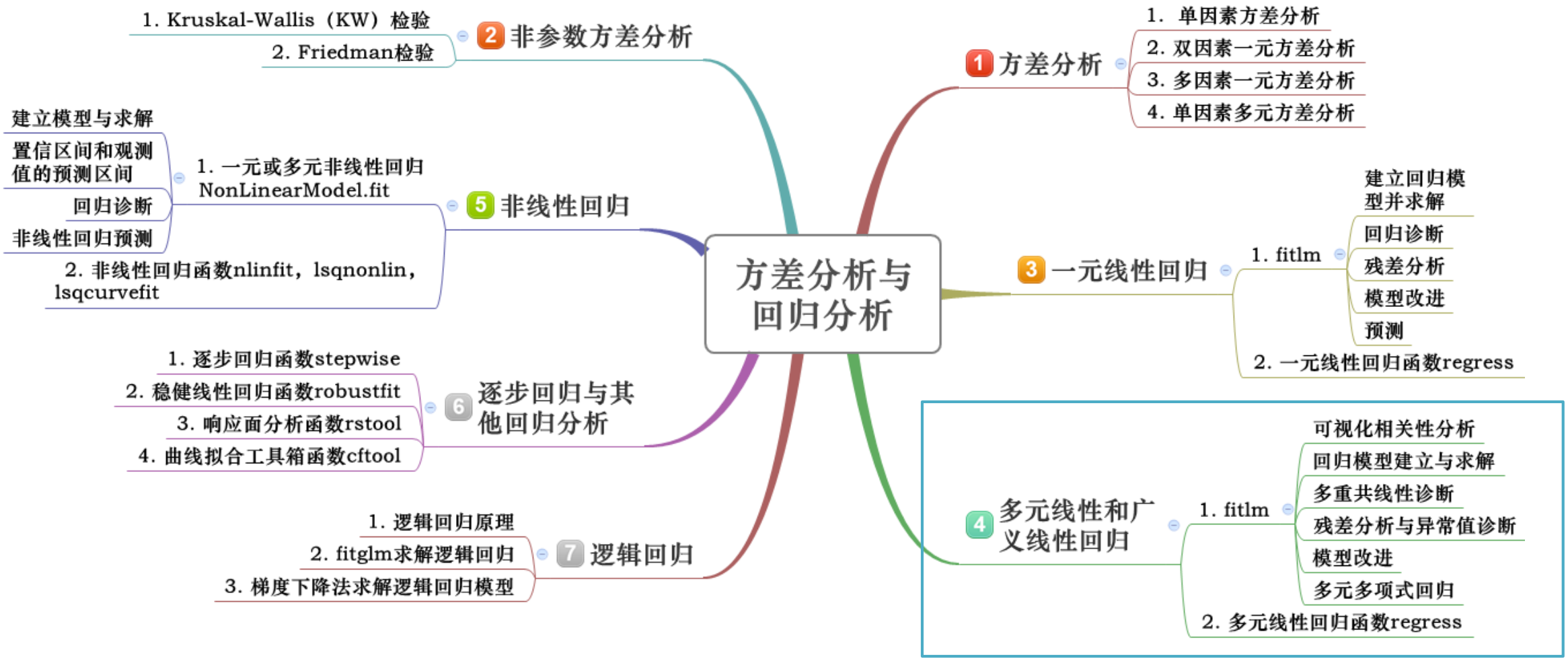
信陽師範学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

第11章 方差分析与回归分析

讲授人：牛言涛

日期：2020年4月16日

第11章 方差分析与回归分析知识点思维导图



一. fitlm多元线性回归函数

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

例1：在有氧锻炼中，人的耗氧能力y是衡量身体状况的重要指标，它可能与以下因素有关：年龄x_1（岁），体重x_2（kg），1500米跑所用的时间x_3（min），静止时心速x_4（次/min），跑步后心速x_5（次/min）．对24名40至57岁的志愿者进行了测试。试根据这些数据建立耗氧能力y与诸因素之间的回归模型。

	A	B	C	D	E	F	G
1	序号	y(ml/min.kg)	x1(岁)	x2(kg)	x3(min)	x4(次/min)	x5(次/min)
2	1	44.6	44	89.5	6.82	62	178
3	2	45.3	40	75.1	6.04	62	185
4	3	54.3	44	85.8	5.19	45	156
5	4	59.6	42	68.2	4.9	40	166
6	5	49.9	38	89	5.53	55	178
7	6	44.8	47	77.5	6.98	58	176
8	7	45.7	40	76	7.17	70	176
9	8	49.1	43	81.2	6.51	64	162
10	9	39.4	44	81.4	7.85	63	174
11	10	60.1	38	81.9	5.18	48	170
12	11	50.5	44	73	6.08	45	168

1. 可视化相关性分析

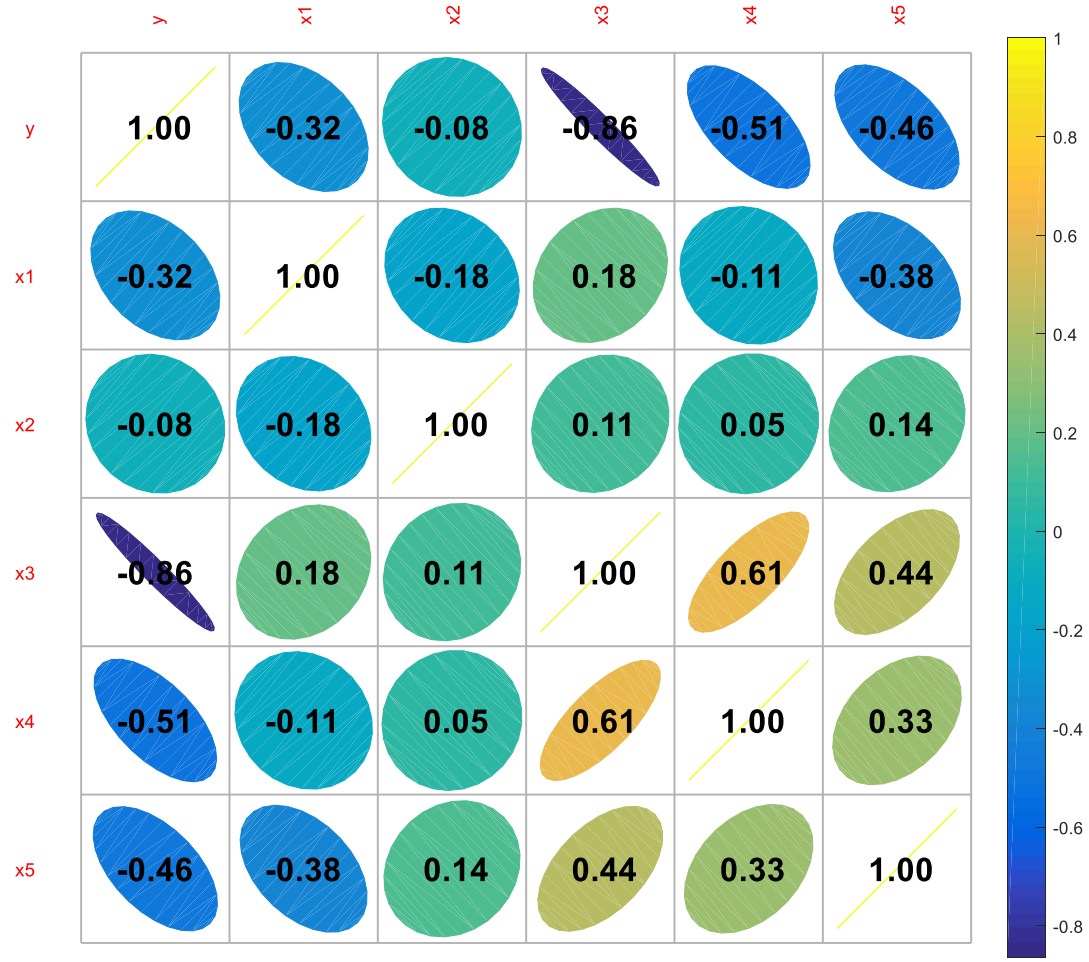
```
>> data = xlsread('body.xls');
>> X = data(:,3:7);
>> y = data(:,2);
>> [R,P] = corrcoef([y,X])  %[R,P] = corrcoef(data(:,2:7))
R =     %相关系数矩阵
    1.0000   -0.3201   -0.0777   -0.8645   -0.5130   -0.4573
   -0.3201    1.0000   -0.1809    0.1845   -0.1092   -0.3757
   -0.0777   -0.1809    1.0000    0.1121    0.0520    0.1410
   -0.8645    0.1845    0.1121    1.0000    0.6132    0.4383
   -0.5130   -0.1092    0.0520    0.6132    1.0000    0.3303
   -0.4573   -0.3757    0.1410    0.4383    0.3303    1.0000
```

```
P =  %线性相关性检验
    1.0000    0.1273    0.7181    0.0000    0.0104    0.0247
    0.1273    1.0000    0.3976    0.3882    0.6116    0.0704
    0.7181    0.3976    1.0000    0.6022    0.8095    0.5111
    0.0000    0.3882    0.6022    1.0000    0.0014    0.0322
    0.0104    0.6116    0.8095    0.0014    1.0000    0.1149
    0.0247    0.0704    0.5111    0.0322    0.1149    1.0000
```

从检验的值矩阵可以看出哪些变量间的线性相关性是显著的，若$\boldsymbol{p \leq 0.05}$，则认为变量间的线性相关性是显著的，反之则认为变量间的线性相关性是不显著的。从P矩阵看出，$\boldsymbol{y}$与$\boldsymbol{x_3, x_4, x_5}$线性相关性是显著的，$\boldsymbol{x_3}$与$\boldsymbol{x_4, x_5}$线性相关性是显著。

1. 可视化相关性分析

```
>> VarNames = {'y','x1','x2','x3','x4','x5'};
>> matrixplot(R,'FigShap','e','FigSize','Auto', 'ColorBar','on','XVar', VarNames,'YVar',VarNames);
```



- 相关系数矩阵图：椭圆色块直观地表示变量间的线性相关程度的大小。
  - 椭圆越扁，变量间相关系数的绝对值越接近于1；
  - 椭圆越圆，变量间相关系数的绝对值越接近于0.
  - 若椭圆的长轴方向是从左下到右上，则变量间为正相关，反之为负相关。

2. 多重共线性诊断

- 多重共线性诊断：基于方差膨胀因子。
  - 考虑自变量 x_i (因变量)关于其余自变量的多元线性回归，计算模型的判定系数，记为 R_i^2 ，定义第i个自变量的方差膨胀因子：

$$VIF_i = \frac{1}{1 - R_i^2}$$

  - 当自变量有依赖于其他自变量的线性关系时， R_i^2 越接近1， VIF_i 越接近于无穷大；反之，接近于0， VIF_i接近于1. VIF_i越大，说明线性依赖关系越严重，即存在共线性；这种多重共线性可能会过度地影响最小二乘估计。
  - 诊断规则： $VIF_i < 5$不存在共线性（或共线性较弱）； $5 \le VIF_i \le 10$中等程度共线性；$VIF_i > 10$共线性严重，必须消除共线性。方法：去除变量，变量变换，岭回归，主成分回归。

2. 多重共线性诊断

方差膨胀因子是指解释变量之间存在多重共线性时的方差与不存在多重共线性时的方差之比。

可根据自变量的相关系数矩阵R_X计算各自变量的方差膨胀因子，自变量x_i的方差膨胀因子VIF_i等于R_X的逆矩阵的对角线上的第i个元素。

```
>> Rx = corrcoef(X);
>> VIF = diag(inv(Rx))
VIF =
    1.5974
    1.0657
    2.4044
    1.7686
    1.6985
```

各自变量的方差膨胀因子均小于5，说明模型不存在多重共线性。

如果存在共线性，可用岭回归。岭回归(ridge regression, Tikhonov regularization)是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。岭回归与多项式回归唯一的不同在于代价函数上的差别。岭回归的代价函数如下（学习率α，正则化参数λ）：

$$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(y^{(i)}-\left(wx^{(i)}+b\right)\right)^2+\frac{\lambda}{2}\|w\|_2^2=\frac{1}{2}MSE(\theta)+\frac{\lambda}{2}\sum_{i=1}^{n}\theta_i^2$$

梯度下降法，参数更新：$\theta=\theta-\left(\frac{\alpha}{m}X^T\left(X\theta-y\right)+\lambda w\right)$

3. 建立回归模型与求解

>> mmdl1 = fitlm(X,y)

```
mmdl1 =
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5

Estimated Coefficients:
                   Estimate        SE        tStat        pValue
                   ________    ________    ________    __________

    (Intercept)      121.17      17.406       6.961    1.6743e-06
    x1              -0.34712     0.14353     -2.4185      0.026406
    x2              -0.016719    0.087353    -0.19139      0.85036
    x3               -4.2903      1.0268     -4.1784    0.00056473
    x4              -0.039917    0.094237    -0.42357      0.67689
    x5               -0.15866    0.078847     -2.0122      0.059407

Number of observations: 24, Error degrees of freedom: 18
Root Mean Squared Error: 2.8
R-squared: 0.816,  Adjusted R-Squared 0.765
F-statistic vs. constant model: 16, p-value = 4.46e-06
```

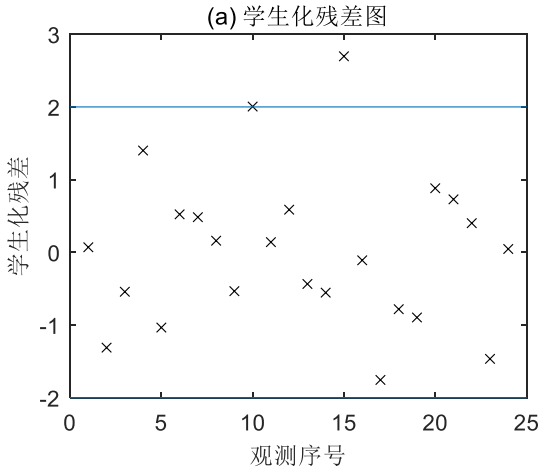
x2、x4系数不显著，其他系数较为显著。R-square即可决系数，反映模型对样本数据的拟合程度。值越大，拟合效果越好。它能否反映总体数据的变化情况，要对模型进行检验。

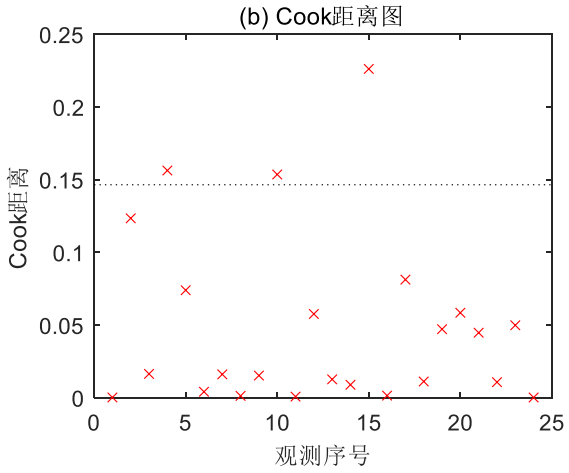
%一元线性回归的离群点诊断代码封装成一个函数，并添加输出离群点对应的行索引。

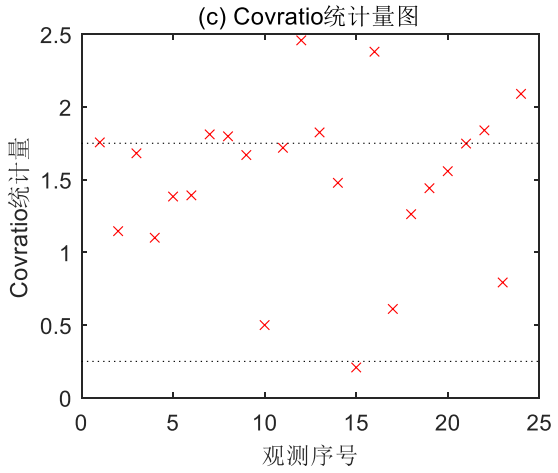
```
Res = model.Residuals;

Res_Stu = Res.Studentized;

Res_Stan = Res.Standardized;

idout = find(abs(Res_Stu) > 2)

md = model.Diagnostics;

idinf = find(md.CooksDistance >
3*mean(md.CooksDistance))

idleve = find(md.Leverage >
2*(model.NumCoefficients+1)/size(model.Residuals,1))
```

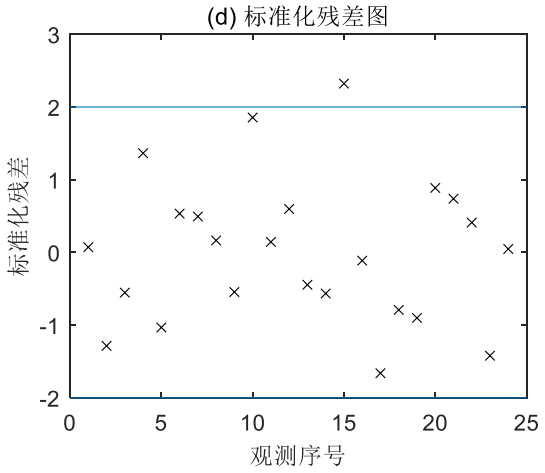
4. 残差分析与异常值诊断

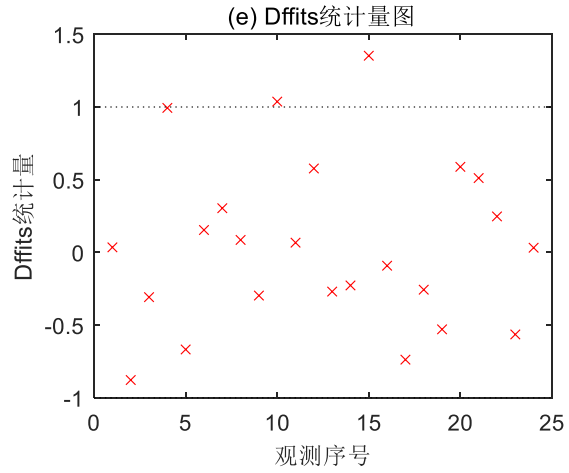
一元线性回归的离群点诊断代码封装成一个函数，调用：plot_outliers(mmdl1)

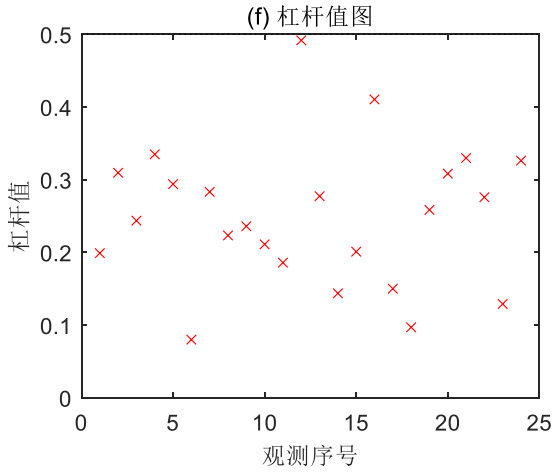












从图中看出，异常点有两个，强影响点有三个，无高杠杆点。程序允许输出：

idout =

10 15

idinf =

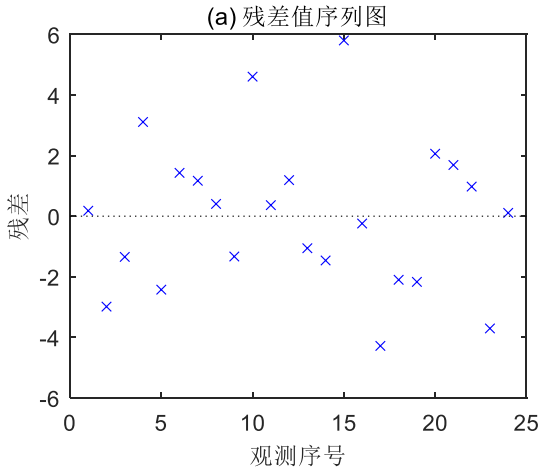
4 10 15

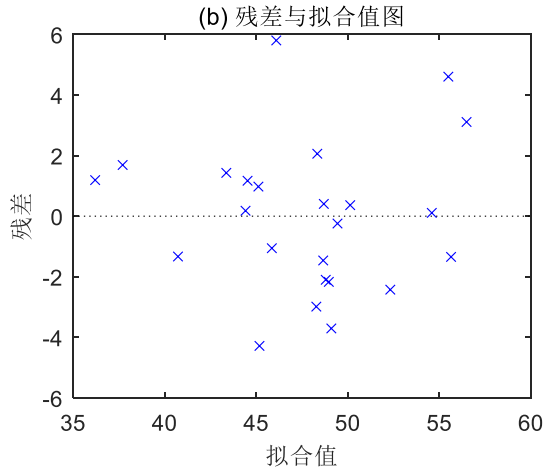
idleve =

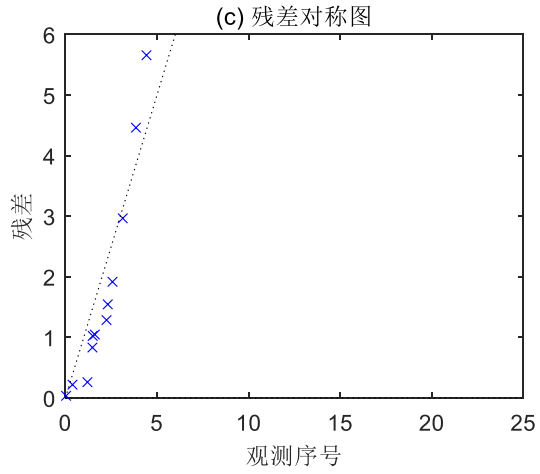
空的 0×1 double 列矢量

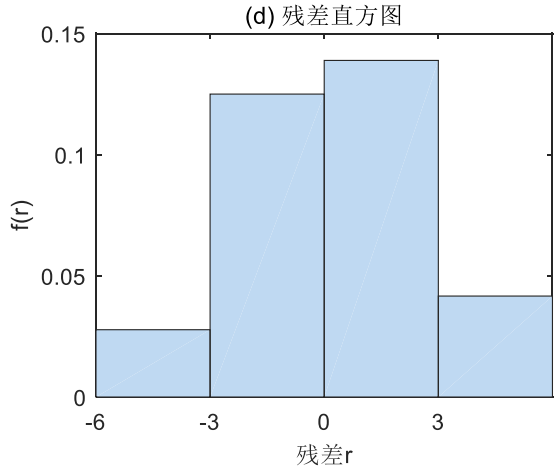
4. 残差分析与异常值诊断

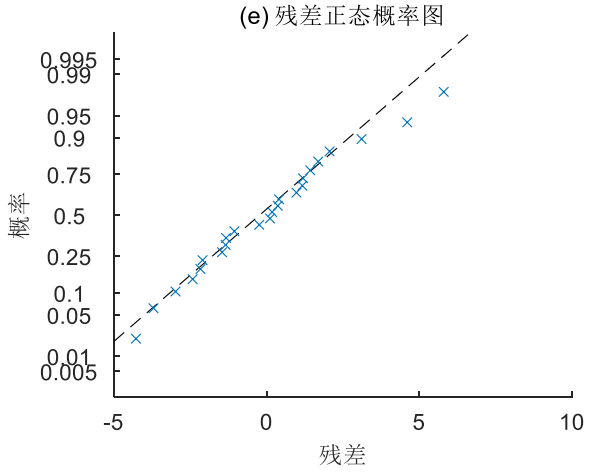
一元线性回归的方差分析代码封装成一个函数，调用：plot_Residuals(mmdl1)

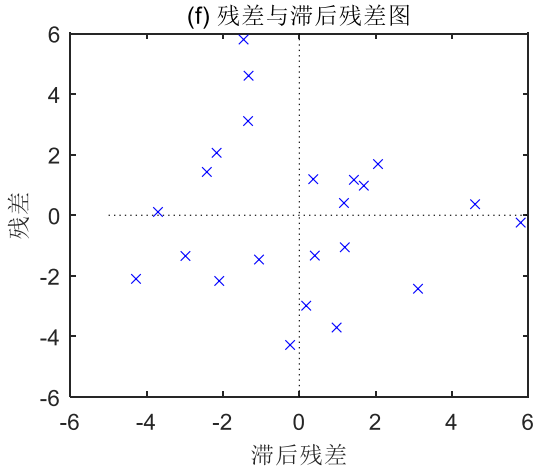












- 从图a、b两图中看出残差在值为0的虚线上下分布比较均匀且满足方差齐性；
- 从c图中看出残差基本满足对称性；
- 从d、e两图中看出残差近似服从正态分布；
- 从f图看出残差无自相关性，独立。

5. 模型改进

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

% 删除系数不显著的变量x2和x4，重新建立模型：

>> Model = 'poly10101';

>> mmdl2 = fitlm (X,y,Model,'Exclude',idout)

```
mmdl2 =
Linear regression model:
    y ~ 1 + x1 + x3 + x5

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    __________
    (Intercept)      119.5       11.81      10.118    7.4559e-09
    x1             -0.36229     0.11272    -3.2141     0.0048108
    x3              -4.0411     0.62858    -6.4289    4.7386e-06
    x5             -0.17739     0.05977    -2.9678     0.0082426

Number of observations: 22, Error degrees of freedom: 18
Root Mean Squared Error: 2.11
R-squared: 0.862,  Adjusted R-Squared 0.84
F-statistic vs. constant model: 37.6, p-value = 5.81e-08
```

mmdl2 = stepwiselm (X,y,'linear','Exclude',idout)%逐步回归

```
1. Removing x4, FStat = 0.16306, pValue = 0.6917
2. Removing x2, FStat = 1.1023, pValue = 0.30845
mmdl2 =
Linear regression model:
    y ~ 1 + x1 + x3 + x5

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    __________
    (Intercept)      119.5       11.81      10.118    7.4559e-09
    x1             -0.36229     0.11272    -3.2141     0.0048108
    x3              -4.0411     0.62858    -6.4289    4.7386e-06
    x5             -0.17739     0.05977    -2.9678     0.0082426

Number of observations: 22, Error degrees of freedom: 18
Root Mean Squared Error: 2.11
R-squared: 0.862,  Adjusted R-Squared 0.84
F-statistic vs. constant model: 37.6, p-value = 5.81e-08
```

5. 模型改进

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

模型的类型：fitlm(X, y, Model)

Character Vector	Model Type
'constant'	模型只包含一个常数（截距）项。
'linear'	模型包含每个预测变量的截距和线性项。
'interactions'	模型包含每个预测变量的截距、线性项以及不同预测变量对的所有乘积（无平方项）。
'purequadratic'	模型包含每个预测变量的截距项、线性项和平方项。
'quadratic'	模型包含截每个预测变量的截距项、线性项和平方项，以及不同预测变量对组的所有乘积。
'polyijk'	模型是一个多项式，其中具有第一个预测变量的 1 到 i 次的所有项，第二个预测变量的 1 到 j 次的所有项，依此类推。请使用数字 0 到 9 指定每个预测变量的最大次数。模型包含交互效应项，但是，每个交互效应项的次数不超过指定次数的最大值。例如，'poly13' 具有截距和 x1、x2、x_2^2、x_2^3、x1*x2 和 x1*x_2^2项，其中 x1 和 x2 分别是第一个和第二个预测变量。

```
model = [0 0 0 0 0  %常数项
         1 0 0 0 0  %x1项
         0 1 0 0 0  %x2项
         0 0 0 0 1  %x5项
         2 0 0 0 0  %x1^2项
         1 1 0 0 0  %x1*x2项
         0 1 1 0 0  %x2*x3项
         1 0 0 1 0  %x1*x4项
         0 0 0 2 0  %x4^2项
         1 0 0 0 1  %x1*x5项
         0 1 0 0 1  %x2*x5项
         0 0 1 0 1  %x3*x5项
         0 0 0 0 2];  %x5^2项
```

6. 多元多项式回归

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

假设理论回归方程为

$$\hat{y} = b_0 + \sum_{i=1}^{5} b_i x_i + \sum_{i=1}^{4} \sum_{j=i+1}^{5} b_{ij} x_i x_j + \sum_{i=1}^{5} b_{ii} x_i^2$$

>> Model = 'poly22222';

% Model = 'quadratic';

>> mmdl3 = LinearModel.fit(X,y,Model)

Number of observations: 24, Error degrees of freedom: 3

Root Mean Squared Error: 0.557

R-squared: 0.999, Adjusted R-Squared 0.991

F-statistic vs. constant model: 123, p-value = 0.00104

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	1804.1	176.67	10.211	0.0020018
x1	-26.768	3.3174	-8.069	0.0039765
x2	-16.422	1.4725	-11.153	0.0015449
x3	-7.2417	17.328	-0.41792	0.70412
x4	1.7071	1.5284	1.1169	0.34543
x5	-5.5878	1.2082	-4.6248	0.019034
x1^2	0.034031	0.02233	1.524	0.22489
x1:x2	0.18853	0.014842	12.702	0.0010526
x2^2	-0.0024412	0.0030872	-0.79075	0.48684
x1:x3	0.23808	0.21631	1.1006	0.35145
x2:x3	-0.56157	0.087918	-6.3874	0.0077704
x3^2	0.68822	0.63574	1.0826	0.35825
x1:x4	0.016786	0.015763	1.0649	0.36502
x2:x4	0.0030961	0.0058481	0.52942	0.63319
x3:x4	-0.065623	0.071279	-0.92065	0.42513
x4^2	-0.016381	0.0047701	-3.4342	0.041411
x1:x5	0.03502	0.011535	3.0359	0.056047
x2:x5	0.067888	0.0063552	10.682	0.0017537
x3:x5	0.17506	0.063871	2.7408	0.071288
x4:x5	-0.0016748	0.0056432	-0.29679	0.78599
x5^2	-0.007748	0.0027112	-2.8577	0.064697

6. 多元多项式回归

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

>> mmdl3sw = stepwiselm(X,y,Model)

1. Removing x4:x5, FStat = 0.088084, pValue = 0.78599

2. Removing x2:x4, FStat = 0.49518, pValue = 0.52043

3. Removing x2^2, FStat = 0.55596, pValue = 0.48944

4. Removing x1:x3, FStat = 2.0233, pValue = 0.20475

5. Removing x3^2, FStat = 1.7938, pValue = 0.22232

6. Removing x3:x4, FStat = 1.7098, pValue = 0.22734

mmdl3sw =

Linear regression model:

y ~ 1 + x1^2 + x1*x2 + x2*x3 + x1*x4 + x4^2 + x1*x5 + x2*x5 + x3*x5 + x5^2

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	1916.6	106.48	17.999	2.2957e-08
x1	-29.485	1.6156	-18.251	2.0321e-08
x2	-15.841	0.92505	-17.124	3.553e-08
x3	3.3267	4.4986	0.7395	0.47845
x4	0.757	0.43986	1.721	0.11936
x5	-6.547	0.69061	-9.4801	5.5705e-06
x1^2	0.060353	0.0051667	11.681	9.6821e-07
x1:x2	0.17622	0.010126	17.403	3.0846e-08
x2:x3	-0.46789	0.050314	-9.2994	6.5277e-06
x1:x4	0.034115	0.0041517	8.2173	1.7857e-05
x4^2	-0.019258	0.0032306	-5.9612	0.00021239
x1:x5	0.045394	0.0050247	9.0342	8.2768e-06
x2:x5	0.063051	0.0043992	14.332	1.6742e-07
x3:x5	0.165	0.025546	6.4588	0.00011693
x5^2	-0.0052175	0.0016766	-3.1119	0.01248

Number of observations: 24, Error degrees of freedom: 9
Root Mean Squared Error: 0.521
R-squared: 0.997, Adjusted R-Squared 0.992
F-statistic vs. constant model: 201, p-value = 1.82e-09

6. 多元多项式回归

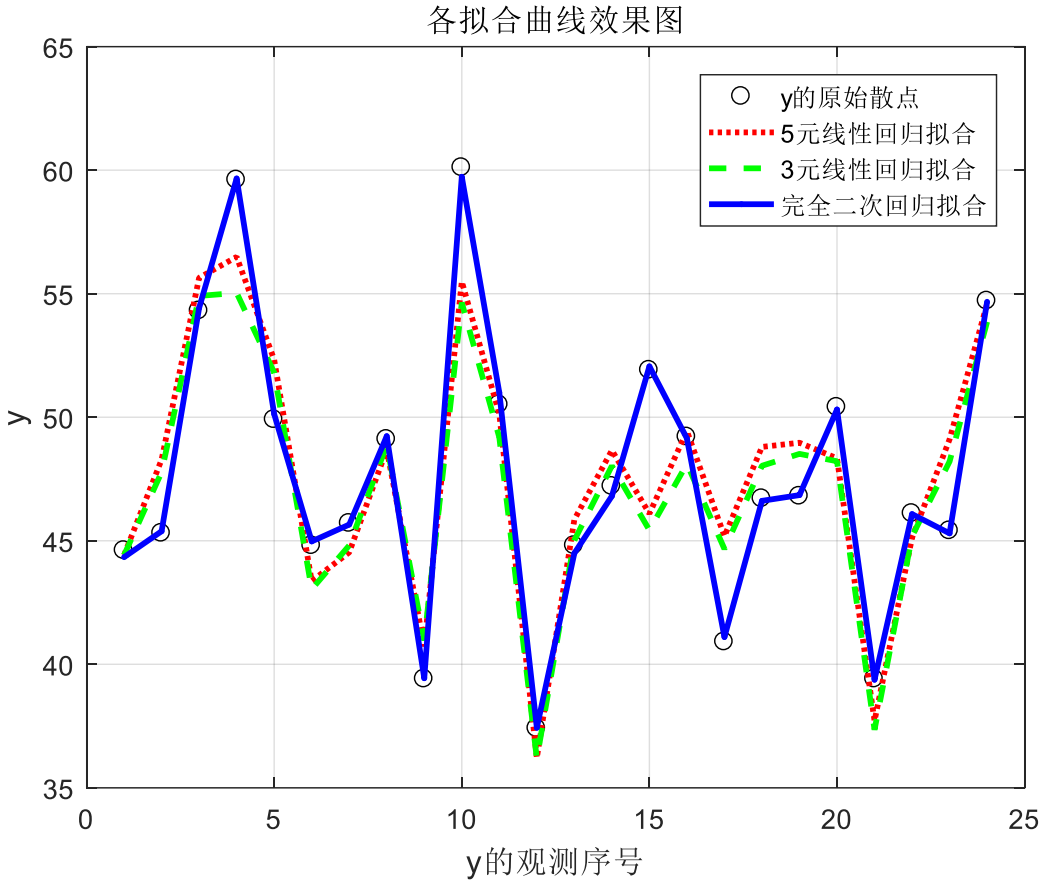
```
>> Model2 = 'quadratic';
>> plot_plotResiduals(mmdl3sw)  %满足残差检验
>> plot_outliers(mmdl3sw)  %识别离群点
idout =
    14   20
idinf =
     7  20
idleve =
  空的 0×1 double 列矢量
>> mdlsw2 = stepwiselm(X,Y,Model2,'Exclude',[14,20])
Linear regression model:
    y ~ 1 + x1*x2 + x1*x3 + x1*x5 + x2*x3 + x2*x5 + x3*x4
+ x3*x5 + x4*x5 + x1^2 + x4^2 + x5^2
```

%删除异常点后的逐步回归，最终系数都比较显著。

```
Estimated Coefficients:
                   Estimate        SE        tStat       pValue
                   ________    _________    _______    __________
    (Intercept)      1705.5      84.994      20.066    5.6817e-06
    x1              -24.921      1.3144      -18.96    7.5213e-06
    x2              -16.383     0.76796      -21.333    4.1964e-06
    x3              -27.454      5.8292      -4.7097      0.005291
    x4                3.933     0.43994        8.94     0.00029181
    x5               -4.9087     0.59623      -8.2329     0.00043076
    x1:x2            0.18629   0.0082196      22.664    3.1091e-06
    x1:x3            0.38512     0.05124      7.5159     0.00065972
    x1:x5            0.02739   0.0046716      5.8631      0.0020465
    x2:x3           -0.51393    0.042023      -12.23    6.4654e-05
    x2:x5            0.06517   0.0037547      17.357    1.1631e-05
    x3:x4           0.088625    0.032028      2.7671      0.039501
    x3:x5            0.23466    0.033038      7.1026     0.00085729
    x4:x5          -0.0096606   0.0021305      -4.5345     0.0061998
    x1^2             0.029544    0.007781      3.7969      0.012669
    x4^2             -0.02416   0.0036799      -6.5655     0.0012293
    x5^2           -0.0078789   0.0015835      -4.9755     0.0041919

Number of observations: 22, Error degrees of freedom: 5
Root Mean Squared Error: 0.379
R-squared: 0.999,  Adjusted R-Squared 0.996
F-statistic vs. constant model: 330, p-value = 1.85e-06
```

7. 各种模型回归对比分析

```
plot(y,'ko');
hold on
plot(mmdl1.predict(X),'r:','LineWidth',2);
plot(mmdl2.predict(X), 'g--','LineWidth',2);
plot(mmdl3.predict(X), 'b.-','LineWidth',2);
legend('y的原始散点','5元线性回归拟合','3元线性回归拟合','完全二次回归拟合');
xlabel('y的观测序号');
ylabel('y');
title('各拟合曲线效果图')
```



二. 多元线性回归函数regress

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

多元线性回归步骤：$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$

1、直观分析，做散点图，初步设定多元线性回归模型的参数个数；

2、[b,bint,r,rint,stats]=regress(Y,X,alpha)，计算参数估计；

3、调用命令rcoplot(r,rint)，分析数据的异常点情况；

4、做显著性检验，若检验通过，则用模型做预测；

5、对模型进一步研究：如残差的正态性检验，残差的异方差检验，残差进行自相关性的检验等。

二. 多元线性回归函数regress

➢ **求回归系数的点估计和区间估计、并检验回归模型：**

[b, bint,r,rint,stats]=regress(Y,X,alpha)

- b：回归系数的区间估计
- r：残差
- rint：置信区间
- stats：用于检验回归模型的统计量，有三个数值：相关系数r^2、F值、与F对应的概率p
- alpha：显著性水平（缺省时为0.05）

相关系数r^2越接近1，说明回归方程越显著；

$F > F_{1-a}(k, n-k-1)$时拒绝H_0，F越大，说明回归方程越显著；

与F对应的概率$p < \alpha$时拒绝H_0，回归模型成立。

画出残差及其置信区间：

rcoplot（r，rint）

多元线性回归案例分析

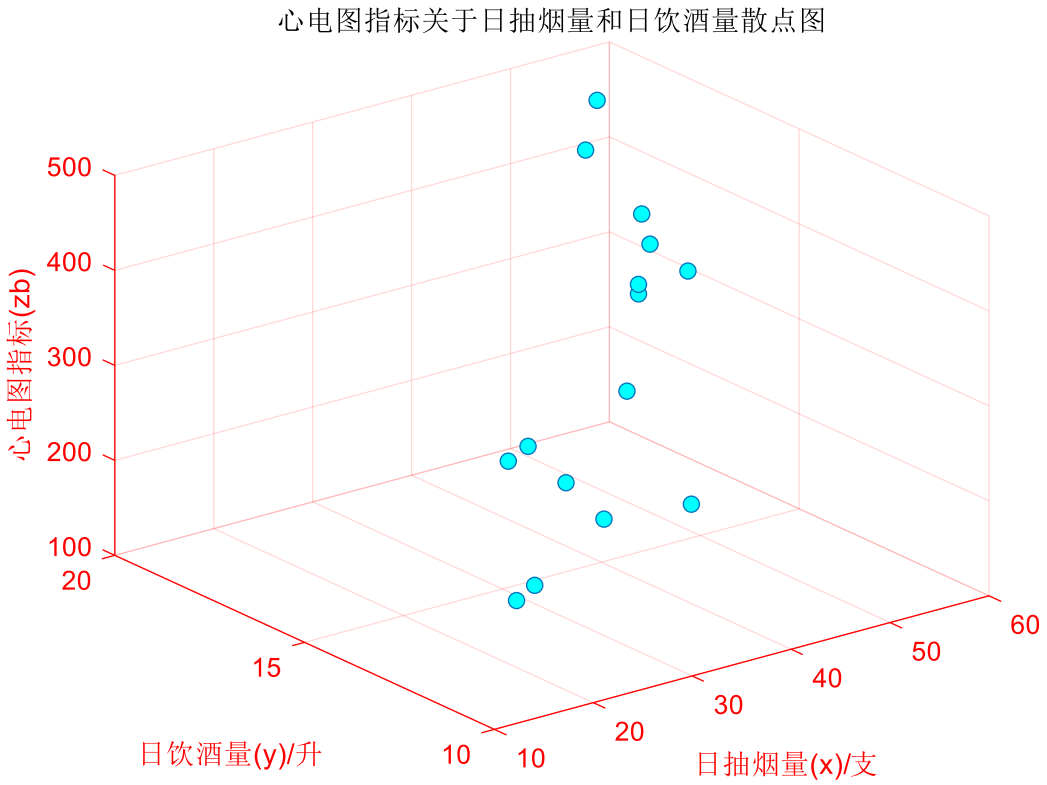
信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

例2：考察15名不同程度的烟民的每日抽烟量、饮酒量（啤酒）与其心电图指标(zb)的对应数据，试建立心电图指标关于日抽烟量和日饮酒量的适合的回归模型。

组别(g)	日抽烟量(x)/支	日饮酒量(y)/升	心电图指标(zb)	组别(g)	日抽烟量(x)/支	日饮酒量(y)/升	心电图指标(zb)
1	30	10	280	2	25	13	300
1	25	11	260	2	23	13	290
1	35	13	330	3	40	14	410
1	40	14	400	3	45	15	420
1	45	14	410	3	48	16	425
2	20	12	170	3	50	18	450
2	18	11	210	3	55	19	470
2	25	12	280				

1. 绘制散点图

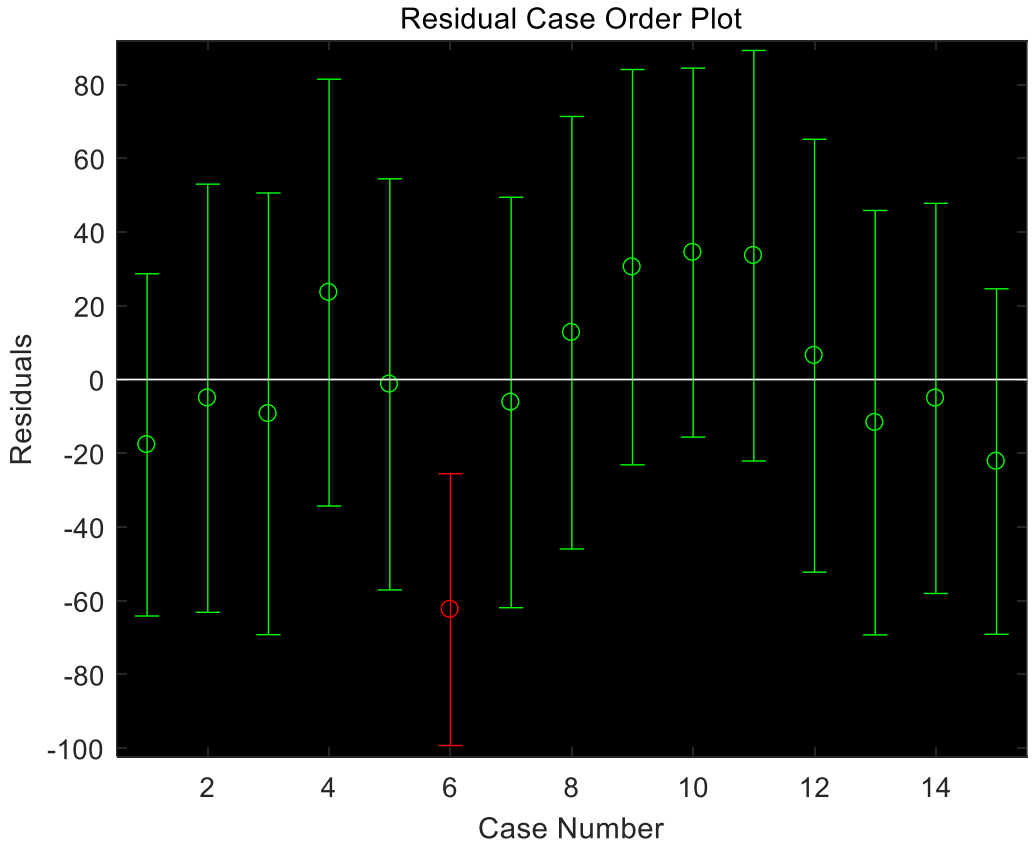
```
>> data=[30  10  280;25  11  260; 35  13  330; 40  14  400;
        45  14  410; 20  12  170; 18  11  210; 25  12  280;
        25 13  300; 23  13  290; 40  14  410; 45  15  420;
        48  16  425; 50  18  450; 55  19  470];
>> plot3(data(:,1), data(:,2), data(:,3),'o','MarkerFaceColor','c')
>> grid on
>> set(gca,'color','none')
>> xlabel('日抽烟量(x)/支');
>> ylabel('日饮酒量(y)/升');
>> zlabel('心电图指标(zb)');
>> title('心电图指标关于日抽烟量和日饮酒量散点图')
>> set(gca,'Xcolor',[1 0 0],'Ycolor',[1 0 0],'Zcolor',[1 0 0])
```



2. 多元线性回归分析

```
>> x= data(:,1);
>> y= data(:,2);
>> z= data(:,3);
>> n=size(x,1);
>> xy=[ones(n,1), x, y];
>> [b,bint,r,rint,stats]=regress(z,xy)
b =
   66.0944
    6.9774
    2.2314
…… ……
stats =
    0.9246   73.5741    0.0000  751.6477
>> rcoplot(r,rint)
```

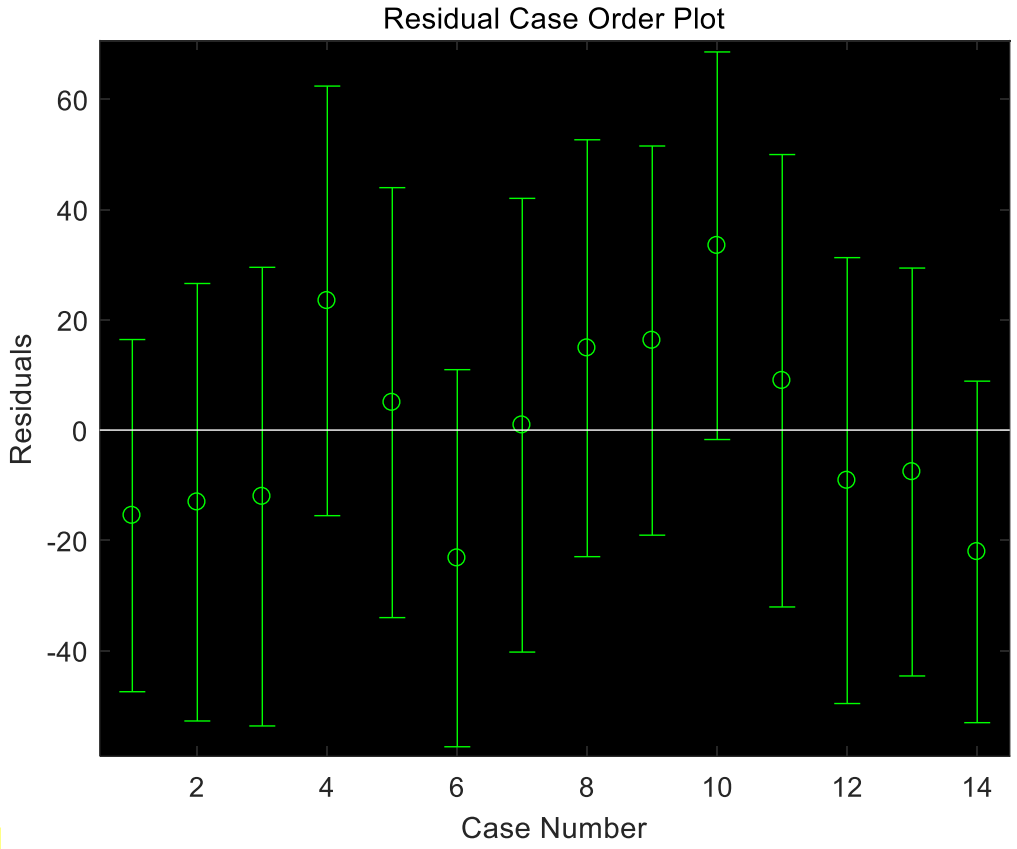
$$z = 66.0944 + 6.9774x + 2.2314y$$



3. 模型改进

删除第6组异常点，重新进行回归分析

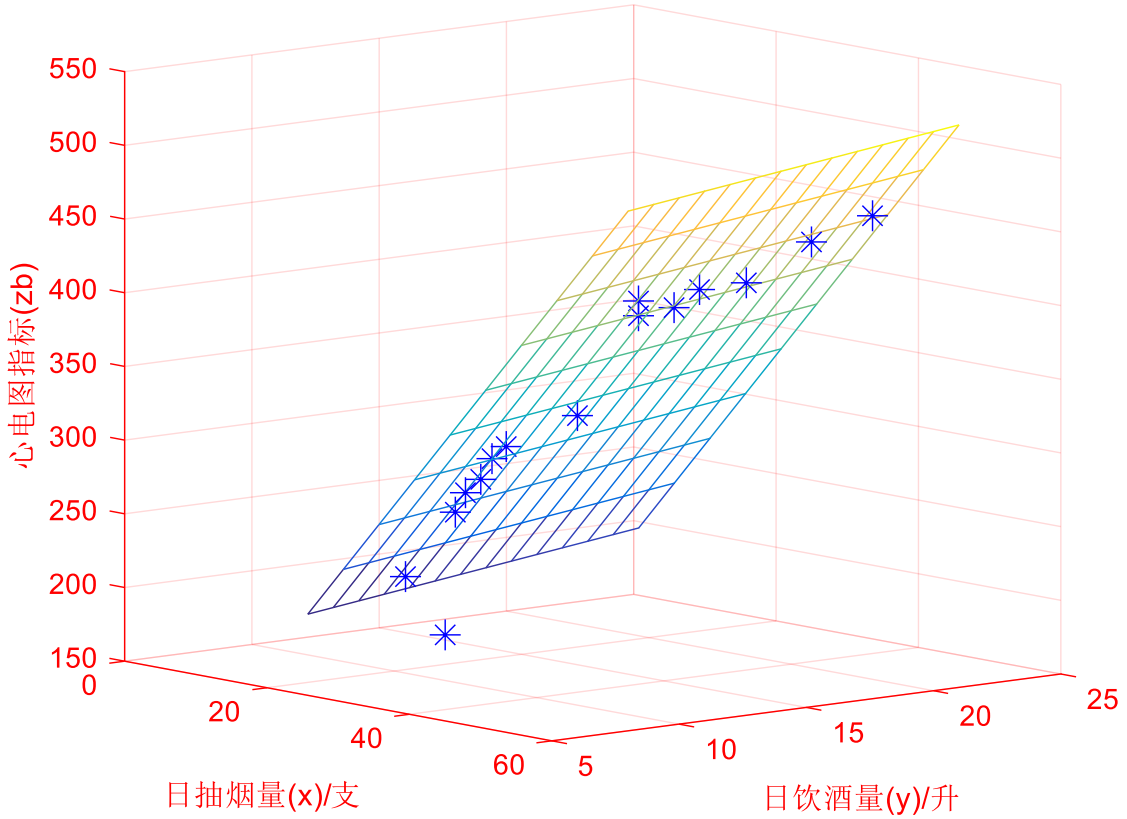
```
>> xy(6,:) = [];
>> z(6) = [];
>> [b,bint,r,rint,stats] = regress(z,xy)
b =
   64.4454
    5.6921
    6.0308
stats =
    0.9550  116.6896    0.0000  362.2955
>> rcoplot(r,rint)
```



stats可决系数提高到0.9550，F统计量观测值提高到116.69；p远小于alpha，残差平方和缩小到362.3

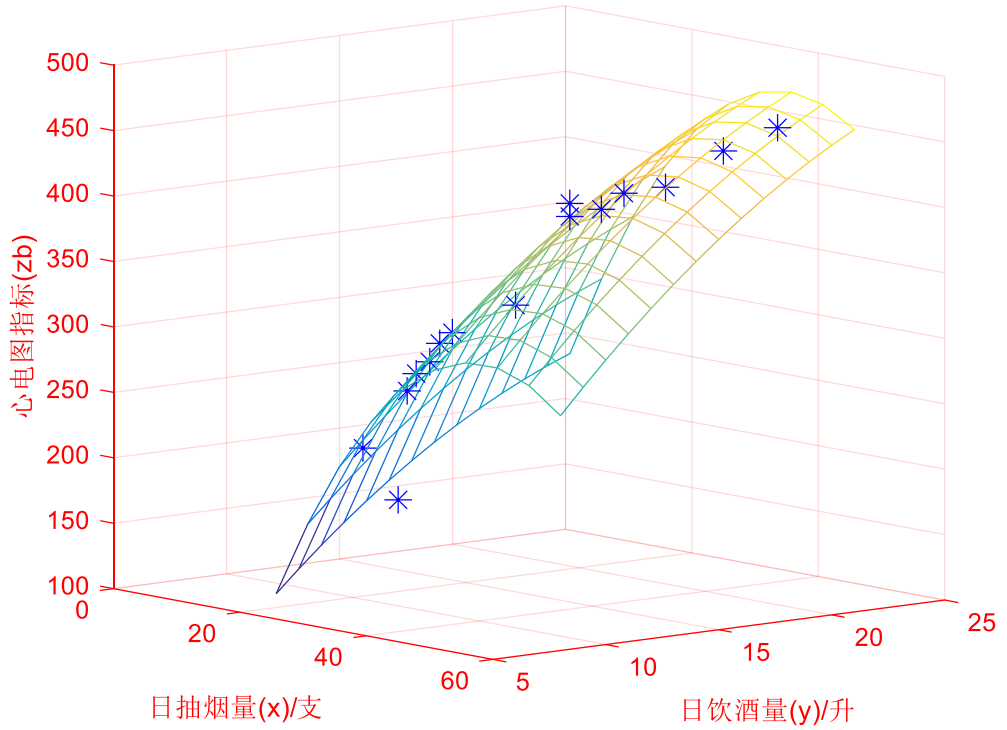
4. 绘制回归平面

```
[xdat,ydat]=meshgrid(15:5:60,8:21);
zdat1=[ones(length(xdat(:)),1) xdat(:) ydat(:)]*b;
zdat1=reshape(zdat1,size(xdat));
mesh(xdat,ydat,zdat1)
alpha(0)    %透明
hold on
plot3(x, y, z,'b*','markersize',10)
xlabel('日抽烟量(x)/支');
ylabel('日饮酒量(y)/升');
zlabel('心电图指标(zb)');
set(gca,'Xcolor',[1 0 0],'Ycolor',[1 0 0],'Zcolor',[1 0 0])
set(gca,'color','none')
view(50,10)
```



5. 二次型回归模型回归曲面

```
xy=[ones(n,1), x, y, x.^2, x.*y, y.^2];
[b,bint,r,rint,stats]=regress(z,xy)
xtemp=xdat(:);ytemp=ydat(:);
zdat2=[ones(length(xtemp),1) xtemp ytemp xtemp.^2 xtemp.*ytemp ytemp.^2]*b;
zdat2=reshape(zdat2,size(xdat));
mesh(xdat,ydat,zdat2)
alpha(0);   hold on
plot3(x, y, z,'b*','markersize',10)
xlabel('日抽烟量(x)/支');  ylabel('日饮酒量(y)/升');
zlabel('心电图指标(zb)');
set(gca,'Xcolor',[1 0 0],'Ycolor',[1 0 0],'Zcolor',[1 0 0])
set(gca,'color','none')
view(50,10)
```



案例分析：销售额预测

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

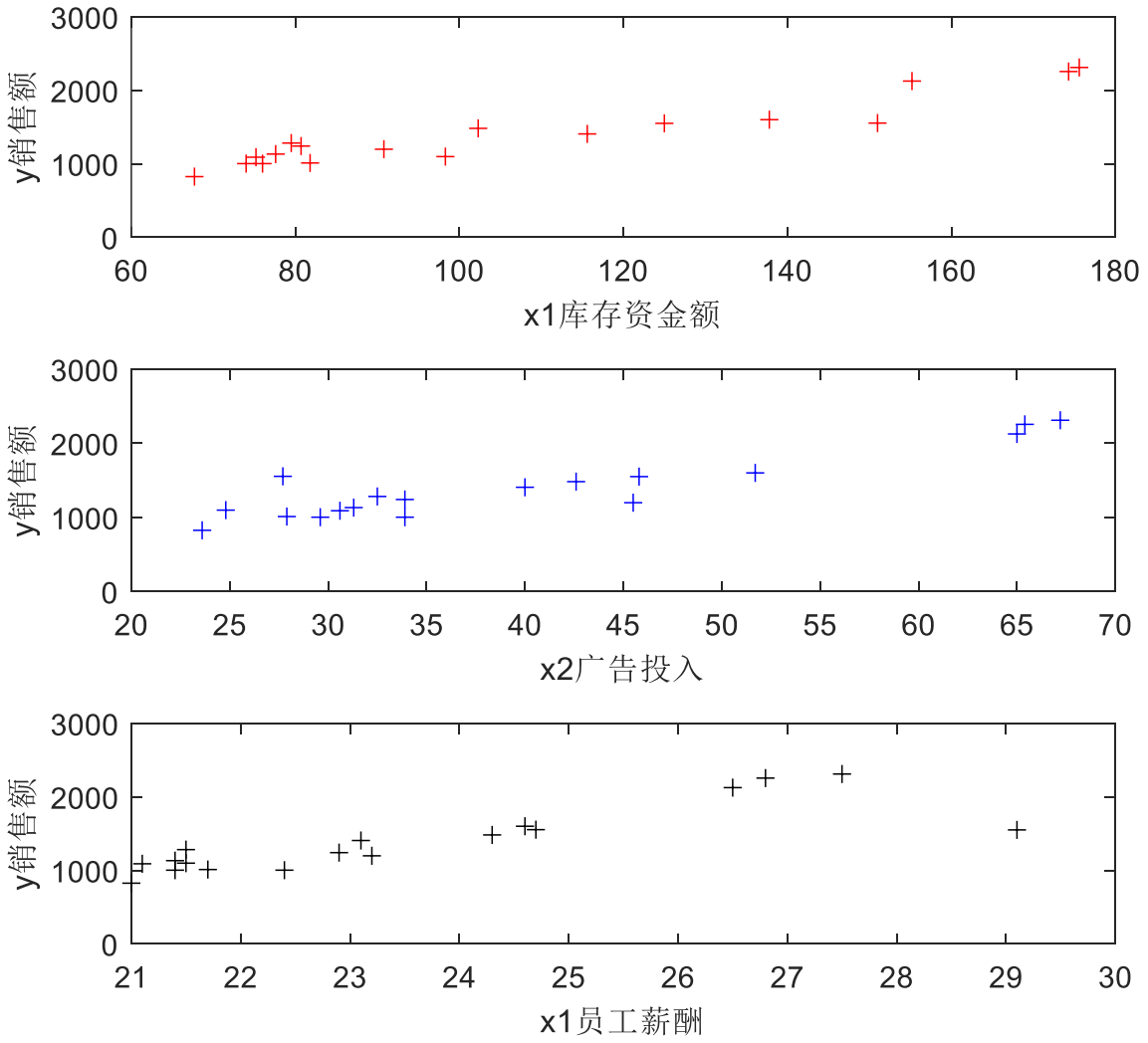
例3：某销售公司将库存占用资金情况、广告投入的费用、员工薪酬以及销售额等方面的数据作了汇总，该公司试图根据这些数据找到销售额与其他变量之间的关系，以便进行销售额预测并为工作决策提供参考依据。

问题：建立销售额的回归模型；如果未来某月库存资金额为150万元，广告投入预算为45万元，员工薪酬总额为27万元，试根据建立的回归模型预测该月的销售额。

月份	库存资金额(X1)	广告投入(X2)	员工薪酬总额(X3)	销售额(y)	月份	库存资金额(X1)	广告投入(X2)	员工薪酬总额(X3)	销售额(y)
1	75.2	30.6	21.1	1090.4	10	151	27.7	24.7	1554.6
2	77.6	31.3	21.4	1133	11	90.8	45.5	23.2	1199
3	80.7	33.9	22.9	1242.1	12	102.3	42.6	24.3	1483.1
4	76	29.6	21.4	1003.2	13	115.6	40	23.1	1407.1
5	79.5	32.5	21.5	1283.2	14	125	45.8	29.1	1551.3
6	81.8	27.9	21.7	1012.2	15	137.8	51.7	24.6	1601.2
7	98.3	24.8	21.5	1098.8	16	175.6	67.2	27.5	2311.7
8	67.7	23.6	21	826.3	17	155.2	65	26.5	2126.7
9	74	33.9	22.4	1003.3	18	174.3	65.4	26.8	2256.5

1. 绘制散点图

```
data = xlsread('saledata.xlsx');
data(:,1) = [];
[m,n] = size(data);
 %分别做散点图
subplot(3,1,1),plot(data(:,1), data(:,4),'r+');
xlabel('x1库存资金额'),ylabel('y销售额')
subplot(3,1,2),plot(data(:,2), data(:,4),'b+');
xlabel('x2广告投入'),ylabel('y销售额')
subplot(3,1,3),plot(data(:,3), data(:,4),'k+');
xlabel('x1员工薪酬'),ylabel('y销售额')
%调用命令regress建立三元线性回归模型
X = [ones(m,1), data(:,1), data(:,2), data(:,3)];
Y = data(:,4);
[b,bint,r,rint,stats] = regress(Y,X)
```



2. 多元线性回归模型

$$y = 162.0632 + 7.2739x_1 + 13.9579x_2 - 4.3996x_3$$

b =

162.0632

7.2739

13.9575

-4.3996

bint =

-580.3603 904.4867

4.3734 10.1743

7.1649 20.7501

-46.7796 37.9805

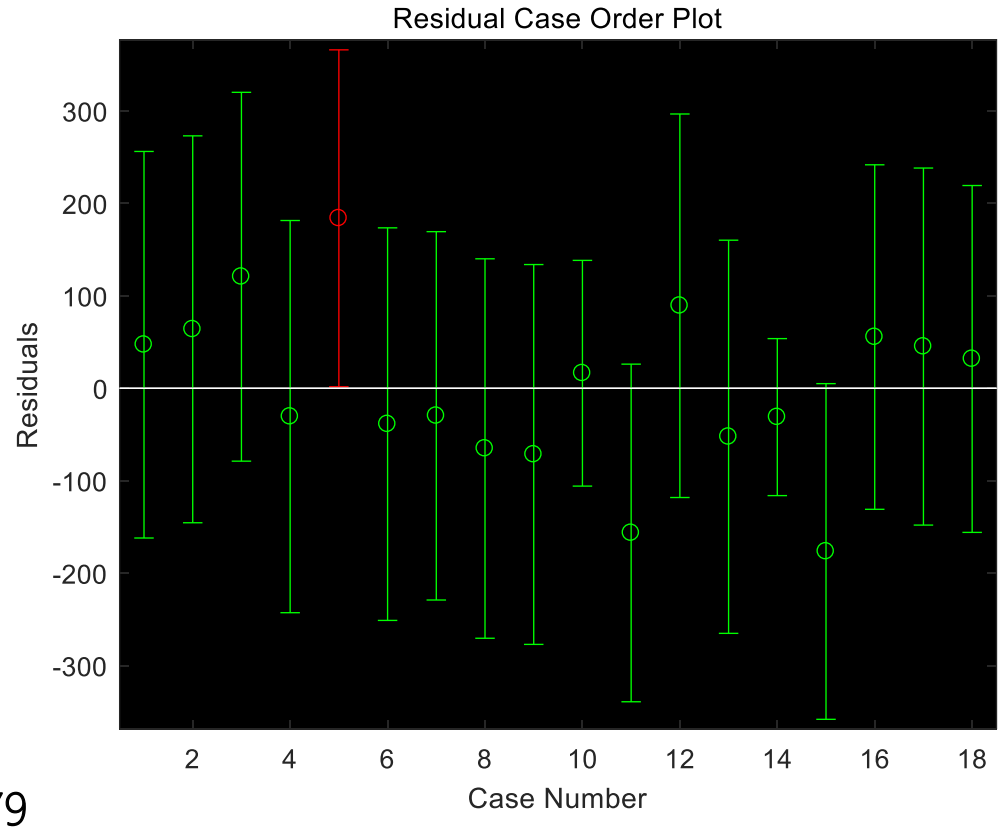
stats =

0.957480405, 105.0866521, 0.000000007746546748, 10077.98679

%绘制残差图

>> rcoplot(r,rint)

第五个点为异常点，从表中可以发现第5个月库存占用资金、广告投入、员工薪酬均比3月份少，为何销售额反而增加？这就可以促使该公司经理找出原因，寻找对策！

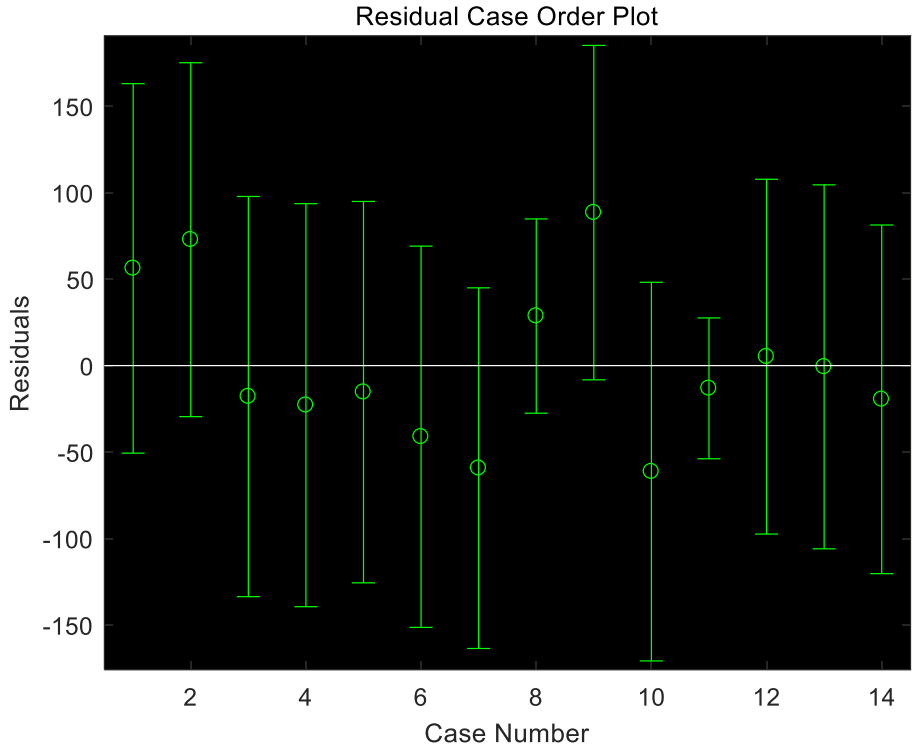


3. 模型改进回归

不断删除异常点后，程序运行如下：

```
>> A = [75.2,30.6,21.1,1090.4;77.6,31.3,21.4,1133; 76,29.6,21.4,1003.2;81.8,27.9,21.7,1012.2; 98.3,24.8,21.5,1098.8;
       67.7,23.6,21,826.3; 74,33.9,22.4,1003.3; 151,27.7,24.7,1554.6; 102.3,42.6,24.3,1483.1; 15.6,40,23.1,1407.1;
       125,45.8,29.1,1551.3; 175.6,67.2,27.5,2311.7;155.2,65,26.5,2126.7;174.3,65.4,26.8,2256.5];
>> [m,n] = size(A);
>> X = [ones(m,1),A(:,1),A(:,2),A(:,3)];
>> Y = A(:,4);
>> [b,bint,r,rint,stats] = regress(Y,X)
  213.7496
    7.6252
   15.8079
  -11.2183
stats =
   0.9913140039, 380.4261456, 0.0000000001333563913, 2774.370745
```

stats可决系数提高到0.9913，F统计量观测值提高到380.42；p远小于alpha，残差平方和缩小到2774.4



三. 广义线性回归

- 广义线性模型（generalized linear model, GLM）由Nelder & Wedderburn(1972)首先提出，是一般线性模型的直接推广。
- GLM通过联结函数建立响应变量的数学期望值与线性组合的预测变量之间的关系，即它使因变量的总体均值通过一个非线性连接函数（link function）而依赖于线性预测值，同时还允许响应概率分布为指数分布族中的任何一员。
- 许多广泛应用的统计模型均属于广义线性模型，如logistic回归模型、Probit回归模型、Poisson回归模型、负二项回归模型等。
- 指数分布族的概率密度（概率函数）可表示为：

$$f(y)=exp\left(\frac{y\theta-b(\theta)}{a(\varphi)}+c(y,\varphi)\right) \qquad E(y)=\mu=b'(\theta),\ \ Var(y)=\varphi\cdot b''(\theta)$$

其中，θ和φ为两个参数，θ称为自然参数，φ为离散参数；a、b、c为函数。

三. 广义线性回归

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

$$f(y)=exp\left(\frac{y\theta-b(\theta)}{a(\varphi)}+c(y,\varphi)\right)$$ ，另一种形式 $$f(y;\theta)=b(y)exp\left(\theta^T T(y)-a(\theta)\right)$$

$$E(y)=\mu=b'(\theta) \quad Var(y)=\varphi\cdot b''(\theta)$$

正态分布:

$$f(y)=\frac{1}{\sqrt{2\pi}\sigma}exp\left(-\frac{y^2}{2\sigma^2}\right)\cdot exp\left(\frac{2\mu y-\mu^2}{2\sigma^2}\right)$$

$$\begin{cases} a(\varphi)=\sigma^2 \\ \theta=\mu \\ b(\theta)=\frac{\mu^2}{2}=\frac{\theta^2}{2} \\ c(y,\sigma^2)=\left(-\frac{y^2}{2\sigma^2}\right)^{\frac{1}{\sqrt{2\pi}\sigma}} \end{cases}$$

各种常见的指数型分布及其主要参数

分　布	θ	$b(\theta)$	φ	$E(y)=b'(\theta)$	$Var(y)=b''(\theta)\varphi$
正态分布	μ	$\theta^2/2$	σ^2	$\mu=\theta$	σ^2
逆高斯分布	$\frac{1}{\mu^2}$	$-(-2\theta)^{1/2}$	σ^2	$\mu=\frac{1}{\sqrt{\theta}}$	$\mu^3\sigma^2$
伽玛分布	$\frac{1}{\mu}$	$-\ln(\theta)$	$\frac{1}{\gamma}$	$\mu=\frac{1}{\theta}$	$\mu^2\gamma$
二项分布	$\ln\frac{p}{1-p}$	$\ln(1+e^\theta)$	1	$p=\frac{e^\theta}{1+e^\theta}$	$p(1-p)$
Poisson 分布	$\ln\lambda$	e^θ	1	$\lambda=e^\theta$	λ
负二项分布	$\ln\lambda$	e^θ	k	$\lambda=e^\theta$	$\lambda+k\lambda^2$

三. 广义线性回归

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

	线性预测	Y的预测值为μ时，Y的分布	联系函数	使用场景
普通线性模型	是	平均值为μ，方差为常数的高斯分布	y	连续变量，比如身高、体重
逻辑回归模型	是	概率为μ的二项分布	$\log\left(\frac{P(y=1)}{1-P(y=1)}\right)$	二分类变量，如购买行为，投票行为
多项逻辑回归模型	是	概率为μ的二项分布	$\log\left(\frac{P(y=i)}{P(y=c)}\right)$	多分类变量，如分类模型
定序回归模型	是	概率为μ的二项分布	$\log\left(\frac{P(y\leq i)}{1-P(y\leq i)}\right)$	定序变量，如人的主观感受
泊松回归模型	是	平均值为μ，方差也为μ的泊松分布	$\log(y)$	计数变量，如销售数量

三. 广义线性回归

- 一个广义线性模型包括以下三个组成部分：

(1) 线性成分(linear component)：$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi}$

(2) 随机成分(random component)：$\varepsilon_i = Y_i - \eta_i$

(3) 连接函数 (link function)：$\eta_i = g(\mu_i)$

连接函数为一单调可微（连续且充分光滑）的函数。

- 故广义线性模型：$g(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_m x_{mi} + \varepsilon_i$
  - 如逻辑回归模型：$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon = w^T x$
  - 如泊松回归模型：$\log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon = w^T x$

三. 广义线性回归

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

fitglm为广义线性回归函数，其调用格式：

`mdl = fitglm(tbl)` returns a generalized linear model fit to variables in the table or dataset array `tbl`. By default, `fitglm` takes the last variable as the response variable.

`mdl = fitglm(X,y)` returns a generalized linear model of the responses `y`, fit to the data matrix `X`.

`mdl = fitglm( ___ ,modelspec)` returns a generalized linear model of the type you specify in `modelspec`.

`mdl = fitglm( ___ ,Name,Value)` returns a generalized linear model with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify which variables are categorical, the distribution of the response variable, and the link function to use.

Canonical Link Function

The default link function for a generalized linear model is the *canonical link function*.

Distribution	Canonical Link Function Name	Link Function	Mean (Inverse) Function
'normal'	'identity'	$f(\mu) = \mu$	$\mu = Xb$
'binomial'	'logit'	$f(\mu) = \log(\mu/(1-\mu))$	$\mu = \exp(Xb) / (1 + \exp(Xb))$
'poisson'	'log'	$f(\mu) = \log(\mu)$	$\mu = \exp(Xb)$
'gamma'	-1	$f(\mu) = 1/\mu$	$\mu = 1/(Xb)$
'inverse gaussian'	-2	$f(\mu) = 1/\mu^2$	$\mu = (Xb)^{-1/2}$


01
IDEA
02
LEARNING
03
CREATIVE
INSPIRATION
EDUCATION
BUSINESS
PHOTO BOOK

感谢聆听