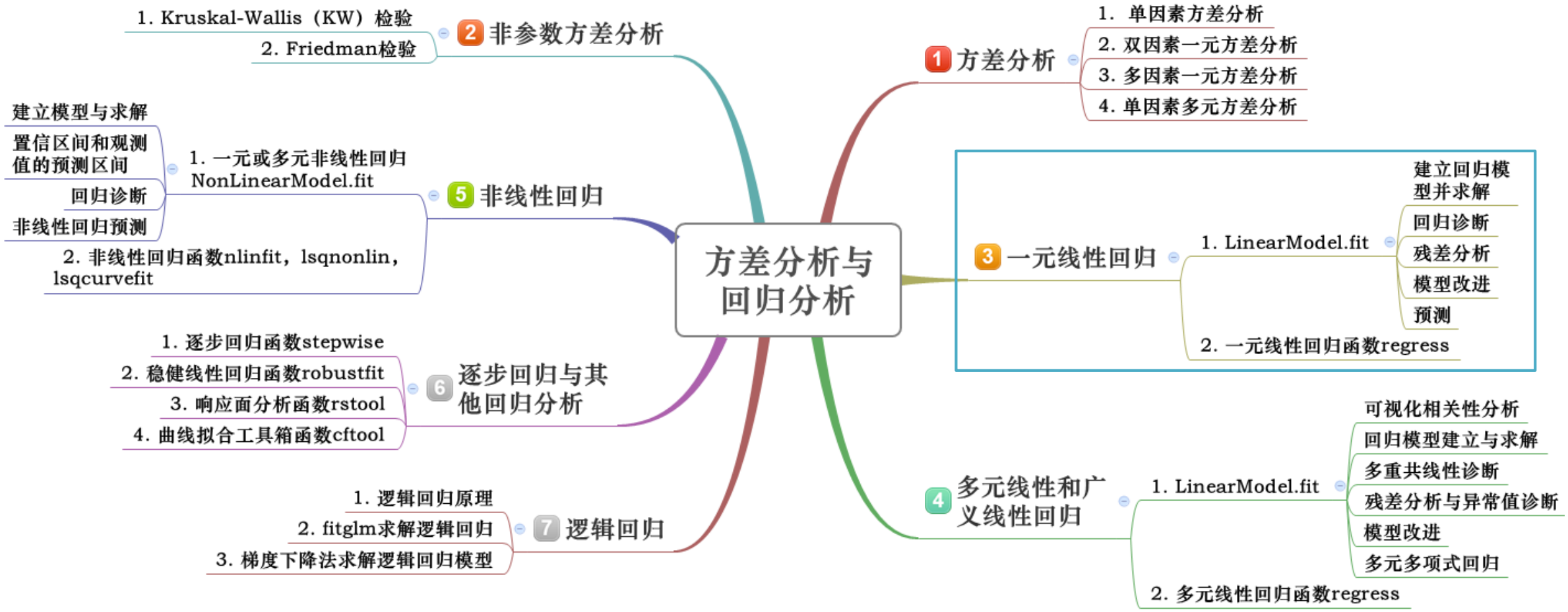
信陽師範学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

第11章 方差分析与回归分析

讲授人：牛言涛

日期：2020年4月15日

第11章 方差分析与回归分析知识点思维导图

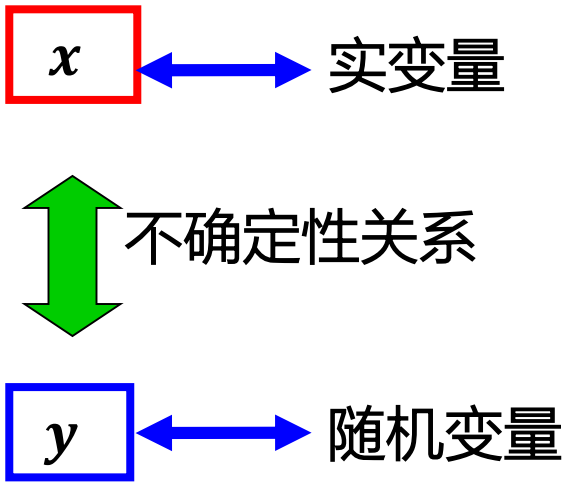


回归分析概述

1、确定性关系与相关关系

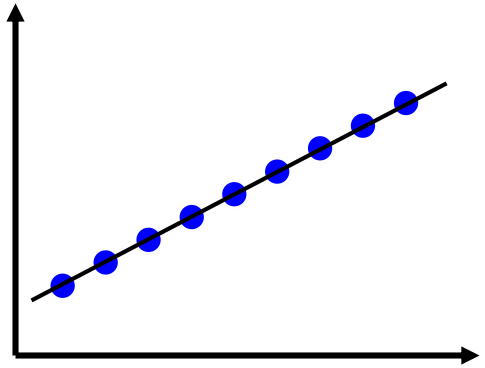
变量间的关系

- 确定性关系或函数关系 $y = f(x)$
- 不确定性关系
  - 人的身高和体重
  - 家庭的收入和消费
  - 商品的广告费和销售额
  - 粮食的产量和施肥量
  - 股票的价格和时间
  - 学生的期中和期末考试成绩,…

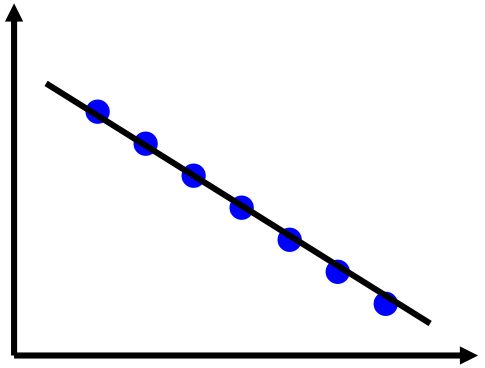


回归分析概述

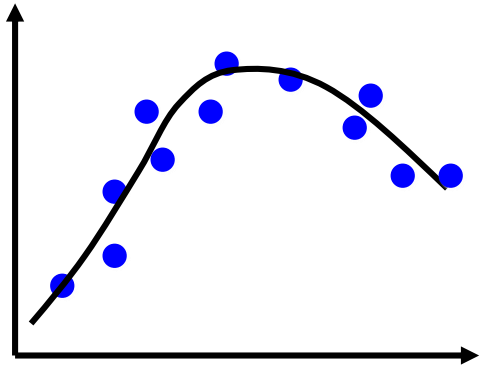
2、相关关系的图示



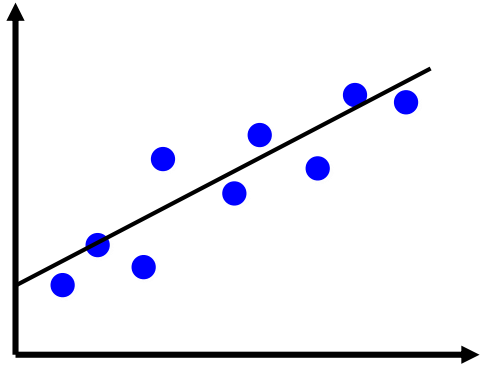
完全正线性相关



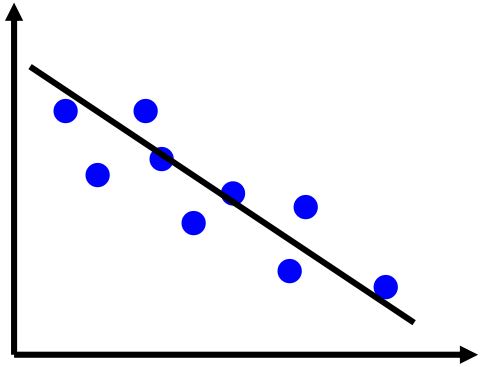
完全负线性相关



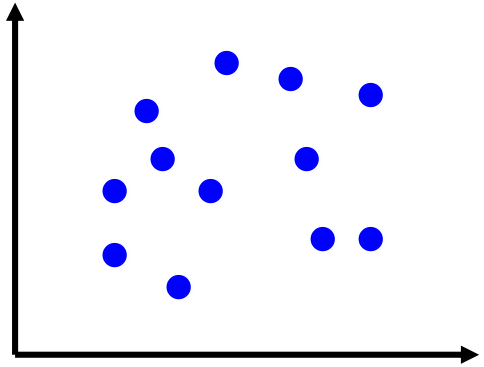
非线性相关



正线性相关



负线性相关



不相关

回归分析概述

信阳师范学院
数学与统计学院

3．什么是回归分析

（1）从一组样本数据出发，确定变量之间的数学关系式；

（2）对这些关系式的可信程度进行各种统计检验，并从影响某一特定变量的诸多变量中找出哪些变量的影响显著，哪些不显著；

（3）利用所求的关系式，根据一个或几个变量的取值来预测或控制另一个特定变量的取值，并给出这种预测或控制的精确程度。

（4）设Y是一个可观测的随机变量，它受到$p(p>0)$个非随机变量因素X_i和随机误差的影响，若Y与X_i有如下线性关系：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p+\varepsilon$$

其中β_i称为回归系数，Y称为多元线性回归模型。

一. fitlm函数回归分析

`mdl = fitlm(tbl)` 返回基于表或数据集数组 `tbl` 中变量拟合的线性回归模型。默认情况下，`fitlm` 将最后一个变量作为响应变量。

`mdl = fitlm(X,y)` 返回基于数据矩阵 `X` 拟合的响应 `y` 的线性回归模型。

`mdl = fitlm(___,modelspec)` 使用上述语法中的任何输入参数组合来定义模型设定。

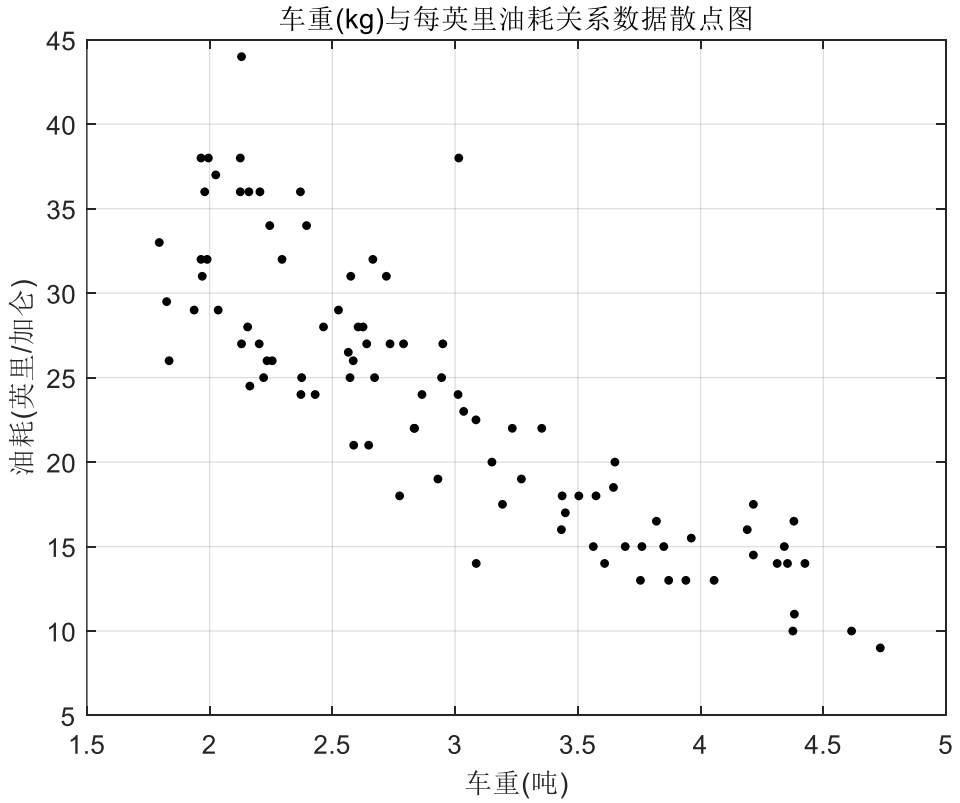
`mdl = fitlm(___,Name,Value)` 使用一个或多个名称-值对组参数指定附加选项。例如，您可以指定哪些变量是分类变量、执行稳健回归或使用观测值权重。

值	模型类型
'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 x_1、x_2、x_2^2、x_2^3、x_1*x_2 和 $x_1*x_2^2$ 项，其中 x_1 和 x_2 分别是第一个和第二个预测变量。

1. 案例分析——汽车数据线性回归

例1：carsmall.mat是MATLAB自带数据集，包括100种汽车数据：Acceleration百公里加速、Cylinders发动机气缸数、 Displacement发动机排量、 Horsepower马力、Mfg制造商、Model车型、Model_Year制造年、MPG每英里加仑油耗、Origin产地、Weight车重。试建立MPG与Weight的关系。

```
% 1. 绘制散点图，分析关系曲线
load carsmall
x = Weight/1000;   y = MPG;
plot(x,y, 'k.', 'Markersize', 10); %绘制散点图
xlabel('车重(吨)');
ylabel('油耗(英里/加仑)');
title('车重(kg)与每英里油耗关系数据散点图')
grid on
```



2. 回归分析求解

mdl1 = LinearModel.fit(x,y) %不推荐使用

mdl1 =

Linear regression model:

y ~ 1 + x1

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	49.238	1.6411	30.002	2.7015e-49
x1	-8.6119	0.5348	-16.103	1.6434e-28

Number of observations: 94, Error degrees of freedom: 92

Root Mean Squared Error: 4.13

R-squared: 0.738, Adjusted R-Squared 0.735

F-statistic vs. constant model: 259, p-value = 1.64e-28

建立的一元线性模型：$\hat{y} = 49.238 - 8.6119x$

$p = 2.7015e-49 < 0.05$、$1.6434e-28 < 0.05$，说明回归系数是显著的。

均方根误差4.13，均方根误差是用来衡量观测值同真值之间的偏差。

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

R-squared（值范围0-1）描述输入变量对输出变量的解释程度。TSS是执行回归分析前，响应变量固有的方差。RSS残差平方和就是，回归模型不能解释的方差。在单变量线性回归中R-squared 越大，说明拟合程度越好。

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / m}{\sum_i (y_i - \bar{y})^2 / m} = 1 - \frac{MSE(\hat{y}, y)}{Var(y)}$$

2. 回归分析求解

mdl2 = fitlm(x,y)　%推荐使用

mdl2 =

Linear regression model:

y ~ 1 + x1

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	49.238	1.6411	30.002	2.7015e-49
x1	-8.6119	0.5348	-16.103	1.6434e-28

Number of observations: 94, Error degrees of freedom: 92

Root Mean Squared Error: 4.13

R-squared: 0.738, Adjusted R-Squared 0.735

F-statistic vs. constant model: 259, p-value = 1.64e-28

对回归直线进行显著性检验，原假设H_0：$\beta_1 = 0$，择备假设 H_1：$\beta_1 \neq 0$. $p = 1.64e-28 < 0.05$，拒绝H_0，可认为线性关系显著。

用R-Square的时候，不断添加变量能让模型的效果提升，而这种提升是虚假的。利用adjusted R-Square，能对添加的非显著变量给出惩罚，也就是说随意添加一个变量不一定能让模型拟合度上升。Adjusted R-Squared 抵消样本数量对 R-Squared 的影响，做到了真正的 0~1，越大越好。n为样本量，k为特征数。

$$R^2_{adj} = 1 - \frac{\left(1-R^2\right)\left(n-1\right)}{n-k-1}$$

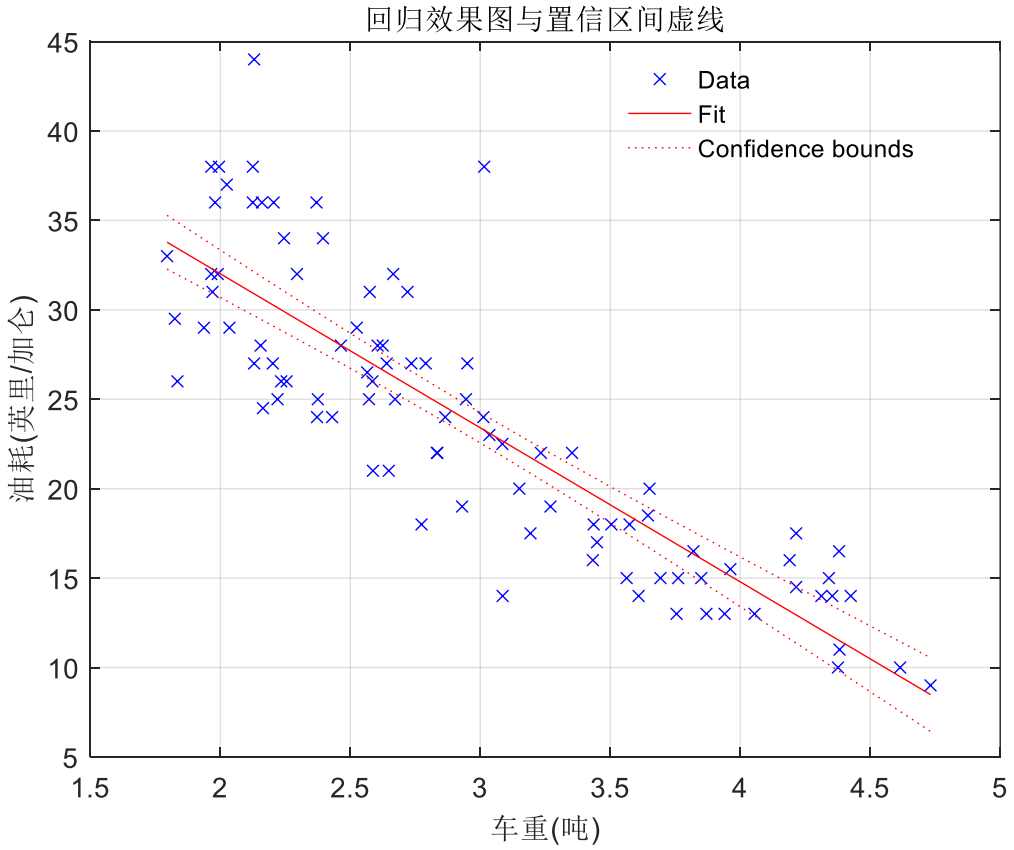
F的值是回归方程的显著性检验，表示的是模型中被解释变量与所有解释变量之间的线性关系在总体上是否显著做出推断。若$F > F_\alpha\,(k-1, n-k)$，则拒绝原假设，即认为列入模型的各个解释变量联合起来对被解释变量有显著影响，反之，则无显著影响。

$$F = \frac{\sum_i \left(\hat{y}_i - \bar{y}\right)^2 \Big/ k}{\sum_i \left(y_i - \hat{y}_i\right)^2 \Big/ \left(n-k-1\right)} \sim F\left(k, n-k-1\right)$$

3. 绘制拟合效果图

LinearModel methods： plot - Summary plot of regression model

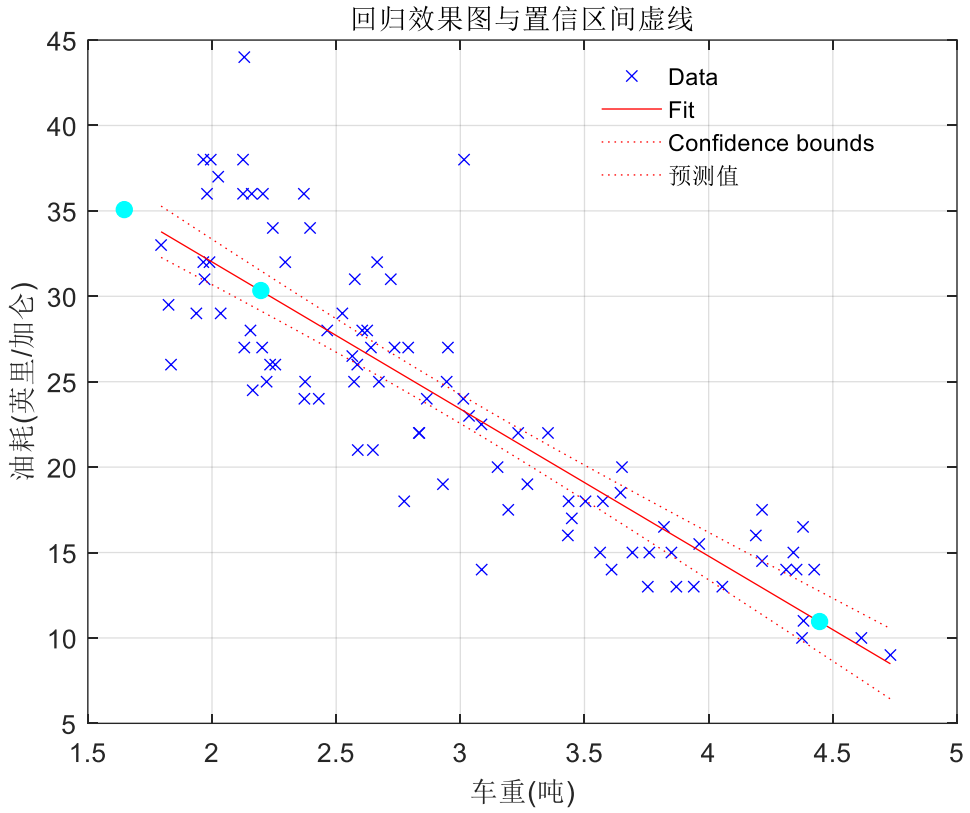
```
>> mdl1.plot   %mdl2.plot
>> xlabel('车重(吨)');
>> ylabel('油耗(英里/加仑)');
>> title('回归效果图与置信区间虚线')
>> grid on
>> legend('boxoff')
```



4. 预测

LinearModel methods： predict - Compute predicted values given predictor values

```
>> xnew = [2.2,1.65,4.45]';
>> ynew = mdl1.predict(xnew)   %ynew = mdl2.predict(xnew)
ynew =
   30.2914
   35.0279
   10.9145
>> hold on
>> plot(xnew,ynew,'co','MarkerFaceColor','c')
>> legend('Data','Fit','Confidence bounds','预测值')
>> legend('boxoff')
```



5. 回归诊断

- 回归系数一般采用“最小二乘估计LS estimator”求解，但是应用中容易忽视的问题是LS估计只有在数据满足相应的条件的情况下才会具有统计描述和推断的优良性质，如要求误差服从正态分布、总体方差相同且相互独立等。
- 当实际数据没有近似满足这些假定时，就会出现异常点(outliers)、杠杆点(leverage point)及影响点(influential observations)，使得结果变得不可靠。
- 回归诊断：
  - 异常点和强影响点诊断：查找数据集中的异常点（离群点）和强影响点，对模型进行改进；
  - 残差分析：用来验证模型的基本假定，包括模型线性诊断、误差正态性诊断、误差方差齐性诊断和误差独立性诊断；
  - 多重共线性诊断：对于多元线性回归，检验自变量之间是否存在共线性关系。

5. 回归诊断

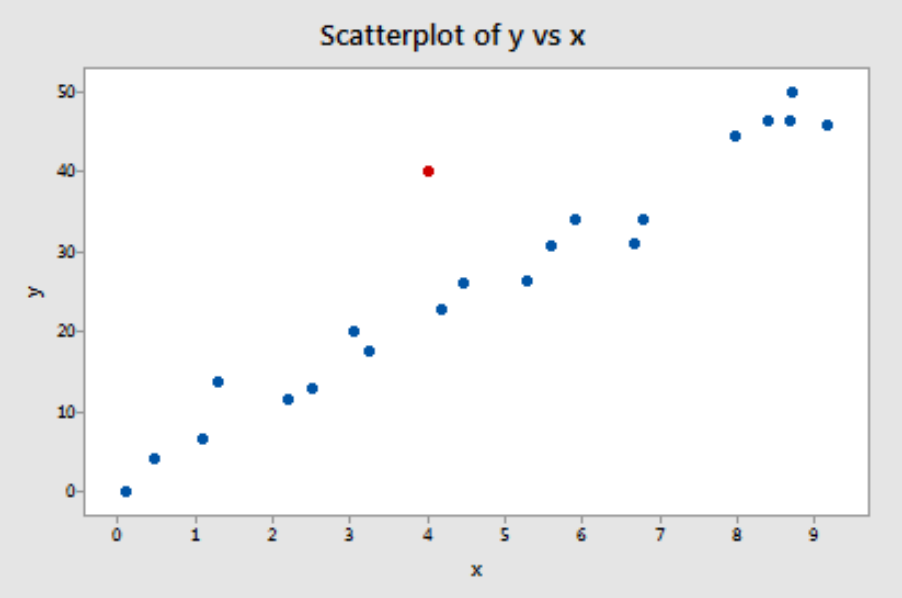
信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

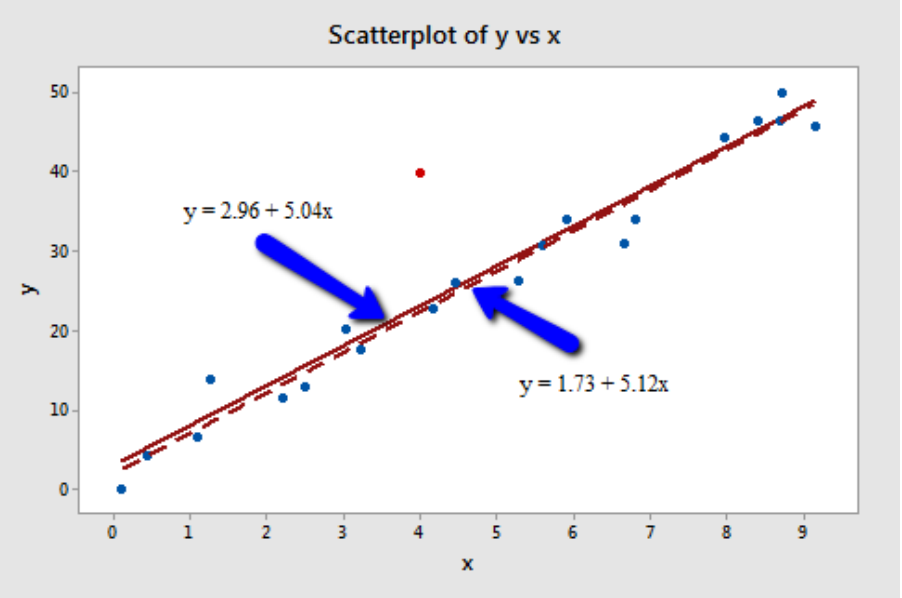
- 异常点的识别与处理，是统计诊断中很重要的内容。异常的的出现会影响分析结果的可信度。但异常点的存在往往蕴涵着重要的信息：
  - 在有些情况下，异常点的出现是因为有新事物出现或者新情况发生，比如经济模型中某种经济政策的出台等，都能表现出异常，这通常是我们的研究兴趣所在，进一步研究。
  - 整体模型变化或局部模型变化，异常点的出现多而且连续，往往蕴涵着机制的变化、新事物的出现或者新局面的形成，大量而且连续的异常点可以用新模型来拟合。对于整个数据集，实质上已经成为一个混合模型。
  - 在另外一些情况下，异常点的出现是由于人为差错或仪器的故障所引起的。此类异常点可删除，改进模型；也可采取容忍的态度，保留异常点。
- 标准化残差和学生化残差用于诊断异常点：标准化残差是指内部的残差为了减少内部差异性制而标准化的残差，便于观察其标准化后的残差做出相应的模型假设判断；学生化的目的是为了减少outliers (influential point)对模型拟合后的残差带来的影响。

5. 回归诊断

(1) 标准化残差（standardized residuals）：相对于普通残差来说，消除了量纲的影响。

(2) 学生化残差（studentized residuals）：相对于标准化残差，还去除了高杠杆值的影响。





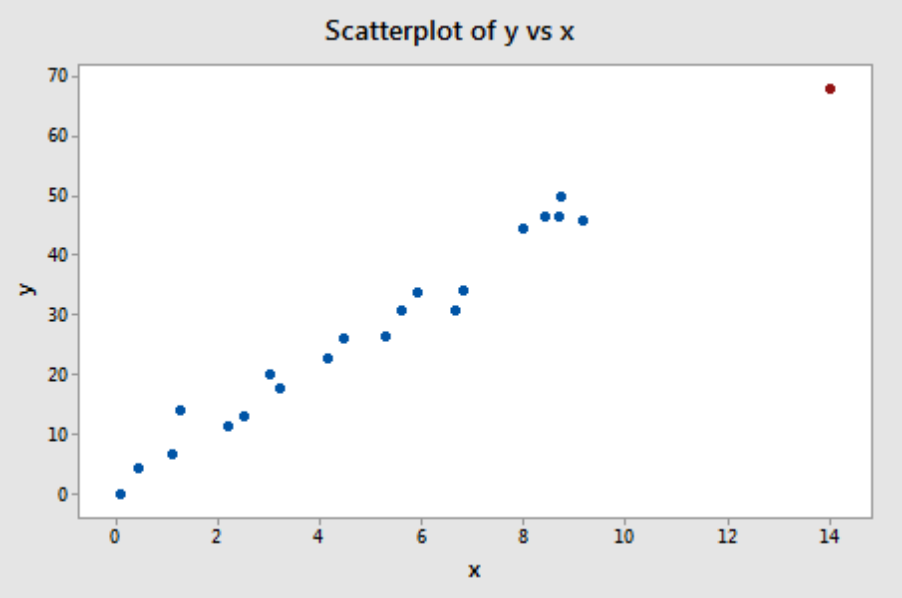
$$z_{e_i} = \frac{e_i}{s_e} = \frac{y_i - \hat{y}_i}{s_e}$$

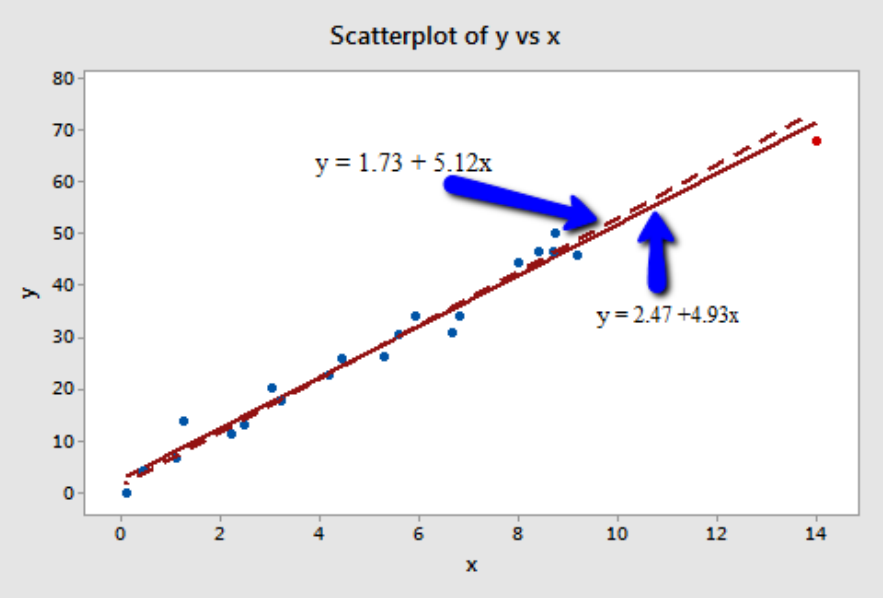
$$t_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

虚线和实线分别是包含红点在内和不包含红点在内训练出来的回归模型。可以看到，两条回归线之间相差不大，因此，该红点不是强影响点。同时，该红点并没有离其他自变量的值很远，因此也不是高杠杆点。但是它离回归线很远（残差大），因此该红点是异常点。

5. 回归诊断

- 高杠杆值观测点，即是与其他预测变量有关的离群点。换句话说，它们是由许多异常的预测变量值组合起来的，与响应变量值没有关系。杠杆点是观测点x是异常的，但是y的值却在合理的预测范围内，杠杆点对模型的拟合影响很大值得关注。



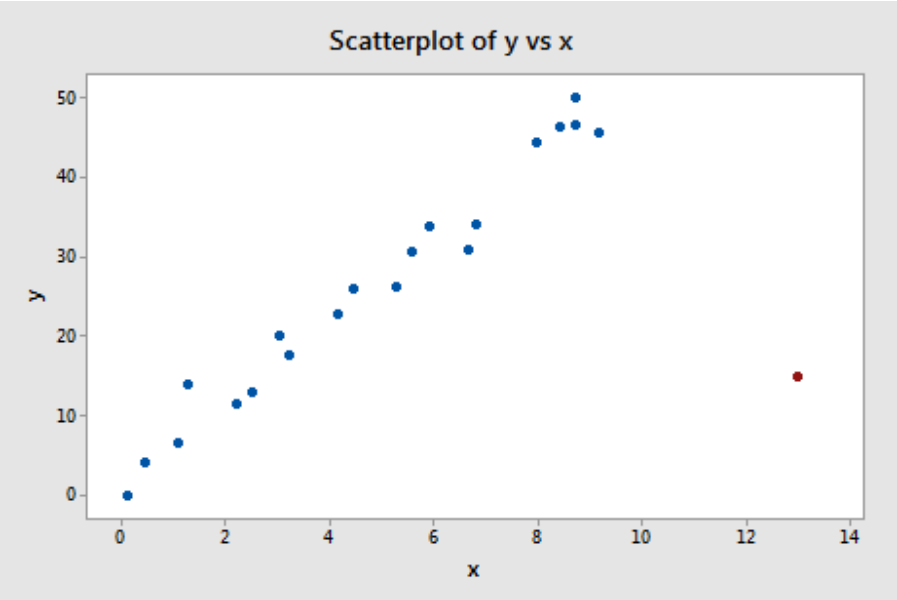


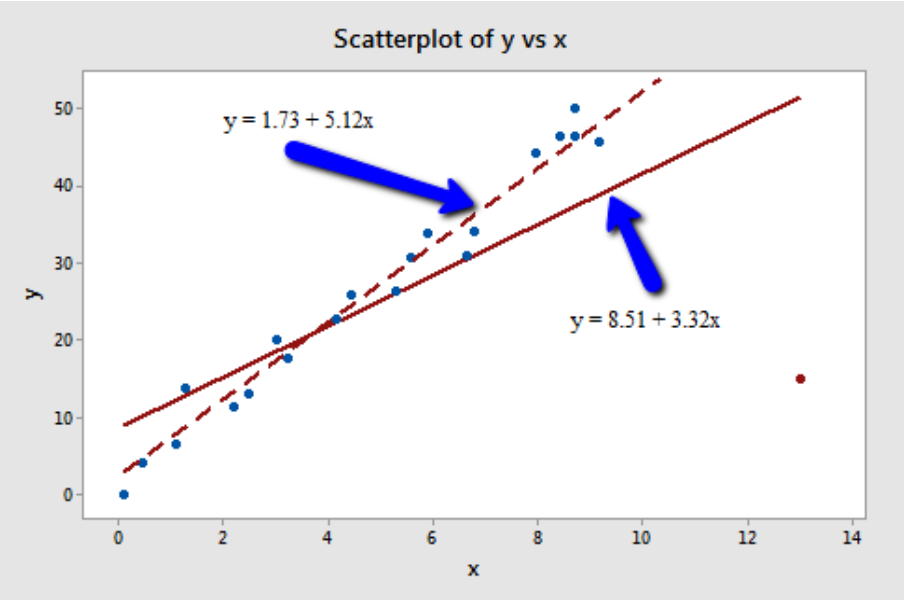
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

虚线和实线分别是包含红点在内和不包含红点在内训练出来的回归模型。可以看到，两条回归线之间相差不大，因此，该红点不是强影响点。同时，该红点离回归线不远，因此也不是异常点。但是它离其他自变量的值很远，因此该红点是高杠杆点。

5. 回归诊断

- 强影响点，即对模型参数估计值影响有些比例失衡的点。例如，若移除模型的一个观测点时模型会发生巨大的改变，那么就需要检测一下数据中是否存在强影响点了。





$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{p \times MSE}$$

虚线和实线分别是包含红点在内和不包含红点在内训练出来的回归模型。可以看到，两条回归线之间相差较大，因此，该红点是强影响点。同时，该红点离其他自变量的值较远，因此是高杠杆点。它离回归线也很远（残差大），因此该红点也是异常点。

5. 回归诊断

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

统计量	定义	判异规则	作用
标准化残差	$Ze_i = e_i/\sqrt{MSE}$	$\lvert Ze_i\rvert > 2$	查找异常值
学生化残差	$Se_i = e_i/\sqrt{MSE(1-h_{ii})}$	$\lvert Se_i\rvert > 2$	
杠杆值	h_{ii}	$h_{ii} > 2\,(p+1)/n$	查找强影响点
Cook距离	$D_i = \frac{e_i^2}{(p+1)MSE} \cdot \frac{h_{ii}}{(1-h_{ii})^2}$	$D_i > 3D$（D为cook平均值）或$4/(n-k-1)$	
Covratio统计量	$C_i = \frac{MSE_{(i)}^{p+1}}{MSE^{p+1}} \cdot \frac{1}{1-h_{ii}}$	$\lvert C_i - 1\rvert > 3\,(p+1)/n$	
Dffits统计量	$Df_i = Se_i\sqrt{\frac{h_{ii}}{1-h_{ii}}}$	$\lvert Df_i\rvert > 2\sqrt{(p+1)/n}$	
Dfbeta统计量	$Db_{ii} = \frac{b_j - b_{j(i)}}{\sqrt{MSE_{(i)}}(1-h_{ii})}$	$\lvert Db_{ii}\rvert > 3\sqrt{n}$	

5. 回归诊断

- n：数据集中观测个数；
- p：为回归模型中自变量个数；
- 第i个观测对应的残差：$\boldsymbol{e_i = y_i - \hat{y}}$；
- 均方残差：$\boldsymbol{MSE = SSE/(n-1-p)}$；
- $\boldsymbol{h_{ii}}$为帽子矩阵$\boldsymbol{H = X(X^TX)^{-1}X^T}$对角线上的第i个元素；
- $\boldsymbol{MSE_{(i)}}$为去掉第i个观测后的均方差残差；
- $\boldsymbol{b_j}$为第j个系数估计值；
- $\boldsymbol{b_{j(i)}}$为去掉第i个观测后的第j个系数估计值。

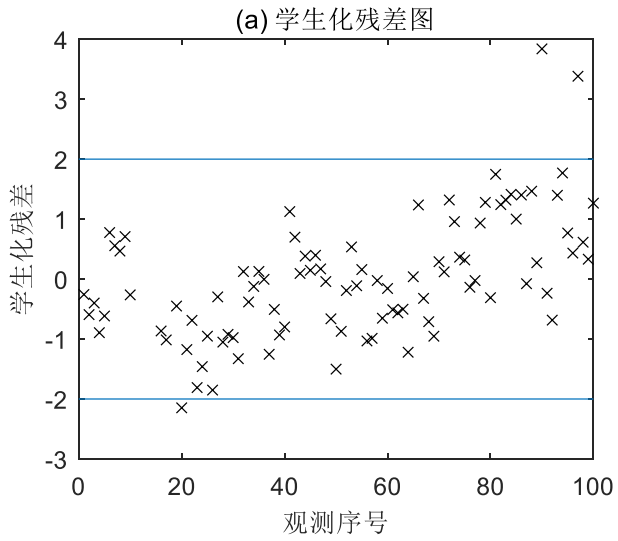
统计量	定义
标准化残差	$Ze_i = e_i/\sqrt{MSE}$
学生化残差	$Se_i = e_i/\sqrt{MSE(1-h_{ii})}$
杠杆值	h_{ii}
Cook距离	$D_i = \frac{e_i^2}{(p+1)MSE} \cdot \frac{h_{ii}}{(1-h_{ii})^2}$
Covratio统计量	$C_i = \frac{MSE_{(i)}^{p+1}}{MSE^{p+1}} \cdot \frac{1}{1-h_{ii}}$
Dffits统计量	$Df_i = Se_i\sqrt{\frac{h_{ii}}{1-h_{ii}}}$
Dfbeta统计量	$Db_{ii} = \frac{b_j - b_{j(i)}}{\sqrt{MSE_{(i)}}(1-h_{ii})}$

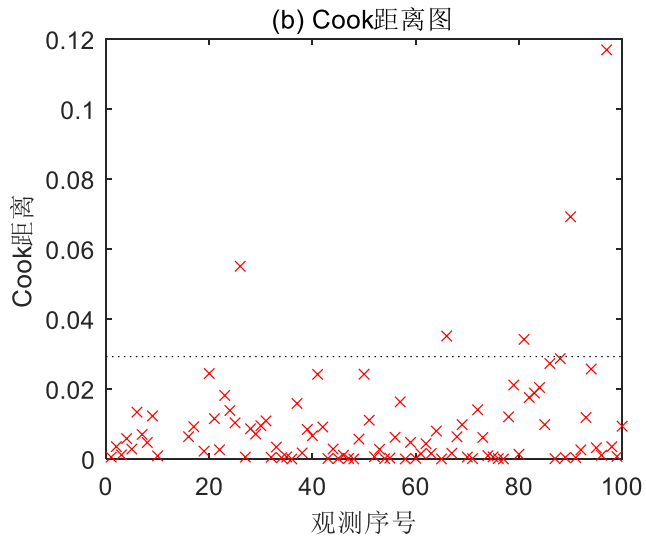
5. 回归诊断

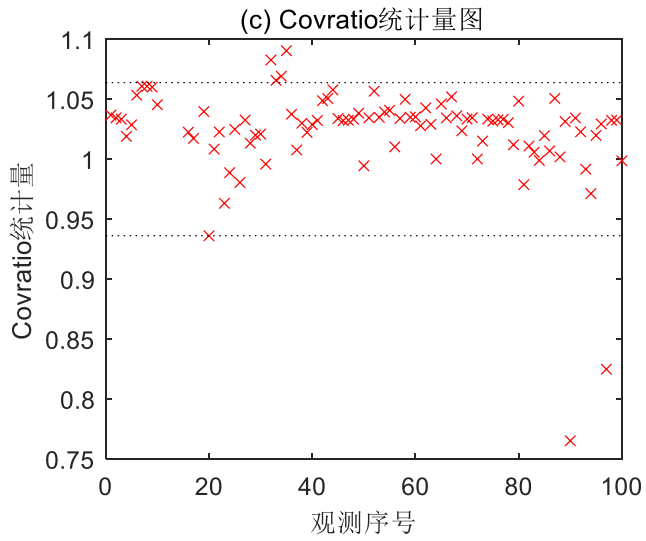
```
Res = mdl2.Residuals;
Res_Stu = Res.Studentized;
Res_Stan = Res.Standardized;
subplot(2,3,1);
plot(Res_Stu,'kx');
refline(0,-2);
refline(0,2);
title('(a) 学生化残差图')
xlabel('观测序号');ylabel('学生化残差');
subplot(2,3,2);
mdl2.plotDiagnostics('cookd');
title('(b) Cook距离图')
xlabel('观测序号');ylabel('Cook距离');
subplot(2,3,3);
mdl2.plotDiagnostics('covratio');
title('(c) Covratio统计量图');
xlabel('观测序号');
ylabel('Covratio统计量');
subplot(2,3,4);
plot(Res_Stan,'kx');
refline(0,-2);
refline(0,2);
title('(d) 标准化残差图');
xlabel('观测序号');
ylabel('标准化残差');
subplot(2,3,5);
mdl2.plotDiagnostics('dffits');
title('(e) Dffits统计量图');
xlabel('观测序号');
ylabel('Dffits统计量');
subplot(2,3,6);
mdl2.plotDiagnostics('leverage');
title('(f) 杠杆值图');
xlabel('观测序号');
ylabel('杠杆值');
```

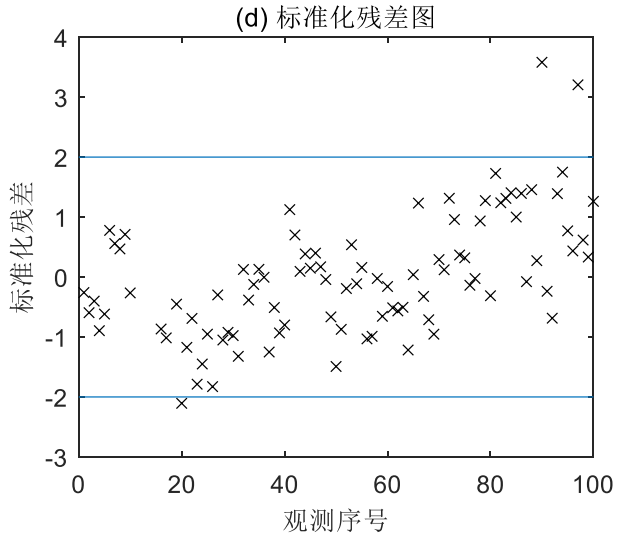
5. 回归诊断

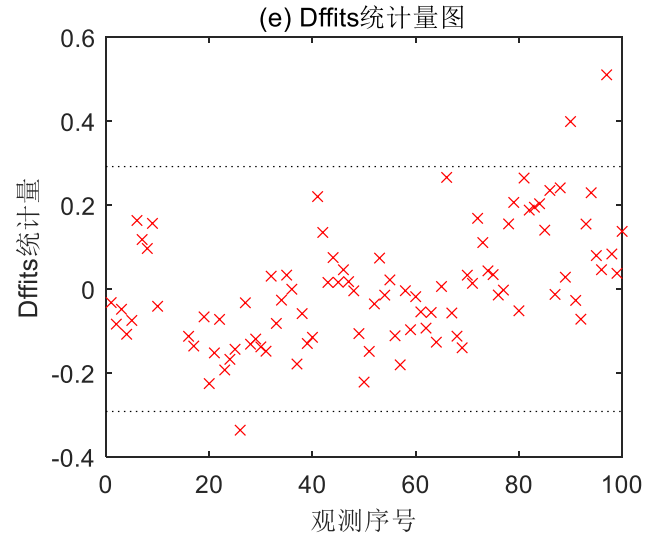
信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

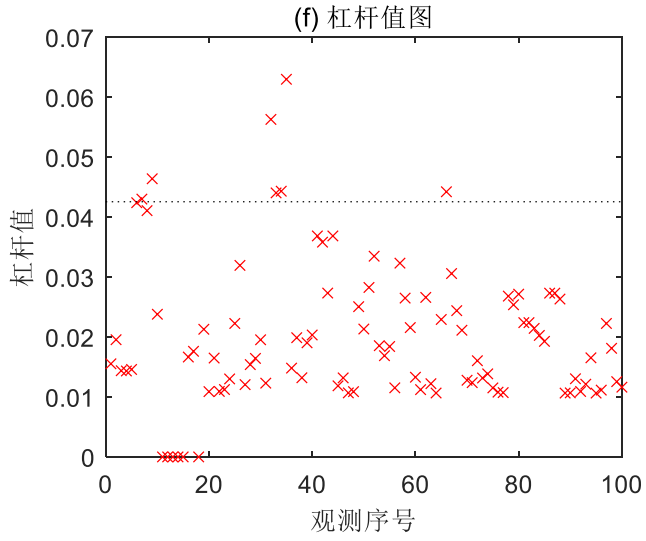








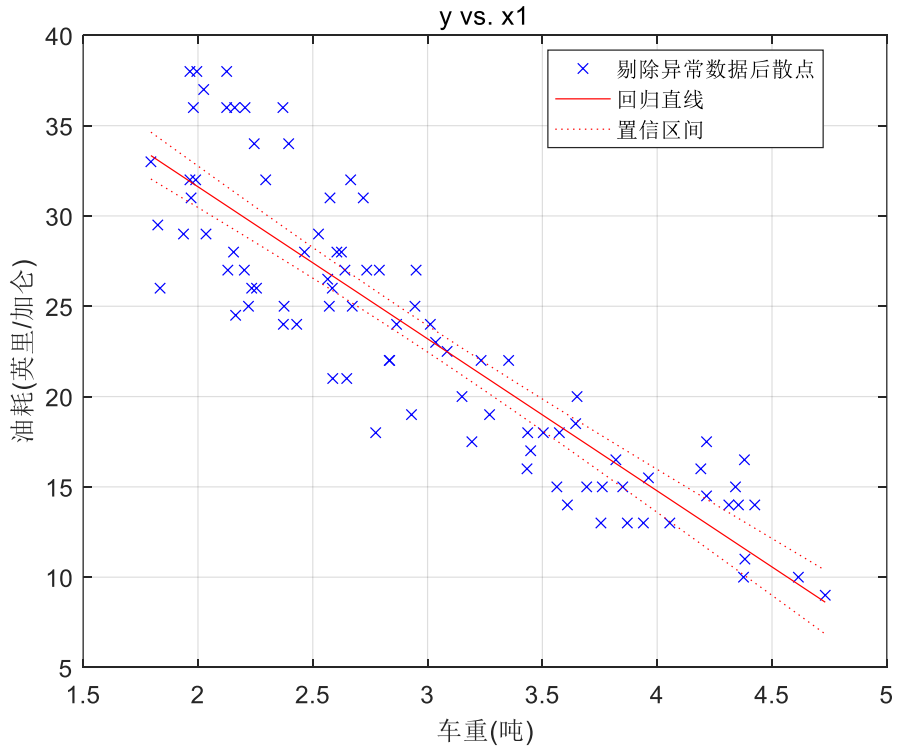




6. 模型改进

```
id = find(abs(Res_Stu)>2);  %查找异常值序号id = 20  90  97
mdl3 = fitlm(x,y, 'Exclude',id)  %删除异常值，重新回归求解
mdl3 =
Linear regression model:
    y ~ 1 + x1
Estimated Coefficients:
                   Estimate      SE       tStat      pValue
                   ________    _______    _____    __________
    (Intercept)     48.455      1.4099    34.369    3.6856e-53
    x1             -8.4195     0.45807   -18.38     4.7501e-32
Number of observations: 91, Error degrees of freedom: 89
Root Mean Squared Error: 3.52
R-squared: 0.791,  Adjusted R-Squared 0.789
F-statistic vs. constant model: 338, p-value = 4.75e-32
```

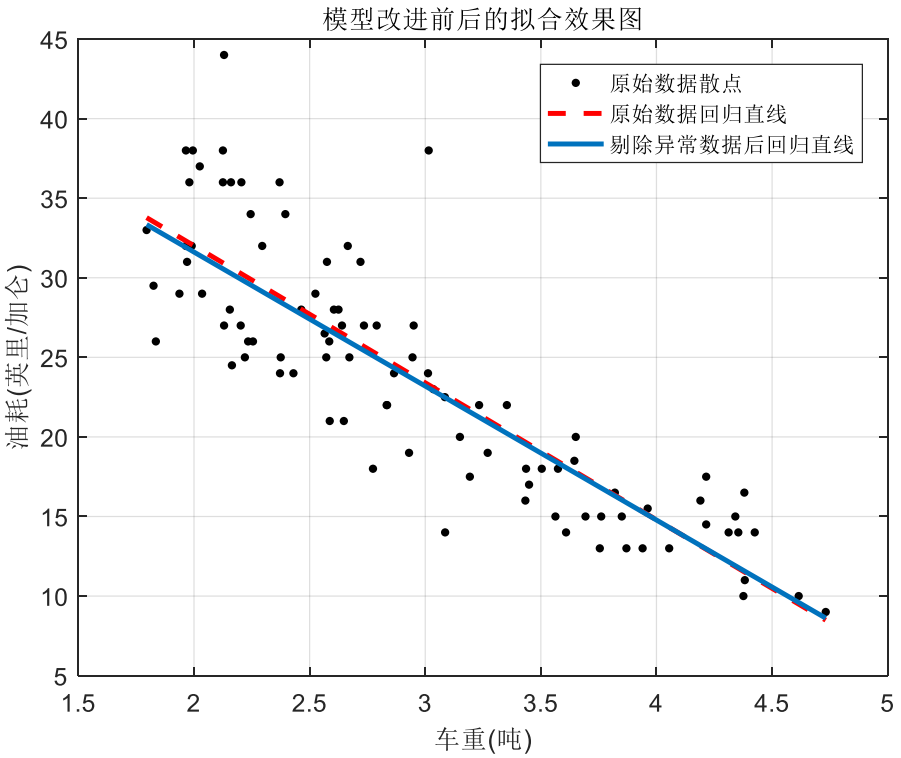
```
>> mdl3.plot;
>> xlabel('车重(吨)');
>> ylabel('油耗(英里/加仑)');
>> legend('剔除异常数据后散点','回归直线','置信区间');
>> grid on
```



6. 模型改进

改进模型后的回归拟合曲线与改进前比较

```
>> plot(x, y, 'k.','MarkerSize',10);   % 画原始数据散点
>> hold on;                             % 图形叠加
>> xnew = sort(x);   % 为了画图的需要将x从小到大排序
>> yhat2 = mdl2.predict(xnew);         % 计算模型2的拟合值
>> yhat3 = mdl3.predict(xnew);         % 计算模型3的拟合值
>> plot(xnew, yhat2, 'r--','linewidth',2);  % 画原始数据对应的回归直线
>> plot(xnew, yhat3, 'linewidth', 2);       % 画剔除异常数据后的回归直线
>> legend('原始数据散点','原始数据回归直线','剔除异常数据后回归直线')
>> xlabel('车重(吨)');
>> ylabel('油耗(英里/加仑)');
>> grid on
>> title('模型改进前后的拟合效果图')
```



7. 残差分析

```
subplot(2,3,1);
mdl3.plotResiduals('caseorder');
title('(a) 残差值序列图');
xlabel('观测序号');
ylabel('残差');
subplot(2,3,2);
mdl3.plotResiduals('fitted');
title('(b) 残差与拟合值图');
xlabel('拟合值');
ylabel('残差');
```

```
subplot(2,3,3);
plot(x,mdl3.Residuals.Raw,'kx');
% refline(0,0)
line([0,25],[0,0],'color','k','linestyle',':');
title('(c) 残差与自变量图');
xlabel('自变量值');
ylabel('残差');
subplot(2,3,4);
mdl3.plotResiduals('histogram');
title('(d) 残差直方图');
xlabel('残差r');
ylabel('f(r)');
```

```
subplot(2,3,5);
mdl3.plotResiduals('probability');
title('(e) 残差正态概率图');
xlabel('残差');
ylabel('概率');
subplot(2,3,6);
mdl3.plotResiduals('lagged');
title('(f) 残差与滞后残差图');
xlabel('滞后残差');
ylabel('残差');
```

7. 残差分析

（a）残差值序列图：各观测对应的残差随机地在水平轴上下无规则地波动，说明残差值间相互独立。如果残差有一定的规律性，则说明残差间不独立。

（b）残差与拟合值图：残差基本分布在上下等宽的水平条带内，说明残差值是等方差的。如果残差分布呈现喇叭口形，则说明残差不满足方差齐性假定，因此应对因变量y作某种变换（如取平方根、取对数、取倒数等），然后重新拟合。

（c）残差与自变量图：残差基本分布在左右等宽的水平条带内，说明线性模型与数据拟合较好。如果残差分布在弯曲的条带内，则说明拟合不好，此时可增加x的非线性项，然后重新拟合。

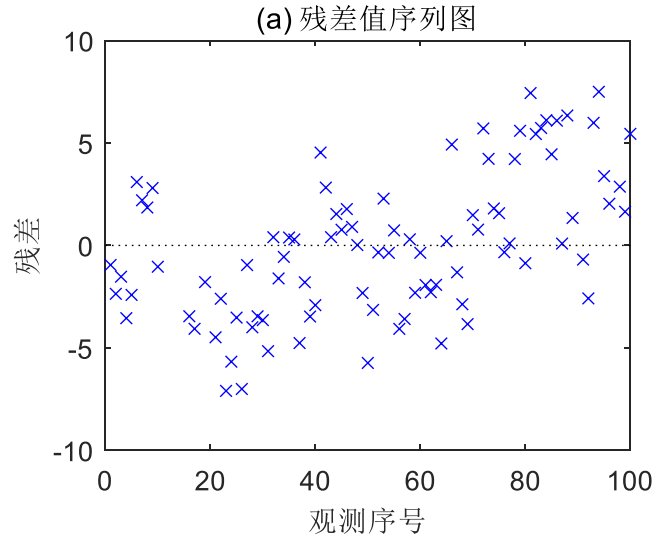
（d）残差直方图：检验残差正态性；

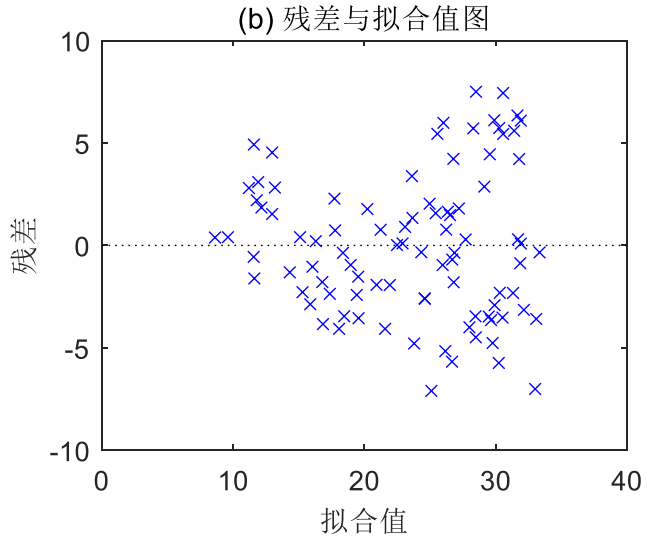
（e）残差正态概率图：检验是否服从正态分布；

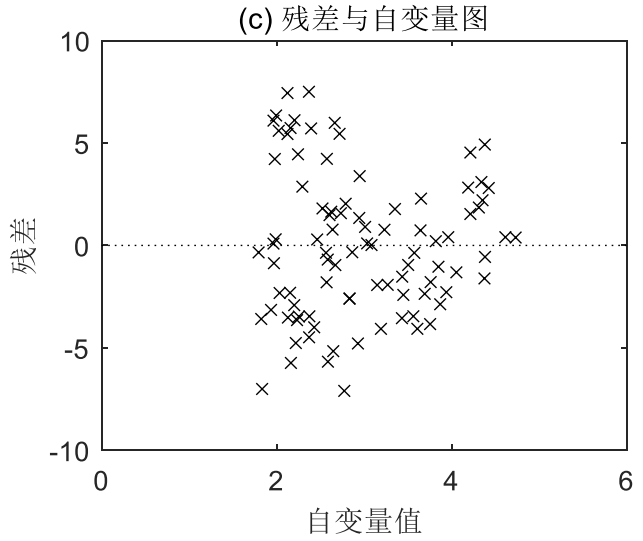
（f）残差与滞后残差图：检验残差间是否存在自相关性。从此图可以看出散点均匀分布在四个象限内，说明残差间不存在自相关性。

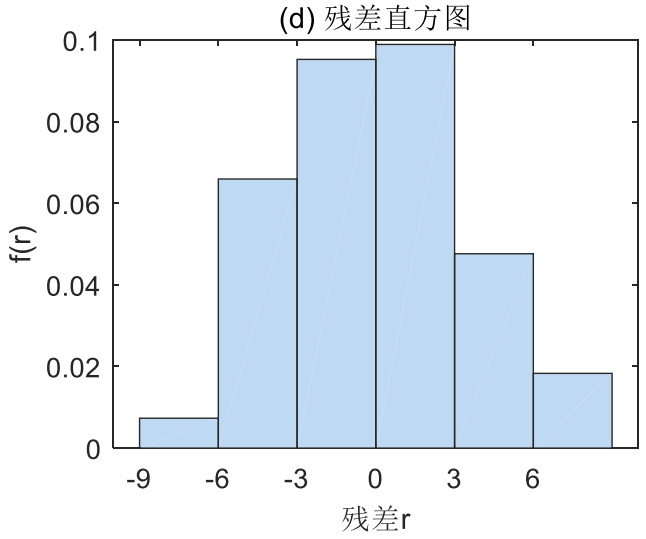
7. 残差分析

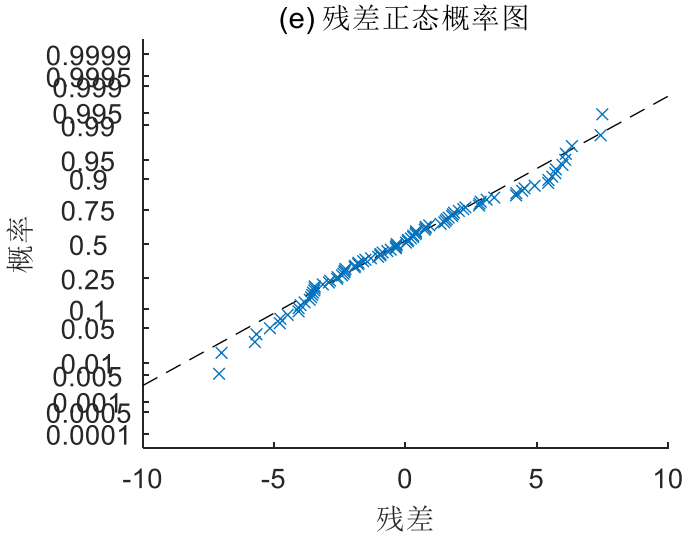
信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

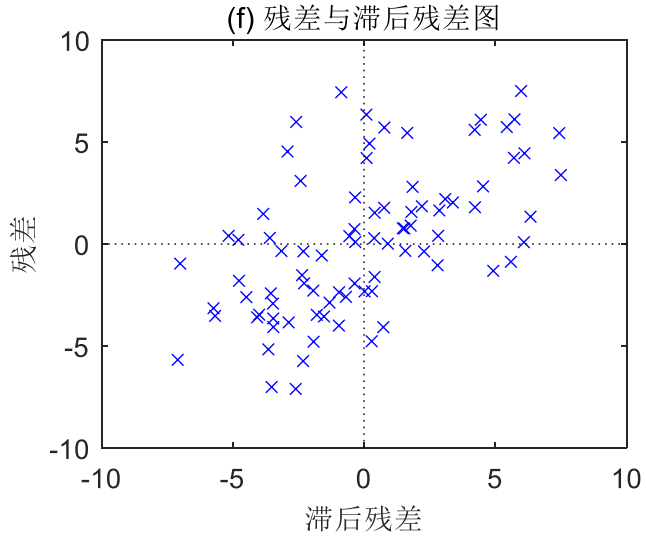












8. 稳健回归

```
>> mdl4 = fitlm(x,y,'RobustOpts','on')
mdl4 =
Linear regression model (robust fit):
    y ~ 1 + x1
Estimated Coefficients:
                   Estimate      SE       tStat       pValue
                   ________    _____    _______    _________
    (Intercept)      48.094     1.5893     30.262    1.3149e-49
    x1              -8.3419     0.5179    -16.107    1.6146e-28
Number of observations: 94, Error degrees of freedom: 92
Root Mean Squared Error: 4
R-squared: 0.739,  Adjusted R-Squared 0.736
F-statistic vs. constant model: 260, p-value = 1.47e-28
```

默认情况下，fitlm函数的'RobustOpts'参数值为'off'，此时采用普通最小二乘法估计模型参数，参数的估计值受异常值影响比较大。

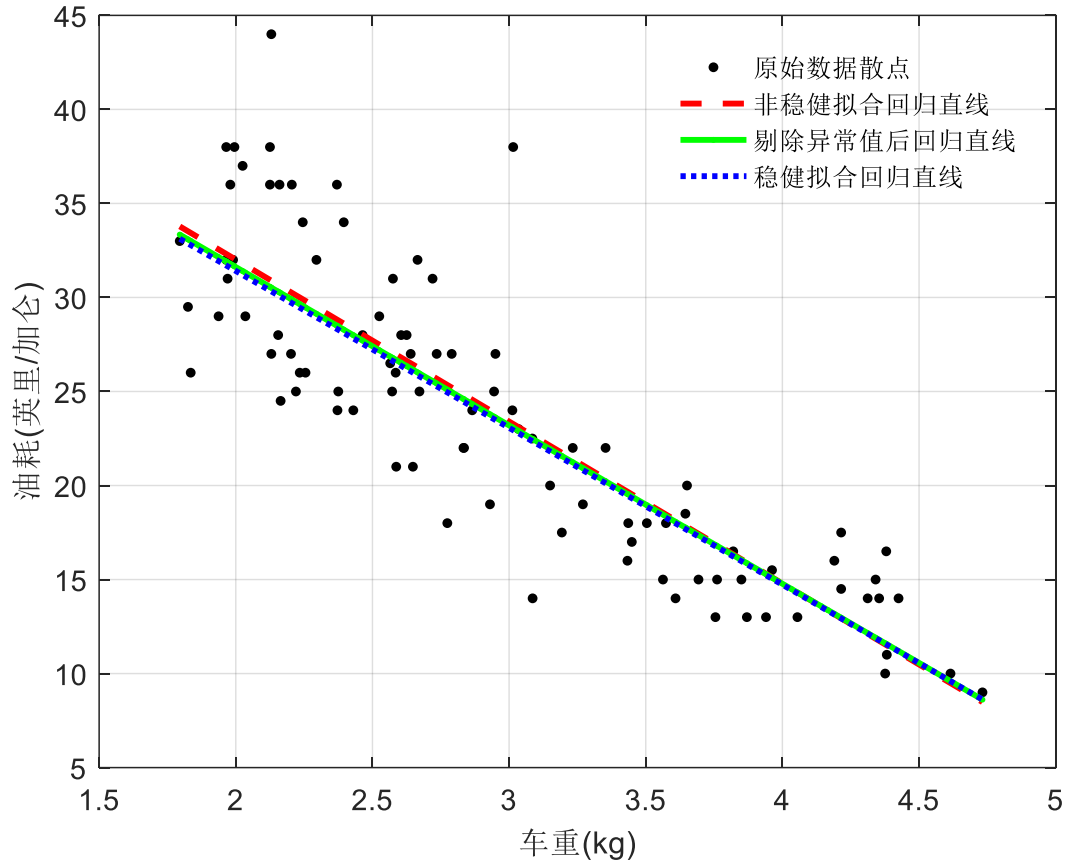
稳健回归，采用加权最小二乘法估计模型参数，结果受异常值影像比较低。

8. 稳健回归

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

绘图对比稳健回归后的拟合直线与非稳健回归直线

```
xnew = sort(x);
yhat2 = mdl2.predict(xnew);
yhat3 = mdl3.predict(xnew);
yhat4 = mdl4.predict(xnew);
plot(x, y, 'k.','MarkerSize',10);    % 画原始数据散点
hold on;
plot(xnew, yhat2, 'r--','linewidth',2);
plot(xnew, yhat3, 'g.-','linewidth',2);
plot(xnew, yhat4, 'b:','linewidth', 2);
legend('原始数据散点','非稳健拟合回归直线','剔除异常值后
回归直线','稳健拟合回归直线');
legend('boxoff')
xlabel('车重(kg)');  ylabel('油耗(英里/加仑)');
grid on
```



二. 回归函数regress

在您只需要函数的输出参数以及要在循环中多次重复拟合模型时，regress 非常有用。如果您需要进一步研究拟合后的回归模型，请使用 fitlm 或 stepwiselm 创建线性回归模型对象 LinearModel。LinearModel 对象提供的功能比 regress 更多。

`b = regress(y,X)` 返回向量 b，其中包含向量 y 中的响应对矩阵 X 中的预测变量的多元线性回归的系数估计值。要计算具有常数项（截距）的模型的系数估计值，请在矩阵 X 中包含一个由 1 构成的列。
`[b,bint] = regress(y,X)` 还返回系数估计值的 95% 置信区间的矩阵 bint。
`[b,bint,r] = regress(y,X)` 还返回由残差组成的向量 r。
`[b,bint,r,rint] = regress(y,X)` 还返回矩阵 rint，其中包含可用于诊断离群值的区间。
`[b,bint,r,rint,stats] = regress(y,X)` 还返回向量 stats，其中包含 R^2 统计量、F 统计量及其 p 值，以及误差方差的估计值。矩阵 X 必须包含一个由 1 组成的列，以便软件正确计算模型统计量。
`[ ___ ] = regress(y,X,alpha)` 使用 100*(1-alpha)% 置信水平来计算 bint 和 rint。您可以指定上述任一语法中的输出参数组合。

1. 案例分析

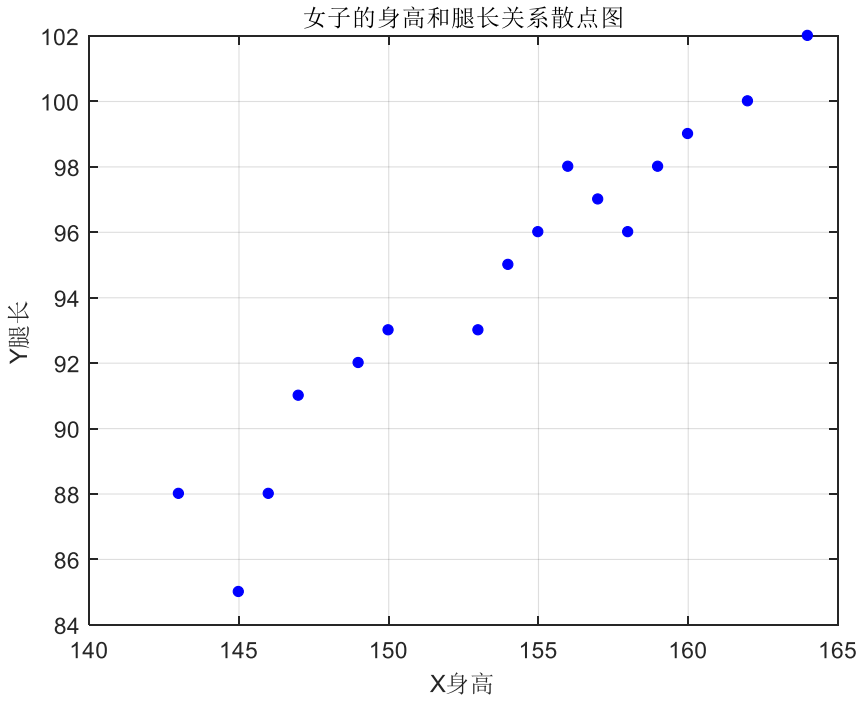
例：测得16名女子的身高和腿长如下(单位:cm):

身高	143	145	146	147	149	150	153	154	155	156	157	158	159	160	162	164
腿长	88	85	88	91	92	93	93	95	96	98	97	96	98	99	100	102

试研究这些数据之间的关系。

% 绘制散点图

```
>> x=[143,145,146,147,149,150,153,154,155,156,157,158,159,160,162,164]';
>> Y=[88,85,88,91,92,93,93,95,96,98,97,96,98,99,100,102]';
>> plot(x,Y,'b.','MarkerSize',15)
>> xlabel('X身高')
>> ylabel('Y腿长')
```



2. 建立回归模型

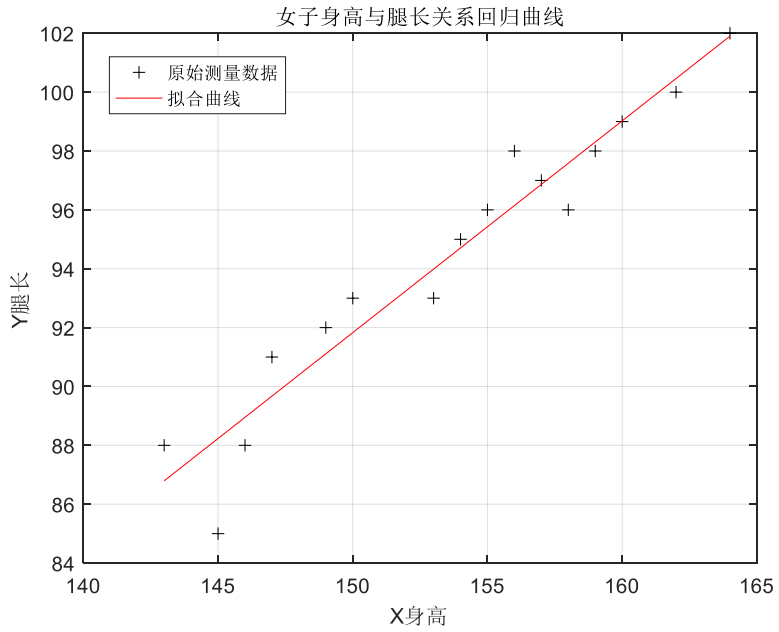
```
>> x=[143,145,146,147,149,150,153,154,155,156,157,158,159,160,162,164]';
>> X=[ones(16,1),x];   %构造系数向量，ones(m,1)为常量系数构造
>> Y=[88,85,88,91,92,93,93,95,96,98,97,96,98,99,100,102]';
>> [b,bint,r,rint,stats]=regress(Y,X)   %回归分析
```

$$\hat{y} = -16.0730 + 0.7194x$$

```
b =
  -16.0730    0.7194
bint =   % 置信区间
  -33.7071    1.5612
    0.6047    0.8340
......
stats =
    0.9282  180.9531    0.0000    1.7437
```

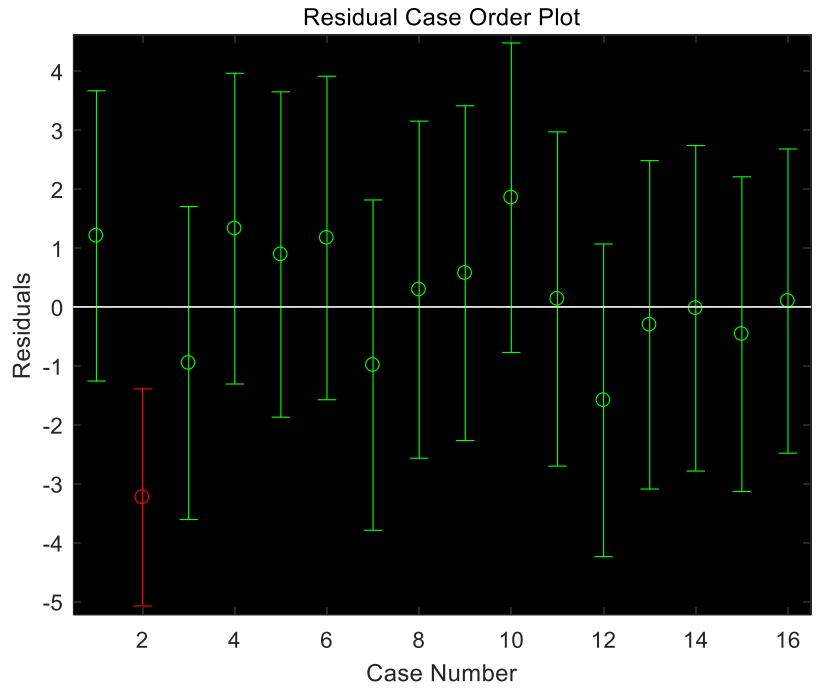
stats第一个数据为模型可决系数，约接近于1说明回归方程越显著；
第二个数据为F统计量的观测值；
第三个数据为概率p，p<alpha，说明显著效果越好；
第四个数据为模型的残差平方和。

```
>> z=b(1)+b(2)*x;
>> plot(x,Y,'k+',x,z,'r')
>> legend('原始测量数据','拟合曲线')
>> grid on
>> xlabel('X身高'); ylabel('Y腿长')
>> title('女子身高与腿长关系回归曲线')
```



3. 残差分析与模型改进

- 残差分析：rcoplot(r,rint)
  - 红色条为异常值
  - 剔除异常值再做回归



从残差图中看出第2个样本存在异常值。

```
>> X(2,:) = []; Y(2) = [];   %删除第2组异常值
>> [b,bint,r,rint,stats]=regress(Y,X)   %回归分析
% b =  -7.2100  0.6633
% stats = 0.9527  261.6389   0.0000   0.8918
>> rcoplot(r,rint)
```


01
IDEA
02
LEARNING
03
CREATIVE
INSPIRATION
EDUCATION
BUSINESS
PHOTO BOOK

感谢聆听