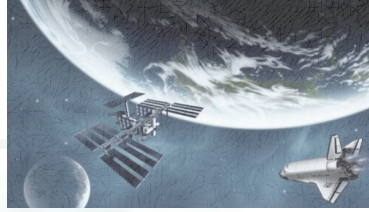


–自动化学院学科核心课–

检测技术与自动化

测量误差与数据处理（5）

本节内容：回归分析

8、回归方程的求法

9、回归方程的检验

回归基本概念

- 回归分析的分类
  一元回归和多元回归
  直线回归和曲线回归

- 本章主要解决以下几方面的问题：

（1） 从一组数据出发，确定这些变量之间的数学表达式——回归方程或经验公式

（2）对回归方程的可信程度进行统计检验

（3）进行因素分析，例如从对共同影响一个变量的许多变量（因素）中，找出哪些是重要因素，哪些是次要因素

回归基本概念

- 函数关系与相关关系

  函数关系：例如，以速度v作匀速运动的物体，走过的距离s与时间t之间 $s = vt$

  相关关系：例如，车床上加工零件误差与零件的直径之间有一定关系，知道零件直径可大致估计其加工误差，但又不能精确地预知加工误差。

- 函数关系和相关关系的对比

  两者可以相互转化

- 相关分析的工具

  回归分析

8. 回归方程的求法 一元线性回归举例

例：确定某段导线的电阻与温度间的关系

温度x	19.1	25.0	30.1	36.0	40.0	46.5	50.0
电阻y	76.30	77.80	79.75	80.80	82.35	83.90	85.10

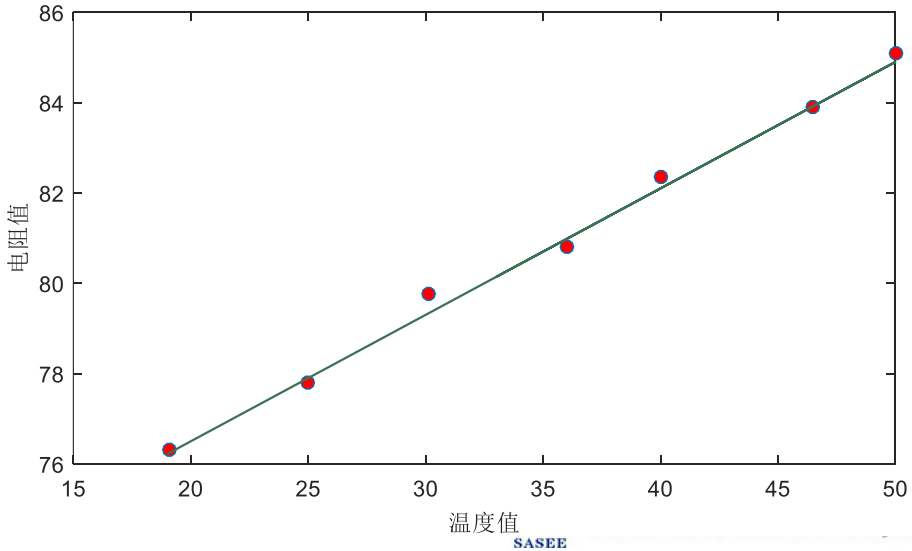
模型： $y = \beta_0 + \beta_1 x + \varepsilon$

求解： $\boldsymbol{A = \left(X^T X\right)^{-1} X^T L}$

建立方程： $L - XA = V$

计算结果：

$\boldsymbol{y = 70.90\Omega + (0.28\,\Omega/^\circ C)x}$



8. 回归方程的求法 一元线性回归举例（续）

例：确定某段导线的电阻与温度间的关系

温度x	19.1	25.0	30.1	36.0	40.0	46.5	50.0
电阻y	76.30	77.80	79.75	80.80	82.35	83.90	85.10

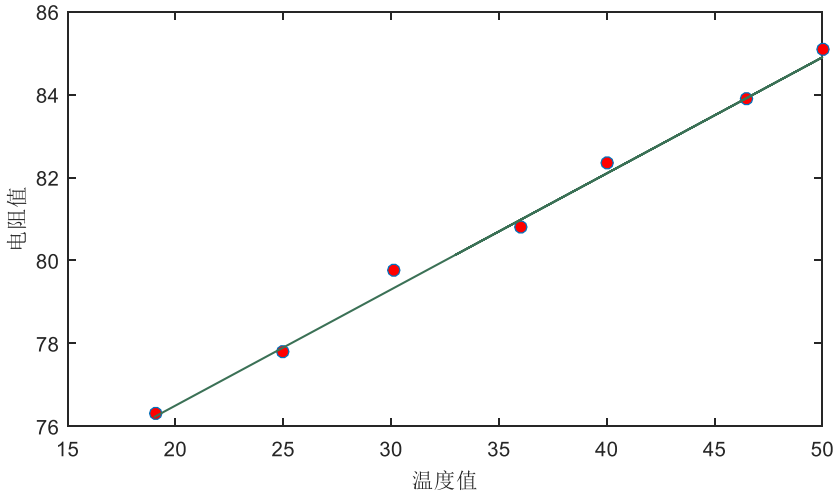
模型： $y = \beta_0 + \beta_1 x + \varepsilon$

对应模型假设：

1）误差项ε是一个期望值为0的随机变量。

2）对于所有的x值，ε的方差σ^2都相同。

3）误差项ε是一个服从正态分布的随机变量，且相互独立。



温度x	19.1	25.0	30.1	36.0	40.0	46.5	50.0
电阻y	76.30	77.80	79.75	80.80	82.35	83.90	85.10

模型： $y = \beta_0 + \beta_1 x + \varepsilon$ $\quad \varepsilon \sim N(0, \sigma^2)$

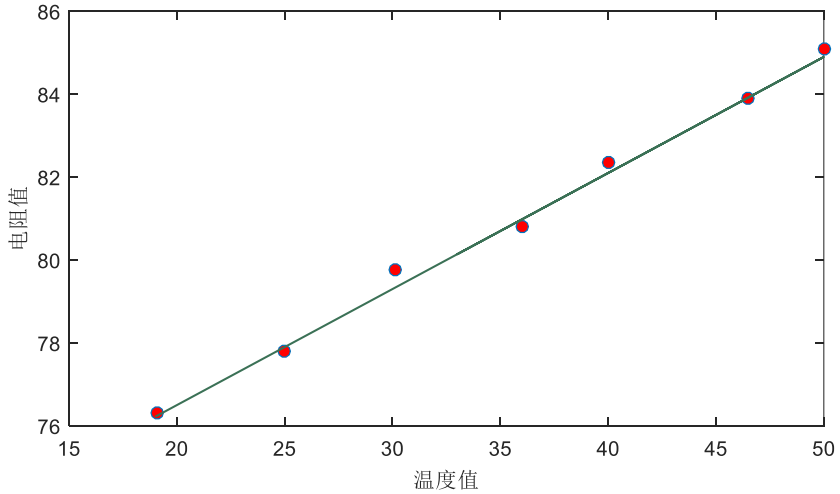
建立方程： $L - XA = V$ 求解： $A = (X^T X)^{-1} X^T L$

计算结果： $y = 70.90\Omega + (0.28\,\Omega/^\circ C)x$

最大似然估计：

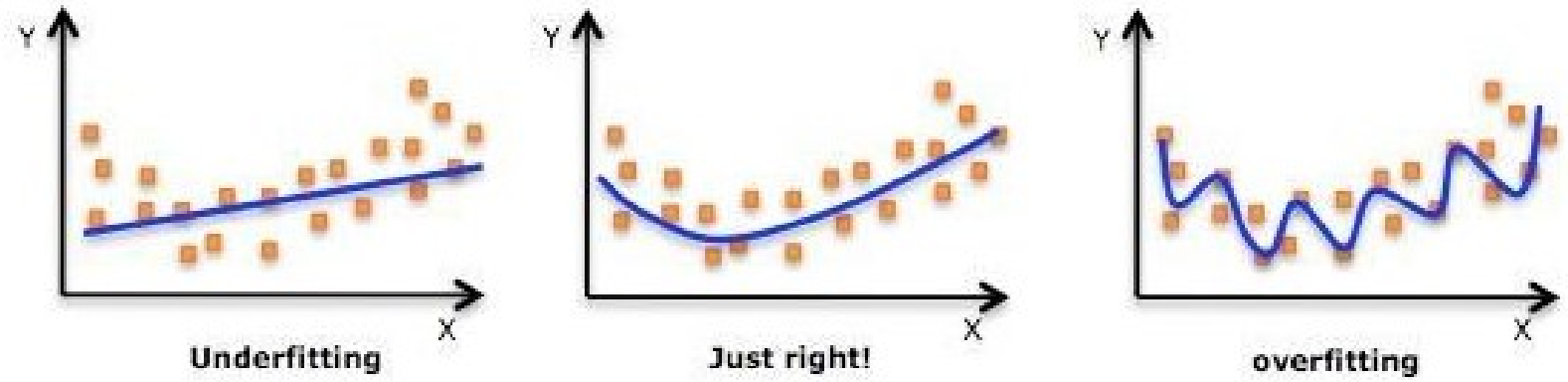
$$\hat{\sigma}^2 = \frac{(L - XA)^T (L - XA)}{n}$$

无偏估计： $\hat{\sigma}^2 = \dfrac{(L - XA)^T (L - XA)}{n - 2}$



推荐阅读：Myung I J. Tutorial on maximum likelihood estimation[J]. Journal of mathematical Psychology, 2003, 47(1): 90-100.

9. 回归方程的检验



（1） 从偏差大小角度

$$S = \sum [y_i - f(x_i)]^2$$

观测值　　拟合值

S 越小越精确

（2）从随机误差角度

不存在过拟合
不存在系统误差

$$y_i = f(x_i) + \varepsilon_i$$

$$E[\varepsilon_i] = 0,$$

$$E[\varepsilon_i \varepsilon_j] = \sigma_i^2 \delta_{ij}.$$

9. 回归方程的检验

（1） 从偏差大小角度

解决办法：方差分析法

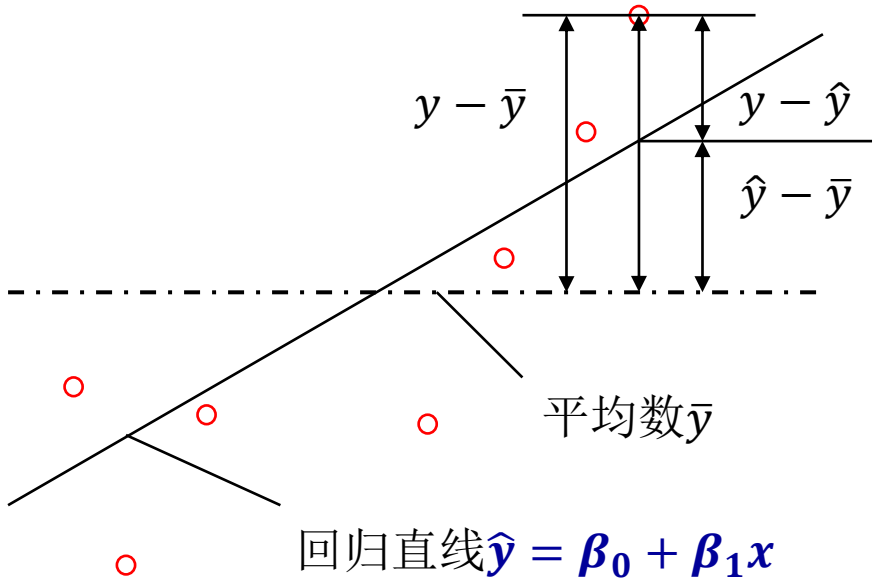
模型：$y = \beta_0 + \beta_1 x + \varepsilon$

总的离差平方和

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$v_s = n-1$$

自由度(degree of freedom, df)指的是计算某一统计量时，取值不受限制的变量个数



可以证明：$SST = SSE + SSR$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$v_U = 1$$

回归平方和

反映y的总变差中由于x和y的线性关系而引起 y 变化的部分

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

$$v_Q = n-2$$

残差平方和

反映所有观测点到回归直线的残余误差，即其它因素对y变差的影响。(除了x对y的)

9. 回归方程的检验

基本思路：方程是否显著取决于SSR和SSE的大小， SSR越大SSE越小，说明y与x的线性关系愈密切。

步骤（1）提出假设：线性关系不显著， $\beta_1 = 0$

步骤（2）计算统计量F

$$F = \frac{SSR/v_U}{SSE/v_Q}$$

对一元线性回归，应为 $F = \frac{SSR/1}{SSE/(n-2)} \sim F_\alpha(1, n-2)$

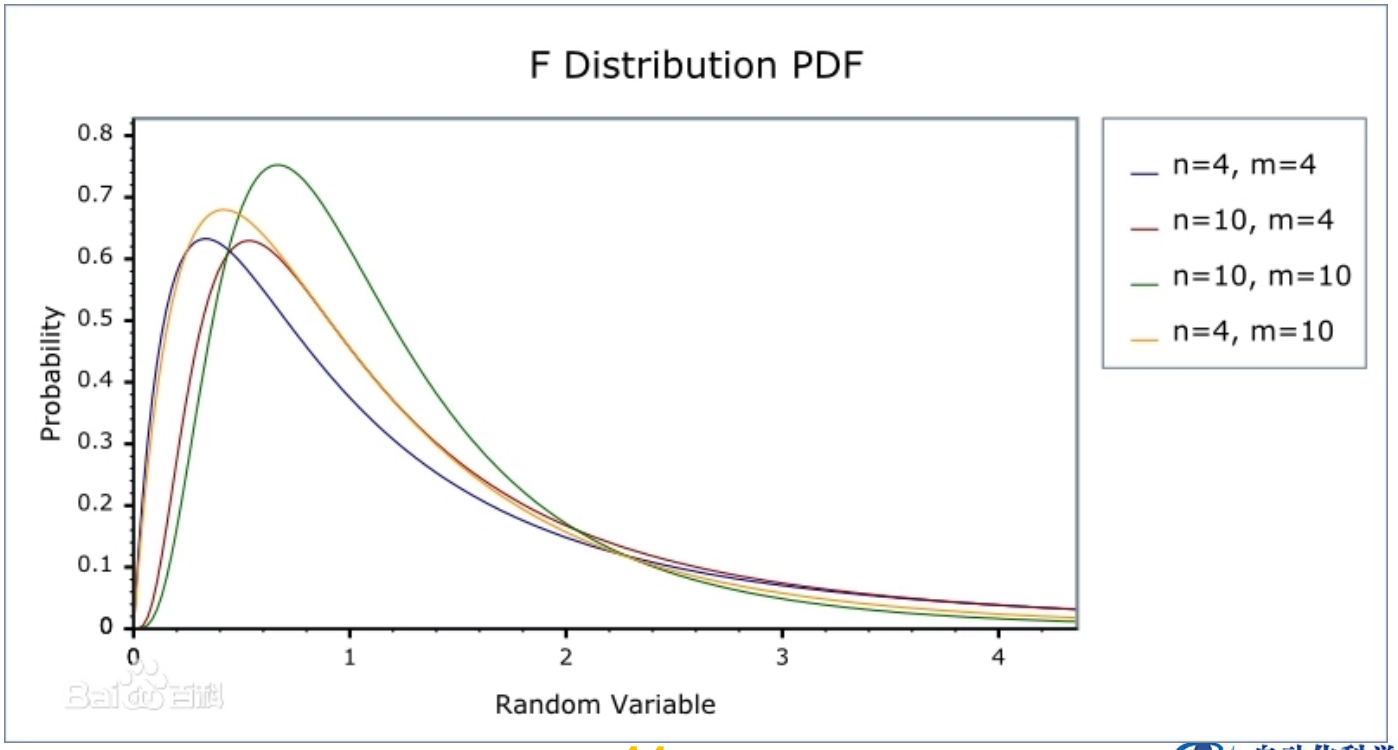
步骤（3）查F分布表，根据给定的显著性水平 和已知的自由度1和n -2进行检验：

F分布定义

若总体$X\sim N(0,1)$，$(X_1,X_2,\ldots,X_{n_1})$和$(Y_1,Y_2,\ldots,Y_{n_2})$为来自X的两个独立样本，设统计量

$$F=\frac{\sum_{i=1}^{n_1}X_i^2}{n_1}\Big/\frac{\sum_{i=1}^{n_2}Y_i^2}{n_2}$$

则称统计量F服从自由度n_1和n_2的F分布，记为$F\sim F(n_1,n_2)$



一元回归分析方差分析表

来源	平方和（sum of squares）	自由度（df）	均方差（mean square）	F	显著性（Sig.）
回归（Regression）	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	1	$SSR/1$	$\frac{SSR/1}{SSE/(n-2)}$	
残差（Residual）	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	n-2	$SSE/(n-2)$		
总计（Total）		n-1			

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10953.203	2	5476.601	12688.741	.000[b]
	Residual	3.021	7	.432		
	Total	10956.224	9			

a. Dependent Variable: y

b. Predictors: (Constant), x2, x1

引申：多元回归分析方差分析表

$$f(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$SST = SSR + SSE \quad \text{其中}$$

$$\mathrm{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$v_U = p \qquad v_Q = n - 1 - p$$

原假设是?

来源	平方和（sum of squares）	自由度（df）	均方差（mean square）	F	显著性（Sig.）
回归（Regression）	$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	p	SSR/p	$\dfrac{SSR/p}{SSE/(n-p-1)}$	Pr(F>检验统计量F值）=P值
残差（Residual）	$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	n-p-1	$SSE/(n-p-1)$		
总计（Total）		n-1			

9. 回归方程的检验

一元线性回归举例（续）

温度x	19.1	25.0	30.1	36.0	40.0	46.5	50.0
电阻y	76.30	77.80	79.75	80.80	82.35	83.90	85.10
电阻估计值	76.25	77.90	79.33	80.98	82.10	83.92	84.90
残差	0.05	-0.1	0.422	-0.18	0.25	-0.02	0.2

$$\boldsymbol{y = 70.90\Omega + (0.28\,\Omega/^\circ C)x}$$

$$\hat{\sigma}^2 = \frac{(L - XA)^T(L - XA)}{n - 2} = 0.0652$$

$$\mathbf{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \mathbf{80.86})^2$$

$$= (-4.61)^2 + (-3.06)^2 + (-1.11)^2 + (-0.06)^2 + (1.49)^2 + (3.04)^2 + (4.24)^2 = 61.2907$$

$$\boldsymbol{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{0.3259}$$

$$F = \frac{SSR/v_U}{SSE/v_Q} = \frac{61.2907/1}{0.3259/5} = 940.33$$

α=0.01

Fα k1 k2	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
2	98.49	99.01	99.17	99.25	99.30	99.33	99.36	99.42	99.46	99.50
3	34.12	30.81	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65

9. 回归方程的检验

多元回归方程显著性检验

- t检验

假设：$\beta_j = 0$，$j = 1, 2, \ldots, p$

记 $(X'X)^{-1} = (c_{ij})$，$i, j = 0, 1, 2, \ldots, p$

$$\mathrm{E}(\hat{\beta}_j) = \beta_j, \quad \mathrm{var}(\hat{\beta}_j) = c_{jj}\sigma^2$$

构造t统计量：

自由度是?

$$t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}}$$

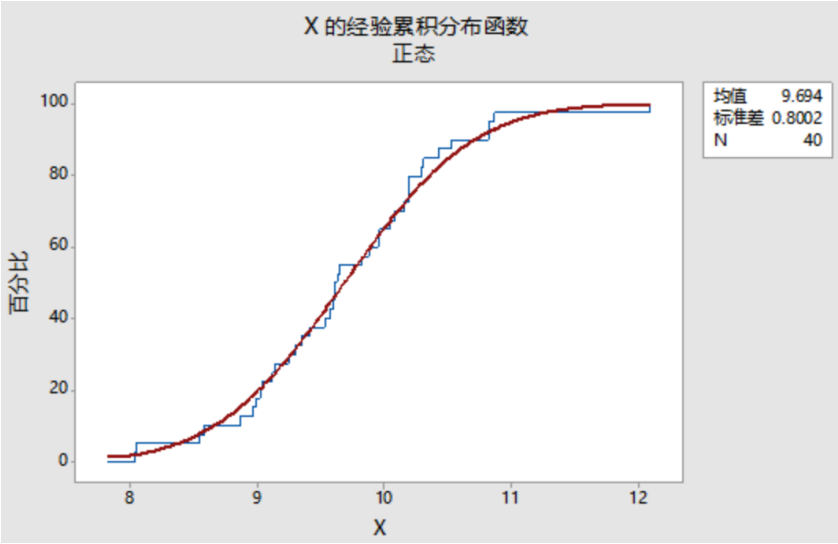
其中，

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

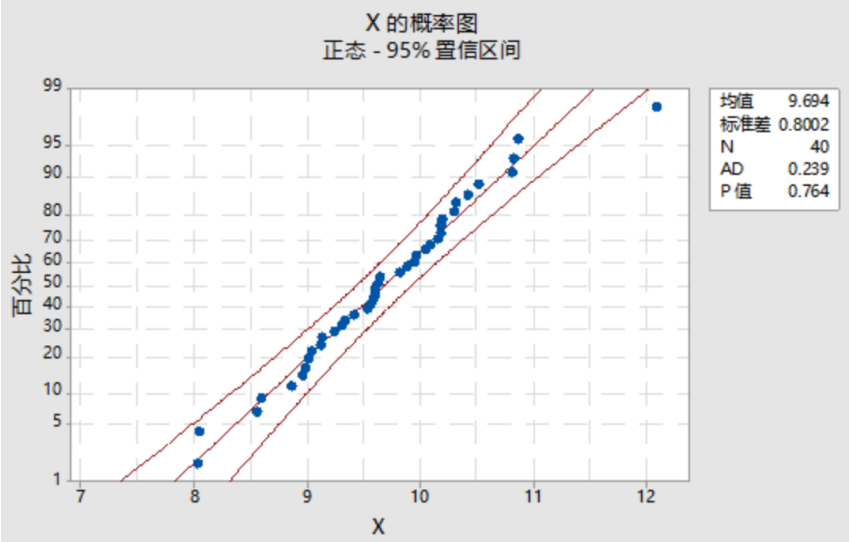
9. 回归方程的检验

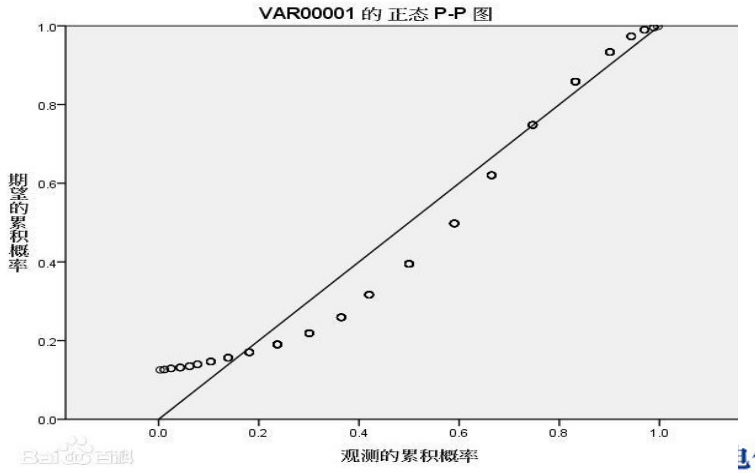
② 从随机误差角度：判断偏差是否是随机误差

观察法：



累积分布函数





P-P图、Q-Q图，…

9. 回归方程的检验

统计学中验证分布的方法

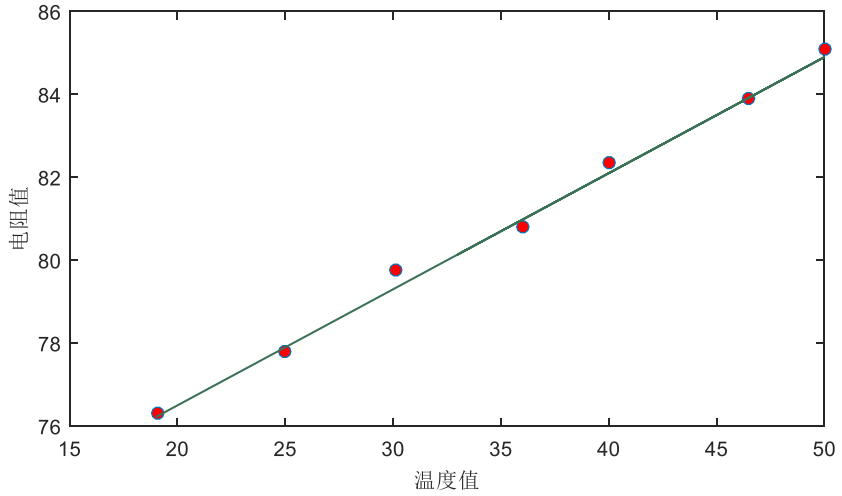
拟合度检验
goodness-of-fit

- 卡方检验
- Kolmogorov-Smirnov D检验，简称K-S检验
- Lilliefors检验
- Anderson – Darling AD检验
- Shapiro – Wilk W检验
- Ryan-Joiner检验

9. 回归方程的检验 一元线性回归举例（续）

温度x	19.1	25.0	30.1	36.0	40.0	46.5	50.0
电阻y	76.30	77.80	79.75	80.80	82.35	83.90	85.10
电阻估计值	76.25	77.90	79.33	80.98	82.10	83.92	84.90
残差	0.05	-0.1	0.422	-0.18	0.25	-0.02	0.2

判断拟合是否合适，看误差分布：
（欠拟合、**过拟合**）
（1）画误差图
（2）误差分布检验



扩展：回归方程的形式

- 关于一个自变量的线性回归方程

$$f(x|a,b)=ax^2+bx+c$$

$$f(x|a,b,c)=a+bexp(-k_1x)+cexp(-k_2x),k_1、k_2已知$$

$$f(x|a,b,c)=ax+b/x+c$$

- 关于多个自变量的线性回归方程

$$f(x,y,z|a,b,c,d)=ax+by+cz+d$$

- 回归方程非线性，但可以线性化

$$\boldsymbol{f(t|a,b) = aexp(-kt)}$$

$$\boldsymbol{\ln f(t|a,b) = -kt + \ln a}$$

应该注意什么？

$$\sigma_{\ln y} = \left|\frac{d\ln y}{dy}\right| \sigma_y = \sigma_y / y_i$$

$$y = \alpha e^{\beta x} \Rightarrow \ln y = \ln\alpha + \beta x$$
$$\Rightarrow y' = \ln\alpha + \beta x$$
$$y = \alpha x^{\beta} \Rightarrow \ln y = \ln a + \beta \ln x$$
$$\Rightarrow y' = \ln\alpha + \beta x'$$
$$y = \frac{x}{\alpha x + \beta} \Rightarrow y' = \alpha + \beta x'$$
$$y = \alpha + \beta \log x \Rightarrow y = \alpha + \beta x'$$
$$y = \frac{1}{\alpha + \beta e^{-x}} \Rightarrow y' = \alpha + \beta x'$$

- 回归方程非线性，且不可以线性化

$$\boldsymbol{f(t|a,b) = aexp(-kt) + b}$$

Nonlinear least-squares fitting procedure

扩展：典型回归技术

1）Linear Regression 线性回归

2）Logistic Regression 逻辑回归

3）Polynomial Regression 多项式回归

4）Stepwise Regression 逐步回归

5）Ridge Regression 岭回归

6）Lasso Regression 套索回归

7）ElasticNet回归

扩展：典型回归技术

变量存在多重共线性问题：

5）Ridge Regression岭回归 $\hat{\beta} = \underset{\beta \in \mathbf{R}^p}{\operatorname{argmin}} \underbrace{\|y - \mathbf{X}\beta\|_2^2}_{Loss} + \lambda \underbrace{\|\beta\|_2^2}_{Penalty}$

6）Lasso Regression套索回归 $\hat{\beta} = \underset{\beta \in \mathbf{R}^p}{\operatorname{argmin}} \underbrace{\|y - \mathbf{X}\beta\|_2^2}_{Loss} + \lambda \underbrace{\|\beta\|_1}_{Penalty}$

系数可能为0，可帮助特征选择

7）ElasticNet回归 综合上述两种回归的优点

$$\hat{\beta} = \underset{\beta \in \mathbf{R}^p}{\operatorname{argmin}} \underbrace{\|y - \mathbf{X}\beta\|_2^2}_{Loss} + \lambda_1 \underbrace{\|\beta\|_2^2}_{Penalty} + \lambda_2 \underbrace{\|\beta\|_1}_{Penalty}$$

谢谢

Q & A