作业1：线性回归

1. 名词解释

人工智能是研究模拟、扩展智能的理论方法及技术应用的科学。人工智能包括模式识别、机器学习等领域，旨在研究智能活动的基本规律，构造具有智能的人工系统。

模式识别是通过计算机用数学技术方法来研究模式的自动处理和判读，环境与客体被统称为“模式”。模式识别需要抽取主要表达特征进行“训练”，识别时再进行比较“匹配”。

机器学习是实现人工智能的一种理论方法，主要设计和分析让计算机可以自动“学习”的算法，从数据中自动分析获得规律，并利用规律对未知数据进行预测。

深度学习是机器学习中一种基于对数据进行表征学习的方法，通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征。

统计学习是关于计算机基于数据构建概率统计模型并运用模型对数据进行预测与分析的一门学科。统计学习方法包括模型、策略和算法，通过学习方法选择最优模型，再利用学习到的最优模型对新数据进行预测和分析。

2. 证明

$$r=\frac{cov(x,y)}{\sigma_x\sigma_y}=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2}\sqrt{\sum(y_i-\overline{y})^2}}$$

因为：

$$\hat{\theta}_1=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}$$

故：

$$r=\hat{\theta}_1\sqrt{\frac{\sum(x_i-\overline{x})^2}{\sum(y_i-\overline{y})^2}}\Rightarrow r^2=\hat{\theta}_1^2\frac{\sum(x_i-\overline{x})^2}{\sum(y_i-\overline{y})^2}$$

又因为：

$$\hat{y}_i=\hat{\theta}_0+\hat{\theta}_1x_i$$
$$\overline{y}=\hat{\theta}_0+\hat{\theta}_1\overline{x}$$
$$\Rightarrow\sum(\hat{y}_i-\overline{y})^2=\sum[\hat{\theta}_1(\hat{x}_i-\overline{x})]^2=\hat{\theta}_1^2\sum(\hat{x}_i-\overline{x})^2$$

因此：

$$R^2=\frac{\sum(\hat{y}_i-\overline{y})^2}{\sum(y_i-\overline{y})^2}=\hat{\theta}_1^2\frac{\sum(x_i-\overline{x})^2}{\sum(y_i-\overline{y})^2}=r^2$$

3. 过拟合问题

3.1 随机生成训练样本数据

设计生成训练样本数据的函数如下，将样本数与标准差作为参数输入，便于分别考虑sigma = 0.5和2，以及训练样本数据为10和100的情况

```
function [x, y]= generate(size,sigma)%输入样本数与标准差
theta1=3;
theta0=6;
x=zeros(size,1);
e=zeros(size,1);
for i=1:size
    x(i)=normrnd(0,1);
    e(i)=normrnd(0,sigma);
end
y=e+theta0+theta1*x;
end
```

3.2 回归与模型预测

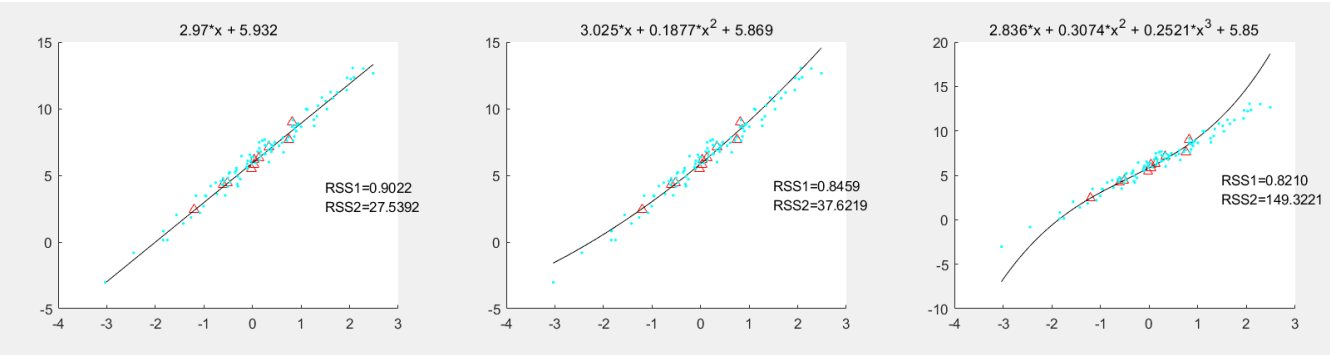
主函数与曲线拟合及模型预测的函数如下：

```
function HM1_3%主函数
clc
clear
[X1,Y1]=generate(10,0.5);%训练数据为10
%[X2,Y2]=generate(100,0.5);%训练数据为100
[X3,Y3]=generate(100,0.5);

[X4,Y4]=generate(10,2);%训练数据为10
%[X5,Y5]=generate(100,2);%训练数据为100
[X6,Y6]=generate(100,2);
r=zeros(6,1);
R=zeros(6,1);
for j=1:3
    subplot(2,3,j);
    hold on;
    [~,r(j),R(j)]=simu(j,X1,Y1,X3,Y3);
    %[~,r(j),R(j)]=simu(j,X2,Y2,X3,Y3);
    hold off;
end
for j=1:3
    subplot(2,3,j+3);
    hold on;
    [~,r(j),R(j)]=simu(j,X4,Y4,X6,Y6);
    %[~,r(j),R(j)]=simu(j,X5,Y5,X3,Y3);
    hold off;
end
end
```

```
function [ p,RSS1,RSS2 ] = simu(index,X,Y,X_n,Y_n)
%曲线拟合
p=polyfit(X,Y,index);
y_hat=polyval(p,X);
err1= Y -y_hat;
RSS1= err1'*err1;%训练样本数据的RSS
err=Y_n-polyval(p,X_n);
RSS2=err'*err;%测试样本数据的RSS
%画图
x1 = linspace(min(min(X,min(X_n))),max(max(X,max(X_n))));
y1 = polyval(p, x1);
plot(X, Y, 'r^', x1, y1, 'k-',X_n,Y_n,'c.');
t=sprintf('RSS1=%.4f\n RSS2=%.4f',RSS1,RSS2);
text(max(max(X,max(X_n)))-1, polyval(p,min(min(X,max(X_n))))+1, t)
tit = char(vpa(poly2sym(p), 4));
title(tit);
end
```

sigma=0.5时散点图与回归曲线如下图所示，其中红色三角表示训练所用的10个随机生成训练样本数据，黑色代表回归得到的拟合曲线，蓝色圆点代表100个随机生成测试样本数据

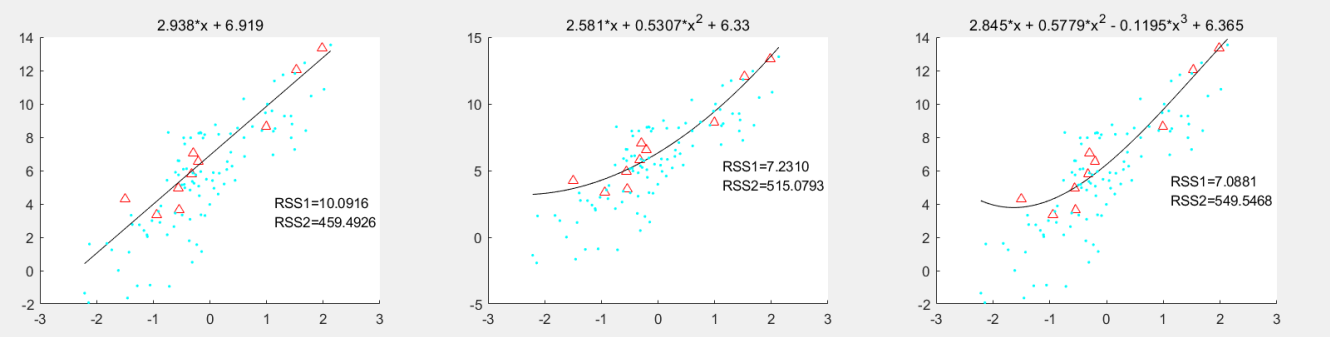


可以看出用线性模型, 一元二次模型，一元三次模型进行回归时，所得模型分别为：

$$y = 2.97x + 5.932,\ RSS = 0.9022$$
$$y = 0.1877x^2 + 3.025x + 5.869, RSS = 0.8459$$
$$y = 0.2521x^3 + 0.3074x^2 + 2.836x + 5.85, RSS = 0.8210$$

从训练样本的RSS可以看出随着模型越来越复杂，训练样本上的回归效果越来越好；以预测样本数据的RSS对模型预测效果进行评价，预测样本的RSS分别为 27.5392, 37.6219, 149.3221，随着模型越来越复杂，测试样本上的预测精确度却越来越差，说明出现过拟合现象。

sigma=2时散点图与回归曲线如下图所示，各图标意义保持不变。



可以看出用线性模型, 一元二次模型，一元三次模型进行回归时，所得模型分别为：

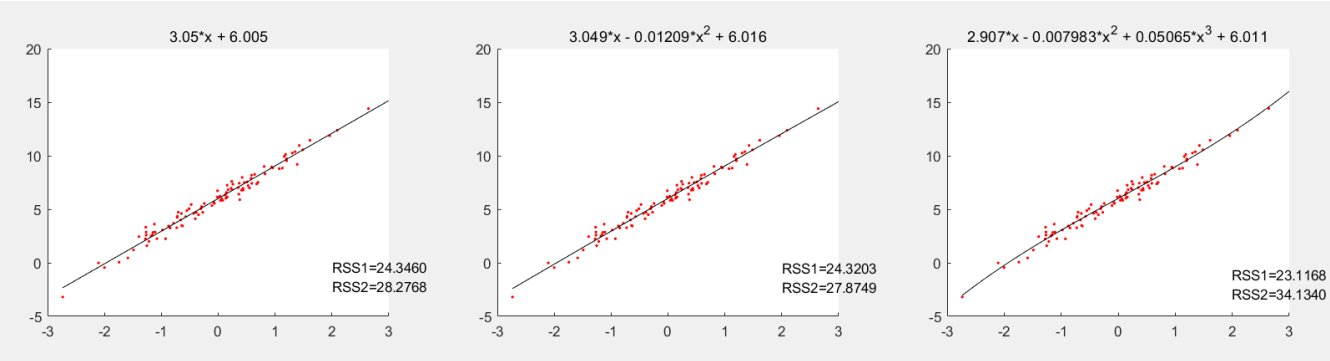
$$y = 2.938x + 6.919，\ RSS = 10.0916$$
$$y = 0.5307x^2 + 2.581x + 6.33, RSS = 7.2310$$
$$y = 0.1195x^3 + 0.5779x^2 + 2.845x + 6.365, RSS = 7.0881$$

仍然可以得出相同结论：从训练样本的RSS可以看出随着模型越来越复杂，训练样本上的回归效果越来越好；预测样本数据的RSS分别为459.4926，515.0793，549.5468，随着模型越来越复杂，测试样本上的预测精确度越来越差，出现过拟合现象。

和sigma=0.5 时相比，训练样本的方差变大后，数据的波动性增大，模型的拟合效果明显下降。

3.2 扩大训练样本重新实验

将随机生成的训练样本数改为100，sigma=0.5时散点图与回归曲线如下图所示，为了便于观察，此时红色圆点表示训练所用的100个随机生成训练样本数据，黑色代表回归得到的拟合曲线，随机生成测试样本数据不再显示

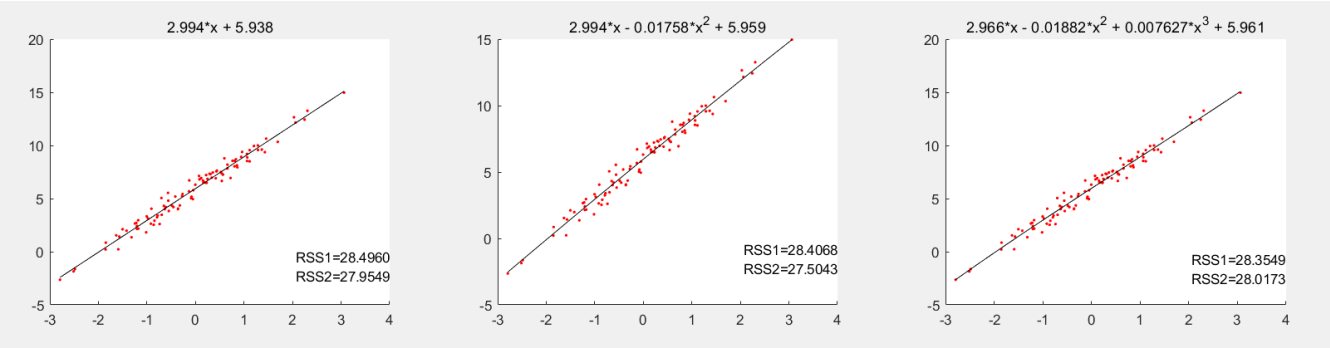


可以看出用线性模型, 一元二次模型，一元三次模型进行回归时，所得模型分别为：

$$y = 3.05x + 6.005，\ RSS = 24.3460$$
$$y = -0.01209x^2 + 3.049x + 6.016, RSS = 24.3203$$
$$y = 0.05065x^3 - 0.007983x^2 + 2.907x + 6.011, RSS = 23.1168$$

此时仍可以得到结论：随着模型越来越复杂，模型对训练样本的拟合程度不断提高，但测试集的残差平方和也不断增大，预测效果变差；在训练样本量增大后，三个模型的拟合效果更加接近，训练 RSS与测试 RSS 的区别减小，总体而言模型的预测效果有所提高。

sigma=2时散点图与回归曲线如下图所示，各图标意义保持不变：



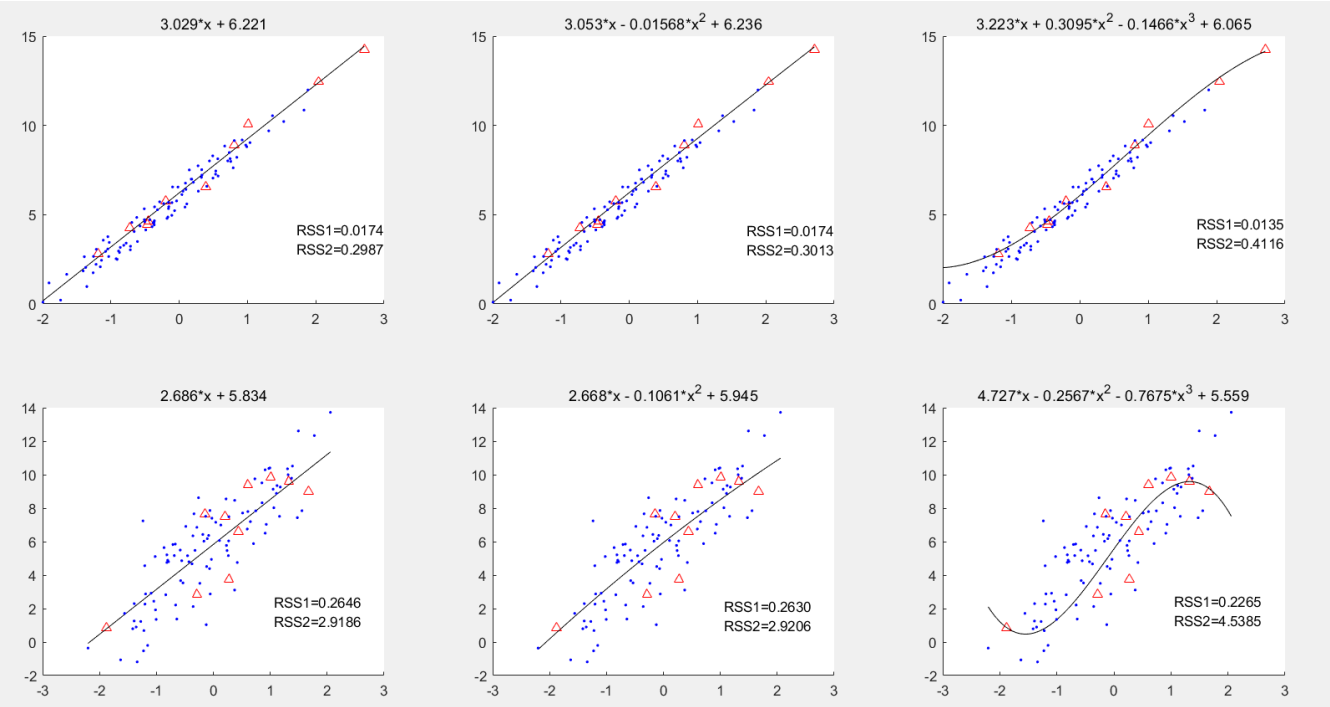
此时用线性模型, 一元二次模型，一元三次模型进行回归时，所得模型分别为：

$$y = 2.994x + 5.938,\ RSS = 28.4960$$
$$y = -0.01758x^2 + 2.994x + 5.959, RSS = 28.4068$$
$$y = 0.007627x^3 - 0.01882x^2 + 2.966x + 5.961, RSS = 28.3549$$

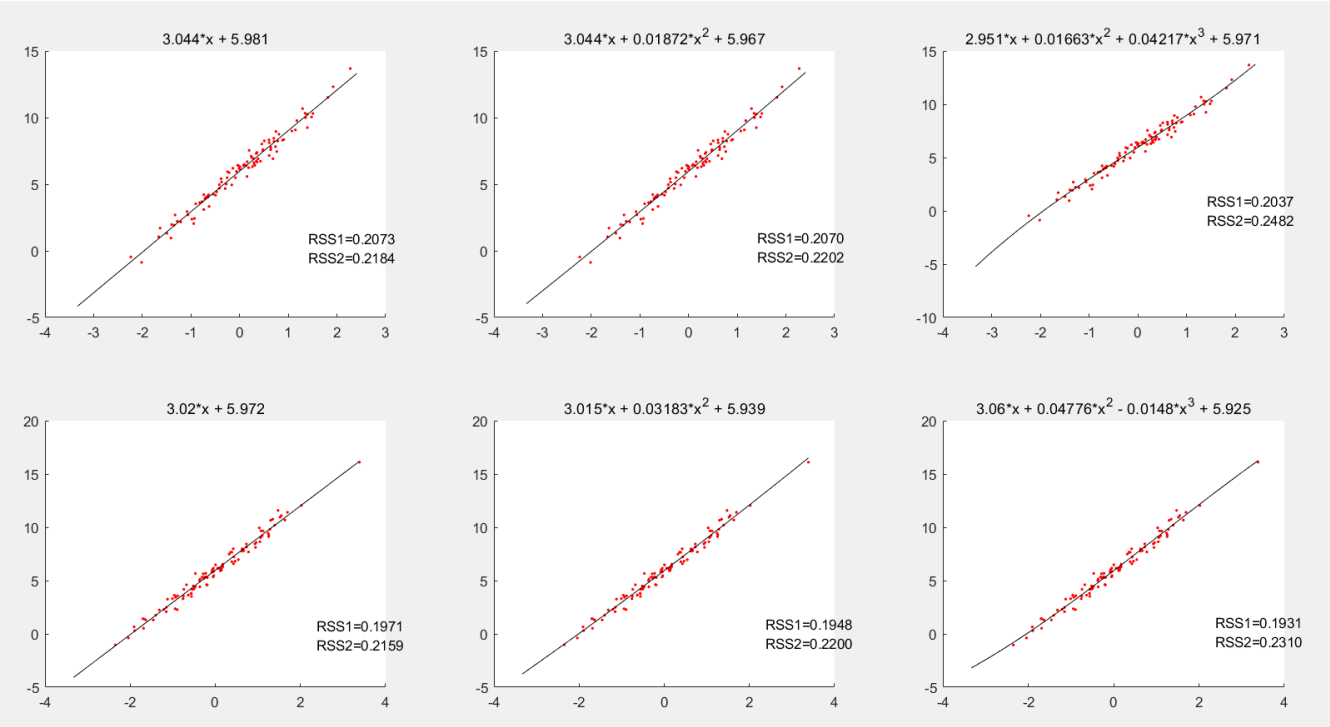
随着模型越来越复杂，训练样本上的回归效果越来越好，测试样本上的预测精确度越来越差，出现过拟合现象；在训练样本量增大后，三个模型的拟合效果更加接近，训练 RSS与测试 RSS 区别减小，总体上模型的预测效果有所提高；和sigma=0.5时相比，训练样本的方差变大，数据的波动性增大，模型的拟合效果略微下降。

4. 对比总结

为了更准确的比较各因素，排除样本数对RSS的影响，将以上实验中RSS均替换为平均RSS进行多次实验，总结实验结论如下：



（sigma=0.5 和 sigma=2时，训练样本为10的回归与预测效果）



（sigma=0.5 和 sigma=2时，训练样本为100的回归与预测效果）

其他因素相同时，sigma越大，训练样本的方差越大，数据的波动性越大，回归与预测的效果较差

其他因素相同时，模型越复杂，越容易出现过拟合现象，模型在训练样本上的回归效果越来越好，但在测试样本上预测精度会越来越差。

其他因素相同时，训练样本量越大，模型在训练样本上的回归效果与测试样本上的预测效果均会提高，且不同模型间的预测的差别减小。

4. 前列腺特异抗原水平预测

1. 多元线性回归

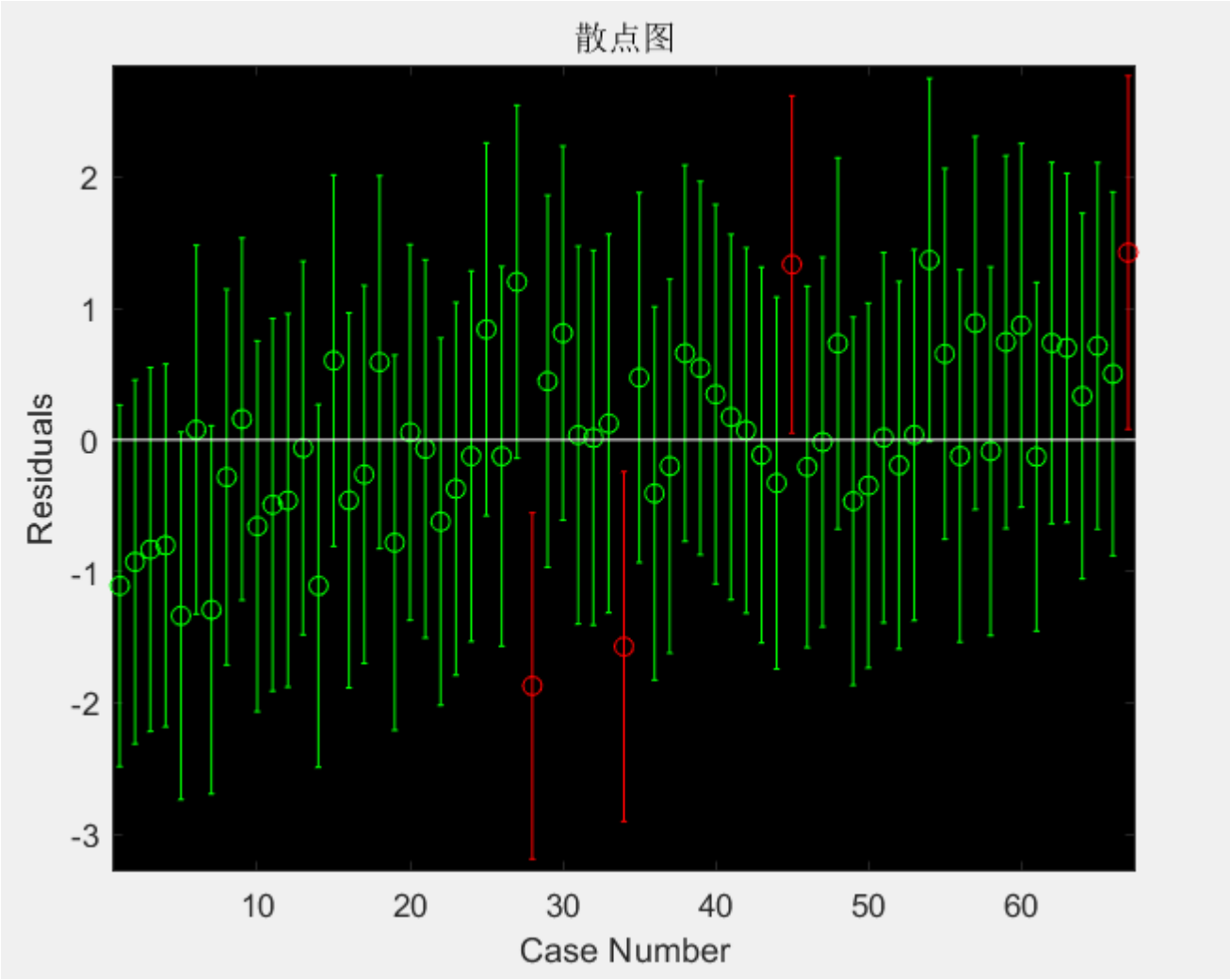
设$lpsa = \hat{\theta}_0 + \hat{\theta}_1 * lcavol + \hat{\theta}_2 * lweight + \hat{\theta}_3 * lbph + \hat{\theta}_4 * svi$

利用regress（）进行多元线性回归的程序如下：

```
clear;
clc;
X1=importdata('prostate_train.txt');
X2=importdata('prostate_test.txt');

x=[ones(size(X1.data(:,1))) X1.data(:,1:4)];
y=X1.data(:,5);
[b,bint,r,rint,stats]=regress(y,x);
rcoplot(r,rint);
title('散点图')
```

可得残差图如下：



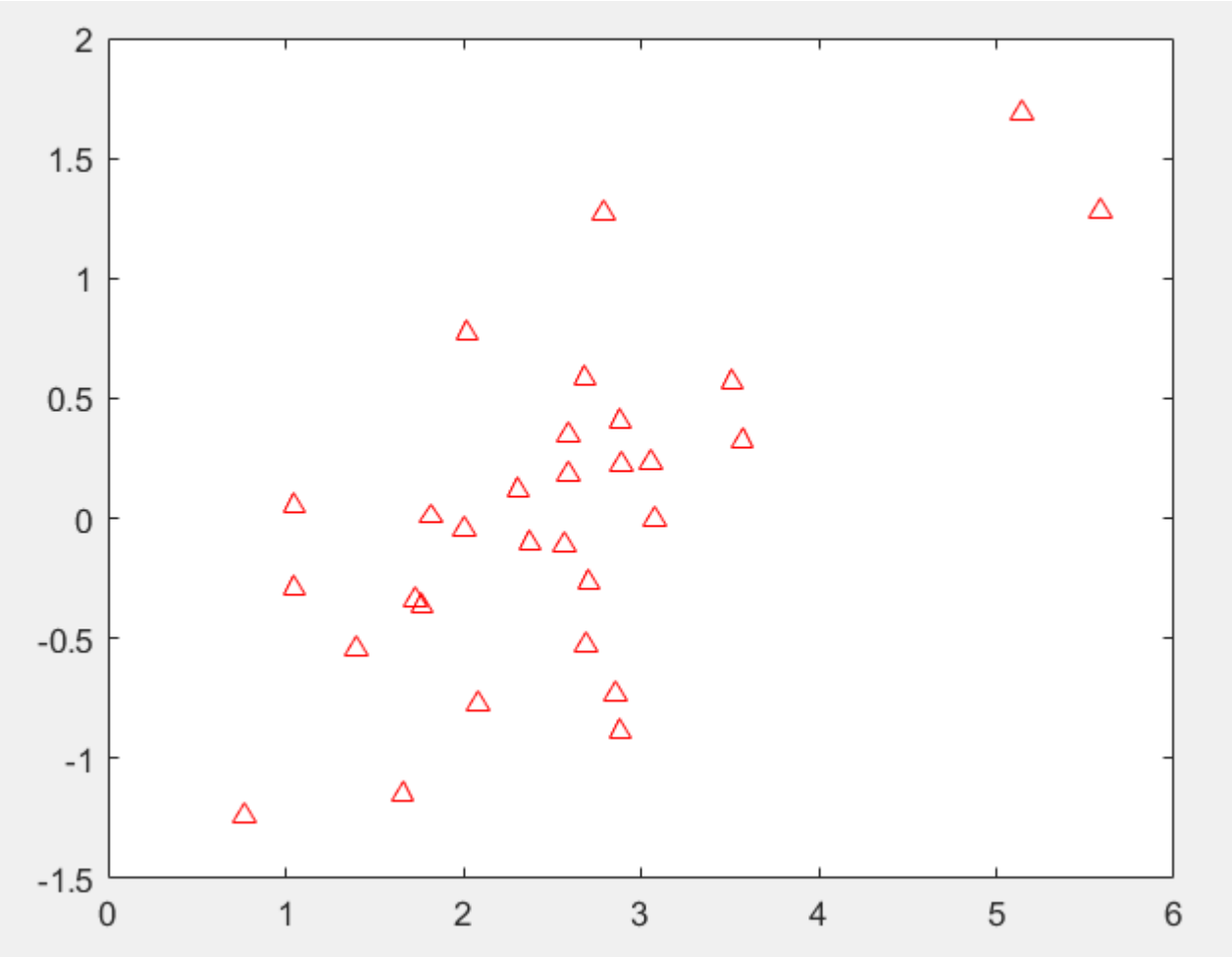
回归系数 $\hat{\theta}_0 = -0.3259, \hat{\theta}_1 = 0.5055, \hat{\theta}_2 = 0.5388, \hat{\theta}_3 = 0.1400$ $\hat{\theta}_4 = 0.6718$

故可得线性回归方程为：

$$lpsa = -0.3259 + 0.5055 * lcavol + 0.5388 * lweight + 0.1400 * lbph + 0.6718 * svi$$

对测试集数据进行预测，可得残差平方和RSS=13.6900，平均残差平方和为0.4563，训练集上平均残差平方和为0.4898，可见预测效果良好。

可以观察测试集上的偏差分布如下：



2.考虑交叉项

考虑二次交叉项时得到的线性回归方程如下：

$$lpsa = -0.8830 + 0.7587 * lcavol + 0.7240 * lweight + 1.2337 * lbph + 4.3586 * svi \\ - 0.0709 * lcavol * lweight + 0.0419 * lcavol * lbph - 0.1956 * lcavol * svi \\ - 0.3007 * lweight * lbph - 0.8711 * lweight * svi - 0.2175 * lbph * svi$$

此时训练集上平均残差平方和为0.4221，和先前相比有所下降

但此时测试集上的RSS=15.9426，残差平方和为0.5341，与不考虑交叉项时相比预测效果变差了。

因此考虑交叉项时样本在训练集上解释的方差更多了，但测试集上预测效果变差了，可能出现了一定程度的过拟合，因而不考虑交叉项时预测效果更好。