y_i \sim N\left(\theta_i, \sigma_i^2\right)$，$\theta_i \sim N\left(\theta + \beta x_i, \sigma_a^2\right)$，则随机效应 Meta 回归模型为 $Y_i \sim N\left(\theta + \beta x_i, \sigma_a^2 + \sigma_i^2\right)$。

我们常常使用 Stata 软件建立 Meta 回归模型，命令如：metareglnor factor 1，factor 2，factor 3，…，wsse (selnor)；如果考虑交互效应时，命令中将交互项(interaction)添加进去即可。

本节通过 Meta 分析可以得到很多可能的分析模型，根据 3.2 中的方法，我们可以得到备选模型集中每个模型所占的权重集合即 $\hat{\omega}$，将这些权重赋给对应的模型，从而将所有可能的模型组合起来最终得到一个组合模型(与模型选择相比，是全模型)，然后就可以运用该模型对所要研究的问题进行分析、预测了。接下来就将上面介绍的模型选择与模型平均方法结合 Meta 分析应用到豆科植物-根瘤菌互利共生合作系统影响因素的实例分析中。

4. 模型选择与模型平均在 Meta 分析中的应用实例分析

针对同一个问题，可能得到不同甚至相左的研究结果，因此，采用 Meta 分析可以整合这些不同研究结果提供的信息，最终分析得到一个统一的结果。所以，通过 Meta 分析方法比使用其他方法得到的研究结果更为客观、全面。基于作者之前采用 AIC 和 AICc 准则进行模型选择做过豆科植物－根瘤菌互利共生合作系统的相关研究，而且 AICc 准则和 S-AIC 准则都是对 AIC 修正衍化得到，因此，在本文中，作者通过 S-AIC 权重选择准则进行模型平均分析，与 AIC 模型选择进行比较，这具有重要的意义。当然，在未来的研究中，我们可以通过其他模型平均方法(如 S-BIC、Mallow 准则等)进行研究分析，然后与 S-AIC 进行比较，这样可以很好的说明这些权重选择准则在模型平均方法中的应用效果。

所以，本文依托 Meta 分析理论和方法，收集大量豆科植物－根瘤菌互利共生合作系统研究的相关文献，根据文献中提供的二次数据建立 Meta 回归模型；基于 AIC 准则考虑模型选择问题；最后基于 S-AIC 准则研究 Meta 分析中的模型平均问题，进而据此分析豆科植物－根瘤菌互利共生系统中的影响因素。结果表明模型平均方法不但可以应用于 Meta 分析中，而且其分析效果优于模型选择。

4.1. 资料来源

本文主要通过 Meta 分析，建立 Meta 回归模型，结合模型选择与模型平均法分别研究豆科植物与根瘤菌组成的互利共生合作系统的影响因素。我们在 ISI Web of Knowledge Web of Science database 和谷歌学术(Google Scholar)上查询国内外学者对豆科植物－根瘤菌互利共生合作系统所做的相关研究，将查询

到的文献所提供的一些信息作为 Meta 分析和 Meta 回归的实例数据。我们将可能对该合作系统造成影响的已知研究中的五个因素：宿主类型(Host Classification)、合作者类型(Cooperator genus)、施肥与否(Fertilization)、控制措施(Measured effect)和接种复杂度(Design class)分别记作 X1、X2、X3、X4、X5。

4.2. 模型选择分析

根据资料提供的数据信息，我们建立 Meta 回归模型进行分析。除了各个单因素即主效应为可能的影响因素外，各个主效应之间的交互作用也可能对豆科植物－根瘤菌互利共生合作系统产生一定的影响。由于我们不能确定包含哪些解释变量可以得到较好的分析结果，因此，建立 Meta 回归模型时应将所有的主效应以及交互效应都考虑进去。用 Stata 软件建立 Meta 回归模型，探索性的建立很多模型，记录各个模型的 F 统计量、P 值以及方差，经过筛选，最终得到 51 个模型。简单列举几个模型：

Model1:metareglnorX3X4X1*X4 X1*X5 X3*X4,wsse (selnor);

Model2:metareglnorX3X4X5 X1*X4 X1*X5 X3*X4,wsse (selnor);

Model3:metareglnorX2 X3X4X1*X4 X3*X4,wsse (selnor);

Model4:metareglnorX3X4X1*X4 X1*X5 X3*X4,wsse (selnor);

Model5:metareglnorX3X4X5X1*X4 X3*X4,wsse (selnor);

… … …

Model51:metareglnorX3X4X5X1*X3X1*X4,wsse(selnor)。

通过 Matlab 软件分别计算这 51 个模型的极大似然函数，再结合公式(1)计算并记录每个模型的 AIC，各个模型的 AIC 值如表 1 所示。

观察表 1，将所有模型的 AIC 值进行排序，得到 AIC 值最小的模型为 Model2，$AIC_{\min} = 54.5$。通过 AIC 模型选择准则，我们从所有候选模型中选择模型 2 对豆科植物－根瘤菌互利共生合作系统的影响因素进行分析，认为该模型在所有模型中的分析效果最好。建立模型 2 的命令为 metareglnorX3X4X5 X1*X4 X1*X5 X3*X4,wsse (selnor)，由此发现该模型包含主效应 X3 (Fertilization)、X4 (Measured effect)以及 X5 (Design class)，此外还有交互效应 X1*X4 (Host Classification* Measured effect)、X1*X5 (Host Classification* Design class)和 X3*X4 (Fertilization* Measured effect)。经过模型选择以及由模型 2，我们得到影响豆科植物-根瘤菌互利共生系统的主要因素为 X3、X4、X5、X1*X4、X1*X5 以及 X3*X4。

4.3. 模型平均分析

根据 4.2 得到的 51 个模型，每个模型含有的变量种类以及个数不同，解释效果也不同，因此，下面通过模型平均整合这些模型提供的信息进行全面分析。根据公式(4)以及 4.2 计算的 AIC 值，我们可以分别计算出各个模型所占的 S-AIC 权重。利用计算出的权重结果将 51 个模型组合起来，最终得到一个组合分析模型。S-AIC 权重结果如表 2 所示。

通过模型平均，我们最终得到一个包含 51 个模型的组合模型，每个模型都有一个权重。由表 2 可知组合模型中权重较大且排在前五的模型有第二、第三、第四、第五和第九个，说明这五个模型在整个组合模型中的所起的解释作用最大。这五个模型的构建命令如下：

Model2:metareglnorX3X4X5 X1*X4 X1*X5 X3*X4,wsse (selnor);

Model3:metareglnorX2 X3X4X1*X4 X3*X4,wsse (selnor);

Model4:metareglnorX3X4X1*X4 X1*X5 X3*X4,wsse (selnor);

Model5:metareglnorX3X4X5X1*X4 X3*X4,wsse (selnor);

Model9:metareglnorX3X4X1*X4 X3*X4,wsse (selnor);

Table 1. Akaike Information Criterion of every model
表 1. 每个模型的 AIC

模型序数	AIC	模型序数	AIC	模型序数	AIC
1	62.6	18	61.4	35	62.3
2	54.5	19	63.1	36	63.6
3	58.4	20	64.4	37	64.8
4	57.3	21	62.2	38	64.8
5	58.8	22	62.2	39	65.9
6	59.7	23	66	40	65
7	59.7	24	66	41	66.3
8	61	25	65.1	42	66.3
9	57.5	26	62.8	43	66.4
10	59.9	27	62.8	44	64.1
11	60.9	28	64	45	65.3
12	63.3	29	65.2	46	65.5
13	60.2	30	63.2	47	65.5
14	61.6	31	64.5	48	65.5
15	61.6	32	65.8	49	65.5
16	63.6	33	65.8	50	65.5
17	63.6	34	64.7	51	65.8

Table 2. The weight of Smoothed Akaike Information Criterion
表 2. S-AIC 权重

模型序数	S-AIC 权重	模型序数	S-AIC 权重	模型序数	S-AIC 权重
1	0.007	18	0.0132	35	0.00812
2	0.4	19	0.00558	36	0.00437
3	0.0577	20	0.00294	37	0.0024
4	0.0992	21	0.00873	38	0.0024
5	0.0488	22	0.00873	39	0.00135
6	0.0309	23	0.00133	40	0.00213
7	0.0309	24	0.0013	41	0.00113
8	0.0155	25	0.00202	42	0.00113
9	0.0908	26	0.0065	43	0.00106
10	0.027	27	0.0065	44	0.00341
11	0.0168	28	0.00358	45	0.00182
12	0.00505	29	0.00193	46	0.00101
13	0.0233	30	0.0054	47	0.00171
14	0.012	31	0.00275	48	0.00171
15	0.0117	32	0.00147	49	0.00169
16	0.00425	33	0.00147	50	0.00165
17	0.00425	34	0.00255	51	0.00143

从整个组合模型来看，所有模型中包含的可能的解释变量对豆科植物－根瘤菌互利共生合作系统或多或少都有影响，只不过所占权重较大的模型中的解释变量影响作用较大。模型平均结果表明：X2、X3、X4、X5、X1*X4、X3*X4、X1*X5 是影响该合作系统的主要因素，其他的如 X1、X1*X3、X3*X5、X4*X5 等因素也有一定的影响，不过影响力较小而已。

4.4. 模型选择与模型平均法所得结果的比较

一方面，利用 AIC 准则我们从 51 个模型中选择了一个最好的模型，得到豆科植物－根瘤菌互利共生系统的主要影响因素有 X3、X4、X5、X1*X4、X1*X5 以及 X3* X4。而通过模型平均方法，使用 S-AIC 给每个模型赋权重，最后所得全模型，该组合模型包含了 51 个模型，解释作用低的模型其权重较小，最终得到的影响因素有 X2、X3、X4、X5、X1*X4、X3*X4、X1*X5、X1、X1*X3、X3*X5、X4*X5 等。与模型选择相比，我们给 51 个模型都赋予了各自的权重，得到的这样一个组合模型包含了数据提供的所有信息，而且模型平均最后得到的组合模型包含模型选择选出的模型 2。因此，我们分析豆科植物-根瘤菌互利共生系统的影响因素，模型平均方法比模型选择更加方便准确，得到的影响因素相对较全面具体。

另一方面，表 2 的分析结果揭示了模型平均方法不但为我们提供了豆科植物－根瘤菌互利共生系统的影响因素，也具体给出了每个模型中的每个因素对该系统影响的重要性。而模型选择仅仅给出了所选的一个最优模型，从这个模型只能发现哪些因素对共生系统有影响，信息量较小，与模型平均相比，遗失了很多有用信息。

5. 结论

综上所述，基于模型选择与模型平均方法，通过对影响豆科植物-根瘤菌互利共生系统的因素实例进行 Meta 分析，由分析结果说明模型平均方法不但可以应用于 Meta 分析中，而且其分析效果优于模型选择。具体结论如下：

(1) 本文基于 Meta 分析，整合了豆科植物－根瘤菌互利共生系统的相关研究，建立了许多 Meta 回归模型。基于所建立的 Meta 回归模型，我们分别采用模型选择和模型平均方法对豆科植物－根瘤菌互利共生系统影响因素的分析模型进行确定。结果表明：模型平均方法得到的组合分析模型，详细的解释了豆科植物－根瘤菌互利共生合作系统的主要影响因素有 X2、X3、X4、X5、X1*X4、X3*X4、X1*X5、X1、X1*X3、X3*X5、X4*X5 等，此外，根据每个模型所占的权重，发现 X4、X3、X1*X4 以及 X3*X4 这四个因素的影响作用最大。而通过模型选择，得到 X3、X4、X5、X1*X4、X1*X5 以及 X3*X4 是该合作系统的影响因素。模型选择得到的模型中没出现的变量并不能说明其对合作系统没有影响，模型平均方法弥补了这一缺陷。

(2) 事实上，线性回归中，模型平均方法的分析效果往往优于模型选择，经本文研究发现：模型选择与模型平均方法除了应用于线性回归中之外，也均可应用于 Meta 分析中，而且模型平均的分析效果优于模型选择，这和线性回归的结论是一致的。因此，未来可将模型平均广泛的应用于统计学以及其他领域。

基金项目

本文得到云南财经大学研究生创新基金项目(2015YUFEYC015)的资助。

参考文献 (References)

[1] Johnson, J.B. and Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology & Evolution*, **19**, 101-108. http://dx.doi.org/10.1016/j.tree.2003.10.013

[2] 张新雨，邹国华 (2011) 模型平均方法及其在预测中的应用. *统计研究*, **6**, 97-102.

[3] Berry, D., Wathen, J.K. and Newell, M. (2009) Bayesian model averaging in meta-analysis: vitamin E supplementation and mortality. *Clinical Trials*, **6**, 28-41. http://dx.doi.org/10.1177/1740774508101279

[4] Li, Q., Wang, S., Huang, C.C., et al. (2014) Meta-analysis based variable selection for gene expression data. *Biometrics*, **70**, 872-880. http://dx.doi.org/10.1111/biom.12213

[5] Buckland, S., Burnham, K. and Augustin, N. (1997) Model selection: An integral part of inference. *Biometrics*, **53**, 603-618. http://dx.doi.org/10.2307/2533961

[6] Hansen, B. (2007) Least squares model averaging. *Econometrica*, **75**, 1175-1189. http://dx.doi.org/10.1111/j.1468-0262.2007.00785.x

[7] 石修权, 王增珍 (2008) Meta 回归与亚组分析在异质性处理中的应用. *中华流行病学杂志*, **5**, 497-501.