简单曲线回归分析及其应用

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】 本文目的是介绍可以直线化的曲线回归分析相关内容及如何使用 SAS 软件来实现。一般来说，采用回归分析研究专业上确实存在联系的两个定量变量之间的依存关系。如果两定量变量之间是直线关系，那么直接采用直线回归分析即可；但在医学实验中，两定量变量之间的关系常常不是直线关系而是曲线关系，此时就应采用曲线回归分析。本文重点讲述可以直线化的曲线回归分析的种类及其 SAS 软件实现。

【关键词】 回归分析；曲线拟合；SAS 软件；曲线直线化
中图分类号：R195.1 文献标识码：A doi：10.11886/j.issn.1007-3256.2017.06.003

Simple curve 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 paper is to introduce how to fit a curve regression equation by the variable transformation and how to perform it by using SAS software. In general, the regression analysis should be applied when there is the relationship between two quantitative variables in profession. If the two variables are linear, then the linear regression analysis can be used directly. However, in medical experiments, the relationship between the two quantitative variables is often not linear, so it is necessary to use curve regression analysis. This article focuses on fitting curve by variable transformation and the corresponding SAS software operation.

【**Keywords**】 Regression analysis; Curve fitting; SAS software; Curve linearization

1 概 述

可直线化的曲线回归分析一般是通过变量变换的方法[1]，将原本是曲线关系的两个定量变量转化为直线关系，再对新变量进行简单线性回归分析得到直线回归方程，最后再回代到原始变量。此方法的关键是找到原始变量的合理变换方法，不恰当的变量变换只会产生错误的结果。

曲线回归分析的步骤：①在直角坐标系内绘制两个定量变量的散点图；②根据散点图全部散点的变化趋势，判断合适的曲线类型；③根据所选的曲线类型，进行变量变换，使变换后的两定量变量之间呈直线变化趋势；④对两个新变量建立直线回归方程，并作假设检验；⑤还原到初始变量，得到曲线回归方程；⑥若同一资料适合多种曲线类型，需要进行曲线拟合优度检验（当自变量只以一次项出现在回归方程中时，也可直接比较对整个回归方程所做的假设检验对应的 F 统计量，大者为优；也可看 R^2，大者为优），比较其差异；⑦选择拟合最好的曲线回归方程，并从专业角度上判断其是否成立。

2 二项式曲线回归分析

当因变量与自变量不是简单的一阶关系，而是与自变量的二阶甚至高阶存在线性关系时，就需要使用多项式回归分析方法。本文介绍因变量与自变量的二阶存在线性关系的曲线拟合问题。散点图一般呈抛物线形状，因此，二次多项式曲线亦称二次抛物线。

【例 1】研究某氧化酶活性与 pH 值之间的关系[2]，数据见表 1。

表 1 酶活性（y）与 pH 值（x）数据

id	pH 值（x）	酶活性（y）
1	6.0	2015
2	6.3	2520
3	6.6	3498
4	6.9	3675
5	7.2	3785
6	7.5	3624
7	7.8	3165
8	8.1	2516
9	8.4	2128

项目基金：国家高技术研究发展计划课题资助（2015AA020102）

【分析与解答】试采用二次抛物线函数来拟合表 1 资料，其所需要的 SAS 程序如下：

```
data pwx; input id x y@@; x2 = x * x; cards;
 1  6.0  2015  2  6.3  2520  3  6.6  3498
 4  6.9  3675  5  7.2  3785  6  7.5  3624
 7  7.8  3165  8  8.1  2516  9  8.4  2128
;
run;
axis1 label = ('PH 值(x)') order = (5.5 to 8.5 by 0.5);
axis2 label = (angle = 90 '(酶活性(y))') order = (1800 to 4000 by 200) minor = none offset = (0.5, 0.5) major = (height = 0.8);
symbol color = black interpol = join value = circle width = 1.5;
proc gplot data = pwx;
plot y * x/haxis = axis1 vaxis = axis2;
run;
proc reg data = pwx; model y = x x2; run;
```

以上程序可分为三部分：第一部分创建临时 SAS 数据集；第二部分绘制散点图；第三部分构建二次抛物线回归方程并进行假设检验。散点图显示，两定量变量之间呈二次抛物线变化趋势。见图 1。

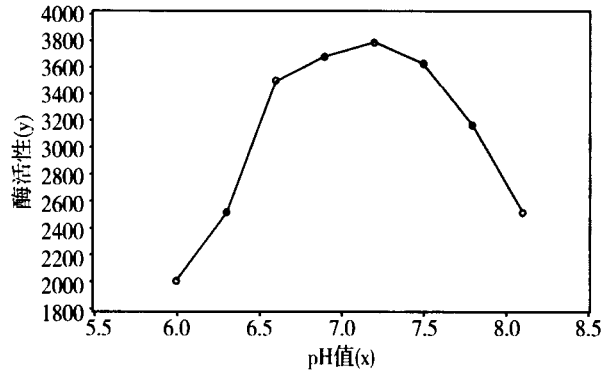

图 1 酶活性(y)与 pH 值(x)之间的散点图

对模型进行假设检验的结果见表 2。

表 2 方差分析

Analysis of Variance					
Source	DF	SumofSquares	MeanSquare	F Value	Pr > F
Model	2	3742099	1871049	59.39	<.0001
Error	6	189013	31502		
Uncorrected Total	8	3931112			

对模型中各参数进行估计和假设检验的结果见表 3。

表 3 参数估计

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-59642	5784.36798	-10.31	<.0001
x	1	17618	1619.93655	10.88	<.0001
x2	1	-1224.51900	112.37048	-10.90	<.0001

假设检验和参数估计的结果见表 2、表 3，在方差分析表中，$F = 59.39$，$P < 0.0001$，说明经直线化变换后的直线回归方程有统计学意义，因此，所求得的二次抛物线回归方程为 $\hat{Y} = -59642 + 17618x - 1224.52x^2$。

3 双曲线形式的曲线回归分析

当因变量与自变量的关系不是直线，而是曲线时，对变量进行适当变换，使曲线直线化。

【例 2】资料来源于《中国卫生统计》的一篇文章，研究钩虫病患者治疗次数与复查阳性率之间的变化规律。见表 4。

【分析与解答】在例 1 中，已给出绘制散点图的程序，读者自行修改即可，此处就不重复了。绘制的散点图见图 2。

表 4 钩虫病患者治疗次数(x)与复查阳性率(y)数据

x	1	2	3	4	5	6	7	8
y	63.9	36.0	17.1	10.5	7.3	4.5	2.8	1.7

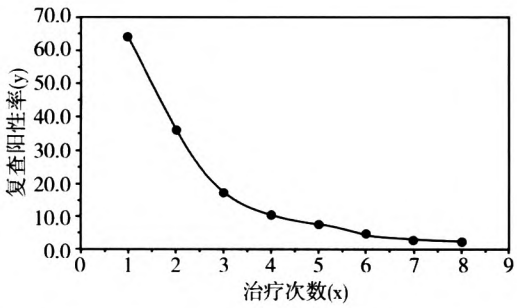

图 2 钩虫病患者治疗次数(x)与复查阳性率(y)散点图

如图 2 所示,资料的散点图不呈直线变化,钩虫阳性率随着治疗次数越多阳性率越小,最后钩虫阳性率趋近于 0。可以对自变量 x 进行倒数变换,重新拟合直线回归方程。

对自变量 x 进行倒数变换后的散点图(在例 1 中,已给出绘制散点图的程序,读者自行修改即可,此处就不重复了)见图 3。

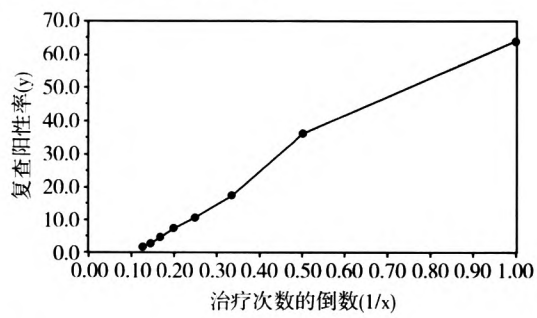

图 3 钩虫病患者治疗次数的倒数(1/x)与复查阳性率(y)散点图

由图 3 可知,基本上实现了曲线直线化。接下来可以拟合双曲线回归方程。

所需的 SAS 程序如下:

```
data gouchong; input x y; x1 = 1/x; x2 = log(x);
y1 = log(y); cards;
 1   63.9   2   36.0   3   17.1   4   10.5
 5    7.3   6    4.5   7    2.8   8    1.7
;
run;
proc reg data = gouchong; model y = x1; run;
```

由 model 语句可知,此处选择了对自变量进行倒数变换的方式。

此处省略了对模型进行假设检验的结果。参数估计结果见表 5。

曲线拟合结果如表 5 所示,模型检验结果 $F = 403.08$, $P < 0.0001$,说明所建立的回归模型有统计学意义,表格略。调整 $R^2 = 0.9829$,由表 5 参数估计结果可得直线回归方程为:$\hat{Y} = -6.88 + 73.16x1$,还原到原始变量,得到曲线回归方程为:$\hat{Y} = -6.88 + \frac{73.16}{x}$。(说明:在使用此回归方程时,自变量 x 不应取 0 值)

表 5 参数估计

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.87959	1.59225	-4.32	0.0050
x1	1	73.15938	3.64399	20.08	<.0001

4 幂函数曲线回归分析

当因变量 y 随着 x 的变化符合幂函数曲线规律时,可以对自变量 x 和因变量 y 同时取对数变换,使幂函数曲线直线化。幂函数的一般形式为:

$y = ax^b + k(a > 0, x > 0)$

当不考虑 k 时,对等号两端同时取对数,

得:$\ln y = \ln a + b\ln x$,即 lny 与 lnx 之间呈直线关系。

【例 3】沿用例 2 的资料,试拟合幂函数曲线回归方程。

【分析与解答】对因变量 y 和自变量 x 都进行对数变换后的散点图(在例 1 中,已给出绘制散点图的程序,读者自行修改即可,此处就不重复了)见图 4。

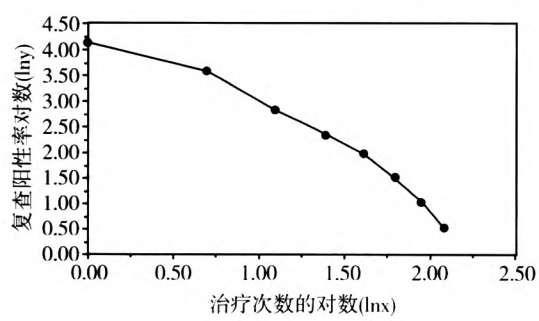

图 4 钩虫病患者治疗次数的对数(lnx)与复查阳性率对数(lny)散点图

由图 4 可知,曲线直线化的效果较好。接下来可以拟合幂函数曲线回归方程。沿用前面的 SAS 数据步程序,现在所需要的 SAS 过程步程序如下:

```
Proc reg data = gouchong; model y1 = x2; run;
```

由 model 语句可知,此处选择了对因变量和自

变量都进行对数变换的方式。

参数估计结果见表6。曲线拟合结果见表6，模型检验结果 $F=115.06$，$P<0.0001$，说明模型有统计学意义，表格略。调整 $R^2=0.9422$，由表6参数估计结果可得直线回归方程为：$\ln y=4.53-1.73x2$，还原到原始变量，得到曲线回归方程为：$\hat{Y}=93.20x^{-1.72508}$。

表6 参数估计

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.53466	0.23800	19.05	<.0001
x2	1	-1.72508	0.16083	-10.73	<.0001

5 指数函数曲线回归分析

当因变量 y 随着 x 的变化符合指数函数曲线规律时，可以对因变量 y 取对数变换，使指数曲线直线化。指数函数的一般形式为：

$y=ae^{bx}+k$ 或 $y=a\exp(bx)+k$

在不考虑 k 时，等号两端同时取对数，得：

$\ln y=\ln a+bx$

如果以 lny 与 x 在直角坐标系内绘制的散点图呈直线变化趋势时，就可以考虑采用指数函数曲线来拟合和解释 y 与 x 之间的关系。

【例4】沿用例2的资料，试拟合指数函数曲线回归方程。

【分析与解答】对因变量 y 进行对数变换后的散点图（在例1中，已给出绘制散点图的程序，读者自行修改即可，此处就不重复了）见图5。

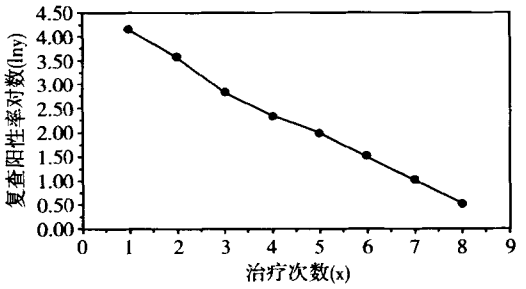

图5 钩虫病患者治疗次数(x)与复查阳性率对数(lny)散点图

由图5可知，曲线直线化的效果较好。接下来可以拟合指数函数曲线回归方程。

沿用前面的 SAS 数据步程序，现在所需要的 SAS 过程步程序如下：

```
Proc reg data = gouchong; model y1 = x; run;
```

由 model 语句可知，此处选择了对因变量进行对数变换的方式。

参数估计结果见表7。

表7 参数估计

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.52605	0.08998	50.30	<.0001
x	1	-0.50625	0.01782	-28.41	<.0001

曲线拟合结果显示，模型检验结果 $F=807.15$，$P<0.0001$，说明模型具有统计学意义，表格略。调整 $R^2=0.9914$，由表7参数估计结果可得直线回归方程为：$\ln y=4.53-0.51x$，还原到原始变量，得到曲线回归方程为：$\hat{Y}=92.43e^{-0.51x}$。

6 对数函数曲线回归分析

当因变量 y 随着 x 的变化符合对数函数曲线规律时，可以对自变量 x 取对数变换，使对数函数曲线直线化。对数函数的一般形式为：

$y=a\ln x+k$

如果以 y 和 lnx 在直角坐标系内绘制的散点图呈直线变化趋势时，就可以考虑采用对数曲线来拟合和解释 y 与 x 之间的关系。

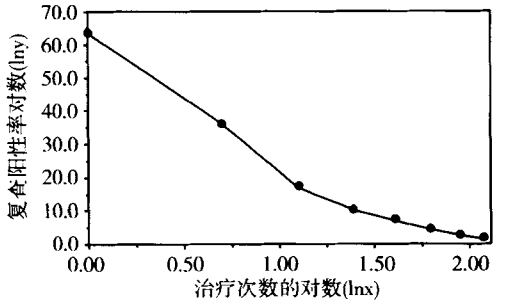

图6 钩虫病患者治疗次数对数(lnx)与复查阳性率(y)散点图

【例5】沿用例2的资料，试拟合对数函数曲线回归方程。

【分析与解答】对自变量 x 进行对数变换后的

散点图(在例 1 中,已给出绘制散点图的程序,读者自行修改即可,此处就不重复了)见图 6。

由图 6 可知,基本上实现了曲线直线化。接下来可以拟合对数函数曲线回归方程。

沿用前面的 SAS 数据步程序,现在所需要的 SAS 过程步程序如下:

```
Proc reg data = gouchong; model y = x2; run;
```

由 model 语句可知,此处选择了对自变量进行对数变换的方式。

表 8 参数估计

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	57.59918	4.43597	12.98	<.0001
X2	1	-29.89206	2.99750	-9.97	<.0001

对模型检验结果 $F = 99.45$,$P < 0.0001$,说明所建立的回归模型有统计学意义,表格略。调整 $R^2 = 0.9336$,由表 8 参数估计结果可得曲线回归方程为:$y = 57.60 - 29.89x2$,还原到原始变量,得到曲线回归方程为:$y = 57.60 - 29.89\ln x$

小结:由本文后四种曲线回归分析可知,有时一组数据可以通过多种变量变换方式得到直线回归方程,此时需要根据各种不同方法拟合的效果来得出最优的变换方式。本文中,指数函数的调整 $R^2 = 0.9914$,是四种曲线类型中最大的,因此应该选择指数函数曲线来进行曲线拟合为宜。

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: 401-426.

[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: 147-149.

[3] 徐勇勇,陈长生,张成岗,等. 曲线拟合中的几个问题[J]. 中国卫生统计,1994,11(2): 58-60.

(收稿日期:2017-12-03)

(本文编辑:陈 霞)