Design of Experiments (Math 556)

FA 351 MWF 8:00am-9:30am,
Office: WH 132
Office hours: MT 3:30-4:30pm
Textbook: Statistics for Experimenters (2nd ed.)
by George Box, J Stuart Hunter and William G. Hunter
Homework due: W in class.
Midterm: Oct. 21 (M), Final: Dec. 12 Thursday 10:25am -12:25pm SW 310. **Changed to WH100E !!**

Each is allowed to bring a piece of paper with anything on it.
Homework assigned during last week is due each Wednesday.
It is on my website: http://www.math.binghamton.edu/qyu/qyu_personal
Remind me if you do not see it by Saturday morning !
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
Grading Policy: 50% hw and quizzes +50% exams, quiz problem: formulas for 447-448.
B = 70 $\pm$

Chapter 1. Introduction

Self-reading.

Chapter 2. Basic

All the concepts of 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 $k-$th sample autocorrelation coefficient of Y_i's are

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

It measures the serial dependence of the data in time.
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) $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$.

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

Chapter 3. Comparing Two Entities

3.1. Consider the test for the difference of the means of two random samples.
H_o: $\mu_Y - \mu_X = \delta$ v.s. H_1: $\mu_Y - \mu_X > \delta$.
Under the assumption that both random samples are from $N(\mu_Y, \sigma^2)$ and $N(\mu_X, \sigma^2)$, then a common test is the t-test

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

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

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

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 the paired t-test

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

where

$t = \frac{\overline{Y}-\overline{X}-\delta}{s\sqrt{1/n}}$,

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - X_i - \overline{Y - X})^2$.

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($\overline{y}_B - \overline{y}_A \geq 1.3$) = ?

Then 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%.

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 ?

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

$$s_p^2 \text{ v.s. } S^2_{Y_B - Y_A}$$

What is the conclusion if we use it ?

Introduce two altertive 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 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 auto-correlation of the data is $\hat{\rho}_1 = -0.29$.
   The data are not independent.
   If one pretends independence, it leads to incorrect conclusion.
2. Normal assumption may not be valid.

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	(1)
yield :	29.2	11.4	26.6	23.7	25.3	28.5	14.2	17.9	16.5	21.1	24.3	

```
> 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 1.69
```

$\overline{y}_B - \overline{y}_A \approx 1.69$.
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)
N=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]])
```

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 ?

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
```

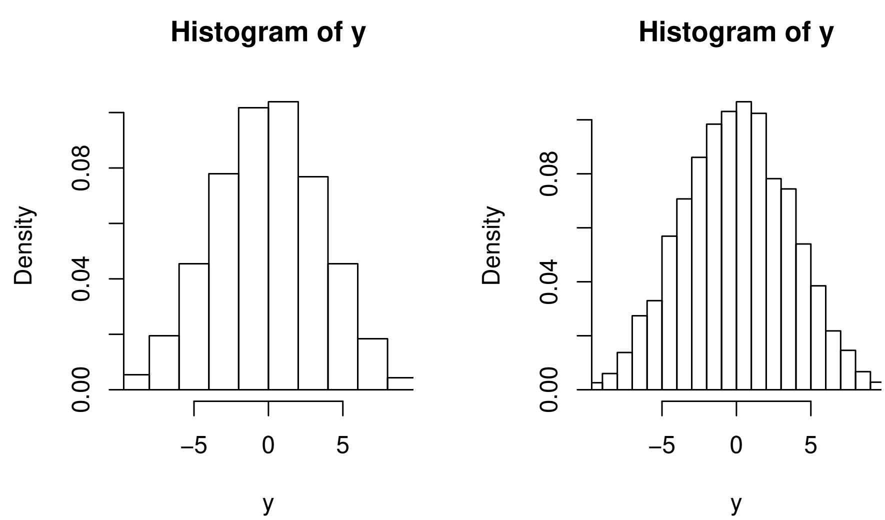


Figure 3.1. Histograms of permutation distribution v.s. simulation one

Remark. If the normal assumption is not valid, the t-test is not applicable. 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 $t_{n_A+n_B-2}$. $P(t_{n_A+n_B-2} \leq \frac{1.69}{s\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}}) \approx 1 - 0.34)$ for the current data.

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. If they had done

```
sample(1:20,10)
[1] 10 4 15 12 13 2 11 6 20 8
```

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 were made of two different materials, A and B. 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 ?

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) # $y_B - y_A$

Then 10 $y_B - y_A$'s yield

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

Should we use 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[i%*%(2**h) +1]=sum(x*((-1)**i))
```

$\underbrace{i\% * \%(2 * *h)}$

$\# \ i1 * 2^0 + i2 * 2^1 + i3 * 2^2 + \cdots + i10 * 2^9,$

\# Examples:

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

\# 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
z=seq(-6,6,0.1)
hist(y)
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)
```

One can see from Figure 3.2 that the simulation distribution is very close to the true permutation distribution.
The P-value using the paired t-test is 0.4% (close enough to 0.7%).

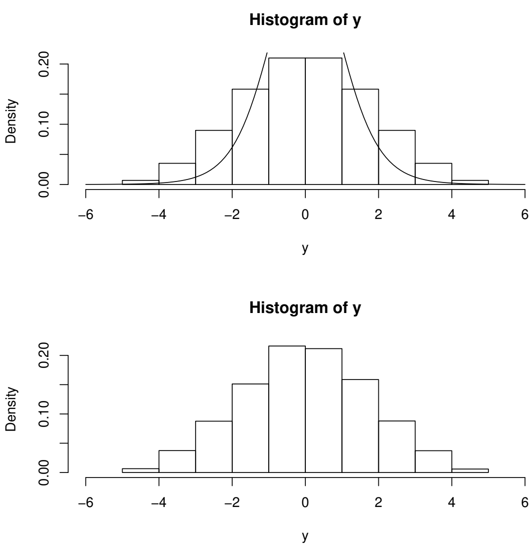


Figure 3.2. Histograms of permutation distribution v.s. simulation one

Any thing wrong with the solution ?
Is it one sided test or two-sided test ?

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 $E(\epsilon|\mathbf{X}) = 0$,
β_i's are parameters,
X_i's and Y are observable,
it is often assumed that x_i's are constant, not random.
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 $\mathbf{Y}' = (Y_1, ..., Y_n)$,

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

The LSE can be obtained by solving the normal equation
$\frac{\partial S}{\partial \beta} = \mathbf{0}$, a p dimensional zero vector.
That is,
$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{0}$. **(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.
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 ??)**
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 $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$?).
Under NID a $(1-\alpha)100\%$ CI of β_i is $\hat\beta_j \pm t_{n-p,\alpha/2}SE$
Exam example: Suppose that $Y_i = \begin{cases} -\beta + W_i & \text{if } i \in \{1, ..., n_-\} \\ \beta + W_i & \text{if } i \in \{n_- + 1, ..., n\} \end{cases}$ and W_1, ..., W_n are i.i.d. from the exponential distribution and $E(W_1) = 1$. β is unknown, Y_i's are observations, but W_i's are not observable, though we know $W_i \sim Exp(1)$. Derive the LSE and the MLE of β based on the regression data (X_1, Y_1), ..., (X_n, Y_n).
Discussion. Let the model be $Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i$.
$p = ?$
Do we know β ?
$W_i = \epsilon_i$?
Do we know X_i ?
Do we know Y_i ?
If we write $Y_i = \alpha + \beta X_i + \epsilon_i$, then $\alpha = ?$
Do we need to estimate α ?
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
$\mathbf{Y} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 9 \end{pmatrix}$,
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.* $\beta_0 = 1$ or $\beta_1 = \beta_2$, etc..

Example 2. One way anova table $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$, $i = 1$, ..., I, and $j = 1$, ..., 3.
Is it a linear regression model ?
$\beta = ?$
LSE $= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

$\mathbf{X} = ?$

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ \vdots \\ Y_{I1} \\ Y_{I2} \\ Y_{I3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & & & \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \epsilon.$$

Indicator variables.
Consider an example that there are three treatments A, B and C. Define

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

$X_{ij}^2 = X_{ij}$,
$X_{i1}X_{i2} = 0$ etc. and
$X_{i1} + X_{i2} + X_{i3} = 1$.

Thus

1. the data cannot be applied to polynomial model of degree 2 or above,
2. 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**).
$\mathbf{X'X}$ 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 ?**

Three modifications:

1. Revise the model. Let $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$
   $lm(Y \sim X_1 + X_2 + X_3 - 1)$
2. Impose a constraint $\alpha_1 = 0$ for the model
   $Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} + \epsilon_i$.
   Then $\beta_i = \mu + \alpha_i$, $i = 1, 2, 3$.
   options(contrasts =c("contr.treatment", "contr.poly"))
   $lm(Y \sim X_1 + X_2 + X_3)$.
3. 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 $\mu + \alpha_i = \beta_i$, $i = 1, 2, 3$.
   options(contrasts =c("contr.sum", "contr.poly"))
   $lm(Y \sim X_1 + X_2 + X_3)$

Example 3 (a simulation study).
(Two way anova table) $Y_{ij} = \mu + \alpha_i + b_j + \epsilon_{ij}$, $i \in \{1, ..., 4\}$, $j \in \{1, ..., 6\}$

```
> 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

> 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

> 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

Relation between these three ?

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 Example 3 ?

$$\mathbf{X}=\begin{pmatrix} 1 \\ \mathbf{1}(a=1) \\ \mathbf{1}(a=2) \\ \mathbf{1}(a=3) \\ \mathbf{1}(a=4) \\ \mathbf{1}(b=1) \\ \mathbf{1}(b=2) \\ \mathbf{1}(b=3) \\ \mathbf{1}(b=4) \\ \mathbf{1}(b=5) \\ \mathbf{1}(b=6) \end{pmatrix} \text{ and } \beta'=(int., a1, ..., a4, b1, ..., b6).$$

The sample size is $n=24$,

$\hat{y}=\mathbf{X}'\hat{\beta}=0.27-0.031\mathbf{1}(a=1)-0.021\mathbf{1}(a=2)+\cdots-0.231\mathbf{1}(b=5)-0.131\mathbf{1}(b=6)$.

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) #1.
    (Output)
    (Intercept) a b
> a=factor(a)
> b=factor(b)
> lm(y∼a+b) #2.
    (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 # 1 ?

What is X and β for $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ in # 2 ?
What is the difference between outcomes # 1 and # 2 ?
Which is the way as in Example 3 ? #1 or #2 ?
What do you expect the estimates $\underline{before}$ **seeing output ?**

$>$ summary(lm(y$\sim$a+b)) # justify the answer to the question
Coefficients:
Estimate Std. Error t value $\Pr(>|t|)$ **= ?**

Model checking
Question. Does a given set of data fits the given model ?
Ans. Test
H_0: $Y = \beta X + \epsilon$ v.s. H_1: $Y \neq \beta X + \epsilon$.
There are two 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
$j = 1, ..., J_i$,
$i = 1, ..., m$, and
$X_{i1} = \cdots = X_{iJ_i}$, with $J_i > 1$ for some i,
and $X_{ij} \neq X_{kh}$ if $i \neq k$,
then a model lack-of-fit test of H_0: $\sigma_L = \sigma_E$ v.s. H_1: $\sigma_L \neq \sigma_E$
$\phi = \mathbf{1}(m_L/m_E > F_{df_L, df_E, \alpha})$, where
$m_E = \frac{1}{df_E}\sum_{i,j}(Y_{ij} - \overline{Y}_{i\cdot})^2$, (unbiased estimator of σ^2 under NID)
$m_L = \frac{1}{df_L}\sum_{i,j}(\overline{Y}_{i\cdot} - \hat{Y}_{ij})^2$, (unbiased estimator of σ^2 under NID and LR Model)
$df_E = \sum_i (J_i - 1)$ $(= n - m)$ and
$df_L = n - p - df_E$ $(= n - (p + df_E) = m - p$, df of residuals)
Here, we make use of
$$\sum_{i,j} Y_{ij}^2$$
$$= \sum_{i,j}(Y_{ij} - \overline{Y})^2 + \sum_{i,j}\overline{Y}^2$$
$$= \sum_{i,j}(Y_{ij} - \hat{Y}_{ij})^2 + \sum_{i,j}(\hat{Y}_{ij} - \overline{Y})^2 + \sum_{i,j}\overline{Y}^2$$
$$= \underbrace{\sum_{i,j}(Y_{ij} - \overline{Y}_{i\cdot})^2}_{m_E \ or \ m_L?} + \underbrace{\sum_{i,j}(\overline{Y}_{i\cdot} - \hat{Y}_{ij})^2}_{m_E \ or \ m_L?} + \sum_{i,j}(\hat{Y}_{ij} - \overline{Y})^2 + \sum_{i,j}\overline{Y}^2.$$
df: (n-m)+(m-p)+(p-1)+1.
We also make use of NID.
Second way. If there is no replication, add another term or another function to the model
$Y = \beta X + \epsilon$,
e.g., consider a new model
$Y = \beta X + \theta X^2 + \epsilon$ (or $Y = \beta X + \theta g(X) + \epsilon$),
and check whether $\theta = 0$, where θ is a $q \times 1$ vector.
That is, set
H_0: $\theta = 0$, v.s. H_1: $\theta \neq 0$.
(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: $\mathbf{C}\gamma = \theta = 0$.

$\hat\gamma = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$,

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

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.

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 Y depended on two factors:

(1) the concentration X_0 of monomer,

(2) the concentration X_1 of dimer.

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

> 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			

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
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) f=1/f 1-pf(f,2,2)
Conclusion Do not reject the model,
and the data fit the linear regression model.
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 ?
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

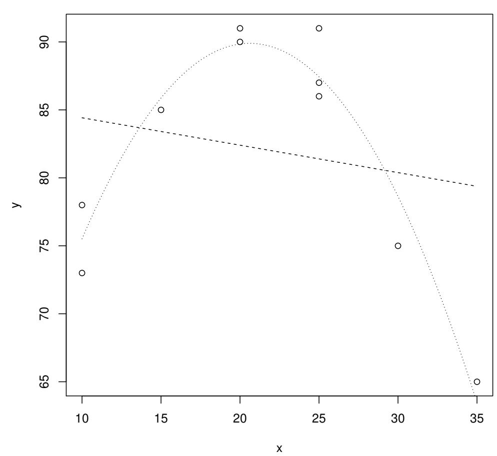
	$Res.Df$	RSS	Df	$Sum\ of\ Sq$	F	$Pr(>F)$
1	4	1.33075				
2	3	0.59044	1	0.74031	3.7615	0.1478

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(x^ 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(x^ 2)+I(x^ 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)
```

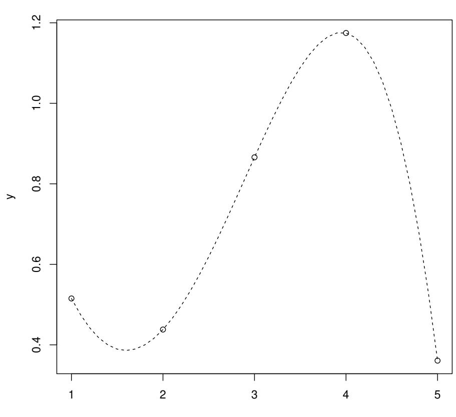
	Df	*Sum Sq*	*Mean Sq*	*F value*	$Pr(>F)$	
x	1	24.50	24.50	3.4608	0.1122	
$I(x^2)$	1	641.20	641.20	90.5674	$7.677e-05$	$***$ **Conclusion ?**
$I(x^3)$	1	2.71	2.71	0.3834	0.5585	
Residuals	6	42.48	7.08			

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.



Example 6. Simulation studies on testing H_0: $Y = \beta X + W$ with the R codes
(summary(y~x+I(sin(x)))$coef[3,4]> 0.05) # test $\phi = \mathbf{1}(p - value \leq 0.05)$
Ideally it tests H_0: $Y = \beta X + W$ v.s. H_1^o: $Y \neq \beta X + W$,
But it **actually** sets H_o v.s. H_1: $Y = \beta X + \theta \sin X + W$, with $\theta \neq 0$ **under NID**.
Simulation 1.
True model: $Y = X + \epsilon$, where $\epsilon \sim N(0,1)$ and $X \sim bin(3, 0.5)$.

Questions:

Is H_1^o true ? How about H_1 ? Is H_0 true ? How about NID ?
What do you expect for the test ?
Sample size= 50, replication= 1000, $\beta = 1$, $\hat{\beta}$= 0.996, sd= 0.17
rate of accepting right H_0 is 0.952. $\hat{P}(H_1|H_0) = 0.048$.
Does the test work as expected ?
What do you expect if $n = 5000$?

Simulation 2.

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

Questions:

Is H_1^o true ? How about H_1 ? Is H_0 true ? How about NID ?
What do you expect for the test ?
Sample size= 50, replication= 100, $\beta = 0$, $\hat{\beta} = -0.003$, sd= 0.03
rate of accepting wrong H_0 is 0.33. $\hat{P}(H_0|H_1) = $?
Does the test work in this case ?
What do you expect if $n = 5000$?

Simulation 3.

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

Questions:

Is H_1^o 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, $\hat{\beta}$= 0.5150, sd= 0.1761
rate of accepting wrong H_0 is 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 ?

Simulation 4.

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

Questions:

Is H_1^o true ? How about H_1 ? Is H_0 true ? How about NID ?
What do you expect for the test ?
Sample size= 5000, replication= 100, $\beta = 0$, $\hat{\beta}$= 0.0149, sd= 0.0141
rate of accepting wrong H_0 is 0.96. $\hat{P}(H_0|H_1) = 0.96$?
Does the test work in this case ?

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 slope = 0
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*	$Pr(>\|t\|)$
$(Intercept)$	-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: $Y = \beta \sin X + W$, H_1: $Y = \beta \sin X + \theta X + W$, with $\theta \neq 0$.

with 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_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. $\hat{P}(H_0|H_1) = 0.795$

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.

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.

For instance, $\xi = \theta g(\mathbf{X})$ 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 $F_{\mathbf{X},Y}$, 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 $n \to \infty$ ($P(H_0|H_1) \to 0$, a consistent test).

In simulations 3 and 4, both H_0 and H_1 are false,

the test can reject H_0 with probability 0 or 0.96.

In Example 7, both H_0 and H_1 are false, as $E(W|X)$ 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 $P(H_1|H_0) = 0.05$.

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 $P(H_1|H_0) \approx 0.33$.
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 terms.

Appendix. The marginal distribution (MD) approach. We shall introduce a new approach for model checking, which can do better than the previous tests most of the time.
A.1. Preliminary. We assume that

$(\mathbf{X}_1, Y_1)$, ..., $(\mathbf{X}_n, Y_n)$ are i.i.d. observations from $F_{\mathbf{X},Y}$, with density function $f_{\mathbf{X},Y}$, where $\mathbf{X}$ is a p-dimensional random vector and Y is a response variable.
Let $F_{Y|\mathbf{X}}$ be the conditional cdf with density function $f_{Y|\mathbf{X}}$. Denote $F_o = F_{Y|\mathbf{X}}(\cdot|\mathbf{0})$, which is called the baseline cdf of $F_{Y|\mathbf{X}}$.

A common regression model is the linear regression (LR) model,

$$Y = \beta'\mathbf{X} + W, \text{ where } (F_W = F_o),\ \beta \in \mathcal{R}^p \tag{1.1}$$

(the p-dimensional Euclidean space), and β' is the transpose of β. The coordinates of $\mathbf{X}$ can be dependent *e.g.*, $\mathbf{X} = (Z, Z^2, ..., Z^p)'$, where Z is a random variable. The LR model is often formulated by

$$Y = \alpha + \beta'\mathbf{X} + \epsilon, \text{ where } E(\epsilon|\mathbf{X}) = 0. \tag{1.2}$$

If the conditional variance $Var(W|\mathbf{X})$ does not depend on $\mathbf{X}$, it is called an ordinary linear regression (OLR) model, otherwise, it is called a weighted linear regression (WLR) model.
Remark 1. *Advantages that the LR model is specified by Eq. (1.1) rather than (1.2) are:*
(1) Eq. (1.2) but not (1.1) requires that $E(Y|\mathbf{X})$ *exists;*
(2) In general, β *but not* α *is identifiable under censorship models (Yu and Wong (2002));*
(3) It is often less important to estimate α *than* β*, the effect of the covariate* $\mathbf{X}$ *on* Y*.*

Under the OLR model, there are several consistent estimators of β if $F_{\mathbf{X},Y} \in \Theta_{lse}$, where

$$\Theta_{lse} = \{F_{\mathbf{X},Y}\text{: } \Sigma_{\mathbf{X}} \text{ is non-singular and } Cov(\mathbf{X}, Y) \text{ exists}\}, \tag{1.3}$$

and $\Sigma_{\mathbf{X}}$ is the $p \times p$ covariance matrix of $\mathbf{X}$. They include
the semi-parametric MLE (SMLE) (if F_o is discontinuous),
the modified SMLE (MSMLE) (see Yu and Wong (2002, 2003 and 2004)),
the least squares estimator (LSE) and
the quantile or median regression estimator.
Yu and Wong (2002) show that
the MSMLE is still consistent if $E(\ln f_W(W))$ exists, and
the SMLE and the MSMLE $\tilde\beta$ satisfy $P(\tilde\beta \neq \beta$ infinitely often$) = 0$ if the cdf F_W is discontinuous.

However, the LSE is inconsistent if $E(|Y||\mathbf{X}) = \infty$.

Given $F_{\mathbf{X},Y} \in \Theta$, the family of all joint cdf of $(\mathbf{X}, Y)$, $F_o = F_{Y|\mathbf{X}}(\cdot|\mathbf{0})$ is well defined, even if $(\mathbf{X}, Y)$ does not satisfy the linear regression model in H_0: $Y = \beta'\mathbf{X} + W$, where W is a random variable that its mean may not exist. We first consider the test of H_0. Let

$$\Theta_0 = \{F_{\mathbf{X},Y} :\ Y = \beta'\mathbf{X} + W, \text{ where } W \perp \mathbf{X},\ \beta \text{ and } F_W \text{ are unknown}\} \tag{2.1}$$

$(F_W = F_o)$. Then H_0: $F_{\mathbf{X},Y} \in \Theta_0$. The next lemma characterizes various LR model and motivating the MD approach for the LR model.

Lemma 1. $F_{Y|\mathbf{X}}$ *is a function of* (F_o, β), $F_Y(t) = E(F_{Y|\mathbf{X}}(t|\mathbf{X}))$. *If* $F_{\mathbf{X},Y} \in \Theta_0$*, then* $F_{Y|\mathbf{X}}(t|x) = F_o(t - \beta'x)$.

For convenience, we write $F_Y(t) = F_Y(t; \beta)$, as F_Y is a function of the unknown parameter β. Given β and $F_{\mathbf{X},Y}$, which may or may not belong to the LR model, define another random variable

$$Y^* = \beta'\mathbf{X} + W^*, \text{ where } F_{W^*}(\cdot) = F_{Y|\mathbf{X}}(\cdot|\mathbf{0}) \text{ and } \mathbf{X} \perp W^*. \tag{2.2}$$

By Lemma 1, the cdf of Y^* is

$$F_{Y^*}(t)\ (= F_{Y^*}(t;\beta)) = E(F_o(t-\beta'\mathbf{X}))\ \text{(denoted also by } F_{Y^*}(t;\beta)). \tag{2.3}$$

Theorem 1. *If $F_{\mathbf{X},Y} \in \Theta_0$ (see Eq. (2.1)), then*
(a) $F_o(\cdot) = F_{Y|\mathbf{X}}(\cdot|\mathbf{0}) = \mathbf{F}_{\mathbf{Y}^*|\mathbf{X}}(\cdot|\mathbf{0})$, (b) $F_{Y|\mathbf{X}} = F_{Y^*|\mathbf{X}}$, and (c) $F_Y = F_{Y^*}$.
If $F_{\mathbf{X},Y} \in \Theta \setminus \Theta_0$, then (e) $F_o(\cdot) = F_{Y|\mathbf{X}}(\cdot|\mathbf{0}) = \mathbf{F}_{\mathbf{Y}^*|\mathbf{X}}(\cdot|\mathbf{0})$, and (d) $F_{Y|\mathbf{X}} \neq F_{Y^*|\mathbf{X}}$.

Notice that if $F_{\mathbf{X},Y} \in \Theta_0$ as in (2.1), $E(Y|\mathbf{X})$ may not exist.

Corollary 1. (1) $F_{\mathbf{X},Y} \in \Theta_0$ iff $F_{Y|\mathbf{X}} = F_{Y^*|\mathbf{X}}$; (2) $F_{\mathbf{X},Y} \in \Theta_0 => F_Y = F_{Y^*}$.

Corollary 1 motivates the MD plot and the MD test. Given data $(\mathbf{X}_i, Y_i)$'s from $F_{\mathbf{X},Y}$, if $F_{\mathbf{X},Y} \in \Theta_0$ in (2.1), then β in $F_{Y^*}(t;\beta)$ is uniquely determined by $F_{\mathbf{X},Y}$. It is often that β in $F_{Y^*}(t;\beta)$ can also be uniquely determined by $F_{\mathbf{X},Y}$ even if $F_{\mathbf{X},Y} \notin \Theta_0$, such as in the case that $F_{\mathbf{X},Y} \in \Theta_{lse}$ (see (1.3)). One estimates β by the LSE if one feels confident that $\Theta_p = \Theta_{lse}$, or by the modified semi-parametric MLE (MSMLE) otherwise. In this course, we only use the LSE.

A.2. The MD plot. The edf of $F_Y(t)$ is $\hat{F}_Y(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(Y_i \leq t)$. We call the 95% pointwise confidence interval of $F_Y(t)$, *i.e.*, $\hat{F}_Y(t) \pm 1.96\sqrt{\hat{F}_Y(t)(1-\hat{F}_Y(t))/n}$, the confidence band (CB) of F_Y. The MD plot is

to plot $y = \hat{F}_{Y^*}(t)$ and $y = \hat{F}_Y(t)$, or together with the 95% CB of F_Y,

or to plot $y = \hat{S}_{Y^*}(t)$ and $y = \hat{S}_Y(t)$, or the CB of S_Y, where $\hat{F}_{Y^*}(t) = \frac{1}{n}\sum_{i=1}^n \hat{F}_o(t-\hat{\beta}'\mathbf{X}_i)$, $\hat{\beta}$ is a consistent estimator of β, $\hat{F}_o$ is a consistent estimator of F_o to be introduced in Eq. (2.5), $S_Y = 1-F_Y$, and $\hat{S}_Y = 1-\hat{F}_Y(t)$, etc. If the two curves are close, in particular, if the curve of $y = \hat{F}_{Y^*}(t)$ lies within the CB of F_Y, then it suggests that the model does fit the data. If the curve of $y = \hat{F}_{Y^*}(t)$ lies outside the CB of F_Y, then it suggests that the model does not fit the data.

The key of our new approach is to construct an estimator of the baseline cdf F_o, say $\hat{F}_o$, which satisfies that for each t, $\hat{F}_o(t) \xrightarrow{P} F_o(t)\ \forall\ F_{\mathbf{X},Y} \in \Theta$. We now explain how to construct the estimators $\hat{F}_o$ and $\hat{F}_{Y^*}$. Since $F_o = F_{Y|\mathbf{X}}(\cdot|\mathbf{0})$, it is desirable that $f_{\mathbf{X}}(\mathbf{0}) > \mathbf{0}$, where $f_{\mathbf{X}}$ is the density function of $\mathbf{X}$, though this is not always true for the given $f_{\mathbf{X}}$. However, since

$$\beta'\mathbf{X} + W = \beta'(\mathbf{X}-a) + \beta'a + W, \text{ and } W \perp \mathbf{X} \text{ iff } W \perp (\mathbf{X}-a), \tag{2.3}$$

without loss of generality (WLOG), we shall assume hereafter that the zero vector satisfies

$$f_{\mathbf{X}}(\mathbf{0}) > \mathbf{0} \text{ and } Y_1, ..., Y_m \text{ are the } Y_i\text{'s where } ||\mathbf{X}_i|| \leq \delta_n,\ \delta_n \to 0\ (e.g.,\ \delta_n = cn^{\frac{-1}{3p}} \tag{2.4}$$

$c = r/2$ and $r = \max_{i,j} ||\mathbf{X}_i - \mathbf{X}_j||$) and $||\cdot||$ is a norm. Otherwise, let $\mathcal{M}$ satisfy $f_{\mathbf{X}}(\mathcal{M}) > 0$, $\check{\mathbf{X}} = \mathbf{X} - \mathcal{M}$ and $\check{W} = \beta'\mathcal{M} + W$, hence $Y = \beta'\mathbf{X} + W$ yields $Y = \beta'\check{\mathbf{X}} + \check{W}$ and $f_{\check{\mathbf{X}}}(\mathbf{0}) = \mathbf{f}_{\mathbf{X}}(\mathcal{M}) > \mathbf{0}$. Moreover, it is desirable, though not necessary, that $\mathcal{M}$ is a mode of $f_{\mathbf{X}}$, so that there are more $\mathbf{X}_i$'s in the neighborhood of $\mathcal{M}$. Recall that $\mathcal{M}$ is a (population) mode of $F_{\mathbf{X}}$ if $P(||\mathbf{X}-\mathcal{M}|| < \delta) \geq P(||\mathbf{X}-a|| < \delta)\ \forall$ small $\delta > 0$ and $\forall\ a \in \mathcal{R}^p$.

Remark 3 (estimating the mode $\mathcal{M}$). Given a random sample of size n with $\beta \in \mathcal{R}^p$, in constructing $\hat{F}_o$, one can estimate $\mathcal{M}$ by $\hat{\mathcal{M}}$ first and then modify $\mathbf{X}_i$ by $\check{\mathbf{X}}_i = \mathbf{X}_i - \hat{\mathcal{M}}$. There are several ways to estimate $\mathcal{M}$. One way is $\hat{\mathcal{M}} = argmax_a \#\{\mathbf{X}_i : ||\mathbf{X}_i - a|| \leq \delta_n\}$ (see Eq. (2.4)), that is, the number of elements in the set $\{\mathbf{X}_i : ||\mathbf{X}_i - a|| \leq \delta_n\}$ is maximized by letting $a = \hat{\mathcal{M}}$. Another way is to first construct a $p-$ dimensional grid with $\lfloor n^{\frac{1}{3}} \rfloor$ cells as follows. Here $\lfloor x \rfloor$ is the largest integer k satisfying $k \leq x$. Suppose the data are contained in a p dimensional box $B = [l_1, r_1] \times \cdots \times [l_o, r_p]$, break each interval $[l_i, r_i]$ into $\lfloor n^{\frac{1}{3p}} \rfloor$ equal intervals. Then there are roughly $\lfloor n^{\frac{1}{3}} \rfloor$ cells in the grid, say B_1, ..., $B_{\lfloor n^{\frac{1}{3}} \rfloor}$. Let B_j be the cell that the number of elements X_i's in B_j achieves the largest value among all possible cells. Then let the center of B_j be $\hat{\mathcal{M}}$. The second way is more convenient. If $n = 100$ and

$p = 3$, the number of elements in the cell B_j would be at least 20. If $p \in \{1, 2\}$, then a mode can be estimated by plotting $\mathbf{X}_i$'s and finding where the data are more concentrated. Notice that both $\mathcal{M}$ and $\hat{\mathcal{M}}$ may not be unique. Hereafter, by Eq. (2.3), WLOG, we assume $\hat{\mathcal{M}} = \mathbf{0}$.

Under the assumption in (2.4), the edf $\hat{F}_o$ based on Y_1, ..., Y_m is a consistent estimator of $F_o(t)$ for all $F_{\mathbf{X},Y} \in \Theta$. Let $\hat{F}_{Y^*}$ be the edf based on the $n \times k$ pseudo observations $\hat{Y}_{ij} = \hat{\beta}'\mathbf{X}_i + Y_j$, $i \in \{1, 2, ..., n\}$ and $j \in \{1, ..., m\}$, where $\hat{\beta}$ is a consistent estimator of β.

$$\hat{F}_{Y^*}(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{F}_o(t - \hat{\beta}'\mathbf{X}_i) = \frac{\frac{1}{n^2}\sum_{i,j} \mathbf{1}(Y_i + \hat{\beta}'\mathbf{X}_j \le t, ||\mathbf{X}_i|| \le \delta_n)}{\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}(||\mathbf{X}_i|| \le \delta)} \tag{2.5}$$

$$(\hat{F}_o(t) = \frac{\sum_i \mathbf{1}(Y_i \le t, ||\mathbf{X}_i|| \le \delta_n)}{\sum_i \mathbf{1}(||\mathbf{X}_i|| \le \delta_n)} \qquad (\delta_n \text{ is as in (2.4))).}$$

Remark 4. *One may wonder whether a naive estimator of F_o is the edf $\check{F}_o$ based on $\hat{W}_i$'s $(= Y_i - \hat{\beta}'\mathbf{X}_i)$. This $\check{F}_o$ is a consistent estimator of F_o if H_0 in Eq. (2.1) is true. Then F_{Y^*} can be estimated by $\check{F}_{Y^*}(t) = \frac{1}{n}\sum_{i=1}^{n} \check{F}_o(t - \hat{\beta}'\mathbf{X}_i)$. The drawback of this naive approach is that if H_0 in Eq. (2.1) is false then $\check{F}_o$ is not consistent. In both Examples 4.1 and 4.2, $\check{F}_{Y^*}$ suggests that the data fit the incorrect models Θ_0. Moreover, it requires $E(|Y||\mathbf{X}) < \infty$. Thus it does not serve our purpose of a diagnostic tool.*

If the curve of $\hat{F}_{Y^*}(t)$ lies either entirely outside or entirely inside the confidence band of $\hat{F}_Y(t)$, then the indication is quite clear. Otherwise, it is quite subjective to say whether the two curves are close. Thus it is desirable to derive certain statistical tests.

A.3. The MD test The MD plotting method leads to a class of tests of H_0: $F_{\mathbf{X},Y} \in \Theta_0$, as follows.

$$T_1 = \int |\hat{F}_Y(t) - \hat{F}_{Y^*}(t)| d\hat{F}_Y(t), \; T_2 = \sup_t |\hat{F}_Y(t) - \hat{F}_{Y^*}(t)|, \tag{2.6}$$

$T_3 = \int \mathcal{W}(t)(\hat{F}_Y(t) - \hat{F}_{Y^*}(t))dG(t)$, or $T_4 = \int \mathcal{W}(t)|\hat{F}_Y(t) - \hat{F}_{Y^*}(t)|^k dG(t)$, where $k \ge 1$, $\mathcal{W}(\cdot)$ is a weight function, and dG is a measure, *e.g.*, dt, $d\hat{F}_o$, $d\hat{F}_Y$ and $d\hat{F}_{Y^*}(t)$. These tests are really testing

$$H_0^{MD}\text{: } F_Y = F_{Y^*}, \text{ v.s. } H_1^{MD}\text{: } F_Y \ne F_{Y^*}, \tag{2.7}$$

where Y^* is defined in Eq. (2.2).

Definition. The tests T_1, ..., T_4 in Eq. (2.6) are called the MD tests.

The percentiles of these T_j's can be estimated by two ways:

A. Derive the asymptotic distribution of these T_i's (see Theorem 3 and §3.6);

B. Make use of the modified bootstrap method as follows.

b1. In view of Remark 3 (estimation of the mode $\mathcal{M}$) and in view of Eq. (2.3), WLOG, we can assume that $\hat{\mathcal{M}} = \mathbf{0}$.

b2. Obtain $\hat{\beta}$, an estimator of β based on $(\mathbf{X}_i, Y_i)$'s under H_0, such as the LSE if it is sure that $F_{\mathbf{X},Y} \in \Theta_{lse}$, or the SMLE if there exist ties in the data, otherwise, the MSMLE.

b3. Take a random sample of size m from the $\mathbf{X}_i$'s in a neighborhood of $\mathbf{0}$, say $N(\mathbf{0}, \delta_\mathbf{n})$, where m and δ_n are as in (2.4), and take another random sample of size $n - m$ from the $\mathbf{X}_i$'s outside $N(\mathbf{0}, \delta_\mathbf{n})$. It yields a sample of $\mathbf{X}_i$'s, say $\mathbf{X}_1^{(1)}$, ..., $\mathbf{X}_n^{(1)}$.

b4. If there is no tie in $\hat{W}_i$'s $(= Y_i - \mathbf{X}_i'\hat{\beta}$'s), construct a continuous distribution function $\tilde{F}_o$ satisfying $\tilde{F}_o(t) = \hat{F}_o(t)$ at the discrete points of $\hat{F}_o$. Otherwise, set $\tilde{F}_o = \hat{F}_o$.

b5. Generate a random sample of size n from $\tilde{F}_o$, say, $W_1^{(1)}$, ..., $W_n^{(1)}$.

b6. Let $Y_i^{(1)} = \hat{\beta}'\mathbf{X}_i^{(1)} + W_i^{(1)}$, $i = 1$, ..., n.

b7. Now, obtain a value of T_1, say $T_1^{(1)}$, based on $(X_i^{(1)}, Y_i^{(1)})$'s and Eq. (2.6).

b8. Repeat the steps b3, ..., b7 a large number of times, say 100 times, obtain $T_1^{(j)}$ for $j = 2$, ..., 100. Thus the desired percentile can be estimated by the edf of these $T^{(j)}$'s.

The MD tests are valid tests of H_0^{MD}: $F_Y = F_{Y^*}$ against H_1^{MD}: $F_Y \ne F_{Y^*}$. It is worth mentioning that even when H_0 in Eq. (2.1) fails and $E(|Y||\mathbf{X}) = \infty$, the asymptotic distribution of the MD test still holds. In particular, if H_0 is not true but $F_{Y^*} = F_Y$, the

MD test would make type I error for testing H_o^{MD} with probability (w.p.) p_o and type II error for testing H_0 in (2.1) w.p.$(1 - p_o)$, where p_o is the size of the MD test. This is not the case for all existing tests.

Example 2.1. We generated data (X_i, Y_i), $i = 1, ..., n$ from the Cox model $h(t|X) = h_o(t)\exp(X)$, where $h_o = 1(t \geq 0)$, $X \sim U(-4/k, 4)$, $k \approx n^{0.7}$, and n is between 60 and 300. We fitted the data to the OLR (or WLR model), that is, H_0: $Y = \beta X + W$ (& $X \perp W$).

The Cox model does not belong to any LR model. The gam test is invalid, as $X \not\perp Y - \beta X - E(Y|X)$. The t-test is also invalid.

For such data with a sample size $n = 200$, the residual plots (see panels (1,2) and (1,3) in Figure 1) and the MD plot (see panel (2,1)) suggest that the OLR model may not fit the data, but a WLR model with a weight function $\sqrt{|(X-4)^3\mathbf{1}(X<3.7)+(X-4.5)^3\mathbf{1}(X\geq 3.7)|}$ might work (see the residual plot in panel (2,2)). However, the MD plot (see panel (2,3)) suggests that the WLR model does not fit the data neither. Thus the MD plots are better.

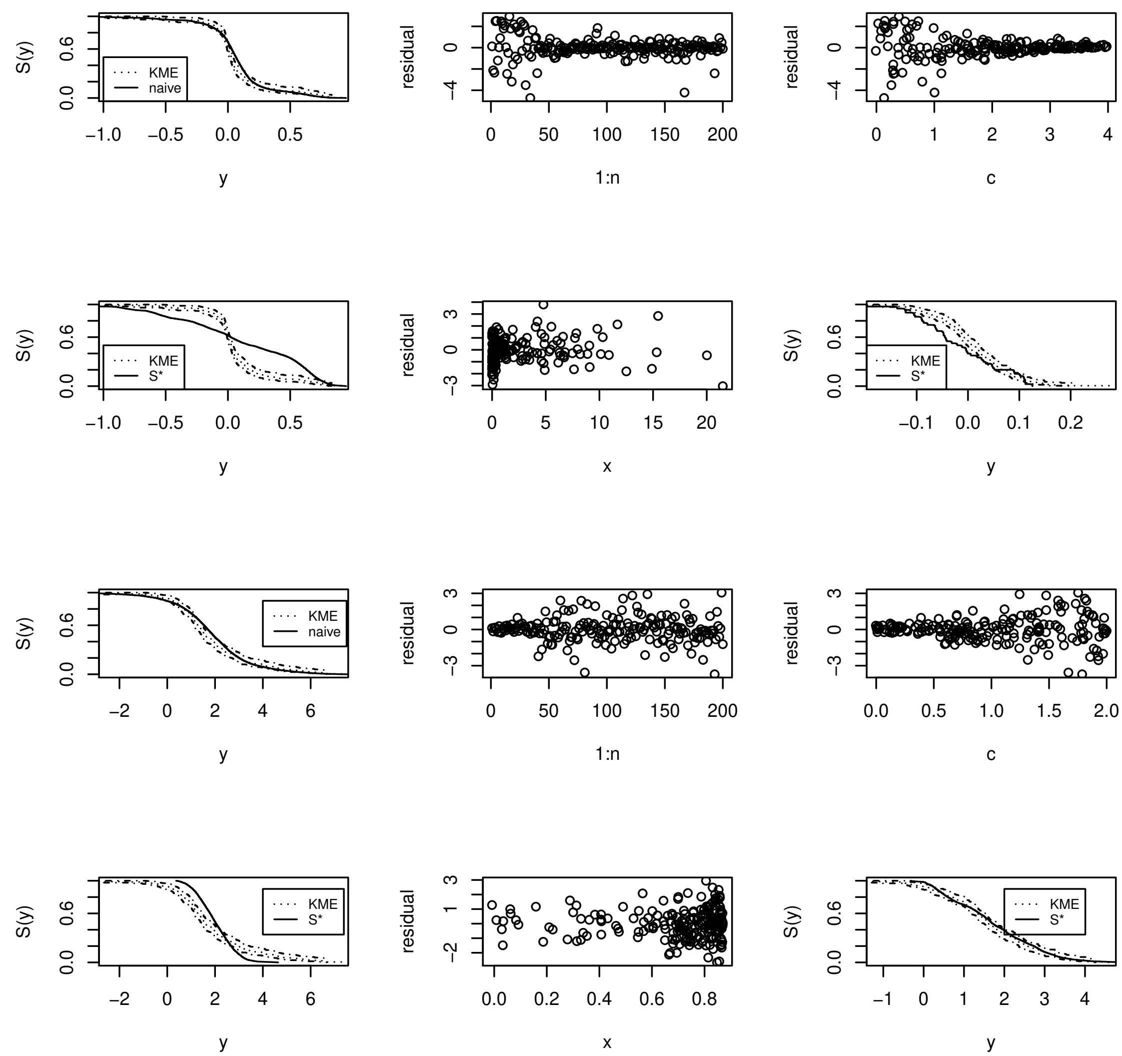
The simulation results suggest that the MD test T_1 performs very well for testing the incorrect OLR model, even when $n = 60$. The MD test can detect that the data do not fit the WLR model for large sample sizes such as $n \geq 200$.

For comparison sake, we also generated random samples from another WLR model:

$$Y = X + W, \text{ where } W \perp X,\ W \sim N(1, X + 0.3) \text{ and } X \sim U(0, 2).$$

Under this model, we carried out two sets of simulation studies. We first fitted the data to the OLR model $Y = \beta X + W$, where $W \perp X$. The residual plots (see panels (3,2) and (3,3)) and the MD plot (see panel (4,1)) of Figure 1 suggest that the OLR model does not fit the data, but a WLR model might work (see panels (4,2) and (4,3)).

The naive estimator $\check{S}^*$ ($= 1 - \check{F}_{Y^*}$ see Remark 4) suggests that the data from the Cox model and from the WLR model all fit the OLR model (see panels (1,1) and (3,1)). Thus it is useless. We also applied the same three tests to the WLR model. Since the data were from the WLR model, thus we estimated $P(H_0|H_1)$ for fitting the OLR model and $P(H_1|H_0)$ for fitting the WLR model, where H_0^2: the model is the WLR model v.s. H_1^2: H_0^2 is not true. The simulation results are presented in the bottom half of Table 1.



Cox data	naive $\check{S}^*$ for OLR	(i, residual)	(X_i,residual) for OLR
	MD plot for OLR	(X_i,residual) for WLR	MD plot for WLR
WLR data	naive $\check{S}^*$ for OLR	(i, residual)	(X_i,residual) for OLR
	MD plot for OLR	(X_i,residual) for WLR	MD plot for WLR

Figure 1. Residuals and MD plots under the Cox Model or the WLR model

Model:		OLR			WLR		
Data	Test:	T_1	SS	gam	T_1	SS	gam
	n	$\hat{p}_{0\|1}$	$\hat{p}_{0\|1}$	$\hat{p}_{0\|1}$	$\hat{p}_{0\|1}$	$\hat{p}_{0\|1}$	$\hat{p}_{0\|1}$
Cox	60	0.01	0.06	1.00	0.78	0.96	1.00
	200	0.00	0.00	1.00	0.19	0.95	1.00
	300	0.00	0.00	1.00	0.02	0.94	1.00
		$\hat{p}_{0\|1}$	$\hat{p}_{0\|1}$	$\hat{p}_{0\|1}$	$\hat{p}_{1\|0}$	$\hat{p}_{1\|0}$	$\hat{p}_{1\|0}$
WLR	60	0.03	0.00	0.59	0.04	0.05	0.08
	120	0.00	0.00	0.60	0.04	0.05	0.07

$\hat{p}_{0|1}$ is the estimate of $P(H_0|H_1)$ and $\hat{p}_{1|0}$ is the estimate of $P(H_1|H_0)$

Table 1. Simulation Results in Example 4.1

Remark. The idea for generating random numbers for a continuous distribution F_X:

$F_X(X) \sim U(0,1)$. Let $Y \sim U(0,1)$, $F_X^{-1}(Y) \sim F_X$.

Example 8. Suppose that F is a piecewise uniform distribution on $(0,1)$ and $(3,4)$ with weights 1/4 and 3/4. A pseudo random number of $n = 10$ can be generated as follows.

```
> n=10
> x=runif(n)
> m=length(x[x<0.25])
> y=runif(m)
> z=runif(n-m)+3
> y [1] 0.08246115 0.76996953
> z [1] 3.848005 3.442600 3.142384 3.670791 3.537500 3.897043 3.558773 3.388922
```

Example 9. Suppose that F is piecewise uniform on $(0, 0.5)$ and $(3,6)$ with weights 1/5 and 3/5 and $F(x) = 1 - 0.2e^{-x+7}$ if $x > 7$. That is $F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.4x & \text{if } x \in [0, 0.5] \\ 0.2 & \text{if } x \in (0.5, 3) \\ 0.2 + 0.2(x-3) & \text{if } x \in [3, 6] \\ 0.8 & \text{if } x \in (6, 7) \\ 1 - 0.2e^{-x+7} & \text{if } x > 7 \end{cases}$

Thus $F^{-1}(t) = \begin{cases} t/0.4 & \text{if } t \in [0, 0.2] \\ \frac{t-0.2}{0.2} + 3 & \text{if } t \in (0.2, 0.8] \\ 7 - \ln\frac{1-t}{0.2} & \text{if } t \in (0.8, 1] \end{cases}$

9 pseudo random numbers can be generated as follows.

```
> (x=sort(runif(9)))
    [1] 0.01509044 0.03312090 0.19840396 0.28440890 0.33304866 0.35577466 0.48100012
    [8] 0.59806993 0.85603151
> y=x
> (k=ceiling(x*5)) # Why ×5 ?
    [1] 1 1 1 2 2 2 3 3 5
> (u=x[k==1]*2.5)
    [1] 0.03772610 0.08280224 0.49600990
> (v=7-log(5*(1-x[k==5])))
    [1] 7.328723
> (x=x[k>1&k<5])
    [1] 0.2844089 0.3330487 0.3557747 0.4810001 0.5980699
> round(c(u,(x-0.2)*5+3,v),2)
    [1] 0.04 0.08 0.50 3.42 3.67 3.78 4.41 4.99 7.33
> y=c(y[k==1]*2.5, 5*(y[k>1&k<5]-0.2)+3, 7-log(5*(1-y[k==5])))
> round(y,2)
    [1] 0.04 0.08 0.50 3.42 3.67 3.78 4.41 4.99 7.33
```

Remark. Given a n distinct Y_i, their edf is $\hat{F}(t) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(Y_i \le t)$. WLOG, assume that $Y_1 < \cdots < Y_m$. A linear interpolation to the discrete $\hat{F}$ is

$$\begin{aligned} \tilde{F}(t) =& \frac{1}{m}\sum_{i=1}^{m}[\frac{t-(Y_i-\epsilon)}{\epsilon}\mathbf{1}(t \in (Y_i-\epsilon, Y_i)) + \mathbf{1}(Y_i \le t)] \\ =& \begin{cases} \frac{j-1}{m} + \frac{t-Y_j+\epsilon}{m\epsilon} & \text{if } t \in (Y_j-\epsilon, Y_j],\ j \in \{1,...,m\} \\ \cdots & \text{if } \cdots, \end{cases} \end{aligned}$$

where $\epsilon = \min_{i<j\le m} |Y_i - Y_j|$,

It seems that there are some errors in my comments in the previous homework solution.

Chapter 4. Comparing a number of entities

4.1. Analysis of Variance (ANOVA)

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

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

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

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

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

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

Remark. The expression of the model is not unique.
(1) $Y_{tj} = \tau_t + \epsilon_{tj}$, $t = 1, ..., I$ and $j = 1, ..., J$.

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

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

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

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

For testing

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

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

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

due to NID and

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

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

The data are

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

, $I = 4,\ J = 6$,

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

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

$(Intercept)$	$treatment2$	$treatment3$	$treatment4$
$6.100e+01$	$5.000e+00$	$7.000e+00$	$-9.999e-15$
$\overline{Y}_{1.}$	$\overline{Y}_{2.} - \overline{Y}_{1.}$	$\overline{Y}_{3.} - \overline{Y}_{1.}$	$\overline{Y}_{4.} - \overline{Y}_{1.}$

> anova(obj)

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

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

	$Estimate$	$Std.\ Error$	$t\ value$	$Pr(>\|t\|)$
$treatment1$	61.0000	0.9661	63.14	$< 2e-16***$
$treatment2$	66.0000	0.9661	68.32	$< 2e-16***$
$treatment3$	68.0000	0.9661	70.39	$< 2e-16***$
$treatment4$	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	$< 2e-16***$
$treatment2$	$5.000e+00$	$1.366e+00$	3.660	$0.00156**$
$treatment3$	$7.000e+00$	$1.366e+00$	5.123	$5.18e-05***$
$treatment4$	$-1.000e-14$	$1.366e+00$	0.000	1.00000

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

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

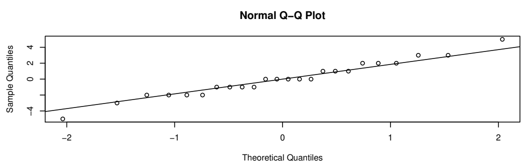
> a=lm(x∼factor(treat)-1)
> summary(a)   # compare "Estimate" in these two summaries.

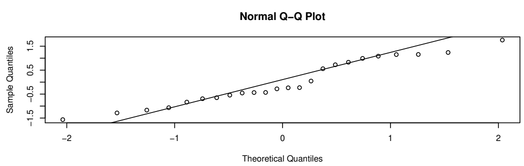
	Estimate	*Std. Error*	*t value*	$Pr(>\|t\|)$
$factor(treat)1$	61.0000	0.6667	91.50	$< 2e-16***$
$factor(treat)2$	66.0000	0.9428	70.00	$< 2e-16***$
$factor(treat)3$	68.0000	0.9428	72.12	$< 2e-16***$

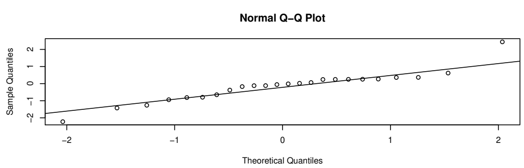
> anova(a) Analysis of Variance Table

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

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







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

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

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

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

We shall do three tests:

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

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

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

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

> 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 ?

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

> aov(x∼treatment+row)

	$treatment$	row	$Residuals$
$Sum\ of\ Squares$	228.0	41.5	70.5
$Deg.\ of\ Freedom$	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] +66treatment[2]+ 68treatment[3]

Derive the LSE directly for two way anova:

$Y_{ij} = \eta + \alpha_i + \gamma_j + \epsilon_{ij}$, $i \in \{1, ..., I\}$, $j \in \{1, ..., J\}$,
$\sum_i \alpha_i = \sum_j \gamma_j = 0$ (contr.sum) (the simplist way).

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

$\overline{Y} = \eta + \overline{\epsilon}$, $=> \hat{\eta} = \overline{Y}$; (due to MME).
$\overline{Y}_{i\cdot} = \eta + \alpha_i + \overline{\epsilon}_{i\cdot}$, $i \in \{1, ..., I\}$, $=> \hat{\alpha}_i = \overline{Y}_{i\cdot} - \overline{Y}$;
$\overline{Y}_{\cdot j} = \eta + \gamma_j + \overline{\epsilon}_{\cdot j}$, $j \in \{1, ..., J\}$, $=> \hat{\gamma}_j = \overline{Y}_{\cdot j} - \overline{Y}$;
$\hat{Y}_{ij} = \hat{\eta} + \hat{\alpha}_i + \hat{\gamma}_j = \overline{Y}_{i\cdot} + \overline{Y}_{\cdot j} - \overline{Y}$.

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

$(\hat{\eta}, \hat{\alpha}_1, ..., \hat{\alpha}_{I-1}, \hat{\gamma}_1, ..., \hat{\gamma}_{J-1})' = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ (contr.sum), $lm(y \sim row + col)$).

The LSE can also be derived by
$(\hat{\eta}, \hat{\alpha}_2, ..., \hat{\alpha}_I, \hat{\gamma}_2, ..., \hat{\gamma}_J)' = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ (default), $lm(y \sim row + col)$).

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 $(\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_I, \hat{\gamma}_2, ..., \hat{\gamma}_J)' = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ (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.

$lm(y \sim)$	$(Intercept)$	$col1$	$col2$	$row1$	$row2$	$row3$
$c + r$	$\overline{Y}_{1\cdot} + \overline{Y}_{\cdot 1} - \overline{Y}$		$\overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 1}$		$\overline{Y}_{2\cdot} - \overline{Y}_{1\cdot}$	$\overline{Y}_{3\cdot} - \overline{Y}_{1\cdot}$
sum	$\overline{Y}$	$\overline{Y}_{\cdot 1} - \overline{Y}$		$\overline{Y}_{1\cdot} - \overline{Y}$	$\overline{Y}_{2\cdot} - \overline{Y}$	
$c + r - 1$		$\overline{Y}_{1\cdot} + \overline{Y}_{\cdot 1} - \overline{Y}$	$\overline{Y}_{1\cdot} + \overline{Y}_{\cdot 2} - \overline{Y}$		$\overline{Y}_{2\cdot} - \overline{Y}_{1\cdot}$	$\overline{Y}_{3\cdot} - \overline{Y}_{1\cdot}$

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	< 1		0.486

What is (β, σ) ?

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 and $k = 4$ experiments were carried out for each blend.

First randomize the experiment by

sample(1:4, replace=F)

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

$blends\backslash treatments$	A	B	C	D
1	89	88	97	94
2	84	77	92	79
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)
(obj=lm(as.vector(x)~T+B))
anova(obj)
```

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			

Conclusion How many statements ?

$(70 + 264)/(3 + 4)/18.833 \approx 2.5$

> 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	.	**Which constraint ?**
$B1$	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		

The model can be simplified as $E(Y|X) = 86 + 6\mathbf{1}(B = 1)$, or

> lm(x∼ factor(B==1))

$(Intercept)\quad factor(B == 1)1$

$88.25\quad -3.75$

$$E(Y|X) = 88.25 - 3.75\mathbf{1}(B \neq 1) \underbrace{+}_{??} 3.75\mathbf{1}(B = 1).$$

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:

Drivers\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*

car and treatment effects are confounded

(2) Simple randomization:

Drivers\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*

due to the 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:

Drivers\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*

, which eliminates the block effects of cars and drivers, as each row and column has A, B, C, D

Compare

	1	2	3	4
I	*A*	*B*	*C*	*D*
II	*B*	*C*	*D*	*A*
III	*C*	*D*	*A*	*B*
IV	*D*	*A*	*B*	*C*

,

	1	2	3	4
I	*A*	*B*	*C*	*D*
III	*B*	*C*	*D*	*A*
II	*C*	*D*	*A*	*B*
IV	*D*	*A*	*B*	*C*

,

	1	2	3	4
I	*A*	*B*	*C*	*D*
II	*C*	*D*	*A*	*B*
III	*B*	*C*	*D*	*A*
IV	*D*	*A*	*B*	*C*

Relation between these 3 ?

The data are put in Table 2.

Drivers\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 3 ?

Table 1 **Table 2**

```
> 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(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	
col	3	24	8.000	1.5	0.307174	
row	3	216	72.000	13.5	0.004466	$**$
$Residuals$	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
$Residuals$	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) as row effect is significant, the model can be simplified as
$E(Y|X) = 23 - 8\mathbf{1}(Drive_3) - 5\mathbf{1}(Drive_4)$?

```
> 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|)	
$factor(D)1$	23.5000	0.9707	24.21	$3.37e-12$	$***$
$factor(D)3$	15.0000	1.3728	10.93	$6.38e-08$	$***$
$factor(D)4$	18.0000	1.3728	13.11	$7.17e-09$	$***$

The model is $E(Y|X) = 23.5\mathbf{1}(Driver\ 1\ or\ 2) + 15\mathbf{1}(Drive_3) + 18\mathbf{1}(Drive_4)$.

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 more equal-level factors with the same level as the treatment (2+1).
One can 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 3 rows of Latin square (5) does not work !

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.
Since $(4+1) \times (4-1) = 15$ dfs are needed, in additional to 1 df for $\overline{Y}$,
two replications are needed **why ??**
Thus 4 additional cycles and 4 additionals emory papers are needed.

The solution to homework 5 is in
http://people.math.binghamton.edu/qyu/ftp/556hwsol

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 ?

1. rep(sample(1:16), 2) ? Or
2. In each replication,
rep(sample(1:4),4) (for 4 pieces of each emory paper in cicles),

rep(sample(1:4),4) (for 4 pieces of each cloth in cicles) and
rep(sample(1:4),2) (for 4 circles).

Note:

In each cicle, 4 experiments are carried out simultaneously, it needs 4 emory papers and 4 pieces of cloth.

The data are as follows.

cycles\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: α, β, γ, δ

replication II
Cycles: c_5, c_6, c_7, c_8
Treatments: A, B, C, D
Holders: 1, 2, 3, 4
Emory paper sheet: ϵ, ξ, θ, κ

What is the property of the arrangement ?

Three Latin squares superimpose together.

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

Notice:

1. $(\alpha,\, A)$, $(\alpha,\, 1)$, $(A,\, 1)$, *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 can be uniquely determined.

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

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

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

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

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
> 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	

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

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(obj)

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$	
T	3	1705.3	568.4	5.3908	0.0212452	$*$
H	3	109.1	36.4	0.3449	0.7937901	
P	3	2217.3	739.1	7.0093	0.0099250	$**$
$C1$	1	3081.1	3081.1	29.2194	0.0004299	$***$
$C2$	1	4134.4	4134.4	39.2077	0.0001476	$***$
$C3$	1	2610.7	2610.7	24.7587	0.0007638	$***$
$C5$	1	968.0	968.0	9.1799	0.0142505	$*$
$C6$	1	1380.2	1380.2	13.0886	0.0055907	$**$
$C7$	1	2596.0	2596.0	24.6190	0.0007787	$***$
$E1$	1	12.5	12.5	0.1185	0.7385293	
$E2$	1	433.5	433.5	4.1110	0.0732215	.
$E3$	1	1656.7	1656.7	15.7115	0.0032854	$**$
$E5$	1	3081.1	3081.1	29.2194	0.0004299	$***$
$E6$	1	287.0	287.0	2.7221	0.1333649	
$E7$	1	638.0	638.0	6.0506	0.0361710	$*$
r	1	603.8	603.8	5.7259	0.0403664	$*$
$Residuals$	9	949.0	105.4			
C	?	?	?	23.35		
E	?	?	?	9.66		
$H+P+C+E+r$	19	23809.4	1253.126	11.88361		

Conclusion ?

```
> 1-pf(9.66,6,9)
[1] 0.00169513
> 1-pf(11.8,19,9)
[1] 0.0003312753

> 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)
> 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			

```
> 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			

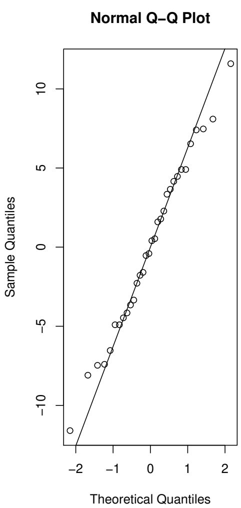
```
> anova(lm(y~T))
```

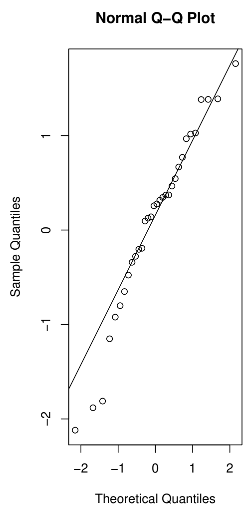
	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$
T	3	1705.3	568.45	0.6429	0.5939
$Residuals$	28	24758.6	884.24		

```
> 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		

```
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 p-value for T 0.02, or 0.008 or 0.6 ?
   It is significantly different from 1.
   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) $D > A > B > C$.
   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	$8.058e-06$

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:

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\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 ? 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)

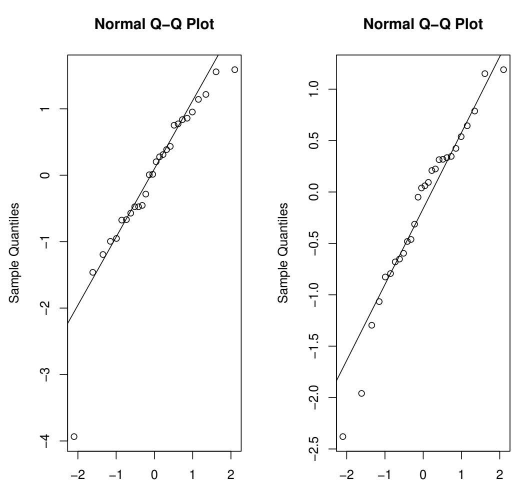
	$Estimate$	$Std.\ Error$	$t\ value$	$Pr(> \|t\|)$	
$(Intercept)$	361.50	20.89	17.309	$6.61e-14$	$***$
TB	210.00	29.54	7.110	$5.17e-07$	$***$
TC	−111.00	29.54	−3.758	0.001157	$**$
TD	−129.50	29.54	−4.384	0.000259	$***$
TE	−176.75	29.54	−5.984	$6.13e-06$	$***$
TF	190.00	29.54	6.433	$2.24e-06$	$***$
TG	−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 ?**
> 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 $n =$?
$\hat\beta_B = \frac{\sum_{i=1}^n y_i \mathbf{1}(T_i=B)}{\sum_{i=1}^n \mathbf{1}(T_i=B)}$, ...
$cov(\hat\beta_A, \hat\beta_B) = E(\hat\beta_A \cdot \hat\beta_B) - E(\hat\beta_A)E(\hat\beta_B)$.
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*	$Pr(>\|t\|)$
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=$(\mathbf{X}'\mathbf{X})^{-1}$.

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**.
Remark. The formulation of the previous table is given as follows

binary number :	0	1	10	11	100	101	110	111
rewritten	000	001	010	011	100	101	110	111
reverse	$---$	$+--$	$-+-$	$++-$	$--+$	$+-+$	$-++$	$+++$
decimal :	0	1	2	3	4	5	6	7
formulation :	1	2	3	4	5	6	7	8

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	$+$	$+$	$+$	$N\ +$	$N\ +$	$Slightly\ ++$
(= order of experiments ?)						
compare to columns				1	3	1&2 *both* $++$

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	$79^{(7)}$	$81^{(15)}$

5.5. Calculation of main effect.
Definition: Main effect of each factor $= \overline{y}_+ - \overline{y}_-$ (see the next tables).

Main effect of T:

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

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			
				$\overline{y}_+$	−	$\overline{y}_-$	$= -5$

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]

$$\begin{matrix} (Intercept) & a & b & c \\ 64.25 & 11.50 & -2.50 & 0.75 \end{matrix}$$ # model 1: $y = \mu + \beta_1 a + \beta_2 b + \beta_3 c + \epsilon$.

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

$$\begin{matrix} a & b & c \\ 23.00 & -5.00 & 1.50 \end{matrix}$$ # main effects

3rd way and the prefer way
> lm(y∼factor(a)+factor(b)+factor(c))$coef[1:4]

$$\begin{matrix} (Intercept) & factor(a)1 & factor(b)1 & factor(c)1 \\ 54.5 & 23.0 & -5.0 & 1.5 \end{matrix}$$ # main effects

Here 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]

$$\begin{matrix} (Intercept) & factor(a)1 & factor(b)1 & factor(c)1 \\ 64.25 & -11.50 & 2.50 & -0.75 \end{matrix}$$ **Here 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:

Given the LSE by the fourth way, how to get the LSE under model 2 (the 3rd way).

Reason: Explanation when $n = 2 = 2^1$ and $y \sim a$ (2nd way) $\mathbf{X} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$,

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$
$$= (\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix})^{-1} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_n \end{pmatrix}$$
$$= \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix}^{-1} \begin{pmatrix} \sum_i Y_i \\ -Y_1 + Y_2 \end{pmatrix}$$
$$= \frac{1}{n} \begin{pmatrix} \sum_i Y_i \\ -Y_1 + Y_2 \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{n} \sum_i Y_i \\ \frac{1}{2} \frac{1}{n/2}(-Y_1 + Y_2) \end{pmatrix}$$
$$= \begin{pmatrix} \overline{Y} \\ \frac{1}{2}(\overline{Y}_+ - \overline{Y}_-) \end{pmatrix}$$

If one uses factor and contr.treatment $\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$

(not $\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$ or $\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}$ as some mistakes in your homework, which leads to NE)

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$
$$= (\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix})^{-1} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_n \end{pmatrix}$$
$$= \begin{pmatrix} n & 1 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_i Y_i \\ Y_n \end{pmatrix}$$
$$= \frac{1}{n-1} \begin{pmatrix} 1 & -1 \\ -1 & n \end{pmatrix} \begin{pmatrix} \sum_i Y_i \\ Y_2 \end{pmatrix}$$
$$= \begin{pmatrix} (\sum_i Y_i - Y_2) \\ (-\sum_i Y_i + nY_2) \end{pmatrix}$$

$= \begin{pmatrix} Y_1 \\ \overline{Y}_2 - \overline{Y}_1 \end{pmatrix}$
$= \begin{pmatrix} Y_1 \\ \overline{Y}_+ - \overline{Y}_- \end{pmatrix}$

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 $\beta X = \beta' X$.
$\hat{\sigma}^2$ 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 reponses
y_{ij}, $i = 1, ..., 8$ and $j = 1, ..., r$, let
$s_i^2 = \frac{1}{r-1}\sum_{j=1}^{r}(y_{ij} - \overline{y}_{i.})^2$, $i = 1, ..., 8$.
If $r = 2$,
$s_i^2 = \frac{(y_{i1}-\overline{y}_{i.})^2+(y_{i2}-\overline{y}_{i.})^2}{2-1} = \frac{(y_{i1}-y_{i2})^2}{2}$, $i = 1, ..., 8$,

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 $y \sim a + b + c + ab + bc + ac + abc$.
No, if under the submodel, *e.g.* $y \sim I(a * b * c)$
(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 ?
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 ?
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 ?
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 ?
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 is larger. $\nu\hat{\sigma}^2/\sigma^2 \sim \chi^2(\nu)$, mean $= \frac{\nu}{2} \cdot 2$ $(= \alpha\beta)$, variance $= \alpha\beta^2 =$? Thus $E(\hat{\sigma}^2) = \sigma^2$ and $V(\hat{\sigma}^2) = \sigma^4/\nu$.

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 SD $\sqrt{2 * Df}$

> w=lm(y~a+b)
> anova(w)

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

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

[1] 0.9183548 # $(= s^2)$

$\boldsymbol{\sigma^2 = 1?}$

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.

Reason: WLOG, assume that

$Y_{ij} = \beta_1 X_i + \beta_2 Z_i + \epsilon_{ij}$, $j = 1, 2$, 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:** $Y_i = \beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \epsilon_i$

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

What is β_0**'s ?**

What is β_1's ?
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

> 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			

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 (normally independently distributed) assumption and 2^3 factorial designs,

$T_o = \frac{\overline{Y}-\beta_0}{\sqrt{s^2/n}} \sim t_{df}$,

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

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

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

For the data in Table 5.3, df=8, $t_{8,0.025} \approx 2.3$, so a 95% confidence interval 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	*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* $\|effect\| \le s$ nearly or too small

> z=lm(y~a+b+c+ab+bc+ac+abc) # (a,b,c)=(T,C,K)
> 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			8			

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

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

> 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 example.

> (z=lm(y∼a*c))

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

> lm(y∼a+c+ac)

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

> 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 $(-1, 1, -1, 1, -1, 1, -1, 1)(-1, -1, -1, -1, -1, 1, -1, 1)' \neq 0 \qquad (X_a' X_{a:c} \neq 0)$

(3) However, the prediction of Y remains the same.

Under control.treatment, the LSE of the intercept is the estimate of the mean response of Y at low level of each factor.

The main effect of "a" or T is the estimate of the contribution of factor T at high level to the mean response Y, or more precise,
the estimate of the change due to factor T changing from the low level to the high level.

The discovery of the experiment is that the suppliers of catalyst K may cause some problem, as they were supposed to produce the same type of catalyst.

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.

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 $Y' = (y_1, ..., y_8)$

Latin Squares deal with $2 + 1$ factors (treatment + 2 block-factors), all with the k levels.
Graeco- Latin Squares deal with $2 + 1 + 1$ factors (treatment + (2+1) block-factors).
Hyper-Graeco- Latin Squares deal with m factors (treatment + (m-1) block-factors), $m \geq 5$.

k=3. Latin Squares : try $\begin{array}{cc} & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \\ 3 & ? & \end{pmatrix} \end{array}$ **Does it work ?**

$\begin{array}{cc} & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{pmatrix} \end{array}$ **is** $\begin{array}{cc} & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{pmatrix} \end{array}$ **a Latin square ?**

$$\rightarrow \text{Graeco-Latin square } \begin{array}{cc} & \begin{array}{ccc}1 & 2 & 3\end{array} \\ \begin{array}{c}1\\2\\3\end{array} & \begin{pmatrix}1,1 & 2,2 & 3,3\\ 2,3 & 3,1 & 1,2\\ 3,2 & 1,3 & 2,1\end{pmatrix}\end{array} \rightarrow \text{Hyper-Graeco-Latin square ?}$$

$$\text{k=5. } \begin{array}{cc} & \begin{array}{ccccc}1 & 2 & 3 & 4 & 5\end{array} \\ \begin{array}{c}1\\2\\3\\4\\5\end{array} & \begin{pmatrix}1 & 2 & 3 & 4 & 5\\ 2 & 3 & 4 & 5 & 1\\ 3 & 4 & 5 & 1 & 2\\ 4 & 5 & 1 & 2 & 3\\ 5 & 1 & 2 & 3 & 4\end{pmatrix}\end{array} \rightarrow \text{Graeco-Latin square } \begin{array}{cc} & \begin{array}{ccccc}1 & 2 & 3 & 4 & 5\end{array} \\ \begin{array}{c}1\\2\\3\\4\\5\end{array} & \begin{pmatrix}1,1 & 2,2 & 3,3 & 4,4 & 5,5\\ 2,3 & 3,4 & 4,5 & 5,1 & 1,2\\ 3,5 & 4,1 & 5,2 & 1,3 & 2,4\\ 4,2 & 5,3 & 1,4 & 2,5 & 3,1\\ 5,4 & 1,5 & 2,1 & 3,2 & 4,3\end{pmatrix}\end{array}$$

$$\text{Does } \begin{array}{cc} & \begin{array}{ccccc}1 & 2 & 3 & 4 & 5\end{array} \\ \begin{array}{c}1\\2\\3\\4\\5\end{array} & \begin{pmatrix}1,1 & 2,2 & 3,3 & 4,4 & 5,5\\ 2,3 & 3,4 & 4,5 & 5,1 & 1,2\\ 3,4 & 4,5 & 5,? & 1 & 2\\ 4,? & 5 & 1 & 2 & 3\\ 5,2 & 1,3 & 2 & 3 & 4\end{pmatrix}\end{array} \text{ work ?}$$

$$\rightarrow \text{Hyper-Graeco-Latin square } \begin{array}{cc} & \begin{array}{ccccc}1 & 2 & 3 & 4 & 5\end{array} \\ \begin{array}{c}1\\2\\3\\4\\5\end{array} & \begin{pmatrix}1,1,1,1 & 2,2,2,2 & 3,3,3,3 & 4,4,4,4 & 5,5,5,5\\ 2,3,4,5 & 3,4,5,1 & 4,5,1,2 & 5,1,2,3 & 1,2,3,4\\ 3,5,2,4 & 4,1,3,5 & 5,2,4,1 & 1,3,5,2 & 2,4,1,3\\ 4,2,5,3 & 5,3,1,4 & 1,4,2,5 & 2,5,3,1 & 3,1,4,2\\ 5,4,3,2 & 1,5,4,3 & 2,1,5,4 & 3,2,1,5 & 4,3,2,1\end{pmatrix}\end{array}$$

$$\text{k=4. } \textbf{Which is correct ?} \begin{array}{cc} & \begin{array}{cccc}1 & 2 & 3 & 4\end{array} \\ \begin{array}{c}1\\2\\3\\4\end{array} & \begin{pmatrix}1 & 2 & 3 & 4\\ 2 & 3 & 4 & 1\\ 3 & 4 & 1 & 2\\ 4 & 1 & 2 & 3\end{pmatrix}\end{array}, \text{ or } \begin{array}{cc} & \begin{array}{cccc}1 & 2 & 3 & 4\end{array} \\ \begin{array}{c}1\\2\\3\\4\end{array} & \begin{pmatrix}1 & 2 & 3 & 4\\ 2 & 1 & 4 & 3\\ 3 & 4 & 1 & 2\\ 4 & 3 & 2 & 1\end{pmatrix}\end{array} \rightarrow \text{G-L square}$$

$$\begin{array}{cc} & \begin{array}{cccc}1 & 2 & 3 & 4\end{array} \\ \begin{array}{c}1\\2\\3\\4\end{array} & \begin{pmatrix}1,1 & 2,2 & 3,3 & 4,4\\ 2,2 & 3,1 & 4,4 & 1,3\\ 3,3 & 4,4 & 1,1 & 2,2\\ 4,4 & 1,3 & 2,2 & 3,1\end{pmatrix}\end{array} \text{ or } \begin{array}{cc} & \begin{array}{cccc}1 & 2 & 3 & 4\end{array} \\ \begin{array}{c}1\\2\\3\\4\end{array} & \begin{pmatrix}1,1 & 2,2 & 3,3 & 4,4\\ 2,4 & 3,1 & 4,2 & 1,3\\ 3,2 & 4,3 & 1,4 & 2,1\\ 4,3 & 1,4 & 2,1 & 3,2\end{pmatrix}\end{array} \text{ or } \begin{array}{cc} & \begin{array}{cccc}1 & 2 & 3 & 4\end{array} \\ \begin{array}{c}1\\2\\3\\4\end{array} & \begin{pmatrix}1,1 & 2,2 & 3,3 & 4,4\\ 2,3 & 1,4 & 4,1 & 3,2\\ 3,4 & 4,3 & 1,2 & 2,1\\ 4,2 & 3,1 & 2,4 & 1,3\end{pmatrix}\end{array} ?$$

A1. Rewrite the codes in page 33 of my lecture notes using contr.treat. Compare the outputs on the LSE's of lm(z~ T) and obj. of your codes to the ones in my notes and make comments.

Note: The LSE's may not be the same, but the prediction and $\hat{\sigma}$ are.

A2. Simulation studies on the H-G-Latin square in pages 31-34 assuming NID:

1. Under the assumption that the H_o^r, ..., H_o holds (see page 35).

2. $Y = \sum_{i=1}^4 \sqrt{i}\mathbf{1}(T = i) + \sum_{j=1}^4 j\mathbf{1}(P = j) + \sum_{k=1}^4 \sin k\mathbf{1}(E = k) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$, and $\sigma = 2$.

Test H_o, the treatment effects are the same, first ignoring block factors and then with blocking factors.

In both cases, $n = 32$, repeat 100 times, then summarize the outcomes.

Note: There are 4 simulation studies, each with 100 times.

From page 35 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 p-value for T 0.02, or 0.008 or 0.6 ?
   It is significantly different from 1.
   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) $D > A > B > C$.
   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 ?

Codes for the 2nd model:

```
    c=gl(8,4,32)
    [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 8 8 8 8
    e=c(1,2,3,4,2,1,4,3,3,4,1,2,4,3,2,1)
    e=c(e,e)
    x=sin(as.numeric(C))+log(e)
for(i in 1:100){
    y=rnorm(32)+x # under H1
    lm(y~T)
    lm(y~T+H+P+C+E+r)
}
```

Answer the following questions:

Is H_0 or H_1 true ?

Is lm(y~T+P+C+E+H+r) true ?

Is lm(y~T) true ?

After the simulation loop ends,

what do you expect ?

LSE=$\hat{\beta}$? Relation between the LSE of the parameter β and their true values ?

Codes for the 1st model:

```
for(i in 1:100){
y=rnorm(32,1,2) # under H_0
lm(y~T)
lm(y~T+P+C+E+H+r)
}
```

Answer the following questions:

Is H_0 or H_1 true ?

Is lm(y∼T+P+C+E+H+r) true ?

Is lm(y∼T) true ?

After the simulation loop ends,

what do you expect ?

Relation between the LSE of the parameter β and their true values ?

For lm(y∼ T), it computes the LSE of the linear regression model:

$Y_{ij} = \eta+\alpha_i+\epsilon_{ij}$, $i = 1, ..., I$, $j = 1, ..., J$, or $Y_k = X_k'\beta+\epsilon_k$, $n = I \times J$, $\epsilon_k \sim N(0, \sigma^2)$.

What is $(\beta, X, \hat{\beta})$ (in control.sum) ?

Two forms: (1) $E(Y) = U\alpha$ for $Y_{ij} = \eta + \alpha_i + \epsilon_{ij}$, (2) for $\hat{\beta} = (X'X)^{-1}X'Y =$, where

$$\alpha = \begin{pmatrix} \eta \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \eta \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, U = \begin{pmatrix} W \\ W \end{pmatrix}, X = \begin{pmatrix} V \\ V \end{pmatrix},$$

$$W = \begin{pmatrix} 1&1&0&0&0 \\ 1&0&1&0&0 \\ 1&0&0&1&0 \\ 1&0&0&0&1 \\ 1&0&1&0&0 \\ 1&1&0&0&0 \\ 1&0&0&0&1 \\ 1&0&0&1&0 \\ 1&0&0&1&0 \\ 1&0&0&0&1 \\ 1&1&0&0&0 \\ 1&0&1&0&0 \\ 1&0&0&0&1 \\ 1&0&0&1&0 \\ 1&0&1&0&0 \\ 1&1&0&0&0 \end{pmatrix}, V = \begin{pmatrix} 1&1&0&0 \\ 1&0&1&0 \\ 1&0&0&1 \\ 1&-1&-1&-1 \\ 1&0&1&0 \\ 1&1&0&0 \\ 1&-1&-1&-1 \\ 1&0&0&1 \\ 1&0&0&1 \\ 1&-1&-1&-1 \\ 1&1&0&0 \\ 1&0&1&0 \\ 1&-1&-1&-1 \\ 1&0&0&1 \\ 1&0&1&0 \\ 1&1&0&0 \end{pmatrix}, \hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} \overline{Y} \\ \overline{Y}_{1\cdot} - \overline{Y} \\ \overline{Y}_{2\cdot} - \overline{Y} \\ \overline{Y}_{3\cdot} - \overline{Y} \end{pmatrix}$$

and $\hat{\alpha}$ has infinitely many solution, due to rank$(U) = 4 < 5$, and $\hat{\beta}$ leads to one $\hat{\alpha} = ...$

For $lm(y \sim T + P + C + E + H + r)$, ...

Remark for Homework 6.

A.1. Verify the previous statement by deriving the elements through the equations in the order given in the table.

$lm(y\sim)$	$(Intercept)$	$col1$	$col2$	$row1$	$row2$	$row3$
$c+r$	1		2		3	4
sum	$\overline{Y}$	$\overline{Y}_{\cdot 1} - \overline{Y}$		$\overline{Y}_{1\cdot} - \overline{Y}$	$\overline{Y}_{2\cdot} - \overline{Y}$	
$c+r-1$		5	6		7	8

In this exercise, $Y_{ij} = \eta+\alpha_i+\beta_j$, $i = 1, 2, 3$ and $j = 1, 2$. The outcomes of $lm(y \sim c+r)$ are different from $lm(y \sim r + c)$.