第八章　方差分析与回归分析

习题 8.1

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

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

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

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 $W=\{F\geq 2.69\}$，

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

得 $S_T=\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y})^2=\sum_{i=1}^{r}\sum_{j=1}^{m}y_{ij}^2-\frac{1}{n}T^2=18439.05-\frac{1}{35}\times788.5^2=675.2714$，

$$S_A=\sum_{i=1}^{r}\sum_{j=1}^{m}(\bar{y}_{i\cdot}-\bar{y})^2=\frac{1}{m}\sum_{i=1}^{r}T_i^2-\frac{1}{n}T^2=\frac{1}{7}\times127185.19-\frac{1}{35}\times788.5^2=405.5343,$$

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

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

并且检验的 p 值 $p=P\{F\ge11.2758\}=1.0527\times10^{-5}<\alpha=0.05$，

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 $\bar{Y}_{i\cdot}$，

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

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

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

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

故 μ_5 的 0.95 置信区间为

$$(\bar{y}_{5\cdot}\pm t_{1-\alpha/2}(f_e)\cdot\frac{\hat{\sigma}}{\sqrt{m}})=(27.9857\pm2.0423\times\frac{2.9985}{\sqrt{7}})=(25.6711,30.3003).$$

习题 8.4

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 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$；

（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 $\hat{\beta}_1 = \dfrac{l_{xy}}{l_{xx}} = \dfrac{25.5218}{0.3024} = 84.3975$， $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 45.7886 - 84.3975 \times 0.125 = 35.2389$，

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：$\beta_1 \neq 0$，

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

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

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

$S_R = \sum(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 l_{xx} = 84.3975^2 \times 0.3024 = 2153.9758$，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 $p = P\{F \geq 108.2863\} = 5.6929 \times 10^{-8} < \alpha = 0.05$，

故拒绝 H_0，接受 H_1，可以认为回归方程显著；

（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

$$(\hat{\beta}_1\pm t_{1-\alpha/2}(n-2)\cdot\sqrt{\frac{S_e}{n-2}}\Big/\sqrt{l_{xx}})=(84.3975\pm 2.1448\times\sqrt{\frac{278.4808}{14}}\Big/\sqrt{0.3024})$$

$$=(67.0023, 101.7927);$$

（6）因 $y = \beta_0 + \beta_1 x + \varepsilon$ 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，$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x=35.2389+84.3975\times 0.15=47.8985$，$y$ 的 $1-\alpha$ 预测区间为

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

$$=(47.8985\pm 2.1448\times\sqrt{\frac{278.4808}{14}}\cdot\sqrt{1+\frac{1}{16}+\frac{(0.15-0.125)^2}{0.3024}})=(38.0288, 57.7683).$$

9. 设回归模型为$\begin{cases} y_i=\beta_0+\beta_1 x_i+\varepsilon_i, \\ \varepsilon_i\sim N(0,\sigma^2), \end{cases}$ 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 $\bar{x}=\dfrac{1}{n}\sum x_i=0.85+\dfrac{1}{15}\times(1.2-1.5)=0.83$，

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

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

$$l_{xy}=\sum x_i y_i-n\bar{x}\bar{y}=32.54+(1.2\times 32.6-1.5\times 32.3)-15\times(0.83\times 25.62-0.85\times 25.60)=30.641,$$

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

（1）$\hat{\beta}_1=\dfrac{l_{xy}}{l_{xx}}=\dfrac{30.641}{19.254}=1.5914$，$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=25.62-1.5914\times 0.83=24.2991$；

（2）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 $\alpha = 0.05$，$n = 15$，$F_{1-\alpha}(1, n-2) = F_{0.95}(1, 13) = 4.6672$，右侧拒绝域 $W = \{F \geq 4.6672\}$，

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

$S_R = \sum (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 l_{xx} = 1.5914^2 \times 19.254 = 48.7624$，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 $F = \dfrac{S_R}{S_e/(n-2)} = \dfrac{48.7624}{2.0816/13} = 304.5278 \in W$，

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

故拒绝 H_0，接受 H_1，可以认为回归方程显著.

（3）因 $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon$ 的 $1 - \alpha$ 预测区间为 $\left(\hat{y}_0 \pm t_{1-\alpha/2}(n-2) \cdot \sqrt{\dfrac{S_e}{n-2}} \cdot \sqrt{1 + \dfrac{1}{n} + \dfrac{(x_0 - \bar{x})^2}{l_{xx}}}\right)$，

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

故在 $x_0 = 1.1$ 时，$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 24.2991 + 1.5914 \times 1.1 = 26.0497$，$y_0$ 的 $1 - \alpha$ 预测区间为

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

$$= \left(26.0497 \pm 2.1604 \times \sqrt{\frac{2.0816}{13}} \cdot \sqrt{1 + \frac{1}{15} + \frac{(1.1 - 0.83)^2}{19.254}}\right) = (25.1552, 26.9441).$$

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