数学实验 · 实验报告10

计算机系　计01　容逸朗　2020010869

13-7 耗氧问题

问题分析与模型建立

题目给出了 5 种可能影响人的耗氧能力的数据，希望找到一个合适的回归模型来说明耗氧能力与诸因素之间的关系。由于缺乏相关的专业知识，因此在这里我使用了最简单的线性回归模型：

$$y = \beta_0 + \sum_{i=1}^{5} \beta_i x_i \tag{1}$$

由于不同小问的变量数要求，可以将模型中部分的 β_i 置为零来得到符合条件的结果。

算法设计

对于第一小问，可以直接利用 **`regress`** 函数计算回归函数，第二、三小问则可以利用 **`stepwise`** 函数逐步回归，第四小问则可以 **`rcoplot`** 来显示残差，方便我们去掉异常点。

除此之外，也可以画出单一变量的散点图，帮助我们找出最有规律的数据，同时检验回归结果的正确性。

程序

第一小问：单变量模型

```
format short

x = [
    1 44.6 44 89.5 6.82 62 178;
    2 45.3 40 75.1 6.04 62 185;
    3 54.3 44 85.8 5.19 45 156;
    4 59.6 42 68.2 4.90 40 166;
    5 49.9 38 89.0 5.53 55 178;
    6 44.8 47 77.5 6.98 58 176;
    7 45.7 40 76.0 7.17 70 176;
    8 49.1 43 81.2 6.51 64 162;
    9 39.4 44 81.4 7.85 63 174;
    10 60.1 38 81.9 5.18 48 170;
    11 50.5 44 73.0 6.08 45 168;
    12 37.4 45 87.7 8.42 56 186;
    13 44.8 45 66.5 6.67 51 176;
    14 47.2 47 79.2 6.36 47 162;
    15 51.9 54 83.1 6.20 50 166;
    16 49.2 49 81.4 5.37 44 180;
    17 40.9 51 69.6 6.57 57 168;
    18 46.7 51 77.9 6.00 48 162;
    19 46.8 48 91.6 6.15 48 162;
```

```
23      20 50.4 47 73.4 6.05 67 168;
24      21 39.4 57 73.4 7.58 58 174;
25      22 46.1 54 79.4 6.70 62 156;
26      23 45.4 52 76.3 5.78 48 164;
27      24 54.7 50 70.9 5.35 48 146;
28      ];
29
30  % 数据处理
31  y = x(:, 2);
32
33  for i = 3: 7
34      % 画图
35      subplot(2, 3, i-2);
36      plot(x(:, i), y, '.');
37      grid;
38      xlabel(sprintf('x_%d', i-2));
39
40      % 数据拟合
41      X = [ones(24, 1), x(:, i)];
42      [b, bint, r, rint, s] = regress(y, X);
43
44      % 对应变量编号及计算结果
45      i-2, b, bint, s
46  end
```

二、三小问：多变量模型

将上面程序的第 33-46 行改为下面的代码即可。

```
% 逐步回归
stepwise(x(:, 3:7), x(:, 2));
```

第四小问：去除异常点

在第一小问的程序第 33-46 行改为下面的代码即可。

```
% 残差分析
[b, bint, r, rint, s] = regress(y, [ones(24, 1), x(:, [3, 5, 7])]);
subplot(1, 1, 1);
rcoplot(r, rint);
```

计算结果与分析

第一小问：单变量模型

分别画出单一变量对应数据的散点图，可以看到只有 x_3 的数据与 y 组成明显的线性关系，其他数据则是难以找出与 y 之间的关系。

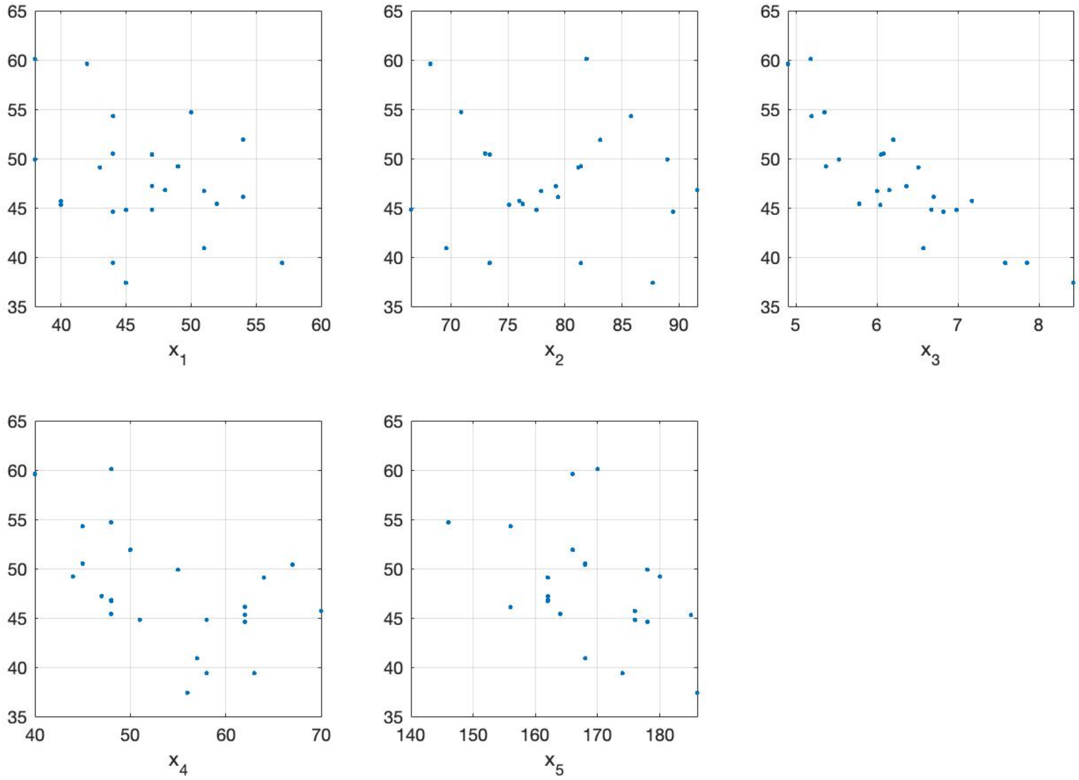

图1　数据散点图

使用公式（1）对应的模型，可以得出如下结果：

变量	β_0	β_1	β_0 置信区间	β_1 置信区间	R^2	F	p	s^2
x_1	64.3812	-0.3599	(42.3913, 86.3711)	(-0.8309, 0.1111)	0.1025	2.5115	0.1273	31.2484
x_2	52.8008	-0.0651	(23.6361, 81.9755)	(-0.4344, 0.3042)	0.0060	0.1337	0.7181	34.6053
x_3	83.4438	-5.6682	(74.1644, 92.7232)	(-7.1252, -4.2112)	0.7474	65.0959	0.0000	8.7943
x_4	67.1094	-0.3599	(52.5706, 81.6483)	(-0.6262, -0.0936)	0.2631	7.8560	0.0104	25.6547
x_5	94.0024	-0.2739	(54.1047, 133.9001)	(-0.5095, -0.0384)	0.2091	5.8169	0.0247	27.5352

从图像和表格的数据可以看见 x_3（1500 米跑用时）与耗氧能力 y 的关联性最大。注意到 x_1 和 x_2 的 β_1 置信区间都包含零点，故应当排除这两种因素（年龄、体重）对耗氧能力 y 的影响。对于 x_4, x_5，虽然对应的 R^2 值也有 0.2 以上，即耗氧能力总变化量的 20% 以上可由自变量确定，但对于单变量模型，显然 x_3 的 R^2 值是更优的，因此应当选择 x_3 来描述 y，这时的模型为：

$$y = \beta_0 + \beta_1 x_3, \quad \beta_0 = 83.4438, \beta_1 = -5.6682. \tag{2}$$

因为 x_3 的 $R^2 = 0.7474$ 且 p 值远小于 $\alpha = 0.05$，由此来看模型是有效的。

第二小问：双变量模型

分别计算变量两两组合的 RMSE 值，发现 x_1, x_3 的组合可以得到最小的 RMSE 值 2.87035。

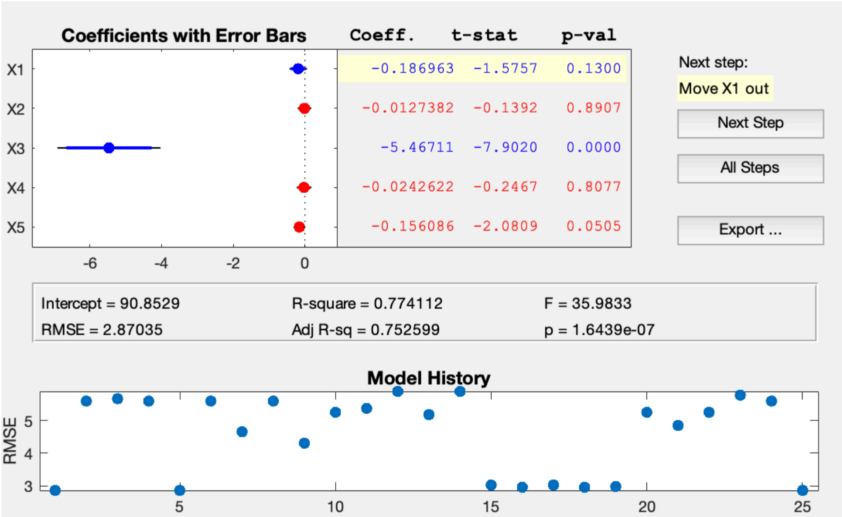

图2　Stepwise Regression

此时 $R^2=0.7741, F=35.9833, p=1.6439e-7$ ，比只使用 x_3 更佳，这时的模型为：

$$y=\beta_0+\beta_1 x_1+\beta_2 x_3, \quad \beta_0=90.8529,\ \beta_1=-0.1870,\ \beta_2=-5.4671 \tag{3}$$

第三小问：多变量模型

以 x_1, x_3 为基础，分别计算增加不同的剩余变量组合而成的 RMSE 值，发现 x_1, x_3, x_5 的组合可以得到最小的 RMSE 值 2.66669。

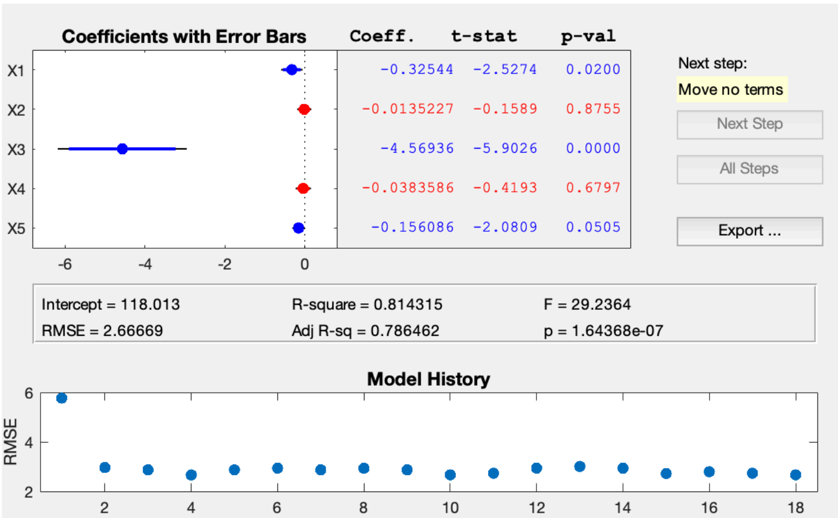

图3　Stepwise Regression

此时 $R^2=0.8143, F=29.2364, p=1.6437\times10^{-7}$ ，比只使用 x_1, x_3 更佳，这时的模型为：

$$\begin{gathered} y=\beta_0+\beta_1 x_1+\beta_2 x_3+\beta_3 x_5, \\ \beta_0=118.013,\ \beta_1=-0.3254,\ \beta_2=-4.5694\ \beta_3=-0.1561 \end{gathered} \tag{4}$$

第四小问：去除异常点

画出残差图，发现有两个异常点，编号分别为 10 和 15.

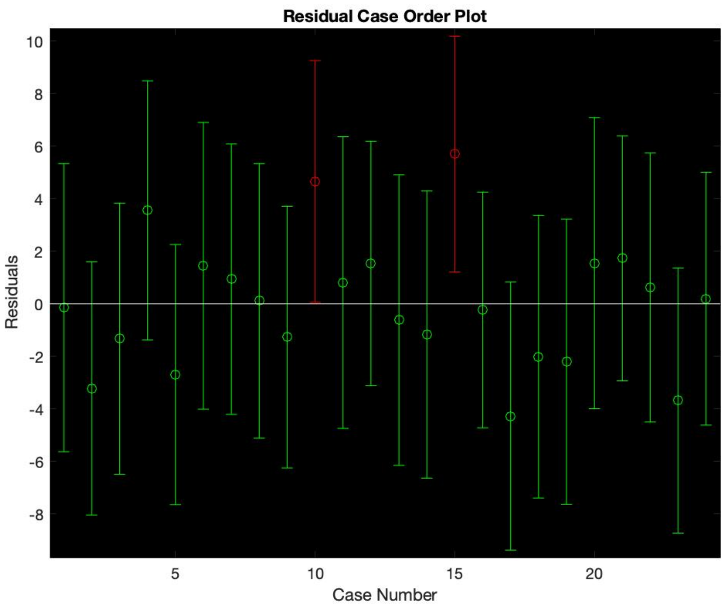

图4 残差图

去除异常点后，最终模型变为的参数如下所示：

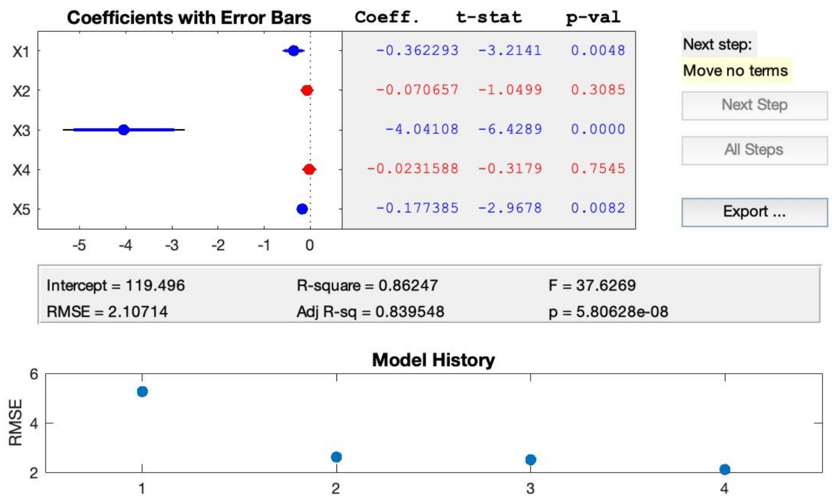

图5 Stepwise Regression

此时模型变为：

$$
\begin{gathered}
y=\beta_0+\beta_1 x_1+\beta_2 x_3+\beta_3 x_5, \\
\beta_0=119.496,\ \beta_1=-0.3623,\ \beta_2=-4.0411\ \beta_3=-0.1774
\end{gathered} \tag{5}
$$

从模型的 $R^2=0.8625$ 和 $p=4.80\times 10^{-8}$ 知模型的效果比去除异常点前的模型更好。

结论

1. 只能选择一个变量时，应该选择 1500 米跑的用时为变量，这时模型为：$y=83.4438+-5.6682x_3$。
2. 若能选择两个变量，则可以选择 1500 米跑的用时和年龄为变量，这时模型为：$y=90.8529-0.1870x_1-5.4671x_3$。
3. 若能选择多个变量，应选择 1500 米跑的用时、年龄和跑步后心率等三种数据为变量，这时模型为 $y=118.013-0.3254x_1-4.5694x_3-0.1561x_5$。

4. 有两个异常点（编号 10，15），剔除后变为 $y = 119.496 - 0.3623x_1 - 4.0411x_3 - 0.1774x_5$。

13-9 洗衣粉试验

问题分析与模型建立

题目希望找出一个线性回归模型来描述洗衣粉泡沫高度 y 与搅拌程度 x_1 与用量 x_2 的关系。由此我们可以建立如下模型：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{6}$$

若搅拌程度视为没有定量关系的 3 个水平，那么我们可以利用两个变量 $k = (k_1, k_2)$ 来代替 x_1，那么 $x_1 = 1$ 可以记为 $k = (0, 0)$，$x_1 = 2$ 则是 $k = (0, 1)$，对 $x_1 = 3$ 则为 $k = (1, 0)$，那么模型变为：

$$y = \beta_0 + \beta_1 k_1 + \beta_2 k_2 + \beta_3 x_2 \tag{7}$$

如果加入交互项，则模型变为：

$$y = \beta_0 + \beta_1 k_1 + \beta_2 k_2 + \beta_3 x_2 + \beta_4 k_1 x_2 + \beta_5 k_2 x_2 \tag{8}$$

算法设计

与上题类似，可以直接利用 **regress** 函数计算模型，然后利用 **rcoplot** 画出残差图分析模型的正确性；

若搅拌程度视为没有定量关系的 3 个水平，同样可以直接利用 **regress** 函数计算回归模型；

如果加入交互项，则可以利用 **stepwise** 函数手动选择合适节点，然后用 **regress** 函数计算回归模型。

程序

第一小问：搅拌程度视作一般变量

```
format short g

x = [
    1 6 28.1;
    1 7 32.3;
    1 8 34.8;
    1 9 38.2;
    1 10 43.5;
    2 6 65.3;
    2 7 67.7;
    2 8 69.4;
    2 9 72.2;
    2 10 76.9;
    3 6 82.2;
    3 7 85.3;
    3 8 88.1;
    3 9 90.7;
    3 10 93.6;
    ];

% 数据处理
```

```
22 y = x(:, 3);
23 n = length(x);
24 
25 % 第一问：定量
26 X = x(:, [1, 2]);
27 
28 % 第二问：非定量
29 
30 % 第三问：逐步回归
31 
32 % 回归计算
33 [b, bint, r, rint, s] = regress(y, [ones(n, 1), X]);
34 b, bint, s
35 subplot(1, 1, 1);
36 rcoplot(r, rint);
37 
```

第二小问：搅拌程度视为没有定量关系的 3 个水平

将第一问代码第 28 行改为下面代码：

```
% 第二问：非定量
X = [];
for i = 1: n
    if x(i, 1) == 1
        X = [X; 0 0];
    elseif x(i, 1) == 2
        X = [X; 0 1];
    elseif x(i, 1) == 3
        X = [X; 1 0];
    end
end
X = [X x(:, 2)];
```

第三小问：引入交互项

在第二问的基础上，再将第一问代码第 30 行改为下面代码：

```
% 第三问：逐步回归
X = [X X(:, 1).*X(:, 3) X(:, 2).*X(:, 3)];
stepwise(X, y);
```

计算结果与分析

第一小问：搅拌程度视作一般变量

直接调用 `regress` 函数解得：$\beta_0=-12.74,\beta_1=26.30,\beta_2=3.0867$ ，三者的置信空间分别为：$(-29.0268,3.5468)$，$(23.1059,29.4941)$，$(1.2426,4.9308)$ ，此时模型为：

$$y=\beta_0+\beta_1x_1+\beta_2x_2,\quad \beta_0=-12.74,\beta_1=26.30,\beta_2=3.0867. \tag{9}$$

该模型的 $R^2=0.9654,F=167.5754,p=1.706\times10^{-9},s^2=21.491$ ，这时作出模型的残差图，得到：

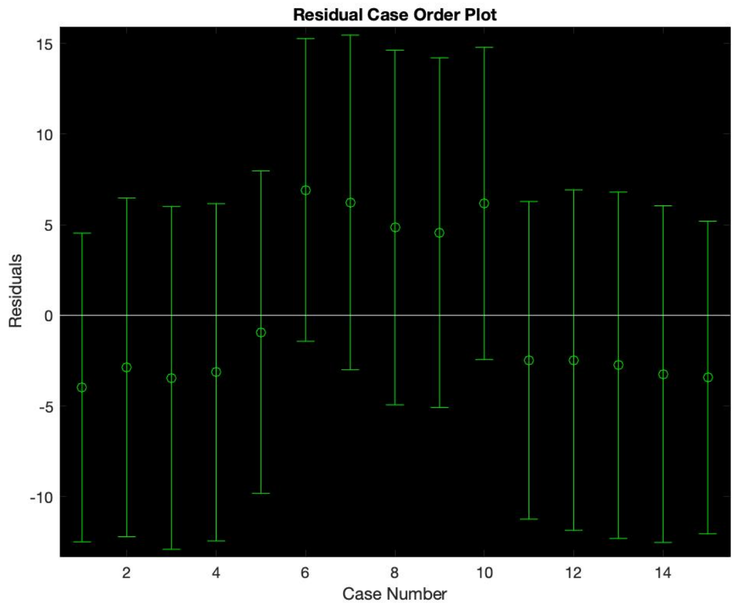

图6　模型残差图

可以看到，搅拌程度为 2 的数据残差与其他数据相比出现了很大的偏差，因此此模型的可信性不大。

第二小问：搅拌程度视为没有定量关系的 3 个水平

同样利用 `regress` 函数可得：$\beta_0=10.687,\beta_1=52.6,\beta_2=34.92,\beta_3=3.0867$ ，对应的置信空间分别为：$(7.4475,13.926)$，$(51.259,53.941)$，$(33.579,36.261)$，$(2.6995,3.4738)$ ，此时模型为：

$$\begin{gathered}y=\beta_0+\beta_1k_1+\beta_2k_2+\beta_3x_2\\ \beta_0=10.687,\beta_1=52.6,\beta_2=34.92,\beta_3=3.0867\end{gathered} \tag{10}$$

该模型的 $R^2=0.9986,F=2675.5,p=5.01\times10^{-16},s^2=0.92824$ ，这时作出模型的残差图，得到：

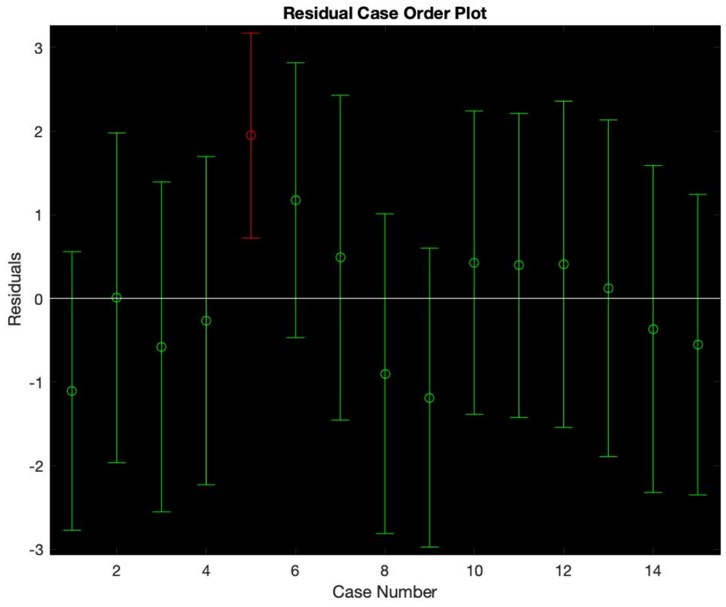

图7　模型残差图

去掉异常点（5号），得到：

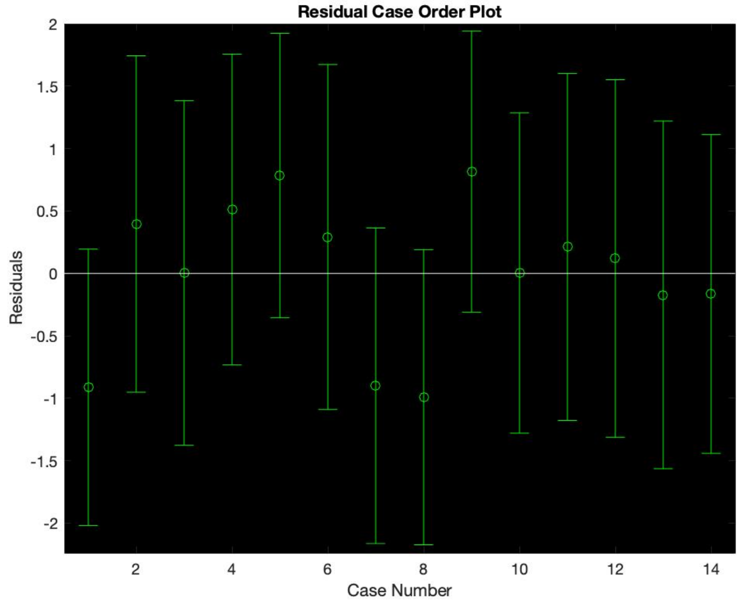

图8　剔走异常点后的残差图

这时模型为：

$$\begin{gathered} y=\beta_0+\beta_1k_1+\beta_2k_2+\beta_3x_2 \\ \beta_0=11.66,\beta_1=53.184,\beta_2=35.504,\beta_3=2.892 \end{gathered} \tag{11}$$

对应的 $R^2=0.99935,F=5141.1,p=3.09\times10^{-16},s^2=0.45264$ ，比第一小问的结果稍优。

第三小问：引入交互项

利用 `stepwise` 函数，可以看到全取所有参数时 RMSE 值最小。

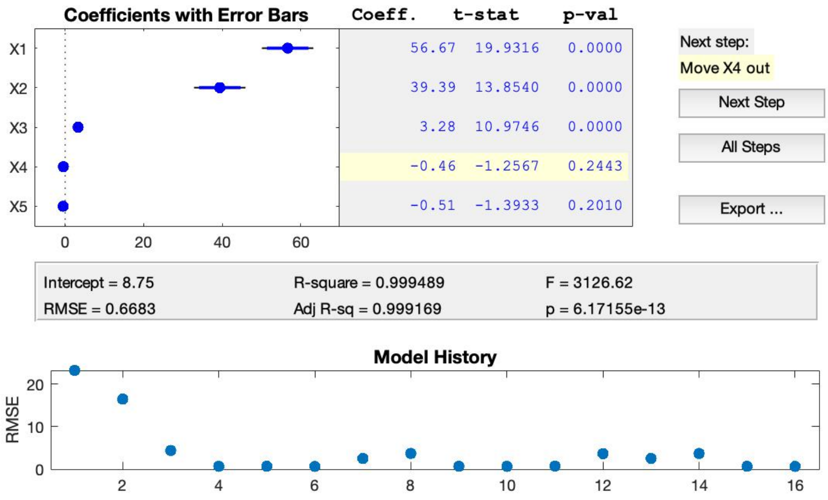

图9 Stepwise Regression

去掉异常点（编号 10）后，模型变为：

$$
\begin{gathered}
y=\beta_0+\beta_1 k_1+\beta_2 k_2+\beta_3 x_2+\beta_4 k_1 x_2+\beta_5 k_2 x_2 \\
\beta_0=6.02,\ \beta_1=59.4,\ \beta_2=45.83,\ \beta_3=3.67,\ \beta_4=-0.85,\ \beta_5=-1.43
\end{gathered} \tag{12}
$$

对应的 $R^2=0.99969, F=5108.1, p=8.67\times 10^{-14}, s^2=0.28562$。

结果

1. 若搅拌程度视作一般变量，模型为 $y=-12.74+26.30x_1+3.0867x_2$ ，但二级搅拌程度的残差与其他的不同。
2. 若搅拌程度视为没有定量关系的 3 个水平，令 $x_1=1$ 记为 $(k_1,k_2)=(0,0)$，$x_1=2$ 则是 $(k_1,k_2)=(0,1)$ ，对 $x_1=3$ 则为 $(k_1,k_2)=(1,0)$ ，此时模型为 $y=11.66+53.184k_1+35.504k_2+2.892x_2$。
3. 引入交互项后，模型效果有所改进，最终模型为：$y=6.02+59.4k_1+45.83k_2+3.67x_2-0.85k_1x_2-1.43k_2x_2$。

13-13 销售趋势

问题分析与模型

题目给定了两种模型及数据，要求找出模型对应参数的估计值。

Logistic 模型：

$$
y=\frac{L}{1+ae^{-kt}} \tag{13}
$$

Gompertz 模型：

$$
y=Le^{-be^{-kt}} \tag{14}
$$

观察可知，若 L 不是固定参数，那么上述模型显然不是可线性化的。若 L 固定，那么 Logistic 模型（13）可以转化为下面的线性模型：

$$\begin{aligned} ae^{-kt} &= \frac{L}{y} - 1 \\ -kt + \ln a &= \ln\left(\frac{L}{y} - 1\right) \end{aligned} \tag{15}$$

这时令 $\beta_0 = \ln a, \beta_1 = -k$ 即可得到线性模型。

同样地，Gompertz 模型（14）也可以转化为下面的线性模型：

$$\ln\left(\ln \frac{y}{L}\right) = ln(-b) - kt \tag{16}$$

这时令 $\beta_0 = \ln(-b), \beta_1 = -k$ 即可得到线性模型。

算法设计

根据（15），（16）式，我们可以利用 **regress** 函数回归计算模型，以此为初值可以利用 **nlinfit** 拟合非线性模型。

程序

```
format short g

% 数据
x = [
    0 43.65;
    1 109.86;
    2 187.21;
    3 312.67;
    4 496.58;
    5 707.65;
    6 960.25;
    7 1238.75;
    8 1560.00;
    9 1824.29;
    10 2199.00;
    11 2438.89;
    12 2737.71;
    ];

% 数据处理
L = 3000;
t = x(:, 1);
y = x(:, 2);
n = length(x);

% 第一问：线性拟合 Logistic 模型
y_log = log(L ./ y - 1);
X = [ones(n, 1) t];
[beta, bint, r, rint, s] = regress(y_log, X);
a = exp(beta(1));
```

```
k = -beta(2);
beta, bint, s, a, k

% 第二问：非线性拟合 Logistic 模型
b0 = [L a k];
[beta, R, J, CovB, MSE] = nlinfit(t, y, @Logistic, b0);
beta, CovB, MSE

% 第三问：非线性拟合 Gompertz 模型
b = 30;
k = 0.4;
L = 3000;
b0 = [L b k];
[beta, R, J, CovB, MSE] = nlinfit(t, y, @Gompertz, b0);
beta, CovB, MSE

function y = Logistic(b, t)
    y = b(1) ./ (1 + b(2) .* exp(-b(3) .* t));
end

function y = Gompertz(b, t)
    y = b(1) .* exp(-b(2) .* exp(-b(3) .* t));
end

```

计算结果与分析

第一问：线性拟合 `Logistic` 模型

当 $L = 3000$ 时，$\beta_0 = 3.8032, \beta_1 = -0.49412$ ，对应的 $a = 44.846, k = 0.49412$ 。

此时模型的 $R^2 = 0.99053, F = 1150.8, p = 1.748 \times 10^{-12}, s^2 = 0.03861$.

第二问：非线性拟合 `Logistic` 模型

代入 $L^{(0)} = 3000, a^{(0)} = 44.846, k^{(0)} = 0.4941$ ，由 **`nlinfit`** 拟合可得:

$L = 3260.4, a = 30.535, k = 0.4148$.

模型的 `MSE` 为 1765.1

第三问：非线性拟合 `Gompertz` 模型

代入 $L^{(0)} = 3000, b^{(0)} = 30, k^{(0)} = 0.4$ ，由 **`nlinfit`** 拟合可得:

$L = 4810.1, b = 4.592, k = 0.17472$.

模型的 `MSE` 为 308.14 ，比 `Logistic` 模型的拟合结果更小，故对于高压锅销售这一问题， `Gompertz` 模型是更好的选择。

结论

1. Logistic 增长曲线模型不是可线性化模型。若给定 $L = 3000$ ，那么有 $a = 44.846, k = 0.49412$.
2. 由 1 计算的结果对 Logistic 模型做非线性回归，得到：$L = 3260.4, a = 30.535, k = 0.4148$.
3. 非线性回归拟合 Gompertz 模型，得到：$L = 4810.1, b = 4.592, k = 0.17472$。比较两者的 MSE 可知，与 Logistic 模型相比，Gompertz 模型是更好的选择。