STA305/1004 Class Notes - Week 8

Nathan Taback

March 11, 2016

1 Factorial Designs at two levels - 2^k designs

Suppose that an investigator is interested in examining three components of a weight loss intervention. The three components are:

1. Keeping a food diary (yes/no)
2. Increasing activity (yes/no)
3. Home visit (yes/no)

The investigator plans to investigate all $2x2x2 = 2^3 = 8$ combinations of experimental conditions. The experimental conditions will be.

Experimental condition number	**Keeping a food diary**	**Increasing physical activity**	**Home visit**	**weight loss**
1	No	No	No	y_1
2	No	No	Yes	y_2
3	No	Yes	No	y_3
4	No	Yes	Yes	y_4
5	Yes	No	No	y_5

6	Yes	No	Yes	y_6
7	Yes	Yes	No	y_7
8	Yes	Yes	Yes	y_8

- To perform a factorial design, you select a fixed number of levels of each of a number of factors (variables) and then run experiments in all possible combinations.
- The factors can be quantitative or qualitative.
- Two levels of a quantitative variable could be two different temperatures or two different concentrations.
- Qualitative factors might be two types of catalysts or the presence and absence of some entity.

The notation 2^3 identifies: - the number of factors (3) - the number of levels of each factor (2) - how many experimental conditions are in the design ($2^3 = 8$)

Factorial experiments can involve factors with different numbers of levels.

1.1 Exercise

Consider a 4^2x3^2x2 design.

a. How many factors?
b. How many levels of each factor?
c. How many experimental conditions (runs)?

Answer: (a) There are 2+2+1=5 factors. (b) Two factors have 4 levels, 2 factors have 3 levels, and 1 factor has 2 levels. (c) There are 288 experimental conditions or runs.

1.2 Difference between ANOVA and Factorial Designs

In ANOVA the objective is to compare the individual experimental conditions with each other. In a factorial experiment the objective is generally to compare combinations of experimental conditions.

Let's consider the food diary study above. What is the effect of keeping a food diary?

Experimental condition number	**Keeping a food diary**	**Increasing physical activity**	**Home visit**	**weight loss**
1	No	No	No	y_1
2	No	No	Yes	y_2
3	No	Yes	No	y_3
4	No	Yes	Yes	y_4
5	Yes	No	No	y_5

6	Yes	No	Yes	y_6
7	Yes	Yes	No	y_7
8	Yes	Yes	Yes	y_8

We can estimate the effect of food diary by comparing the mean of all conditions where food diary is set to NO (conditions 1-4) and mean of all conditions where food diary set to YES (conditions 5-8). This is also called the **main effect** of food diary, the adjective *main* being a reminder that this average is taken over the levels of the other factors.

The main effect of food diary is:

$$\frac{y_1 + y_2 + y_3 + y_4}{4} - \frac{y_5 + y_6 + y_7 + y_8}{4}.$$

The main effect of physical activity is:

$$\frac{y_1 + y_2 + y_5 + y_6}{4} - \frac{y_3 + y_4 + y_7 + y_8}{4}.$$

The main effect of home visit is:

$$\frac{y_1 + y_3 + y_5 + y_7}{4} - \frac{y_2 + y_4 + y_6 + y_8}{4}.$$

All experimental subjects are used, but are rearranged to make each comparison. Subjects are recycled to measure different effects. This is one reason why factorial experiments are more efficient.

2 Factorial designs at two levels

To perform a factorial design:

1. Select a fixed number of levels of each factor.
2. Run experiments in all possible combinations.

We will discuss designs where there are just two levels for each factor. Factors can be quantitative or qualitative. Two levels of quantitative variable could be two different temperatures or concentrations. Two levels of a quantitative variable could be two different types of catalysts or presence/absence of some entity.

The following example if from Box, Hunter, and Hunter (2005).

An experiment employed a 2^3 factorial design with two quantitative factors - temperature (T) and concentration (C) - and a single qualitative factor - type of catalyst K.

Temerature T (C°) has two levels: $160\mathrm{C}^\circ$, and $180\mathrm{C}^\circ$. These are coded as -1 and +1 respectively.

Concentration C (%) has two levels: 20 and 40. These are coded as -1 and +1 respectively.

Catalyst K has two levels: A and B. These are coded as -1 and +1 respectively.

Each data value recorded is for the response yield y averaged over two duplicate runs.

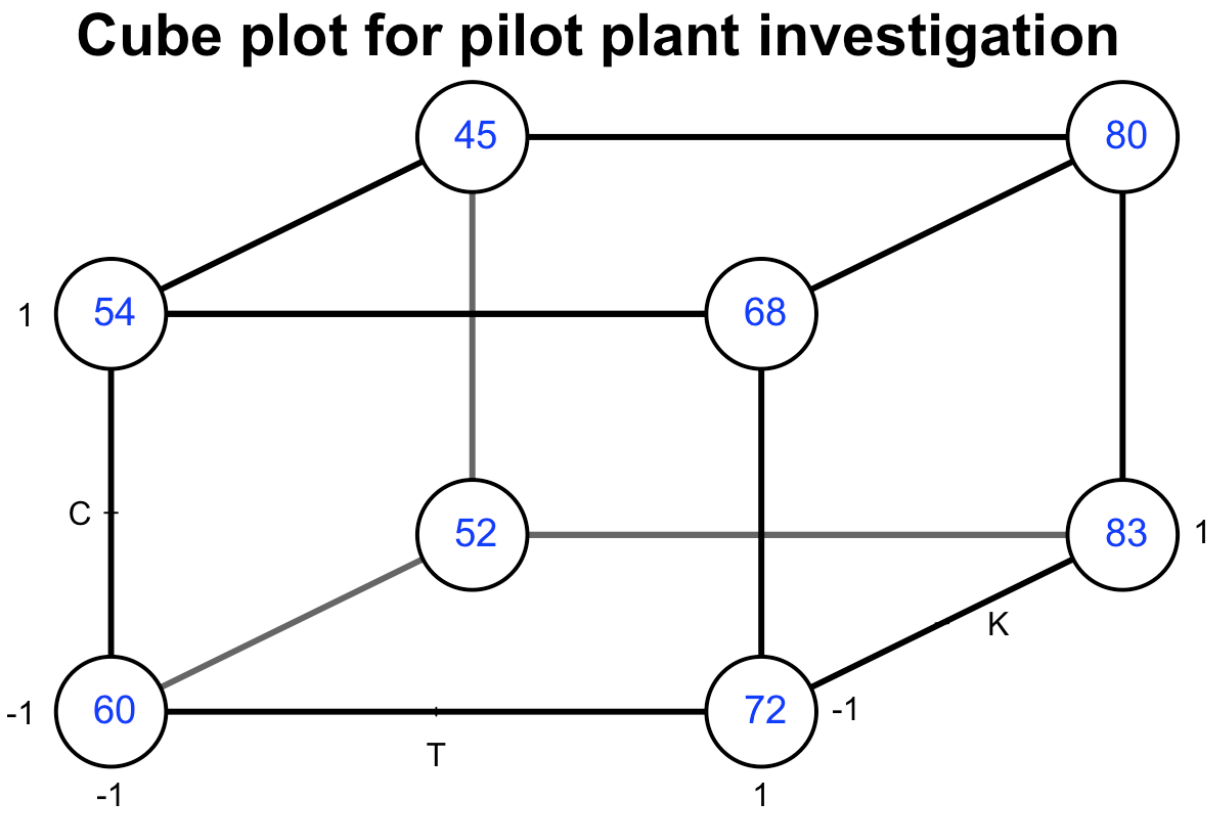
run	T	C	K	y
1	-1	-1	-1	60
2	1	-1	-1	72
3	-1	1	-1	54
4	1	1	-1	68
5	-1	-1	1	52
6	1	-1	1	83
7	-1	1	1	45
8	1	1	1	80

2.1 Cube plots

The figure below shows the value of y for the various combinations of factors T, C, and K at the corners of a cube. For example, $y = 54$ was obtained from the run 3 when T=-1, C=1, and K=-1.

- The cube shows how this design produces 12 comparisons along the 12 edges of the cube: four measures of the effect of temperature change; four measures of the effect of concentration change; four measures of the effect of catalyst change.
- On each edge of the cube only one factor is changed with the other two held constant.

```
library("FrF2")
bhh54 <- lm(y~T*C*K,data=tab0502)
cubePlot(bhh54,"T","K","C",main="Cube plot for pilot plant investigatio
n")
```



2.2 Factorial effects

2.2.1 Main effects

The effects of runs 1 and 2 differ only becuase of tempertaure since concentration is 20% and type of catalyst is A. The difference 72-60=12 supplies one measure of the temperature effect with the remaining factors held fixed. There are four such measures of the temperture effect one for each of the four combinations of concentration and catalyst.

C	K	Effect of changing T from 160 to 180
20	A	$y_2 - y_1 = 72 - 60 = 12$
40	A	$y_4 - y_3 = 68 - 54 = 14$
20	B	$y_6 - y_5 = 83 - 52 = 31$
40	B	$y_8 - y_7 = 80 - 45 = 35$

The main (average) effect of T is

$$T = \frac{12 + 14 + 31 + 35}{4} = 23$$

There are a similar set of measures for the concentration C. In each of these the levels T and K are kept constant. The main effect for concentration C is:

T	K	Effect of changing C from 20 to 40
160	A	$y_3 - y_1 = 54 - 60 = -6$
180	A	$y_4 - y_2 = 68 - 72 = -4$
160	B	$y_7 - y_5 = 45 - 52 = -7$
180	B	$y_8 - y_6 = 80 - 83 = -3$

The main (average) effect of C is

$$C = \frac{(-6) + (-4) + (-7) + (-3)}{4} = -5$$

The main effect for K is

T	C	Effect of changing K from A to B
160	20	$y_5 - y_1 = 52 - 60 = -8$
180	20	$y_6 - y_2 = 83 - 72 = 11$
160	40	$y_7 - y_3 = 45 - 54 = -9$
180	40	$y_8 - y_4 = 80 - 68 = 12$

The main (average) effect of K is

$$K = \frac{(-8) + (11) + (-9) + (12)}{4} = 1.5$$

All 8 runs are used to estimate each of the main effects. This is the reason that factorial designs are more effcient comapred to examining one factor at a time.

In general the main effects are the differences between two averages:

$$\text{Main Effect} = \bar{y}_+ - \bar{y}_-.$$

Where $\bar{y}_+$ is the average response corresponding to the +1 level of the factor and $\bar{y}_-$ is the average response corresponding to the -1 level of the factor.

$$T = \frac{72 + 68 + 83 + 80}{4} - \frac{60 + 54 + 52 + 45}{4} = 23$$
$$C = \frac{54 + 68 + 45 + 80}{4} - \frac{60 + 72 + 52 + 83}{4} = -5$$
$$K = \frac{52 + 83 + 45 + 80}{4} - \frac{60 + 72 + 54 + 68}{4} = 1.5$$

2.2.2 Interaction effects

2.2.2.1 Two factor interactions

When the catalyst K is A the temperature effect is:

$$\frac{68+72}{2} - \frac{60+54}{2} = 70 - 57 = 13.$$

When the catalyst K is B the temperature effect is:

$$\frac{83+80}{2} - \frac{52+45}{2} = 81.5 - 48.5 = 33.$$

The average difference between these two average differences is called the **interaction** between temperature and catalyst denoted by TK. This is the interaction between the two factors temperature and catalyst - the two factor interaction between temperature and catalyst.

$$TK = \frac{33-13}{2} = 10$$

This can also be seen on the cube plot: the average temperature effect is greater on the back face of the cube (33) compared to the front face of the cube (13).

2.2.2.2 Three factor interactions

The temperature by concentration interaction when the catalyst is B (at it's +1 level) is:

$$\text{Interaction TC} = \frac{(y_8 - y_7) - (y_6 - y_5)}{2} = \frac{(80-45)-(83-52)}{2} = 2.$$

The temperature by concentration interaction when the catalyst is A (at it's -1 level) is:

$$\text{Interaction TC} = \frac{(y_4 - y_3) - (y_2 - y_1)}{2} = \frac{(68-54)-(72-60)}{2} = 1.$$

The difference between these two interactions measures how consistient the temperature-by-concentration interaction for the two catalysts. Half this difference is defined as the three factor interaction of temperature, concentration, and catalyst denoted by TCK.

$$\text{TCK} = \frac{2-1}{2} = \frac{1}{2}.$$

3 Replication in factorial designs

The ouctome y of the pilot plant experiment was the average of two replicated runs. The two seperate runs are shown in the table below. The run order was randomized. For example, runs 6 and 13 are two replicates under the same settings for T, C, and K (T=-1, C=-1, K=-1).

run	T	C	K	y
6	-1	-1	-1	59
2	1	-1	-1	74
1	-1	1	-1	50
5	1	1	-1	69
8	-1	-1	1	50

9	1	-1	1	81
3	-1	1	1	46
7	1	1	1	79
13	-1	-1	-1	61
4	1	-1	-1	70
16	-1	1	-1	58
10	1	1	-1	67
12	-1	-1	1	54
14	1	-1	1	85
11	-1	1	1	44
15	1	1	1	81

Replicating a run is not always feasible. The pilot plant experiment run involved cleaning the reactor, inserting the appropriate catalyst charge, and running the apparatus at a given temperature at a given feed concentration for 3 hours to a1low the process to settle down at the chosen experimental conditions, and (4) sampling the output every 15 minutes during the final hours of running. (Box, Hunter, Hunter, 2005)

run1	run2	T	C	K	y1	y2	diff
6	13	-1	-1	-1	59	61	-2
2	4	1	-1	-1	74	70	4
1	16	-1	1	-1	50	58	-8
5	10	1	1	-1	69	67	2
8	12	-1	-1	1	50	54	-4
9	14	1	-1	1	81	85	-4
3	11	-1	1	1	46	44	2
7	15	1	1	1	79	81	-2

Suppose that the variance of each measurement is σ^2 . The estimated variance at each set of conditions is:

$$s_i^2 = \frac{(y_{i1} - y_{i2})^2}{2} = \frac{\text{diff}^2}{2},$$

where y_{i1} is the first outcome from ith run. In the table above $\text{diff}_i = (y_{i1} - y_{i2})$. A pooled estimate of σ^2 is

$$s^2 = \frac{\sum_{i=1}^{8} s_i^2}{8} = \frac{64}{8} = 8.$$

The estimate of the variance with one degree of freedom for a duplicated run is $s_i^2 = (y_{i1} - y_{i2})$. The average of these yields single degree-of-freedom estiamtes yields a pooled estimate $s^2 = 8$ with 8 degrees of freedom.

3.1 Estimate of the error variance and standard error of effects from replicated runs

Each estimated effect such as T, C, K, TC, etc. is a difference between two averages of 8 observations. The variance of a factoril effect for duplicated runs is

$$Var\,(\text{effect}) = \left(\frac{1}{8} + \frac{1}{8}\right) s^2 = \frac{8}{4} = 2$$

So, the standard error of any factorial effect is:

$$se\,(\text{effect}) = \sqrt{2} = 1.4.$$

4 Interpretation of results

Which effects are real and which can be explained by chance? A rough rule of thumb is any effect that is 2-3 times their standard error are not easily explained by chance alone.

If we assume that the observations are independent and normally distributed then

$$\text{effect}/se\,(\text{effect}) \sim t_8.$$

So a 95% confidence interval can be calculated as:

$$\text{effect} \pm t_{8,.05/2} sese\,(\text{effect}).$$

where $t_{8,.05/2}$ is the 97.5th percentile of the t_8. This is obtained in R via the `qt()` function.

```
qt(p = 1-.025,df = 8)
```

```
## [1] 2.306004
```

So, a 95% confidence interval for a factorial effect is

$$\text{effect} \pm 2.3 \times 1.4 = \text{effect} \pm 3.2.$$

A 95% confidence interval for T is

```
23-3.2 #lower limit
```

```
## [1] 19.8
```

```
23+3.2 #upper limit
```

```
## [1] 26.2
```

A 95% confidence interval for K is

```
1.5-3.2 #lower limit
```

```
## [1] -1.7
```

```
1.5+3.2 #upper limit
```

```
## [1] 4.7
```

The effect due to temperature is probably not due to chance, but chance cannot be rules for the effect due to catalyst.

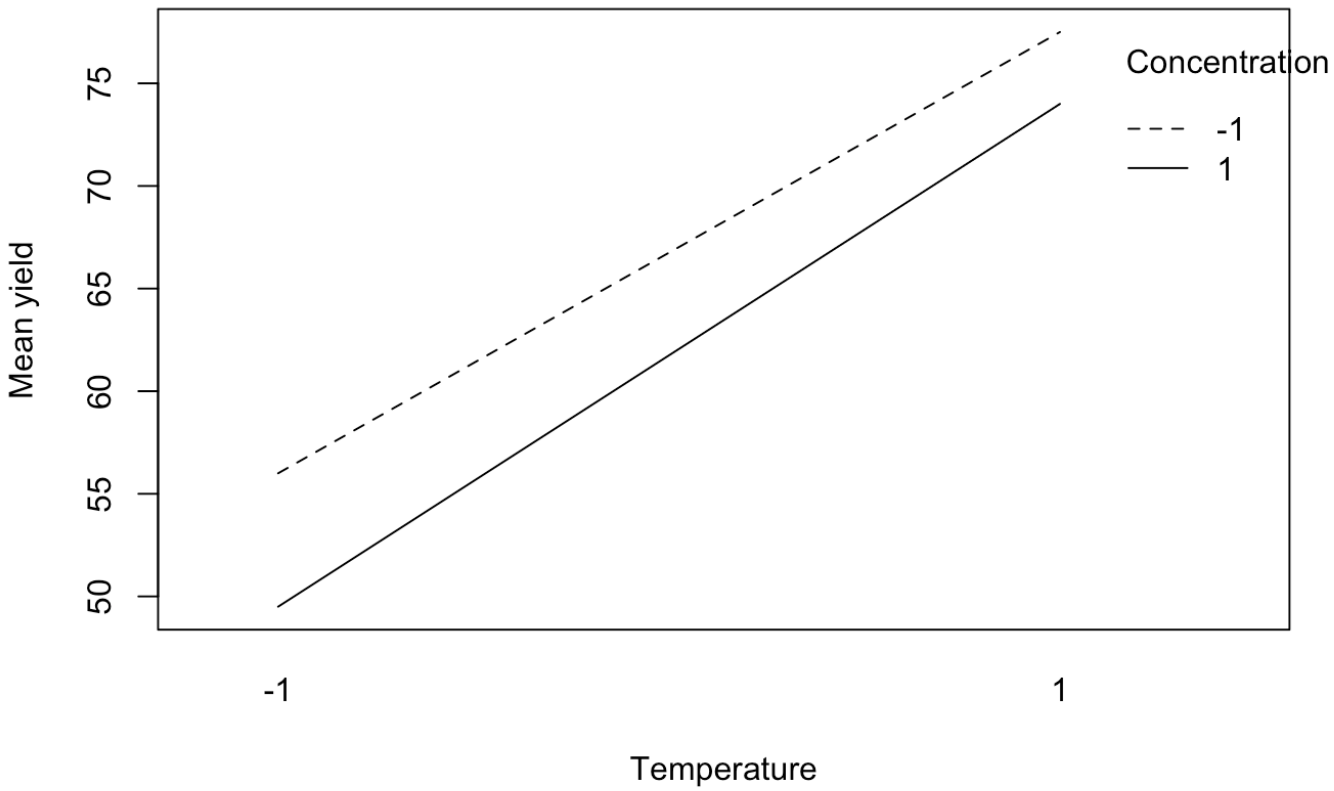
The main effect of a factor should be individually interpreted only if there is no evidence that the factor interacts with other factors.

4.1 Interaction plots

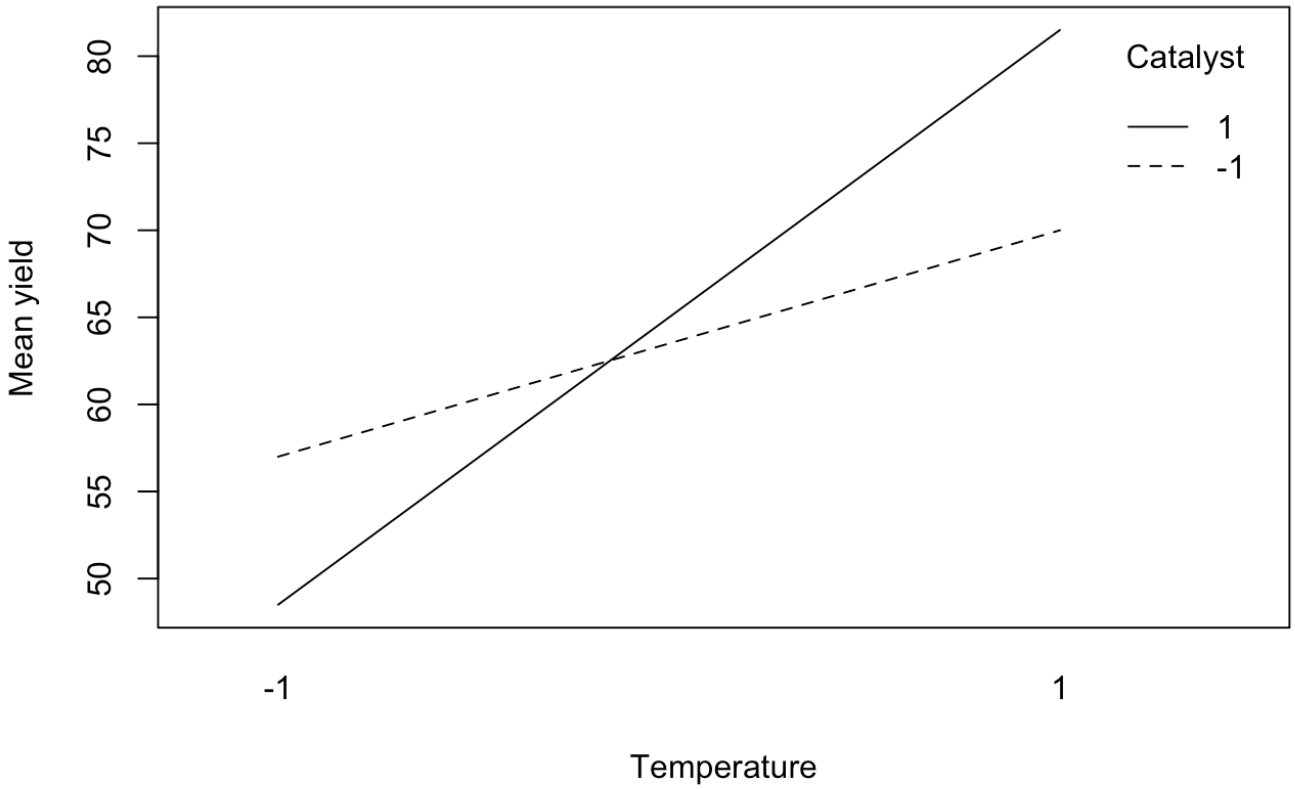
The plots below show the mean yield for each pair of factors TC, TK, CK (i.e., each factor-level combination of these factors). These plots are often called interaction plots. If the two lines are parallel then this indicates no interaction, and if the lines cross or looare close to crossing then this indicates that an interaction might be present.

The plots below indicate a two-way interaction between catalyst and temperature.

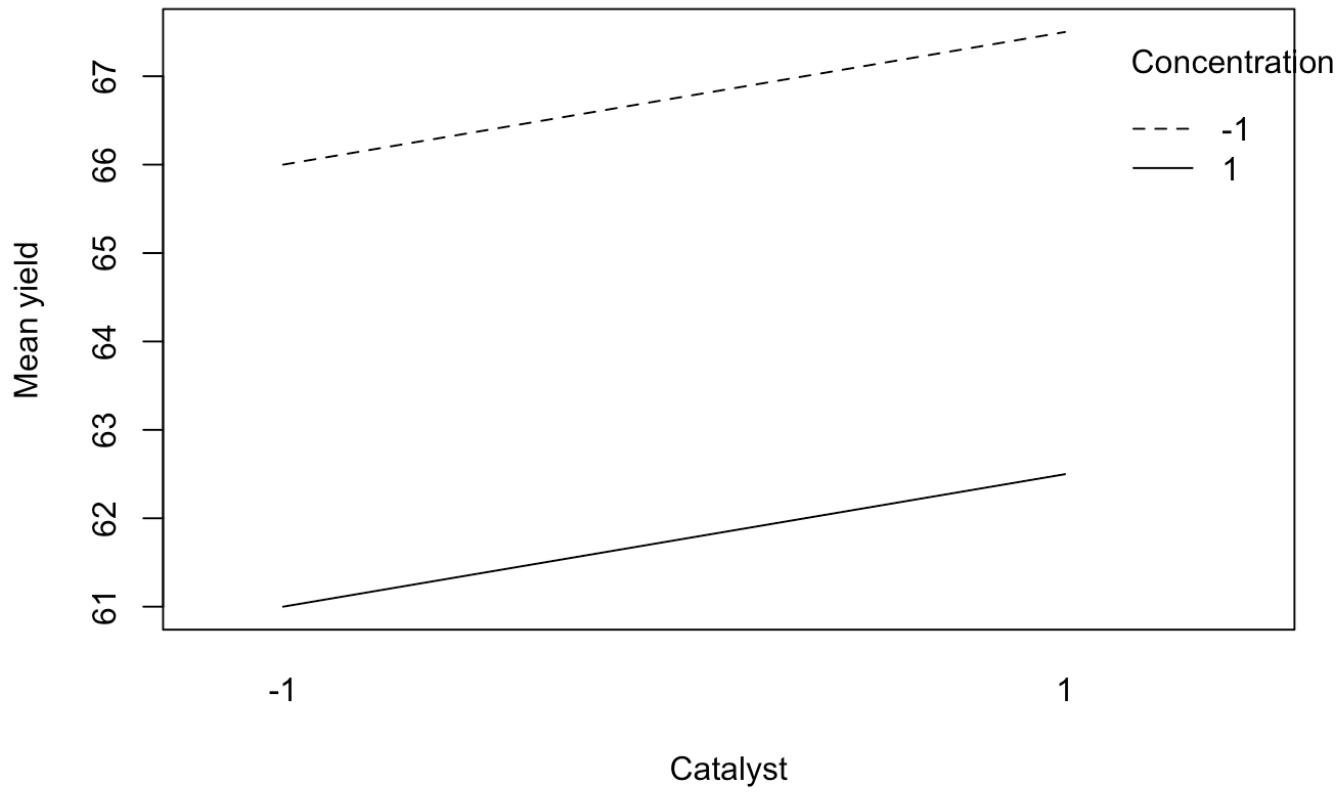
```
interaction.plot(tab0502$T,tab0502$C,tab0502$y, type="l",
                 xlab="Temperature",trace.label="Concentration",
                 ylab="Mean yield")
```




```
interaction.plot(tab0502$T,tab0502$K,tab0502$y, type="l",
                 xlab="Temperature",trace.label="Catalyst",
                 ylab="Mean yield")
```




```
interaction.plot(tab0502$K,tab0502$C,tab0502$y, type="l",
                 xlab="Catalyst",trace.label="Concentration",
                 ylab="Mean yield")
```



5 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:

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

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.

The parameter estimates are obtained via the `lm()` function in R.

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

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    64.38        0.7   92.50     0.00
## T              11.63        0.7   16.70     0.00
## K               0.88        0.7    1.26     0.24
## C              -2.37        0.7   -3.41     0.01
## T:K             5.12        0.7    7.36     0.00
## T:C             0.87        0.7    1.26     0.24
## K:C             0.12        0.7    0.18     0.86
## T:K:C           0.38        0.7    0.54     0.60
```

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

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

If the column of T is mutiplied by the yield average and divided by 4 then the main effect of T is obtained.

- $T = \frac{-60+72-54+68-52+83-45+80}{4} = 23$. The divisor of 4 transforms the constrast into a difference between two averages.
- Signs for interaction contrasts obtained by multiplying signs of their respective factors.
- Each column perfectly balanced (equal numbers of positives and negatives) with respect to other columns.
- Balanced (orthogonal) design ensures each estimated effect is unaffected by magnitude and signs of other effects.

The estimated least squares coefficients are one-half the factorial estimates, and the intercept β_0 is the sample mean. Therefore, the factorial estimates are twice the least squares coefficients. For example,

$$\hat{\beta}_1 = 11.63 \Rightarrow T = 2 \times 11.63 = 23.26$$
$$\hat{\beta}_2 = 0.88 \Rightarrow K = 2 \times 0.88 = 1.75$$
$$\hat{\beta}_4 = 5.12 \Rightarrow TK = 2 \times 5.12 = 10.25$$

The least squares estimates can be multiplied by 2 in R.

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

```
## (Intercept)           T           K           C         T:K         T:
C
##      128.75       23.25        1.75       -4.75       10.25        1.7
5
##         K:C       T:K:C
##        0.25        0.75
```

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 $\beta_1 = 0 \Rightarrow T = 0 \Rightarrow \mu_{T+} - \mu_{T-} = 0 \Rightarrow \mu_{T+} = \mu_{T-}$,

where 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°. The p-value for temperature is small (Pr(>|t|)=0). This means that there is evidence that the mean yield is different at 180° compared to 160°.

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()`.

```
2*confint.lm(fact.mod)
```

```
##                   2.5 %     97.5 %
## (Intercept) 125.540178 131.959822
## T            20.040178  26.459822
## K            -1.459822   4.959822
## C            -7.959822  -1.540178
## T:K           7.040178  13.459822
## T:C          -1.459822   4.959822
## K:C          -2.959822   3.459822
## T:K:C        -2.459822   3.959822
```

The 95% confidence interval for the main effect of concentration is (-8.0,-1.5), and the two-way interaction between temperature and concentration has 95% confidence interval (-1.46,4.96).

6 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.

7 Questions

1. (Adapted from BHH question 8, pg. 227) A chemist performed an experiment with temperature at 130 and 150 degrees celcius and two catalysts. The chemist performed three runs randomizing the order of runs within each week.

Run number	**Temperature**	**Catalyst**
1	130	1
2	130	2
3	150	1

Is this a factorial experiment? If it is not a factorial experiment then is it possible to turn this design into a factorial design? Explain?

2. What is the table of contrasts for a 2^3 factorial design.
3. A 2^2 factorial design involved two factors A and B. The main effects for A and B are 10 and 12 respectively. The lab technician that ran the experiment discovered that he made an error in recording the experimental results: whenever factor A was set to the + level the measurement should be increased by 5 (i.e., if y_i is the measurement when A is set to the + level then the measurement should have been recorded as $y_i + 5$) What are the correct main effects for A and B?
4. Suppose that you were studying two factors A and B each at two levels in a 2^2 factorial design.

a. Write a linear model with parameters that correpond to the main effects.

b. Use least squares to estimate the parameters in part (a).

5. (Box, Hunter, and Hunter problem 5.6) A study was conducted to determine the effects of individual bathcrs on the fecal and total coliform bacterial populations in water. The variables of interest were the time since the subject's last bath, the vigor of the subject's activity in the water, and the subject's sex. The experiments were perfom1ed in a 100-gallon polyethylene tub using dechlorinated tap water at 38°C. The bacterial contribution of each bather was determined by subtracting the bacterial concentration measured al 15 and 30 minutes from that measured initially.

A replicated 2^3 factorial design was used for this experiment.

Code	Name	Low Level	High Level
x_1	Time since last bath	1 hour	24 hour
x_2	Vigor of bathing activity	Lethargic	Vigorous
x_3	Sex of bather	Female	Male

Code	Name
y_1	Fecal coliform contribution after 15 minutes (organisms/100 mL)
y_2	Fecal coliform contribution after 30 minutes (organisms/100 mL)
y_3	Total coliform contribution after 15 minutes (organisms/100 mL)
y_4	Total coliform contribution after 30 minutes (organisms/1OO mL)

The data are shown in the table below.

run	x1	x2	x3	y1	y2	y3	y4
1	-1	-1	-1	1	1	3	7
2	1	-1	-1	12	15	57	80
3	-1	1	-1	16	10	323	360
4	1	1	-1	4	6	183	193
5	-1	-1	1	153	170	426	590
6	1	-1	1	129	148	250	243
7	-1	1	1	143	170	580	450
8	1	1	1	113	217	650	735

9	-1	-1	-1	2	4	10	27
10	1	-1	-1	37	39	280	250
11	-1	1	-1	21	21	33	53
12	1	1	-1	2	5	10	87
13	-1	-1	1	96	67	147	193
14	1	-1	1	390	360	1470	1560
15	-1	1	1	300	377	665	810
16	1	1	1	280	250	675	795

a. Calculate main and interaction effects on fecal and total coliform populations after 15 and 30 minutes.

b. Use R to calculate the main and interaction effects for y_3.

c. Interpret the main effects and interaction effects in (a). Do you think that the effects are real or noise? Explain.

8 Answers to questions

1. No this is not a factorial experiment since not all factor-level combinations were run. If the chemist added a fourth run setting temperature at 150 and catalyst at 2 then the design would be a 2^2 factorial design.

2. Multiply columns AB, AC, BC, and ABC to obtain interactions.

A	B	C	AB	AC	BC	ABC
-1	-1	-1	1	1	1	-1
1	-1	-1	-1	-1	1	1
-1	1	-1	-1	1	-1	1
1	1	-1	1	-1	-1	-1
-1	-1	1	1	-1	-1	1
1	-1	1	-1	1	-1	-1
-1	1	1	-1	-1	1	-1
1	1	1	1	1	1	1

3. Using the standard design matrix. The main effect for A is $\frac{y_2+y_4-y_1-y_3}{2} = 10.$ and the main effect for B is $\frac{y_3+y_4-y_1-y_2}{2} = 10.$ But the main effect for A should be

$$\frac{(y_2+5)+(y_4+5)-y_1-y_3}{2} = \frac{y_2+y_4-y_1-y_3}{2} + \frac{10}{2} = 10 + \frac{10}{2} = 15.$$

The main effect for B should be:

$$\frac{-(y_2+5)+(y_4+5)-y_1+y_3}{2} = \frac{-y_2+y_4-y_1+y_3}{2} = 12.$$

So the main effect of A changes, but the main effect of B remains unchanged.

4. a. Let A and B be the two factors with two levels denoted + and -.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

where, $i = 1, 2, 3, 4$,

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

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

b. The parameter estimates can be obtained by minimizing

$$L(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{4} \epsilon_i^2 = \sum_{i=1}^{4} (y_i - \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})^2.$$

This can be done directly, by directly solving $\frac{\partial L}{\partial \beta_i} = 0$. It's also acceptable to use the fact the $\hat{\beta} = (X'X)^{-1}X'y$, where

$$X = \begin{bmatrix} 1 & -1 & -1 \\ 1 & +1 & -1 \\ 1 & -1 & +1 \\ 1 & +1 & +1 \end{bmatrix}.$$

Use R to calculate $(X'X)^{-1}X'$

```
X <- matrix(c(1,1,1,1,-1,1,-1,1,-1,-1,1,1),nrow=4,ncol=3)
solve(t(X)%*%X)%*%t(X)
```

```
##      [,1]  [,2]  [,3] [,4]
## [1,]  0.25  0.25  0.25 0.25
## [2,] -0.25  0.25 -0.25 0.25
## [3,] -0.25 -0.25  0.25 0.25
```

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} .25 & .25 & .25 & .25 \\ -.25 & .25 & -.25 & .25 \\ -.25 & -.25 & .25 & .25 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^{4} y_i}{4} \\ \frac{y_2+y_4-y_1-y_3}{4} \\ \frac{y_3+y_4-y_1-y_2}{4} \end{bmatrix}$$

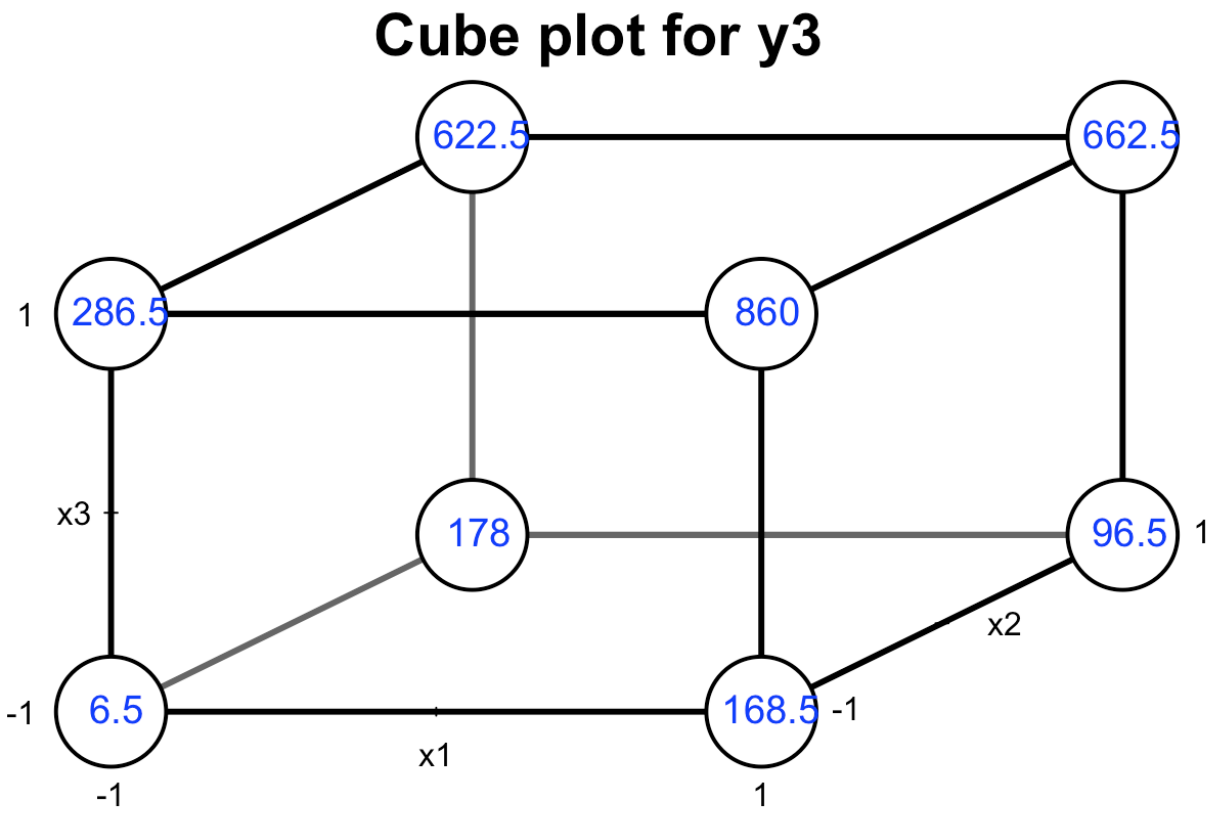
Notice that $\hat{\beta_0} = \bar{y}, \hat{\beta_1} = \frac{1}{2}\hat{A}$, and $\hat{\beta_2} = \frac{1}{2}\hat{B}$, where $\hat{A}$ and $\hat{B}$ are the estiamted factorial effects of A and B.

5. a.

```
prb0506 <- read.table("~/Dropbox/Docs/sta305/BHHData/BHH2-Data/prb0506.da
t", header=TRUE, quote="\"")
fact.mod <- lm(y3~x1*x2*x3,data=prb0506)
knitr::kable(summary(fact.mod)$coefficients)
```

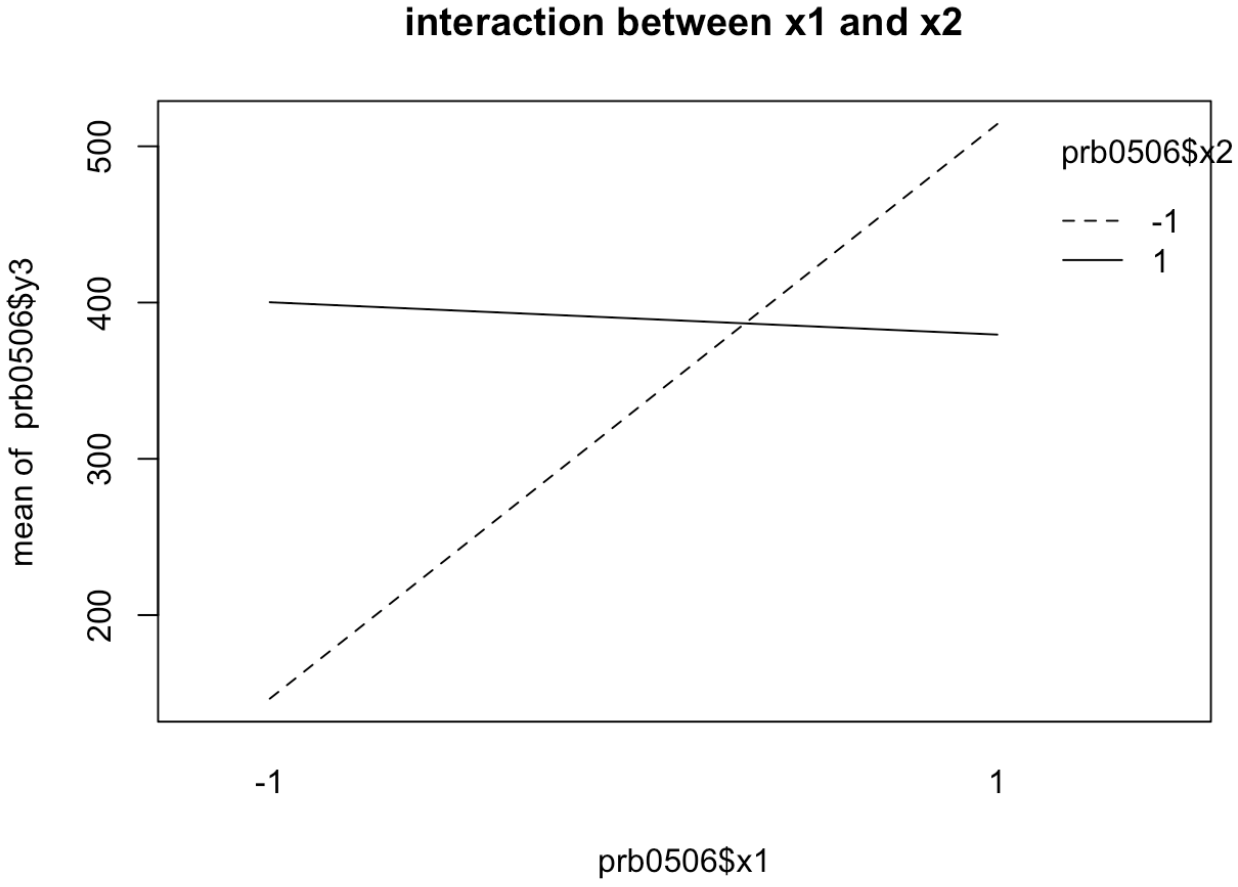
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	360.125	82.39335	4.3708018	0.0023779
x1	86.750	82.39335	1.0528762	0.3231570
x2	29.750	82.39335	0.3610728	0.7273943
x3	247.750	82.39335	3.0069174	0.0168927
x1:x2	-97.125	82.39335	-1.1787966	0.2723484
x1:x3	66.625	82.39335	0.8086211	0.4421145
x2:x3	4.875	82.39335	0.0591674	0.9542701
x1:x2:x3	-36.250	82.39335	-0.4399627	0.6716081

```
cubePlot(fact.mod,"x1","x2","x3",round=1,modeled=F)
```

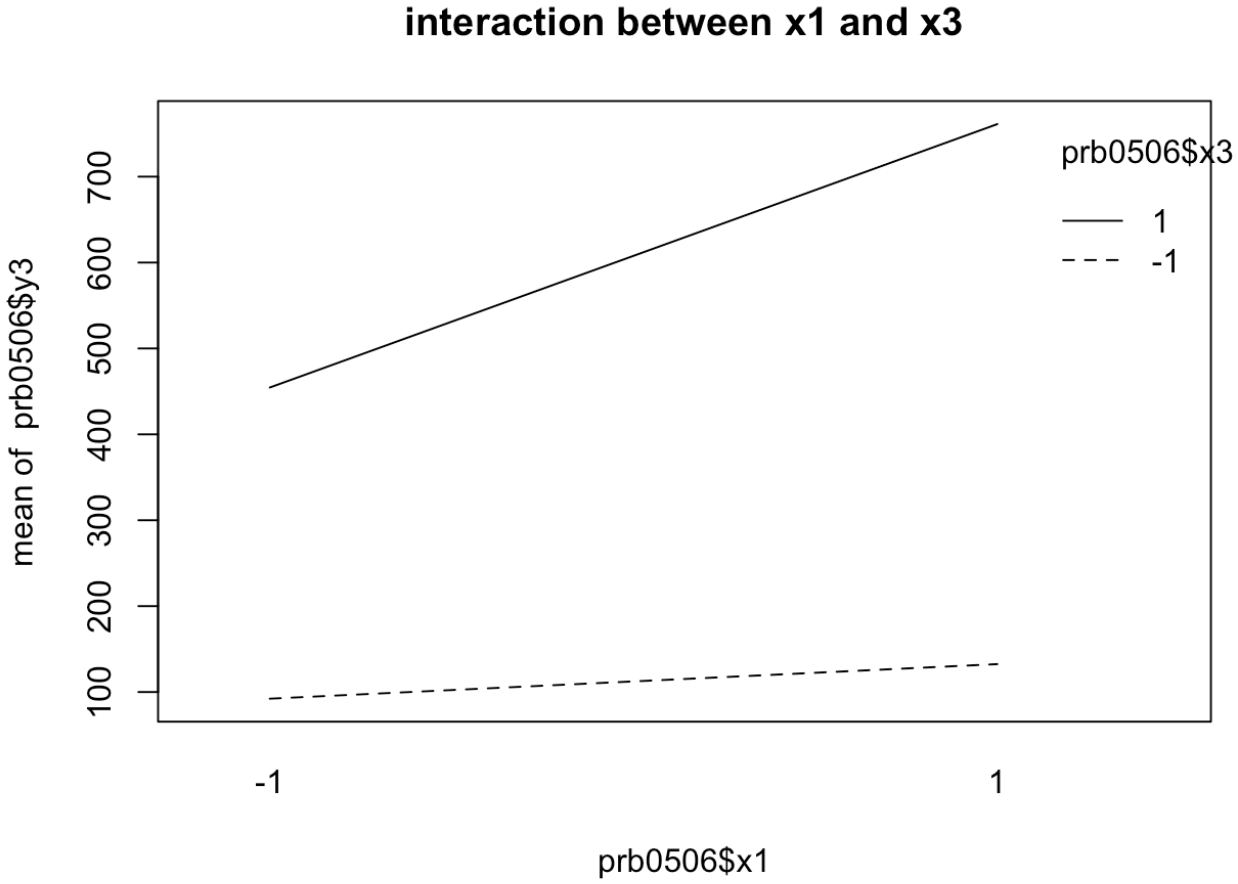


modeled = FALSE

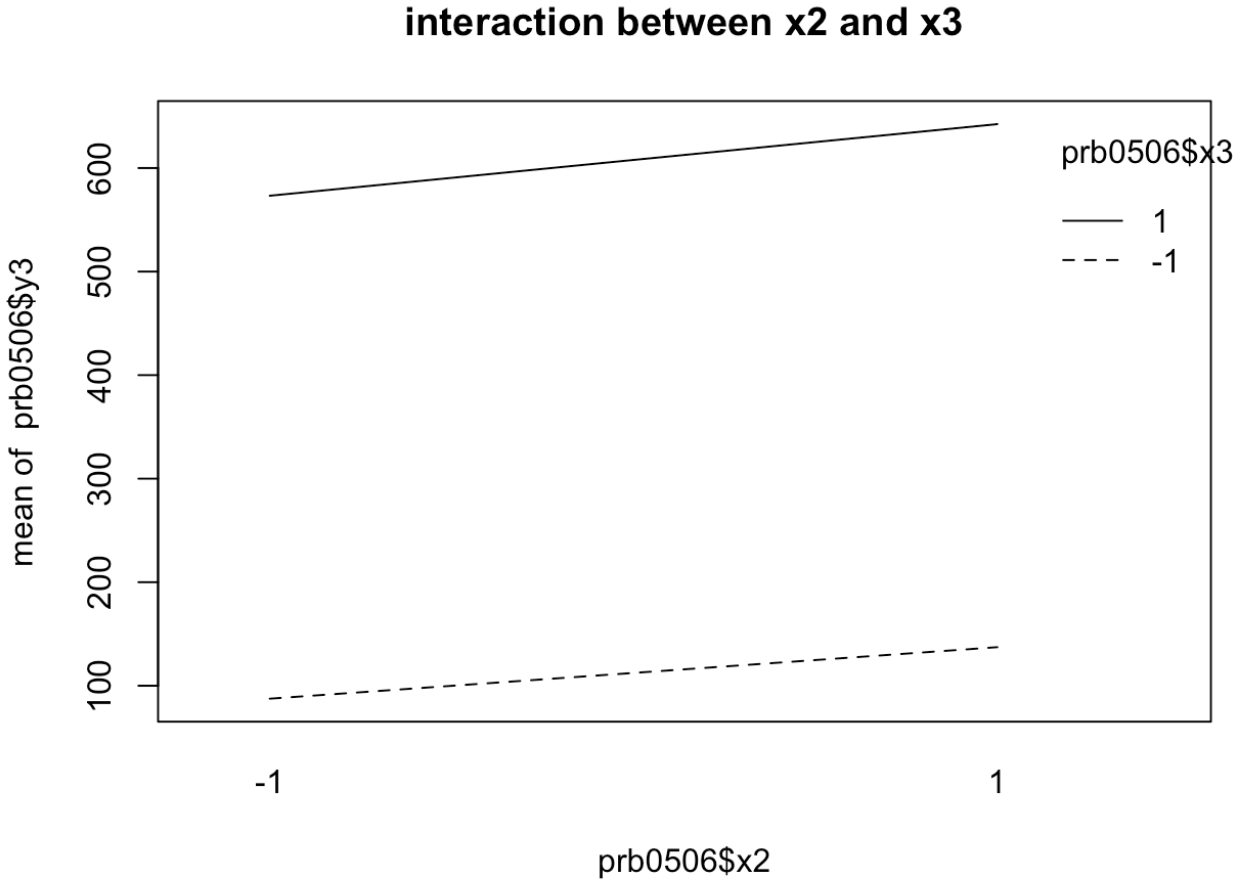
```
interaction.plot(prb0506$x1,prb0506$x2,prb0506$y3,main="interaction betwe
en x1 and x2")
```




```
interaction.plot(prb0506$x1,prb0506$x3,prb0506$y3,main="interaction betwe
en x1 and x3")
```




```
interaction.plot(prb0506$x2,prb0506$x3,prb0506$y3,main="interaction betwe
en x2 and x3")
```



The main effects and interactions are obtained by multiplying the regression coeffcicents by 2

(Intercept)	720.25
x1	173.50
x2	59.50
x3	495.50
x1:x2	-194.25
x1:x3	133.25
x2:x3	9.75
x1:x2:x3	-72.50

b. The main effect for x_1 is interpreted as the change in total coliform contribution after 15 minutes when time since last bath x_1 increases from 1 hour to 24 hours.

The interaction between x_1 and x_2 is interpreted as the difference in the change in total coliform contribution after 15 minutes when time since last bath x_1 increases from 1 hour to 24 hours for the two different types of bathing activity.

The interaction plot for x_1 and x_2 indicates that there is an interaction, but the hypothesis test doesn't support this conclusion (P-value=0.27). The other interaction plots are consistient with the hypothesis tests.

The main effect for x_3 is the only effect that seems real (P-value=0.02). The cube plot also supports the large effect that x_3 has on y_3. The other effects might be due to noise.