·专家论坛·

Logistic 回归在职业流行病学研究中的应用

北京医科大学第三医院中心实验室（100083）　赵一鸣

Logistic 回归是近年来发展起来的一种具有广泛应用前途的多因素统计方法。我们在职业流行病学研究中应用这种新的研究手段、结合专业特点提出了一些新的方法，解决了一些难题，提高了职业流行病学研究的质量和水平。现将体会、收获简述如下。

1 校正混杂因素的干扰，正确评价职业暴露与疾病的联系

在职业流行病学研究中混杂因素是普遍存在的。通常解决混杂因素干扰的方法有配对设计、分层分析等。这些方法可以在一定程度上避免混杂因素的干扰，但在实际工作中这些方法存在许多不足之处，如配对研究时不易配到足够的对子数、分层分析时需要较大的样本量等，Logistic 回归在校正混杂因素干扰方面可以克服这些缺点。我们在研究噪声与高血压病关系时了解到许多研究报告的结果很不一致。通过分析发现，混杂因素干扰是造成结果不一致的主要原因。我们在应用Logistic 回归校正年龄、遗传、盐摄入量等因素的影响之后观察到噪声暴露与高血压的联系强度略低于单因素分析的结果。与过去许多研究相比，这一结果处于中间水平。因此许多学者认为我们的研究结果比较符合实际情况。在教练员与高脂血症关系的调查中发现，由于存在多个混杂因素的干扰，教练员的高脂血症患病率虽然略高于对照组，但是无显著差异。而用Logistic 回归校正多个混杂因素干扰后教练员这种特殊的人群和职业与高脂血症的联系就明显增大。这些工作表明，用Logistic 回归可以校正混杂因素的干扰，使我们能够客观地评价职业暴露与疾病的联系。

2 剂量-反应关系研究

职业暴露与疾病之间是否存在剂量-反应关系是职业流行病学研究中确定职业暴露与疾病联系的重要依据。以往研究常采用分层分析的方法，但结果比较粗糙。现有的Logistic 回归分析软件允许连续变量进入模型。同时Logit曲线是S型的，它的不同部分可以近似地拟合直线、对数和指数曲线，非常适用于剂量-反应关系研究。利用这些特点分析噪声强度和累积噪声暴露量与高血压关系时发现，在校正了多种因素影响后工人接触噪声的强度每增加 1dB(A)时高血压的 OR 为 1.033。如果一个工人从 70dB(A) 的环境调换到 100dB(A)的环境中长期工作，他患高血压的危险性 (OR) = $(1.033)^{100-70} = 2.65$。用这种方法可以计算出任何噪声强度暴露时高血压的OR，而分层分析不能进行这种细致的分析。

3 比较多种危险因素在疾病发生中的作用强度和相对比例，确定职业危害在疾病发生中的地位

职业流行病学研究非常希望了解职业暴露造成的危害究竟有多大，它与其它危险因素相比哪一个更重要，以指导选择合理的措施来预防疾病的发生。例如，噪声可以引起高血压，盐摄入量过高也可以引起高血压，如果两者各自都可以引起高血压而且危险程度相似，不论降低噪声或减少盐摄入量都可以达到降低高血压病发生的目的。在制定防治措施时人们可以根据实际情况决定采取什么方法进行预防。但是，要了解不同危险因素在疾病发生中的作用强度和相对比例有实际困难，如危险因素的变量类型可能不同（连续变量、等级变量和二分变量），变量的内涵不同。如果没有一个共同的标准，实际上不可能进行这类比较。因此需要采用标化的方法来做这类分析。Logistic 回归可以采用三种不同的标化方法比较各危险因素在疾病发生中的作用强度和相对比例，最大似然值差值、标化偏回归系数平方和标化偏回归系数。在噪声与高血压的研究中采用最大似然值差值比较各危险因素引起高血压的危险性，发现年龄的危险性约为噪声的10倍，父母高血压史和盐摄入量过高分别为噪声的2倍。这一结果为我们今后选择预防噪声危害的措施提供了重要的依据。

4 对职业暴露估计模型的优劣进行评价

由于Logistic 回归可以接受连续变量作为预报变量、可以校正混杂因素的干扰，因此可用于评价不同职业暴露估计模型的优劣。其原理是，当模型与实际暴露情况愈接近时其估计的剂量-反应关系就拟合得愈好，Logistic 回归中的最大似然值愈大。用这种方法分析了按等能量原理和等效应学说建立起来的累积噪声暴露模型，发现当模型的转换系数(k) 为2.9时最大似然值最大，表明此时噪声暴露与高血压的剂量-反应关系拟合得最好。这一数值与等能量原理 $k=3$ 的转换系数非常接近，提示稳态噪声暴露与高血压的关系符合等能量原理。

5 区分年龄与工龄在疾病发生中的作用

在稳定的职业人群中年龄与工龄往往存在着高度相关的关系，这是职业流行病学研究中常见的现象。当某一疾病的发生与年龄有关，同时又与职业暴露长短（工龄）有关，无论采用现有的单因素分析或多因素分析方法都不能区分两者各自对疾病危险性的大小，Logistic 回归也没有这种能力。但在某些情况下，可以利用合并暴露强度和暴露时间来消除年龄与工龄高度相关的关系，进而利用Logistic 回归分析年龄和职业暴露各自在疾病发生中的作用。在噪声与高血压关系的研究中工人的年龄与工龄高度相关（$r=0.97$），造成这两个变量在 Logistic 回归中产生多元共线性，致使无法客观地分析年龄和噪声暴露年限与高血压的关系。在进一步研究中利用等能量公式将噪声强度与噪声暴露年限合并为累积噪声暴露量，此时年龄与累积噪声暴露量的相关性明显降低（$r=0.41$），再用Logistic 回归就可以同时分析年龄与噪声暴露各自与高血压的关系。

6 对职业流行病学研究设计的影响

Logistic 回归作为一种具有广泛用途的多因素统计方法，必然会影响职业流行病学研究设计。它的影响首先表现为设计时应考虑采用多因素分析方法，设计调查表时要全面考虑，以保证在分析资料时既可以采用单因素分析方法，又可以采用多因素分析方法。样本量设计的原则是：Logistic 回归研究所需的样本量以每一个危险因素与所观察的疾病进行单因素分析够用即可。Logistic 回归可以完全代替分层分析的方法而不需要增加样本量。设计时应尽量避免配对设计，因为配对研究的设想很好，但实际应用时往往难以找到合适的对子，使许多资料无法使用。同时，配对设计可以控制的混杂因素一般不超过三个，过多的控制因素可以造成“过度配对”，也会造成偏倚。一般情况下选择成组设计比较合理，只要保证各组在主要混杂因素方面保持基本均衡即可。但特别要注意设计调查表时一定要包括所有可能存在的混杂因素，以便在分析时考察和校正。采用成组设计的方法对研究对象的选择可以相对放宽，只要职业暴露情况清楚、各项资料齐全都可以作为研究对象。相对小的样本量、比较宽松的选择研究对象的条件和避免做配对调查，是Logistic 回归给职业流行病学研究设计带来的有利条件，可以使我们在更加宽阔的领域内更加自由地开展研究。

7 Logistic回归的缺点和应注意的事项

Logistic 回归有许多优点，但它并不是十全十美的。多元Logistic 回归要求每一个进入模型的研究对象的每一个项目都必须有可靠的信息。如果出现缺失值，该研究对象将被删除。因此，缺失很多的数据不宜用 Logistic 回归进行分析。Logistic回归模型允许不同类型的变量作为危险因素同时进入模型，但只允许二分变量作为结果在模型中出现。如研究高血压时每个人的血压都有实际测量值，但在Logistic 回归中必须按照一定的标准将研究对象区分为高血压和非高血压。这样做实际上丢失了一部分信息，使研究所需要的样本量增大。Logistic 回归特别适用于多种危险因素造成一种疾病的研究。但某一个危险因素与疾病的联系太强烈、样本量又相对较小时，这个危险因素不存在时可能没有一个研究对象得病，不能计算 OR 值，不能用Logistic回归。在做 Logistic 回归分析之前要做好单因素分析，因为多元Logistic 回归分析往往要参考单因素分析的结果。某些Logistic 回归分析程序可以自动根据一定的规则筛选变量，但这种方法容易出现失误，最好采用人工干预的方法一步一步分析。这样虽然慢一些，但可以结合单因素分析结果和专业知识，分析的结果比较符合实际情况。

Logistic 回归在职业流行病学研究中的应用刚刚开始，还有许多问题有待探索。我们希望能够合理地使用这种方法，以达到加快职业流行病学研究的速度、提高研究水平的目的。

（参考文献　略）

（上接第17页）

苯浓度均与其个体白细胞水平密切相关。笔者以为当日下午班中或班末终末呼出气及班末尿中苯浓度可视为当日的接触水平，第二天班前呼出气中的苯浓度可视为苯在机体内的蓄积水平。

我们的调查还发现在此环境中工作的52名作业工人的白细胞计数均值呈逐年渐降趋势。在工龄≥5年后的168人次检查中，WBC$<4\times10^9$/L者11人次，$4\times10^9\sim4.5\times10^9$/L者26人次，其均值水平比工龄<5年值明显降低（$P<0.01$）。本次调查的班末、第二天班前终末呼出气及班末尿中苯浓度均值均高于笔者曾提出的生物接触限值。为了更好地保护作业者健康，有必要研究制订苯及其代谢物在呼出气及尿液中的生物接触限值，配合最高容许浓度值一起运用。