```
[p h] = signtest(u, 3, 0.05)
p = 0.0026                    h = 1
[p h] = signtest(u, 3, 0.01)
p = 0.0026                    h = 1
```

其它符号检验法如分布的检验,游程检验等可参阅有关文献编写检验函数.

第四节　方差分析与回归分析

方差分析与线性回归分析同属线性模型,是数理统计的重要组成部分.

一、方差分析

方差分析法因其基于统计数据的总变动成分与随机误差成分的比较而得名,是基于不太多的统计数据,定量地分析一个或多个因素对某个(些)响应变量的影响和作用的显著性,这种显著性是基于一定概率条件下而言的,其前提条件是在各因素的作用下,响应变量的分布具有正态性和等方差性.

1. 单因素方差分析

单因素方差分析的基本问题是比较和估计多个等方差正态总体的均值,其基本模型为

$$y_{ij} = \alpha_j + \varepsilon_{ij},$$

其中 y_{ij} 为数据观测值,α_j 是各正态总体的数学期望,ε_{ij} 为各数据的随机误差.在MATLAB中提供了单因素方差分析函数 anoval( ), 其使用格式为

```
p = anoval(X)
```

或

```
p = anoval(X,g)
```

该函数返回无效假设成立的概率,第一种格式中 $\boldsymbol{X}$ 为一矩阵,函数将矩阵的每一列当作一个总体,矩阵的行数即为样本重复数.若函数返回的概率值接近于零,则无效假设值得怀疑,表明各列的均值事实上是不同的.第二种格式中的 $\boldsymbol{X}$ 为一向量,$\boldsymbol{g}$ 是与 $\boldsymbol{X}$ 同长度的向量,且 $\boldsymbol{g}$ 中的值为整数,最小值为 1,最大值为数据的组数,每一组至少有一个数,但并不要求每组中元素个数相同.因此,第一种格式用于等重复的单因素方差分析,第二种格式用于重复数不等的单因素方差分析.方差分析同时还显示一个图与一张表,表即为方差分析表,它与一般教科书所列方差分析表一般无二;而图则给出了各列数据的 box 图,这种图为一"盒子"形状,因而取名 box 图,其特征为:① 盒子的上底与下底间为内四分位间距;盒子的上、下两条线分别为样本的 25% 和 75% 分位数.② 盒子的中间线为样本的中位数,如果中位数不在盒子中间,表明样本存在一定偏度.③ 虚线贯穿盒子上下,显示了样本其余部分,如果没有奇异值,则样本的最大值为虚线顶点,

最小值为虚线底端.默认奇异值为距盒子底端和顶端超过 1.5 倍内四分位间距的点.在图中,奇异值为超出虚线底端的点,用"+"表示一个奇异值.④ 切口是样本中位数的置信区间,可用 box 图对样本进行多重比较.

例 1 考察四种不同药剂处理的水稻种子对苗高的影响,所得试验数据见矩阵 $\boldsymbol{X}$,各重复四次,$\boldsymbol{X}$ 的每一列为一个处理,则有(图 5-11 为 box 图)

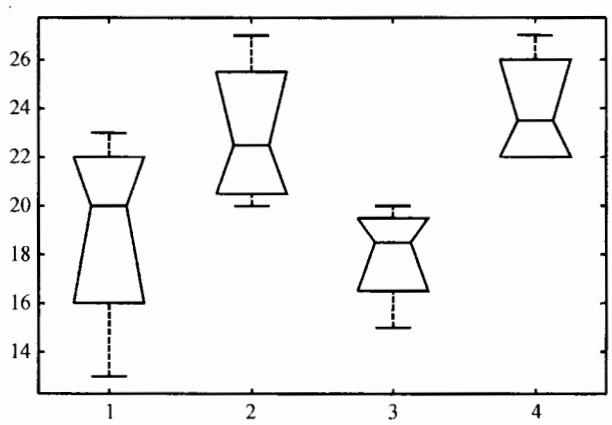

图 5-11 不同药剂处理的水稻种子对苗高影响的 box 图

```
x = 23  21  20  22
    19  24  18  25
    21  27  19  27
    13  20  15  22
p = anoval(x)
p = 0.0487
```

ANOVA Table

Source	SS	df	MS	F
Columns	104	3	34.67	3.525
Error	118	12	9.833	
Total	222	15		

由于返回的概率值 $p=0.0487<0.05$,但 $p>0.01$,故不同药剂对种子的处理的差异达到了显著水平,但并不是极显著.

例 2 将三种不同菌型的伤寒病毒分别接种于 10 只、9 只和 11 只小白鼠上,观察存活天数,结果用下述 x 和 g 表示,x 为存活天数,g 为组标识,表明数据 x 属于哪一组的值.结果如下(图 5-12 为方差分析的 box 图)

```
x = [2 4 3 2 4 7 7 2 5 4 5 6 8 5 10 7 12 6 6 7 11 6 6 7 9 5 10 6 3 10];
```

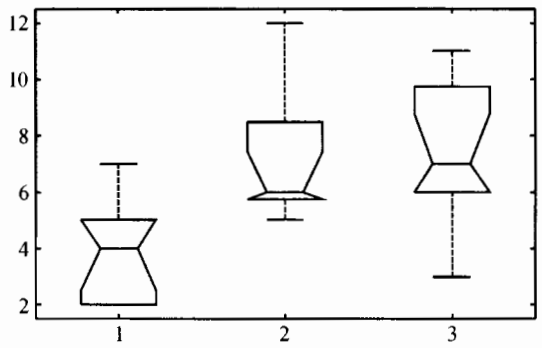

图 5－12　不同菌型伤寒病毒对小白鼠存活天数的影响

```
g = [1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3];
p = anova1(x, g)
p = 0.0038
```

ANOVA Table

Source	SS	df	MS	F
Columns	70.43	2	35.21	6.903
Error	137.7	27	5.101	
Total	208.2	29		

由方差分析结果 $p=0.003\,8<0.01$ 知，不同菌型的伤寒病毒对小白鼠存活天数的影响差异极显著.

2. 双因素方差分析

双因素方差分析是一种两因素多水平析因试验数据的统计分析方法，目的在于确定来自不同组的数据是否具有相同的均值. 其基本模型为

$$y_{ijk}=\mu+\alpha_j+\beta_i+\gamma_{ij}+\varepsilon_{ijk},$$

其中 y_{ijk} 为数据观测值，μ 为样本总均值，α_j 为因素 A 各列的均值，β_i 为因素 B 各行的均值，γ_{ij} 是因素 A 与因素 B 的交互作用(当没有重复时，该项并入随机误差)，ε_{ijk} 为观测值的随机误差. MATLAB 提供有等重复试验的双因素方差分析函数 anova2( )，其使用格式为

```
p = anova2(X, r)
```

其中 $\boldsymbol{X}$ 为数据观测值，因素 A 按列放. 因素 B 按行放，即矩阵 $\boldsymbol{X}$ 的列为因素 A 的水平. 当重复数为 1 时，矩阵 $\boldsymbol{X}$ 的行为因素 B 的水平；当重复数大于 1 时，前面 r 行为因素 B 的第一水平，$r+1$ 行到 $2r$ 行为因素 B 的第二水平，依此类推，r 为重复数. 矩阵 $\boldsymbol{X}$ 的行数应为 B 的水平数乘以重复数 r. 该函数返回一个向量 $\boldsymbol{p}$，当 $r=1$ 时，$\boldsymbol{p}$ 中有两个元素，第一个为因素 A 各水平均值相等的概率，第二

个为因素 B 各水平均值相等的概率;当 $r>1$ 时,p 中有三个元素,第三个元素为 A 与 B 交互作用无显著作用的概率.该函数除返回概率值外,还显示一个方差分析表.

例 3 三名学生对四个品种的稻米含氮量(mg) 各作了一次分析,数据如下述 X,每行为一个学生,每列为一品种稻米,作方差分析有

```
x = 2.2000   2.3000   2.6000   2.7000
    2.2000   2.0000   2.5000   2.7000
    2.0000   2.3000   2.7000   2.8000
p = anova2(x, 1)
p = 0.0021    0.4472
```

ANOVA Table

Source	SS	df	MS	F
Columns	0.7833	3	0.2611	18.08
Rows	0.02667	2	0.01333	0.9231
Error	0.08667	6	0.01444	
Total	0.8967	11		

由上述结论知,A 因素(列)均值相等的概率为 $p=0.002\ 1<0.01$,故差异极显著;而 B 因素(行)均值相等的概率为 $0.447\ 2>0.05$,因而差异不显著.

例 4 三种肥料(因素 A)施于三种土壤(因素 B),设三次重复,得小麦产量见下述数据 Y,A 的水平为列,B 的水平为行,则有:

```
y = 21.4000   12.0000   12.8000
    21.2000   14.2000   13.8000
    20.1000   12.1000   13.7000
    19.6000   13.0000   14.2000
    18.8000   13.7000   13.6000
    16.4000   12.0000   13.7000
    17.6000   13.3000   12.0000
    16.6000   14.0000   14.6000
    17.5000   13.9000   14.0000
p = anova2(y, 3)
p = 0.0000   0.1542   0.0052
```

ANOVA Table

Source	SS	df	MS	F
Columns	178.1	2	89.05	97.23
Rows	3.807	2	1.903	2.078

Interaction	19.55	4	4.887	5.335
Error	16.49	18	0.9159	
Total	217.9	26		

由上述结论知,因素 A 即不同肥料间均值相等的概率 $p=0.000\ 0$, 因而差异极显著;因素 B 即不同土壤间均值相等的概率为 $0.154\ 2>0.05$, 故差异不显著;交互效应作用不显著的概率为 $0.005\ 2<0.01$, 故差异极显著.

对于不等重复的双因素方差分析,可参阅有关文献编写方差分析函数.

二、回归分析

回归分析是研究变量之间相关关系的一种统计方法,即利用统计数据来寻求变量间关系的近似表达式(经验公式),并利用所得公式进行统计描述、分析和推断,以解决预测、优化和控制问题.这里仅就线性回归进行实验.

线性回归包括一元线性回归和多元线性回归,这种回归关于未知参数和回归变量都是线性关系,其变量关系为

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m+\varepsilon,$$

其中 $\beta_i(i=0,1,2,\cdots,m)$ 是回归系数,ε 为随机误差,且有 $E(\varepsilon)=0, D(\varepsilon)=\sigma^2$;当 $m=1$ 时为一元线性回归,$m>1$ 时称为多元线性回归. MATLAB 中提供了一个多元线性回归函数 regress( ),其使用格式为

```
[b,bint,r,rint,stats] = regress(y,x,a)
```

其中 $\boldsymbol{Y}$ 为观测得到的随机变量,$\boldsymbol{X}$ 为自变量矩阵.若回归关系中包括常数项,则 $\boldsymbol{X}$ 的第一列应全部为 1,$\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的行数应相等,$\boldsymbol{X}$ 的列数等于参数的个数;当 $\boldsymbol{X}$ 为两列且第一列全为 1 时为一元线性回归. a 为输出各种置信区间用的显著水平.输出结果中有五个项:b 为参数的点估计;$bint$ 为参数的区间估计;r 为残差的点估计;$rint$ 为残差的区间估计,当点估计落在区间估计之外时,拒绝无效假设;$stats$ 中包含了三个项,第一个是回归方程的决定系数 R^2,第二个是回归方程的 F 统计量,第三个是拒绝无效假设的概率.

例 5　对某地区生产同一产品的 8 个不同规模乡镇企业进行生产费用调查,得产量 $\boldsymbol{X}$ 和生产费用 $\boldsymbol{Y}$ 的数据如下述 x' 和 y',进行回归分析(图 5-13)有

```
y' = 5.6  6.6  7.2  7.8  10.1  10.8  13.5  16.5
x' = 1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0
     1.5  2.0  3.0  4.5  7.5  9.1  10.5  12.0
[b bint r rint s] = regress(y, x, 0.05)
[b bint r rint s] = regress(y, x, 0.01)
b = 4.1575    0.8950
bint = 2.4692    5.8458        bint = 1.5994    6.7156
```

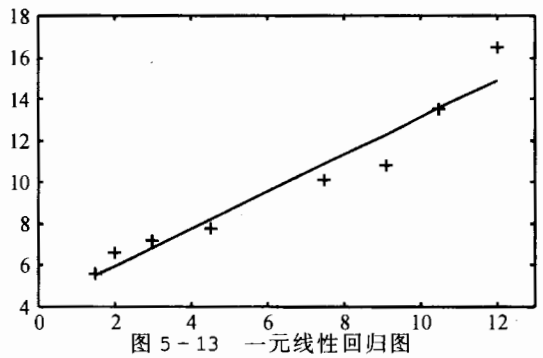

图 5-13 一元线性回归图

```
          0.6644       1.1256                  0.5456       1.2444
    r = 0.1000   0.6525   0.3575   -0.3850   -0.7701   -1.5021
-0.0551   1.6024
    rint = -2.1283    2.3283          rint = -3.2762    3.4762
           -1.5282    2.8332                 -2.6516    3.9566
           -2.0074    2.7223                 -3.2256    3.9406
           -2.8445    2.0744                 -4.1115    3.3414
           -3.1401    1.5999                 -4.3610    2.8208
           -3.2935    0.2893                 -4.2163    1.2121
           -2.3519    2.2417                 -3.5351    3.4249
           0.4846     2.7201                 -0.0912    3.2959
    s = 0.9376    90.1871    0.0001
```

由上述回归结果知，$\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的决定系数达到了 $R^2=0.9376$，回归系数及误差均达到了显著水平，因而得到的回归方程

$$\hat{y}=4.1575+0.8950\ x$$

是可以信赖的.

例 6 对 18 个土样中的无机磷含量(x_1)，溶于 K_2CO_3 并为溴化物水解的有机磷含量(x_2)，溶于 K_2CO_3 而不为溴化物水解的有机磷含量(x_3)和玉米吸收磷的含量(y)进行测定，得到如下数据 y 和 x，进行多元回归分析.

```
y = 64      x = 1.0000   0.4000   53.0000   158.0000
    60          1.0000   0.4000   23.0000   163.0000
    71          1.0000   3.1000   19.0000   37.0000
    61          1.0000   0.6000   34.0000   157.0000
    54          1.0000   4.7000   24.0000   59.0000
    77          1.0000   1.7000   65.0000   123.0000
    81          1.0000   9.4000   44.0000   46.0000
```

```
    93          1.0000   10.1000   31.0000   117.0000
    93          1.0000   11.6000   29.0000   173.0000
    51          1.0000   12.6000   58.0000   112.0000
    76          1.0000   10.9000   37.0000   111.0000
    96          1.0000   23.1000   46.0000   114.0000
    77          1.0000   23.1000   50.0000   134.0000
    93          1.0000   21.6000   44.0000   73.0000
    95          1.0000   23.1000   56.0000   168.0000
    54          1.0000   1.9000    36.0000   143.0000
   168          1.0000   26.8000   58.0000   202.0000
    99          1.0000   29.9000   51.0000   124.0000
[b bint r rint s] = regress(y,x,0.05)
b = 43.6522    1.7848    -0.0834    0.1611
bint =   5.0241    82.2803
          0.6315    2.9380
        -0.9793    0.8125
        -0.0784    0.4006
(r 与 rint 略去)
s = 0.5493    5.6885    0.0092
```

于是得到回归方程

$$\hat{y} = 43.652\,2 + 1.784\,8x_1 - 0.083\,4x_2 + 0.1611x_3$$

由于 $R^2 = 0.549\,3$，$F = 5.688\,5$，$p = 0.009\,2 < 0.01$，可见方程是极显著的.但未必每一个变量对 y 的贡献都是极显著的.为了求出有显著贡献的变量，MATLAB还提供了一个逐步回归的函数 stepwise( )，该函数是交互式图形工具函数，其格式为

```
stepwise(x,y,q,d)
```

其中 $\boldsymbol{X}$ 为自变量矩阵，列数即为自变量个数.若含有常数项，则 $\boldsymbol{X}$ 的第一列全为 1. $\boldsymbol{Y}$ 为因变量向量，q 为输入模式，即由起始选入的自变量序号组成.如起始选择 x_1 和 x_3，则 $q = [1,3]$.当 q 省略时，默认为全部变量，d 为显著水平，即剔除和选入变量用的显著水平.该函数执行后显示三个图文框.第一个是参数估计值及置信区间，决定系数 R^2、F 统计量等图文框.第二个是逐步回归历史图.第三个是逐步回归诊断表，凡系数估计值距 0 线很近，且其置信区间穿越 0 线(由虚线绘出)，则该系数与零无显著差异，属可剔除的变量；凡系数结果值远离 0 线，且其置信区间不穿越 0 线(由实线绘出)，则该系数与零有显著差异，该变量应保留，直至留下的变量全部显著为止.如对于上例，执行 stepwise(x,y)，则有

(图5－14 至图 5－16)

		Confidence Intervals	
Column #	Parameter	Lower	Upper
1	1.785	0.3258	3.244
2	-0.0834	-1.217	
3	0.1611	-0.1419	0.4641

RMSE	R-square	F	P
19.97	0.5493	5.689	0.009224

Close Help

图 5－14

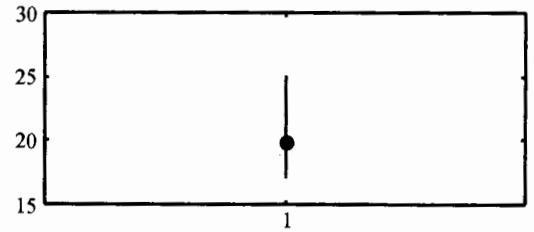

图 5－15

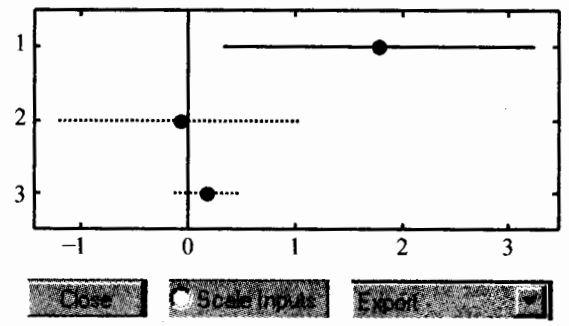

图 5－16

由上图知 x_2 和 x_3 的估计值与零无显著差异,可先剔除 x_2,这里必须一个变量一个变量剔除,不可一次剔除多个变量(图 5－17 至图 5－19).

		Confidence Intervals	
Column #	Parameter	Lower	Upper
1	1.737	0.4817	2.993
2	-0.0834	-1.217	
3	0.1548	-0.1239	0.4336

RMSE	R-square	F	P
19.32	0.5481	9.095	0.002589

Close Help

图 5－17

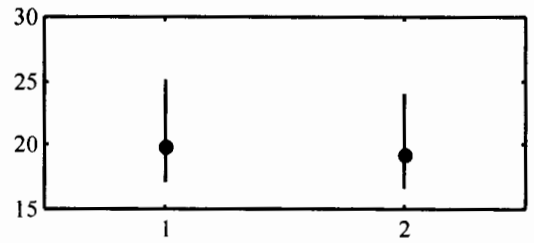

图 5－18

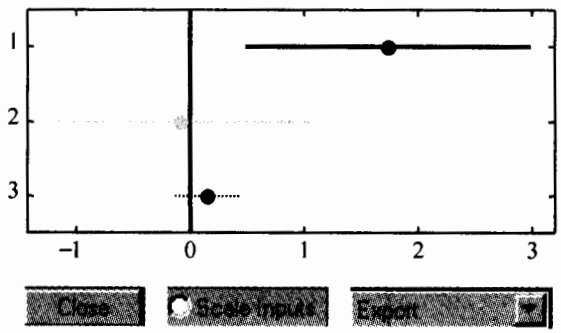

图 5－19

由上图知，剔除 x_2 后，x_3 仍不显著，就应继续剔除. 再将 x_3 剔除(图 5－20 至图 5－22)，现在只剩下 x_1 没有被剔除，而 x_1 是显著的，不能剔除。

		Confidence Inte rvals	
Column #	Parameter	Lower	Upper
1	1.843	0.5653	3.122
2	0.08665	-1.029	1.203
3	0.1548	-0.1239	0.4336

RMSE	R-square	F	P
20.05	0.4808	14.82	0.001417

Close　　Help

图 5－20

x_2 和 x_3 不再显著，因而也不能再引入，故最终获得的最优回归关系为

$$\hat{y} = 1.843\,4x_1 + 59.259\,0$$

它可由下述命令得到

```
x1 = [ones(18,1) x(:,1)];
[b bint r rint s] = regress(y,x1)
```

```
b = 59.2590    1.8434
bint = 43.5293    74.9886
       0.8282     2.8587
(r 与 rint 略去)
s = 0.4808    14.8171    0.0014.
```

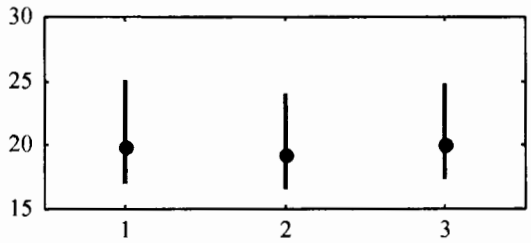

图 5－21

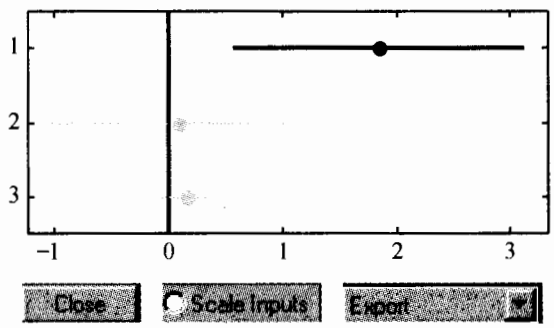

图 5－22

第五节 概率统计模型举例

一、气象观测站的优化

某地区有 12 个气象观测站，为了节省开支，计划减少气象观测站数目，已知该地区 12 个气象观测站的位置(图 5－23)，以及 10 年来各站测得的年降水量(表 5－1).