简单线性回归分析及其应用

谷恒明[1]，胡良平[1,2*]

（1.军事医学科学院生物医学统计学咨询中心，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：lphu812@sina.com）

【摘要】 本文目的是介绍简单线性回归分析的前提条件、种类、实现计算的SAS程序及结果解释，并说明数据是否值得做直线回归分析以及如何选择正确的直线回归分析类型。简单线性回归分析有三种具体情形，分别是：简单直线回归分析、加权直线回归分析和具有重复试验的直线回归分析。进一步通过实例来阐述如何进行不同的简单线性回归分析，并给出实现这些直线回归分析所需要的SAS程序及输出结果。

【关键词】 简单线性回归分析；SAS程序；加权直线回归分析；重复试验的线性回归分析
中图分类号：R195.1 文献标识码：A doi：10.11886/j.issn.1007-3256.2017.06.002

Simple linear regression analysis and its application

Gu Hengming[1], *Hu Liangping*[1,2*]

(1. *Consulting Center of Biomedical Statistics, Academy of Military Medical Sciences, 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: lphu812@sina.com*)

【**Abstract**】 The aim of this article is to introduce the preconditions, categories, SAS programs and the results interpretation of the simple linear regression analysis to illustrate how to choose the correct regression model and whether the data is worth regression analysis. There are three kinds of simple linear regression analyses: simple linear regression analysis, weighted linear regression analysis and repeated experimental linear regression analysis. The following examples are used to illustrate different simple linear regression analyses and the corresponding SAS programs required to perform these linear regression analyses and their results.

【**Keywords**】 Simple linear regression analysis; SAS Program; Weighted linear regression analysis; Repeated experimental linear regression analysis

1 概 述

简单线性回归分析是研究两定量变量之间依存变化关系的一种最常用最简单的方法。如何正确实现简单线性回归分析，需要考察以下两组前提条件。

第一组前提条件，即从数理统计学角度考量所归纳出来的前提条件[1]：①自变量X可以是普通变量，也可以是随机变量，但因变量Y必须是随机变量；②线性，即因变量Y与自变量X之间的关系为线性关系，在直角坐标系内绘制关于X与Y的散点图，可以看出线性关系；③独立性，即各个观察对象之间必须是相互独立的；④正态性，即给定X的取值后，Y服从正态分布；⑤等方差性，即不同X值对应的Y的分布具有相同的方差。

第二组前提条件，即从基本常识角度考量所归纳出来的前提条件：①对于两个定量变量而言，所有受试对象应具有同质性；②所研究的两个定量变量在专业上应具有一定的联系；③在直角坐标系中绘制(X,Y)的全部散点，全部散点应呈现直线变化趋势；④散点图上不存在下列两类可疑的异常点，第一类，在垂直于横坐标轴方向上的可疑异常点，第二类，在假定的理想直线的左右两端的延长线方向上的可疑异常点。

事实上，上述的第二组前提条件更有实用价值，它也是进行简单直线回归分析的基本步骤。在此基础上，再计算直线回归方程中的参数并对参数进行假设检验；最后，再将所获得的简单直线回归方程用于“预测（给定自变量的数值去计算因变量的取值）”或“控制（给定因变量的取值去估算自变量的变化范围）”。

2 简单直线回归分析

简单直线回归模型为：

$$y=\alpha+\beta x+\varepsilon \quad (1)$$

简单线性回归分析的任务：其一，估计式(1)中参数α和β的数值；其二，假设检验，包括对截距、斜率和整个直线回归方程的检验。

【例1】研究20名儿童的血红蛋白(y)与血铁

项目基金：国家高技术研究发展计划课题资助(2015AA020102)

(x)之间的关系[2],数据见表 1。

表 1　20 名儿童的血红蛋白(y)与血铁(x)的测定资料

n	y(mg/dL)	x(ug/dL)	*n*	y(mg/dL)	x(ug/dL)
1	13.5	518.7	11	10.2	409.8
2	13	467.3	12	10	384.1
3	11	469.8	13	9.5	356.3
4	14.3	456.6	14	9.4	388.6
5	12.5	448.7	15	8.8	325.9
6	12.5	424.1	16	6.3	292.8
7	11.8	405.6	17	7.3	332.8
8	11.5	446	18	7.8	283
9	11	416.7	19	7.3	312.5
10	10.7	430.8	20	7	294.7

【分析与解答】对表 1 资料进行简单直线回归分析所需要的 SAS 程序如下:

```
data jz; input n y x@@; cards;
1 13.5 518.7 2 13 467.3 3 11 469.8
4 14.3 456.6 5 12.5 448.7 6 12.5 424.1
7 11.8 405.6 8 11.5 446 9 11 416.7
10 10.7 430.8 11 10.2 409.8 12 10 384.1
13 9.5 356.3 14 9.4 388.6 15 8.8 325.9
16 6.3 292.8 17 7.3 332.8 18 7.8 283
19 7.3 312.5 20 7 294.7
; run;
proc gplot data = jz; plot x * y = 's'; run;
proc reg data = jz; model y = x/noint; run;
```

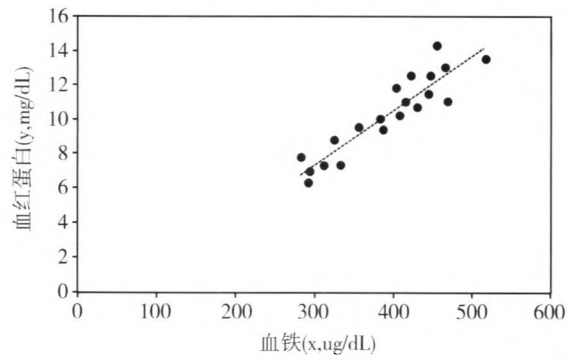

图 1　20 名儿童的血红蛋白(y,mg/dL)与血铁(x,ug/dL)的散点图

简单直线回归分析的假设检验结果见表 2、表 3。

表 2　方差分析

Analysis of Variance					
Source	DF	SumofSquares	MeanSquare	F Value	Pr > F
Model	1	2194.08323	2194.08323	2273.44	<0.0001
Error	19	18.33677	0.96509		
Uncorrected Total	20	2212.42000			

表 3　参数估计

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
x	x	1	0.02626	0.00055080	47.68	<0.0001

由图 1 中全部散点的分布情况可以认为两定量变量之间存在线性变化趋势。y 依 x 变化的直线关系是否成立,其结果见表 2、表 3:经方差分析,得 $F=2273.44$,$P<0.0001$,说明该直线回归方程有统计学意义;在参数估计中,由于截距项(intercept)与 0 之间差异无统计学意义,$P>0.05$,将 SAS 程序中的model语句由 y = x 改为 y = x/noint,选项“noint”的作用是拟合不含截距项的简单直线回归方程。最终的结果为:斜率与 0 之间的差异有统计学意义,$t=47.68$,$P<0.0001$,因此,求得的直线回归方程为 $\hat{Y}=0.02626x$。

3　加权直线回归分析

医学或药学试验中经常需要计算引起试验动物总体中半数动物产生某种反应所需的药物(或毒物)剂量,即半数有效量,需要使用到加权直线回归分析[1]。

【例 2】SAS 9.3 帮助文档中 Probit 过程中第一个例子,研究不同剂量药物下小鼠反应数。数据见表 4。

表 4 不同剂量药物下小鼠反应数

dose	n	response	dose	n	response
1	10	1	5	12	8
2	12	2	6	10	8
3	10	4	7	10	10
4	10	5	.	.	.

注:dose 代表剂量,n 代表每个剂量组的动物数,response 代表每个剂量组的阳性反应动物数

【分析与解答】对表 4 资料进行加权直线回归分析所需要的 SAS 程序如下:

```
data a; input Dose N Response @@ ; datalines;
1 10 1 2 12 2 3 10 4
4 10 5 5 12 8 6 10 8
7 10 10
;
run;
Proc probit log10 data = a; model Response/N = Dose / lackfit inversecl itprint;
output out = B p = Prob std = std xbeta = xbeta; run;
```

参数估计结果见表 5。

表 5 参数估计

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi - Square	Pr > ChiSq
Intercept	1	-1.8127	0.4493	-2.6934	-0.9320	16.27	<.0001
Log10(Dose)	1	3.4181	0.7455	1.9569	4.8794	21.02	<.0001

所求得该药物的半数反应剂量为 3.39096。见表 6。

表 6 半数反应剂量

Probit Analysis on Dose			
Probability	Dose	95% Fiducial Limits	
0.50	3.39096	2.61175	4.27138

4 具有重复试验的直线回归分析

在同一试验条件下进行多次重复试验,研究因变量与自变量之间是否存在直线关系时可以用具有重复试验的直线回归分析。

具有重复试验的直线回归分析与无重复试验的直线回归分析的区别在于:前者可以对“失拟(即直线回归方程所不能解释的那部分信息)”进行假设检验,仅当“失拟”的检验结果无统计学意义时,可将其视为无重复试验的情形,但试验点数为不同 X 值个数乘以重复试验次数(各 X 值对应的重复试验次数相等);否则,应选择合适的曲线类型,进行曲线回归分析。

【例 3】研究不同血液浓度与血红蛋白含量之间的关系[1]。数据见表 7。

表 7 不同血液浓度下血红蛋白的测定值

血液浓度 X(%)	血红蛋白测定值 Y(g/dL)		
	1	2	3
10	3.2	3.1	3.3
20	6.2	6.2	6.2
30	9.2	9.3	9.2
40	12.3	12.4	12.2
50	15.6	15.2	15.4
60	18.3	18.2	18.3
70	21.1	21.3	21.3
80	23.9	23.8	23.7
90	26.5	26.4	26.4
100	29.0	29.1	28.9

【分析与解答】对表 7 资料进行具有重复试验的直线回归分析所需要的 SAS 程序如下:

```
data b; input x n@@ ; g = _n_; do i = 1 to n; input y @@ ; output; end; cards;
10 3 3.2 3.1 3.3 20 3 6.2 6.2 6.2
30 3 9.2 9.3 9.2 40 3 12.3 12.4 12.2
50 3 15.6 15.2 15.4 60 3 18.3 18.2 18.3
70 3 21.1 21.3 21.3 80 3 23.9 23.8 23.7
90 3 26.5 26.4 26.4 100 3 29.0 29.1 28.9
```

```
;
run;
proc glm data = b; class g;model y = x g/ss1; run;
proc reg data = b; model y = x; run;
```

具有重复试验的直线回归分析较简单直线回归分析多了失拟检验，目的是考察仅采用直线回归方程是否可以较好地拟合给定的资料。失拟检验的结果见表 8。

表 8 本例资料的失拟检验结果

源	自由度	I 型 平方和	均方	*F* 值	*Pr* > *F*
x	1	2062.667677	2062.667677	199613	<0.0001
g	8	2.444323	0.305540	29.57	<0.0001

在表 8 中，只需看最后一行，对分组变量 g（它在本质上就是自变量不同取值的个数）进行检验即可，这里 $F = 29.57$，$P < 0.0001$，说明失拟平方和基本上是由模型分组因素造成，也就是说，该资料未通过失拟检验，不能直接拟合简单直线回归方程，而需要根据散点图中全部散点的分布趋势和形态，选定合适的曲线类型并拟合之。

究竟如何进一步处理此资料，请读者阅读完本期中的下一篇文章《简单曲线回归分析及其应用》后，自己去把它完成。提示：若采用二次抛物线、对数函数曲线、指数函数曲线或幂函数曲线来分别拟合本例资料，从模型的假设检验的 *F* 值和复相关系数平方（即 R^2）的数值越大越好以及残差图中散点分布情况（全部散点在残差为 0 的水平线上下随机波动且无明显变化趋势为好）来全面考量，相对来说，本例资料拟合二次抛物线最好。

参考文献

[1] 胡良平. 科研设计与统计分析[M]. 北京：军事医学科学出版社，2012:381-400.

[2] 徐天和，柳青. 中国医学统计百科全书 多元统计分册[M]. 北京：人民卫生出版社，2004：2.

（收稿日期:2017-12-03）

（本文编辑:陈 霞）