数学实验第十次实验报告

计 76　张翔　2017011568

2020 年 6 月 1 日

1　实验目的

1. 了解回归分析的基本原理，掌握 MATLAB 实现的方法;
2. 练习用回归分析解决实际问题。

2　Ch13-P7 耗氧能力

2.1　模型建立

本题不同小问需要建立不同模型，由于缺乏专业的先验知识，并且提供的数据量不大，因此采用简单的线性回归模型处理该问题。首先可以绘制年龄、体重、1500 米跑所用时间、静止心速与跑后心速这些单一变量对于耗氧能力的影响，确定函数选取。这里使用的线性回归模型的基本形式如下，当相应项不需要时，将系数 β 置为 0 即可

$$y = \beta_0 + \sum_{i=1}^{5} \beta_i x_i + \sum_{1 \le i,\, j \le 5} \beta_{ij} x_i x_j + \epsilon$$

在分析时，从最简单的一次函数关系入手，当它的拟合效果较差时，再考虑引入上述式子中的高次项或是交互项 $\beta_{ij} x_i x_j$。

根据题目中的不同小问限制的选取变量个数，这里需要采用多种不同模型，具体的模型将在下面计算结果部分单独说明，这里不再赘述。当模型选定后，可以通过残差和置信区间寻找异常点，将其剔除后重新计算，以使得模型更加准确。

2.2　算法设计

对于第 (1) 小问，要求只能使用一种变量，可以使用 MATLAB 的 `regress` 函数对每个变量单独进行回归，得到回归系数估计值、置信区间及统计量等。

对于第 (2) 问，可选择 2 个变量，可以使用 `rstool` 进行分析，考虑交互项和二次项的影响。

对于第 (3) 问，在此基础上可以使用 `stepwise` 进行交互式逐步分析，得到剩余方差最小的模型。在第 (4) 问中，可以使用 `rcplot` 作出残差图，判断异常点并移除。

2.3　Matlab 程序

数据输入处理的代码如下

```
%% import data
data = reshape(readmatrix('p7.txt'), 24, []);
y = data(:, 1);
```

```
x1 = data(:, 2);
x2 = data(:, 3);
x3 = data(:, 4);
x4 = data(:, 5);
x5 = data(:, 6);
```

第 (1) 问作图与拟合的代码如下

```
%% plot scatter
for i=1:5
    subplot(2, 3, i), plot(data(:, i + 1), y, '+'), grid, xlabel(sprintf('
        x_%d', i)), ylabel('y');
end

%% regress
X=[ones(size(x1)), x1];
[b,bint,r,rint,s]=regress(y, X);
```

第 (2) 问使用逐步回归的代码如下

```
stepwise([x1, x2, x3, x4, x5], y);
rstool([x1, x3], y, 'linear');
```

第 (3)(4) 问仍然使用 `stepwise` 进行交互式处理，不再赘述。

2.4 计算结果与分析

2.4.1 单变量模型

首先作出 5 个变量的散点图，如下

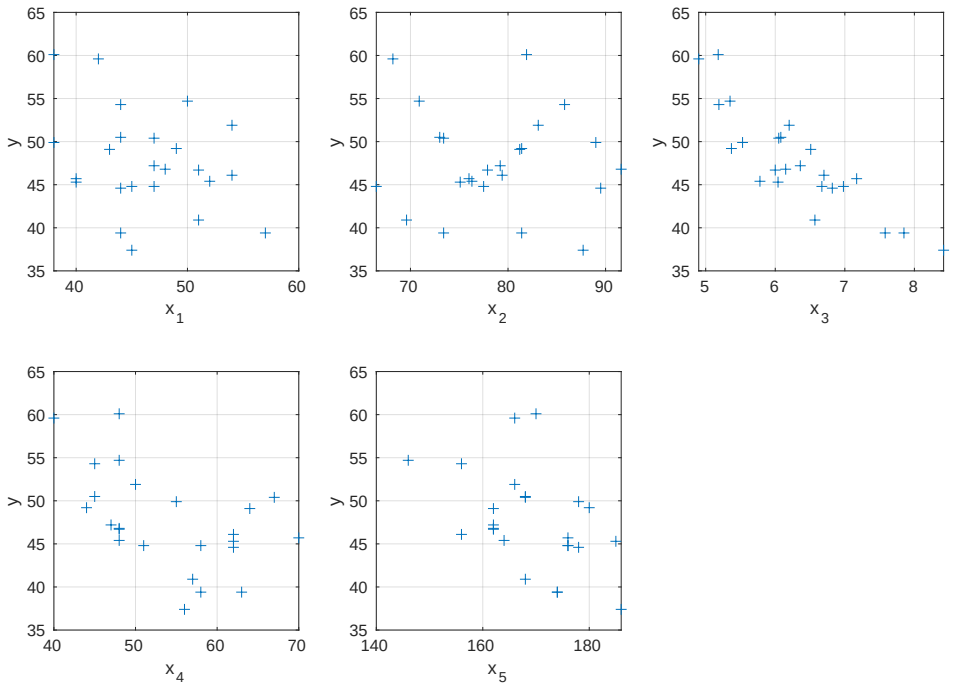

图 1: 单一变量作用时的散点图

可以看出只有 x_3 与 y 之间有较为明显的线性关系（负相关），对于其他自变量都难以直接观测出其对于因变量的影响。根据上述结果，可以假设自变量 x_3 与耗氧能力直接相关。使用只有一次项的模型 $y = \beta_0 + \beta_1 x + \epsilon$，对各个自变量进行单参数回归如下：

自变量	β_0	β_1	β_0 置信区间	β_1 置信区间	$\mathbf{R^2}$	**F**	**p**	$\mathbf{s^2}$
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	[23.6261, 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	65.0909	5.13×10^{-8}	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

表 1: 单变量回归结果

从上述回归结果可以证明 x_3（1500m 跑时间）反映 y（耗氧能力）的能力最强。x_1, x_2 的 β_1 置信区间均包含了 0，说明 y 可能与这两个参数无关，且它们的 p 值也明显大于 $\alpha = 0.05$，可以不选择这两个自变量。比较 x_3, x_4, x_5，可以发现 x_3 的决定系数 R^2 明显大于后两者，其对因变量的决定作用更大，并且用其拟合的 s^2 与 p 更小，可以确定单变量的情况下 x_3 反映 y 的能力最好，这与散点图的观测结果吻合。

使用 Polytool 可以检验 x_3 的高次项情况，结果如下

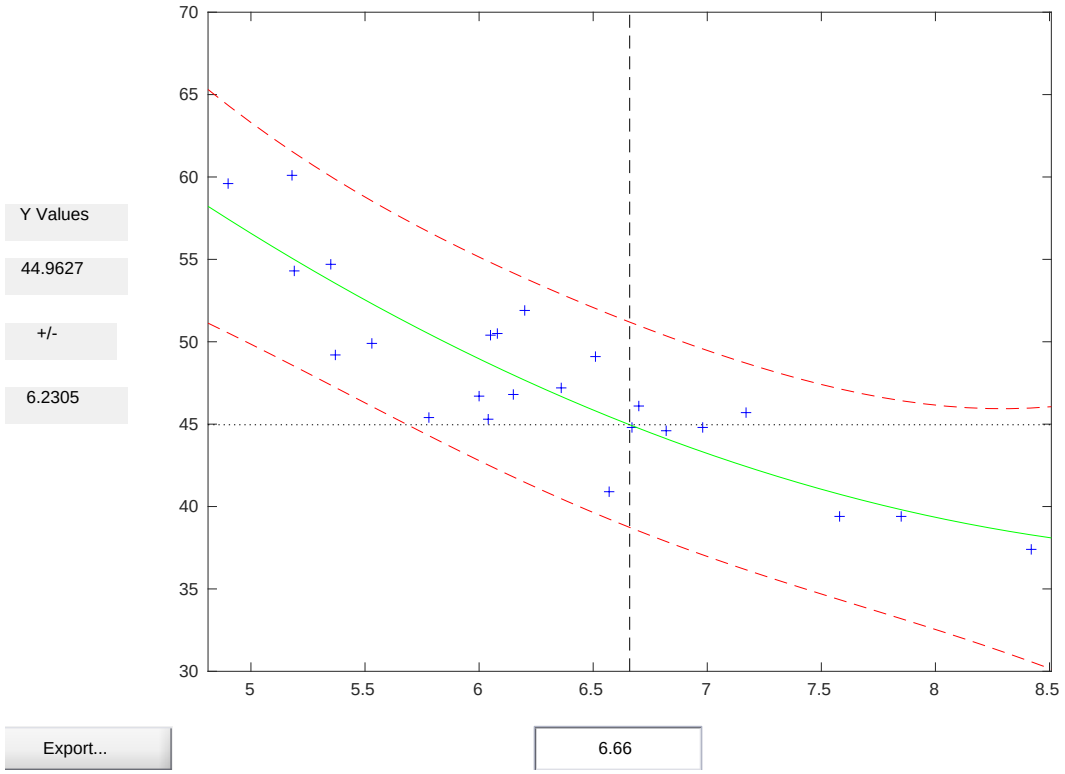

图 2: 使用 Polytool 检验 x_3 的二次项

- $\beta_0 = 122.7242$，置信区间 [67.1878, 178.2605]
- $\beta_1 = -17.9072$，置信区间 [-35.0387, -0.7757]
- $\beta_2 = 0.9356$，置信区间 [-0.3695, 2.2408]

可以看出，相比一次函数拟合的结果，β_0, β_1 的置信区间明显变宽，而 β_2 的置信区间包含 0，说明引入二次项是不必要的。

综上所述，只能使用单变量时，使用如下模型描述 y 是最准确的：

$$y = \beta_0 + \beta_1 x_3 \quad \text{其中 } \beta_0 = 83.4438,\ \beta_1 = -5.6682$$

2.4.2 双变量模型

在 (1) 问的基础上，认为 x_3 对因变量有较大影响，预先选择该变量，之后使用 `stepwise` 从 x_1, x_2, x_4, x_5 中选取，结果如下

变量	x_1	x_2	x_4	x_5
RMSE	2.87035	3.03307	3.03247	2.98927

表 2: 选取不同变量时的 RMSE

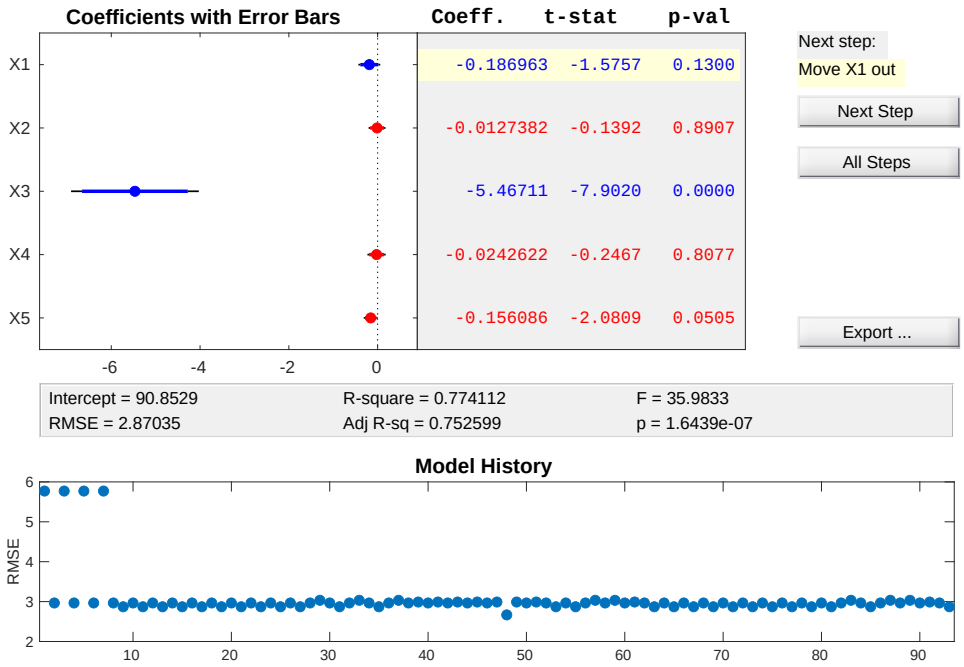

图 3: 逐步回归示例 (选择 x_3, x_5)

由上述结果可看出，选定 x_3 后，添加 x_1 变量得到的 s 最小，且输出的模型的 F, P 可以通过有效性检验，故双变量的情况下应选择 x_1, x_3。使用 `rstool` 进行高次项和交互项检验，结果如下

	β_0	$\beta_1 x_1$	$\beta_2 x_3$	$\beta_3 x_1^2$	$\beta_4 x_3^2$	$\beta_5 x_1 x_3$	RMSE
purequadratic	142.8835	-1.1718	-14.7911	0.0109	0.7111		2.9028
quadratic	144.4666	-1.0199	-16.4515	0.6818	0.0450	0.0062	2.9786
interaction	120.1929	-0.8364	-10.1096			0.1025	2.9033
linear	90.8529	-0.1870	-5.4671				2.8704

表 3: 选择 x_3, x_1 变量的 4 个模型的输出

可以看出，使用纯线性的模型得到的 RMSE 最小，且高次项和相关项的系数都非常小，说明它们

对 y 的影响不大，因此最终选择的模型如下

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

2.4.3 不限制变量时的模型

根据上述分析，可知本题中的交互项与高次项对于 y 的影响应较小，因此最终模型中不再考虑它们，而是仅使用一次项进行回归分析。类似于上述两变量的情况，使用 `stepwise` 逐个选取变量，取 RMSE 最小的模型，选取过程如下

- 基于上述双变量的最佳模型，选取 x_1, x_3，此时 $RMSE = 2.87035$；
- 增加变量 x_5，此时 $RMSE = 2.66669$；
- 如果再增加剩余的两个变量之一，均会使得 RMSE 上升，故不再选取更多变量。

最终选取了 x_1, x_3, x_5 变量做为最终模型，即

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_5 \quad \text{其中 } \beta_0 = 118.013,\ \beta_1 = -0.3254,\ \beta_2 = -4.5694,\ \beta_3 = -0.1561$$

其中 $\beta_1, \beta_2, \beta_3$ 的置信区间分别为 [-0.594, -0.0568], [-6.1842, -2.9546], [-0.3126, 0.0004]。$R^2 = 0.8143$, $F = 29.2364$, $RMSE = 2.66669$, $P = 1.6437 \times 10^{-7}$。

2.4.4 残差观察、剔除异常点

将最终模型的残差可视化，结果如下

图 4: 初始时模型的残差

可以看出 10 号与 15 号数据点异常，剔除后再次观察，结果如下

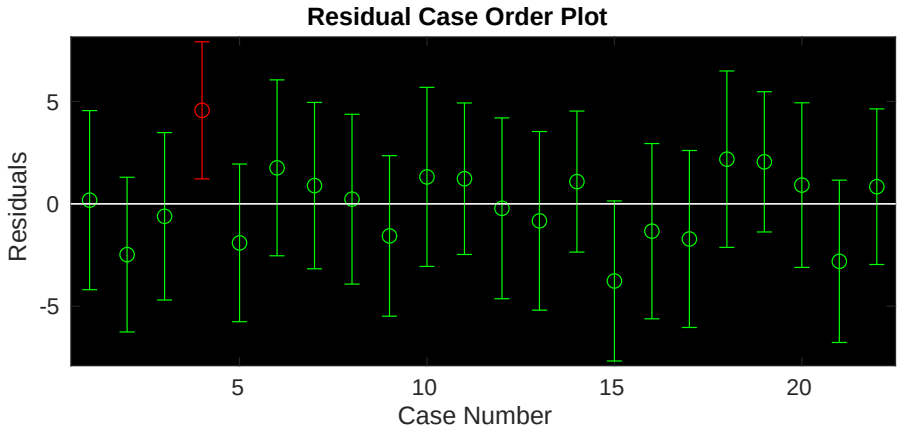

图 5: 残差（剔除 10，15 号）

可以看出 4 号数据异常，再次剔除，观察残差图仍有异常点。总共操作 4 次后，去掉 5 个数据点（4,10,15,17,23）后得到没有异常点的模型：

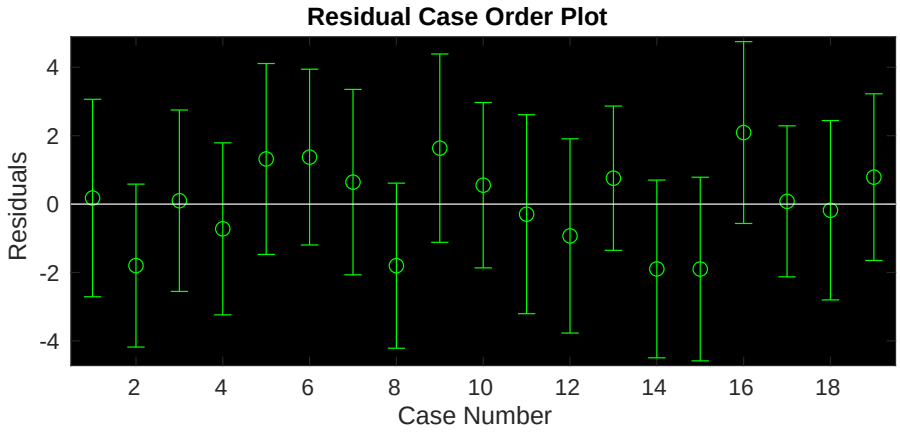

图 6: 残差（无异常点）

此时有

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_5 \quad \text{其中 } \beta_0 = 109.3470,\ \beta_1 = -0.2230,\ \beta_2 = -3.7694,\ \beta_3 = -0.1652$$

可以看出，这些系数与 (3) 的结果相比有了较大差异。不过，考虑到剔除异常点后虽然能在一定程度上降低其对整体的干扰，但同时也放大了其他正常点的异常性，使得新的异常点不断产生。剔除过多时，数据量过少会降低模型反映总体的能力，因此最终考虑只进行一次剔除，即去除 10，15 号数据点，得到最终结果

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_5 \quad \text{其中 } \beta_0 = 119.4955,\ \beta_1 = -0.3623,\ \beta_2 = -4.0411,\ \beta_3 = -0.1774$$

参数的置信区间依次为 [94.6827, 144.3084], [-0.5991, -0.1255], [-5.3617, -2.7205], [-0.3030, -0.0518]。$R^2 = 0.8625$, $F = 37.6269$, $s^2 = 4.44$, $P = 0.0000$。与剔除异常点之前的模型相比，此模型的结果更加准确了。

2.5 结论

1. 若只能选择 1 个变量，应建立模型为：$y = \beta_0 + \beta_1 x_3 \quad$ 其中 $\beta_0 = 83.4438,\ \beta_1 = -5.6682$，即 1500 米跑所用时间是耗氧能力的决定因素；

2. 若选择 2 个变量，应建立模型为：$y = \beta_0 + \beta_1 x_1 + \beta_2 x_3$　其中 $\beta_0 = 90.8529$, $\beta_1 = -0.1870$, $\beta_2 = -5.4671$，即 1500 米跑所用时间、年龄均对耗氧能力有一定影响；

3. 若不限制变量个数，剔除异常点后得到模型：$y = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_5$　其中 $\beta_0 = 119.4955$, $\beta_1 = -0.3623$, $\beta_2 = -4.0411$, $\beta_3 = -0.1774$，即认为跑步后心速也是影响耗氧能力的因素；

4. 通过模型可以看出，其他条件相同时，年龄越大，耗氧能力越低；1500 米跑所用时间越长，耗氧能力越低；而跑步后心速越慢，说明耗氧能力越高。

3 Ch13-P9 泡沫高度

3.1 模型建立与算法设计

根据题面描述，本题可以使用线性回归模型来处理，使用搅拌程度 x_1 与洗衣粉用量 x_2，将搅拌程度视为普通变量时，可以建立如下模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

为了求解上述模型，可以选择 MATLAB 的 `rstool`，根据剩余方差的情况判断是否需要加入交互项。得到模型结果后，可以使用 `rcoplot` 作出残差图，判断模型是否合理。

如果将搅拌程度视为没有定量关系的 3 个水平，此时可以选用 2 个 0-1 变量 z_0, z_1 来描述它们，其中 (0, 0) 表示搅拌程度 1，(0, 1) 表示搅拌程度 2，(1, 0) 表示搅拌程度 3，此时只需要将上述模型中的一次项和交互项更改为使用 z_0, z_1 变量即可。该模型的处理方法与上述模型相同，具体过程在计算结果部分描述。

3.2 Matlab 程序

第 (1) 问代码　如下

```
%% import data
data = reshape(readmatrix('p9.txt'), 15, []);
x1 = data(:, 1);
x2 = data(:, 2);
y = data(:, 3);

%% plot
figure(1);
subplot(1, 2, 1), plot(x1, y, '*'), grid, xlabel('Level of Stirring'),
    ylabel('Height of Foams');
subplot(1, 2, 2), plot(x2, y, '*'), grid, xlabel('Amount of Detergent'),
    ylabel('Height of Foams');

[b, bint, r, rint, s] = regress(y, [ones(size(x1)), x1, x2]);
figure(2);
rcoplot(r, rint);
```

第 (2) 问代码 使用如下代码将 x_1 转换为 0-1 变量并回归

```
% binary vars
z0 = zeros(size(x1));
z1 = zeros(size(x1));
z0 = z0 + (x1 == 3);
z1 = z1 + (x1 == 2);

%% choose vars
stepwise([z0, z1, x2], y);

%% regress
[b, bint, r, rint, s] = regress(y, [ones(size(z0)), z0, z1, x2]);
figure(2);
rcoplot(r, rint);
```

第 (3) 问代码 如下

```
rstool([x1, x2], y, 'interaction');
```

3.3 计算结果与分析

3.3.1 搅拌程度 x_1 视为普通变量

首先可以作出 x_1, x_2 作为单一变量时与因变量 y 的图线，如下

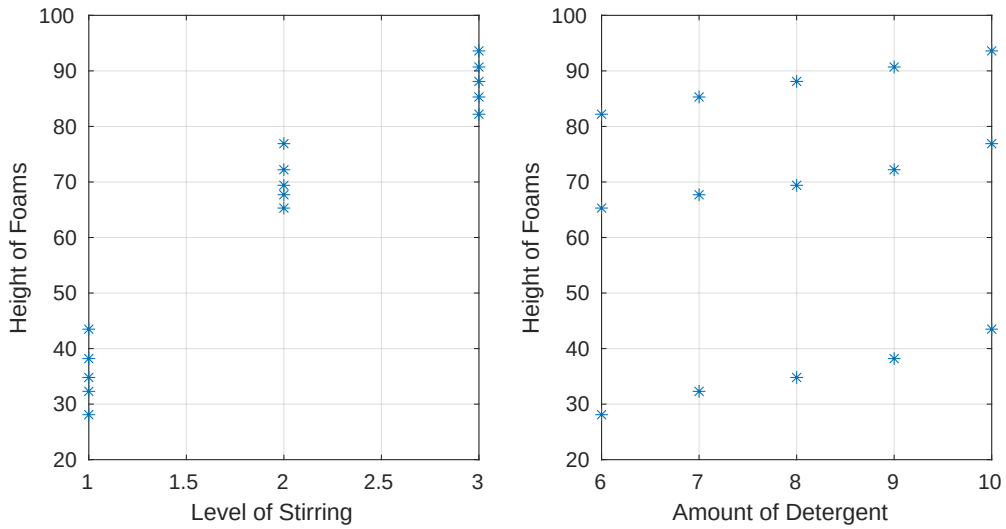

图 7: y 关于 x_1, x_2 的图线

从图中可以看出，y 与 x_1, x_2 之间应有近似线性的关系。使用 `regress` 进行回归，得到如下结果

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \text{ 其中 } \beta_0 = -12.74,\ \beta_1 = 26.3,\ \beta_2 = 3.0867$$

三个参数的置信区间分别为 [-29.0268, 3.5468], [23.1059, 29.4941], [1.2426, 4.9308]。$R^2 = 0.9654$, $F = 167.5754$, $P = 1.706 \times 10^{-9}$, $s^2 = 21.49$。使用 `rcoplot` 得到残差图，如下

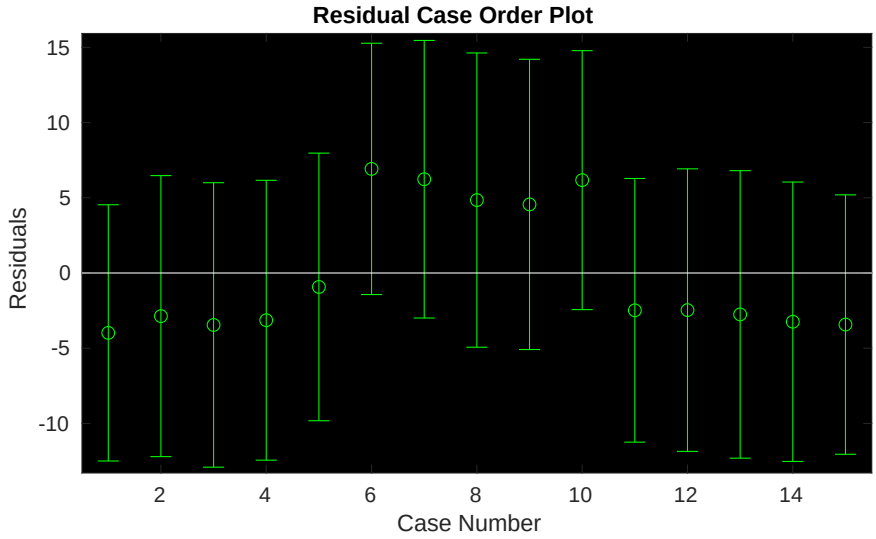

图 8: 将 x_1 视为普通变量时的残差图

从上述图片可以发现一个明显的问题：搅拌程度为 2 的残差与其他不同，说明该模型存在问题，下面尝试将 x_1 视为没有定量关系的 3 个水平。

3.3.2 将搅拌程度 x_1 视为没有定量关系的 3 个水平

使用 2 个 0-1 变量 z_0, z_1 来描述搅拌程度，其中 $(0, 0)$ 表示搅拌程度 1，$(0, 1)$ 表示搅拌程度 2，$(1, 0)$ 表示搅拌程度 3。

使用 `stepwise` 可知，当 z_0, z_1, x_2 均使用时，模型的 RMSE 最小，此时模型为

$$y = \beta_0 + \beta_1 z_0 + \beta_2 z_1 + \beta_3 x_2, \text{ 其中 } \beta_0 = 10.6867, \beta_1 = 52.60, \beta_2 = 34.92, \beta_3 = 3.0867$$

模型的 $R^2 = 0.9986$, $F = 2675.5$, $P = 5.01 \times 10^{-16}$, $s^2 = 0.9282$。可以看到 s^2 与 R^2 相比前一个模型有了显著提升。使用 `rcoplot` 得到残差图，如下

图 9: 将 x_1 视为没有定量关系的水平时的残差图

可以看出 5 号数据为异常点，剔除后再运行程序，得到

$$y = \beta_0 + \beta_1 z_0 + \beta_2 z_1 + \beta_3 x_2, \text{ 其中 } \beta_0 = 11.66,\ \beta_1 = 53.184,\ \beta_2 = 35.504,\ \beta_3 = 2.892$$

模型的 $R^2 = 0.9994$, $F = 5141.1$, $P = 3.09 \times 10^{-16}$, $s^2 = 0.4526$，可以看出，结果相比剔除异常点前有了进一步提升。此时残差图为

图 10: 剔除异常点后的残差图

3.3.3 引入交互项

当 x_1 视为普通变量时，使用 `rstool` 容易验证，引入交互项后 RMSE 增大 ($3.87 \to 4.55$)，故最佳模型仍为只有一次项。

当 x_1 为没有定量关系的 3 个水平时，使用 `stepwise` 手动选取引入哪些交互项，容易得到引入 z_0x_2, z_1x_2 之后模型的 RMSE 最小，使用 `regress` 回归后作出残差图，得到 10 号点异常，去除后结果如下

$$y = \beta_0 + \beta_1 z_0 + \beta_2 z_1 + \beta_3 x_2 + \beta_4 z_0 x_2 + \beta_5 z_1 x_2,$$

$$\text{其中 } \beta_0 = 6.02,\ \beta_1 = 59.4,\ \beta_2 = 45.83,\ \beta_3 = 3.67,\ \beta_4 = -0.85,\ \beta_5 = -1.43$$

模型的 $R^2 = 0.9997$, $F = 5108.1$, $P = 8.67 \times 10^{-14}$, $s^2 = 0.2856$。

可以看出，此模型的剩余方差比前面的其他模型都要小，是比较好的模型。

3.4 结论

- 将搅拌程度 x_1 视为没有定量关系的 3 个水平的建模方式在本题的情境中最为合理；
- 引入交互项后，仅对 x_1 视为没有定量关系的 3 个水平的模型有效，且引入交互项为表示搅拌程度的变量与洗衣粉用量，说明二者之间存在明显的相互作用；
- 最佳模型为引入交互项的 x_1 视为没有定量关系的 3 个水平的模型，具体数值参见计算结果部分。

4 Ch13-P13 高压锅销量

4.1 模型建立与算法设计

题目中给定了两种可用于拟合高压锅销量的模型，分别为 Logistic 模型

$$y_t = \frac{L}{1 + ae^{-kt}}$$

与 Gompertz 模型

$$y_t = Le^{-be^{-kt}}$$

当 L 不是固定参数时，上述两种模型显然不是可线性化的。当 L 给定后，Logistic 模型可以转化为如下的线性模型

$$-kt + \ln a = \ln(\frac{L}{y_t} - 1)$$

令 $\beta_0 = \ln a$, $\beta_1 = -k$，则

$$\ln(\frac{L}{y_t} - 1) = \beta_0 + \beta_1 t$$

此模型的因变量 $\ln(\frac{L}{y_t} - 1)$ 相对于参数 β_0, β_1 是线性的。

同理，L 给定后，Gompertz 模型也可以转化为如下的模型

$$\ln(\ln \frac{y_t}{L}) = \ln(-b) - kt$$

令 $\beta_0 = \ln(-b)$, $\beta_1 = -k$，得到的因变量对新参数也是线性的。

模型转化为线性模型后，可以使用 MATLAB 自带的 `regress` 函数进行回归。通过线性模型估计出非线性模型的参数后，以此为初值，可以对非线性模型进行拟合，这可以使用 MATLAB 的 `nlinfit` 或 `nlintool` 实现。

4.2 Matlab 程序

第 (1)(2) 问 代码如下

```
%% import data
data = reshape(readmatrix('p13.txt'), 13, []);
t = data(:, 1);
y = data(:, 2);

%% linearize Logistics
L = 3000;
Y = log(L ./ y - 1);
[b,bint,r,rint,stats] = regress(Y, [ones(size(t)), t]);
a = exp(b(1));
k = -b(2);


%% non-linear Logistics
b0 = [L, a, k];
nlintool(t, y, @logistics, b0);

```

```
%% model functions
function y = logistics(b, x)
    y = b(1) ./ (1 + b(2) .* exp(-b(3) .* x));
end
```

第 (3) 问 只需要修改拟合函数如下即可

```
function y = gompertz(b, x)
    y = b(1) .* exp(-b(2) .* exp(-b(3) .* x));
end
```

4.3 计算结果与分析

4.3.1 线性化 Logistic 模型回归

根据上面的分析，当 L 未知时，模型不可线性化；当 L 给定后，Logistic 模型是可线性化的。此时模型为

$$Y = \ln\left(\frac{L}{y_t} - 1\right) = \beta_0 + \beta_1 t,\ \beta_0 = \ln a,\ \beta_1 = -k$$

使用 `regress` 可以得到

$$\beta_0 = 3.8032,\ \beta_1 = -0.4941$$

线性化模型的 $R^2 = 0.9905$, $F = 1150.8$, $P = 1.7485 \times 10^{-12}$, $s^2 = 0.0386$。由上述结果得到估计的 $a = 44.8463$, $k = 0.4941$。

4.3.2 Logistic 模型非线性回归

使用上述线性模型得到的结果作为初值 $a^{(0)} = 44.8463$, $k^{(0)} = 0.4941$, $L^{(0)} = 3000$，利用 `nlintool` 进行拟合，结果如下

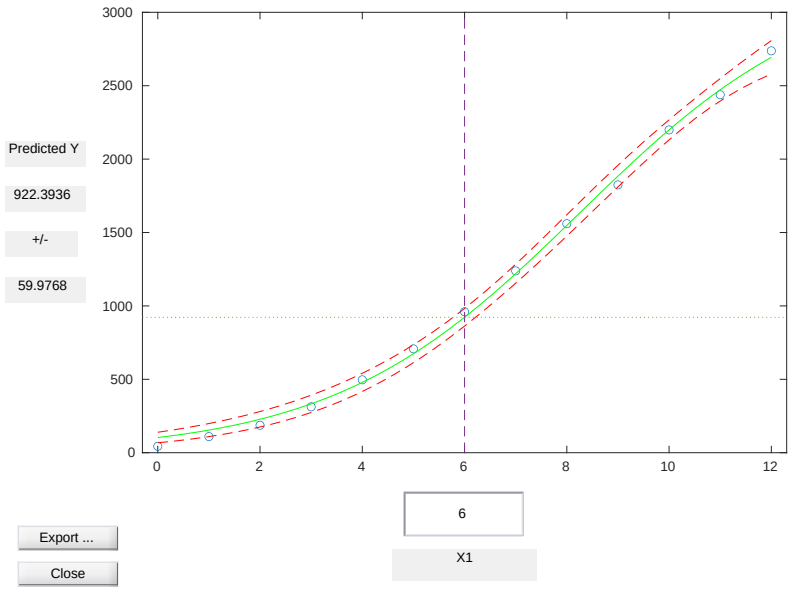

图 11: `nlintool` 交互式拟合 Logistics 模型的结果

拟合得到参数为

$$L = 3260.4,\ a = 30.5351,\ k = 0.4148$$

参数置信区间分别为 [2996.7, 3524.1], [24.8155, 36.2548], [0.3743, 0.4553], RMSE=42.0134。

4.3.3 Gompertz 模型非线性回归

使用初值 $b^{(0)} = 30$, $k^{(0)} = 0.4$, $L^{(0)} = 3000$ 拟合 Gompertz 模型，结果如下

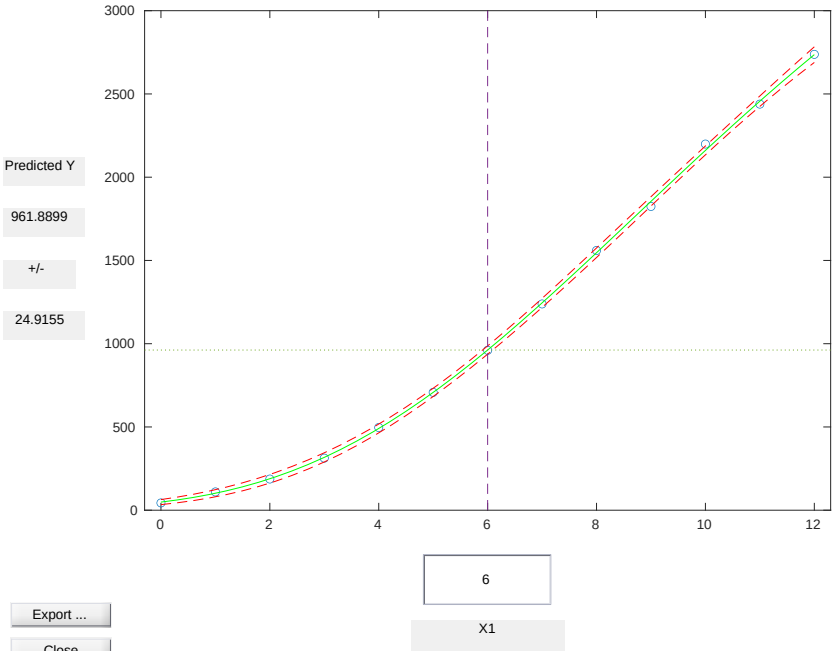

图 12: `nlintool` 交互式拟合 Gompertz 模型的结果

拟合得到参数为

$$L = 4810.1,\ b = 4.5920,\ k = 0.1747$$

参数置信区间分别为 [4428.9, 5191.3], [4.4429, 4.7410], [0.1622, 0.1873], RMSE=17.5539。

与 Logistics 模型相比，Gompertz 模型的 RMSE 明显下降，从拟合图线中也可以直观地观察到该模型比 Logistics 模型拟合效果更好，说明题给的情境更适用 Gompertz 模型。

4.4 结论

1. 使用 Logistics 与 Gompertz 模型拟合得到的结果见前一部分；
2. 题中所给的高压锅销量问题更适合使用 Gompertz 模型拟合。

5 收获与建议

通过这次实验，我掌握了使用 MATLAB 求解回归问题的一般方法，并对相关知识有了更深刻的理解。非常感谢本学期老师和助教的辛苦付出。