数据采集方法作业

姓名：蒋贵豪　　**学号：**B+X9bo

2021 年 11 月 15 日

题目 1.　在某一个工业试验中，限定其他试验条件，只考虑温度这个因素对产品产量的影响，并记为 A。选定 5 个水平，分别为 $A_1 = 60°\text{C}$、$A_2 = 70°\text{C}$、$A_3 = 80°\text{C}$、$A_4 = 90°\text{C}$、$A_5 = 100°\text{C}$。在每个水平下试验的重复次数都为 3。结果如表1所示，其总均值 $\bar{y}_{..} = 68.2$。对表1的数据进行失拟检验，判断一次和二次模型能否较好的拟合给定的数据。

表 1: 单因素试验

温度 (x)	60°C	70°C	80°C	90°C	100°C
产量 (y)	37	80	91	81	53
	40	77	93	83	49
	43	74	92	79	51
平均产量 ($\bar{y}_{i.}$)	40	77	92	81	51

解答.　(1) 判断一次函数的拟合效果：

我们根据编写的 **Matlab** 代码进行失拟检验，得到线性函数的重复试验的失拟检验表。表中的各个参数的计算方法如下：

$$\mathrm{SS}_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 \tag{1}$$

其中，y_{ij} 为试验得到的响应值，$\bar{y}_{..}$ 为所有响应的均值，n_i 为第 i 个区组试验次数，k 为区组数。

$$\mathrm{SS}_R = \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})^2 \tag{2}$$

$$
\begin{aligned}
\mathrm{SS}_E &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\hat{y}_i\right)^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2 + \sum_{i=1}^{k} n_i\left(\bar{y}_{i.}-\hat{y}_i\right)^2 \\
&\equiv \mathrm{SS}_{PE} + \mathrm{SS}_{LOF},
\end{aligned} \tag{3}
$$

其中，$\hat{y}_i$ 为回归模型对第 i 个水平响应的估计值。为了判断模型是否正确，有：

$$
F = \frac{\mathrm{SS}_{LOF}/(k-p)}{\mathrm{SS}_{PE}/(n-k)} = \frac{\mathrm{MS}_{LOF}}{\mathrm{MS}_{PE}} \tag{4}
$$

其中 n 为总试验次数，p 为回归多项式的次数。若模型正确，F 应服从自由度 $k-p$ 和 $n-k$ 的 F 分布。最后，我们计算 p 值：

$$
p = 1 - \Phi_{k-p.n-k}(F) \tag{5}
$$

其中 $\Phi_{k-p.n-k}$ 为自由度 $k-p$ 和 $n-k$ 的 F 分布的累积分布函数。再将计算出的 p 值与我们给定的显著性水平 α 比较，若 $p \leq \alpha$, 则我们拒绝 H_0，认为检验显著。若 $p > \alpha$，我们接受 H_0, 认为检验不显著。

我们得到的失拟检验表2所示：

表 2: 线性模型重复试验的失拟检验

方差来源	自由度	平方和	均方	F 值	p 值
回归	1	161.2	161.2		
失拟	3	3662.4	1220.8	381.5	1.315×10^{-10}
纯误差	10	32	3.2		
总和	14	4102.8			

我们设立的检验水平 $\alpha = 0.05$, 由表2可知，$p < \alpha$。于是，检验显著，我们拒绝 H_0, 意味着失拟部分和纯误差部分有显著区别。进一步对回归部分和存误差部分做 F 检验，如表3所示：

表 3: 线性模型重复试验的失拟检验 (续)

方差来源	自由度	平方和	均方	F 值	p 值
回归	1	161.2	161.2	50.375	3.3051×10^{-5}
纯误差	10	32	3.2		

由表3可知，$p < \alpha$, 我们拒绝 H_0，也就是意味着**线性可用但不是很好，需要改进**。

(2) 判断二次函数的拟合效果：

同一次函数失拟检验的过程，我们给出二次函数的失拟检验表4。

我们设立的检验水平 $\alpha = 0.05$, 由表4可知，$p > \alpha$。于是，检验不显著，我们接受 H_0, 意味着失拟部分和纯误差部分没有显著区别。进一步对失拟部分和存误差部分的自由度及平方和各自相加，重新计算均方，再对回归平方和做 F 检验，如表5所示：

表 4: 二次函数重复试验的失拟检验

方差来源	自由度	平方和	均方	F 值	p 值
回归	2	5688.514	2844.257		
失拟	2	7.886	3.943	0.730	0.506
纯误差	10	54	5.4		
总和	14	5750.4			

表 5: 二次函数重复试验的失拟检验 (续)

方差来源	自由度	平方和	均方	F 值	p 值
回归	2	5688.514	2844.257	551.518	1.5536×10^{-12}
误差	12	61.886	5.157		
总和	14	5690.4			

于是，$p < \alpha$, 我们拒绝 H_0，也就是意味着**二次模型拟合效果较好**。最后，我们得到的回归方程为：

$$\hat{y}_i = -0.1143x^2 + 18.5457x - 661.1714 \tag{6}$$

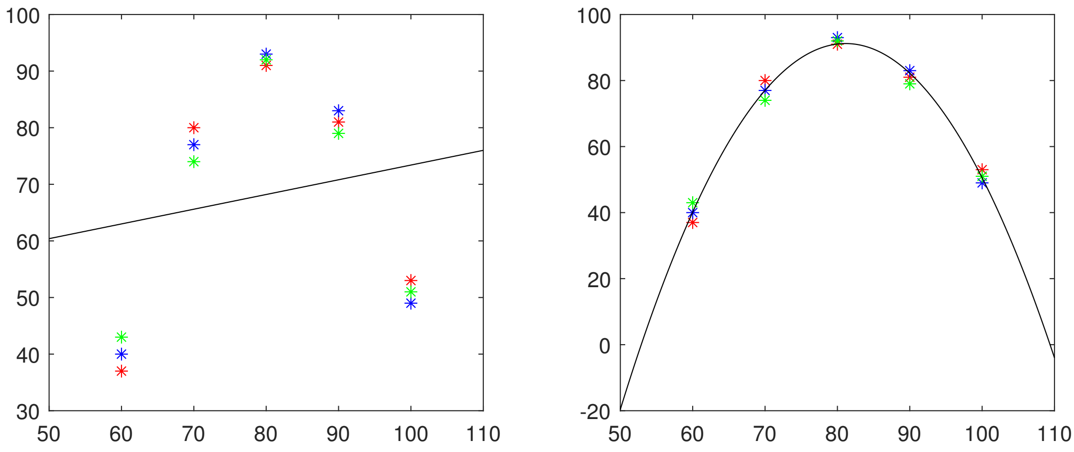

图 1: 单因素试验的线性和二次回归拟合模型

如图1所示，线性拟合模型的效果较差，二次回归拟合具有良好的效果。

本题中使用到的代码如下：

(1) 拟合模型的代码：

```
function a = leastsquare(x, y, ...
        N)%最小二乘法，输入x和y，N为拟合模型多项式次数
n = length(x);
for i = 1:N+1
for j = 1:N+1
H(i, j) = sum((x.^(i-1)).*(x.^(j-1)));
end
end
for i = 1:N+1
d(i) =  sum((x.^(i-1)).*(y));
end
A = H^(-1)*d';
for i = 1:N+1
a(i) = A(N+2-i);
end
```

(2)ANOVA 各个参数计算代码：

```matlab
function [SS_PE, MS_PE, SS_T, SS_LOF, MS_LOF, SS_E, MS_E, SS_R, ...
        MS_R, F, P, a] = ANOVA(x, y, N)
alpha = 0.05;
S = size(x);
x_vector = x(:);
y_vector = y(:);
y_bar = mean(y');
a = leastsquare(x_vector, y_vector, N);
k = S(1);
n = length(x_vector);
p = N+1;%回归自由度+1
Y = polyval(a,x)%回归值
for i = 1:k
for j = 1:p
ss_PE(i, j) = (y(i,j)-y_bar(i))^2;
ss_T(i, j) = (y(i,j)-mean(y_bar))^2;
ss_LOF(i,j) = (Y(i,j)-y_bar(i))^2;
ss_E(i,j) = (Y(i,j)-y(i,j))^2;
end
end
SS_PE = sum(sum(ss_PE))
MS_PE = SS_PE/(n-k)
SS_T = sum(sum(ss_T))
SS_LOF = sum(sum(ss_LOF))
MS_LOF = SS_LOF/(k-p)
SS_E = sum(sum(ss_E))
MS_E = SS_E/(n-p)
SS_R = SS_T-SS_E
MS_R = SS_R/(p-1)
F = MS_LOF/MS_PE
```

```
P = 1-fcdf(F,k-p,n-k)
```

(3) 计算和绘图代码

```
x = [60 60 60
70 70 70
80 80 80
90 90 90
100 100 100];
y = [37 40 43
80 77 74
91 93 92
81 83 79
53 49 51];
N = 1%回归多项式系数
[SS_PE, MS_PE, SS_T, SS_LOF, MS_LOF, SS_E, MS_E, SS_R, MS_R, F, P, ...
        a_1] = ANOVA(x, y, N);
x_1 = x';
y_1 = y';
subplot(1,2,1)
plot(x_1(1,:),y_1(1,:),'*r')
hold on
plot(x_1(2,:),y_1(2,:),'*b')
hold on
plot(x_1(3,:),y_1(3,:),'*g')
hold on
X_1 = linspace(50,110,601);
Y_1 = polyval(a_1,X_1);
plot( X_1, Y_1, 'k')
N = 2%回归多项式系数
[SS_PE, MS_PE, SS_T, SS_LOF, MS_LOF, SS_E, MS_E, SS_R, MS_R, F, P, ...
```

```
            a_2] = ANOVA(x, y, N);
subplot(1,2,2)
plot(x_1(1,:),y_1(1,:),'*r')
hold on
plot(x_1(2,:),y_1(2,:),'*b')
hold on
plot(x_1(3,:),y_1(3,:),'*g')
hold on
X_2 = linspace(50,110,601);
Y_2 = polyval(a_2,X_2);
plot( X_2, Y_2, 'k')
```

题目 2. 某品尝小组欲比较五种不同品牌的冰淇淋 (A、B、C、D、E)。然而为了正确的评判和比较这些口味，每位品尝专家至多只能品尝 3 种品牌。这种情况下最好使用哪种类型的试验设计？能否构造出有 5 个品尝专家的这种设计？

解答. 由于该情况下，区组所含的试验单元数最多为 $t = 3$, 因素水平数为 $q = 5$。区组所含的试验单元数小于因素的水平数，因此**我们采用平衡不完全随机区组设计**。

在此题的情况中，区组数 $b = 5$, 设每个水平的试验次数为 r, 任一水平在同一区组内同时出现的次数为 λ。则由平衡不完全区组设计存在的必要条件：

$$bt = qr \tag{7}$$

$$\lambda(q-1) = r(t-1) \tag{8}$$

$$b \geq q \tag{9}$$

由式7可知，$r = 3$。再结合式8，可知 $\lambda = 1.5$ 不是整数，因此无法满足平衡性。

当 $t = 2$ 时，由式7可知，$r = 2$。再结合式8，可知 $\lambda = 0.5$ 也不是整数，也无法满足平衡性。于是，**无法构造出有 5 个品尝专家的这种设计**。

题目 3. 有一化学试验，其目的是比较一种新的催化剂 B 是否比原来的催化剂 A 更能提高产量。试验在六批不同的原材料上进行，已知各批次的原材料相互之间是不同的。每批原材料分为两部分，分别随机地使用 A 和 B 做试验，结果数据在表中给出。

(a) 说明所用的试验设计。

(b) 进行适当的 t 检验。

(c) 构造 A 和 B 之间差异的 95% 的置信区间。

表 6: 化学反应试验的产出数据

催化剂	1	2	3	4	5	6
A	9	19	28	22	18	8
B	10	22	30	21	23	12

解答. (a) 所用的试验设计为**配对比较设计**。

(b) 由题意知，试验次数 $N=6$, 设 A 与 B 的结果差异：

$$d_i = y_{i2} - y_{i1} \tag{10}$$

计算 d 的均值及标准差：

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i = 2.3333 \tag{11}$$

$$s_d = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2} = 2.1602 \tag{12}$$

我们构造的 t 统计量为：

$$t_{paired} = \frac{\sqrt{N}\bar{d}}{s_d} = 2.6458 \tag{13}$$

我们设定显著水平 $\alpha = 0.05$, 于是 $t_{0.975,5} = 2.5706 < t_{paired}$, 于是我们**认为 A 和 B 催化剂有显著差异**。

(c) 结合 (b)，**A 和 B 之间差异的 95% 置信区间为：**

$$\begin{aligned}&[\bar{d} - t_{1-\frac{\alpha}{2},N-1}s_d/\sqrt{N}, \bar{d} + t_{1-\frac{\alpha}{2},N-1}s_d/\sqrt{N}] \\ &= [2.3333 \pm 2.5706 \times 2.1602/\sqrt{6}] = [0.0663, 4.6003]\end{aligned} \tag{14}$$

题目 4. 有一试验用于比较一个处理的四个不同水平，共安排了 40 个水平组合的试验，每天最多 8 个。试验人员怀疑天与天之间变化的影响可能存在。

(a) 问试验是否应分区组？什么因子会与作此决策有关？

(b) 无论你的决策如何，简述你如何安排此试验。

解答. (a) 试验**需要分区组**，因为天与天之间的变化可能存在对试验结果的影响，所以可以把试验的天数看作是一个区组。

(b) 我们将试验分为 5 天进行，也就是按天分为了 5 个区组。每天对 4 个不同的水平各做两次试验，在每天的 8 次试验中，4 个水平的顺序随机生成。如此做 5 天试验，可以得到同一天内同一水平的一组试验数据，4 个水平共 4 组，可以用来判断不同水平的影响是否显著。而对于 5 天的区组，可以得到 4 个不同水平在 5 天内的不同响应数据，来判断天与天之间的变化对试验结果的影响是否显著。

设四个不同的水平为 A、B、C、D，表7给出了一种此试验设计的安排：

表 7: 试验设计表

天数	1	2	3	4	5
试验安排	A	B	C	D	A
	B	C	D	A	C
	C	D	A	B	D
	D	A	B	C	B
	D	A	B	C	B
	C	D	A	B	D
	B	C	D	A	C
	A	B	C	D	A