# 第10章

# 统计回归模型

当人们对研究对象的内在特性和各因素间的关系有比较充分的认识时，一般用机理分析方法建立数学模型，本书前面讨论的绝大多数模型都是如此. 如果由于客观事物内部规律的复杂性及人们认识程度的限制，无法分析实际对象内在的因果关系，建立合乎机理规律的数学模型，那么通常的办法是搜集大量的数据，基于对数据的统计分析去建立模型，本章只介绍用途非常广泛的一类随机模型——统计回归模型.

与专门讲述统计方法的教材不同，这里将不涉及回归分析的数学原理和方法，而是通过几个实例讨论如何选择不同类型的模型，以及怎样对软件得到的结果进行分析. 没有学过这部分数学知识的读者只要不追究这些结果的统计学上的意义，就仍然可以在这里学到用回归模型解决实际问题的基本方法.

## 10.1 牙膏的销售量

**问题** 某大型牙膏制造企业为了更好地拓展产品市场，有效地管理库存，公司董事会要求销售部门根据市场调查，找出公司生产的牙膏销售量与销售价格、广告投入等之间的关系，从而预测出在不同价格和广告费用下的销售量. 为此，销售部的研究人员收集了过去30个销售周期(每个销售周期为4周)公司生产的牙膏的销售量、销售价格、投入的广告费用，以及同期其他厂家生产的同类牙膏的市场平均销售价格，见表1. 试根据这些数据建立一个数学模型，分析牙膏销售量与其他因素的关系，为制订价格策略和广告投入策略提供数量依据[13].

**表1 牙膏销售量与销售价格、广告费用等数据**

(其中价格差指其他厂家平均价格与公司销售价格之差)

| 销售周期 | 公司销售价格/元 | 其他厂家平均价格/元 | 广告费用/百万元 | 价格差/元 | 销售量/百万支 |
|---|---|---|---|---|---|
| 1 | 3.85 | 3.80 | 5.50 | -0.05 | 7.38 |
| 2 | 3.75 | 4.00 | 6.75 | 0.25 | 8.51 |

续表

| 销售周期 | 公司销售价格/元 | 其他厂家平均价格/元 | 广告费用/百万元 | 价格差/元 | 销售量/百万支 |
|---|---|---|---|---|---|
| 3 | 3.70 | 4.30 | 7.25 | 0.60 | 9.52 |
| 4 | 3.70 | 3.70 | 5.50 | 0 | 7.50 |
| 5 | 3.60 | 3.85 | 7.00 | 0.25 | 9.33 |
| 6 | 3.60 | 3.80 | 6.50 | 0.20 | 8.28 |
| 7 | 3.60 | 3.75 | 6.75 | 0.15 | 8.75 |
| 8 | 3.80 | 3.85 | 5.25 | 0.05 | 7.87 |
| 9 | 3.80 | 3.65 | 5.25 | -0.15 | 7.10 |
| 10 | 3.85 | 4.00 | 6.00 | 0.15 | 8.00 |
| 11 | 3.90 | 4.10 | 6.50 | 0.20 | 7.89 |
| 12 | 3.90 | 4.00 | 6.25 | 0.10 | 8.15 |
| 13 | 3.70 | 4.10 | 7.00 | 0.40 | 9.10 |
| 14 | 3.75 | 4.20 | 6.90 | 0.45 | 8.86 |
| 15 | 3.75 | 4.10 | 6.80 | 0.35 | 8.90 |
| 16 | 3.80 | 4.10 | 6.80 | 0.30 | 8.87 |
| 17 | 3.70 | 4.20 | 7.10 | 0.50 | 9.26 |
| 18 | 3.80 | 4.30 | 7.00 | 0.50 | 9.00 |
| 19 | 3.70 | 4.10 | 6.80 | 0.40 | 8.75 |
| 20 | 3.80 | 3.75 | 6.50 | -0.05 | 7.95 |
| 21 | 3.80 | 3.75 | 6.25 | -0.05 | 7.65 |
| 22 | 3.75 | 3.65 | 6.00 | -0.10 | 7.27 |
| 23 | 3.70 | 3.90 | 6.50 | 0.20 | 8.00 |
| 24 | 3.55 | 3.65 | 7.00 | 0.10 | 8.50 |
| 25 | 3.60 | 4.10 | 6.80 | 0.50 | 8.75 |
| 26 | 3.65 | 4.25 | 6.80 | 0.60 | 9.21 |
| 27 | 3.70 | 3.65 | 6.50 | -0.05 | 8.27 |
| 28 | 3.75 | 3.75 | 5.75 | 0 | 7.67 |
| 29 | 3.80 | 3.85 | 5.80 | 0.05 | 7.93 |
| 30 | 3.70 | 4.25 | 6.80 | 0.55 | 9.26 |

**分析与假设** 由于牙膏是生活必需品，对大多数顾客来说，在购买同类产品的牙膏时更多地会在意不同品牌之间的价格差异，而不是它们的价格本身. 因

此，在研究各个因素对销售量的影响时，用价格差代替公司销售价格和其他厂家平均价格更为合适.

记牙膏销售量为 $y$，其他厂家平均价格与公司销售价格之差（价格差）为 $x_1$，公司投入的广告费用为 $x_2$，其他厂家平均价格和公司销售价格分别为 $x_3$ 和 $x_4$，$x_1 = x_3 - x_4$. 基于上面的分析，我们仅利用 $x_1$ 和 $x_2$ 来建立 $y$ 的预测模型.

**基本模型** 为了大致地分析 $y$ 与 $x_1$ 和 $x_2$ 的关系，首先利用表 1 的数据分别作出 $y$ 对 $x_1$ 和 $x_2$ 的散点图（见图 1 和图 2 中的圆点）.

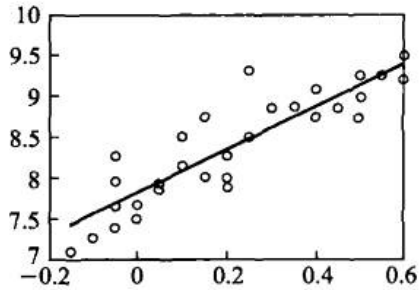


图 1 $y$ 对 $x_1$ 的散点图

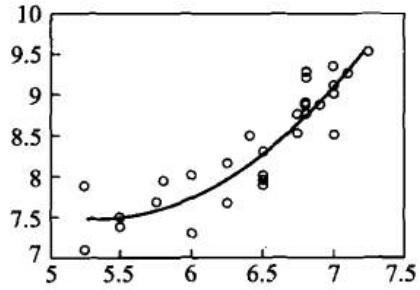


图 2 $y$ 对 $x_2$ 的散点图

从图 1 可以发现，随着 $x_1$ 的增加，$y$ 的值有比较明显的线性增长趋势，图中的直线是用线性模型

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \tag{1}$$

拟合的（其中 $\varepsilon$ 是随机误差）. 而在图 2 中，当 $x_2$ 增大时，$y$ 有向上弯曲增加的趋势，图中的曲线是用二次函数模型

$$y = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \varepsilon \tag{2}$$

拟合的.

综合上面的分析，结合模型（1）和（2）建立如下的回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon \tag{3}$$

（3）式右端的 $x_1$ 和 $x_2$ 称为回归变量（自变量），$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$ 是给定价格差 $x_1$、广告费用 $x_2$ 时，牙膏销售量 $y$ 的平均值，其中的参数 $\beta_0, \beta_1, \beta_2, \beta_3$ 称为回归系数，由表 1 的数据估计，影响 $y$ 的其他因素作用都包含在随机误差 $\varepsilon$ 中. 如果模型选择得合适，$\varepsilon$ 应大致服从均值为 0 的正态分布.

**模型求解** 直接利用 MATLAB 统计工具箱中的命令 regress 求解，使用格式为

```
[b,bint,r,rint,stats] = regress(y,x,alpha)
```

其中输入 $y$ 为模型（3）中 $\boldsymbol{y}$ 的数据（$n$ 维向量，$n = 30$），$x$ 为对应于回归系数 $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ 的数据矩阵 $[\mathbf{1} \quad \boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \boldsymbol{x}_2^2]$（$n \times 4$ 矩阵，其中第 1 列为全 1 向量），*alpha* 为置信水平 $\alpha$（缺省时 $\alpha = 0.05$）；输出 b 为 $\boldsymbol{\beta}$ 的估计值，常记作 $\hat{\boldsymbol{\beta}}$，

bint 为 b 的置信区间，r 为残差向量 $y-x\hat{\beta}$，rint 为 r 的置信区间，stats 为回归模型的检验统计量，有 4 个值，第 1 个是回归方程的决定系数 $R^2$（$R$ 是相关系数），第 2 个是 $F$ 统计量值，第 3 个是与 $F$ 统计量对应的概率值 $p$，第 4 个是剩余方差 $s^2$.

得到模型(3)的回归系数估计值及其置信区间（置信水平 $\alpha=0.05$）、检验统计量 $R^2, F, p, s^2$ 的结果见表 2.

**表 2　模型(3)的计算结果**

| 参数 | 参数估计值 | 参数置信区间 |
|---|---|---|
| $\beta_0$ | 17.324 4 | [5.728 2, 28.920 6] |
| $\beta_1$ | 1.307 0 | [0.682 9, 1.931 1] |
| $\beta_2$ | −3.695 6 | [−7.498 9, 0.107 7] |
| $\beta_3$ | 0.348 6 | [0.037 9, 0.659 4] |
| $R^2=0.905\ 4$　$F=82.940\ 9$　$p<0.000\ 1$　$s^2=0.049\ 0$ | | |

**结果分析**　表 2 显示，$R^2=0.905\ 4$ 指因变量 $y$（销售量）的 90.54% 可由模型确定，$F$ 值远远超过 $F$ 检验的临界值，$p$ 远小于 $\alpha$，因而模型(3)从整体来看是可用的.

表 2 的回归系数给出了模型(3)中 $\beta_0, \beta_1, \beta_2, \beta_3$ 的估计值，即 $\hat{\beta}_0=17.324\ 4$，$\hat{\beta}_1=1.307\ 0$，$\hat{\beta}_2=-3.695\ 6$，$\hat{\beta}_3=0.348\ 6$. 检查它们的置信区间发现，只有 $\beta_2$ 的置信区间包含零点（但区间右端点距零点很近），表明回归变量 $x_2$（对因变量 $y$ 的影响）不是太显著的，但由于 $x_2^2$ 是显著的，我们仍将变量 $x_2$ 保留在模型中.

**销售量预测**　将回归系数的估计值代入模型(3)，即可预测公司未来某个销售周期牙膏的销售量 $y$，预测值记作 $\hat{y}$，得到模型(3)的预测方程

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_2^2 \tag{4}$$

只需知道该销售周期的价格差 $x_1$ 和投入的广告费用 $x_2$，就可以计算预测值 $\hat{y}$.

值得注意的是公司无法直接确定价格差 $x_1$，而只能制定公司该周期的牙膏售价 $x_4$，但是同期其他厂家的平均价格 $x_3$ 一般可以通过分析和预测当时的市场情况及原材料的价格变化等估计出. 模型中引入价格差 $x_1=x_3-x_4$ 作为回归变量，而非 $x_3, x_4$ 的好处在于，公司可以更灵活地来预测产品的销售量（或市场需求量），因为 $x_3$ 的值不是公司所能控制的. 预测时只要调整 $x_4$ 达到设定的回归变量 $x_1$ 的值，比如公司计划在未来的某个销售周期中，维持产品的价格差为 $x_1$

=0.2 元，并将投入 $x_2=6.5$ 百万元的广告费用，则该周期牙膏销售量的估计值为 $\hat{y}=17.324\ 4+1.307\ 0\times0.2-3.695\ 6\times6.5+0.348\ 6\times6.5^2=8.292\ 8$ 百万支.

回归模型的一个重要应用是，对于给定的回归变量的取值，可以以一定的置信度预测因变量的取值范围，即预测区间. 比如当 $x_1=0.2, x_2=6.5$ 时可以算出①，牙膏销售量的置信度为95%的预测区间为[7.823 0, 8.763 6]，它表明在将来的某个销售周期中，如公司维持产品的价格差为 0.2 元，并投入 650 万元的广告费用，那么可以有 95% 的把握保证牙膏的销售量在 7.823 到 8.763 6 百万支之间. 实际操作时，预测上限可以用来作为库存管理的目标值，即公司可以生产(或库存)8.763 6 百万支牙膏来满足该销售周期顾客的需求；预测下限则可以用来较好地把握(或控制)公司的现金流，理由是公司对该周期销售 7.823 百万支牙膏十分自信，如果在该销售周期中公司将牙膏售价定为3.70元，且估计同期其他厂家的平均价格为 3.90 元，那么董事会可以有充分的依据知道公司的牙膏销售额应在 $7.823\times3.7\approx29$ 百万元以上.

**模型改进** 模型(3)中回归变量 $x_1$ 和 $x_2$ 对因变量 $y$ 的影响是相互独立的，即牙膏销售量 $y$ 的均值与广告费用 $x_2$ 的二次关系由回归系数 $\beta_2$ 和 $\beta_3$ 确定，而不依赖于价格差 $x_1$，同样，$y$ 的均值与 $x_1$ 的线性关系由回归系数 $\beta_1$ 确定，不依赖于 $x_2$. 根据直觉和经验可以猜想，$x_1$ 和 $x_2$ 之间的交互作用会对 $y$ 有影响，不妨简单地用 $x_1, x_2$ 的乘积代表它们的交互作用，于是将模型(3)增加一项，得到

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_2^2+\beta_4x_1x_2+\varepsilon \tag{5}$$

在这个模型中，$y$ 的均值与 $x_2$ 的二次关系为 $(\beta_2+\beta_4x_1)x_2+\beta_3x_2^2$，由系数 $\beta_2, \beta_3, \beta_4$ 确定，并依赖于价格差 $x_1$.

下面让我们用表 1 的数据估计模型(5)的系数. 利用 MATLAB 的统计工具箱得到的结果见表 3.

**表 3 模型(5)的计算结果**

| 参数 | 参数估计值 | 参数置信区间 |
|---|---|---|
| $\beta_0$ | 29.113 3 | [13.701 3, 44.525 2] |
| $\beta_1$ | 11.134 2 | [1.977 8, 20.290 6] |
| $\beta_2$ | −7.608 0 | [−12.693 2, −2.522 8] |
| $\beta_3$ | 0.671 2 | [0.253 8, 1.088 7] |
| $\beta_4$ | −1.477 7 | [−2.851 8, −0.103 7] |
| $R^2=0.920\ 9$ | $F=72.777\ 1$ | $p<0.000\ 1$ $s^2=0.042\ 6$ |

① 具体计算参见[91](4.77)～(4.79)式，用 MATLAB 统计工具箱中现成的程序结果与此不同.

表3与表2的结果相比，$R^2$ 有所提高，说明模型(5)比模型(3)有所改进. 并且，所有参数的置信区间，特别是 $x_1, x_2$ 的交互作用项 $x_1x_2$ 的系数 $\beta_4$ 的置信区间不包含零点，所以有理由相信模型(5)比模型(3)更符合实际.

用模型(5)对公司的牙膏销售量作预测. 仍设在某个销售周期中，维持产品的价格差 $x_1 = 0.2$ 元，并将投入 $x_2 = 6.5$ 百万元的广告费用，则该周期牙膏销售量 $y$ 的估计值为 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 + \hat{\beta}_4 x_1 x_2 = 29.113\,3 + 11.134\,2 \times 0.2 - 7.608\,0 \times 6.5 + 0.671\,2 \times 6.5^2 - 1.477\,7 \times 0.2 \times 6.5 = 8.325\,3$ 百万支，置信度为95%的预测区间为[7.895 3, 8.759 2]，与模型(3)的结果相比，$\hat{y}$ 略有增加，而预测区间长度短一些.

在保持广告费用 $x_2 = 6.5$ 百万元不变的条件下，分别对模型(3)和(5)中牙膏销售量的均值 $\hat{y}$ 与价格差 $x_1$ 的关系作图，见图3和图4.

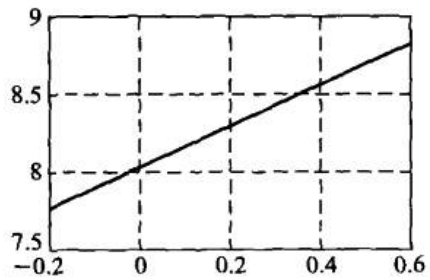


图3 模型(3)$\hat{y}$ 与 $x_1$ 的关系

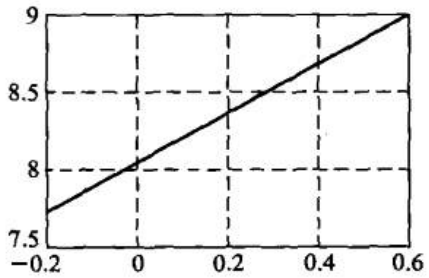


图4 模型(5)$\hat{y}$ 与 $x_1$ 的关系

在保持价格差 $x_1 = 0.2$ 元不变的条件下，分别对模型(3)和(5)中牙膏销售量的均值 $\hat{y}$ 与广告费用 $x_2$ 的关系作图，见图5和图6.

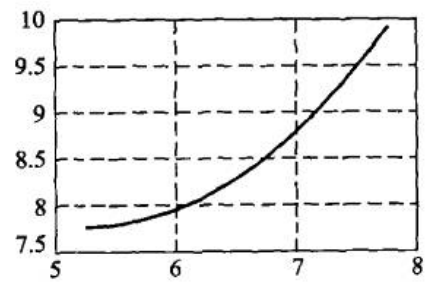


图5 模型(3)$\hat{y}$ 与 $x_2$ 的关系

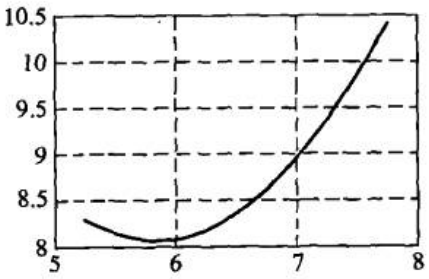


图6 模型(5)$\hat{y}$ 与 $x_2$ 的关系

可以看出，交互作用项 $x_1x_2$ 加入模型，对 $\hat{y}$ 与 $x_1$ 的关系稍有影响，而 $\hat{y}$ 与 $x_2$ 的关系有较大变化，当 $x_2 < 6$ 时 $\hat{y}$ 出现下降，$x_2 > 6$ 以后 $\hat{y}$ 上升则快得多.

**进一步的讨论** 为进一步了解 $x_1$ 和 $x_2$ 之间的交互作用，考察模型(5)的预测方程

$$\hat{y} = 29.113\,3 + 11.134\,2x_1 - 7.608\,0x_2 + 0.671\,2x_2^2 - 1.477\,7x_1x_2 \qquad (6)$$

如果取价格差 $x_1=0.1$ 元，代入(6)可得

$$\hat{y}\Big|_{x_1=0.1}=30.226\,7-7.755\,8x_2+0.671\,2x_2^2 \tag{7}$$

再取 $x_1=0.3$ 元，代入(6)得

$$\hat{y}\Big|_{x_1=0.3}=32.453\,6-8.051\,3x_2+0.671\,2x_2^2 \tag{8}$$

它们均为 $x_2$ 的二次函数，其图形见图 7，且

$$\hat{y}\Big|_{x_1=0.3}-\hat{y}\Big|_{x_1=0.1}=2.226\,9-0.295\,5x_2 \tag{9}$$

由(9)式可得，当 $x_2<7.536\,0$ 时，总有 $\hat{y}\Big|_{x_1=0.3}>\hat{y}\Big|_{x_1=0.1}$，即若广告费用不超过大约 7.5 百万元，价格差定在 0.3 元时的销售量，比价格差定在 0.1 元的大，也就是说，这时的价格优势会使销售量增加.

由图 7 还可以发现，虽然广告投入的增加会使销售量增加(只要广告费用超过大约 6 百万元)，但价格差较小时增加的速率要更大些. 这些现象都是由于引入了交互作用项 $x_1x_2$ 后产生的. 价格差较大时，许多消费者是受价格的驱动来购买公司的产品，所以可以较少地依赖广告投入的增加来提高销售量. 价格差较小时，则更需要靠广告来吸引更多的顾客.

另外，当公司牙膏的售价在市场中明显处于弱势时，$x_1$ 和 $x_2$ 之间的交互作用项不见得就是乘积项 $x_1x_2$ 了，可能出现其他形式的组合.

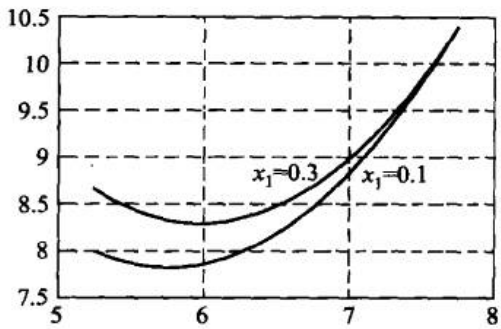


图 7　$\hat{y}$ 与 $x_2$ 的关系((7)与(8)的图形)

**完全二次多项式模型**　与 $x_1$ 和 $x_2$ 的完全二次多项式模型

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2+\beta_4x_1^2+\beta_5x_2^2+\varepsilon \tag{10}$$

相比，模型(5)只少 $x_i^2$ 项，我们不妨增加这一项，建立模型(10). 这样做的好处之一是 MATLAB 统计工具箱中有直接的命令 rstool 求解，并且以交互式画面给出 $y$ 的估计值 $\hat{y}$ 和预测区间. 这个命令的输出如图 8，从左下方的输出 Export 可以得到模型(10)的回归系数的估计值为

$$\hat{\beta}=(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\beta}_4,\hat{\beta}_5)=(32.098\,4,14.743\,6,$$

$-8.636\ 7, -2.103\ 8, 1.107\ 4, 0.759\ 4)$

用鼠标移动交互式画面中的十字线，或在图下方的窗口内输入，可改变 $x_1$ 和 $x_2$ 的数值，图中当 $x_1 = 0.2, x_2 = 6.5$ 时，左边的窗口显示 $\hat{y} = 8.302\ 9$，预测区间为 $8.302\ 9 \pm 0.255\ 8 = [8.047\ 1, 8.558\ 7]$. 这些结果与模型(5)相差不大.

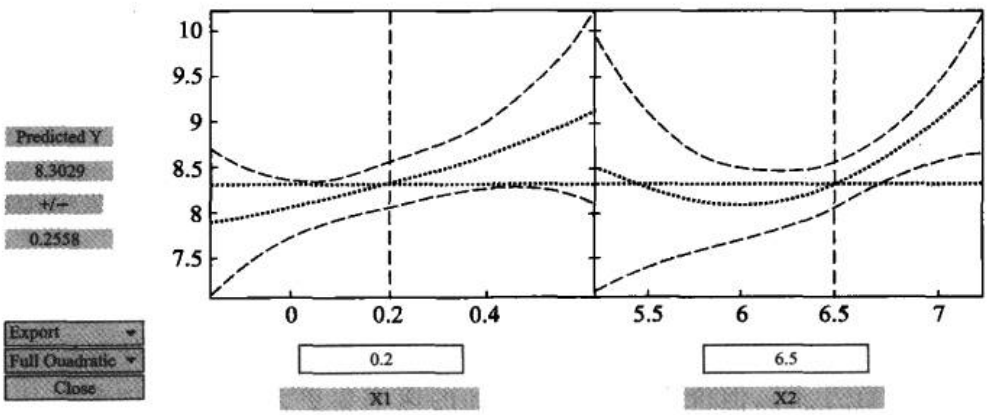


图 8　完全二次多项式模型(10)的输出

**评注**　从这个实例我们看到，建立回归模型可以先根据已知的数据，从常识和经验进行分析，辅以作图(如图 1，图 2 的散点图)，决定取哪几个回归变量，及它们的函数形式(如线性的、二次的). 用软件(如 MATLAB 统计工具箱)求解后，作统计分析：$R^2$，$F$，$p$ 值，$s^2$ 的大小是对模型整体的评价，每个回归系数置信区间是否包含零点，可以用来检验对应的回归变量对因变量的影响是否显著(若包含零点则不显著). 如果对结果不够满意，则应改进模型，如添加二次项、交互项等.

对因变量进行预测，经常是建立回归模型的主要目的之一，本节提供了预测的方法，以及对结果作进一步讨论的实例.

## 10.2　软件开发人员的薪金

**问题**　一家高技术公司人事部门为研究软件开发人员的薪金与他们的资历、管理责任、教育程度等因素之间的关系，要建立一个数学模型，以便分析公司人事策略的合理性，并作为新聘用人员薪金的参考. 他们认为目前公司人员的薪金总体上是合理的，可以作为建模的依据，于是调查了 46 名软件开发人员的档案资料，如表 1，其中资历一列指从事专业工作的年数，管理一列中 1 表示管理人员，0 表示非管理人员，教育一列中 1 表示中学程度，2 表示大学程度，3 表示更高程度(研究生)[19].

表 1　软件开发人员的薪金与他们的资历、管理责任、教育程度之间的关系

| 编号 | 薪金 | 资历 | 管理 | 教育 | 编号 | 薪金 | 资历 | 管理 | 教育 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 13 876 | 1 | 1 | 1 | 24 | 22 884 | 6 | 1 | 2 |
| 2 | 11 608 | 1 | 0 | 3 | 25 | 16 978 | 7 | 1 | 1 |
| 3 | 18 701 | 1 | 1 | 3 | 26 | 14 803 | 8 | 0 | 2 |
| 4 | 11 283 | 1 | 0 | 2 | 27 | 17 404 | 8 | 1 | 1 |
| 5 | 11 767 | 1 | 0 | 3 | 28 | 22 184 | 8 | 1 | 3 |
| 6 | 20 872 | 2 | 1 | 2 | 29 | 13 548 | 8 | 0 | 1 |
| 7 | 11 772 | 2 | 0 | 2 | 30 | 14 467 | 10 | 0 | 1 |
| 8 | 10 535 | 2 | 0 | 1 | 31 | 15 942 | 10 | 0 | 2 |
| 9 | 12 195 | 2 | 0 | 3 | 32 | 23 174 | 10 | 1 | 3 |
| 10 | 12 313 | 3 | 0 | 2 | 33 | 23 780 | 10 | 1 | 2 |
| 11 | 14 975 | 3 | 1 | 1 | 34 | 25 410 | 11 | 1 | 2 |
| 12 | 21 371 | 3 | 1 | 2 | 35 | 14 861 | 11 | 0 | 1 |
| 13 | 19 800 | 3 | 1 | 3 | 36 | 16 882 | 12 | 0 | 2 |
| 14 | 11 417 | 4 | 0 | 1 | 37 | 24 170 | 12 | 1 | 3 |
| 15 | 20 263 | 4 | 1 | 3 | 38 | 15 990 | 13 | 0 | 1 |
| 16 | 13 231 | 4 | 0 | 3 | 39 | 26 330 | 13 | 1 | 2 |
| 17 | 12 884 | 4 | 0 | 2 | 40 | 17 949 | 14 | 0 | 2 |
| 18 | 13 245 | 5 | 0 | 2 | 41 | 25 685 | 15 | 1 | 3 |
| 19 | 13 677 | 5 | 0 | 3 | 42 | 27 837 | 16 | 1 | 2 |
| 20 | 15 965 | 5 | 1 | 1 | 43 | 18 838 | 16 | 0 | 2 |
| 21 | 12 366 | 6 | 0 | 1 | 44 | 17 483 | 16 | 0 | 1 |
| 22 | 21 352 | 6 | 1 | 3 | 45 | 19 207 | 17 | 0 | 2 |
| 23 | 13 839 | 6 | 0 | 2 | 46 | 19 346 | 20 | 0 | 1 |

**分析与假设**　按照常识，薪金自然随着资历（年）的增长而增加，管理人员的薪金应高于非管理人员，教育程度越高薪金也越高. 薪金记作 $y$，资历（年）记作 $x_1$，为了表示是否管理人员，定义

$$x_2 = \begin{cases} 1, \text{管理人员} \\ 0, \text{非管理人员} \end{cases}$$

为了表示3种教育程度，定义

$$x_3 = \begin{cases} 1, \text{中学} \\ 0, \text{其他} \end{cases} \qquad x_4 = \begin{cases} 1, \text{大学} \\ 0, \text{其他} \end{cases}$$

这样，中学用 $x_3 = 1, x_4 = 0$ 表示，大学用 $x_3 = 0, x_4 = 1$ 表示，研究生则用 $x_3 = 0$，$x_4 = 0$ 表示.

为简单起见，我们假定资历(年)对薪金的作用是线性的，即资历每加一年，薪金的增长是常数；管理责任、教育程度、资历诸因素之间没有交互作用，建立线性回归模型.

**基本模型** 薪金 $y$ 与资历 $x_1$，管理责任 $x_2$，教育程度 $x_3, x_4$ 之间的多元线性回归模型为

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + \varepsilon \tag{1}$$

其中 $a_0$，$a_1$，…，$a_4$ 是待估计的回归系数，$\varepsilon$ 是随机误差.

利用MATLAB的统计工具箱可以得到回归系数及其置信区间(置信水平 $\alpha = 0.05$)、检验统计量 $R^2, F, p, s^2$ 的结果，见表2.

**表2 模型(1)的计算结果**

| 参数 | 参数估计值 | 参数置信区间 |
|---|---|---|
| $a_0$ | 11 032 | [10 258, 11 807] |
| $a_1$ | 546 | [484, 608] |
| $a_2$ | 6 883 | [6 248, 7 517] |
| $a_3$ | −2 994 | [−3 826, −2 162] |
| $a_4$ | 148 | [−636, 931] |
| $R^2 = 0.957$ $F = 226$ $p < 0.000\ 1$ $s^2 = 1.057 \times 10^6$ | | |

**结果分析** 从表2知 $R^2 = 0.957$，即因变量(薪金)的95.7%可由模型确定，$F$ 值远远超过 $F$ 检验的临界值，$p$ 远小于 $\alpha$，因而模型(1)从整体来看是可用的. 比如，利用模型可以估计(或预测)一个大学毕业、有2年资历、非管理人员的薪金为

$$\hat{y} = \hat{a}_0 + \hat{a}_1 \times 2 + \hat{a}_2 \times 0 + \hat{a}_3 \times 0 + \hat{a}_4 \times 1 = 12\ 272$$

模型中各个回归系数的含义可初步解释如下：$x_1$ 的系数为546，说明资历每增加1年，薪金增长546；$x_2$ 的系数为6 883，说明管理人员的薪金比非管理人员

多 6 883；$x_3$ 的系数为 -2 994，说明中学程度的薪金比研究生少 2 994；$x_4$ 的系数为 148，说明大学程度的薪金比研究生多 148，但是应该注意到 $a_4$ 的置信区间包含零点，所以这个系数的解释是不可靠的.

需要指出，以上解释是就平均值来说，并且，一个因素改变引起的因变量的变化量，都是在其他因素不变的条件下才成立的.

**进一步的讨论** $a_4$ 的置信区间包含零点，说明基本模型(1)存在缺点. 为寻找改进的方向，常用残差分析方法(残差 $\varepsilon$ 指薪金的实际值 $y$ 与用模型估计的薪金 $\hat{y}$ 之差，是模型(1)中随机误差 $\varepsilon$ 的估计值，这里用了同一个符号). 我们将影响因素分成资历与管理 - 教育组合两类，管理 - 教育组合的定义如表 3.

**表 3　管理 - 教育组合**

| 组合 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 管理 | 0 | 1 | 0 | 1 | 0 | 1 |
| 教育 | 1 | 1 | 2 | 2 | 3 | 3 |

为了对残差进行分析，图 1 给出 $\varepsilon$ 与资历 $x_1$ 的关系，图 2 给出 $\varepsilon$ 与管理 $x_2$ - 教育 $x_3$，$x_4$ 组合间的关系.

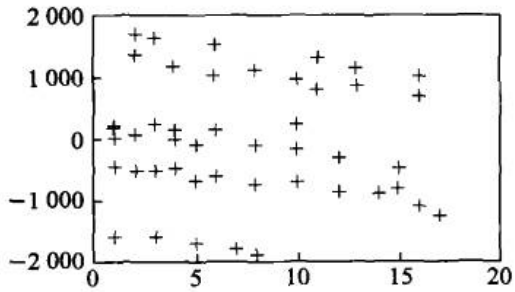


图 1　模型(1)$\varepsilon$ 与 $x_1$ 的关系

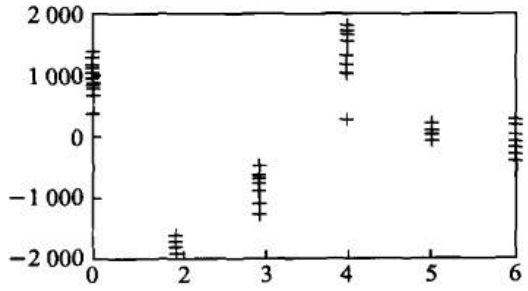


图 2　模型(1)$\varepsilon$ 与 $x_2 - x_3, x_4$ 组合的关系

从图 1 看，残差大概分成 3 个水平，这是由于 6 种管理 - 教育组合混在一起，在模型中未被正确反映的结果；从图 2 看，对于前 4 个管理 - 教育组合，残差或者全为正，或者全为负，也表明管理—教育组合在模型中处理不当.

在模型(1)中管理责任和教育程度是分别起作用的，事实上，二者可能起着交互作用，如大学程度的管理人员的薪金会比二者分别的薪金之和高一点.

以上分析提示我们，应在基本模型(1)中增加管理 $x_2$ 与教育 $x_3$，$x_4$ 的交互项，建立新的回归模型.

**更好的模型**　增加 $x_2$ 与 $x_3$，$x_4$ 的交互项后，模型记作

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_2x_3 + a_6x_2x_4 + \varepsilon \tag{2}$$

利用 MATLAB 的统计工具箱得到的结果如表 4.

**表 4　模型(2)的计算结果**

| 参数 | 参数估计值 | 参数置信区间 |
|---|---|---|
| $a_0$ | 11 204 | [11 044, 11 363] |
| $a_1$ | 497 | [486, 508] |
| $a_2$ | 7 048 | [6 841, 7 255] |
| $a_3$ | −1 727 | [−1 939, −1 514] |
| $a_4$ | −348 | [−545, −152] |
| $a_5$ | −3 071 | [−3 372, −2 769] |
| $a_6$ | 1 836 | [1 571, 2 101] |
| $R^2 = 0.998\ 8$　　$F = 5\ 545$　　$p < 0.000\ 1$　　$s^2 = 3.004\ 7 \times 10^4$ | | |

由表 4 可知,模型(2)的 $R^2$ 和 $F$ 值都比模型(1)有所改进,并且所有回归系数的置信区间都不含零点,表明模型(2)是完全可用的.

与模型(1)类似,作模型(2)的两个残差分析图(图 3,图 4),可以看出,已经消除了图 1、图 2 中的不正常现象,这也说明了模型(2)的适用性.

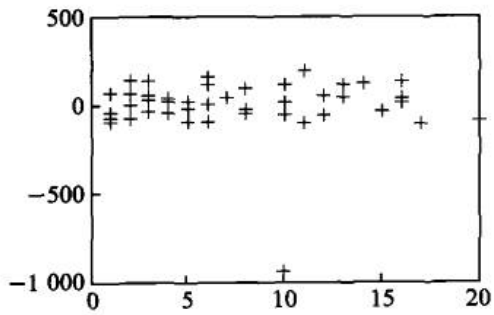


图 3　模型(2)$\varepsilon$ 与 $x_1$ 的关系

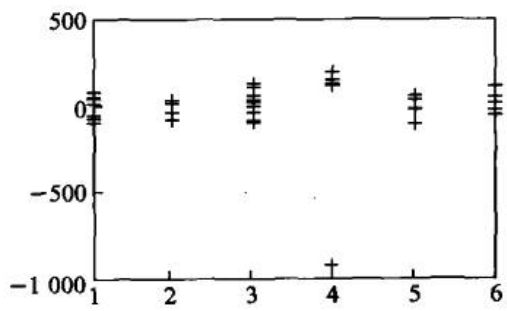


图 4　模型(2)$\varepsilon$ 与 $x_2 - x_3, x_4$ 组合的关系

从图 3、图 4 还可以发现一个异常点:具有 10 年资历、大学程度的管理人员(从表 1 可以查出是 33 号),他的实际薪金明显地低于模型的估计值,也明显低于与他有类似经历的其他人的薪金. 这可能是由我们未知的原因造成的. 为了使个别的数据不致影响整个模型,应该将这个异常数据去掉,对模型(2)重新估计回归系数,得到的结果如表 5,残差分析图见图 5,图 6. 可以看出,去掉异常数据后结果又有改善.

表 5　模型(2)去掉异常数据后的计算结果

| 参数 | 参数估计值 | 参数置信区间 |
|---|---|---|
| $a_0$ | 11 200 | [11 139, 11 261] |
| $a_1$ | 498 | [494, 503] |
| $a_2$ | 7 041 | [6 962, 7 120] |
| $a_3$ | -1 737 | [-1 818, -1 656] |
| $a_4$ | -356 | [-431, -281] |
| $a_5$ | -3 056 | [-3 171, -2 942] |
| $a_6$ | 1 997 | [1 894, 2 100] |
| $R^2 = 0.999\ 8$　$F = 36\ 701$　$p < 0.000\ 1$　$s^2 = 4.347 \times 10^3$ | | |

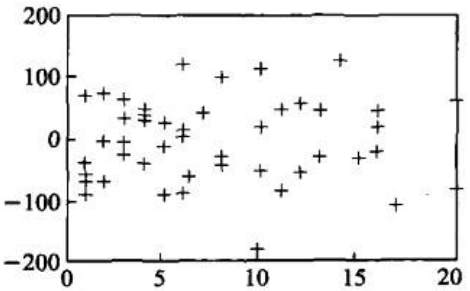


图 5　模型(2)去掉异常数据后 $\varepsilon$ 与 $x_1$ 的关系

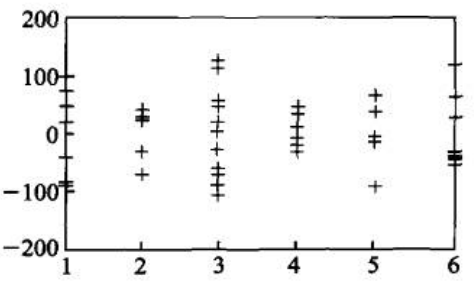


图 6　模型(2)去掉异常数据后 $\varepsilon$ 与 $x_2 - x_3, x_4$ 组合的关系

**模型应用**　对于回归模型(2),用去掉异常数据(33 号)后估计出的系数,得到的结果是满意的. 作为这个模型的应用之一,不妨用它来"制订"6 种管理 - 教育组合人员的"基础"薪金(即资历为零的薪金,当然,这也是平均意义上的). 利用模型(2)和表 5 容易得到表 6.

表 6　6 种管理 - 教育组合人员的"基础"薪金

| 组合 | 管理 | 教育 | 系数 | "基础"薪金 |
|---|---|---|---|---|
| 1 | 0 | 1 | $a_0 + a_3$ | 9 463 |
| 2 | 1 | 1 | $a_0 + a_2 + a_3 + a_5$ | 13 448 |
| 3 | 0 | 2 | $a_0 + a_4$ | 10 844 |
| 4 | 1 | 2 | $a_0 + a_2 + a_4 + a_6$ | 19 882 |
| 5 | 0 | 3 | $a_0$ | 11 200 |
| 6 | 1 | 3 | $a_0 + a_2$ | 18 241 |

可以看出,大学程度的管理人员的薪金比研究生程度的管理人员的薪金高,而大学程度的非管理人员的薪金比研究生程度的非管理人员的薪金略低. 当然,这是根据这家公司实际数据建立的模型得到的结果,并不具普遍性.

**评注** 从建立回归模型的角度我们通过本例介绍了以下内容:

1. 对于影响因变量的定性因素(管理、教育),可以引入0-1变量来处理,0-1变量的个数可比定性因素的水平少1(如教育程度有3个水平,引入2个0-1变量).

2. 用残差分析方法可以发现模型的缺陷,引入交互作用项常常能够给予改善.

3. 若发现异常值应剔除,有助于结果的合理性.

在本例中我们由简到繁,先分别引进管理和教育因素,再进入交互项. 实际上,可以直接对6种管理-教育组合引入5个0-1变量,读者不妨试一下,看看结果如何.

## 10.3 酶促反应

**背景和问题** 酶是一种具有特异性的高效生物催化剂,绝大多数的酶是活细胞产生的蛋白质. 酶的催化条件温和,在常温、常压下即可进行. 酶催化的反应称为酶促反应,要比相应的非催化反应快$10^3 \sim 10^{17}$倍. 酶促反应动力学简称酶动力学,主要研究酶促反应的速度与底物(即反应物)浓度以及其他因素的关系. 在底物浓度很低时酶促反应是一级反应;当底物浓度处于中间范围时,是混合级反应;当底物浓度增加时,向零级反应过渡.

某生化系学生为了研究嘌呤霉素在某项酶促反应中对反应速度与底物浓度之间关系的影响,设计了两个实验,一个实验中所使用的酶是经过嘌呤霉素处理的,而另一个实验所用的酶是未经嘌呤霉素处理过的,所得的实验数据见表1. 试根据问题的背景和这些数据建立一个合适的数学模型,来反映这项酶促反应的速度与底物浓度以及嘌呤霉素处理与否之间的关系[9].

**表1 嘌呤霉素实验中的反应速度与底物浓度数据**

| 底物浓度/ppm | | 0.02 | | 0.06 | | 0.11 | | 0.22 | | 0.56 | | 1.10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 反应速度 | 处理 | 76 | 47 | 97 | 107 | 123 | 139 | 159 | 152 | 191 | 201 | 207 | 200 |
| | 未处理 | 67 | 51 | 84 | 86 | 98 | 115 | 131 | 124 | 144 | 158 | 160 | — |

注:1 ppm = 0.001‰.

**分析与假设** 记酶促反应的速度为$y$,底物浓度为$x$,二者之间的关系写作$y=f(x,\boldsymbol{\beta})$,其中$\boldsymbol{\beta}$为参数. 由酶促反应的基本性质可知,当底物浓度较小时,反

应速度大致与浓度成正比(即一级反应);而当底物浓度很大,渐进饱和时,反应速度将趋于一个固定值——最终反应速度(即零级反应).下面的两个简单模型具有这种性质:

Michaelis-Menten 模型

$$y = f(x,\boldsymbol{\beta}) = \frac{\beta_1 x}{\beta_2 + x} \tag{1}$$

指数增长模型

$$y = f(x,\boldsymbol{\beta}) = \beta_1(1 - e^{-\beta_2 x}) \tag{2}$$

图 1 和图 2 分别是表 1 给出的经过嘌呤霉素处理和未经处理的反应速度 $y$ 与底物浓度 $x$ 的散点图,可以知道,模型(1),(2)与实际数据得到的散点图是大致符合的.下面只对模型(1)进行详细的分析,将模型(2)留给有兴趣的读者(习题 4).

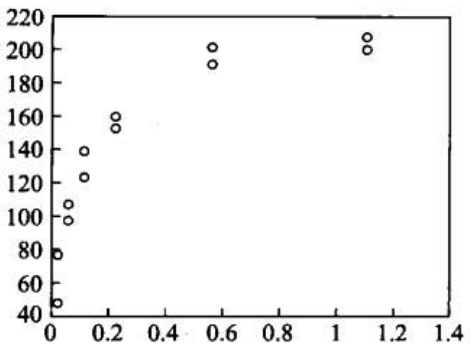


图 1 $y$ 对 $x$(经处理)的散点图

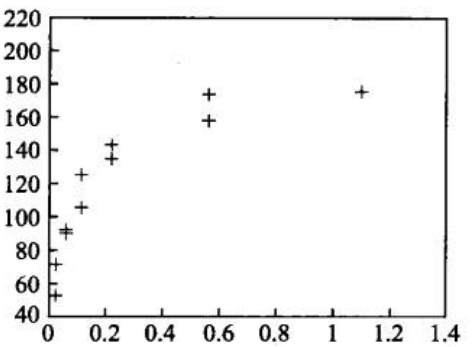


图 2 $y$ 对 $x$(未经处理)的散点图

首先对经过嘌呤霉素处理的实验数据进行分析(未经处理的数据可同样分析),在此基础上,再来讨论是否有更一般的模型来统一刻画处理前后的数据,进而揭示其中的联系.

**线性化模型** 模型(1)对参数 $\boldsymbol{\beta} = (\beta_1, \beta_2)$ 是非线性的,但是可以通过下面的变量代换化为线性模型

$$\frac{1}{y} = \frac{1}{\beta_1} + \frac{\beta_2}{\beta_1}\frac{1}{x} = \theta_1 + \theta_2 u \tag{3}$$

模型(3)中的因变量 $1/y$ 对新的参数 $\boldsymbol{\theta} = (\theta_1, \theta_2)$ 是线性的.

对经过嘌呤霉素处理的实验数据,作出反应速度的倒数 $1/y$ 与底物浓度的倒数 $u = 1/x$ 的散点图(图 3),可以发现在 $1/x$ 较小时有很好的线性趋势,而 $1/x$ 较大时则出现很大的起落.

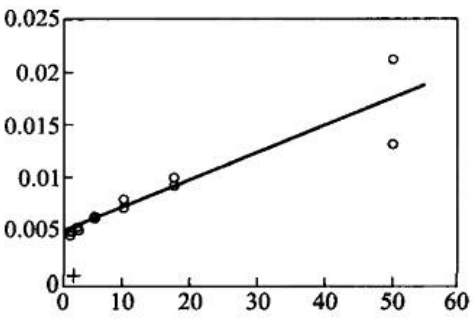


图 3 $1/y$ 与 $1/x$ 的散点图和回归直线

如果单从线性回归模型的角度作计算，很容易得到线性化模型(3)的参数 $\theta_1,\theta_2$ 的估计和其他统计结果(见表2)以及 $1/y$ 与 $1/x$ 的拟合图(图3). 再根据(3)式中 $\beta$ 与 $\theta$ 的关系得到 $\beta_1=1/\theta_1,\beta_2=\theta_2/\theta_1$，从而可以算出 $\beta_1$ 和 $\beta_2$ 的估计值分别为 $\hat{\beta}_1=195.802\ 0$ 和 $\hat{\beta}_2=0.048\ 40$.

**表2　线性化模型(3)参数的估计结果**

| 参数 | 参数估计值($\times10^{-3}$) | 参数置信区间($\times10^{-3}$) |
|---|---|---|
| $\theta_1$ | 5.107 2 | [3.538 6, 6.675 8] |
| $\theta_2$ | 0.247 2 | [0.175 7, 0.318 8] |
| $R^2=0.855\ 7$　$F=59.297\ 5$　$p<0.000\ 1$　$s^2=3.580\ 6\times10^{-6}$ | | |

将经过线性化变换后最终得到的 $\beta$ 值代入原模型(1)，得到与原始数据比较的拟合图(图4). 可以发现，在 $x$ 较大时 $y$ 的预测值要比实际数据小，这是因为在对线性化模型作参数估计时，底物浓度 $x$ 较低($1/x$ 很大)的数据在很大程度上控制了回归参数的确定，从而使得对底物浓度 $x$ 较高数据的拟合，出现较大的偏差.

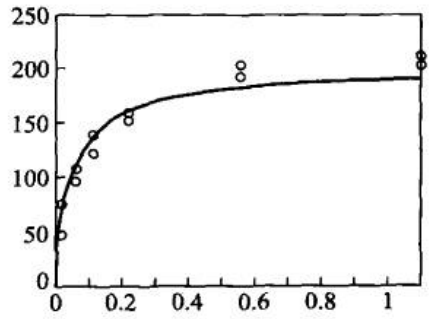

图4　用线性化得到的原始数据拟合图

为了解决线性化模型中拟合欠佳的问题，我们直接考虑非线性模型(1).

**非线性模型及求解**　可以用非线性回归的方法直接估计模型(1)中的参数 $\beta_1$ 和 $\beta_2$. 模型的求解可利用 MATLAB 统计工具箱中的命令进行，使用格式为

```
[beta,R,J]=nlinfit(x,y,'model',beta0)
```

其中输入 x 为自变量数据矩阵，每列一个变量；y 为因变量数据向量；model 为模型的 M 函数文件名，M 函数形式为 y =f(beta,x)，beta 为待估计参数；beta0 为给定的参数初值. 输出 beta 为参数的估计值，R 为残差，J 为用于估计预测误差的 Jacobi 矩阵. 参数 beta 的置信区间用命令 nlparci(beta,R,J)得到.

我们用线性化模型(3)得到的 $\beta$ 作为非线性模型参数估计的初始迭代值，将实际数据 $x,y$ 输入后执行以下程序：

```
beta0=[195.802  0.0484];
[beta,R,J]=nlinfit(x,y,'huaxue',beta0);
betaci=nlparci(beta,R,J);
beta,betaci
```

```
yy = beta(1) * x./(beta(2) + x);
plot(x,y,'o',x,yy,'+'),pause
nlintool(x,y,'huaxue',beta)

function yhat = huaxue(beta,x)
yhat = beta(1) * x./(beta(2) + x);
```

得到的数值结果见表 3.

**表 3　模型(1)参数的估计结果**

| 参数 | 参数估计值 | 参数置信区间 |
|---|---|---|
| $\beta_1$ | 212.681 8 | [197.202 8, 228.160 8] |
| $\beta_2$ | 0.064 12 | [0.046 7, 0.082 57] |

拟合的结果直接画在原始数据图(图 5)上. 程序中的 nlintool 用于给出一个交互式画面(图 6),拖动画面中的十字线可以改变自变量 $x$ 的取值,直接得到因变量 $y$ 的预测值和预测区间,同时通过左下方 Export 下拉式菜单,可输出模型的统计结果,如剩余标准差等,本例中剩余标准差 $s = 10.933\ 7$.

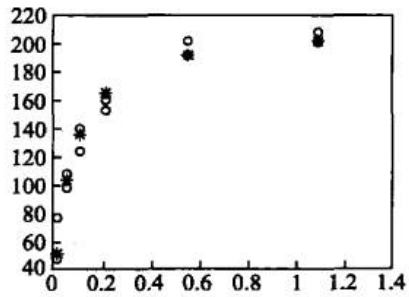


图 5　模型(1)的预测图

(o—原始数据;+—拟合结果)

从上面的结果可以知道,对经过嘌呤霉素处理的实验数据,在用 Michaelis-Menten 模型(1)进行回归分析时,最终反应速度为 $\hat{\beta}_1 = 212.681\ 8$. 还容易得到,反应的"半速度点"(达到最终反应速度一半时的底物浓度 $x$ 值)恰为 $\hat{\beta}_2 = 0.064\ 12$. 以上结果对这样一个经过设计的实验(每个底物浓度做两次实验)已经很好地达到了要求.

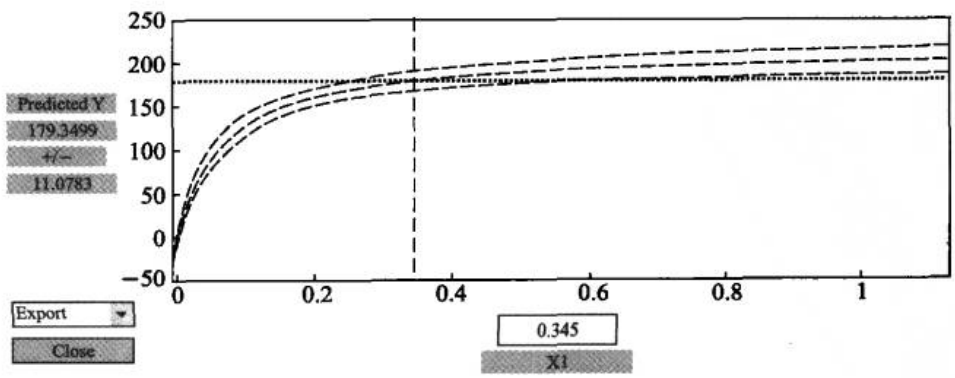


图 6　模型(1)的预测及结果输出

**混合反应模型** 酶动力学知识告诉我们，酶促反应的速度依赖于底物浓度，并且可以假定，嘌呤霉素的处理会影响最终反应速度参数 $\beta_1$，而基本不影响半速度参数 $\beta_2$. 表 1 的数据（图 1、图 2 更为明显）也印证了这种看法. 模型（1）的形式可以分别描述经过嘌呤霉素处理和未经处理的反应速度与底物浓度的关系（两个模型的参数 $\boldsymbol{\beta}$ 会不同），为了在同一个模型中考虑嘌呤霉素处理的影响，我们采用对未经嘌呤霉素处理的模型附加增量的方法，考察混合反应模型

$$y = f(\boldsymbol{x}, \boldsymbol{\beta}) = \frac{(\beta_1 + \gamma_1 x_2) x_1}{(\beta_2 + \gamma_2 x_2) + x_1} \tag{4}$$

其中自变量 $x_1$ 为底物浓度（即模型（3）中的 $x$），$x_2$ 为一示性变量（0 - 1 变量），用来表示是否经嘌呤霉素处理，令 $x_2 = 1$ 表示经过处理，$x_2 = 0$ 表示未经处理；参数 $\beta_1$ 是未经处理的最终反应速度，$\gamma_1$ 是经处理后最终反应速度的增长值，$\beta_2$ 是未经处理的反应的半速度点，$\gamma_2$ 是经处理后反应的半速度点的增长值（为一般化起见，这里假定嘌呤霉素的处理也会影响半速度点）.

**混合模型的求解和分析** 仍用 MATLAB 统计工具箱中的命令 nlinfit 来计算模型（4）的回归系数 $\beta_1$，$\beta_2$，$\gamma_1$ 和 $\gamma_2$. 为了给出合适的初始迭代值，从实验数据我们注意到，未经处理的反应速度的最大实验值为 160，经处理的最大实验值为 207，于是可取参数初值 $\beta_1^0 = 170$，$\gamma_1^0 = 60$；又从数据可大致估计未经处理的半速度点约为 0.05，经处理的半速度点约为 0.06，我们取 $\beta_2^0 = 0.05$，$\gamma_2^0 = 0.01$.

与模型（1）的编程计算相似，得到混合模型（4）的回归系数的估计值与其置信区间（表 4）、拟合结果（图 7）、残差图（图 8），及预测和结果输出图（图 9），模型的剩余标准差 $s = 10.400\ 0$.

**表 4 模型（4）参数的估计结果**

| 参数 | 参数估计值 | 参数置信区间 |
|---|---|---|
| $\beta_1$ | 160.280 2 | [145.846 6, 174.713 7] |
| $\beta_2$ | 0.047 7 | [0.030 4, 0.065 0] |
| $\gamma_1$ | 52.403 5 | [32.413 0, 72.394 1] |
| $\gamma_2$ | 0.016 4 | [-0.007 5, 0.040 3] |

然而，从表 4 可以发现，$\gamma_2$ 的置信区间包含零点，这表明参数 $\gamma_2$ 对因变量 $y$ 的影响并不显著，这一结果与前面的说法（即嘌呤霉素的作用不影响半速度参数）是一致的. 因此，可以考虑简化模型

$$y = f(\boldsymbol{x}, \boldsymbol{\beta}) = \frac{(\beta_1 + \gamma_1 x_2) x_1}{\beta_2 + x_1} \tag{5}$$

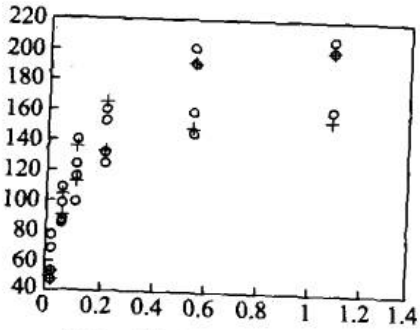

图 7　模型(4)的预测图

(○—原始数据；+—拟合结果)

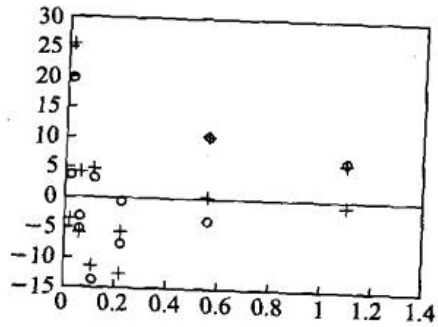

图 8　模型(4)残差图

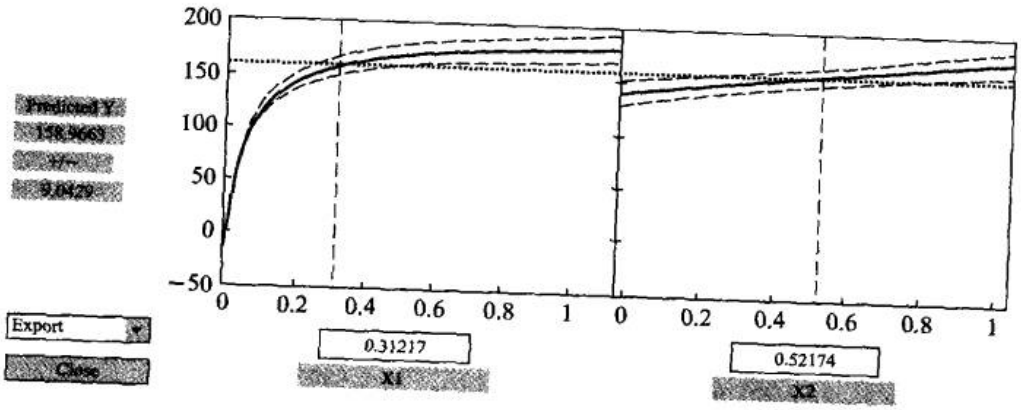


图 9　模型(4)的预测及结果输出

采用与模型(4)类似的计算、分析方法，模型(5)的结果概括在表 5 和表 6，以及图 10、图 11 和图 12 中，模型(5)的剩余标准差 $s = 10.585\ 1$.

**表 5　模型(4)参数的估计结果**

| 参数 | 参数估计值 | 参数置信区间 |
|---|---|---|
| $\beta_1$ | 166.602 5 | [154.488 6，178.716 4] |
| $\beta_2$ | 0.058 0 | [0.045 6，0.070 3] |
| $\gamma_1$ | 42.025 2 | [28.941 9，55.108 5] |

**表 6　模型(4)与模型(5)预测值与预测区间的比较(预测区间为预测值 ± Δ)**

| 实际数据 | 模型(4)预测值 | Δ(模型(4)) | 模型(5)预测值 | Δ(模型(5)) |
|---|---|---|---|---|
| 67 | 47.344 3 | 9.207 8 | 42.735 8 | 5.444 6 |
| 51 | 47.344 3 | 9.207 8 | 42.735 8 | 5.444 6 |
| 84 | 89.285 6 | 9.571 0 | 84.735 6 | 7.047 8 |
| 86 | 89.285 6 | 9.571 0 | 84.735 6 | 7.047 8 |

续表

| 实际数据 | 模型(4)预测值 | Δ(模型(4)) | 模型(5)预测值 | Δ(模型(5)) |
|---|---|---|---|---|
| 98 | 111.793 8 | 7.754 6 | 109.105 3 | 7.028 1 |
| 115 | 111.793 8 | 7.754 6 | 109.105 3 | 7.028 1 |
| 131 | 131.716 6 | 7.500 7 | 131.858 6 | 7.587 8 |
| 124 | 131.716 6 | 7.500 7 | 131.858 6 | 7.587 8 |
| 144 | 147.697 3 | 10.372 9 | 150.974 3 | 9.442 3 |
| 158 | 147.697 3 | 10.372 9 | 150.974 3 | 9.442 3 |
| 160 | 153.617 6 | 12.119 7 | 158.262 3 | 10.562 1 |
| 76 | 50.566 | 7.691 4 | 53.515 8 | 6.740 9 |
| 47 | 50.566 | 7.691 4 | 53.515 8 | 6.740 9 |
| 97 | 102.811 | 9.564 3 | 106.110 1 | 8.236 8 |
| 107 | 102.811 | 9.564 3 | 106.110 1 | 8.236 8 |
| 123 | 134.361 6 | 8.252 2 | 136.627 0 | 7.422 3 |
| 139 | 134.361 6 | 8.252 2 | 136.627 0 | 7.422 3 |
| 159 | 164.684 7 | 7.029 4 | 165.119 7 | 7.059 5 |
| 152 | 164.684 7 | 7.029 4 | 165.119 7 | 7.059 5 |
| 191 | 190.832 9 | 9.148 4 | 189.057 4 | 8.843 8 |
| 201 | 190.832 9 | 9.148 4 | 189.057 4 | 8.843 8 |
| 207 | 200.968 8 | 11.044 7 | 198.183 7 | 10.181 2 |
| 200 | 200.968 8 | 11.044 7 | 198.183 7 | 10.181 2 |

混合模型(4)和(5)不仅有类似于模型(1)的实际解释,同时把嘌呤霉素处理前后酶促反应的速度之间的变化体现在模型之中,因此它们比单独的模型具有更实际的价值.另外,虽然模型(5)的某些统计指标可能没有模型(4)的好,比如模型(5)的剩余标准差略大于模型(4),但由于它的形式更简单明了,易于实际中的操作和控制,而且从表6中数据可以发现,虽然两个模型的预测值相差不大,但模型(5)预测区间的长度明显比模型(4)的短.因此,总体来说模型(5)更为优良.

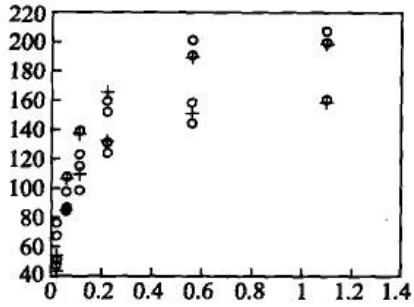


图 10　模型(5)的预测图

(○—原始数据；+—拟合结果)

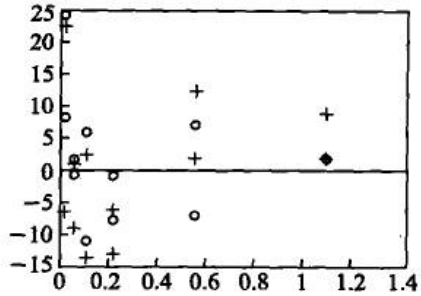


图 11　模型(5)残差图

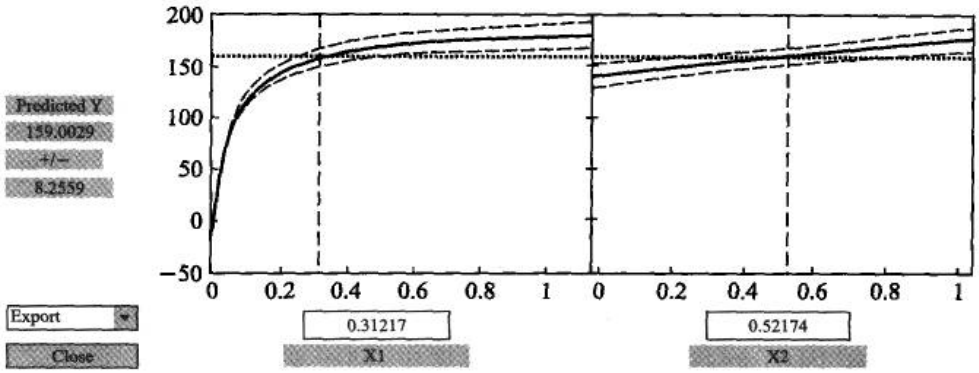


图 12　模型(5)的预测及结果输出

**可进一步研究的模型**　假如在实验中当底物浓度增加到一定程度后，反应速度反而有轻微的下降(在本例中只有一个数据点如此)，那么可以考虑模型

$$y = f(\boldsymbol{x}, \boldsymbol{\beta}) = \frac{\beta_1 x}{\beta_2 + x + \beta_3 x^2} \tag{6}$$

或引入混合模型

$$y = f(\boldsymbol{x}, \boldsymbol{\beta}) = \frac{(\beta_1 + \gamma_1 x_2) x_1}{(\beta_2 + \gamma_2 x_2) + x_1 + (\beta_3 + \gamma_3 x_2) x_2^2} \tag{7}$$

有兴趣的读者可以尝试一下，会发现用这些模型可以改善模型(4)和(5)的残差图(图 8、图 11)中表现出来的在各个浓度下残差散布不均匀的现象.

**评注**　无论从机理分析，还是从实验数据看，酶促反应中反应速度与底物浓度及嘌呤霉素的作用之间的关系都是非线性的，本节我们先用线性化模型来简化参数估计，如果这样能得到满意的结果，当然很好，但是由于变量的代换已经隐含了误差扰动项的变换，因此，除非变换后的误差项仍具有常数方差，一般情况下我们还需要采用原始数据做非线性回归，而把线性化模型的参数估计结果作为非线性模型参数估计的迭代初值.

应该指出，在非线性模型参数估计中，用不同的参数初值进行迭代，可能得到差别很大的结果（它们都是拟合误差平方和的局部极小点），也可能出现收敛速度等问题，因此，合适的初值是非常重要的.

另外，评价线性回归模型拟合程度的统计检验无法直接用于非线性模型. 例如，$F$ 统计量不能用于非线性模型拟合程度的显著性检验，因为即使误差项服从均值为 0 的正态分布，也无法从回归残差得到误差方差的一个无偏估计. 但是 $R^2$ 和剩余标准差 $s$ 仍然可以在通常意义下用于非线性回归模型拟合程度的度量.

从本例还可以看到，通过引入示性变量，能够描述定性上不同的处理水平对模型参数的影响，这是一种直接明了的建模方法.

## 10.4 投资额与生产总值和物价指数

**问题** 为研究某地区实际投资额与国民生产总值（GNP）及物价指数的关系，收集了该地区连续 20 年的统计数据（见表 1），目的是由这些数据建立一个投资额的模型，根据对未来国民生产总值及物价指数的估计，预测未来的实际投资额.

表 1 的数据是以时间为序的，称**时间序列**. 由于投资额、国民生产总值、物价指数等许多经济变量均有一定的滞后性，比如，前期的投资额对后期投资额一般有明显的影响. 因此，在这样的时间序列数据中，同一变量的顺序观测值之间的出现相关现象（称自相关）是很自然的. 然而，一旦数据中存在这种自相关序列，如果仍采用普通的回归模型直接处理，将会出现不良后果，其预测也会失去意义，为此，我们必须先来诊断数据是否存在自相关，如果存在，就要考虑自相关关系，建立新的回归模型[59,79].

**表 1 某地区实际投资额（亿元）与国民生产总值（亿元）及物价指数数据**

| 年份序号 | 投资额 | 国民生产总值 | 物价指数 | 年份序号 | 投资额 | 国民生产总值 | 物价指数 |
|---|---|---|---|---|---|---|---|
| 1 | 90.9 | 596.7 | 0.716 7 | 11 | 229.8 | 1 326.4 | 1.057 5 |
| 2 | 97.4 | 637.7 | 0.727 7 | 12 | 228.7 | 1 434.2 | 1.150 8 |
| 3 | 113.5 | 691.1 | 0.743 6 | 13 | 206.1 | 1 549.2 | 1.257 9 |
| 4 | 125.7 | 756.0 | 0.767 6 | 14 | 257.9 | 1 718.0 | 1.323 4 |
| 5 | 122.8 | 799.0 | 0.790 6 | 15 | 324.1 | 1 918.3 | 1.400 5 |
| 6 | 133.3 | 873.4 | 0.825 4 | 16 | 386.6 | 2 163.9 | 1.504 2 |
| 7 | 149.3 | 944.0 | 0.867 9 | 17 | 423.0 | 2 417.8 | 1.634 2 |
| 8 | 144.2 | 992.7 | 0.914 5 | 18 | 401.9 | 2 631.7 | 1.784 2 |
| 9 | 166.4 | 1 077.6 | 0.960 1 | 19 | 474.9 | 2 954.7 | 1.951 4 |
| 10 | 195.0 | 1 185.9 | 1.000 0 | 20 | 424.5 | 3 073.0 | 2.068 8 |