Chapter4ANOVAII

January 23, 2023

1 Chapter 4 Analysis of Variance

1.1 4.2 Two-way ANOVA

1.1.1 4.2.1 Motivating example and model

Consider the Hey Fever dataset. Nine compounds for Hay Fever Relief are made by varying levels of two basic ingredients. Ingredient 1 (denoted factor A) has $a = 3$ levels: low ($i = 1$), medium ($i = 2$) and high ($i = 3$). And Ingredient 2 (factor B) has $b = 3$ levels: low ($j = 1$), medium ($j = 2$) and high ($j = 3$). A total of 36 subjects suffering from hay fever are selected and each of the 9 compounds are given to randomly selected $n = 4$ individuals.

The cell means model for a *balanced* two-way ANOVA takes the following form

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk},\ k = 1, \ldots, n, j = 1, \ldots, b, i = 1, \ldots, a,$$

where $\{\epsilon_{ijk}\}$ are i.i.d. $N(0, \sigma^2)$. There are ab unknown means and one unknown σ in the cell means model.

In practice, we often prefer the fhe factor effect form

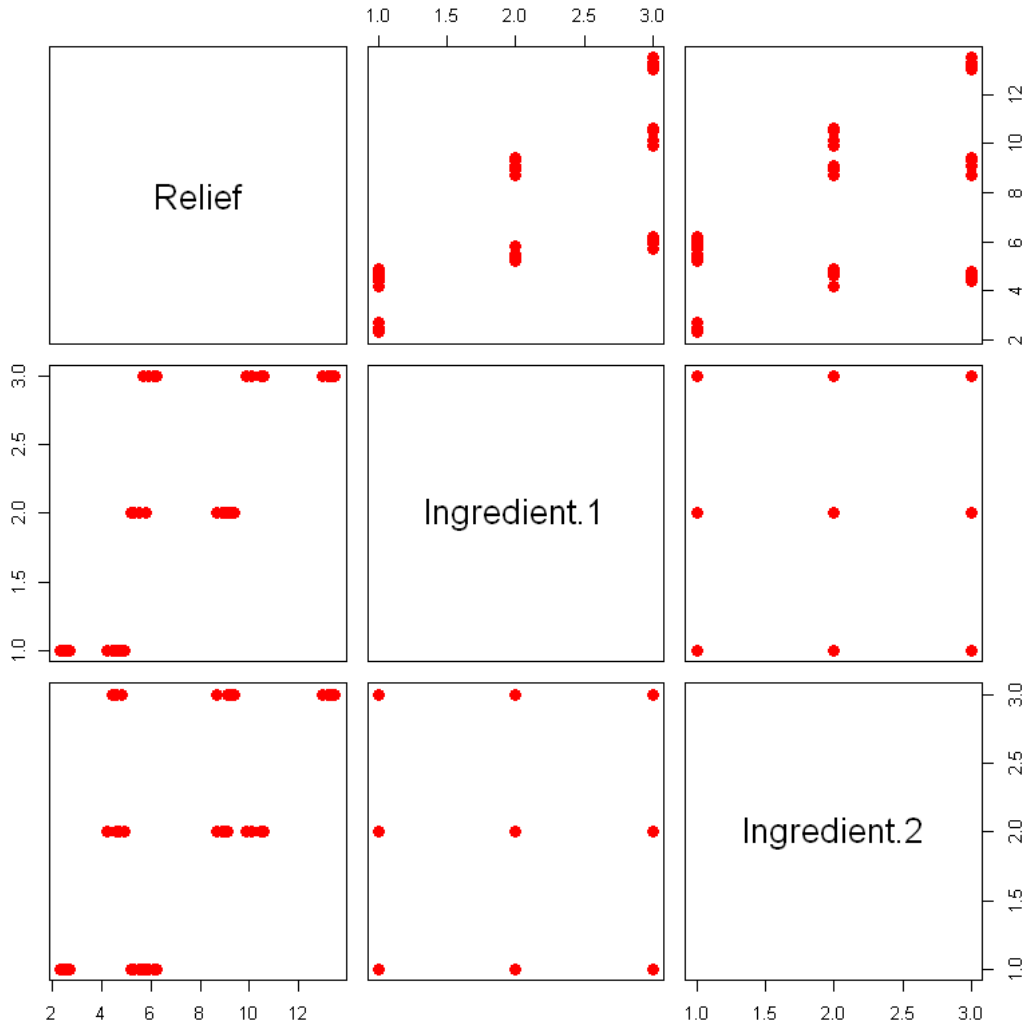
$$Y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},\ k = 1, \ldots, n, j = 1, \ldots, b, i = 1, \ldots, a,$$

where $\{\epsilon_{ijk}\}$ are i.i.d. $N(0, \sigma^2)$. A quick count shows that there are a αs, b βs, ab interaction terms, one μ and one σ in the factor effect form, which amount to $ab + a + b + 2$ parameters. Indeed, the parameters in the factor effect form are no longer "free" as in the cell means model. A set of constraints are in placed is in place with the factor effect form, which will be discussed in Section 4.2.2.

Remark: Note that $(\alpha\beta)_{ij}$ is a new parameter standing for the interaction effect, but **not** the product of α_i and β_j. We can replace it with γ_{ij} to avoid confusion. However, the advantage of using $(\alpha\beta)_{ij}$ is self-explanatory; that is the expression $(\alpha\beta)_{ij}$ shows that it is an effect jointly controlled by factor A and factor B.

[1]:
```
Hay <- read.csv(file="../Data/HayFever.csv", header=TRUE, sep=",")

# Use a slightly different visualization:
pairs(Hay,pch=16,col='red',cex=1.5)
```



1.1.2 4.2.2 Constraints on parameters

In the two-way model, the mapping between the cell means form and the factor effects form is slightly more complicated that that in the one-way case. We consider the balanced design here for simplicity.

The cell means $\{\mu_{ij} : i = 1, \ldots, a, j = 1, \ldots, b\}$ are naturally defined as the population mean (or expectation) within each cell determined by one unique combination of two factors. We proceed to define the factor effects using the cell means.

We first look at the over mean $\mu_{\cdot\cdot}$

$$\mu_{\cdot\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}/(ab),\ \mu_{i\cdot} = \sum_{j=1}^{b}\mu_{ij}/b,\ \mu_{\cdot j} = \sum_{i=1}^{a}\mu_{ij}/a.$$

We can then define the factor effects as

$$\alpha_i = \mu_{i\cdot} - \mu_{\cdot\cdot},\ \beta_j = \mu_{\cdot j} - \mu_{\cdot\cdot},\ (\alpha\beta)_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot}.$$

From these definitions, we can show that there are natural constraints on these effects

$$\sum \alpha_i = \sum \beta_j = 0 \tag{1}$$

$$\sum_{i=1}^{a}(\alpha\beta)_{ij} = \sum_{j=1}^{b}(\alpha\beta)_{ij} = 0 \tag{2}$$

We can now write down the factor effects form with constraints

$$Y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},\ k = 1, \ldots, n, j = 1, \ldots, b, i = 1, \ldots, a,$$

where $\{\epsilon_{ijk}\}$ are i.i.d. $N(0, \sigma^2)$ and

$$\sum_{i} \alpha_i = \sum_{j} \beta_j = 0 \tag{3}$$

$$\sum_{i=1}^{a}(\alpha\beta)_{ij} = \sum_{j=1}^{b}(\alpha\beta)_{ij} = 0 \tag{4}$$

Very often you may see an additive model $Y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + \epsilon_{ijk}$, where the interactions $\{(\alpha\beta)_{ij}\}$ are dropped. Dropping the interaction terms reduces the number of unknown parameters to estimate, which improves efficiency if the reduced model does not fall too far away from the truth. However, a reduced model places additional assumptions which may not be true in the real world. In practice, we need to carefully decide the form of model based on, but not limited to, background, questions of interest, prior knowledge, exploratory analysis, etc.

1.1.3 4.2.3 Estimation

We can estimate the cell means as before $\hat{\mu}_{ij} = \bar{Y}_{ij\cdot}, \hat{\mu}_{\cdot\cdot} = \bar{Y}_{\cdot\cdot\cdot}, \hat{\mu}_{i\cdot} = \bar{Y}_{i\cdot\cdot}, \hat{\mu}_{\cdot j} = \bar{Y}_{\cdot j\cdot}$. Noting that the effects are linear combinations of cell means, we can estimate them using the equalities in Section 4.2.2

$$\hat{\alpha}_i = \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot}, \hat{\beta}_j = \bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdot\cdot\cdot}, \hat{(\alpha\beta)}_{ij} = \bar{Y}_{ij} - (\bar{Y}_{\cdot\cdot\cdot} + \hat{\alpha}_i + \hat{\beta}_j).$$

As a result, the same set of constraints still hold on the estimators.

For each Y_{ijk}, the fitted value is $\bar{Y}_{ij\cdot}$. The residual is thus $e_{ijk} \equiv Y_{ijk} - \bar{Y}_{ij\cdot}$. We can show that the fitted value is independent with the residual. Just like in one-way ANOVA, this independence is the backbone for testing.

The sum of squares are as follow.

Residual sum of squares.

$$\text{SSE} = \sum_i \sum_j \sum_k e_{ijk}^2 = \sum_i \sum_j \sum_k \left(Y_{ijk} - \bar{Y}_{ij\cdot}\right)^2, df(\text{SSE}) = n_T - ab = (n-1)ab.$$

As before, we can define the mean squared errors as MSE = SSE/df(SSE), where we know that $\mathbb{E}[\text{MSE}] = \sigma^2$.

Total sum of squares.

$$\text{SSTO} = \sum_i \sum_j \sum_k \left(Y_{ijk} - \bar{Y}_{\cdots}\right)^2, \ df(\text{SSTO}) = n_T - 1 = nab - 1.$$

Sum of squares due to the main effect of factor A.

$$\text{SSA} = \sum_i \sum_j \sum_k \hat{\alpha}_i^2 = nb \sum_{i=1}^a \hat{\alpha}_i^2, \ df(\text{SSA}) = a - 1.$$

Sum of squares due to the main effect of factor B.

$$\text{SSB} = \sum_i \sum_j \sum_k \hat{\beta}_j^2 = na \sum_{j=1}^a \hat{\beta}_j^2, \ df(\text{SSB}) = b - 1.$$

Sum of squares due to the interaction effects.

$$\text{SSAB} = \sum_i \sum_j \sum_k \hat{(\alpha\beta)}_{ij}^2 = n \sum_i \sum_j \hat{(\alpha\beta)}_{ij}^2, \ df(\text{SSAB}) = (a-1)(b-1).$$

We have the following equalities

$$\text{SSTO} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}$$

and

$$df(\text{SSTO}) = df(\text{SSA}) + df(\text{SSB}) + df(\text{SSAB}) + df(\text{SSE}).$$

For the mean squares, when ϵs are i.i.d. $N(0, \sigma^2)$, we have $\hat{Y}_{ijk} \perp e_{ijk}$, which yields MSE $\perp$ MSA, MSB, MSAB.

Properties 1. $\mathbb{E}[\text{MSE}] = \sigma^2$; 2. $\mathbb{E}[\text{MSA}] = \sigma^2 + nb\sum_i \alpha_i^2/(a-1)$; 3. $\mathbb{E}[\text{MSB}] = \sigma^2 + na\sum_j \beta_j^2/(b-1)$; 4. $\mathbb{E}[\text{MSAB}] = \sigma^2 + n\sum_i \sum_j (\alpha\beta)_{ij}^2/(a-1)(b-1)$.

1.1.4 4.2.4 Inference

Tests for two-way ANOVA models are essentially identical to those for one-way ANOVA model. We discuss only the F-tests for the interactions and one main effect here.

F-test for interactions. Consider the following null and alternative hypotheses.

$$H_0 : (\alpha\beta)_{ij} = 0 \ \forall i, j \quad \text{v.s.} \quad H_1 : \text{ not all } (\alpha\beta)_{ij} \text{ are zero.}$$

The F-statistics is F^* = MSAB/MSE. Under the null, F^* follows an F-distribution with $df = ((a-1)(b-1), (n-1)ab)$.

Test for the main effect of factor A. Consider the following null and alternative hypotheses.

$$H_0 : \alpha_i = 0 \ \ \forall i \quad \text{v.s.} \quad H_1 : \text{ not all } \alpha_i \text{ are zero.}$$

We have F^* = MSA/MSE that follows an F-distribution with $df = (a-1, (n-1)ab)$.

Similar results hold for factor B. Theory and justifications for these tests are almost identical to those for one-way ANOVA.

MSE without interactions. In the absence of interactions, the MSE defined in the previous section no long makes sense. Instead, we can define $\hat{Y}_{ijk} = \hat{\mu}_{..} + \hat{\alpha}_i + \hat{\beta}_j$. This leads to $e'_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$. Consequently, SSE $= \sum\sum\sum \left(Y_{ijk} - \hat{Y}_{ijk}\right)^2$ and df(SSE) $= nab - a - b + 1$. Finally, we have that MSE = SSE$/(nab - a - b + 1)$.

Methods for testing linear combinations and simultaneous inference generalize to the case with two-way, and higher order, ANOVA model.

1.1.5 4.2.5 Special case: one observation per case

When there is only one observation per cell, the number of observations (ab) equals to the number of unknown means (ab). It is not possible to estimate the full model with interactions with proper measures of uncertainty. However, the number of observed cases does not alter the truth underneath the data. We may still want to test whether interactions are present.

Tukey's test for additivity. Consider the following model

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

where ϵ_{ij} are i.i.d. $N(0, \sigma^2)$. In this model, the fourth term on the right-hand side attempts to capture the interaction effects with limited capacity. The specific form $D\alpha_i\beta_j$ does not capture all possible interactions, but it is a work-around with limited observations. In this model, we can test

$$H_0 : D = 0 \quad \text{v.s.} \quad H_a : D \neq 0.$$

We can derive that

$$\hat{D} = \frac{\sum_i \sum_j Y_{ij}\hat{\alpha}_i\hat{\beta}_j}{\sum_i \hat{\alpha}_i^2 \sum_j \hat{\beta}_j^2}.$$

We can construct the F-test for $H_0 : D = 0$ in a similar manner as in Section 4.2.4. You can read more about this test in, e.g., here.

1.1.6 4.2.6 Imbalanced two-way ANOVA

A two-way ANOVA when the numbers of observations vary across cells takes the following form

$$Y_{ijk} = \mu_{ijk} + \epsilon_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \ k = 1, \ldots, n_{ij}, j = 1, \ldots, b, i = 1, \ldots, a,$$

where $\{\epsilon_{ijk}\}$ are i.i.d. $N(0, \sigma^2)$. The same constraints hold on the parameters.

The factor effects are still defined as follows

$$\mu_{\cdot\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}/(ab),\ \mu_{i\cdot} = \sum_{j=1}^{b}\mu_{ij}/b,\ \mu_{\cdot j} = \sum_{i=1}^{a}\mu_{ij}/a.$$

And furthermore, we have

$$\alpha_i = \mu_{i\cdot} - \mu_{\cdot\cdot},\ \beta_j = \mu_{\cdot j} - \mu_{\cdot\cdot},\ (\alpha\beta)_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot}.$$

Estimators follow the same form by plugging the sample means in the equations above.

However, it is **no longer true** that

$$\hat{\mu}_{\cdot\cdot} = \bar{Y}_{\cdot\cdot\cdot},\ \hat{\mu}_{i\cdot} = \bar{Y}_{i\cdot\cdot},\ \hat{\mu}_{\cdot j} = \bar{Y}_{\cdot j\cdot},\ \text{SSTO} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}.$$

To see this, we just need to expand $\bar{Y}_{\cdot\cdot\cdot}$ as weighted average of the sample cell means.

Suppose that we are interested in testing the presence of interactions.

$$H_0 : (\alpha\beta)_{ij} = 0 \text{ v.s. } H_1 : \text{not all } (\alpha\beta)_{ij} \text{ are zero.}$$

We now need to use the following framework for testing. * Full model: $Y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$ * Reduced model: $Y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + \epsilon_{ijk}$

The F-statistic is then

$$F^* = \frac{[\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}}]/[df_{\text{red}} - df_{\text{full}}]}{\text{SSE}_{\text{full}}/df_{\text{full}}},$$

where $F^* \sim F((a-1)(b-1), n_T - ab)$ under the null hypothesis.

Note: you can use the `Anova()` function in the `car` package if you want to fit an imbalanced ANOVA model.

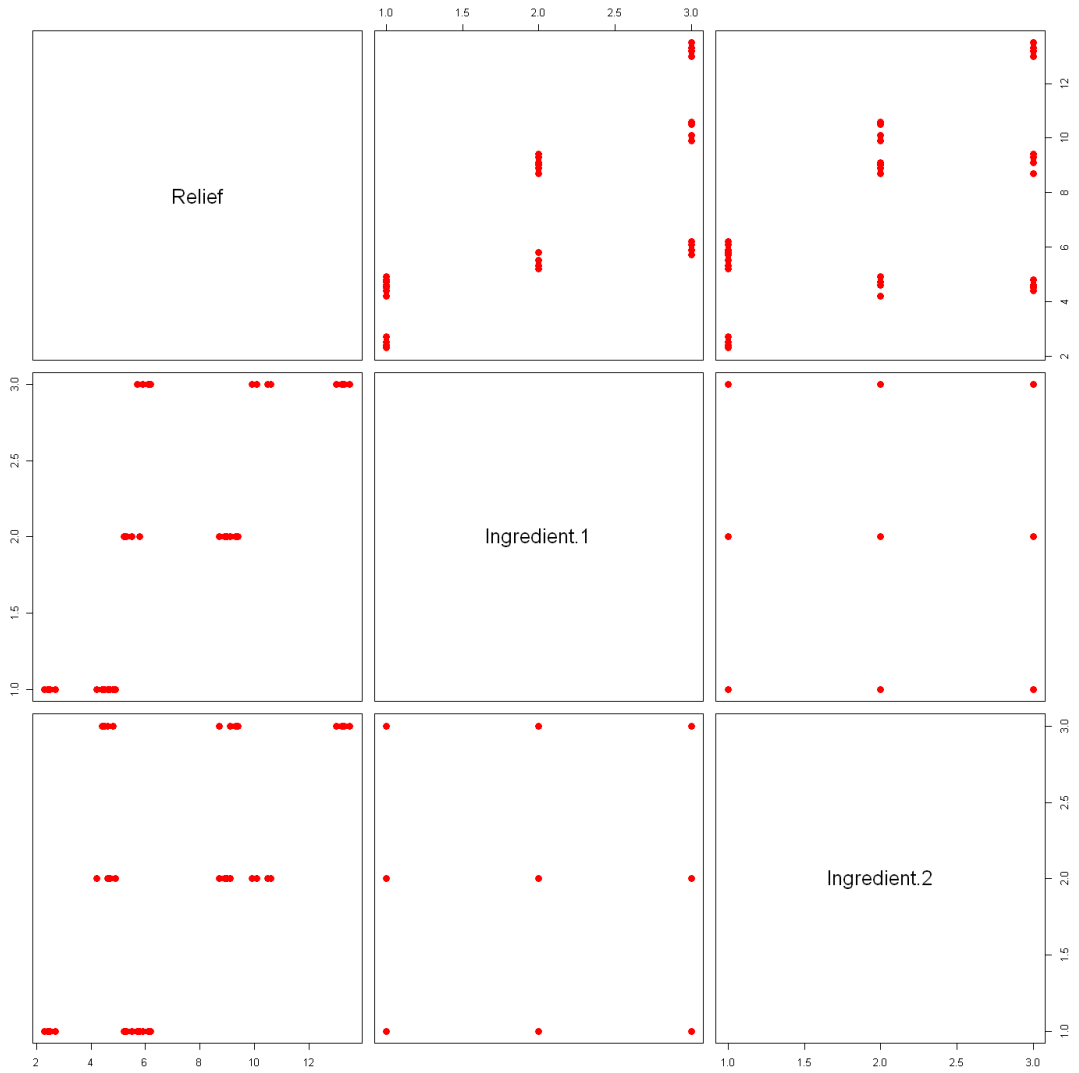
1.1.7 4.2.7 Example: Hay Fever data

Let's return to the Hay Fever data. If an analysis plan has been determined, as is usually the case for randomized control trials, we need to strictly follow the analysis plan. The pre-determined analysis plan makes sure that there won't be data mining attempts that inflates the type I errors.

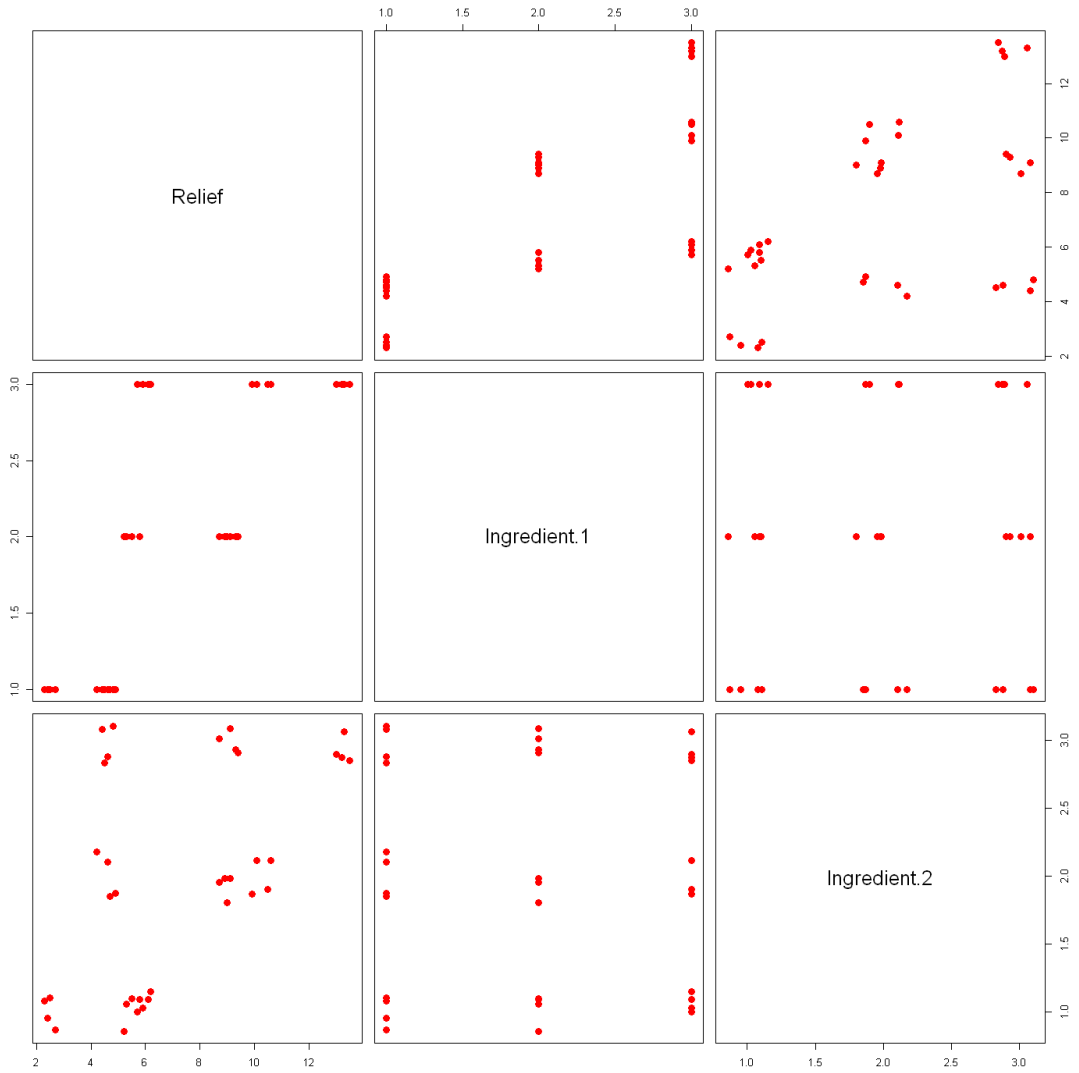
In this example, we will pretend that we have no such plans. Instead, our task is to explore and investigate the effectiveness of the two ingradients. We still, however, need to be mindful with type I errors from our inference. Think about the type I error as what determines the *weight* your results carry: the larger the type I error is, the more negligible your statistical inference is.

[22]:
```
Hay <- read.csv(file="../Data/HayFever.csv", header=TRUE, sep=",")

# Pairwise scatter plot
pairs(Hay,pch=16,col='red',cex=1.5)
```



[24]:
```
Hay.jittered = Hay;
Hay.jittered$Ingredient.2 = jitter(Hay.jittered$Ingredient.2)
# Jittered one factor
pairs(Hay.jittered,pch=16,col='red',cex=1.5)
```

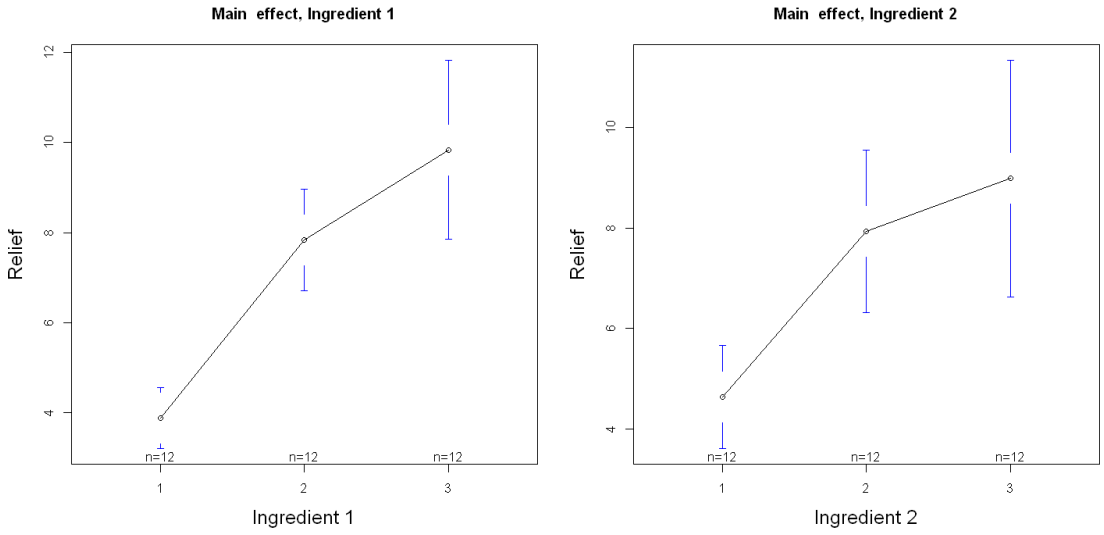


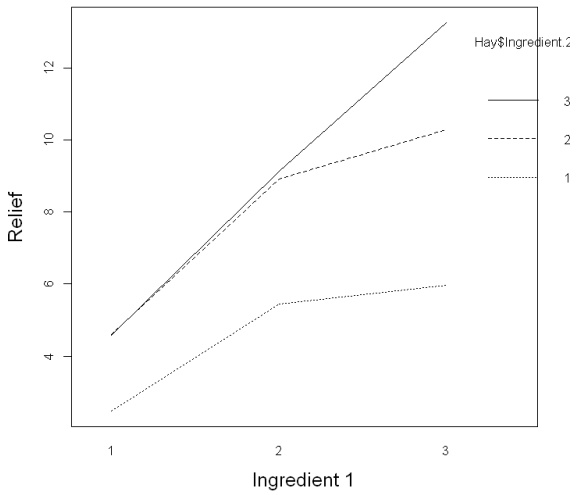
There are two factors with a total of 9 combinations in this data with 36 samples. It is natural to consider a two-way ANOVA model here. The next question is then whether the interaction effects are present, i.e., whether we want to use a two-way ANOVA model with or without interactions.

[25]:
```
# Exploratory analysis
library(gplots)
Hay$Ingredient.1=as.factor(Hay$Ingredient.1)
Hay$Ingredient.2=as.factor(Hay$Ingredient.2)

options(repr.plot.width=12, repr.plot.height=12)
par(mfrow=c(2,2))
# Main effect plot for ingredient 1
plotmeans(Relief~Ingredient.1,data=Hay,xlab="Ingredient 1",ylab="Relief",
```

```
          main="Main  effect, Ingredient 1",cex.lab=1.5)
# Main effect plot for ingredient 2
plotmeans(Relief~Ingredient.2,data=Hay,xlab="Ingredient 2",ylab="Relief",
          main="Main  effect, Ingredient 2",cex.lab=1.5)
#Interaction plot
interaction.plot(Hay$Ingredient.1, Hay$Ingredient.2, Hay$Relief
                 ,cex.lab=1.5,ylab="Relief",xlab='Ingredient 1')
par(mfrow=c(1,1))
```





[26]:
```
# Test for interactions
full_model=lm(Relief~as.factor(Ingredient.1)+as.factor(Ingredient.2)+as.
 ↪factor(Ingredient.1)*as.factor(Ingredient.2),data=Hay);
```

```
reduced_model=lm(Relief~as.factor(Ingredient.1)+as.factor(Ingredient.
↪2),data=Hay);
anova(reduced_model,full_model)
# Conclusion?
```

A anova: 2 × 6

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	31	31.050	NA	NA	NA	NA
2	27	1.625	4	29.425	122.2269	6.972083e-17

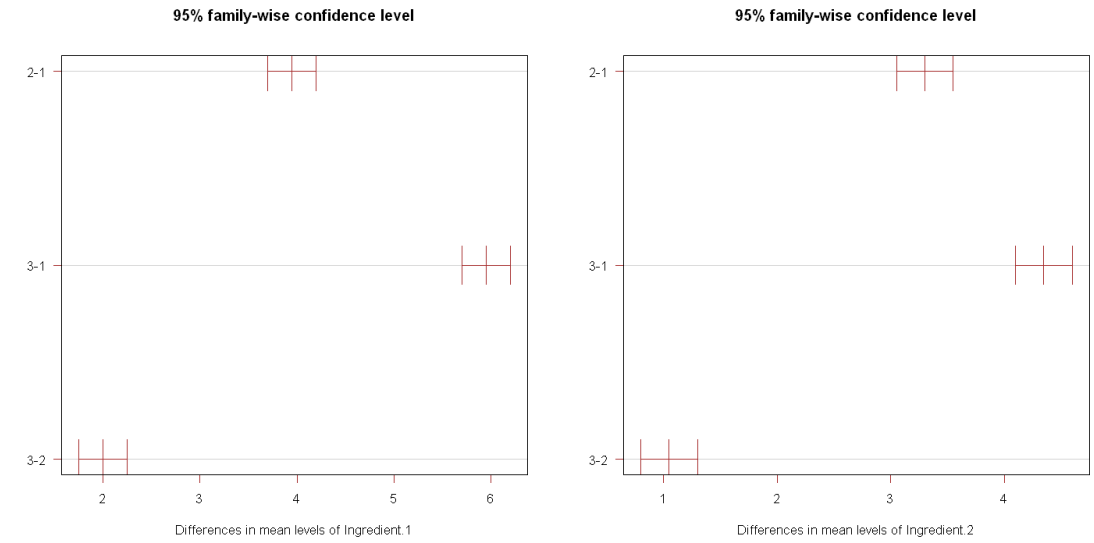
[27]:
```
# Fit the chosen model:
library(stats)
sig.level=0.05;
anova.fit<-aov(Relief~Ingredient.1+Ingredient.2+Ingredient.1*Ingredient.
↪2,data=Hay)
summary(anova.fit)
```

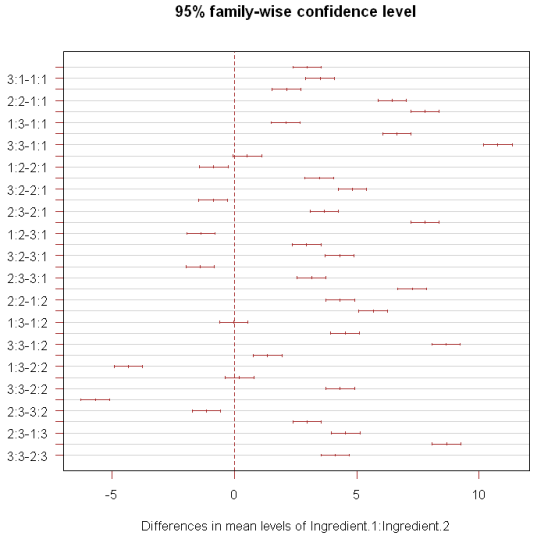
```
                         Df Sum Sq Mean Sq F value Pr(>F)
Ingredient.1              2 220.02  110.01  1827.9 <2e-16 ***
Ingredient.2              2 123.66   61.83  1027.3 <2e-16 ***
Ingredient.1:Ingredient.2 4  29.43    7.36   122.2 <2e-16 ***
Residuals                27   1.63    0.06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It seems that both integredients and their interaction may have (statistically) significant impact on the response variable. We may want to explore what the best combination of ingredients is, or whether such combination exists. The search of "best combination" can be translated as finding the cell with the highest cell mean compared to other cells, which leads us to the Tukey-Kramer method.

[32]:
```
# Find the best combination

T.ci=TukeyHSD(anova.fit,conf.level = 1-sig.level)
par(mfrow=c(2,2))
plot(T.ci, las=1 , col="brown")
par(mfrow=c(1,1))
```





[33]:
```
# We only need to pay attention to the differences of the two largest means
idx=list();
idx[[1]]=Hay$Ingredient.1;idx[[2]]=Hay$Ingredient.2;
(means.comb=tapply( Hay$Relief, INDEX=idx,mean))
# From this table, the two cells are (3,3) and (3,2)
```

A matrix: 3 × 3 of type dbl

	1	2	3
1	2.475	4.600	4.575
2	5.450	8.925	9.125
3	5.975	10.275	13.250

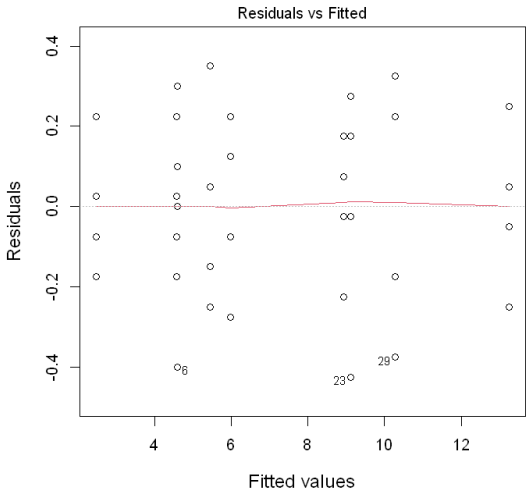
[39]:
```
# Find the confidence interval among the many comparisons...
T.ci[['Ingredient.1:Ingredient.2']]
```

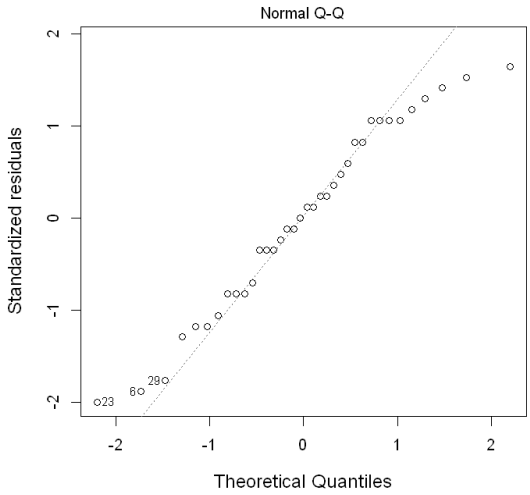
	diff	lwr	upr	p adj
2:1-1:1	2.975	2.3913187	3.5586813	2.142730e-14
3:1-1:1	3.500	2.9163187	4.0836813	4.440892e-15
1:2-1:1	2.125	1.5413187	2.7086813	5.300671e-11
2:2-1:1	6.450	5.8663187	7.0336813	4.218847e-15
3:2-1:1	7.800	7.2163187	8.3836813	4.218847e-15
1:3-1:1	2.100	1.5163187	2.6836813	6.947942e-11
2:3-1:1	6.650	6.0663187	7.2336813	4.218847e-15
3:3-1:1	10.775	10.1913187	11.3586813	4.218847e-15
3:1-2:1	0.525	-0.0586813	1.1086813	1.033088e-01
1:2-2:1	-0.850	-1.4336813	-0.2663187	1.142444e-03
2:2-2:1	3.475	2.8913187	4.0586813	4.551914e-15
3:2-2:1	4.825	4.2413187	5.4086813	4.218847e-15
1:3-2:1	-0.875	-1.4586813	-0.2913187	7.862241e-04
2:3-2:1	3.675	3.0913187	4.2586813	4.218847e-15
3:3-2:1	7.800	7.2163187	8.3836813	4.218847e-15
1:2-3:1	-1.375	-1.9586813	-0.7913187	5.283997e-07
2:2-3:1	2.950	2.3663187	3.5336813	2.520206e-14
3:2-3:1	4.300	3.7163187	4.8836813	4.218847e-15
1:3-3:1	-1.400	-1.9836813	-0.8163187	3.744504e-07
2:3-3:1	3.150	2.5663187	3.7336813	8.659740e-15
3:3-3:1	7.275	6.6913187	7.8586813	4.218847e-15
2:2-1:2	4.325	3.7413187	4.9086813	4.218847e-15
3:2-1:2	5.675	5.0913187	6.2586813	4.218847e-15
1:3-1:2	-0.025	-0.6086813	0.5586813	1.000000e+00
2:3-1:2	4.525	3.9413187	5.1086813	4.218847e-15
3:3-1:2	8.650	8.0663187	9.2336813	4.218847e-15
3:2-2:2	1.350	0.7663187	1.9336813	7.475322e-07
1:3-2:2	-4.350	-4.9336813	-3.7663187	4.218847e-15
2:3-2:2	0.200	-0.3836813	0.7836813	9.596929e-01
3:3-2:2	4.325	3.7413187	4.9086813	4.218847e-15
1:3-3:2	-5.700	-6.2836813	-5.1163187	4.218847e-15
2:3-3:2	-1.150	-1.7336813	-0.5663187	1.305447e-05
3:3-3:2	2.975	2.3913187	3.5586813	2.142730e-14
2:3-1:3	4.550	3.9663187	5.1336813	4.218847e-15
3:3-1:3	8.675	8.0913187	9.2586813	4.218847e-15
3:3-2:3	4.125	3.5413187	4.7086813	4.218847e-15

A matrix: 36 × 4 of type dbl

To wrap up this exploratory analysis, we want to investigate the plausibility of the assumptions we make. In `R`, the basic diagnostic plots can be conveniently called using the `plot()` function once we feed it with a fitted model. If you find any red flags, you may proceed to conduct formal testings or sensitivity analysis (i.e., fitting alternative models).

[52]:
```
# Diagnostic plots
options(repr.plot.width=12, repr.plot.height=6)
par(mfrow=c(1,2))
plot(anova.fit,cex.lab=1.2,which=1:2)
par(mfrow=c(1,1))
```


Residuals vs Fitted
Residuals
Fitted values
6
23
29


Normal Q-Q
Standardized residuals
Theoretical Quantiles
23
6
29