In [1]:

```
library(repr)

# Change default plot size
options(repr.plot.width=18, repr.plot.height=8)
```

Statistical Analysis Assignment - Christian Gilson - January 20th 2020

Q1) Two-stage nested (hierarchical) design

Model equation:

$z_{lmn} = \mu + \delta_l + \gamma_{(l)m} + \epsilon_{(lm)n}$.

Sum to zero constraints: $\sum_{l=1}^{a} \delta_l = 0$.

a)

Conditions that characterise two-stage nested design data collection:

- The z_{lmn} responses can be classified according to levels of factors L and M, where M is a nested factor *within* L. You can see a tabular representation of how this nested design would look in the table below.
- The levels of factor M ($m = 1, ..., b$), for a given level of factor L ($l = 1, ..., a$) are not identical to the corresponding levels of factor M at another level of factor L, *but they are similar in nature*.

1			l			a		
1	m	b	1	m	b	1	m	b
z_{111}	z_{1m1}	z_{1b1}	z_{l11}	z_{lm1}	z_{lb1}	z_{a11}	z_{am1}	z_{ab1}
z_{11n}	z_{1mn}	z_{1bn}	z_{l1n}	z_{lmn}	z_{lbn}	z_{a1n}	z_{amn}	z_{abn}
z_{11r}	z_{1mr}	z_{1br}	z_{l1r}	z_{lmr}	z_{lbr}	z_{a1r}	z_{amr}	z_{abr}

Right hand side terms of model equation:

- μ is the overall mean, which is estimated by the grand average $\hat{\mu} = \bar{z}$.
- δ_l is the effect of the l-th level of factor L (a constant that is estimated by taking the mean of all responses in level l and correcting by the grand mean).
- $\gamma_{(l)m}$ is the effect of the m-th level of factor M nested within the l-th level of factor L. In our case, this is a random term where $\gamma_{(l)m} \sim \text{NID}(0, \sigma^2_M)$).
- $\epsilon_{(lm)n}$ is an error term, where each error term is assumed to be i.i.d. and following a normal with mean 0 and variance σ^2.

Scenario where this experimental design may be appropriate:

- Measuring the quality of teaching of schools and the teachers within those schools, where the responses would be external evaluations (e.g. Ofsted) of lessons taught by those teachers. There would be r lessons evaluated per teacher.
- Factor L in this scenario would be schools, where you would investigate a fixed number of a schools.
  - δ_l would therefore be a measure of the effect of the quality of teaching at the school level.
  - Large positive δ_l terms would indicate school l has a higher quality of teaching with respect to the other $a - 1$ schools.
- Factor M in this case would be teachers: b randomly selected teachers from each of the a schools would be evaluated.
- z_{lmn} would therefore be the evaluation response for the n-th lesson evaluated for teacher m at school l.

b)

$F = \text{SS}_R$ (summing up the square of the residuals between each response and the average of the r responses for each (l,m) nested layer)

$G = \text{SS}_{M(L)}$ (summing up the square averages of r values for each (l,m) nested layer of factor M nested within factor L, corrected for the average of the l layer of factor L)

$H = \text{SS}_L$ (summing up square averages of each l layer of factor L, corrected for the grand average)

$K = \text{SS}_T$ (summing the squares of each response corrected for the grand average)

Therefore we can write: $\text{SS}_T = \text{SS}_L + \text{SS}_{M(L)} + \text{SS}_R$ as:

$K = H + G + F$

c) Conditions: fixed L (sum to zero on δ_l), random M ($\gamma_{(l)m} \sim \text{NID}(0, \sigma^2_M)$), $\epsilon_{(lm)n}$ error terms independent of the sum to zero constraint and the distribution of the $\gamma_{(l)m}$'s.

i) Since $H = \text{SS}_L$, this is $\mathbb{E}[\text{SS}_L \,/\, (a-1)] = \mathbb{E}[\text{MS}_L]$

$$\mathbb{E}[\text{MS}_L] = \sigma^2 + r\sigma^2_M + \frac{br}{a-1}\sum_{l=1}^{a} \delta_l^2$$

ii) Since $G = \text{SS}_{M(L)}$, this is $\mathbb{E}[\text{SS}_{M(L)} \,/\, a(b-1)] = \mathbb{E}[\text{MS}_{M(L)}]$

$$\mathbb{E}[\text{MS}_{M(L)}] = \sigma^2 + r\sigma^2_M$$

iii) Since $F = \text{SS}_R$, this is $\mathbb{E}[\text{SS}_R \,/\, ab(r-1)] = \mathbb{E}[\text{MS}_R]$

$$\mathbb{E}[\text{MS}_R] = \sigma^2$$

d) Hypothesis testing

$H_{0L} : \delta_1 = \delta_2 = \ldots = \delta_a = 0$

$H_{0M} : \sigma^2_M = 0$

How do the expectations in **c)** change:

i) If H_{0L} is true, then $\frac{br}{a-1} \sum_{l=1}^{a} \delta_l^2 = 0$ and therefore:

$$\mathbb{E}[\mathrm{MS}_L] = \sigma^2 + r\sigma^2_M;$$

$$\mathbb{E}[\mathrm{MS}_{M(L)}] = \sigma^2 + r\sigma^2_M;$$

$$\mathbb{E}[\mathrm{MS}_R] = \sigma^2,$$

and hence $\mathbb{E}[\mathrm{MS}_L] = \mathbb{E}[\mathrm{MS}_{M(L)}]$.

ii) If H_{0M} is true, then $r\sigma^2_M = 0$ and therefore:

$$\mathbb{E}[\mathrm{MS}_L] = \sigma^2 + \frac{br}{a-1} \sum_{l=1}^{a} \delta_l^2;$$

$$\mathbb{E}[\mathrm{MS}_{M(L)}] = \sigma^2;$$

$$\mathbb{E}[\mathrm{MS}_R] = \sigma^2,$$

and hence $\mathbb{E}[\mathrm{MS}_{M(L)}] = \mathbb{E}[\mathrm{MS}_R]$.

iii) If both H_{0L} and H_{0M} are true, then $\frac{br}{a-1} \sum_{l=1}^{a} \delta_l^2 = 0$ and $r\sigma^2_M = 0$, and therefore:

$$\mathbb{E}[\mathrm{MS}_L] = \sigma^2;$$

$$\mathbb{E}[\mathrm{MS}_{M(L)}] = \sigma^2;$$

$$\mathbb{E}[\mathrm{MS}_R] = \sigma^2,$$

and hence $\mathbb{E}[\mathrm{MS}_L] = \mathbb{E}[\mathrm{MS}_{M(L)}] = \mathbb{E}[\mathrm{MS}_R]$.

e) Pharmaceutical company, 6 different labs, producing same medicine. Looking for sources of variation of Tetrasodium Pyrophosphate in the medicine.

- 6 laboratories, so a = 6;
- 3 *random* batches per lab, so b = 3;
- 2 repeated measurements per batch, so r = 2;
- 6x3x2 = 36 observations in total.
- Similar to above, we have laboratory factor is fixed, and batch factor is random.

Let's call factor L the laboratory factor, and factor M the medicine batch factor, to keep the notation consistent with the previous parts of question 1.

Experiment data would look like:

	Laboratory 1			**Laboratory l**			**Laboratory 6**		
Batch 1	**Batch 2**	**Batch 3**	**Batch 1**	**Batch 2**	**Batch 3**	**Batch 1**	**Batch 2**	**Batch 3**
z_{111}	z_{121}	z_{131}	z_{l11}	z_{l21}	z_{l31}	z_{611}	z_{621}	z_{631}
z_{112}	z_{122}	z_{132}	z_{l12}	z_{l22}	z_{l32}	z_{612}	z_{622}	z_{632}

Sources of variation:

The output of the F scores in R's ANOVA summary assume that both the laboratory factor and the batch factors are **fixed**. In this experiment, the batches of medicine are randomly selected, and so we have a fixed laboratory factor and a random batch factor.

We therefore calculate the F scores ourselves, using the MS values from the ANOVA table.

$F_L = MS_L \,/\, MS_{M(L)}$, (not $MS_L \,/\, MS_R$ as calculated in R), where under H_{0A}, $F_L \sim F_{5,12}$.

> $$F_L = 10.767 \,/\, 4.800 = 2.243$$

$F_M = MS_{M(L)} \,/\, MS_R$, (the same as calculated in R), where under H_{0M}, $F_M \sim F_{12,18}$.

> $$F_M = 4.800 \,/\, 1.922 = 2.497$$

As expected, we get precisely the same F_M as in the R analysis, but we get a much lower F_L value than in the R analysis, which had an associated p-value that was statistically significant at the 1% significance level (p-value = 0.00276).

To test the significance of whether the variation in the amount of Tetrasodium Pyrophosphate in the medicine is due to the difference in laboratory, we now calculate the p-value associated with $F_L = 2.243$.

In [2]:

```
# Getting p-value for F_L from F distribution with (5,12) degrees of freedom
pf(2.243, 5, 12, lower.tail = FALSE)
```

0.116817008463374

In [3]:

```
# F_L score required to meet a 5% significance level
qf(0.05, 5, 12, lower.tail = FALSE)
```

3.10587523908412

With a p-value of **0.12**, the effect of different laboratories is therefore not significant at either the 1% or even the 5% significance level. We would tell the pharmaceutical company that the evidence suggests variation in Tetrasodium Pyrophosphate in the medicine is not due to the experiments being ran in different laboratories.

There is evidence at the 5% confidence level that batches from each laboratory are reasonably variable.

This suggests that a significant cause in variation in Tetrasodium Pyrophosphate originates *between batches at a laboratory*, rather than between laboratories.

An estimate for the variance of the batches is:

$$\hat{\sigma}^2_M = \frac{MS_{M(L)} - MS_R}{r} = \frac{4.800 - 1.922}{3} = 0.96.$$

When discussing these results with the pharmaceutical company, I'd ask them to lay out the systematic process as to how the medicine is produced, especially the parts involving Tetrasodium Pyrophosphate.

I'd look to identify areas where human error / measurement error could be minimised, or where additional quality control processes could be implemented - such as quality assurance checks and sanity tests at various stages of the production process.

Any time a sanity check produced a result outside of a pre-determined threshold, the laboratory should be alerted and the underlying cause of the issue found and fixed (such as a piece of apparatus requiring re-calibration).

Q2) Completely randomised 2-factor factorial design

a) Conditions that characterise data collection. Explain how to plant the potatoes.

Conditions that characterise the 2-factorial design data collection:

- The responses are denoted $\{y_{ijk}\}$ where y_{ijk} represents the k-th replicate at the *i*-th level of **variety**, and the j-th level of **location**, where there are 3 replicates per (i, j) tuple, with k = 1, 2, 3, i = A, B, C, j = 1, 2, 3, 4.
- There must be at least **2** replications per (i, j) pair in a completely randomised two-factor factorial design. This criteria is satisfield as we have **3**.
- The responses are taken in a random order.

Method to fulfil randomisation criteria when planting

- There are 3 plots per potato variety per location, therefore there are 3 x 3 x 4 = 36 planting plot units in total. I would first randomly assign the 3 replicates for each of the 3 varieties to the 9 units per location, for each of the 4 locations. With the co-ordinates randomly assigned to each variety per location, I'd then plot in a random order by randomly assigning each plot with an integer between 1 to 36.

Example below:

In [4]:

```
# each plot is assigned a unique identifier
plot.id <- 1:36

# assigning locations to each unique plot.id
plot.location <- rep(rep(1:4), rep(9,4))

# bag of seeds per location
variety.seeds.per.location <- rep(c('A','B','C'), 3)

# randomly sampling bag of seeds per location
plot.seeds <- rep(sample(variety.seeds.per.location, 9, replace = FALSE), 4)

# randomising the plotting order
plot.order <- sample.int(36, 36, replace = FALSE)

df.plot <- data.frame(plot.id, plot.location, plot.seeds, plot.order)

# dataframe ordered by the plotting order
df.plot[order(plot.order),]
```

A data.frame: 36 × 4

	plot.id	plot.location	plot.seeds	plot.order
	<int>	<int>	<fct>	<int>
36	36	4	C	1
32	32	4	B	2
18	18	2	C	3
35	35	4	A	4
24	24	3	A	5
9	9	1	C	6
5	5	1	B	7
14	14	2	B	8
26	26	3	A	9
21	21	3	A	10
10	10	2	B	11
28	28	4	B	12
23	23	3	B	13
6	6	1	A	14
3	3	1	A	15
29	29	4	C	16
15	15	2	A	17
25	25	3	C	18
12	12	2	A	19
31	31	4	B	20
20	20	3	C	21
4	4	1	B	22
33	33	4	A	23
17	17	2	A	24
30	30	4	A	25
8	8	1	A	26
27	27	3	C	27
19	19	3	B	28
16	16	2	C	29
2	2	1	C	30
13	13	2	B	31
1	1	1	B	32
7	7	1	C	33
22	22	3	B	34
11	11	2	C	35

	plot.id	plot.location	plot.seeds	plot.order
	<int>	<int>	<fct>	<int>
34	34	4	C	36

b) Fill in missing R code:

```
* <- rep(rep(1:4), rep(9,4))
** <- rep(rep(c('A','B','C'), rep(3,3)), 4)
```

In [5]:

```
yield <- c(15, 19, 12, 20, 24, 18, 22, 17, 14, 17, 10, 13, 24, 18, 22,
           +26, 19, 21, 9, 12, 6, 12, 15, 10, 10, 5, 8, 14, 8, 11, 21, 16, 14,
           +19, 15, 12)
```

In [6]:

```
# want each location from 1,2,3,4 to be repeated 9 times
loc <- factor(rep(rep(1:4), rep(9,4)))

# want A,B,C to be repeated 3 times (once per variety)
#, where each character is repeated 3 times (once per repeated observation per v
ariety),
# , then the whole sequence repeated 4 times (once per location)
variety <- factor(rep(rep(c('A','B','C'), rep(3,3)), 4))

potato <- data.frame(yield, loc, variety)

# selecting top 10
potato[1:10,]
```

A data.frame: 10 × 3

yield	loc	variety
<dbl>	<fct>	<fct>
15	1	A
19	1	A
12	1	A
20	1	B
24	1	B
18	1	B
22	1	C
17	1	C
14	1	C
17	2	A

c) Describe the form of the statistical model, corresponding least squares estimates for the parameters (under sum to zero constraints) and structure of ANOVA table if:

i) an interaction term is included:

The linear model will have the following form:

$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}$

where $i = A, B, C$, $j = 1, 2, 3, 4$, and $k = 1, 2, 3,$

and μ is the overall mean, τ_i is the effect of the i-th variety of potato, β_j is the effect of the j-th location, and $(\tau\beta)_{ij}$ is the interaction term between the i-th variety and j-th location, and ϵ_{ijk} is a random error term.

When analysing the data including the interaction term, one would pay particular attention to the $(\tau\beta)_{ij}$ term, where the parameter estimates are defined as:

$\hat{(\tau\beta)}_{ij} = \frac{y_{ij.}}{3} - \hat{\tau}_i - \hat{\beta}_j - \hat{\mu},$

where the interaction effect is the mean of the (i, j) groups corrected for the effects for variety and location, and also the grand average. A large interaction effect would indicate a synergy between a specific potato variety and a specific location. (Definitions for $\hat{\tau}_i$, $\hat{\beta}_j$, and $\hat{\mu}$ are found in **b)**, below).

In the ANOVA table, one would pay specific attention to the F value of the interaction term and the

ii) *no* interaction term is included:

With no interaction term, we lose the $(\tau\beta)_{ij}$ term in the model, which takes the following form:

$y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ijk}$,

with i, j, k having the same values as in **i)**.

With no interaction term included in the model, one would pay particular attention to the τ_i and β_j effect terms in the model.

Under the *sum to zero* constraints (four constraints where the τ_i's sum to zero over all i, the β_j's sum to zero over all j, and the $(\tau\beta)_{ij}$'s sum to zero both over all i for each j, and over all j for each i), the parameter estimates are:

$\hat{\mu} = \frac{y_{...}}{abn} = \bar{y}_{...}$ = grand average = $\frac{548}{36} = 15.22$

$\hat{\tau_i} = \frac{y_{i..}}{bn} - \hat{\mu} = \bar{y}_{i..} - \bar{y}_{...}$, for $i = A, B, C$, where $b = 4$ (total number of locations) and $n = 3$ (number of replicates). These effect terms are the means of the variety groups (A, B, and C) corrected for the grand average.

$$\hat{\tau_A} = \frac{146}{12} - 15.222 = -3.06;$$

$$\hat{\tau_B} = \frac{214}{12} - 15.222 = 2.61;$$

$$\hat{\tau_C} = \frac{188}{12} - 15.222 = 0.44.$$

The $\hat{\beta_j}$ terms are defined as:

$\hat{\beta_j} = \frac{y_{.j.}}{an} - \hat{\mu} = \bar{y}_{.j.} - \bar{y}_{...}$, for $j = 1, 2, 3, 4$ and $a = 3$ (the total number of potato varieties). These effect terms are the means of the location groups (1, 2, 3, and 4) corrected for the grand average.

$$\hat{\beta_1} = \frac{161}{9} - 15.222 = 2.67;$$

$$\hat{\beta_2} = \frac{170}{9} - 15.222 = 3.67;$$

$$\hat{\beta_3} = \frac{87}{9} - 15.222 = -5.56;$$

$$\hat{\beta_4} = \frac{130}{9} - 15.222 = -0.78.$$

Large effect terms positive (such as $\hat{\tau_B}$ and $\hat{\beta_2}$) and negative (such as $\hat{\tau_A}$ and $\hat{\beta_3}$) indicate specific potato varieties and plot locations that yield above or below the grand average yield, respectively.

In the ANOVA table, one would pay particular attention to the p-values associated with F_{variety} and F_{location}. If these F values are statistically significant, then we can treat the effects of potato variety or location as significant factors that impact bushel yield.

d) ANOVA with and without interaction terms

With interaction terms:

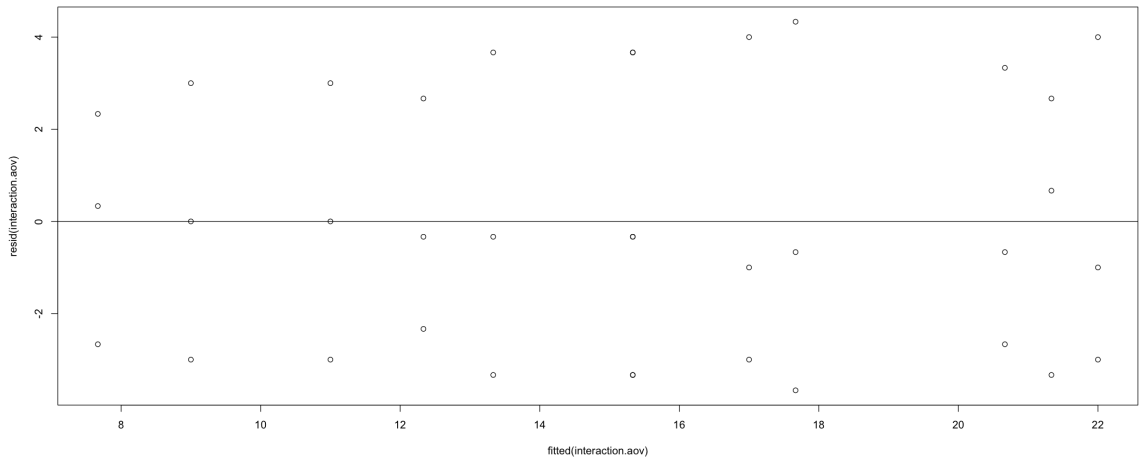
In [7]:

```
interaction.aov <- aov(yield ~ loc + variety + loc:variety, potato)
summary(interaction.aov)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
loc          3  468.2  156.07  14.556 1.31e-05 ***
variety      2  196.2   98.11   9.150  0.00111 **
loc:variety  6   78.4   13.07   1.219  0.33090
Residuals   24  257.3   10.72
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In [8]:

```
plot(fitted(interaction.aov), resid(interaction.aov))
abline(h=0)
```



Without interaction terms:

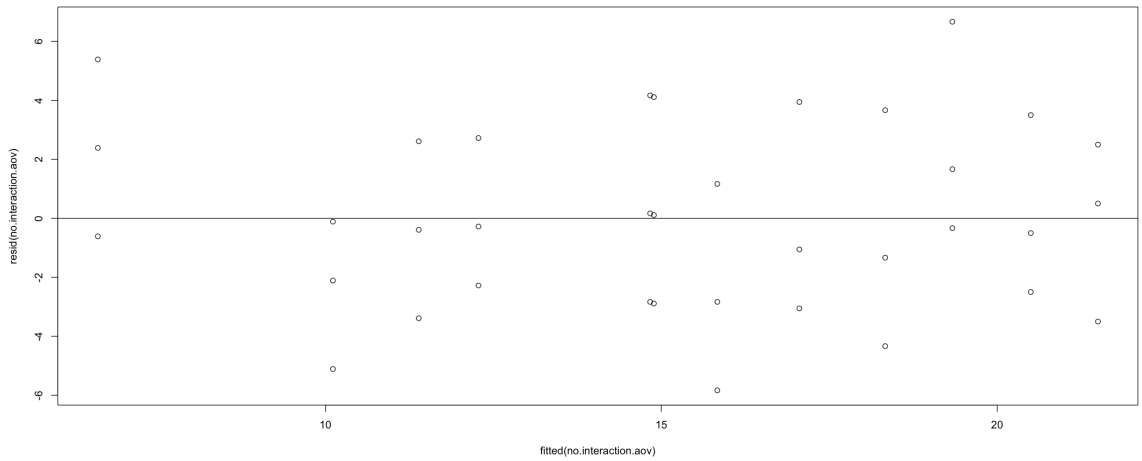
In [9]:

```
no.interaction.aov <- aov(yield ~ loc + variety, potato)
summary(no.interaction.aov)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
loc          3  468.2  156.07  13.944 7.16e-06 ***
variety      2  196.2   98.11   8.766    0.001 **
Residuals   30  335.8   11.19
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In [10]:

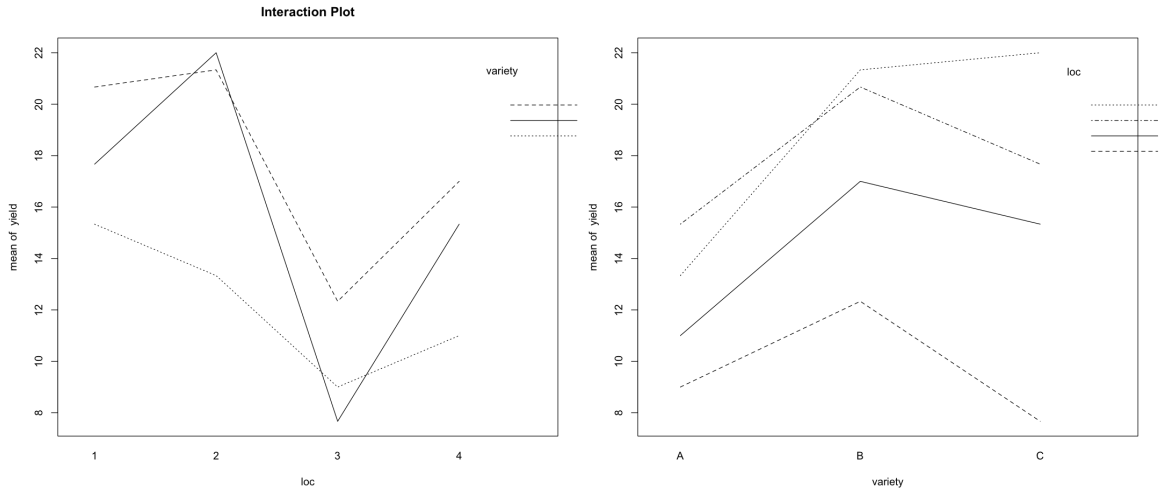
```
plot(fitted(no.interaction.aov), resid(no.interaction.aov))
abline(h=0)
```



In [11]:

```
par(mfrow = c(1, 2))

# not quite sure why the labels aren't showing properly...
interaction.plot(loc, variety, yield, main = "Interaction Plot")
interaction.plot(variety, loc, yield)
```



(i) Which model do we endorse?

By first looking at the ANOVA summary table for the *interaction* model, we see that F values for both variety and location, F_{variety} and F_{location} respectively, are significant at the 1% level. The p-value attributed to the F score for the interaction term however is 0.33, and therefore is not statistically significant.

As a result of the interaction term not being deemed statistically significant, we remove the interaction effect from the model, which leaves the *no-interaction* model. We can also visually inspect the interactions in *interaction.plot* above. There doesn't appear to be a clear effect of one of the variables having an impact on the other.

For the *no-interaction* model, F_{location} is significant at the 0.1% level and F_{variety} is significant at the 1% level (having a p-value exactly on the 0.1% border).

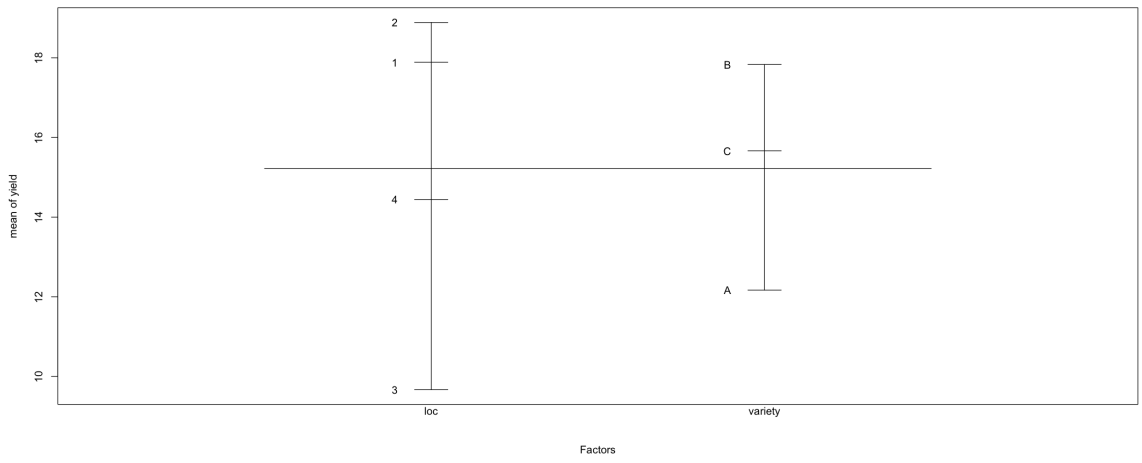
Looking at the plot of residuals against fitted values for the no-interaction model, there appears to be no trend or systematic pattern in the residuals, and the points appear to be scattered evenly around zero.

As both variety and location factors are statistically significant at the 1% level, and a visual check of model adequacy of the residuals shows no discernable systematic pattern, with the residuals randomly scattered around zero, we endorse the no-interaction model.

(ii) Recommendation to farmer to maximise crop yield

In [12]:

```
plot.design(yield ~ loc + variety + loc:variety, potato)
```



After endorsing the conclusions that potato variety and location are both statistically significant, we inspect the β_j and τ_i values calculated in **c) ii)** and also presented in *model.tables* below, we can see from the β_j's that location 2 appears to be the best, followed by location 1. Location 3 is clearly the worst, with an average of 9.67 bushels compared to location 1 with 18.89.

By inspecting the τ_i's, potato variety B appears to be the best with respect to maximising yield, with variety A being the worst.

These observations can also clearly be seen in *plot.design* above.

If we look at the interaction averages at each (i, j) plot in *model.tables* below, we see that it's actually variety C, not B, that produces the highest yield at location 2. Due to variety B being better in all other locations, and the interaction term not being statistically significant, I would recommend to the farmer to plot variety B to maximise crop yield, where the farmer will see best results at location 1 and 2.

I'd definitely be interested to see more historical crop yield data if possible, and to hear if there were any systematic differences in preparation for locations 1 and 2 versus 3 given how drastic the yields are for that particular season.

In [13]:

```
model.tables(interaction.aov)
```

```
Tables of effects

 loc
loc
     1      2      3      4
 2.667  3.667 -5.556 -0.778

 variety
variety
      A       B       C
-3.0556  2.6111  0.4444

 loc:variety
   variety
loc A       B       C
  1  0.5000  0.1667 -0.6667
  2 -2.5000 -0.1667  2.6667
  3  2.3889  0.0556 -2.4444
  4 -0.3889 -0.0556  0.4444
```

In [14]:

```
model.tables(interaction.aov, type='means')
```

```
Tables of means
Grand mean

15.22222

 loc
loc
     1      2      3      4
17.889 18.889  9.667 14.444

 variety
variety
     A      B      C
12.167 17.833 15.667

 loc:variety
   variety
loc A      B      C
  1 15.333 20.667 17.667
  2 13.333 21.333 22.000
  3  9.000 12.333  7.667
  4 11.000 17.000 15.333
```

-- End of Assignment --