STA 305 Project/RMarkdown Tutorial
Today in BA 1130, 12-1
Tomorrow in SS 2135, 12-1

STA305/1004-Class 15

Reminder: Complete practice problems in course notes.

October 31, 2019

Project draft due on Nov. 11

Today's Class

- R Data Frames for Factorial Experiments
- Linear model for factorial design
- Estimating Factorial Effects using Linear Regression
- Inference for Factorial Effects using Linear Regression

R Data Frames for Factorial Experiments

- One option is to use a spreadsheet program such as Excel to save your data.
- R can read data from saved in many different formats.
- For example, if your data is saved as an Excel file (e.g., `pilotplant.xlsx`) then use the `readxl` library to read the file into an R data frame.

```
library(readxl)
tab0503.1 <- read_excel("pilotplant.xlsx")
tab0503.1
```

If readxl is not installed then you must install library.

dataframe.

```
## # A tibble: 8 x 7
##    run1  run2     T     C     K    y1    y2
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     6    13    -1    -1    -1    59    61
## 2     2     4     1    -1    -1    74    70
## 3     1    16    -1     1    -1    50    58
## 4     5    10     1     1    -1    69    67
## 5     8    12    -1    -1     1    50    54
## 6     9    14     1    -1     1    81    85
## 7     3    11    -1     1     1    46    44
## 8     7    15     1     1     1    79    81
```

- Each row contains data for two runs.
- If we want to estimate factorial effects using linear regression then we need one row per run.

run = an experimental condition

e.g., T = -1, C = -1, K = -1.

R Data Frames for Factorial Experiments

- The data that we saw at the beginning of last class used the average `y` of `y1` and `y2` (from the previous data set).
- The data was stored in a different tab-delimited file `tab0502.dat`

```
tab0502 <- read.csv("tab0502.dat", sep = "")
tab0502
```

```
##   run  T  C  K  y
## 1   1 -1 -1 -1 60
## 2   2  1 -1 -1 72
## 3   3 -1  1 -1 54
## 4   4  1  1 -1 68
## 5   5 -1 -1  1 52
## 6   6  1 -1  1 83
## 7   7 -1  1  1 45
## 8   8  1  1  1 80
```

each row corresponds to a run

R Data Frames for Factorial Experiments

Please try this code yourself!

- To create a 2^k factorial design matrix (defined later) in R.
- The sequence of -1 and +1 can be created using the `rep()` function in R.
- For example: `rep(c(-1, 1) 2)` repeats the vector `(-1, 1)` twice to produce a vector `(-1, 1, -1, 1)`.

  the vector (-1,1)
- A 2^3 design matrix could be generated by the following code.

```
x1 <- rep(c(-1, 1), 4)  # repeat (-1,1) 4 times.
x2 <- rep(rep(c(rep(-1, 2), rep(1, 2)), 2))
x3 <- c(rep(-1, 4), rep(1, 4))
mydat <- data.frame(x1, x2, x3, "x1*x2" = x1*x2, "x1*x3" = x1*x3, "x2*x3" = x2*
           "x1*x2*x3" = x1*x2*x3)
mydat
```

-1
-1
1
1 } repeat twice

Cut off.

try to generate a 2^2 design or a 2^4 design.

```
##   x1 x2 x3 x1.x2 x1.x3 x2.x3 x1.x2.x3
## 1 -1 -1 -1     1     1     1       -1
## 2  1 -1 -1    -1    -1     1        1
## 3 -1  1 -1    -1     1    -1        1
## 4  1  1 -1     1    -1    -1       -1
## 5 -1 -1  1     1    -1    -1        1
## 6  1 -1  1    -1     1    -1       -1
## 7 -1  1  1    -1    -1     1       -1
## 8  1  1  1     1     1     1        1
```

```
#write.csv(mydat, "mydat.csv") #write the data to a csv file
```

Linear model for factorial design

Let y_i be the yield from the i^{th} run,

$$x_{i1} = \begin{cases} +1 & \text{if } T = 180 \\ -1 & \text{if } T = 160 \end{cases}$$

$$x_{i2} = \begin{cases} +1 & \text{if } C = 40 \\ -1 & \text{if } C = 20 \end{cases}$$

$$x_{i3} = \begin{cases} +1 & \text{if } K = B \\ -1 & \text{if } K = A \end{cases}$$

A linear model for a 2^3 factorial design is:

$$y_i = \beta_0 + \overbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}^{\text{main effects}} + \overbrace{\beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3}}^{\text{Two-way interactions}} + \overbrace{\beta_7 x_{i1} x_{i2} x_{i3}}^{\text{Three way Interaction.}} + \epsilon_i.$$

The variables $x_{i1} x_{i2}$ is the interaction between temperature and concentration, $x_{i1} x_{i3}$ is the interaction between temperature and catalyst, etc.

Linear model for factorial design

The **table of contrasts** for a 2^3 design is the **design matrix** X from the linear model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \beta_7 x_{i1} x_{i2} x_{i3} + \epsilon_i,$$

linear model.

where x_{ij} are defined on the previous slide.

```
fact.mod1 <- lm(y~T*K*C,data = tab0502)
X <- model.matrix(fact.mod1)
X[,-9] # -9 remove last column (dependent variable) to only show X
```

The X matrix in $X\beta + \varepsilon$

```
##   (Intercept)  T  K  C T:K T:C K:C T:K:C
## 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
```

Linear model for factorial design

- Ave yield when T=1 − Ave yield when T=−1 = main effect of T
- (72+68+83+80)/4 − (60+54+52+45)/4

Mean	T	K	C	T:K	T:C	K:C	T:K:C	yield average
1	-1	-1	-1	1	1	1	-1	60
1	1	-1	-1	-1	-1	1	1	72
1	-1	-1	1	1	-1	-1	1	54
1	1	-1	1	-1	1	-1	-1	68
1	-1	1	-1	-1	1	-1	1	52
1	1	1	-1	1	-1	-1	-1	83
1	-1	1	1	-1	-1	1	-1	45
1	1	1	1	1	1	1	1	80

- T:K interaction: $(60+54+83+80)/4 - \frac{(72+68+52+45)}{4}$

- All factorial effects can be calculated from this table.
- Signs for interaction contrasts obtained by multiplying signs of their respective factors.
- Each column perfectly balanced with respect to other columns.
- Balanced (orthogonal) design ensures each estimated effect is unaffected by magnitude and signs of other effects.
- Table of signs obtained similarly for any 2^k factorial design.

Linear model for factorial design

What is the table of contrasts for a 2^2 factorial design?

4 Rows

mean	x_1	x_2	$x_1 x_2$	y
1	-1	-1	1	y_1
1	$+1$	-1	-1	y_2
1	-1	$+1$	-1	y_3
1	$+1$	$+1$	$+1$	y_4

$$x_1 = \begin{cases} 1 \\ -1 \end{cases}$$

$$x_2 = \begin{cases} 1 \\ -1 \end{cases}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon .$$

$\boxed{\text{Do this for a } 2^4 \text{ design}}$

Linear model for factorial design - calculating factorial effects from parameter estimates

- The parameter estimates are obtained via the `lm()` function in R.
- Estimated least squares coefficients are one-half the factorial estimates.
- Therefore, the factorial estimates are twice the least squares coefficients.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$$ for a 2^2 design.

Factorial effect for x_1 would be

$$\left(\frac{y_2 + y_4}{2}\right) - \left(\frac{y_1 + y_3}{2}\right)$$

$$\mu_1 = E(y_1 \mid x_1 = -1, x_2 = -1) = \beta_0 - \beta_1 - \beta_2 + \beta_3$$

$$\mu_2 = E(y_2 \mid x_1 = 1, x_2 = -1) = \beta_0 + \beta_1 - \beta_2 - \beta_3$$

$$\mu_3 = E(y_3 \mid x_1 = -1, x_2 = 1) = \beta_0 - \beta_1 + \beta_2 - \beta_3$$

$$\mu_4 = E(y_4 \mid x_1 = 1, x_2 = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

$$\left(\frac{\mu_2 + \mu_4}{2}\right) - \left(\frac{\mu_1 + \mu_3}{2}\right) = 2\beta_1$$

factorial effect $= 2 \times$ regression coefficient for x_1

Linear model for factorial design - calculating factorial effects from parameter estimates

```r
fact.mod <-lm(y~T*K*C,data=tab0502)
round(summary(fact.mod)$coefficients,2)
```

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     64.25        NaN     NaN      NaN
T               11.50        NaN     NaN      NaN
K                0.75        NaN     NaN      NaN
C               -2.50        NaN     NaN      NaN
T:K              5.00        NaN     NaN      NaN
T:C              0.75        NaN     NaN      NaN
K:C              0.00        NaN     NaN      NaN
T:K:C            0.25        NaN     NaN      NaN
```

$\hat{\beta}_0$ $\hat{\beta}_1$ $\hat{\beta}_2$ $\hat{\beta}_3$

$$\hat{\beta}_1 = 11.50 \Rightarrow T = 2 \times 11.50 = 23.26$$

$$\hat{\beta}_2 = 0.75 \Rightarrow K = 2 \times 0.75 = 1.5$$

$$\hat{\beta}_4 = 5.00 \Rightarrow TK = 2 \times 5.00 = 10.00$$

- Why is the `Std. Error` column `NaN`?

Inference for Factorial Effects using Linear Regression

- In order for `lm()` to calculate standard errors at least two runs per experimental run are needed.
- Data format: each row should correspond to an experimental run.
- The data is stored this way in `tab0503.dat`.

```
library(tidyverse)
tab0503 <- read.csv("tab0503.dat", sep="")
glimpse(tab0503)
```

```
## Observations: 16
## Variables: 5
## $ run <int> 6, 2, 1, 5, 8, 9, 3, 7, 13, 4, 16, 10, 12, 14, 11, 15
## $ T   <int> -1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1
## $ C   <int> -1, -1, 1, 1, -1, -1, 1, 1, -1, -1, 1, 1, -1, -1, 1, 1
## $ K   <int> -1, -1, -1, -1, 1, 1, 1, 1, -1, -1, -1, -1, 1, 1, 1, 1
## $ y   <int> 59, 74, 50, 69, 50, 81, 46, 79, 61, 70, 58, 67, 54, 85, 44...
```

Inference for Factorial Effects using Linear Regression

- When there are replicated runs we also obtain p-values and confidence intervals for the factorial effects from the regression model.
- For example, the p-value for β_1 corresponds to the factorial effect for temperature

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0.$$

If the null hypothesis is true then $\beta_1 = 0 \Rightarrow T = 0 \Rightarrow \mu_{T+} - \mu_{T-} = 0 \Rightarrow \mu_{T+} = \mu_{T-}$.

- μ_{T+} is the mean yield when the temperature is set at 180° and μ_{T-} is the mean yield when the temperature is set to 160°.

Inference for Factorial Effects using Linear Regression

To obtain 95% confidence intervals for the factorial effects we multiply the 95% confidence intervals for the regression parameters by 2. This is easily done in R using the function `confint.lm()`.

```
fact.mod <-lm(y~T*K*C,data=tab0503)
round(2*confint.lm(fact.mod),2)
```

```
              2.5 % 97.5 %
(Intercept) 125.24 131.76
T            19.74  26.26
K            -1.76   4.76
C            -8.26  -1.74
T:K           6.74  13.26
T:C          -1.76   4.76
K:C          -3.26   3.26
T:K:C        -2.76   3.76
```

Which 95% CI contain 0?
The CI that do not contain 0 mean that the factor has a significant (5%) effect on yield

Advantages of factorial designs over one-factor-at-a-time designs

- Suppose that one factor at a time was investigated. For example, temperature is investigated while holding concentration at 20% (-1) and catalyst at B (+1).
- In order for the effect to have more general relevance it would be necessary for the effect to be the same at all the other levels of concentration and catalyst.
- In other words there is no interaction between factors (e.g., temperature and catalyst).
- If the effect is the same then a factorial design is more efficient since the estimates of the effects require fewer observations to achieve the same precision.
- If the effect is different at other levels of concentration and catalyst then the factorial can detect and estimate interactions.