重 庆 大 学

学 生 实 验 报 告

实验课程名称 数学实验

开课实验室 DS1407

学 院 计算机学院 年级 2021 专业班

计卓 2 班，计科 2 班，信安 1 班

学 生 姓 名 文红兵 学 号 20214590

学 生 姓 名 张奎元 学 号 20214358

学 生 姓 名 高志朋 学 号 20214141

开 课 时 间 2022 至 2023 学年第 二 学期

总 成 绩	

数 统 学 院 制

开课学院、实验室： DS1407　　实验时间 ： 2023 年　4 月　9 日

课程名称	数学实验	实验项目名　　称	回归模型	实验项目类型				
				验证	演示	综合	设计	其他
指导教师	肖剑	成　　绩				√		

题目

汽车销售商认为汽车销售量与汽油价格、贷款利率有关，两种类型汽车（普通型和豪华型）18 个月的调查资料如下表，其中 y_1 是普通型汽车销售量（千辆），y_2 是豪华型汽车销售量（千辆），x_1 是汽油价格（美元/加仑），x_2 是贷款利率（%）

序号	y_1	y_2	x_1	x_2
1	22.1	7.2	1.89	6.1
2	15.4	5.4	1.94	6.2
⋮	⋮	⋮	⋮	⋮
18	44.3	15.6	1.68	2.3

（1）对普通型和豪华型汽车分别建立如下模型：

$$y_1=\beta_0^{(1)}+\beta_1^{(1)}x_1+\beta_2^{(1)}x_2,\quad y_2=\beta_0^{(2)}+\beta_1^{(2)}x_1+\beta_2^{(2)}x_2$$

给出 β 的估计值和置信区间，决定系数 R^2，F 值及剩余方差等.

（2）用 $x_3=0$，1 表示汽车类型，建立统一模型：$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$，给出 β 的估计值和置信区间，决定系数 R^2，F 值及剩余方差等.以 $x_3=0$，1 代入统一模型，将结果与（1）的两个模型的结果比较，解释二者的区别.

（3）对统一模型就每种类型汽车分别作 x_1 和 x_2 与残差的散点图，有什么现象，说明模型有何缺陷？

（4）对统一模型增加二次项和交互项，考察结果有什么改进.

完整数据如下表：

序号	y_1	y_2	x_1	x_2
1	22.1	7.2	1.89	6.1
2	15.4	5.4	1.94	6.2
3	11.7	7.6	1.95	6.3
4	10.3	2.5	1.82	8.2
5	11.4	2.4	1.85	9.8
6	7.5	1.7	1.78	10.3
7	13	4.3	1.76	10.5
8	12.8	3.7	1.76	8.7
9	14.6	3.9	1.75	7.4
10	18.9	7	1.74	6.9
11	19.3	6.8	1.7	5.2
12	30.1	10.1	1.7	4.9
13	28.2	9.4	1.68	4.3
14	25.6	7.9	1.6	3.7
15	37.5	14.1	1.61	3.6
16	36.1	14.5	1.64	3.1
17	39.8	14.9	1.67	1.8
18	44.3	15.6	1.68	2.3

第一小问

模型 1.1

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

程序 1.1

```
clc; clear ;
y1=[22.1,15.4,11.7,10.3,11.4,7.5,13,12.8,14.6,18.9
,19.3,30.1,28.2,25.6,37.5,36.1,39.8,44.3]';
y2=[7.2,5.4,7.6,2.5,2.4,1.7,4.3,3.7,3.9,7,6.8,10.1
,9.4,7.9,14.1,14.5,14.9,15.6]';
x1=[1.89,1.94,1.95,1.82,1.85,1.78,1.76,1.76,1.75,1
.74,1.7,1.7,1.68,1.6,1.61,1.64,1.67,1.68]';
x2=[6.1,6.2,6.3,8.2,9.8,10.3,10.5,8.7,7.4,6.9,5.2,
4.9,4.3,3.7,3.6,3.1,1.8,2.3]';
X = [ones(size(x1)) x1 x2] ;
[b,bint,r,rint,stats] = regress(y1,X) ;

rcoplot(r,rint) ; % 残差图
```

```
%去除反常数据后代码

clc; clear ;
y1=[22.1,15.4,11.7,10.3,11.4,7.5,13,12.8,14.6,18.9
,30.1,28.2,37.5,36.1,39.8]';
x1=[1.89,1.94,1.95,1.82,1.85,1.78,1.76,1.76,1.75,1
.74,1.7,1.68,1.61,1.64,1.67]';
x2=[6.1,6.2,6.3,8.2,9.8,10.3,10.5,8.7,7.4,6.9,4.9,
4.3,3.6,3.1,1.8]';

X = [ones(size(x1)) x1 x2] ;
[b,bint,r,rint,stats] = regress(y1,X) ;

rcoplot(r,rint) ; % 残差图
```

结果 1.1

原始数据

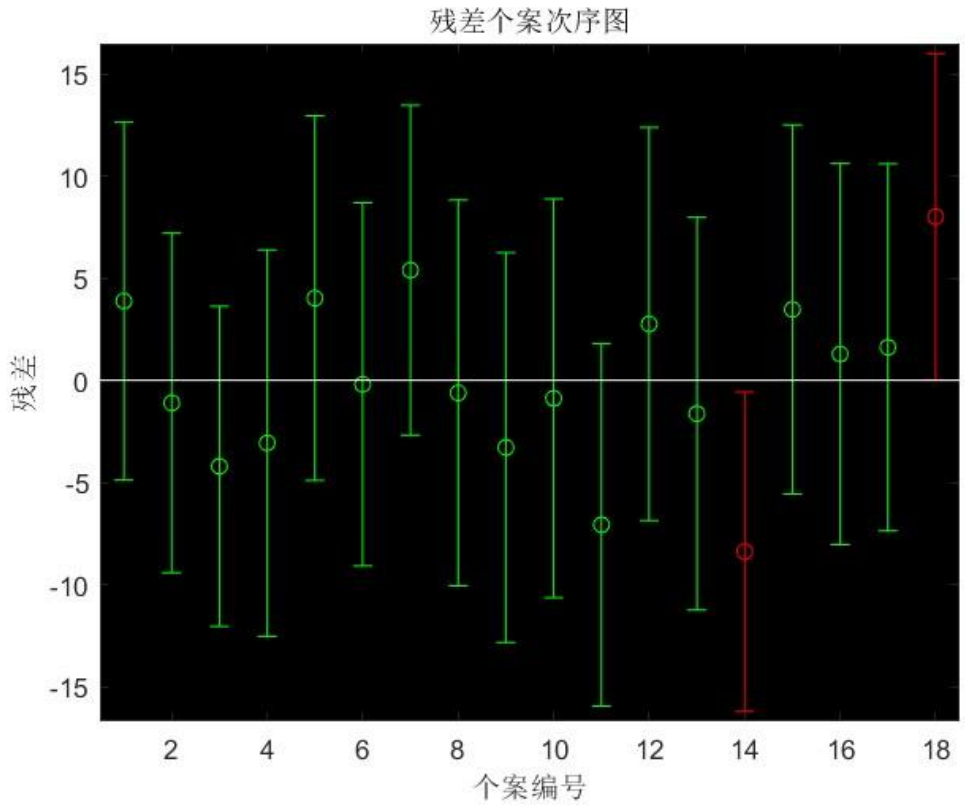
参数估计结果

	估计值	置信区间
β_1	90.1814	[46.1971 ,134.1656]
β_2	-27.6588	[-54.5542 ,-0.7634]
β_3	-3.2283	[-4.2747 ,-2.1819]

显著性分析

	估计值
R^2	0.8593
F	45.7992
σ^2	20.7910
P	0.0000

残差图



剔除反常数据后

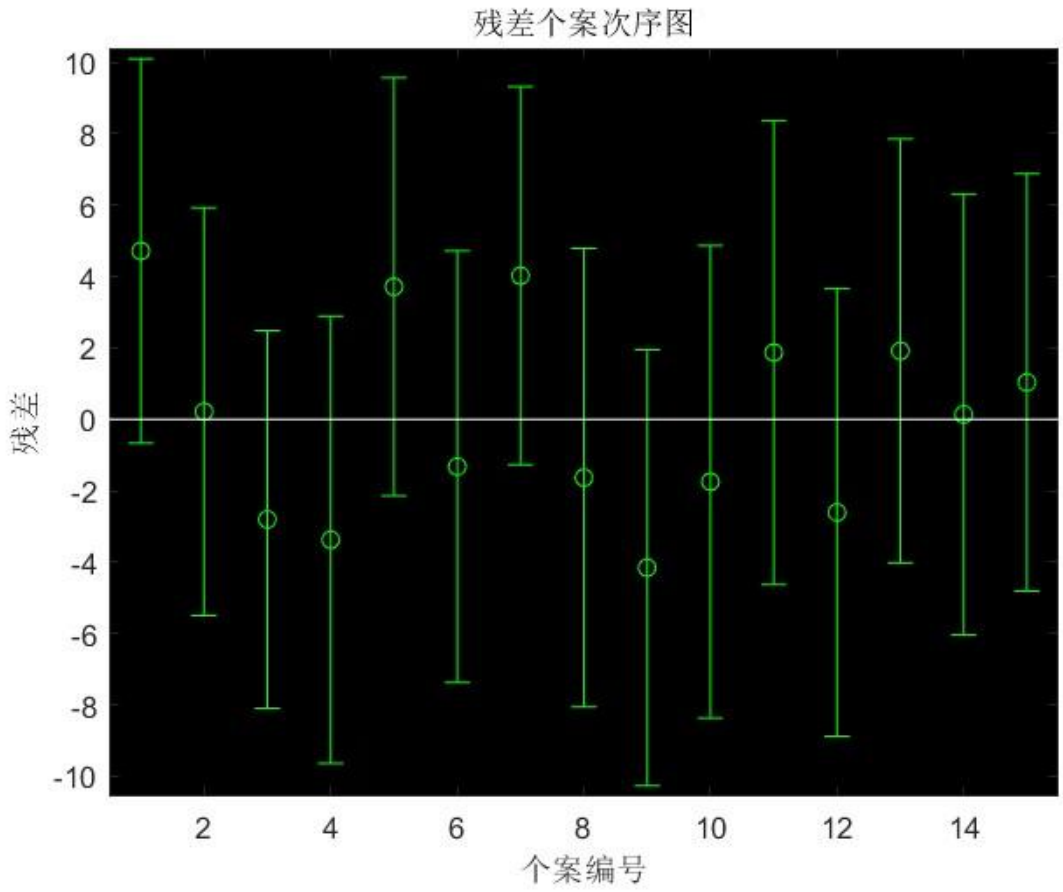
参数估计结果

	估计值	置信区间
β_1	107.5601	[75.3160,139.8042]
β_2	-37.9283	[-57.2842,-18.5723]
β_3	-3.0314	[-3.7862,-2.2767]

显著性分析

	估计值
R^2	0.9334
F	84.0758
σ^2	9.2746
P	0.0000

残差图



分析 1.1

通过对已知数据进行线性回归分析，对参数进行估计得到的结果如上述表格所示

其中 R 方的值为 0.8593，解释程度较高，F 估计值为 45.7992，远大于所设定置信度下的目标值，p 为 0，模型与被解释变量相关性较大。且参数置信区间中不含零点，因此该回归模型合适。

模型 1.2

$$y_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

程序 1.2

```
clc; clear ;%导入数据
y2=[7.2,5.4,7.6,2.5,2.4,1.7,4.3,3.7,3.9,7,6.8,10.1
,9.4,7.9,14.1,14.5,14.9,15.6]';
x1=[1.89,1.94,1.95,1.82,1.85,1.78,1.76,1.76,1.75,1
.74,1.7,1.7,1.68,1.6,1.61,1.64,1.67,1.68]';
x2=[6.1,6.2,6.3,8.2,9.8,10.3,10.5,8.7,7.4,6.9,5.2,
4.9,4.3,3.7,3.6,3.1,1.8,2.3]';
X = [ones(size(x1)) x1 x2] ;
[b,bint,r,rint,stats] = regress(y2,X);
rcoplot(r,rint) ; % 残差图
```

```
% 剔除反常数据后
clc; clear ;
y2=[7.2,5.4,7.6,2.5,2.4,1.7,3.7,3.9,7,10.1,9.4,14.
1,14.5,14.9,15.6]';
x1=[1.89,1.94,1.95,1.82,1.85,1.78,1.76,1.75,1.74,1
.7,1.68,1.61,1.64,1.67,1.68]';
x2=[6.1,6.2,6.3,8.2,9.8,10.3,8.7,7.4,6.9,4.9,4.3,3
.6,3.1,1.8,2.3]';
X = [ones(size(x1)) x1 x2] ;
[b,bint,r,rint,stats] = regress(y2,X) ;
rcoplot(r,rint) ; % 残差图
```

结果 1.2

原始数据

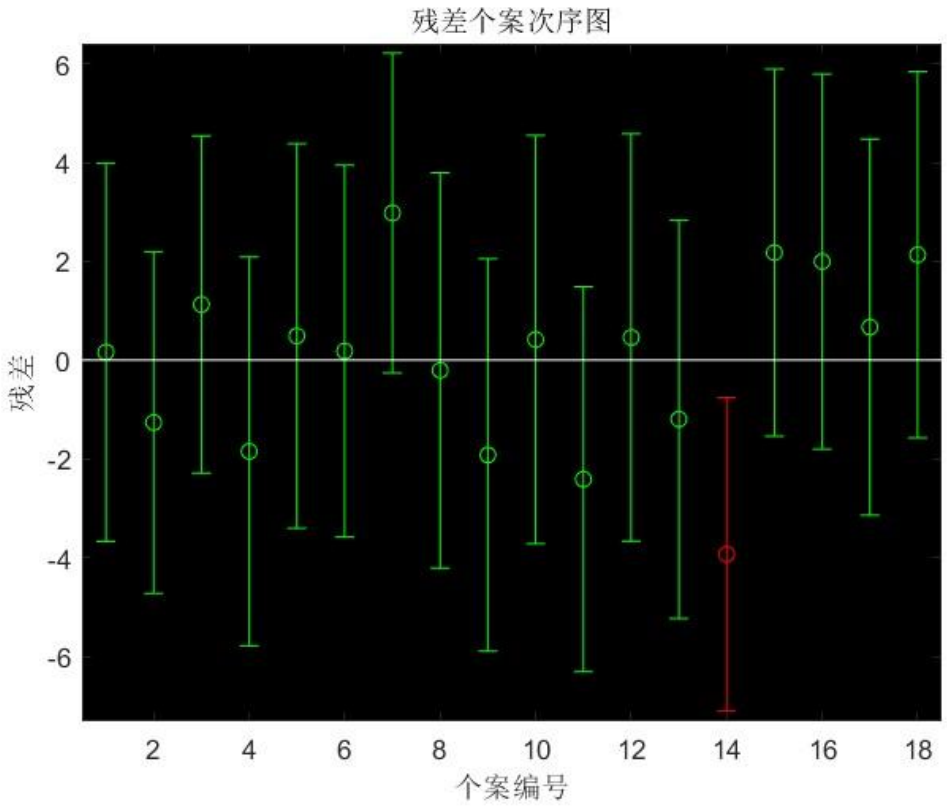
参数估计

	估计值	置信区间
β_1	24.5471	[5.9501,43.1740]
β_2	-4.6285	[-16.0184,6.7615]
β_3	-1.4360	[-1.8792,-0.9929]

显著性分析

	估计值
R^2	0.8402
F	39.4474
σ^2	3.7288
P	0.0000

残差图



剔除反常数据后

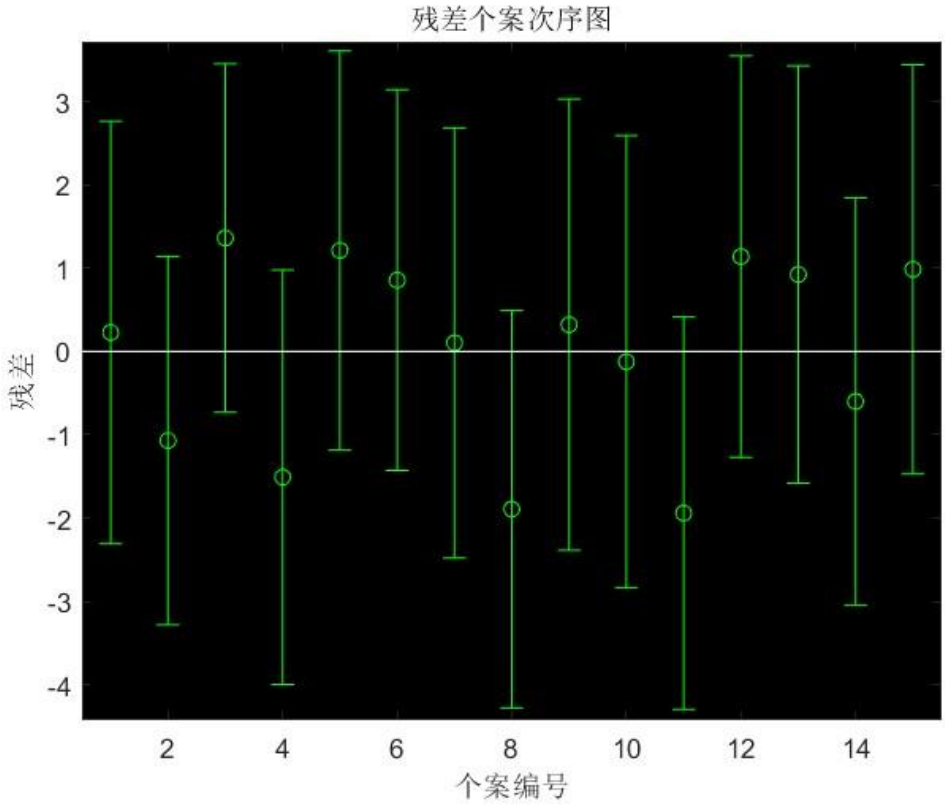
参数估计结果：

	估计值	置信区间
β_1	29.7583	[16.2864,43.2303]
β_2	-6.7738	[-14.9774 ,1.4299]
β_3	-1.6367	[-1.9680 ,-1.3054]

显著性分析结果：

	估计值
R^2	0.9450
F	103.1152
σ^2	1.5413
P	0.0000

残差图:



分析 1.2

通过对已知数据进行线性回归分析，对参数进行估计得到的结果如上述表格所示

其中 R 方的值为 0.8402，解释程度较高，F 估计值为 39.4474，远大于所设定置信度下的目标值，p 为 0，模型与被解释变量相关性较大。且参数置信区间中不含零点，因此该回归模型合适。

第二小问

模型 2

$$y_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

程序 2

```
clc ; clear ;
n = 18 ;
%开始插入数据
y=[22.1,15.4,11.7,10.3,11.4,7.5,13,12.8,14.6,18.9,19.3,3
0.1,28.2,25.6,37.5,36.1,39.8,44.3,7.2,5.4,7.6,2.5,2.4,1.
7,4.3,3.7,3.9,7,6.8,10.1,9.4,7.9,14.1,14.5,14.9,15.6]';
x1 =
[1.89,1.94,1.95,1.82,1.85,1.78,1.76,1.76,1.75,1.74,1.7,1
.7,1.68,1.6,1.61,1.64,1.67,1.68,1.89,1.94,1.95,1.82,1.85
,1.78,1.76,1.76,1.75,1.74,1.7,1.7,1.68,1.6,1.61,1.64,1.6
7,1.68]';
x2 =
[6.1,6.2,6.3,8.2,9.8,10.3,10.5,8.7,7.4,6.9,5.2,4.9,4.3,3
.7,3.6,3.1,1.8,2.3,6.1,6.2,6.3,8.2,9.8,10.3,10.5,8.7,7.4
,6.9,5.2,4.9,4.3,3.7,3.6,3.1,1.8,2.3]';
x3 = [zeros(1,n) ones(1,n)]' ;
X = [ones(size(x1)) x1 x2 x3] ;
[b,bint,r,rint,stats] = regress(y,X) ;
rcoplot(r,rint) ; % 残差图
```

结果 2

参数估计：

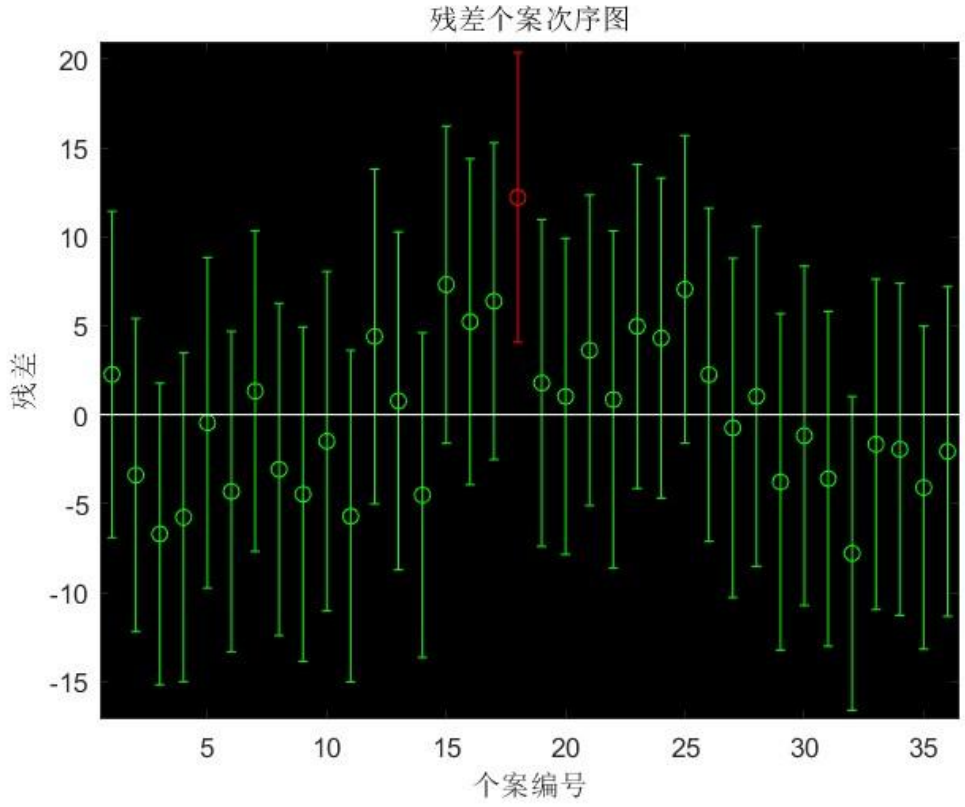
	估计值	置信区间
β_1	64.5753	[33.5007,95.6499]
β_2	-16.1436	[-35.1193 ,2.8320]
β_3	-2.3322	[-3.0705 ,-1.5939]
β_4	-14.4222	[-17.6546,-11.1898]

显著性分析：

	估计值
R^2	0.8366
F	54.6111

σ^2	22.6642
P	0.0000

残差图



分析 2

显著性分析：R 方的值为 0.8366，解释程度较高，F 估计值为 54.6111，远大于所设定置信度下的目标值，p 为 0，模型与被解释变量相关性较大。但参数置信区间中 x2 参数区间中含有零点，这也为后面对交互项的分析留下原因，但是总体而言该回归模型相对合适。

加入 x3 变量后，线性回归得到的结果与原独立模型下结果并不一致。统一模型中的参数估计接近原独立模型中参数的平均值。分析原因如下

1、 基于分析问题的不同，独立模型使用数据集中样本数为 18，统一模型数据集中样本数则为 36，且样本维度不同，因此结果会有较大的差异。

2、 回归分析的原理是基于最小的误差平方和，从宏观上来看统一模型所用的正是两个独立模型的平均式，因此 x1 与 x2 的系数在统一模型中会相互接近。

3、 回归分析的结果本身是一个随机变量，因此在自变量相同的情况下，问题不同，所得到的模型就不会完全相同。

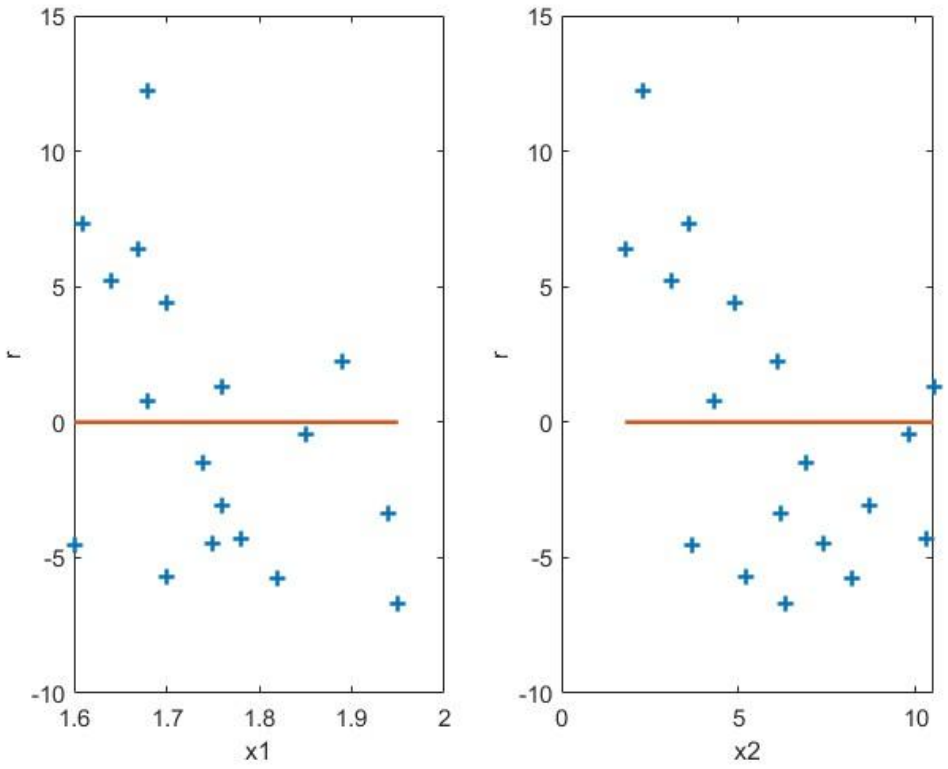
第三小问

程序 3

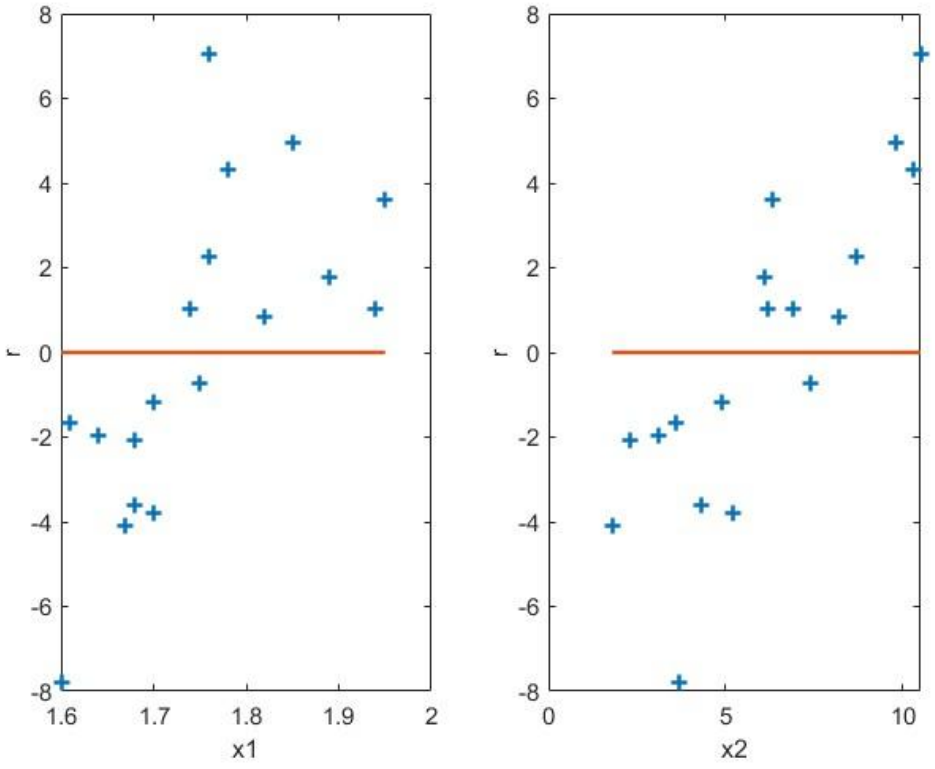
```
clc; clear ;
n = 18 ;
%数据的导入
y=[22.1,15.4,11.7,10.3,11.4,7.5,13,12.8,14.6,18.9,19.3,30.1,28.
2,25.6,37.5,36.1,39.8,44.3,7.2,5.4,7.6,2.5,2.4,1.7,4.3,3.7,3.9,
7,6.8,10.1,9.4,7.9,14.1,14.5,14.9,15.6]';
x1 =
[1.89,1.94,1.95,1.82,1.85,1.78,1.76,1.76,1.75,1.74,1.7,1.7,1.68
,1.6,1.61,1.64,1.67,1.68,1.89,1.94,1.95,1.82,1.85,1.78,1.76,1.7
6,1.75,1.74,1.7,1.7,1.68,1.6,1.61,1.64,1.67,1.68]';
x2 =
[6.1,6.2,6.3,8.2,9.8,10.3,10.5,8.7,7.4,6.9,5.2,4.9,4.3,3.7,3.6,
3.1,1.8,2.3,6.1,6.2,6.3,8.2,9.8,10.3,10.5,8.7,7.4,6.9,5.2,4.9,4
.3,3.7,3.6,3.1,1.8,2.3]';
x3 = [zeros(1,n) ones(1,n)]' ;
X = [ones(size(x1)) x1 x2 x3] ;
[b,bint,r,rint,stats] = regress(y,X) ;
figure(1) ;%普通汽车残差图像

title('普通汽车') ;
subplot(1,2,1) ; z = 0 * x1 ;
plot(x1(1:18), r(1:18),'+',x1,z,'LineWidth',1.5) ;
xlabel('x1') ; ylabel('r') ;
subplot(1,2,2) ;
plot(x2(1:18), r(1:18),'+',x2,z,'LineWidth',1.5) ;
xlabel('x2') ; ylabel('r') ;
figure(2) ;%豪华汽车残差图像

title('豪华汽车') ;
subplot(1,2,1) ; z = 0 * x1 ;
plot(x1(19:36), r(19:36),'+',x1,z,'LineWidth',1.5) ;
xlabel('x1') ; ylabel('r') ;
subplot(1,2,2) ;
plot(x2(19:36), r(19:36),'+',x2,z,'LineWidth',1.5) ;
xlabel('x2') ; ylabel('r') ;
```

结果 3

普通车残差图像：



豪华车残差图像：



分析 3

理想情况下的残差分析应该服从均值为零的同方差正态分布。残差点应该比较均匀的分布在 y=0 直线两侧，且数值接近 0 的点应该占据多数。而得到的独立模型结果却与理想假设差别较大。利用统一模型做出的结果同样不够理想，说明原假设 x1 与 x2 独立对 y 作用是不合适的，两个变量之间存在相互作用，且在 x1 与 x2 固定的情况下，x3 取值的不同也会得到不一样的残差结果，这同样说明汽车类型对油价，贷款利率等有相互作用。

第四小问

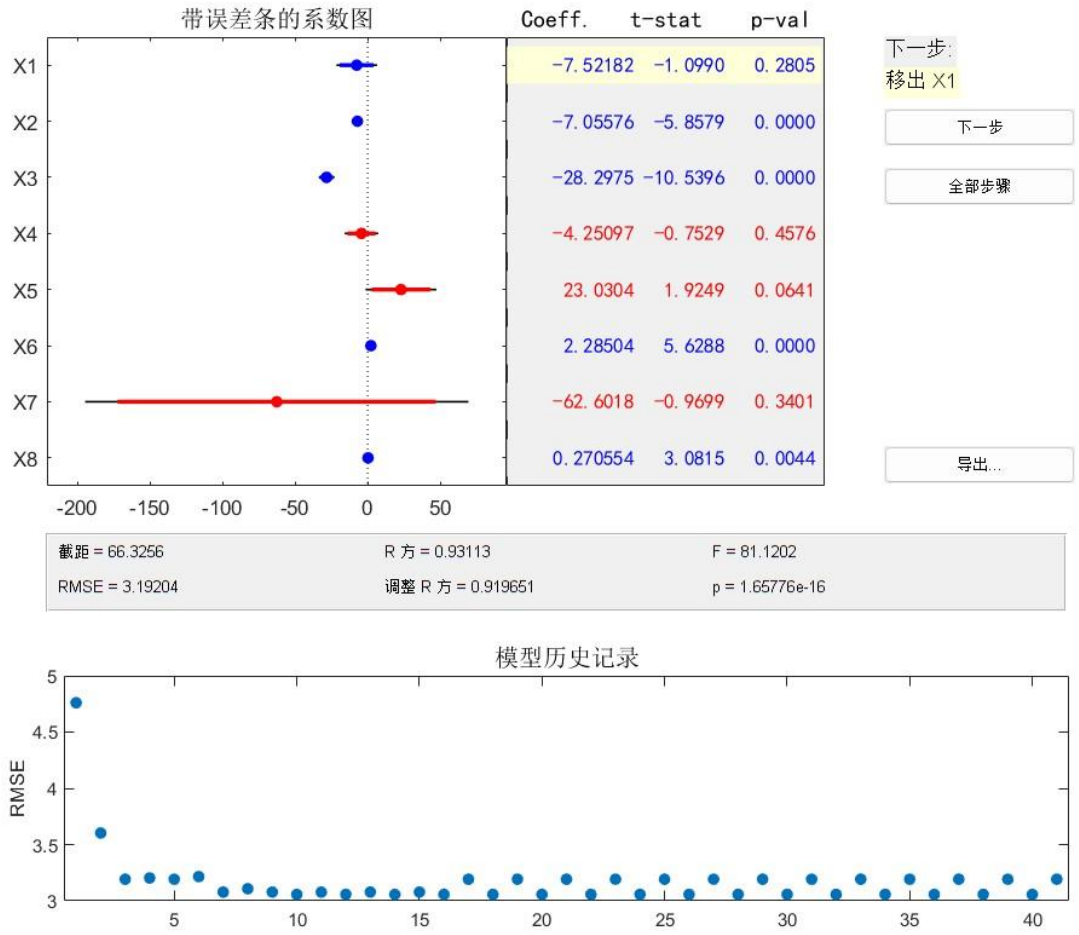
模型 4

$$y = -1.099x_1 - 5.8579x_2 - 10.5396x_3 + 5.6288x_2x_3 + 3.0815x_2^2$$

程序 4

```
clc; clear ;
n = 18 ;
y =
[22.1,15.4,11.7,10.3,11.4,7.5,13,12.8,14.6,18.9,19.3,30.1,
28.2,25.6,37.5,36.1,39.8,44.3,7.2,5.4,7.6,2.5,2.4,1.7,4.3,
3.7,3.9,7,6.8,10.1,9.4,7.9,14.1,14.5,14.9,15.6]';
x1 =
[1.89,1.94,1.95,1.82,1.85,1.78,1.76,1.76,1.75,1.74,1.7,1.7
,1.68,1.6,1.61,1.64,1.67,1.68,1.89,1.94,1.95,1.82,1.85,1.7
8,1.76,1.76,1.75,1.74,1.7,1.7,1.68,1.6,1.61,1.64,1.67,1.68
]';
x2 =
[6.1,6.2,6.3,8.2,9.8,10.3,10.5,8.7,7.4,6.9,5.2,4.9,4.3,3.7
,3.6,3.1,1.8,2.3,6.1,6.2,6.3,8.2,9.8,10.3,10.5,8.7,7.4,6.9
,5.2,4.9,4.3,3.7,3.6,3.1,1.8,2.3]';
x3 = [zeros(1,n) ones(1,n)]' ;
x4 = x1 .* x2 ;
x5 = x1 .* x3 ;
x6 = x2 .* x3 ;
x7 = x1 .^ 2 ;
x8 = x2 .^ 2 ;
X = [x1 x2 x3 x4 x5 x6 x7 x8] ;

stepwise(X, y, [1,2,3]) % [1,2,3]表示 x1 x2 x3 均保留在模型中
```

结果 4



分析 4

在保留 x1 x2 x3 的情况下通过调整变量的加入和清除使得 F 的值变大， p 变小以得到非劣解的情况下最终得到的回归方程是：$y = -1.099x_1 - 5.8579x_2 - 10.5396x_3 + 5.6288x_2x_3 + 3.0815x_2^2$

相比于没有加入交互项和二次项后 F 值为 81.1202，p 值为 0，相比于独立模型和原统一模型均有所优化，方差减小，模型的精度得以显著提高。因此对交互项和二次项的分析和调整是有必要的。