Design of Experiments (Math 556)

MWF 8:00am-9:30am, WH 329
Office: WH 132
Office hours: M, T 7-8pm. Through Zoom
    https://binghamton.zoom.us/j/8265526594?pwd=d3l6OGx1cmZ4M3cxZEJwVGd1RGcrUT09
    Meeting ID: 826 552 6594
    Passcode: 031320
Textbook: Statistics for Experimenters (2nd ed.)
    by George Box, J Stuart Hunter and William G. Hunter
Quiz: Once a week at a random day,
quiz problems: formulas for Math 447-448 (see my website)
Midterm: March 20 (M) Final May 11 5:40-7:40pm CW 314 **Changed to WH329 !!**
    Each is allowed to bring a piece of paper with anything you prefer on it.
Homework assigned during a week is due next Wednesday before 8:00am.
    Email me at qyu@math.binghamton.edu before 8:00am on Wednesday.
    HW is on my website: http://www.math.binghamton.edu/qyu/qyu_personal
Remind me if you do not see it by Saturday morning !
    Try to use Latex in homework. Otherwise, take a picture and convert it to a pdf file.

There will be homework due this Friday, as well as quiz !!!

The lecture note is also on my website
    http://www.math.binghamton.edu/qyu/qyu_personal
    note and note2 are updated one,
Grading Policy: 40% hw and quizzes +60% exams,
    B = 70 ±

Chapter 1. Introduction

Self-reading.

Chapter 2. Basic

All concepts in this chapter have been introduced in 501, except autocorrelation.
Recall
    X and Y are random variables, with observations (X_i, Y_i), $i = 1, ..., n$.
Population covariance and correlation:
    $Cov(X,Y) = E(XY) - E(X)E(Y)$,
    $\rho = \rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$,
Sample Covariance $\hat{Cov}(X,Y) = \overline{XY} - \overline{X} \cdot \overline{Y}$
Sample correlation $\hat\rho = r = \frac{\overline{XY} - \overline{X}\cdot\overline{Y}}{\hat\sigma_X \hat\sigma_Y}$, where $\hat\sigma^2_X = \overline{XX} - (\overline{X})^2$,
Note that the sample variance of X is often refer to $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$
(S^2 is also denoted by s^2 in the textbook. **Which is a better notation ?**)
Definition. The lag-k sample autocorrelation coefficient of Y_i's is

$$r_k = \frac{\sum_{i>k}^n (Y_i - \overline{Y})(Y_{i-k} - \overline{Y})}{\sum_{i=1}^n (Y_i - \overline{Y})^2}, \ k = 1, 2, ...$$

It measures the serial dependence of the data in time.

If $r_k \neq 0$, are the data i.i.d. ?

If $r_k > 0$ significantly, are the data i.i.d. ?

```
> x=rnorm(20)
```

> cor(x[1:19],x[2:20]) [1] -0.1431549
> cor.test(x[1:19],x[2:20])
t = -0.59639, df = 17, p-value = 0.5588
cor -0.1431549

Theorem 1. If X_1, ..., X_n are i.i.d. from $N(\mu, \sigma^2)$, then
(a) $\overline{X} \perp S^2$;
(b) $\overline{X} \sim N(\mu, \sigma^2/n)$;
(c) $(n-1)S^2/\sigma^2 = n\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-1)$.

Chapter 3. Comparing Two Entities

3.1. Consider the test for the difference of the means of two random samples X_i's and Y_j's.

H_o: $\mu_Y - \mu_X = \delta$ v.s. H_1: $\mu_Y - \mu_X > \delta$.

Two-samples test: Under the assumption that (1) two samples are independent, (2) X_i's are from $N(\mu_X, \sigma^2)$ and (3) Y_j's are $N(\mu_Y, \sigma^2)$, then a common test is

$$\phi = \mathbf{1}(t > t_{\alpha, n_Y+n_X-2}), \text{ where} \tag{1}$$

$$t = \frac{\overline{Y} - \overline{X} - \delta}{s_p\sqrt{1/n_X + 1/n_Y}} \text{ and } s_p^2 = \frac{\sum_{i=1}^{n_X}(X_i - \overline{X})^2 + \sum_{j=1}^{n_Y}(Y_j - \overline{Y})^2}{n_Y + n_X - 2}.$$

This is due to

(a) $T = \frac{N(0,1)}{\sqrt{\chi^2(\nu)/\nu}} \sim$ distribution ?, where $N(0,1) \perp \chi^2(\nu)$

(b) $t = \frac{\overline{Y}-\overline{X}-\delta}{\sigma\sqrt{1/n_X+1/n_Y}} / \sqrt{s_p^2/\sigma^2}$,

(c) $\frac{\sum_{i=1}^{n_X}(X_i-\overline{X})^2}{\sigma^2} \sim \chi^2(n_X - 1)$, $\frac{\sum_{i=1}^{n_Y}(Y_i-\overline{Y})^2}{\sigma^2} \sim$?

(d) $\frac{\sum_{i=1}^{n_X}(X_i-\overline{X})^2}{\sigma^2} + \frac{\sum_{i=1}^{n_Y}(Y_i-\overline{Y})^2}{\sigma^2} \sim$ what distribution ?

The paired t-test: Under the paired random sample of size n from $N(\mu_Y, \sigma_Y^2)$ and $N(\mu_X, \sigma_X^2)$, then a common test is

$$\phi_p = \mathbf{1}(t > t_{\alpha, n-1}), \text{ where}$$

$$t = \frac{\overline{Y}-\overline{X}-\delta}{s\sqrt{1/n}} \text{ and } s^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - X_i - \overline{Y - X})^2.$$

Importance of the independent normally distributed assumptions in both tests.

Chemical Example in Table 3.2. An experiment was performed on a factory by making in sequence 10 batches of chemical using a standard production method (A) followed by 10 batches of a chemical using a modified method (B). The data are

A: 89.7, 81.4, ..., 84.5
B: 84.7, 86.1, ..., 88.5

See Table 3.1 on page 69.

Summary:
$n_A = n_B = 10$,
$\overline{y}_A = 84.24$,
$\overline{y}_B = 85.54$,
$s_p^2 = 10.8727$,

H_o: $\mu_B - \mu_A = 0$, v.s. H_1: $\mu_B - \mu_A > 0$.
$\overline{y}_B - \overline{y}_A = 1.3$.
Is it significant ? **What does it mean ?**

We need to
(1) set α (= 0.05), and

(2) compute P($\overline{y}_B - \overline{y}_A \geq 1.3$) = ? **what is it called ?**
Then conclude that if P($\overline{y}_B - \overline{y}_A \geq 1.3$) $< \alpha$ ...

One often uses the two-sample t-test in Eq. (1), then the P-value is 19% here.
Is it significant ?
Do we reject H_o ?
Can we use paired t-test ?
Does the SD become larger or smaller if we use it ?

$\sigma^2/(2n-2)$ v.s. $\sigma^2/(n-1)$.

$s_p^2 = \frac{\sum_{i=1}^{n_X}(X_i-\overline{X})^2+\sum_{j=1}^{n_Y}(Y_j-\overline{Y})^2}{n_Y+n_X-2}$ v.s. $S^2_{Y_B-Y_A}$

What is the conclusion if we use it ? P-value=P($T \geq \frac{\overline{y}_B-\overline{y}_A}{SE}$)
Introduce two alternative approaches next.
External Reference Distribution.
Old data. 210 batches of the chemical products recorded in time order before the 20 data:
x_1, ..., x_{210}
The old data (see p.120) provide an external reference distribution.
Under H_o, the 20 data can be viewed as a sample from the population of the 210 data.

Compute
$D_t = \sum_{i=t+10}^{t+19} x_i/10 - \sum_{i=t}^{t+9} x_i/10$, $t = 1, ..., 191$.
See the histogram Figure 3.3 on page 70.

$P(\overline{y}_B - \overline{y}_A \geq 1.3) = 9/191 \approx 0.047$. Is it significant ?
Recall that if one uses t-test, the P-value is 19%. **Anything wrong ?**

1. The lag-1 sample auto-correlation of the data is $r_1 = \hat{\rho}_1 = -0.29$.
   The data are not independent.
   If one pretends independence, it leads to incorrect conclusion.
2. Normal assumption may not be valid **(do we need to check it ?)**

Internal Reference distribution. Random sampling distribution.
A randomized design in the comparison of standard and modified fertilizer mixtures for tomato plants. 11 plants in a row. 5 with standard (A), 6 with modified (B). One way is to apply A to the first 5 and B to the next 6 in a row. There are correlation between locations and it is not a good idea without randomization.
Randomizing the order in the row (sample(1:11,5) = ?) resulting

location :	1	2	3	4	5	6	7	8	9	10	11
fertilizer :	A	A	B	B	A	B	B	B	A	A	B
yield :	29.2	11.4	26.6	23.7	25.3	28.5	14.2	17.9	16.5	21.1	24.3

(1)

Remark: Role of a statistician:
(1) randomization before an experiment (DOE);
(2) make inferences after the experiment.

```
> x=c(29.2,11.4,26.6,23.7,25.3,28.5,14.2,17.9,16.5,21.1, 24.3)
> z=c(3,4,6,7,8,11)
> mean(x[z])-mean(x[subset=−z])
```
results in $\overline{y}_B - \overline{y}_A \approx 1.69$.

To test H_o: $\mu_B - \mu_A = 0$ against H_1: $\mu_B - \mu_A > 0$.
Need to compute $P(\overline{y}_B - \overline{y}_A \geq 1.69) = ?$
Rather than using t-test, which needs normal assumption, and equal variance,
we make use of the
Permutation distribution.

Table (1) is one combination of selecting 5 out of 11.

1	2	3	4	5	6	7	8	9	10	11	
A	A	A	A	A	B	B	B	B	B	B	(2)
29.2	11.4	26.6	23.7	25.3	28.5	14.2	17.9	16.5	21.1	24.3	

is another combination under H_o: $\mu_B - \mu_A = 0$.

```
> mean(x[6:11])−mean(x[1:5])                # results in −2.82
```

Eq.(2) yields $\overline{y}_B - \overline{y}_A \approx -2.82$; while
Eq.(1) yields $\overline{y}_B - \overline{y}_A \approx 1.69$.

There are $\binom{11}{5} = \frac{11!}{5!6!} = 11 \cdot 3 \cdot 2 \cdot 7 = 462$ such combinations.

```
> P=combn(1:11,6)
> P[,1:10]
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	1	1	1	1	1	1	1	1	1	1
[2,]	2	2	2	2	2	2	2	2	2	2
[3,]	3	3	3	3	3	3	3	3	3	3
[4,]	4	4	4	4	4	4	4	4	4	4
[5,]	5	5	5	5	5	5	6	6	6	6
[6,]	6	7	8	9	10	11	7	8	9	10

Thus these 462 combinations yield 462 $\overline{y}_B - \overline{y}_A$ values.
These 462 values form a (discrete) distribution called the **permutation distribution**.

```
    x=c(29.2,11.4,26.6,23.7,25.3,28.5,14.2,17.9,16.5,21.1, 24.3)
```

Either use loop

```
    N=choose(11,6)                 # =462
    y=1:N
    P=combn(1:11,6)                # Can we use combn(1:11,5) ?
    for(i in 1:N)
        y[i]=mean(x[P[,i]])−mean(x[−P[,i]])
    length(y[y>=1.69])/N                  # result is 0.3203463
```

Or without loop:

```
    y=x[P]
    dim(y)=c(6,462)
    B=apply(y,2,sum)
    y=B/6−(sum(x)−B)/5
    length(y[y>=1.69])/N                  # result is 0.3203463
```

What is the conclusion of the test ?

```
    library(jmuOutlier)      (another codes)
    y=runif(16,0,1)
    x=runif(20,0,1)
    perm.test(y,x,alternative=c("two.sided", "less", "greater"), mu=0, paired=FALSE,
all.perms=TRUE, plot=FALSE, stat=sum)
```

The permutation distribution can also be simulated by the R code as follows.

```
    x=c(29.2,11.4,26.6,23.7,25.3,28.5,14.2,17.9,16.5,21.1, 24.3)
    N=10000
    y=rep(0,N)
    for(i in 1:N){
        u=sample(x)
        y[i]=mean(u[1:6])-mean(u[7:11])
    }
    length(y[y>=1.69])/N # result is 0.3209
```

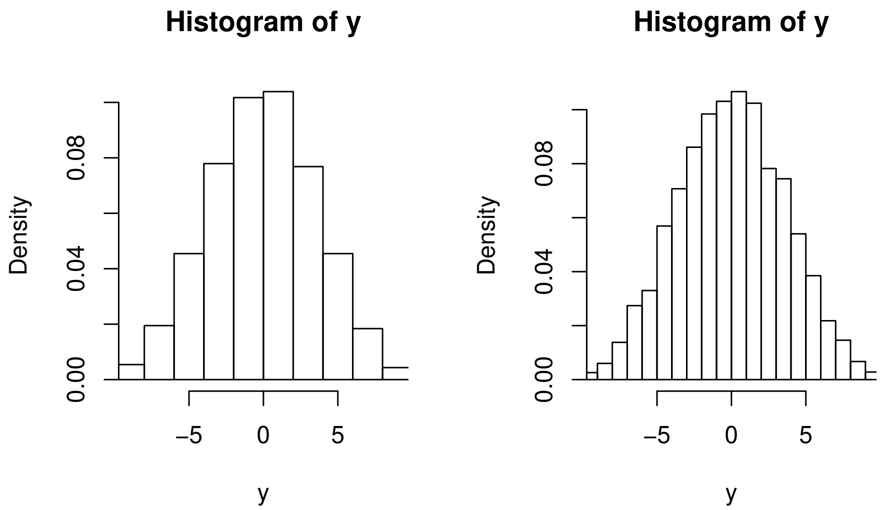


Figure 3.1. Histograms of permutation distribution v.s. simulation one

Should we use simulation here ?

Remark. The two-samples t-test $P(t_{n_A+n_B-2} > \frac{1.69}{s\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}}) \approx 0.34$) for the current data. If the normal assumption is not valid, the t-test is not applicable (though it happens to be close to 0.32

The permutation distribution is based on a different sample space from the sample space where the data come from. But if $n_A + n_B$ is large, the permutation distribution of $\overline{Y}_B - \overline{Y}_A$ is very close to $t_{n_A+n_B-2}$, whereas the two-sample t-test may not have the $t_{n_A+n_B-2}$ distribution (*e.g.* if the random variables satisfy $X_1 = \cdots = X_{n_A}$ and $Y_1 = \cdots = Y_{n_B}$), thus they are not independent).

Can we say $n_A + n_B$ is large here ?

Is it appropriate to apply randomization distribution in the chemical example ?

Remark. In the fertilizer example, the data are resulted from randomization, whereas in the previous chemical example, the data are in sequence.

A A A A A A A A A A B B B B B B B B B B

We use the External Reference distribution (old data) to get the P-value.

Can we use the permutation distribution to get the P-value in that example ?

No. If they had done

sample(1:20,10)

for the order of 10 batches of chemical using method A, then the permutation distribution would be valid.

3.2. Randomized paired comparison design: Boys shoes example. The shoe soles can be made of two different materials, A and B. To find out whether there is a difference between them, ten boys were chosen randomly to compare the shoe wear. Each boy wore a special pair of shoes. The decision as to whether the left or right sole was made with A or B was determined by

(1) convenience,

(2) by flipping a coin (or rbinom(n,1,0.5)).

Which result in a random sample ? (Took 2 steps in DOE. **Which 2 ?**)

The randomization results $\begin{pmatrix} boy: & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ material\ A & L & L & R & L & R & L & L & L & R & L \end{pmatrix}$

The experiment results in

x=(0.8,0.6,0.3,-0.1,1.1,-0.2,0.3,0.5,0.5,0.3) $\quad\quad$ # $y_B - y_A$

Then 10 $y_B - y_A$'s yield

mean(x) # $\overline{y}_B - \overline{y}_A = 0.41$

Should we use two-sample t-test or paired t-test ?

What assumptions do we need in order to use one of them ?

Another way to compute P-value for $\overline{y}_B - \overline{y}_A \geq 0.41$ is the permutation distribution.
Under H_o: $\mu_B - \mu_A = 0$, a combination could be

(R L R L R L L L R L)

$$\begin{pmatrix} & R & L & R & L & R & L & L & L & R & L \\ real & L & L & R & L & R & L & L & L & R & L \end{pmatrix}$$

Then the data become

x=(-0.8,0.6,0.3,-0.1,1.1,-0.2,0.3,0.5,0.5,0.3)

Compare to the real data:

x=(0.8,0.6,0.3,-0.1,1.1,-0.2,0.3,0.5,0.5,0.3)

The randomized reference distribution under H_o: $\mu_A = \mu_B$ can be obtained as follows.

```
x=c(0.8,0.6,0.3,-0.1,1.1,-0.2,0.3,0.5,0.5,0.3)
sum(x) # result=4.1
y=1:1024 # initialize y
for(i1 in 0:1)
for(i2 in 0:1)
for(i3 in 0:1)
for(i4 in 0:1)
for(i5 in 0:1)
for(i6 in 0:1)
for(i7 in 0:1)
for(i8 in 0:1)
for(i9 in 0:1)
for(i10 in 0:1){
    i=c(i1,i2,i3,i4,i5,i6,i7,i8,i9,i10)
    h=0:9
```

y[$\underbrace{i\% * \%(2 * *h)}$ +1]=sum(x*((-1)**i))

\# $i1*2^0 + i2*2^1 + i3*2^2 + \cdots + i10*2^9$, $(0, ..., 0)(2^0, 2^1, ..., 2^9)' + 1 = 1$,
...

\# Examples:

\# binary number $1110 = 1 * 2^3 + 1 * 2^2 + 1 * 2^1 + 0 * 2^0 = 14$

\# ternary number $2101 = 2 * 3^3 + 1 * 3^2 + 0 * 3^1 + 1 * 3^0 = 64$

\# decimal number $2101 = 2 * 10^3 + 1 * 10^2 + 0 * 10^1 + 1 * 10^0$

```
}
length(y[y>= 4.1])/1024          # result = 0.0068
hist(y); z=seq(-6,6,0.1);lines(z,dt(z,9))
```

The randomized reference distribution under H_o: $\mu_A = \mu_B$ can be approximated by simulation as follows.

```
N=10000
y=rep(0,N)
x=c(0.8,0.6,0.3,-0.1,1.1,-0.2,0.3,0.5,0.5,0.3)
for (i in 1:N) {
s=rbinom(10,1,0.5)
z=(-1)**s
y[i]=sum(x*z)
}
length(y[y>=4.1])/N          #0.0063
hist(y,xlim=c(-6,6), breaks=12, freq=F)
```

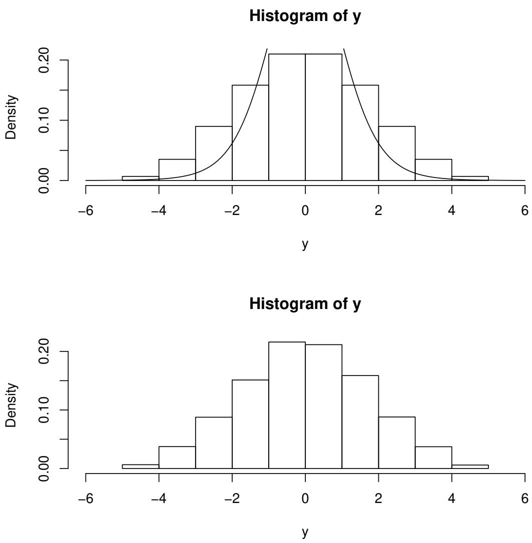


Figure 3.2. Histograms of permutation distribution v.s. simulation one

One can see from Figure 3.2 that the simulation distribution is very close to the true permutation distribution. The density of t_9 is displayed at the top of Fig. 3.2.

The P-value using the 1-sided paired t-test is 0.4%.

Any thing wrong with the solution 0.0068 or 0.4 ?

Is it one sided test or two-sided test ?

Can we mimic P=combn(1:10, ?) to write a code to replace the 1st one ?

Bashar, Mohamed A. 2 Chaikin, Kassidy 3 Phillips, Bruce 4 Zhao, Zhongyuan

Chapter 10. Linear regression models.

10.1. Main assumption:

$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, or

$E(Y|\mathbf{X}) = \beta_1 X_1 + \cdots + \beta_p X_p$, where

ϵ is unobservable random variable with $\underline{E(\epsilon|\mathbf{X}) = 0}$ (no assumption on $V(\epsilon|\mathbf{X})$ yet),

β_i's are parameters,

X_i's and Y are observable.

Given (independent) observations $(Y_i, x_{i1}, ..., x_{ip})$, $i = 1, ..., n$,

we shall make inference about β_i's.

Remark. A special case of the linear regression model is

$Y = \alpha + \beta X + \epsilon$.

Least squares estimator (LSE) minimizes

$S(\beta) = \sum_{i=1}^n (Y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$ where $\beta = (\beta_1, ..., \beta_p)'$.

Notice that $S(\beta)$ can be written as a matrix form

$S(\beta) = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$

where $\mathbf{Y}' = (Y_1, ..., Y_n)$,

$$\mathbf{X} = (x_{ij})_{n\times p} = \begin{pmatrix} x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \qquad \neq X$$

The LSE can be obtained by solving the normal equation

$\frac{\partial S}{\partial \beta} = \mathbf{0}$, a $p \times 1$ zero vector. $\qquad \frac{\partial S}{\partial \beta'} = ?$

That is,

$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{0}$. $\qquad$ (**Why not** $(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{X} = \mathbf{0}$ **?**)

The LSE has the form

$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ if $\mathbf{X}'\mathbf{X}$ is invertible,

otherwise, the solution to LSE is not unique,

one often imposes further constraints to get a unique solution.

If ϵ is normal, then $\hat\beta$ is the MLE. Otherwise, it is a semi-parametric estimator.
Fitted value $\hat y_i = (x_{i1}, ..., x_{ip})\hat\beta$. ($= \hat E(Y|\mathbf{x})$)
Residuals $y_i - \hat y_i$, $i = 1, ..., n$.
If one further assumes that $V(\epsilon_i) = \sigma^2 \ \forall\ i$, then

$\hat\sigma^2 = \frac{1}{n-p}\sum_{i=1}^n (y_i - \hat y_i)^2$ is an unbiased estimator of σ^2,
and conditional on $\mathbf{X}$ (if one assumes $\mathbf{X}$ is random),
$V(\hat\beta) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ or $V(\hat\beta|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ **(are they both correct ??)**
Is $V(\hat\beta)$ variance or covariance matrix ?

SE of $\hat\beta_j$ is $\sqrt{v}$, where v is obtained by the j-th diagonal element of $\hat\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

(why not $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$? SD = SE ? Are they r.v.'s ?)

Under NID a $(1-\alpha)100\%$ CI of β_j is $\hat\beta_j \pm t_{n-p,\alpha/2}SE$

Example 0: Suppose that $Y_i = \begin{cases} -\gamma + W_i & \text{if } i \in \{1, ..., n_-\} \\ \gamma + W_i & \text{if } i \in \{n_- + 1, ..., n,\} \end{cases}$ $n_- > 1$, and $W_1, ..., W_n$ are i.i.d. from the exponential distribution and $E(W_1) = 1$. γ and W_i;s are unknown, though we know $W_i \sim Exp(1)$. Y_i's are observations. Derive the LSE and the MLE of γ based on these regression data.

Discussion. The typical linear regression model is

$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i = X_i'\beta + \epsilon_i$ with $E(\epsilon_i) = 0$.
$p = ?$
Do we observe Y_i ?
Do we observe $(X_{i1}, ..., X_{ip})$?
$W_i = \epsilon_i$?
Do we know β ? or $(\beta_1, ..., \beta_p)$?
If we rewrite th model as $Y_i = \alpha + \gamma X_i + \epsilon_i$, then $\alpha = ?$
Do we need to estimate α ?

Homework 10.1. Find the MLE and the LSE of β under the assumptions above.

Polynomial model: $Y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_k x_i^k + \epsilon_i$, $i = 1, ..., n$.

k can be as large as $n-1$ if x_i's are all distinct.

Example 1. Data: (X_i, Y_i): (1,2), (3,4). The LSE $\hat\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ under the models:

$Y = \beta_0 + \epsilon$, $\mathbf{X} = ?$
$Y = \beta_1 x + \epsilon$, $\mathbf{X} = ?$
$Y = \beta_0 + \beta_1 x + \epsilon$. $\mathbf{X} = ?$

If one fits model $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$. Then $\mathbf{Y} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$, $\mathbf{X} = \begin{array}{c} \begin{array}{ccc} \beta_0 & \beta_1 & \beta_2 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 9 \end{pmatrix} \end{array}$

rank of $\mathbf{X}'\mathbf{X}$ is 2. $\mathbf{X}'\mathbf{X}$ is not invertible. The LSE is not uniquely determined. We say that the parameter is not identifiable.

Possible modification: Add a constraint to β_i's, *e.g.* $\beta_0 = 0$ or $\beta_1 = \beta_2$, etc.:

models\X type :	*original*	*in model*	$\mathbf{X}$ *in LSE formula*	β
$Y = \beta_0 + \epsilon$	$1, 3$	$1, 1$	$(1, 1)'$	β_0
$Y = \beta_1 x + \epsilon$	$1, 3$	$1, 3$	$(1, 3)'$	β_1
$Y = \beta_0 + \beta_1 x + \epsilon$	$1, 3$	$(1, 1), (1, 3)$	$\begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}$	$(\beta_0, \beta_1)'$
$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$	$1, 3$	$(1, 1, 1), (1, 3, 9)$	?	$(\beta_0, \beta_1, \beta_2)'$
$Y - 1 = \beta_1 x + \beta_2 x^2 + \epsilon$	$1, 3$	$(1, 1), (3, 9)$	?	$(\beta_1, \beta_2)'$
$Y = \beta_0 + \beta_1(x + x^2) + \epsilon$	$1, 3$	$(1, 2), (1, 12)$	?	$(\beta_0, \beta_1)'$

Example 2. One way anova table

$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$, $i = 1, ..., 4$, and $j = 1, 2, 3$.

Consider an example that there are three treatments A, B and C. There are I (=4) groups, each consists of 3 patients. Total of 12 patients. In each group, the 3 patients receive 3 different treatments separately. The result for the jth patient in the ith group is Y_{ij}.

Is it a linear regression model ?
$\beta = ?$
LSE $= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.
$\mathbf{X} = ?$ One possibility is based on $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$,

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ \vdots \\ Y_{41} \\ Y_{42} \\ Y_{43} \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_{10} \\ Y_{11} \\ Y_{12} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & & & \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \mathbf{e} \qquad Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$$= \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} \\ \vdots & & & \\ 1 & X_{n1} & X_{n2} & X_{n3} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \mathbf{e} = \mathbf{X}\beta + \mathbf{e}, \quad \mathbf{e} = (\epsilon_1, ..., \epsilon_{12})',$$

X_{i1}=$\mathbf{1}$(treatment=A for the $i-$th patient).
X_{i2}=$\mathbf{1}$(treatment=B for the $i-$th patient).
X_{i3}=$\mathbf{1}$(treatment=C for the $i-$th patient).
Notice that $X_{i1} + X_{i2} + X_{i3} = 1$.
$\mathbf{X}'\mathbf{X}$ is not invertible as
$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} \end{pmatrix}$ is of rank at most 3, not 4,
as $\begin{pmatrix} X_{11} + X_{12} + X_{13} \\ \vdots \\ X_{n1} + X_{n2} + X_{n3} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ **why ?**
Thus the LSE for $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$ is not unique.
(We say that the parameters are not **identifiable**).

Three modifications:
M1. Revise the model. Let $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$ with $\mu = 0$,
$\beta = (\alpha_1, \alpha_2, \alpha_3)'$, $\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & X_{13} \\ \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & X_{n3} \end{pmatrix}$, $\hat\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ works.
R codes: $lm(Y \sim X_1 + X_2 + X_3 - 1)$
Interpretation: β_i is the effect of treatment i (T_i).
M2. Impose a constraint $\alpha_1 = 0$ for the model
$Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} + \epsilon_i$. $\qquad Y_i = (1, X_{i2}, X_{i3})(\mu, \alpha_2, \alpha_3)'$.
Let $(\beta_1, \beta_2, \beta_3)$ be as in M1.
Then $\beta_i = \mu + \alpha_i$, $i = 1, 2, 3$.
e.g., if $X_{i1} = 1$, then $Y_i = \beta_1 + \epsilon_i = \mu + \alpha_1 + \epsilon_i$, where $\alpha_1 = 0$, $\mu = \beta_1$, ...
i.e., μ is the effect of treatment 1, but α_i is the additional effect of T_i to T_1.
options(contrasts =c("contr.treatment", "contr.poly"))
$lm(Y \sim X_1 + X_2 + X_3)$.
M3. Impose another constraint $\sum_i \alpha_i = 0$ ($\alpha_3 = -\alpha_1 - \alpha_2$) for the model
$Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} + \epsilon_i$
then $\mu + \alpha_i = \beta_i$, $i = 1, 2, 3$. $\qquad Y_i = (1, X_{i1} - X_{i3}, ????)(\mu, \alpha_1, \alpha_2)'$
i.e., μ is the average treatment effect, α_i is the additional effect of T_i.
options(contrasts =c("contr.sum", "contr.poly"))
$lm(Y \sim X_1 + X_2 + X_3)$

Example 3 (a simulation study on the Two way anova table).

$Y_{ij} = \mu + a_i + b_j + \epsilon_{ij}$, $i \in \{1, ..., 4\}$, $j \in \{1, ..., 6\}$

```
> y=rnorm(24)
> a=gl(4,6,24)
> b=gl(6,1,24)
> a
[1] 1 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4 4 Levels: 1 2 3 4
> b
[1] 1 2 3 4 5 6 1 2 3 4 5 6 1 2 3 4 5 6 1 2 3 4 5 6 Levels: 1 2 3 4 5 6
> lm(y~a+b-1) #1.
```

$a1$	$a2$	$a3$	$a4$	$b2$	$b3$	$b4$	$b5$	$b6$
0.266	0.235	0.246	-0.379	-0.102	-0.319	0.913	-0.227	-0.125

$\hat{\mu}$ and $b1$= ? $(\mu, a_1, ..., a_4, b_1, ..., b_6) = ?$

```
> lm(y~a+b) #2
# or
# lm(y~a+b, contrasts =c("contr.treatment", "contr.poly"))
```

$(Intercept)$	$a2$	$a3$	$a4$	$b2$	$b3$	$b4$	$b5$	$b6$
0.268	-0.031	-0.020	-0.645	-0.102	-0.319	0.913	-0.227	-0.125

$a1$,$b1$?

```
> options(contrasts =c("contr.sum", "contr.poly")) #3.
> lm(y~a+b)
```

$(Intercept)$	$a1$	$a2$	$a3$	$b1$	$b2$	$b3$	$b4$	$b5$
0.115	0.174	0.143	0.154	-0.023	-0.125	-0.342	0.890	-0.250

$a4$, $b6$?

Relation between these three ?

$\hat{E}(Y_{ij})$ = intercept$+ai + bj$ = same ?

1.		$int. = 0$	$b1 = 0$							
int	$a1$	$a2$	$a3$	$a4$	$b1$	$b2$	$b3$	$b4$	$b5$	$b6$
0	0.27	0.24	0.25	-0.38	0	-0.10	-0.32	0.91	-0.23	-0.13
2.		$a1 = 0$	$b1 = 0$							
0.27	0	-0.03	-0.02	-0.65	0	-0.10	-0.32	0.91	-0.23	-0.13
3.		$a4 =?$	$b6 =?$							
0.12	0.17	0.14	0.15	-0.46	-0.02	-0.13	-0.34	0.89	-0.25	-0.15

$$\hat{E}(Y_{11}) = \begin{cases} 0 + 0.266 + 0 & \text{from \#1} \\ 0.268 + 0 + 0 & \text{from \#2} \quad \textbf{Are they the same ?} \\ 0.115 + 0.174 - 0.023 = 0.266 & \text{from \#3.} \end{cases}$$

What is X, β and $\hat{\beta}$ in the model $Y = X'\beta + \epsilon$ for lm($y \sim a + b$) in Ex. 3 ?

$$X = \begin{pmatrix} 1 \\ \mathbf{1}(a=1) \\ \mathbf{1}(a=2) \\ \mathbf{1}(a=3) \\ \mathbf{1}(a=4) \\ \mathbf{1}(b=1) \\ \mathbf{1}(b=2) \\ \mathbf{1}(b=3) \\ \mathbf{1}(b=4) \\ \mathbf{1}(b=5) \\ \mathbf{1}(b=6) \end{pmatrix} \text{ or } \begin{pmatrix} 1 \\ \mathbf{1}(a=2) \\ \mathbf{1}(a=3) \\ \mathbf{1}(a=4) \\ \mathbf{1}(b=2) \\ \mathbf{1}(b=3) \\ \mathbf{1}(b=4) \\ \mathbf{1}(b=5) \\ \mathbf{1}(b=6) \end{pmatrix} \text{ ? } \beta' = (\mu, a_2, ..., a_4, b_2, ..., b_6), \text{ and } \hat{\beta}' = \cdots \text{ ?}$$

The sample size is $n = 24$, $\hat{y} = X'\hat{\beta} = 0.27 + 0\mathbf{1}(a=1) - 0.031\mathbf{1}(a=2) -$
$0.021\mathbf{1}(a=3) + \cdots + 0\mathbf{1}(b=1) + \cdots - 0.231\mathbf{1}(b=5) - 0.131\mathbf{1}(b=6)$

What is β and $\mathbf{X}$ for $\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$?

$\beta = (\mu, a_2, a_3, a_4, b_2, ..., b_6)'$ and

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccccccc} int & a2 & a3 & a4 & b2 & \cdots & b6 \end{array} \\ \left(\begin{array}{ccccccc} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & & & \\ 1 & 0 & 0 & 1 & 0 & \cdots & 1 \end{array} \right)_{n\times 9} \end{array} \quad \textbf{Why ??}$$

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

Homework. 10.2. What is X and β for $\hat\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ in $lm(y \sim a+b-1)$ in Example 3 ?

Example 3 (continued).
Another way to generate the same type of data:

```
> y=rnorm(24)
> a=rep(1,6)
> a=c(a,a+1,a+2,a+3)
> b=rep(1:6,4)
> lm(y~a+b) # #A
     (Output)
     (Intercept) a b
> a=factor(a)
> b=factor(b)
> lm(y~a+b) # #B
     (Output)
     (Intercept) a2 a3 a4 b2 b3 b4 b5 b6
```

What is X and β for $\hat\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ in # A ?
What is X and β for $\hat\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ in # B ?
What is the difference between outcomes # A and # B ?

```
     > lm(y~a+b−1) #1.
     > lm(y~a+b) #2
     > options(contrasts =c("contr.sum", "contr.poly")) #3.
     > lm(y~a+b)
```

Which is the way same as in Example 3 ? #A or #B ?
What do you expect the estimates $\underline{before}$ **seeing output ?**

```
> summary(lm(y~a+b))                    # justify the answer to the question
Coefficients:
     Estimate  Std. Error  t value  Pr(> |t|)
```

How to find the P-value to justify the answer to the previous question?
Is it Pr($>|t|$) ?

Homework 10.3.
1. Repeat Example 3 once yourself and answer the questions there.
2. Mimic Example 3 (continued) by inserting y=1+2*a+y right after b=rep(1:6,4) (not before each lm(y∼a+b),as $2 * factor(a)$ does not work).
   Then ask yourself relevant questions and answer them.

Hw due Wednesday before class. **Late hw -3, submit both .tex file and .pdf file !**

In regression analysis, there are several issues:
1. What is model for the data ? (LR, non-LR, Cox, Parametric) model ?
2. Can the model be simplifies ?
3. Does the model fit the data ?

Model checking
Question. Does a given set of data fits the given model (LR, non-LR, Cox, Lehmann, Parametric) ?

Ans. Various diagnostic plots, QQplots, residual plots, and model tests.
For example, for question about the LR model, test

H_0: $Y = \beta X + \epsilon$ v.s. H_1: $Y \neq \beta X + \epsilon$, $\qquad X \in \mathcal{R}^p$.

Two common approaches.

1. A check of model fit. If there are replications in X_i's, that is, the model

$Y_i = \beta' X_i + \epsilon_i$, $i = 1, ..., n$,

can be written as

$Y_{ij} = \beta \mathbf{X}_{ij} + \epsilon_{ij}$, where
$j = 1, ..., J_i$,
$i = 1, ..., m$,
$X_{i1} = \cdots = X_{iJ_i}$, with $J_i > 1$ for some i,
and $X_{ij} \neq X_{kh}$ if $i \neq k$, $\qquad$ *e.g*, $(X_1, ..., X_6) = (2, 2, 2, 1, 3, 3)$.

then a model lack-of-fit test of H_0^l: $\sigma_L = \sigma_E$ v.s. H_1^l: $\sigma_L \neq \sigma_E$
$\phi = \mathbf{1}(m_L/m_E > F_{df_L, df_E, \alpha})$, where
$m_E = \frac{1}{df_E} \sum_{i,j} (Y_{ij} - \overline{Y}_{i\cdot})^2$, (unbiased estimator of σ^2 under NID $(E(Y_{ij}) = \alpha_i)$
$m_L = \frac{1}{df_L} \sum_{i,j} (\overline{Y}_{i\cdot} - \hat{Y}_{ij})^2$, (unbiased estimator of σ^2 under NID and LR Model)
$df_E = \sum_i (J_i - 1)$ $(= n - m)$ and
$df_L = m - p$, df of residuals) $\qquad = n - p - df_E = n - (p + df_E)$

Here, we make use of

$\sum_{i,j} Y_{ij}^2$
$= \sum_{i,j} (Y_{ij} - \overline{Y})^2 + \sum_{i,j} \overline{Y}^2 \qquad = \sum_{i,j} Y_{ij}^2 - 2\overline{Y} \sum_{i,j} Y_{ij} + \sum_{i,j} (\overline{Y})^2 + \sum_{i,j} \overline{Y}^2$
$= \sum_{i,j} (Y_{ij} - \hat{Y}_{ij})^2 + \sum_{i,j} (\hat{Y}_{ij} - \overline{Y})^2 + \sum_{i,j} \overline{Y}^2$

$$= \underbrace{\sum_{i,j} (Y_{ij} - \overline{Y}_{i\cdot})^2}_{\textit{relate to } m_E \textit{ or } m_L?} + \underbrace{\sum_{i,j} (\overline{Y}_{i\cdot} - \hat{Y}_{ij})^2}_{m_E \textit{ or } m_L?} + \sum_{i,j} (\hat{Y}_{ij} - \overline{Y})^2 + \sum_{i,j} \overline{Y}^2.$$

df: $(n - m) \qquad +(m - p) \qquad +(p - 1) + 1$.

We also make use of NID.

Second way. If there is no replication, add another function to the model

$Y = \beta X + \epsilon$,

e.g., consider a new model

$Y = \beta X + \theta X^2 + \epsilon$ (or $Y = \beta X + \theta g(X) + \epsilon$, *e.g.*, $g(x) = (x^3, x^2)$),

and check whether $\theta = 0$, where $\theta \in \mathcal{R}$ (or $\mathcal{R}^q$ if $g(X) \in \mathcal{R}^q$).
That is, set

H_0^t: $\theta = 0$, v.s. H_1^t: $\theta \neq 0$.

(a) One test is t-test (if $q = 1$):

$\phi = \mathbf{1}(|\hat{\theta}|/\hat{\sigma}_{\hat{\theta}} > t_{n-p,\alpha/2})$.

If n is large and p is not so, the statistic does not rely on $\epsilon \sim N(\mu, \sigma^2)$.

(b) Another test is F-test:

Assuming $E(Y|X) = \beta' X + \theta' g(X)$, H_0: $\theta = 0$ v.s. H_1: $\theta \neq 0$.
Write $\mathbf{Y}_{n \times 1} = \mathbf{Z}_{n \times (p+q)} \gamma + \mathbf{e}$, where $\mathbf{Y} = (Y_1, ..., Y_n)'$,

$$\mathbf{Z} = \begin{pmatrix} X_1' & g(X_1)' \\ \cdots & \\ X_n' & g(X_n)' \end{pmatrix},$$

$$\mathbf{X} = \begin{pmatrix} X_1' \\ \cdots \\ X_n' \end{pmatrix},$$

$$\gamma = \begin{pmatrix} \beta \\ \theta \end{pmatrix}.$$

Let $\mathbf{C} = (\underbrace{\mathbf{0}}_{q \times p} \ \underbrace{I}_{q \times q})$, where I is an identity matrix.

The original H_0 becomes

H_0^f: $\mathbf{C}\gamma = \theta = 0$.
$\hat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{Y}$,
$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$,

SSE= $\mathbf{Y}'\mathbf{Y} - \hat{\gamma}'\mathbf{Z}'\mathbf{Y}$ (= $||\mathbf{Y} - \hat{\gamma}'\mathbf{Z}||^2$), df= ?
SSW= $\mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y}$ (= $||\mathbf{Y} - \hat{\beta}'\mathbf{X}||^2$), df= ?

An F test is

$\phi = \mathbf{1}(\frac{\frac{SSW-SSE}{q}}{\frac{SSE}{n-p-q}} > F_{q,n-p-q,\alpha})$,

where q is the dimension of θ, which is 1 most of the time.

F-test relies on NID.

3 tests are introduced: (1) Lack of fit test if there are ties in X_i's, (2) t-test or F-test.

Q: 1. If there exist ties in X_i's, can we use all three approaches ?
2. If there do not exist ties in X_i's, can we use all three approaches ?

Impurity data. An experiment to determine how the initial rate of formation of an undesirable impurity (wu1dian3) Y depended on two factors:
(1) the concentration X_0 of monomer, (dan1ti3)
(2) the concentration X_1 of dimer. (shuang1ti3)

The relation is expected to be

$Y = \beta_0 X_0 + \beta_1 X_1 + \epsilon$.

The data are as follows.

order in experiment	X_0	X_1	Y	i	ij
3	0.34	0.73	5.75	1	11
6	0.34	0.73	4.79	2	12
2	0.58	0.69	5.44	3	21
4	1.26	0.97	9.09	4	31
1	1.26	0.97	8.59	5	32
5	1.82	0.46	5.09	6	41
why ordered ?					
define	$\mathbf{X}_0$	$\mathbf{X}_1$	$\mathbf{Y}$		

Can we use **all three approaches** for checking H_0: $E(Y|\mathbf{X}) = \beta_0 X_0 + \beta_1 X_1$?

Notice: $n = 6$, $i = 4$, $J_1 = J_3 = 2$ and $J_2 = J_4 = 1$.

$m_E = \frac{1}{df_E}\sum_{i,j}(Y_{ij} - \overline{Y}_{i\cdot})^2 = \frac{1}{2}(\frac{(5.75-4.79)^2}{2} + \frac{(9.09-8.59)^2}{2})$ **why ??**

((5.75-4.79)**2+ (9.09-8.59)**2)/4

[1] 0.2929

$m_L = \frac{1}{df_L}\sum_{i,j}(\overline{Y}_{i\cdot} - \hat{Y}_{ij})^2 = ?$

$\hat{Y}_{ij} = \hat{\beta}\mathbf{X}_{i,j}$= ?

$$\hat{\beta} = \begin{pmatrix} \mathbf{X}_0'\mathbf{X}_0 & \mathbf{X}_0'\mathbf{X}_1 \\ \mathbf{X}_0'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_1 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_0'\mathbf{Y} \\ \mathbf{X}_1'\mathbf{Y} \end{pmatrix}$$

$$= \frac{1}{\mathbf{X}_0'\mathbf{X}_0\mathbf{X}_1'\mathbf{X}_1 - (\mathbf{X}_0'\mathbf{X}_1)^2} \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & -\mathbf{X}_0'\mathbf{X}_1 \\ -\mathbf{X}_0'\mathbf{X}_1 & \mathbf{X}_0'\mathbf{X}_0 \end{pmatrix} \begin{pmatrix} \mathbf{X}_0'\mathbf{Y} \\ \mathbf{X}_1'\mathbf{Y} \end{pmatrix}$$

Too tedious, thus use R

```
> x=c(0.34,0.73,5.75,
    0.34,0.73,4.79,
    0.58,0.69,5.44,
    1.26,0.97,9.09,
    1.26,0.97,8.59,
    1.82,0.46,5.09)
> dim(x)=c(3,6)
# y=lm(x[3,]~x[1,]+x[2,]−1)
> x=t(x)
> y=lm(x[,3]~x[,1]+x[,2]−1)
> y
    Coefficients:
```

$x[,1]$	$x[,2]$
1.207	7.123

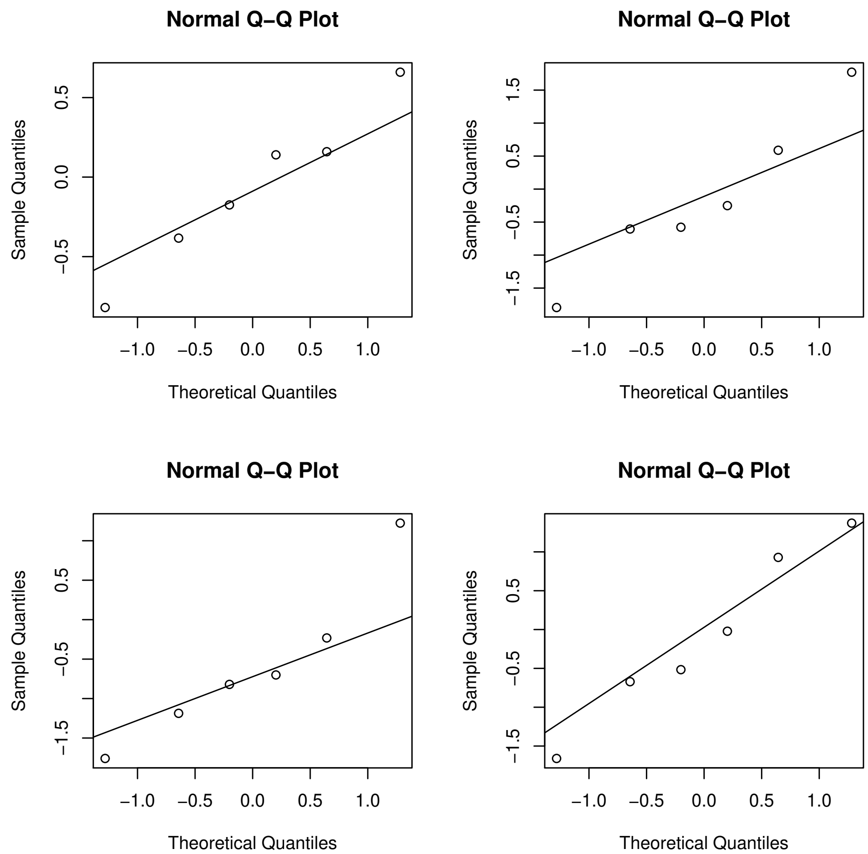


Fig. 10.1. QQ-plot data and rnorm(6) (3 times) Why do this ?

```
par(mfrow=c(2,2))
qqnorm(y$resid)
qqline(y$resid)
z=rnorm(6)
qqnorm(z)
qqline(z)
......
```

```
> anova(y)
```

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$	
$x[,1]$	1	207.693	207.693	624.29	$1.523e-05$	$***$
$x[,2]$	1	58.901	58.901	177.05	0.0001844	$***$
$Residuals$	4	1.331	0.333			

Lack of fit test $\phi = \mathbf{1}(F > F_{I-1,I(J-1),\alpha})$.

```
> z=c(1,1,2,3,3,4)
#z=factor(x[,1])
> Y=lm(x[,3]~x[,1]+x[,2]+factor(z)−1)
> anova(Y)
```

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$	
$x[,1]$	1	207.693	207.693	709.0903	0.001407	$**$
$x[,2]$	1	58.901	58.901	201.0966	0.004936	$**$
$factor(z)$	2	0.745	$m_L = 0.372$	$\frac{m_L}{n_E} = 1.2717$	0.440202	$p-value > 0.05?$
$Residuals$	2	0.586	$m_E = 0.293$			

f=1/1.27

1-pf(f,2,2)

Conclusion Do not reject the model,

and the data fit the linear regression model.

Are we done ?

The second way: H_o: $Y = \beta X + \epsilon$ v.s. H_1: $Y = \beta X + \theta g(X) + \epsilon$ with $\theta \neq 0$.

```
> z=lm(x[,3]~x[,1]+x[,2]+x[,1]*x[,2]−1)
> summary(z)
```

	Estimate	*Std. Error*	*t value*	$Pr(>\|t\|)$	
$x[,1]$	0.3844	0.5171	0.743	0.51120	
$x[,2]$	6.4990	0.5226	12.437	0.00112	$**$
$x[,1]:x[,2]$	1.6812	0.8668	1.939	0.14779	$p-value > 0.05$?

Conclusion ?

What else needs to be done ?

The third way:

```
> anova(z)
```

	Df	*Sum Sq*	*Mean Sq*	*F value*	$Pr(>F)$	
$x[,1]$	1	207.693	207.693	1055.2702	$6.411e-05$	$***$
$x[,2]$	1	58.901	58.901	299.2726	0.0004209	$***$
$x[,1]:x[,2]$	1	0.740	0.740	3.7615	0.1477919	$p-value > 0.05$?
Residuals	3	0.590	0.197			

Conclusion ?

Another code:

```
> anova(y,z)
```

Model 1: x[, 3] ∼ x[, 1] + x[, 2] - 1

Model 2: x[, 3] ∼ x[, 1] + x[, 2] + x[, 1] * x[, 2] - 1

	Res.Df	*RSS*	*Df*	*Sum of Sq*	*F*	$Pr(>F)$
1	4	1.33075				
2	3	0.59044	1	0.74031	3.7615	0.1478

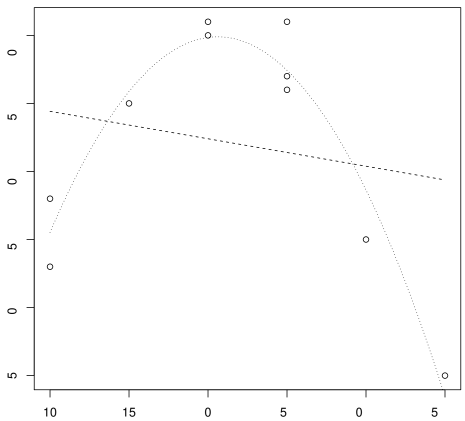
Example 4 (Growth rate data). The data in Table 10.7 is for the growth rate of rats (denoted by Y) fed various doses of a dietary supplement (denoted by X). From similar investigation, it was believed that the relation could be roughly linear. We shall test two models: a simple linear model and a quadratic model.

H_o: $E(Y|X) = \alpha + \beta X$ v.s. H_1: $E(Y|X) \neq \alpha + \beta X$.

```
y=c(73,78,85,90,91,87,86,91,75,65) # rate
x=c(10,10,15,20,20,25,25,25,30,35) # dose
a=factor(c(1,1,2,3,3,4,4,4,5,6))
#a=factor(x)
z=lm(y~x)
plot(x,y)
v=(100:350)/10
u=z$coef[1]+z$coef[2]*v
lines(v,u,lty=2)
z=lm(y~x+I(x^2))
z=z$coef
u=z[1]+z[2]*v+z[3]*v^2
lines(v,u,lty=3)
z=lm(y~x+a)
anova(z) # lack of fit test
```

	Df	*Sum Sq*	*Mean Sq*	*F value*	$Pr(>F)$	
x	1	24.5	24.502	3.6299	0.12946	
a	4	659.4	164.850	24.4222	0.00452	$**$
Residuals	4	27.0	6.750			

Conclusion: The linear regression model does not fit the data.



Now consider

H_o: $E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$ v.s. H_1: $E(Y|X) \neq \beta_0 + \beta_1 X + \beta_2 X^2$

First way, lack of fit.

z=lm(y∼x+I(xˆ 2)+a)

anova(z)

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$	
x	1	24.50	24.50	3.6299	0.1294567	
$I(x^2)$	1	641.20	641.20	94.9933	0.0006207	$***$
a	3	18.19	6.06	0.8985	0.5156739	
$Residuals$	4	27.00	6.75			

Conclusion: The quadratic regression model does fit the data.

Second way: H_o: $\beta_3 = 0$ v.s. H_1: $\beta_3 \neq 0$.

assuming $E(Y|X) = \beta_o + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$.

z=lm(y∼x+I(xˆ 2)+I(xˆ 3))

summary(z)

	$Estimate$	$Std.Error$	$tvalue$	$Pr(>\|t\|)$	
$(Intercept)$	24.007599	19.712021	1.218	0.2690	
x	7.198068	3.179330	2.264	0.0642	. **Conclusion ?**
$I(x^2)$	-0.222267	0.153348	-1.449	0.1974	
$I(x^3)$	0.001409	0.002276	0.619	0.5585	

Third way: H_o: $\beta_3 = 0$ v.s. H_1: $\beta_3 \neq 0$.

> anova(z)

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$	
x	1	24.50	24.50	3.4608	0.1122	
$I(x^2)$	1	641.20	641.20	90.5674	$7.677e-05$	**Conclusion ?**
$I(x^3)$	1	2.71	2.71	0.3834	0.5585	
$Residuals$	6	42.48	7.08			

It seems that the data fit $Y \sim x^2$. How to check it ?

> Z=lm(y∼I(xˆ 2)+a)

> anova(Z)

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$	
$I(x^2)$	1	91.42	91.421	13.544	0.021200	**Conclusion ?**
a	4	592.48	148.120	21.944	0.005533	
$Residuals$	4	27.00	6.750			

Compare to z=lm(y∼x+I(xˆ 2)+a)

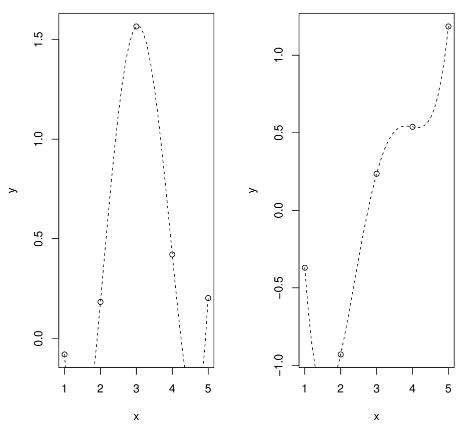
Q: Is it true that

the regression model fits the data if the regression curve fits the data well ?
The answer can be found from the next example.

Example 5. Consider the model $logY \sim N(0,1)$.

```
x=1:5
y=exp(rnorm(5))
plot(x,y)
z=lm(y~x+I(x^ 2)+I(x^ 3)+I(x^ 4))
z=z$coef
v=(10:50)/10
u=z[1]+v*z[2]+z[3]*v^ 2+z[4]*v^ 3+z[5]*v^ 4
lines(v,u,lty=2)
```

Then $\hat{Y}_i = Y_i$ for all i. However, the data do not fit the polynomial regression model.
If we do it again the equation is totally different.



Model checking: Given a regression data set (X_i, Y_i)'s, to make statistical inferences on μ_Y, σ_Y, $Y = g(X) + \epsilon$ etc. we need to assume a certain model:
parametric models: normal ? exponential ? uniform ? etc.,
semi-parametric models: LR, NLR, Cox, Lehmann, among others.
which of them is appropriate ? Or none of them is ?
For example, if one choose LR model, say

$$Y_i = \beta X_i + \epsilon_i, \text{ and } \epsilon \sim N(0, \sigma^2), \quad (1)$$

and we can fit the data to the model and get the LSE of β, $F_{Y|X}$, SE of β, CI of β, and do testing about β and $F_{Y|X}$.
After these, we should ask
is the model in Eq. (1) appropriate for the data ?
is NID valid ? etc....
This is model checking. The tools are model diagnostic plots and model checking tests.
Example 6. Simulation studies on testing H_0: $Y = \beta X + W$ **(or with NID)** v.s. H_1: H_0 is not true (*i.e.* $Y \neq \beta X + W$ or NID is not true) with the R codes
(summary(y~x+I(sin(x)))$coef[3,4]> 0.05) # test $\phi = \mathbf{1}(p - value \leq 0.05)$ (1)
Ideally it tests H_0: $Y = \beta X + W$ **or with NID** v.s. H_1: $Y \neq \beta X + W$,
actually it tests H_0': $\theta = 0$ v.s. H_1': $\theta \neq 0$, under the assumption

$$Y = \beta X + \theta \sin X + W \textbf{ and NID}. \quad (6.1)$$

Simulation 1.

True model: $Y = X + \epsilon$, where $\epsilon \sim N(0,1)$ and $X \sim bin(3, 0.5)$.

Questions:

Is H_1' true ? How about H_1 ? Is H_0 true ? How about NID ?

What do you expect for the test ?

Sample size= 50, replication= 1000, $\beta = 1$, $\overline{\hat{\beta}}$= 0.996, sd= 0.17

Rate of accepting right H_0 is 0.952, $\hat{P}(H_1|H_0) = 0.048$, $\hat{P}(H_1'|H_0) = 0.048$, $P(H_1|H_0) = ?$

Does the test work as expected ?

What do you expect if $n = 5000$?

$\hat{P}(H_0|H_1)$= 0.952 ? $\qquad\qquad\qquad = ?$

Simulation 2.

True model: $Y = \sin X + \epsilon$, where $X \sim bin(3, 0.5)$ and $\epsilon \sim N(0,1)$.

H_0: $Y = \beta X + W$, H_1: $Y \neq \beta X + W$, H_1': $\theta \neq 0$ under assumption (6.1).

Questions: Is H_1' true ? How about H_1 ? Is H_0 true ? How about NID ?

What do you expect for the test ?

Sample size= 50, replication= 100, $\beta = 0$, $\overline{\hat{\beta}} = -0.003$, sd= 0.03

Rate of accepting wrong H_0 is 0.33. $\hat{P}(H_0|H_1) = ?$ $\hat{P}(H_0|H_1') = ?$

$\hat{P}(H_1|H_0) = 1 - 0.33$?

Does the test work in this case ?

What do you expect if $n = 5000$ **?** $\hat{P}(H_0|H_1) \to 0$? $\quad \hat{P}(H_0|H_1) \to 1$?

Simulation 3.

True model: $Y = X^{1/2} + \epsilon$, where $\epsilon \sim N(0,1)$ and $X \sim bin(3, 0.5)$.

H_0: $Y = \beta X + W$, H_1: $Y \neq \beta X + W$, H_1': $\theta \neq 0$ under assumption (6.1).

Questions: Is H_1' true ? How about H_1 ? Is H_0 true ? How about NID ?

What do you expect for the test ?

Sample size= 5000, replication= 100, β= 0, $\overline{\hat{\beta}}$= 0.5150, sd= 0.1761

Rate of accepting wrong H_0 is 0.00. $\hat{P}(H_0|H_1) = 0.00$ **??** $\hat{P}(H_0|H_1') = 0.00$ **??**

It says that H_1' is true, the model is $Y = \alpha + \beta X + \theta \sin X + \epsilon$.

Does the test work in this case ?

Both H_0 and H_1' are wrong, though H_1 is true. It happens to work.

Simulation 4.

True model $Y = X^{1/2} + \epsilon$, where $X \sim B * |W|$, $B \sim U(0,3)$, $B \perp W$, and ϵ and $W \sim Cauchy$.

H_0: $Y = \beta X + W$, H_1: $Y \neq \beta X + W$, H_1': $Y = \beta X + \theta \sin X + W$, $\theta \neq 0$.

Questions: Is H_1' true ? How about H_1 ? Is H_0 true ? How about NID ?

What do you expect for the test ?

Sample size= 5000, replication= 100, $\beta = 0$, $\overline{\hat{\beta}}$= 0.0149, sd= 0.0141

rate of accepting wrong H_0 is 0.96. $\hat{P}(H_0|H_1) = 0.96$? $\hat{P}(H_0|H_1') = 0.96$?

Does the test work in this case ?

Remark. The homework solution is in my website. Quiz on 447 and 448 on Friday.

The codes for simulations 1-4 are as follows.

```
n=5000 # need to adjust for input sample
beta=1
NN=100 # No. of simulation replication
swb = 1 # switch for binomial covariant
swn = 0 # switch for normal error
sww = 1 # switch for wrong LR model
p=0 # No. of accepting H0
b=0 # LSE
s=0 # SD of LSE
for (N in 1:NN) {
    c=rbinom(n,3,0.5)
    if (swb == 0)
        c=abs(rcauchy(n))*c
```

```
c=sort(c)
e=rcauchy(n)
if (swn == 1)
    e=rnorm(n)
y=beta*c+e
if (sww == 1)
    y=beta*sqrt(c)+e
z=lm(y~c+I(sin(c)))
b=b+z$co[2]
s=s+z$co[2]*z$co[2]
p=p+(summary(z)$coef[3,4]>0.05) }
(p=p/NN)
(b=b/NN)
(s=sqrt(s/NN-b*b))
summary(z)$coef
```

	Estimate	*Std. Error*	*t value*	$Pr(>\vert t\vert)$	(needs NID)
$(Intercept)$	-4.8670290	3.324966	-1.4637829	0.14331617	
c	2.8889273	1.472481	1.9619452	0.04982429	
$I(sin(c))$	-0.5723319	3.625139	-0.1578786	0.87455884	

Example 7. Simulation on testing
H_0: $Y = \beta \sin X + W$ v.s. H_1: H_0 is false.
H_1': $Y = \beta \sin X + \theta X + W$, with $\theta \neq 0$, under NID.
The R codes
(summary(y~x+I(sin(x)))$coef[2,4]> 0.05) # test $\phi = \mathbf{1}(p{-}value \leq 0.05)$
True model $Y = X + W$, where $X \sim bin(3, 0.5)$ and $W \sim |Cauchy|$.
Questions:
Is H_1 true ?
Is H_1' true ?
Is H_0 true ?
What do you expect for the test ?
Sample size= 50, replication= 1000, $\theta = 1$, $\hat{\theta}$= 2.324, sd= 47.06,
rate of accepting H_0 is 0.795. $\hat{P}(H_0|H_1) = 0.795$? $\hat{P}(H_0|H_1') = 0.795$? $\hat{P}(H_1|H_0) = 1 - 0.795$?

Summary on the simulation studies.
There are many regression models:
the linear regression models,
the logistic regression models,
the generalized linear (regression) models,
the generalized additive models, etc..
Given a data set, one needs to check which model fits the data. This is to test
H_0: the data fits a given model, *e.g.*, $E(Y|X) = \beta' X$, v.s. H_1: H_0 is false.
To implement, people design H_1' instead. If H_0 and H_1' are not properly designed, then the previous 3 model checking tests can be misleading as in simulations 3 and 4 of Ex. 6.
The existing model checking tests are the tests of
H_0^t: $\xi(\cdot) = 0$, v.s. H_1^t: $\xi(\cdot) \neq 0$, where $\xi(\mathbf{X}) = E(Y|\mathbf{X}) - \beta'\mathbf{X}$ has a certain form with NID.
e.g., $\xi = \theta g(\mathbf{X})$ in the 3 aforementioned model checking tests. In order to establish the distribution theories for the tests, each of these tests imposes certain regularity conditions on $F_{\mathbf{X},Y}$ such as NID, which specifies a parameter space for $F_{\mathbf{X},Y}$, say Θ_p, under which the test is valid. The Θ_p depends on the specific test and is a certain common regression model that contains Θ_0. For instance, in Example 6,

$$\Theta_p = \{F_{\mathbf{X},Y} : \ Y = \alpha + \beta X + \theta \sin X + \epsilon, \ \epsilon \sim N(0, \sigma^2), \ X \perp \epsilon\}.$$

Thus $\Theta_p \neq \Theta$, the family of all cdfs $F_{\mathbf{X},Y}$. If $F_{\mathbf{X},Y} \notin \Theta_p$, these tests are <u>*invalid*</u> in

the sense that the (asymptotic) distributions specified for these tests are false.
In simulation 1, H_0 is true and the model assumptions holds,

the test can either reject H_0 or do not reject. But $P(H_1|H_0) = 0.05$.
In simulation 2, H_1 is true, and the assumptions for the t-test hold.

The test rejects H_0 with probability $\to 1$ as $n \to \infty$
($P(H_0|H_1) \to 0$, $P(H_0|H_1') \to 0$, a consistent test). $P(H_1|H_0)$= ?
In simulations 3 and 4, both H_0 and H_1' are false, thus no $P(H_0|H_1')$.

The test can reject H_0 with probability 0 or 0.96, *i.e.*, $P(H_0|H_1)$ can be ≈ 0 or 0.96.

In Example 7, both H_0 and H_1' are false, as NID is false and $E(W|X)$ does not exist.

An estimate of $P(H_0|H_1)$ is ≈ 0.8.

Remark. Type I error, denoted by $H_1|H_0$, implies that H_0 is true. In Simulation 1 of Ex.6, it is true that $P(H_1|H_0) = 0.05$. It works as expected.
Type II error, denoted by $H_0|H_1$, implies that H_1 is true. In Simulation 2 of Ex.6, it is true that $P(H_0|H_1) \approx 0.33 \to 0$, as $n \to \infty$. It works as expected.
In Simulations 3 and 4 of Ex.6 and in Example 7, neither H_0 nor H_1' is true. Thus neither $P(H_1|H_0)$ nor $P(H_0|H_1')$ is a proper term. The test is based on invalid assumption in (6.1) Thus the test is not valid. Just like a random guess. $\hat{P}(H_0|H_1)$ can be ≈ 0, 0.8, 1.

Interpretation of one way anova $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $i = 1, 2, 3$; $j = 1, ..., 10$.
Another way:

$$E(Y_{ij}|X) = \mu + \sum_{i=1}^{3} \alpha_i \mathbf{1}(\text{Treatment= i for the } ij\text{−th person}) \tag{1}$$

There are 3 treatments, each is applied to 10 people. α_i is the effect of treatment i.
From Eq. (1), there are 3 equations and 4 unknown variables (due to $i \in \{1, 2, 3\}$.

$E(Y_{1j}|X) = \mu + \alpha_1$,
$E(Y_{2j}|X) = \mu + \alpha_2$,
$E(Y_{3j}|X) = \mu + \alpha_3$, j=1,...,10.

(1) $\mu = 0$. α_i is the average effect of treatment i.
(2) $\alpha_1 = 0$. μ is the average effect of treatment 1. α_i is the deviation effect of treatment i from treatment 1. (Obviously $\alpha_1 = 0$).
(3) $\sum_{i=1}^{3} \alpha_i = 0$. μ is the average effect of the 3 treatments. α_i is the deviation effect of treatment i from the average.

```
x=1:3
x=rep(x,10)
y=4*x+rnorm(30)
lm(y~x)
lm(y~ factor(x)-1)
lm(y~ factor(x))
options(contrasts =c("contr.sum", "contr.poly"))
lm(y~ factor(x))
```

$lm(y \sim x)$	$(Intercept)$	x		
$\hat{\beta}$	-0.5422	4.2430		
β	0	4		
$lm(y \sim factor(x) - 1)$		$factor(x)1$	$factor(x)2$	$factor(x)3$
$\hat{\beta}$		3.757	7.832	12.243
β		4	8	12
$lm(y \sim factor(x))$	$(Intercept)$		$factor(x)2$	$factor(x)3$
$\hat{\beta}$	3.757	?	4.075	8.486
β	4	0	4	8
$contr.sum$				
$lm(y \sim factor(x))$	$(Intercept)$	$factor(x)1$	$factor(x)2$	
$\hat{\beta}$	7.9439	-4.1871	-0.1119	?
β	8	-4	0	4

Remark.

Once contr.sum is applied, it remains there unless we apply
options(contrasts =c(”contr.treatment”, ”contr.poly”))

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 is the averages of the $t-$th sample (a parameter).

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

If $I = 2$, we use $t-$ test.

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

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

$X_{tj} = (X_{tj1}, X_{tj2}, X_{tj3})$, where $X_{tjk} = \mathbf{1}(t = k)$.

Remark. The model is $E(Y_{tj}) = \tau_t$, $t \in \{1, 2, 3\}$, $j \in \{1, 2\}$, which is often written as

$$E(Y_{tj}) = \tau_t = \eta + \alpha_t \tag{1}$$
$$E(Y_{tj}) = \tau_t = \eta + \alpha_t \text{ with } \alpha_1 = 0 \tag{2}$$
$$E(Y_{tj}) = \tau_t = \eta + \alpha_t \text{ with } \sum_{j=1}^{3} \alpha_j = 0 \tag{3}$$
$$E(Y_{tj}) = \tau_t = \alpha_t \text{ with } \eta = 0 \tag{4}$$

We say the parameters in Eq. (1) are not identifiable, as $\exists$ infinitely many solutions, *e.g.*,

$(\eta, \alpha_1, \alpha_2, \alpha_3) = (0, \tau_1, \tau_2, \tau_3)$,
$(\eta, \alpha_1, \alpha_2, \alpha_3) = (\tau_1, \tau_1 - \tau_1, \tau_2 - \tau_1, \tau_3 - \tau_1)$
$(\eta, \alpha_1, \alpha_2, \alpha_3) = (\overline{\tau}, \tau_1 - \overline{\tau}, \tau_2 - \overline{\tau}, \tau_3 - \overline{\tau})$ $\qquad (\overline{\tau} = \sum_{t=1}^{3} \tau_t/3)$

are 3 solutions to Eq. (1).

Since the parameters in Eq. (1) are not identifiable, the LSE cannot be uniquely determined. Thus we either set $\eta = 0$, or $\alpha_1 = 0$ or $\sum_{t=1}^{I} \alpha_t = 0$.

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.

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

due to NID and

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

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

(What is the output in the following Table ?)

$$\text{The data are} \quad \begin{array}{cccc}
A & B & C & D \\
62^{(20)} & 63^{(12)} & 68^{(16)} & 56^{(23)} \\
60^{(2)} & 67^{(9)} & 66^{(7)} & 62^{(3)} \\
63^{(11)} & 71^{(15)} & 71^{(1)} & 60^{(6)} \\
59^{(10)} & 64^{(14)} & 67^{(17)} & 61^{(18)} \\
63^{(5)} & 65^{(4)} & 68^{(13)} & 63^{(22)} \\
59^{(24)} & 66^{(8)} & 68^{(21)} & 64^{(19)}
\end{array}, \; I = 4,\; J = 6,$$

$$\begin{array}{ccccc}
\textit{Source of variation} & \textit{sum of squares} & \textit{df} & \textit{mean square} & F \\
\textit{Between treatments} & S_T = 228 & \nu_T = 3 & m_T = 76 & \\
\textit{Within treatments} & S_R = 112 & \nu_R = 20 & m_R = 5.6 & 13.57 \\
 & & & & \\
\textit{Between treatments} & S_T = \sum_{t,j}(\overline{Y}_{t,\cdot} - \overline{Y})^2 & \nu_T = I-1 & m_T = \frac{S_T}{\nu_T} & \\
\textit{Within treatments} & S_R = \sum_{t,j}(Y_{tj} - \overline{Y}_{t,\cdot})^2 & \nu_R = I(J-1) & m_R = \frac{S_R}{\nu_R} & \frac{m_T}{m_R}
\end{array}$$

```
> 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))
```

$$\begin{array}{cccc}
(\textit{Intercept}) & \textit{treatment}2 & \textit{treatment}3 & \textit{treatment}4 \\
6.100e+01 & 5.000e+00 & 7.000e+00 & -9.999e-15 \\
\overline{Y}_{1\cdot} & \overline{Y}_{2\cdot} - \overline{Y}_{1\cdot} & \overline{Y}_{3\cdot} - \overline{Y}_{1\cdot} & \overline{Y}_{4\cdot} - \overline{Y}_{1\cdot}
\end{array}$$

> 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 do we know it ?)**

P-values is 0.00005.

There is real difference between the mean coagulation times for the four different diets.

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
$(\hat{\eta}, \hat{\alpha}_1, ..., \hat{\alpha}_{I-1})' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

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 ? (see summary(lm(x~treatment-1)))
> 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 * **

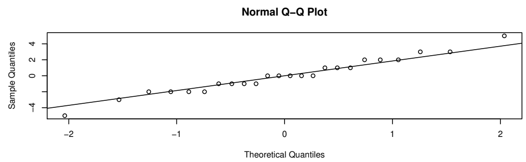
> a=lm(x~factor(treat)-1)
> summary(a) # compare "Estimate" in these two summaries.

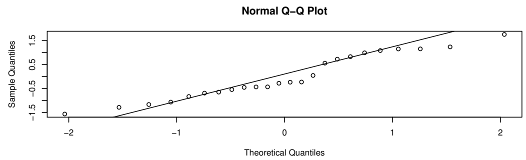
	Estimate	*Std. Error*	*t value*	*Pr(> |t|)*
*factor(treat)*1	61.0000	0.6667	91.50	< 2e − 16 * **
*factor(treat)*2	66.0000	0.9428	70.00	< 2e − 16 * **
*factor(treat)*3	68.0000	0.9428	72.12	< 2e − 16 * **

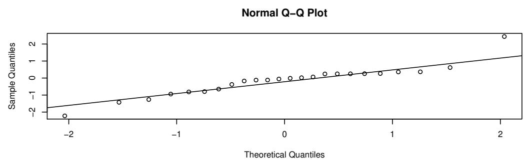
> anova(a) Analysis of Variance Table

	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, and 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	$S_R =$	$\nu_R =$	$m_R = \frac{S_R}{\nu_R}$	
and blocks	$\sum_{t,j}(Y_{t,j} - \overline{Y}_{t,\cdot} - \overline{Y}_{\cdot,j})^2$	$(k-1)(n-1)$		
Total about $\overline{Y}$	$S_D = \sum_{t,j}(Y_{tj} - \overline{Y})^2$	$\nu_D = kn-1$		
H_0^*	$S_B + S_T$	$\nu_B + \nu_T$		$\frac{\frac{S_B+S_T}{\nu_B+\nu_T}}{m_R}$

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

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))

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

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

$\hat{Y}_{11} = ?$

$\hat{Y}_{21} = ?$

$\overline{Y}_{2\cdot} + \overline{Y}_{\cdot 1} - \overline{Y}$

> summary(tr) Call: lm(formula = x ~ treatment + row)

	Estimate	*Std. Error*	*t value*	*Pr(> |t|)*
(Intercept)	$5.925e+01$	$1.328e+00$	44.630	$< 2e-16$ ***
treatment2	$5.000e+00$	$1.252e+00$	3.995	0.00117 **
treatment3	$7.000e+00$	$1.252e+00$	5.593	$5.14e-05$ ***
treatment4	$-1.088e-14$	$1.252e+00$	0.000	1.00000
row2	$1.500e+00$	$1.533e+00$	0.978	0.34335
row3	$4.000e+00$	$1.533e+00$	2.609	0.01973*
row4	$5.000e-01$	$1.533e+00$	0.326	0.74881
row5	$2.500e+00$	$1.533e+00$	1.631	0.12374
row6	$2.000e+00$	$1.533e+00$	1.305	0.21167

> u=lm(x~1)

> anova(u,tr) # **which null hypothesis does it test ?**

Model 1: x ~ 1

Model 2: x ~ treatment + row

	Res.Df	*RSS*	*Df*	*Sum of Sq*	*F*	*Pr(> F)*	
1	23	340.0					
2	15	70.5	8	269.5	7.1676	0.0005797	***

> anova(tr)

	Df	*Sum Sq*	*Mean Sq*	*F value*	*Pr(> F)*
treatment	3	228.0	76.0	16.170	$5.745e-05$
row	5	41.5	8.3	1.766	0.1806
Residuals	15	70.5	4.7		

how many tests ?

$\frac{228+41.5}{3+5}/4.7 = 7.167553$ indent

> aov(x~treatment+row)

	treatment	*row*	*Residuals*	
Sum of Squares	228.0	41.5	70.5	(see columns 1&2 of anova(tr))
Deg. of Freedom	3	5	15	

Residual standard error: 2.167948 ($= \sqrt{4.7}$).

Ans: Reject H_0^* and H_0, but not H_0', the row effect is not significant, the model should be $x \sim$ treatment

$x = 61$treatment[1 or 4] $+66$treatment[2]$+$ 68treatment[3]

($x = 61 \cdot \mathbf{1}$(treatment is type 1 or 4)$+66 \cdot \mathbf{1}$(treatment is type 2)$+68 \cdot \mathbf{1}$(treatment is type 3)

Derive the LSE directly for two way anova:

$Y_{ij} = \eta + \alpha_i + \gamma_j$, $i \in \{1, ..., I\}$, $j \in \{1, ..., J\}$,

$\sum_i \alpha_i = \sum_j \gamma_j = 0$ (contr.sum) (the simplest way).

$$=> \quad \sum_i \sum_j Y_{ij}/n = \sum_i \sum_j (\eta + \alpha_i + \gamma_j)/n = \eta + \sum_j \sum_i \alpha_i/n + \sum_i \sum_j \gamma_j/n.$$

$\overline{Y} = \eta$, $=> \hat\eta = \overline{Y}$; (due to MME).

$\overline{Y}_{i\cdot} = \eta + \alpha_i$, $i \in \{1, ..., I\}$, $=> \hat\alpha_i = \overline{Y}_{i\cdot} - \overline{Y}$;

$\overline{Y}_{\cdot j} = \eta + \gamma_j$, $j \in \{1, ..., J\}$, $=> \hat\gamma_j = \overline{Y}_{\cdot j} - \overline{Y}$;

$\hat{Y}_{ij} = \hat\eta + \hat\alpha_i + \hat\gamma_j = \overline{Y}_{i\cdot} + \overline{Y}_{\cdot j} - \overline{Y}$.

For instance, if $(I, J) = (3, 2)$, $\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{21} \\ Y_{31} \\ Y_{12} \\ Y_{22} \\ Y_{32} \end{pmatrix}$, $\beta = \begin{pmatrix} \eta \\ \alpha_1 \\ \alpha_2 \\ \gamma_1 \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 0 & -1 \\ 1 & 0 & 1 & -1 \\ 1 & -1 & -1 & -1 \end{pmatrix}$

$(\hat\eta, \hat\alpha_1, ..., \hat\alpha_{I-1}, \hat\gamma_1, ..., \hat\gamma_{J-1})' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ (contr.sum), $lm(y \sim row + col)$.

The LSE can also be derived by

$(\hat\eta, \hat\alpha_2, ..., \hat\alpha_I, \hat\gamma_2, ..., \hat\gamma_J)' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ (default), $lm(y \sim row + col)$.

$$\text{If } (I,J)=(3,2),\ \mathbf{Y}=\begin{pmatrix} Y_{11}\\ Y_{21}\\ Y_{31}\\ Y_{12}\\ Y_{22}\\ Y_{32}\end{pmatrix},\ \beta=\begin{pmatrix}\eta\\ \alpha_2\\ \alpha_3\\ \gamma_2\end{pmatrix},\ \mathbf{X}=\begin{pmatrix}1&0&0&0\\1&1&0&0\\1&0&1&0\\1&0&0&1\\1&1&0&1\\1&0&1&1\end{pmatrix}$$

How about $(\hat\alpha_1,\hat\alpha_2,...,\hat\alpha_I,\hat\gamma_2,...,\hat\gamma_J)'=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ (default) $(lm(y\sim row+col-1))$?

In two-way anova, $\hat Y_{ij}=\hat\eta+\hat\alpha_i+\hat\gamma_j=\overline{Y}_{i\cdot}+\overline{Y}_{\cdot j}-\overline{Y}$ is valid for the 3 models. It is easiest to derive the LSE through control.sum model, then to yield the other LSE's.

$lm(y\sim)$	$(Intercept)$	$r1$	$r2$	$r3$	$c1$	$c2$
$r+c$	$\overline{Y}_{1\cdot}+\overline{Y}_{\cdot 1}-\overline{Y}$	0	$\overline{Y}_{2\cdot}-\overline{Y}_{1\cdot}$	$\overline{Y}_{3\cdot}-\overline{Y}_{1\cdot}$	0	$\overline{Y}_{\cdot 2}-\overline{Y}_{\cdot 1}$
sum	$\overline{Y}$	$\overline{Y}_{1\cdot}-\overline{Y}$	$\overline{Y}_{2\cdot}-\overline{Y}$	$\overline{Y}_{3\cdot}-\overline{Y}$	$\overline{Y}_{\cdot 1}-\overline{Y}$	$\overline{Y}_{\cdot 2}-\overline{Y}$
$r+c-1$	0	$\overline{Y}_{1\cdot}+\overline{Y}_{\cdot 1}-\overline{Y}$	?	?	0	$\overline{Y}_{\cdot 2}-\overline{Y}_{\cdot 1}$

Key: $\hat Y_{ij}$ are the same in 3 forms. It is equivalent to the identifying the parameters in $E(Y_{ij})=\eta+\alpha_i+\gamma_j$:

	η	α_1	α_2	α_3	γ_1	γ_2
$i+j$	$E(Y_{1\cdot})+E(Y_{\cdot 1})-E(Y_{\cdot\cdot})$	0	$E(Y_{2\cdot})-E(Y_{1\cdot})$	$E(Y_{3\cdot})-E(Y_{1\cdot})$	0	$E(Y_{\cdot 2})-E(Y_{\cdot 1})$
sum	$E(Y_{\cdot\cdot})$					
$i+j-1$	0					

A simulation for understanding the estimates.

```
> y=rnorm(6)
> (col=gl(2,3,6))
   [1] 1 1 1 2 2 2
> (row=gl(3,1,6))
   [1] 1 2 3 1 2 3
> x=y
> dim(x)=c(3,2)
> (a=mean(x))
   [1] -0.3406383
> mean(x[1,])-a
   [1] 0.6422441
> mean(x[2,])-a
   [1] -0.2224916
> mean(x[,1])-a
   [1] -0.07302435
> options(contrasts =c("contr.sum", "contr.poly"))
> lm(y~row)
```

$(Intercept)$	$row1$	$row2$
-0.3406	0.6422	-0.2225
$\overline{Y}$	$\overline{Y}_{1\cdot}-\overline{Y}$	$\overline{Y}_{2\cdot}-\overline{Y}$

```
> lm(y~col)
```

$(Intercept)$	$col1$
-0.34064	-0.07302
$\overline{Y}$	$\overline{Y}_{\cdot 1}-\overline{Y}$

```
> lm(y~row+col)
```

$(Intercept)$	$row1$	$row2$	$col1$
-0.34064	0.64224	-0.22249	-0.07302
$\overline{Y}$	$\overline{Y}_{1\cdot}-\overline{Y}$	$\overline{Y}_{2\cdot}-\overline{Y}$	$\overline{Y}_{\cdot 1}-\overline{Y}$

> anova(lm(y~row+col)) **#What do you expect ?**

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$	$\hat\sigma$
row	2	0.63053	0.315267	1.2241	0.4496	0.561
col	1	0.07823	0.078233	0.3038	0.6369	0.279
$Residuals$	2	0.51508	0.257542			0.507
$row+col$	3	0.708	0.236	$\approx 1-$		0.486

What is (p, β, σ) in lm(y∼row+col) ?
What are the conclusions about H_0, H_0' and H_0^* ?
Are these null hypotheses really true ?

4.2. Randomized Block Designs

Penicillin Yield Example.
Yield due to 4 variants of the process A, B, C and D was obtained.
The raw experiment material (corn steep liquor) varied considerably.
Each blend of materials can make 4 runs.
So n=5 blends were prepared (ideally, randomly select 5 blends from possible more in storage), and $k = 4$ experiments were carried out for each blend.
First randomize the experiment by

```
rep(sample(1:4,replace=F),5)
```

which is the order to use processes A, B, C and D for the 5 blends. The data are given as follows.

blends\treatments	*A*	*B*	*C*	*D*
1	$89^{(1)}$	$88^{(3)}$	$97^{(2)}$	$94^{(4)}$
2	$84^{(4)}$	$77^{(2)}$	$92^{(3)}$	$79^{(1)}$
3	$81^?$	$87^?$	$87^?$	$85^?$
4	$87^?$	$92^?$	$89^?$	$84^?$
5	$79^?$	$81^?$	$80^?$	$88^?$

```
x=c(89,84,81,87,79, 88,77,87,92,81, 97,92,87,89,80, 94,79,85,84,88)
dim(x)=c(5,4)
# x=matrix(c(89,84,81,87,79, 88,77,87,92,81, 97,92,87,89,80, 94,79,85,84,88),ncol=4)
T = factor(as.vector(col(x))) # T=gl(4,5,20)
B = factor(as.vector(row(x))) # B=gl(5,1,20)
options(contrasts =c(”contr.sum”, ”contr.poly”))
(obj=lm(as.vector(x)~T+B))
anova(obj)
```

Consider 3 hypotheses:

H_o: $\tau_A = \cdots = \tau_D$ v.s. H_1: at least one inequality.
H_o': $\gamma_1 = \cdots = \gamma_5$ v.s. H_1': at least one inequality.
H_o^*: $\tau_A = \cdots = \tau_D$ and $\gamma_1 = \cdots = \gamma_5$ v.s. H_1^*: at least one inequality.

	Df	*Sum Sq*	*Mean Sq*	*F value*	*Pr(> F)*	
T	3	70	23.333	1.2389	0.33866	
B	4	264	66.000	3.5044	0.04075	*
Residuals	12	226	18.833			

$F\ value = (70 + 264)/(3 + 4)/18.833 \approx 2.5$ P-value ?

```
> 1-pf(2.5,7,12)
[1] 0.07821256
```

Conclusion How many statements ?

```
> summary(obj)
```

	Estimate	*Std. Error*	*t value*	*Pr(> |t|)*	
(Intercept)	86.0000	0.9704	88.624	$< 2e-16$	***
T1	−2.0000	1.6808	−1.190	0.25708	
T2	−1.0000	1.6808	−0.595	0.56292	
T3	3.0000	1.6808	1.785	0.09956	.
B1	6.0000	1.9408	3.092	0.00934	**
B2	−3.0000	1.9408	−1.546	0.14812	
B3	−1.0000	1.9408	−0.515	0.61573	
B4	2.0000	1.9408	1.031	0.32310	

Which constraint ?

The model can be simplified.

Should the model be $E(Y|X) = 86 + 6\mathbf{1}(B = 1)$? Or

```
> lm(x~ factor(B==1))
```

(Intercept)	*factor(B == 1)1*
88.25	−3.75

$$\hat{E}(Y|X) = 88.25 - 3.75 \underbrace{\mathbf{1}(B \neq 1)}_{\textbf{why not } =??}.$$

4.3. is skipped.

4.4. Latin squares Latin squares deal with the case that there are 2 more equal-level factors with the same level as the treatment. (R, C, T) v.s. (R, T).

Car Emissions Data. 4 drivers using 4 different cars to test the feasibility of reducing air pollution by modifying a gas mixture with very small amounts of certain chemicals A, B, C and D. There are 4 cars and 4 drivers. For randomization, randomly select cars and drivers. Then there are several ways to carry out the experiments.

(1) Convenient way: $\begin{array}{ccccc} Drivers\backslash cars & 1 & 2 & 3 & 4 \\ I & A & B & C & D \\ II & A & B & C & D \\ III & A & B & C & D \\ IV & A & B & C & D \end{array}$ car and treatment effects are confounded

(2) Simple randomization: $\begin{array}{ccccc} Drivers\backslash cars & 1 & 2 & 3 & 4 \\ I & D & A & C & B \\ II & D & A & B & C \\ III & C & B & A & D \\ IV & A & C & D & B \end{array}$ based on R output below

> rep(sample(c(“A”,“B”,“C”,“D”)),4)
[1] “D” “A” “C” “B” “D” “A” “B” “C” “C” “B” “A” “D” “A” “C” “D” “B”

(3) Latin Square: $\begin{array}{ccccc} Drivers\backslash cars & 1 & 2 & 3 & 4 \\ I & A & B & C & D \\ II & B & C & D & A \\ III & C & D & A & B \\ IV & D & A & B & C \end{array}$, which eliminates the block effects of cars and drivers, as each row and column has A, B, C, D

Compare $\begin{array}{ccccc} & 1 & 2 & 3 & 4 \\ I & A & B & C & D \\ II & D & A & B & C \\ III & B & C & D & A \\ IV & C & D & A & B \end{array}$, $\begin{array}{ccccc} & 1 & 2 & 3 & 4 \\ I & A & B & C & D \\ II & C & D & A & B \\ III & B & C & D & A \\ IV & D & A & B & C \end{array}$, $\begin{array}{ccccc} & 1 & 2 & 3 & 4 \\ I & A & B & C & D \\ II & C & D & A & B \\ III & D & A & B & C \\ IV & B & C & D & A \end{array}$

Relation between these 3 ?

The data are put in Table 2.

$Drivers\backslash cars$	1	2	3	4
I	A	B	D	C
	19	24	23	26
II	D	C	A	B
	23	24	19	30
III	B	D	C	A
	15	14	15	16
IV	C	A	B	D
	19	18	19	16

which pattern of the above 3 ?

> y=c(19, 24, 23, 26, 23, 24, 19, 30, 15, 14, 15, 16, 19, 18, 19, 16)
> (col=gl(4,1,16))
[1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4
Levels: 1 2 3 4
> (row=gl(4,4,16))
[1] 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4
Levels: 1 2 3 4
> T=c(A,B,D,C,D,C,A,B,B,D,C,A,C,A,B,D) # **Does it work ?**
> T=c(”A”,”B”,”D”,”C”,”D”,”C”,”A”,”B”,”B”,”D”,”C”,”A”,”C”,”A”,”B”,”D”)

> T=c(1,2,4,3,4,3,1,2,2,4,3,1,3,1,2,4)
> T=factor(T)
> (obj=lm(y∼col+row+T))

$(Intercept)$	$col2$	$col3$	$col4$	$row2$	$row3$
$2.000e+01$	$1.000e+00$	$-1.088e-15$	$3.000e+00$	$1.000e+00$	$-8.000e+00$
$row4$	$T2$	$T3$	$T4$		
$-5.000e+00$	$-4.000e-01$	$3.000e-01$	$1.000e+00$		

> anova(obj)

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$		
T	3	40	13.333	2.5	0.156490		<0.5
col	3	24	8.000	1.5	0.307174		car
row	3	216	72.000	13.5	0.004466	$**$	$driver$
$Residuals$	6	32	5.333				

> $(40+24+216)/9/5.333$
[1] 5.8
> 1−pf$(5.8, 9, 6)$ **what does it mean ?**
[1] 0.023
> (ob=lm(y∼T))

$(Intercept)$	$T2$	$T3$	$T4$
18	4	3	1

> anova(ob)

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$	
T	3	40	13.333	0.5882	0.6343	>0.5
$Residuals$	12	272	22.667			

H_o: $\tau_A = \tau_B = \tau_C = \tau_D$ v.s. H_1: H_o is false.
> summary(lm(y∼row))

	$Estimate$	$Std.\ Error$	$t\ value$	$Pr(>\|t\|)$	
$(Intercept)$	23.000	1.414	16.26	$1.54e-09$	$***$
$row2$	1.000	2.000	0.50	0.62612	
$row3$	-8.000	2.000	-4.00	0.00176	$**$
$row4$	-5.000	2.000	-2.50	0.02792	$*$

Conclusion ? Based on anova(obj) or anova(ob) ?
Ans: Based on anova(obj). The P-value of T is smaller.
Also row effect is significant, the model can be simplified as
$\hat{E}(Y|X) = 23 - 8\mathbf{1}(Drive_3) - 5\mathbf{1}(Drive_4)$? $\mathbf{1}(Drive_3) = \mathbf{1}$(driver is #3)
> D=rep(1,4)
> D=c(D,D,D+2,D+3)
> D [1] 1 1 1 1 1 1 1 1 3 3 3 3 4 4 4 4
> summary(lm(y∼factor(D)-1))

	$Estimate$	$Std.\ Error$	$t\ value$	$Pr(>\|t\|)$	
$(Intercept)$	23.5000	0.9707	24.209	$3.37e-12$	$***$
$factor(D)3$	-8.5000	1.6813	-5.055	0.00022	$***$
$factor(D)4$	-5.5000	1.6813	-3.271	0.00608	$**$

The model is $\hat{E}(Y|X) = 23.5 - 8.5\mathbf{1}(Drive_3) - 5.5\mathbf{1}(Drive_4)$.
Graeco-Latin Squares deal with the case that
there are 3 block factors with levels equal the level of the treatment factor (3+1),
whereas Latin squares deal with the case that
there are 2 equal-level factors with the same level as the treatment (2+1).
One may try to superimpose two Latin Squares together.
Which of the following two can eliminate confounding effect ?

1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
2	1	4	3	2	1	4	3	3	4	1	2	3	4	1	2	2	3	4	1
3	4	1	2	3	4	1	2	4	3	2	1	2	1	4	3	3	4	1	2
4	3	2	1	4	3	2	1	2	1	4	3	4	3	2	1	4	1	2	3

(1) latin sq., (2) replication, (3) permute 3 rows, (4) permute 2 rows, (5) different.

1−	−2			1−	−3			1−	−4			1−	−5		
11	22	33	44	11	22	33	44	11	22	33	44	11	22	33	44
22	11	44	33	23	14	41	32	23	14	41	32	22	13	44	31
33	44	11	22	34	43	12	21	32	41	14	23	33	44	11	22
44	33	22	11	42	31	24	13	44	33	22	11	44	31	22	13

Conclusion ?

1. Permute 3 rows of Latin square (1) works;

2. Permute 2 rows of Latin square or superimpose (5) does not work !

How to tell ?

No pair of numbers occurs twice.

Hyper-Graeco-Latin Squares deal with the case that there are 4 block factors with levels equal the level of the treatment factor (4+1).

A Hyper-Graeco-Latin Square used in a Martindale wear tester.
The martindale wear tester is a machine used for testing the wearing quality of types of cloth or other such materials.

* 4 pieces of cloth may be compared simultaneously in one machine cycle.
* The response is the weight loss in tenths of a milligram suffered by the test piece when it is rubbed again a standard grade of emory paper for 1000 revolutions of the machine.
* Specimens of the four different types of cloth (treatments) A, B, C, D whose wearing qualities are to be compared are mounted in 4 different specimen holders 1, 2, 3, 4.
* Each holder can be in any of the 4 positions P_1, P_2, P_3, P_4 on the machine.
* Each emory paper sheet α, β, γ, δ was cut into 4 quarters and each quarter used to complete a single cycle c_1, c_2, c_3 and c_4 of 1000 revolutions.

The object of the experiment:
(1) to make a more accurate comparison of the treatments
(2) to discover how much a total variability was contributed by the various factors:
holders, positions, emory paper and cycles.
One replication has 16 df.
Under control-sum, $1 + (4 + 1) \times (4 - 1) = 16$ dfs are needed, thus
two replications are needed **why ??**
Thus 4 additional cycles and 4 additional emory papers are needed.

So there are 32 experiments. It is important to consider randomizing the 32 experiments. In the first 16 runs, each run involves 5 conditions: (4+1) factors, each with 4 levels.
How to order them for randomization ?

In each circle, 4 experiments are carried out simultaneously, it needs 4 types of emory papers and 4 types of cloth. Each holder, position and circle are one unit, respectively. Each cloth and emory paper are cut to 4 pieces. If the qualify of cloth and emory papers are uniform, then no need to randomize (the textbook does not bother). Otherwise, in each replication of 16 experiments, we can randomize as follows.

for (i in 1:4) sample (1:4) (for 4 pieces of each emory paper in 4 circles),

for (i in 1:4) sample (1:4) (for 4 pieces of each type of cloth in 4 circles).
The data are as follows.

$cycles\backslash position$	P_1	P_2	P_3	P_4
c_1	$\alpha A1$	$\beta B2$	$\gamma C3$	$\delta D4$
	320	297	299	313
c_2	$\beta C4$	$\alpha D3$	$\delta A2$	$\gamma B1$
	266	227	260	240
c_3	$\gamma D2$	$\delta C1$	$\alpha B4$	$\beta A3$
	221	240	267	252
c_4	$\delta B3$	$\gamma A4$	$\beta D1$	$\alpha C2$
	301	238	243	290
c_5	$\epsilon A1$	$\xi B2$	$\theta C3$	$\kappa D4$
	285	280	331	311
c_6	$\xi C4$	$\epsilon D3$	$\kappa A2$	$\theta B1$
	268	233	291	280
c_7	$\theta D2$	$\kappa C1$	$\epsilon B4$	$\xi A3$
	265	273	234	243
c_8	$\kappa B3$	$\theta A4$	$\xi D1$	$\epsilon C2$
	306	271	270	272

replication I
Cycles: c_1, c_2, c_3, c_4
Treatments: A, B, C, D
Holders: 1, 2, 3, 4
Emory paper sheet: $\alpha,\ \beta,\ \gamma,\ \delta$

replication II
Cycles: c_5, c_6, c_7, c_8
Treatments: A, B, C, D
Holders: 1, 2, 3, 4
Emory paper sheet: $\epsilon,\ \xi,\ \theta,\ \kappa$

What is the property of the arrangement ?

Three Latin squares superimpose together twice.

$$\begin{array}{cccc} \alpha A1 & & & \\ 111 & 222 & 333 & 444 \\ 234 & 143 & 412 & 321 \\ 342 & 431 & 124 & 213 \\ 423 & 314 & 241 & 132 \end{array} \qquad \begin{array}{c} pattern \\ 1 \\ 234 \\ 34 \\ 4 \end{array} \text{, but not } \begin{array}{cccc} \alpha A1 & & & \\ 111 & 222 & 333 & 444 \\ 222 & 111 & 444 & 333 \\ 333 & 444 & 111 & 222 \\ 444 & 333 & 222 & 111 \end{array}$$

Notice:

1. (α, A), (α, 1), (A, 1), *etc.* will not occur twice.
2. Rows 2, 3, 4 belongs to $\{(2,1,4,3), (3,4,1,2), (4,3,2,1)\}$ in the order $\begin{array}{c} 111 \\ 234 \\ 34c \\ 4ab \end{array}$
3. The element (a,b,c) in the table is uniquely determined.

It is easier to set the Hyper-Graeco-Latin Square this way:

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & \\ 3 & 4 & 1 & \\ 4 & & & 1 \end{array} \rightarrow \text{LS=} \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \\ 3 & 4 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{array} \rightarrow \begin{array}{c} 1 \\ 234 \\ 34 \\ 4 \end{array} \rightarrow \begin{array}{c} 111 \\ 234 \\ 34 \\ 4\textbf{?} \end{array} \rightarrow \begin{array}{c} 111 \\ 234 \\ 34? \\ 42 \end{array} \rightarrow \begin{array}{c} 111 \\ 234 \\ 342 \\ 42? \end{array} \rightarrow \begin{array}{c} 111 \\ 234 \\ 342 \\ 423 \end{array}$$

What does it mean ?
$$\begin{array}{lcl} 111 & = & (1st, 1st, 1st)\ row\ of\ LSquare \\ 234 & = & (2nd, 3rd, 4th)\ row\ of\ LS \\ 342 & = & (3rd, 4th, 2nd)\ row\ of\ LS \\ 423 & = & (4th, 2nd, 3rd)\ row\ of\ LS \end{array}$$

$$\rightarrow \begin{array}{cccc} 111 & 2 & 3 & 4 \\ 234 & 1 & 4 & 3 \\ 342 & 4 & 1 & 2 \\ 423 & 3 & 2 & 1 \end{array} \rightarrow \begin{array}{cccc} \alpha A1 & & & \\ 111 & 222 & 333 & 444 \\ 234 & 143 & 412 & 321 \\ 342 & 431 & 124 & 213 \\ 423 & 314 & 241 & 132 \end{array}$$

This may not work for other dimension, say 5.

Consider model

$Y \sim replication_1 + cycle_6 + position_3 + Emory_6 + holder_3 + treatment_3$, or
$Y = X\beta + \epsilon$, where Y is a 32×1 vector, β is a vector in $\mathcal{R}^{23}$ $(1 + 1 + 6 + 3 + 6 + 3 + 3 = 23)$, and X is a matrix of dimension 32×23.

```
> y=c(320, 297, 299, 313, 266, 227, 260, 240, 221, 240, 267, 252, 301, 238, 243,
290)
> z=c(y, 285, 280, 331, 311, 268, 233, 291, 280, 265, 273, 234, 243, 306, 271, 270,
272)
> options(contrasts =c("contr.sum", "contr.poly"))
> (P=gl(4,1,32))
```

```
  [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 1 2 3 4 1 2 3 4
  Levels: 1 2 3 4
> r=gl(2,16,32) # replication index
> T=c(1,2,3,4,3,4,1,2,4,3,2,1,2,1,4,3)
> T=factor(c(T,T))
> H=c(1,2,3,4,4,3,2,1,2,1,4,3,3,4,1,2) # holder
> H=factor(c(H,H))
> (C1=c(rep(1,4),rep(0,8),rep(-1,4)))
  [1] 1 1 1 1 0 0 0 0 0 0 0 0 -1 -1 -1 -1 # why −1 ?
> (C2=c(rep(0,4),rep(1,4),rep(0,4),rep(-1,4)))
  [1] 0 0 0 0 1 1 1 1 0 0 0 0 -1 -1 -1 -1
> (C3=c(rep(0,8),rep(1,4),rep(-1,4)))
  [1] 0 0 0 0 0 0 0 0 1 1 1 1 -1 -1 -1 -1
> C5=c(rep(0,16),C1)
> C6=c(rep(0,16),C2)
> C7=c(rep(0,16),C3)
> C1=c(C1,rep(0,16)) # C1 is a factor or numerical variable ?
> C2=c(C2,rep(0,16))
> C3=c(C3,rep(0,16))
> E1=c(1,0,0,-1,0,1,-1,0,0,-1,1,0,-1,0,0,1) # emory
# α, β, γ, δ, β, α, δ, γ, γ, δ, α, β, δ, γ, β, α
> E2=c(0,1,0,-1,1,0,-1,0,0,-1,0,1,-1,0,1,0)
> E3=c(0,0,1,-1,0,0,-1,1,1,-1,0,0,-1,1,0,0)
> E5=c(rep(0,16),E1)
> E6=c(rep(0,16),E2)
> E7=c(rep(0,16),E3)
> E1=c(E1,rep(0,16))
> E2=c(E2,rep(0,16))
> E3=c(E3,rep(0,16))
> obj=lm(z ∼ T+H+P+C1+C2+C3 +C5+C6+C7 +E1+E2+E3+E5+E6+E7+r))

> (ob=lm(z ∼ T))
```

(Intercept)	T1	T2	T3
271.469	−1.469	4.156	8.406

```
> obj
```

$(Intercept)$	$T1$	$T2$	$T3$	$H1$	$H2$
271.4688	−1.4688	4.1563	8.4063	−2.5938	0.5313
$H3$	$P1$	$P2$	$P3$	$C1$	$C2$
2.5313	7.5312	−14.0938	2.9063	40.1250	−18.8750
$C3$	$C5$	$C6$	$C7$	$E1$	$E2$
−22.1250	25.9375	−7.8125	−22.0625	8.8750	−2.6250
$E3$	$E5$	$E6$	$E7$	$r1$	
−17.6250	−19.8125	−10.5625	10.9375	−4.3438	

(1)

Remark. LSE of treatment effects of two models are the same.

```
> C=c(C1,C2,C3,C5,C6,C7)
> dim(C)=c(32,6)
> E=c(E1,E2,E3,E5,E6,E7)
> dim(E)=c(32,6)
```

>lm(z~T+H+P+C+E+r) difference between Eq.(1) and Eq.(2) ?

$(Intercept)$	$T1$	$T2$	$T3$	$H1$	$H2$
271.4688	−1.4688	4.1563	8.4063	−2.5938	0.5313
$H3$	$P1$	$P2$	$P3$	$C1$	$C2$
2.5313	7.5312	−14.0937	2.9063	40.1250	−18.8750
$C3$	$C4$	$C5$	$C6$	$E1$	$E2$
−22.1250	25.9375	−7.8125	−22.0625	8.8750	−2.6250
$E3$	$E4$	$E5$	$E6$	$r1$	
−17.6250	−19.8125	−10.5625	10.9375	−4.3438	

(2)

Main concern: H_o: $\tau_A = \tau_B = \tau_C = \tau_D$ v.s. H_1: H_o fails.

> anova(lm(z~T))

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$
T	3	1705.3	568.45	0.6429	0.5939
$Residuals$	28	24758.6	884.24		

Conclusion ?

> anova(lm(z~T+H+P+C+E+r))

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$
T	3	1705.3	568.45	5.3908	0.021245
H	3	109.1	36.36	0.3449	0.793790
P	3	2217.3	739.11	7.0093	0.009925
C	6	14770.4	2461.74	23.3455	$5.273e-05$
E	6	6108.9	1018.16	9.6555	0.001698
r	1	603.8	603.78	5.7259	0.040366
$Residuals$	9	949.0	105.45		

Conclusion ?

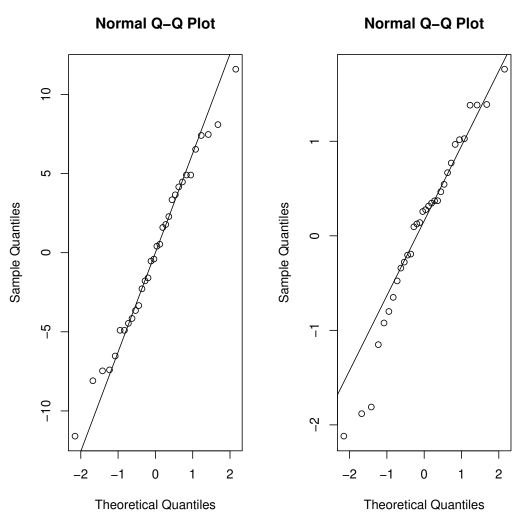
> anova(lm(z~T+P+C+E+r))

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$	
T	3	1705.3	568.45	6.4467	0.0075703	$**$
P	3	2217.3	739.11	8.3822	0.0028332	$**$
C	6	14770.4	2461.74	27.9181	$2.221e-06$	$***$
E	6	6108.9	1018.16	11.5467	0.0002213	$***$
r	1	603.8	603.78	6.8474	0.0225196	$*$
$Residuals$	12	1058.1	88.18			

> anova(lm(y~T+r))

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$
T	3	1705.3	568.45	0.6354	0.5987
r	1	603.8	603.78	0.6749	0.4185
$Residuals$	27	24154.8	894.62		

Why different ?



qqnorm(obj$resid)

```
qqline(obj$resid)
z=rnorm(32)
qqnorm(z)
qqline(z)
```

Summary:

1. H_o^r: No difference in replication. ??
2. H_o^c: No difference in cycles. ??
3. H_o^H: No difference in specimen holder. P-value = 0.8 > 0.05.
4. H_o^P: No difference in positions. ??
5. H_o^e: No difference in emory papers. ??
6. H_o: $\tau_A = \tau_B = \tau_C = \tau_D$ v.s. H_1: H_o fails.
   Is the p-value for T 0.594, or 0.021 or 0.008, or 0.599 ?
   The difference is very significant.
   Reject H_o, and the treatment effect are not equal.
   Notice that without blocking factor P, C and E, the conclusion is different, even with replications.
   p-value for T is 0.59 > $\alpha = 0.05$.
7. Preference of treatments (weight loss) $D > A > B > C$. $\begin{matrix} (Int) & T1 & T2 & T3 \\ 271 & -1 & 4 & 8 \end{matrix}$

There are several models:

(1) lm(y∼T)
(2) lm(y∼T+r)
(3) lm(y∼T+H+P+C+E+r)
(4) lm(y∼T+P+C+E+r)

Which of them is appropriate ?
What is the connection between the previous question and goodness-of-fit test ?

H_o: $E(Y|\mathbf{X}) = \beta'\mathbf{X}$ v.s. H_1: $E(Y|\mathbf{X}) = \beta'\mathbf{X} + \theta g(\mathbf{X})$.
Model (1) is a special case of Models (2), (3) and (4).
Does anova suggests that it can be simplified ?

Which of them is better ?

> anova(obj,ob)
Model 1: z ∼ T + P + C+ E + r
Model 2: z ∼ T

	Res.Df	*RSS*	*Df*	*Sum of Sq*	*F*	*Pr(> F)*
1	12	1058.1				
2	28	24758.6	−16	−23700	16.799	8.058e − 06

> summary(obj)

	Estimate	*Std.Error*	*tvalue*	*Pr(> |t|)*	
(*Intercept*)	271.4688	1.8153	149.546	< 2e − 16	∗ ∗ ∗
*T*1	−1.4688	3.1442	−0.467	0.651505	
*T*2	4.1563	3.1442	1.322	0.218814	
*T*3	8.4063	3.1442	2.674	0.025471	∗
*H*1	−2.5938	3.1442	−0.825	0.430726	
*H*2	0.5313	3.1442	0.169	0.869561	
*H*3	2.5313	3.1442	0.805	0.441531	
*P*1	7.5312	3.1442	2.395	0.040206	∗
*P*2	−14.0937	3.1442	−4.483	0.001527	∗∗
*P*3	2.9063	3.1442	0.924	0.379429	
*C*1	40.1250	4.4465	9.024	8.35e − 06	∗ ∗ ∗
⋮					

Q: Under model $lm(z \sim T)$ under control sum, if we write in the standard LR model form $Y_i = \beta X_i + \epsilon_i$, (Y_i, X_i, β)= ?

$Y = 271.5 - 1.5T1 + 4.2T2 + 8.4T3$?

β =(T1,T2,T3) ?

Or try

lm(formula = z ~ T - 1)

$T1$	$T2$	$T3$	$T4$
270.0	275.6	279.9	260.4

$Y_i = z_i$,
$X_i' = (1, \mathbf{1}(T = 1), \mathbf{1}(T = 2), \mathbf{1}(T = 3), \mathbf{1}(T = 4))$, or more accurately,
$X_i' = (1, \mathbf{1}(T \textit{ is cloth } A), \mathbf{1}(T \textit{ is cloth } B), \mathbf{1}(T \textit{ is cloth } C), \mathbf{1}(T \textit{ is cloth } D))$,
$\beta' = (\beta_0, \beta_1, \beta_2, \beta_3, -\sum_{i=1}^3 \beta_i)$.
$\hat\beta' = (271.5, -1.5, 4.2, 8.4, -11.1)$.
Interpretation:

The mean wearing effect on the 4 cloths is 271.5 units,

effect on cloth A is 1.5 units lower,

effect on cloth B is 4.2 units higher,

effect on cloth C is 8.4 units higher,

effect on cloth D is 11.1 units lower.

Homework 4.1. 1. Suppose that each emory paper α, β, γ, δ can be cut into 8 pieces rather than 4 quaters and each piece is used to complete a single cicle c_1, ..., c_8 of 1000 revolutions. That is $(\epsilon, \xi, \theta, \kappa)$ are replaced by (α, β, γ, δ). Pretend the data remain the same. Revise the codes and do data analysis again.

4.5. Balanced incomplete block designs. The Martindale wear tester example is a complete block design. There are 4 treatment, and block size (Emory paper) is also 4.
If # of treatments > block size, then we have incomplete block designs, *e.g.*, if there are 4 treatment, and block size (Emory paper) is 3, then it is an incomplete block design.

A **balanced** incomplete block design:

circle of 10^3 *revolutions*		A	B	C	D
	1	α	β	γ	
	2	β	γ		α
	3	γ		α	β
	4		α	β	γ

Its properties:
1. Within block of cycles, every pair of treatments appears twice. *e.g.* (A,B) occurs at blocks (circles) 1 and 2, and (A,D) occurs at blocks 2 and 3.
2. Every row contains each of α, β and γ.
3. Every column contains each of α, β and γ.
Thus each of α, β and γ block contains $\{A, B, C, D\}$ and circle $\{1, 2, 3, 4\}$.

Youden Squares: A second wear testing example. There are 7 treatment, and block size of emory paper is still 4, a balanced incomplete block design is as follows.

cycles\treatment	A	B	C	D	E	F	G	
1		α627		β248		γ563	δ252	DG
2	α344		β233			δ442	γ226	
3			α251	γ211	δ160		β297	DG
4	β337	δ537			γ195		α300	AB
5		γ520	δ278		β199	α595		
6	γ369			δ196	α185	β606		
7	δ396	β602	γ240	α273				AB

Within block of cycles, every pair of treatments appears twice.
e.g. In the block of cycles (A,B) occurs at blocks 4 and 7.
Each row and column contains α, β, γ, δ.

Does

$cycles\backslash treatment$	A	B	C	D	E	F	G
1	α	β	γ	δ			
2	β	γ	δ				α
3	γ	δ				α	β
4	δ				α	β	γ
5				α	β	γ	δ
6			α	β	γ	δ	
7		α	β	γ	δ		

work ?

> y=c(627,248,563,252, 344,233,442,226, 251,211,160,297, 337,537,195,300,
520,278,199,595, 369,196,185,606, 396,602,240,273)

> T=c(”B”,”D”,”F”,”G”, ”A”,”C”,”F”,”G”, ”C”,”D”,”E”,”G”, ”A”,”B”,”E”,”G”,
”B”,”C”,”E”,”F”, ”A”,”D”,”E”,”F”, ”A”,”B”,”C”,”D”)

> e=c(”a”,”b”,”r”,”d”, ”a”,”b”,”d”,”r”, ”a”,”r”,”d”,”b”, ”b”,”d”,”r”,”a”,
”r”,”d”,”b”,”a”, ”r”,”d”,”a”,”b”, ”d”,”b”,”r”,”a”)

> c= gl(7,4,28)

[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 7 7 7 7

> (z=lm(y∼T)) # **What is the LSE of TA ?**

What is the LSE of wearing effect on cloth A ?

$(Intercept)$	TB	TC	TD	TE	TF	TG
361.50	210.00	-111.00	-129.50	-176.75	190.00	-92.75

> (x=lm(y∼T+e+c))

$(Intercept)$	TB	TC	TD	TE	TF
408.429	191.357	-111.571	-147.643	-184.500	188.429
TG	eb	ed	er	$c2$	$c3$
-87.571	-7.571	-44.857	-35.857	-72.429	-23.786
$c4$	$c5$	$c6$	$c7$		
-23.929	-9.286	-11.429	8.357		

> anova(x)

	Df	$Sum\ q$	$Mean\ Sq$	$Fvalue$	$Pr(>F)$	
T	6	589623	98271	96.4619	$1.899e-09$	$***$
e	3	9846	3282	3.2217	0.06125	.
c	6	14570	2428	2.3837	0.09445	.
$Residuals$	12	12225	1019			
			Can	we	$simplify$	?
				$Delete$	$e\ or\ c$	?
$e+c$	9	24416	2712.9	2.6623	0.0583	
					$pf(2.67,9,12)$	

> anova(x,z)

Model 1: y ∼ T + e + c

Model 2: y ∼ T

	$Res.Df$	RSS	Df	$Sum\ of\ Sq$	F	$Pr(>F)$	
1	12	12225					
2	21	36641	-9	-24416	2.663	0.05828	.

> summary(z)

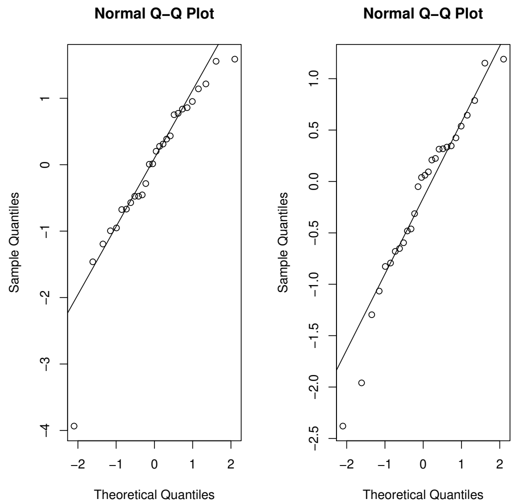
	$Estimate$	$Std.\ Error$	$t\ value$	$Pr(>\|t\|)$	
$(Intercept)$	361.50	20.89	17.309	$6.61e-14$	$***$
TB	210.00	29.54	7.110	$5.17e-07$	$***$
TC	-111.00	29.54	-3.758	0.001157	$**$
TD	-129.50	29.54	-4.384	0.000259	$***$
TE	-176.75	29.54	-5.984	$6.13e-06$	$***$
TF	190.00	29.54	6.433	$2.24e-06$	$***$
TG	-92.75	29.54	-3.140	0.004943	$**$

> qqnorm(studres(z))

> qqline(studres(z))

> u=rnorm(28)

> qqnorm(u)



> qqline(u)

Summary:

H_o: $\tau_A = \tau_B = \tau_C = \tau_D = \tau_E = \tau_F = \tau_G$ v.s. H_1: H_o fails.

p-value for T is $< 0.001 < \alpha = 0.05$.

The treatments are significantly different.

Reject H_o, and the treatment effect are not equal.

Which of the two model is appropriate ?

(1) $E(Y|\mathbf{X}) = \alpha + \beta_1' T$,

(2) $E(Y|\mathbf{X}) = \alpha + \beta_1' T + \beta_2' e + \beta_3' c$

Preference in treatments: $E > D > C > G > A > F > B$. **Why ?**

Interpretation of α under model (1) ?

The average effect of the 7 treatments ?

The average effect of Treatment A ?

Interpretation of β_i ?

Interpretation of α under model (2) ?

The average effect of the 7 treatments ?

The average effect of Treatment A ?

Interpretation of β_i ?

> names(summary(z))

[1] "call" "terms" "residuals" "coefficients"

[5] "aliased" "sigma" "df" "r.squared"

[9] "adj.r.squared" "fstatistic" "cov.unscaled"

> summary(lm(y∼T-1))$cov

	TA	TB	TC	TD	TE	TF	TG
TA	0.25	0.00	0.00	0.00	0.00	0.00	0.00
TB	0.00	0.25	0.00	0.00	0.00	0.00	0.00
TC	0.00	0.00	0.25	0.00	0.00	0.00	0.00
TD	0.00	0.00	0.00	0.25	0.00	0.00	0.00
TE	0.00	0.00	0.00	0.00	0.25	0.00	0.00
TF	0.00	0.00	0.00	0.00	0.00	0.25	0.00
TG	0.00	0.00	0.00	0.00	0.00	0.00	0.25

Residual standard error: 41.77 on 21 degrees of freedom

> (U=summary(lm(y∼T))$cov)

	$(Intercept)$	TB	TC	TD	TE	TF	TG
$(Intercept)$	0.25	−0.25	−0.25	−0.25	−0.25	−0.25	−0.25
TB	−0.25	0.50	0.25	0.25	0.25	0.25	0.25
TC	−0.25	0.25	0.50	0.25	0.25	0.25	0.25
TD	−0.25	0.25	0.25	0.50	0.25	0.25	0.25
TE	−0.25	0.25	0.25	0.25	0.50	0.25	0.25
TF	−0.25	0.25	0.25	0.25	0.25	0.50	0.25
TG	−0.25	0.25	0.25	0.25	0.25	0.25	0.50

Why is there such a big difference ?

Under the model $y \sim T - 1$,

$\hat\beta_A = \frac{\sum_{i=1}^n y_i \mathbf{1}(T_i = A)}{\sum_{i=1}^n \mathbf{1}(T_i = A)}$, where $n = ?$ 7 treatments and 4 blocks.

$\hat\beta_B = \frac{\sum_{i=1}^n y_i \mathbf{1}(T_i = B)}{\sum_{i=1}^n \mathbf{1}(T_i = B)}$, ...

$cov(\hat\beta_A, \hat\beta_B) = E(\hat\beta_A \cdot \hat\beta_B) - E(\hat\beta_A)E(\hat\beta_B)$.

Under the model $y \sim T$,

$\hat\beta_0 = \frac{\sum_{i=1}^n y_i \mathbf{1}(T_i = A)}{\sum_{i=1}^n \mathbf{1}(T_i = A)}$, $\hat\beta_A = 0$, $\hat\beta_B = \frac{\sum_{i=1}^n y_i \mathbf{1}(T_i = B)}{\sum_{i=1}^n \mathbf{1}(T_i = B)} - \hat\beta_0$, ...

> summary(lm(y~T-1))

	Estimate	*Std. Error*	*t value*	$Pr(>\lvert t\rvert)$
TA	361.50	20.89	17.309	$6.61e-14***$
TB	571.50	20.89	27.363	$< 2e-16***$
TC	250.50	20.89	11.994	$7.35e-11***$
TD	232.00	20.89	11.108	$2.98e-10***$
TE	184.75	20.89	8.846	$1.59e-08***$
TF	551.50	20.89	26.406	$< 2e-16***$
TG	268.75	20.89	12.868	$1.99e-11***$

Is U really a covariance matrix ?

$\hat\Sigma_{\hat\beta} = \hat\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

cov.unscaled=$(\mathbf{X}'\mathbf{X})^{-1}$.

> 0.25*41.77**2

[1] 436.1832

> 20.89**2

[1] 436.3921

Chapter 5. Factorial Designs at two levels

We shall look at 3 examples. Two are qualitative and one is quantitative.

5.2. Example 1: The effect of 3 factors on clarity of film.

An experiment to determine how the cloudiness of a floor wax is affected when certain changes are introduced into the formula for its preparation.

1 response: cloudiness of a floor.

3 factors each with two levels:

amount of emulsifier A (low, high) or (-,+),

amount of emulsifier B (low, high) or (-,+),

catalyst concentration C (low, high) or (-,+).

There are $2^3 = 8$ combinations and one needs 8 (random) runs of experiments. They are called 2^3 factorial designs.

run#	A	B	C	*results*(N/Y)	*or*$(-/+)$
1	−	−	−	*No*	−
2	+	−	−	*No*	−
3	−	+	−	*Yes*	+
4	+	+	−	*Yes*	+
5	−	−	+	*No*	−
6	+	−	+	*No*	−
7	−	+	+	*Yes*	+
8	+	+	+	*Yes*	+
compare					*same as B*

Results can also be given as **visual display** in Figure 5.1 (in the textbook). One can see from Figure 5.1 that cloudy is mainly due to high amount of emulsifier B. Factors A nd C are called **inert**.

Is the run number the order of experiments ? Then ?

5.3. The effects of 3 factors on 3 physical properties of a polymer solution.
In the previous example, there is just one response.
There are 3 responses in the current experiment
3 responses: Is the polymer solution
milky ? (y_1),
viscous ? (y_2),
yellow color ? (y_3).
3 factors each with two levels in the formulation of the solution:
amount of a reactive monomer (10,30)% or $(-,+)$,
the type of chain length regulator (A,B) or $(-,+)$,
amount of chain length regulator (1,3)% or $(-,+)$.

run#	1	2	3	*milky*?	*viscous*?	*yellow*?
1	$-$	$-$	$-$	$Y\ -$	$Y\ -$	$N\ -$
2	$+$	$-$	$-$	$N\ +$	$Y\ -$	$N\ -$
3	$-$	$+$	$-$	$Y\ -$	$Y\ -$	$N\ -$
4	$+$	$+$	$-$	$N\ +$	$Y\ -$	*Slightly* $++$
5	$-$	$-$	$+$	$Y\ -$	$N\ +$	$N\ -$
6	$+$	$-$	$+$	$N\ +$	$N\ +$	$N\ -$
7	$-$	$+$	$+$	$Y\ -$	$N\ +$	$N\ -$
8	$+$	$+$	$+$	$N\ +$	$N\ +$	*Slightly* $++$
compare to columns				1	3	Y if 1&2 both ++ *ow N*

See Figures 5.2 and 5.3 for visual display of the results.
Pay attention to the row of "compare columns" to the figures.

Notice that the response is qualitative in the previous two examples.
The factorial design can tell which factor do what to which response.

5.4. A pilot investigation.
1 response: yields of the experiment (numerical).
3 factors: temperature T (160, 180) or $(-,+)$,
concentration C (20,40) or $(-,+)$,
type of catalyst K (A,B) or $(-,+)$.
There are duplicate runs (8+8=16).

run#	T	C	K	*average yields of* 2 *runs*	$y_{i1}^{(order)}$	$y_{i2}^{(order)}$
1	$-$	$-$	$-$	60	$59^{(6)}$	$61^{(13)}$
2	$+$	$-$	$-$	72	$74^{(2)}$	$70^{(4)}$
3	$-$	$+$	$-$	54	$50^{(1)}$	$58^{(16)}$
4	$+$	$+$	$-$	68	$69^{(5)}$	$67^{(10)}$
5	$-$	$-$	$+$	52	$50^{(8)}$	$54^{(12)}$
6	$+$	$-$	$+$	83	$81^{(9)}$	$85^{(14)}$
7	$-$	$+$	$+$	45	$46^{(3)}$	$44^{(11)}$
8	$+$	$+$	$+$	80	$79^{(7)}$	$81^{(15)}$

Table 5.3

Remark. Using average is only for the convenience of computing the main effects, not for anova.

5.5. Calculation of main effect.
Definition: Main effect of each factor $= \overline{y}_+ - \overline{y}_-$ (see the next tables).
Interpretation: the average difference between level 2 (+) of a factor and level 1 (-) (same as control.treatment)

Main effect of T: **Main effect of C:**

$run\#$	T	C	K	y_+		y_-	$yields$
1	$-$	$-$	$-$			60	
2	$+$	$-$	$-$	72			
3	$-$	$+$	$-$			54	
4	$+$	$+$	$-$	68			
5	$-$	$-$	$+$			52	
6	$+$	$-$	$+$	83			
7	$-$	$+$	$+$			45	
8	$+$	$+$	$+$	80			
				$\overline{y}_+$	$-$	$\overline{y}_-$	$=23$

$run\#$	T	C	K	y_+		y_-	$yields$
1		$-$	$-$			60	
2		$-$	$-$			72	
3		$+$	$-$	54			
4		$+$	$-$	68			
5		$-$	$+$			52	
6		$-$	$+$			83	
7		$+$	$+$	45			
8		$+$	$+$	80			
				$\overline{y}_+$	$-$	$\overline{y}_-$	$=-5$

Main effect of K:

$run\#$	T	C	K	y_+		y_-	$yields$
1			$-$			60	
2			$-$			72	
3			$-$			54	
4			$-$			68	
5			$+$	52			
6			$+$	83			
7			$+$	45			
8			$+$	80			
				$\overline{y}_+$	$-$	$\overline{y}_-$	$=1.5$

Four ways to compute with R-code:

```
> y=c(60,72,54,68,52,83,45,80)
> (a=rep(c(-1,1),4))
    [1] -1 1 -1 1 -1 1 -1 1
> (b=rep(c(-1,-1,1,1),2))
    [1] -1 -1 1 1 -1 -1 1 1
> c=rep(-1,4)
> (c=c(c,-c))
    [1] -1 -1 -1 -1 1 1 1 1
```

\# **First way to compute effects**

```
> (v=c(y% *% a/4, y% *% b/4, y% *% c/4))
    [1] 23.0 −5.0 1.5 # main effects
```

\# **2nd way to compute effects**

```
> W=lm(y∼a+b+c)
> W$coef[1:4]
```

$(Intercept)$	a	b	c
64.25	11.50	-2.50	0.75

\# model 1: $y = \mu + \beta_1 a + \beta_2 b + \beta_3 c + \epsilon$.

```
> c( 2*W$coef[2:4])
```

a	b	c
23.00	-5.00	1.50

\# main effects

\# **3rd way and the prefer way**

```
> lm(y∼factor(a)+factor(b)+factor(c) )$coef[1:4]
```

$(Intercept)$	$factor(a)1$	$factor(b)1$	$factor(c)1$
54.5	23.0	-5.0	1.5

\# main effects

\# **factor(a)1 refers to 1(a=1)**

\#model 2: $y = \mu + \beta_1 \mathbf{1}(T=+) + \beta_2 \mathbf{1}(C=+) + \beta_3 \mathbf{1}(K=+) + \epsilon$.

```
> mean(y)
    [1] 64.25
```

The fourth way:

```
> options(contrasts =c("contr.sum", "contr.poly"))
>U= lm(y∼factor(a)+factor(b)+factor(c))$coef[1:4]
```

$(Intercept)$	$factor(a)1$	$factor(b)1$	$factor(c)1$
64.25	-11.50	2.50	-0.75

factor(a)1 refers to 1(a=-1)

\#model 3: $y = \mu + \beta_1(\mathbf{1}(T=-) - \mathbf{1}(T=+)) + \beta_2(\mathbf{1}(C=-) - \mathbf{1}(C=+))$
$+ \beta_3(\mathbf{1}(K=-) - \mathbf{1}(K=+)) + \epsilon$. (somewhat opposite to model 1).

```
> −2*U[2:4] # main effects
```

Remark. $\hat{\mathbf{Y}}$ remains unchanged in the last three ways.
Homework problem 5.5: Given the LSE by the fourth way, how to get the LSE under model 2 (the 3rd way) ?

5.6. Interaction.

Two-factor interaction for TC,

$run\#$	T	C	K	y_+		y_-	$yields$
1	−	−		60			
2	+	−				72	
3	−	+				54	
4	+	+		68			
5	−	−		52			
6	+	−				83	
7	−	+				45	
8	+	+		80			
				$\overline{y}_+$	−	$\overline{y}_-$	= 1.5

Two-factor interaction for TK,

$run\#$	T	C	K	y_+		y_-	$yields$
1	−		−	60			
2	+		−			72	
3	−		−	54			
4	+		−			68	
5	−		+			52	
6	+		+	83			
7	−		+			45	
8	+		+	80			
				$\overline{y}_+$	−	$\overline{y}_-$	= 10

Two-factor interaction for CK,

$run\#$	T	C	K	y_+		y_-	$yields$
1		−	−	60			
2		−	−	72			
3		+	−			54	
4		+	−			68	
5		−	+			52	
6		−	+			83	
7		+	+	45			
8		+	+	80			
				$\overline{y}_+$	−	$\overline{y}_-$	= 1.5

Three-factor interaction

$run\#$	T	C	K	y_+		y_-	$yields$
1	−	−	−			60	
2	+	−	−	72			
3	−	+	−	54			
4	+	+	−			68	
5	−	−	+	52			
6	+	−	+			83	
7	−	+	+			45	
8	+	+	+	80			
				$\overline{y}_+$	−	$\overline{y}_-$	= 0

R commands:

```
ab=a*b
ac=a*c
bc=b*c
abc=ab*c
a=factor(a)
b=factor(b)
c=factor(c)
ab=factor(ab) # why not ab=a*b ?
ac=factor(ac)
bc=factor(bc)
abc=factor(abc)
lm(y~a+b+c+ab+ac+bc+abc)
```

5.7. Estimation of variance of replicate runs.
(1) Under the i.i.d. N(μ, σ^2) assumption,

$\hat{\sigma}^2 = \frac{1}{df}\sum_{i=1}^n (Y_i - \hat{\beta}X_i)^2$ if df> 0, where $\beta X \overset{def}{=} \beta' X$.

$\hat{\sigma}^2$ is the unbiased estimator using mean squared residuals, under the null hypothesis H_o: $E(Y|\mathbf{X}) = \beta\mathbf{X}$.
If there is no replicate runs ($r = 1$ under the full model), then

$\sum_{i=1}^n (Y_i - \hat{\beta}X_i)^2 = 0$, as there are 8 parameters and 8 observations y_{ij}'s.

Thus it is not a proper estimator in such case.
(2) If there are r replicate runs in a 2^3 factorial design, with responses

y_{ij}, $i = 1, ..., 8$ and $j = 1, ..., r$, let
$s_i^2 = \frac{1}{r-1}\sum_{j=1}^r (y_{ij} - \overline{y}_{i\cdot})^2$, $i = 1, ..., 8$.

If $r = 2$,

$s_i^2 = \frac{(y_{i1}-\overline{y}_{i\cdot})^2 + (y_{i2}-\overline{y}_{i\cdot})^2}{2-1} = \frac{(y_{i1}-y_{i2})^2}{2}$, $i = 1, ..., 8$,

where $\overline{y}_{i\cdot} = \frac{y_{i1}+y_{i2}}{2}$.

$s^2 = \sum_{i=1}^8 s_i^2/8$ is an (unbiased) estimator of σ^2.

$$s^2 = \hat{\sigma}^2?$$

Yes, if under the full model $y \sim a + b + c + ab + bc + ac + abc$. $(p = 8)$
No, if under the submodel, *e.g.* $y \sim I(a * b * c)$ $(p = 2)$.
(3) Is the Mean Sq in each row of anova(), unbiased estimator of σ^2 ?
How about (Residual standard error)^2 in summary(lm()) ?
How about Residual Mean Sq in anova() ?

Simulation example 5.7.1

> a=rep(c(-1,1),4)
> b=rep(c(-1,-1,1,1),2)
> c=rep(-1,4)
> c=c(c,-c)
> a=c(a,a)
> b=c(b,b)
> c=c(c,c)
> ab=a*b
> ac=a*c
> bc=b*c
> e=rnorm(16)
> y=a+2*b−3*c+16*ab+bc+e
> (z=lm(y∼a+b+c+ab+bc)) $\begin{pmatrix} (Intercept) & a & b & c & ab & bc \\ -0.12 & 0.98 & 2.34 & -3.36 & 15.89 & 0.97 \end{pmatrix}$

Let $Y = \beta'\mathbf{X} + \epsilon$, where $\beta \in \mathcal{R}^p$. $p = ?$ $\beta = ?$ $\hat{\beta} = ?$

> anova(z)

	Df	*Sum Sq*	*Mean Sq*	*F value*	*Pr(> F)*	
a	1	14.7	14.7	15.914	0.002562	**
b	1	47.1	47.1	51.107	3.109e − 05	***
c	1	159.6	159.6	173.258	1.219e − 07	***
ab	1	3994.3	3994.3	4335.153	1.590e − 14	***
bc	1	39.7	39.7	43.034	6.391e − 05	***
Residuals	10	9.2	0.92			

5 possible null hypotheses:
H_o^i: $\beta_i = 0$ for an $i \in \{1, ..., 5\}$.

Is H_o^i true ?
Is the model true (under the NID) ?
What can be said about the Mean Sq in anova table ??
Do they look like $\sigma^2 = 1$?

> z=lm(y∼a+b+c)

> anova(z)

	Df	*Sum Sq*	*Mean Sq*	*F value*	*Pr(> F)*
a	1	14.7	14.66	0.0435	0.8383
b	1	47.1	47.09	0.1398	0.7150
c	1	159.6	159.63	0.4738	0.5043
Residuals	12	4043.2	336.93		

Three possible null hypotheses:
H_o^1: $\beta_1 = 0$.
H_o^2: $\beta_2 = 0$.
H_o^3: $\beta_3 = 0$.

Is H_o^i true ?
Is the model true (under the NID) ?
What can be said about the Mean Sq in anova table ??
Do they look like $\sigma^2 = 1$?

> mean((y[1:8]-y[9:16])**2/2)
[1] 1.130107 # $(= s^2 \approx \sigma^2$??)

Remark. If the model is wrong, s^2 is an unbiased estimators of σ^2,
but not $\hat{\sigma}^2$ and other mean squares in anova.
If the model is correct, both $\hat{\sigma}^2$ and s^2 are unbiased.

Simulation example 5.7.2

> y=rnorm(16)

> z=lm(y∼a+b+c)

> anova(z)

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
a	1	0.0212	0.02121	0.0282	0.869
b	1	0.8812	0.88119	1.1730	0.300
c	1	0.1444	0.14441	0.1922	0.668
Residuals	12	9.0148	0.75123		

3 possible null hypotheses:

H_o^i: $\beta_i = 0$ for an $i \in \{1, ..., 3\}$.

Is H_o^i true ?

Is the model true (under the NID) ?

What can be said about the mean squares in anova table ??

Do they look like $\sigma^2 = 1$?

> z=lm(y∼a+b+c+ab+bc)

> anova(z)

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
a	1	0.0212	0.02121	0.0254	0.8765
b	1	0.8812	0.88119	1.0563	0.3283
c	1	0.1444	0.14441	0.1731	0.6861
ab	1	0.0176	0.01762	0.0211	0.8873
bc	1	0.6553	0.65531	0.7856	0.3963
Residuals	10	8.3419	0.83419		

> mean((y[1:8]-y[9:16])**2/2)

[1] 0.986949

5 possible null hypotheses:

H_o^i: $\beta_i = 0$ for an $i \in \{1, ..., 5\}$.

Is H_o^i true ?

Is the model true (under the NID) ?

What can be said about the mean squares in anova table ??

Do they look like $\sigma^2 = 1$?

Remark. If the model is correct and H_o is correct, all mean squares are unbiased estimators of σ^2. But $\hat{\sigma}^2$ has smaller variance than the other Mean Sq., as its degree of freedom (Df) is larger. $\nu\hat{\sigma}^2/\sigma^2 \sim \chi^2(\nu)$ (with mean $= \frac{\nu}{2} \cdot 2$ $(= \alpha\beta)$, variance $= \alpha\beta^2 =$? Thus $E(\hat{\sigma}^2) = \sigma^2$ and $V(\hat{\sigma}^2) = 2\sigma^4/\nu$.

Simulation example 5.7.3

> n=100

> a=rexp(n)

> b=rbinom(n,5,0.5)

> a=c(a,a)

> b=c(b,b)

> e=rnorm(2*n)

> y=2+a+b+e

> z=lm(y∼a)

> anova(z)

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
a	1	215.36	215.365	111.09	$< 2.2e-16 ***$
Residuals	198	383.86	1.939		
			σ^2= 1± ??		

Note: $SS/\sigma^2 \sim \chi^2(Df)$ with Var $2 * Df$. Thus $1 \pm 2\sqrt{2/Df} \approx 1 \pm 0.2$

> w=lm(y∼a+b)

> anova(w)

$$\begin{pmatrix} & Df & Sum\ Sq & Mean\ Sq & F\ value & Pr(>F) \\ a & 1 & 215.37 & 215.365 & 210.83 & < 2.2e-16*** \\ b & 1 & 182.62 & 182.621 & 178.77 & < 2.2e-16*** \\ Residuals & 197 & 201.24 & 1.022 & & \\ & & & \sigma^2 = \mathbf{1??} & & \end{pmatrix}$$

> mean((y[1:n]-y[(n+1):$(2*n)$])$**$2/2)

[1] 0.9183548 # (= s^2)

σ^2 **= 1?**

Conclusion:

1. If the model is correct, $\frac{1}{n-p}\sum_i (Y_i - \hat{Y}_i)^2$ is an unbiased estimator of σ^2.
2. If the model is correct, $\beta_i = 0$, the corresponding Mean Sq is unbiased.
3. If there are replications, s^2 is unbiased.

If the model is incorrect, $\frac{1}{n-p}\sum_i (Y_i - \hat{Y}_i)^2$ is not an unbiased estimator of σ^2. This can be proved by a counterexample as follows.

Counterexample : Let

$Y_{ij} = \beta_1 X_i + \beta_2 Z_i + \epsilon_{ij}$, $j = 1, 2$, and $i = 1, ..., m$,

where X_i, Z_i and ϵ_{ij} are independent $\sim N(0, \sigma^2)$.

$s^2 = \frac{1}{m}\sum_{i=1}^m \frac{(Y_{i1}-Y_{i2})^2}{2}$

$= \frac{1}{m}\sum_{i=1}^m \frac{(\epsilon_{i1}-\epsilon_{i2})^2}{2}$

$= \frac{1}{m}\sum_{i=1}^m (\frac{\epsilon_{i1}-\epsilon_{i2}}{\sqrt{2}\sigma})^2\sigma^2$.

$\sum_{i=1}^m (\frac{\epsilon_{i1}-\epsilon_{i2}}{\sqrt{2}\sigma})^2 \sim \chi^2(m)$.

$=> E(s^2) = \sigma^2$. (Abusing notation, treating s^2 as a r.v.).

Now if the model is chosen incorrectly, say, consider model,

$Y_{ij} = \beta_1 X_i + W_{ij}$, where $W_{ij} = \beta_2 Z_i + \epsilon_{ij} \sim N(0, (\beta_2^2 + 1)\sigma^2)$,

$\tilde{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^m \sum_{j=1}^2 (Y_{ij} - \hat{Y}_{ij})^2$ is an unbiased estimator of $(\beta_2^2+1)\sigma^2 \quad \neq \sigma^2$.

$n = ?$ $p = ?$

$W_{i1} \perp W_{i2}$ **???**

$E(W_{11}W_{12}) = E(\beta_2^2 Z_1^2 + \beta_2 Z_1(\epsilon_{11}+\epsilon_{12}) + \epsilon_{11}\epsilon_{12}) = E(\beta_2^2 Z_1^2) = \beta_2^2 E(Z_1^2)$

$E(W_{11})E(W_{12}) = \beta_2^2 (E(Z_1))^2$. ...

Simulation example 5.7.4.

```
> a=rep(c(-1,1),4)
> b=rep(c(-1,-1,1,1),2)
> c=rep(-1,4)
> c=c(c,-c)
> n=80
> e=rnorm(n)
> a=rep(a,10)
> b=rep(b,10)
> c=rep(c,10)
> y=2*a−5*b+e
> a=factor(a)
> b=factor(b)
> c=factor(c)
> z=lm(y~a+b+c)
> summary(z)
```

Note that a, b and c are all factors.

Using Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \epsilon_i$

and under control treatment, X_{i1}**= ?**

What is β_0 **?**

What is β_1 **?**

Where to find $\hat{\beta}_j$**'s ?**

	Estimate	*Std. Error*	*t value*	$Pr(>\|t\|)$	
$(Intercept)$	3.2353??	0.2014	16.064	$< 2e-16$	$***$
$a1$	3.8619	0.2014	19.176	$< 2e-16$	$***$
$b1$	-10.1350	0.2014	-50.323	$< 2e-16$	$***$
$c1$	-0.2773	0.2014	-1.377	0.173	

Residual standard error: 0.9007 on 76 degrees of freedom

$\hat{Y} \approx \beta_0 = -2 + 5 = 3$ if $a = -1 = b = c$ under control.treat.

$\hat{Y} \approx \beta_0 + \beta_1 + \beta_2 + \beta_3 = 3 + 4 - 10 + 0 = -3$ if $a = 1 = b = c$ under control.treat.

$\beta_0 \approx 0 \approx \overline{Y}$ under control.sum.

> anova(z)

	Df	*Sum Sq*	*Mean Sq*	*F value*	$Pr(>F)$	
a	1	298.29	298.29	367.7073	$< 2e-16$	$***$
b	1	2054.37	2054.37	2532.4402	$< 2e-16$	$***$
c	1	1.54	1.54	1.8952	0.1727	
Residuals	76	61.65	0.81			

5.7.5. Homework. Carry out the simulations in §5.7 yourself with different parameters and $rnorm(n, 1, 2)$, then summarize the results and address the questions.

5.8. Interpretation of results.

Under NID assumption and 2^3 factorial designs,

$T_o = \frac{\overline{Y} - \beta_0}{\sqrt{s^2/n}} \sim t_{df}$, $\qquad (= N(0,1)/\sqrt{\chi^2(df)/df})$

$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2(\frac{1}{4r} + \frac{1}{4r})}} \sim t_{df}$, where

$df = 2^k(r-1)$ for s^2 in 2^k factorial design with r replicates and under the full model. $\hat{\beta}_j$ refers to one of the 7 effects.

Remark. In the linear regression, if the model is correct, then we have

$T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim t_{n-p}$, where

$\hat{\sigma}_j^2$ is the $j-$th diagonal element of $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, and

$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{j=1}^n (Y_j - \hat{Y}_j)^2$.

Notice $n = 2^k r$ and $p = 2^k$ in the previous case.

Example of pilot study in §5.4.

The data presented in §5.4 are the 8 averages of 2 replications in a 2^3 factorial design. The 16 data rather than the averages are as follows.

> y=c(59,74,50,69,50,81,46,79, 61,70,58,67,54,85,44,81) # yield of experiments in Table 5.3

> mean((y[1:8]-y[9:16])**2/2)

[1] 8 $\qquad = s^2$

$V(effect) = V(\overline{y}_+ - \overline{y}_-) = \sigma^2(\frac{1}{4r} + \frac{1}{4r})$

$SE = \sqrt{\frac{8}{4r} + \frac{8}{4r}} \approx 1.4.$

For the data in Table 5.3, df=8, $t_{8,0.025} \approx 2.3$, so a 95% confidence interval (CI) is

$\hat{\beta}_j \pm 2.3 \times 1.4$ (or $\hat{\beta}_j \pm 3.2$).

In practice, people prefer $\hat{\beta}_j \pm SE$, *i.e.*,

$\hat{\beta}_j \pm 1.4$, as it is more conservative (not relying on NID).

effects	$70\%CI$		
T	23.0 ± 1.4		*temperature* $(160, 180)$
C	-5.0 ± 1.4		*concentration* $(20, 40)$
K		1.5 ± 1.4	*catalyst* (A, B)
TC		1.5 ± 1.4	
TK	10.0 ± 1.4		
CK		0.0 ± 1.4	
TCK		0.5 ± 1.4	
	important	*ignorable*	*if* $\|effect\| \leq s$ nearly or too small

> z=lm(y~a+b+c+ab+bc+ac+abc) #**(a,b,c)=(T,C,K)** (are factors)

> anova(z)

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$	
a	1	2116	2116	264.500	$2.055e-07$	$***$
b	1	100	100	12.500	0.007670	$**$
c	1	9	9	1.125	0.319813	
ab	1	9	9	1.125	0.319813	
bc	1	0	0	0.000	1.000000	
ac	1	400	400	50.000	0.000105	$***$
abc	1	1	1	0.125	0.732810	
$Residuals$	8	64	8	$(=\hat{\sigma}^2 = s^2)$		

Implication:

> w=lm(y∼a+b+ac)

> anova(w,z)

Model 1: y ∼ a + b + ac

Model 2: y ∼ a + b + c + ab + bc + ac + abc

	$Res.Df$	RSS	$Df\ Sum\ of\ Sq$	F	$Pr(>F)$
1	12	83			
2	8	64	419	0.5938	0.6772

> anova(w)

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$	
a	1	2116	2116.00	305.928	$6.631e-10$	$***$
b	1	100	100.00	14.458	0.002519	$**$
ac	1	400	400.00	57.831	$6.292e-06$	$***$
$Residuals$	12	83	6.92			

Estimator of σ^2 can be 6.92 rather than 8.

Summary. Recall that a, b, ..., abc are factors defined in §5.6.

What does the main effect mean ?

lm(y∼a+b+c+bc) <=>

$E(Y|\mathbf{X}) = \beta_0 + \beta_1\mathbf{1}(a=1) + \beta_2\mathbf{1}(b=1) + \beta_3\mathbf{1}(c=1) + \beta_4\mathbf{1}(bc=1)$,

where $\mathbf{X'} = (1, \mathbf{1}(a=1), \mathbf{1}(b=1), \mathbf{1}(c=1), \mathbf{1}(b=c\in\{-1,1\}))$ or

$E(Y|\mathbf{X}) = \beta_0 + \beta_{-1}\mathbf{1}(a=-1) + \beta_1\mathbf{1}(a=1) + \beta_{-2}\mathbf{1}(b=-1) + \beta_2\mathbf{1}(b=1) + \cdots$

$\mathbf{X'} = (1, \mathbf{1}(a=-1), \mathbf{1}(a=1), \mathbf{1}(b=-1), \mathbf{1}(b=1), ...)$ with $\beta_{-1} = 0 = \beta_{-2} = ...)$

lm(y∼a+b*c) <=>

$E(Y|\mathbf{X}) = \beta_0 + \beta_1\mathbf{1}(a=1) + \beta_2\mathbf{1}(b=1) + \beta_3\mathbf{1}(c=1) + \beta_4\mathbf{1}(b*c=1)$,

where $\mathbf{X'} = (1, \mathbf{1}(a=1), \mathbf{1}(b=1), \mathbf{1}(c=1), \mathbf{1}(b=c=1))$. (Compare to bc).

> lm(y∼a+b*c)

$(Intercept)$	$a1$	$b1$	$c1$	$b1:c1$
$5.450e+01$	$2.300e+01$	$-5.000e+00$	$1.500e+00$	$3.553e-15$

> lm(y∼a+b+c+bc)

$(Intercept)$	$a1$	$b1$	$c1$	bc
$5.450e+01$	$2.300e+01$	$-5.000e+00$	$1.500e+00$	$4.441e-16$

Remark. It is easier to see the difference through the next model.

> lm(y∼a+c+ac)

$(Intercept)$	$a1$	$c1$	$ac1$
47.0	23.0	1.5	10.0

$\mathbf{1}(ac=1) = \mathbf{1}(a=c\in\{-1,1\})$

> (z=lm(y∼a*c))

$(Intercept)$	$a1$	$c1$	$a1:c1$
57.0	13.0	-8.5	20.0

$\mathbf{1}(a:c=1) = \mathbf{1}(a=c=1)$

> predict(z,newdata=data.frame(a="-1",c="-1"))

57 # $= \underbrace{57+0+0+0}_{y\sim a*c} = \underbrace{47+0+0+10}_{y\sim a+c+ac}$

> predict(z,newdata=data.frame(a="1",c="-1"))

70 # $= 57+13+0+0 = 47+23+0+0$

> predict(z,newdata=data.frame(a="-1",c="1"))

48.5 # $= 57+0-8.5+0 = 47+0+1.5+0$

> predict(z,newdata=data.frame(a="1",c="1"))

81.5 # $= 57+13-8.5+20 = 47+23+1.5+10$

Observations:

(1) If one changes the model from $y \sim a+b+c+ab+ac+bc+abc)$ to $y \sim a+c+ac$,
the LSE of (β_a, β_c), remains the same,
due to the vectors in the table of contrast are orthogonal.

(2) If one changes the model from $y \sim a+b+c+ab+ac+bc+abc)$ to $y \sim a+c+a:c)$,
the LSE of (β_a, β_c) may not be the same,
as $(-1,1,-1,1,-1,1,-1,1)(-1,-1,-1,-1,-1,1,-1,1)' \neq 0 \quad (X_a' X_{a:c} \neq 0)$

$$\left(\begin{array}{cccccccccccc} a & =(& -1, & 1, & -1, & 1, & -1, & 1, & -1, & 1)' & \\ c & =(& -1, & -1, & -1, & -1, & 1, & 1, & 1, & 1)' & a'c=0 \\ a:c & =(& -1, & -1, & -1, & -1, & -1, & 1, & -1, & 1)' & \\ a*(a:c) & = & 1 & -1 & 1 & -1 & 1 & 1 & 1 & 1 & =4 \end{array}\right)$$

(3) However, the prediction of Y remains the same.
Under control.treatment,
the intercept is the estimate of the mean response of Y at the low levels of factors.
The main effect is the est. of the change due to the factor changing from - to +.

The conclusion of the experiment: $Y \approx 47 + 23T - 8.5K + 20TK$.
To get high yields of the product, set

1. the temperature $T = 180$ (high);
2. the concentration at C=20 (low);
3. It was thought that the suppliers of catalyst K do not matter and they were supposed to produce the same type of catalyst. In fact c (or K) is not significant. However, they now notice that TK is significant. Further study of the data yields

run#	*T*	*C*	*K*	*outputs :*			
1	−		−		60		
2	+		−			72	
3	−		−		54		
4	+		−			68	
5	−		+	52			
6	+		+				83
7	−		+	45			
8	+		+				80
mean				48.5	57	70	81.5

They should select the better supplier (who supplies catalyst B (K+)).

Remark. A 2^2 factorial design $\begin{array}{ccc} - & - & y_1 \\ - & + & y_2 \\ + & - & y_3 \\ + & + & y_4 \end{array}$ can be viewed as an additive model for one-way ANOVA or two-way ANOVA.

For one-way anova: $Y_{ij} = \eta + \tau_i + \epsilon_{ij}$, $i, j \in \{1, 2\}$, where $(Y_{11}, Y_{12}, Y_{21}, Y_{22}) = (y_1, y_2, y_3, y_4)$.

For two-way anova: $Y_{ij} = \eta + \tau_i + \theta_j + \epsilon_{ij}$, $i, j \in \{1, 2\}$.
In particular, under two-way anova, one can write

$Y_{ij} = \eta + \tau_i + \theta_j + \epsilon_{ij}$, $i, j \in \{0, 1\}$.

$Y_{ij} = \eta + \tau_0 \mathbf{1}(i = 0) + \tau_1 \mathbf{1}(i = 1) + \theta_0 \mathbf{1}(j = 0) + \theta_1 \mathbf{1}(j = 1) + \epsilon_{ij}$, $i, j \in \{0, 1\}$.

$Y_{ij} = \eta + \tau_1 \mathbf{1}(i = 1) + \theta_1 \mathbf{1}(j = 1) + \epsilon_{ij}$, $i, j \in \{0, 1\}$, under control.treatment.
Q: $\tau_1 =$? if (1) under default; (2) under control.sum with H_0: $\tau_0 = \tau_1$.

5.9. Table of contrast.

Yates number	*mean*	*a*	*b*	*c*	*ab*	*ac*	*bc*	*abc*
1	1	−1	−1	−1	1	1	1	−1
2	1	1	−1	−1	−1	−1	1	1
3	1	−1	1	−1	−1	1	−1	1
4	1	1	1	−1	1	−1	−1	−1
5	1	−1	−1	1	1	−1	−1	1
6	1	1	−1	1	−1	1	−1	−1
7	1	−1	1	1	−1	−1	1	−1
8	1	1	1	1	1	1	1	1
df	8	4	4	4	4	4	4	4

Notice that $Y'(a, b, c, ab, ac, bc, abc)/4 =$(7 effects), where $Y' = (y_1, ..., y_8)$

5.10. Misuse of the ANOVA for 2^k factorial experiments.
If there is no replicate runs ($r = 1$), then ANOVA may not be very helpful (see explanation before Simulation example 5.7.1).
Skip the rest of the section.

5.11. Eyeing the data. In some special case, the interactions are negligible. Then the main factors are orthogonal, and one can do contour eyeballing.

Example of Testing worsted yarn. (jing fang mao xian) Table 5.6 shows part of the data from an investigation on the strength of the particular type of yarn under cycles of repeated loading. This is a 2^3 factorial design with 3 factors:
Length of specimen (A) ((250,350) mm),
amplitude of load cycle (B) ((8,10) mm),
load (C) ((40,50) g).

Yates #	*A*	*B*	*C*	*durance y*
1				28
2				36
3				22
4				31
5				25
6				33
7				19
8				26

Table 5.6

The effects are

mean	*A*	*B*	*C*	*AB*	*AC*	*BC*	*ABC*
27.5	8	−6	−3.5	0	−0.5	−0.5	−0.5

Notice the interaction effects are all negligible ($|-0.5| \leq |main\ effect|/7$).
$\vec{A}$, $\vec{B}$ and $\vec{C}$ are essentially orthogonal.
The direction of steepest ascent is then $(8, -6, -3.5)$.

The contour plane of a durance 25 is a hyperplane $25 = (8, -6, -3.5)\begin{pmatrix} A_o \\ B_o \\ C_o \end{pmatrix}$

where $A_o = (x_1-250)/(350-250)$, $B_o = (x_2-8)/(10-8)$, $C_o = (x_3-40)/(50-40)$, (see Figure 5.6). **Where are (250,350) come from ?**

The contour plane of a durance y is a hyperplane $y = (8, -6, -3.5)\begin{pmatrix} A_o \\ B_o \\ C_o \end{pmatrix} = f(\vec{x})$

5.12. Dealing with more than one response: A pet food experiment.
The manufacturer of pet food had received complaints that packages of food pellets received by the customers contained an unsatisfactorily large amount of powder.

The factory did a 2^3 factorial design to investigate it. All in two levels.
A. Conditioning Temperature: 80% at max, or max
B: Flow: 80% at max, or max
C: Compression zone: 2, or 2.5

Responses:
Y_1 – powder in product;
Y_2 – powder in plant;
Y_3 – a measure of yield;
Y_4 – energy consumed.

Y_1 was obtained after the same process as if a customer would eventually get it.
They tried to find out
the relation between Y_1 and Y_2, as well as
how to control the response Y_2 by adjusting the factors,
without losing too much in yield Y_3 and energy Y_4.

Responses in standard (Yates) order:
y1=c(132,107,117,122,102,92,107,104)
y2=c(166,162,193,185,173,192,196,164)
y3=c(83, 85, 99, 102, 59, 75, 80,73)
y4=c(235,224,255,250,233,223,250,249)

Rough estimates of errors are obtained through previous duplicated runs:
$\hat{\sigma}_1 = 5.6$ (for Y_1),
$\hat{\sigma}_{effect_1} = \hat{\sigma}_1\sqrt{\frac{1}{4}+\frac{1}{4}} = \hat{\sigma}_1/\sqrt{2} = 5.6/\sqrt{2} = 4.0$,
$\hat{\sigma}_2 = ...\ \hat{\sigma}_3 = ...\ \hat{\sigma}_4 = ...$

$$\hat{\sigma}_{effect_i} = \begin{cases} 4.0 & \text{if } i=1 \text{ (for } Y_1) \\ 7.4 & \text{if } i=2 \text{ (for } Y_2) \\ 4.9 & \text{if } i=3 \text{ (for } Y_3) \\ 1.1 & \text{if } i=4 \text{ (for } Y_4) \end{cases}$$

Finding:

There is no serious correlation between Y_1 and Y_2 by plotting (Y_1, Y_2) and
> cor(y1,y2)
[1] -0.1686297
> summary(lm(y2∼y1))

	Estimate	*Std.Error*	*tvalue*	$Pr(>\lvert t\rvert)$	
(Intercept)	199.7701	50.1472	3.984	0.00725	$=> \hat{y2} = 200 + 0 \times y1$
$y1$	-0.1893	0.4518	-0.419	0.68976	

		powder in product Y_1	*powder in plant* Y_2	*yield* Y_3	*energy* Y_4
	$\hat{\sigma}_i$	4	7.4	4.9	1.1
	temp, A	-8.2	-6.3	3.5	$-6.8*$
	flow, B	4.3	11.3	$13*$	$22.3*$
relate to	*zone, C*	$-18.2*$	4.8	$-20.5*$	-2.3
powder	*AB*	$9.3*$	-13.7	-5.5	$3.8*$
inert	*AC*	1.7	-0.3	1	1.3
inert	*BC*	4.3	-13.7	-3.5	-0.8
inert	*ABC*	-5.7	-11.7	-6	0.8

=> adjustment should set { zone at + or 2.5, from row C,
temp*flow at −, from row AB.

How to choose the levels from A and B ?

$(AB)-$	$energy(\hat{Y}_4)$	$\hat{Y}_1$	$yield(\hat{Y}_3)$
$(A+, B-)$	$-6.8 - 22.3$	$-8.2 - 4.3$	
$(A-, B+)$		$8.2 + 4.3$	$-3.5 + 13$

If energy saving is more important: $(A+, B-)$, which also decreases Y_1,
otherwise $(A-, B+)$ $(\hat{Y}_3, \hat{Y}_1) = (-3.5 + 13, 8.2 + 4.3)$.

5.13. A 2^4 factorial design: Process development study

Often there are more factors to be investigated than can conventionally be accommodated within the time and budget available, but you will find that usually you can separate genuine effects from noise without replication. In a pilot study, if one plans 16 runs for a replicated 2^3 factorial design with 3 factors, it can be replaced by a 2^4 factorial design with 4 factors.

A process development study.

Factors:

1. (a). Catalyst charge (lb) (10,15) or (−,+), (yongliang)
2. (b). Temperature (^oC) (220,240) or (−,+),
3. (c). Pressure (psi) (50,80) or (−,+),
4. (d). Concentration (%) (10,12) or (−,+),

Table 5.10a. Data

Yates run #	1	2	3	4	*conversion*(%)	*random order*
1	−	−	−	−	70	8
2	+	−	−	−	60	2
3	−	+	−	−	89	10
4	+	+	−	−	81	4
5	−	−	+	−	69	15
6	+	−	+	−	62	9
7	−	+	+	−	88	1
8	+	+	+	−	81	13
9	−	−	−	+	60	16
10	+	−	−	+	49	5
11	−	+	−	+	88	11
12	+	+	−	+	82	14
13	−	−	+	+	60	3
14	+	−	+	+	52	12
15	−	+	+	+	86	6
16	+	+	+	+	79	7

```
x=c(70,60,89,81,69,62,88,81,60,
    49,88,82,60,52,86,79)
a=rep(c(-1,1),8)
b=rep(c(-1,-1,1,1),4)
c=rep(-1,4)
c=c(c,-c,c,-c)
d=c(rep(-1,8),rep(1,8))
ab=a*b
ac=a*c
ad=a*d
bc=b*c
bd=b*d
cd=c*d
abc=ab*c
abd=ab*d
acd=ac*d
bcd=bc*d
abcd=ab*cd
mean(x)
lm(x~factor(a))$coef[2]      # = ?
sum(x*a)/8                   # = ?
x%*%a/8                      # = ?
    round(lm(x~a*b*c*d)$coef[2:16],2)*2
```

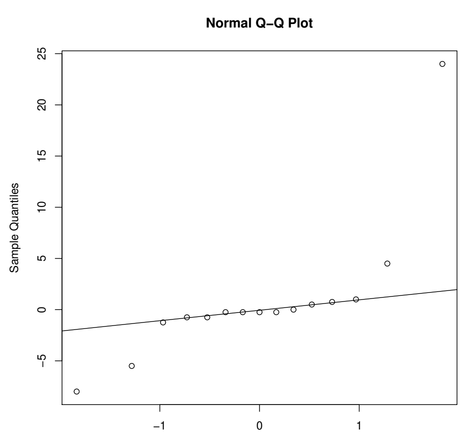


Figure 5.10

The average is 72.25.

The effects are

a	b	c	d	ab	ac	ad	bc	bd	cd
-8.00	24.00	-0.25	-5.50	1.00	0.75	0.00	-1.25	4.50	-0.25

abc	abd	acd	bcd	$abcd$
-0.75	0.50	-0.25	-0.75	-0.25

In a 2^3 factorial design with r replicates,
σ^2 is estimated by

$\tilde{\sigma}^2 = \frac{1}{2^3}\sum_{i=1}^{2^3} s_i^2$, where

$s_i^2 = \frac{1}{r-1}\sum_{h=1}^{r}(Y_{ih} - \overline{Y}_{i\cdot})^2$,

df of $\tilde{\sigma}^2$ is $2^3(r-1)$.

$V(effect) = \tilde{\sigma}^2(\frac{1}{4r} + \frac{1}{4r})$, as effect $= \overline{y}_+ - \overline{y}_-$.

A CI for effect is

effect $\pm t_{df,0.025}\sqrt{\tilde{\sigma}^2(\frac{1}{4r} + \frac{1}{4r})}$.

In this example, there is no replication (16 runs with 16 parameters).
The 5 3-factor and 4-factor interaction effects can be viewed as errors.
A conservative estimate of the SE of effect ($\sqrt{V(effect)}$ is
$\sqrt{\frac{\sum_{i=11}^{15} effect_i^2}{5}} \approx 0.55$ (treating each of the 5 effect$_i$'s as a variation of the **effect**).
The CI of effect is then

effect $\pm t_{5,0.025}0.55$ ($= 2.57 * 0.55$).

It can be justified by qq-plot.
The significant effects can be found out:

a	b	c	d	ab	ac	ad	bc	bd	cd	
		-0.25		1.00	0.75	0.00	-1.25		-0.25	
-8.00	24.00		-5.50					4.50		$**$
abc	abd	acd	bcd	$abcd$						
	-0.75	0.50	-0.25	-0.75	-0.25					

Remark. Factor c and the interactions related to c are inert. It becomes a 2^3 factorial design, with replication $c = 2$. Thus we can use s^2 to estimate σ^2.

It is interesting to see from Figure 5.10 (see last page) that the significant effects can be detected by the qq-plot against normal distribution.

It is also interesting to see from the following stem-and-leaf plot that all but the 4 significant effects appear normal distribution.

Thus one may use all but 4 effects to estimate V(effect)

u=c(-8.00,24.00,-0.25,-5.50 ,1.00 ,0.75, 0.00, -1.25, 4.50,-0.25,-0.75, 0.50,-0.25,-0.75,-0.25)

u=u[abs(u)<2] **what is it ?**

sort(u)

[1] -1.25 -0.75 -0.75 -0.25 -0.25 -0.25 -0.25 0.00 0.50 0.75 1.00

stem(u)

The decimal point is at the |

-1 | 3
-0 | 88
-0 | 3333
 0 | 0
 0 | 58
 1 | 0

sqrt(mean(u*u))

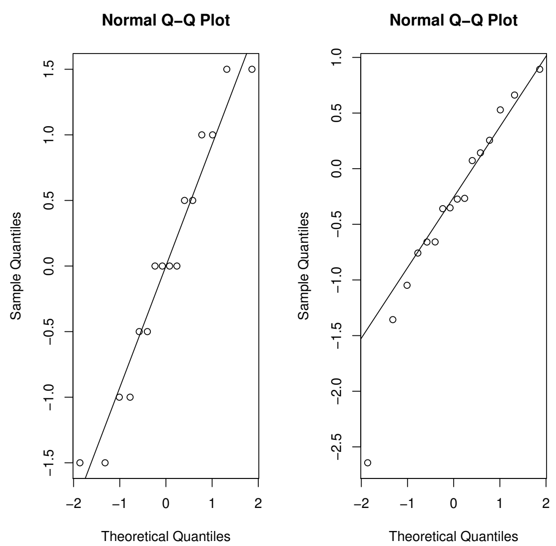
[1] 0.6571287 **what is it ?**

> summary(lm(x~a*b*d))

	Estimate	*Std. Error*	*t value*	$Pr(>\|t\|)$	
(*Intercept*)	$7.225e+01$	$3.307e-01$	218.463	$< 2e-16$	$***$
a	$-4.000e+00$	$3.307e-01$	-12.095	$2.02e-06$	$***$
b	$1.200e+01$	$3.307e-01$	36.285	$3.65e-10$	$***$
d	$-2.750e+00$	$3.307e-01$	-8.315	$3.30e-05$	$***$
$a:b$	$5.000e-01$	$3.307e-01$	1.512	0.169020	
$a:d$	$-3.955e-16$	$3.307e-01$	0.000	1.000000	
$b:d$	$2.250e+00$	$3.307e-01$	6.803	0.000137	$***$
$a:b:d$	$2.500e-01$	$3.307e-01$	0.756	0.471362	

Residual standard error: 1.323 on 8 degrees of freedom.

Remark. $s^2 = 1.323^2 = \hat{\sigma}^2$. The 3rd $\hat{\sigma}_{effect} = 1.332\sqrt{\frac{1}{4r} + \frac{1}{4r}}$ **What are the first two ?**



How to explain ties ?

Interpretation of the data.

1. Conversion changes -8% if catalyst charge switches from 10 to 15.
2. Pressure is inert.
3. To increase conversion set catalyst charge at 10 lb, temperature at 240^oC , concentration at 10%. It causes 33% increase.
4. Interaction between temperature and concentration can be seen from **Figure 5.11** in the textbook.
5. It reduces to a duplicated 2^3 FD.
6. $x = 72 - 4a + 12b - 2.75d + 2.25bd$. **Do we need to run lm() again for the LSE ?**

   (a). Catalyst charge (lb) (10,15) or $(-,+)$, (yongliang)

   (b). Temperature (^oC) (220,240) or $(-,+)$,

   (d). Concentration (%) (10,12) or $(-,+)$,

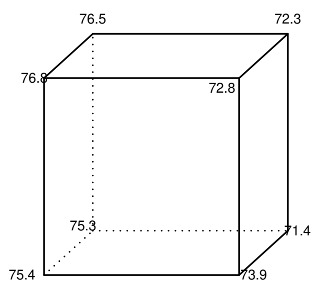
5.14. A first look at sequential assembly. The process of investigation includes interactive deduction and induction. Running an experiment can gain improvement on the production, but also indicates the possibility of even further advance and shows where additional runs needed to be made. It is called **sequential assembly**.

Experiment by Hill and Wiles (1975).
The object: to increase the disappointingly low yields of a chemical product.
3 factorial designs were run in sequence but only the first will be described here.
In phase I, a 2^3 factorial design was run.

3 factors:
concentration C,
rate of reaction R,
temperature T.

$run\#$	C	R	T	y_i $(yields)$
$phase\ I$				
1	$-$	$-$	$-$	
2	$+$	$-$	$-$	
3	$-$	$+$	$-$	
4	$+$	$+$	$-$	
5	$-$	$-$	$+$	
6	$+$	$-$	$+$	
7	$-$	$+$	$+$	
8	$+$	$+$	$+$	

```
> y=c(75.4,73.9,76.8,72.8,75.3,71.4,76.5,72.3)
```



Visual Display suggests that C is significant

```
> C=rep(c(-1,1),4), R=rep(c(-1,-1,1,1),2), c=rep(-1,4), T=c(c,-c)
> z=lm(y~ C*R*T)$coef
> c(z[1],2*z[2:8]) # (no need  to define factors) The effects are
```

$\overline{y}$	C	R	T	CR	CT	RT	CRT
74.3	-3.4	0.6	-0.85	-0.7	-0.65	0.45	0.55
$intercept$	$factors$						
76.3	??						
	$significant$						
	Why ?						

> H=lm(y∼ factor(C)) **Do we need to update the estimates ?**

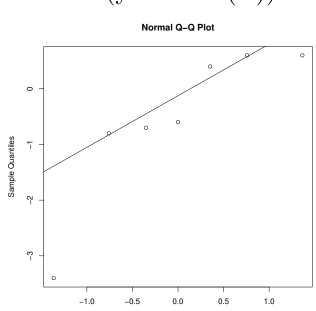


Fig. 1 QQplot

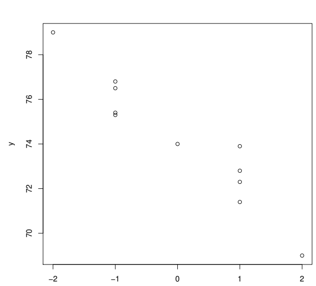


Fig. 2 (plot(a,y))

```
> qqnorm(z[2:8])
> qqline(z[2:8])
```

> sqrt(mean((2*z[3:8])**2)) [1] 0.6454972 # $\hat{\sigma} = 0.65$ or $\hat{\sigma}_{effect} = 0.65$?

> summary(H)$coef[2,2] [1] 0.6454972

> sqrt(anova(H)[2,3]/2) # $\sqrt{\sigma^2(\frac{1}{4}+\frac{1}{4})}$ [1] 0.6454972

Thus the model becomes $Y = \alpha + \beta\mathbf{1}(C=1) + \epsilon$ or $\hat{Y} = 76.3 - 3.4\mathbf{1}(C=1)$.

In phase II since C is significant, 3 runs were further made.

Purpose: To check whether the next new model is appropriate:

lm(y~ C) # $\hat{y} = 74.3 - 1.7C$

Moreover, whether further improvement can be made.

run#	*C*	*R*	*T*	*phase II*
9	-2	0	0	
10	0	0	0	
11	2	0	0	

> y=c(y,79,74,69)

> a=c(C,-2,0,2)

> plot(a,y) # See above Fig. 2. **What does it suggest ?**

> (u=lm(y~a))

(Intercept) a

74.22 -2.10

fitted equation: $\hat{Y}_u = 74.22 - 2.1a$.

Compare to the original simplified fitted equation:

$\hat{Y} = 76.0 - 3.4\mathbf{1}(C=1)$ or $\hat{Y} = 74.3 - 1.7C$.

Can we further improve the yield by reducing the concentration C ?

Possible further experiment design ?

Remark. The difference between

$lm(y \sim a + b)$ and $lm(y \sim factor(a) + b)$, and 2^k factorial designs.

> n=20

> a=rbinom(n,3,0.5)

> b=rbinom(n,3,0.5)

> y=74−2*a+rnorm(n,0,2)

> x=factor(a)

> lm(y~a) # True model: $E(Y|X) = (\beta_0, \beta_1)(1, a)^t = 74 - 2a$

$(Intercept)$	a
74.011	-1.943

> lm(y~x) # True model: $E(Y|X) = (\beta_1, ..., \beta_4)X$

$$= 74 - 2\mathbf{1}(a=1) - 4\mathbf{1}(a=2) - 6\mathbf{1}(a=3)$$

$(Intercept)$	$x1$	$x2$	$x3$
73.764	-1.580	-3.824	-5.494

> lm(y~x+b) # True model: $Y = (\beta_1, ..., \beta_5)(1, \mathbf{1}(a=1), \mathbf{1}(a=2), \mathbf{1}(a=3), b)^t + \epsilon$

$= 74 - 2\mathbf{1}(a=1) - 4\mathbf{1}(a=2) - 6\mathbf{1}(a=3) + 0 \cdot b + \epsilon$

$(Intercept)$	$x1$	$x2$	$x3$	b
73.5866	-1.6907	-3.9429	-5.4584	0.1777

The LSE and prediction are all different now.

5.16. Blocking the 2^k factorial designs.

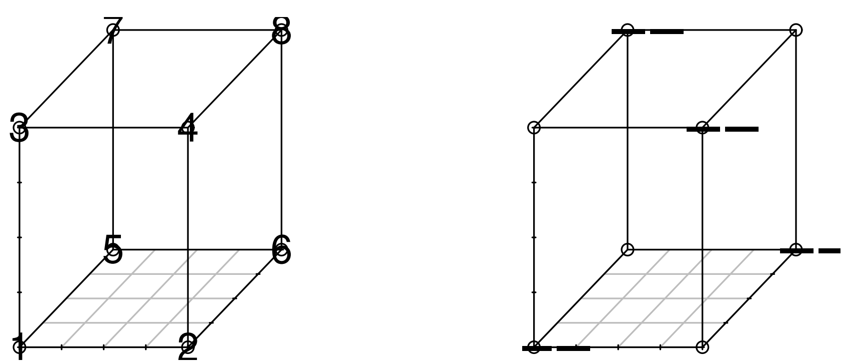
In a trial to be conducted using a 2^k factorial design, one either use 2^k different batches of raw materials or one batch of the same material. Otherwise, one may need blocking idea. Block sizes can be 2, 2^2, ..., 2^{k-1}. *e.g.*, for 2^3 FD, the block sizes are 2 and 4.

Block of size 4 for 2^3 FD. If one batch of raw material is only enough for 4 experiment, then partition according to 123 (or abc) $= \pm 1$.

	mean	*a*	*b*	*c*	*ab*	*ac*	*bc*	*abc*
1	1	−1	−1	−1	1	1	1	−1
2	1	1	−1	−1	−1	−1	1	1
3	1	−1	1	−1	−1	1	−1	1
4	1	1	1	−1	1	−1	−1	−1
5	1	−1	−1	1	1	−1	−1	1
6	1	1	−1	1	−1	1	−1	−1
7	1	−1	1	1	−1	−1	1	−1
8	1	1	1	1	1	1	1	1
df	8	4	4	4	4	4	4	4

1	1	−1	−1	−1	1	1	1	−1	
4	1	1	1	−1	1	−1	−1	−1	
6	1	1	−1	1	−1	1	−1	−1	
7	1	−1	1	1	−1	−1	1	−1	*block* 1
2	1	1	−1	−1	−1	−1	1	1	*block* 2
3	1	−1	1	−1	−1	1	−1	1	
5	1	−1	−1	1	1	−1	−1	1	
8	1	1	1	1	1	1	1	1	

It leads to two sets of the run #:
$\{1, 4, 6, 7\}$ and $\{2, 3, 5, 8\}$
Drawback: It cannot estimate 3-factor interaction.
abc = block variable.
Advantage: See Figure 5.16
the batch I (1,4,6,7) and the batch II (2,3,5,8) are in 4 opposite vertices.



Remark. If we ignore the confounding effect, then what happans ? For example, in the previous case, suppose $Y = \beta_1 \mathbf{1}(a = 1) + \beta_2 \mathbf{1}(abc = 1) + \epsilon$, where β_2 is the confounding effect of different batch of raw materials and interaction abc. If we ignore confounding effects, and set $Y = \beta_1 \mathbf{1}(a = 1) + \epsilon_m$, then

$$\sigma^2_{\epsilon_m} = Var(\beta_2 \mathbf{1}(abc = 1) + \epsilon) = \beta_2^2 pq + \sigma^2 \qquad \textbf{Why ??}$$

Hence NID fails. If $\beta_2^2 + \sigma^2 > 2\beta_1$, then β_1 is likely to become insignificant.

Can we use ab, or ac, or bc ?

> Yes, but it is often that abc is inert. Moreover, it is not like abc which form 2 pair of opposite vertices. *e.g.* bc leads to (1,2,7,8), v.s. (3,4,5,6).

Can we use a or b or c as a partition factor ?

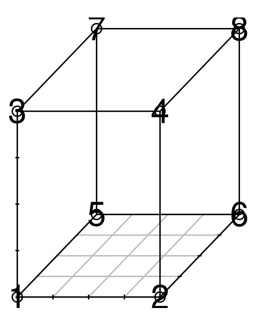
> No, we need to estimate the main effect, which is often more important than other effects, do not let it be confounded with the block factor.

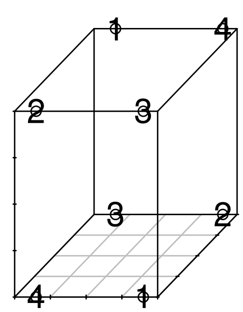
Block of size 2 for 2^3 FD. If a batch of raw material can only be used in two experiments, partition according to (12,13) (or (ab,ac)).
It leads to 4 sets of the run due to $--$, $-+$, $+-$, $++$:

run #	**1**	**2**	**3**	**4 = 12**	**5 = 13**	*block#*	
1	−	−	−	+	+	*IV*	
2	+	−	−	−	−	*I*	
3	−	+	−	−	+	*II*	
4	+	+	−	+	−	*III*	
5	−	−	+	+	−	*III*	*Table* 1
6	+	−	+	−	+	*II*	
7	−	+	+	−	−	*I*	
8	+	+	+	+	+	*IV*	

2	+	−	−	−	−	*I*	
7	−	+	+	−	−	*I*	
3	−	+	−	−	+	*II*	
6	+	−	+	−	+	*II*	
4	+	+	−	+	−	*III*	*Table* 2
5	−	−	+	+	−	*III*	
1	−	−	−	+	+	*IV*	
8	+	+	+	+	+	*IV*	

The two block positions can be viewed as factors **4** and **5**, together with the original 3 factors **1**, **2**, and **3** (or a, b, c). Each pair is on the opposite vertex of the cube.





Thus there is no confounding (main) effects.

		runs	
variable **5**	+	(3, 6)	(1, 8)
	−	(2, 7)	(4, 5)
		−	+
		variable **4**	

The advantage of this approach is that the 3 factors a, b, c are all in different values (see Table 2).

How about let $(4,5) = (12,23)$? or (13,23) ?

Patterns not to partition. **(4,5)**= (**123**,**23**) (or (abc,bc)), due to $--, +-, -+, ++$.

run #	**1**	**2**	**3**	**4 = 123**	**5 = 23**	*block#*
1	−	−	−	−	+	*III*
2	+	−	−	+	+	*IV*
3	−	+	−	+	−	*II*
4	+	+	−	−	−	*I*
5	−	−	+	+	−	*II*
6	+	−	+	−	−	*I*
7	−	+	+	−	+	*III*
8	+	+	+	+	+	*IV*

It leads to 4 sets of the run #:

	*			*		
4	+	+	−	−	−	*I*
6	+	−	+	−	−	*I*
3	−	+	−	+	−	*II*
5	−	−	+	+	−	*II*
1	−	−	−	−	+	*III*
7	−	+	+	−	+	*III*
2	+	−	−	+	+	*IV*
8	+	+	+	+	+	*IV*

The drawback of this approach is that factor a is the same in each block.
Not to partition according to (1,123) (or (a,abc), due to $--, +-, -+, ++$.

run #	**1**	**2**	**3**	**4 = 123**	*block#*
1	−	−	−	−	*I*
2	+	−	−	+	*IV*
3	−	+	−	+	*III*
4	+	+	−	−	*II*
5	−	−	+	+	*III*
6	+	−	+	−	*II*
7	−	+	+	−	*I*
8	+	+	+	+	*IV*

It leads to 4 sets of the run #:

	*			*	
1	−	−	−	−	*I*
7	−	+	+	−	*I*
4	+	+	−	−	*II*
6	+	−	+	−	*II*
3	−	+	−	+	*III*
5	−	−	+	+	*III*
2	+	−	−	+	*IV*
8	+	+	+	+	*IV*

The drawback of this approach is that factor a is the same in each block.
In the above two cases, the block factors confounded with a.

Generators and defining relations.

Recall the table of contrast:
$$\begin{array}{c} \begin{array}{cccccccc} I & a & b & c & ab & ac & bc & abc \end{array} \\ \begin{pmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array}.$$
It can be viewed as a 8×8 matrix, with each column being an 8×1 vector, say

$(I\quad \mathbf{1}\quad \mathbf{2}\quad \mathbf{3}\quad \mathbf{12}\quad \mathbf{13}\quad \mathbf{23}\quad \mathbf{123})$ or
$\big(I\quad \vec{a}\quad \vec{b}\quad \vec{c}\quad \vec{ab}\quad \vec{ac}\quad \vec{bc}\quad \vec{abc}\big)$.
Then I=**11**=**22**=**33**=**44**=**55**,
$\mathbf{1}I = \mathbf{1} = I\mathbf{1}$,
2I=**2**=I**2**, ... , as how we get ab, ac, ...
Recall in R, a*a=$(1, ..., 1)'$ and $a\% * \%a = ?$

The defining relations **4**=**12** and **5**=**13** for two new factors in the previous cases are also called generators.

Then I=**2345** as **2345**= **231213**=I (=**124135**) and I=**124**=**135**
Namely, **45**=**23**, or **4** and **5** are confounded with **23**, **12**, **13** (none is a main effect), in the sense that each element in {**4**,**5**,**45**,**23**,**12**,**13**,**125**,**134**} is either **4** or **5** or **45**.

On the other hand, if we let **4**=**123** and **5**=**23**, and form 4 blocks (out of 8 runs) by $(\mathbf{4}, \mathbf{5})$,

then I=**451** as I=**1234235**=**451**. Also I=**1234**=**235**.
That is, **45**=**1**, or **4** and **5** are confounded with **123**, **23**, **1** (with one main effect).

What happens to (**4**,**5**)=(**12**,**23**) or (**13**,**23**) ?

Finally, if we let **4**=**123** and form 4 blocks by $(\mathbf{1}, \mathbf{4})$,
Then **1** and **4** are clearly confounded with the block factor **1** (as well as, **4**, **123**, **23**, **14**). How about **4**=**13** and form 4 blocks by (**1**,**4**) ?

5.16.2. Homework. Answer the previous two question marks.
Connection between defining relations and blocking:

1. Use higher order interaction if possible.
2. The new defining factors have interaction of higher order.

For more details, see Table 5A.1 as follow.
Table 5A.1. Blocking Arrangements for 2^k FD.

k	block size	block generator	
3	4	123	*how about*
	2	12, 13	21, 23?
4	8	1234	
	4	124, 134	
	2	12, 23, 34	
5	16	12345	
	8	123, 345	
	4	125, 235, 345	
	2	12, 13, 34, 45	
6	32	123456	
	16	1236, 3456	
	8	135, 1256, 1234	
	4	126, 136, 346, 456	
	2	12, 23, 34, 45, 56	

Examples of 2^6 FD, with block size 8.
The first example (which is in the table).
Define B_1 = **135**, B_2=**1256** and B_3=**1234**. Then
B_1B_2=**1351256**=**236**,
B_1B_3=**1351234**=**245**,
B_3B_2=**12341256**=**3456**,
$B_1B_2B_3$=**13512561234**=**146** (no replication of numbers).
Thus B_1, B_2 and B_3 are confounded with **135**, **1256**, **1234**, **236**, **245**, **3456**, **146**.
Interpretation:
These effects **135**, **1256**, **1234**, **236**, **245**, **3456**, **146** cannot be estimated.
Their order (of interaction): 3+.

Another example (not in the table).
Define A_1=**12456**, B_2=**1256** and B_3=**1234**. Then

A_1B_2=**4**,
A_1B_3=**356**,
B_3B_2=**3456**,
$A_1B_2B_3$=**123**.

Interpretation:
These effects **12456**, **1256**, **1234**, **4**, **356**, **3456**, **123** cannot be estimated.

Their order: 1+

Is it appropriate ?

How about (**246**, **1236**, **2345**) ?
The third example. Define A_2=**1245**, B_2=**1256** and B_3=**1234**. Then

A_2B_2=**46**,
A_2B_3=**35**,
B_3B_2=**3456**,
$A_2B_2B_3$=**1236**.

Interpretation:
These effects **1245**, **1256 1234**, **46**, **35**, **3456**, **1236** cannot be estimated.

Their order: 2+.

Is it appropriate ?
Which is the best among these three ?

Capter 6 Fractional Factorial Designs

We shall introduce the concept through examples.

6.1. Experiment on effects of 5 factors on six properties of films in 8 runs.

Factors :		$-$	$+$		
A :	*catalyst*(%)	1	1.5		
B :	*additive*(%)	0.25	0.5		
C :	*emulsifier P* (%)	2	3	*ruhuaji*	or (a,b) $\subset (2,3)$?
D :	*emulsifier Q* (%)	1	2		why not 2 & 3 ?
E :	*emulsifier R* (%)	1	2		
Response :	(*qualitative*)				
y_1 :	*hazy*?				
y_2 :	*adhere*?				
y_3 :	grease on top of film ?				
y_4 :	grease under film ?				
y_5 :	dull, adjusted pH				
y_6 :	dull, original pH				

A standard FD in such a case is 2^5 design with $n \geq 32$ experiments.
But it is done by a fractional factorial design in n=8 runs. The data are as follows.

run #	1	2	3	$4 = 123$	$5 = 23$	y_1	y_2	y_3	y_4	y_5	y_6
	A	B	C	E	D						
1	$-$	$-$	$-$	$-$	$+$	*no*	*no*	*yes*	*no*	*slightly*	*yes*
2	$+$	$-$	$-$	$+$	$+$	*no*	*yes*	*yes*	*yes*	*s*	*yes*
3	$-$	$+$	$-$	$+$	$-$	*no*	*no*	*no*	*yes*	*no*	*no*
4	$+$	$+$	$-$	$-$	$-$	*no*	*yes*	*no*	*no*	*no*	*no*
5	$-$	$-$	$+$	$+$	$-$	*yes*	*no*	*no*	*yes*	*no*	*s-no*
6	$+$	$-$	$+$	$-$	$-$	*yes*	*yes*	*no*	*no*	*no*	*no*
7	$-$	$+$	$+$	$-$	$+$	*yes*	*no*	*yes*	*no*	*s*	*yes*
8	$+$	$+$	$+$	$+$	$+$	*yes*	*yes*	*yes*	*yes*	*s*	*yes*
res	$y_2 \uparrow$		$y_1 \uparrow$	$y_4 \uparrow$	$y_3, y_5, y_6 \uparrow$	C	A	D	E	D	D

It is called a 2^{5-2} design, or a quarter fraction of the full 2^5 design.
The results in the last row in the table are the purpose of the experiment:
which level yields the desired result.
The set-ups of the two quantitative levels are based on the experience of engineers.
The values are not uniquely determined, at least in this experiment.

Notice: This is different from blocking, where (**12**,**13**) is used.
Justification: (High order) interactions are often negligible.
Can we choose **4=123** and **5=12** ?

Main difference between blocking factor and fractional FD.

The former tries to avoid confounding with other effects.

The latter focuses on main effects assuming higher order interaction is not significant.

Why FD and FFD ? There are two types of covariates: categorical and numerical.

Categorical variables are naturally factorial.

Numerical variable can also be specified as factor variables as in §6.1. The purpose in FD is to find the tendency for desired results, not necessarily to find the linear relation. The FFD is try to use less experiments to find the tendency of more factors.

6.1.2. Homework. 1. Discuss a statistician what are the possible randomization steps for the experiment in §6.1 using the fractional FD. Notice that the raw materials include films, catalysts, additives, emulsifiers, among others.

6.2. Stability of new product, 4 factors in 8 runs (a half fractional FD).
A chemist in a lab was trying to formulate a household liquid product using a new process.
The product had some nice properties but he had not found the value of factors to achieve the desired value y of stability at 25 or above.
So he carried out another experiment as follows.

Factors :		$-$	$+$
A :	*acid concentration*	20%	30%
B :	*catalyst concentration*	1%	2%
C :	*temperature*	100	150
D :	*monomer concentration* :	25%	50%

Let $D = 4 = 123$. The data according to Yates order are

y=c(20,14,17,10,19,13,14,10).

It was disappointed that the value $y \geq 25$ is not achieved in any of the cases. However,
the experiment provided a trend for it.
The main effects:

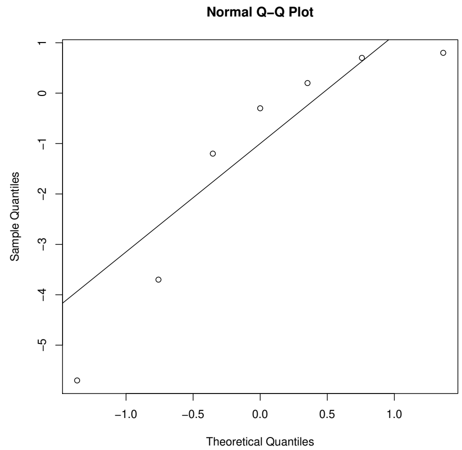
intercept	A	B	C	D			
19.25	-5.75	-3.75	-1.25	0.75	0.25	0.75	-0.25
$\hat{\beta}_0$	$\hat{\beta}_A$	$\hat{\beta}_B$	$\hat{\beta}_C$	$\hat{\beta}_D$			

It occurs that the effect of D (or maybe C) is negligible,
> x=c(0.75,0.25,0.75,-0.25)
> round(2.33*sqrt(mean(x*x)),2) 1.64 1.96 2.33 2.58

[1] 1.30

or as displayed in the qqnorm() of the 7 effects.
The simplified relation becomes

$\hat{Y}_i = \hat{\beta}_o - 5.75\mathbf{1}(factor(A_i) = 1) - 3.75\mathbf{1}(factor(B_i) = 1)$ if $A_i \in \{20, 30\}$ & $B_i \in \{1, 2\}$.



run #	1	2	3	4 = 123	*y or R*
	A	*B*	*C*	*D*	
5	−	−	+	+	19
1	−	−	−	−	20
6	+	−	+	−	13
2	+	−	−	+	14
7	−	+	+	−	14
3	−	+	−	+	17
4	+	+	−	−	10
8	+	+	+	+	10

Ignoring columns C and D, the first two columns become a replicate 2^2 factorial design, as in Figure 6.1 (see Textbook p.238).
It seems from Figure 6.1 that one may simplify the relation as

$$y \approx \underbrace{\frac{20+19}{2}}_{how?} \underbrace{-5.75}_{where?} \underbrace{\frac{A-20}{10}}_{how?} -3.75(B-1) \text{ (in the unit of \%)} \quad (1)$$

$$y \approx \overline{y} - \frac{5.75}{2}\frac{A-25}{5} - \frac{3.75}{2}\frac{B-1.5}{0.5} \text{ in control.sum with } \overline{y} = 14.625 \quad (2)$$

Factors :		−	+
A :	*acid concentration*	20	30
B :	*catalyst concentration*	1	2

Eq. (1) is a *guess*, not from the LSE.
Roughly speaking, from Fig. 6.1,
if A=15 (%) and $B = 0.5$ (%), then Eq. (1) yields
$y = \frac{20+19}{2} - (5.75 + 3.75)(-0.5) = 24.25$.
Thus the stability value $y = 25$ can be reached if
acid concentration is set less than 15% and
catalyst concentration is set less than 0.5%.
The LSE:
> u=lm(y∼ a+b)$coef

$(Intercept)$	$a1$	$b1$
19.37	−5.75	−3.75
$\neq 19.5$	(*see Eq.*(1)	

> v=c(1, 0, −0.5, −0.6)
> u[1]+v*(u[2]+u[3]) $\qquad 19.37 - 5.75\frac{A-20}{10} - 3.75(B-1) \geq 25?$
[1] 9.875 19.375 24.125 25.075
$\frac{A-20}{10} = -0.6$ => A=14
$B - 1 = -0.6$ => $B = 0.4$
The example illustrates:
1. How a fractional design was used for screening purposes to isolate 2 factors out of 4.
2. How a desirable direction in which to carry out further experiment was discovered.

6.2.2. Homework. What is the set up for the further experiment to serve the chemist's original plan ? How many experiments would you suggest ? Why ?

Factors :		$-$	$+$
A :	*acid concentration*	20%	30%
B :	*catalyst concentration*	1%	2%
C :	*temperature*	100	150
D :	*monomer concentration* :	25%	50%

Remark related to the last homework. EQ.(2) is the equation for the contour plane.

$$y \approx 14.63 - \frac{5.75}{2}\frac{A-25}{5} - \frac{3.75}{2}\frac{B-1.5}{0.5} \qquad \text{in control.sum} \tag{2}$$

$(A, B) = (14, 0.4)$ is a point on the contour plane

$$25.075 = 14.625 - \frac{5.75}{2}\frac{A-25}{5} - \frac{3.75}{2}\frac{B-1.5}{0.5}$$

It is a straght line on the AB plane (with A-axis and B-axis). $(A, B) = (20, 1.5)$ is a point on the contour plane $0 = \frac{5.75}{2}\frac{A-25}{5} + \frac{3.75}{2}\frac{B-1.5}{0.5}$ or $B = 1.5 - \frac{5.75}{3.75}\frac{A-25}{10}$ (from

$$14.625 = 14.625 - \frac{5.75}{2}\frac{A-25}{5} - \frac{3.75}{2}\frac{B-1.5}{0.5}).$$

It is another straght line on the AB plane.

6.3. Another Half-fraction FD example. The modification of a bearing.
A manufacturer of bearing tries to improve their product of bearing.
A project team conjectured that they might need to modify
4 factors:

A: a particular characteristic of the manufacturing process for the balls in the bearing,

B: the cage design,

C: the type of grease,

D: the amount of grease.

A 2^{4-1} half fractional FD was carried out with D corresponding to abc.

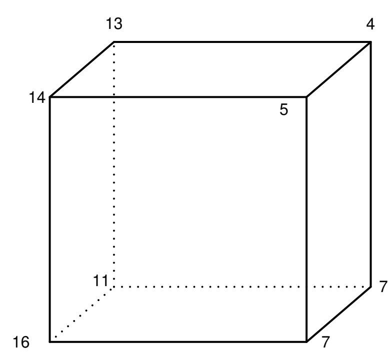
The results are

Yates run #	1	2	3	4	5	6	7	8	
failure rate y	16	7	14	5	11	7	13	4	%
		abc							

R yields effects:

A	B	C	D	ab	ac	bc
-7.7	-1.2	-1.7	-1.2	-1.3	1.2	0.7

The cube plots is



By experience, they suspected that interactions are inert (has little effect), then it reduces to a 2^3 or duplicated 2^2 design (see Figure 6.2).
From this half fraction FD design experiment,

they found the major factors A and C to improve their bearing.

$y = 14.37 - 7.75 factor(a) - 1.75 factor(c)$
Both should be set at the "+" level (**why ?**)

Remark. The discussion on homework is in 556.2.pdf

6.4. The anatomy of the half fraction.

A complete 2^4 factorial design can estimate 16 independent quantities:
average, 4 main effects: A, B, C, D, and the interaction effects: AB, ..., ABCD.
A half fraction design using ABC to accommodate factor D.
Thus the main effect of D cannot distinguished from ABC interaction.
The main effect D is really
$l_D = \frac{1}{4}(-1,1,1,-1,1,-1,-1,1)\cdot(y_1,...,y_8)$ (abc%*%y/4).
Thus l_D is really estimate the sum of the effects D and ABC, denoted by
$l_D \to D + ABC$.
The reason we said l_D is the main effect of D is that the 3-factor interactions are **often** negligible. For instance, in the example of §6.3, knowing ABC is inert by experience,

$effects$	$estimates$	effects assuming D	and its interactions are inert
A	-7.7	A	
B	-1.2	B	
C	-1.7	C	
D	-1.2	$D(\&noise)$	(as ABCD=ABC(ABC)=I).
$AB+CD$	-1.3	AB	
$AC+BD$	1.2	AC	
$AD+BC$	0.7	BC	

The effects D and ABC are said to be **confounded**.
ABC is called an **alias** of D.
Under this design we also have
$l_A \to A + BCD$, $l_{AB} \to AB + CD$
$l_B \to B + ACD$, $l_{AC} \to AC + BD$
$l_C \to C + ABD$, $l_{BC} \to BC + AD$.

Table 1

Why ?

Recall AB represents interaction of A and B,
corresponding to their coordinates multiplying separately.
The AA corresponds to a vector with coordinates being all +1,
denoted by
I=AA=BB=CC=DD
Notice under the fractional factorial design
D=ABC (called the **generating relation**).

$$\text{I=DD=ABCD, } \overbrace{\text{A=BCD, B=ACD, C=ABD, AB=CD, AC=BD, AD=BC.}}^{combinations\ of\ ABCD\ in\ 2\ groups}$$

I=ABCD is also called the **generating relation** of the fractional FD.

The 2^{4-1} fractional factorial design used here is said to be of **resolution** 4, as the generating relation is
I=ABCD with 4 letters,
and no other products of less than **4 distinct** letters lead to I.
It is also denoted by 2^{4-1}_{IV} or "2 to the four minus 1, resolution four".

Remark. 2^{k-1} FFD may not be resolution k. For instance, if the generating relation is
D=AB, (2^{4-1})
then
I=ABD.
Also A=BD, B=AD, $\underbrace{\text{AC=BCD, BC=ACD, CD=ABC, I=ABD}}_{3\ letters\ on\ the\ right}$, C=$\underbrace{\text{ABCD}}_{4\ letters}$.

No other products of less than 3 distinct letters lead to I.
The half fraction FD has a resolution 3, and is a 2^{4-1}_{III} (not 2^{4-1}_{IV}).
Thus 3-factor interaction may be confounded with 2-factor interaction. (AC=BCD)
If D=ABC, then 2-factor interaction confounded with a 2-factor interaction (see Table 1 ↑).

Projectivity. Look at the next example of 4-run design in factors A, B, C:

	a	b	ab
run #	A	B	C
1	−	−	+
2	+	−	−
3	−	+	−
4	+	+	+

a		ab
A	*run #*	C
−	3	−
+	2	−
−	1	+
+	4	+
1		2

ab		b
C	*run #*	B
−	2	−
+	1	−
−	3	+
+	4	+
1		2

or (C,B) as (1,2).

The design is a 2^{3-1}_{III}, as ABC=I.
If you drop one of the factor, you obtain a 2^2 FD in the remaining 2 factors.
It is said of projectivity P=2. The 2^2 FD in the next table can be viewed as 2^{3-1}_{III}:

Yates run #	A	B	C
1	−	−	−
4	+	+	−
6	+	−	+
7	−	+	+

(See also Fig. 6.3 (p.244))

In general,

P=resolution of the design−1.

Denoted by

P=R−1.

Remark 6.3. See the previous example of 2^{4-1}_{IV} in Figure 6.2 with generating relation D=ABC. Then P=4 − 1=3. The geometric interpretation is clear from the figure, as well as the next table:

run #	1	2	3	$4=123$	*y or R*	*run #*	1	2	3	$4=123$	*y or R*	
	A	B	C	D			A	B	C	D		
	a	b	c				a	b		c		
1	−	−	−	−	20	1	−	−	−	−	20	*front*
2	+	−	−	+	14	6	+	−	+	−	13	
3	−	+	−	+	17	7	−	+	+	−	14	
4	+	+	−	−	10	4	+	+	−	−	10	
5	−	−	+	+	19	5	−	−	+	+	19	*back*
6	+	−	+	−	13	2	+	−	−	+	14	
7	−	+	+	−	14	3	−	+	−	+	17	
8	+	+	+	+	10	8	+	+	+	+	10	

run #	1	2	3	$4=123$	*y or R*		*run #*	1	2	3	$4=123$	*y or R*	
	A	B	C	D				A	B	C	D		
	a		b	c					a	b	c		
1	−	−	−	−	20	*bottom*	1	−	−	−	−	20	L
4	+	+	−	−	10		4	+	+	−	−	10	
7	−	+	+	−	14		6	+	−	+	−	13	
6	+	−	+	−	13		7	−	+	+	−	14	
3	−	+	−	+	17	*top*	2	+	−	−	+	14	R
2	+	−	−	+	14		3	−	+	−	+	17	
5	−	−	+	+	19		5	−	−	+	+	19	
8	+	+	+	+	10		8	+	+	+	+	10	

How to understand Figure 6.2 ? (see explanation figure on blackboard as well).

On the other hand, if the generating relation is D=AB, then P=3−1=2.
Dropping one variable does not **always** reduce to a 2^3 FD (see Table 3 below).

run #	*a*	*b*	*c*	*ab*
	A	B	C	D
	1	2	3	
1	−	−	−	+
2	+	−	−	−
3	−	+	−	−
4	+	+	−	+
5	−	−	+	+
6	+	−	+	−
7	−	+	+	−
8	+	+	+	+

− − − ?

run #	*a*	*b*	*c*	*ab*
	A	B	C	D
		1	2	3
2	+	−	−	−
3	−	+	−	−
6	+	−	+	−
7	−	+	+	−
1	−	−	−	+
4	+	+	−	+
5	−	−	+	+
8	+	+	+	+

Yes

run #	*a*	*b*	*ab*
	A	B	D
1	−	−	+
2	+	−	−
3	−	+	−
4	+	+	+
5	−	−	+
6	+	−	−
7	−	+	−
8	+	+	+

?

Dropping two variables <u>*does*</u> reduce to a (replicated) 2^2 FD.

Keep D:

run #	*a*	*b*	*c*	*ab*
	A	B	C	D
	1			2
3	−	+	−	−
2	+	−	−	−
1	−	−	−	+
4	+	+	−	+
7	−	+	+	−
6	+	−	+	−
5	−	−	+	+
8	+	+	+	+

run #	*a*	*b*	*c*	*ab*
	A	B	C	D
		1		2
2	+	−	−	−
3	−	+	−	−
1	−	−	−	+
4	+	+	−	+
6	+	−	+	−
7	−	+	+	−
5	−	−	+	+
8	+	+	+	+

run #	*a*	*b*	*c*	*ab*
	A	B	C	D
			1	2
2	+	−	−	−
6	+	−	+	−
1	−	−	−	+
5	−	−	+	+
3	−	+	−	−
7	−	+	+	−
4	+	+	−	+
8	+	+	+	+

Otherwise ?

6.5. The 2^{7-4}_{III} design: a bicycle example.

7 Factors:

- A: seat (up,down),
- B: dynamo (generator) (off, on),
- C: handlebars (up, down),
- D: gear (low, median),
- E: raincoat (on, off),
- F: breakfast (yes, no),
- G: tires (hard, soft),

Response: y, climb hill in seconds.

run #	*a*	*b*	*c*	*ab*	*ac*	*bc*	*abc*	y
	A	B	C	D	E	F	G	
1	−	−	−	+	+	+	−	69
2	+	−	−	−	−	+	+	52
3	−	+	−	−	+	−	+	60
4	+	+	−	+	−	−	−	83
5	−	−	+	+	−	−	+	71
6	+	−	+	−	+	−	−	50
7	−	+	+	−	−	+	−	59
8	+	+	+	+	+	+	+	88

Table 6.4

Estimates of effects:

- $l_A = 3.5$ seat (up,down),
- $l_B = 12$ dynamo (generator) (off, on),
- $l_C = 2.5$ handlebars (up, down), **typo in the textbook**
- $l_D = 22.5$ gear (low, median),
- $l_E = 1$ raincoat (on, off),
- $l_F = 0.5$ breakfast (yes, no),
- $l_G = 1.0$ tires (hard, soft),

Average=66.5

From previous experiments on the bicycle example, an estimate of the SD of repeated runs is 3. So the SE of the estimated effects is

$$\sqrt{\frac{3^2}{4}+\frac{3^2}{4}}=2.1.$$

Thus there are only two factors which are distinguishable from noise. They are dynamo B and gear D. Or roughly, one can determine by

```
> z=c( 3.5, 12.0, 2.5, 22.5, 1.0, 0.5, 1.0)
> qqnorm(z)
> qqline(z)
```

Or

```
> stem(z)
```

The decimal point is 1 digit(s) to the right of the |

```
0 | 11133
0 |
1 | 2
1 |
2 | 2
```

This fractional design can reduce the number of runs and present a replicated 2^2 FD for factors B and D, which is not very clear before the experiment.

Notice that (ignoring interactions of order 3+)

$l_I \to average.$

$l_A \to A+BD+CE+FG$

$l_B \to B+AD+CF+EG,$

$l_C \to C+AE+BF+DG,$

$l_D \to D+AB+EF+CG,$

$l_E \to E+AC+DF+BG,$

$l_F \to F+BC+DE+AG,$

$l_G \to G+CD+BE+AF,$

How are they obtained ?

The Defining Relations. The 4 generators

D=AB, E=AC, F=BC, G=ABC

yield 4 defining relations:

(1) $\binom{4}{1}=4$ I=ABD=ACE=BCF=ABCG.

which lead to A=BD=CE=ABCF=BCG (not l_A).

To find all defining relations and to find all aliases, we need to add all words,

there are $\sum_{i=1}^4 \binom{4}{i}=15$ defining relations: (from ABD=ACE=BCF=ABCG (=I)),

(2) $\binom{4}{2}=6$: (from ABD=ACE=BCF=ABCG (=I)),

I=(ABD)(ACE)=BCDE

I=(ABD)(BCF)=ACDF

I=(ABD)(ABCG)=CDG

I=(ACE)(BCF)=ABEF

I=(ACE)(ABCG)=BEG

I=(BCF)(ABCG)=AFG

(3) $\binom{4}{3}$ from ABD=ACE=BCF=ABCG(=I),

I=DEF (=(ABD)(ACE)(BCF))

I=ADEG (=(ABD)(ACE)(ABCG))

I=BDFG (=(ABD)(BCF)(ABCG))

I=CEFG (=(ACE)(BCF)(ABCG))

(4) $\binom{4}{4}$ from $\underbrace{ABD=ACE=BCF=ABCG}_{I=ABCDEFG}$(=I)

I=ABCDEFG (only choose 2 words from the 4 cases).

Remark. (1) In each of the 15 words, the letters are all distinct.

(2) Now it is clear why

$l_A \to A + BD + CE + FG$
due to I=ABD=ACE =AFG =ABCG=ACDF=ABEF =ADEG=ABCEDFG

The shortest "word" in the 15 defining relations among (1) – (4) is 3.
It is called a 2^{7-4}_{III} FD.
Remark. This is different from the definition of resolution in half FD. But latter can be rephraced as this new one.
The 2^{7-4}_{III} can be viewed as a replicate 2^2 FD.

	a	b	c	ab	ac	bc	abc	y
new run # in (D, B)	A	B	C	D	E	F	G	
2	−	−	−	+	+	+	−	69
1	+	−	−	−	−	+	+	52
3	−	+	−	−	+	−	+	60
4	+	+	−	+	−	−	−	83
6	−	−	+	+	−	−	+	71
5	+	−	+	−	+	−	−	50
7	−	+	+	−	−	+	−	59
8	+	+	+	+	+	+	+	88
new order		2		1				

Median	69	71	– – – – – –	83	88
Gear, D		|			|
low	52	50	– – – – – –	60	59
		off	*generator* , *B*		*on*

6.6. Eight-run designs Table 6.4 ignoring the response y can be used to produce the 2^3, the 2^{4-1}_{IV}, or the 2^{7-4}_{III} designs.
The latter two (not the first one) are called **nodal** designs, in the sense that
for a given number of runs, the nodal design includes the largest number of factors at a given resolution.
The resolution R = the smallest # of distinct letters in the product.
There are
7 factors in 2^{7-4}_{III} design (where R=3).
4 factors in 2^{4-1}_{IV} design (where R=4).

There are 3 2^{4-1}_{III}:

a	b	c	ab	ac	bc	abc
A	B	C	D			
A	B	C		E		
A	B	C			F	

Remark. It does not matter whether one calls the factors A, B, C, D, or A, B, C, E.
These 3 2^{4-1}_{III} generating relations are I=ABD, I=ACE and I=BCF, respectively.
R=3 **Why ?**

2^{4-1}_{IV}

a	b	c	ab	ac	bc	abc
A	B	C				G

The generating relation I=ABCG. R=4 **Why ?**
Between 2^{7-4}_{III} and 2^{4-1}_{IV}, we have 8-run 2^{5-2}_{III} and 2^{6-3}_{III} designs, but they are not nodal designs.
For example, if one considers a 5-factor design 2^{5-2}, there are 6 of them:

run #	*a*	*b*	*c*	*ab*	*ac*	*bc*	*abc*
	A	B	C	D	E		
	A	B	C	D		F	
	A	B	C		E	F	
	A	B	C	D			G
	A	B	C		E		G
	A	B	C			F	G
1	−	−	−	+	+	+	−
2	+	−	−	−	−	+	+
3	−	+	−	−	+	−	+
4	+	+	−	+	−	−	−
5	−	−	+	+	−	−	+
6	+	−	+	−	+	−	−
7	−	+	+	−	−	+	−
8	+	+	+	+	+	+	+

Table 6.6.

Note that I=ABCG=BCF in the last row of Table 6.6.

R=4 or R=3 in Table 6.6 ?

In Table 6.6, each of the 6 FD has either ABD=I, or I=ACE, or I=BCF and no product of two distinct letters = I, thus its resolution R=3.

Do we have 2^{8-5}_{II} design ?

run #	*a*	*b*	*c*	*ab*	*ac*	*bc*	*abc*	
	A	B	C	D	E	F	G	Then resolution = 3 or 2 ?
							H	

Comments:

The fractional FG is used to screen out significant factors from a larger group of factors.

It is hopeful to reduce to 2 factors by 2^{7-4}_{III}

It is hopeful to reduce to 3 factors by 2^{4-1}_{IV}

2^{8-5}_{II} can not even reduce to 1 factor, as G and H cannot be distinguished.

6.7. Using Table 6.6. An illustration.

$$\begin{pmatrix} & a & b & c & ab & ac & bc & abc & Projectivity & P \\ 2^3 & A & B & C & & & & & & \\ 2^{4-1}_{IV} & A & B & C & & & & G & 3 & \\ 2^{7-4}_{III} & A & B & C & D & E & F & G & 2 & \end{pmatrix}$$

For 2^{4-1}_{IV} design, ignoring 3-factor interaction,

$l_A \to A,$

$l_B \to B,$

$l_C \to C,$

$l_D \to AB + CG,$ (due to ABCG=I)

$l_E \to AC + BG,$

$l_F \to BC + AG,$

$l_G \to G.$

For 2^{7-4}_{III} design, ignoring 3-factor interaction,

$l_A \to A + BD + CE + FG,$

$l_B \to B + AD + CF + EG,$

$l_C \to C + AE + BF + DG,$

$l_D \to D + AB + EF + CG,$ (as discussed in $\S 6\dot{5}$)

$l_E \to E + AC + DF + BG,$

$l_F \to F + BC + DE + AG,$

$l_G \to G + CD + BE + AF.$

An Experiment. In the early stages of a lab experiment, 5 factors are given as follows.

$factors$		-1	$+1$	
1 :	$concentration\ of \gamma$	94	96%	
2 :	$proportion\ of\ \gamma\ to\ \alpha$	3.85	4.15	mol/mol
3 :	$amount\ of\ solvent$	280	310	cm^3
4 :	$proportion\ of\ \beta\ to\ \alpha$	3.5	5.5	mol/mol
5 :	$reaction\ time$	2	4	hr

The best conditions known at that time were thought to be far from optimal
and the main effects were believed to be dominant,
but the interaction AC were thought to be active and needs to be avoid.
So the column corresponds to AC needs to be dropped in section column from Table 6.6.

One way is to select columns A, B, C, D, G: G should be selected as abc is of higher order of interaction. That is, the 5th factor is not denoted by E, but by G.

$run\ \#$	a	b	c	ab	ac	bc	abc	
	A	B	C	D			G	y
1	$-$	$-$	$-$	$+$	$+$	$+$	$-$	77.1
2	$+$	$-$	$-$	$-$	$-$	$+$	$+$	68.9
3	$-$	$+$	$-$	$-$	$+$	$-$	$+$	75.5
4	$+$	$+$	$-$	$+$	$-$	$-$	$-$	72.5
5	$-$	$-$	$+$	$+$	$-$	$-$	$+$	67.9
6	$+$	$-$	$+$	$-$	$+$	$-$	$-$	68.5
7	$-$	$+$	$+$	$-$	$-$	$+$	$-$	71.5
8	$+$	$+$	$+$	$+$	$+$	$+$	$+$	63.7

The estimates are

l_A	l_B	l_C	l_D	l_E	l_F	l_G	
-4.5	0.2	-5.6	-0.8	1.0	-0.8	-3.4	
s		s				s	$why?$

Do we have factors E and F ?

l_F can be viewed as a noise, then so does l_E in view of l_F.
The optimal yields might be obtained by moving in a direction such that the concentration of γ (A), the amount of solvent (C) and the reaction time (G) were all reduced. A serious of further experiments lead to a yield of 84% (v.s. 77.1%) for the chemical manufacturing process.

6.8. Sign switching, foldover and sequential assembly. Further runs may needed when fractional designs yield ambiguity.

A strategy is **Foldover**:
A single column foldover:

multiply one selected column by -1, or switching sign of the column.
An example of Bicycle experiment, where B and D are significant effects.

run #	a	b	c	ab	ac	bc	abc	y
	A	B	C	D	E	F	G	
1	−	−	−	+	+	+	−	69
2	+	−	−	−	−	+	+	52
3	−	+	−	−	+	−	+	60
4	+	+	−	+	−	−	−	83
5	−	−	+	+	−	−	+	71
6	+	−	+	−	+	−	−	50
7	−	+	+	−	−	+	−	59
8	+	+	+	+	+	+	+	88
switch				$\times(-1)$				
9	−	−	−	−	+	+	−	47
10	+	−	−	+	−	+	+	74
11	−	+	−	+	+	−	+	84
12	+	+	−	−	−	−	−	62
13	−	−	+	−	−	−	+	53
14	+	−	+	+	+	−	−	78
15	−	+	+	+	−	+	−	87
16	+	+	+	−	+	+	+	60

Effect: These 16 runs provide unaliased estimates of the main effect D and all two-factor interactions involving D.

Reason:

$1st\ 2^3:$	l_A	l_B	l_C	l_D	l_E	l_F	l_G	l_I
	3.5	12	1	22.5	0.5	1.0	2.5	66.5
$2nd\ 2^3:$	l'_A	l'_B	l'_C	l'_D	l'_E	l'_F	l'_G	l'_I
	0.7	10.2	2.7	25.2	1.7	2.2	−0.7	68.125

The first 8 runs yield (ignoring higher order interactions):

$l_A \to A + BD + CE + FG,$
$l_B \to B + AD + CF + EG,$
$l_C \to C + AE + BF + DG,$
$l_D \to D + AB + EF + CG,$
$l_E \to E + AC + DF + BG,$
$l_F \to F + BC + DE + AG,$
$l_G \to G + CD + BE + AF,$

The second 8 runs yield (ignoring higher order interactions):

$l'_A \to A - BD + CE + FG,$
$l'_B \to B - AD + CF + EG,$
$l'_C \to C + AE + BF - DG,$
$l'_D \to D - AB - EF - CG,$
$l'_E \to E + AC - DF + BG,$
$l'_F \to F + BC - DE + AG,$
$l'_G \to G - CD + BE + AF,$

Then ignoring three or high order interactions,

$0.5(l_A + l'_A) = 2.1 \to A + CE + FG,$
$0.5(l_B + l'_B) = 11.1 \to B + CF + EG,$
$0.5(l_C + l'_C) = 1.9 \to C + AE + BF,$
$0.5(l_D + l'_D) = 23.9 \to D,$
$0.5(l_E + l'_E) = -0.6 \to E + AC + BG,$
$0.5(l_F + l'_F) = 1.6 \to F + BC + AG,$
$0.5(l_G + l'_G) = 0.9 \to G + BE + AF,$

In fact, $0.5(l_A + l'_A) = 2.1 \to A + CE + FG + BCG + BEF$, as

$$\begin{aligned}0.5(l_A + l'_A) \to &[A + BD + CE + FG + BCG + BEF + CDF + DEG + BCDEFG\\ &+(A - BD + CE + FG + BCG + BEF - CDF - DEG - BCDEFG)]/2\\ &= A + CE + FG + BCG + BEF\end{aligned}$$

Recall for 2^{7-4}_{III} design, the 15 ($= \sum_{i=1}^{4}\binom{4}{i}$) defining relations are
I=ABD =CDG =DEF =ACE=BCF =BEG=AFG

=BCDE =ACDF =ADEG =BDFG =ABCG =ABEF =CEFG
=ABCDEFG

Foldover yields

I=−ABD =−CDG =−DEF =ACE=BCF =BEG=AFG
=−BCDE =−ACDF =−ADEG =−BDFG =ABCG =ABEF =CEFG
=−ABCDEFG

Average them yields

I=ACE=BCF =BEG =AFG =ABCG =ABEF =CEFG

Moreover,

$0.5(l_A - l'_A) = 1.4 \to BD$,
$0.5(l_B - l'_B) = 0.9 \to AD$,
$0.5(l_C - l'_C) = -0.9 \to DG$,
$0.5(l_D - l'_D) = -1.4 \to AB + EF + CG$,
$0.5(l_E - l'_E) = 1.1 \to DF$,
$0.5(l_F - l'_F) = -0.6 \to DE$,
$0.5(l_G - l'_G) = 1.6 \to CD$, as

$$
\begin{aligned}
0.5(l_A - l'_A) \to & [A + BD + CE + FG + BCG + BEF + CDF + DEG + BCDEFG \\
& -(A - BD + CE + FG + BCG + BEF - CDF - DEG - BCDEFG)]/2 \\
& = BD + CDF + DEG + BCDEFG \\
(l_D - l'_D)/2 \to & [D + AB + EF + CG + BCE + ACF + AEG + BFG + ABCEFG \\
& +(-D + AB + EF + CG + BCE + ACF + AEG + BFG + ABCEFG)]/2 \\
& = [AB + EF + CG + BCE + ACF + AEG + BFG + ABCEFG \\
(l_D + l'_D)/2 \to & [D + AB + EF + CG + BCE + ACF + AEG + BFG + ABCEFG \\
& -(-D + AB + EF + CG + BCE + ACF + AEG + BFG + ABCEFG)]/2 \\
& = D
\end{aligned}
$$

Notice that now D is not aliased with any 2 or 3-factor interaction ...
The column D foldover "de-alias" the main effect D and all its interaction with other effects.

So, $0.5(l_I + l'_I) = 67.3 \to$ average,
$0.5(l_I - l'_I) = -1.6 \to$ block effect (**which blocks ?**)

How to implement in R ?

```
> y=c(69, 52, 60, 83, 71, 50, 59, 88, 47, 74, 84, 62, 53, 78, 87, 60)
> a=rep(c(-1,1),4)
> b=rep(c(-1,-1,1,1),2)
> c=rep(-1,4)
> c=c(c,-c)
> D=a*b
> E=a*c
> F=b*c
> G=a*F
>z=lm(y[1:8]~a*b*c)$coef
>(z=c(z[1],z[2:8]*2))
    66.5 3.5 12.0 1.0 22.5 0.5 1.0 2.5
>D=−D
>x=lm(y[9:16]~a+b+c+D+E+F+G)$coef
>(x=c(x[1],x[2:8]*2))
    68.125 0.750 10.250 2.750 25.250 −1.750 −2.250 −0.750
> (z+x)/2
    67.3125 2.1250 11.1250 1.8750 23.8750 −0.6250 −0.6250 0.8750        (1)

> (z−x)/2
    1.375 0.875 −0.875 −1.375 1.125 1.625 1.625 −0.8125 1.375 0.875 −0.875 −1.375
```

1.125 1.625 1.625

> a=c(a,a)
> b=c(b,b)
> c=c(c,c)
> D=c(−D,D)   why not C(D,−D) ?
> E=c(E,E)
> F=c(F,F)
> G=c(G,G)
> lm(y∼a+b+c+D+E+F+G)$coef[2:8]*2

2.125 11.125 1.875 $\underline{23.875}$ −0.625 −0.625 0.875

lm(y∼factor(a)+factor(b)+factor(c)+factor(D)+factor(E)+factor(F)+factor(G))$coef[2:8]

2.125 11.125 1.875 $\underline{23.875}$ −0.625 −0.625 0.875

Can we apply it to other column ?

The foldover is part of sequential process of scientific learning,
in contrast to the "one-shot" experiment we have learned so far.
In the previous example, the first 8 run is the first shot.
If we stop, then it is a one-shot experiment.
In experimental design, we have initial informed guesses:
- what factors to include ?
- what response to measure ?
- where to locate the experimental region ?
- by how much to vary the factors ?
- after the data are available, how to proceed ?

We do not expect to find answers to all the question in one-shot.
We can try smaller experiment to reduce the unknown possibilities gradually by making second guesses. Foldover is one of such strategy.

Does the total of the 16 runs consist of a 2^{7-3} FD ?   Where to find D ?

run #	*a*	*b*	*c*	*ab*	*ac*	*bc*	*abc*	*y*
	A	*B*	*C*	*D*	*E*	*F*	*G*	
1	−	−	−	+	+	+	−	69
2	+	−	−	−	−	+	+	52
3	−	+	−	−	+	−	+	60
4	+	+	−	+	−	−	−	83
5	−	−	+	+	−	−	+	71
6	+	−	+	−	+	−	−	50
7	−	+	+	−	−	+	−	59
8	+	+	+	+	+	+	+	88
switch				×(−1)				
9	−	−	−	−	+	+	−	47
10	+	−	−	+	−	+	+	74
11	−	+	−	+	+	−	+	84
12	+	+	−	−	−	−	−	62
13	−	−	+	−	−	−	+	53
14	+	−	+	+	+	−	−	78
15	−	+	+	+	−	+	−	87
16	+	+	+	−	+	+	+	60

16 runs FFD

run #	*a*	*b*	*c*	?	*ac*	*bc*	*abc*	*y*
	A	*B*	*C*	*D*	*E*	*F*	*G*	
9	−	−	−	−	+	+	−	47
2	+	−	−	−	−	+	+	52
3	−	+	−	−	+	−	+	60
12	+	+	−	−	−	−	−	62
13	−	−	+	−	−	−	+	53
6	+	−	+	−	+	−	−	50
7	−	+	+	−	−	+	−	59
16	+	+	+	−	+	+	+	60
1	−	−	−	+	+	+	−	69
10	+	−	−	+	−	+	+	74
11	−	+	−	+	+	−	+	84
4	+	+	−	+	−	−	−	83
5	−	−	+	+	−	−	+	71
14	+	−	+	+	+	−	−	78
15	−	+	+	+	−	+	−	87
8	+	+	+	+	+	+	+	88

2^{7-3} FFD

6.9. Multiple-column foldover. Its effect is that all main effects can be unaliased with two-factor interactions. (A single column (say D) foldover unaliases D with all interactions).

Chemical plants experiment. A number of similar chemical plants in different locations had been operated successfully for years. In a newly constructed plant certain filtration cycle took twice as long as the other plants. In order to find the reason, 7 factors are identified and a 2^{7-4}_{III} fractional design was carried out.

	Factors	$-$	$+$	
$A:$	*water supply*	*town reservoir*	*well*	
$B:$	*raw material*	*on site*	*other*	
$C:$	*temperature*	*low*	*high*	
$D:$	*recycle*	*yes*	*no*	
$E:$	*caustic soda*	*fast*	*slow*	*(kexingna)*
$F:$	*filter cloth*	*new*	*old*	
$G:$	*holdup time*	*low*	*high*	

	a	b	c	ab	ac	bc	abc	y
run #	A	B	C	D	E	F	G	
1	$-$	$-$	$-$	$+$	$+$	$+$	$-$	68.4
2	$+$	$-$	$-$	$-$	$-$	$+$	$+$	77.7
3	$-$	$+$	$-$	$-$	$+$	$-$	$+$	66.4
4	$+$	$+$	$-$	$+$	$-$	$-$	$-$	81.0
5	$-$	$-$	$+$	$+$	$-$	$-$	$+$	78.6
6	$+$	$-$	$+$	$-$	$+$	$-$	$-$	41.2
7	$-$	$+$	$+$	$-$	$-$	$+$	$-$	68.7
8	$+$	$+$	$+$	$+$	$+$	$+$	$+$	38.7

foldover	$-a$	$-b$	$-c$	$-ab$	$-ac$	$-bc$	$-abc$	y
run #	A	B	C	D	E	F	G	
9	$+$	$+$	$+$	$-$	$-$	$-$	$+$	66.7
10	$-$	$+$	$+$	$+$	$+$	$-$	$-$	65.0
11	$+$	$-$	$+$	$+$	$-$	$+$	$-$	86.4
12	$-$	$-$	$+$	$-$	$+$	$+$	$+$	61.9
13	$+$	$+$	$-$	$-$	$+$	$+$	$-$	47.8
14	$-$	$+$	$-$	$+$	$-$	$+$	$+$	59.0
15	$+$	$-$	$-$	$+$	$+$	$-$	$+$	42.6
16	$-$	$-$	$-$	$-$	$-$	$-$	$-$	67.6

For the 2^{7-4}_{III} design, the defining relations:

I= ABD=ACE=BCF=ABCG=BCDE=ACDF=CDG=ABEF=BEG=AFG
=DEF=ADEG=BDFG=CEFG= ABCDEFG.

Estimates:

$l_A = \underline{-10.9} \to A + BD + CE + FG$,
due to I=ABD=ACE =AFG,
and ignoring high order interactions: +BCG +CDF +BEF +DEG +BCDEFG.

(ABCF, ABCDE, ACDG , ABEG , ADEF , ABDFG, ACEFG are **not** defined as interactions of A **from now on** even though it was defined in the textbook originally.

$l_B = -2.8 \to B + AD + CF + EG$,
due to I=ABD=BCF =BEG

$l_C = \underline{-16.6} \to C + AE + BF + DG$,
due to I=ACE=BCF =CDG

$l_D = 3.2 \to D + AB + EF + CG$,
due to I=ABD =CDG=DEF

$l_E = \underline{-22.8} \to E + AC + DF + BG$,
due to I=ACE =BEG=DEF

$l_F = -3.4 \to F + BC + DE + AG$,
due to I=BCF =AFG=DEF

$l_G = 0.5 \to G + CD + BE + AF$,
due to I=CDG=BEG=AFG.

The estimates and summary(y~a*b+b*c+a*c) (why ignore ℓ_G ?) suggest that the causes are factors A, C and E. To further investigate, another 8 runs were made.

The defining relation for the foldover is

I= −ABD=−ACE=−BCF =−CDG =−BEG=−AFG=−DEF = −ABCDEFG.

=ABCG=BCDE=ACDF =ABEF =ADEG=BDFG=CEFG

$l'_A \to A - BD - CE - FG$,

due to I=−ABD=−ACE =−AFG,

and ignoring high order interactions: BCG+CDF+BEF+DEG−BCDEFG

$l'_B \to B - AD - CF - EG$,
$l'_C \to C - AE - BF - DG$,
$l'_D \to D - AB - EF - CG$,
$l'_E \to E - AC - DF - BG$,
$l'_F \to F - BC - DE - AG$,
$l'_G \to G - CD - BE - AF$.

Then

$0.5(l_A + l'_A) = \underline{-6.7} \to A$, ignoring BCG+CDF+BEF+DEG,
$0.5(l_B + l'_B) = -3.9 \to B$,
$0.5(l_C + l'_C) = -0.4 \to C$,
$0.5(l_D + l'_D) = 2.7 \to D$,
$0.5(l_E + l'_E) = \underline{-19.2} \to E$,
$0.5(l_F + l'_F) = -0.1 \to F$,
$0.5(l_G + l'_G) = -4.3 \to G$,
$0.5(l_I + l'_I) = 63.6$.
$0.5(l_A - l'_A) = -4.2 \to BD + CE + FG$,
$0.5(l_B - l'_B) = 1.1 \to AD + CF + EG$,
$0.5(l_C - l'_C) = \underline{-16.2} \to AE + BF + DG$,
$0.5(l_D - l'_D) = 0.5 \to AB + EF + CG$, (no high order interaction, except ABCEFG)
$0.5(l_E - l'_E) = -3.6 \to AC + DF + BG$,
$0.5(l_F - l'_F) = -3.4 \to BC + DE + AG$,
$0.5(l_G - l'_G) = 4.8 \to CD + BE + AF$,
$0.5(l_I - l'_I) = 3.0$.

```
x=c(-6.7,-3.9,-0.4,2.7,-19.2,-0.1,-4.3,0.5,-3.6,1.1,-16.2, 4.8,-3.4,-4.2,3)
round(x,0)
    [1] -7 -4 0 3 -19 0 -4 0 -4 1 -16 5 -3 -4 3
sort(x)
    [1] -19.2 -16.2 -6.7 -4.3 -4.2 -3.9 -3.6 -3.4 -0.4 -0.1 0.5 1.1 2.7 3.0 4.8
stem(x)
    −1 | 96
    −1 |
    −0 | 7
    −0 | 4444300
    +0 | 1133
    +0 | 5
```

Thus it suggests that rather than A, C and E in the first 2^{7-4}_{III} FD, the multiple foldover finds that the main causes are A, E and AE **(why not BF+DG ?)**, C is a noise.

On one hand, the stem and leaf plot suggests that
the main causes are E and AE, but A is also a noise.

On the other hand, the Analysis of Variance Table suggests that A is marginally significant.

Model 1: y ∼ E + I(A * E)
Model 2: y ∼ A + E + I(A * E)

	Res.Df	*RSS*	*Df*	*Sum of Sq*	*F*	*Pr(> F)*	
1	13	645.94					
2	12	467.05	1	178.89	4.5962	0.05322	.

$y = 63.6 - 9.6E - 8.1AE + \epsilon$ (or $\hat{y} = 63.6 - 9.6E - 8.1AE$, where $\overline{y} = 63.6$) ?

$y = 63.6 - 19.2\mathbf{1}(E = 1) - 16.2\mathbf{1}(AE = 1) + \epsilon$?
Improvement can be obtained by E=+1 (caustic soda slow) and A=+1 (well water).

An economic alternative to total foldover.
The foldover took 8 runs to find out the previous conclusion,
but there were simpler ways to do it.
Since the 8-run 2^{7-4}_{III} experiment indicates that A, C and E are **possibly** not inert,
we can consider 2^3 FD with factor A, C and E.

	A	C	E
1	$-$	$-$	$+$
2	$+$	$-$	$-$
3	$-$	$-$	$+$
4	$+$	$-$	$-$
5	$-$	$+$	$-$
6	$+$	$+$	$+$
7	$-$	$+$	$-$
8	$+$	$+$	$+$

run #	*a*	*b*	*c*	*ab*	*ac*	*bc*	*abc*	*y*
	A		C		E			
2	$+$		$-$		$-$			77.7
4	$+$		$-$		$-$			81.0
5	$-$		$+$		$-$			78.6
7	$-$		$+$		$-$			68.7
1	$-$		$-$		$+$			68.4
3	$-$		$-$		$+$			66.4
6	$+$		$+$		$+$			41.2
8	$+$		$+$		$+$			38.7

	1	2	3
?	$-$	$-$	$-$
2, 4	$+$	$-$	$-$
5, 7	$-$	$+$	$-$
?	$+$	$+$	$-$
1, 3	$-$	$-$	$+$
?	$+$	$-$	$+$
?	$-$	$+$	$+$
6, 8	$+$	$+$	$+$

Then runs 1-8 can be treated as 4 pairs of replicates, leading to estimate of SD

$s^2 = \frac{\sum_{i=1}^4 d_i^2/2}{4} = 14.9$ with df 4. ($= \hat{V}(\epsilon)$? or $\hat{V}(effect)$? $d_i = ?$)

Reason: Recall under model $Y = \beta' X + \epsilon$ with $\beta \in \mathcal{R}^p$,

$\frac{1}{n-p} \sum_{j=1}^n (Y_j - \hat{\beta}' X_j)^2$ has df $n - p$.

$d_i^2/2 = \frac{1}{n-p} \sum_{j=1}^n (Y_j - \hat{\beta}' X_j)^2$, under model $Y_j = \mu + \epsilon$, $n = 2$ and $p = ?$

Thus it has df =1.

Add 4 more runs (instead of 8) yields

run #	*a*	*b*	*c*	*ab*	*ac*	*bc*	*abc*	*y*	
	A		C		E				
16	$-$		$-$		$-$			67.6	
2, 4	$+$		$-$		$-$			77.7	81.0
5, 7	$-$		$+$		$-$			78.6	68.7
11	$+$		$+$		$-$			86.4	
1, 3	$-$		$-$		$+$			68.4	66.4
13	$+$		$-$		$+$			47.8	
10	$-$		$+$		$+$			65.0	
6, 8	$+$		$+$		$+$			41.2	38.7

The 12 runs lead to estimates (through $lm(y \sim a * c + a * e + c * e)\$coef[2:7] * 2$)

l_A	l_C	l_E	l_{AC}	l_{AE}	l_{CE}
-5.0	0.7	-21.7	-1.1	-17.3	-5.8
		s		s	?

with $\hat{\sigma}_{effect} = s\sqrt{\frac{1}{6} + \frac{1}{6}} = 2.23$ (why /6 ?) and $t_{0.025,4} \approx 2.8$.

$2.8 \times 2.23 \approx 6.24$.

The effect is not significant if $|effect| < 6.24$. (P-value =0.06).

Thus, the main causes are E and AE, same as the 16-run results. Moreover, A, C and CE are not significant.

Remark. Notice that under the economic alternative design with 12 runs, l_A *etc.* are derived from $lm(y \sim \cdots)$, not from $\overline{y}_+ - \overline{y}_-$, as can be seen from the table.

```
> mean(y[c(2,4,11,13,6,8)])-mean(y[c(16,5,7,1,3,10)])
  [1] -6.983333
```

	l_a	l_c	l_e	l_{ac}	l_{ae}	l_{ce}
$\overline{y}_+ - \overline{y}_-$	-6.98	-5.05	-22.08	-8.35	-17.05	-7.52
$lm(y \sim \cdot)\$coef[-1] * 2$	-5.04	-0.71	-21.71	-1.11	-17.29	-5.83
foldover	-6.7	-0.4	-19.2	-3.6	-16.2	-4.2
	$l_A - l'_A$	$l_C - l'_C$	$l_E - l'_E$	$l_E + l'_E$	$l_C + l'_C$	$l_A + l'_A$

Thus $\overline{y}_+ - \overline{y}_-$ is not applicable in foldover. Moreover, unlike the 2^k FD, given lm(y∼a*c*e) the final model needs to be estimated again. See the next R outputs.

> lm(y∼factor(a) + factor(c) + factor(e) + factor(a*e)+factor(c*e))$coef

$(Intercept)$	$factor(a)1$	$factor(c)1$	$factor(e)1$	$factor(c*e)1$	$factor(a*e)1$
90.39167	−5.03750	0.71250	−22.08333	−5.83750	−17.28750

> x=lm(y∼a*c+a*e+c*e)$coef
> c(x[1],x[−1]*2)

$(Intercept)$	a	c	e	$a:c$	$a:e$	$c:e$
65.89375	−5.03750	0.71250	−21.71250	−1.11250	−17.28750	−5.83750

It turns out to be different from $\overline{y}_+ - \overline{y}_-$.
> w=lm(y∼e+I(a*e)+I(c*e))$coef
> c(w[1],w[-1]*2)

$(Intercept)$	e	$I(a*e)$	$I(c*e)$
65.625000	−22.083333	−17.050000	−7.516667

> V=lm(y∼e +I(a*e))$coef
> c(V[1],V[2:3]*2)

$(Intercept)$	e	$I(a*e)$
65.62500	−22.08333	−17.05000

It turns out to be the same as $\overline{y}_+ - \overline{y}_-$.
> anova(h,z)

Analysis of Variance Table
Model 1: y ∼ e + I(a * e)
Model 2: y ∼ e + I(a * e) + I(c * e)

	$Res.Df$	RSS	Df	$Sum\ of\ Sq$	F	$Pr(>F)$	
1	9	308.3					
2	8	138.8	1	169.5	9.7672	0.01412	$*$

Thus the final model is $\hat{y} = 65.6 - 22.1e - 17.1a*e - 7.5c*e$, where $c*e$ is (marginal ??) significant.

R program for computing effects.

```
y1=c(68.4, 77.7, 66.4, 81.0, 78.6, 41.2, 68.7, 38.7)
a=rep(c(-1,1),4)
b=rep(c(-1,-1,1,1),2)
c=rep(-1,4)
c=c(c,-c)
z=lm(y1~a*b*c)$coef
summary(lm(y1~a*b+a*c+b*c))
```

	$Estimate$	$Std.Error$	$tvalue$	$Pr(>\|t\|)$	
$(Intercept)$	65.0875	0.2625	247.952	0.00257	$**$
a	−5.4375	0.2625	−20.714	0.03071	$*$
b	−1.3875	0.2625	−5.286	0.11903	
c	−8.2875	0.2625	−31.571	0.02016	$*$
$a:b$	1.5875	0.2625	6.048	0.10432	
$a:c$	−11.4125	0.2625	−43.476	0.01464	$*$
$b:c$	−1.7125	0.2625	−6.524	0.09683	.

Analysis of Variance Table
Model 1: y1 ∼ a * b + a * c + b * c
Model 2: y1 ∼ a + c + I(a * c)

	$Res.Df$	RSS	Df	$Sum\ of\ Sq$	F	$Pr(>F)$
1	1	0.551				
2	4	59.575	−3	−59.024	35.691	0.1223

```
A=a
B=b
C=c
```

```
ab=a*b
ac=a*c
bc=b*c
abc=ab*c
a=−a
b=−b
c=−c
ab=−ab
ac=−ac
bc=−bc
abc=−abc
y2=c(66.7, 65.0, 86.4, 61.9, 47.8, 59.0, 42.6, 67.6)
```
lm(y2∼ −A*B*C) #**Does it work ?**

(Intercept)
62.12

(x=lm(y2∼a*b*c)$coef) #**Does it work ?**

$(Intercept)$	a	b	c	$a:b$	$a:c$
62.125	−1.250	−2.500	7.875	−1.125	7.800

$b:c$	$a:b:c$
−1.650	−4.575

(x=lm(y2∼a+b+c+ab+ac+bc+abc)$coef)

$(Intercept)$	a	b	c	ab	ac
62.125	−1.250	−2.500	7.875	1.125	−7.800

bc	abc
1.650	−4.575

u=round(z+x,1) # l_A, l_B, l_C, l_D, l_E, l_F, ...
v=round(z-x,1)

sort(c(v,u[1]/2,u[-1]))

$a:c$	c	a	$a:b:c$	a	b
−19.2	−16.2	−6.7	−4.3	−4.2	−3.9
$a:c$	$b:c$	c	$b:c$	$a:b$	b
−3.6	−3.4	−0.4	−0.1	0.5	1.1
$a:b$	$(Intercept)$	$a:b:c$	$(Intercept)$		
2.7	3.0	4.8	63.6		

x=lm(y2∼a*b*c)$coef
u=round(z+x,1)
v=round(z-x,1)

sort(c(v,u[1]/2,u[-1]))

$a:c$	c	a	$a:b:c$	a	b
−19.2	−16.2	−6.7	−4.3	−4.2	−3.9
$a:c$	$b:c$	c	$b:c$	$a:b$	b
−3.6	−3.4	−0.4	−0.1	0.5	1.1
$a:b$	$(Intercept)$	$a:b:c$	$(Intercept)$		
2.7	3.0	4.8	63.6		

d=c(1,2,16)
(s=sqrt(mean(x[-d]**2))) # treating the other estimates as noises.

[1] 3.526929

s*qt(0.975,13)

[1] 7.619468 # cut point for significance, which suggests that a= −6.7 is not significant.

Another way:

```
a=c(A,a)
c=c(C,c)
b=c(B,b)
ab=c(A*B,ab)
ac=c(A*C,ac)
bc=c(B*C,bc)
abc=c(A*B*C,abc)
```

y=c(y1,y2)

AA=c(A,A)
BB=c(B,B)
CC=c(C,C)
EE=AA*CC
sort(round(lm(y~a+b+c+ab+ac+bc+abc +AA*BB*CC)$coef[-1]*2,1)) # effects after foldover

E	AE	A					
ac	CC	a	abc	AA	b	AA : CC	BB : CC
−19.2	−16.2	−6.7	−4.3	−4.2	−3.9	−3.6	−3.4

c	bc	AA : BB	BB	ab	AA : BB : CC
−0.4	−0.1	0.5	1.1	2.7	4.8

round((lm(y~ AA*BB*CC)$coef*2),1) # effects unchanged

(Intercept)	AA	BB	CC	AA : BB	AA : CC	BB : CC	AA : BB : CC
127.2	−4.2	1.1	−16.2	0.5	−3.6	−3.4	4.8

(lm(y~ e+a:e)$coef*2)

(Intercept)	e	e : a
127.2125	−19.2125	−16.1625

(lm(y~ a+e+a:e)$coef*2)

(Intercept)	a	e	a : e
127.2125	−6.6875	−19.2125	−16.1625

(lm(y~a+c+e+AA+CC+EE)$coef*2)

(Intercept)	a	c	e	AA	CC	EE
127.2125	−6.6875	−0.4125	−19.2125	−4.1875	−16.1625	−3.6125

Analysis of Variance Table
Model 1: y ~ a + c + e + AA + CC + EE
Model 2: y ~ e + CC

	Res.Df	RSS	Df	Sum of Sq	F	Pr(> F)
1	9	344.03				
2	13	645.94	−4	−301.91	1.9745	0.1822

Analysis of Variance Table
Model 1: y ~ a + c + e + AA + CC + EE
Model 2: y ~ a + e + CC

	Res.Df	RSS	Df	Sum of Sq	F	Pr(> F)
1	9	344.03				
2	12	467.05	−3	−123.02	1.0728	0.4083

Analysis of Variance Table
Model 1: y ~ e + CC
Model 2: y ~ a + e + CC

	Res.Df	RSS	Df	Sum of Sq	F	Pr(> F)	
1	13	645.94					
2	12	467.05	1	178.89	4.5962	0.05322	.

An economic alternative:
d=c(16,11,13,10)
a=c(A,a[d])
c=c(C,c[d])
e=c(A*C,ac[d])
d=c(8,3,5,2) # d+8=c(16,11,13,10)
y=c(y1,y2[d])
x=lm(y~a*c*e)
summary(x)
d=c(16,11,13,10)-8
a=c(A,-A[d])
c=c(C,-C[d])
e=c(A*C,-A[d]*C[d])

y=c(y1,y2[d])
x=lm(y~a*c*e)
x=lm(y~a*c*e)
summary(x)

	Estimate	*Std Error*	*t value*	$Pr(>\|t\|)$	
$(Intercept)$	65.8938	1.1816	55.764	$6.19e-07$	$***$
a	-2.5188	1.1816	-2.132	0.10003	
c	0.3562	1.1816	0.301	0.77807	
e	-10.8562	1.1816	-9.187	0.00078	$***$
$a:c$	-0.5562	1.1816	-0.471	0.66235	
$a:e$	-8.6438	1.1816	-7.315	0.00186	$**$
$c:e$	-2.9188	1.1816	-2.470	0.06894	.
$a:c:e$	-0.8062	1.1816	-0.682	0.53251	

Residual standard error: 3.859 on 4 degrees of freedom
$s = \sqrt{14.9} = 3.86$ computed by replications s^2, and matching summary().

u=lm(y~a*c+e*c+a*e)
summary(u)
anova(u,x)

	Estimate	*Std Error*	*t value*	$Pr(>\|t\|)$	
$(Intercept)$	65.6250	1.0528	62.331	$2.01e-08$	$***$
a	-2.5188	1.1167	-2.256	0.073765	.
c	0.3562	1.1167	0.319	0.762610	
e	-10.8562	1.1167	-9.722	0.000196	$***$
$a:c$	-0.5562	1.1167	-0.498	0.639536	
$c:e$	-2.9188	1.1167	-2.614	0.047457	$*$
$a:e$	-8.6438	1.1167	-7.740	0.000575	$***$

Residual standard error: 3.647 on 5 degrees of freedom

If ignoring ace, $\hat\sigma = \sqrt{\frac{1}{n-p}\sum_{i=1}^{12}(Y_i - \hat Y_i)^2} = 3.65$ with df 5.

Analysis of Variance Table
Model 1: y ~ a * c + e * c + a * e
Model 2: y ~ e + I(a * e)

	Res.Df	*RSS*	*Df*	*Sum of Sq*	*F*	$Pr(>F)$	
1	5	66.509					
2	9	308.334	-4	-241.82	4.545	0.06398	.

Analysis of Variance Table
Model 1: y ~ a * c + e * c + a * e
Model 2: y ~ e + I(a * e) + I(c * e)

	Res.Df	*RSS*	*Df*	*Sum of Sq*	*F*	$Pr(>F)$
1	5	66.509				
2	8	138.833	-3	-72.325	1.8124	0.2617

Analysis of Variance Table
Model 1: y ~ e + I(a * e)
Model 2: y ~ e + I(a * e) + I(c * e)

	Res.Df	*RSS*	*Df*	*Sum of Sq*	*F*	$Pr(>F)$		
1	9	308.33						
2	8	138.83	1	169.5	9.7672	0.01412	$*$	**conclusion**?

Conclusion: AE and E are significant and CE is significant.
Remark: There are three conclusion based on whether CE is significant:
(1) In last ANOVA (economic alternative), AE, E and CE are all significant.
(2) In full foldover, CE is on the boundary (p-value=0.053) (in contrast to the conclusion in (1).
(3) In the ecomonic alternative, recall

l_A	l_C	l_E	l_{AC}	l_{AE}		l_{CE}
-5.0	0.7	-21.7	-1.1	-17.3		-5.8
		s		s	$p-value = 0.06$ *using replications*	

The last conclusion (0.06 based on $\hat{\sigma}_{effect} = 2.23$) might be more reliable, as it relies on replications and does not rely on N(.,.).
Of course, if NID is true, statement (1) is more reliable.

Estimation of SD of effects.

```
y1=c(68.4, 77.7, 66.4, 81.0, 78.6, 41.2, 68.7, 38.7)
a=rep(c(-1,1),4)
b=rep(c(-1,-1,1,1),2)
c=rep(-1,4)
c=c(c,-c)
(z=lm(y1~a*b*c)$coef[-1]*2)
```

a	b	c	$a:b$	$a:c$	$b:c$	$a:b:c$
-10.875	-2.775	-16.575	3.175	-22.825	-3.425	0.525
s		s		s		

```
x=z$coef[c(2,4,6,7)]
sqrt(mean(x**2)) # SD of effects
    [1] 2.733587
x=lm(y1~a*c)
summary(x)
```

	$Estimate$	$Std.Error$	$tvalue$	$Pr(>\|t\|)$	
$(Intercept)$	65.087	1.364	47.702	$1.16e-06$	$***$
a	-5.437	1.364	-3.985	0.01633	$*$
c	-8.287	1.364	-6.074	0.00371	$**$
$I(a*c)$	-11.412	1.364	-8.364	0.00112	$**$
$2\times effects$	-10.8	2.734			
	-16.57	2.734			
	-22.82	2.734			
		$SD_{effects}$			

Residual standard error: 3.859 on 4 degrees of freedom

$3.859\sqrt{\frac{1}{4}+\frac{1}{4}} = 2.734$

sqrt(2*mean(x**2)) # compare to Residual standard error in summary

[1] 3.865876 # SD of errors.

Effect $=\overline{y}_+ - \overline{y}_-$.

$Var(effect) = \sigma_\epsilon^2(\frac{1}{4}+\frac{1}{4})$.

Recall that runs 1-8 can be treated as 4 pairs of replicates corresponding to factors a and c, leading to estimate of SD

$\hat{S}_\epsilon^2 = s^2 = \frac{\sum_{i=1}^4 d_i^2/2}{4} = 14.9$ with df 4 and $\sqrt{14.9} = 3.86$.

6.10. Increasing design resolution from III to IV by foldover.

The foldover in §6.9 increases the resolution from III to IV.
Before multiple column foldover, there are 15 defining relations:

I=ABD =ACE=BCF =CDG =DEF =BEG=AFG =ABCDEFG
=ABCG =BCDE =ACDF =ADEG =BDFG =ABEF =CEFG

Adding the foldover, only 4-letter defining relations remain and the resolution becomes 4.
I= ABCG=BCDE=ACDF =ABEF =ADEG=BDFG=CEFG. (Total of 7, not $\binom{7}{4} = 35$).
If we choose 3 letters from ABCG, we can form a replicated 2^3 FD, based on I=ABCG.

e.g. (A,B,C), (A,B,G), (A,C,G), (B,C,G)

So total of $4\times 7 = 28$ combinations,
out of total of $\binom{7}{3} = 35$.

This is the advantage of such an approach.
Notice that a single column foldover yield defining relation

I=ACE=BCF =BEG =ABCG =ABEF =CEFG

The resolution is ?

Recall that the single column foldover results in a 2^{7-3}_{III} design.
Look at the previous 16-run experiment. Is it a 2^{7-3} design ? If so, it is a 2^{7-3}_{IV} design.

	run #	*a*	*b*	*c*	*ab*	*ac*	*bc*	*abc*	*y*	(*table of contrast*)
		A	*B*	*C*	*D*	*E*	*F*	*G*		(*defining factors*)
	1	−	−	−	+	+	+	−	68.4	
	2	+	−	−	−	−	+	+	77.7	
	3	−	+	−	−	+	−	+	66.4	
	4	+	+	−	+	−	−	−	81.0	
	5	−	−	+	+	−	−	+	78.6	
	6	+	−	+	−	+	−	−	41.2	
	7	−	+	+	−	−	+	−	68.7	
Original order	8	+	+	+	+	+	+	+	38.7	
	run #	$-a$	$-b$	$-c$	$-ab$	$-ac$	$-bc$	$-abc$	*y*	
	9	+	+	+	−	−	−	+	66.7	
	10	−	+	+	+	+	−	−	65.0	
	11	+	−	+	+	−	+	−	86.4	
	12	−	−	+	−	+	+	+	61.9	
	13	+	+	−	−	+	+	−	47.8	
	14	−	+	−	+	−	+	+	59.0	
	15	+	−	−	+	+	−	+	42.6	
	16	−	−	−	−	−	−	−	67.6	

		a	*b*	*c*	$-ab$	$-ac$	$-bc$	*abc*	
	16	−	−	−	−	−	−	−	67.6
	15	+	−	−	+	+	−	+	42.6
	14	−	+	−	+	−	+	+	59.0
Reverse the last 8	13	+	+	−	−	+	+	−	47.8
	12	−	−	+	−	+	+	+	61.9
	11	+	−	+	+	−	+	−	86.4
	10	−	+	+	+	+	−	−	65.0
	9	+	+	+	−	−	−	+	66.7

The main component of 2^4 (or 2^{7-3}) FD: a,b,c,d columns.

		A	*B*	*C*	*E*		*A*	*B*	*C*	*F*
	16	−	−	−	−	16	−	−	−	−
	2	+	−	−	−	15	+	−	−	−
	14	−	+	−	−	3	−	+	−	−
	4	+	+	−	−	4	+	+	−	−
	5	−	−	+	−	5	−	−	+	−
	11	+	−	+	−	6	+	−	+	−
	7	−	+	+	−	10	−	+	+	−
Are these 2^4 FD ?	9	+	+	+	−	9	+	+	+	−
	1	−	−	−	+	1	−	−	−	+
	15	+	−	−	+	2	+	−	−	+
	3	−	+	−	+	14	−	+	−	+
	13	+	+	−	+	13	+	+	−	+
	12	−	−	+	+	12	−	−	+	+
	6	+	−	+	+	11	+	−	+	+
	10	−	+	+	+	7	−	+	+	+
	8	+	+	+	+	8	+	+	+	+

	run #	*a*	*b*	*c*	*ab*	*ac*	*bc*	*abc*	*y*	(*table of contrast*)
Is A B C D 2^4 FD ?		*A*	*B*	*C*	*D*	*E*	*F*	*G*		(*defining factors*)

What are the FD patterns for I=ABCG=BCDE=ACDF =ABEF =ADEG=BDFG=CEFG?
What are the FD patterns for I=ABCG=BCDE=ACDF =ABEF =ADEG=BDFG=CEFG?

They are not 2^4 FD, but replicated 2^{4-1} fractional FD:

I=ABCG

	A	B	C	G
16	−	−	−	−
15	+	−	−	+
14	−	+	−	+
13	+	+	−	−
12	−	−	+	+
11	+	−	+	−
10	−	+	+	−
9	+	+	+	+
1	−	−	−	−
2	+	−	−	+
3	−	+	−	+
4	+	+	−	−
5	−	−	+	+
6	+	−	+	−
7	−	+	+	−
8	+	+	+	+

	1	2	123	3
	A	B	C	G
16	−	−	−	−
11	+	−	+	−
10	−	+	+	−
13	+	+	−	−
12	−	−	+	+
15	+	−	−	+
14	−	+	−	+
9	+	+	+	+
1	−	−	−	−
6	+	−	+	−
7	−	+	+	−
4	+	+	−	−
5	−	−	+	+
2	+	−	−	+
3	−	+	−	+
8	+	+	+	+

$(A, C, G) = (1, 2, 3)$
$(B, C, G) = (1, 2, 3)$

?

I=ABEF

run #	a	b			ABF	c
	A	B	C	D	E	F
16	−	−				−
15	+	−				−
10	−	+				−
9	+	+				−
12	−	−				+
11	+	−				+
14	−	+				+
13	+	+				+
5	−	−				−
6	+	−				−
3	−	+				−
4	+	+				−
1	−	−				+
2	+	−				+
7	−	+				+
8	+	+				+

I=ADEG

run #	1	2	3		123
	A	D	E	F	G
16	−	−	−		−
9	+	−	−		+
14	−	+	−		+
11	+	+	−		−
12	−	−	+		+
13	+	−	+		−
10	−	+	+		−
15	+	+	+		+
7	−	−	−		−
2	+	−	−		+
5	−	+	−		+
4	+	+	−		−
3	−	−	+		+
6	+	−	+		−
1	−	+	+		−
8	+	+	+		+

or I=BCDE=ACDF =BDFG=CEFG

This is a 2^{7-3}_{IV} design. It can scan 3 out of 7 factors to form a replicated 2^3 FD for all $\binom{7}{3} = 35$ patterns, though $7 \times 4 = 28$ of them cannot form 2^4 FD.

6.10.2. Homework. Prove or disprove it. Can (B, D, F) be chosen as (a,b,c) ? How about (A, B, D) ?

6.11. 16-run design.

For computation purpose, instead define the orthogonal array in Table 6.14a, one can use

z=lm(y∼ a*b*c*d)

2*z$coef[2:16]

where y is the response variable and a, b, c, d are variables defined as follows.

a=c$(-1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1)$

b=c$(-1, -1, 1, 1, -1, -1, 1, 1, -1, -1, 1, 1, -1, -1, 1, 1)$

c=c$(-1, -1, -1, -1, 1, 1, 1, 1, -1, -1, -1, -1, 1, 1, 1, 1)$

d=c($-1,-1,-1,-1,-1,-1,-1,-1,1,1,1,1,1,1,1,1$)

nodal designs	*a*	*b*	*c*	*d*	*ab*	*ac*	*ad*	*bc*	*bd*	*cd*	*abc*	*abd*	*acd*	*bcd*	*abcd*
2^4	A	B	C	D											
2_V^{5-1}	A	B	C	D											P
2_{IV}^{8-4}	A	B	C	D							L	M	N	O	
2_{III}^{15-11}	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P

2^4 is not a nodal design. Neither is the 2_{IV}^{7-3} in §6.10. Note that

for convenience, in 2_V^{5-1} design, the 5-th factors are denoted by P, instead of by E.

I is not used as it denote columns with all +'s.

The alias structure for 16-run nodal designs are given in Table 6.14c.

Table 6.14c. Alias Structure for Sixteen Run Nodal Designs

$$\left(\begin{array}{cccc}
 & 2_V^{5-1} & 2_{IV}^{8-4} & 2_{III}^{15-11} \\
a & A & A & A+BE+\cdots \\
b & B & B & B+AE+\cdots \\
c & C & C & C+AF+\cdots \\
d & D & D & D+AG+\cdots \\
ab & AB & AB+CL+DM+NO & E+AB+\cdots \\
ac & AC & AC+BL+DN+MO & F+AC+\cdots \\
ad & AD & AD+BM+CN+LO & G+AD+\cdots \\
bc & BC & AL+BC+DO+MN & H+AL+\cdots \\
bd & BD & AM+BD+CO+LN & J+AM+\cdots \\
cd & CD & AN+BO+CD+LM & K+AN+\cdots \\
abc & DP & L & L+AH+\cdots \\
abd & CP & M & M+AJ+\cdots \\
acd & BP & N & N+AK+\cdots \\
bcd & AP & O & O+AP+\cdots \\
abcd & P & AO+BN+CM+DL & P+AO+\cdots
\end{array}\right)$$

Notice that the higher order interactions are ignored in the table.
It is due to the defining relation:

2_V^{5-1}: I=ABCDP,

$$2_{IV}^{8-4}: I=\underbrace{ABCL=ABDM=ACDN=BCDO}_{\binom{4}{1}}=\underbrace{...=ADLO}_{\binom{4}{2}}=\underbrace{ALMN=...=DMNO}_{\binom{4}{3}}$$

$$=\underbrace{ABCDLMNO}_{\binom{4}{4}}, \text{ total of } 2^4-1=15.$$

2_{IV}^{15-11}: I=ABE=..., total of $\binom{11}{1}+\binom{11}{2}+\cdots+\binom{11}{11}=2^{11}-1$

6.12. The nodal half replicate of 2^5 FD.

Reactor example. Table 6.15 shows the data and estimates from a complete 2^5 factorial design in factor A, B, C, D, E.

factor		$-$	$+$	
A :	*feed rate (L/min)*	10	15	
B :	*catalyst*(%)	1	2	
C :	*agitation(rpm)*	100	120	*jiaodong*
D :	*temperature*(^oC)	140	180	
E :	*concentration*	3	4	

a=c(a,a)
b=c(b,b)
c=c(c,c)

```
d=c(d,d)
e=rep(-1,16)
e=c(e,-e)
y=c(61, 53, 63, 61, 53, 56, 54, 61, 69, 61, 94, 93, 66, 60, 95, 98,
    56, 63, 70, 65, 59, 55, 67, 65, 44, 45, 78, 77, 49, 42, 81, 82)
(x=sort(round(lm(y~a*b*c*d*e)$coef[2:32]*2,1)))
```

$d:e$	e	$a:c:e$	$a:b:e$	a	$a:d$	$a:c:d$	$b:c:d:e$
−11.00	−6.25	−2.50	−1.87	−1.37	−0.88	−0.75	−0.63
c	$a:b:c:d:e$	$b:d:e$	$a:b:c:d$	$a:e$	$b:c:e$	$c:d:e$	$a:b:d:e$
−0.62	−0.50	−0.25	0.00	0.12	0.13	0.13	0.62
$a:d:e$	$a:c$	$b:c$	$c:e$	$a:c:d:e$	$b:c:d$	$a:b$	$a:b:d$
0.63	0.75	0.87	0.87	1.00	1.13	1.37	1.38
$a:b:c$	$a:b:c:e$	$b:e$	$c:d$	d	$b:d$	b	
1.50	1.50	2.00	2.12	10.75	13.25	19.50	

```
stem(x)
−1 || 1
−0 || 6
−0 || 321111110
+0 || 00001111111112222
+0 ||
+1 || 13
+1 ||
+2 || 0
```

The stem and leaf plot or the normal plot (Fig.6.7a) of the 31 estimates indicates that

over the ranges studied, only the estimates of the main effects B, D and E, and the interactions BD and DE are distinguishable from the noise.

```
summary(lm(y~b*d*e))
```

	Estimate	*Std. Error*	*t value*	$Pr(>\|t\|)$	
(*Intercept*)	65.5000	0.5774	113.449	$< 2e-16$	***
b	9.7500	0.5774	16.887	$8.00e-15$	***
d	5.3750	0.5774	9.310	$1.95e-09$	***
e	−3.1250	0.5774	−5.413	$1.47e-05$	***
$b:d$	6.6250	0.5774	11.475	$3.14e-11$	***
$b:e$	1.0000	0.5774	1.732	0.0961	.
$d:e$	−5.5000	0.5774	−9.526	$1.26e-09$	***
$b:d:e$	−0.1250	0.5774	−0.217	0.8304	

```
z=lm(y~b*d*e)
w=lm(y~b*d+d*e)
anova(w,z) Analysis of Variance Table
```

Model 1: y ~ b * d + d * e

Model 2: y ~ b * d * e

	Res.Df	*RSS*	*Df*	*Sum of Sq*	*F*	$Pr(>F)$
1	26	288.5				
2	24	256.0	2	32.5	1.5234	0.2383

Model $(E(Y|X) = 65.5 + 9.8b + 5.4d - 3.1e + 6.6b*d - 5.5d*e$; or $(E(Y|X)$ =**??**$+19.6I(b = 1) + 10.8I(d = 1) - 6.2I(e = 1) + 13.2I(b*d = 1) - 11.0I(d*e = 1)$.

run #	A	B	C	D	E			
1	−	−	−	−	−	1	17*	+
2*	+	−	−	−	−			−
3*	−	+	−	−	−			−
4	+	+	−	−	−	4	20*	+
5*	−	−	+	−	−			−
6	+	−	+	−	−	6	22*	+
7	−	+	+	−	−	7	23*	+
8*	+	+	+	−	−			−
9*	−	−	−	+	−			−
10	+	−	−	+	−	10	26*	+
11	−	+	−	+	−	11	27*	+
12*	+	+	−	+	−			−
13	−	−	+	+	−	13	29*	+
14*	+	−	+	+	−			−
15*	−	+	+	+	−			−
16	+	+	+	+	−	16	32*	+
17*	−	−	−	−	+		17	+
18	+	−	−	−	+			−
19	−	+	−	−	+			−
20*	+	+	−	−	+		20	+
21	−	−	+	−	+			−
22*	+	−	+	−	+		22	+
23*	−	+	+	−	+		23	+
24	+	+	+	−	+			−
25	−	−	−	+	+			−
26*	+	−	−	+	+		26	+
27*	−	+	−	+	+		27	+
28	+	+	−	+	+			−
29*	−	−	+	+	+		29	+
30	+	−	+	+	+			−
31	−	+	+	+	+			−
32*	+	+	+	+	+		32	+
new design of 2_V^{5-1}	a	b	c	d		*replace E*	*by*	$e = abcd$

Table 6.15

The full 2^5 requires 32 runs. If the experimenter had chosen instead to just make the 16 runs marked with asterisks in Table, 6.15, then only the data of the next Table would have been available.

$y_1 \sim y_{16}$	53	63	53	61	69	93	60	95
runs	2	3	5	8	9	12	14	15
$y_{17} \sim y_{32}$	56	65	55	67	45	78	49	82
runs	17	20	22	23	26	27	29	32

The generating relation is E=ABCD. The defining relation is I=ABCDE.

```
s=c(2,3,5,8,9,12,14,15,17,20,22,23,26,27,29,32)
(x=round(lm(y[s]~a[s]*b[s]*c[s]*d[s])$coef[2:16]*2,1))
# Or s=c(17,2,3,20,5,22,23,8,9,26,27,12,29,14,15,32)
# t=1:16
# (x=round(lm(y[s]~a[t]*b[t]*c[t]*d[t])$coef[2:16]*2,1))
```

a	b	c	d	$a:b$	$a:c$	$b:c$	$a:d$	$b:d$	$c:d$
−2.0	20.5	0.0	12.2	1.5	0.5	1.5	−0.7	10.8	0.3
	B		D					BD	

$a:b:c$	$a:b:d$	$a:c:d$	$b:c:d$	$a:b:c:d$
−9.5	2.2	1.2	1.2	−6.2
DE				E

Note that the main effects and 2-factor interaction effects are not very different from those from the full 2^5 design.

	a	b	c	d	$a:b$	$a:c$	$b:c$	$a:d$	$b:d$	$c:d$
$2^5:$	-1.4	19.5	-0.6	10.8	1.4	0.7	0.9	-0.9	13.2	2.1
$2^{5-1}:$	-2.0	20.5	0.0	12.2	1.5	0.5	1.5	-0.7	10.8	0.3
		B		D					BD	
	$d:e$		$a:b:c$	$a:b:d$	$a:c:d$	$b:c:d$			e	$\cdots$
$2^5:$	$(-11$	$+$	$1.5)$	1.4	-0.7	1.1			-6.2	$\cdots$
$2^{5-1}:$	$(=)$	-9.5		2.2	1.2	1.2			-6.2	?
		DE							$ABCD$	

ABC is aliased with DE due to I=ABCDE.
Moreover, the normal plot shows the similar pattern.
stem(x,3)

```
-0 | 06
-0 | 21
 0 | 00111222
 0 |
 1 | 12
 1 |
 2 | 1
```

sort(x)

DE	E								
$a:b:c$	$a:b:c:d$	a	$a:d$	c	$c:d$	$a:c$	$a:c:d$	$b:c:d$	$a:b$
-9.5	-6.2	-2.0	-0.7	0.0	0.2	0.5	1.3	1.3	1.5
$b:c$	$a:b:d$	$b:d$	d	b					
1.5	2.2	10.7	12.2	20.5					

summary(lm(y1~b*d+e+a*b*c))

	Estimate	*Std. Error*	*t value*	$Pr(>\|t\|)$	
$(Intercept)$	$6.525e+01$	$6.638e-01$	98.298	$2.07e-09$	$***$
b	$1.025e+01$	$6.638e-01$	15.441	$2.07e-05$	$***$
d	$6.125e+00$	$6.638e-01$	9.227	0.000251	$***$
e	$-3.125e+00$	$6.638e-01$	-4.708	0.005300	$**$
a	$-1.000e+00$	$6.638e-01$	-1.506	0.192295	
c	$5.412e-16$	$6.638e-01$	0.000	1.000000	
$b:d$	$5.375e+00$	$6.638e-01$	8.097	0.000466	$***$
$b:a$	$7.500e-01$	$6.638e-01$	1.130	0.309803	
$a:c$	$2.500e-01$	$6.638e-01$	0.377	0.721908	
$b:c$	$7.500e-01$	$6.638e-01$	1.130	0.309803	
$b:a:c$	$-4.750e+00$	$6.638e-01$	-7.156	0.000828	$***$

Residual standard error: 2.655 on 5 degrees of freedom
Multiple R-squared: 0.9894, Adjusted R-squared: 0.9683
F-statistic: 46.75 on 10 and 5 DF, p-value: 0.0002622

The 2_V^{5-1} can be used as a factor screen. In this example, factors A and C are inert. They can be checked from the half fraction factorial design. It is called a factor screen of order [16,5,4] (16 runs, 5 factors and projectivity 4). If one wants the full design, it can be obtained by foldover (on which factors ?)

6.13. The 2_{IV}^{8-4} nodal 16th fraction of a 2^8 factorial. This design is useful to screen 3 out of 8 factors in this 16-run design. A $[16, 8, 3]$ factor screen for 16 runs, 8 factors at projectivity 3. There are $\binom{8}{3} = 56$ ways.

nodal designs	a	b	c	d	ab	ac	ad	bc	bd	cd	abc	abd	acd	bcd	$abcd$
2_{IV}^{8-4}	A	B	C	D							L	M	N	O	
	A	B	C	D							E	F	G	H	??

There are $(\binom{4}{1} + \binom{4}{2} + \binom{4}{3} + \binom{4}{4} = 15)$ defining relations:

$$
\begin{aligned}
I &= \underbrace{ABCL = ABDM = ACDN = BCDO}_{\binom{4}{1}} \\
&= \underbrace{CDLM = BDLN = ADLO = BCMN = ACMO = ABNO}_{\binom{4}{2}} \\
&= \underbrace{ALMN = BLMO = CLNO = DMNO}_{\binom{4}{3}} = ABCDLMNO
\end{aligned}
$$

A Paint trial. In developing a paint for certain vehicles a customer required that the paint have high glossiness (y_1 on a scale 1 to 100) (guang-ze-du) and acceptable abrasion resistance (y_2 on a scale of 1 to 10) (nai-mo-xing). They believe that there are two main factors, say A and B. However, the factors A and B

either produce high glossiness but low abrasion,

or produce low glossiness but acceptable abrasion,

For instance,

A	$-$	$+$	$-$	$+$	$-$	$+$	$-$	$+$	
B	$-$	$-$	$+$	$+$	$-$	$-$	$+$	$+$	*ideal*
y_1 :	53	78	48	78	68	61	70	65	≥ 65
y_2 :	6.3	2.1	6.9	2.5	3.1	4.3	3.4	3.0	≥ 5

According to the paint technologist, there are 6 more factors, C, D, E, F, G, H. They want to find out how to select the factors to obtain high glossiness and high abrasion,

The experiments results in data as follows.

y_1 = c(53,60,68,78,48,67,55,78,49,68,61,81,52,70,65,82)

y_2 = c(6.3,6.1,5.5,2.1,6.9,5.1,6.4,2.5,8.2,3.1,4.3,3.2,7.1,3.4,3.0,2.8)

lm(y_1 ~ a*b*c*d)$coef[2:16]*2

Effects are

	A	B	C	D	E	F	G	H							
y_1	**16.6**	**12.6**	-0.1	2.6	-0.1	-0.9	-3.6	1.9	0.9	2.6	1.9	-1.9	-0.1	2.6	-0.4
y_2	**-2.4**	**-2.0**	-0.2	-0.7	0.1	**1.6**	0.6	-0.3	0.3	0.0	-0.1	0.1	-0.1	-0.4	-0.2

Sorting the effects of y_1:

-3.6 -1.9 -0.9 -0.4 -0.1 -0.1 -0.1 0.9 1.9 1.9 2.6 2.6 2.6 12.6 16.6;

```
-0 | 4210000
 0 | 122333
 0 |
 1 | 3
 1 | 7
```

Sorting the effects of y_2

-2.4 -2.0 -0.7 -0.4 -0.3 -0.2 -0.2 -0.1 -0.1 0.0 0.1 0.1 0.3 0.6 1.6

```
-2 | 40
-1 |
-1 |
-0 | 7
-0 | 432211
 0 | 0113
 0 | 6
 1 |
 1 | 6
```

Analysis of Variance Table

Model 1: y[1,] ~ a + b + I(a * b * d)

Model 2: y[1,] ~ a * b * d

	Res.Df	*RSS*	*Df*	*Sum of Sq*	*F*	*Pr(> F)*
1	12	181.25				
2	8	136.50	4	44.75	0.6557	0.6394

The ANOVA and stem-and-leaf plot suggest that

effects on y_1 are not significant different from error if they are in the range $(-3.6, 2.6)$,

effects on y_2 are not significantly different from error if they are in the range $(-0.7, 0.6)$.
Thus in addition to factors A and B, factor F is also an important factor.

Hence we screen 3 factors out of 8.

x=lm(y[1,]∼a+b+I(a*b*d))

		Estimate	*Std Error*	*t value*	$Pr(>\|t\|)$	
	$(Intercept)$	64.6875	0.9413	68.719	$< 2e-16$	$***$
summary(x)	a	8.3125	0.9413	8.831	$7.46e-07$	$***$
	b	6.3125	0.9413	6.706	$1.46e-05$	$***$
	$I(a*b*d)$	-0.4375	0.9716	-0.450	0.661	

$Y_1 = 64.7 + 8.3a + 6.3b + \epsilon.$

$Y_1 = 50.1 + 16.6 \times \mathbf{1}(a = 1) + 12.6 \times \mathbf{1}(b = 1) + \epsilon$ **how**?

x=lm(y[2,]∼a+b+I(a*b*d))

		Estimate	*Std Error*	*t value*	$Pr(>\|t\|)$	
	$(Intercept)$	4.7500	0.1647	28.842	$1.88e-12$	$***$
summary(x)	a	-1.2125	0.1647	-7.362	$8.71e-06$	$***$
	b	-1.0250	0.1647	-6.224	$4.42e-05$	$***$
	$I(a*b*d)$	0.8000	0.1647	4.858	0.000393	$***$

$Y_2 = 4.8 - 1.2a - 1.0b + 0.8a * b * d + \epsilon.$

$Y_2 = 6.2 - 2.4 \times \mathbf{1}(a = 1) - 2.1 \times \mathbf{1}(b = 1) + 1.6 \times \mathbf{1}(a * b * d = 1) + \epsilon.$

How to obtain high y_1 (≥ 65) and acceptable y_2 (≥ 5) if A and B **can be numerical ?**

The contour plots (based on Eq.s (1) and (2)) in Figure 6.9 (in the textbook) suggest that

at + level of A, at − level of B, addition of F (+ level) would make possible of substantial improvement in (high) glossiness y_1 (≥ 65) while maintaining an acceptable level of abrasion resistance y_2 (≥ 5).

$Y_1 = 64.7 + 8.3a + 6.3b + \epsilon.$

$Y_2 = 4.8 - 1.2a - 1.0b + 0.8a * b * d + \epsilon.$

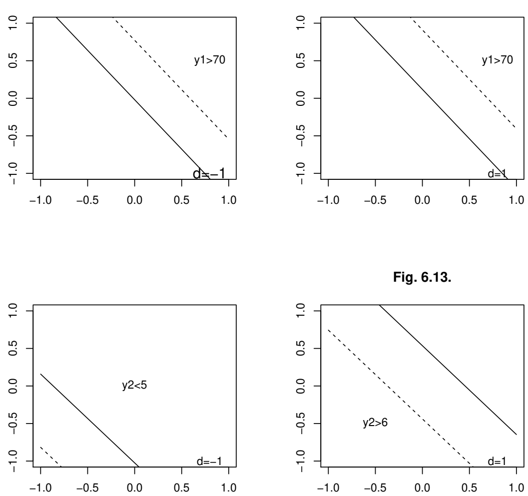


Fig. 6.13.

a,b-axis, (F=abd= ±1)

```
A=c(-1,1)
x=lm(y[1,]∼a+b+I(a*b*d))$coef
B=(65-x[1]-x[2]*A+x[4])/x[3] # (65=x[1]+x[2]*A +x[3]*B -x[4])
plot(A,B, ylim=c(-1,1), type="l", lty=1)
B=(70-x[1]-x[2]*A+x[4])/x[3]
lines(A,B, type="l", lty=2)
text(0.8,-1,"(F=-1)") # (a,b,d)∈ {(1, −1, −1), (−1, 1, −1)}
```

```
text(0.8,0.5,”y1>70”)
B=(65-x[1]-x[2]*A-x[4])/x[3] #+level
plot(A,B, ylim=c(-1,1), type=”l”, lty=1)
B=(70-x[1]-x[2]*A-x[4])/x[3]
lines(A,B, type=”l”, lty=2)
text(0.8,-1,”(F=1)”) # (a,b,d)∈ {(1,−1,1),(−1,1,1)}
text(0.8,0.5,”y1>70”)
x=lm(y[2,]∼a+b+I(a*b*d))$coef
B=(5-x[1]-x[2]*A+x[4])/x[3]
plot(A,B, ylim=c(-1,1), type=”l”, lty=1)
B=(6-x[1]-x[2]*A+x[4])/x[3]
lines(A,B, type=”l”, lty=2)
text(0.8,-1,”(F=-1)”)
text(0.0,0.0,”y2<5”)
B=(5-x[1]-x[2]*A-x[4])/x[3] #+level
plot(A,B, ylim=c(-1,1), type=”l”, lty=1)
B=(6-x[1]-x[2]*A-x[4])/x[3]
lines(A,B, type=”l”, lty=2)
text(0.8,-1,”(F=1)”)
text(-0.5,-0.5,”y2>6”)
```

6.13.2. Homework. Draw the contour plots for the region in (A,B) with $F = +1$ such that both y_1 and y_2 acceptable.

6.14. 2^{15-11}_{III} **design, an example.**

The design can be used to screen for two factors amount 15 factors.
Postextrusion shrinkage of a speedometer casing had produced undesirable noise.
The objective of the experiment was to find a way to reduce the shrinkage.
A considerable length (in > 300 meters) of product was made during each run and measurements were made at 4 equally spaced points, the responses are
the averages and log variances of the 4 measurements.

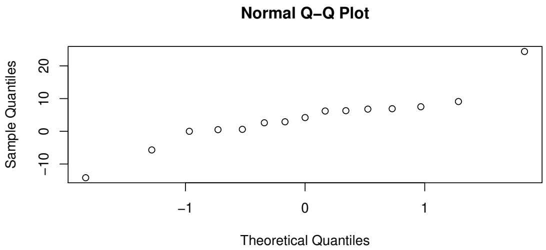
```
y=c(48.5,57.5,8.8,17.5,18.5,14.5,22.5,17.5,12.5,12,45.5,53.5,17,27.5,34.2,58.2) #mean

s=c(-0.8,-0.8,0.4,0.3,-0.2,0.1,-0.3,1.1,-0.1,-0.2,0.7,-0.9,0.7,-0.9,0.1,0.7,0.6) #run log variance
```

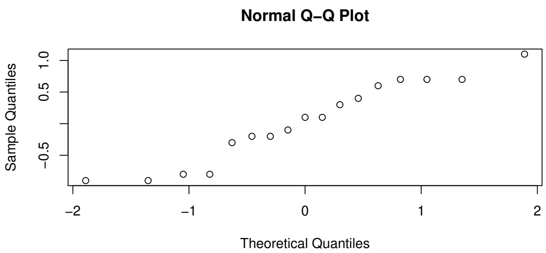
The 15 factors are

A: liner tension, (chen ban zhangli)
B: liner line speed, (ban lun xian speed)
C: liner die, (ban lun mo ju)
D: liner outsider diameter, (chen guan outsider diameter)
E: melt temperature,
F: coating material,
G: liner temperature,
H: braid tension, (bian zhi tension)
J: wire braid type, (xian bian zhi lei xing)
K: liner material,
L: cooling method,
M: screen pack,
N: coating die type, (tu cheng mo ju lei xing)
O: wire diameter,
P: line speed.

```
x=lm(y∼a*b*c*d)$coef[2:16]*2
```

a	b	c	d	$a:b$	$a:c$	$b:c$	$a:d$	$b:d$	$c:d$
6.3	6.2	-5.7	6.9	2.6	0.0	7.5	4.2	$\underline{24.4}$	9.1
A	B	C	D	E	F	G	H	**J**	K
$a:b:c$	$a:b:d$	$a:c:d$	$b:c:d$	$a:b:c:d$					
0.5	2.9	6.7	$\underline{-14.2}$	0.7					
L	M	N	**O**	P					





Normal plot of mean and log variance of the measurements

The second normal plot about s did not reveal any important factors.

The stem-and-leaf plot of y reveals important factors.

−1| 4
−0| 6
−0|
+0| 011334
+0| 667789
+1|
+1|
+2| 4

It turns out from the normal plot of the effects due to averages that
the factors O, J and C are important factors.

nodal			c						bd					bcd	
designs	a	b	c	d	ab	ac	ad	bc	bd	cd	abc	abd	acd	bcd	$abcd$
2^{15-11}_{III}	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P

J=b*d
O=J*c
x=lm(y~factor(c)+factor(J)+factor(O))
summary(x)

	Estimate	*Std.Error*	*tvalue*	$Pr(>\|t\|)$			
$(Intercept)$	26.863	5.351	5.020	0.000299	$***$		
$factor(c)1$	-5.737	5.351	-1.072	0.304693		$O*J$	(1)
$factor(J)1$	24.388	5.351	4.558	0.000657	$***$	$b*d$	
$factor(O)1$	-14.163	5.351	-2.647	0.021305	$*$	$J*c$	

Residual standard error: 10.7 on 12 degrees of freedom
Multiple R-squared: 0.7068, Adjusted R-squared: 0.6335
F-statistic: 9.643 on 3 and 12 DF, p-value: 0.001612
anova(u,w)
Model 1: y ~ J*O

Model 2: y ∼ J + O

	Res.Df	*RSS*	*Df*	*Sum of Sq*	*F*	*Pr(> F)*
1	12	1374.3				
2	13	1506.0	−1	−131.68	1.1497	0.3047

Notice that C is alias with the interaction of OJ, *i.e.*, I=OJC, as O=bcd and J=bd.
Conclusion: The model is

$$y = 26.87 + 24.38 \times \mathbf{1}(J = 1) - 14.16 \times \mathbf{1}(O = 1) \text{ (see (1)) ? Or}$$

$$y = \underbrace{23.99}_{=26.87-5.74/2} +24.38 \times \mathbf{1}(J = 1) - 14.16 \times \mathbf{1}(O = 1), \text{ or}$$

$$y = 29.11 + 12.19 \times J - 7.08 \times O,\ J, O = \pm 1.$$

23.99+12.19-7.08=29.10
In order to reduce the shrinkage, set factors J and O at levels -1 and +1, respectively.

Remark. The results using summary(lm(y ∼ factor(c) + factor(J) + factor(O))) can be derived directly as follows. The 2^{15-11}_{III} design can be viewed as a 4 replicated 2^2 factorial design with $(\mathbf{1}, \mathbf{2}) = (J, O)$, $\begin{array}{ccccc} j & - & + & - & + \\ o & - & - & + & + \\ type\ \# & 1 & 2 & 3 & 4 \end{array}$ with their average of y_1 .

From the table of contract of 2^{15-11}_{III}, we have

J	+	+	−	−	+	+	−	−	−	−	+	+	−	−	+	+
O	−	−	+	+	+	+	−	−	+	+	−	−	−	−	+	+
types	2	2	3	3	4	4	1	1	3	3	2	2	1	1	4	4
run #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

```
> s=c(7,8,13,14,1,2,11,12,3,4,9,10,5,6,15,16) # where are they come from ?
> mean(y[s[1:4]) # 21.1
> mean(y[s[5:8]) # 51.3
> mean(y[s[9:12]) # 12.7
> mean(y[s[13:16]) # 31.4
> mean(y)
[1] 29.10625
```

The results lead to $\begin{array}{ccc} O+ & 12.7 & 31.4 \\ O- & 21.1 & 51.3 \\ & J- & J+ \end{array}$

$$=> \begin{array}{cccc} \hat\beta_J = 24.45 & \overline{y}_{j-} & \overline{y}_{j+} & \\ O+ & 12.7 & 31.4 & \overline{y}_{o+} \\ O- & 21.1 & 51.3 & \overline{y}_{o-} \\ & J- & J+ & -14.15 = \hat\beta_O \end{array}$$

```
> sqrt((var(y[s[1:4]])+var(y[s[5:8]])+var(y[s[9:12]])+var(y[s[13:16]]))/4)
[1] 10.70166 # estimating residual SD directly, same as in summary(x)
> y=c(21.1,51.3,12.7,31.4)
> v=lm(y~J+O)$coef
> c(v[1],2*v[2:3])
```

$(\overline{y})$	J	O
29.10625	24.38750	−14.16250

So the 2^{15-11}_{III} fractional FD successfully screens 2 factors from 15 factors.

6.15. Constructing other two-level fractions. Adding a factor to a nodal design. Recall Table 6.14b for 16-run nodal designs.

nodal designs	a	b	c	d	ab	ac	ad	bc	bd	cd	abc	abd	acd	bcd	$abcd$
2^4	A	B	C	D											
2_V^{5-1}	A	B	C	D											P
2_{IV}^{8-4}	A	B	C	D							L	M	N	O	
2_{III}^{15-11}	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P
$non-nodal$															
2_{III}^{9-5}	A	B	C	D							L	M	N	O	P

One can choose a factor which is most likely to be inert and it is likely to be factor P.
The alias structure becomes

	2_{IV}^{8-4} *each has* 16 *aliases*	$1st\ 2_{III}^{9-5}$ *each has* 32 *aliases*
a	A	$A+OP$
b	B	$B+NP$
c	C	$C+MP$
d	D	$D+LP$
ab	$AB+CL+DM+NO$	
ac	$AC+BL+DN+MO$	
ad	$AD+BM+CN+LO$	*same*
bc	$AL+BC+DO+MN$	*as*
bd	$AM+BD+CO+LN$	2_{IV}^{8-4}
cd	$AN+BO+CD+LM$	
abc	L	$L+DP$
abd	M	$M+CP$
acd	N	$N+BP$
bcd	O	$O+AP$
$abcd$	$AO+BN+CM+DL$	$P+AO+BN+CM+DL$

6.15.2. Remark. The 1st non-nodal 2^{9-5} design is of resolution III. The reason is as follows.

The generating relation is I=ABCDP, together with 4 generating relations from 2_{IV}^{8-4} FFG. There are $2^5-1=31$ defining relations. 15 of them are the same as the 2_{IV}^{8-4} FFD, which has either 4 letters or 8 letters.

$$2_{IV}^{8-4}\text{: } I=\underbrace{\text{ABCL=ABDM=ACDN=BCDO}}_{\binom{4}{1}}=\underbrace{...=ADLO}_{\binom{4}{2}}=\underbrace{ALMN=...=DMNO}_{\binom{4}{3}}$$

$$=\underbrace{ABCDLMNO}_{\binom{4}{4}},\text{ total of } 2^4-1=15.$$

Another 15 are due to ABCDP times each of the previous 15. Since each of these 4-letter words does not contain all of ABCD, their products with ABCDP have lengths ≥ 3. *e.g.*, The 1st one is ABCL. ABCDP(ABCL)=DLP. Moreover, ABCDP(ABCDLMNO)=LMNOP.

Are there other non-nodal 2^{9-5} design of resolution III ? Consider the next example.

(1) I =ABCL=ABDM=ACDN=BCDO =ABE (replacing P=ABCD).

It's resolution is III, the reason is as follows.

There are 31 defining relations, which consists of original 15 from the 2_{IV}^{7-4} FFD + ABE, and ABE times each of the original 14 4-letter words, which do not contain E. Thus, the shortest one of the latter 15 products is ABE(ABXY)=EXY. Moreover,

ABCDLMNO(ABE)=DELMNO.

6.16. Elimination of block effects. Fractional designs may be run in blocks with suitable contrast used as "block variable". A design in $\underline{2^q}$ blocks is defined by

q independent contrast. All effects (including aliases) associated with these chosen contrasts and all their interactions are confounded with blocks. Consider the 2_V^{5-1} design as follows.

run #	a	b	c	d	$e = abcd$	ab	ac	ad	bc	abc
1	−	−	−	−	+	+	+	+	+	−
2	+	−	−	−	−	−	−	−	+	+
3	−	+	−	−	−	−	+	+	−	+
4	+	+	−	−	+	+	−	−	−	−
5	−	−	+	−	−	+	−	+	−	+
6	+	−	+	−	+	−	+	−	−	−
7	−	+	+	−	+	−	−	+	+	−
8	+	+	+	−	−	+	+	−	+	+
9	−	−	−	+	−	+	+	−	+	−
10	+	−	−	+	+	−	−	+	+	+
11	−	+	−	+	+	−	+	−	−	+
12	+	+	−	+	−	+	−	+	−	−
13	−	−	+	+	+	+	−	−	−	+
14	+	−	+	+	−	−	+	+	−	−
15	−	+	+	+	−	−	−	−	+	−
16	+	+	+	+	+	+	+	+	+	+

$q = 1$. A 2_V^{5-1} in two blocks of either runs. (*e.g.* male or female patients). If one believes that AC is most likely to be negligible, then 2^1 blocks can be decided as follows.

1. the 8 runs 2, 4, 5, ..., 15, having − in the AC column;
2. the other 8 runs having + in the AC column.

The block contrast is AC. AC is confounded with the block factor, say **6**, with 2 levels.

$q = 2$. A 2_V^{5-1} design in 4 blocks of 4 runs. (*e.g.* a pack of raw material enough for 4 runs). If one uses AC and BC to define blocks, then the sign (- -), (- +), (+ -) and (+ +) can be the 4 blocks.

(−−): runs 4, 5, 12, 13;

(−+): runs 2, 7, 11, 15;

......

In this case, AC and BC are confound with the block factor **6** with 4 levels (or 2 new block factors F=BC and G=BC.

run #	a	b	c	d	$abcd$	ab	ac	ad	bc	abc
8	+	+	+	−	−	+	+		+	+
9	−	−	−	+	−	+	+		+	−
16	+	+	+	+	+	+	+		+	+
1	−	−	−	−	+	+	+		+	−
2	+	−	−	−	−	−	−		+	+
15	−	+	+	+	−	−	−		+	−
7	−	+	+	−	+	−	−		+	−
10	+	−	−	+	+	−	−		+	+
14	+	−	+	+	−	−	+		−	−
3	−	+	−	−	−	−	+		−	+
6	+	−	+	−	+	−	+		−	−
11	−	+	−	+	+	−	+		−	+
12	+	+	−	+	−	+	−		−	−
13	−	−	+	+	+	+	−		−	+
4	+	+	−	−	+	+	−		−	−
5	−	−	+	−	−	+	−		−	+

$q = 3$. Is it possible to use ab, ac, bc for the case of $q = 3$?

How about other combinations ?
Ans. (ab, ac, ad) works; and (ab, ac, abc) works.

Remark. Homework solution is in my website

Minimum-Aberration 2^{k-p} designs. Before given its definition, consider first 2_{IV}^{7-2} designs.

Table 6.21. 3 choices for a 2_{IV}^{7-2} fractional FD

	$design(a)$	$design(b)$	$design(c)$
	$share\ 2\ \#$	$share\ 1\ \#$	$share\ 3\ \#$
$2\ generators$	$6=123,\ 7=234$	$6=123,\ 7=145$	$6=1234,\ 7=1235$
$3\ defining$	$I=1236=2347=1467$	$I=1236=1457$	$I=4567$
$relations$		$=234567$	$=12346=12357$
$\binom{4}{2}$ $alliases\ from\ 1st$	$12+36$	$12+36$	$45+67$
$(with\ 2\ letters)$	$13+26$	$13+26$	$46+57$
	$16+23$	$16+23$	$47+56$
$\binom{4}{2}$ $alliases\ from\ 2nd$	$23+47$	$14+57$	
$(with\ 2\ letters)$	$24+37$	$15+47$	
	$27+34$	$17+45$	
$\binom{4}{2}$ $alliases\ from\ 3rd$	$14+67$		
$(with\ 2\ letters)$	$16+47$		
	$17+46$		
$distinct\ patterns$	$12+36$	$12+36$	$45+67$
	$13+26$	$13+26$	$46+57$
	$16+23+47$	$16+23$	$47+56$
		$14+57$	
	$24+37$	$15+47$	
	$27+34$	$17+45$	
	$14+67$		
	$17+46$		
total # words:	15	12	6

How about 6=12345 and $\underbrace{7=12}_{R<IV}$ or $\underbrace{123}_{R=IV}$ or $\underbrace{1234}_{R<IV}$?

(4) 6=12345 and 7=123 => I=123456=1237=4567 is similar to design (b).
Which pattern has the least # of shortest words among defining relations ?
Definition. The minimum-aberration design is the one that minimizes the number of words in the defining relation having minimum length.

See for examples, 2_{IV}^{7-2}, 2_{V}^{5-1} and 2^4 designs as follows.

2_{IV}^{7-2} **fractional FD design**: There are several types of them. In each type, Factors 1, 2, 3, 4, 5 (as well as 6, 7) are aliased with 3-factor or high order interactions, Table 6.21 above gives an example of each of three types, where
2-factor interactions which are aliased with (only) 2-factor interactions are given there.
Which design is better ?

Design (c), as it has the least number of 2 factor interactions.

Design (c) is the minimum-aberration 2_{IV}^{7-2} design, which has 1 word of length 4.

Design (b) or (a) has 2 or 3 words of length 4.

Is it the unique minimum-aberration 2_{IV}^{7-2} design ?

2_{V}^{5-1} **fractional FD designs:** There is just one (a nodal design), with a unique defining generator $5 = 1234$. So it is the minimum-aberration 2_{V}^{5-1} design.

2^4 **FD.** There is no defining generator. So it is also a minimum-aberration 2^4 design.

Table 6.22 is for general 2^{k-p} FD with Table 6.21 as the special case.
Explain it via the next table.

	# of variables k (or factors)	
# of runs	5	6
4		
8	$\frac{1}{4}$ *Fractional FD of* 2^5	$\frac{1}{8}$ *Fractional FD of* 2^6
16	$\frac{1}{2}$ *Fractional FD of* 2^5	$\frac{1}{4}$ *Fractional FD of* 2^6
32	1 *FD of* 2^5	$\frac{1}{2}$ *Fractional FD of* 2^6
64	2 *replicated FD of* 2^5	1 *FD of* 2^6
128	4 *replicated FD of* 2^5	2 *replicated FD of* 2^6

$123456789\overline{10} = ABCDEFGHJK$

What are the nodal designs in row 1 ?

What are the nodal designs in row 2 ?

What are the nodal designs in row 3 ?