530 陈斯杰 电子信息工程 第18次作业

一、回归分析

1.

利用MATLAB求解：

```
%q1_1
clear;clc;
X=[74.3 78.8 68.8 78.0 70.4 80.5 80.5 69.7 71.2 73.5 ...
    79.5 75.6 75.0 78.8 72.0 72.0 72.0 74.3 71.2 72.0 ...
    75.0 73.5 78.8 74.3 75.8 65.0 74.3 71.2 69.7 68.0 ...
    73.5 75.0 72.0 64.3 75.8 80.3 69.7 74.3 73.5 73.5 ...
    75.8 75.8 68.8 76.5 70.4 71.2 81.2 75.0 70.4 68.0];
M=mean(X)%均值
V=var(X)%方差
S=std(X)%标准差
P=max(X)-min(X)%极差
SE=S/sqrt(50)%标准误
CV=S/M%变异系数
SK=skewness(X)%偏度
KU=kurtosis(X)%峰读
%得到结果：
M = 73.5740
V = 15.4424
S = 3.9297
P = 16.9000
SE = 0.5557
CV = 0.0534
SK = -0.0370
```

```
KU = 2.7072
```

2 解：根绝题意，使用matlab的mle函数解决最大似然估计问题，matlab代码如下：

```
data = ones(1,1000);
for i=1:1000
    if i <= 365
        data(i)=5;
    elseif i <= 610
        data(i) = 15;
    elseif i <= 760
        data(i) = 25;
    elseif i <= 860
        data(i) = 35;
    elseif i <= 930
        data(i) = 45;
    elseif i <= 975
        data(i) = 55;
    else
        data(i) = 65;
    end
end
[paramhat,paramint]=mle(data,'distribution', '
    exponential')
```

求出的结果为：

$$\lambda = 20$$

3.利用matlab运行以下代码：

```
X = [140,137,136,140,145,148,140,135,144,141];
Y = [135,118,115,140,128,131,130,115,131,125];
x = mean(X);y = mean(Y);
s1 = var(X);s2 = var(Y);
t = tinv(0.975,18);
sw = sqrt(9*(s1+s2)/18);
ph1 = x-y-t*sw*sqrt(1/10+1/10)
ph2 = x-y+t*sw*sqrt(1/10+1/10)
%result: ph1 = 7.5363
%         ph2 = 20.0637
```

所以置信区间为[7.5363,20.0637].

4.由极大似然估计法可知，

$\widehat{\lambda} = \overline{x} = 0.805$

i	n_i	$\widehat{p}_i$	$n\widehat{p}_i$	$\frac{(n_i - n\widehat{p}_i)^2}{n\widehat{p}_i}$
0	92	0.4471	89.42	0.0744
1	68	0.3599	71.98	0.2201
2	28	0.1449	28.98	0.0331
3	11	0.0389	7.78	1.3327
4	1	0.0078	1.56	0.2010
5	0	0.0013	0.26	0.2600
总和	200	1.0000	200	2.1213

$\chi^2 = 2.1213 < \chi^2_{0.1}(4) = 7.779$

所以可以认为每分钟顾客人数服从卡方分布。

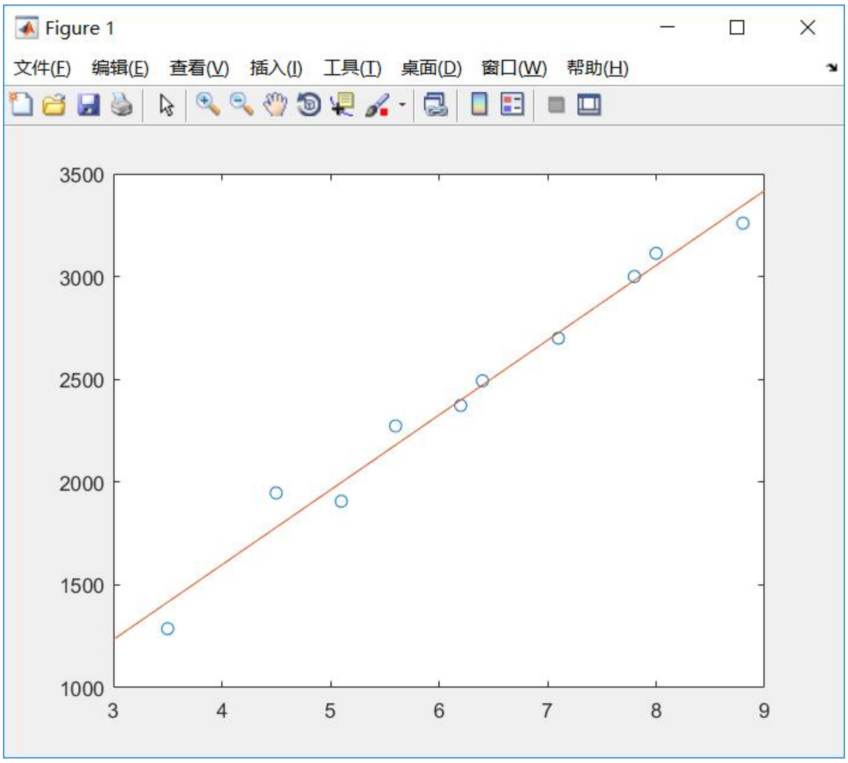
5.使用matlab的ranksum函数得到如下结果：

p = 1.8267e-04,h = 1 ,即拒绝H_0,认为超过5%的概率两者无关.

6.代码如下：

```
clear all;clc;
xi=3:0.01:9;
x=[5.1 3.5 7.1 6.2 8.8 7.8 4.5 5.6 8.0 6.4];
y=[1907 1287 2700 2373 3260 3000 1947 2273 3113 2493];
p=polyfit(x,y,1);
yi=p(1)*xi+p(2);
plot(x,y,'o',xi,yi)
fprintf('回归方程为y=%.2fx+%.2f',p(1),p(2))
[H,P,CI]=ttest2(x,y)
X=0.7;
Y=p(1)*X+p(2)
```

散点图为：



Y与X没有线性关系。

Y关于X的一元线性回归方程为：$y = 364.18x + 140.95$

显著性检验的结果为H=1，即表面零假设呗拒绝，X与Y在统计上认为是来自不同分布的数据，有区分度。

预测今年灌溉面积为$395.88hm^2$。

7.解：根据题意有R程序如下：

```
X1<-c(0.4,0.4,3.1,0.6,4.7,1.7,9.4,10.1,11.6,12.6,10.9,23.1,23.1,21.6,23.1,1.
;
X2<-c(52,23,19,34,24,65,44,31,29,58,37,46,50,44,56,36,58,51);
X3<-c(158,163,37,157,59,123,46,117,173,112,111,114,134,73,168,143,202,124)
;
Y<-c(64,60,71,61,54,77,81,93,93,51,76,96,77,93,95,54,168,99);
lm.sol<-lm(Y~X1+X2+X3)
summary(lm.sol)
```

(1).根据运行结果分析Y关于X1、X2、X3的线性回归方程为：

$$Y = 43.65007 + 1.78534 * X1 - 0.08329 * X2 + 0.16102 * X3$$

(2).由上述的结果可以得知方程的常量与X1高度显著；X2,X3不显著。回归方程的显著性检验不通过检验，相关系数的显著性检验通过检验(3).在源代码中加入下列代码：

```
> lm.step<-step(lm.sol)
> summary(lm.step)
```

根据运行结果分析可得最优回归方程：

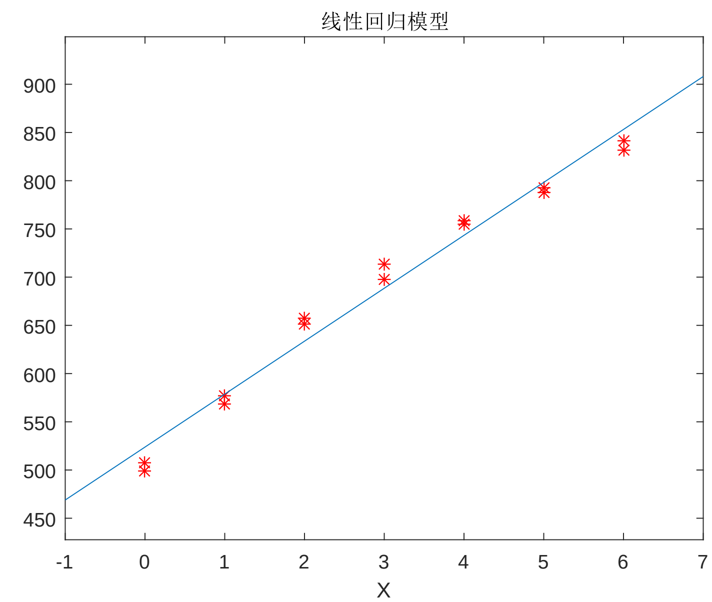
$$Y = 41.4794 + 1.7374 * X1 + 0.1548 * X3$$

8.解：根绝题意，使用matlab的regress函数解决一元线性回归和多项式回归模型，matlab代码如下：

```
    x1 = [0 0 1 1 2 2 3 3 4 4 5 5 6 6]';
    y = [508.1 498.4 568.2 577.3 651.7 657.0 713.4
        697.5 755.3 758.9 787.6 792.1 841.4 831.8]';
```

```
plot(x1',y','r*');
hold on;
%线性回归模型
x=[ ones(14,1) x1 ] ;
[b1,bint1,r1,rint1] = regress(y,x);
syms X
f = b1(1) + b1(2) * X;
ezplot(f,[-1,7]);
hold on;
%多项式回归模型
x=[ ones(14,2) x1 ] ;
[b2,bint2,r2,rint2] = regress(y,x);
syms Y
f = b2(1) + b2(2) * Y + b2(3) * Y^2;
ezplot(f,[-1,7]);
```

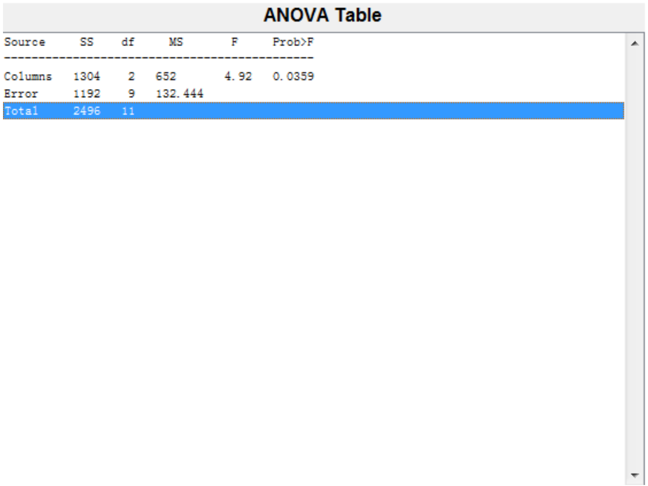
因为根据散点图可以看出，这个是个线性回归模型，所以不考虑其他多项式回归模型，画出的散点图和拟合曲线如下：



9.(1) 用matlab中anova1函数做单因素方差分析：

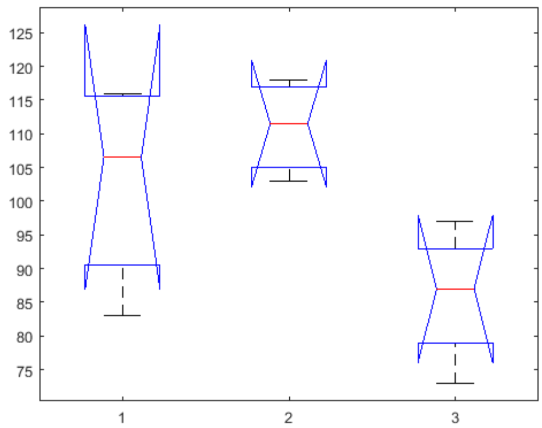
```
x = [115,103,73;
     116,107,89;
     98,118,85;
     83,116,97];
p = anova1(x)
```

得以下结果：



ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	1304	2	652	4.92	0.0359
Error	1192	9	132.444		
Total	2496	11			



所以$p = 0.0359 < \alpha = 0.05$,故拒绝$H_0$,即三个工厂生产的零件强度有显著差异. (2) 利用matlab运行以下代码：

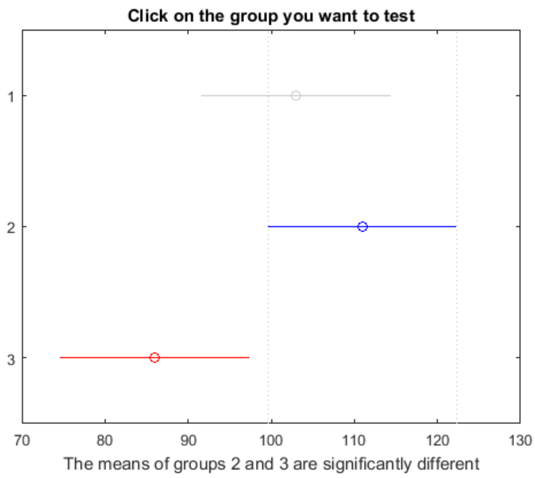
```
[m1,s1,mf1,sf1] = normfit(x(:,1),0.05);
[m2,s2,mf2,sf2] = normfit(x(:,2),0.05);
[m3,s3,mf3,sf3] = normfit(x(:,3),0.05);
%m1 = 103          m2 = 111          m3 = 86
%mf1 = 78.0426     mf2 = 99.5993     mf3 = 70.0878
%      127.9574          122.4007          101.9122
```

所以三个工厂的均值和$\alpha = 0.05$置信区间依次为：

甲：$\bar{x_1} = 103$,区间为[78.0426,127.9574];

乙：$\bar{x_2} = 111$,区间为[99.5993,122.4007]; 丙：$\bar{x_3} = 86$,区间为[70.0878,101.9122]. (3) matlab调用以下代码：

```
[p,table,stats] = anova1(x);
c = multcompare(stats);
%c = 1.0000    2.0000   -30.7205   -8.0000   14.7205   0.6050
%    1.0000    3.0000    -5.7205   17.0000   39.7205   0.1471
%    2.0000    3.0000     2.2795   25.0000   47.7205   0.0323
```



由图以及矩阵c的结果可得，乙厂和丙厂具有显著差异.

10.由matlab可求出，条件数$\kappa = 1376.9$,所以多重共线性较强。

由逐步回归法，可得最终的回归模型为

$\widehat{y} = 103.097 + 1.440x_1 - 0.614x_2$

二、方差分析

1.样本均值192.15,样本方差1783.9.

使用matlab的ttest函数得到如下结果：

p = 0.0025,h = 1 ,即拒绝H_0,认为油漆工的血小板较正常人有显著下降.

2.利用matlab求解，代码如下：

```
clear  all;clc;
xin=[126 125 136 128 123 138 142 116 110 108 115 140];
dui=[162 176 177 170 175 152 159 160 162];
[H1,P1,LSTAT1,CV1]=lillietest(xin,0.05)
[H2,P2,LSTAT2,CV2]=lillietest(dui,0.05)
var1=var(xin)
var2=var(dui)
mean1=mean(xin)
mean2=mean(dui)
```

结果得，两组数据都服从正态分布，且两组方程不相同。

从均值上看对照组大于新药组，且方差小于新药组，即对照组药效更好且更稳定。

3.

(1).

```
df<-data.frame(x=c(115,116,98,83,103,107,118,116,73,89,85,97)
    ,
                 A=gl(3,4))
fit.aov<-aov(x~A,data = df)
summary(fit.aov)
```

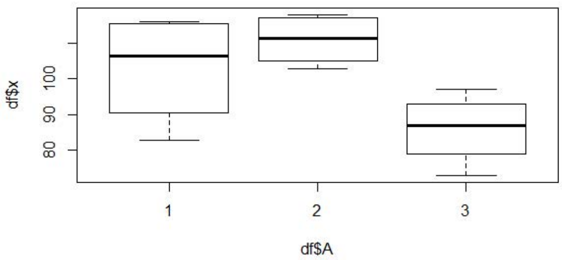
结果为：

```
            Df Sum Sq Mean Sq F value Pr(>F)  
A            2   1304   652.0   4.923 0.0359 *
Residuals    9   1192   132.4                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

由于p值为0.0359，小于0.05，因此认为三个工厂零件的强度有显著差异

```
plot(df$x ~ df$A)
```

结果为：



从图中也可以看出三个工厂零件的强度有显著差异(2).

```
> mean.1 <- mean(df$x[df$A == 1])
> mean.1.t <- t.test(df$x[df$A == 1], conf.level = 0.95)
> mean.2 <- mean(df$x[df$A == 2])
> mean.2.t <- t.test(df$x[df$A == 2], conf.level = 0.95)
> mean.3 <- mean(df$x[df$A == 3])
> mean.3.t <- t.test(df$x[df$A == 3], conf.level = 0.95)
>
> reulst <- data.frame(A = c(1, 2, 3), mean = c(mean.1, mean
    .2, mean.3), t.test.down = c(mean.1.t[4]$conf.int[1],
+
    mean.2.t[4]$conf.int[1], mean.3.t[4]$conf.int[1]), t.test.
    up = c(mean.1.t[4]$conf.int[2],
+
    mean.2.t[4]$conf.int[2], mean.3.t[4]$conf.int[2]))
> reulst
```

结果为:

	A	mean	t.test.down	t.test.up
1	1	103	78.04264	127.9574
2	2	111	99.59932	122.4007
3	3	86	70.08777	101.9122

(3).

```
pairwise.t.test(df$x, df$A, p.adjust.method = "none")
```

结果为：

```
        Pairwise comparisons using t tests with pooled SD

data:  df$x and df$A

  1     2
2 0.351 -
3 0.066 0.013

P value adjustment method: none
```

可以看到1和2的p值为0.351，可认为两水平无差异

1和3的P值0.066有差异，不显著；2和3的P值0.013，差异显著

结论为工厂丙的零件强度与其他两个厂有显著差异，甲乙两厂无差异

4.解：分析四个厂家生产产品的变化率代码为：

```
    x = [20 18 19 17 15 16 13 18 22 17
         26 19 26 28 23 25 23 23 25 25
         24 25 18 22 27 24 24 24 25 25
         12 14 12 14 12 14 12 14 12 14];
    p = anova1(x)
```

得出结果为：

$$p = 0.9995 > 1-\alpha = 0.95,\text{故拒绝}H_0,\text{即四个厂家生产产品变化率有差别。}$$

分析国内外厂家生产产品的变化率代码为：

```
    x = [20 18 19 17 15 16 13 18 22 17 26 19 26 28 23
      25 24 25 18 22 27 24
         12 14 12 14 12 14 12 14 12 14 12 14 12 14 12
            14 12 14 12 14 12 14];
    p = anova1(x)
```

得出结果为：

$p = 0.9994 > 1 - \alpha = 0.95$,故拒绝$H_0$,即国内外厂家生产产品变化率有差别。

分析国内各厂家生产产品的变化率代码为：

```
x = [20 18 19 17 15 16 13 18 22 17
     26 19 26 28 23 25 23 23 25 25
     24 25 18 22 27 24 24 24 25 25];
p = anova1(x)
```

得出结果为：

$p = 0.9898 > 1 - \alpha = 0.95$,故拒绝$H_0$,即国内各厂家生产产品变化率有差别。

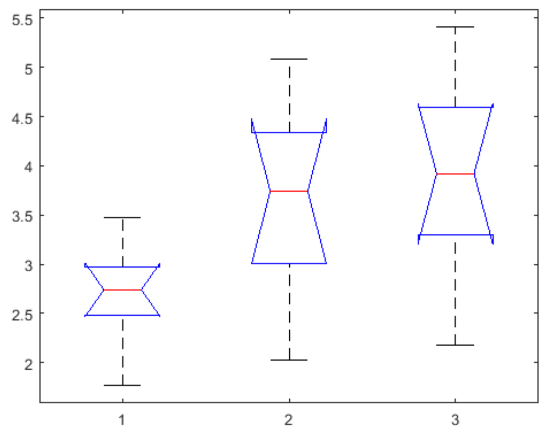
5.用matlab中anova1函数做单因素方差分析：

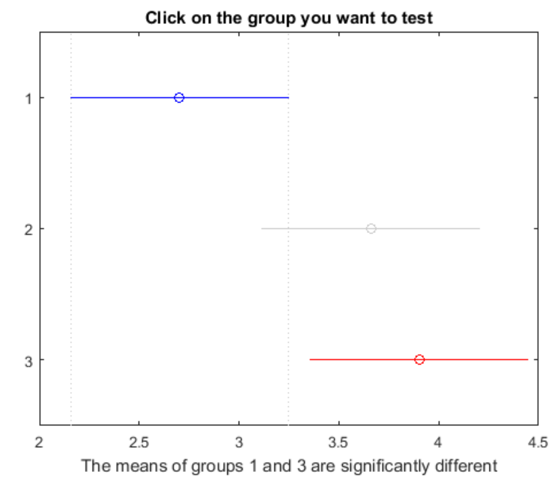
```
x = [2.79,3.83,5.41;
     2.69,3.15,3.47;
     3.11,4.70,4.92;
     3.47,3.97,4.07;
     1.77,2.03,2.18;
     2.44,2.87,3.13;
     2.83,3.65,3.77;
     2.52,5.09,4.26];
[p,table,stats] = anova1(x);
c = multcompare(stats)
```

结果如下：

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	6.4368	2	3.2184	4.28	0.0275
Error	15.7757	21	0.75123		
Total	22.2125	23			





所得$p = 0.0275 < \alpha = 0.05$,故拒绝假设$H_0$, 所以3 种不同处理的诱导作用不同，其中酚层RNA组与对照组差别最显著.

6.

```
s=[29.6  27.3  5.8  21.6  29.3
     24.3  32.6  6.2  17.4  32.8
```

```
    28.5  30.8  11.0  18.3  25.0
    32.0  34.8  8.3  19.0  24.2]
p=anova1(s)
```

可求出p值为$6.81 \times 10^{-8} < 0.05$

所以这些百分比均值由显著差异，即不同抗生素与血浆蛋白质结合的百分比有显著不同。

7.代码如下：

```
x0=[173,172,173  174,176,178  177,179,176  172,173,174
175,173,176  178,177,179  174,175,173  170,171,172
177,175,176  174,174,175  174,173,174  169,169,170];
x1=x0(:,1:3:10);
x2=x0(:,2:3:11);
x3=x0(:,3:3:12);
for  i=1:3
x(2*i-1,:)=x1(i,:);x(2*i,:)=x2(i,:);x(3*i,:)=x3(i,:);
end
[p,  t,  st]=anova2(x,  3)
x
```

得到问题的解为p_1=0.9989,p_2=0.0000,p_3=1.0000.

即认为化肥之间的差异对小麦的产量有显著影响，小麦品种的差异对小麦的产量无显著影响，两者的交互作用也不明显。

8.代码如下：

```
clear  all;clc;
x=[23.1  57.6  10.5  23.6  11.9  54.6  21.0  20.3
   22.7  53.2  9.7  19.6  13.8  47.1  13.6  23.6
```

```
    22.5  53.7  10.8  21.1  13.7  39.2  13.7  16.3
    22.6  53.1  8.3  21.6  13.3  37.0  14.8  14.8];
plot(x(1,:))
hold on
plot(x(2,:))
hold on
plot(x(3,:))
hold on
plot(x(4,:))
legend('恐惧','愉快','忧虑','平静')
```

结果图为：