第八章　方差分析与回归分析

习题 8.1

1．在一个单因子试验中，因子 A 有三个水平，每个水平下各重复 4 次，具体数据如下：

水平	数据			
一水平	8	5	7	4
二水平	6	10	12	9
三水平	0	1	5	2

试计算误差平方和 S_e、因子 A 的平方和 S_A、总平方和 S_T，并指出它们各自的自由度．

解：设 $T=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m}y_{ij}$，$T_i=\sum\limits_{j=1}^{m}y_{ij}$，

有 $\bar{y}=\dfrac{1}{n}T$，$\bar{y}_{i\cdot}=\dfrac{1}{m}T_i$，

因 $r=3$，$m=4$，$n=rm=12$，y_{ij} 及计算结果如下表：

水平	数据				T_i	T_i^2	$\sum\limits_{j=1}^{m}y_{ij}^2$
一水平	8	5	7	4	24	576	154
二水平	6	10	12	9	37	1369	361
三水平	0	1	5	2	8	64	30
Σ					69	2009	545

故 $S_T=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m}(y_{ij}-\bar{y})^2=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m}y_{ij}^2-\dfrac{1}{n}T^2=545-\dfrac{1}{12}\times69^2=148.25$，自由度 $f_T=n-1=11$；

$S_A=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m}(\bar{y}_{i\cdot}-\bar{y})^2=\dfrac{1}{m}\sum\limits_{i=1}^{r}T_i^2-\dfrac{1}{n}T^2=\dfrac{1}{4}\times2009-\dfrac{1}{12}\times69^2=105.5$，自由度 $f_A=r-1=2$；

$S_e=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2=S_T-S_A=148.25-105.5=42.75$，自由度 $f_e=n-r=9$．

2．在一个单因子试验中，因子 A 有 4 个水平，每个水平下重复次数分别为 5, 7, 6, 8．那么误差平方和、A 的平方和及总平方和的自由度各是多少？

解：因 $r=4$，$m_1=5$，$m_2=7$，$m_3=6$，$m_4=8$，有 $n=m_1+m_2+m_3+m_4=26$，

故总平方和 $S_T=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m_i}(y_{ij}-\bar{y})^2$ 的自由度是 $f_T=n-1=25$；

A 的平方和 $S_A=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m_i}(\bar{y}_{i\cdot}-\bar{y})^2=\sum\limits_{i=1}^{r}m_i(\bar{y}_{i\cdot}-\bar{y})^2$ 的自由度是 $f_A=r-1=3$；

误差平方和 $S_e=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m_i}(y_{ij}-\bar{y}_{i\cdot})^2$ 的自由度是 $f_e=n-r=22$．

3．在单因子试验中，因子 A 有 4 个水平，每个水平下各重复 3 次试验，现已求得每个水平下试验结果的样本标准差分别为 1.5, 2.0, 1.6, 1.2，则其误差平方和为多少？误差的方差 σ^2 的估计值是多少？

解：每个水平下的样本标准差为$s_i=\sqrt{\dfrac{1}{m-1}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2}$，有$\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2=(m-1)s_i^2$，

故误差平方和$S_e=\sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{i\cdot})^2=\sum_{i=1}^{r}(m-1)s_i^2=2\times(1.5^2+2.0^2+1.6^2+1.2^2)=20.5$；

因$r=4$，$m=3$，$n=rm=12$，误差平方和S_e的自由度是$f_e=n-r=8$，

故误差的方差σ^2的估计值$\hat{\sigma}^2=\dfrac{S_e}{n-r}=\dfrac{20.5}{8}=2.5625$.

4. 在单因子方差分析中，因子A有三个水平，每个水平下各做4次重复试验，请完成下列方差分析表，并在显著性水平$\alpha=0.05$下对因子A是否显著作出检验.

方差分析表

来源	平方和	自由度	均方和	F比	p值
因子A	4.2				
误差e	2.5				
和T	6.7				

解：因$r=3$，$m=4$，$n=rm=12$，$S_A=4.2$，$S_e=2.5$，$S_T=6.7$，

则自由度$f_A=r-1=2$，$f_e=n-r=9$，$f_T=n-1=11$，

均方和$MS_A=\dfrac{S_A}{f_A}=\dfrac{4.2}{2}=2.1$，$MS_e=\dfrac{S_e}{f_e}=\dfrac{2.5}{9}=0.2778$，

F比$F=\dfrac{S_A/f_A}{S_e/f_e}=\dfrac{2.1}{0.2778}=7.56$，

方差分析表

来源	平方和	自由度	均方和	F比	p值
因子A	4.2	2	2.1	7.56	0.0118
误差e	2.5	9	0.2778		
和T	6.7	11			

假设H_0：$a_1=a_2=a_3=0$，

选取统计量$F=\dfrac{S_A/f_A}{S_e/f_e}\sim F(f_A,f_e)$，

显著性水平$\alpha=0.05$，$F_{1-\alpha}(f_A,f_e)=F_{0.95}(2,9)=4.26$，右侧拒绝域$W=\{F\geq 4.26\}$，

因$F=\dfrac{S_A/f_A}{S_e/f_e}=\dfrac{2.1}{0.2778}=7.56\in W$，并且检验的$p$值$p=P\{F\geq 7.56\}=0.0118<\alpha=0.05$，

故拒绝H_0，接受H_1，可以认为因子A显著.

5. 用4种安眠药在兔子身上进行试验，特选24只健康的兔子，随机把它们均分为4组，每组各服一种安眠药，安眠时间如下所示.

安眠药试验数据

安眠药	安眠时间/h					
A_1	6.2	6.1	6.0	6.3	6.1	5.9
A_2	6.3	6.5	6.7	6.6	7.1	6.4
A_3	6.8	7.1	6.6	6.8	6.9	6.6
A_4	5.4	6.4	6.2	6.3	6.0	5.9

在显著性水平$\alpha = 0.05$下对其进行方差分析，可以得到什么结果？

解：假设 H_0：$a_1 = a_2 = a_3 = a_4 = 0$，

选取统计量$F = \dfrac{S_A/f_A}{S_e/f_e} \sim F(f_A, f_e)$，

显著性水平$\alpha = 0.05$，$r = 4$，$m = 6$，$n = 24$，有$f_A = r - 1 = 3$，$f_e = n - r = 20$，

则$F_{1-\alpha}(f_A, f_e) = F_{0.95}(3, 20) = 3.10$，右侧拒绝域$W = \{F \geq 3.10\}$，

安眠药	安眠时间/h						T_i	T_i^2	$\sum_{j=1}^{m} y_{ij}^2$
A_1	6.2	6.1	6.0	6.3	6.1	5.9	36.6	1339.56	223.36
A_2	6.3	6.5	6.7	6.6	7.1	6.4	39.6	1568.16	261.76
A_3	6.8	7.1	6.6	6.8	6.9	6.6	40.8	1664.64	277.62
A_4	5.4	6.4	6.2	6.3	6.0	5.9	36.2	1310.44	219.06
Σ							153.2	5882.8	981.8

得$S_T = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{r}\sum_{j=1}^{m} y_{ij}^2 - \dfrac{1}{n}T^2 = 981.8 - \dfrac{1}{24} \times 153.2^2 = 3.8733$，

$$S_A = \sum_{i=1}^{r}\sum_{j=1}^{m}(\bar{y}_{i\cdot} - \bar{y})^2 = \frac{1}{m}\sum_{i=1}^{r} T_i^2 - \frac{1}{n}T^2 = \frac{1}{6} \times 5882.8 - \frac{1}{24} \times 153.2^2 = 2.54\,,$$

$$S_e = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2 = S_T - S_A = 3.8733 - 2.54 = 1.3333\,,$$

方差分析表

来源	平方和	自由度	均方和	F 比	p 值
因子 A	2.54	3	0.8467	12.7	7.1641×10^{-5}
误差 e	1.3333	20	0.0667		
和 T	3.8733	23			

有$F = \dfrac{S_A/f_A}{S_e/f_e} = \dfrac{2.54/3}{1.3333/20} = \dfrac{0.8467}{0.0667} = 12.7 \in W$，

并且检验的 p 值 $p = P\{F \geq 12.7\} = 7.1641 \times 10^{-5} < \alpha = 0.05$，

故拒绝 H_0，接受 H_1，可以认为因子 A 显著，即 4 种安眠药对兔子的安眠作用有明显差别.

6．为研究咖啡因对人体功能的影响，特选 30 名体质大致相同的健康的男大学生进行手指叩击训练，此外咖啡因选三个水平：

$A_1 = 0$ mg，　　$A_2 = 100$ mg，　　$A_3 = 200$ mg.

每个水平下冲泡 10 杯水，外观无差别，并加以编号，然后让 30 位大学生每人从中任选一杯服下，2 h 后，请每人做手指叩击，统计员记录器每分钟叩击次数，试验结果统计如下表：

咖啡因含量	叩击次数									
A_1：0 mg	242	245	244	248	247	248	242	244	246	242
A_2：100 mg	248	246	245	247	248	250	247	246	243	244
A_3：200 mg	246	248	250	252	248	250	246	248	245	250

请对上述数据进行方差分析，从中可得到什么结论（取$\alpha = 0.05$）？

解：假设 H_0：$a_1 = a_2 = a_3 = 0$，

选取统计量$F=\dfrac{S_A/f_A}{S_e/f_e}\sim F(f_A,f_e)$，

显著性水平$\alpha=0.05$，$r=3$，$m=10$，$n=30$，有$f_A=r-1=2$，$f_e=n-r=27$，
则$F_{1-\alpha}(f_A,f_e)=F_{0.95}(2,27)=3.3541$，右侧拒绝域$W=\{F\ge 3.3541\}$，
将叩击次数的原始数据减去 240 次，列计算表

咖啡因含量	叩击次数 -240										T_i	T_i^2	$\sum\limits_{j=1}^{m}y_{ij}^2$
A_1：0 mg	242	245	244	248	247	248	242	244	246	242	48	2304	282
A_2：100 mg	248	246	245	247	248	250	247	246	243	244	64	4096	448
A_3：200 mg	246	248	250	252	248	250	246	248	245	250	83	6889	733
Σ											195	13289	1463

得$S_T=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m}(y_{ij}-\bar{y})^2=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m}y_{ij}^2-\dfrac{1}{n}T^2=1463-\dfrac{1}{30}\times 195^2=195.5$，

$$S_A=\sum_{i=1}^{r}\sum_{j=1}^{m}(\bar{y}_{i\cdot}-\bar{y})^2=\frac{1}{m}\sum_{i=1}^{r}T_i^2-\frac{1}{n}T^2=\frac{1}{10}\times 13289-\frac{1}{30}\times 195^2=61.4\text{，}$$

$$S_e=\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2=S_T-S_A=195.5-61.4=134.1\text{，}$$

方差分析表

来源	平方和	自由度	均方和	F 比	p 值
因子 A	61.4	2	30.7	6.1812	0.0062
误差 e	134.1	27	4.9667		
和 T	195.5	29			

有$F=\dfrac{S_A/f_A}{S_e/f_e}=\dfrac{61.4/2}{134.1/27}=\dfrac{30.7}{4.9667}=6.1812\in W$，

并且检验的p值$p=P\{F\ge 12.7\}=0.0062<\alpha=0.05$，
故拒绝 H_0，接受 H_1，可以认为因子A显著，即咖啡因的不同剂量对手指叩击次数有明显影响.

7. 某粮食加工厂试验三种储藏方法对粮食含水率有无显著影响．现取一批粮食分成若干份，分别用三种不同的方法储藏，过一段时间后测得的含水率如下表：

储藏方法	含水率数据				
A_1	7.3	8.3	7.6	8.4	8.3
A_2	5.4	7.4	7.1	6.8	5.3
A_3	7.9	9.5	10.0	9.8	8.4

（1）假定各种方法储藏的粮食的含水率服从正态分布，且方差相等，试在$\alpha=0.05$水平下检验这三种方法对含水率有无显著影响；

（2）对每种方法的平均含水率给出置信水平为 0.95 的置信区间.

解：（1）假设 H_0：$a_1=a_2=a_3=0$，

选取统计量$F=\dfrac{S_A/f_A}{S_e/f_e}\sim F(f_A,f_e)$，

显著性水平$\alpha=0.05$，$r=3$，$m=5$，$n=15$，有$f_A=r-1=2$，$f_e=n-r=12$，

则 $F_{1-\alpha}(f_A, f_e) = F_{0.95}(2, 12) = 3.89$，右侧拒绝域 $W = \{F \ge 3.89\}$，

储藏方法	含水率数据					T_i	T_i^2	$\sum_{j=1}^{m} y_{ij}^2$
A_1	7.3	8.3	7.6	8.4	8.3	39.9	1592.01	319.39
A_2	5.4	7.4	7.1	6.8	5.3	32	1024	208.66
A_3	7.9	9.5	10.0	9.8	8.4	45.6	2079.36	419.26
Σ						117.5	4695.37	947.31

得 $S_T = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{r}\sum_{j=1}^{m} y_{ij}^2 - \frac{1}{n}T^2 = 947.31 - \frac{1}{15}\times 117.5^2 = 26.8933$，

$$S_A = \sum_{i=1}^{r}\sum_{j=1}^{m}(\bar{y}_{i\cdot} - \bar{y})^2 = \frac{1}{m}\sum_{i=1}^{r} T_i^2 - \frac{1}{n}T^2 = \frac{1}{5}\times 4695.37 - \frac{1}{15}\times 117.5^2 = 18.6573，$$

$$S_e = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2 = S_T - S_A = 26.8933 - 18.6573 = 8.236，$$

方差分析表

来源	平方和	自由度	均方和	F 比	p 值
因子 A	18.6573	2	9.3287	13.5920	8.2496×10^{-4}
误差 e	8.236	12	0.6863		
和 T	26.8933	14			

有 $F = \dfrac{S_A/f_A}{S_e/f_e} = \dfrac{18.6573/2}{8.236/12} = \dfrac{9.3287}{0.6863} = 13.5920 \in W$，

并且检验的 p 值 $p = P\{F \ge 12.7\} = 8.2496 \times 10^{-4} < \alpha = 0.05$，

故拒绝 H_0，接受 H_1，可以认为因子 A 显著，即三种储藏方法对粮食含水率有显著影响；

（2）估计平均含水率 μ_i，$i = 1, 2, 3$，

选取枢轴量 $T = \dfrac{\bar{Y}_{i\cdot} - \mu_i}{\hat{\sigma}/\sqrt{m}} \sim t(f_e)$，其中 $\hat{\sigma} = \sqrt{\dfrac{S_e}{f_e}}$，置信区间为 $(\bar{Y}_{i\cdot} \pm t_{1-\alpha/2}(f_e)\cdot\dfrac{\hat{\sigma}}{\sqrt{m}})$，

因 $m = 5$，$\bar{y}_{1\cdot} = \dfrac{T_1}{m} = \dfrac{39.9}{5} = 7.98$，$\bar{y}_{2\cdot} = \dfrac{T_2}{m} = \dfrac{32}{5} = 6.4$，$\bar{y}_{3\cdot} = \dfrac{T_3}{m} = \dfrac{45.6}{5} = 9.12$，

置信水平 $1 - \alpha = 0.95$，$t_{1-\alpha/2}(f_e) = t_{0.975}(12) = 2.1788$，$\hat{\sigma} = \sqrt{\dfrac{S_e}{f_e}} = \sqrt{0.6863} = 0.8285$，

故 μ_1 的 0.95 置信区间为 $(\bar{y}_{1\cdot} \pm t_{1-\alpha/2}(f_e)\cdot\dfrac{\hat{\sigma}}{\sqrt{m}}) = (7.98 \pm 2.1788\times\dfrac{0.8285}{\sqrt{5}}) = (7.1728, 8.7872)$；

μ_2 的 0.95 置信区间为 $(\bar{y}_{2\cdot} \pm t_{1-\alpha/2}(f_e)\cdot\dfrac{\hat{\sigma}}{\sqrt{m}}) = (6.4 \pm 2.1788\times\dfrac{0.8285}{\sqrt{5}}) = (5.5928, 7.2072)$；

μ_3 的 0.95 置信区间为 $(\bar{y}_{3\cdot} \pm t_{1-\alpha/2}(f_e)\cdot\dfrac{\hat{\sigma}}{\sqrt{m}}) = (9.12 \pm 2.1788\times\dfrac{0.8285}{\sqrt{5}}) = (8.3128, 9.9272)$.

8．在入户推销上有五种方法，某大公司相比较这五种方法有无显著的效果差异，设计了一项实验：从应

聘的且无推销经验的人员中随机挑选一部分人，将他们随机地分为五个组，每一组用一种推销方法进行培训，培训相同时间后观察他们在一个月内的推销额，数据如下：

组别	推销额/千元						
第一组	20.0	16.8	17.9	21.2	23.9	26.8	22.4
第二组	24.9	21.3	22.6	30.2	29.9	22.5	20.7
第三组	16.0	20.1	17.3	20.9	22.0	26.8	20.8
第四组	17.5	18.2	20.2	17.7	19.1	18.4	16.5
第五组	25.2	26.2	26.9	29.3	30.4	29.7	28.2

（1）假定数据满足进行方差分析的假定，对数据进行分析，在$\alpha = 0.05$下，这五种方法在平均月推销额上有无显著差异？

（2）那种推销方法的效果最好？试对该种方法一个月的平均月推销额求置信水平为 0.95 的置信区间．

解：（1）假设 H$_0$：$a_1 = a_2 = a_3 = a_4 = a_5 = 0$，

选取统计量$F = \dfrac{S_A/f_A}{S_e/f_e} \sim F(f_A, f_e)$，

显著性水平$\alpha = 0.05$，$r = 5$，$m = 7$，$n = rm = 35$，有$f_A = r - 1 = 4$，$f_e = n - r = 30$，

则$F_{1-\alpha}(f_A, f_e) = F_{0.95}(4, 30) = 2.69$，右侧拒绝域$W = \{F \ge 2.69\}$，

组别	推销额/千元							T_i	T_i^2	$\sum\limits_{j=1}^{m} y_{ij}^2$
第一组	20.0	16.8	17.9	21.2	23.9	26.8	22.4	149	22201	3243.3
第二组	24.9	21.3	22.6	30.2	29.9	22.5	20.7	172.1	29618.41	4325.25
第三组	16.0	20.1	17.3	20.9	22.0	26.8	20.8	143.9	20707.21	3030.99
第四组	17.5	18.2	20.2	17.7	19.1	18.4	16.5	127.6	16281.76	2334.44
第五组	25.2	26.2	26.9	29.3	30.4	29.7	28.2	195.9	38376.81	5505.07
Σ								788.5	127185.19	18439.05

得$S_T = \sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m}(y_{ij} - \bar{y})^2 = \sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m} y_{ij}^2 - \dfrac{1}{n}T^2 = 18439.05 - \dfrac{1}{35}\times 788.5^2 = 675.2714$，

$$S_A = \sum_{i=1}^{r}\sum_{j=1}^{m}(\bar{y}_{i\cdot} - \bar{y})^2 = \frac{1}{m}\sum_{i=1}^{r} T_i^2 - \frac{1}{n}T^2 = \frac{1}{7}\times 127185.19 - \frac{1}{35}\times 788.5^2 = 405.5343,$$

$$S_e = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2 = S_T - S_A = 675.2714 - 405.5343 = 269.7371,$$

方差分析表

来源	平方和	自由度	均方和	F 比	p 值
因子 A	405.5343	4	101.3836	11.2758	1.0527×10^{-5}
误差 e	269.7371	30	8.9912		
和 T	675.2714	34			

有$F = \dfrac{S_A/f_A}{S_e/f_e} = \dfrac{405.5343/4}{269.7371/30} = \dfrac{101.3836}{8.9912} = 11.2758 \in W$，

并且检验的p值$p = P\{F \ge 11.2758\} = 1.0527\times 10^{-5} < \alpha = 0.05$，

故拒绝 H$_0$，接受 H$_1$，可以认为因子A显著，即五种方法在平均月推销额上有显著差异；

（2）因平均月推销额μ_i的点估计为$\overline{Y}_{i\cdot}$，

有 $\hat{\mu}_1 = \overline{y}_{1\cdot} = \dfrac{T_1}{m} = \dfrac{149}{7} = 21.2857$，$\hat{\mu}_2 = \overline{y}_{2\cdot} = \dfrac{T_2}{m} = \dfrac{172.1}{7} = 24.5857$，

$\hat{\mu}_3 = \overline{y}_{3\cdot} = \dfrac{T_3}{m} = \dfrac{143.9}{7} = 20.5571$，$\hat{\mu}_4 = \overline{y}_{4\cdot} = \dfrac{127.6}{7} = 18.2286$，$\hat{\mu}_5 = \overline{y}_{5\cdot} = \dfrac{195.9}{7} = 27.9857$，

即 $\hat{\mu}_4 < \hat{\mu}_3 < \hat{\mu}_1 < \hat{\mu}_2 < \hat{\mu}_5$，从点估计来看，第 5 种推销方法的效果最好，

估计μ_i，选取枢轴量$T = \dfrac{\overline{Y}_{i\cdot} - \mu_i}{\hat{\sigma}/\sqrt{m}} \sim t(f_e)$，其中$\hat{\sigma} = \sqrt{\dfrac{S_e}{f_e}}$，置信区间为$(\overline{Y}_{i\cdot} \pm t_{1-\alpha/2}(f_e) \cdot \dfrac{\hat{\sigma}}{\sqrt{m}})$，

置信水平 $1-\alpha = 0.95$，$t_{1-\alpha/2}(f_e) = t_{0.975}(30) = 2.0423$，$\hat{\sigma} = \sqrt{\dfrac{S_e}{f_e}} = \sqrt{8.9912} = 2.9985$，$m = 7$，

故μ_5的 0.95 置信区间为

$$(\overline{y}_{5\cdot} \pm t_{1-\alpha/2}(f_e) \cdot \frac{\hat{\sigma}}{\sqrt{m}}) = (27.9857 \pm 2.0423 \times \frac{2.9985}{\sqrt{7}}) = (25.6711, 30.3003) .$$

习题 8.2

1． 采用习题 8.1 中第 7 题的数据，对三种储藏方法的平均含水率在$\alpha = 0.05$下作多重比较.

解：因子 A 显著且三个水平下重复数相等，用 T 法检验，

假设H_0^{ij}：$\mu_i = \mu_j,\quad i, j = 1, 2, 3$，

选取 t 化极差统计量$q(r, f_e) = \max\limits_i\{\dfrac{\overline{Y}_{i\cdot} - \mu}{\hat{\sigma}/\sqrt{m}}\} - \min\limits_j\{\dfrac{\overline{Y}_{j\cdot} - \mu}{\hat{\sigma}/\sqrt{m}}\}$，

显著性水平$\alpha = 0.05$，$q_{1-\alpha}(r, f_e) = q_{0.95}(3, 12) = 3.77$，$\hat{\sigma} = \sqrt{\dfrac{S_e}{f_e}} = \sqrt{\dfrac{8.236}{12}} = 0.8285$，$m = 5$，

有$c = q_{1-\alpha}(r, f_e) \cdot \hat{\sigma}/\sqrt{m} = 3.77 \times 0.8285/\sqrt{5} = 1.3968$，

拒绝域$W = \{q(r, f_e) \ge 3.77\} = \{|\overline{y}_{i\cdot} - \overline{y}_{j\cdot}| \ge 1.3968\}$，

因$\overline{y}_{1\cdot} = \dfrac{T_1}{m} = \dfrac{39.9}{5} = 7.98$，$\overline{y}_{2\cdot} = \dfrac{T_2}{m} = \dfrac{32}{5} = 6.4$，$\overline{y}_{3\cdot} = \dfrac{T_3}{m} = \dfrac{45.6}{5} = 9.12$，

则$|\overline{y}_{1\cdot} - \overline{y}_{2\cdot}| = |7.98 - 6.4| = 1.58 > 1.3968$，可以认为$\mu_1$与$\mu_2$有显著差异，

$|\overline{y}_{1\cdot} - \overline{y}_{3\cdot}| = |7.98 - 9.12| = 1.14 < 1.3968$，可以认为$\mu_1$与$\mu_3$没有显著差异，

$|\overline{y}_{2\cdot} - \overline{y}_{3\cdot}| = |6.4 - 9.12| = 2.72 > 1.3968$，可以认为$\mu_2$与$\mu_3$有显著差异，

故可以认为第一种与第三种储藏方法在粮食的含水率方面差别不明显，但它们与第二种储藏方法的差别显著.

2． 采用习题 8.1 中第 8 题的数据，对五种推销方法在$\alpha = 0.05$下作多重比较.

解：因子 A 显著且五个水平下重复数相等，用 T 法检验，

假设H_0^{ij}：$\mu_i=\mu_j$, $i,j=1,2,3,4,5$，

选取 t 化极差统计量$q(r,f_e)=\max\limits_i\{\dfrac{\overline{Y}_{i\cdot}-\mu}{\hat{\sigma}/\sqrt{m}}\}-\min\limits_j\{\dfrac{\overline{Y}_{j\cdot}-\mu}{\hat{\sigma}/\sqrt{m}}\}$，

显著性水平$\alpha=0.05$，$q_{1-\alpha}(r,f_e)=q_{0.95}(5,30)=4.10$，$\hat{\sigma}=\sqrt{\dfrac{S_e}{f_e}}=\sqrt{\dfrac{269.7371}{30}}=2.9985$，$m=7$，

有$c=q_{1-\alpha}(r,f_e)\cdot\hat{\sigma}/\sqrt{m}=4.10\times2.9985/\sqrt{7}=4.6467$，

拒绝域$W=\{q(r,f_e)\ge4.10\}=\{|\bar{y}_{i\cdot}-\bar{y}_{j\cdot}|\ge4.6467\}$，

因$\bar{y}_{1\cdot}=\dfrac{T_1}{m}=\dfrac{149}{7}=21.2857$，$\bar{y}_{2\cdot}=\dfrac{T_2}{m}=\dfrac{172.1}{7}=24.5857$，$\bar{y}_{3\cdot}=\dfrac{T_3}{m}=\dfrac{143.9}{7}=20.5571$，
$\bar{y}_{4\cdot}=\dfrac{T_4}{m}=\dfrac{127.6}{7}=18.2286$，$\bar{y}_{5\cdot}=\dfrac{T_5}{m}=\dfrac{195.9}{7}=27.9857$，

则$|\bar{y}_{1\cdot}-\bar{y}_{2\cdot}|=|21.2857-24.5857|=3.3<4.6467$，可以认为$\mu_1$与$\mu_2$没有显著差异，

$|\bar{y}_{1\cdot}-\bar{y}_{3\cdot}|=|21.2857-20.5571|=0.7286<4.6467$，可以认为$\mu_1$与$\mu_3$没有显著差异，

$|\bar{y}_{1\cdot}-\bar{y}_{4\cdot}|=|21.2857-18.2286|=3.0571<4.6467$，可以认为$\mu_1$与$\mu_4$没有显著差异，

$|\bar{y}_{1\cdot}-\bar{y}_{5\cdot}|=|21.2857-27.9857|=6.7>4.6467$，可以认为$\mu_1$与$\mu_5$有显著差异，

$|\bar{y}_{2\cdot}-\bar{y}_{3\cdot}|=|24.5857-20.5571|=4.0286<4.6467$，可以认为$\mu_2$与$\mu_3$没有显著差异，

$|\bar{y}_{2\cdot}-\bar{y}_{4\cdot}|=|24.5857-18.2286|=6.3571>4.6467$，可以认为$\mu_2$与$\mu_4$有显著差异，

$|\bar{y}_{2\cdot}-\bar{y}_{5\cdot}|=|24.5857-27.9857|=3.4<4.6467$，可以认为$\mu_2$与$\mu_5$没有显著差异，

$|\bar{y}_{3\cdot}-\bar{y}_{4\cdot}|=|20.5571-18.2286|=2.3286<4.6467$，可以认为$\mu_3$与$\mu_4$没有显著差异，

$|\bar{y}_{3\cdot}-\bar{y}_{5\cdot}|=|20.5571-27.9857|=7.4286>4.6467$，可以认为$\mu_3$与$\mu_5$有显著差异，

$|\bar{y}_{4\cdot}-\bar{y}_{5\cdot}|=|18.2286-27.9857|=9.7571>4.6467$，可以认为$\mu_4$与$\mu_5$有显著差异，

故可以认为第一种、第三种、第四种推销方法之间在平均月推销额上没有显著差异，它们都与第五种推销方法有显著差异，第二种与第五种推销方法没有显著差异，第二种、第一种、第三种推销方法之间也没有显著差异，但第二种与第四种推销方法有显著差异.

3．有七种人造纤维，每种抽 4 根测其强度，得每根纤维的平均强度及标准差如下：

i	1	2	3	4	5	6	7
$\bar{y}_{i\cdot}$	6.3	6.2	6.7	6.8	6.5	7.0	7.1
s_i	0.81	0.92	1.22	0.74	0.88	0.58	1.05

假定各种纤维的强度服从等方差的正态分布.

（1）试问七种纤维强度间有无显著差异（取$\alpha=0.05$）？

（2）若各种纤维的强度间无显著差异，则给出平均强度的置信水平为 0.95 的置信区间；若各种纤维的强度间有显著差异，请进一步在$\alpha = 0.05$下进行多重比较，并指出哪种纤维的平均强度最大，同时给出该种纤维平均强度的置信水平为 0.95 的置信区间.

解：（1）假设 H_0：$a_1 = a_2 = \cdots = a_7 = 0$，

选取统计量$F = \dfrac{S_A/f_A}{S_e/f_e} \sim F(f_A, f_e)$，

显著性水平$\alpha = 0.05$，$r = 7$，$m = 4$，$n = rm = 28$，有$f_A = r - 1 = 6$，$f_e = n - r = 21$，

则$F_{1-\alpha}(f_A, f_e) = F_{0.95}(6, 21) = 2.5727$，右侧拒绝域 $W = \{F \ge 2.5727\}$，

i	1	2	3	4	5	6	7	Σ
$\bar{y}_{i\cdot}$	6.3	6.2	6.7	6.8	6.5	7.0	7.1	
$T_i = m\bar{y}_{i\cdot}$	25.2	24.8	26.8	27.2	26	28	28.4	186.4
T_i^2	635.04	615.04	718.24	739.84	676	784	806.56	4974.72
s_i	0.81	0.92	1.22	0.74	0.88	0.58	1.05	
$Q_i = (m-1)s_i^2$	1.9683	2.5392	4.4652	1.6428	2.3232	1.0092	3.3075	17.2554

得$S_A = \sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m}(\bar{y}_{i\cdot} - \bar{y})^2 = \dfrac{1}{m}\sum\limits_{i=1}^{r}T_i^2 - \dfrac{1}{n}T^2 = \dfrac{1}{4}\times 4974.72 - \dfrac{1}{28}\times 186.4^2 = 2.7886$，

$$S_e = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2 = \sum_{i=1}^{r}Q_i = 17.2554，$$

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij} - \bar{y})^2 = S_A + S_e = 2..7886 + 17.2554 = 20.0440，$$

方差分析表

来源	平方和	自由度	均方和	F 比	p 值
因子 A	2.7886	6	0.4648	0.5656	0.7529
误差 e	17.2554	21	0.8217		
和 T	20.0440	27			

有$F = \dfrac{S_A/f_A}{S_e/f_e} = \dfrac{2.7886/6}{17.2554/21} = \dfrac{0.4648}{0.8217} = 0.5656 \notin W$，

并且检验的 p 值 $p = P\{F \ge 0.5656\} = 0.7529 > \alpha = 0.05$，

故接受 H_0，拒绝 H_1，可以认为因子 A 不显著，即七种纤维强度间没有显著差异；

（2）由于各种纤维的强度间无显著差异，可以认为所有样品来自于同一个总体，

未知σ^2，估计μ，选取枢轴量$T = \dfrac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t(n-1)$，置信区间为$[\bar{Y} \pm t_{1-\alpha/2}(n-1)\dfrac{S}{\sqrt{n}}]$，

置信度 $1 - \alpha = 0.95$，$n = 28$，$t_{1-\alpha/2}(n-1) = t_{0.975}(27) = 2.0518$，

$$\bar{y} = \frac{T}{n} = \frac{186.4}{28} = 6.6571，s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij} - \bar{y})^2} = \sqrt{\frac{S_T}{n-1}} = \sqrt{\frac{2.7886 + 17.2554}{27}} = 0.8616，$$

故μ的 95%置信区间为$[\bar{y} \pm t_{1-\alpha/2}(n-1)\dfrac{s}{\sqrt{n}}] = [6.6571 \pm 2.0518 \times \dfrac{0.8616}{\sqrt{28}}] = [6.3231,\ 6.9912]$.

4． 一位经济学家对生产电子计算机设备的企业收集了在一年内生产力提高指数(用 0 到 100 内的数表示）并按过去三年间在科研和开发上的平均花费分为三类：

A_1：花费少，　　A_2：花费中等，　　A_3：花费多.

生产力提高的指数如下表所示：

水平	生产力提高指数											
A_1	7.6	8.2	6.8	5.8	6.9	6.6	6.3	7.7	6.0			
A_2	6.7	8.1	9.4	8.6	7.8	7.7	8.9	7.9	8.3	8.7	7.1	8.4
A_3	8.5	9.7	10.1	7.8	9.6	9.5						

请列出方差分析表，并进行多重比较.（取$\alpha = 0.05$）

解：假设 H_0：$a_1 = a_2 = a_3 = 0$，

选取统计量$F = \dfrac{S_A/f_A}{S_e/f_e} \sim F(f_A, f_e)$，

取显著性水平$\alpha = 0.05$，$r = 3$，$m_1 = 9$，$m_2 = 12$，$m_3 = 6$，$n = m_1 + m_2 + m_3 =27$，

则$f_A = r - 1 = 2$，$f_e = n - r = 24$，有 $F_{1-\alpha}(f_A, f_e) = F_{0.95}(2, 24) = 3.4028$，右侧拒绝域 $W = \{F \ge 3.4028\}$，

水平	生产力提高指数												m_i	T_i	$\dfrac{T_i^2}{m_i}$	$\sum\limits_{j=1}^{m} y_{ij}^2$
A_1	7.6	8.2	6.8	5.8	6.9	6.6	6.3	7.7	6.0				9	61.9	425.7344	431.03
A_2	6.7	8.1	9.4	8.6	7.8	7.7	8.9	7.9	8.3	8.7	7.1	8.4	12	97.6	793.8133	800.12
A_3	8.5	9.7	10.1	7.8	9.6	9.5							6	55.2	507.84	511.6
Σ													27	214.7	1727.3878	1742.75

得$S_T = \sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m_i}(y_{ij} - \bar{y})^2 = \sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m_i} y_{ij}^2 - \dfrac{1}{n}T^2 = 1742.75 - \dfrac{1}{27} \times 214.7^2 = 35.4874$，

$$S_A = \sum_{i=1}^{r}\sum_{j=1}^{m_i}(\bar{y}_{i\cdot} - \bar{y})^2 = \sum_{i=1}^{r}\frac{T_i^2}{m_i} - \frac{1}{n}T^2 = 1727.3878 - \frac{1}{27} \times 214.7^2 = 20.1252\text{，}$$

$$S_e = \sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{i\cdot})^2 = S_T - S_A = 35.4874 - 20.1252 = 15.3622\text{，}$$

方差分析表

来源	平方和	自由度	均方和	F 比	p 值
因子 A	20.1252	2	10.0626	15.7205	4.3307×10^{-5}
误差 e	15.3622	24	0.6401		
和 T	35.4874	26			

有$F = \dfrac{S_A/f_A}{S_e/f_e} = \dfrac{20.1252/2}{15.3622/24} = \dfrac{10.0626}{0.6401} = 15.7205 \in W$，

并且检验的p 值 $p = P\{F \ge 15.7205\} = 4.3307 \times 10^{-5} < \alpha = 0.05$，

故拒绝 H_0，接受 H_1，可以认为因子 A 显著，即在科研和开发上的花费对生产力提高指数有显著影响；

因子 A 显著且三个水平下重复数不相等，用 S 法检验，

假设H_0^{ij}：$\mu_i=\mu_j$ vs H_1^{ij}：$\mu_i\neq\mu_j$ $i,j=1,2,3$，

选取统计量$F=\dfrac{1}{r-1}\max\limits_{1\le i<j\le r}F_{ij}=\dfrac{1}{r-1}\max\limits_{1\le i<j\le r}\dfrac{(\bar{y}_{i\cdot}-\bar{y}_{j\cdot})^2}{\left(\dfrac{1}{m_i}+\dfrac{1}{m_j}\right)\hat{\sigma}^2}\dot{\sim}F(r-1,f_e)$，

显著性水平$\alpha=0.05$，$F_{1-\alpha}(r-1,f_e)=F_{0.95}(2,24)=3.4028$，$\hat{\sigma}^2=\dfrac{S_e}{f_e}=\dfrac{15.3622}{24}=0.6401$，

且$m_1=9$，$m_2=12$，$m_3=6$，

有$c_{12}=\sqrt{(r-1)F_{1-\alpha}(r-1,f_e)\cdot\left(\dfrac{1}{m_1}+\dfrac{1}{m_2}\right)\hat{\sigma}^2}=\sqrt{2\times3.4028\times\left(\dfrac{1}{9}+\dfrac{1}{12}\right)\times0.6401}=0.9203$，

$c_{13}=\sqrt{(r-1)F_{1-\alpha}(r-1,f_e)\cdot\left(\dfrac{1}{m_1}+\dfrac{1}{m_3}\right)\hat{\sigma}^2}=\sqrt{2\times3.4028\times\left(\dfrac{1}{9}+\dfrac{1}{6}\right)\times0.6401}=1.1000$，

$c_{23}=\sqrt{(r-1)F_{1-\alpha}(r-1,f_e)\cdot\left(\dfrac{1}{m_2}+\dfrac{1}{m_3}\right)\hat{\sigma}^2}=\sqrt{2\times3.4028\times\left(\dfrac{1}{12}+\dfrac{1}{6}\right)\times0.6401}=1.0436$，

右侧拒绝域$W=\{F\ge3.4028\}=\{|\bar{y}_{1\cdot}-\bar{y}_{2\cdot}|\ge0.9203\}\cup\{|\bar{y}_{1\cdot}-\bar{y}_{3\cdot}|\ge1.1000\}\cup\{|\bar{y}_{2\cdot}-\bar{y}_{3\cdot}|\ge1.0436\}$，

因$\bar{y}_{1\cdot}=\dfrac{T_1}{m_1}=\dfrac{61.9}{9}=6.8778$，$\bar{y}_{2\cdot}=\dfrac{T_2}{m_2}=\dfrac{97.6}{12}=8.1333$，$\bar{y}_{3\cdot}=\dfrac{T_3}{m_3}=\dfrac{55.2}{6}=9.2$，

则$|\bar{y}_{1\cdot}-\bar{y}_{2\cdot}|=|6.8778-8.1333|=1.2556>0.9203$，可以认为$\mu_1$与$\mu_2$有显著差异，

$|\bar{y}_{1\cdot}-\bar{y}_{3\cdot}|=|6.8778-9.2|=2.3222>1.1000$，可以认为$\mu_1$与$\mu_3$有显著差异，

$|\bar{y}_{2\cdot}-\bar{y}_{3\cdot}|=|8.1333-9.2|=1.0667>1.0436$，可以认为$\mu_2$与$\mu_3$有显著差异，

故可以认为在科研和开发上花费的每一种水平下对生产力提高指数都有显著差异.

习题 8.3

1．采用例 8.1.1 的数据，在显著性水平$\alpha=0.05$下用 Hartley 检验考察三个总体方差是否彼此相等.

解：三个水平下重复数相等，用 Hartley 检验，假设 H$_0$：$\sigma_1^2=\sigma_2^2=\sigma_3^2$，

选取统计量$H=\dfrac{\max\{S_1^2,S_2^2,S_3^2\}}{\min\{S_1^2,S_2^2,S_3^2\}}$，

显著性水平$\alpha=0.05$，$r=3$，$m=8$，$H_{1-\alpha}(r,m-1)=H_{0.95}(3,7)=6.94$，右侧拒绝域$W=\{H\ge6.94\}$，

饲料 A	鸡重/g – 1000								T_i	T_i^2	$\sum\limits_{j=1}^{m}y_{ij}^2$
A_1	73	9	60	1	2	12	9	28	194	37636	10024
A_2	107	92	−10	109	90	74	122	1	585	342225	60355
A_3	93	29	80	21	22	32	29	48	354	125316	20984

因 $s_i^2=\dfrac{1}{m-1}\sum\limits_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2=\dfrac{1}{m-1}(\sum\limits_{i=1}^{m}y_{ij}^2-\dfrac{1}{m}T_i^2)$，

则 $s_1^2=\dfrac{1}{7}\times(10024-\dfrac{1}{8}\times37636)=759.9286$， $s_2^2=\dfrac{1}{7}\times(60355-\dfrac{1}{8}\times342225)=2510.9821$，

$s_3^2=\dfrac{1}{7}\times(20984-\dfrac{1}{8}\times125316)=759.9286$，

有 $H=\dfrac{\max\{s_1^2,s_2^2,s_3^2\}}{\min\{s_1^2,s_2^2,s_3^2\}}=\dfrac{2510.9821}{759.9286}=3.3042\notin W$，

故接受 H$_0$，拒绝 H$_1$，可以认为三个总体方差彼此相等.

2． 在安眠药试验（见习题 8.1 第 5 题）中已求得四个样本方差

$$s_1^2=0.02,\ s_2^2=0.08,\ s_3^2=0.036,\ s_4^2=0.1307\ .$$

请用 Hartley 检验在显著性水平 $\alpha=0.05$ 下考察四个总体方差是否彼此相等.

解：四个水平下重复数相等，用 Hartley 检验，

假设 H$_0$：$\sigma_1^2=\sigma_2^2=\sigma_3^2=\sigma_4^2$，

选取统计量 $H=\dfrac{\max\{S_1^2,S_2^2,S_3^2,S_4^2\}}{\min\{S_1^2,S_2^2,S_3^2,S_4^2\}}$，

显著性水平 $\alpha=0.05$，$r=4$，$m=6$，$H_{1-\alpha}(r,m-1)=H_{0.95}(4,5)=13.7$，右侧拒绝域 $W=\{H\ge13.7\}$，

因 $s_1^2=0.02,\ s_2^2=0.08,\ s_3^2=0.036,\ s_4^2=0.1307$，

有 $H=\dfrac{\max\{s_1^2,s_2^2,s_3^2,s_4^2\}}{\min\{s_1^2,s_2^2,s_3^2,s_4^2\}}=\dfrac{0.1307}{0.02}=6.535\notin W$，

故接受 H$_0$，拒绝 H$_1$，可以认为四个总体方差彼此相等.

3． 在生产力提高的指数研究中（见习题 8.2 第 4 题）已求得三个样本方差，它们是

$$s_1^2=0.663,\ s_2^2=0.574,\ s_3^2=0.752\ .$$

请用 Bartlett 检验在显著性水平 $\alpha=0.05$ 下考察三个总体方差是否彼此相等.

解：三个水平下重复数不相等，用 Bartlett 检验，

假设 H$_0$：$\sigma_1^2=\sigma_2^2=\sigma_3^2$，

选取统计量 $B=\dfrac{1}{C}(f_e\ln MS_e-\sum\limits_{i=1}^{r}f_i\ln S_i^2)\dot{\sim}\chi^2(r-1)$，其中 $C=1+\dfrac{1}{3(r-1)}\left(\sum\limits_{i=1}^{r}\dfrac{1}{f_i}-\dfrac{1}{f_e}\right)$，

显著性水平 $\alpha=0.05$，$r=3$，$\chi_{1-\alpha}^2(r-1)=\chi_{0.95}^2(2)=5.9915$，右侧拒绝域 $W=\{B\ge5.9915\}$，

因 $m_1=9$，$m_2=12$，$m_3=6$，$n=m_1+m_2+m_3=27$，
有 $f_1=m_1-1=8$，$f_2=m_2-1=11$，$f_3=m_3-1=5$，$f_e=n-r=24$，

则 $C=1+\dfrac{1}{3(r-1)}\left(\sum\limits_{i=1}^{r}\dfrac{1}{f_i}-\dfrac{1}{f_e}\right)=1+\dfrac{1}{3\times2}\times\left(\dfrac{1}{8}+\dfrac{1}{11}+\dfrac{1}{5}-\dfrac{1}{24}\right)=1.0624$，

且 $MS_e = \frac{1}{f_e}\sum_{i=1}^{r}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{i\cdot})^2 = \frac{1}{n-r}\sum_{i=1}^{r}(m_i-1)s_i^2 = \frac{1}{24}\times(8\times 0.663+11\times 0.574+5\times 0.752)=0.6408$，

有 $B=\frac{1}{1.0624}\times[24\times\ln 0.6408-(8\times\ln 0.663+11\times\ln 0.574+5\times\ln 0.752)]=0.1285\notin W$，

故接受 H$_0$，拒绝 H$_1$，可以认为三个总体方差彼此相等.

4. 在入户推销效果研究中（见习题 8.1 第 8 题），分别用 Hartley 检验和 Bartlett 检验在显著性水平 $\alpha = 0.05$ 下对五个总体作方差齐性检验.

解：用 Hartley 检验，

假设 H$_0$：$\sigma_1^2=\sigma_2^2=\sigma_3^2=\sigma_4^2=\sigma_5^2$，

选取统计量 $H=\frac{\max\{S_1^2,S_2^2,S_3^2,S_4^2,S_5^2\}}{\min\{S_1^2,S_2^2,S_3^2,S_4^2,S_5^2\}}$，

显著性水平 $\alpha = 0.05$，$r = 5$，$m = 7$，$H_{1-\alpha}(r, m-1) = H_{0.95}(5, 6) = 12.1$，右侧拒绝域 $W = \{H \ge 12.1\}$，

组别	推销额/千元							T_i	T_i^2	$\sum_{j=1}^{m} y_{ij}^2$
第一组	20.0	16.8	17.9	21.2	23.9	26.8	22.4	149	22201	3243.3
第二组	24.9	21.3	22.6	30.2	29.9	22.5	20.7	172.1	29618.41	4325.25
第三组	16.0	20.1	17.3	20.9	22.0	26.8	20.8	143.9	20707.21	3030.99
第四组	17.5	18.2	20.2	17.7	19.1	18.4	16.5	127.6	16281.76	2334.44
第五组	25.2	26.2	26.9	29.3	30.4	29.7	28.2	195.9	38376.81	5505.07
Σ								788.5	127185.19	18439.05

因 $s_i^2=\frac{1}{m-1}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2=\frac{1}{m-1}(\sum_{i=1}^{m}y_{ij}^2-\frac{1}{m}T_i^2)$，

则 $s_1^2=\frac{1}{6}\times(3243.3-\frac{1}{7}\times 22201)=11.9548$，$s_2^2=\frac{1}{6}\times(4325.25-\frac{1}{7}\times 29618.41)=15.6748$，

$s_3^2=\frac{1}{6}\times(3030.99-\frac{1}{7}\times 20707.21)=12.1362$，$s_4^2=\frac{1}{6}\times(2334.44-\frac{1}{7}\times 16281.76)=1.4124$，

$s_5^2=\frac{1}{6}\times(5505.07-\frac{1}{7}\times 38376.81)=3.7781$，

有 $H=\frac{\max\{s_1^2,s_2^2,s_3^2,s_4^2,s_5^2\}}{\min\{s_1^2,s_2^2,s_3^2,s_4^2,s_5^2\}}=\frac{15.6748}{1.4124}=11.0981\notin W$，

故接受 H$_0$，拒绝 H$_1$，可以认为五个总体方差彼此相等；

用 Bartlett 检验，

假设 H$_0$：$\sigma_1^2=\sigma_2^2=\sigma_3^2=\sigma_4^2=\sigma_5^2$ vs H$_1$：$\sigma_1^2,\sigma_2^2,\sigma_3^2,\sigma_4^2,\sigma_5^2$ 不全相等，

选取统计量 $B=\frac{1}{C}(f_e\ln MS_e-\sum_{i=1}^{r}f_i\ln S_i^2)\overset{\cdot}{\sim}\chi^2(r-1)$，其中 $C=1+\frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{f_i}-\frac{1}{f_e}\right)$，

显著性水平 $\alpha = 0.05$，$r = 5$，$\chi_{1-\alpha}^2(r-1)=\chi_{0.95}^2(4)=9.4877$，右侧拒绝域 $W = \{B \ge 9.4877\}$，

因 $m_1 = m_2 = m_3 = m_4 = m_5 = m = 7$，$n = rm = 35$，有 $f_1 = f_2 = f_3 = f_4 = f_5 = m - 1 = 6$，$f_e = n - r = 30$，

则$C = 1 + \dfrac{1}{3(r-1)}\left(\sum_{i=1}^{r}\dfrac{1}{f_i} - \dfrac{1}{f_e}\right) = 1 + \dfrac{1}{3\times 4}\times\left(\dfrac{1}{6}+\dfrac{1}{6}+\dfrac{1}{6}+\dfrac{1}{6}+\dfrac{1}{6}-\dfrac{1}{30}\right) = 1.0667$，

$$MS_e = \frac{1}{f_e}\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2 = \frac{1}{n-r}(\sum_{i=1}^{r}\sum_{j=1}^{m}y_{ij}^2 - \frac{1}{m}\sum_{i=1}^{r}T_i^2) = \frac{1}{30}\times(18439.05 - \frac{1}{7}\times 127185.19) = 8.9912,$$

有$B = \dfrac{1}{1.0667}\times[30\times\ln 8.9912 - 6\times(\ln 11.9548 + \ln 15.6748 + \ln 12.1362 + \ln 1.4124 + \ln 3.7781)]$

$= 8.8728 \notin W$，

故接受 H$_0$，拒绝 H$_1$，可以认为五个总体方差彼此相等；两种检验方法的结果是一致的.

5. 在对粮食含水率的研究中（见习题 8.1 第 7 题）已求得 3 个水平下的组内平方和：

$Q_1 = 1.148$，　$Q_2 = 2.237$，　$Q_3 = 2.407$.

请用修正的 Bartlett 检验在显著性水平$\alpha = 0.05$ 下考察三个总体方差是否彼此相等.

解：用修正的 Bartlett 检验，

假设 H$_0$：$\sigma_1^2 = \sigma_2^2 = \sigma_3^2$　vs　H$_1$：$\sigma_1^2, \sigma_2^2, \sigma_3^2$不全相等，

选取统计量$B' = \dfrac{n_2 BC}{n_1(A - BC)} \stackrel{\cdot}{\sim} F(n_1, n_2)$，$A = \dfrac{n_2}{2 - C + 2/n_2}$，

其中$B = \dfrac{1}{C}(f_e \ln MS_e - \sum_{i=1}^{r} f_i \ln S_i^2)$，$C = 1 + \dfrac{1}{3(r-1)}\left(\sum_{i=1}^{r}\dfrac{1}{f_i} - \dfrac{1}{f_e}\right)$，$n_1 = r - 1$，$n_2 = \dfrac{r+1}{(C-1)^2}$，

因 $m_1 = m_2 = m_3 = m = 5$，$n = rm = 15$，有$f_1 = f_2 = f_3 = m - 1 = 4$，$f_e = n - r = 12$，

则$C = 1 + \dfrac{1}{3(r-1)}\left(\sum_{i=1}^{r}\dfrac{1}{f_i} - \dfrac{1}{f_e}\right) = 1 + \dfrac{1}{3\times 2}\times\left(\dfrac{1}{4}+\dfrac{1}{4}+\dfrac{1}{4}-\dfrac{1}{12}\right) = 1.1111$，

$n_1 = r - 1 = 2$，$n_2 = \dfrac{r+1}{(C-1)^2} = \dfrac{3+1}{(1.1111-1)^2} = 324$，$A = \dfrac{324}{2 - 1.1111 + 2/324} = 361.9862$，

显著性水平$\alpha = 0.05$，$F_{1-\alpha}(n_1, n_2) = F_{0.95}(2, 324) = 3.0236$，右侧拒绝域 $W = \{B' \ge 3.0236\}$，

储藏方法	含水率数据					T_i	T_i^2	$\sum_{j=1}^{m} y_{ij}^2$
A_1	7.3	8.3	7.6	8.4	8.3	39.9	1592.01	319.39
A_2	5.4	7.4	7.1	6.8	5.3	32	1024	208.66
A_3	7.9	9.5	10.0	9.8	8.4	45.6	2079.36	419.26
Σ						117.5	4695.37	947.31

因$Q_i = \sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2 = \sum_{i=1}^{m} y_{ij}^2 - \dfrac{T_i^2}{m}$，$s_i^2 = \dfrac{Q_i}{f_i}$，

则$Q_1 = 319.39 - \dfrac{1}{5}\times 1592.01 = 0.988$，$s_1^2 = \dfrac{0.988}{4} = 0.247$，

$Q_2 = 208.66 - \dfrac{1}{5}\times 1024 = 3.86$，$s_2^2 = \dfrac{3.86}{4} = 0.965$，

$Q_3 = 419.26 - \dfrac{1}{5}\times 2079.36 = 3.388$，$s_3^2 = \dfrac{3.388}{4} = 0.847$，

且$MS_e=\frac{1}{f_e}\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2=\frac{1}{n-r}(\sum_{i=1}^{r}\sum_{j=1}^{m}y_{ij}^2-\frac{1}{m}\sum_{i=1}^{r}T_i^2)=\frac{1}{12}\times(947.31+\frac{1}{5}\times4695.37)=0.6863$，

有$B=\frac{1}{1.1111}\times[12\times\ln0.6863-4\times(\ln0.247+\ln0.965+\ln0.847)]=1.6951$，

得$B'=\frac{n_2BC}{n_1(A-BC)}=\frac{324\times1.6951\times1.1111}{2\times(361.9862-1.6951\times1.1111)}=0.8473\notin W$，

故接受 H_0，拒绝 H_1，可以认为三个总体方差彼此相等.

注：此题所给的 3 个水平下的组内平方和与原题的计算结果不一致，

原题可计算得 $Q_1=0.988$，$Q_2=3.86$，$Q_3=3.388$，此题是 $Q_1=1.148$，$Q_2=2.237$，$Q_3=2.407$，

若按此题条件应为：$s_1^2=\frac{Q_1}{f_1}=\frac{1.148}{4}=0.287$，$s_2^2=\frac{Q_2}{f_2}=\frac{2.237}{4}=0.5593$，$s_3^2=\frac{Q_3}{f_3}=\frac{2.407}{4}=0.6018$，

且$MS_e=\frac{1}{f_e}\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2=\frac{1}{n-r}\sum_{i=1}^{r}Q_i=\frac{1}{12}\times(1.148+2.237+2.407)=0.4827$，

有$B=\frac{1}{1.1111}\times[12\times\ln0.4827-4\times(\ln0.287+\ln0.5593+\ln0.6018)]=0.5474$，

得$B'=\frac{n_2BC}{n_1(A-BC)}=\frac{324\times0.5474\times1.1111}{2\times(361.9862-0.5474\times1.1111)}=0.2727\notin W$，

故接受 H_0，拒绝 H_1，可以认为三个总体方差彼此相等.

6．针对糖果包装研究的数据（见例 8.1.4），请用修正的 Bartlett 检验在显著性水平$\alpha=0.05$下考察四个总体方差是否满足方差齐性假定.

解：用修正的 Bartlett 检验，

假设 H_0：$\sigma_1^2=\sigma_2^2=\sigma_3^2=\sigma_4^2$，

选取统计量$B'=\frac{n_2BC}{n_1(A-BC)}\dot{\sim}F(n_1,n_2)$，$A=\frac{n_2}{2-C+2/n_2}$

其中$B=\frac{1}{C}(f_e\ln MS_e-\sum_{i=1}^{r}f_i\ln S_i^2)$，$C=1+\frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{f_i}-\frac{1}{f_e}\right)$，$n_1=r-1$，$n_2=\frac{r+1}{(C-1)^2}$，

因 $m_1=2$，$m_2=3$，$m_3=3$，$m_4=2$，$n=m_1+m_2+m_3+m_4=10$，

有 $f_1=m_1-1=1$，$f_2=m_2-1=2$，$f_3=m_3-1=2$，$f_4=m_4-1=1$，$f_e=n-r=6$，

则$C=1+\frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{f_i}-\frac{1}{f_e}\right)=1+\frac{1}{3\times3}\times\left(\frac{1}{1}+\frac{1}{2}+\frac{1}{2}+\frac{1}{1}-\frac{1}{6}\right)=1.3148$，

$n_1=r-1=3$，$n_2=\frac{r+1}{(C-1)^2}=\frac{4+1}{(1.3148-1)^2}=50.4498$，

$A=\frac{50.4498}{2-1.3148+2/50.4498}=69.6024$，

显著性水平$\alpha=0.05$，$F_{1-\alpha}(n_1,n_2)=F_{0.95}(3,50.4498)=2.7883$，右侧拒绝域 $W=\{B'\ge2.7883\}$，

包装类型	销售量数据			m_i	T_i	$\dfrac{T_i^2}{m_i}$	$\sum\limits_{j=1}^{m} y_{ij}^2$
A_1	12	18		2	30	450	468
A_2	14	12	13	3	39	507	509
A_3	19	17	21	3	57	1083	1091
A_4	24	30		2	54	1458	1476
Σ				10	180	3498	3544

因 $s_i^2=\dfrac{1}{m_i-1}\sum\limits_{j=1}^{m_i}(y_{ij}-\bar{y}_{i\cdot})^2=\dfrac{1}{m_i-1}(\sum\limits_{i=1}^{m_i}y_{ij}^2-\dfrac{T_i^2}{m_i})$，

则 $s_1^2=468-450=18$，$s_2^2=\dfrac{1}{2}\times(509-507)=1$，$s_3^2=\dfrac{1}{2}\times(1091-1083)=4$，$s_4^2=1476-1458=18$，

且 $MS_e=\dfrac{1}{f_e}\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2=\dfrac{1}{n-r}(\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m}y_{ij}^2-\sum\limits_{i=1}^{r}\dfrac{T_i^2}{m_i})=\dfrac{1}{6}\times(3544-3498)=7.6667$，

有 $B=\dfrac{1}{1.3148}\times[6\times\ln 7.6667-(1\times\ln 18+2\times\ln 1+2\times\ln 4+1\times\ln 18)]=2.7897$，

得 $B'=\dfrac{n_2BC}{n_1(A-BC)}=\dfrac{50.4498\times 2.7897\times 1.3148}{3\times(69.6024-2.7897\times 1.3148)}=0.9355\notin W$，

故接受 H_0，拒绝 H_1，可以认为四个总体方差满足方差齐性假定.

习题 8.4

1．假设回归直线过原点，即一元线性回归模型为

$$y_i=\beta x_i+\varepsilon_i,\quad i=1,\cdots,n,$$

$E(\varepsilon_i)=0$，$Var(\varepsilon_i)=\sigma^2$，诸观测值相互独立.

（1）写出 β,σ^2 的最小二乘估计；

（2）对给定的 x_0，其对应的因变量均值的估计为 $\hat{y}_0$，求 $Var(\hat{y}_0)$.

解：（1）取 $Q=\sum\varepsilon_i^2=\sum(y_i-\beta x_i)^2$，

令 $\dfrac{\partial Q}{\partial\beta}=\sum 2(y_i-\beta x_i)\cdot(-x_i)=0$，有 $\beta\sum x_i^2=\sum x_iy_i$，

故 β 的最小二乘估计为 $\hat{\beta}=\dfrac{\sum x_iy_i}{\sum x_i^2}$；

因 $y_i=\beta x_i+\varepsilon_i$，且 $E(\varepsilon_i)=0$，$Var(\varepsilon_i)=\sigma^2$，有 $E(y_i)=\beta x_i$，$Var(y_i)=\sigma^2$，

则 $E(\hat{\beta})=\dfrac{\sum x_i E(y_i)}{\sum x_i^2}=\dfrac{\sum x_i\cdot\beta x_i}{\sum x_i^2}=\dfrac{\beta\sum x_i^2}{\sum x_i^2}=\beta$，$Var(\hat{\beta})=\dfrac{\sum x_i^2 Var(y_i)}{(\sum x_i^2)^2}=\dfrac{\sum x_i^2\cdot\sigma^2}{(\sum x_i^2)^2}=\dfrac{\sigma^2}{\sum x_i^2}$，

设 $S_T=\sum y_i^2$，$S_R=\sum\hat{y}_i^2$，$S_e=\sum(y_i-\hat{y}_i)^2$，$\hat{y}_i=\hat{\beta}x_i$，

因 $\hat{\beta}$ 是正规方程 $\sum(y_i-\beta x_i)x_i=0$ 的解，即 $\sum(y_i-\hat{\beta}x_i)x_i=0$，有 $\sum(y_i-\hat{\beta}x_i)\hat{y}_i=0$，

则$S_T=\sum(y_i-\hat{y}_i+\hat{y}_i)^2=\sum(y_i-\hat{y}_i)^2+\sum\hat{y}_i^2+2\sum(y_i-\hat{y}_i)\hat{y}_i=S_e+S_R$，

因$\mathrm{E}(\hat{y}_i)=\mathrm{E}(\hat{\beta})x_i=\beta x_i$， $\mathrm{Var}(\hat{y}_i)=\mathrm{Var}(\hat{\beta})x_i^2=\dfrac{x_i^2\sigma^2}{\sum x_i^2}$，

则$\mathrm{E}(S_e)=\sum\mathrm{E}(y_i^2)-\sum\mathrm{E}(\hat{y}_i^2)=\sum\{\mathrm{Var}(y_i)+[\mathrm{E}(y_i)]^2\}-\sum\{\mathrm{Var}(\hat{y}_i)+[\mathrm{E}(\hat{y}_i)]^2\}$

$$=\sum(\sigma^2+\beta^2x_i^2)-\sum(\frac{x_i^2\sigma^2}{\sum x_i^2}+\beta^2x_i^2)=n\sigma^2-\frac{\sum x_i^2\sigma^2}{\sum x_i^2}=(n-1)\sigma^2,$$

故σ^2的无偏估计为$\hat{\sigma}^2=\dfrac{S_e}{n-1}=\dfrac{1}{n-1}\sum(y_i-\hat{y}_i)^2$；

（2）因$\hat{y}_0=\hat{\beta}x_0$，故$\mathrm{Var}(\hat{y}_0)=x_0^2\,\mathrm{Var}(\hat{\beta})=\dfrac{x_0^2\sigma^2}{\sum x_i^2}$.

2． 设回归模型为

$$\begin{cases}y_i=\beta_0+\beta_1x_i+\varepsilon_i,\quad i=1,2,\cdots,n;\\ \text{各}\varepsilon_i\text{独立同分布，其分布为}N(0,\sigma^2).\end{cases}$$

试求β_0,β_1,σ^2的最大似然估计，它们与其最小二乘估计一致吗？

解：因$y_i\sim N(\beta_0+\beta_1x_i,\sigma^2)$，密度函数为$p_i(y_i)=\dfrac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(y_i-\beta_0-\beta_1x_i)^2}{2\sigma^2}}$，$i=1,2,\cdots,n$，

有$L(\beta_0,\beta_1,\sigma^2)=p_1(y_1)p_2(y_2)\cdots p_n(y_n)=\dfrac{1}{(\sqrt{2\pi\sigma^2})^n}\mathrm{e}^{-\frac{1}{2\sigma^2}\sum(y_i-\beta_0-\beta_1x_i)^2}$，

即$\ln L=-\dfrac{n}{2}\ln(2\pi)-\dfrac{n}{2}\ln(\sigma^2)-\dfrac{1}{2\sigma^2}\sum(y_i-\beta_0-\beta_1x_i)^2$，

$$\text{令}\begin{cases}\dfrac{\partial\ln L}{\partial\beta_0}=-\dfrac{1}{2\sigma^2}\sum2(y_i-\beta_0-\beta_1x_i)\cdot(-1)=-\dfrac{1}{\sigma^2}(n\beta_0+\beta_1\sum x_i-\sum y_i)=0,\\ \dfrac{\partial\ln L}{\partial\beta_1}=-\dfrac{1}{2\sigma^2}\sum2(y_i-\beta_0-\beta_1x_i)\cdot(-x_i)=-\dfrac{1}{\sigma^2}(\beta_0\sum x_i+\beta_1\sum x_i^2-\sum x_iy_i)=0,\\ \dfrac{\partial\ln L}{\partial\sigma^2}=-\dfrac{n}{2}\cdot\dfrac{1}{\sigma^2}+\dfrac{1}{2\sigma^4}\sum(y_i-\beta_0-\beta_1x_i)^2=-\dfrac{1}{2\sigma^4}[n\sigma^2-\sum(y_i-\beta_0-\beta_1x_i)^2]=0.\end{cases}$$

$$\text{即}\begin{cases}n\beta_0+\beta_1\sum x_i=\sum y_i,\\ \beta_0\sum x_i+\beta_1\sum x_i^2=\sum x_iy_i,\\ n\sigma^2=\sum(y_i-\beta_0-\beta_1x_i)^2,\end{cases}$$

故$\hat{\beta}_1=\dfrac{\sum x_iy_i-n\overline{xy}}{\sum x_i^2-n\bar{x}^2}=\dfrac{l_{xy}}{l_{xx}}$，$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$，$\hat{\sigma}_M^2=\dfrac{1}{n}\sum(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2=\dfrac{1}{n}\sum(y_i-\hat{y}_i)^2=\dfrac{S_e}{n}$，

显然β_0,β_1的最大似然估计与其最小二乘估计一致；

因$\mathrm{E}(\hat{\sigma}_M^2)=\dfrac{\mathrm{E}(S_e)}{n}=\dfrac{(n-2)}{n}\sigma^2\neq\sigma^2$，即$\sigma^2$的最大似然估计$\hat{\sigma}_M^2$不是$\sigma^2$的无偏估计，

故可取$\hat{\sigma}^2=\dfrac{S_e}{n-2}$，$\hat{\sigma}^2$是$\sigma^2$的无偏估计.

3． 在回归分析计算中，常对数据进行变换：

$$\widetilde{y}_i = \frac{y_i - c_1}{d_1},\ \widetilde{x}_i = \frac{x_i - c_2}{d_2},\quad i = 1, \cdots, n\,,$$

其中 $c_1, c_2, d_1\ (d_1 > 0), d_2\ (d_2 > 0)$ 是适当选取的常数.

（1）试建立由原始数据和变换后数据得到的最小二乘估计、总平方和、回归平方和以及残差平方和之间的关系.

（2）证明：由原始数据和变换后数据得到的 F 检验统计量的值保持不变.

解：（1）因 $\bar{\widetilde{x}} = \dfrac{1}{n}\sum\dfrac{x_i - c_2}{d_2} = \dfrac{1}{n}\cdot\dfrac{\sum x_i - nc_2}{d_2} = \dfrac{\bar{x} - c_2}{d_2}$， $\bar{\widetilde{y}} = \dfrac{1}{n}\sum\dfrac{y_i - c_1}{d_1} = \dfrac{1}{n}\cdot\dfrac{\sum y_i - nc_1}{d_1} = \dfrac{\bar{y} - c_1}{d_1}$，

$$l_{\widetilde{x}\widetilde{x}} = \sum(\widetilde{x}_i - \bar{\widetilde{x}})^2 = \sum\left(\frac{x_i - c_2}{d_2} - \frac{\bar{x} - c_2}{d_2}\right)^2 = \frac{1}{d_2^2}\sum(x_i - \bar{x})^2 = \frac{1}{d_2^2}l_{xx}\,,$$

$$l_{\widetilde{x}\widetilde{y}} = \sum(\widetilde{x}_i - \bar{\widetilde{x}})(\widetilde{y}_i - \bar{\widetilde{y}}) = \sum\left(\frac{x_i - c_2}{d_2} - \frac{\bar{x} - c_2}{d_2}\right)\left(\frac{y_i - c_1}{d_1} - \frac{\bar{y} - c_1}{d_1}\right) = \frac{1}{d_1 d_2}\sum(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{d_1 d_2}l_{xy}\,,$$

$$l_{\widetilde{y}\widetilde{y}} = \sum(\widetilde{y}_i - \bar{\widetilde{y}})^2 = \sum\left(\frac{y_i - c_1}{d_1} - \frac{\bar{y} - c_1}{d_1}\right)^2 = \frac{1}{d_1^2}\sum(y_i - \bar{y})^2 = \frac{1}{d_1^2}l_{yy}\,,$$

故 $$\tilde{\hat{\beta}}_1 = \frac{l_{\widetilde{x}\widetilde{y}}}{l_{\widetilde{x}\widetilde{x}}} = \frac{\frac{1}{d_1 d_2}l_{xy}}{\frac{1}{d_2^2}l_{xx}} = \frac{d_2}{d_1}\cdot\frac{l_{xy}}{l_{xx}} = \frac{d_2}{d_1}\hat{\beta}_1\,,$$

$$\tilde{\hat{\beta}}_0 = \bar{\widetilde{y}} - \tilde{\hat{\beta}}_1\bar{\widetilde{x}} = \frac{\bar{y} - c_1}{d_1} - \frac{d_2}{d_1}\hat{\beta}_1\cdot\frac{\bar{x} - c_2}{d_2} = \frac{\bar{y} - c_1 - \hat{\beta}_1\bar{x} + \hat{\beta}_1 c_2}{d_1} = \frac{\hat{\beta}_0 - c_1 + \hat{\beta}_1 c_2}{d_1}\,,$$

$$\widetilde{S}_T = \sum(\widetilde{y}_i - \bar{\widetilde{y}})^2 = l_{\widetilde{y}\widetilde{y}} = \frac{1}{d_1^2}l_{yy} = \frac{1}{d_1^2}S_T\,,$$

$$\widetilde{S}_R = \sum(\hat{\widetilde{y}}_i - \bar{\widetilde{y}})^2 = \tilde{\hat{\beta}}_1^2 l_{\widetilde{x}\widetilde{x}} = \left(\frac{d_2}{d_1}\hat{\beta}_1\right)^2\cdot\frac{1}{d_2^2}l_{xx} = \frac{1}{d_1^2}\hat{\beta}_1^2 l_{xx} = \frac{1}{d_1^2}S_R\,,$$

$$\widetilde{S}_e = \sum(\widetilde{y}_i - \hat{\widetilde{y}}_i)^2 = \widetilde{S}_T - \widetilde{S}_R = \frac{1}{d_1^2}S_T - \frac{1}{d_1^2}S_R = \frac{1}{d_1^2}(S_T - S_R) = \frac{1}{d_1^2}S_e\,;$$

（2）$\widetilde{F} = \dfrac{\widetilde{S}_R}{\widetilde{S}_e/(n-2)} = \dfrac{\frac{1}{d_1^2}S_R}{\frac{1}{d_1^2}S_e\Big/(n-2)} = \dfrac{S_R}{S_e/(n-2)} = F$，故 F 检验统计量的值保持不变.

4． 对给定的 n 组数据 $(x_i, y_i), i = 1, 2, \cdots, n$ ，若我们关心的是 y 如何依赖 x 的取值而变动，则可以建立如

下回归方程

$$\hat{y}=a+bx\ .$$

反之，若我们关心的是 x 如何依赖 y 的取值而变动，则可以建立另一个回归方程

$$\hat{x}=c+dy\ .$$

试问这两条直线在直角坐标系中是否重合？为什么？若不重合，它们有无交点？若有，试给出交点的坐标.

解：对于回归方程 $\hat{y}=a+bx$，有 $b=\dfrac{l_{xy}}{l_{xx}}$， $a=\overline{y}-b\overline{x}$，可化为 $y-\overline{y}=\dfrac{l_{xy}}{l_{xx}}(x-\overline{x})$，

而对于回归方程 $\hat{x}=c+dy$，有 $d=\dfrac{l_{xy}}{l_{yy}}$， $c=\overline{x}-d\overline{y}$，可化为 $x-\overline{x}=\dfrac{l_{xy}}{l_{yy}}(y-\overline{y})$，

当且仅当 $\dfrac{l_{xy}}{l_{xx}}\cdot\dfrac{l_{xy}}{l_{yy}}=1$，即 $l_{xy}^2=l_{xx}l_{yy}$ 时，这两条直线在直角坐标系中才重合，

因相关系数 $r^2=\dfrac{l_{xy}^2}{l_{xx}l_{yy}}$，且 $r^2=1$ 时，数据 $(x_i,y_i), i=1,2,\cdots,n$ 完全在一条直线上，

故相关系数 $r^2=1$，即数据 $(x_i,y_i), i=1,2,\cdots,n$ 完全在一条直线上时，这两条直线才重合；
一般情况下，这两条直线不重合，它们要过点 $(\overline{x},\overline{y})$，即它们有交点 $(\overline{x},\overline{y})$.

5. 为考察某种维尼纶纤维的耐水性能，安排了一组实验，测得其甲醛浓度 x 及相应的“缩醇化度” y 数据如下：

x	18	20	22	24	26	28	30
y	26.86	28.35	28.75	28.87	29.75	30.00	30.36

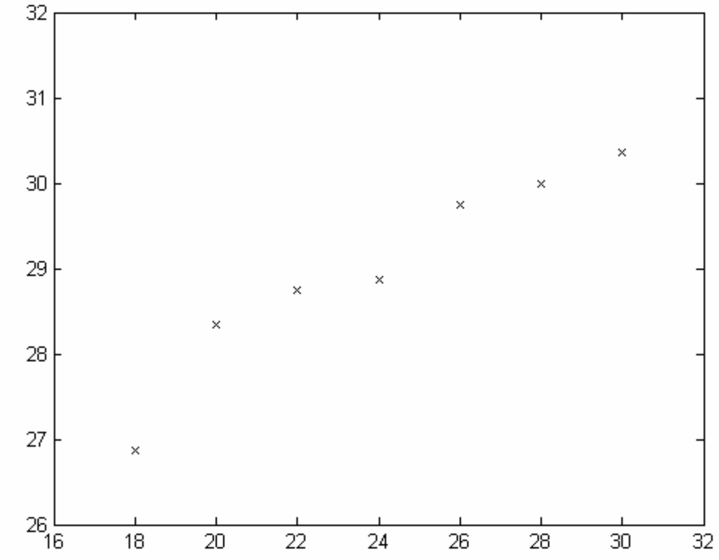
（1）作散点图；

（2）求样本相关系数；

（3）建立一元线性回归方程；

（4）对建立的回归方程作显著性检验（$\alpha=0.01$）.

解：（1）



从散点图中看到这些点大致在一条直线上；

（2）

x_i	18	20	22	24	26	28	30	168
y_i	26.86	28.35	28.75	28.87	29.75	30.00	30.36	202.94
x_i^2	324	400	484	576	676	784	900	4144
$x_i y_i$	483.48	567	632.5	692.88	773.5	840	910.8	4900.16
y_i^2	721.4596	803.7225	826.5625	833.4769	885.0625	900	921.7296	5892.0136

则$\bar{x}=\dfrac{1}{n}\sum x_i=\dfrac{1}{7}\times 168=24$，$\bar{y}=\dfrac{1}{n}\sum y_i=\dfrac{1}{7}\times 202.94=28.9914$，

$$l_{xx}=\sum x_i^2-\frac{1}{n}(\sum x_i)^2=4144-\frac{1}{7}\times 168^2=112\text{，}$$

$$l_{xy}=\sum x_i y_i-\frac{1}{n}\sum x_i\sum y_i=4900.16-\frac{1}{7}\times 168\times 202.94=29.6\text{，}$$

$$l_{yy}=\sum y_i^2-\frac{1}{n}(\sum y_i)^2=5892.0136-\frac{1}{7}\times 202.94^2=8.4931\text{，}$$

故样本相关系数$r=\dfrac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}=\dfrac{29.6}{\sqrt{112\times 8.4931}}=0.9597$；

（3）$\hat{\beta}_1=\dfrac{l_{xy}}{l_{xx}}=\dfrac{29.6}{112}=0.2643$，$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=28.9914-0.2643\times 24=22.6486$，

故一元线性回归方程为$\hat{y}=22.6486+0.2643x$；

（4）假设 H_0：$\beta_1=0$ vs H_1：$\beta_1\neq 0$，

选取统计量$F=\dfrac{S_R}{S_e/(n-2)}\sim F(1,n-2)$，

显著性水平$\alpha=0.01$，$F_{1-\alpha}(1,n-2)=F_{0.99}(1,5)=16.26$，拒绝域 $W=\{F\geq 16.26\}$，

因$S_R=\sum(\hat{y}_i-\bar{y})^2=\hat{\beta}_1^2 l_{xx}=0.2643^2\times 112=7.8229$，$S_T=\sum(y_i-\bar{y})^2=l_{yy}=8.4931$，

则 $S_e=S_T-S_R=8.4931-7.8229=0.6702$，有$F=\dfrac{S_R}{S_e/(n-2)}=\dfrac{7.8229}{0.6702/5}=58.3596\in W$，

故拒绝 H_0，接受 H_1，可以认为 x 与 y 线性关系显著.

也可进行相关系数检验，假设 H_0：$\beta_1=0$ vs H_1：$\beta_1\neq 0$，

选取统计量$r=\dfrac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}$，

显著性水平$\alpha=0.01$，$r_{1-\alpha}(n-2)=r_{0.99}(5)=0.874$，拒绝域 $W=\{|r|\geq 0.874\}$，

因样本相关系数$r=\dfrac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}=\dfrac{29.6}{\sqrt{112\times 8.4931}}=0.9597\in W$，

故拒绝 H_0，接受 H_1，可以认为 x 与 y 线性关系显著.

6． 测得一组弹簧形变 x（单位：cm）和相应的外力 y（单位：N）数据如下：

y	1	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.8	3.0
x	3.08	3.76	4.31	5.02	5.51	6.25	6.74	7.40	8.54	9.24

由胡克定律知 $y = kx$，试估计 k，并在 $x = 2.6\,\text{cm}$ 试给出相应的外力 y 的 0.95 预测区间.

解：因 k 的最小二乘估计为 $\hat{k} = \dfrac{\sum x_i y_i}{\sum x_i^2} \sim N(k, \dfrac{\sigma^2}{\sum x_i^2})$，有 $\hat{y} = \hat{k}x \sim N(kx, \dfrac{x^2\sigma^2}{\sum x_i^2})$，

且 $y = kx + \varepsilon \sim N(kx, \sigma^2)$，$\hat{y}$ 与 y 相互独立，有 $y - \hat{y} \sim N(0, (1 + \dfrac{x^2}{\sum x_i^2})\sigma^2)$，

则 $\dfrac{y - \hat{y}}{\sigma\sqrt{1 + \dfrac{x^2}{\sum x_i^2}}} \sim N(0, 1)$，

因 $\dfrac{S_e}{\sigma^2} = \dfrac{\sum (y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi^2(n-1)$，得 $\dfrac{y - \hat{y}}{\sqrt{\dfrac{S_e}{n-1}}\sqrt{1 + \dfrac{x^2}{\sum x_i^2}}} \sim t(n-1)$，

则 y 的 $1 - \alpha$ 预测区间为 $(\hat{y} \pm t_{1-\alpha/2}(n-1) \cdot \sqrt{\dfrac{S_e}{n-1}}\sqrt{1 + \dfrac{x^2}{\sum x_i^2}})$，

x_i	3.08	3.76	4.31	5.02	5.51	6.25	6.74	7.40	8.54	9.24	59.85
y_i	1	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.8	3.0	19.4
x_i^2	9.4864	14.1376	18.5761	25.2004	30.3601	39.0625	45.4276	54.76	72.9316	85.3776	395.3199
$x_i y_i$	3.08	4.512	6.034	8.032	9.918	12.5	14.828	17.76	23.912	27.72	128.296
y_i^2	1	1.44	1.96	2.56	3.24	4	4.84	5.76	7.84	9	41.64

故 k 的最小二乘估计为 $\hat{k} = \dfrac{\sum x_i y_i}{\sum x_i^2} = \dfrac{128.296}{395.3199} = 0.3245$；

因 $S_e = \sum (y_i - \hat{y}_i)^2 = \sum y_i^2 - \sum \hat{y}_i^2 = \sum y_i^2 - \hat{k}^2 \sum x_i^2 = 41.64 - 0.3245^2 \times 395.3199 = 0.0032$，

且 $1 - \alpha = 0.95$，$t_{1-\alpha/2}(n - 1) = t_{0.975}(9) = 2.2622$，

故在 $x = 2.6\,\text{cm}$ 时，$\hat{y} = \hat{k}x = 0.3245 \times 2.6 = 0.8438$，$y$ 的 $1 - \alpha$ 预测区间为

$$(\hat{y} \pm t_{1-\alpha/2}(n-1) \cdot \sqrt{\frac{S_e}{n-1}}\sqrt{1 + \frac{x^2}{\sum x_i^2}}) = (0.8438 \pm 2.2622 \times \sqrt{\frac{0.0032}{9}} \times \sqrt{1 + \frac{2.6^2}{395.3199}})$$

$$= (0.8009,\ 0.8867).$$

7． 设由 (x_i, y_i), $i = 1, \cdots, n$ 可建立一元线性回归方程，$\hat{y}_i$ 是由回归方程得到的拟合值，证明样本相关系数 r 满足如下关系

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2},$$

上式也称为回归方程的决定系数.

证：因样本相关系数$r=\dfrac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}$，

故$r^2=\dfrac{l_{xy}^2}{l_{xx}l_{yy}}=\dfrac{\left(\dfrac{l_{xy}}{l_{xx}}\right)^2\cdot l_{xx}}{l_{yy}}=\dfrac{\hat\beta_1^2 l_{xx}}{l_{yy}}=\dfrac{S_R}{S_T}=\dfrac{\sum\limits_{i=1}^n(\hat y_i-\bar y)^2}{\sum\limits_{i=1}^n(y_i-\bar y)^2}$.

8．现收集了 16 组合金钢的碳含量x及强度y的数据，求得

$\bar x=0.125,\ \bar y=45.7886,\ l_{xx}=0.3024,\ l_{xy}=25.5218,\ l_{yy}=2432.4566$.

（1）建立y关于x的一元线性回归方程$\hat y=\hat\beta_0+\hat\beta_1 x$；

（2）写出$\hat\beta_0$和$\hat\beta_1$的分布；

（3）求$\hat\beta_0$和$\hat\beta_1$的相关系数；

（4）列出对回归方程做显著性检验的方差分析表（$\alpha=0.05$）；

（5）给出β_1的 0.95 置信区间；

（6）在$x=0.15$时求对应的y的 0.95 预测区间.

解：（1）因$\hat\beta_1=\dfrac{l_{xy}}{l_{xx}}=\dfrac{25.5218}{0.3024}=84.3975$，$\hat\beta_0=\bar y-\hat\beta_1\bar x=45.7886-84.3975\times0.125=35.2389$，

故y关于x的一元线性回归方程为$\hat y=35.2389+84.3975x$；

（2）因$\hat\beta_0\sim N(\beta_0,\left(\dfrac{1}{n}+\dfrac{\bar x^2}{l_{xx}}\right)\sigma^2)$，$\hat\beta_1\sim N(\beta_1,\dfrac{\sigma^2}{l_{xx}})$，$\dfrac{1}{n}+\dfrac{\bar x^2}{l_{xx}}=\dfrac{1}{16}+\dfrac{0.125^2}{0.3024}=0.1142$，$\dfrac{1}{l_{xx}}=3.3069$，

故$\hat\beta_0\sim N(\beta_0,0.1142\sigma^2)$，$\hat\beta_1\sim N(\beta_1,3.3069\sigma^2)$；

（3）因$\mathrm{Cov}(\hat\beta_0,\hat\beta_1)=-\dfrac{\bar x}{l_{xx}}\sigma^2=-\dfrac{0.125}{0.3024}\sigma^2=-0.4134\sigma^2$，$\mathrm{Var}(\hat\beta_0)=0.1142\sigma^2$，$\mathrm{Var}(\hat\beta_1)=3.3069\sigma^2$，

故$\hat\beta_0$和$\hat\beta_1$的相关系数$\mathrm{Corr}(\hat\beta_0,\hat\beta_1)=\dfrac{\mathrm{Cov}(\hat\beta_0,\hat\beta_1)}{\sqrt{\mathrm{Var}(\hat\beta_0)}\sqrt{\mathrm{Var}(\hat\beta_1)}}=\dfrac{-0.4134\sigma^2}{\sqrt{0.1142\sigma^2}\sqrt{3.3069\sigma^2}}=-0.6727$；

（4）假设 H$_0$：$\beta_1=0$ vs H$_1$：$\beta_1\neq0$，

选取统计量$F=\dfrac{S_R}{S_e/(n-2)}\sim F(1,n-2)$，

显著性水平$\alpha=0.05$，$n=16$，$F_{1-\alpha}(1,n-2)=F_{0.95}(1,14)=4.60$，右侧拒绝域$W=\{F\ge4.60\}$，

因$S_T=\sum(y_i-\bar y)^2=l_{yy}=2432.4566$，自由度为$n-1=15$，

$S_R=\sum(\hat y_i-\bar y)^2=\hat\beta_1^2 l_{xx}=84.3975^2\times0.3024=2153.9758$，自由度为 1，

$S_e=\sum(y_i-\hat y_i)^2=S_T-S_R=2432.4566-2153.9758=278.4808$，自由度为$n-2=14$，

方差分析表

来源	平方和	自由度	均方和	F 比	p 值
回归 R	2153.9758	1	2153.9758	108.2863	5.6929×10^{-8}
误差 e	278.4808	14	19.8915		
和 T	2432.4566	15			

有 $F=\dfrac{S_R}{S_e/(n-2)}=\dfrac{2153.8758}{278.4808/14}=108.2863\in W$，

并且检验的 p 值 $p=P\{F\ge108.2863\}=5.6929\times10^{-8}<\alpha=0.05$，
故拒绝 H$_0$，接受 H$_1$，可以认为回归方程显著；

（5）因 $\hat\beta_1\sim N(\beta_1,\dfrac{\sigma^2}{l_{xx}})$，有 $\dfrac{\hat\beta_1-\beta_1}{\sigma/\sqrt{l_{xx}}}\sim N(0,1)$，且 $\dfrac{S_e}{\sigma^2}=\dfrac{\sum(y_i-\hat y_i)^2}{\sigma^2}\sim\chi^2(n-2)$，

则 $\dfrac{\hat\beta_1-\beta_1}{\sqrt{\dfrac{S_e}{n-2}}\Big/\sqrt{l_{xx}}}\sim t(n-2)$，有$\beta_1$ 的 $1-\alpha$ 置信区间为 $(\hat\beta_1\pm t_{1-\alpha/2}(n-2)\cdot\sqrt{\dfrac{S_e}{n-2}}\Big/\sqrt{l_{xx}})$，

显著性水平$\alpha=0.05$，$t_{1-\alpha/2}(n-2)=t_{0.975}(14)=2.1448$，
故β_1 的 0.95 置信区间为

$$(\hat\beta_1\pm t_{1-\alpha/2}(n-2)\cdot\sqrt{\frac{S_e}{n-2}}\Big/\sqrt{l_{xx}})=(84.3975\pm2.1448\times\sqrt{\frac{278.4808}{14}}\Big/\sqrt{0.3024})$$
$$=(67.0023,\ 101.7927);$$

（6）因 $y=\beta_0+\beta_1x+\varepsilon$ 的 $1-\alpha$ 预测区间为 $(\hat y\pm t_{1-\alpha/2}(n-2)\cdot\sqrt{\dfrac{S_e}{n-2}}\cdot\sqrt{1+\dfrac{1}{n}+\dfrac{(x-\bar x)^2}{l_{xx}}})$，

且 $1-\alpha=0.95$，$t_{1-\alpha/2}(n-2)=t_{0.975}(14)=2.1448$，

故在 $x=0.15$ 时，$\hat y=\hat\beta_0+\hat\beta_1x=35.2389+84.3975\times0.15=47.8985$，$y$ 的 $1-\alpha$ 预测区间为

$$(\hat y\pm t_{1-\alpha/2}(n-2)\cdot\sqrt{\frac{S_e}{n-2}}\cdot\sqrt{1+\frac{1}{n}+\frac{(x-\bar x)^2}{l_{xx}}})$$

$$=(47.8985\pm2.1448\times\sqrt{\frac{278.4808}{14}}\cdot\sqrt{1+\frac{1}{16}+\frac{(0.15-0.125)^2}{0.3024}})=(38.0288,57.7683).$$

9. 设回归模型为$\begin{cases}y_i=\beta_0+\beta_1x_i+\varepsilon_i,\\ \varepsilon_i\sim N(0,\sigma^2),\end{cases}$ 现收集了 15 组数据，经计算有

$$\bar x=0.85,\ \bar y=25.60,\ l_{xx}=19.56,\ l_{xy}=32.54,\ l_{yy}=46.74,$$

后经核对，发现有一组数据记录错误，正确数据为 (1.2, 32.6)，记录为 (1.5, 32.3).

（1）求 $\hat\beta_0,\hat\beta_1$ 的 LSE；

（2）对回归方程做显著性检验（$\alpha=0.05$）；

（3）若 $x_0=1.1$，给出对应响应变量的 0.95 预测区间.

解：对计算的中间结果进行修正，

有 $\bar{x}=\frac{1}{n}\sum x_i=0.85+\frac{1}{15}\times(1.2-1.5)=0.83$，

$\bar{y}=\frac{1}{n}\sum y_i=25.60+\frac{1}{15}\times(32.6-32.3)=25.62$，

$l_{xx}=\sum x_i^2-n\bar{x}^2=19.56+(1.2^2-1.5^2)-15\times(0.83^2-0.85^2)=19.254$，

$l_{xy}=\sum x_i y_i-n\overline{xy}=32.54+(1.2\times32.6-1.5\times32.3)-15\times(0.83\times25.62-0.85\times25.60)=30.641$，

$l_{yy}=\sum y_i^2-n\bar{y}^2=46.74+(32.6^2-32.3^2)-15\times(25.62^2-25.60^2)=50.844$，

（1）$\hat{\beta}_1=\frac{l_{xy}}{l_{xx}}=\frac{30.641}{19.254}=1.5914$，$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=25.62-1.5914\times0.83=24.2991$；

（2）假设 H_0：$\beta_1=0$ vs H_1：$\beta_1\neq0$，

选取统计量 $F=\frac{S_R}{S_e/(n-2)}\sim F(1,n-2)$，

显著性水平 $\alpha=0.05$，$n=15$，$F_{1-\alpha}(1,n-2)=F_{0.95}(1,13)=4.6672$，右侧拒绝域 $W=\{F\geq4.6672\}$，

因 $S_T=\sum(y_i-\bar{y})^2=l_{yy}=50.844$，自由度为 $n-1=14$，

$S_R=\sum(\hat{y}_i-\bar{y})^2=\hat{\beta}_1^2 l_{xx}=1.5914^2\times19.254=48.7624$，自由度为 1，

$S_e=\sum(y_i-\hat{y}_i)^2=S_T-S_R=50.844-48.7624=2.0816$，自由度为 $n-2=13$，

方差分析表

来源	平方和	自由度	均方和	F 比	p 值
回归 R	48.7624	1	48.7624	304.5278	2.1063×10^{-10}
误差 e	2.0816	13	0.1601		
和 T	50.844	14			

有 $F=\frac{S_R}{S_e/(n-2)}=\frac{48.7624}{2.0816/13}=304.5278\in W$，

并且检验的 p 值 $p=P\{F\geq304.5278\}=2.1063\times10^{-10}<\alpha=0.05$，
故拒绝 H_0，接受 H_1，可以认为回归方程显著.

（3）因 $y_0=\beta_0+\beta_1x_0+\varepsilon$ 的 $1-\alpha$ 预测区间为 $(\hat{y}_0\pm t_{1-\alpha/2}(n-2)\cdot\sqrt{\frac{S_e}{n-2}}\cdot\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}})$，

且 $1-\alpha=0.95$，$t_{1-\alpha/2}(n-2)=t_{0.975}(13)=2.1604$，

故在 $x_0=1.1$ 时，$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_0=24.2991+1.5914\times1.1=26.0497$，$y_0$ 的 $1-\alpha$ 预测区间为

$$(\hat{y}_0\pm t_{1-\alpha/2}(n-2)\cdot\sqrt{\frac{S_e}{n-2}}\cdot\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}})$$

$$=(26.0497\pm2.1604\times\sqrt{\frac{2.0816}{13}}\cdot\sqrt{1+\frac{1}{15}+\frac{(1.1-0.83)^2}{19.254}})=(25.1552,26.9441).$$

10．在生产中积累了 32 组某种铸件在不同时间 x 下腐蚀深度 y 的数据，求得回归方程为

$$\hat{y} = -0.4441 + 0.002263x ,$$

且误差方差的无偏估计为 $\hat{\sigma}^2 = 0.001452$，总偏差平方和为 0.1246．

（1）对回归方程做显著性检验（$\alpha = 0.05$），列出方差分析表；

（2）求样本相关系数；

（3）若腐蚀时间 $x = 870$，试给出 y 的 0.95 近似预测区间．

解：（1）假设 H_0：$\beta_1 = 0$ vs H_1：$\beta_1 \neq 0$，

选取统计量 $F = \dfrac{S_R}{S_e/(n-2)} \sim F(1, n-2)$，

显著性水平 $\alpha = 0.05$，$n = 32$，$F_{1-\alpha}(1, n-2) = F_{0.95}(1, 30) = 4.17$，右侧拒绝域 $W = \{F \geq 4.17\}$，

因 $S_T = 0.1246$，自由度为 $n - 1 = 31$，且 $\hat{\sigma}^2 = \dfrac{S_e}{n-2} = 0.001452$，

有 $S_e = (n-2)\hat{\sigma}^2 = 30 \times 0.001452 = 0.04356$，自由度为 $n - 2 = 30$，

$S_R = S_T - S_e = 0.1246 - 0.04356 = 0.08104$，自由度为 1

方差分析表

来源	平方和	自由度	均方和	F 比	p 值
回归 R	0.08104	1	0.08104	55.8127	2.5101×10^{-8}
误差 e	0.04356	30	0.001452		
和 T	0.1246	31			

有 $F = \dfrac{S_R}{S_e/(n-2)} = \dfrac{0.08104}{0.01452} = 55.8127 \in W$，

并且检验的 p 值 $p = P\{F \geq 55.8127\} = 2.5101 \times 10^{-8} < \alpha = 0.05$，

故拒绝 H_0，接受 H_1，可以认为回归方程显著；

（2）因 $|r| = \dfrac{|l_{xy}|}{\sqrt{l_{xx}l_{yy}}} = \sqrt{\dfrac{S_R}{S_T}} = \sqrt{\dfrac{0.08104}{0.1246}} = 0.8065$，$\hat{\beta}_1 = \dfrac{l_{xy}}{l_{xx}} = 0.002263 > 0$，有 $l_{xy} = \hat{\beta}_1 l_{xx} > 0$，

故 $r = \dfrac{l_{xy}}{\sqrt{l_{xx}l_{yy}}} > 0$，即 $r = 0.8065$；

（3）因并且检验的 p 值 $p = P\{F \geq 304.5278\} = 2.1063 \times 10^{-10} < \alpha = 0.05$，

$y = \beta_0 + \beta_1 x + \varepsilon$ 的 $1 - \alpha$ 近似预测区间为 $(\hat{y} \pm u_{1-\alpha/2} \cdot \sqrt{\dfrac{S_e}{n-2}})$，

且 $1 - \alpha = 0.95$，$u_{1-\alpha/2} = u_{0.975} = 1.96$，

故在 $x = 870$ 时，$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -0.4441 + 0.002263 \times 870 = 1.5247$，$y$ 的 $1 - \alpha$ 近似预测区间为

$$(\hat{y} \pm u_{1-\alpha/2} \cdot \sqrt{\frac{S_e}{n-2}}) = (1.5247 \pm 1.96 \times \sqrt{\frac{0.04356}{30}}) = (1.4500, 1.5994) .$$

11．我们知道营业税税收总额 y 与社会商品零售总额 x 有关．为能从社会商品零售总额去预测税收总额，需要了解两者之间的关系．现收集了如下 9 组数据（单位：亿元）：

序号	社会商品零售总额	营业税税收总额
1	142.08	3.93
2	177.30	5.96
3	204.68	7.85
4	242.68	9.82
5	316.24	12.50
6	341.99	15.55
7	332.69	15.79
8	389.29	16.39
9	453..40	18.45

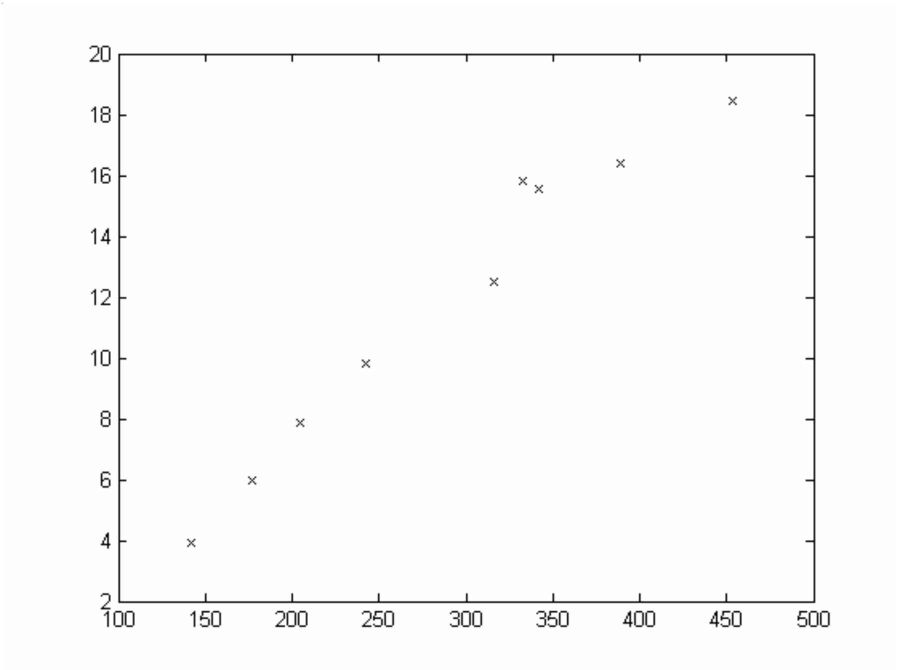
（1）画散点图；

（2）建立一元线性回归方程，并做显著性检验（取$\alpha = 0.05$），列出方差分析表；

（3）若已知某年社会商品零售总额为 300 亿元，试给出营业税税收总额的概率为 0.95 的预测区间；

（4）若已知回归直线过原点，试求回归方程，并在显著性水平 0.05 下作显著性检验.

解：（1）



从散点图中看到这些点大致在一条直线上；

（2）

序号	社会商品零售总额 x_i	营业税税收总额 y_i	x_i^2	x_iy_i	y_i^2
1	142.08	3.93	20186.7264	558.3744	15.4449
2	177.30	5.96	31435.2900	1056.7080	35.5216
3	204.68	7.85	41893.9024	1606.7380	61.6225
4	242.68	9.82	58893.5824	2383.1176	96.4324
5	316.24	12.50	100007.7376	3953.0000	156.2500
6	341.99	15.55	116957.1601	5317.9445	241.8025
7	332.69	15.79	110682.6361	5253.1751	249.3241
8	389.29	16.39	151546.7041	6380.4631	268.6321
9	453.40	18.45	205571.5600	8365.2300	340.4025
$\sum$	2600.35	106.24	837175.2991	34874.7507	1465.4326

则 $\bar{x}=\dfrac{1}{n}\sum x_i=\dfrac{1}{9}\times 2600.35=288.9278$，$\bar{y}=\dfrac{1}{n}\sum y_i=\dfrac{1}{9}\times 106.24=11.8044$，

$$l_{xx}=\sum x_i^2-\frac{1}{n}(\sum x_i)^2=837175.2991-\frac{1}{9}\times 2600.35^2=85861.9522,$$

$$l_{xy}=\sum x_i y_i-\frac{1}{n}\sum x_i\sum y_i=34874.7507-\frac{1}{9}\times 2600.35\times 106.24=4179.0636,$$

$$l_{yy}=\sum y_i^2-\frac{1}{n}(\sum y_i)^2=1465.4326-\frac{1}{9}\times 106.24^2=211.3284,$$

有 $\hat{\beta}_1=\dfrac{l_{xy}}{l_{xx}}=\dfrac{4179.0636}{85861.9522}=0.04867$，$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=11.8044-0.04867\times 288.9278=-2.2582$，

故一元线性回归方程为 $\hat{y}=-2.2582+0.04867x$；

假设 H_0：$\beta_1=0$ vs H_1：$\beta_1\neq 0$，

选取统计量 $F=\dfrac{S_R}{S_e/(n-2)}\sim F(1,n-2)$，

显著性水平 $\alpha=0.05$，$n=9$，$F_{1-\alpha}(1,n-2)=F_{0.95}(1,7)=5.59$，右侧拒绝域 $W=\{F\geq 5.59\}$，

因 $S_T=\sum(y_i-\bar{y})^2=l_{yy}=211.3284$，自由度为 $n-1=8$，

$S_R=\sum(\hat{y}_i-\bar{y})^2=\hat{\beta}_1^2 l_{xx}=0.04867^2\times 85861.9522=203.4029$，自由度为 $n-2=7$，

$S_e=\sum(y_i-\hat{y}_i)^2=S_T-S_R=211.3284-203.4029=7.9255$，自由度为 1，

方差分析表

来源	平方和	自由度	均方和	F 比	p 值
回归 R	203.4029	1	203.4029	179.6507	3.0172×10^{-6}
误差 e	7.9255	7	1.1322		
和 T	211.3284	8			

有 $F=\dfrac{S_R}{S_e/(n-2)}=\dfrac{203.4029}{7.9255/7}=179.6507\in W$，

并且检验的 p 值 $p=P\{F\geq 179.6507\}=3.0172\times 10^{-6}<\alpha=0.05$，
故拒绝 H_0，接受 H_1，可以认为回归方程显著；

（3）因 $y=\beta_0+\beta_1 x+\varepsilon$ 的 $1-\alpha$ 预测区间为 $\left(\hat{y}\pm t_{1-\alpha/2}(n-2)\cdot\sqrt{\dfrac{S_e}{n-2}}\cdot\sqrt{1+\dfrac{1}{n}+\dfrac{(x-\bar{x})^2}{l_{xx}}}\right)$，

且 $1-\alpha=0.95$，$t_{1-\alpha/2}(n-2)=t_{0.975}(7)=2.3646$，

故在 $x=300$ 时，$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x=-2.2582+0.04867\times 300=12.3434$，$y$ 的 $1-\alpha$ 预测区间为

$$\left(\hat{y}\pm t_{1-\alpha/2}(n-2)\cdot\sqrt{\frac{S_e}{n-2}}\cdot\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{l_{xx}}}\right)$$

$$=\left(12.3434\pm 2.3646\times\sqrt{\frac{7.9255}{7}}\cdot\sqrt{1+\frac{1}{9}+\frac{(300-288.9278)^2}{85861.9522}}\right)=(9.6895,14.9972);$$

（4）若回归直线过原点，即回归模型为 $y = kx + \varepsilon$，

则 k 的最小二乘估计为 $\hat{k} = \dfrac{\sum x_i y_i}{\sum x_i^2} = \dfrac{34874.7507}{837175.2991} = 0.04166$，

故回归直线为 $\hat{y} = 0.04166x$；

因 $S_T = \sum y_i^2 = 1465.4326$，自由度为 $n = 9$，

$S_R = \sum \hat{y}_i^2 = \hat{k}^2 \sum x_i^2 = 0.04166^2 \times 837175.2991 = 1452.8000$，自由度为 1，

$S_e = \sum (y_i - \hat{y}_i)^2 = S_T - S_R = 1465.4326 - 1452.8000 = 12.6326$，自由度为 $n - 1 = 8$，

假设 H_0：$\beta_1 = 0$ vs H_1：$\beta_1 \neq 0$，

选取统计量 $F = \dfrac{S_R}{S_e/(n-1)} \sim F(1, n-1)$，

显著性水平 $\alpha = 0.05$，$n = 9$，$F_{1-\alpha}(1, n-1) = F_{0.95}(1, 8) = 5.32$，右侧拒绝域 $W = \{F \geq 5.32\}$，

方差分析表

来源	平方和	自由度	均方和	F 比	p 值
回归 R	1452.8000	1	1452.8000	920.0290	1.5152×10^{-9}
误差 e	12.6326	8	1.5971		
和 T	1465.4326	9			

有 $F = \dfrac{S_R}{S_e/(n-1)} = \dfrac{1452.8000}{12.6326/8} = 920.0290 \in W$，

并且检验的 p 值 $p = P\{F \geq 920.0290\} = 1.5152 \times 10^{-9} < \alpha = 0.05$，

故拒绝 H_0，接受 H_1，可以认为回归方程显著.

习题 8.5

1. 设曲线函数形式为 $y = a + b\ln x$，试给出一个变换将之化为一元线性回归的形式.

解：设 $u = \ln x$，$v = y$，即得一元线性回归形式 $v = a + bu$.

2. 设曲线函数形式为 $y = a + b\sqrt{x}$，试给出一个变换将之化为一元线性回归的形式.

解：设 $u = \sqrt{x}$，$v = y$，即得一元线性回归形式 $v = a + bu$.

3. 设曲线函数形式为 $y - 100 = a\,\mathrm{e}^{-x/b}\ (b > 0)$，试给出一个变换将之化为一元线性回归的形式.

解：因可化为 $\ln(y - 100) = \ln a - \dfrac{x}{b}$，设 $u = x$，$v = \ln(y - 100)$，且 $a^* = \ln a$，$b^* = -\dfrac{1}{b}$，

故得一元线性回归形式 $v = a^* + b^* u$.

4. 设曲线函数形式为 $y = a + \mathrm{e}^{bx}\ (b > 0)$，问能否找到一个变换将之化为一元线性回归的形式，若能，试给出；若不能，说明理由.

解：不能，因不可能将其改写为 $g(y) = a^* + b^* f(x)$ 的形式，其中 $f(x)$ 与 $g(y)$ 不含未知参数.

5. 设曲线函数形式为 $y = \dfrac{1}{a + b\,\mathrm{e}^{-x}}$，问能否找到一个变换将之化为一元线性回归的形式，若能，试给出；若不能，说明理由.

解：能，因可化为 $\dfrac{1}{y}=a+b\mathrm{e}^{-x}$，设 $u=\mathrm{e}^{-x}$，$v=\dfrac{1}{y}$，即得一元线性回归形式 $v=a+bu$.

6．设曲线函数形式为 $y=a\,\mathrm{e}^{b/x}$，问能否找到一个变换将之化为一元线性回归的形式，若能，试给出；若不能，说明理由.

解：能，因可化为 $\ln y=\ln a+\dfrac{b}{x}$，设 $u=\dfrac{1}{x}$，$v=\ln y$，且 $a^*=\ln a$，即得一元线性回归形式 $v=a^*+bu$.

7．为了检验 X 射线的杀菌作用，用 200 kV 的 X 射线照射杀菌，每次照射 6 min，照射次数为 x，照射后所剩细菌数为 y，下表是一组试验结果

x	y	x	y	x	y
1	783	8	154	15	28
2	621	9	129	16	20
3	433	10	103	17	16
4	431	11	72	18	12
5	287	12	50	19	9
6	251	13	43	20	7
7	175	14	31		

根据经验知道 y 关于 x 的曲线回归方程形如

$$\hat{y}=a\,\mathrm{e}^{bx},$$

试给出具体的回归方程，并求其对应的决定系数 R^2 和剩余标准差 s.

解：因可化为 $\ln y=\ln a+bx$，设 $u=x$，$v=\ln y$，且 $a^*=\ln a$，即得一元线性回归形式 $v=a^*+bu$，

$u=x$	y	$v=\ln y$	u^2	uv	v^2	$\hat{y}_i$	$(y_i-\hat{y}_i)^2$	$(y_i-\overline{y})^2$
1	783	6.6631	1	6.6631	44.3973	821.2819	1465.5064	360300.0625
2	621	6.4313	4	12.8627	41.3620	641.3262	413.1548	192063.0625
3	433	6.0707	9	18.2122	36.8539	500.8016	4597.0574	62625.0625
4	431	6.0661	16	24.2644	36.7977	391.0681	1594.5538	61628.0625
5	287	5.6595	25	28.2974	32.0297	305.379	337.7872	10868.0625
6	251	5.5255	36	33.1527	30.5306	238.4657	157.1093	4658.0625
7	175	5.1648	49	36.1535	26.6750	186.2141	125.7563	60.0625
8	154	5.0370	64	40.2956	25.3709	145.4117	73.7591	826.5625
9	129	4.8598	81	43.7383	23.6178	113.5497	238.7114	2889.0625
10	103	4.6347	100	46.3473	21.4807	88.6692	205.3720	6360.0625
11	72	4.2767	121	47.0433	18.2899	69.2404	7.6155	12265.5625
12	50	3.9120	144	46.9443	15.3039	54.0687	16.5546	17622.5625
13	43	3.7612	169	48.8956	14.1466	42.2214	0.6062	19530.0625
14	31	3.434	196	48.0758	11.7923	32.9701	3.8811	23028.0625
15	28	3.3322	225	49.9831	11.1036	25.7458	5.0814	23947.5625
16	20	2.9957	256	47.9317	8.9744	20.1045	0.0109	26487.5625
17	16	2.7726	289	47.1340	7.6872	15.6993	0.0904	27805.5625
18	12	2.4849	324	44.7283	6.1748	12.2593	0.0672	29155.5625
19	9	2.1972	361	41.7473	4.8278	9.5731	0.3285	30189.0625
20	7	1.9459	400	38.9182	3.7866	7.4755	0.2261	30888.0625
210	3655	87.2250	2870	751.3889	421.2027		9243.2298	943197.75

则$\overline{u}=\frac{1}{n}\sum u_i=\frac{1}{20}\times 210=10.5$，$\overline{v}=\frac{1}{n}\sum v_i=\frac{1}{20}\times 87.2250=4.3612$，

$$l_{uu}=\sum u_i^2-\frac{1}{n}(\sum u_i)^2=2870-\frac{1}{20}\times 210^2=655\text{，}$$

$$l_{uv}=\sum u_i v_i-\frac{1}{n}\sum u_i\sum v_i=751.3889-\frac{1}{20}\times 210\times 87.2250=-164.4733\text{，}$$

$$l_{vv}=\sum v_i^2-\frac{1}{n}(\sum v_i)^2=421.2027-\frac{1}{20}\times 87.2250^2=40.7930\text{，}$$

有$\hat{b}=\frac{l_{uv}}{l_{uu}}=\frac{-164.4733}{655}=-0.2473$，$\hat{a}*=\overline{v}-\hat{b}\overline{u}=4.3612-(-0.2473)\times 10.5=6.9582$，

故一元线性回归方程为$\hat{v}=6.9582-0.2473u$，即$\hat{y}=1051.7331\mathrm{e}^{-0.2473x}$，

决定系数$R^2=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\overline{y})^2}=1-\frac{9243.2298}{943197.75}=0.9902$，

剩余标准差$s=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{\frac{9243.2298}{18}}=22.6608$.