STAT 3119

Week 7: 10/8/2019 @GWU

Outline

- Two factor studies- one case per treatment ($n = 1$)
- Model and Estimation
- Tukey Test for Additivity
- Intro: Block design

Two-factor (Two-way) ANOVA model

In previous Chapter,

- The response variable Y is continuous
- There are two categorical explanatory variables (factors), called Factor A (a levels) and Factor B (b levels).
- A particular combination of levels is called a **treatment** or a **cell**. There are **ab** treatments.
- Y_{ijk}: index i denotes the level of the factor A, j denotes the level of the factor B, k denotes the kth observation for treatment (i, j), $k = 1$ to n, with $n > 1$:
- The df for the MSE (estimator of variance) is $ab(n - 1)$.

Here, due to constraints on cost, time, materials, etc, some studies only allow one replication/case per treatment

- $n = 1$, then $Y_{ij} \equiv Y_{ijk}$ (there is no need to use k in the subscript).
- We cannot use the full ANOVA model with A, B and interaction A:B in last Chapter, since df for MSE=0, i.e, we can no longer estimat the variance separately for each treament.
- The impact is that, if we fit a ANOVA model to the data, we will not be able to estimate the interaction terms; we will have to **assume** no interaction in the model.
- We will discuss "the Tukey test" to test this additive assumption, and remedial measure.

No Interaction model (Ch 20.1)

Factor Effects Model

$$Y_{ij} = \mu_{..} + \alpha_i + \beta_j + \epsilon_{ij}$$

- $\mu_{..}$ is the overall mean
- α_i is the main effect of A with $\sum_i \alpha_i = 0$.
- β_j is the main effect of B with $\sum_j \beta_j = 0$.

Because we have only one case per treatment, we do not have enough information to estimate the interaction in the usual way. We assume no interaction.

SS partition and ANOVA table

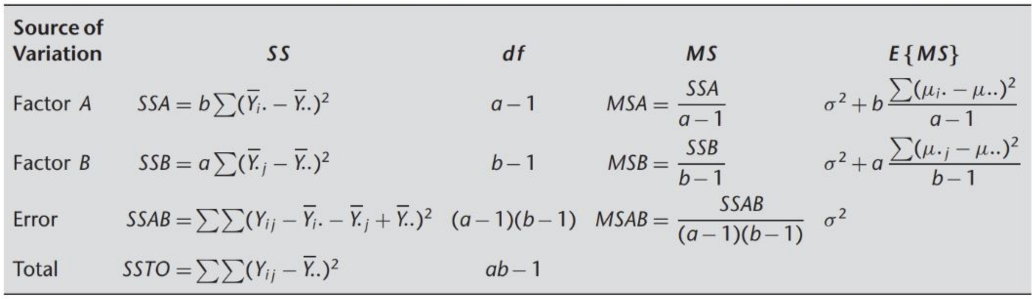
$$SS = SSA + SSB + SSAB$$

$$MSAB = SSAB/(a-1)(b-1)$$

may be used to estimate the variance σ^2 since we assume no interaction is present.

The ANOVA model for two-factor study with $n = 1$ per treatment is:

TABLE 20.1 **ANOVA Table for No-Interaction Two-Factor Model (20.1) with Fixed Factor Levels, $n = 1$.**



Source of Variation	SS	df	MS	$E\{MS\}$
Factor A	$SSA = b\sum(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$a-1$	$MSA = \frac{SSA}{a-1}$	$\sigma^2 + b\frac{\sum(\mu_{i\cdot} - \mu_{\cdot\cdot})^2}{a-1}$
Factor B	$SSB = a\sum(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2$	$b-1$	$MSB = \frac{SSB}{b-1}$	$\sigma^2 + a\frac{\sum(\mu_{\cdot j} - \mu_{\cdot\cdot})^2}{b-1}$
Error	$SSAB = \sum\sum(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2$	$(a-1)(b-1)$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	σ^2
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$	$ab-1$		

Hypothesis testing (utilize $MSAB$ in the denominator, instead of MSE):

- Testing Factor A main effect:

$$F^* = MSA/MSAB \sim F(a-1, (a-1)(b-1))$$

.

- Testing Factor B main effect:

$$F^* = MSB/MSAB \sim F(b-1, (a-1)(b-1))$$

.

- We reject the null hypothesis of no main effect (A or B) when the corresponding F^* is large.

Tukey Test for Additivity

- If we want to examine whether or not the two factors in a two-factor study interact, we can use *Tukey one degree of freedom test*.
- Without using all the df to estimate the interaction, we use the following special interaction model

$$Y_{ij} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + D\alpha_i\beta_j + \epsilon_{ij}$$

We use one degree of freedom to estimate D, leaving the rest of the df to estimate experiment error. It can be shown now that we can have the following SS partition

$$SSTO = SSA + SSB + SSAB^\star + SSRem^\star$$

with modified interaction SS

$$SSAB^{*} = \frac{\left[\sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}) Y_{ij}\right]^2}{\sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \sum_j (\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2} \qquad \textbf{(20.9)}$$

and

$$SSRem^{\star} = SSTO - SSA - SSB - SSAB$$

Then under the null $H_0 : D = 0$ (no interactions of present), the test statistic (Tukey test of additivity)

$$F^{*} = \frac{SSAB^{*}/1}{SSRem^{*}/(ab - a - b)} \sim F(1, ab - a - b)$$

Large values of F lead to the rejection of H_0 and the conclusion of H_a (interactions present).

Example of Two-factor (one case per treatment) study

Example (page 882): An analyst in an insurance commissioner's office studied the premiums for automobile insurance (Y) charged by an insurance company in six cities. The **six** cities were selected to represent different **regions of the state** and different **sizes of cities**. Table 20.2a shows the **amounts of three-month premiums** (Y) charged by the automobile insurance firm for a specific type and amount of coverage in a given risk category for each of the six cities, classified by

- size of city (factor A) : 3 levels (small, medium and large)
- geographic region (factor B): 2 levels (East, West)
- Note there is only one observation per cell, namely, the amount of the premium charged in the city for each factor level combination.
- The analyst wished to evaluate the effects of city size and geographic region on the amount of the premium.

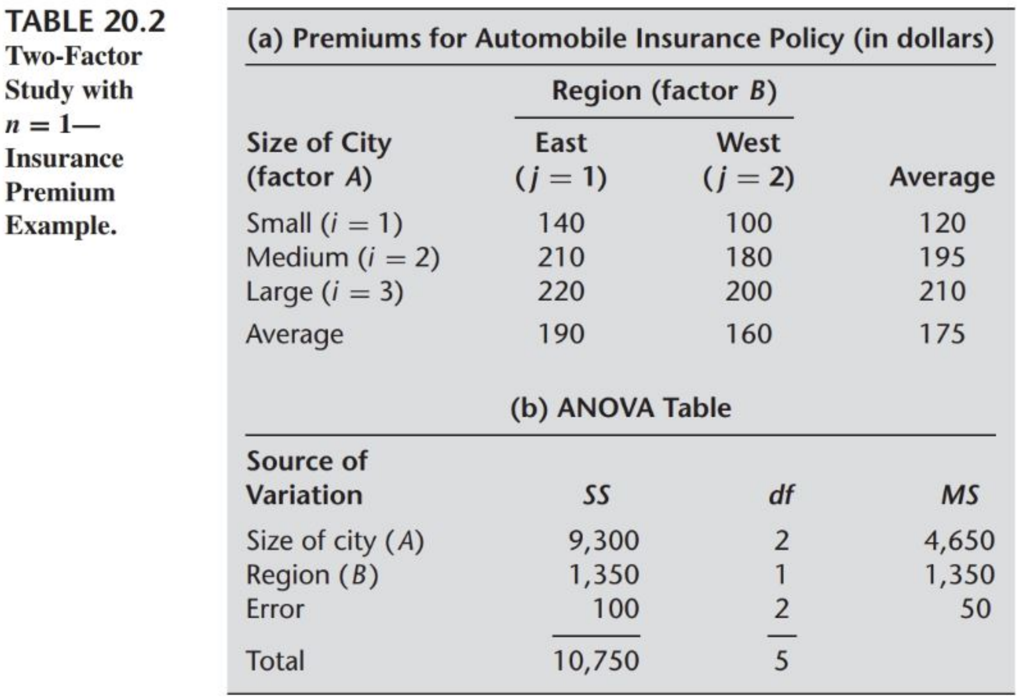


TABLE 20.2 Two-Factor Study with $n = 1$— Insurance Premium Example.

(a) Premiums for Automobile Insurance Policy (in dollars)

Size of City (factor A)	Region (factor B) East ($j = 1$)	 West ($j = 2$)	Average
Small ($i = 1$)	140	100	120
Medium ($i = 2$)	210	180	195
Large ($i = 3$)	220	200	210
Average	190	160	175

(b) ANOVA Table

Source of Variation	SS	df	MS
Size of city (A)	9,300	2	4,650
Region (B)	1,350	1	1,350
Error	100	2	50
Total	10,750	5	

Data Analysis with R

1. Read the data

```
# read data from week5 folder online
Ex20  =read.table(
        url("https://raw.githubusercontent.com/npmldabook/Stat3119/master/Week7/CH20TA02.txt"))

Ex20
```

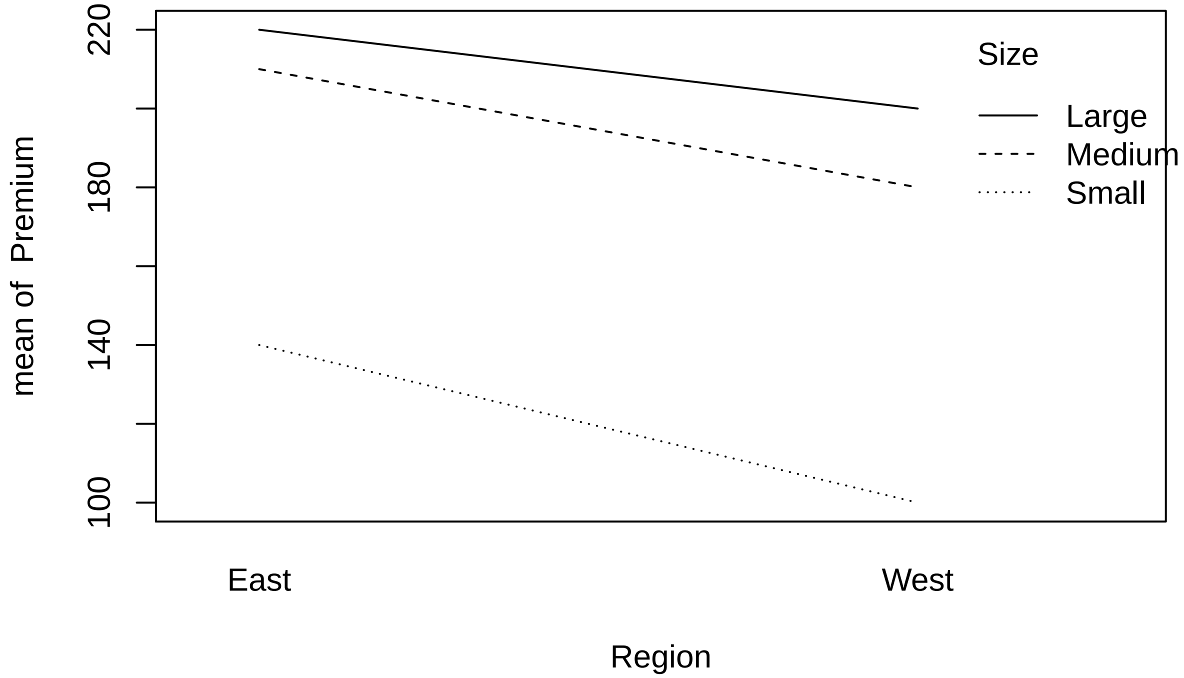
```
##    V1 V2 V3
## 1 140  1  1
## 2 100  1  2
## 3 210  2  1
## 4 180  2  2
## 5 220  3  1
## 6 200  3  2
```

```
names(Ex20) =  c("Premium", "Size", "Region")

# make  categorical variables for factor A and B
Ex20$Size    =  as.factor(Ex20$Size)
Ex20$Region =  as.factor(Ex20$Region)

levels(Ex20$Size)= c("Small", "Medium" , "Large")
levels(Ex20$Region)= c("East", "West")

Ex20
```

```
##   Premium   Size Region
## 1     140  Small   East
## 2     100  Small   West
## 3     210 Medium   East
## 4     180 Medium   West
## 5     220  Large   East
## 6     200  Large   West
```

2. Plot the observations by treatment (i,j)

```
with(Ex20, interaction.plot(x.factor = Region, trace.factor = Size, response = Premium))
```



- The plot does not show strong effect of interaction (treatment line crossing, converging or diverging)
- Since there is only one observation per treatment, the moderate lack of parallelism in the response lines could simply be the result of measurement errors within each treatment cell.

3. Run the Tukey test of additivity

- **Approach 1**: we can calculate the SS from the formula and F-statistic to do the Tukey test.
- **Approach 2**: Use a R-function *tukey.test()* from package **additivityTests**.

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

library(additivityTests)

## make the data into matrix A*B Form (dim a * b)
(Ex20m =  matrix( Ex20$Premium, nrow=3, ncol=2, byrow=T ))
```

```
##      [,1] [,2]
## [1,]  140  100
## [2,]  210  180
## [3,]  220  200
```

```
##
tukey.test( Ex20m, alpha = 0.05)
```

```
##
## Tukey test on 5% alpha-level:
##
## Test statistic: 6.75
## Critival value: 161.4
## The additivity hypothesis cannot be rejected.
```

4. Run two way ANOVA with main effects only

```
fit = aov(Premium~ Size+  Region, data= Ex20 )
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Size         2   9300    4650      93 0.0106 *
## Region       1   1350    1350      27 0.0351 *
## Residuals    2    100      50
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results: If the significant level 0.05 is considered,

1. for the main effect of city size, F-statistic = 93, $P = 0.0106$, we reject the null and conclude that city size effects are present.
2. for the main efect of geographic region , F-statistic = 27, $P = 0.0351$, we reject the null and conclude that geographic region effects are present.

Estimation of model parameters

Setting: when no-interaction ANOVA model (20.1) is used, one case per treatment

1. Comparisons of factor A and factor B level means:

Similar to what have learned in the last chapter for the case with no interaction between A and B in chapter 19, but we simply replace MSE in all of the earlier results with $MSAB$ as the estimator of the error variance σ^2 and modify the degrees of freedom accordingly, i.e., using $df = (a-1)(b-1)$ instead..

2. Estimation of Treatment Mean

- Sometimes, there is some interest in estimating a treatment mean μ_{ij}. If $n > 1$, the best estimator is the sample mean with the minimum variance. But when $n = 1$, we don't want to use single observation Y_{ij} observed at level (i,j), because there is a more efficient (minimum variance) estimator based on the factor level and overall means.

$$\hat{\mu}_{ij} = \overline{Y}_{i.} + \overline{Y}_{j.} - \overline{Y}_{..}$$

Example (page 884) : Estimate the treatment means, SE and CI.

```
(Overall.mean = mean(Ex20$Premium))
```

```
## [1] 175
```

```
(FactorA.mean = with(Ex20, by(Premium, Size,  mean )))
```

```
## Size: Small
## [1] 120
## ------------------------------------------------------------
## Size: Medium
## [1] 195
## ------------------------------------------------------------
## Size: Large
## [1] 210
```

```
(FactorB.mean = with(Ex20, by(Premium, Region, mean )))
```

```
## Region: East
## [1] 190
## ------------------------------------------------------------
## Region: West
## [1] 160
```

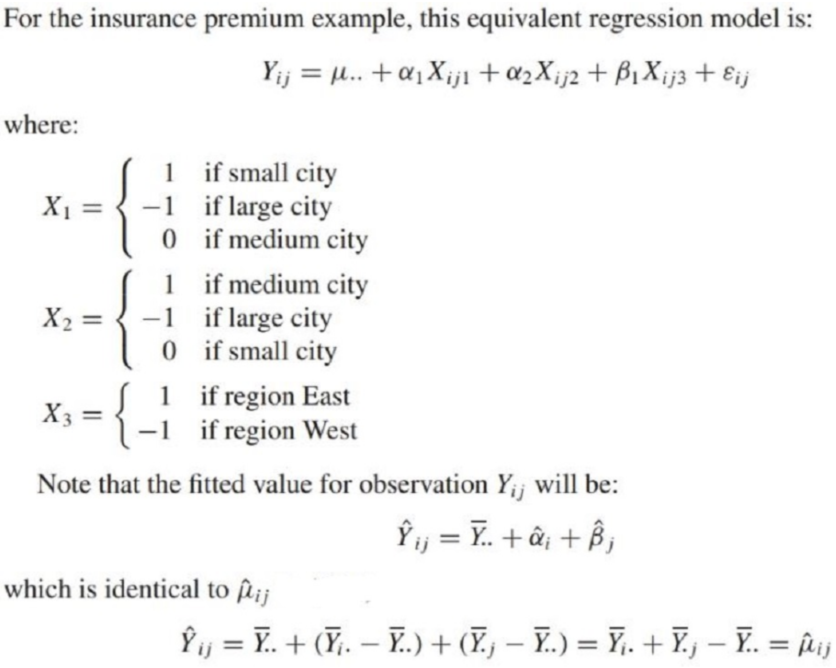
```
(All = data.frame(Ex20, Overall.mean, A.mean=rep(FactorA.mean,c(2,2,2)),B.mean= rep(FactorB.mean, 3)))
```

```
##   Premium   Size Region Overall.mean A.mean B.mean
## 1     140  Small   East          175    120    190
## 2     100  Small   West          175    120    160
## 3     210 Medium   East          175    195    190
## 4     180 Medium   West          175    195    160
## 5     220  Large   East          175    210    190
## 6     200  Large   West          175    210    160
```

```
Treatment.mean = All$A.mean+  All$B.mean - All$Overall.mean
data.frame(Ex20, Treatment.mean)
```

```
##   Premium   Size Region Treatment.mean
## 1     140  Small   East            135
## 2     100  Small   West            105
## 3     210 Medium   East            210
## 4     180 Medium   West            180
## 5     220  Large   East            225
## 6     200  Large   West            195
```

- Equivalently, we can estimate and obtain standard error of the treatment means from the equivalent regression model (page 185)



For the insurance premium example, this equivalent regression model is:

$$Y_{ij} = \mu_{..} + \alpha_1 X_{ij1} + \alpha_2 X_{ij2} + \beta_1 X_{ij3} + \varepsilon_{ij}$$

where:

$$X_1 = \begin{cases} 1 & \text{if small city} \\ -1 & \text{if large city} \\ 0 & \text{if medium city} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if medium city} \\ -1 & \text{if large city} \\ 0 & \text{if small city} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if region East} \\ -1 & \text{if region West} \end{cases}$$

Note that the fitted value for observation Y_{ij} will be:

$$\hat{Y}_{ij} = \bar{Y}_{..} + \hat{\alpha}_i + \hat{\beta}_j$$

which is identical to $\hat{\mu}_{ij}$

$$\hat{Y}_{ij} = \bar{Y}_{..} + (\bar{Y}_{i.} - \bar{Y}_{..}) + (\bar{Y}_{.j} - \bar{Y}_{..}) = \bar{Y}_{i.} + \bar{Y}_{.j} - \bar{Y}_{..} = \hat{\mu}_{ij}$$

Implementation in R

```r
# The indicator functions are set up to satisfy the zero sum contrast for main effects

Factor1<- (Ex20$Size=="Small")*1  + (Ex20$Size=="Large")*(-1)
Factor2<- (Ex20$Size=="Medium")*1 + (Ex20$Size=="Large")*(-1)
Factor3<- (Ex20$Region=="East")*1 + (Ex20$Region=="West")*(-1)

LMfit<-lm( Premium~ Factor1+Factor2+ Factor3, data=Ex20 )
summary(LMfit)
```

```
##
## Call:
## lm(formula = Premium ~ Factor1 + Factor2 + Factor3, data = Ex20)
##
## Residuals:
##          1          2          3          4          5          6
##  5.000e+00 -5.000e+00 -7.550e-15  8.882e-15 -5.000e+00  5.000e+00
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  175.000      2.887  60.622 0.000272 ***
## Factor1      -55.000      4.082 -13.472 0.005465 ** 
## Factor2       20.000      4.082   4.899 0.039231 *  
## Factor3       15.000      2.887   5.196 0.035099 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.071 on 2 degrees of freedom
## Multiple R-squared:  0.9907, Adjusted R-squared:  0.9767 
## F-statistic:    71 on 3 and 2 DF,  p-value: 0.01392
```

```
LMfit$fitted.values
```

```
##   1   2   3   4   5   6 
## 135 105 210 180 225 195
```

Note: The fitted value from regression model are the same as the estimated treatment means as before.

- Then, to obtain the confidence interval of the treatment means, we can use *predict()* function for the lm fit.

```
# Our treatment levels
new <- data.frame(Factor1, Factor2, Factor3)
(pred.CI <- predict(LMfit, new, interval="confidence", level = 0.95, se.fit = TRUE))
```

```
## $fit
##   fit       lwr      upr
## 1 135 110.15862 159.8414
## 2 105  80.15862 129.8414
## 3 210 185.15862 234.8414
## 4 180 155.15862 204.8414
## 5 225 200.15862 249.8414
## 6 195 170.15862 219.8414
## 
## $se.fit
##        1        2        3        4        5        6 
## 5.773503 5.773503 5.773503 5.773503 5.773503 5.773503 
## 
## $df
## [1] 2
## 
## $residual.scale
## [1] 7.071068
```

- To verify the above results, we can calculate SE of the fitted value using matrix based formula (6.58 in chapter 6).

From the regression model, $\hat{Y} = \mathbf{X}\hat{\beta}$, where $\mathbf{X}$ is the design matrix. We can obtain the variance of the fitted $\hat{Y}$ using the matrix formula $var(\hat{Y}) = \mathbf{X}var(\beta)\mathbf{X}'$, where $var(\beta)$ is the variance and covariance matrix of the coefficient estimates.

```r
(DesignX = cbind(1, Factor1, Factor2, Factor3))
```

```
##        Factor1 Factor2 Factor3
## [1,] 1       1       0       1
## [2,] 1       1       0      -1
## [3,] 1       0       1       1
## [4,] 1       0       1      -1
## [5,] 1      -1      -1       1
## [6,] 1      -1      -1      -1
```

```r
# variance- covariance matrix of model coefficient
(Var_beta = vcov(LMfit))
```

```
##             (Intercept)       Factor1       Factor2       Factor3
## (Intercept)    8.333333  0.000000e+00  0.000000e+00  0.000000e+00
## Factor1        0.000000  1.666667e+01 -8.333333e+00 -5.341563e-16
## Factor2        0.000000 -8.333333e+00  1.666667e+01  1.068313e-15
## Factor3        0.000000 -5.341563e-16  1.068313e-15  8.333333e+00
```

```r
# variance- covariance matrix of fitted Y
(Var_Yhat = DesignX %*% Var_beta %*% t(DesignX))
```

```
##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] 33.333333 16.666667  8.333333 -8.333333  8.333333 -8.333333
## [2,] 16.666667 33.333333 -8.333333  8.333333 -8.333333  8.333333
## [3,]  8.333333 -8.333333 33.333333 16.666667  8.333333 -8.333333
## [4,] -8.333333  8.333333 16.666667 33.333333 -8.333333  8.333333
## [5,]  8.333333 -8.333333  8.333333 -8.333333 33.333333 16.666667
## [6,] -8.333333  8.333333 -8.333333  8.333333 16.666667 33.333333
```

```r
# standard deviation of Y_hat
(sd_Yhat  = sqrt(diag(Var_Yhat)))
```

```
## [1] 5.773503 5.773503 5.773503 5.773503 5.773503 5.773503
```

```r
# confidence interval is based on the t-distribution, df=(a-1)(b-1)=2
qt(.975, 2)
```

```
## [1] 4.302653
```

```r
data.frame( LCI= LMfit$fitted.values-4.302653*5.773503, UCI=LMfit$fitted.values+4.302653*5.773503 )
```

```
##         LCI      UCI
## 1 110.15862 159.8414
## 2  80.15862 129.8414
## 3 185.15862 234.8414
## 4 155.15862 204.8414
## 5 200.15862 249.8414
## 6 170.15862 219.8414
```

Remedial Actions if Interaction Effects Are Present

- When the Tukey test indicates the presence of interaction effects in an analysis of variance application where $n = 1$, efforts should be made to remove the interactions to use methods discussed in this chapter
- One possibility is to try simple transformations of the response variable, such as a square root or a logarithmic transformation
- Or try the Box-Cox power transformation. For each value of λ, the Tukey test statistic is then obtained. If a λ value leads to a nonsignificant Tukey test statistic, a transformation will then have been found that removes the interaction effect. Often, we choose a simple round-off λ value.

New Topic: Randomized Complete Block Designs (Ch 21)

- When the available experimental units are not homogeneous, grouping the experimental units into **blocks** of homogeneous units will reduce the experimental error variance, and more precise estimates about the treatment effects.
- Nuisance Factor may be present in experiment, has effect on response but its effect is not of interest
  - If unknown - Protecting experiment through randomization
  - If known and controllable, we use **blocking**. e.g. In a randomized multicenter clinical trials to study a new treatment vs. standard treament on a certain disease. The center/site/hospital would be the nuisance factor, blocks, and we should randomize different treatment within each block to remove its confounding effect.
- block designs
  - Complete block designs: each block consists of one complete replication of the set of treatments (Ch 21).
  - Incomplete block designs: When the number of experimental units available in a block is lessthan the number of treatments (Ch 28)

Randomized complete block design

In a **randomized complete block design**, the experimental units are first sorted into homogeneous groups, called **blocks**, and all treatment combinations are then assigned at random to experimental units within the blocks.

The key objective in blocking the experimental units is to make them as homogeneous as possible within blocks with respect to the response variable under study, and to make the different blocks as heterogeneous as possible with respect to the response variable.

- Advantages : reduce the experimental error variance, and more precise estimates about the treatment effects
- Disadvantages:
  - missing observations within a block makes the analysis more complicated
  - Because the blocking variable is an observational factor and not an experimental factor (not randomize into blocks), cause-and-effect inferences concerning the relationship between the blocking variable.

- Example 1: In an experiment on the effects of **four levels of newspaper advertising** saturation on sales volume, the experimental unit is a city, and 16 cities are available for the study. **Size of city** usually is highly correlated with the response variable, sales volume. Hence, it is desirable to block the 16 cities into four groups of four cities each, according to population size. Within each block, the four treatments are then assigned at random to the four cities.
- Example 2: A chemist is studying the reaction rate of **five chemical agents**. Only five agents can be analyzed effectively per day. Since day-to-day differences may affect the reaction rate, **each day is used as a block**, and all five chemical agents are tested each day in independently randomized orders.

Criteria for Blocking

Two types of blocking criteria:

- Characteristics associated with the unit—for persons: gender, age, income, intelligence, education, job experience, attitudes, etc.; for geographic areas: population size, average income, etc.
- Characteristics associated with the experimental setting—observer, time of processing, machine, batch of material, measuring instrument, etc.
- There is no need to use only a single blocking criterion; several may be employed if the experimental error can be further reduced by doing so.
- The design of an effective randomized block experiment requires the ability to anticipate potential sources of variation—the blocking variables—in advance of experimentation. These variables are then held constant within blocks as the experiment is conducted in order to reduce the experimental error variability.

We will discuss the model and analysis of RCBD (Randomized Complete Block Designs) in the next class

Summary this week

- Quiz (#3) This Thurs (10/10): main concept of chapters 18-19

1. ANOVA assumptions, diagnostics, and remedial measures
2. Two-way ANOVA: 2 models, ANOVA table (ss partition, df, MS= SS/df), F-tests for A, B, and A:B, and intepretation of test results.

- Reading: Chapter 20
- HW: 20.2, 20.3, 20.4