生物统计学（VIII）
第八章：可直线化的非线性回归分析

西安交通大学数学与统计学院

May, 2018

- 直线关系是两个变量间最简单的一种关系，但其常在变量取值范围较小时适用
- 当变量取值范围增大时，变量间就不一定存在直线关系，而是表现出多种多样的曲线关系，根据图示法和直线化法可确定曲线的类型
- 对于倒数函数，幂函数及 logistic 生长曲线等非线性方程，均可以通过数据转换变形为线性方程，通过直线回归的方法进行线性回归方程的推导和回归关系的显著性检验，然后在反转为相应的曲线函数
- 数据转换后的线性回归模型的各种假定和计算公式，仅适用于转换后的线性化数据，不适用于原始数据。
- 曲线回归方程拟合程度的好坏，可用相关指数来进行衡量。

- 直线关系是两个变量间最简单的一种关系。理论上，如果缩小研究取值范围，任意非线性关系都可以用线性关系来近似；
- 但研究范围缩小在使用上不方便，一是不能对变量间关系有一个整体的把握，二是在不同取值范围内还要换用不同的方差，因此在许多情况下考虑非线性关系是必要的。
- 非线性关系广泛存在，可以用来表示两个变量间曲线关系种类很多
  - 有些曲线类型可以通过数据转换而变形为直线形式——直线化（rectification）。
  - 转换后的数据可建立直线回归方程，然后再反转为曲线回归方程（curvilinear regression equation）——可直线化的非线性回归分析。
- 非线性回归可分为已知曲线（公式）类型和未知曲线类型（公式）。

一、曲线类型的确定

- 生物学中变量间的曲线关系通常有倒数函数曲线、指数函数曲线、对数函数曲线、幂函数曲线、S 形曲线等多种形式
- 找出两个变量间恰当的非线性回归方程形式是不易的，需要专业知识；且求得的方程的统计数及由此导出的统计表，不应和有关专业实践有矛盾；可从专业理论中得到解释，并使这些解释严密化和数量化。
- 确定曲线类型是非线性回归分析的关键，通常有几种方法：
  - 图示法：根据所获得的试验资料的自然尺度绘出散点图，然后按照散点图的趋势绘出能够反映他们间变化规律的曲线，与已知曲线比较，找出与之较为相似的曲线图形，该曲线即为选定的类型
  - 直线化法：根据散点图进行直观的比较，选出一种曲线类型，并将原始数据进行转换，将曲线方程直线化，用转换后的数据绘出散点图，若该图形是直线趋势，表明恰当，否则需重新选择。
- 确定曲线类型，需从多方面考虑。曲线配合的好坏，通常以所配曲线与实测点吻合程度的高低来衡量，这取决于离回归平方和 $\sum(y-\hat{y})^2$ 与 y 的总平方和 $\sum(y-\bar{y})^2$ 的比例大小。

相关指数

我们把 1 与离回归平方和 $\sum(y-\hat{y})^2$ 与 y 的总平方和 $\sum(y-\bar{y})^2$ 的比值之差定义为相关指数（correlation exponential），记为 R^2:

$$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

相关指数 R^2 的大小反映了回归曲线拟合度的高低，表示利用曲线回归方程进行估测的可靠程度的高低。

对于同一组实测数据，根据散点图的形状，可用若干相近的曲线进行拟合，同时建立若干曲线回归方程，然后根据 R^2 的大小和生物学知识，选择既符合生物学规律，又有较高拟合度的曲线回归方程来描述两个变量的曲线回归关系。

二、数据变换的方法

在进行非线性回归曲线的直线化时，对原数据进行转换的方法通常有两种

(1) 直接引入新变量，使变量间为直线关系

- 对数函数曲线方程为

$$\hat{y} = a + b \lg x$$

引入新变量 $x' = \lg x$，原方程转换为直线形式，即

$$\hat{y} = a + bx'$$

(2) 方程变换后再引入新变量。这种方法是将原曲线方程进行一定的数学变换后，再引入新变量，从而使变量间的关系变为直线关系

- 幂函数曲线方程为 $\hat{y} = ax^b$，两边取对数，得

$$\lg \hat{y} = \lg a + b \lg x$$

引入新变量 $y' = \lg y, a' = \lg a, x' = \lg x$，原方程转换为直线形式，即

$$\hat{y}' = a' + bx'$$

曲线回归方程	经尺度转换的新变量及参数 y'	x'	a'	直线化的方程
$\hat{y}=\frac{a+bx}{x}$	$y'=yx$			$\hat{y}'=a+bx$
$\hat{y}=\frac{1}{a+bx}$	$y'=1/y$			$\hat{y}'=a+bx$
$\hat{y}=\frac{x}{a+bx}$	$y'=x/y$			$\hat{y}'=a+bx$
$\hat{y}=ax+bx^2$	$y'=y/x$			$\hat{y}'=a+bx$
$\hat{y}=a+b\ln x$		$x'=\ln x$		$\hat{y}'=a+bx'$
$\hat{y}=a+b\lg x$		$x'=\lg x$		$\hat{y}'=a+bx$
$\hat{y}=ax^b$	$y'=\ln y$	$x'=\ln x$	$a'=\ln a$	$\hat{y}'=a+bx'$
$\hat{y}=ae^{bx}$	$y'=\ln y$		$a'=\ln a$	$\hat{y}'=a+bx$
$\hat{y}=axe^{bx}$	$y'=\ln y/x$		$a'=\ln a$	$\hat{y}'=a+bx$
$\hat{y}=1/(ax^b)$	$y'=\ln 1/y$	$x'=\ln x$	$a'=\ln a$	$\hat{y}'=a+bx'$

统计上已经证明，对于倒数函数曲线、指数函数曲线、对数函数曲线、幂函数曲线、S 形曲线这几种曲线方程，只要变量转换后的直线化回归关系达到显著，变量反转后的曲线方程回归关系也较好。

倒数函数（reciprocal function）常见的表达式有以下几种：

$$
\begin{aligned}
\hat{y} &= \frac{a+bx}{x} \\
\hat{y} &= \frac{1}{a+bx} \\
\hat{y} &= \frac{x}{a+bx}
\end{aligned}
$$

引入不同的新变量，这些函数都可以转换为直线方程。本节将以第一个方程来讨论倒数函数方程的拟合过程。

- 如果两个变量在散点图上类似于倒数函数的曲线图，可以试配 $\hat{y} = \frac{a+bx}{x}$ 型的回归方程。
- 首先引入 $y' = xy$，然后用 y' 与 x 进行直线回归分析，求得 a 和 b，经过数据还原即可得到倒数函数方程
- 应用 $\hat{y} = \frac{a+bx}{x}$ 的条件如下
  - x 的观测值没有 0 值。
  - xy 应具有专业意义，而不是抽象的量。
  - 以 y' 和 x 为坐标绘制的散点图有明显的直线性。
  - y' 和 x 的相关系数显著。

例 8.1

测定 ‘苏品 1 号’ 玉米在不同密度下的平均株重 (x) 和经济系数 (y) 的关系，结果如下表。试对该资料进行回归分析。

密度	x	y	$y' = xy$	x^2	xy'	y'^2	$\hat{y}$	$y-\hat{y}$
1	399	0.380	151.620	159201	60496.380	22988.624	0.384	- 0.004
2	329	0.379	124.691	108241	41023.339	15547.845	0.376	0.003
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
8	90	0.248	22.320	8100	2008.800	498.182	0.259	-0.011

- 绘制 x, y 的散点图，发现与倒数函数 $a < 0, b > 0$ 的曲线相似
- 绘制 $x, y' = xy$ 的散点图，可以看到存在直线关系，表明选择该倒数函数曲线是合适的。
- 配合倒数函数方程：利用 y' 与 x 进行直线回归分析，得到 a 和 b
- 直线化回归方程的显著性检验
- 计算曲线回归方程的相关指数。

具体计算步骤如下

(2) 直线回归分析

- 第一步：计算 6 个一级数据：
  $\sum x, \sum x^2, \sum y', \sum y'^2, \sum xy', n = 7$
- 第二步：由一级数据计算 5 个二级数据：$SS_x, SS_{y'}, SP_{xy'}, \bar{x}, \bar{y}'$
- 第三步：计算 b 和 a：$b = \frac{SP_{xy'}}{SS_x}$，$a = \bar{y}' - b\bar{x}$
- 第四步：将 a 和 b 带入，得到曲线方程 $\hat{y} = \frac{-14.5353+0.4206x}{x}$

(3) 直线化回归方程的显著性检验

- 计算 x 与 y' 的相关系数并对其进行显著性检验：
  $r_{xy'} = \frac{SP_{xy'}}{\sqrt{SS_x \cdot SS_y}} = 0.9996$ 查表得 $r_{0.01(5)} = 0.874, |r_{xy'}| > r_{0.01}$，
  表明 x 与 y' 之间存在极显著的直线回归关系。

(4) 计算曲线回归方程的相关指数

- 将 x 的每一取值代入倒数方程，得到相应的 $\hat{y}$，其结果列于上表。从每一对 y 与 $\hat{y}$ 的差异大小可以看出每一对观测值受随机误差影响的大小，差异越大，随机误差越大。
- 本例中倒数函数的相关指数：$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = 0.9829$ 表明玉米株重与经济系数的回归关系可用倒数函数方程进行描述，其相关指数为 0.9829，该倒数函数拟合较好。

指数函数（exponential function）常见的有两种形式

$$
\begin{aligned}
\hat{y} &= ae^{bx} \\
\hat{y} &= ab^{x}
\end{aligned}
$$

在指数函数中，x 是作为指数出现的，系数 b 用以描述增长或衰减速率。

- 对指数函数 $\hat{y} = ae^{bx}$, $b > 0$ 表示增长曲线；$b < 0$ 表示衰减曲线
- 对指数函数 $\hat{y} = ab^{x}$, $b > 1$ 表示凹增长曲线，$0 < b < 1$ 表示凸增长曲线; $b < 0$ 表示衰减曲线
- 将 $\hat{y} = ae^{bx}$ 两边取自然对数，得到 $\ln \hat{y} = \ln a + bx$，令 $y' = \ln y; a' = \ln a$, 我们得到

$$\hat{y}' = a' + bx$$

此即直线化后的表达式。

例 8-2

棉花红铃虫产卵数与温度有关，试根据下表数据建立棉花红铃虫产卵数与文件的回归方程。

表: 棉花红铃虫产卵数与温度的关系

温度	21	23	25	27	29	32	35
产卵数	7	11	21	24	66	115	325
$y' = \ln y$	1.9459	2.3979	3.0445	3.1781	4.1897	4.7449	5.7838

(1) 绘制散点图，判断曲线类型: 散点图与 $b > 0$ 时的指数函数曲线相似，选择 $\hat{y} = ae^{bx}$ 进行曲线拟合；以 x 与 y' 的数据绘制散点图，其直线趋势比较明显

(2) 配合指数曲线方程，以 x 与 y' 进行直线回归分析，求出 a' 与 b，并将 a' 转换为 a，代入 $\hat{y} = ae^{bx}$ 即可。

具体步骤如下

(2) 直线回归分析

- 第一步：计算 6 个一级数据:
  $\sum x, \sum x^2, \sum y', \sum y'^2, \sum xy', n = 7$
- 第二步：由一级数据计算 5 个二级数据:$SS_x, SS_{y'}, SP_{xy'}, \bar{x}, \bar{y}'$
- 第三步：计算 b 和 a'：$b = \frac{SP_{xy'}}{SS_x}$，$a' = \bar{y}' - b\bar{x}$
- 第四步：计算 a 值：将 a, b 代入 $\hat{y} = ae^{bx}$，得到曲线方程

$$\hat{y} = 0.0213e^{0.2720x}$$

(3) 直线化回归方程的显著性检验

- 检验 y' 与 x 的回归关系，计算回归平方和 U 与离回归平方和 Q，

$$F = \frac{U}{Q} \times n - 2$$

对数函数 (logarithmic function) 的一般表达式为

$$\hat{y} = a + b \lg x$$

若令 $\lg x = x'$，则上式可写成

$$\hat{y} = a + bx'$$

此即为上式的直线化表达式。

- 如果两个变量的成对观测值在坐标系中的散点图分布趋势类似于对数函数曲线，可配合对数曲线方程 $\hat{y} = a + b \lg x$
- 首先将 x 变量的每一观测值取对数转换为新变量 x'，然后将 x' 与 y 进行直线回归分析，求出 a, b，即得到对数函数方程。

例 8-3

水稻育秧，塑料薄膜青苗床内空气最高温度 y 和室外空气最高温度 x 资料，试求它们之间的函数关系。

表: 表 8-5: 苗床内最高温度与空气最高温度的关系

序号	x	y	$x' = \lg x$	x'^2	$x'y$	y^2	$\hat{y}$
1	7.2	13.8	0.8573	0.7350	11.8312	190.44	16.79
2	7.9	21.4	0.8976	0.8057	19.2092	457.96	18.79
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
20	31.4	50.4	1.4969	2.2408	75.4453	2540.16	48.57
合计	420.1	767.5	25.8350	33.9997	1022.5905	31174.43	

(1) 绘制散点图，判断曲线类型：散点图与 $b > 0$ 时的对数函数曲线相似；为了能够作出更准确的判断，绘出 X' 与 y 的散点图，x' 与 y 间存在直线关系，直线趋势比较明显

(2) 配合对数曲线方程，以 x' 与 y 进行直线回归分析，求出 a 与 b，代入 $\hat{y} = a + b\lg x$ 即得对数函数方程。具体步骤如下

- 第一步：计算 6 个一级数据：$\sum x', \sum x'^2, \sum y, \sum y^2, \sum x'y, n = 7$
- 第二步：由一级数据计算 5 个二级数据：$SS_{x'}, SS_y, SP_{x'y}, \bar{x'}, \bar{y}$
- 第三步：计算 b 和 a'：$b = \frac{SP_{x'y}}{SS_{x'}}$，$a = \bar{y} - b\bar{x'}$
- 第四步：计算 a 值：将 a, b 代入 $\hat{y} = ae^{bx}$，得到曲线方程

$$\hat{y} = -25.8184 + 49.6930\lg x$$

(3) 直线化回归方程的显著性检验

- 检验 y 与 x' 的回归关系，计算回归平方和 $U = bSP_{x'y}$ 与离回归平方和 $Q = SS_y - U$，

$$F = \frac{U}{Q} \times n - 2$$

利用曲线回归方程进行预测时，可将每一个 x 值转换为 x'，再通过回归分析方法预测出 y 所在区间

幂函数 (power function) 曲线是 y 为 x 某次幂的函数曲线，其方程为

$$\hat{y} = ax^b$$

两边取对数，得到

$$\lg \hat{y} = \lg a + b \lg x$$

令 $y' = \lg y, x' = \lg x, a' = \lg a$，得到直线方程

$$\hat{y}' = a' + bx'$$

如果两个变量的散点图类似于幂函数的某条曲线，说明可用幂函数曲线进行拟合。为了进一步验证，可将 x 与 y 都去对数，然后绘出 x' 与 y' 的散点图，如果散点图呈直线趋势，即说明选用幂函数曲线是适合的。

例 8-4

为了研究 CO_2 对变黄期烟叶叶绿素降解的影响，在 30 倍 CO_2 浓度下测定了不同烘烤时间 (x,h) 下叶绿素含量（y, 占干重%)，其结果列于下表，试对该资料进行回归分析。

表: 表 8-7: 烘烤时间与叶绿素含量的关系

x	y	$x' = \lg x$	$y' = \lg y$
12	0.17430	1.0792	-0.7587
15	0.11080	1.1761	-0.9555
⋮	⋮	⋮	⋮
58	0.03533	1.7634	-1.4519

(1) 绘制散点图，判断曲线类型：散点图与 $b > 0$ 时的幂函数曲线相似；为了能够作出更准确的判断，绘出 x' 与 y' 的散点图，x' 与 y' 间存在直线关系，直线趋势比较明显

(2) 配合幂函数曲线方程，以 x' 与 y' 进行直线回归分析，求出 a 与 b，代入 $\hat{y} = ax^b$ 即得幂函数方程。具体步骤如下

- 第一步：计算 6 个一级数据：$\sum x', \sum x'^2, \sum y', \sum y'^2, \sum x'y', n = 11$
- 第二步：由一级数据计算 5 个二级数据:$SS_{x'}, SS_{y'}, SP_{x'y'}, \bar{x'}, \bar{y'}$
- 第三步：计算 b 和 a'：$b = \frac{SP_{x'y'}}{SS_{x'}} v = -0.9631$, $a' = \bar{y'} - b\bar{x'} = 0.1464$
- 第四步：将 a' 转换为 a：将得到幂函数曲线方程

$$a = 10^{a'} = 1.4009; \hat{y} = 1.4009x^{-0.9631}$$

(3) 直线化回归方程的显著性检验

- 检验 y 与 x' 的回归关系，计算回归平方和 $U = bSP_{x'y'}$ 与离回归平方和 $Q = SS_{y'} - U$,

$$F = \frac{U}{Q} \times n - 2$$

为了更好的说明例 8.4 中资料符合幂函数曲线关系，而不是其他类型曲线或直线关系，可将其进行不同曲线方程的拟合，并计算直线化后两个变量间的相关系数

- 幂函数方程 $\hat{y} = 1.4009x^{-0.9631}$, $\lg x$ 与 $\lg y$ 之间的 $r = -0.9417$
- 对数函数方程 $\hat{y} = 0.3147 - 0.0743\ln x$, $\ln x$ 与 y 之间的 $r = -0.8562$
- 指数函数方程 $\hat{y} = 0.1457e^{-0.0304x}$, x 与 $\ln y$ 之间的 $r = -0.7562$
- 直线函数方程 $\hat{y} = 0.1372 - 0.0023x$，$x$ 与 y 之间的 $r = -0.7652$

以上 4 个方程，其直线化后两个变量的相关系数均达到极显著水平 $r_{0.01(9)} = 0.735$。而幂函数曲线直线化的相关程度最高，表明配合出的幂函数曲线 $\hat{y} = 1.4009x^{-0.9631}$ 是最适宜的。

一、Logistic 生长曲线的由来和基本特征

- Logistic 生长曲线（logistic growth curve）最初由比利时学家 P.F. Verhulst 于 1838 年推导出来，但长期被忽略，直到 20 世纪初 20 年代才被生物学家和统计学家 R. Pearl 和 L.J. Reed 重新发现
- 目前被广泛用于动植物的饲养、栽培、资源、生态、环保等方面的模拟研究
- 其函数特点是开始增长缓慢，而后在某一范围内迅速增长，达到某限度后，增长又缓慢起来。曲线呈拉长的 S 形，也称为 S 形曲线

$$\hat{y} = \frac{K}{1 + ae^{-bx}}$$

- 当 $x = 0$, $\hat{y} = \frac{K}{1+a}$；当 $x \to \infty, \hat{y} = K$, 所以时间为 0 的起始量是 $\frac{K}{1+a}$，时间无限延长的终极量为 K
- 生长曲线的二阶导数 $\frac{d^2y}{dx^2} = 0$ 时，$x = \frac{-\ln\frac{1}{a}}{b}$, $\hat{y} = \frac{K}{2}$, 这说明 y 是随 x 的增加而增加的。当 $x = \frac{-\ln\frac{1}{a}}{b}$ 时，曲线有一个拐点 (knee point)，这时 $\hat{y} = \frac{K}{2}$，恰为终极量 K 的一半，所以 x 在 $\left[0, x = \frac{-\ln\frac{1}{a}}{b}\right]$ 区间内，曲线下凹；x 在 $\left[x = \frac{-\ln\frac{1}{a}}{b}, \infty\right]$ 区间内，曲线上凸。

二、Logistic 生长曲线方程的配合

- 只要确定了 K 值，就可利用直线化方法求出方程的两个统计数 a 和 b，经移项后可得

$$\frac{K-\hat{y}}{\hat{y}} = ae^{-bx}$$

两边取自然对数后的

$$\ln\frac{K-\hat{y}}{\hat{y}} = \ln a - bx$$

若令 $\hat{y}' = \ln\frac{K-\hat{y}}{\hat{y}}, a' = \ln a, b' = -b$, 我们得到

$$y' = a' + b'x$$

- 因此可以将每一个 y 观测值转换为 y'，用 y' 与 x 进行直线回归分析，即可求出 a' 与 b'。在转换时，必须先确定 K 值，K 的确定方法有下面两种：
  (1) 如果依变量 y 是累积频率，则 y 无限增大的终极量应为 100%，可用 $K = 100$ 来表示
  (2) 当依变量 y 是生长量或者繁殖量时，可取 3 对自变量为等间距的观测值 $(x_1, y_1), (x_2, y_2), (x_3, y_3)$，并将其带入生长函数公式，得到联立方程组

$$\begin{cases} \frac{K-y_1}{y_1} = ae^{-bx_1} \\ \frac{K-y_2}{y_2} = ae^{-bx_2} \\ \frac{K-y_3}{y_3} = ae^{-bx_3} \end{cases}$$

消去方程组中的 a, b，得

$$\frac{y_2(K-y_1)}{y_1(K-y_2)} = \left[\frac{y_3(K-y_2)}{y_2(K-y_3)}\right]^{\frac{x_1-x_2}{x_2-x_3}}$$

如果令 $x_2 = \frac{x_1+x_3}{2}$，可得

$$K = \frac{y_2^2(y_1+y_3) - 2y_1y_2y_3}{y_2^2 - y_1y_3}$$

例 8-5

下表是测定某种肉鸡在良好的生长条件下生长过程的数据资料，试配合 Logistic 生长曲线方程。

表: 表 8-9: 肉鸡生长过程的资料

x	y	$\frac{2.827-y}{y}$	$y' = \ln\frac{2.827-y}{y}$
2	0.30	8.4233	2.1310
4	0.86	2.2872	0.8273
⋮	⋮	⋮	⋮
14	2.80	0.0096	-4.6415

分析步骤如下

(1) 求终极量 K：取 x 为等间距的 $x_1 = 2, x_2 = 8, x_3 = 14$ 的生长量 $y_1 = 0.30, y_2 = 2.20, y_3 = 2.80$，可计算得 $K = 2.827(kg)$

(2) 将 y 转换为 y'：计算出 $\frac{2.827-y}{y}$ 和 $y' = \ln\frac{2.827-y}{y}$ 分别列于表中

(3) 计算 x 和 y' 回归分析得二级数据 $SS_x, SS_{y'}, SP_{xy'}, \bar{x}, \bar{y'}$

(4) 计算 x 和 y' 的相关系数 $r_{xy'}$

$$r_{xy'} = \frac{-58.2363}{\sqrt{112 \times 30.8067}} = -0.9914$$

$|r_{xy'}| > r_{0.01(5)} = 0.874$ 达到极显著水平，说明 x 与 y' 的直线关系是极显著的，配合 Logistic 生长曲线方程是适合的

(5) 求 a' 和 b'，并转换为 a 和 b

$$\begin{aligned} b' &= \frac{SP_{xy'}}{SS_x} = -0.5200 \to a = e^{a'} = 19.9654 \\ a' &= \bar{y'} - b'\bar{x'} = 2.9940 \to b = -b' = 0.5200 \end{aligned}$$

所以有

$$\hat{y} = \frac{2.827}{1 + 19.9654e^{-0.5200x}}$$

且在 $\hat{y} = \frac{K}{2} = 1.4135$, $x = \frac{-\ln\frac{1}{a}}{b} = 5.76$ 时，是曲线的拐点，生长速率从越来越快开始变为越来越慢，这是此种肉鸡生长的关键时期。

Question?