合理进行多重线性回归分析——结合倾向性评分分析

胡纯严[1]，胡良平[1,2*]

（1. 军事科学院研究生院，北京 100850；
2. 世界中医药学会联合会临床科研统计学专业委员会，北京 100029
*通信作者：胡良平，E-mail：lphu927@163.com）

【摘要】 本文目的是介绍如何结合倾向性评分分析，合理地进行多重线性回归分析的方法。第一，介绍了与倾向性评分分析有关的3个基本概念。第二，介绍了倾向性评分分析的核心内容，即3种匹配方法。第三，通过一个流行病学的调查实例，介绍了如何用SAS软件进行分析的全过程，内容如下：①针对原始数据集，检验协变量在处理组与对照组之间的差异是否具有统计学意义；②针对原始数据集，直接进行多重线性回归分析；③采用倾向性评分分析产生匹配后的数据集；④针对匹配后的数据集，检验协变量在处理组与对照组之间的差异是否具有统计学意义；⑤针对匹配后的数据集，合理进行多重线性回归分析。

【关键词】 处理变量；倾向性评分分析；匹配方法；Logistic 回归模型；多重线性回归模型

中图分类号：R195.1 **文献标识码**：A doi：10.11886/scjsws20221113004

Reasonably conduct the multiple linear regression analysis combined with the propensity score analysis

Hu Chunyan[1], *Hu Liangping*[1,2*]

（1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
*Corresponding author: *Hu Liangping*, *E-mail*: *lphu*927@163.*com*）

【**Abstract**】 The purpose of this paper was to introduce how to combine the propensity score analysis to reasonably carry out multiple linear regression analysis. Firstly, it introduced 3 basic concepts related to the propensity score analysis. Secondly, it presented the core contents of the propensity score analysis, that was, three matching methods. Thirdly, through an epidemiological survey example, it gave the whole process of how to use SAS software for the analysis. The contents were as follows: ① for the original data set, test whether the difference of covariates between the treatment group and the control group was statistically significant; ② directly implement the multiple linear regression analysis for the original data set; ③ the propensity score analysis was used to generate the matched data set; ④ for the matched data set, test whether the difference of covariates between the treatment group and the control group was statistically significant; ⑤ a reasonable multiple linear regression analysis was used for the matched data set.

【**Keywords**】 Treatment variable; Propensity score analysis; Matching method; Logistic regression model; Multiple linear regression model

在开展临床研究和流行病学研究中，研究者收集到的资料通常都是多因素多指标的资料。当结果变量为计量变量时，研究者常选择多重线性回归模型来拟合资料[1-2]。当资料中存在处理变量时，研究者最关注的是处理变量对结果变量所产生的因果效应[3]。若资料为基于前瞻性随机对照研究设计收集到的大样本临床资料，协变量在处理组与对照组之间的可比性较好，通常可直接进行多重线性回归分析。然而，若资料为基于回顾性研究设计或前瞻性调查研究设计收集到的资料，即使样本含量非常大，也很难保证大多数协变量在处理组与对照组之间具有很好的可比性。直接基于可比性差的资料进行任何统计分析，其结果的可信度都会大打折扣。本文将介绍如何结合倾向性评分分析[3-4]，合理进行多重线性回归分析。

1 基本概念

1.1 处理变量

在新药的临床试验中，常规的做法是将某疾病患者随机均分入试验组与对照组，试验组患者接受新药治疗，对照组患者接受对照药物（通常为安慰剂）治疗。这里，“药物种类”就是研究者关注并考查的重要影响因素，它在统计学上常被称为“处理

变量”或“试验因素”。在流行病学的调查研究中，也常涉及处理变量。例如，在研究吸烟与患肺癌关系的调查研究中，“是否吸烟”就是一个处理变量；又例如，在戒烟与体重改变量之间关系的调查研究中，“是否戒烟”就是一个处理变量。

1.2 处理变量两水平组之间协变量的可比性

来自受试者的各种非试验因素常被称为协变量，例如年龄、性别、种族、血型、职业、生活方式、行为习惯、身高、体重、体重指数、血压和血脂等。处理组与对照组中的受试者若在诸多协变量上的取值不同，对结果变量的影响也必然不同。因此，即使研究者观察到两组受试者在某些结果变量上存在明显差别，也不完全是由处理水平与对照水平之间的效应之差所致。

由统计学的理论可知，在样本含量足够大的随机对照临床试验中，处理变量T的两个水平组（T=1代表处理组、T=0代表对照组）的协变量之间通常是具有可比性的；而在观察性研究中，处理变量两个水平组之间协变量的可比性通常较差。一旦两组的协变量之间出现严重的不均衡性，若直接基于这样的资料进行统计分析，其结果的可信度就会大打折扣。

1.3 倾向性评分分析法

消除非随机对照试验资料或观察性研究资料中处理变量两个水平组协变量之间不均衡性的一种有效方法为倾向性评分分析法。该方法的基本思想是使试验组与对照组中的受试对象形成配对组，处在同一个配对组中的受试对象在所有协变量上的取值尽可能接近。形成配对组的依据是个体进入相应组的概率相等或差距在事先确定的一个很小的区间内。

具体做法：将处理变量视为结果变量，将所有需要纳入分析的协变量视为自变量，构建关于“T=1”的二值结果变量的多重Logistic回归模型，计算数据集中每个个体的发生概率$P_i(i=1,2,\cdots,n)$。当个体属于处理组时，保持P_i不变；当个体属于对照组时，取其概率为$1-P_i$。有鉴于此，可理解Rosenbaum和Rubin对倾向性评分的定义，即根据一组观察到的基线协变量分配处理的概率[3]。

2 匹配方法

2.1 匹配的必要性

在随机对照试验研究中，人们普遍接受这样的假定：即任何一个个体都以同等的机会进入试验组或对照组。由基本常识可知，当样本含量足够大时，被随机分配到处理组或对照组中的个体在所有协变量上的取值是基本均衡的、可比的。这就意味着具有相同协变量向量的个体被分配进入处理组或对照组的概率是基本相同的，由此可知，将处理组与对照组中具有相同概率或概率值近似相等的个体配成对子，就意味着这些个体在协变量上的取值接近。基于这一思路形成的匹配后的新数据集，就类似于前瞻性随机对照试验所得到的数据集。值得一提的是，匹配会损失掉一些信息，导致匹配后新数据集的样本含量明显小于原始数据集。

SAS/STAT的proc psmatch过程中介绍了三种匹配方法，即贪婪的最近邻匹配、替换匹配和最优匹配。当用户指定match语句时，proc psmatch过程使用下文中详细描述的三种匹配方法之一，将对照组与处理组的观察个体进行匹配[3,5]。

2.2 三种匹配方法

2.2.1 贪婪的最近邻匹配

贪婪的最近邻匹配由method=greedy选项请求，选择倾向性评分与每个处理单元（即处理组中的个体）的倾向性评分最匹配的控制单元（即对照组中的个体）。贪婪的最近邻匹配是按顺序进行匹配，不进行替换。以下条件可用于贪婪最近邻匹配：①与每个处理单元匹配的控制单元数量（用户可以在K=子选项中指定此数量）；②处理单元倾向性评分的顺序，可以是升序、降序或随机（用户可以在order=子选项中指定顺序）。

2.2.2 替换匹配

替换匹配由method=replace选项请求，通过替换选择倾向性评分最接近每个处理单元倾向性评分的控制单元。用户可以在K=子选项中指定要与每个处理单元匹配的控制单元数。

2.2.3 最优匹配

最优匹配同时选择所有匹配项，不进行替换，以最小化所有匹配项中倾向性评分的绝对总差异。用户可以请求以下最优匹配方法：①固定比率匹配，由method=optimal选项请求，将固定数量的控制单元与每个处理单元相匹配；②可变比率匹配，由method=varratio选项请求，将一个或多个控制单元与每个处理单元相匹配；③完全匹配，由method=full

选项请求，将每个处理单元与一个或多个控制单元匹配，或将每个控制单元与一个或多个处理单元匹配，通过另外指定 kmean=、ncontrol=或 pctcontrol=subsoptions，用户可以请求约束完全匹配，其中匹配的控制单元数小于可用的控制单元的总数。

通常，用户按倾向性评分值进行匹配，但用户还可以依据基于倾向性评分算得的 logit 值匹配，或者依据马氏距离来匹配。

3 实例与 SAS 实现

3.1 问题与数据结构

【例 1】研究者分析戒烟对个体体重的影响。本例数据是 Hernán 和 Robins 的 NHANES I 流行病学随访研究（NHEFS）数据的子集。在这项研究中，医疗和行为信息是在最初的身体检查期间收集的，大约十年后再次在后续访谈中收集[3]。数据集 SmokingWeight（样本含量 n=1 746）中部分数据显示在下面的数据步程序中，因篇幅所限，详细数据见文献[3]。

```
data SmokingWeight;
input Sex Age Race Education Exercise BaseWeight
Weight Change Activity YearsSmoke PerDay Quit;
datalines;
0 42 1 1 2 79.04 68.95 -10.09 0 29 30 0
0 36 0 2 0 58.63 61.23 2.60 0 24 20 0
1 56 1 2 2 56.81 66.22 9.41 0 26 20 0
0 68 1 1 2 59.42 64.41 4.99 1 53 3 0
0 40 0 2 1 87.09 92.08 4.99 1 19 20 0
……（此处省略了 1 735 行数据）
0 45 0 1 0 63.05 64.41 1.36 0 29 40 0
1 47 0 1 0 57.72 61.23 3.51 0 31 20 0
1 51 0 3 0 62.71 . . 0 30 40 0
0 68 0 1 1 52.39 57.15 4.76 1 46 15 0
0 26 0 . 0 86.75 87.54 0.79 0 9 20 0
0 29 0 2 1 90.83 106.59 15.76 1 14 30 1
;
run;
```

数据中的变量含义如下，Activity：日常活动水平，取值分别为 0、1、2；Age：1971 年时的年龄；BaseWeight：1971 年时的体重（kg）；Change：随访和基线访谈中的体重差异（kg）；Education：教育水平，取值分别为 0、1、2、3、4；Exercise：定期休闲运动量，取值分别为 0、1、2；PerDay：1971 年每天吸烟数量；Quit：如果个体在初次访谈和后续访谈之间戒烟，取值为 1，否则，取值为 0；Race：白人取值为 0，非白人取值为 1；Sex：男性取值为 0，女性取值为 1；Weight：随访时的体重（kg）；YearsSmoke：吸烟的年数。

试构建因变量 Change 关于全部自变量的多重线性回归模型，着重考查处理变量 Quit 对因变量的影响程度和方向。

3.2 原始数据集中处理组与对照组协变量之间可比性比较

3.2.1 将 10 个协变量按性质划分成两组

已知变量 Quit 为处理变量，Quit=1 与 Quit=0 分别代表处理组与对照组；变量 Change 为结果变量，其他 10 个变量（Sex、Age、Race、Education、Exercise、BaseWeight、Weight、Activity、YearsSmoke、PerDay）均为协变量。由于 Change=Weight−BaseWeight，故应该将 Weight 和 BaseWeight 两个变量从协变量的集合中删除。为了显示 8 个协变量在处理组与对照组中的可比性，需要将它们分成两组：第一组为定性的协变量，包括 Sex、Race、Education、Exercise、Activity；第二组为定量的协变量，包括 Age、YearsSmoke、PerDay。

3.2.2 比较 5 个定性协变量在两组之间的差别

设所需要的 SAS 程序如下：

```
proc freq data=smokingweight;
tables quit*(Sex Race Education Exercise Activity)/chisq;
run;
```

【SAS 主要输出结果及解释】两组性别（χ^2=8.810，P=0.003）、种族（χ^2=6.820，P=0.009）、教育水平（χ^2=12.678，P=0.013）构成之间的差别均有统计学意义；两组运动量（χ^2=4.490，P=0.106）、日常活动水平（χ^2=2.681，P=0.262）构成之间的差别均无统计学意义。由此可知：有 3 个定性协变量在两组之间的差异具有统计学意义。

3.2.3 比较 3 个定量协变量在两组之间的差别

设所需要的 SAS 程序如下：

```
proc ttest data=smokingweight;
class quit;
var Age YearsSmoke PerDay;
```

run;

【SAS 主要输出结果及解释】两组年龄(t=5.530,P<0.01)、吸烟的年数(t=3.360,P=0.001)、1971 年每天吸烟数量(t=3.990,P<0.01)均值之间的差异均有统计学意义。由此可知:3 个定量协变量在两组之间的差异均有有统计学意义。

3.2.4 两组协变量可比性的总结

由“第 3.2.2 节”和“第 3.2.3 节”的分析结果可知,本例资料中处理组(Quit=1)与对照组(Quit=0)之间有 6 个协变量不具有可比性。因此,基于这样的资料构建多重线性回归模型,所得结果的可信度不高。

3.3 对原始数据集采用传统的多重线性回归分析

定量的结果变量为 Change,着重考查的协变量 Quit 为处理变量,其他 8 个变量(Sex、Race、Education、Exercise、Activity、Age、YearsSmoke、PerDay)为一般协变量,构建 Change 关于 9 个变量的多重线性回归模型。设所需要的 SAS 程序如下:

```
proc reg data=smokingweight;
model Change=quit Sex Race Education Exercise Activity Age YearsSmoke PerDay
/selection=stepwise sle=0.50 sls=0.05;
run;
```

【SAS 主要输出结果及解释】由对整个回归模型的假设检验结果可知,回归模型有统计学意义(F=37.060,P<0.01);复相关系数平方的数值比较小,即 R^2=0.087。回归模型中各参数的假设检验结果见表 1。

表 1 回归模型中各参数的假设检验结果

Table 1 Hypothesis test results of the parameters in the regression model

变 量	参数估计	标准误差	II 型 SS	F	Pr>F
Intercept	10.190	0.844	8 287.976	145.780	<0.010
Quit	3.193	0.441	2 985.719	52.520	<0.010
Activity	−0.702	0.296	319.971	5.630	0.018
Age	−0.220	0.032	2 682.197	47.180	<0.010
YearsSmoke	0.069	0.032	267.020	4.700	0.030

以上是回归模型中各参数的假设检验结果,截距项和 4 个协变量与 0 之间的差异均有统计学意义。值得一提的是:Quit 的回归系数为 3.193>0,表明戒烟使受试者的体重平均增加 3.193 kg。

3.4 采用 proc psmach 过程进行倾向性评分分析

3.4.1 调用倾向性评分过程

倾向性评分分析的目的是使所有的协变量在处理组与对照组之间的差别尽可能缩小,程序输出结果主要是匹配后的数据集。设所需要的 SAS 程序如下:

```
ods graphics on;
proc psmatch data=smokingweight region=treated;
class quit Sex Race Education Exercise Activity;
psmodel quit(Treated='1')=Sex Race Education Exercise Activity
Age YearsSmoke PerDay;
match distance=lps method=greedy(k=1) exact=Sex caliper=0.5
weight=none;
assess lps var=(Sex);
output out(obs=match)=OutEx4 matchid=_MatchID;
run;
```

3.4.2 基于倾向性评分分析的结果进行后续分析

倾向性评分分析的结果储存在输出数据集 OuttEx4 中,主要内容是在原始数据集中增加了一个新变量“MatchID”,即匹配变量。比较匹配后的数据集中协变量在处理组与对照组之间的差别。设所需要的 SAS 程序如下:

```
data abc;
set OutEx4;
if _MatchID='.' then delete;
run;
proc freq data=abc;
tables quit*(Sex Race Education Exercise Activity)/chisq;
run;
proc ttest data=abc;
class quit;
var Age YearsSmoke PerDay;
run;
```

【SAS 主要输出结果及解释】两组性别(χ^2=0.000,P=1.000)、种族(χ^2=0.014,P=0.905)、教育水平(χ^2=0.521,P=0.971)、运动量(χ^2=0.057,P=0.972)、日常活动水平(χ^2=0.327,P=0.849)构成之间的差异均无统计学意义。两组年龄(t=−0.150,

P=0.879）、吸烟的年数（t=0.280，P=0.783）、1971 年每天吸烟数量（t=0.230，P=0.821）均值之间的差异均无统计学意义。

由此可知：8 个协变量在两组之间的差异均无统计学意义。值得一提的是，本例原始数据集的 8 个协变量中有 6 个协变量在两组之间的差异有统计学意义（见前文第 3.2.4 节）。这说明采用倾向性评分分析方法进行匹配后获得的数据集，所考查的协变量在处理组与对照组之间均达到了均衡性要求。

3.5 基于匹配后的数据集再拟合多重线性回归模型

3.5.1 直接基于匹配后的数据集建模

设所需要的 SAS 程序如下：

```
proc reg data=abc;
model Change=quit Sex Race Education Exercise Activity
Age YearsSmoke PerDay
/selection=stepwise sle=0.50 sls=0.05;
run;
```

【SAS 主要输出结果及解释】由对整个回归模型的假设检验结果可知，回归模型有统计学意义（F=33.450，P<0.01）；复相关系数平方的数值 R^2=0.110 较之前的 R^2=0.087 略有增大。回归模型中各参数的假设检验结果见表 2。

表 2 回归模型中各参数的假设检验结果

Table 2 Hypothesis test results of the parameters in the regression model

变 量	参数估计	标准误差	II 型 SS	F	Pr>F
Intercept	10.194	1.184	4 997.083	74.180	<0.010
Quit	3.346	0.576	2 276.174	33.790	<0.010
Activity	−0.953	0.446	307.706	4.570	0.033
Age	−0.180	0.024	3 869.537	57.440	<0.010

以上是回归模型中各参数的假设检验结果，截距项和 3 个协变量与 0 之间的差异均有统计学意义。值得一提的是：Quit 的回归系数为 3.346>0，表明戒烟使受试者的体重平均增加 3.346 kg。

3.5.2 在匹配后的数据集中引入派生自变量后再建模

引入 3 个派生自变量（age1、age2、age3）后重新构建多重线性回归模型[6-7]，设所需要的 SAS 程序如下：

```
data a1;
set abc;
age1=age*quit;age2=age*activity;
age3=age*age;
proc reg data=a1;
model Change=quit Sex Race Education Exercise Activity
Age YearsSmoke PerDay age1 age2 age3
/noint selection=stepwise sle=0.50 sls=0.05;
run;
```

【SAS 主要输出结果及解释】由对整个回归模型的假设检验结果可知，回归模型有统计学意义（F=55.260，P<0.01）；复相关系数平方的数值 R^2=0.214 较之前的 R^2=0.110 有较大增加。回归模型中各参数的假设检验结果见表 3。

表 3 回归模型中各参数的假设检验结果

Table 3 Hypothesis test results of the parameters in the regression model

变 量	参数估计	标准误差	II 型 SS	F	Pr>F
Quit	3.269	0.569	2 184.033	33.080	<0.010
Age	0.288	0.031	5 606.896	84.920	<0.010
age2	−0.022	0.009	361.887	5.480	0.020
age3	−0.005	0.001	5 301.123	80.290	<0.010

以上是回归模型中各参数的假设检验结果，无截距项，4 个协变量与 0 之间的差异均有统计学意义。值得一提的是：Quit 的回归系数为 3.269>0，表明戒烟使受试者的体重平均增加了 3.269 kg。

3.6 结论

由“第 3.2 节”的比较结果可知，原始数据集中处理组与对照组协变量之间的可比性较差。因此，“第 3.3 节”基于原始数据集构建多重线性回归模型所得结果和结论的可信度不高。由“第 3.4 节”的比较结果可知，匹配后的数据集中处理组与对照组协变量之间的可比性很好。因此，“第 3.5 节”基于匹配后的数据集构建多重线性回归模型所得结果和结论的可信度有了较大提高（模型的复相关系数的平方由 R^2=0.087 提高到 R^2=0.214）。

4 讨论与小结

4.1 讨论

倾向性评分分析法是对非随机试验资料或调查研究资料进行预处理的有效方法，它特别适用于具有一个二水平处理变量的多因素资料。基于匹配后的数据集，可以采用多重线性回归分析（要求结果变量为计量变量）或多重 Poisson 回归分析（要求结果变量为计数变量）或多重 Logistic 回归分析（要求

结果变量为定性变量)或多重 Cox 比例风险回归分析(要求结果变量为生存时间变量)。

匹配后的数据集样本含量的大小取决于所采用的匹配方法,一般来说,匹配方法越严格,匹配后的数据集样本含量就越小。就本例而言,原始数据集的样本含量 n=1 746,采用贪婪的最近邻匹配法得到的匹配后数据集样本含量 n=854,采用替换匹配法得到的匹配后数据集样本含量 n=741,采用最优匹配法得到的匹配后数据集样本含量 n=1 498。值得一提的是,基于不同的匹配后数据集进行多重线性回归分析,其结果可能会不同。

4.2 小结

本文介绍了与倾向性评分分析有关的基本概念、三种匹配方法以及基于一个实例的 SAS 实现。基本概念的内容涉及处理变量、处理变量两水平组之间协变量的可比性和倾向性评分分析法;三种匹配方法分别是贪婪的最近邻匹配、替换匹配和最优匹配;运用 SAS 软件对一个实例进行了全面分析,包括对原始数据集基本情况的描述、基于倾向性评分分析产生匹配后的数据集、分别对原始数据集和匹配后的数据集进行多重线性回归分析,并对基于不同数据集所得到的分析结果进行比较和总结。

参考文献

[1] 赵文, 雷于佳, 翟瑞, 等. 重大公共危机事件下四川省在校学生安全感与焦虑状况[J]. 四川精神卫生, 2022, 35(1): 62-65.
Zhao W, Lei YJ, Zhai R, et al. Analysis of sense of security and anxiety of students in Sichuan Province under major public crisis [J]. Sichuan Mental Health, 2022, 35(1): 62-65.

[2] 沈雪梅, 杜春红, 王欢. 绵阳市区成人睡眠质量现状及影响因素[J]. 四川精神卫生, 2022, 35(1): 66-69.
Shen XM, Du CH, Wang H. Current status and influencing factors of sleep quality among adult residents in urban areas of Mianyang[J]. Sichuan Mental Health, 2022, 35(1): 66-69.

[3] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 2301-2364, 2365-2424, 8093-8214.

[4] 邓伟, 贺佳. 临床试验设计与统计分析[M]. 北京: 人民卫生出版社, 2012: 237-241.
Deng W, He J. Design and statistical analysis of clinical trials [M]. Beijing: People's Medical Publishing House, 2012: 237-241.

[5] 胡安宁. 应用统计因果推论[M]. 上海: 复旦大学出版社, 2020: 94-110.
Hu AN. Applied statistical causal inference [M]. Shanghai: Fudan University Press, 2020: 94-110.

[6] 谷恒明, 胡良平. 基于经典统计思想实现多重线性回归分析[J]. 四川精神卫生, 2018, 31(1): 7-11.
Gu HM, Hu LP. Realization of a multiple linear regression analysis based on the classical statistical thought [J]. Sichuan Mental Health, 2018, 31(1): 7-11.

[7] Armitage P, Colton T. Encyclopedia of biostatistics [M]. 2nd edition. New York: John Wiley & Sons, 2005: 4626-4630.

(收稿日期:2022-11-13)
(本文编辑:戴浩然)