第9章 非线性回归

1. 可化为线性回归的曲线回归
2. 多项式回归
3. 非线性模型

9.1 可化为线性回归的曲线回归

曲线回归模型

$$y=\beta_0+\beta_1 e^{bx}+\varepsilon\ ,\ (\text{b已知})\quad (9.1)$$

只须令$x'=e^{bx}$ 即可化为 y 对 x' 是线性的形式

$$y=\beta_0+\beta_1 x'+\varepsilon$$

需要指出的是，新引进的自变量**只能依赖于原始变量**，而不能与未知参数有关。

9.1 可化为线性回归的曲线回归

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\ldots+\beta_p x^p+\varepsilon \qquad (9.2)$$

令$x_1=x, x_2=x^2, \ldots, x_p=x^p$，于是得到$y$关于$x_1, x_2, \ldots, x_p$的**线性表达式**

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p+\varepsilon$$

(9.2)式本来只有一个自变量x，是一元p次多项式回归，在线性化后，**变为p元线性回归**。

线性回归的“线性”是针对未知参数$\beta_j, j=1,\ldots p$而言的。对于回归解释变量的线性是非本质的，因为解释变量是非线性时，总可以通过变量的替换把它**转化成线性的**。

9.1 可化为线性回归的曲线回归

可线性化的曲线回归模型，也称为**本质线性回归模型**

$$y=ae^{bx}e^{\varepsilon} \qquad (9.3)$$

对等式两边同时取自然对数，得：

$$\ln y=\ln a+bx+\varepsilon$$

令$y'=\ln y$，$\beta_0=\ln a$，$\beta_1=b$，

于是得到y'关于x的**一元线性回归模型**

$$y'=\beta_0+\beta_1 x+\varepsilon$$

9.1 可化为线性回归的曲线回归

不可线性化的曲线回归模型，如

$$y = ae^{bx} + \varepsilon \qquad (9.4)$$

不能通过对等式两边同时取自然对数的方法将回归模型线性化，只能用非线性最小二乘方法求解。

(9.3)式的误差项称为**乘性误差项**。

(9.4)式的误差项称为**加性误差项**。

一个非线性回归模型是否可以线性化，不仅与回归函数的形式有关，而且与误差项的形式有关。

9.1 可化为线性回归的曲线回归

在对非线性回归模型线性化时，总是假定误差项的形式就是能够使回归模型线性化的形式，为了方便，常常省去误差项，仅写出回归函数的形式。例如把回归模型

(9.3)式 $y=ae^{b\,x}e^{\varepsilon}$ 简写为 $y=ae^{bx}$。

(9.3)式与(9.4)式的回归参数的估计值是有差异的。对误差项的形式，首先应该由**数据的经济意义来确定**，然后由回归拟合效果做检验。过去，由于没有非线性回归软件，人们总是希望非线性回归模型可以线性化，因而误差项的形式就假定为可以把模型线性化的形式。现在利用计算机软件可以容易的解决**非线性回归问题**，因而对误差项形式应该做正确的选择。

9.1 可化为线性回归的曲线回归

10种常见的可线性化的曲线回归方程

表 9-1

英文名称	中文名称	方程形式
Linear	线性函数	$y = b_0+b_1t$
Logarithm	对数函数	$y = b_0+b_1\ln t$
Inverse	逆函数	$y = b_0+b_1/t$
Quadratic	二次曲线	$y = b_0+b_1t+b_2t^2$
Cubic	三次曲线	$y = b_0+b_1t+b_2t^2+b_3t^3$
Power	幂函数	$y = b_0\,t^{b_1}$
Compound	复合函数	$y = b_0\,b_1^t$
S	S 形函数	$y = \exp(b_0+b_1/t)$
Logistic	逻辑函数	$y = \dfrac{1}{1+e^{-t}}$
Growth	增长曲线	$y = \exp(b_0+b_1t)$
Exponent	指数函数	$y = b_0\exp(b_1t)$

9.1 可化为线性回归的曲线回归

除了上述 10 种常用的曲线外，还有几种常用的曲线如下。

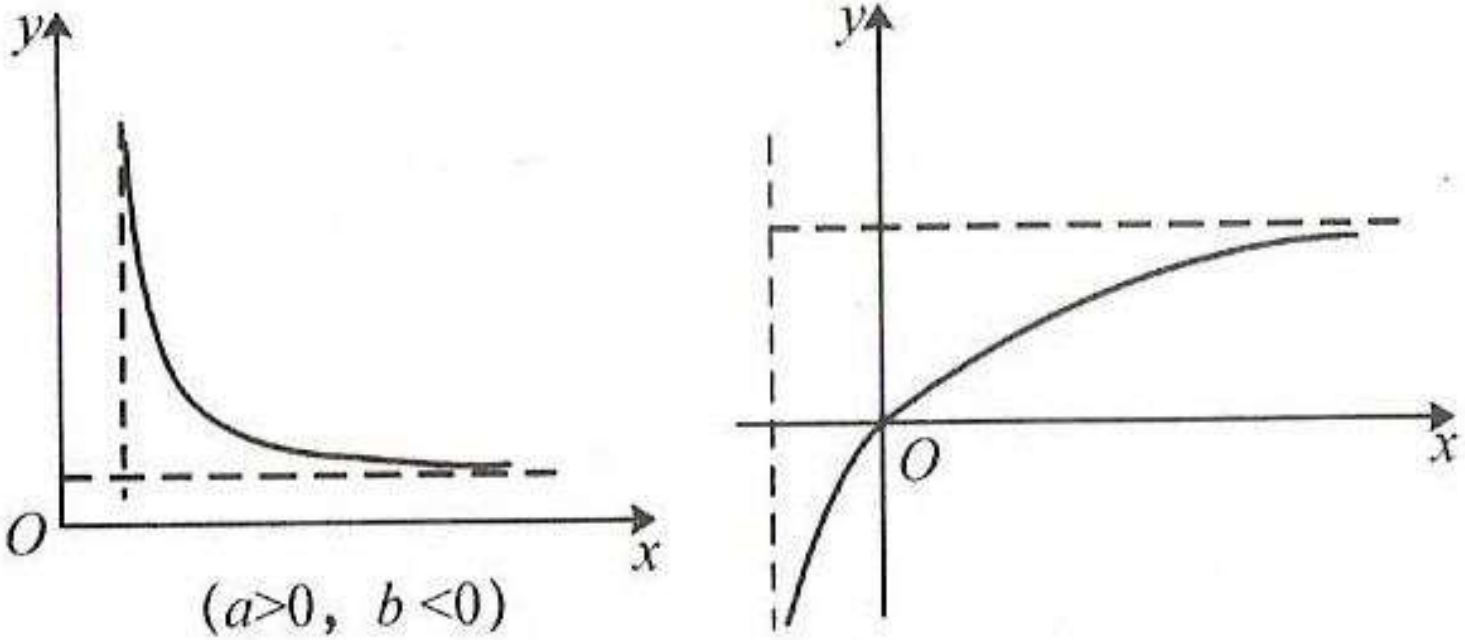
1. 双曲函数

$$y=\frac{x}{ax+b}$$

或等价地表示为

$$\frac{1}{y}=a+b\frac{1}{x}$$

9.1 可化为线性回归的曲线回归



(a) 双曲函数

9.1 可化为线性回归的曲线回归

2. S型曲线II

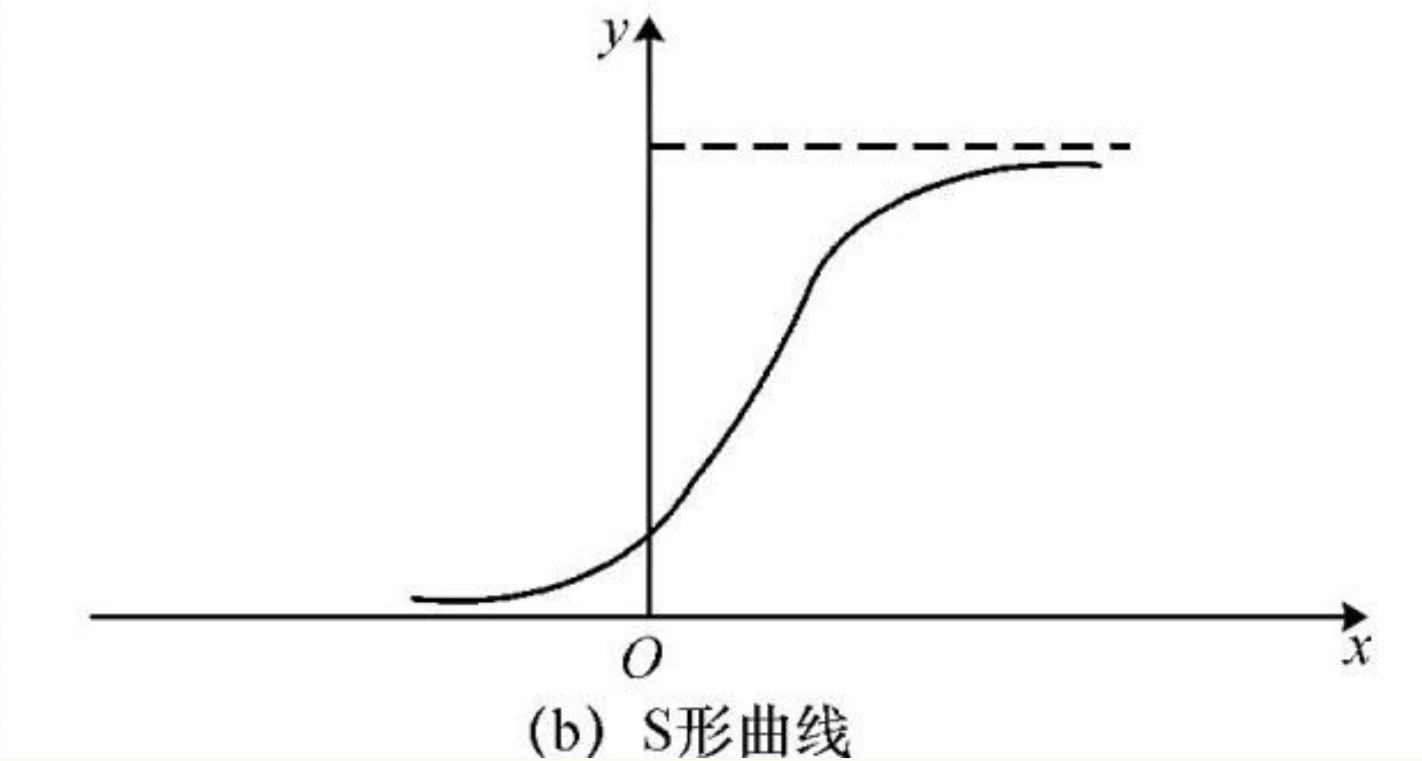
$$y=\frac{1}{a+be^{-x}}$$

此S型曲线II当$a>0$，$b>0$时，是 x 的增函数。 当 $x\to+\infty$时，$y\to 1/a$； $x\to-\infty$时，$y\to 0$。 $y=0$ 与 $y=1/a$是这条曲线的两条渐进线。

S型曲线有多种，其共同特点是曲线**首先是缓慢增长**，在达**到某点后迅速增长**， 在超过某点后**又变为缓慢增长**， 并且**趋于一个稳定值。**

S型曲线在社会经济等很多领域都有应用，例如某种产品的销售量与时间的关系，树木、农作物的生长与时间的关系等。

9.1 可化为线性回归的曲线回归



(b) S形曲线

9.1 可化为线性回归的曲线回归

例9-1 对国内生产总值(GDP)的拟合。我们选取GDP指标为因变量，单位为亿元，拟合GDP关于时间t的趋势曲线。以1991年为基准年，取值为t=1，2013年t=23，1991—2013年的数据如表9-2所示:

9.1 可化为线性回归的曲线回归

表 9-2

年　份	t	y	$y'=\ln y$	$\hat{y}$	e
1991	1	21 781.5	9.99	28 546.3	−6 764.8
1992	2	26 923.5	10.20	32 779.5	−5 856.0
1993	3	35 333.9	10.47	37 640.4	−2 306.5
1994	4	48 197.9	10.78	43 222.2	4 975.7
1995	5	60 793.7	11.02	49 631.7	11 162.0
1996	6	71 176.6	11.17	56 991.7	14 184.9
1997	7	78 973.0	11.27	65 443.1	13 529.9
1998	8	84 402.3	11.34	75 147.9	9 254.4
1999	9	89 677.1	11.40	86 291.7	3 385.4
2000	10	99 214.6	11.51	99 088.1	126.5
2001	11	109 655.0	11.61	113 782.1	−4 127.1
2002	12	120 333.0	11.70	130 655.1	−10 322.1
2003	13	135 823.0	11.82	150 030.3	−14 207.3
2004	14	159 878.0	11.98	172 278.6	−12 400.6
2005	15	184 937.0	12.13	197 826.2	−12 889.2
2006	16	216 314.0	12.28	227 162.3	−10 848.3
2007	17	265 810.0	12.49	260 848.8	4 961.2
2008	18	314 045.0	12.66	299 530.6	14 514.4
2009	19	340 507.0	12.74	343 948.7	−3 441.7
2010	20	401 513.0	12.90	394 953.7	6 559.3
2011	21	473 104.0	13.07	453 522.3	19 581.7
2012	22	519 470.0	13.16	520 776.2	−1 306.2
2013	23	568 845.0	13.25	598 003.3	−29 158.3

9.1 可化为线性回归的曲线回归

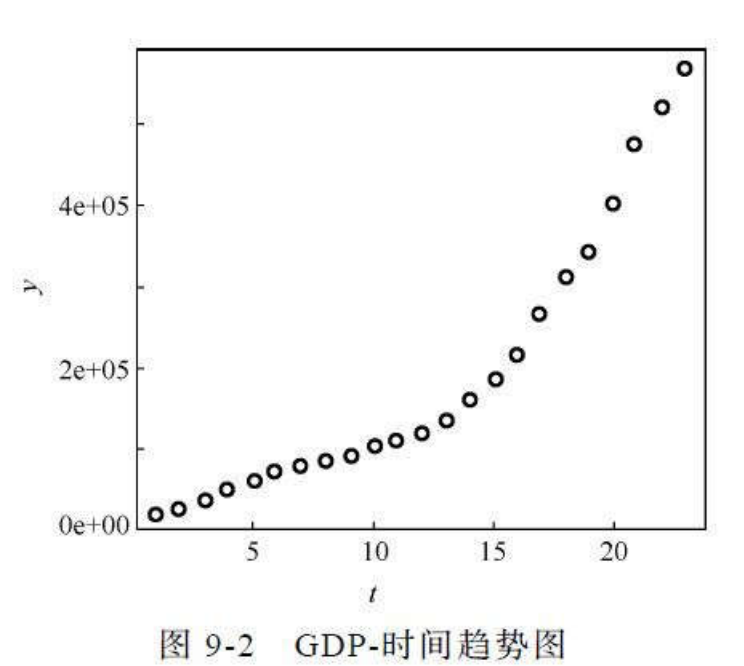


图 9-2　GDP-时间趋势图

从散点图中看到，GDP 随时间t的变化趋势大致为**指数函数形式**，从经济学角度看，当 GDP 的年增长速度大致相同时，其趋势线就是指数函数形式。易看出复合函数 $y = b_0 b_1^t$，增长曲线 $y = \exp\left(b_0 + b_1 t\right)$，指数函数 $y = b_0 \exp\left(b_1 t\right)$ 这三个曲线方程实际上是**等价的**。在本例中，复合函数 $y = b_0 b_1^t$ 的形式与经济意义更吻合。

9.1 可化为线性回归的曲线回归

以时间 t 为自变量，对数据进行拟合，我们考虑建立简单线性回归模型和复合函数回归模型，其中**复合函数** $y = b_0 b_1^t$ 是可线性化的，只需要对式子两边同时取对数即可将其化为 $\ln y$ 关于 t 的线性函数。因此，在建立复合函数回归模型前需要计算 $\ln y$ 的值，见表9-2。

建立简单线性回归模型和复合函数回归模型的计算代码如下，其运行结果如输出结果9.1 和图9-3 所示。

9.1 可化为线性回归的曲线回归

计算代码

```
lm9.1<-lm(y~t,data=data9.1)     #做简单线性回归
summary(lm9.1)
anova(lm9.1)
ly<-log(y)                      #对因变量 y 取对数并赋给 ly
lm9.12<-lm(ly~t)                #做 ly 关于 t 的线性回归
summary(lm9.12)
anova(lm9.12)
plot(data9.1)                   #画散点图
lines(data9.1$t, exp(predict(lm9.12)), col='red')     #画拟合曲线
abline(lm9.1)                   #添加拟合的直线
detach(data9.1)
```

9.1 可化为线性回归的曲线回归

输出结果 9.1

```
> summary(lm9.1)
Call:
lm(formula = y ~ t, data = data9.1)

Residuals:
   Min     1Q     Median     3Q     Max
-79390 -53910    -11187  42650 126163

Coefficients:
             Estimate   Std. Error t value  Pr(>|t|)
(Intercept)    -80498     27318       -2.947   0.0077 **
t               22747      1992       11.417   1.81e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 63380 on 21 degrees of freedom
Multiple R-squared:  0.8612,    Adjusted R-squared:  0.8546
```

9.1 可化为线性回归的曲线回归

```
F-statistic: 130.3 on 1 and 21 DF,  p-value: 1.814e-10

> anova(lm9.1)
Analysis of Variance Table
Response: y
            Df      Sum Sq     Mean Sq       F value    Pr(>F)
t            1   5.2363e+11   5.2363e+11    130.35    1.814e-10 ***
Residuals   21   8.4361e+10   4.0172e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(lm9.12)
Call:
lm(formula = ly ~ t)

Residuals:
      Min           1Q         Median         3Q        Max
-0.270465    -0.065304    -0.002511  0.044795    0.222258

Coefficients:
                  Estimate   Std.Error  t value  Pr(>|t|)
(Intercept)      10.121005   0.052062    194.40    <2e-16 ***
t                 0.138276   0.003797     36.42    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1208 on 21 degrees of freedom
Multiple R-squared:  0.9844,    Adjusted R-squared:  0.9837
F-statistic:  1326 on 1 and 21 DF,  p-value: < 2.2e-16
```

9.1 可化为线性回归的曲线回归

```
> anova(lm9.12)
                Analysis of Variance Table
Response: ly
             Df     Sum Sq    Mean Sq  F value    Pr(>F)
t             1    19.3497   19.3497   1326.2   < 2.2e-16 ***
Residuals    21     0.3064    0.0146
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

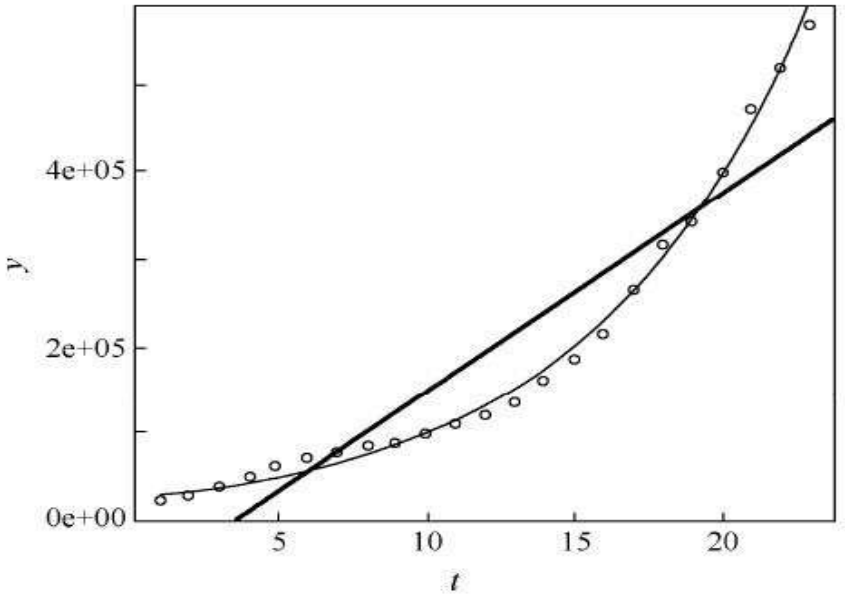


图 9-3　例 9-1 的运行结果

9.1 可化为线性回归的曲线回归

由输出结果9.1可知，线性回归的**决定系数R^2 = 0.8612**，残差平方和SSE=8.4361e+10，复合函数回归的**决定系数R^2 = 0.9844**，残差平方和SSE=0.3064是按线性化后的回归模型计算的，两者的残差不能直接相比。为了与线性回归的拟合效果直接相比，可以先存储复合函数**y的预测值**$\hat{y} = \exp(\hat{y})$，计算残差序列e(见表9-2)，然后计算出复合函数回归的SSE=3.005e+9，可知复合函数拟合效果明显优于线性回归。

另外，从模型拟合图中，也可直观得到这一结论，故在解决此类问题时应采用复合函数回归。

9.1 可化为线性回归的曲线回归

根据输出结果9.1 中线性化后复合函数的回归系数，可以计算得到复合函数回归系数分别为$b_0 = 24859.62$，等比系数$b_1 = 1.148$，因此回归方程为

$$\hat{y} = 24859.62 \times (1.148)^t$$

式中 $b_1 = 114.8\%$表示GDP 的平均发展速度，平均增长速度为14.8%。这里GDP用的是**当年现价**，包含物价上涨因素在内。本例只是作为计算非线性回归的示例。在实际工作中， 如果需要对GDP做趋势拟合或预测，应对此模型做一些改进，例如用不变价格代替现价，对误差项的自相关做相应的处理；考虑到GDP的年增长速度会有减缓趋势，可以给回归函数增加适当的**阻尼因子**，或采用**S 形曲线拟合**等改进方法。

9.2 多项式回归

9.2.1 几种常见的多项式回归模型

一元二次多项式模型 $y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i$ 的回归函数 $y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2$ 是一条抛物线方程，通常称为二项式回归函数。

回归系数 β_1 为线性效应系数，β_{11} 为二次效应系数。

相应地，回归模型

$$y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3 + \varepsilon_i$$

称为**一元三次多项式模型**。

9.2 多项式回归

称回归模型

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^{\ 2} + \beta_{22} x_{i2}^{\ 2} + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

为二元二阶多项式回归模型。它的回归系数中分别含有两个自变量的**线性项系数** β_1 和 β_2, **二次项系数** β_{11} 和 β_{22} 并含有**交叉乘积项系数** β_{12}。交叉乘积项表示 x_1 与 x_2 的交互作用，系数 β_{12} 通常称为交互影响系数。

9.2 多项式回归

9.2.2 应用实例

例9-2　表9-3列出的数据是关于18个35岁～44岁经理的:

前两年平均年收入 x_1（千美元）

风险反感(意识)度 x_2

人寿保险额 y（千美元）

风险反感度是根据发给每个经理的标准调查表估算得到的，它的数值越大，风险反感就越厉害。

9.2 多项式回归

研究人员想研究给定年龄组内的经理年平均收入，风险反感度和人寿保险额的关系。研究者预计，在经理的收入和人寿保险额之间成立着二次关系，并有把握认为风险反感度对人寿保险额只有线性效应，而没有二次效应。但是，研究者对两个自变量是否对人寿保险额有交互效应，心中没底。因此，研究者拟合了一个**二阶多项式回归模型**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

并打算先检验是否有**交互效应**，然后检验风险反感的**二次效应。**

9.2 多项式回归

表 9-3

序　　号	x_{i1}	x_{i2}	y_i
1	66.290	7	196
2	40.964	5	63
3	72.996	10	252
4	45.010	6	84
5	57.204	4	126
6	26.852	5	14
7	38.122	4	49
8	35.840	6	49
9	75.796	9	266
10	37.408	5	49
11	54.376	2	105
12	46.186	7	98
13	46.130	4	77
14	30.366	3	14
15	39.060	5	56
16	79.380	1	245
17	52.766	8	133
18	55.916	6	133

9.2 多项式回归

回归采用逐个引入自变量的方式，这样可以清楚地看到各项对回归的贡献，使显著性检验更加明确。依次引入自变量 $x_1, x_2, x_1^2, x_2^2, x_1x_2$ 以查看各变量对回归的贡献，计算代码如下：

```
data9.2<-read.csv("D:/data9.2.csv",head=TRUE)
lm9.21<-lm(y~x1,data=data9.2)
lm9.22<-lm(y~x1+x2,data=data9.2)
lm9.23<-lm(y~x1+x2+I(x1^2),data=data9.2)    #I(x1^2)表示变量 x1 的二次项
lm9.24<-lm(y~x1+x2+I(x1^2)+I(x2^2),data=data9.2)
lm9.25<-lm(y~x1+x2+I(x1^2)+I(x2^2)+I(x1*x2),data=data9.2)
#I(x1*x2)表示变量 x1 与 x2 的交互项
anova(lm9.21)
anova(lm9.22)
anova(lm9.23)
anova(lm9.24)
anova(lm9.25)
```

9.2 多项式回归

上述计算程序，首先是建立依次引入各变量后的回归模型，然后依次输出各模型的方差分析表，根据方差分析表中的结果，我们将运行结果所得的依次引入各变量后的偏平方和以及残差平方和进行整理并计算偏 F 值，得到方差分析表见表9-4，其中取显著性水平为0.05。

表 9-4

变量	偏平方和	残差平方和	检验系数	偏 F 值
x_1	104 474	3 567	β_1	—
$x_2 \mid x_1$	2 284	1 283	β_2	—
$x_1^2 \mid x_1, x_2$	1 238	45	β_{11}	1 238/(45/14)=385
$x_2^2 \mid x_1, x_2, x_1^2$	3	42	β_{22}	3/(42/13)=0.93
$x_1x_2 \mid x_1, x_2, x_1^2, x_2^2$	6	36	β_{12}	6/(36/12)=2.00
合计	108 005			

9.2 多项式回归

全模型的SST = 108041，SSE = 36，SSE 的自由度 $df = n - p - 1 = 18 - 5 - 1 = 12$。采用式(3.42)的偏 F 检验，对交互影响系数 β_{12} 的显著性检验的偏 F 值 = 2.00，临界值 $F_{0.05}(1,12) = 4.75$，交互影响系数 β_{12} **不能通过显著性检验**，认为 $\beta_{12}=0$，回归模型中不应该包含交互作用项 x_1x_2。这个结果与人们的经验相符，有了此结果，两个自变量的效应也就容易解释了。此时，研究者暂时决定使用无交互效应的模型

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \varepsilon_i$$

9.2 多项式回归

但仍想检验风险反感度的二次效应是否存在。这相当于检验二次效应**系数 β_{22} 的显著性**，这个检验的偏 F 值等于 0.93，临界值 $F_{0.05}(1,13) = 4.67$，二次效应系数 β_{22} 不能通过显著性检验，认为 $\beta_{22} = 0$，回归模型中不应该包含二次效应项 x_2^2 。此时，研究者决定使用简化的回归模型

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \varepsilon_i \quad i$$

9.2 多项式回归

进一步检验年平均收入的二次效应是否存在，这相 当于检验二次效应系数 β_{11} 的显著性，这个检验的偏 F 值等于385，临界值 $F_{0.05}(1,14) = 4.60$，二次效应系数 β_{11} 通过了显著性检验，认为 β_{11} 不为零，回归模型中应该包含二次效应项 x_1^2。得最终的回归方程为

$$\hat{y} = -62.349 + 0.840x_1 + 5.685x_2 + 0.0371x_1^2$$
$$(0.164) \quad (0.164) \quad (0.785)$$

其中，括号中的数值是标准化回归系数。这样，研究者可用这个回归方程来进一步研究经理的年平均收入和风险反感度对人寿保险额的效应。从标准化回归系数看到，年平均收入的二次效应对人寿保险额的影响程度最大。

9.3 非线性模型

9.3.1 非线性最小二乘

非线性回归模型一般可记为：

$$y_i = f(x_i, \boldsymbol{\theta}) + \varepsilon_i,\ i = 1, 2, \ldots, n \qquad (9.8)$$

其中，**y_i是因变量**，

非随机向量$x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})'$是自变量，

$\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_p)'$是未知参数向量，

ε_i是随机误差项并且满足GM假定，即

$$\begin{cases} E(\varepsilon_i) = 0, \quad i = 1, 2, \cdots, n \\ \operatorname{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} \end{cases} \qquad (i, j = 1, 2, \cdots, n)$$

9.3 非线性模型

如果$f(\boldsymbol{x}_i, \boldsymbol{\theta})=\beta_0+\beta_1 X_1+\ldots+\beta_p X_p$，那么式 (9.8)就是前面讨论的线性模型，而且必然有 $k=p$；对于**一般情况的非线性模型**，参数的数目与自变量的数目并没有一定的对应关系，**不要求 $k=p$**。

对非线性回归模型式(9.8)，仍使用**最小二乘法估计**参数$\boldsymbol{\theta}$，即求使

$$Q(\boldsymbol{\theta})=\sum_{i=1}^{n}(y_i-f(x_i,\boldsymbol{\theta}))^2$$

达到最小的 $\hat{\boldsymbol{\theta}}$，称 $\hat{\boldsymbol{\theta}}$为**非线性最小二乘估计**。

9.3 非线性模型

在假定 f 函数对参数 θ 连续可微时，可以利用微分法，建立正规方程组，求解使 Q(θ)达最小的 $\hat{\theta}$。

将 f 函数对参数 θ_j 求导，并令为 0，得 p+1 个方程：

$$\left.\frac{\partial Q}{\partial \theta_j}\right|_{\theta_j=\hat{\theta}_j} = -2\sum_{i=1}^{n}(y_i - f(x_i,\hat{\theta}))\left.\frac{\partial f}{\partial \theta_j}\right|_{\theta_j=\hat{\theta}_j} = 0$$

$$j = 0,1,2,\cdots,p$$

称为**非线性最小二乘估计的正规方程组**

也可以用数值法极小化残差平方和 Q(θ)，求出未知参数 θ 的非线性最小二乘估计 $\hat{\theta}$。

9.3 非线性模型

对于**非线性最小二乘估计**，我们仍然需要做参数的区间估计、显著性检验、回归方程的显著性检验等回归诊断，这需要知道有关统计量的分布。在非线性最小二乘中，一些精确分布是很难得到的，在大样本时，可以得到近似的分布。计算机软件在求出参数$\boldsymbol{\theta}$的非线性最小二乘估计值的同时，还给出近似的回归诊断结果。

在非线性回归中，**平方和分解式SST=SSR+SSE不再成立。**类似于线性回归中的复判定系数，定义非线性回归的相关比（也称为相关指数）为：

$$R^2 = 1 - \frac{SSE}{SST}$$

9.3 非线性模型

9.3.2 非线性回归模型的应用

例9-3 一位药物学家使用下面的非线性模型对药物反应拟合回归模型：

$$y_i = c_0 - \frac{c_0}{1+\left(\frac{x_i}{c_2}\right)^{c_1}} + \varepsilon_i$$

自变量 x 是药剂量，用级别表示；

因变量 y 是药物反应程度，用百分数表示。

3个参数c_0、c_1、c_2都是非负的，根据专业知识，c_0的上限是100%， 3个参数的初始值取为c_0=100，c_1=5，c_2=4.8。测得9个反应数据如表9-5：

9.3 非线性模型

表 9-5 反应数据

x	1	2	3	4	5	6	7	8	9
y(%)	0.5	2.3	3.4	24.0	54.7	82.1	94.8	96.2	96.4

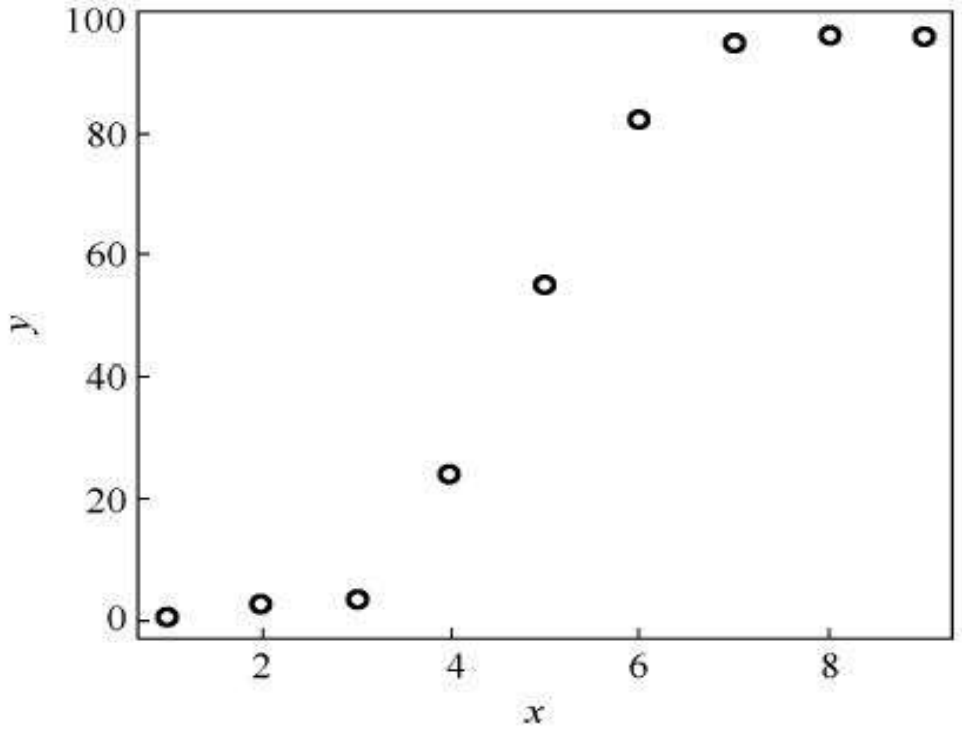


图 9-4 药物反应程度散点图

9.3 非线性模型

通过图 9-4 可以看出，y 与 x 之间确实呈非线性关系，因此需要对数据进行非线性回归分析。R 软件中做非线性回归的函数为 nls(formula,data,start,…)，formula 部分为非线性模型的函数表达式，start 为模型中未知参数的初始值，对例 9.3 中的数据进行非线性回归分析的计算代码如下，运行结果见输出结果9.2。

9.3 非线性模型

计算代码

```
x=c(1:9)
y=c(0.5,2.3,3.4,24,54.7,82.1,94.8,96.2,96.4)
nls9.3<-nls(y~a-a/(1+(x/c)^b),start=list(a=100,b=5,c=4.8))
#非线性回归，其中未知参数的初始值分别为 100,5,4.8
summary(nls9.3)
e<-resid(nls9.3)                #计算残差赋给变量 e
ebar<-mean(e)                   #残差 e 的均值
SE<- deviance(nls9.3)
#残差平方和，由于 e 的均值不等于 0，所以 SE 不等于残差的离差平方和
SSE<-sum((e-ebar)^2)            #残差的离差平方和

prey<-fitted(nls9.3)            #y 的预测值
pybar<-mean(prey)               #y 的预测值的均值
SSR<-sum((prey-pybar)^2)        #回归离差平方和
ybar<-mean(y)                   #y 的均值
SST<-sum((y-ybar)^2)            #总离差平方和
Rsquare<- 1-SE/SST              #相关指数(仿照线性回归中的计算公式)
Rsquare
```

9.3 非线性模型

输出结果 9.2

```
> summary(nls9.3)
Formula: y ~ a - a/(1 + (x/c)^b)
Parameters:
     Estimate    Std. Error    t value    Pr(>|t|)
a    99.54052     1.56733       63.51     1.02e-09 ***
b     6.76125     0.42198       16.02     3.75e-06 ***
c     4.79964     0.05017       95.68     8.79e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.834 on 6 degrees of freedom
Number of iterations to convergence: 6
Achieved convergence tolerance: 2.485e-06
> Rsquare
[1] 0.9986467
```

9.3 非线性模型

由以上输出结果可知，对参数的估计经过**6 步迭代后收敛**，而且相关指数 $R^2 = 0.9986$，说明非线性回归拟合效果很好。同时，上述输出结果中对参数的显著性检验显示参数均通过显著性检验。但是，在样本量较小的情况下，不可线性化的非线性回归的**残差通常不满足正态性**，进而使用 t 分布进行检验也是无效的，因此**显著性检验的结果并不具有重要意义**。另外，由上述代码可以计算出 y 的预测值、残差、残差平方和、回归平方和、总离差平方和等，将这些计算结果列于表中，具体可见表9-6。

9.3 非线性模型

表 9-6

序号	x	y	$\hat{y}$	e	$\hat{y}-\bar{y}$
1	1	0.5	0.00	0.5	−50.488 89
2	2	2.3	0.27	2.03	−50.218 89
3	3	3.4	3.98	−0.58	−46.508 89
4	4	24.0	22.48	1.52	−28.008 89
5	5	54.7	56.61	−1.91	6.121 11
6	6	82.1	81.52	0.58	31.031 11
7	7	94.8	92.34	2.46	41.851 11
8	8	96.2	96.49	−0.29	46.001 11
9	9	96.4	98.14	−1.74	47.651 11
均值	5	50.488 89	50.203 33	0.285 556	−0.285 56
离差平方和	60	14 917.89	15 156.55	19.431 62	15 156.55
平方和	285	37 860.04	37 839.85	20.188 03	15 157.28

9.3 非线性模型

本例回归离差平方和SSR=15156.55，而总离差平方和SST=14917.89<SSR，可见对非线性回归**不再满足平方和分解式**，即

$$SST \neq SSR+SSE$$

另外，非线性回归的残差和一般不等于零，本例残差均值为**0.285556≠0**。当然，如果回归拟合的效果好，残差的均值会接近于零的。

通过以上分析可以认为药物反应程度y与药剂量x符合以下非线性回归方程：

$$\hat{y}=99.541-\frac{99.541}{1+\left(\frac{x}{4.7996}\right)^{6.7612}}$$

9.3 非线性模型

例9-4 龚珀兹（Gompertz）模型是计量经济中的一个常用模型，用来拟合社会经济现象发展趋势，龚珀兹曲线形式为：

$$y_t = k \cdot a^{b^t}$$

其中 k 为变量的增长上限，$0<a<1$ 和 $0<b<1$是未知参数。当 k 未知时，龚珀兹模型**不能线性化**，可以用非线性最小二乘法求解。表 9-7 的数据是我国民航国内航线里程数据，以下用龚珀兹模型拟合这个数据。

9.3 非线性模型

表 9-7　我国民航国内航线里程数据　　单位：万公里

年　份	t	y	年　份	t	y
1980	1	11.41	1993	14	68.21
1981	2	13.55	1994	15	69.37
1982	3	13.28	1995	16	78.08
1983	4	12.92	1996	17	78.02
1984	5	15.28	1997	18	92.06
1985	6	17.12	1998	19	100.14
1986	7	21.67	1999	20	99.89
1987	8	24.02	2000	21	99.45
1988	9	24.55	2001	22	103.67
1989	10	30.55	2002	23	106.32
1990	11	34.04	2003	24	103.42
1991	12	38.17	2004	25	115.52
1992	13	53.36			

9.3 非线性模型

使用 R 软件对表 9-7 中的数据进行拟合，建立非线性模型，其中需要确定**未知参数的初始值**。由于初始值要求不是很准确，所以很多时候可以凭经验给定，对于本例题，龚珀兹中的参数 k 是变量的发展上限，应该取其初始值略大于最大观测值。本题最大观测值是115.52，不妨取 k 的初始值为120。a 和 b 都是 0～1 之间的数，可以取其初始值为0.5，非线性回归的计算代码如下。

```
data9.4<-read.csv("D:/data9.4.csv",head=TRUE)
y<-data9.4[,3]
t<-data9.4[,2]
model<-nls(y~k*(a^(b^t)),start=list(a=0.5,b=0.5,k=120))
```

9.3 非线性模型

按上述代码进行运算会出现产生**无限值不收敛**的情况，这是由于回归迭代过程中的**参数取值超出了范围**，可以通过对参数的取值增加一些限制来解决。因此，将参数 k 的初始值调整为130，另外对其上下限也做出限制，最小值取为116 即大于样本的最大观测值115.52，此时 nls 函数中的算法 algorithm 不能使用默认的高斯-牛顿迭代算法，需改为port，重新运行以下代码，得到输出结果9.3，并画出国内航线里程趋势预测图，如图9-5 所示。

什么是高斯-牛顿迭代法？？

9.3 非线性模型

```
model<-nls(y~k* (a^(b^t)),start=list(a=0.5,b=0.5,k=120),lower=c(0,0,116),
           upper=c(1,1,10000),algorithm="port")      #做非线性回归
summary(model)
c<-coef(model)                                       #将模型的回归系数赋给 c
tt<-c(1:30)
yp<-c[3]*(c[1]^(c[2]^tt))                            #计算时间取值为 tt 时对应的 y 的预测值
t1=t+1979                                            #计算相应的年份值赋给 t1
t2<-tt+1979
plot(t1,y,type="o",ann=FALSE,ylim=c(0,160),xlim=c(1975,2015))
#画样本的散点图
lines(t2,yp)                                         #画预测值
```

9.3 非线性模型

输出结果 9.3

```
Formula: y ~ k * (a^(b^t))

Parameters:
   Estimate      Std. Error   t value   Pr(>|t|)
a 1.243e-02      6.066e-03     2.050     0.0525 .
b 8.927e-01      1.475e-02    60.526  < 2e-16 ***
k 1.500e+02      1.581e+01     9.483     3.15e-09 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6.104 on 22 degrees of freedom
Algorithm "port", convergence message: relative convergence (4)
```

9.3 非线性模型

用非线性最小二乘法求得的三个参数估计值为

k=150.0，a=0.012，b=0.893

其中k=150.0为回归模型估计的国内航线里程增长上限。

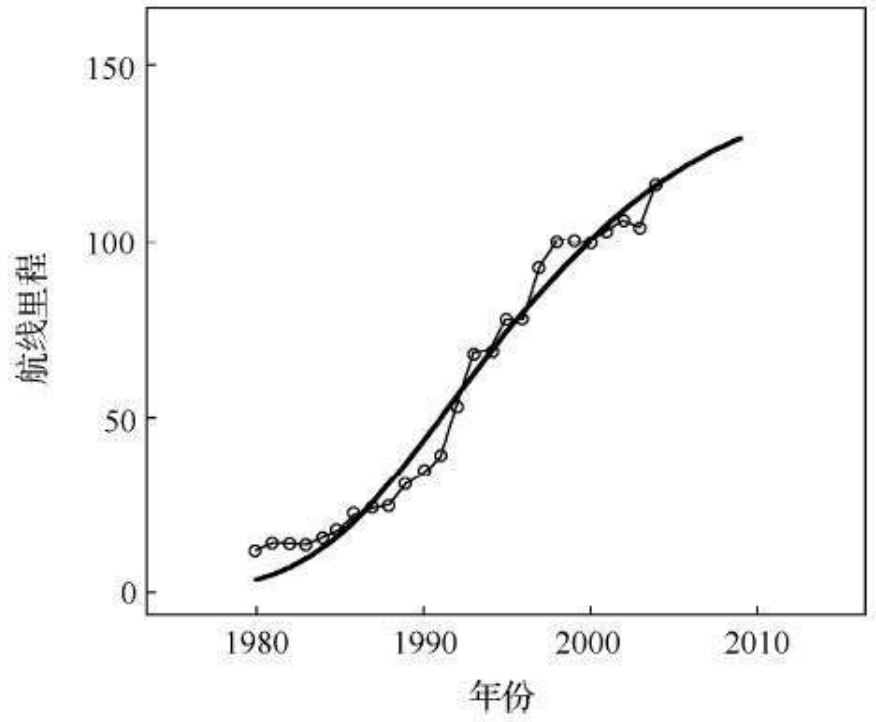


图 9-5　龚珀兹曲线拟合国内航线里程趋势图

如图9-5 中，圆圈代表观测值，光滑曲线为拟合曲线，从图中可以直观地看到，龚珀兹曲线能够较好刻画数据的变化趋势。

9.3 非线性模型

例9-5 下表9-8是我国从1950——2013年历年大陆总人口数，试用威布尔（Weibull）曲线拟合数据并做预测。威布 尔曲线如下：

$$y = k - a \cdot b^{t^c}$$

其中参数 k 是变量发展的上限，参数 $a>0$，$0<b<1$，$c>0$。

表 9-8 我国历年大陆总人口数 单位：亿人

年 份	t	y	年 份	t	y
1950	1	5.519 6	1958	9	6.599 4
1951	2	5.630 0	1959	10	6.720 7
1952	3	5.748 2	1960	11	6.620 7
1953	4	5.879 6	1961	12	6.585 9
1954	5	6.026 6	1962	13	6.729 5
1955	6	6.146 5	1963	14	6.917 2
1956	7	6.282 8	1964	15	7.049 9
1957	8	6.465 3	1965	16	7.253 8

9.3 非线性模型

续表

年　份	t	y	年　份	t	y
1966	17	7.454 2	1990	41	11.433 3
1967	18	7.636 8	1991	42	11.582 3
1968	19	7.853 4	1992	43	11.717 1
1969	20	8.067 1	1993	44	11.851 7
1970	21	8.299 2	1994	45	11.985 0
1971	22	8.522 9	1995	46	12.112 1
1972	23	8.717 7	1996	47	12.238 9
1973	24	8.921 1	1997	48	12.362 6
1974	25	9.085 9	1998	49	12.476 1
1975	26	9.242 0	1999	50	12.578 6
1976	27	9.371 7	2000	51	12.674 3
1977	28	9.497 4	2001	52	12.762 7
1978	29	9.625 9	2002	53	12.845 3
1979	30	9.754 2	2003	54	12.922 7
1980	31	9.870 5	2004	55	12.998 8
1981	32	10.007 2	2005	56	13.075 6
1982	33	10.154 1	2006	57	13.144 8
1983	34	10.249 5	2007	58	13.212 9
1984	35	10.347 5	2008	59	13.280 2
1985	36	10.453 2	2009	60	13.345 0
1986	37	10.572 1	2010	61	13.409 1
1987	38	10.724 0	2011	62	13.473 5
1988	39	10.897 8	2012	63	13.540 4
1989	40	11.270 4	2013	64	13.607 2

9.3 非线性模型

根据人口学的专业预测，我国人口上限为16亿人，因此取 k 的初值=16，取 b 的初值=0.5，取 c 的初值=1。

对以上初值把 t=1 时（即1950年）$y_1 = 5.5196$代入，得 $a = 2(k - y_1) \approx 2(16 - 5.5) = 21$。用21作为 a 的初值，做非线性最小二乘，相应的计算代码如下，其运行结果见输出结果9.4。

```
data9.5<-read.csv("D:/data9.5.csv",head=T)
y<-data9.5[,3]
t<-data9.5[,2]
model<-nls(y~k-(a*(b^(t^c))),start=list(a=21,b=0.5,c=1,k=16),
     lower=c(0,0,0,0),upper=c(10000,1,10000,10000),algorithm="port",
     control=nls.control(maxiter=1000,tol=1e-1000))
#对参数的上下限做了限制，另外参数 control 部分为控制迭代的次数及收敛标准
summary(model)
c<-coef(model)                          #将模型的回归系数赋给 c
tt<-c(1:70)
yp<-c[4]-(c[1]*(c[2]^(tt^c[3])))  #计算时间取值为 tt 时对应的 y 的预测值
```

9.3 非线性模型

```
t1=t+1949    #计算年份并赋给 t1
t2<-tt+1949
plot(t1,y,type="o",xlab="年份",ylab="大陆总人口数",ylim=c(5,16),
     xlim=c(1950,2020),cex=0.75)      #画样本的散点图
lines(t2,yp,col="red")                #添加拟合曲线
```

输出结果 9.4

```
Formula: y ~ k - (a * (b^(t^c)))
Parameters:
   Estimate     Std. Error   t value      Pr(>|t|)
a 9.237e+00    2.874e-01     32.14       <2e-16 ***
b 9.978e-01    3.564e-04     2799.24     <2e-16 ***
c 1.637e+00    5.349e-02     30.60       <2e-16 ***
k 1.491e+01    2.514e-01     59.29       <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.1258 on 60 degrees of freedom
Algorithm "port", convergence message: both X-convergence and
relative convergence (5)
```

9.3 非线性模型

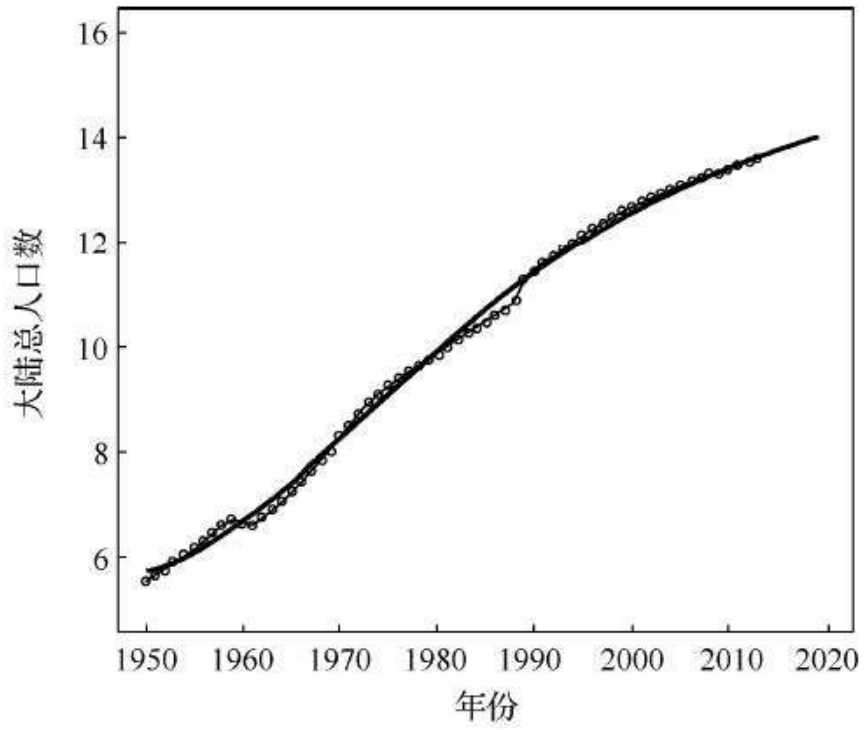


图 9-6　威布尔模型预测我国人口趋势图

从输出结果中看到，人口上限$k = 14.91$ 亿人，这与人口学预测的人口上限有一些差异，这是因为人口数会受到国家政策等许多因素的影响。如图 9-6 所示是绘制的人口趋势预测图，其中圆圈代表观测值，曲线代表预测值，其中预测 2020 年的人口数**约为 14 亿**。

9.3 非线性模型

例9-6　柯布—道格拉斯**生产函数**研究。在计量经济学中有一种熟知的C-D（Cobb—Douglas）生产函数

$$y = AK^{\alpha}L^{\beta}$$

其中，y 为产出，K(资本)、L(劳力)为两个投入要素，$A>0$ 为效率系数，α 和 β 是 K 和 L 的产出弹性，A，α，β 都是待估参数。

9.3 非线性模型

α是产出对资本投入的**弹性系数**，度量在劳动投入保持不变时资本投入增加1%时产出增加的百分比。

β是产出对劳动投入的**弹性系数**，度量在资本投入保持不变时劳动投入增加1%时产出增加的百分比。

两个弹性系数之和 $\alpha+\beta$ 表示**规模报酬**（returns to scale）。

$\alpha+\beta=1$表示规模报酬不变，即1倍的投入带来1倍的产出；

$\alpha+\beta<1$表示规模报酬递减，即1倍的投入带来少于1倍的产出；

$\alpha+\beta>1$表示规模报酬递增，即1倍的投入带来大于1倍的产出。

9.3 非线性模型

可以按两种形式设定随机误差项：

(1) 乘性误差项，模型形式为 $y = AK^{\alpha}L^{\beta}\mathrm{e}^{\varepsilon}$ 。

(2) 加性误差项，模型形式为 $y = AK^{\alpha}L^{\beta} + \varepsilon$

对乘法误差项模型可通过两边取对数转化成线性模型。

$$\ln y = \ln A + \alpha \ln K + \beta \ln L$$

令 $y' = \ln y$, $\beta_0 = \ln A$, $x_1 = \ln K$, $x_2 = \ln L$, 则转化为线性回归方程：

$$y' = \beta_0 + \alpha x_1 + \beta x_2 + \varepsilon$$

9.3 非线性模型

表 9-9

年　份	t	GDP	K	L	lnGDP	ln$K$	ln$L$
1978	1	3 624.1	1 377.9	40 152	8.1953 61	7.228 316	10.600 43
1979	2	4 038.2	1 474.2	41 024	8.3035 54	7.295 871	10.621 91
1980	3	4 517.8	1 590.0	42 361	8.4157 80	7.371 489	10.653 98
1981	4	4 862.4	1 581.0	43 725	8.4892 87	7.365 813	10.685 68
1982	5	5 294.7	1 760.2	45 295	8.5744 62	7.473 183	10.720 95
1983	6	5 934.5	2 005.0	46 436	8.6885 38	7.603 399	10.745 83
1984	7	7 171.0	2 468.6	48 197	8.8778 00	7.811 406	10.783 05
1985	8	8 964.4	3 386.0	49 873	9.1010 16	8.127 405	10.817 24
1986	9	10 202.2	3 846.0	51 282	9.2303 59	8.254 789	10.845 10
1987	10	11 962.5	4 322.0	52 783	9.3895 32	8.371 474	10.873 94
1988	11	14 928.3	5 495.0	54 334	9.6110 14	8.611 594	10.902 91
1989	12	16 909.2	6 095.0	55 329	9.7356 13	8.715 224	10.921 05
1990	13	18 547.9	6 444.0	64 749	9.8281 12	8.770 905	11.078 27
1991	14	21 617.8	7 517.0	65 491	9.9812 72	8.924 922	11.089 67

9.3 非线性模型

续表

年　份	t	GDP	K	L	lnGDP	ln$K$	ln$L$
1992	15	26 638.1	9 636.0	66 152	10.190 10	9.173 261	11.099 71
1993	16	34 634.4	14 998.0	66 808	10.452 60	9.615 672	11.109 58
1994	17	46 759.4	19 260.6	67 455	10.752 77	9.865 817	11.119 22
1995	18	58 478.1	23 877.0	68 065	10.976 41	10.080 67	11.128 22
1996	19	67 884.6	26 867.2	68 950	11.125 56	10.198 66	11.141 14
1997	20	74 462.6	28 457.6	69 820	11.218 05	10.256 17	11.153 68
1998	21	78 345.2	29 545.9	70 637	11.268 88	10.293 70	11.165 31
1999	22	82 067.5	30 701.6	71 394	11.315 30	10.332 07	11.175 97
2000	23	89 468.1	32 611.4	72 085	11.401 64	10.392 42	11.185 60
2001	24	97 314.8	37 460.8	73 025	11.485 71	10.531 05	11.198 56
2002	25	105 172.3	42 355.4	73 740	11.563 36	10.653 85	11.208 30

9.3 非线性模型

其中，y 是国内生产总值GDP (单位：亿元)， K是资金投入，包括固定资产投资和库存占用资金(单位：亿元)，L是就业总人数(单位：万人)。

(1)假设随机误差项为相乘的，我们可以用两边取对数的办法，对数变换后的数据见表9-9，用 R 软件做**线性回归**的代码如下，运行代码得到输出结果9.5。

```
data9.6<-read.csv("D:/data9.6.csv",head=TRUE)
#data9.6 中存储的为表 9-9 中的数据，变量名依次记为 t，y，k，l，ly，lk，ll
model1<-lm(ly~lk+ll,data9.6)
summary(model1)
anova(model1)
```

9.3 非线性模型

输出结果 9.5

```
> summary(model1)
Call:
lm(formula = ly ~ lk + ll, data = data9.6)

Residuals:
     Min       1Q   Median       3Q      Max
-0.144098 -0.023947  0.005014  0.030900  0.076601

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.08589    1.90325  -1.096   0.2849
lk           0.90239    0.03489  25.862   <2e-16 ***
ll           0.36054    0.20099   1.794   0.0866 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9.3 非线性模型

```
Residual standard error: 0.05219 on 22 degrees of freedom
Multiple R-squared:  0.9981,    Adjusted R-squared:  0.998
F-statistic:  5918 on 2 and 22 DF,  p-value: < 2.2e-16

> anova(model1)
                Analysis of Variance Table
Response: ly
           Df   Sum Sq   Mean Sq      F value        Pr(>F)
lk          1   32.228   32.228   11831.9985       <2e-16 ***
ll          1    0.009    0.009       3.2178       0.0866 .
Residuals  22    0.060    0.003
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9.3 非线性模型

得两个弹性系数估计值为α=0.902，β=0.361，**资金的贡献率大于劳动力的贡献率。**

规模报酬$\alpha+\beta$=0.902+0.361=1.263＞1表示规模报酬递增。

效率系数A=0.1242。

其中系数β的显著性概率P值=0.087，显著性较弱。

得乘性误差项的C-D生产函数为：

$$y = 0.1242K^{0.902}L^{0.361}$$

9.3 非线性模型

（2）对加性误差项模型，不能通过变量变换转化成线性模型，只能用非线性最小二乘求解未知参数。以上面乘性误差项的**参数为初始值**做非线性最小二乘，计算代码如下所示，得到的运行结果见输出结果9.6。

```
model2<-nls(y~A*((k^a)*(l^b)),data9.6,start=list(A=2,a=0.9,b=0.3),
        lower=c(0,0,0),upper=c(10000,100,100),algorithm="port",control=
        nls.control(maxiter=1000,tol=1e-1000))
summary(model2)
```

9.3 非线性模型

由输出结果9.6 可知，参数 β 仍未通过显著性检验，与乘性误差项模型的检验结果一致，因此不能认为 β 非零。另外，得加性误差项的 C-D 生产函数为

$$\hat{y} = 0.02K^{0.922}L^{0.505}$$

输出结果 9.6

```
Formula: y ~ A * ((k^a) * (l^b))
Parameters:
     Estimate   Std. Error   t value   Pr(>|t|)
A    0.02047     0.10418     0.196      0.846
a    0.92237     0.06446    14.309      1.26e-12 ***
b    0.50486     0.51094     0.988      0.334
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2194 on 22 degrees of freedom
Algorithm "port", convergence message: relative convergence (4)
```

9.3 非线性模型

乘性误差项模型和加性误差项模型所得的结果**还是有较大差异的**，其中**乘性误差项模型认为 y_t 本身是异方差的**，而$\ln y_t$ 是等方差的。

加性误差项模型认为 y_t 是等方差的。

从统计性质看两者的差异，前者淡化了**y_t值大的项**（近期数据）的作用，强化了y_t值小的项（早期数据）的作用，对早期数据拟合的效果较好。

而后者则对**近期数据拟合的效果较好**。

9.3 非线性模型

9.3.3 其他形式的非线性回归

非线性最小二乘是使**残差平方和**

$$Q(\theta)=\sum_{i=1}^{n}(y_i-f(x_i,\theta))^2$$

达极小的方法, 其最大的缺点是缺乏稳健性。当数据存在异常值时，**参数的估计效果变得很差**。因而在一些场合，我们希望用一些更稳健的残差损失函数代替平方损失函数，例如绝对值损失函数。**绝对值残差损失函数**为

$$Q(\theta)=\sum_{i=1}^{n}|y_i-f(x_i,\theta)|$$

备注：即非线性分位数回归τ=0.5的特例。