Comparative Studies involving Multiple Factors – Factorial Designs

Tim Robinson

Factorial Treatment Structure

- Up to this point, we have looked at completely randomized designs where g treatments are assigned at random to N experimental units.

- Factorial treatment structure exists when the g treatments are the combinations of levels of two or more factors.

- Terminology is best defined via an example. Suppose we would like to study the effect of aspect and habitat type on abundance of ocelots. Aspect is broken into three categories (NE, SE, and W). Habitat types are: tropical deciduous forest (tdf), semi-deciduous forest (sdf), and grasslands (grass). Abundance is measured in ocelots per 1000 m^2.

Factors: Habitat type and Aspect

Levels of Habitat: TDF, SDF, Grass

Levels of Aspect: NE, SE, W

g = 9 combinations of Habitat and Aspect

Experimental Design with Multiple Factors

- Burn time is an important quality measure for candles and burn time is thought to be influenced by potentially many factors…some are given below

y =burn time of candle (continuous variable)

X1 = amount of fragrance oil

X2 = wick diameter

X3 = dye concentration

X4 = wax type

$$y = f(X_1, X_2, ..., X_5)$$ $$$ constraints

Experimental Design with Multiple Factors

- Candles are made from wax in large vats…suppose we consider the following levels of each factor:

X1 = amount of fragrance oil = Low,High
X2 = wick diameter = Low,High
X3 = dye concentration = Low,High
X4 = wax type – 3 suppliers

How would you set up a completely randomized design?

Candle Design Set-Up

Factorial Designs - Notation

- We use the notation y_{ijk} to indicate responses in the two-way factorial experiment. In this notation, y_{ijk} is the k^{th} response in the treatment formed from the i^{th} level of factor A and the j^{th} level of factor B.

- Data from a two-way factorial is often represented in a table with rows corresponding to the levels of one factor and columns corresponding to the levels of another factor.

Factorial Designs – *Ocelot1.csv*

Ex. Ocelot1.csv

Factor 1 Habitat Type	Factor 2 (Aspect)			
	NE	SE	W	Average
TDF	6.6, 6.2	8.1, 7.9	6.3, 5.8	6.817
SDF	5.5, 5.1	7.0, 6.5	5.2, 4.9	5.7
Grass	2.1, 1.7	2.6, 2.4	0.8, 0.6	1.7
Average	4.533	5.75	3.933	4.739

Main Effects and Interaction Questions

- Effects involving the individual factors are knows as 'main effects'

    1. Are ocelots more abundant in certain habitats?
    2. Are ocelots more abundant in certain aspects?

- An 'interaction effect' is present when the effect of one factor depends upon the effect of another factor

    3. If there is a difference in abundance across habitats, does the difference depend upon aspect? Similarly, if there is a difference in aspects, does this difference depend upon habitat type?

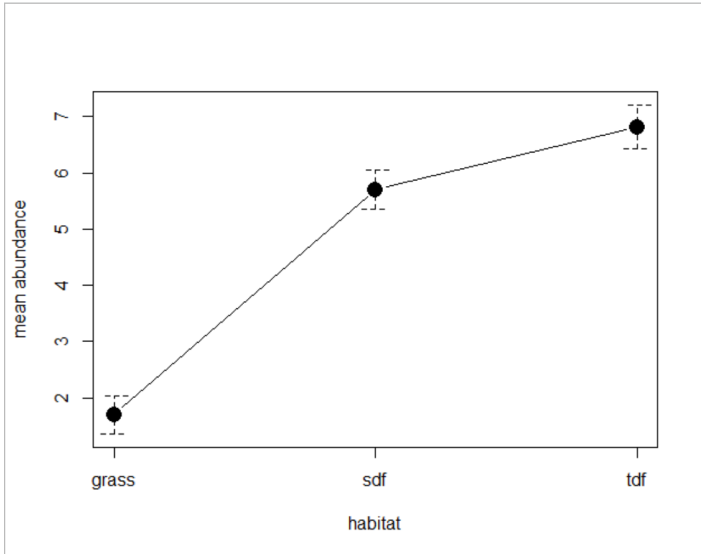
Plotting Your Data

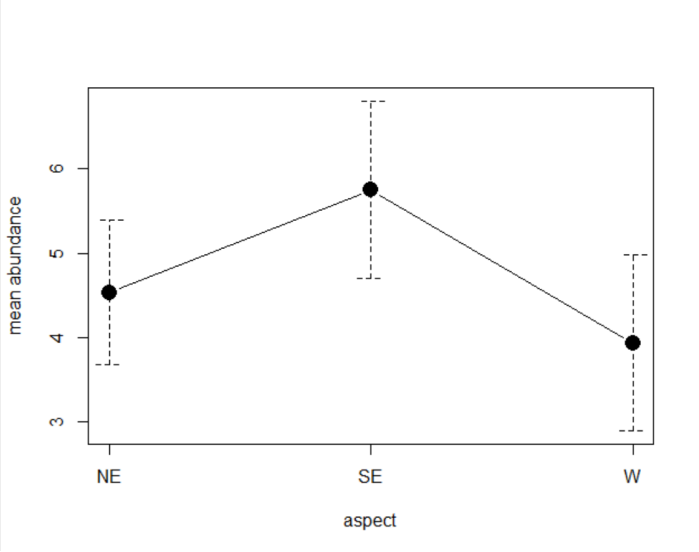
- Before conducting a statistical analysis of the data, you should always plot your data
  - Main effects plots show how the response changes across the levels of the individual factors – you plot the mean response for each level of your factor
  - Interaction plots show how the response changes over the levels of one factor for each setting of the other factor – if the response profile for a particular factor is consistent over the levels of the other factor, the two factors do NOT interact

Main Effects Plots

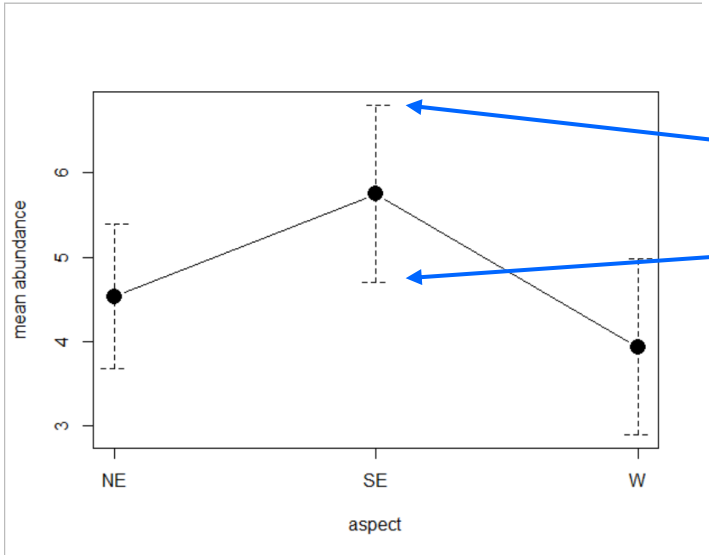
R code:

```
plotMeans(ocelot1$abund, ocelot1$aspect,
          error.bars="se",main="", xlab="aspect",ylab="mean abundance")

plotMeans(ocelot1$abund, ocelot1$habitat,
          error.bars="se",main="",xlab="habitat",ylab="mean abundance")
```





Standard Error Bars in plotMeans



$$\bar{y}_{SE..} \pm \frac{s_{SE}}{\sqrt{6}} = 5.75 \pm \frac{2.585}{\sqrt{6}}$$

$$= 5.75 \pm 1.055$$

$$= [4.695, 6.805]$$

error.bars="se"

$$\bar{y}_{SE..} \pm t_{6-1} \frac{s_{SE}}{\sqrt{6}} = 5.75 \pm 2.57 \frac{2.585}{\sqrt{6}}$$

$$= 5.75 \pm 2.57 * 1.055$$

$$= 5.75 \pm 2.712$$

$$= [3.038, 8.462]$$

error.bars="conf.int"

aspect	mnhab	sdhab	count	stderrmn
NE	4.533	2.110	6.000	0.861
SE	5.750	2.585	6.000	1.055
W	3.933	2.552	6.000	1.042

Interpretation of Main Effects Plots

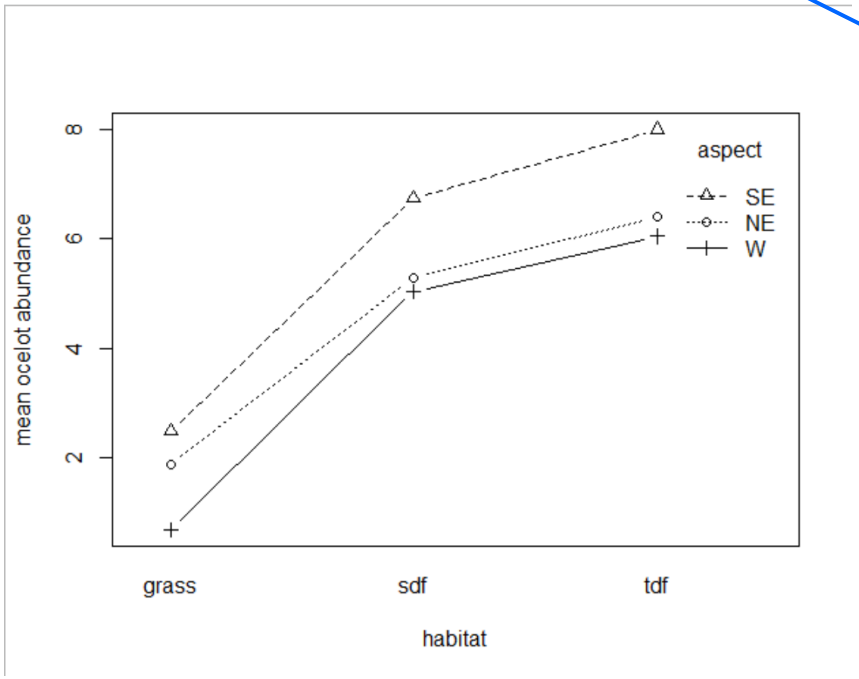
Interaction Effect Plots

R code:

```
interaction.plot(ocelot1$habitat, ocelot1$aspect,
                 ocelot1$abund, fun = mean, type = "b",
                 pch=c(1:3), legend = TRUE, trace.label = "aspect",
                 fixed = FALSE, xlab = "habitat",ylab = "mean ocelot abundance")
```

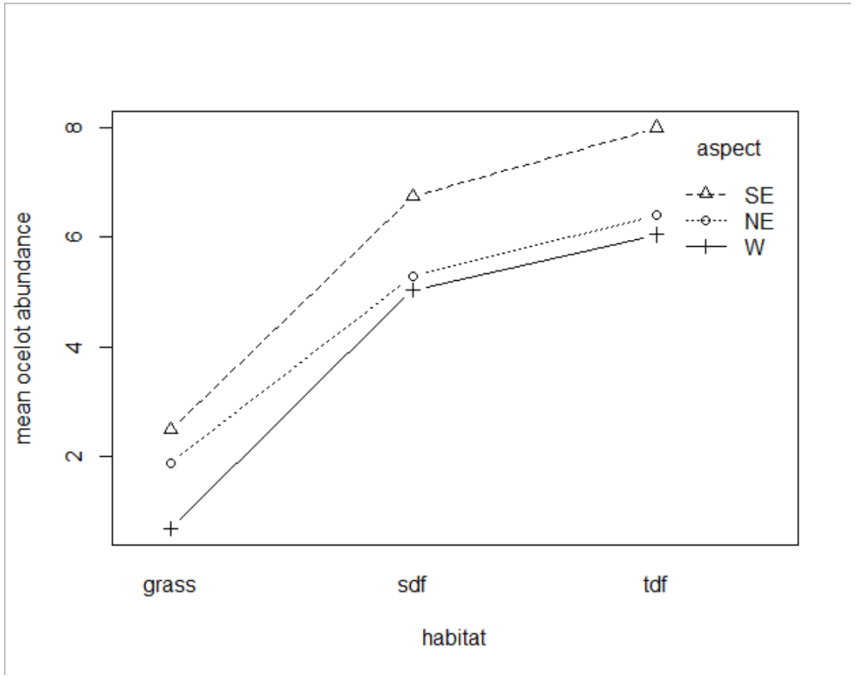
Plots 'both' points and lines

pch option indicates that we want 3 different plot symbols – one for each aspect



Computes the mean response at each combination of habitat and aspect

Interaction Plot Interpretation

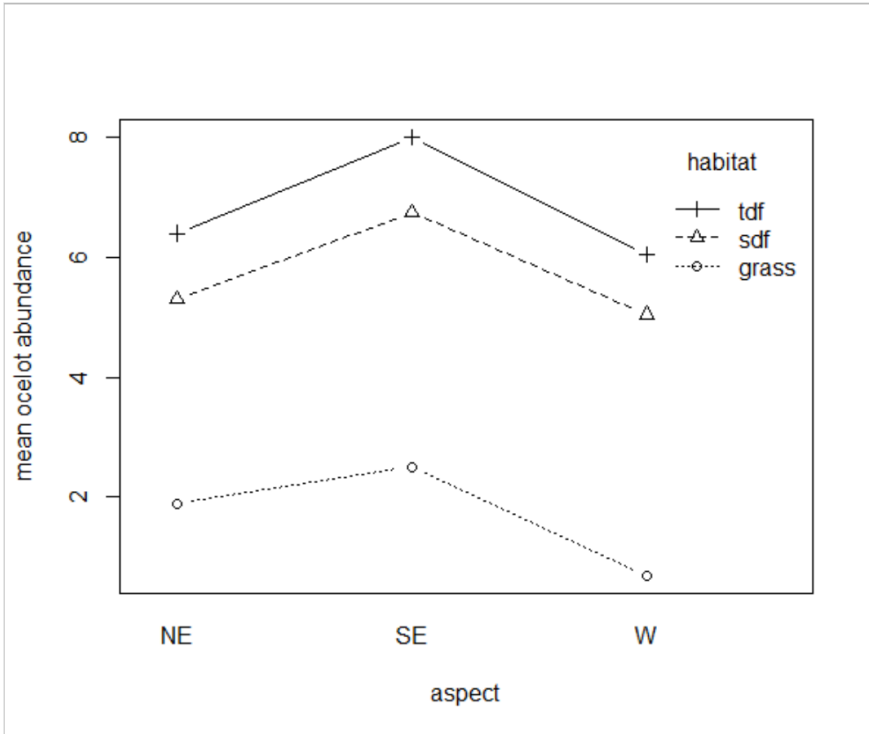


Note that the story of abundance for each habitat (grass, sdf, and tdf) is the same for each aspect

Less ocelots in grass than sdf than in tdf, regardless of the aspect

Since the effect of habitat is the same for each aspect, we observe that there appears to be NO interaction between habitat and aspect

Interaction Plot Interpretation



Note that the story of abundance for each aspect (NE, SE, and W) is the same for each habitat

More ocelots on SE slope for any habitat type that you choose

NO interaction between habitat and aspect

Factorial Study ANOVA

- The logic in the factorial ANOVA is the same as that in the one-way ANOVA. We seek to determine if the variation due to effect A is more than what we would expect due to random error? Is the variation due to effect B more than what we would expect due to random error? Finally, is the variability due to the interaction more than what we would expect due to random error?

Factorial Design Model

Statistical Model:

Interaction term

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

$\alpha_i = \mu_{i.} - \mu$ Effect of factor A (difference in row means)

$\beta_j = \mu_{.j} - \mu$ Effect of factor B (difference in column means)

$(\alpha\beta)_{ij} = \mu_{ij.} - \mu_{i.} - \mu_{.j} + \mu$

ε_{ijk} experimental error variance

Interpreting Experimental Error

- The interpretation of experimental error is identical to what it was in One-Way studies – it represents the natural variation among experimental units which are exposed to the same treatment combination
- Here, a treatment combination is any combination of Habitat and Aspect – ex. TDF, NE is a treatment combination.

Factor 1 Habitat Type	Factor 2 (Aspect)			
	NE	SE	W	Average
TDF	6.6, 6.2	8.1, 7.9	6.3, 5.8	6.817
SDF	5.5, 5.1	7.0, 6.5	5.2, 4.9	5.7
Grass	2.1, 1.7	2.6, 2.4	0.8, 0.6	1.7
Average	4.533	5.75	3.933	4.739

Assume variation in experimental units exposed to the same treatment combination is the same for each treatment combination

Partition of Sources of Variation

- Just as in the one-factor setting, we can partition the sources of variation as follows:

y_{ijk} = k^{th} response at the i^{th} setting of factor A and the j^{th} setting of factor B

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(y_{ijk} - \bar{y}_{...}\right)^2 = bn\sum_{i=1}^{a}\left(\bar{y}_{i..} - \bar{y}_{...}\right)^2 + an\sum_{j=1}^{a}\left(\bar{y}_{.j.} - \bar{y}_{...}\right)^2$$

$$+n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}...\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(y_{ijk} - \bar{y}_{ij.}\right)^2$$

$SS_{Total} = SS_A + SS_B + SS_{AB} + SS_{Error}$

**Observe what happens when n=1

Effects - Changes Between Rows and Columns

- The analysis of this model works the same way as the analysis of the one factor experiment. We ask ourselves, is the variation due to effect A more than what we would expect due to random error? Is the variation due to effect B more than what we would expect due to random error?

- Before we proceed to the analysis of the model on the previous page, we have to discuss the notion of the *interaction* of factors A and B.

- The main effect of rows tells us how the response changes when we move from one row to another, averaged across all columns. The main effect for columns tells us how the response changes when we move from one column to another, averaged across all rows.

- The interaction tells us how the change in response depends on columns when moving between rows, or vice versa. When interaction exists, the change in response when moving across the levels of factor A depends on which level of factor B is under consideration.

Variance between levels of factor A = $SS_A / df\ A = MS_A$

Variance between levels of factor B = $SS_B / df\ B = MS_B$

Variance between levels of factor AB = $SS_{AB} / df\ AB = MS_{AB}$

Experimental Error Variance = $SS_{Error} / df\ Error = MS_{Error}$

To do appropriate tests we use: $F = MS_{Factor} / MS_{Error}$

The partition of the total variability (SS Total) into the variability due to Factor A (Habitat), Factor B (Aspect), Factor AB (Interaction) and Experimental Error is arranged in an ANOVA table just as in the one-factor case. The layout is shown below:

Source	Sum of Squares	Df	MS
Habitat	SS Habitat	a - 1	MS_A
Aspect	SS Aspect	b – 1	MS_B
Habitat*Aspect	SS Interact	(a-1)*(b-1)	MS_{AB}
Error	SS Error	ab(n-1)	MS_{Error}
Total	SS Total	N - 1	

Factorial Analysis

R Code

```
m1 <- lm(abund~habitat + aspect + habitat:aspect ,data=ocelot1)
anova(m1)
```

R Output

```
> anova(m1)
Analysis of Variance Table

Response: abund
               Df  Sum Sq Mean Sq F value    Pr(>F)
habitat         2  86.854  43.427 656.8824 1.768e-10 ***
aspect          2  10.281   5.141  77.7563 2.095e-06 ***
habitat:aspect  4   0.772   0.193   2.9202    0.0838 .
Residuals       9   0.595   0.066
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Factorial Analysis with Indicators

R Code

```
m1 <- lm(abund~habitat + aspect + habitat:aspect ,data=ocelot1)
summary(m1)
```

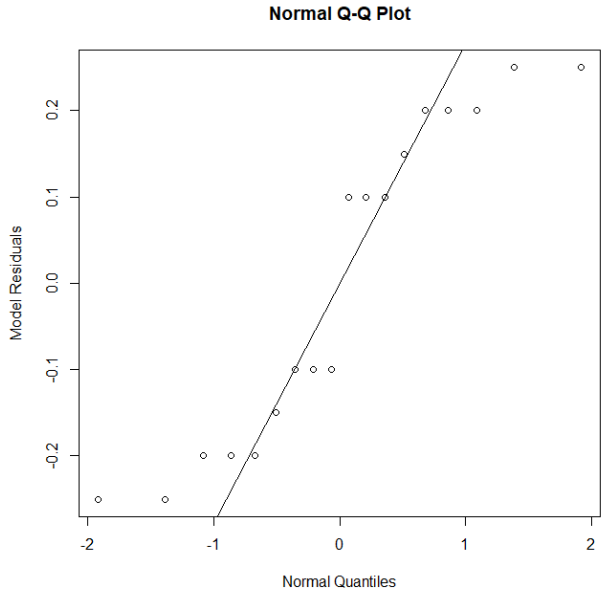
R Output

```
Coefficients:
                           Estimate Std. Error t value      Pr(>|t|)
(Intercept)                  1.9000     0.1818  10.450 0.0000024769 ***
habitat[T.sdf]               3.4000     0.2571  13.223 0.0000003354 ***
habitat[T.tdf]               4.5000     0.2571  17.502 0.0000000294 ***
aspect[T.SE]                 0.6000     0.2571   2.334       0.04449 *
aspect[T.W]                 -1.2000     0.2571  -4.667       0.00117 **
habitat[T.sdf]:aspect[T.SE]  0.8500     0.3636   2.338       0.04420 *
habitat[T.tdf]:aspect[T.SE]  1.0000     0.3636   2.750       0.02247 *
habitat[T.sdf]:aspect[T.W]   0.9500     0.3636   2.613       0.02815 *
habitat[T.tdf]:aspect[T.W]   0.8500     0.3636   2.338       0.04420 *
```

Interpreting the Indicators

Model assumptions

- Just as we assumed with one-way ANOVA, the multi-way ANOVA assumes that your data comes from a Normal distribution – this should be checked via a Normal QQ-plot on the residuals AND with the shapiro-wilk test

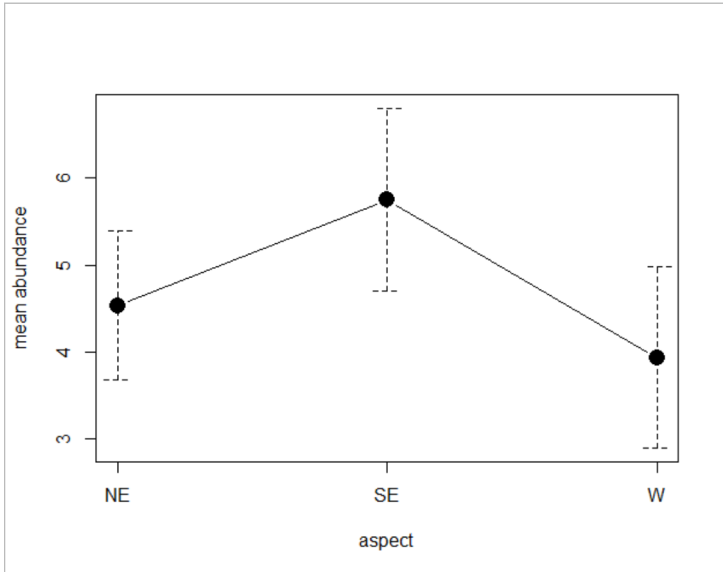



```
> shapiro.test(ocelot1$residsm1)

        Shapiro-Wilk normality test

data:  ocelot1$residsm1
W = 0.87006, p-value = 0.01784
```

This data needs a transformation OR Non-parametric approach…later

Aspect is deemed statistically significant – a paradox?

Perhaps unexpectedly we observed a significant difference in Ocelot abundance across the different Aspects – unexpected since the main effects plot suggested that the 1 standard error bars in the main effects plots overlapped for each Aspect?



One standard error bars overlapped for each of the three Aspects

Factorial MSE vs. One-Factor MSE

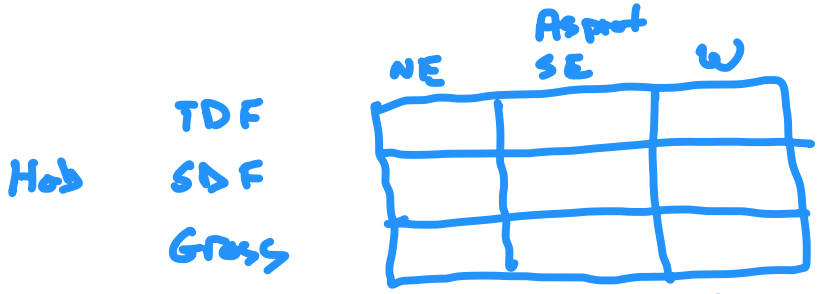
One Factor Analysis

```
> m2_aspect <- lm(abund~aspect,data=ocelot1)
> anova(m2_aspect)
Response: abund
           Df  Sum Sq Mean Sq F value Pr(>F)
aspect      2  10.281  5.1406   0.874 0.4375
Residuals  15  88.222  5.8814
```

Two Factor Analysis

```
Response: abund
              Df  Sum Sq Mean Sq F value    Pr(>F)
habitat        2  86.854  43.427 656.8824 1.768e-10 ***
aspect         2  10.281   5.141  77.7563 2.095e-06 ***
habitat:aspect 4   0.772   0.193   2.9202    0.0838 .
Residuals      9   0.595   0.066
```

Reduced Model



- When the two-factor interaction is unimportant, we can dump this factor into the residual term by specifying a main effects only model

```
> m3 <- lm(abund~habitat+aspect,data=ocelot1)
> anova(m3)
Analysis of Variance Table

Response: abund
           Df  Sum Sq Mean Sq F value    Pr(>F)
habitat     2  86.854  43.427 412.920 1.725e-12 ***
aspect      2  10.281   5.141  48.878 8.959e-07 ***
Residuals 13   1.367   0.105
```

we are only interested in ctr mean comparisons when interaction is significant

3 row means to be compared

3 column means to compare

= total of 6 pairwise comparisons of interest

Final ANOVA Conclusions*

- There is a statistically significant (p-value <0.001) difference in the mean abundance of ocelots across the three habitats.
- There is a statistically significant (p-value<0.0001) difference in the mean abundance of ocelots across the three levels of aspect.
- The mean abundance of ocelots across the three habitats does not appear to depend upon the level of aspect (p-value > 0.08)
  …interaction term dumped into residual for better precision in our multiple comparisons
- **Next step multiple comparisons!**
- **A transformation or Nonparametric approach should be explored – later!!!**

Multiple Comparisons -Factorials

Recall that Tukey's procedure makes use of the distribution of the **studentized range statistic**

$$T_\alpha = \frac{q_\alpha(a, f)}{\sqrt{2}} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

where *a* denotes the number of groups being compared and *f* denotes the error degrees of freedom

For the two-way problem, *a* refers to the total number of means that are being compared. Here, we have that aspect and habitat have significant effects – our multiple comparisons will involve 3 comparisons for habitat and 3 for aspect so a = 6.

Tukey's test declares two means are significantly different if the absolute value of their sample difference exceeds T_α…equivalently, we could construct a set of $100(1-\alpha)$ percent confidence intervals for all pairs of means as follows

$$\bar{y}_{i.} - \bar{y}_{j.} - \frac{q_\alpha(a,f)}{\sqrt{2}}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)},$$

$$\bar{y}_{i.} - \bar{y}_{j.} + \frac{q_\alpha(a,f)}{\sqrt{2}}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

We can use the following R code for conducting the pairwise Tukey comparisons…we need the **multcomp** library to do these comparisons

Comparing habitats:

$$T_\alpha = \frac{q_\alpha(a,f)}{\sqrt{2}}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

$$= \frac{q_{0.05}(6,13)}{\sqrt{2}}\sqrt{0.105\left(\frac{1}{6}+\frac{1}{6}\right)}$$

$$= \frac{4.689}{\sqrt{2}}(0.187) = 0.62$$

The pairwise habitat means must have an absolute difference greater than 0.62 to be declared significantly different at the 0.05 level

You will see that this is the larger than the Tukey margin of error used on the confidence interval construction on mean differences in the TukeyHSD() function

To compare the mean abundances of TDF to those of SDF, we would have the Tukey 95% confidence interval constructed as:

$$(\bar{y}_{TDF..} - \bar{y}_{SDF..}) \pm \frac{q_{\alpha}(6,13)}{\sqrt{2}} \sqrt{MSE\left(\frac{1}{6}+\frac{1}{6}\right)}$$

$$(6.817 - 5.7) \pm 0.62$$

$$[0.55, 1.79]$$

R only looks at the number of levels of the term in the model and thus has 3 instead of 6 for number of groups

$$(\bar{y}_{TDF..} - \bar{y}_{SDF..}) \pm \frac{q_{\alpha}(3,13)}{\sqrt{2}} \sqrt{MSE\left(\frac{1}{6}+\frac{1}{6}\right)}$$

$$(6.817 - 5.7) \pm \frac{3.734}{\sqrt{2}}(0.1872)$$

$$(6.817 - 5.7) \pm 0.494$$

$$[0.62, 1.61]$$

exactly the C.I. produced by R in TukeyHSD( )

Tukey 95% Confidence Intervals from R

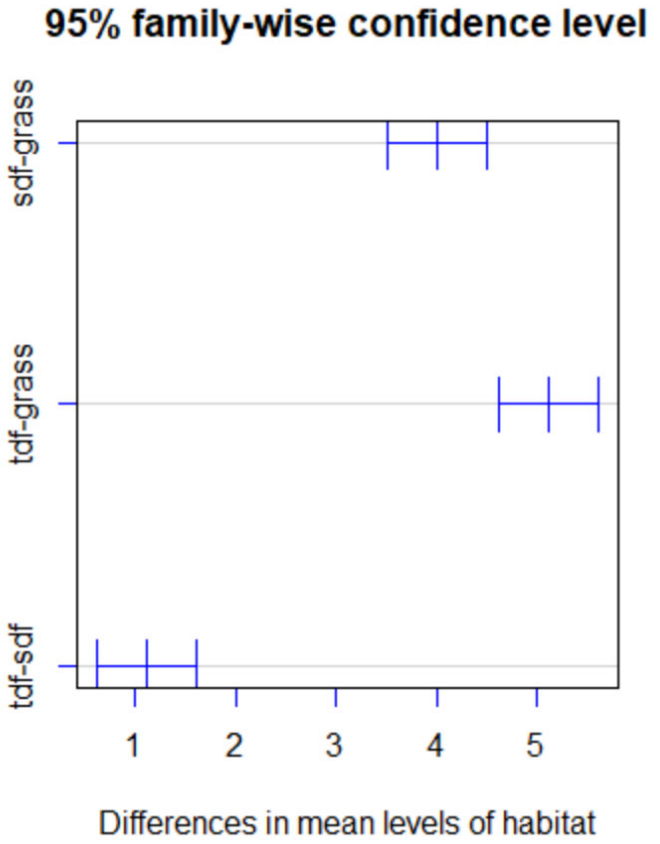
```
> TukeyHSD(aov(m3),conf.level=0.95)
  Tukey multiple comparisons of means 95% family-wise confidence level

$habitat
             diff       lwr       upr     p adj
sdf-grass 4.000000 3.5056175 4.494382 0.0000000
tdf-grass 5.116667 4.6222842 5.611049 0.0000000
tdf-sdf   1.116667 0.6222842 1.611049 0.0001302

$aspect
             diff       lwr        upr     p adj
SE-NE   1.216667 0.7222842  1.7110492 0.0000558
W-NE   -0.600000 -1.0943825 -0.1056175 0.0176863
W-SE   -1.816667 -2.3110492 -1.3222842 0.0000007
```

TukeyHSD() uses the wrong number of groups when both main effects are statistically important

Visualizing the Tukey's Habitat Comparisons

R Code:

```
hab_compare <- TukeyHSD(aov(m3),"habitat",
                                    conf.level=0.95)

#to view the Tukey's comparisons

library(multcompView)
plot(hab_compare , col="blue")
```



Example 2 – Guinea Pig Tooth Growth

- 'Tooth Growth' is a built-in R data set which comprises the measurement of tooth growth length on 60 guinea pigs. Guinea pigs were randomly assigned to combinations of two methods of vitamin C distribution (orange juice or ascorbic acid) and three different dosages (0.5, 1 or 2 mg per day).

- Data is in the file *pigs.csv* on WyoCourses

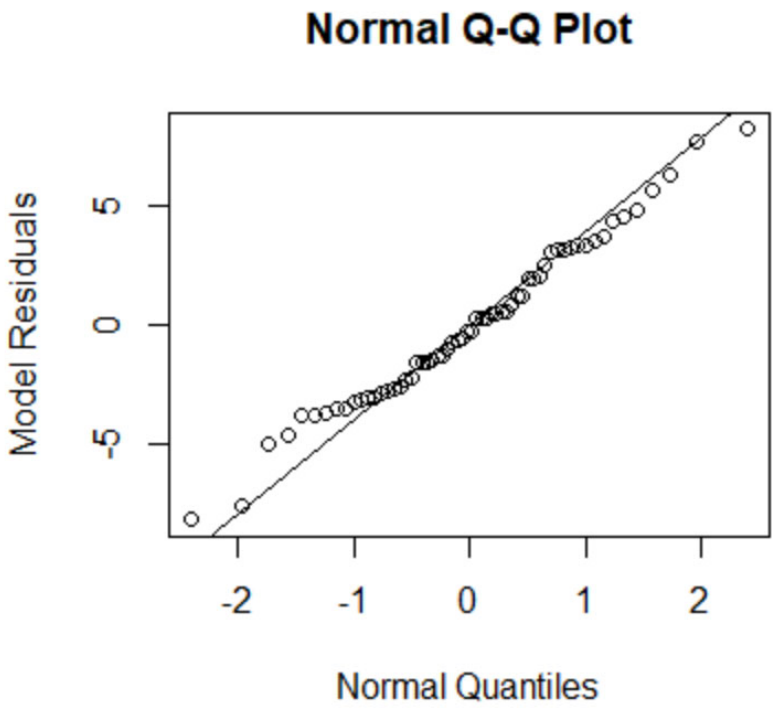
- Want to know how tooth growth is influenced by vitamin C distribution and dosage.

```
> m1 <- lm(len~supp + dose + supp:dose,data=guinea)
> anova(m1)
Analysis of Variance Table

Response: len
            Df  Sum Sq  Mean Sq F value    Pr(>F)
supp         1  205.35   205.35  15.572 0.0002312 ***
dose         2 2426.43  1213.22  92.000 < 2.2e-16 ***
supp:dose    2  108.32    54.16   4.107 0.0218603 *
Residuals   54  712.11    13.19
```

We need to compare all means (6 of them)

Given that there is a significant interaction, main effects **should not** be interpreted!

Residual diagnostics and equal variance check



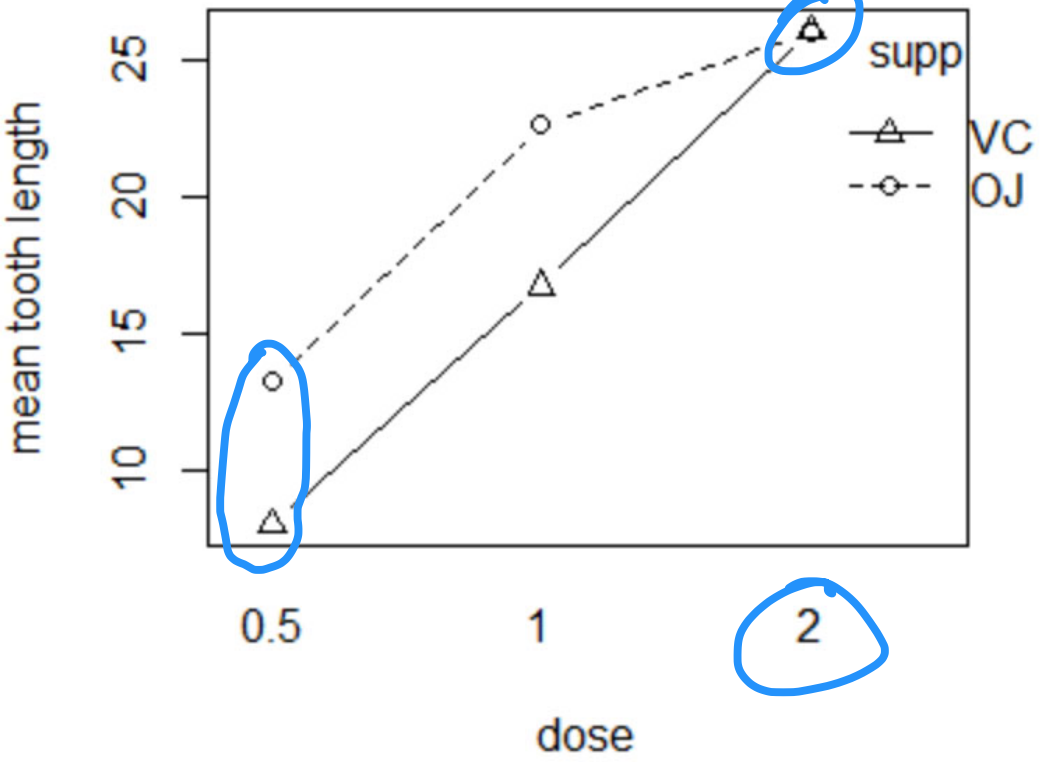

```
> shapiro.test(guinea$residsm1)

        Shapiro-Wilk normality test

data:  guinea$residsm1
W = 0.98499, p-value = 0.6694
```

```
guinea_sum <- guinea %>%
        group_by(supp,dose)%>%
        summarise(varlen = var(len))

var_rat <- max(guinea_sum$varlen)/
        min(guinea_sum$varlen)
var_rat
```

var_rat = 3.63

will contain the 6 sample variances - one for each combo of supp and dose

Interaction Plot



Note that OJ has higher mean tooth growth than VC for low to mid dosages (i.e. 0.5 to 1 mg) but no difference in OJ and VC for the highest dosage

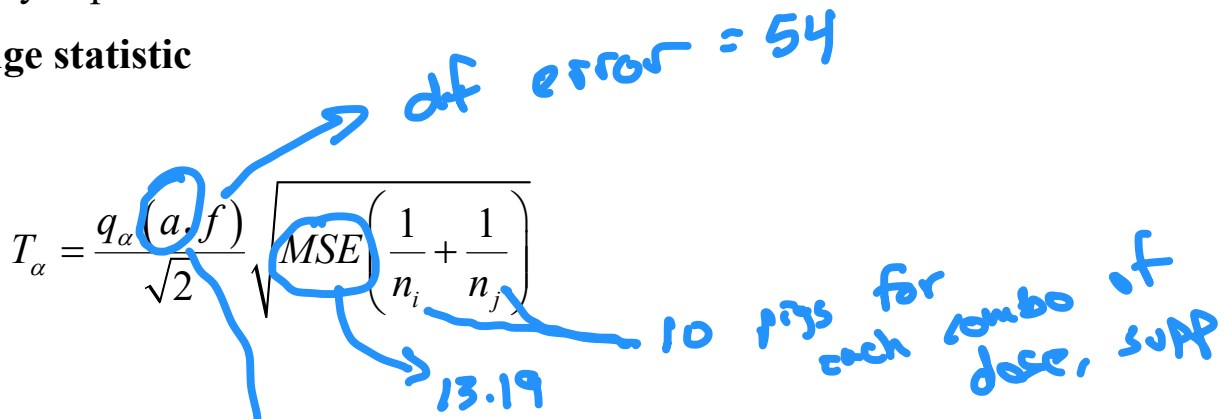
Post-hoc comparisons after significant F-

- Once we determine that there is as significant interaction term we can investigate differences among cell means – either use Tukey's HSD for all pairwise comparisons or a contrast approach which looks at specific differences

- If interaction is not important, focus should turn to investigating which marginal means are different – Tukey's and contrast approach can also be used here

Tukey's for cell mean comparisons

Recall that Tukey's procedure makes use of the distribution of the **studentized range statistic**



$$T_\alpha = \frac{q_\alpha(a, f)}{\sqrt{2}} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

where a denotes the number of groups being compared and f denotes the error degrees of freedom

For the two-way problem, a refers to the total number of means that are being compared. Here, we have dose (0.5,1, and 2) and type of vitamin C (OJ and VC) so a = 6 cell means.

Tukey's for cell mean comparisons

Comparing cell means:

$$T_\alpha = \frac{q_\alpha(a,f)}{\sqrt{2}}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

$$= \frac{q_{0.05}(6,54)}{\sqrt{2}}\sqrt{13.19*\left(\frac{1}{10}+\frac{1}{10}\right)}$$

$$= \frac{4.178}{\sqrt{2}}(1.624) = 4.798$$

qtukey(p=.95, 6, 54) = 4.178

Cell means must differ by at least 4.798 mm in order for them to be declared 'significantly different' at the 0.05 level. to be declared significantly different at the 0.05 level

$\bar{y}_{OJ,1} - \bar{y}_{OJ,0.5} = 9.47$

Tukey's cell mean comparisons in R

R Code

```
cell_mn_guinea <- TukeyHSD(aov(m1),"supp:dose",conf.level=0.95)
cell_mn_guinea
```

compares mean growth for 1 mg, OJ vs 0.5 mg, OJ

9.47 - 4.79B

You should be able to compute the lower and upper values by hand

?

	diff	lwr	upr	p adj
VC:0.5-OJ:0.5	-5.25	-10.048124	-0.4518762	0.0242521
OJ:1-OJ:0.5	9.47	4.671876	14.2681238	0.0000046
VC:1-OJ:0.5	3.54	-1.258124	8.3381238	0.2640208
OJ:2-OJ:0.5	12.83	8.031876	17.6281238	0.0000000
VC:2-OJ:0.5	12.91	8.111876	17.7081238	0.0000000
OJ:1-VC:0.5	14.72	9.921876	19.5181238	0.0000000
VC:1-VC:0.5	8.79	3.991876	13.5881238	0.0000210
OJ:2-VC:0.5	18.08	13.281876	22.8781238	0.0000000
VC:2-VC:0.5	18.16	13.361876	22.9581238	0.0000000
VC:1-OJ:1	-5.93	-10.728124	-1.1318762	0.0073930
OJ:2-OJ:1	3.36	-1.438124	8.1581238	0.3187361
VC:2-OJ:1	3.44	-1.358124	8.2381238	0.2936430
OJ:2-VC:1	9.29	4.491876	14.0881238	0.0000069
VC:2-VC:1	9.37	4.571876	14.1681238	0.0000058
VC:2-OJ:2	0.08	-4.718124	4.8781238	1.0000000

Tukey's comparison of marginal means

- If the interaction term is statistically important (i.e. p-value < 0.05), you should not use Tukey's to make marginal mean comparisons since Tukey's will do all pairwise comparisons for the factor of interest

- If you want to compare a set of marginal means, do so with contrasts ONLY if the conclusion makes sense based on cell means.

Contrasts for Comparing Cell Means

- When there is a significant interaction, focus turns on comparing means for specific treatment combinations
- When comparing OJ to VC, is the mean tooth growth different at a dosage of 1 mg per day?
- Differences between means are known as 'contrasts'
- Contrasts can be used to compare main effect means as well as to compare cell means

BEWARE - I'm going to show you how to do marginal mean contrast in guinea pig ex. for illustratory only since you should not do marginal mean comparisons when the interaction is important

Contrasts

- Here are three possible contrasts from the guinea pig data:

$c_1 = 1$ $c_2 = -1$

$\mu_{1..} - \mu_{2..}$ (Mean growth at 1 mg - Mean growth at 2 mg)

$\mu_{.OJ.} - \mu_{.VC.}$ (Mean growth for OJ - Mean growth for VC)

} 2 marginal comparisons

$\mu_{1,OJ,.} - \mu_{1,VC,.}$ (Mean growth for OJ at 1 mg - Mean growth for VC at 1 mg)

} cell mean contrast

- Contrasts are written as a weighted combination of means

$$C = \sum_{i=1}^{r} c_i \mu_i \quad \text{where } r \text{ denotes the number of means compared}$$

$= c_1 M_1 + c_2 M_2 + \dots + c_R M_R$

c_i's are just numbers

Contrasts

contrasts in R involve linear combinations of β's. I'm going to show you how to translate contrasts in β's to contrasts in means

- Contrast estimates and standard errors are given by

$$\hat{C} = \sum_{i=1}^{r} c_i \bar{y}_i \qquad se\left(\hat{C}\right) = \sqrt{\text{MSE} * \sum_{i=1}^{r} \frac{c_i^2}{n_i}}$$

- A confidence interval on a contrast is given by

$$\sum_{i=1}^{r} c_i \bar{y}_i \pm t_{\frac{\alpha}{2}, df \text{ error}} * \sqrt{MSE \sum_{i=1}^{r} \frac{c_i^2}{n_i}}$$

Suppose we wish to determine if the mean tooth length for guinea pigs treated with 1 mg per day is different than the mean tooth length when treated with 2 mg per day.

This set of hypotheses seeks to determine if avg tooth growth for 1 mg differs from avg growth of 2 mg?

$$H_o: \mu_{.1.} - \mu_{.2.} = 0$$
$$H_1: \mu_{.1.} - \mu_{.2.} \neq 0$$

translate this into β's

marginal mean comparison of 1 mg dose to 2 mg dose

To test this and use R, we will need to think of the regression model that R is analyzing when using the lm() function.

$$y = \beta_0 + \beta_1 suppVC + \beta_2 dose1 + \beta_3 dose2 + \beta_4 suppVC:dose1 + \beta_5 suppVC:dose2$$

How to determine what model R is fitting in terms of β's? → use summary(m1)

$\mu_{OJ,1} = \beta_0 + \beta_2$

$\mu_{VC,1} = \beta_0 + \beta_1 + \beta_2 + \beta_4$

? β_2

R Code:

```
m1 <- lm(len~supp + dose + supp:dose,data=guinea)
Summary(m1)
```

R Output:

```
> summary(m1)

Call:
lm(formula = len ~ supp + dose + supp:dose, data = guinea)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       13.230      1.148  11.521 3.60e-16 ***
suppVC            -5.250      1.624  -3.233  0.00209 **
dose1              9.470      1.624   5.831 3.18e-07 ***
dose2             12.830      1.624   7.900 1.43e-10 ***
suppVC:dose1      -0.680      2.297  -0.296  0.76831
suppVC:dose2       5.330      2.297   2.321  0.02411 *
```

To test $H_o: \mu_{.1.} - \mu_{.2.} = 0$ vs. $H_1: \mu_{.1.} - \mu_{.2.} \neq 0$, we need to translate the means into the $\beta's$.

$$\mu_{.1.} = \frac{1}{2}\left(\mu_{OJ,1,.} + \mu_{VC,1,.}\right) = \frac{1}{2}\left[\left(\beta_0 + \beta_2\right) + \left(\beta_0 + \beta_1 + \beta_2 + \beta_4\right)\right]$$

$$= \beta_0 + \beta_2 + \frac{1}{2}\beta_1 + \frac{1}{2}\beta_4$$

$$\mu_{.2.} = \frac{1}{2}\left(\mu_{OJ,2,.} + \mu_{VC,2,.}\right) = \frac{1}{2}\left[\left(\beta_0 + \beta_3\right) + \left(\beta_0 + \beta_1 + \beta_3 + \beta_5\right)\right]$$

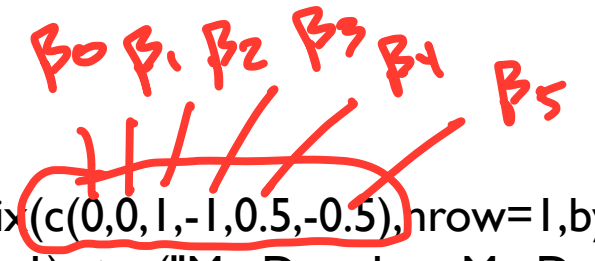
$$= \beta_0 + \beta_3 + \frac{1}{2}\beta_1 + \frac{1}{2}\beta_5$$

$H_o: \mu_{.1.} - \mu_{.2.} = 0$ is equivalent to

$$H_o: \left[\beta_0 + \beta_2 + \frac{1}{2}\beta_1 + \frac{1}{2}\beta_4\right] - \left[\beta_0 + \beta_3 + \frac{1}{2}\beta_1 + \frac{1}{2}\beta_5\right] = 0$$

$$H_o: \beta_2 + \frac{1}{2}\beta_4 - \beta_3 - \frac{1}{2}\beta_5 = 0$$

Contrast Using R



R Code:

```
library(multcomp)
contrast1 <- matrix(c(0,0,1,-1,0.5,-0.5), nrow=1,byrow=T)
rownames(contrast1) <- c("Mn Dose1 vs. Mn Dose2")
colnames(contrast1)<-names(coef(aovm1))
contrast1

summary(glht(m1, linfct = contrast1))
confint(glht(m1, linfct = contrast1))
```

Contrast Using R

R Output

```
> summary(glht(m1, linfct = contrast1))
```

Simultaneous Tests for General Linear Hypotheses

```
                                  Estimate Std. Error t value Pr(>|t|)
Mn Dose1 vs. Mn Dose2 == 0   -6.365      1.148    -5.543  9.12e-07 ***
```

Simultaneous Confidence Intervals

```
Quantile = 2.0049
95% family-wise confidence level

Linear Hypotheses:
                                  Estimate     lwr        upr
Mn Dose1 vs. Mn Dose2 == 0 -6.3650  -8.6673 -4.0627
```

Contrasts by Hand

- Contrast estimates and standard errors are given by

$$\hat{C} = \sum_{i=1}^{r} c_i \bar{y}_i = \bar{y}_{.1.} - \bar{y}_{.2.} = 19.735 - 26.10 = -6.365$$

$$se\left(\hat{C}\right) = \sqrt{\text{MSE} * \sum_{i=1}^{r} \frac{c_i^2}{n_i}} = \sqrt{13.19 * \left(\frac{1}{20} + \frac{1}{20}\right)} = 1.148$$

- A confidence interval on a contrast is given by

$$\sum_{i=1}^{r} c_i \bar{y}_i \pm \text{t}_{\frac{\alpha}{2}, df \text{ error}} * \sqrt{MSE \sum_{i=1}^{r} \frac{c_i^2}{n_i}}$$

$$-6.365 \pm t_{0.975,54} * 1.148$$

$$-6.365 \pm 2.0048 * 1.148$$

$$\left[-8.667, -4.06\right]$$

Suppose we wish to determine if the mean tooth length for guinea pigs treated with orange juice (1 mg per day) is different than the mean tooth length when treated with VC (1 mg per day).

$$H_o : \mu_{OJ,1.} - \mu_{VC,1,.} = 0$$
$$H_1 : \mu_{OJ,1.} - \mu_{VC,1,.} \neq 0$$

1. Write the above in terms of the beta's from below

$$y = \beta_0 + \beta_1 suppVC + \beta_2 dose1 + \beta_3 dose2 + \beta_4 suppVC : dose1 + \beta_5 suppVC : dose2$$

$\mu_{OJ,1.} =$

$\mu_{VC,1.} =$

Suppose we wish to determine if the mean tooth length for guinea pigs treated with orange juice (1 mg per day) is different than the mean tooth length when treated with VC (1 mg per day).

$$H_o : \mu_{OJ,1.} - \mu_{VC,1,.} = 0$$
$$H_1 : \mu_{OJ,1.} - \mu_{VC,1,.} \neq 0$$

2. R Code for testing the above using a contrast

Suppose we wish to determine if the mean tooth length for guinea pigs treated with orange juice (1 mg per day) is different than the mean tooth length when treated with VC (1 mg per day).

$$H_o: \mu_{OJ,1.} - \mu_{VC,1,.} = 0$$
$$H_1: \mu_{OJ,1.} - \mu_{VC,1,.} \neq 0$$

3. Testing the above by hand, obtain $\hat{C}$ and $se\left(\hat{C}\right)$

$$\hat{C} = \sum_{i=1}^{r} c_i \bar{y}_{i1.} = \bar{y}_{OJ1.} - \bar{y}_{VC1.} =$$

$$se\left(\hat{C}\right) = \sqrt{\text{MSE}^* \sum_{i=1}^{r} \frac{c_i^2}{n_i}} =$$

Response Surface Methodology

Introduction to Experimental Designs for Prediction Models

2^k Designs and 2^k Designs with Center Runs

2 refers to the number of levels for each factor

k = # of factors (explanatory variables)

2^K = # of exp. design locations at which data will be collected

Experimental Designs for Prediction Models

- Everything that we have talked about so far has had interest in comparing groups – explanatory variables are categorical. There are many applications where the interest is in prediction rather than group comparison.
- When your are interested in prediction, we use regression analysis. In regression, explanatory variables can be both continuous and categorical.
- The application that we are concerned with is when we want to understand a system and we want to collect data to understand the system – the issue is that data collection is often expensive and we want as cheap of an experiment as possible
- We want to understand how the response is related to the explanatory variables while simultaneously keeping costs low!

$y = f(X_1, X_2, ..., X_5)$ $$$ constraints

y is a function of explanatory variables $x_1, x_2, ..., x_5$

Experimental Designs for Prediction Models

- Candle example – want to optimize 'burn time' of jar candles . . . which combinations of the explanatory variables yield the highest burn time?
- Note from below, there are 5 explanatory variables and an experiment could get large fast!

y =burn time of candle (continuous variable)

X1 = amount of fragrance oil

X2 = wick diameter

X3 = dye concentration

X4 = wax type

X5 = stirring rate

$y \stackrel{?}{=} f(x_1, x_2 \cdots x_r)$

Sequential Nature of RSM

- **Beginning**: long list of variables ($X_1, X_2, \ldots, X_k$) that *could be* important in explaining the response…some factors likely to interact with each other

- First step in 'design' is to select the important variables

- Cheap design and a simple model

$$y = f(X_1, X_2, \ldots, X_5) \approx \beta_0 + \sum_{i=1}^{5} \beta_i X_i + \sum_{j>i} \sum_{i=1}^{5} X_i X_j + \varepsilon$$

linear approximation

Lord Taylor

Two-Level Factorial Designs: A Regression Approach

1 factor 2 levels
1 factor 3 levels
$2^1 \times 3^1$

A study wished to investigate the impact of the reduction in air flow volume (A) and furnace temperature (B) on the concentration of CO^2 (y) in coal burning emissions.

The researchers studied air flow reduction in terms of a percentage ranging from 0% and 22.2%. Furnace temperatures of 2000 C and 2500 C were used.

scale to -1, 1

Treatment Combination	A	B	Conc. of CO^2 in Emissions Replicate 1	Replicate 2
(1)	--	--	20.3	20.4
a	+	--	13.6	14.8
b	--	+	15.0	15.1
ab	+	+	9.7	10.7

"Design" - prescription for data collection

2^2 → A, B
2 → +, -

y = CO^2 concentration

x_1 = A = reduction in air flow % (0%, 22.2%)

x_2 = B = furnace temp (2000, 2500)

Coded Factor Levels

- It is often convenient to "code" the factors, with -1 representing the low level and +1 the high level. Coded levels obtained by subtracting midpoint of variable range and dividing by half of variable range. Here, we have:

$$\text{coded A} = \frac{airflow - 11.1}{11.1}, \quad \text{coded B} = \frac{temp - 2250}{250}$$

	coded levels	
	-1	1
A	0	22.2
B	2000	2500

Natural Levels

The low and high levels of A and B are denoted '--' and '+' respectively – coding allows to compare magnitude of effects from one factor to another (interpretation doesn't depend on scale of each X); intercept is mean response when each value of X is at its center;.

The four treatment combinations are represented by lowercase letters.

More explicitly, *a* represents the combination of factor levels with A at the high level and B at the low level, *b* represents A at the low level and B at the high level and *ab* represents both factors being run at the high level.

By convention, (1) is used to denote A and B each run at the low level.

Regression Analysis Coal

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	14.9500	0.1961	76.254	1.77e-07 ***
A_coded	-2.7500	0.1961	-14.027	0.00015 ***
B_coded	-2.3250	0.1961	-11.859	0.00029 ***
A_coded:B_coded	0.3250	0.1961	1.658	0.17272

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5545 on 4 degrees of freedom
Multiple R-squared: 0.9884, Adjusted R-squared: 0.9797
F-statistic: 113.4 on 3 and 4 DF, p-value: 0.0002523

$$CO^2 = 14.95 - 2.75\,Acode - 2.32\,Bcode + 0.325\,Acode : Bcode$$

Predicted Emissions: A=10; B=2100

$$\text{coded A} = \frac{airflow - 11.1}{11.1}, \quad \text{coded B} = \frac{temp - 2250}{250}$$

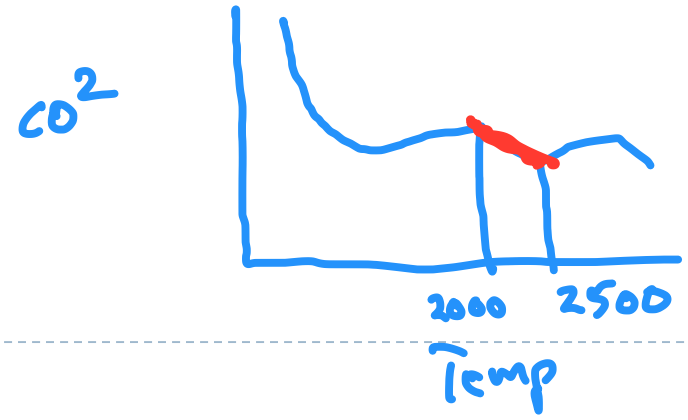
$$\text{coded A} = \frac{10 - 11.1}{11.1}, \quad \text{coded B} = \frac{2100 - 2250}{250}$$

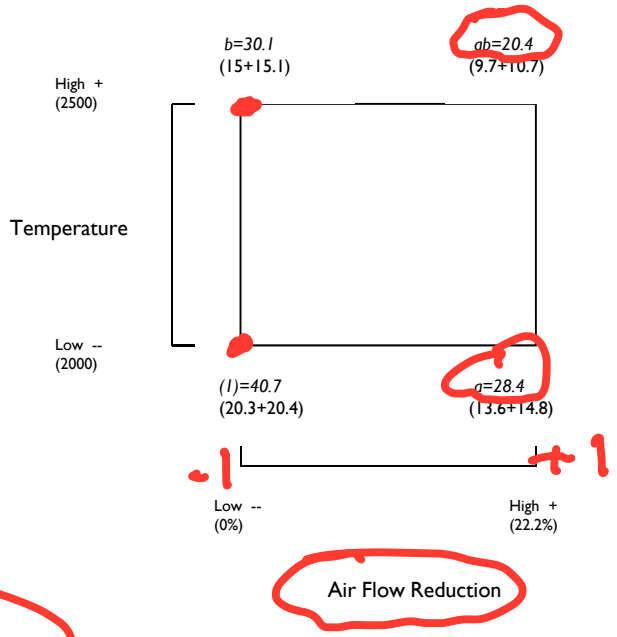
$$\text{coded A} = -0.099, \quad \text{coded B} = -0.6$$

$$CO^2 = 14.95 - 2.75\,Acode - 2.32\,Bcode + 0.325\,Acode : Bcode$$

$$= 14.95\text{-}2.75(\text{-}0.099)\text{-}2.32(\text{-}0.6)\text{+}0.325(\text{-}0.099)(\text{-}0.6)$$

$$= 16.63$$





$\bar{y}_{A^+} =$ diff in sum high to low / # levels of A * # reps

A *main effect* is defined as the difference in the average response at the high level of a factor and the low level of a factor. If this effect is positive, the response increases as we go from the low level of the factor to the high level.

main effect factor A = -5.5

Thus, in general for the 2^2 factorial designs, the main effects for factors A and B are defined as:

$$A = \bar{y}_{A^+} - \bar{y}_{A^-} = \frac{1}{2n}[ab + a - b - (1)], \qquad \text{Flow} = \bar{y}_{A^+} - \bar{y}_{A^-} = \frac{1}{2*2}[20.4 + 28.4 - 30.1 - 40.7] = \frac{-22}{4} = -5.5$$

Right Left

$$B = \bar{y}_{B^+} - \bar{y}_{B^-} = \frac{1}{2n}[ab + b - a - (1)]. \qquad \text{Temp} = \bar{y}_{B^+} - \bar{y}_{B^-} = \frac{1}{2*2}[20.4 + 30.1 - 28.4 - 40.7] = \frac{-18.6}{4} = -4.65$$

reps

2 → # of design points where B high (B low)

The interaction effect, AB, is the difference in the average of the right to left diagonal design points in the figure on the last page 4 [*ab* and (1)] minus the left to right diagonal design points [*b* and *a*], or

$$\text{AB} = \frac{1}{2n}\left[ab + (1) - a - b\right]$$

From these equations, it is evident that each effect in a 2^2 design is a function of a *contrast* of design points. Each contrast represents a single degree of freedom. The sum of squares for any contrast is equal to the contrast squared divided by the number of observations making up each contrast. Thus we can write the sum of squares in the 2^2 design as

$$\text{SS}_\text{A} = \frac{\left[ab + a - b - (1)\right]^2}{4n} \qquad \text{SS}_\text{B} = \frac{\left[ab + b - a - (1)\right]^2}{4n} \qquad \text{SS}_\text{AB} = \frac{\left[ab + (1) - a - b\right]^2}{4n}$$

In the example, the first order regression model with interaction is written

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

By definition, a slope is simply the change in 'y' divided by the change in 'x'. As a result, the slope associated with X_1 (or factor A) is given by

$$\beta_1 = \frac{\bar{y}_{A^+} - \bar{y}_{A^-}}{2} = \frac{\text{Main Effect for A}}{2}$$

In general, any slope parameter in the 2^2 design is just the main effect of the particular factor divided by two.

Coal Emissions Results – R (slopes)

R Code:

```
mcode_coal <- lm(emissions~A_coded+B_coded+A_coded:B_coded,data=coal)
summary(mcode_coal)
```

R Output:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	14.9500	0.1961	76.254	1.77e-07	***
A_coded	-2.7500	0.1961	-14.027	0.00015	***
B_coded	-2.3250	0.1961	-11.859	0.00029	***
A_coded:B_coded	0.3250	0.1961	1.658	0.17272	

Note that the estimates above are slopes = rate of change over 1 unit

Effects are changes over the entire range of the explanatory variable (i.e. from -1 to 1)

Coal Emissions Results – R (ANOVA)

R Code:

```
mcode_coal <- lm(emissions~A_coded+B_coded+A_coded:B_coded,data=coal)
anova(mcode_coal)
```

R Output:

```
> anova(mcode_coal)
Analysis of Variance Table

Response: emissions
                Df Sum Sq Mean Sq F value    Pr(>F)
A_coded          1 60.500  60.500 196.748 0.0001499 ***
B_coded          1 43.245  43.245 140.634 0.0002895 ***
A_coded:B_coded  1  0.845   0.845   2.748 0.1727182
Residuals        4  1.230   0.308
```

Response Surface Model

The intercept, β_o, is the grand mean of the responses. For the coal example, the fitted regression model is then given by

$$\hat{y} = 14.95 + \left(\frac{-5.5}{2}\right)\text{Acode} + \left(\frac{-4.65}{2}\right)\text{Bcode} + \left(\frac{0.65}{2}\right)Acode:Bcode$$

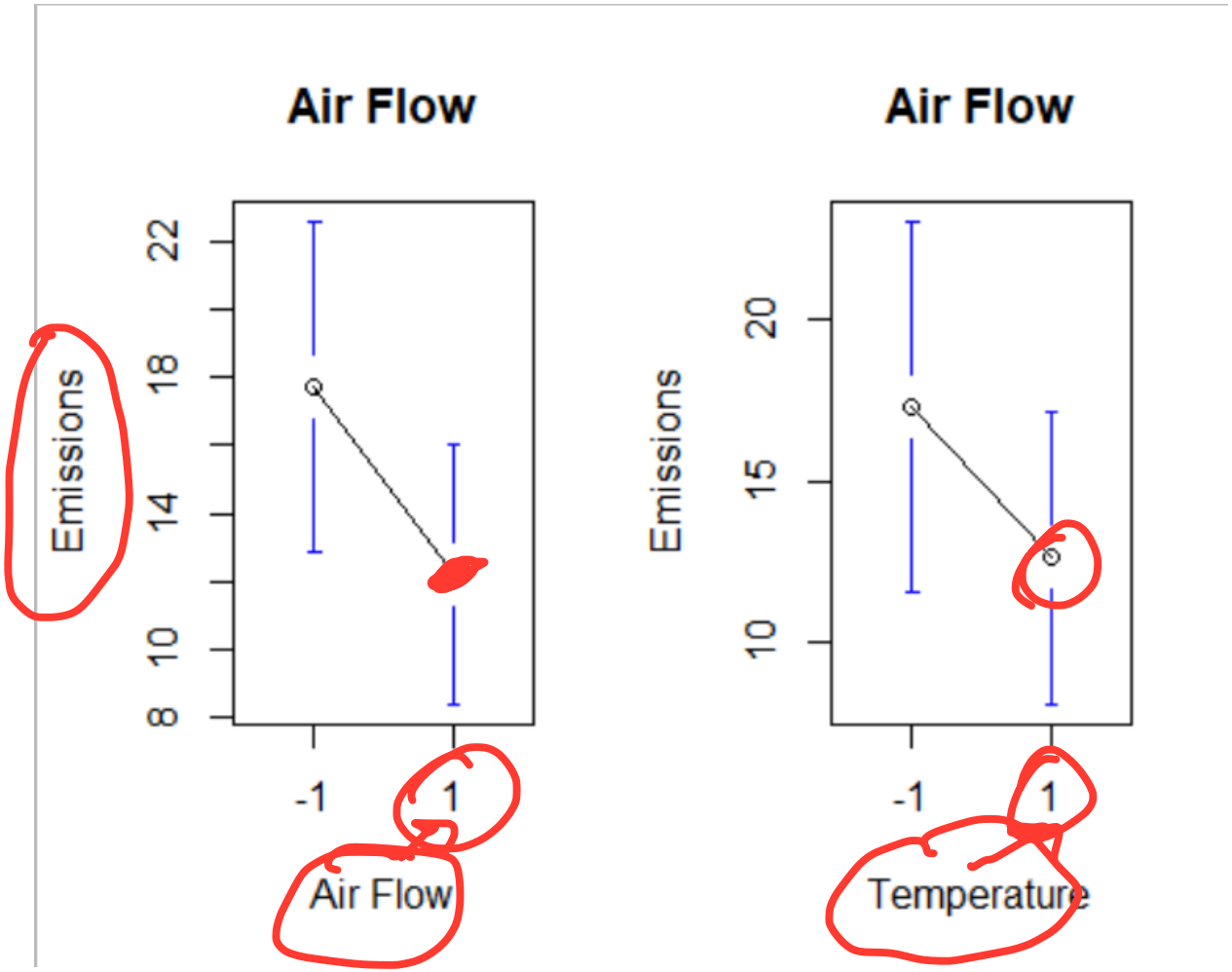
$$= 14.95 - 2.75\text{Acode} - 2.325\text{Bcode} + 0.325\text{Acode:Bcode}.$$

Typically the fitted regression model is found through ordinary least squares.

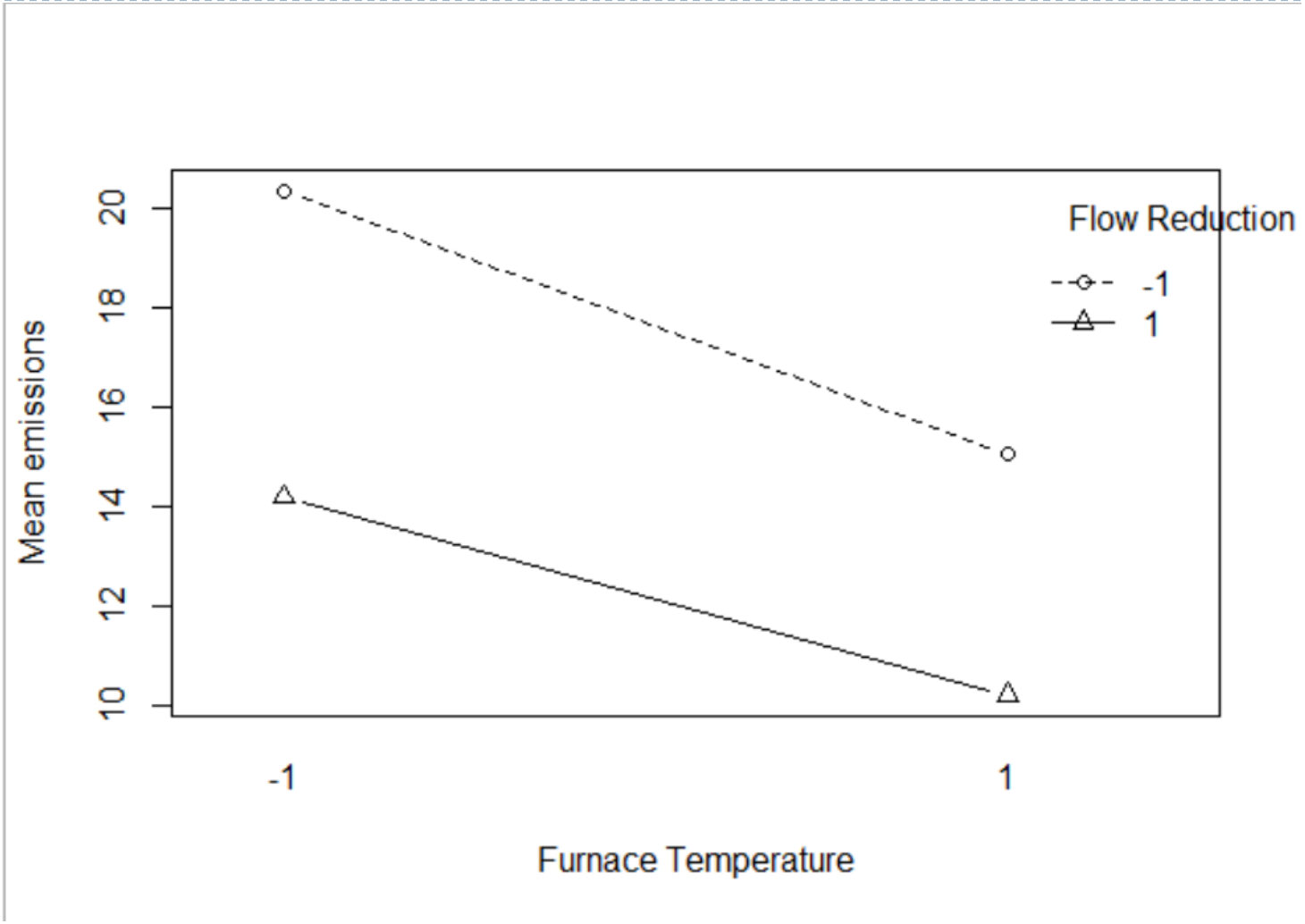
The 'least squares' estimates of the slopes and intercept are what is provided in the code: summary(m1).

This equation is the *response surface* equation

Plotting the Main Effects

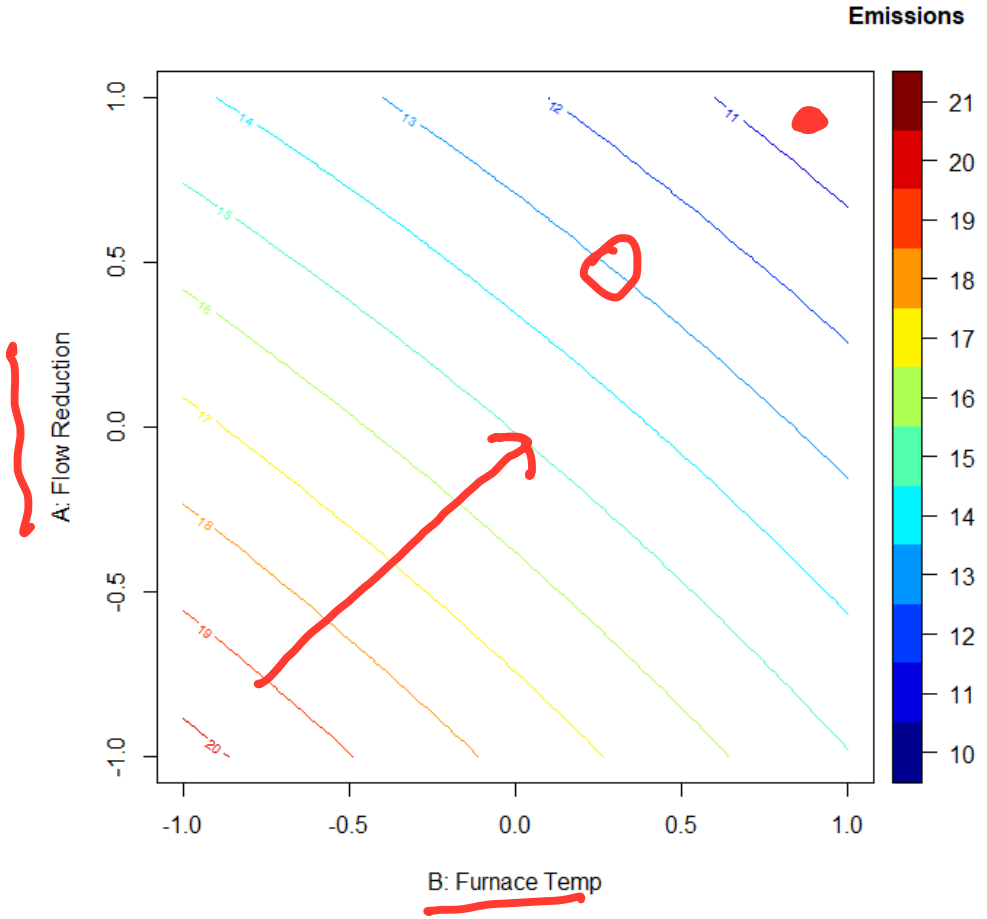


Plotting the Interaction



Contour plot

- Recall goal: Find combination of furnace temp and air flow which yields lowest CO^2 emission

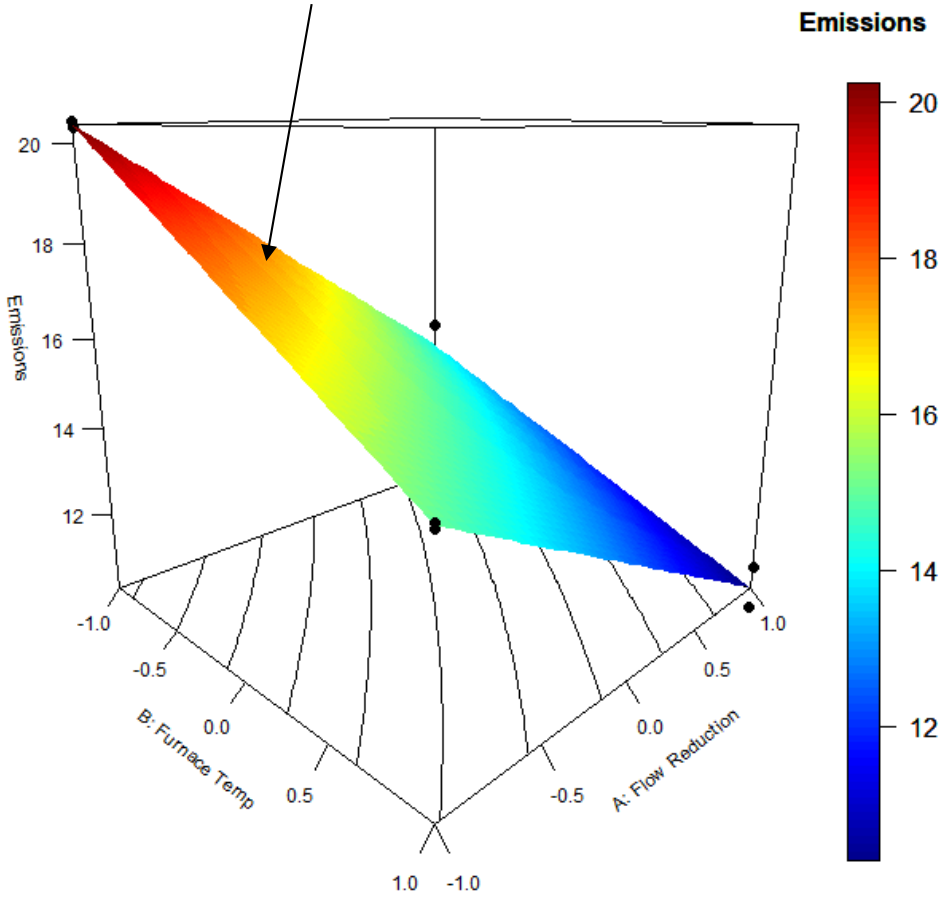


$\hat{y} = 14.95 - 2.75A - 2.725B + 1.325AB$

provides contours

named mcode_coal in our R code Line 19

$$\hat{y} = 14.95 - 2.75x1 - 2.325x2 + 0.325x1x2$$



2^3 Factorial Designs

- Let's consider the following 2^3 design

- Effects of Polysorbate 80 (X_1), Propylene glycol (X_2) and Invert sucrose concentration (X_3) on the turbidity of an oral solution.

  - Polysorbate 80 (X_1) : 3.7% to 4.3%
  - Propylene glycol (X_2): 17% to 23%
  - Invert sucrose conc. (X_3): 49 mL to 61 mL

Want to find model such that $turbidity = f(polysorb, propylene, sucrose)$

2^k Factorial Designs

- Screening Designs for Small Number of Variables
- All **Main** effects and **Interactions** are given 'single degree of freedom'
- $2^k = 1 + k + k(k-1)/2 + k(k-1)(k-2)/3 + \ldots + 1$
- Example: $2^4 = 1 + 4 + 6 + 4 + 1$
- All linear and linear x linear interactions can be estimated/tested

Coded Factor Levels

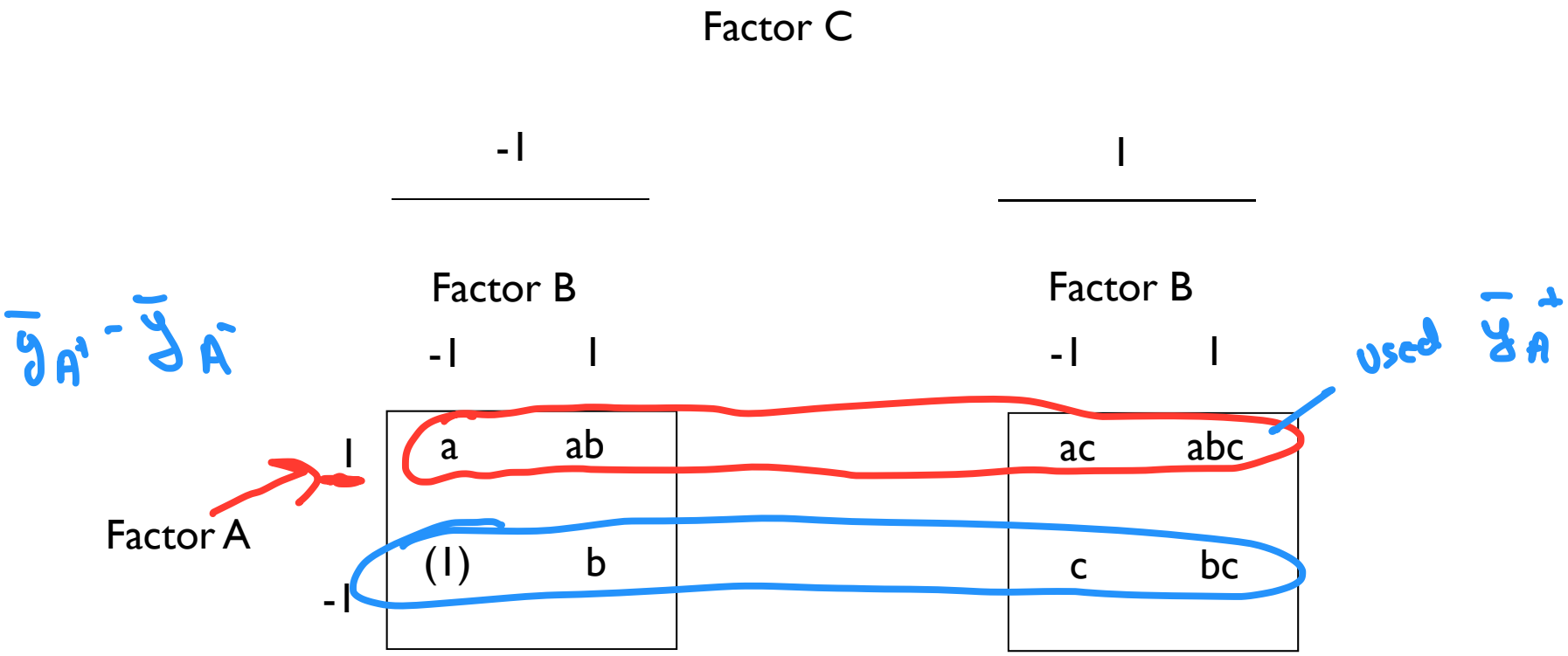
- Here is the factor coding for the 2^3 Design:

$$\text{coded A} = \frac{polysorb - 4.0}{0.3}, \quad \text{coded B} = \frac{propylene - 20}{3}, \quad \text{coded C} = \frac{sucrose - 55}{6}$$

	coded levels	
	-1	1
A	3.7	4.3
B	17	23
C	49	61

Natural Levels

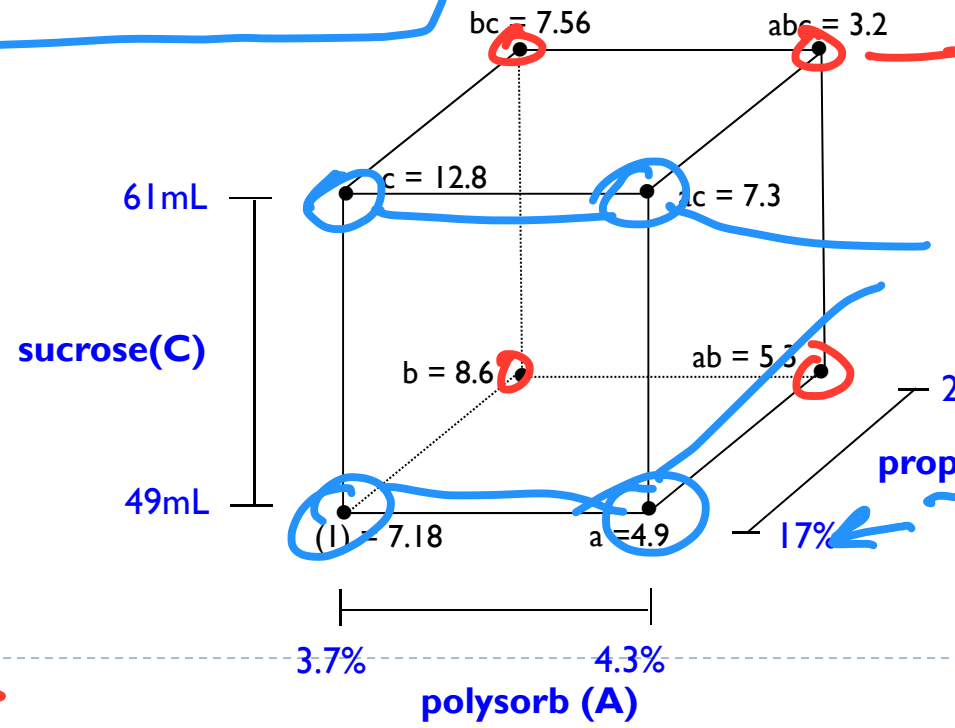
2^3 Factorial Experiment
3 *Factors*, Each at 2 *Levels*



A (polysorb)	B (propylene)	C (sucrose)	y (turbidity (ppm)) Rep.1	y (turbidity (ppm)) Rep. 2	Factor Comb.
-1	-1	-1	3.1	4.08	(1)
1	-1	-1	2.8	2.1	a
-1	1	-1	3.9	4.7	b
1	1	-1	3.1	2.2	ab
-1	-1	1	6.0	6.8	c
1	-1	1	3.4	3.9	ac
-1	1	1	3.5	4.06	bc
1	1	1	1.8	1.4	abc

Design protocol

Design protocol for how to formulate recipe of rough syrup



$\bar{y}_{B^+}$

$\bar{y}_{B^-}$

effect = $\bar{y}_{B^+} - \bar{y}_{B^-}$

$SS = \frac{(\text{effect})^2}{2^3 * n}$

Factor Combination	Factorial Effect I	A	B	AB	C	AC	BC	ABC
(1)	+	-	-	+	-	+	+	-
a	+	+	-	-	-	+	-	+
b	+	-	+	-	-	-	+	+
ab	+	+	+	+	-	-	-	-
c	+	-	-	+	+	+	+	+
ac	+	+	-	-	+	+	-	-
bc	+	-	+	-	+	-	+	-
abc	+	+	+	+	+	+	+	+

Contrast for Effect A = -(1) + a – b + ab – c + ac – bc + abc $= \bar{y}_{A+} - \bar{y}_{A-}$

Contrast for Effect B = -(1) – a + b + ab - c – ac + bc + abc $= \bar{y}_{B+} - \bar{y}_{B-}$

etc.

contrast

(a + ab + ac + abc) - ((1) + b + c + bc)

Main Effect A = $\frac{\text{contrast}}{4n}$ for 2^3 design

"2" was # design points where A low and where A high

$= \frac{\text{contrast}}{2n}$

compare to p.68

The full model fit to the data:

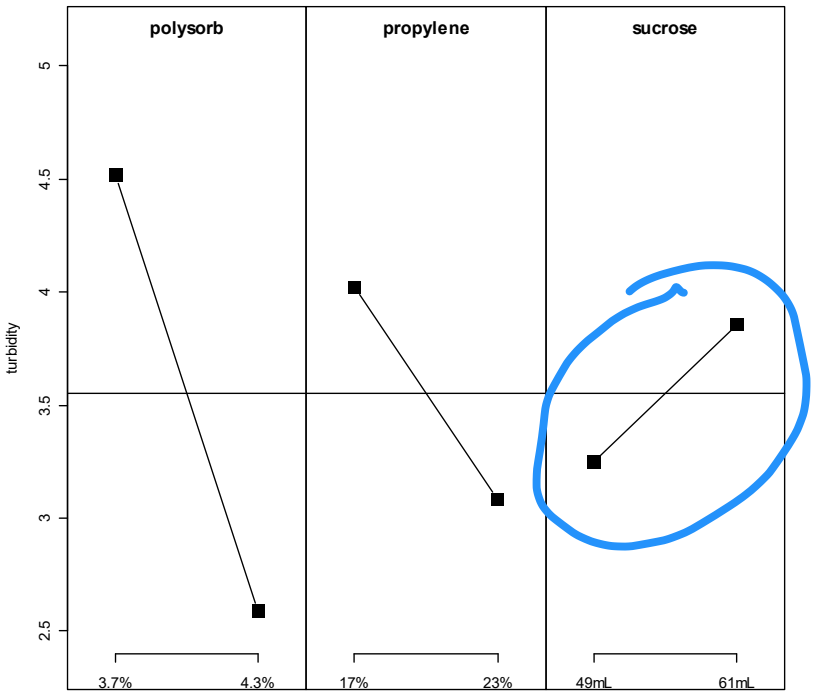
$$turbidity = \beta_0 + \beta_1 polysorb + \beta_2 propylene + \beta_3 sucrose + \beta_{12} polysorb * propylene + \beta_{13} polysorb * sucrose + \beta_{23} propylene * sucrose + \beta_{123} polysorb * propylene * sucrose + \varepsilon$$

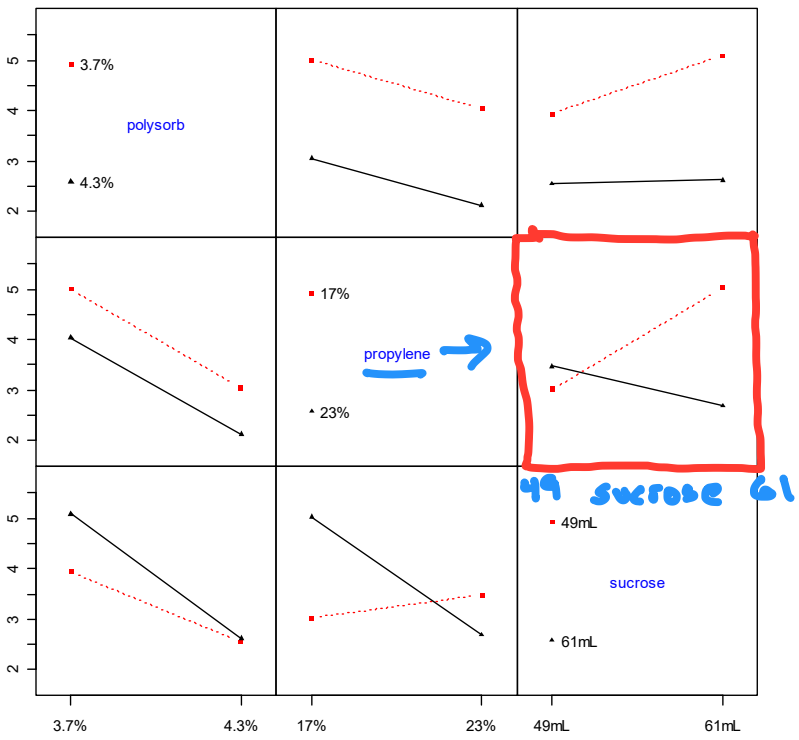
The intercept, β_0, is the mean turbidity when all factors are at their center locations.

β_1 represents mean change in turbidity as polysorb% increases from 3.7% to 4.0% (assuming propylene% and sucrose are fixed)

β_2 represents mean change in turbidity as propylene% increases from 17% to 20% (assuming polysorb% and sucrose are fixed)

β_3 represents mean change in turbidity as sucrose increases from 49mL to 55mL (assuming polysorb% and propylene% are fixed)


Main effects plot for turbidity
polysorb
propylene
sucrose
turbidity
5
4.5
4
3.5
3
2.5
3.7%
4.3%
17%
23%
49mL
61mL


Interaction plot matrix for turbidity
3.7%
polysorb
4.3%
17%
propylene
23%
49mL
sucrose
61mL
3.7%
4.3%
17%
23%
49mL
61mL

```
> anova(m1)
Analysis of Variance Table

Response: turbidity
                           Df  Sum Sq Mean Sq F value    Pr(>F)
polysorb                    1 14.8996 14.8996 55.9084 7.079e-05 ***
propylene                   1  3.5344  3.5344 13.2623 0.0065724 **
sucrose                     1  1.4884  1.4884  5.5850 0.0457251 *
polysorb:propylene          1  0.0009  0.0009  0.0034 0.9550840
polysorb:sucrose            1  1.1449  1.1449  4.2961 0.0719273 .
propylene:sucrose           1  7.7841  7.7841 29.2086 0.0006426 ***
polysorb:propylene:sucrose  1  0.2916  0.2916  1.0942 0.3261193
Residuals                   8  2.1320  0.2665
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Why 8 df for 'Residuals'? The content of the residual SS is 'pure error'

Next step is to fit a reduced model with only the important effects…remember to be somewhat liberal in including effects…why?

polysorb + propylene + sucrose + poly:prop + poly:sucrose + prop:sucrose

```
m2 <- lm(turbidity ~ (polysorb + propylene + sucrose)^2 - polysorb:propylene,
  data=sucrose)
summary(m2)
anova(m2)
```

m2 ← lm(turbidity ~ poly + prop + sucr + prop:sucr + poly:suc)

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.5525     0.1231  28.859 5.81e-11 ***
polysorb1            -0.9650     0.1231  -7.839 1.41e-05 ***
propylene1           -0.4700     0.1231  -3.818 0.003384 **
sucrose1              0.3050     0.1231   2.478 0.032671 *
polysorb1:sucrose1   -0.2675     0.1231  -2.173 0.054885 .
propylene1:sucrose1  -0.6975     0.1231  -5.666 0.000208 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4924 on 10 degrees of freedom
Multiple R-squared: 0.9225, Adjusted R-squared: 0.8837
F-statistic:  23.8 on 5 and 10 DF,  p-value: 2.971e-05


Analysis of Variance Table

Response: turbidity
                  Df  Sum Sq Mean Sq F value    Pr(>F)
polysorb           1 14.8996 14.8996 61.4543 1.406e-05 ***
propylene          1  3.5344  3.5344 14.5779 0.0033841 **
sucrose            1  1.4884  1.4884  6.1390 0.0326712 *
polysorb:sucrose   1  1.1449  1.1449  4.7222 0.0548850 .
propylene:sucrose  1  7.7841  7.7841 32.1060 0.0002078 ***
Residuals         10  2.4245  0.2424
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

fit = 3.55 − .965 poly − .47 prop + .305 sucr − .2675 poly:suc − .6975 prop:sucr

evaluated at poly = 1, prop = 1, suc = −1

Regression Model Building and Prediction of Response

From the R output we were given the following estimates of the coefficients for the *coded factors*:

Intercept	Poly.	Prop.	Sucrose	Poly.*Sucr.	Prop.*Sucr.
3.55	-0.96	-0.47	0.30	-0.27	-0.70

The fitted regression model is then:

$$\hat{y} = 3.55 - 0.96Poly. - 0.47Prop. + 0.3Sucr. - 0.27Poly.*Sucr. - 0.70Prop.*Sucr.$$

Regression function is useful in obtaining prediction at a certain factor combination. For example, the predicted turbidity with 4.3% polysorbate, 17% propylene glycol and 49 mL of sucrose (i.e. Polysorb = +1, Propylene = -1, and Sucrose = -1) is:

$$\begin{aligned}\hat{y} &= 3.55 - 0.96*(+1) - 0.47*(-1) + 0.3*(-1) \\ &\quad - 0.27*(+1)*(-1) - 0.70*(-1)*(-1) \\ &= 2.33 \text{ ppm}\end{aligned}$$

The fitted regression function can be used for **prediction** of response values (turbidity values) or estimation of mean response. To check the quality (or lack of quality) of predicton is important. Analysis of **residuals** is part of the check.

After checking for model diagnostics, you can then use the estimated mean function and optimize this function over the design space.