数学实验第十次实验报告

计算机系　计 43　2014011330　黄家晖

2017 年 5 月 29 日

1　实验目的

- 了解回归分析的基本原理，并学会解决实际问题；
- 练习使用 MATLAB 编程解决问题。

2　计算题

2.1　CH13-5 犯罪率研究

2.1.1　算法设计

对于该社会学问题，我们没有非常强的形式上的先验知识，因此采用简单方便的线性回归模型来处理这个问题。我们首先绘制收入情况、失业率和人口总数这些单一变量，观察他们单独对犯罪率的影响，确定函数的选取，并通过回归系数的置信区间、R^2 和 p 值、以及异常点个数来综合衡量模型的好坏。在分析的时候，从题目出发，由最简单的一次函数关系入手。如果效果不好或与图形观察结果不符合，再考虑交互项或是高次项。

根据这种思想，对于第一问，回归模型设计为：

$$y = \beta_0 + \beta_1 x_A + \beta_2 x_B + \epsilon$$

其中 x_A 和 x_B 需要尝试 x_1, x_2, x_3 的不同组合，可以采用 MATLAB 中提供的 `regress`函数进行尝试。

第二问引入了所有的自变量，但解法与第二问基本相同，对应的回归模型设计为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

在进行残差观察的时候，可以采用 MATLAB 提供的 `rcoplot`函数。

2.1.2　MATLAB 程序

主程序如下：

```
%% Math Exp Homework 10 13-T5
% Crime rate problem

%% Input data
y = [11.2 13.4 40.7 5.3 24.8 12.7 20.9 35.7 8.7 9.6 14.5 26.9 15.7 36.2 18.1 28.9 14.9 25.8
    21.7 25.7];
```

```
x1 = [16.5 20.5 26.3 16.5 19.2 16.5 20.2 21.3 17.2 14.3 18.1 23.1 19.1 24.7 18.6 24.9 17.9
    22.4 20.2 16.9];
x2 = [6.2 6.4 9.3 5.3 7.3 5.9 6.4 7.6 4.9 6.4 6 7.4 5.8 8.6 6.5 8.3 6.7 8.6 8.4 6.7];
x3 = [587 643 635 692 1248 643 1964 1531 713 749 7895 762 2793 741 625 854 716 921 595 3353 ];
n = length(y);

%% Draw simple plot
figure;
subplot(1, 3, 1);
scatter(x1, y, '+');
xlabel('x1'); ylabel('y');
subplot(1, 3, 2);
scatter(x2, y, '+');
xlabel('x2'); ylabel('y');
subplot(1, 3, 3);
scatter(x3, y, '+');
xlabel('x3'); ylabel('y')

%% Try out xA and xB
X = [ones(n, 1), x1', x2'];
[b, bint, r, rint, s] = regress(y', X);

%% Try out x1,x2,x3
X = [ones(n, 1), x1', x2', x3'];
[b, bint, r, rint, s] = regress(y', X)

%% Now draw plots
rcoplot(r, rint);

%% Eliminate bad points and recalculate
newX = X([1:7, 9:19], :);
y = y'; newY = y([1:7, 9:19], :);
[b, bint, r, rint, s] = regress(newY, newX);
```

2.1.3 计算结果和分析

第一问 首先画出收入情况、失业率和人口总数这些单一变量对犯罪率的影响图如图 1所示。从图中可以粗略看出，收入情况和失业率与犯罪率的关系大致成线性，但是人口总数与犯罪率在该图中并不能看出明显的线性相关性，可能对于模型的贡献不大。

使用 MATLAB 统计出的线性回归模型相关数值写在各个表中（表 1：$x_A = x_1, x_B = x_2$，表 2：$x_A = x_2, x_B = x_3$，表 3：$x_A = x_1, x_B = x_3$）。从表中可以看出，使用 x_1 和 x_2 两个变量得出的模型 R^2 值最高，且对应的剩余方差 s^2 最小。虽然其他的参数选择组合也接受了 x_3 对 y 有影响的假设，但是从数值上可以看出 x_3 系数 β 的置信区间包含 0。因此只选择两个变量，最好的模型参数应该由表 1决定，回归模型为：

$$y = -34.07 + 1.22x_1 + 4.4x_2$$

第二问 采用三变量的模型对回归模型进行计算，具体模型的形式请参见算法设计部分，计算结果如表 4所示。

从表中能够看出，使用了所有三个变量之后无论是 R^2 还是 F 值都比只使用 x_1 和 x_2 两个变量更高一些，代表了模型对问题描述的更加准确，且剩余方差有所降低，然而所有量降

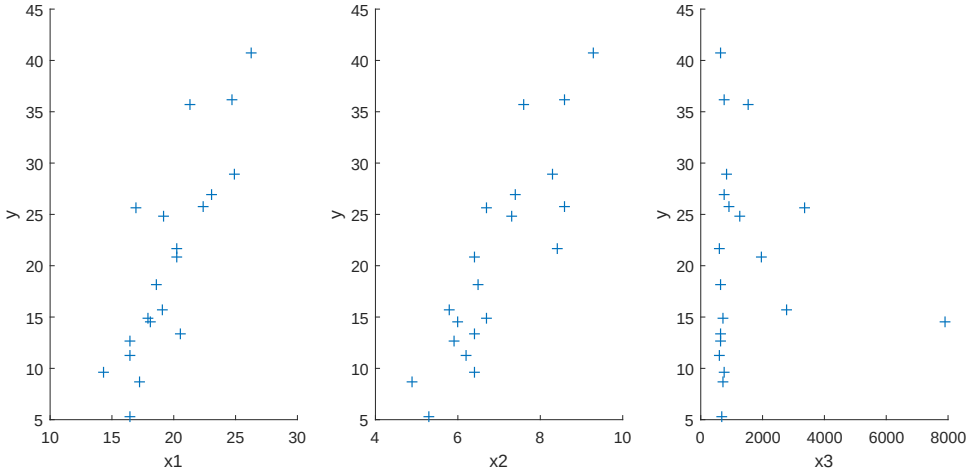

图 1: 收入情况、失业率和人口总数对犯罪率的影响

回归系数	回归系数估计值	回归系数置信区间
β_0	-34.07	$[-48.26, -19.88]$
β_1	1.22	$[0.03, 2.42]$
β_2	4.40	$[1.18, 7.62]$
$R^2 = 0.80, F = 34.43, p < 10^{-6}, s^2 = 21.61$		

表 1: 使用了 x_1 和 x_2 的线性回归模型计算结果

回归系数	回归系数估计值	回归系数置信区间
β_0	-31.60	$[-46.84, -16.36]$
β_1	7.35	$[5.27, 9.43]$
β_2	0.0	$[-6 \times 10^{-4}, 2 \times 10^{-3}]$
$R^2 = 0.76, F = 28.01, p < 4 \times 10^{-6}, s^2 = 25.41$		

表 2: 使用了 x_2 和 x_3 的线性回归模型计算结果

回归系数	回归系数估计值	回归系数置信区间
β_0	-31.22	$[-48.73, -13.70]$
β_1	2.60	$[1.74, 3.45]$
β_2	0.0004	$[-0.001, 0.002]$
$R^2 = 0.71, F = 20.84, p < 2 \times 10^{-5}, s^2 = 31.61$		

表 3: 使用了 x_1 和 x_3 的线性回归模型计算结果

回归系数	回归系数估计值	回归系数置信区间
β_0	-36.76	$[-51.63, -21.90]$
β_1	1.19	$[0.0015, 2.38]$
β_2	4.7198	$[1.48, 7.96]$
β_3	0.0008	$[-0.0006, 0.0021]$
$R^2 = 0.82, F = 24.02, p < 3 \times 10^{-6}, s^2 = 21.07$		

表 4: 使用了所有三个变量之后的线性回归模型计算结果

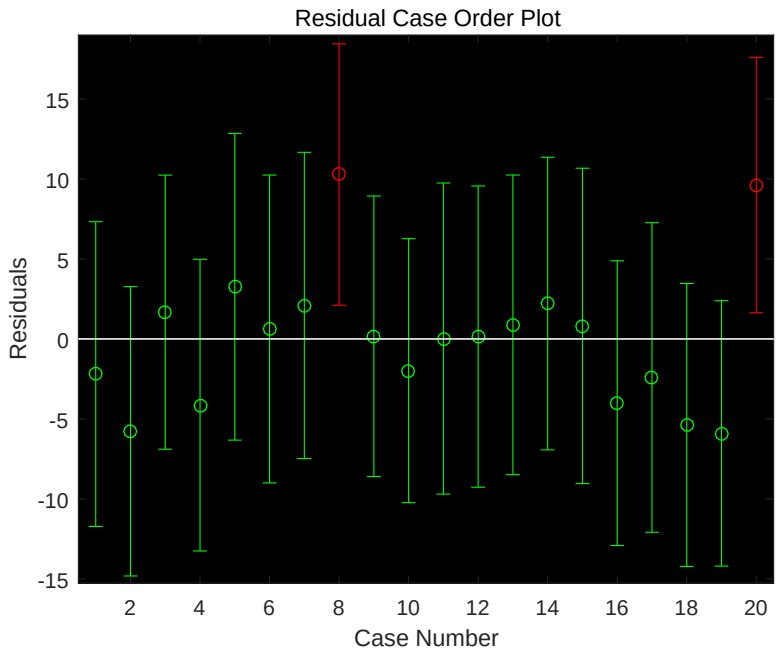

图 2: 使用第一问确定的最优模型残差图

低或增加的都不多，这代表着增加 x_3 对于该回归模型的贡献并不大。从 β_3 的数值和置信区间可以看出，x_3 仅仅与 y 存在很弱的相关性，置信区间包含 0，实际意义并不大。

从另一个方面来看，使用第一问确定的最优模型残差图如图 2所示，使用三个变量模型的残差图如图 3所示。对比这两幅图，可以看出三个变量残差图中异常点个数增多，说明使用了这种模型不容易拟合所有的数据点。考虑到三个变量模型的复杂性以及新加入的 x_3 对模型较小的贡献，我们依然认为最终模型为：

$$y = -34.07 + 1.22x_1 + 4.4x_2$$

第三问 最终模型的残差图如图 2所示，发现共有 2 个异常点（编号为 8 和 20），将他们去除之后，新回归模型的计算结果如表 5所示，残差图如图 4所示。

回归系数	回归系数估计值	回归系数置信区间
β_0	-35.71	$[-45.26, -26.16]$
β_1	1.60	$[0.78, 2.43]$
β_2	3.39	$[1.22, 5.57]$
$R^2 = 0.91, F = 78.39, p < 1 \times 10^{-8}, s^2 = 9.18$		

表 5: 去除了异常点之后的回归模型计算结果

对比图 2和图 4，以及表 1和表 5，能够看出：在去除了异常点之后残差图所有点的误差区间均包含 0，没有进一步出现异常点，此外，R^2 和 F 值也进一步增加，剩余方差大大减小，因此去除了这两个异常点可以算是合理选择。

去掉异常点之后的模型为：

$$y = -35.71 + 1.60x_1 + 3.39x_2$$

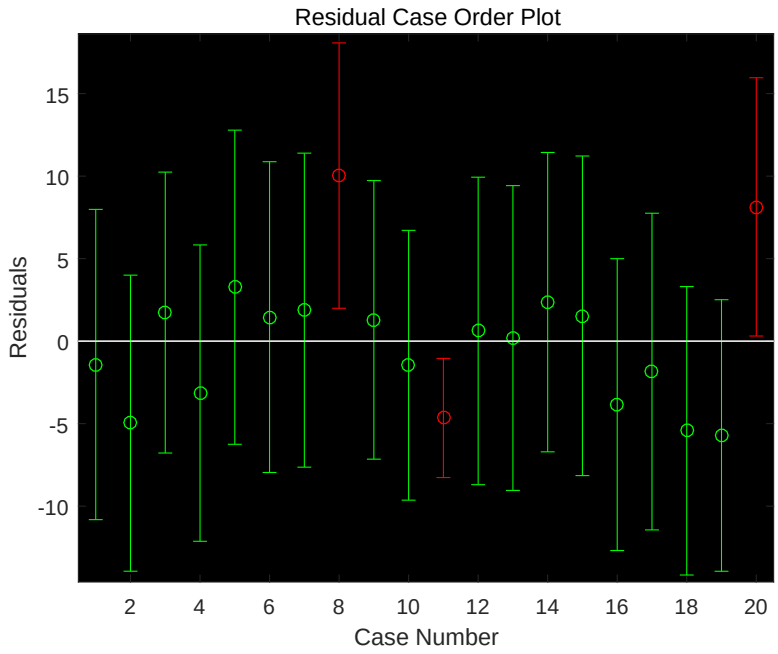

图 3: 使用三个变量模型的残差图

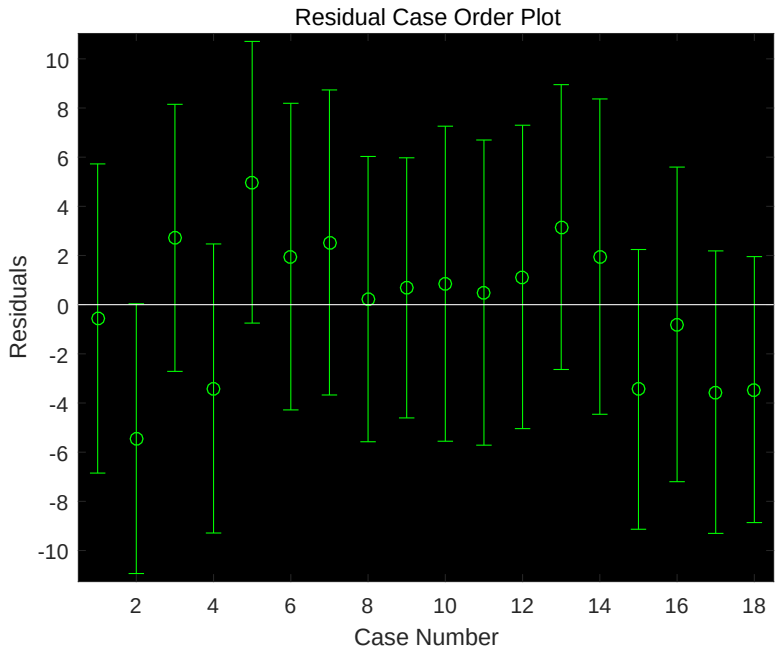

图 4: 使用第一问确定的最优模型去除异常点之后的残差图

从本题的分析可以看出，犯罪率的确与收入、失业率和人口规模相关，但是和收入与失业率的相关度最大。具体来说，犯罪率与年收入低于 5000 美元家庭的百分比成正比关系，与失业率成正比关系。

2.2 CH13-10 人寿保险额的影响因素

2.2.1 算法设计

根据题面描述，本题可以选择的自变量有 $x_1, x_2, x_1x_2, x_1^2, x_2^2$，因此可以给出二次函数回归模型：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon$$

为了求解上述模型可以使用 MATLAB 自带的 `rstool`功能的 `quadratic`等选项进行计算，选择剩余方差最小的模型，同时也可以选择 MATLAB 自带的 `stepwise`进行交互式逐步分析。本题采用 `rstools`进行分析，而下一题采用 `stepwise`进行分析。

2.2.2 MATLAB 程序

主程序如下：

```
%% Math Exp Homework 10 13—T10
% Insurance problem

%% Input data
y = [196 63 252 84 126 14 49 49 266 49 105 98 77 14 56 245 133 133];
x1 = [66.290 40.964 72.996 45.010 57.204 26.852 38.122 35.840 75.796 ...
    37.408 54.376 46.186 46.130 30.366 39.060 79.380 52.766 55.916];
x2 = [7 5 10 6 4 5 4 6 9 5 2 7 4 3 5 1 8 6];
n = length(y);

%% Test rstool
rstool([x1' x2'], y', 'quadratic');

%% Use the prediction to do regression
X = [ones(n, 1), x1', x2', x1' .* x2', x1' .^2, x2' .^ 2];
X_new = X([1:2, 4, 6, 8:18],:);
y_new = y'; y_new = y_new([1:2, 4, 6, 8:18],:);
%[b, bint, r, rint, s] = regress(y', X)
[b, bint, r, rint, s] = regress(y_new, X_new)

%% Barplot
rcoplot(r, rint);
```

2.2.3 计算结果和分析

分别使用 `rstool`提供的四种模型计算结果如表 6所示。各种模型计算图形输出见图 5, 6, 7, 8所示。

能够看出，使用了 `(full) quadratic`模型计算得到的 RMSE 最小，最合适作为最终模型，因此使用这种方法得到的最终模型为：

$$y = -65.39 + 1.02x_1 + 5.22x_2 + 0.036x_1^2 + 0.17x_2^2 - 0.0196x_1x_2$$

	β_0	β_1	β_2	β_3	β_4	β_5	s
linear	-158.77	4.84	5.20				9.25
interaction	-149.72	4.71	3.32	0.29			9.52
purequadratic	-60.91	0.93	4.45	0.04	0.11		1.81
quadratic	-65.39	1.02	5.22	-0.02	0.036	0.17	1.74

表 6: 使用 `rstool`的计算结果

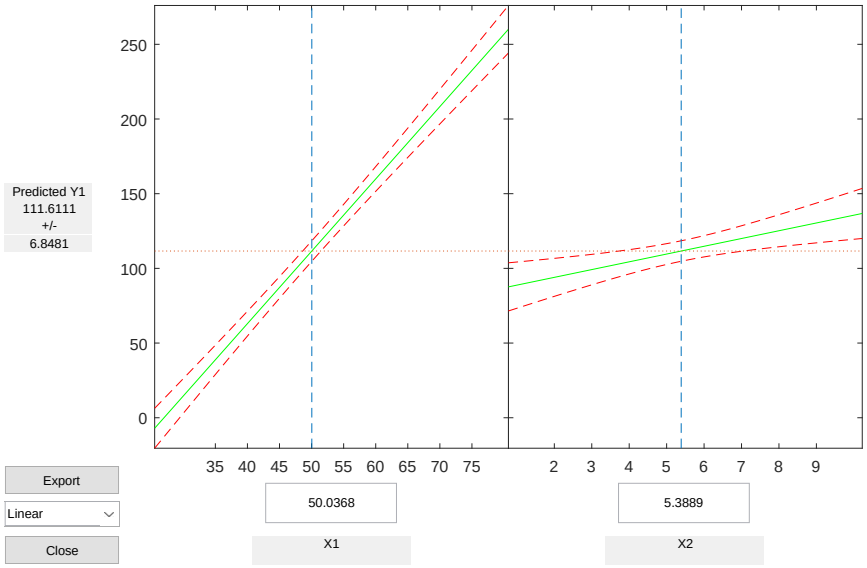

图 5: 在 rstool 中使用 linear 模型得到的计算输出

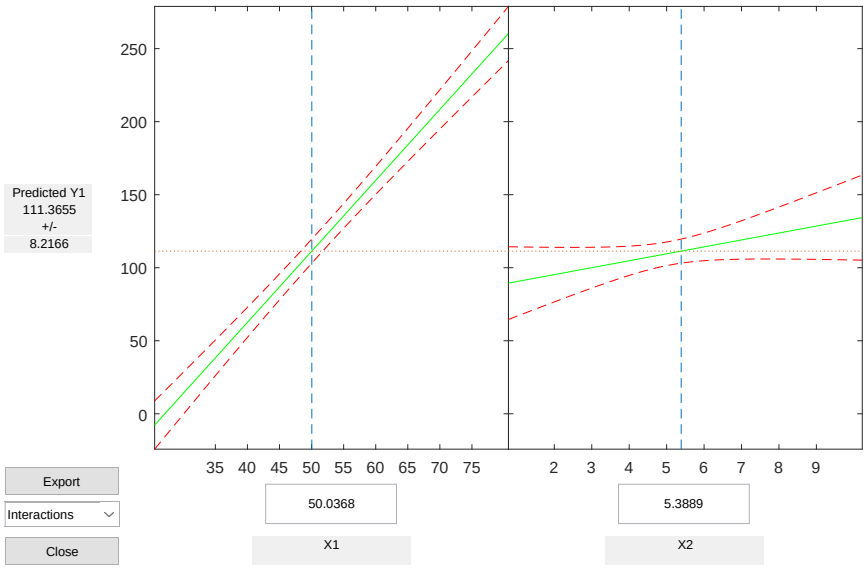

图 6: 在 rstool 中使用 interaction 模型得到的计算输出

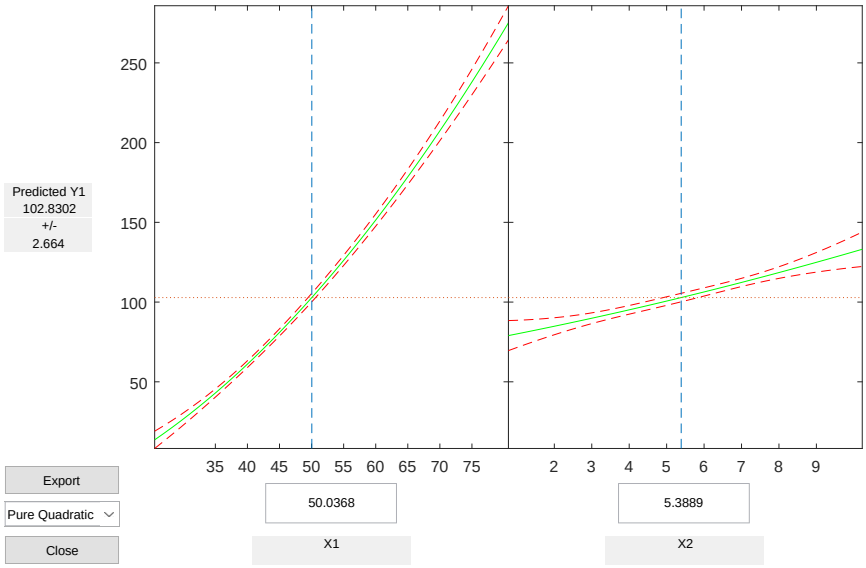

图 7: 在 rstool 中使用 purequadratic 模型得到的计算输出

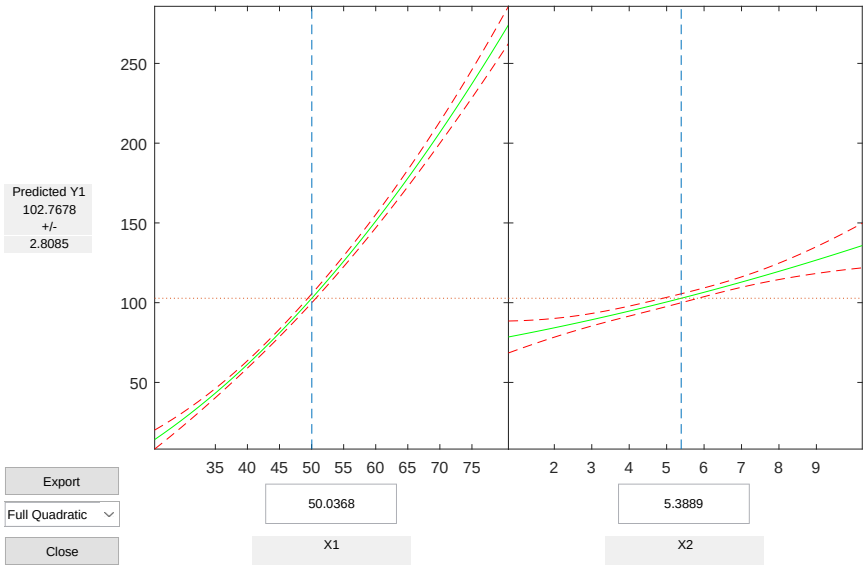

图 8: 在 rstool 中使用 quadratic 模型得到的计算输出

另外，继续使用 `regress`工具来计算每个参数的置信区间，得出的结果如表 7所示。另外使用该模型画出的残差图如图 9。从图中能够观察出 2 个异常点，不断去掉异常点直到不再出现，总共去除了 3 个点（分别为第 3、5、7 个数据点）之后新的模型结果如表 8所示，残差图如图 10所示。去除了异常点之后得到的模型为：

$$y = -64.7762 + 1.1074x_1 + 3.8722x_2 + -0.0160x_1x_2 + 0.0347x_1^2 - 0.2916x_2^2$$

回归系数	回归系数估计值	回归系数置信区间
β_0	-65.3856	$[-78.7265, -52.0446]$
β_1	1.0171	$[0.5202, 1.5141]$
β_2	5.2171	$[2.2785, 8.1558]$
β_3	-0.0195	$[-0.0501, 0.0109]$
β_4	0.0357	$[0.0310, 0.0405]$
β_5	0.1661	$[-0.0956, 0.4279]$
$R^2 = 0.9997, F = 7110.2023, p < 10^{-20}, s^2 = 3.0381$		

表 7: 去除异常点之前的回归模型计算结果

回归系数	回归系数估计值	回归系数置信区间
β_0	-64.7762	$[-72.7255, -56.8270]$
β_1	1.1074	$[0.8381, 1.3768]$
β_2	3.8722	$[1.7991, 5.9453]$
β_3	-0.0160	$[-0.0331, 0.001]$
β_4	0.0347	$[0.0321, 0.0373]$
β_5	0.2917	$[0.0961, 0.4871]$
$R^2 = 0.99991, F = 20637.1319, p < 10^{-20}, s^2 = 0.8105$		

表 8: 去除异常点之后的回归模型计算结果

根据上述分析可以看出，这些经理的人寿保险额不仅仅和年均收入以及风险偏好度有较强的线性关系，还对两个变量有着一定的二次关系和交互效应，本题的计算结果也从某种角度验证了研究人员心中的疑问。通过回归模型确认风险偏好度对人寿保险额有着二次效应、两个自变量对人寿保险额也有一定的交互效应。

2.3 CH13-11 止痛剂的疗效

2.3.1 算法设计

本题的思路与前两题类似，对于药物的实际疗效，并没有确切的数学关系来进行表达，所以需要进行回归分析。

设病痛减轻时间为 y，用药剂量为 x_1，性别为 x_2，血压组别为 x_3。假设 y 的影响因素有很多，例如线性项 x_1, x_2, x_3、二次项 x_1^2, x_2^2, x_3^2 和交互项 x_1x_2, x_1x_3, x_2x_3 等等。上述变量作为候选集合非常庞大，因此可以采用逐步回归的方式依次选择引入和剔除的变量，MATLAB 自带的工具箱函数 `stepwise`就提供了这样的功能，可以直接进行使用。

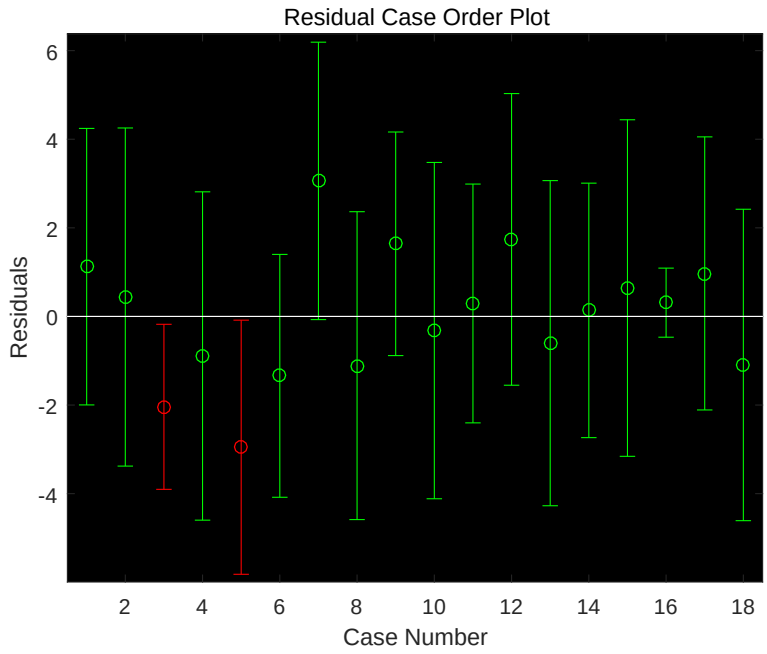

图 9: 去除异常点之前的回归模型残差图

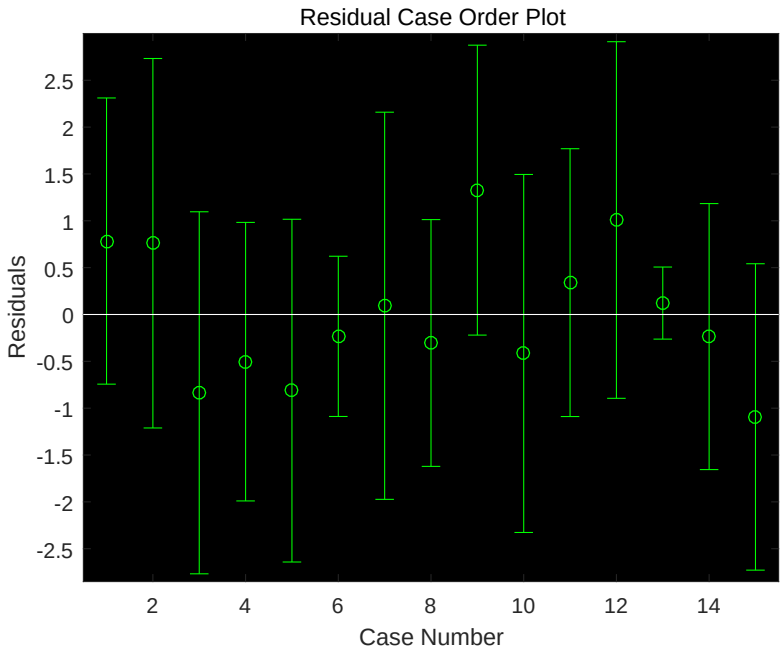

图 10: 去除异常点之后的回归模型残差图

2.3.2 MATLAB 程序

主程序如下：

```
%% Math Exp Homework 10 13—T11
% Drugs and pains

%% Input data
y = [35,43,55,47,43,57,26,27,28,29,22,29,19,11,14,23,20,22,13,8,3,27,26,5]';
x1 = [2,2,2,2,2,2,5,5,5,5,5,5,7,7,7,7,7,7,10,10,10,10,10,10]';
x2 = [0,0,0,1,1,1,0,0,0,1,1,1,0,0,0,1,1,1,0,0,0,1,1,1]';
x3 = [0.25,0.50,0.75,0.25,0.50,0.75,0.25,0.50,0.75,0.25,0.50,0.75,0.25,0.50,...
    0.75,0.25,0.50,0.75,0.25,0.50,0.75,0.25,0.50,0.75]';
n = length(y);

%% Draw intuitive plot
figure;
subplot(1, 3, 1);
scatter(x1, y, '+');
xlabel('x1'); ylabel('y');
subplot(1, 3, 2);
scatter(x2, y, '+');
xlabel('x2'); ylabel('y');
subplot(1, 3, 3);
scatter(x3, y, '+');
xlabel('x3'); ylabel('y')

%% Prepare some features
x12 = x1 .* x2; % x4
x23 = x2 .* x3; % x5
x13 = x1 .* x3; % x6
xx1 = x1 .^ 2; % x7
xx2 = x2 .^ 2; % x8
xx3 = x3 .^ 2; % x9
X = [x1 x2 x3 x12 x23 x13 xx1 xx2 xx3];

%% Interactive stepwise
stepwise(X, y);

%% Use the prediction to do regression
filter = [1:3, 5:7, 9:18, 20:22];
X = [ones(n, 1), x1, x12, x13, xx1, xx3];
X_new = X(filter,:);
y_new = y(filter,:);
[b, bint, r, rint, s] = regress(y_new, X_new)

%% Barplot
rcoplot(r, rint);
```

2.3.3 计算结果和分析

首先画出用药剂量、性别和血压这些单一变量对病痛减轻时间的影响图如图 11所示。从图中可以粗略看出，用药剂量与病痛减轻时间的关系大致成线性，但是其他变量与病痛减轻时间不能看出明显的线性或是二次关系，这说明在下面的测试中需要尝试引入交互项。

由于本题使用的工具箱 `stepwise`是一个交互式变量选择的函数，因此下面将以图的方式展示变量选择的过程：

1. 初始状况如图 12所示；

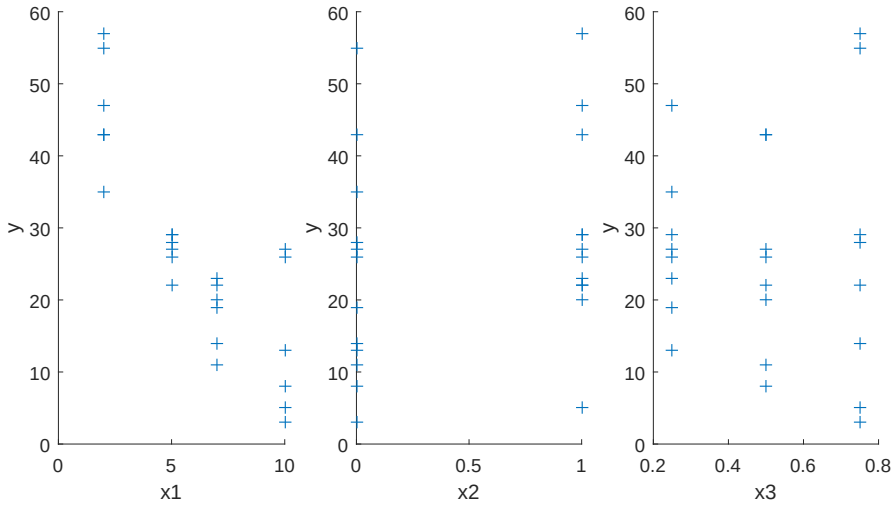

图 11: 用药剂量、性别和血压对病痛减轻时间的影响

2. 首先使用程序默认的 Next Step 按钮进行自动调节，最终调节的结果选择了变量 x_1, x_1x_2, x_1^2，如图 13所示，自动调节所能到达的最小 RMSE 为 6.298；
3. 根据图 11的分析，x_3 本身可能和 y 无明显线性关系，但是其交互项可能会对模型有直接影响，因此考虑手动加入变量 x_1x_3，如图 14所示，RMSE 降低到 5.954；
4. 进一步根据提示进行调整，加入变量 x_3^2，如图 15所示，此时 RMSE 降低到 4.016，经过其他尝试已经无法降低 RMSE，此即为模型的最优 RMSE。

最终得到的模型为：

$$y = 52.8084 - 7.0608x_1 + 0.9551x_1x_2 - 7.3746x_1x_3 + 0.5111x_1^2 + 42.5282x_3^2$$

另外，继续使用 `regress`工具来计算每个参数的置信区间，得出的结果如表 9所示。另外使用该模型画出的残差图如图 16。从图中能够观察出 2 个异常点，不断去掉异常点直到不再出现，总共去除了 4 个点（分别为第 4、8、23、24 个数据点）之后新的模型结果如表**??**所示，残差图如图 17所示。去除了异常点之后得到的模型为：

$$y = 46.7237 - 5.611x_1 + 0.7045x_1x_2 - 9.1549x_1x_3 + 0.4938x_1^2 + 55.1979x_3^2$$

通过此题也能够看出病痛的减轻时间不仅仅和用药剂量有关，还和患者本身的性别和血压有关系，在进行预测的时候，还可以参考其他 MATLAB 函数给出预测病痛减缓时间的区间分布，作为医疗参考。此外，对于患者来说，一定也急于了解病痛的减缓时间，医疗工作者也可以将预测信息提供给患者做进一步参考。

3 收获与建议

通过这次的实验，我学会了使用 MATLAB 求解回归分析问题的一般方法，并对回归分析的知识有了更深的理解。希望在之后的课堂上老师能够当堂进行相关的技巧演示并给出题目的分步解答。

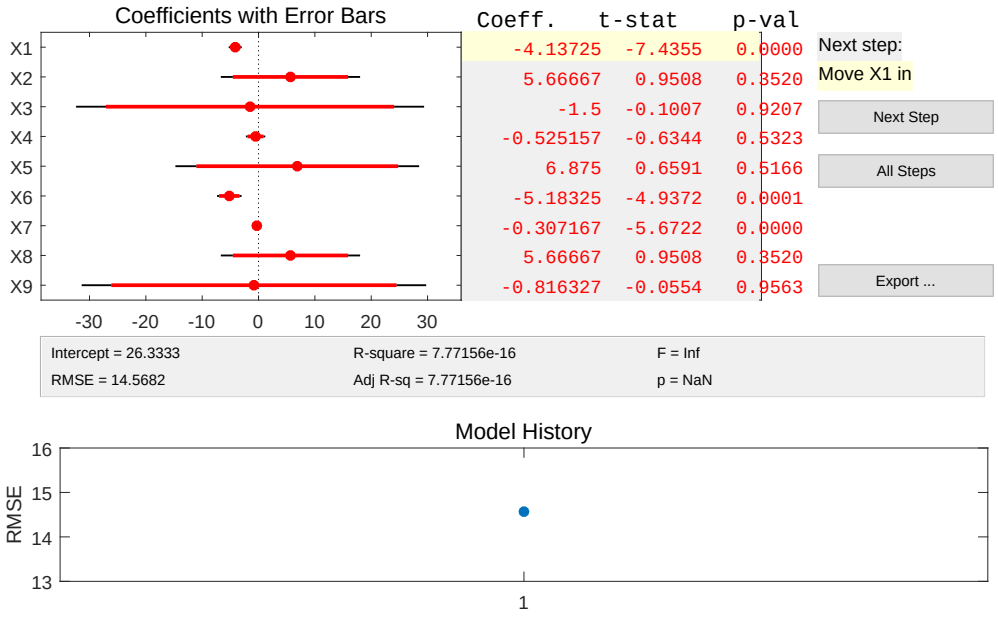

图 12: 使用 stepwise 进行调整时的初始状态

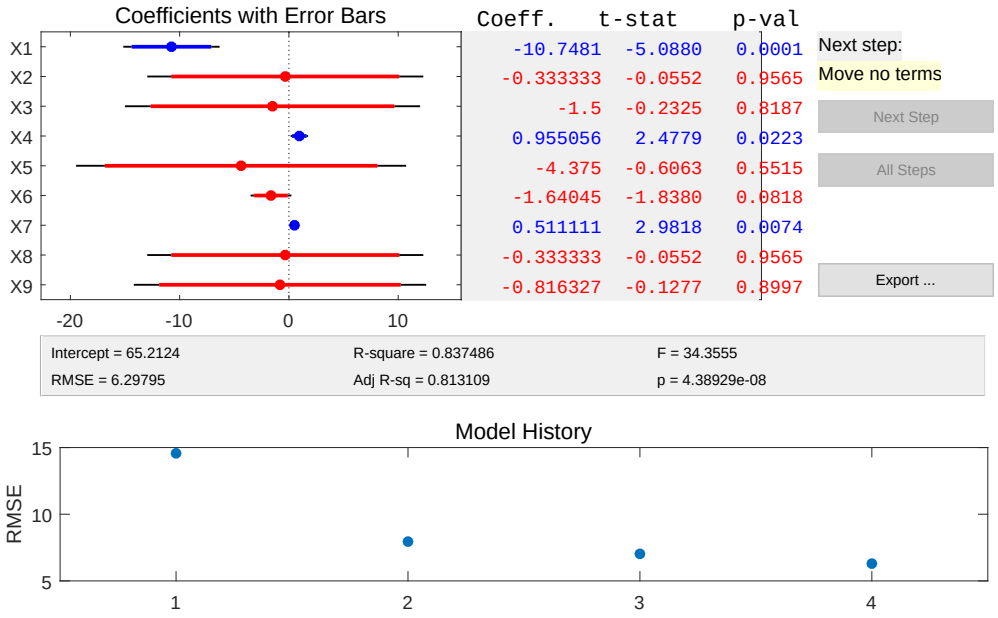

图 13: 使用 stepwise 进行第一步调整

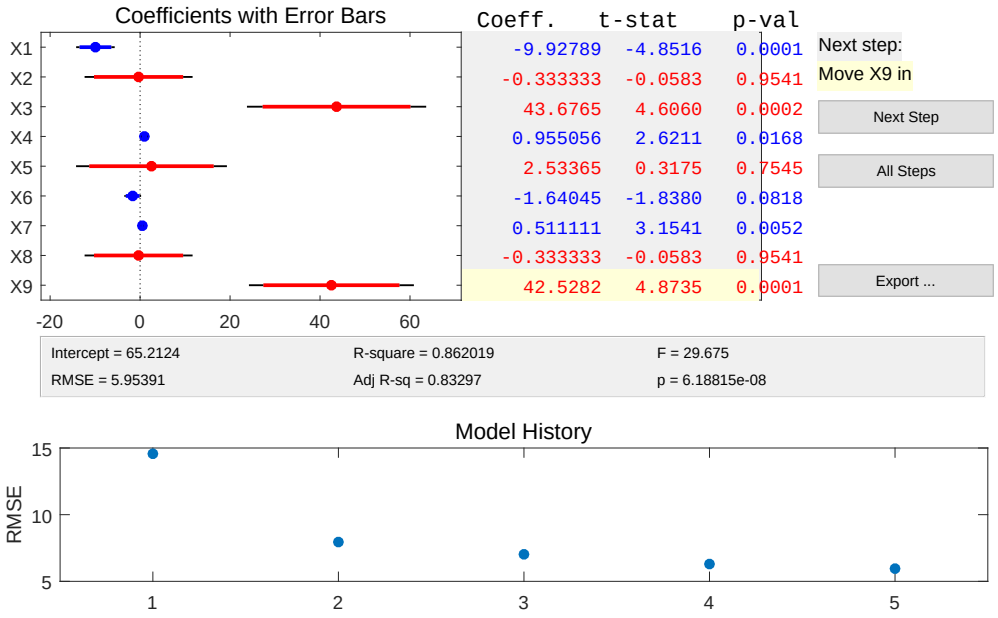

图 14: 使用 stepwise 进行第二步调整

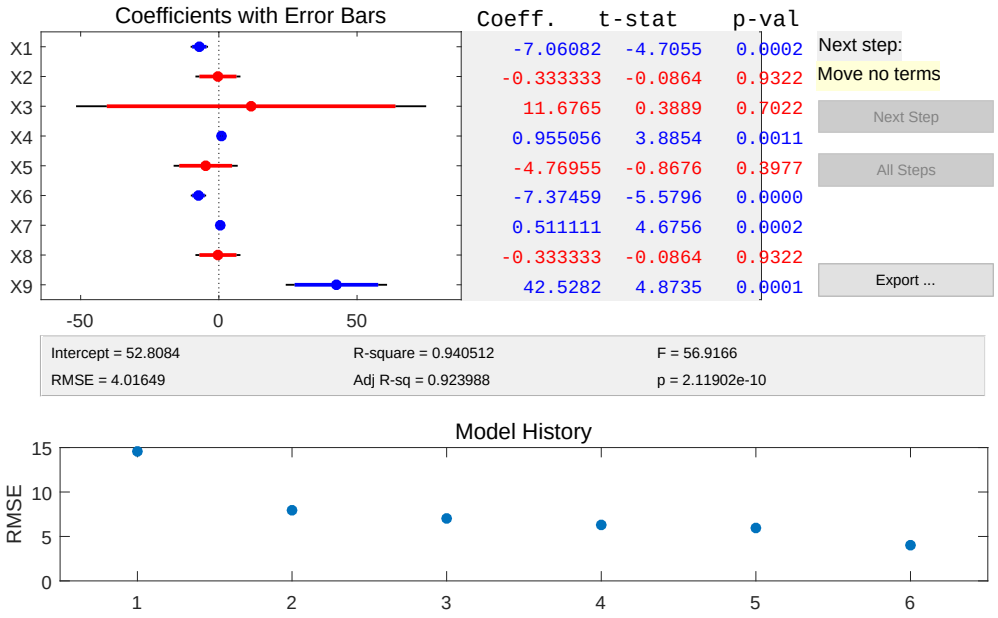

图 15: 使用 stepwise 进行第三步调整

回归系数	回归系数估计值	回归系数置信区间
β_0	52.8084	$[43.6424, 61.9743]$
β_1	-7.0608	$[-10.2133, -3.9082]$
β_2	0.9551	$[0.4386, 1.4714]$
β_3	-7.3746	$[-10.1514, -4.5978]$
β_4	0.5111	$[0.2814, 0.7408]$
β_5	42.5282	$[24.1946, 60.8618]$
$R^2 = 0.9405, F = 56.9166, p < 2 \times 10^{-10}, s^2 = 16.1321$		

表 9: 去除异常点之前的回归模型计算结果

回归系数	回归系数估计值	回归系数置信区间
β_0	46.7237	$[40.8625, 52.5849]$
β_1	-5.6106	$[-7.6900, -3.5313]$
β_2	0.7045	$[0.3433, 1.0657]$
β_3	-9.1549	$[-11.0728, -7.2368]$
β_4	0.4938	$[0.3533, 0.6341]$
β_5	55.1979	$[43.8045, 66.5913]$
$R^2 = 0.9863, F = 187.3349, p < 1 \times 10^{-11}, s^2 = 4.0129$		

表 10: 去除异常点之后的回归模型计算结果

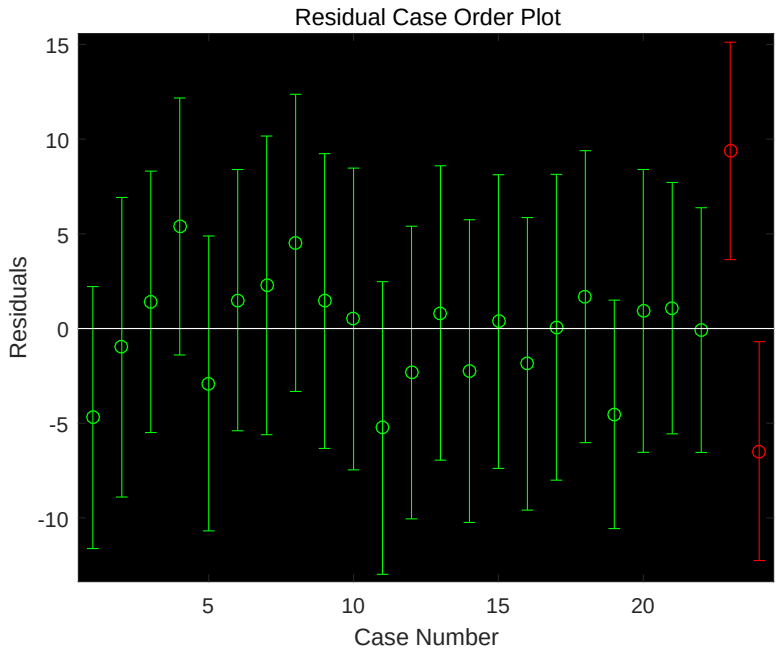

图 16: 去除异常点之前的回归模型残差图

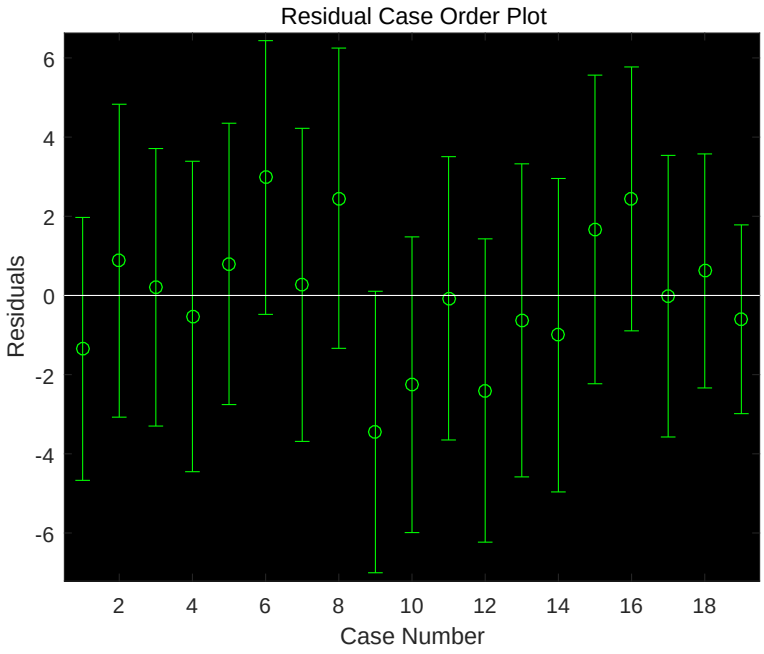

图 17: 去除异常点之后的回归模型残差图

本讲的内容实际上和目前很火的机器学习领域关系非常密切，希望老师也能够适当引入这方面的内容，与时俱进。非常感谢老师和助教这学期的辛苦付出。