STAT 3119

Week 8: 10/15/2019 @GWU

Outline

- Block design: more than one blocking variable (Ch 21.6)
- Block design with more than one replication in each block (Ch 21.7)
- Factorial Treatment (Ch 21.8)
- Planning a RCBD Experiment (Ch 21.9)
- New topic for ANCOVA (chapter 22: not test in mid-term)
- `This Thursday: Mid-term review (chapters 15-21); Mid-term (10/24)`

Block Design: more than one blocking Variable (1) (Ch 21.6)

Sometimes, a substantial reduction in the experimental error variability can only be obtained by utilizing more than one variable for determining blocks. For instance, both age and gender may be related to the response and both need to be controlled in testing some treatment factor within the age/gender groups.

- If we have the age factor defined in young or old group, then in combination with gender, there will be four age-gender blocks: i.e., young male, old male, young female and old female.
- In general, block factor 1 has n_1 levels and block factor 2 has n_2 levels, then in combination, we will have $n_b = n_1 n_2$ block levels.
- Then the same ANOVA model used to model the data can be written as

$$Y_{ij} = \mu_{..} + \rho_i + \tau_j + \epsilon_{ij}$$

with $j = 1, ..., n_b (= n_1 n_2)$ levels. The total sample size will be $n_T = n_1 n_2 r$, with r= no. of treatment levels.

Block Design: more than one blocking Variable (2)

- The $n_b = n_1 n_2$ is large to randomize the set of r treatment within each block, e.g. there are two blocks, each with 4 levels, and there are four treatment levels, then $4 \times 4 \times 4 = 64$ experiment units are needed, compared to $n_T = 16$ (if only one block factor is used). If two blocks of six levels each, then there will be 36 blocks.
- In such cases, when two block factors have the same number of levels and equal to the number of treatments, the design table is like *square*, we may use a **Latin square designs** (Chapter 28), which permits the use of a much smaller number of experimental units while still preserving the full benefits of error variance reduction by using both all levels of blocking variables each .

Example for 4 by 4 Latin square:

Previously, in our tire brands example: we are testing the difference in four brands of tires (A, B, C, D) on tread wear.

- `Design 3 for Tire Brand Test : one block factor`

We put an restriction on the randomization, requires that each brand be used only once om one each car to get the **randomized complete block design**. Now four brands are placed on a car at random and each car gets one tire of each brand. This provides a more homogeneous condition to test the four brands of tires.

	Car			
	I	II	III	IV
Brand distribution	B	D	A	C
	C	C	B	D
	A	B	D	A
	D	A	C	B

- `Design 4 for Tire Brand Test : two block factors`
- However, there may be a possible position effect in testing the tire brands. Experiences show that rear tires get different wear than front tires and even different sides of the same car may show different amounts of tread wear. We would like to impose another restriction on the randomization in such a way that each brand is not only used once on **each car** but also only once in **each of the four possible positions**: left front, left rear, right front, right rear.
- Such design in which each treatment appears once and only once in each row (position) and once in each columns (cars) is called a **Latin square design**: two restrictions the on randomization. (*Note*: Design 3 is not balanced by the rows.)

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

- Latin square is not unique, say, you can randomly permute the columns or rows. There are other design layouts. For this given problem, you can choose a Latin square design table: then assign the cars or tire positions randomly to the rows and columns, and also assign the treatments (four brands) to letters A,B,C, D in a random fashion.
- It is hard to memorize the design tables. We can create a Latin Square Design in **R** for example with the function **design.lsd** of the add-on package **agricolae**.

```
Rpackage= "agricolae"
if (! Rpackage  %in% installed.packages()) install.packages(Rpackage)
```

```r
library(agricolae)
## Generate a 3 by 3 Latin square
design.lsd(LETTERS[1:3])$sketch
```

```
##      [,1] [,2] [,3]
## [1,] "B"  "C"  "A"
## [2,] "C"  "A"  "B"
## [3,] "A"  "B"  "C"
```

```r
## Generate a 4 by 4 Latin square
design.lsd(LETTERS[1:4])$sketch
```

```
##      [,1] [,2] [,3] [,4]
## [1,] "C"  "D"  "B"  "A"
## [2,] "A"  "B"  "D"  "C"
## [3,] "D"  "A"  "C"  "B"
## [4,] "B"  "C"  "A"  "D"
```

```r
## Generate a 5 by 5 Latin square
design.lsd(LETTERS[1:5])$sketch
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,] "C"  "A"  "E"  "B"  "D"
## [2,] "A"  "D"  "C"  "E"  "B"
## [3,] "E"  "C"  "B"  "D"  "A"
## [4,] "B"  "E"  "D"  "A"  "C"
## [5,] "D"  "B"  "A"  "C"  "E"
```

```r
## Generate a 6 by 6 Latin square
design.lsd(LETTERS[1:6])$sketch
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,] "B"  "D"  "F"  "E"  "A"  "C"
## [2,] "D"  "F"  "B"  "A"  "C"  "E"
## [3,] "E"  "A"  "C"  "B"  "D"  "F"
## [4,] "C"  "E"  "A"  "F"  "B"  "D"
## [5,] "F"  "B"  "D"  "C"  "E"  "A"
## [6,] "A"  "C"  "E"  "D"  "F"  "B"
```

- Similar ANOVA model can be used with two block factors

$$Y_{ijk} = \mu_{..} + \rho_i^{(1)} + \rho_j^{(2)} + \tau_k + \epsilon_{ijk}$$

- The Latin square design is balanced for two blocking factors, and our usual estimators and sums of squares are also "working". In *R*, we would use the model formula Y ~ Block1 + Block2 + Treatment.

Block design with more than one replication (Ch 21.7)

- In a **generalized randomized block design**, it is similar to a randomized block design except that d experimental units are assigned to each treatment within a block. This generalized design increases the size of a block from r units for a randomized block design to dr units. This design allows the interaction effects of the treatment and blocks to be investigated.
- As we shall demonstrate by an example, a *generalized randomized block design* is analyzed like an *ordinary two-factor study* where blocks are one factor. Hence, no new problems are encountered with a generalized randomized block design in testing for treatment effects or in estimating them. Now, we can calculate *MSE* and use it as an estimator of the error variance σ^2.

Model:

$$Y_{ijk} = \mu.. + \rho_i + \tau_j + (\rho\tau)_{ij} + \varepsilon_{ijk} \qquad \textbf{(21.10)}$$

where:

$\mu..$ is a constant
ρ_i, τ_j, are constants subject to the restrictions $\sum \rho_i = \sum \tau_j = 0$
$(\rho\tau)_{ij}$ are constants subject to the restrictions that the sums over any subscript are zero
ε_{ijk} are independent $N(0, \sigma^2)$
$i = 1, \ldots, n_b; j = 1, \ldots, r; k = 1, \ldots, d$

We shall refer to model (21.10) as the *generalized randomized block model*.

Example:

Table 21.4 contains the data for a single-factor experiment in which

- the *effects of distraction level* (factor A: low distraction, high distraction) on *the time required to complete a task* were studied, using eight **men** and eight **women**, Four men were assigned at random to each of the r = 2 treatments, and independently four women were assigned at random to each treatment.
- Here gender is the blocking variable. Each block contains eight persons, with four randomly assigned to each treatment within the block. $r = 2, n_b = 2, d = 4,$ then $n_T = 16$.

TABLE 21.4 **Data on Completion Times for Generalized Randomized Block Design with $d = 4$— Task Completion Example.**

	Block (gender)	
	Male	Female
Low Distraction:		
	12	3
	8	9
	7	5
	5	9
High Distraction:		
	14	11
	16	9
	15	10
	13	14

R analysis: 1. Read the data

```
# read data from weekly folder online
Ex21  =read.table(
        url("https://raw.githubusercontent.com/npmldabook/Stat3119/master/Week8/CH21TA04.txt"))

Ex21
```

```
##    V1 V2 V3 V4
## 1  12  1  1  1
## 2   8  1  1  2
## 3   7  1  1  3
## 4   5  1  1  4
## 5   3  2  1  1
## 6   9  2  1  2
## 7   5  2  1  3
## 8   9  2  1  4
## 9  14  1  2  1
## 10 16  1  2  2
## 11 15  1  2  3
## 12 13  1  2  4
## 13 11  2  2  1
## 14  9  2  2  2
## 15 10  2  2  3
## 16 14  2  2  4
```

```
names(Ex21) =  c("Time", "Gender", "Distraction", "Units")
```

```r
# make  categorical variables for factor A and B
Ex21$Gender    =  as.factor(Ex21$Gender)
Ex21$Distraction =  as.factor(Ex21$Distraction)

str(Ex21)
```

```
## 'data.frame':    16 obs. of  4 variables:
##  $ Time       : int  12 8 7 5 3 9 5 9 14 16 ...
##  $ Gender     : Factor w/ 2 levels "1","2": 1 1 1 1 2 2 2 2 1 1 ...
##  $ Distraction: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 2 2 ...
##  $ Units      : int  1 2 3 4 1 2 3 4 1 2 ...
```

2. Apply the regular two-way ANOVA model: treat gender as a block factor

```r
fit = aov(Time~ Gender + Distraction + Distraction:Gender , data= Ex21 )
summary(fit)
```

```
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## Gender              1     25      25   4.167 0.063851 .
## Distraction         1    121     121  20.167 0.000738 ***
## Gender:Distraction  1      4       4   0.667 0.430127
## Residuals          12     72       6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Findings: Here the residual MS is the MSE as in the textbook. The F-statistics are derived from the ratio of MSE and the MS (distraction) and MS (gender). For the significant level of .01,

1. The p-value for interaction =0.43, suggesting that blocks (gender) do not interact with treatments (distraction level).
2. The p-value for treatment effects (difference in distraction levels) is 0.00074, suggesting a significant treatment effect. The time to complete a task differ by the two distraction levels.
3. The p-value for block effect (gender difference) =0.06, which is not significant at the significant level of 0.01. There is no strong evidence suggesting the responses differ by gender.

Block design: Factorial Treatment (Ch 21.8)

- Randomized complete block designs can also be used when the treatments have a factorial structure. This means treatments contains the combination of all the levels of two or more factors, with the treatment levels $r = ab$ for two factors A (a levels) and B (b levels) . The factorial treatment levels are randomized within the blocks.
- Similar to the model for block design, now we can partition of the treatment effects into the effects due to factor A, B and their interactions in the model, and partition SSTR (df) into these 3 parts as well.
- Model and SS partition

When factorial treatments are employed, the ANOVA model can be modified by showing the component factor effects in place of the treatment effect. For a two-factor study, we have:

$$Y_{ijk} = \mu... + \rho_i + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \qquad \textbf{(21.11)}$$

where the terms in the model have the usual meaning and (j, k) corresponds to the treatment mean $\mu_{.jk}$. In the analysis of variance, we proceed as always by decomposing the treatment sum of squares *SSTR* into sums of squares for the factor main effects and interactions. This is shown in Table 21.5 for a two-factor study, the factors having a and b levels, respectively. The decomposition is done in the usual fashion, as explained in Section 19.4, utilizing the relation in (19.39):

$$SSTR = SSA + SSB + SSAB$$

Note:

$$df.Tr = r - 1 = ab - 1 \equiv df.A + df.B + df.AB = (a-1) + (b-1) + (a-1)(b-1)$$

- Data layout and ANOVA table

FIGURE 21.7
Layout for a Two-factor Study in a Randomized Complete Block Design.

	A_1		A_2	
	B_1	B_2	B_1	B_2
Block 1	Y_{111}	Y_{112}	Y_{121}	Y_{122}
2	Y_{211}	Y_{212}	Y_{221}	Y_{222}
3	Y_{311}	Y_{312}	Y_{321}	Y_{322}

TABLE 21.5
ANOVA Table for a Two-Factor Study in a Randomized Complete Block Design—Randomized Block Model (21.11).

Source of Variation	*SS*	*df*	*MS*
Blocks	*SSBL*	$n_b - 1$	*MSBL*
Treatments	*SSTR*	$r - 1$	*MSTR*
Factor *A*	*SSA*	$a - 1$	*MSA*
Factor *B*	*SSB*	$b - 1$	*MSB*
AB interactions	*SSAB*	$(a-1)(b-1)$	*MSAB*
Error	*SSBL.TR*	$(n_b - 1)(r - 1)$	*MSBL.TR*
Total	*SSTO*	$n_b r - 1$	

Note: $r = ab$

- Formulas (19.39a, b, c) are still appropriate for calculating the component SS. Tests for factor effects are conducted as usual, and no new problems are encountered in the estimation of fixed factor effects.
- In *R*, we would use the model formula Y ~ Block + A*B; or Y ~ Block + A+B +A:B to study the effect of treatments from both factors and their interactions.

Sample size design for RCBD (1)

- Planning the sample sizes for a randomized complete block design is very similar to that for a completely randomized design. Without the replications within blocks, calculation of the needed number of blocks n_b is similar to the calculation of the number of units (replications) needed per treatment for a balanced one factor study (chapter 16 or 17).
- Sample size planning is either based on (a) power and type I errror; or (b) the estimation precision. In either case, we need to specify or estimate in advance the experimental error variance σ^2.

Example 1: Power approach (page 910)

In the risk premium example (3 treatment levels), suppose that the number of blocks had not yet been determined and the experimenter desired the following risk protections:

1. Type I error is to be controlled at $\alpha = .05$.
2. If any two treatment means differ by three or more rating points, i.e., if the minimum range of the treatment means is $\Delta = 3$, the risk of concluding that there are no treatment effects should not exceed $\beta = .20$.
3. The experimenter anticipates that the experimental error standard deviation when executives are grouped by age will be approximately $\sigma = 2$.

The power is based on F-test, we can use similar R code as in Chapter 16.

```
Delta = 3
sigma = 2
r = 3

# specify the correct type I error and power
power.anova.test(groups = r, between.var = Delta^2/(2*(r-1)), within.var = sigma^2, power = 0.8, sig.le
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##          groups = 3
##               n = 9.636519
##     between.var = 2.25
##      within.var = 4
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

Results: We use the next smallest integer , $n = 10$. Thus, We find $n_b = 10$. Thus, the experimenter requires approximately 10 blocks of three executives each in order to obtain the desired protection against incorrect decisions.

Example 2: sample size based on precision (page 910)

For the risk premium example, all pairwise comparisons are of interest. The desired width of the confidence intervals is +/-1.5. The Tukey procedure is to be used with a 95 percent family confidence coefficient. A planning value of $\sigma = 2$ is reasonable.

Solution:

We can calculate the anticipated half-width of the confidence interval as $T\sigma(\hat{L})$ where T is the Tukey multiple

$$T = 1/\sqrt{2}\, q\big(1-\alpha; r, (n_b-1)(r-1)\big)$$

and the variance for the paired difference

$$\sigma(\hat{L}) = \sqrt{\sigma^2(1/n_b + 1/n_b)} = \sigma\sqrt{2/n_b}$$

We can try a series of sample sizes, then plug in the formula to see if it is adequate to achieve the estimate precision.

R example:

Since there are lots of *repeated* calculations, we can generate our own function with n_b as the input parameter.

```
CIwidth <- function(nb){
  r=3
  newdf= (nb-1)*(r-1)
  sigma= 2
  alpha=0.05

  # Tukey mutiple
  T =1/sqrt(2)* qtukey(1-alpha, nm=r, newdf)

  # SE(L)
  SE.L= sqrt(2/nb)*sigma

  # CI width
  unlist(list(T=T, SE.L= SE.L,  CIwidth= T*SE.L))
}
```

Then we can call this function with different input n_b to search a n_b to meet the estimation precision.

```
CIwidth(nb=10)
```

```
##         T      SE.L   CIwidth
## 2.5521631 0.8944272 2.2827241
```

```
CIwidth(nb=20)
```

```
##         T      SE.L   CIwidth
## 2.4388261 0.6324555 1.5424490
```

```
CIwidth(nb=21)
```

```
##         T      SE.L   CIwidth
## 2.4339193 0.6172134 1.5022476
```

```
CIwidth(nb=22)
```

```
##         T      SE.L   CIwidth
## 2.4294938 0.6030227 1.4650399
```

Results: n_b=22 blocks are required. (The textbook used $n_b = 21$ as a round-off answer.)

RCBD planning (2): Efficiency of the Blocking variable

- Once a randomized complete block experiment has been run, we often wish to estimate the efficiency of the blocking variable for guidance in future experimentation.
- This can be done by comparing the variances from the two difference design

$$E = \sigma_r^2/\sigma_b^2$$

  where σ_r^2 is the error variances from the completely randomized design (CRD), and σ_b^2 from a randomized complete block design (RCBD).
- The measure E indicates how much larger the replications need be with a completely randomized design as compared to a randomized complete block design in order that the variance of any estimated treatment contrast be the same.
- From the observed data, we can estimate the σ_r^2, then estimate the efficiency E as follows:

$$s_r^2 = \frac{(n_b - 1)MSBL + n_b(r-1)MSBL.TR}{n_b r - 1} \qquad \textbf{(21.13)}$$

Hence, we estimate E as follows:

$$\hat{E} = \frac{s_r^2}{MSBL.TR} = \frac{(n_b - 1)MSBL + n_b(r-1)MSBL.TR}{(n_b r - 1)MSBL.TR} \qquad \textbf{(21.14)}$$

- Since formula (21-14) is slightly overestimate E, there is a frequently used modification

$$\hat{E}' = \frac{(df_2 + 1)(df_1 + 3)}{(df_2 + 3)(df_1 + 1)}\hat{E} \qquad \textbf{(21.15)}$$

where

- df_1 = the degrees of freedom for estimating σ_r^2 in completely randomized design,
- df_1 = the degrees of freedom for estimating σ_b^2 in randomized complete block design.

Example (page 912) : Estimate the efficiency of blocking by age of executives in the risk premium example.

First we get the ANOVA table for RCBD:

TABLE 21.3 ANOVA Table for Randomized Complete Block Design—Risk Premium Example of Table 21.1.

Source of Variation	*SS*	*df*	*MS*
Blocks	171.3	4	42.8
Methods for risk premium specification	202.8	2	101.4
Error	23.9	8	2.99
Total	398.0	14	

Then

- From the Table, we get $MSBL = 42.8$, $MSBL.Tr =$ Error (residual) MS $= 2.99$, $nb = 5$ blocks, $r = 3$ treatment levels, then the efficiency

$$\hat{E} = 4(42.8) + 5(2)(2.99)/(14 * 2.99) = 4.8$$

- For the modified formula, $df_1 = (n_T - r) = 15 - 3 = 12$, $df_2 = (n_b - 1)(r - 1) = 8$,

$$\hat{E}' = (8 + 1)(12 + 3)/((8 + 3)(12 + 1)) * 4.8 = 4.5$$

Results: RCBD design is more efficient. Either formula gives similar conclusion. We would have required more than four times as many replications per treatment with a completely randomized design to achieve the same variance of any estimated contrast as is obtained with blocking by age.

New Topic : ANCOVA- Analysis of Covariance

Analysis of covariance (ANCOVA) is a technique that combines features of analysis of variance (ANOVA) and regression.

- The basic idea is to augment the analysis of variance model containing the factor effects with **one or more additional quantitative (continuous) variables** that are related to the response variable.
- This augmentation is intended to reduce the variance of the error terms in the model, i.e., to make the analysis and estimate the treatment effects more precise.
- In addition, by controlling the known related covariates (concomitant variable), we gain greater insight into the effects of the treatment factor on the response.

Examples

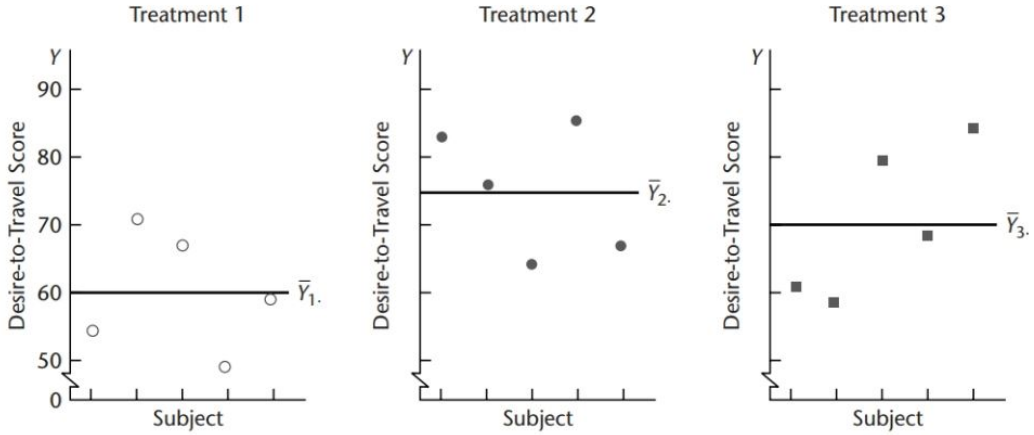
Example 1. Use covariates to reduce the large error term variance

Consider a study in which the effects of **three different films** (treatment factor) promoting travel in a state are studied.

- A subject receives an initial questionnaire to elicit information about the subject's attitudes toward the state.
- The subject is then shown one of the three five-minute films, and immediately afterwards is questioned about the film, about desire to travel in the state. The response variable of **desire-to-travel score** are collected.
- $n = 15$ subjects are recruited for this study.

FIGURE 22.1 Illustration of Error Variability Reduction by Covariance Analysis.

(a) Error Variability with Single-factor Analysis of Variance Model

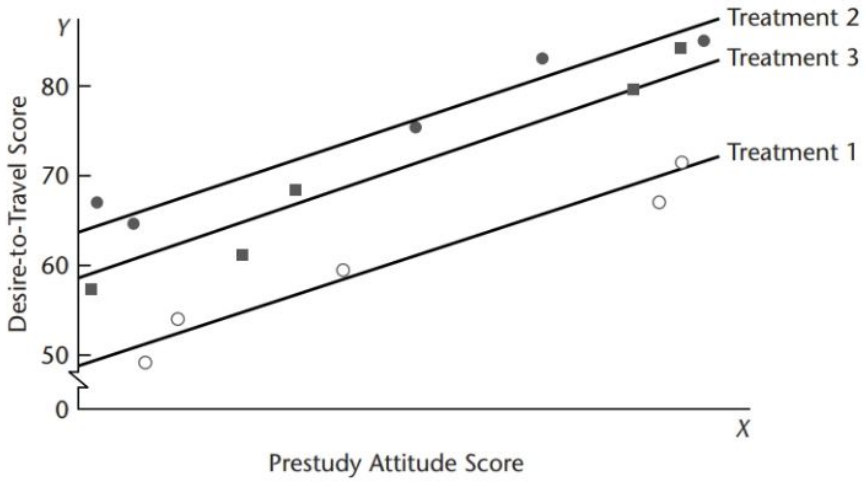


Q: What can we find from the scatterplot by treatment ($n = 5$ each) ?

If we use one-factor ANOVA model, it is evident that the error terms, as shown by the scatter around the estimated treatment means are fairly large.

- Suppose now that we can utilize also the subjects' initial attitude scores. We plot in the desire-to-travel score (obtained after exposure to the film) against the initial attitude score for each of the 15 subjects by the 3 treatment levels.

(b) Error Variability with Covariance Analysis Model



Q: What can we find from the plot?

Also note that the scatter around the treatment regression lines in Fig (b) is much less than the scatter in Fig (a) around the treatment means, as a result of the desire-to-travel scores being highly linearly related to the initial attitude scores. This shows the consider the covariate (initial attitude score) in the model will reduce the residual error variability and make the study more efficient for comparing treatment effects.

Choice of Covariates in ANCOVA

- We consider continuous covariate related to the response. If other categorical variables are considered, we will use the regular two-factor or multi-factor ANOVA model.
- Choice of covariates: If such variables have no relation to the response variable, nothing is to be gained by covariance analysis, and one might as well use a simpler ANOVA model.
- Covariates frequently used with human subjects include pre-study or baseline information, baseline age, lab values and socioeconomic status before the clinical treatments. When retail stores are used as study units, concomitant variables might be last period's sales or number of employees.
- For a clear interpretation of the results, a concomitant variable should be observed before the study; or if observed during the study, it should NOT be influenced by the treatments in any way. The interactions of treatment factor and covariate will make the interpretation more difficult.

Example of Single factor ANCOVA (Ch 22.3)

Example(page 926): A company studied the effects of **three different types of promotions** on *sales of its crackers*:

- Treatment 1: Sampling of product by customers in store and regular shelf space
- Treatment 2: Additional shelf space in regular location
- Treatment 3: Special display shelves at ends of aisle in addition to regular shelf space
- **Fifteen stores** were selected for the study, and a **completely randomized experimental design** was utilized. Each store was randomly assigned one of the promotion types, with five stores assigned to each type of promotion.
- Other relevant conditions under the control of the company, such as price and advertising, were kept the same for all stores in the study. Data on the **number of cases of the product sold during the promotional period (Y)**, and the **sales of the product in the preceding period (X- covariate of interest)** are presented in Table 22.1.

TABLE 22.1 Data—Cracker Promotion Example (number of cases sold).

Treatment	Store (j)									
	1		2		3		4		5	
i	Y_{i1}	X_{i1}	Y_{i2}	X_{i2}	Y_{i3}	X_{i3}	Y_{i4}	X_{i4}	Y_{i5}	X_{i5}
1	38	21	39	26	36	22	45	28	33	19
2	43	34	38	26	38	29	27	18	34	25
3	24	23	32	29	31	30	21	16	28	29

R analysis

1. Read the data

```
# read data from week5 folder online
Ex22  =read.table(
        url("https://raw.githubusercontent.com/npmldabook/Stat3119/master/Week8/CH22TA01.txt"))

Ex22
```

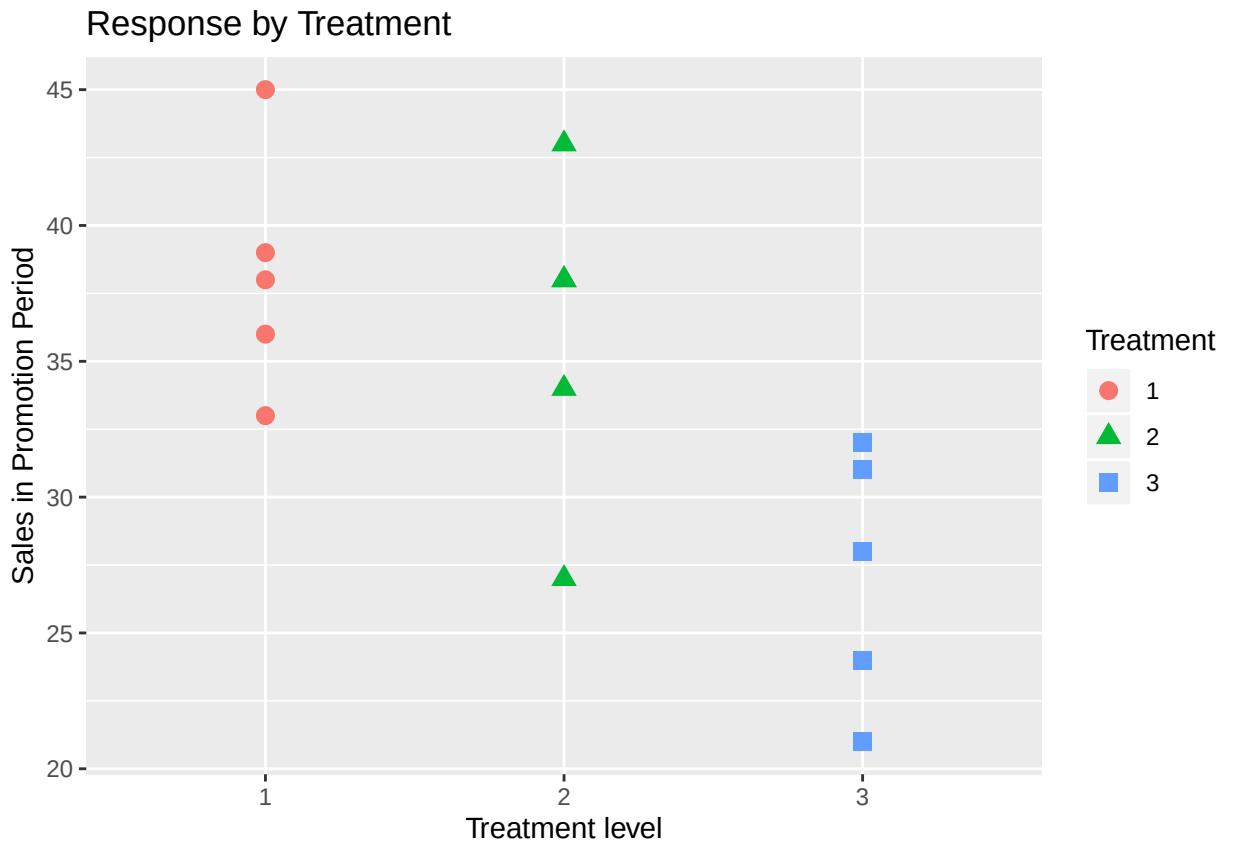
```
##    V1 V2 V3 V4
## 1  38 21  1  1
## 2  39 26  1  2
## 3  36 22  1  3
## 4  45 28  1  4
## 5  33 19  1  5
## 6  43 34  2  1
## 7  38 26  2  2
## 8  38 29  2  3
## 9  27 18  2  4
## 10 34 25  2  5
## 11 24 23  3  1
## 12 32 29  3  2
## 13 31 30  3  3
## 14 21 16  3  4
## 15 28 29  3  5
```

```
names(Ex22) =  c("New_Sales", "Old_Sales", "Treatment","Units")

# make  categorical variables for treatment
Ex22$Treatment =  as.factor(Ex22$Treatment)

Ex22
```

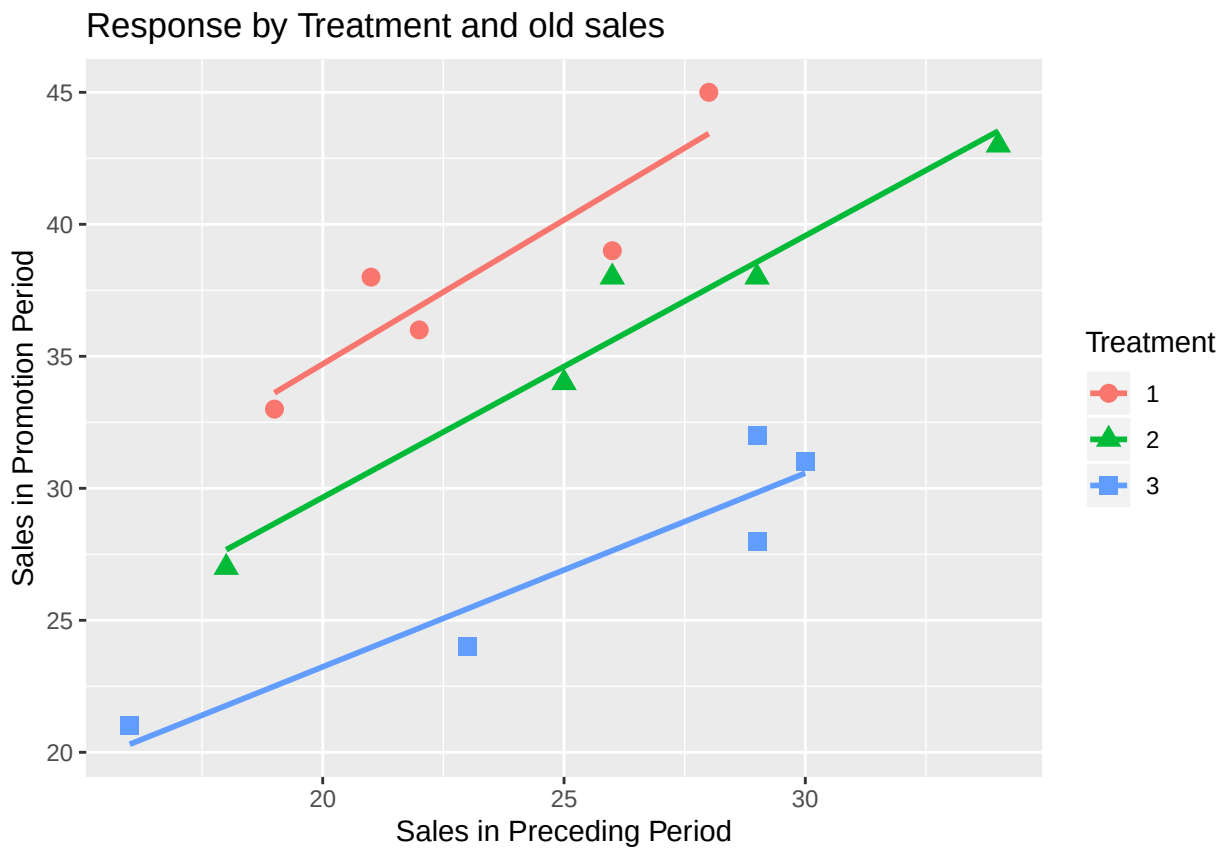
```
##    New_Sales Old_Sales Treatment Units
## 1         38        21         1     1
## 2         39        26         1     2
## 3         36        22         1     3
## 4         45        28         1     4
## 5         33        19         1     5
## 6         43        34         2     1
## 7         38        26         2     2
## 8         38        29         2     3
## 9         27        18         2     4
## 10        34        25         2     5
## 11        24        23         3     1
## 12        32        29         3     2
## 13        31        30         3     3
## 14        21        16         3     4
## 15        28        29         3     5
```

2. Read the data

```
library(ggplot2)

ggplot(data=Ex22, aes(x= Treatment, y= New_Sales,color= Treatment, shape=  Treatment  ))+
  geom_point(size=3) +
  labs(title="Response by Treatment",
       x="Treatment level", y="Sales in Promotion Period")
```




```
ggplot(Ex22, aes(x= Old_Sales, y= New_Sales, color= Treatment, shape=  Treatment ))+
  geom_point(size=3) +
  geom_smooth(method = lm, se= F)+
  labs(title="Response by Treatment and old sales",
       x="Sales in Preceding Period", y="Sales in Promotion Period")
```



3. compare the models with or without X

```
fit.noX <- aov(New_Sales~ Treatment, data =  Ex22)
summary(fit.noX)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Treatment    2  338.8  169.40   6.609 0.0116 *
## Residuals   12  307.6   25.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit.X <- aov(New_Sales~ Treatment + Old_Sales, data =  Ex22)
summary(fit.X)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Treatment    2  338.8  169.40   48.31 3.57e-06 ***
## Old_Sales    1  269.0  269.03   76.72 2.73e-06 ***
## Residuals   11   38.6    3.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit.noX, fit.X)
```

```
## Analysis of Variance Table
##
## Model 1: New_Sales ~ Treatment
## Model 2: New_Sales ~ Treatment + Old_Sales
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     12 307.600
## 2     11  38.571  1    269.03 76.723 2.731e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results:

- Residual error MSE in one factor ANOVA is 25.63 , compared to the residual error =3.51 in the ANCOVA model.
- There is a significant association of Sales in Preceding Period (old_sales) vs. Sales in Promotion Period (New_sales).
- The p-value to test the treatment effect was 0.0116 in ANOVA but p= 3.57e-06 in ANCOVA model, suggesting ANCOVA model can estimate the treatment effects more efficiently.

We will go over the statistical model, model diagnostics, inferences after the mid-term exam.

Summary this week

- Reading: Chapter 21 (RCBD)
- HW for Chapter 21: 21.5, 21.6, 21.14, 21.15, 21.18 (Due 10/24 by 6 pm before the class)