Regression Models

Statistics and Actuarial Science

中国人民大学统计与精算系列教材

回归模型

孟生旺　编著

中国人民大学出版社

Regression Models

Statistics and Actuarial Science

中国人民大学统计与精算系列教材

回归模型

孟生旺　编著

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

回归模型/孟生旺编著. —北京：中国人民大学出版社，2015.10
中国人民大学统计与精算系列教材
ISBN 978-7-300-22064-2

Ⅰ.①回… Ⅱ.①孟… Ⅲ.①回归分析-高等学校-教材 Ⅳ.①O212.1

中国版本图书馆 CIP 数据核字（2015）第 248055 号

中国人民大学统计与精算系列教材
回归模型
孟生旺 编著
Huigui Moxing

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010－62511242（总编室）	010－62511770（质管部）	
	010－82501766（邮购部）	010－62514148（门市部）	
	010－62515195（发行公司）	010－62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com(人大教研网)		
经　　销	新华书店		
印　　刷	北京昌联印刷有限公司		
规　　格	185 mm×260 mm　16 开本	**版　　次**	2015 年 11 月第 1 版
印　　张	15.5 插页 1	**印　　次**	2015 年 11 月第 1 次印刷
字　　数	380 000	**定　　价**	32.00 元

前　言

回归模型是应用最为广泛的数据分析方法之一，它的核心思想是建立若干个解释变量与一个因变量的函数关系，并通过这个函数关系对因变量的变化规律进行解释和预测。

回归模型包括线性回归模型、广义线性模型、广义可加模型、线性混合模型、广义线性混合模型以及它们的各种扩展模型。本书以线性回归模型和广义线性模型为主，介绍回归模型的基本原理和应用技巧。

线性回归模型主要建立在因变量服从正态分布等一系列重要假设之上，所以不能完全满足解决某些特定问题的实际需求。譬如，损失次数是非负的整数，损失金额是大于零的实数，都不符合线性回归模型关于因变量服从正态分布假设的前提条件。在广义线性模型中，因变量可以服从指数分布族中的任意分布，如正态分布、二项分布、泊松分布、伽马分布、逆高斯分布和 Tweedie 分布等。这些分布非常适合描述保险损失数据，如损失次数可以用二项分布、泊松分布或负二项分布进行描述，损失金额可以用伽马分布或逆高斯分布进行描述。事实上，几乎所有分布假设下的广义线性模型都可以在保险损失数据的分析中找到用武之地，也很难找到另一个数据分析需要用到所有的广义线性模型的实际领域。有鉴于此，本书主要使用汽车保险的损失数据来说明广义线性模型的建模原理，但读者可以毫无困难地将其应用到其他领域的数据分析中。

由于实际数据结构的复杂性，广义线性模型还在不断发展。譬如，在广义线性模型的线性预测项中引入平滑函数，就可以建立广义可加模型；在线性预测项中增加随机效应，就得到了广义线性混合效应模型；如果不仅对分布的均值参数建立回归模型，同时对其他参数建立回归模型，就可以将广义线性模型推广到关于位置参数、尺度参数和形状参数的广义可加模型（GAMLSS）。

本书将以线性回归模型和广义线性模型为主，介绍回归模型的基本原理和建模技术。书中虽然有一定的理论介绍，但更加侧重于回归模型的实际应用。第 1 章介绍线性回归模型的基本原理，并通过模拟数据介绍线性回归模型的建模技术。第 2 章介绍广义线性模型的统计理论，为后面各章的实际应用奠定基础。第 3～7 章根据因变量的不同类型，分别介绍基于二分类因变量、计数型因变量、连续型因变量和混合型因变量建立回归模型的基本方法，包括参数估计、统计推断、残差分析和模型检验等内容。此外，作为对广义线性模型的进一步拓展，还介绍了基于多分类因变量的回归模型、有限混合回归模型、零膨胀回归模型、零调整逆高斯回归模型和广义可加模型。第 8 章简要介绍贝叶斯视角下回归模型的参数估计方法。第 9 章应用我国一家财产保险公司的车损险数据，讨论如何建立索赔发生概率、索赔频率、索赔强度和纯保费的回归模型。

本书在建模过程中使用了两种类型的数据，一类是模拟数据；另一类是实际数据。基于模拟数据的建模过程可以清晰地展示数据的生成机理和回归模型的误差，而基于实际数据的建模过程可以揭示回归模型在应用中可能遇到的问题和解决方法。

常用的统计软件都包含建立回归模型的模块，如 R，SAS 和 STATA 等。本书主要使用

了 R 软件中的 lm 函数、glm 函数和 glm.nb 函数，以及 gamlss，tweedie，cplm，car，rstan，penalized，data.table，ggplot2，knitr 等程序包。在此谨向这些程序包的开发者表示感谢。

本书适合统计、精算、经济、金融、保险和管理等相关专业的学生使用。本书案例分析和练习题中使用的数据集可以从孟生旺的新浪博客 http://blog.sina.com.cn/mengshw 下载。

本书在编写过程中得到了中国人民大学统计学院部分研究生的帮助，他们是李政宵、杨亮、王选鹤、王明高、刘新红、陈静仁、邱子真、隋凤艳、徐轩、龙骧、叶适，其中李政宵和杨亮直接参与了部分章节初稿的编写，在此向他们表示衷心感谢。

孟生旺

教师教学服务说明

中国人民大学出版社工商管理分社以出版经典、高品质的工商管理、财务会计、统计、市场营销、人力资源管理、运营管理、物流管理、旅游管理等领域的各层次教材为宗旨。

为了更好地为一线教师服务，近年来工商管理分社着力建设了一批数字化、立体化的网络教学资源。教师可以通过以下方式获得免费下载教学资源的权限：

在“人大经管图书在线”（www. rdjg. com. cn）注册，下载“教师服务登记表”，或直接填写下面的“教师服务登记表”，加盖院系公章，然后邮寄或传真给我们。我们收到表格后将在一个工作日内为您开通相关资源的下载权限。

如您需要帮助，请随时与我们联络：

中国人民大学出版社工商管理分社

联系电话：010－62515735，62515749，62515987

传　　真：010－62515732，62514775　　电子邮箱：rdcbsjg@crup. com. cn

通讯地址：北京市海淀区中关村大街甲 59 号文化大厦 1501 室（100872）

教师服务登记表

姓 名		□先生 □女士	职　　称		
座机/手机			电子邮箱		
通讯地址			邮　　编		
任教学校			所在院系		
所授课程	课程名称	现用教材名称	出版社	对象（本科生/研究生/MBA/其他）	学生人数
需要哪本教材的配套资源					
人大经管图书在线用户名					

院/系领导（签字）：

院/系办公室盖章

目　录

第 1 章　线性回归模型

线性回归模型是应用最为广泛的统计分析方法之一，也是广义线性模型的基础。本章主要介绍线性回归模型的基本原理，包括模型设定、参数估计、假设检验、模型诊断、模型评价和比较等。

1.1　模型结构和假设

假设我们感兴趣的变量是 y，希望建立它与其他 k 个解释变量 x_1，x_2，…，x_k 之间的函数关系。显然，最一般的函数形式可以表示为：

$$y=f(x_1, x_2, \cdots, x_k)+\varepsilon$$

式中，ε 为用函数 $f(x_1, x_2, \cdots, x_k)$ 表示 y 产生的随机误差。函数 $f(x_1, x_2, \cdots, x_k)$ 通常是未知的，但在线性回归模型中，我们假设它是一个线性函数，即

$$f(x_1, x_2, \cdots, x_k)=\beta_0+\beta_1 x_1+\cdots+\beta_k x_k$$

式中，参数 β_0，β_1，…，β_k 是未知的，需要进行估计。

由此可得线性回归模型的一般形式为：

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_k x_k+\varepsilon$$

如果对因变量和解释变量有 n 次观测，第 i 次观察值记为 y_i 和 x_{1i}，x_{2i}，…，x_{ki}，则相应的线性回归模型可以表示为：

$$y_i=\beta_0+\beta_1 x_{1i}+\cdots+\beta_k x_{ki}+\varepsilon_i \tag{1.1}$$

式中，ε_i 是随机误差项。

若令

$$\boldsymbol{\beta}=[\beta_0, \beta_1, \cdots, \beta_k]^{\mathrm{T}}$$
$$\boldsymbol{x}_i^{\mathrm{T}}=[1, x_{1i}, \cdots, x_{ki}]$$

则式（1.1）的线性回归模型也可以表示为：

$$y_i=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon_i \tag{1.2}$$

式（1.1）和（1.2）中的线性回归模型也可以表示为下述矩阵形式：

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \tag{1.3}$$

式中，$\boldsymbol{y}$ 表示因变量向量；$\boldsymbol{X}$ 表示设计矩阵；$\boldsymbol{\beta}$ 表示回归系数向量；$\boldsymbol{\varepsilon}$ 表示误差向量。它们的具体形式如下：

$$\boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1},\ \boldsymbol{X}=\begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}_{n\times(k+1)},\ \boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1},\ \boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

线性回归模型的参数估计建立在一系列基本假设之上，这些假设包括：

(1) 误差项的均值为零，且与解释变量相互独立，即有

$$E(\boldsymbol{\varepsilon})=\mathbf{0} \tag{1.4}$$

$$E(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon})=\mathbf{0} \tag{1.5}$$

(2) 误差项独立同分布，即每个误差项之间相互独立：

$$\mathrm{Cov}(\varepsilon_i,\ \varepsilon_j)=0,\ i\neq j \tag{1.6}$$

且每个误差项的方差都相等：

$$\mathrm{Var}(\varepsilon_i)=\sigma_i^2=\sigma^2,\ i=1,2,\cdots,n \tag{1.7}$$

独立同分布假设可以用矩阵形式表示为：

$$\mathrm{Var}(\boldsymbol{\varepsilon})=\sigma^2\mathbf{I}$$

式中，$\mathbf{I}$ 是 $n\times n$ 的单位矩阵。

(3) 设计矩阵 $\boldsymbol{X}$ 是满秩矩阵，其秩 $\mathrm{rk}(\boldsymbol{X})=k+1=p$ 等于回归参数的个数。模型有 k 个解释变量，对应 k 个参数，再加上截距项，共有 $k+1$ 个参数。满秩假设意味着解释变量之间没有线性相关关系。

(4) 正态假设，即假设误差项服从正态分布：

$$\varepsilon_i\sim N(0,\ \sigma^2) \tag{1.8}$$

正态假设使得回归参数的最小二乘估计等价于极大似然估计。该假设主要用于对回归参数的估计值进行统计检验，并求得回归参数的置信区间。在大样本条件下，由中心极限定理可知，即使没有正态分布假设，也可以对回归参数的估计值进行统计推断。

在上述假设之下，很容易求得

$$E(y_i)=E(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon_i)=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}=\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki}$$

$$\mathrm{Var}(y_i)=\mathrm{Var}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon_i)=\mathrm{Var}(\varepsilon_i)=\sigma^2$$

$$\mathrm{Cov}(y_i,\ y_j)=\mathrm{Cov}(\varepsilon_i,\ \varepsilon_j)=0$$

这就意味着因变量服从正态分布，且有

$$y_i\sim N(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta},\ \sigma^2) \tag{1.9}$$

如果表示为矩阵形式，则有

$$E(\boldsymbol{y})=\boldsymbol{X}\boldsymbol{\beta}$$

$$\mathrm{Cov}(\boldsymbol{y})=\sigma^2\mathbf{I}$$

$$\boldsymbol{y}\sim N(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2\mathbf{I})$$

注意，在古典线性回归模型中，解释变量仅对因变量的均值产生影响，而与因变量的方差或协方差无关。

1.2 解释变量

在线性回归模型中，解释变量可以是连续变量，也可以是分类变量，还可以是若干个变

量的乘积（即交互效应），或一个变量的函数变换（如幂变换）。这些都是建立线性回归模型必须考虑的现实问题，本节简要介绍这些问题的处理方法。

1.2.1　分类解释变量

如果回归模型中的解释变量包含分类变量，如性别、职业、车型、地区等，则在建模过程中需要将分类解释变量转化为“哑变量”或“虚拟变量”。为了避免解释变量之间出现完全的共线性，即防止一个解释变量可以表示为其他解释变量的线性组合，每个分类解释变量转化为虚拟变量的个数与分类变量的水平数有关，即等于分类变量的水平数减去 1。譬如，车型是一个分类解释变量，有“A”，“B”，“C”和“D”四个水平，则可以转化为三个虚拟变量 x_1，x_2，x_3，它们的定义如表 1—1 所示。

表 1—1　　分类解释变量转化为虚拟变量

车型	x_1	x_2	x_3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

用 x_1 表示车型“A”，即当车型为“A”时，$x_1=1$，在其他情况下，$x_1=0$。

用 x_2 表示车型“B”，即当车型为“B”时，$x_2=1$，在其他情况下，$x_2=0$。

用 x_3 表示车型“C”，即当车型为“C”时，$x_3=1$，在其他情况下，$x_3=0$。

在 x_1，x_2，x_3 的取值确定之后，车型就可以完全确定。譬如，当 $x_1=0$，$x_2=0$，$x_3=0$ 时，车型必然为“D”。换言之，4 个水平的解释变量只需 3 个虚拟变量，再增加一个虚拟变量会导致完全的共线性。不妨增加一个虚拟变量 x_4 表示车型“D”，则当车型为“D”时，$x_4=1$，在其他情况下，$x_4=0$。此时，模型的 4 个虚拟变量将是完全线性相关的，即 $x_1+x_2+x_3+x_4=1$。

假设车型是模型中唯一的解释变量，则线性回归模型的拟合值可以表示为：

$$\mu=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3 \tag{1.10}$$

模型（1.10）的设计矩阵为：

$$\boldsymbol{X}=\begin{pmatrix}1 & 1 & 0 & 0\\ 1 & 0 & 1 & 0\\ 1 & 0 & 0 & 1\\ 1 & 0 & 0 & 0\end{pmatrix}$$

根据模型（1.10），可以求得不同车型条件下对因变量的拟合值为：

$$\mu=\begin{cases}\beta_0+\beta_1，\text{车型}=\text{A}\\ \beta_0+\beta_2，\text{车型}=\text{B}\\ \beta_0+\beta_3，\text{车型}=\text{C}\\ \beta_0，\text{车型}=\text{D}\end{cases}$$

在模型（1.10）中，车型 D 是基准水平（base level），也称作参照水平（reference level）。在该水平下，对因变量的拟合值为 β_0，在其他车型水平下，对因变量的拟合值都是对 β_0

的调整，即在车型 A，B，C 条件下对因变量的拟合值分别是在 β_0 的基础上加上 β_1，β_2，β_3。

从理论上讲，可以把车型的任意一个水平设定为基准水平。譬如，如果要把车型“A”设定为基准水平，只需把表 1—1 中的虚拟变量 x_1，x_2，x_3 分别用于表示车型 B，C 和 D 即可。

在常见的统计软件中，分类变量的基准水平是根据某种特定顺序确定的。譬如，SAS 软件把按照字母顺序或数值顺序排列在最后的一个水平确定为基准水平，而 R 软件把排列在最前面的一个水平确定为基准水平。对于表 1—1 中的车型变量，如果用 R 中的 model. matrix 函数生成设计矩阵，则车型 A 会被自动确定为基准水平。为了把车型“D”设定为基准水平，可以将其改为“0D”，此时，车型“0D”将排在最前面，从而会被自动设定为基准水平，此时的 R 程序代码及其输出结果如下所示。

```
type = factor(c("A", "B", "C", "0D"))
model.matrix( ~ type)
##   (Intercept)  typeA  typeB  typeC
## 1           1      1      0      0
## 2           1      0      1      0
## 3           1      0      0      1
## 4           1      0      0      0
```

在实际应用中，基准水平应该根据具体需要来选定。为了预测结果的稳定性，通常会把观测值较多的水平确定为基准水平。对基准水平的拟合值等于回归模型的常数项，而对其他水平的拟合值就是对常数项的调整。如果选定的基准水平观测值太少，参数估计结果将缺乏稳定性。不妨考虑一种极端情况，即基准水平没有观测数据，也就是在模型（1.10）中，车型 D 没有观测值。假设实际观测数据如表 1—2 所示，其中 y 表示因变量，x_1，x_2，x_3 分别表示车型 A，B 和 C 的虚拟变量。

表 1—2　　观测数据

y	车型	x_1	x_2	x_3
120	A	1	0	0
230	B	0	1	0
460	C	0	0	1
110	A	1	0	0
240	B	0	1	0
500	C	0	0	1

在表 1—2 中，车型 D 没有观测值，所以回归模型（1.10）的设计矩阵为：

$$\boldsymbol{X}^* = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

显然，上述设计矩阵是一个奇异矩阵，即后面三列之和等于第一列，所以无法求得回归系数的估计值及其标准误。

更加现实的情况是，观测数据中的基准水平只有很少的观测值，此时，设计矩阵将非常接近一个奇异矩阵。在这种情况下，虽然可以求得回归系数的估计值，但它们的标准误会很大，表明模型的精度较低，结果的稳定性欠佳。

1.2.2　交互效应

交互效应是指一个解释变量对因变量的影响与另一个解释变量有关。譬如，在汽车保险中，通常需要预测汽车保险的索赔频率，即平均每辆汽车在一个保险期间发生的期望索赔次数。索赔频率受多种因素的影响，例如，驾驶人的年龄会对汽车保险的索赔频率产生影响，但这种影响还与驾驶人的性别有关，即不同性别的驾驶人，其年龄对索赔频率的影响是不同的，此时就称年龄和性别之间存在交互效应。

假设用 y 表示汽车保险的索赔频率；用 x_1 表示驾驶人的年龄，是一个连续变量；用 x_2 表示驾驶人的性别，是一个虚拟变量，$x_2=1$ 表示男性驾驶员，$x_2=0$ 表示女性驾驶员；用 x_1x_2 表示年龄和性别的交互效应，则线性回归模型的拟合值可以表示为：

$$\mu=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2 \tag{1.11}$$

在有交互效应的线性回归模型中，年龄 x_1 每增加一个单位，对索赔频率拟合值的影响将不再是常数 β_1，而是一个与性别 x_2 有关的值，即为：

$$\frac{\partial y}{\partial x_1}=\beta_1+\beta_3x_2$$

在上式中，当性别 $x_2=0$ 时，年龄 x_1 每增加一个单位，对索赔频率拟合值的影响为 β_1；而当性别 $x_2=1$ 时，年龄 x_1 每增加一个单位，对索赔频率拟合值的影响为 $\beta_1+\beta_3$。

类似地，性别对索赔频率的影响也受驾驶人年龄的影响。令 $x_2=0$，由式（1.11）可得女性驾驶人的索赔频率拟合值为 $\beta_0+\beta_1x_1$。令 $x_2=1$，可得男性驾驶人的索赔频率拟合值为 $\beta_0+\beta_1x_1+\beta_2+\beta_3x_1$。由此可见，男性驾驶人与女性驾驶人的索赔频率拟合值之差 $\beta_2+\beta_3x_1$ 与驾驶人的年龄 x_1 有关。

1.2.3　变量的标准化

线性回归模型中回归系数的估计值依赖于解释变量的度量单位，譬如，如果以元为单位的居民收入作为解释变量，回归系数是 1.24，则以万元为单位时，回归系数将变为 1 240。为了得到唯一的回归系数，同时使得不同解释变量的回归系数之间具有可比性，可以考虑对变量进行标准化处理。

所谓标准化，是指从原始变量中减去其均值后再除以标准差，即

$$\tilde{y}=\frac{y-\bar{y}}{s_y},\ \tilde{x}_j=\frac{x_j-\bar{x}_j}{s_j} \tag{1.12}$$

式中

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i,\ \bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ji}$$

分别是因变量和第 j 个解释变量的样本均值；

$$s_y=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}},\ s_j=\sqrt{\frac{\sum_{i=1}^{n}(x_{ji}-\bar{x}_j)^2}{n-1}}$$

分别是因变量和第 j 个解释变量的样本标准差。

标准化变量的均值为 0，标准差为 1。

假设基于原始数据建立的回归模型为：

$$y_i=\beta_0+\beta_1 x_{1i}+\cdots+\beta_k x_{ki} \tag{1.13}$$

基于标准化以后的数据建立的回归模型为：

$$\tilde{y}_i=\theta_0+\theta_1\tilde{x}_{1i}+\cdots+\theta_k\tilde{x}_{ki} \tag{1.14}$$

则回归系数 $\hat{\beta}_i$ 与 $\hat{\theta}_i$ 之间具有下述关系：

$$\hat{\beta}_i=\frac{s_y}{s_i}\hat{\theta}_i,\ i=1,2,\cdots,k$$

$$\hat{\beta}_0=\bar{y}-\sum_{i=1}^{k}\hat{\beta}_i\bar{x}_i$$

变量经过标准化处理以后，回归系数的解释将与原来的回归模型有所不同。在基于原始数据的回归模型中，$\hat{\beta}_i$ 表示在其他解释变量固定不变的条件下，x_i 每变化一个单位所引致的因变量 y 的变化量。在基于标准化变量的回归模型中，$\hat{\theta}_i$ 表示 x_i 的标准化变量每增加一个单位，标准化的因变量 y 的变化量。

在回归模型（1.14）中，标准化回归系数的绝对值大小度量了解释变量的相对重要性。一个解释变量的标准化回归系数的绝对值越大，表明该解释变量对因变量的影响越大。

1.2.4 变量变换

在最简单的线性回归模型中，通常假设因变量与解释变量之间是线性关系，即解释变量对因变量的影响不随解释变量取值的变化而变化，但在实际问题中，很有可能出现解释变量与因变量之间是非线性关系的情况。

在这种情况下，可以考虑对解释变量进行变换或建立多项式回归模型。对大于零的解释变量进行变换的常用方式包括对数变换和平方根变换，它们都可以压缩解释变量的取值范围，从而有助于改进模型的拟合效果。譬如，假设原始解释变量的取值范围在 1～10 000 之间，则自然对数变换以后的取值范围将在 0～9.21 之间，而平方根变换以后的取值范围将在 1～100 之间。

多项式回归是把一个解释变量的幂变换作为新的解释变量引入回归模型。在多项式回归中，虽然因变量与解释变量的关系是非线性的，但模型的回归系数仍然是线性的，所以多项式回归也属于线性回归模型。

为简化表述，不妨假设只有一个原始解释变量 x，则 m 次多项式回归模型的基本形式如下：

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_m x^m+\varepsilon \tag{1.15}$$

在上式中，x^2，…，x^m 可以看作基于原始解释变量 x 生成的新解释变量。m 越大，多项

式回归模型对观测数据的拟合会越好。如果观测值的个数为 n，则当 $m=n-1$ 时，回归模型中的参数个数将等于观测值的个数，此时，多项式回归模型可以完全拟合观测数据。但是，这种模型只是再现了原始观测数据，预测能力很差，没有实际应用价值。

在实际应用中建立多项式回归模型时，可以从一阶多项式开始，顺序增加多项式的阶数，直至新增加的多项式阶数不再显著为止；也可以首先建立高阶的多项式模型，然后从最高阶项开始，顺序删除不显著的多项式阶数，直至模型中剩余的最高阶项是显著的为止。

在普通的多项式回归中，多项式的阶数每改变一次，就要重新对模型的参数进行一次估计。多项式的阶数不同，参数估计结果也不同。为了克服这种缺陷，可以使用正交多项式回归模型。譬如，在三阶多项式回归模型中，可以把原来的解释变量 x，x^2，x^3 转化为三个新的解释变量 z_1，z_2，z_3，使得 z_1，z_2，z_3 之间是正交的。

三阶正交多项式回归模型可以表示为：

$$y=\beta_0+\beta_1 z_1+\beta_2 z_2+\beta_3 z_3+\varepsilon \tag{1.16}$$

在 R 软件中，用函数 poly 可以构造正交多项式。由于 z_i 之间是正交的，所以在正交多项式回归模型中删除某个 z_i 以后，其他解释变量的回归系数估计值保持不变。

下面通过一组模拟数据来说明多项式回归的建模过程。下面的 R 程序代码模拟了 20 个观测值。模拟数据的真实模型是二次多项式，即

$$y=2+x+x^2+\varepsilon \tag{1.17}$$

基于该组模拟数据，至少可以建立 3 种回归模型：一元线性回归模型、二次多项式回归模型和 19 次多项式回归模型。图 1—1 是应用 3 种回归模型对模拟数据的拟合结果。从图 1—1 可以看出，19 次多项式回归模型可以完美拟合观测数据，二次多项式回归模型较好地反映了观测数据的基本趋势，而一元线性回归模型对观测数据的拟合较差。

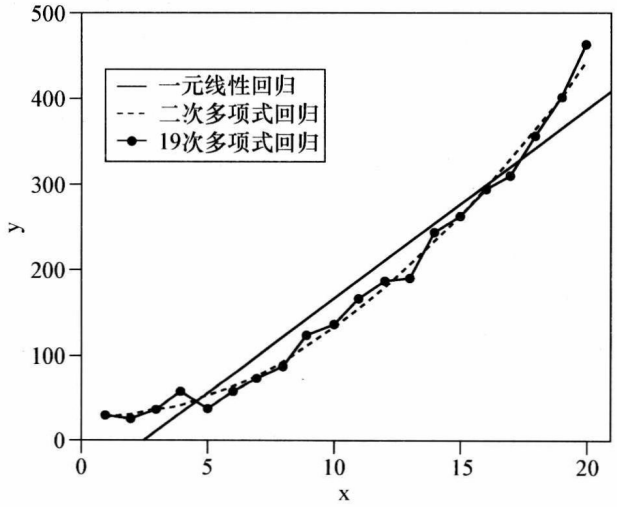

图 1—1　多项式回归的拟合效果

模拟数据和绘制图 1—1 的 R 程序代码如下所示。

```
#设定模拟的种子
set.seed(10)

# 解释变量
x = 1:20

# 模拟因变量
y = 2 + x + x^2 + runif(20) * 50

#一元线性回归
mod1 = lm(y ~ x)

# 二阶正交多项式回归
mod2 = lm(y ~ poly(x, 2))

# 19 阶正交多项式回归
mod3 = lm(y ~ poly(x, 19))

#绘图
plot(y ~ x, yaxs = 'i', pch = 19, ylim = c(0, 500), xlim = c(0, 21), xaxs = 'i', las = 1)
abline(mod1)
points(x, fitted(mod2), col = 2, type = 'l', lty = 2, pch = ' ')
points(x, fitted(mod3), col = 4, type = 'l', lty = 3, pch = ' ')
legend(1, 450, c('一元线性回归', '二次多项式回归', '19 次多项式回归'), lty = c(1, 2,
3), col = c(1, 2, 4))
```

在建立多项式回归模型时，如果解释变量 x 的取值较大，模型中包含 x 的高次项可能会导致计算溢出（overflow），从而使得对其参数的估计值出现下溢（underflow）。解决这一问题的常用方法是对 x 进行下述变换：

$$x^* = \frac{2x - x_{\max} - x_{\min}}{x_{\max} - x_{\min}} \tag{1.18}$$

式中，$x_{\max}$表示最大的观测值；$x_{\min}$表示最小的观测值。容易看出，上述变换使得 x^* 的取值范围落在$-1\sim1$之间。

如果多项式回归模型中包含多个自变量，譬如两个，则其二阶多项式回归模型可以表示为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \tag{1.19}$$

在多项式回归模型中，如果已经包含了高次项，则所有的低次项通常也要保留在模型中。譬如，在二次多项式回归模型中，如果删除不显著的一次项，仅保留二次项，则相应的模型可以表示为：

$$y=\beta_0+\beta_2 x^2+\varepsilon \tag{1.20}$$

在上述模型中，如果对解释变量 x 进行一次线性尺度变换，即把 x 变形为 $x+a$，则上述模型中又会出现 x 的一次项。通常情况下，我们不希望尺度变换改变模型的形式。这就解释了在多项式回归中，为什么一旦保留了高次项，就必须同时保留低次项。同样的道理，对于式（1.19）中的二次多项式回归模型，如果删除交互项 x_1x_2，仅仅保留 x_1^2 和 x_2^2，则对解释变量空间的任何旋转都将导致模型中重新出现交互项。

1.3　参数估计

线性回归模型的参数可以使用最小二乘法或极大似然法进行估计。在正态分布假设下，最小二乘估计等价于极大似然估计（Charnes，1976）。

1.3.1　最小二乘估计

回归参数的最小二乘估计可以在下面的残差平方和最小化的条件下求得（Aldrich，1998）：

$$S=\sum_{i=1}^{n}(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 \tag{1.21}$$

上式的残差平方和用矩阵形式可以表示为：

$$\begin{aligned}S&=(\boldsymbol{y}-\boldsymbol{X\beta})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X\beta})\\&=(\boldsymbol{y}^{\mathrm{T}}-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}})(\boldsymbol{y}-\boldsymbol{X\beta})\\&=\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}-\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X\beta}+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta}\\&=\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}-2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X\beta}+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta}\end{aligned} \tag{1.22}$$

在上式最后一行的变形过程中应用了下述结论：$\boldsymbol{\beta}^{\mathrm{T}}$ 是一个 $1\times(k+1)$ 的矩阵，$\boldsymbol{X}^{\mathrm{T}}$ 是一个 $(k+1)\times n$ 的矩阵，$\boldsymbol{y}$ 是一个 $n\times 1$ 的矩阵，所以，$\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ 是一个 1×1 的矩阵，与其转置矩阵 $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X\beta}$ 相等。

对式（1.22）中的残差平方和 S 关于参数 $\boldsymbol{\beta}$ 求偏导，并令其等于零，即得

$$\frac{\partial S}{\partial \boldsymbol{\beta}}=-2\,\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}+2\,\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta}=\boldsymbol{0}\Rightarrow\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{1.23}$$

解上述方程，可得回归参数的最小二乘估计值为：

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{1.24}$$

前述的推导过程表明，回归参数的最小二乘估计并不需要假设误差项服从正态分布。但是，在进行统计推断时，仍然需要使用到正态分布假设。

1.3.2　极大似然估计

回归参数的极大似然估计需要假设误差项服从正态分布，即 $\varepsilon_i\sim N(0,\ \sigma^2)$。

在正态分布假设下，因变量 y_i 的密度函数为：

$$f(\boldsymbol{y}_i;\ \boldsymbol{\beta},\ \sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2}{2\sigma^2}\right] \tag{1.25}$$

对数似然函数为：