Chapter 4. Comparing a number of entities

4.1. Analysis of Variance (ANOVA)

One-way ANOVA is to check the difference between several samples, in contrast to the t-test which is to check the difference between two samples.

Suppose that

$Y_{tj} = \tau_t + \epsilon_{tj}$, $t = 1, ..., I$ and $j = 1, ..., J$,

where $\epsilon_{tj} \sim N(0, \sigma^2)$, and τ_t are parameters.

H_o: $\tau_1 = \cdots = \tau_I$ v.s. H_1: at least one inequality.

Example 3. Let $I = 3$, $J = 2$, $\begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \\ Y_{31} & Y_{32} \end{pmatrix}$, then $n = 6$, $p = 3$,

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \mathbf{X}\beta + \mathbf{e},\ \mathbf{X} = ??\ \beta = ??\ \mathbf{X} \perp \mathbf{e}\ ?$$

Remark. The expression of the model is not unique.

(1) $Y_{tj} = \tau_t + \epsilon_{tj}$, $t = 1, ..., I$ and $j = 1, ..., J$.

R command: $lm(Y \sim treatment - 1)$

(2) $Y_{tj} = \eta + \tau_t + \epsilon_{tj}$, $t = 1, ..., I$ and $j = 1, ..., J$.

R command: $lm(Y \sim treatment)$

Under Model (2), if we do not impose constraint to the parameters, then the parameters are not identifiable, that is, the LSE is not uniquely determined. Thus we either set $\tau_1 = 0$ or $\sum_{t=1}^{I} \tau_t = 0$.

For testing

H_o: $\tau_1 = \cdots = \tau_I$ v.s. H_1: H_o is false.

The test is $\phi = \mathbf{1}(F > F_{I-1,I(J-1),\alpha}$, where F is given in the ANOVA table.

Source of variation	*sum of squares*	*df*	*mean square*	F
Between treatments	$S_T = \sum_{t,j}(\overline{Y}_{t,\cdot} - \overline{Y})^2$	$\nu_T = I - 1$	$m_T = \frac{S_T}{\nu_T}$	
Within treatments	$S_R = \sum_{t,j}(Y_{tj} - \overline{Y}_{t,\cdot})^2$	$\nu_R = I(J-1)$	$m_R = \frac{S_R}{\nu_R}$	$\frac{m_T}{m_R}$
(hint)	$\sum_i (Y_i - \hat{Y}_i)^2$	$n - p$		$\uparrow\uparrow$
Total about $\overline{Y}$	$S_D = \sum_{i,j}(Y_{ij} - \overline{Y})^2$	$\nu_D = IJ - 1$		

due to NID and

$$= \underbrace{\sum_{t,j}(Y_{tj} - \overline{Y})^2}_{S_D} + \textstyle\sum_{t,j} \overline{Y}^2$$

$$= \underbrace{\sum_{t,j}(Y_{tj} - \overline{Y}_{t\cdot})^2}_{?} + \underbrace{\sum_{t,j}(\overline{Y}_{t\cdot} - \overline{Y})^2}_{?} + \textstyle\sum_{t,j} \overline{Y}^2.$$

Blood Coagulation Time Example.

Table 4.1 gives coagulation times for sample blood drawn from 24 animals receiving 4 different diets A, B, C and D.

Question: Is there evidence to indicate any real difference between the mean coagulation times for the four different diets ?

To randomized the outcomes, in addition to randomly select 24 animals, one may randomly put them into four groups by (1) number them, and (2) use

> sample(1:24,replace=F)

[1] 7 11 19 16 20 2 — 8 5 9 23 1 21 — 3 12 15 22 24 13 — 6 17 10 14 4 18

A	B	C	D
62	63	68	56
60	67	66	62
63	71	71	60
59	64	67	61
63	65	68	63
59	66	68	64

The data are , $I = 4$, $J = 6$,

Source of variation	*sum of squares*	*df*	*mean square*	F
Between treatments	$S_T = 228$	$\nu_T = 3$	$m_T = 76$	
Within treatments	$S_R = 112$	$\nu_R = 20$	$m_R = 5.6$	13.57

> x=c(62 , 63 , 68 , 56, 60 , 67 , 66 , 62, 63 , 71 , 71 , 60, 59 , 64 , 67 , 61, 63 , 65 , 68 , 63, 59 , 66 , 68 , 64)
> (treatment=gl(4,1,24))
[1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4
Levels: 1 2 3 4
> (obj=lm(x∼treatment))

(*Intercept*)	*treatment*2	*treatment*3	*treatment*4
$6.100e+01$	$5.000e+00$	$7.000e+00$	$-9.999e-15$
$\overline{Y}_{1\cdot}$	$\overline{Y}_{2\cdot} - \overline{Y}_{1\cdot}$	$\overline{Y}_{3\cdot} - \overline{Y}_{1\cdot}$	$\overline{Y}_{4\cdot} - \overline{Y}_{1\cdot}$

> anova(obj)

	Df	*Sum Sq*	*Mean Sq*	*F value*	*Pr*(> *F*)	
treatment	3	228	76.0	13.571	$4.658e-05$	$***$
Residuals	20	112	5.6			

Summary:
H_o: $\tau_1 = \cdots = \tau_4$ v.s. H_1: at least one inequality.
Conclusion: Yes, reject H_o, as F is far away from 1 **(where can we get it ?**
P-values is 0.00005.
There is real difference between the mean coagulation times for the four different diets.
Reason for one way anova (under control.sum):
$Y_{ij} = \eta + \alpha_i + \epsilon_{ij}$, $i \in \{1, ..., I\}$, $j \in \{1, ..., J\}$,
$\sum_i \alpha_i = 0$
=> $\overline{Y} = \eta + \overline{\epsilon}$,
$\overline{Y}_{i\cdot} = \eta + \alpha_i + \overline{\epsilon}_{i\cdot}$, $i \in \{1, ..., I\}$. One can also explain by
$(\hat{\eta}, \hat{\alpha}_2, ..., \hat{\alpha}_I)' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.
Blood Coagulation Time Example (continued).
> summary(lm(x∼treatment-1))

	Estimate	*Std. Error*	*t value*	*Pr*(> |*t*|)
*treatment*1	61.0000	0.9661	63.14	$< 2e-16 ***$
*treatment*2	66.0000	0.9661	68.32	$< 2e-16 ***$
*treatment*3	68.0000	0.9661	70.39	$< 2e-16 ***$
*treatment*4	61.0000	0.9661	63.14	$< 2e-16 ***$

> dim(x)=c(4,6); x=t(x)
> apply(x,2,mean)
[1] 61 66 68 61

> summary(lm(x∼treatment))

	Estimate	*Std. Error*	*t value*	*Pr*(> |*t*|)
(*Intercept*)	$6.100e+01$	$9.661e-01$	63.141	$< 2e-16 ***$
*treatment*2	$5.000e+00$	$1.366e+00$	3.660	$0.00156 **$
*treatment*3	$7.000e+00$	$1.366e+00$	5.123	$5.18e-05 ***$
*treatment*4	$-1.000e-14$	$1.366e+00$	0.000	1.00000

> treat=rep(c(1,2,3,1),6) # what does 4→ 1 mean ?
> a=lm(x∼factor(treat))
> summary(a)

	Estimate	*Std. Error*	*t value*	$Pr(>\|t\|)$
$(Intercept)$	61.0000	0.6667	91.500	$< 2e-16 **$
$factor(treat)2$	5.0000	1.1547	4.330	$0.000295 **$
$factor(treat)3$	7.0000	1.1547	6.062	$5.14e-06 **$

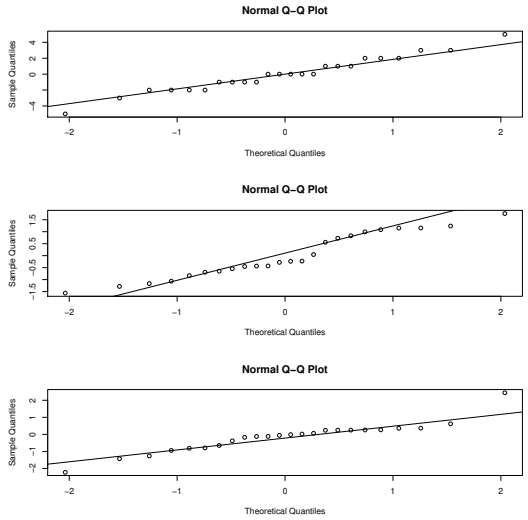
> a=lm(x∼factor(treat)-1)
> summary(a) # compare "Estimate" in these two summaries.

	Estimate	*Std. Error*	*t value*	$Pr(>\|t\|)$
$factor(treat)1$	61.0000	0.6667	91.50	$< 2e-16 **$
$factor(treat)2$	66.0000	0.9428	70.00	$< 2e-16 **$
$factor(treat)3$	68.0000	0.9428	72.12	$< 2e-16 **$

> anova(a) Analysis of Variance Table

	Df	*Sum Sq*	*Mean Sq*	*F value*	$Pr(>F)$
$factor(treat)$	3	98532	32844	6158.2	$< 2.2e-16 **$
Residuals	21	112	5		

> qqnorm(a$resid)
> qqline(a$resid)
> b=rnorm(24)
> qqnorm(b)
> qqline(b) # repeat the last 3 lines one or two times **why ?**



Two-way ANOVA is to check the difference between several samples, as well as between blocks.

Suppose that

$Y_{tj} = \eta + \tau_t + \beta_j + \epsilon_{tj}$, $t = 1, ...,k$ and $j = 1, ..., n$,

where $\epsilon_{tj} \sim N(0, \sigma^2)$, η, τ_t and β_j are parameters, subject to

$\tau_1 = 0 = \beta_1$ (or $\sum_t \tau_t = \sum_j \beta_j = 0$).

We shall do three tests:

H_o^*: $\tau_1 = \cdots = \tau_k$ and $\beta_1 = \cdots = \beta_n$ v.s. H_1^*: at least one inequality.

H_o: $\tau_1 = \cdots = \tau_k$ v.s. H_1: at least one inequality.

H_o': $\beta_1 = \cdots = \beta_n$ v.s. H_1': at least one inequality.

Source of variation	*sum of squares*	*df*	*mean squares*	F
Between blocks	$S_B = \sum_{j=1}^n (\overline{Y}_{\cdot,j} - \overline{Y})^2$	$\nu_B = n-1$	$m_B = \frac{S_B}{\nu_B}$	$\frac{m_B}{m_R}$
Between treatments	$S_T = \sum_{t=1}^k (\overline{Y}_{t,\cdot} - \overline{Y})^2$	$\nu_T = k-1$	$m_T = \frac{S_T}{\nu_T}$	$\frac{m_T}{m_R}$
Within treatments	$S_R =$ $\sum_{t,j}(Y_{t,j} - \overline{Y}_{t,\cdot} - \overline{Y}_{\cdot,j})^2$	$\nu_R =$ $(k-1)(n-1)$	$m_R = \frac{S_R}{\nu_R}$	
Total about $\overline{Y}$	$S_D = \sum_{t,j}(Y_{tj} - \overline{Y})^2$	$\nu_D = kn-1$		

$\sum_{t,j} Y_{tj}^2 = S_D + \sum_{t,j} \overline{Y}^2 = S_B + S_T + S_R + \sum_{t,j} \overline{Y}^2$.

Blood Coagulation Time Example (continued).

H_o^*: $\tau_1 = \cdots = \tau_k$, $\beta_1 = \cdots = \beta_n$ v.s. v.s. H_1^*: at least one inequality.

H_o: treatment effects: $\tau_1 = \cdots = \tau_k$ v.s. H_1: at least one inequality.

H_o': row effects $\beta_1 = \cdots = \beta_n$ v.s. H_1': at least one inequality.

> (row=gl(6,4,24))

[1] 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6

Levels: 1 2 3 4 5 6

> (tr=lm(x∼treatment+row))

$(Intercept)$	$treatment2$	$treatment3$	$treatment4$	$row2$	$row3$
$5.925e+01$	$5.000e+00$	$7.000e+00$	$1.285e-14$	$1.500e+00$	$4.000e+00$
$row4$	$row5$	$row6$	↑		
$5.000e-01$	$2.500e+00$	$2.000e+00$	↓		

$(Intercept)$	$treatment2$	$treatment3$	$treatment4$	$row2$	$row3$
	$\overline{Y}_{2\cdot} - \overline{Y}_{1\cdot}$	$\overline{Y}_{3\cdot} - \overline{Y}_{1\cdot}$	$\overline{Y}_{4\cdot} - \overline{Y}_{1\cdot}$	$\overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 1}$	$\overline{Y}_{\cdot 3} - \overline{Y}_{\cdot 1}$