Fall 2015

Vitis Gene Expression Profiling Using Mixed Models

Yin Yin

VITIS GENE EXPRESSION PROFILING USING MIXED MODELS

A Master Thesis

Presented to

The Graduate College of

Missouri State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science, Mathematics

By

Yin Yin

December 2015

***VITIS* GENE EXPRESSION PROFILING USING MIXED MODELS**

Mathematics

Missouri State University, December 2015

Master of Science

Yin Yin

ABSTRACT

The purpose of this thesis is to analyze gene expression in grapevine under different treatments using a mixed linear statistical model. The experiment involves two *Vitis* species (*V. vinifera* Cabernet sauvignon and *V. aestivalis* Norton) and applies two different treatments to them (inoculation with *Erysiphe necator* conidiospores and mock inoculation). There are three biological replicates measured at each of the following six assigned time points: 0, 4, 8, 12, 24, and 48 hours. By setting up split-plot model for the data, statistical hypotheses concerning gene expressions, especially gene expressions in terms of treatment effect, are tested. The result of the analysis identify those genes expressed differently, and further experiments will indicate biological properties of those specific genes. After performing the analysis by using the split-plot model, discussions about another possible model, repeated measures design, is introduced at the end of this thesis in order to incorporate the potential biological property, such as diurnal pattern, into the modeling and analysis. By these series of analysis, certain genes are found with different expression, such as the gene with ID 000002 and ID 000004. This result will be useful in further biological researches.

KEYWORDS: gene expression profiling, residual analysis, data transformation and normalization, split-plot experiment design, repeated-measures experiment design, hypothesis testing.

This abstract is approved as to form and content

_____

Yingcai Su, Ph. D.
Chairperson, Advisory Committee
Missouri State University

VITIS GENE EXPRESSION PROFILLING USING MIXED MODELS

By

Yin Yin

A Master Thesis
Submitted to the Graduate College
Of Missouri State University
In Partial Fulfillment of the Requirements
For the Degree of Master of Science, Mathematics

December 2015

Approved:

_____
Yingcai Su, Ph. D.

_____
Laszlo G. Kovacs, Ph. D.

_____
Wenping Qiu, Ph.D.

_____
Julie Masterson, Ph.D: Dean, Graduate College

ACKNOWLEDGEMENTS

It would never have been possible for me to finish my thesis without the guidance of my committee members, help from professors in Mathematics department, Missouri State University.

I would like to express my deepest gratitude to my three committee members Dr. Laszlo G. Kovacs, Dr. Wenping Qiu, and Dr. Yingcai Su (chair) for their guidance, caring, patience, and providing me with an excellent atmosphere for doing research. Especially, I would like to thank Dr. Qiu for helping me get access to their valuable grapevine research and experiment data. I would also like to thank Dr. Kovacs for guiding my research and helping me to develop my background in biology.

In addition, I would like to thank all professors, Dr. George Mathew and Dr. Songfeng Zheng, for giving me statistical background. I would like to thank Dr. Shouchuan Hu, Dr. Paula Kemp, Dr. Mark W. Rogers, Dr. Kishor Shah, for giving me pleasant working experience.

My sincere thanks also goes to Dr. William Bray for offering me the opportunities of working for mathematics department and professors and giving me flexibility of time on doing research and projects.

I dedicate this thesis to my Chinese family, Yin's, and my American host family, Biser's, for supporting me spiritually throughout my school years at Missouri State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Understanding how the genome of an organism function is primary focus of modern biology research. The high cost of generating functional genomics data requires researchers to use experimental material as efficiently as possible, and expectations are that the data are analyzed with great precision. Usually, genomic studies involve a large number of observations, and the downstream analysis and interpretation of the resulting vast datasets is challenging. In order to gather information from data, substantial biology knowledge and statistical technique have to be applied in combination.

Functional genomics is a powerful tool to improve the health and productivity of agricultural plants. Researchers conducted an experiment involving two *Vitis* species, two treatments, and made three measurements at each of six assigned time points. This experiment involves the following:

1. Two *Vitis* species: *V. vinifera* Cabernet sauvignon (CS) and *V. aestivalis* Norton (N);

2. Two treatments applied to each species: inoculation with *E. necator* conidiospores (INC) and mock inoculation (MOC);

3. Observations at six time points for each species under each treatment: 0 hour, 4 hours, 8 hours, 12 hours, 24 hours, 48 hours;

4. Three biological replicates are made at each time point for each species under each treatment.

The main purpose of this experiment is to monitor the expression levels of 16436 *Vitis* genes across the two treatments. Statistical methods are applied to analyze the signal levels of each of these 16436 genes to assess the significance of the differences between these two treatments. The ultimate goal is to differentiate those genes that express at

significantly different level the treatments. In this procedure, we have to set up an appropriate statistical model from the data observed in experiment. Statistical hypothesis testing will be used, as well. For this certain experiment, we will use split-plot experiment design to model the data researchers got from experiment. In Chapter 2, we will use observations of one gene to explain how split-plot experiment design works and define related concepts, such as whole-plot and split-plot, and then we will interpret what the general split-plot model looks like. The split-plot model will be applied to all genes' observations to analyze gene expressions in Chapter 3. Chapter 4 will discuss whether there are other experiment design models which would be more appropriate to be applied, such as repeated measurement (experiment design) model, which will be discussed later in this thesis. And finally, Chapter 5 will contain a summary of all results and conclusions got from previous chapters.

CHAPTER 2: SPLIT-PLOT DESIGN

Split-plot designs are extremely popular in experimental design since they cover a common case in real world, which is when you have a factor that you want to study but cannot change as often as your other factors. In a factorial experimental arrangement, however, it is not always possible or practical to completely randomize the order of experiment and to obtain genuine independent replicates. These practical conditions or restrictions lead to a split-plot design. The following data are the signal values of a gene (called GENE00001) in the grapevine study which will be used to explain split-plot model.

Section 2.1 Example

Let us use the single stage Split-plot design as an example; models for multiple stages cases are derived similarly.

Consider observations of first gene (with ID GENE00001). We select the first replicates only at each time point as an example here to explain how the split-plot model work. Recall that we have two species, Cabernet sauvignon (CS) and Norton (N); two treatments, Inoculation with *Erysiphe - necator* conidiospores (INO) and mock inoculation (MOC); and that there are six assigned time points (0, 4, 8, 12, 24, 48 hours) for each species and each treatment.

Table 2.1.1 shows the sample data. It is clear that the four combinations of species and treatments, CS and INO, CS and MOC, N and INO, N and MOC, make four small groups or blocks. This gives a general randomized block structure. Six time points divide

each small group into six subgroups. These three factors, SPECIES, TREATMENT, and TIME are fixed factors. By viewing Table 2.1.1, you might think of this experiment as a 2×2×6 factorial design with one replication per cell. However, in a completely randomized factorial experiment, researchers would have to grow, for instance, 60 Cabernet sauvignon (CS) plants and apply Inoculation with *E. necator* conidiospores (INO). After the inoculation, they need to randomly select a time point from the hours 0, 4, 8, 12, 24, and 48, say 8, and then at the 8-hour time point, they need to harvest 10 leaves, one from each plant, to make one GeneChip in order to obtain observations. Similar procedures have to be done six times, to get all the six observations. Of course, this is not possible or practically not feasible! We would argue that other ways of carrying out the randomization is too expensive.

Table 2.1.1: Observations of GENE00001 (with First Replicate Signals Only)

Species	Treatment	Time	Observation	Species	Treatment	Time	Observation
CS	INO	0	2416.7	N	INO	0	2711.2
CS	INO	4	2519.7	N	INO	4	2745.4
CS	INO	8	2704	N	INO	8	2769.7
CS	INO	12	2434.6	N	INO	12	2347.3
CS	INO	24	2834	N	INO	24	2439.2
CS	INO	48	2680	N	INO	48	2662.2
CS	MOC	0	2819.2	N	MOC	0	2507.6
CS	MOC	4	3151.8	N	MOC	4	2587.9
CS	MOC	8	2262.5	N	MOC	8	2255.9
CS	MOC	12	2254.5	N	MOC	12	2453.9
CS	MOC	24	3105.1	N	MOC	24	2476.5
CS	MOC	48	3019.2	N	MOC	48	2678.3

What really happened is the experiment was done in a split-plot design setting. After the inoculation, experimenters randomly select 10 plants to produce the first GeneChip at time 0 (hour), at time 4 (4 hours later) randomly select another set of 10

plants to produce the second GeneChip, and so on. The entire experiment was replicated for other three treatment combinations (CS and MOC, N and INO, and N and MOC). These four treatment combinations are the whole plots, and time points (0, 4, 8, 12, 24, and 48) are the split-plots. The effects of the whole plot are confounded with the replicate effects.

Apart from those we discussed above, we might think that time effects are nested within each species and treatment combination, but actually they are not since each time point crossed with all species and treatment combinations.

Taking all effects into consideration, the split-plot model for this experiment is:

Model 1: $Y_{ijk} = \mu + \alpha_i + \theta_j + (\alpha\theta)_{ij} + \beta_k + (\alpha\beta)_{ik} + (\theta\beta)_{jk} + (\alpha\theta\beta)_{ijk} + \varepsilon_{ijk}$

Where

μ is the grand mean;

α_i is the species effect for the i^{th} species;

θ_j is the treatment for the j^{th} treatment;

$(\alpha\theta)ij$ is the interaction effect between the i^{th} species and the j^{th} treatment;

β_k is the time effect for the k^{th} time point;

$(\alpha\beta)_{ik}$ is the interaction effect for the i^{th} species at the k^{th} time point;

$(\theta\beta)_{jk}$ is the interaction effect for the j^{th} treatment at the k^{th} time point;

$(\alpha\theta\beta)_{ijk}$ is the interaction effect for the i^{th} species with the j^{th} treatment applied at the k^{th} time point;

ε_{ijk} is the random error term.

With those assumptions listed above, we can start to derive Sum of Squares and prepare for ANOVA table to do further analysis. For i = 1, 2; j = 1, 2; k = 1, 2. y_{ijk} is

observations. $\overline{y_{...}}$ is defined as overall mean. $\overline{y_{i..}}$,$\overline{y_{.j.}}$, $\overline{y_{..k}}$ represent mean of all observations of i^{th} species, mean of all observations of j^{th} treatment, and mean of all observations of k^{th} time points, respectively. $\overline{y_{ij.}}$,$\overline{y_{i.k}}$, $\overline{y_{.jk}}$ are mean of all observations of i^{th} species applied with j^{th} treatment, mean of all observations of i^{th} species at k^{th} time points and mean of all observations of j^{th} treatment at k^{th} time points, respectively.

Assumptions are involved. $\theta j \sim N(0, \sigma\theta 2)$, $(\alpha\theta)ij \sim N(0, \sigma\alpha\theta 2)$, $(\theta\beta)jk \sim N(0, \sigma\theta\beta 2)$, $(\alpha\theta\beta)ijk \sim N(0, \sigma\alpha\theta\beta 2)$, $\varepsilon ijk \sim N(0, \sigma\varepsilon 2)$ and they are mutually independent for i = 1, 2; j = 1, 2; k = 1, 2,…, 6.

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{6}(y_{ijk}-\overline{y_{...}})^2$$

$$=2\times 6\sum_{i=1}^{m}(\overline{y_{i..}}-\overline{y_{...}})^2+2\times 6\sum_{j=1}^{2}(\overline{y_{.j.}}-\overline{y_{...}})^2+6\sum_{i=1}^{2}\sum_{j=1}^{2}(\overline{y_{ij.}}-\overline{y_{i..}}-\overline{y_{.j.}}+\overline{y_{...}})^2$$

$$+2\times 2\sum_{k=1}^{6}(\overline{y_{..k}}-\overline{y_{...}})^2+2\sum_{i=1}^{2}\sum_{k=1}^{6}(\overline{y_{i.k}}-\overline{y_{i..}}-\overline{y_{..k}}+\overline{y_{...}})^2$$

$$+2\sum_{j=1}^{2}\sum_{k=1}^{6}(\overline{y_{.jk}}-\overline{y_{.j.}}-\overline{y_{..k}}+\overline{y_{...}})^2$$

$$+\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{6}(y_{ijk}-\overline{y_{ij.}}-\overline{y_{i.k}}-\overline{y_{.jk}}+\overline{y_{i..}}+\overline{y_{.j.}}+\overline{y_{..k}}+\overline{y_{...}})^2$$

That is, $SS_{Total} = SS_{Species} + SS_{Treatment} + SS_{Species\times Treatment} + SS_{Time} + SS_{Species\times Time} + SS_{Treatment\times Time} + SS_{Species\times Treatment\times Time}$. Specifically, we can get these results:

$SS_{Total} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{6}(y_{ijk}-\overline{y_{...}})^2$; d.f. = 2×2×6 - 1

$SS_{Species} = 2\times 6\sum_{i=1}^{2}(\overline{y_{i..}}-\overline{y_{...}})^2$; d.f. = 2-1

$SS_{Treatment} = 2\times 6\sum_{j=1}^{2}(\overline{y_{.j.}}-\overline{y_{...}})^2$; d.f. = 2-1

$SS_{Species\times Treatment} = 6\sum_{i=1}^{2}\sum_{j=1}^{2}(\overline{y_{ij.}} - \overline{y_{i..}} - \overline{y_{.j.}} + \overline{y_{...}})^2$; d.f. = (2-1) (2-1)

$SS_{Time} = 2 \times 2\sum_{k=1}^{6}(\overline{y_{..k}} - \overline{y_{...}})^2$; d.f. = 6-1

$SS_{Species\times Time} = 2\sum_{i=1}^{2}\sum_{k=1}^{6}(\overline{y_{i.k}} - \overline{y_{i..}} - \overline{y_{..k}} + \overline{y_{...}})^2$; d.f. = (2-1) (6-1)

$SS_{Treatment\times Time} = 2\sum_{j=1}^{2}\sum_{k=1}^{6}(\overline{y_{.jk}} - \overline{y_{.j.}} - \overline{y_{..k}} + \overline{y_{...}})^2$; d.f. = (2-1) (6-1)

$SS_{Species\times Treatment\times Time} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{6}(y_{ijk} - \overline{y_{ij.}} - \overline{y_{i.k}} - \overline{y_{.jk}} + \overline{y_{i..}} + \overline{y_{.j.}} + \overline{y_{..k}} + \overline{y_{...}})^2$;

d.f. = (2-1) (2-1) (6-1)

On the procedure of deriving, we cannot find any expression to estimate the sum of square of our error term ε_{ijk}; in other words, the error term in the model above is not estimable. Split-plot design has both whole plots and split-plots, which leads to multiple error term for different plot levels. In terms of this, when doing analysis and hypothesis testing, we should figure out which term to use as error term.

Since we define SPECIES and TREATMENT as whole-plot factors, we can consider the interaction effect of SPECIES and TREATMENT as the error term within whole-plot level. We consider the interaction between treatment and species to be a random effect since the interaction effect is different for each species and each treatment. We can only control the treatment and the species, but we cannot control or predict the effect of this interaction. Applying similar logic to split-plot level, interaction effect of SPECIES, TREATMENT and TIME will be a possible error term to use in doing analysis. In fact, there exists statistical reason to doing this prediction. How to determine these error is shown in Table 2.1.2. Details of each term in the EMS column will be discussed in Appendix II. Table 2.1.2 classifies all factors into whole-plot factors and split-plot factors and gives degree of freedom directly, which is convenient. We have another thing to do before analysis, which is, normalizing data. Due to all normal

assumptions we made above for our model, we have to check if our data is (approximate) normally distributed, or not. For raw data listed in Table 2.1.1, we can get histogram as what shows in Figure 2.1.

Table2.1.2: EMS (Error of Mean Squares) of model for GENE00001 sample (with first replicates only)

	Sources	*d.f.*	2 F i	2 R j	6 F k	EMS
Whole plot	α_i	2-1	0	2	6	$\sigma_\varepsilon^2 + 6\sigma_{\alpha\theta}^2 + 2\times 6\frac{\sum_i \alpha_i^2}{2-1}$
	θ_j	2-1	2	1	6	$\sigma_\varepsilon^2 + 2\times 6\frac{\sum_j \theta_j^2}{2-1}$
	$(\alpha\theta)_{ij}$	(2-1)(2-1)	0	1	6	$\sigma_\varepsilon^2 + 6\sigma_{\alpha\theta}^2$
Split plot	β_k	6-1	0	2	0	$\sigma_\varepsilon^2 + 2\,\sigma_{\alpha\theta\beta}^2 + 2\frac{\sum_j\sum_k(\theta\beta)_{jk}^2}{(2-1)(6-1)} + 2\times 2\frac{\sum_k \beta_k^2}{6-1}$
	$(\alpha\beta)_{ik}$	(2-1)(6-1)	0	2	0	$\sigma_\varepsilon^2 + 2\,\sigma_{\alpha\theta\beta}^2 + 2\frac{\sum_i\sum_j(\alpha\beta)_{ij}^2}{(2-1)(6-1)}$
	$(\theta\beta)_{jk}$	(2-1)(6-1)	0	1	0	$\sigma_\varepsilon^2 + 2\,\sigma_{\alpha\theta\beta}^2$
	$(\alpha\theta\beta)_{ijk}$	(2-1)(2-1)(6-1)	0	1	0	$\sigma_\varepsilon^2 + \sigma_{\alpha\theta\beta}^2$
	ε_{ijk}		1	1	1	σ_ε^2 *(not estimable)*
	Total	2×2×6-1				

The raw data of this sample is approximately bell-shaped (referring to Figure 2.1), except that the data range from 2000 to 3400. The magnitude of raw data is too large, which increases noise effects. Normalization is necessary here. Usually, two method will be used. One is making log-transformation. Log-transformation decrease the magnitude of the data. The other is finding z-scores for each of the observations. Z-scores calculation forces data to be normally distributed. This approach usually is used when the

histogram of data is not close to bell-shaped at all. For our sample data, the original histogram is already close to the histogram of normal distribution, so log-transformation should be enough. The basic idea is transform our raw data into $\log_2$ (raw data + 1). The reason for adding 1 is to avoid the case when some observations are zeros. After doing this, we will get the data and its histogram as shown in Table 2.1.3.

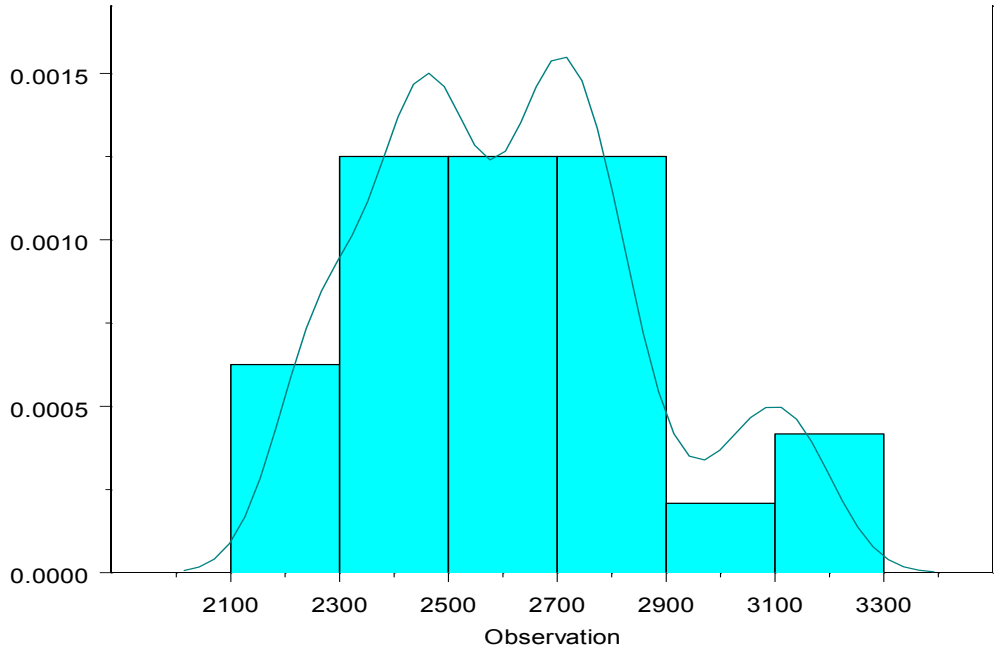

Figure 2.1 Histogram for raw data of GENE00001 (first replicate only)

Notice that this histogram in Figure 2.2 has two peaks, which is different from the unimodal normal distribution. For most experiment, this is acceptable. We can use Normal QQ-plot to support our statement. In Figure 2.3, we notice all points are extremely close to the straight line. Once the line made by points is approximately a straight line, we claim that the data is approximately normal, which is totally enough for experimental modeling.

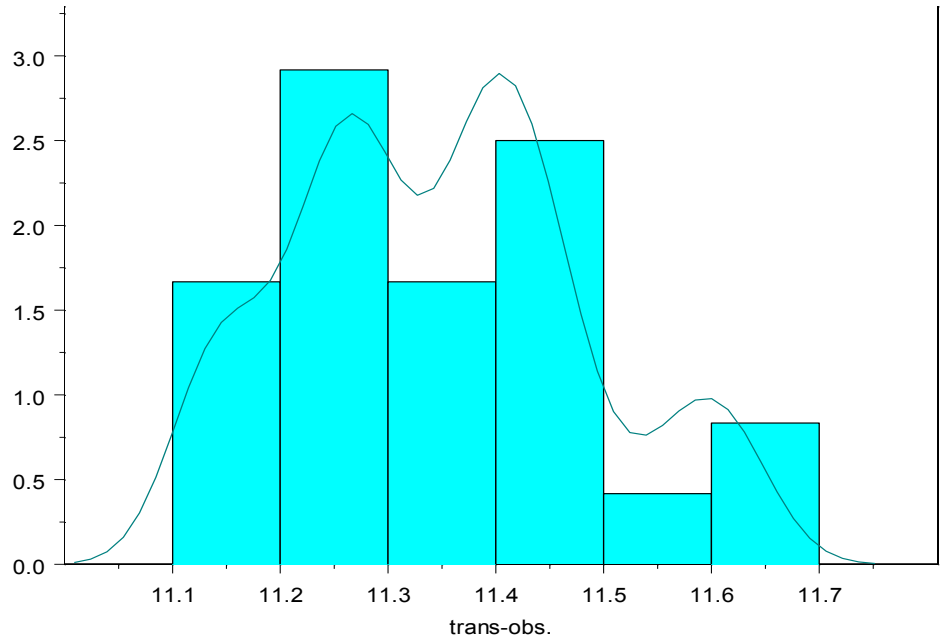

Figure 2.2 Histogram for transformed data of GENE00001 (first replicate only)

Table 2.1.3: Transformed Observations of GENE00001 sample (with first replicate signals only)

Species	Treatment	Time	Observation	Specie	Treatment	Time	Observation
CS	INO	0	11.2394	N	INO	0	11.4052
CS	INO	4	11.2996	N	INO	4	11.4233
CS	INO	8	11.4014	N	INO	8	11.4360
CS	INO	12	11.2500	N	INO	12	11.1974
CS	INO	24	11.4691	N	INO	24	11.2527
CS	INO	48	11.3885	N	INO	48	11.3789
CS	MOC	0	11.4615	N	MOC	0	11.2926
CS	MOC	4	11.6224	N	MOC	4	11.3381
CS	MOC	8	11.1443	N	MOC	8	11.1401
CS	MOC	12	11.1392	N	MOC	12	11.2614
CS	MOC	24	11.6008	N	MOC	24	11.2746
CS	MOC	48	11.5604	N	MOC	48	11.3876

Now the data are ready for analysis. By using statistical software, we can get the following ANOVA table (referring Table 2.1.4). We can read all results of sum of squares we listed earlier in this chapter from the Sum of Square column, directly. Notice that this ANOVA table is not classified by whole-plot factor and split-plot factor, so we

need to be really careful when determined error terms to apply. Recall from the discussion above that, we consider interaction of SPECIES and TREATMENT as whole plot error, and interaction among SPECIES, TREATMENT and TIME is the split-plot error.

Therefore, MSE of whole-plot error is 0.03224905 and all whole-plot factors should be tested with using this value. Similarly, MSE of split-plot error is 0.01101595, and it is used to test split-plot factor. Of course, degree of freedoms should all be matched to tests. For example, if we want test SPECIES factor with hypothesis statement as

$$H_0: \alpha_i = 0 \; vs. H_a: \alpha_i \neq 0$$

The F-ratio should be $\frac{0.02591636}{0.03224905}$ and the critical value read from F distribution table should be with degree of freedom (1, 1). The results obviously show that SPECIES factor is significant.

Checking effects factor by factor, it is clear that all term in our model are significant, which implies that this model is fit for our sample data. Some useful index can display this fitness well, residual, for example. If we can get an approximately horizontal residual plot, we can claim that the model we set up for specific set of data is fit well. Due to the original experiment design and the way we pick our sample data, the residuals for this model is zeros, which is possible but rare in most cases.

Until now, we analyze our sample data precisely with a lot of details from different aspects and get some knowledge about split-plot design by this specific case. All those ideas and concerns can be applied to general split-plot case. Now, let us see the general split-plot looks like.

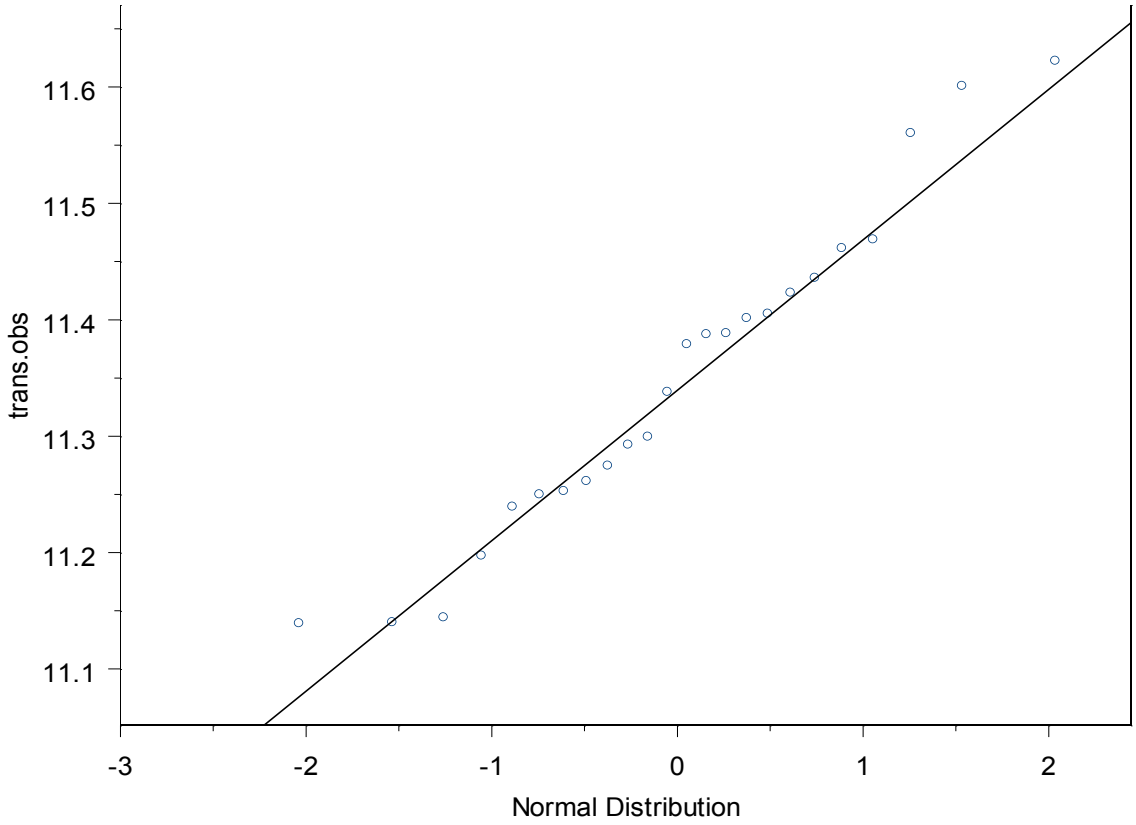

Figure 2.3 Normal QQ-plot of model for GENE00001 sample

Table 2.1.4 ANOVA table of Split-plot model for GENE00001 Sample

Sources	Degree of Freedom	Sum of Squares (SS)	Mean Square Error (MSE)
Species	1	0.0259164	0.02591636
Treatment	1	0.0002777	0.00027767
Time	5	0.1501550	0.03003100
Species×Treatment	1	0.0322491	0.03224905
Species×Time	5	0.0638854	0.01277709
Treatment×Time	5	0.1078854	0.02157709
Species×Treatment×Time	5	0.0550798	0.01101595

Section 2.2 Split-plot Design

What is split-plot design? Split-plot experimental designs were originally developed by Fisher (1966) for use of in agriculture experiments. Split-plot designs are used when treatment factors can be categorized into two groups. Group A includes any factors that are relatively stable during the experiment, and Group B contains other relatively unstable factors. Formally speaking, Group A factors define whole-plot and Group B factors creates split-plots. It is not hard to notice that split plots are nested within whole plots. Usually, Split-plot design is nested within standard designs, such as Completely Randomized Design (CRD), Randomized Block Design (RBD) and Latin Square Design (LSD). This special structure makes Split-plot designs different from other standard designs. Depending on the needs of experiment, whole plot can split up multiple times to make multiple stages of split plots, for instance, Split-split-plot design. The model of Split-plot designs depends on how many stages of split plots the experiment has and which standard design it works with.

Continuing using those notation from our previous example, the general split-plot model usually as following:

Model 1: $Y_{ijk} = \mu + \alpha_i + \theta_j + (\alpha\theta)_{ij} + \beta_k + (\alpha\beta)_{ik} + (\theta\beta)_{jk} + (\alpha\theta\beta)_{ijk} + \varepsilon_{ijk}$

if there were only two whole-plot factor and one split-plot factor as the case in our example. Similarly, all these terms in the model is defined as: μ is the grand mean; α_i is the species effect for i^{th} species; θ_j is the treatment for j^{th} treatment; $(\alpha\theta)ij$ is the interaction effect between i^{th} species and j^{th} treatment applied; β_k is the time effect for k^{th} time point; $(\alpha\beta)_{ik}$ is the interaction effect for i^{th} species and k^{th} time point; $(\theta\beta)_{jk}$ is the

interaction effect for j^{th} treatment and k^{th} time point ; $(\alpha\theta\beta)_{ijk}$ is the interaction effect among i^{th} species, j^{th} treatment and k^{th} time point ; and ε_{ijk} is the random error term .

The assumptions are $\theta_j \sim N(0, \sigma_\theta^2)$, $(\alpha\theta)_{ij} \sim N(0, \sigma_{\alpha\theta}^2)$, $(\theta\beta)_{jk} \sim N(0, \sigma_{\theta\beta}^2)$, $(\alpha\theta\beta)_{ijk} \sim N(0, \sigma_{\alpha\theta\beta}^2)$, $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$ and they are mutually independent for i = 1, 2, …, I; j = 1, 2, …, J; k = 1, 2,…, K. General split plot model still follows the sum of squares rule:

$SS_{Total} = SS_{Species} + SS_{Treatment} + SS_{Species\times Treatment} + SS_{Time} + SS_{Species\times Time} + SS_{Treatment\times Time} + SS_{Species\times Treatment\times Time}$.

Generalizing what we have already gotten from specific example above, the formulas for sum of squares can be calculated as

$SS_{Total} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \overline{y_{...}})^2$;

$SS_{Species} = JK\sum_{i=1}^{I}(\overline{y_{i..}} - \overline{y_{...}})^2$;

$SS_{Treatment} = IK\sum_{j=1}^{J}(\overline{y_{.j.}} - \overline{y_{...}})^2$;

$SS_{Species\times Treatment} = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{y_{ij.}} - \overline{y_{i..}} - \overline{y_{.j.}} + \overline{y_{...}})^2$;

$SS_{Time} = IJ\sum_{k=1}^{K}(\overline{y_{..k}} - \overline{y_{...}})^2$;

$SS_{Species\times Time} = J\sum_{i=1}^{I}\sum_{k=1}^{K}(\overline{y_{i.k}} - \overline{y_{i..}} - \overline{y_{..k}} + \overline{y_{...}})^2$;

$SS_{Treatment\times Time} = I\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{y_{.jk}} - \overline{y_{.j.}} - \overline{y_{..k}} + \overline{y_{...}})^2$;

$SS_{Species\times Treatment\times Time} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \overline{y_{ij.}} - \overline{y_{i.k}} - \overline{y_{.jk}} + \overline{y_{i..}} + \overline{y_{.j.}} + \overline{y_{..k}} + \overline{y_{...}})^2$;

Sum of square of error term ε_{ijk} is still not estimable, even for the general model. There are many reasons leading to this kind of estimation problem, but the typical one is that the error term has already been partitioned by other random effects in the model. In terms of this concern, we need to modify the existing model to make it standard. Expected Mean Squares (EMS) is an effective tool. Furthermore, for standard designs, all

factor effects are treated as fixed effects. Here in Split-plot design, some factors are random and they may be randomized in different stages. This randomization structure leaves a challenge to data analysis and related testing. Especially in F-tests, we have to be careful about the denominators since they may not be the MSE as what we have been used in standard design. To determine the appropriate denominators, we need to know how to write the EMS for all sources of variation.

The specific method and rules about how to calculate EMS column in Table 2.2.1 will be shown in Appendix II with i = 1, 2,…, I; j = 1, 2,…, J; k = 1, 2,…, K. This is a serious disadvantage for full model as Model 1 above since Mean square error plays main role in many kinds of testing. Combining with what we get from Table 2.2.1, $\sigma_\varepsilon^2 + K\sigma_{\alpha\theta}^2$ is treated as the EMS of whole-plot error and $\sigma_\varepsilon^2 + \sigma_{\alpha\theta\beta}^2$ is taken as the EMS of split-plot error.

Back to the Model 1 we set up, the interaction term $(\alpha\theta)_{ik}$ is often referred to as the whole plot error representing by $\varepsilon^w_{i(j)}$ and $\varepsilon^w_{i(j)}$'s ~ $i.i.d.N(0, \sigma_W^2)$.The usual assumption is that this interaction does not exist, that this term is really an estimate of the error within the whole plot. The term $(\alpha\theta\beta)_{ijk}$ is referred to as the split plot error representing by $\varepsilon^s_{k(ij)}$ and $\varepsilon^s_{k(ij)}$'s ~ $i.i.d.\ N(0, \sigma_S^2)$.

Sometimes the $(\alpha\beta)_{ik}$ or $(\theta\beta)_{jk}$ is also considered to be nonexistent depending on experiment data, and is combined with $(\alpha\theta\beta)_{ijk}$ as a part of error term ε_{ijk} (Hicks et al, 1999). It's clear that there is a nested blocking structure: whole-plots are nested within blocks and split plots are nested within whole plots. This structure leads to two kinds of randomization. One source of randomization are the whole plots, the other is the split plots.

Table 2.2.1 EMS for Split-plot design based on GENE00001 sample

	Sources	*d.f.*	*I* *F* *i*	*J* *R* *j*	*K* *F* *k*	EMS
	α_i	I-1	0	J	K	$\sigma_\varepsilon^2 + K\sigma_{\alpha\theta}^2 + JK\frac{\sum_i \alpha_i^2}{m-1}$
Whole plot	θ_j	J-1	I	1	K	$\sigma_\varepsilon^2 + IK\sigma_\theta^2$
	$(\alpha\theta)_{ij}$	(I-1)(J-1)	0	1	K	$\sigma_\varepsilon^2 + K\sigma_{\alpha\theta}^2$
	β_k	K-1	0	J	0	$\sigma_\varepsilon^2 + \sigma_{\alpha\theta\beta}^2 + J\sigma_{\theta\beta}^2 + IJ\frac{\sum_k \beta_k^2}{b-1}$
	$(\alpha\beta)_{ik}$	(I-1)(K-1)	0	J	0	$\sigma_\varepsilon^2 + \sigma_{\alpha\theta\beta}^2 + J\frac{\sum_i \sum_j (\alpha\beta)_{ij}^2}{(m-1)(b-1)}$
Split plot	$(\theta\beta)_{jk}$	(J-1)(K-1)	0	1	0	$\sigma_\varepsilon^2 + J\sigma_{\theta\beta}^2$
	$(\alpha\theta\beta)_{ijk}$	(I-1)(J-1)(K-1)	0	1	0	$\sigma_\varepsilon^2 + \sigma_{\alpha\theta\beta}^2$
	ε_{ijk}		1	1	1	σ_ε^2 *(not estimable)*
	Total	IJK-1				

After all this preliminary work, we can finally build up the ANOVA table for Model 1 is as Table 2.2.2.

Moreover, setting up EMS table as Table 2.2.1 is extremely helpful in determining which effects are comparable and is valid to be tested by using hypothesis testing. We can compare EMS directly from Table 2.3.2. For example, EMSs of α_i and $(\alpha\theta)_{ij}$ are $\sigma_\varepsilon^2 + K\sigma_{\alpha\theta}^2 + JK\frac{\sum_i \alpha_i^2}{m-1}$and $\sigma_\varepsilon^2 + K\sigma_{\alpha\theta}^2$, respectively. They share first two terms and $JK\frac{\sum_i \alpha_i^2}{m-1}$ is the only thing that makes the difference, which gives the clue that it may be easy to do testing for α_i, and that actually, it represents the fixed effect in whole-plot stage. Similarly, it is not hard to notice that all fixed factors and their interaction effect can be easily tested by our common used F-test. There is one important aspect needs to

be pointed out when doing F-test for this kind of Split-plot design. We have referred to the term "stages "many times; it is significant to figure out which stage the effect you want to test located in since the "stage" determine the denominator you would use if you need to do an F-test.

Table 2.2.2 ANOVA Table for General Split-plot Model

Source of Variance	D.F.	S.S.	M.S.	F-Ratio
Species	I-1	$SS_{Species}$	$MS_{Species}$	$MS_{Species}/MS_E{}^W$
Treatment	J-1	$SS_{Treatment}$		
Whole-plot error	(I-1)(J-1)	$SS_E{}^W$	$MS_E{}^W$	
Time	K-1	SS_{Time}	MS_{Time}	$MS_{Time}/MS_E{}^S$
Interaction effect between Species&Treatment	(I-1)(K-1)	$SS_{Species\times Treatment}$	$MS_{Species\times Treatment}$	$MS_{Species\times Treatment}/MS_E{}^S$
Interaction effect between Treatment & Time	(J-1)(K-1)	$SS_{Treatment\times Time}$	$MS_{Treatment\times Time}$	
Split-plot error	(I-1)(J-1)(K-1)	$SS_E{}^S$	$MS_E{}^S$	
Total	IJK-1	SSTO		

Based on those values in Table 2.2.2, we can do the following tests for fixed effects:

H_0: $\alpha_i = 0\ vs. H_a$: $\alpha_i \neq 0$, by using $F_0 = {MS_{Species}}/{MS_E{}^W}$

H_0: $\beta_k = 0\ vs. H_a$: $\beta_k \neq 0$, by using $F_1 = {MS_{Time}}/{MS_E{}^S}$

H_0: $(\alpha\beta)_{ij} = 0\ vs. H_a$: $(\alpha\beta)_{ij} \neq 0$, by using $F_2 = {MS_{Species\times Time}}/{MS_E{}^S}$

Notice that the interaction effect between Time and Treatment is not valid to do an F-test and the term $(\alpha\theta\beta)_{ijk}$ is used as split-plot error for F-test. We can start to think about if the related terms in Model 1 is really necessary to list out separately or not. In statistics, it is always advantageous to use fewer parameters than more. Consider to modify Model 1 as this:

$$\text{Model 2: } Y_{ijk} = \mu + \alpha_i + \theta_j + (\alpha\theta)_{ij} + \beta_k + (\alpha\beta)_{ik} + \varepsilon_{ijk}$$

where

α_i: whole-plot factor (Species) main effect;

θ_j: Treatment effect;

$(\alpha\theta)_{ij}$: whole-plot error;

β_k: split-plot factor (Time) main effect;

$(\alpha\beta)_{ik}$: interaction effect between Species and Time;

ε_{ijk}: split-plot error;

Assumptions are hold for θ_j, $(\alpha\theta)_{ij}$ and ε_{ijk} as what we stated in Model 1. This simplified model combines interaction effects of TIME and TREATMENT, and SPECIES and Treatment into ε_{ijk}, and it satisfies entirely to cases if those two interaction effects listed above are not of the interest. The related procedures for sum of squares and EMS are nothing but a little bit change in whole plot terms as what we did for Model 1. Here, it no need to write down again.

Section 2.3 Matrix Form

For both mixed model and linear model, it is more convenient to use matrix form to rewrite the specific models. Matrix form formula is helpful to distinguish fixed

variables from random variable, comparing with regular model formulas as what we used previously in this chapter. For mixed model, such as split-plot model here, model allows the existence of both fixed and random effects. But in specific case, one of them may missing. Like the model we used for our sample data, all factors are fixed actually. So the model we generalized here don't have random variables. In order to apply matrix notation to general split-plot model, let ignore the original experiment design and assume θ_j are random, and others are fixed.

Usually, $\vec{Y}$ represents matrix of all depending variable, such as Y_{ijk} in our model. $\vec{\beta}$ contains all fixed effect parameter vectors, for instance, $\overrightarrow{\alpha_i}$, $\overrightarrow{\beta_k}$, $\overrightarrow{\alpha\beta_{ik}}$ in Model 2. $\vec{Z}$ is a matrix filled with 1's and 0's depending on specific model and experiment requirement. $\vec{U}$ is a matrix containing all random effect variable vectors, such as $\overrightarrow{\theta_j}$, and $\vec{e}$ is matrix of all error terms. The details and properties of those matrices above will be shown in Appendix III. Our reduced Model 2, here, can be written as

$\vec{Y} = \vec{X}\vec{\beta} + \vec{Z}\vec{U} + \vec{e}$, where

$$\vec{Y} = \begin{pmatrix} y_{111} \\ \vdots \\ y_{116} \\ y_{121} \\ \vdots \\ y_{126} \\ y_{211} \\ \vdots \\ y_{216} \\ y_{221} \\ \vdots \\ y_{226} \end{pmatrix}, \vec{X}\vec{\beta} = \begin{pmatrix} \alpha_1 & \beta_1 & \alpha\beta_{11} \\ \vdots & \vdots & \vdots \\ \vdots & \beta_6 & \alpha\beta_{16} \\ \vdots & \beta_1 & \alpha\beta_{11} \\ \vdots & \vdots & \vdots \\ \alpha_1 & \beta_6 & \alpha\beta_{16} \\ \alpha_2 & \beta_1 & \alpha\beta_{11} \\ \vdots & \vdots & \vdots \\ \vdots & \beta_6 & \alpha\beta_{16} \\ \vdots & \beta_1 & \alpha\beta_{11} \\ \vdots & \vdots & \vdots \\ \alpha_2 & \beta_6 & \alpha\beta_{16} \end{pmatrix}, \vec{Z}\vec{U} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_2 \\ \theta_1 \\ \vdots \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_2 \end{pmatrix}, \text{ and } \vec{e} = \begin{pmatrix} \varepsilon_{111} \\ \vdots \\ \varepsilon_{116} \\ \varepsilon_{121} \\ \vdots \\ \varepsilon_{126} \\ \varepsilon_{211} \\ \vdots \\ \varepsilon_{216} \\ \varepsilon_{221} \\ \vdots \\ \varepsilon_{226} \end{pmatrix}$$

Section 2.4 SAS for Split-plot Design

The proper analysis of a split-plot design must account for the fact that treatments applied to main plots are subject to larger experimental error than those applied to subplots. Hence, different means squares must be used as denominators for the corresponding F-ratios. Also, many means comparisons of potential interest have error terms that are linear combinations of mean squares. While PROC GLM is useful for determining expected mean squares, PROC MIXED is better suited to analyze split-plot design. (Littell 2006)

Based on general analysis in last section, TREATMENT is considered as random and Model 2 is appropriate to use here. Both PROC GLM and PROC MIXED can be used. In last section, EMS table (Table 2.2.1) and ANOVA table (Table 2.2.2) have been set up by pure structure analysis and computation. In SAS, those complex procedures can be done with several statements.

Doing analysis with PROC GLM in SAS.

```
proc glm;
class species treatment time;
model Y= species treatment species*treatment time species*time/ss3;
test h=species e=species*treatment;          /* random treatment species*treatment/test; */
run;
```

In PROC GLM, the TEST option "test h=species e=species*treatment" is used to do hypothesis for species (whole-plot factor) and it implies "species*treatment" is the term used as error term (whole-plot error).

The MSE for GENE00001 sample is shown on fourth column of Table 2.4.1. The result matches what we got in Section 2.1. In the Source column, the results point out which mean squares we need for further analysis and testing. Note that SPECIES×TREATMENT is listed as a source, which means the F-ratio $\frac{MS_{Species}}{MS_{Specie*Treatment}}$ need to be computed for testing species.

However, the F-ratio is not valid statistically in default due to the special error term SPECIES×TREATMENT is used. It would be better to add RANDOM statement here to fix this default problem. After adding RANDOM statement, SAS would output the correct ANOVA table.

Table 2.4.1 SAS output of split-plot GENE00001 sample

Source	DF	Type III SS	Mean Square	F Value	Pr > F
species	1	0.0259	0.0259	0.81	0.5331
treatment	1	0.0002	0.0002	0.03	0.8882
Error	1	0.1501	0.0300		
species*treatment	1	0.0322	0.0322	3.43	0.1232
time	5	0.0638	0.0127	2.65	0.1539
species*time	5	0.1078	0.0215	1.22	0.4150
treatment*time	5	0.0550	0.0110	1.89	0.2505
Error: MS(Error)	5	0.0259	0.0259		

As the conclusion made from ANOVA of standard designs, F Values indicate significance of all listed sources.

Since the model contains fixed factors and random factor, it is a perfect mixed model. PROC MIXED is recommended. It can be done by some even simpler statements.

```
proc mixed;
class species treatment time;
model Y= species treatment species*treatment/ddfm=satterth;
random time species*time;
run;
```

In PROC MIXED procedure, only fixed terms are listed in MODEL statement and error terms go in the RANDOM statement. SAS don't need to be told which error term to be used for testing.

The specifics of a given design should determine whether to consider the blocking criterion as fixed or random, but no need here. Furthermore, more statements, such as CONTRAST or ESTIMATE, can be added appropriately to do further analysis as needed.

CHAPTER 3: SPLIT-PLOT APPLICATION IN GRAPEVINE EXPERIMENT

After discussing split-plot and its related designs, we can start to work on RAWDATA. In this chapter, we will use split-plot design to model the grapevine experiment, and then we will use an appropriate model expression to examine gene behavior with other statistical techniques.

Section 3.1 Experiment Description

Wild grapevines represent important genetic resources. During the past century, a broad range of wild bunch grape species (subgenus *Euvitis*) form North America and East Asia have been used to introgress genes for pest and pathogen resistance and environmental tress tolerance into cultivated scion and rootstock varieties (Alleweldt et al, 1990). Current trends toward reducing chemical input and relying more on biological-based disease resistance in grape cultivation makes the biological diversity of the wild grapevines even more relevant (Bisson et al, 2002). Understanding the molecular basis of this diversity will accelerate progress in harnessing the biological resources in *Vitis*. To our knowledge, genome-scale transcript level variation has not been examined in different *Vitis* species or different *Vitis vinifera* genotypes. An assessment of gene expression differences in various grapevine species will provide information about the role transcriptional regulation may play in phenotypic variation and adaptation. *Vitis* is highly heterozygous genus (Reisch and Pratt, 1996). At the molecular level, heterozygozity manifests itself in DNA sequence divergence among the different species and between haplotypes of V. vinifera, as evidenced by results from molecular marker-

based genotyping, and form sequencing of allelic variants for genes and bacterial artificial chromosome inserts (Aradhya et al, 2003; Salmaso et al, 2004; Adam-Blondon et al, 2005).

In present work, we conducted comparative mRNA abundance measurements in two grapevine genotypes, *V. aestivalis* Norton and *V. vinifera* Cabernet sauvignon. *V. aestivalis* Norton is a cultivated variety of North American origin. It is typical representative of North American grapevine species in that it is highly resistant to such economically important diseases as black rot, powdery mildew, and downy mildew, and is able to tolerate the insect pest phylloxera. *V. vinifera* Cabernet, on the other hand, is a Eurasian grapevine species which highly susceptibility to the above disease and pest. Furthermore, the two grapevines are distinctly different in a number of other features, including morphologically and physiological traits as well as fruit characteristics.

To test the tolerance of disease for both species, *V. vinifera* Cabernet sauvignon and *V. aestivalis* Norton, researchers raised two sets of plants in two separate growth chambers, controlling all conditions the same. In each set of plants, there are 60 *V. vinifera* Cabernet sauvignon and 60 *V. aestivalis* Norton. To observe the plants' behavior under different treatments, researchers applied inoculation with *E. necator* conidiospores to all plants raised in growth chamber A and mock inoculation was applied to those plants in growth chamber B. For each group containing 60 plants as we described above, they assigned numbers to each plants, then randomly chose 10 numbers without replacement and picked leaves from plants assigned with those 10 numbers to make a batch. An observation make on each batch yields a GeneChip. This procedure did six times at 0, 4, 8, 12, 24, 48 hours. On each chip, measurements for each of the 16436

genes were obtained. Here, each GeneChip generates a typical RNA microarray, and genes in the microarray are addressed by a unique identification number. We call them as gene ID'S. Throughout, CR1, CR2, and CR3 stands for three replicates of *V. vinifera* Cabernet sauvignon and NR1, NR2, and NR3 are those for the three replicates of *V. aestivalis* Norton. Therefore, researchers got 2×2×6×3 = 72 observations for each of 16436 genes. The data set for this experiment is denoted as RAWDATA.

Section 3.2 Preliminary Data Modifications

Normalization. For some data, we don't need to normalize the observations. Why do we perform normalization here? As what we mentioned in section 3.1, RAWDATA is a Microarray data. Microarray data are noisy due to the co-existence of genuine biological variations (signal) and noise (Wu 2005). Signal is what we desire to use to distinguish one sample from another, but noise can be raised by any step of experiment, which may hide useful information and mislead the further data analysis, somehow. To make downstream analysis more precise, we have to minimize the effect of noise.

The technique used to minimize noise here is normalization. There is variety of methods to normalize data. In RAWDATA, there are 72 samples due to the original experimental design of researchers. Here, we choose a two-step method to normalize data sample by sample. Firstly, using log2-transformaiton to decrease magnitudes of signals. Raw data may be extremely skewed. Doing log2-transformation, the magnitude of data decreases, which minimizes part of the noise and forces transformed data to be more normally distributed. For log2-transformation, $log_2(signal)$ is enough for most cases, here we prefer to modify it to be $log_2(signal + 1)$ since there are zeros in RAWDATA.

To check if this transformation method works or not, we pick all observations of the first biological replicate at time point 0 of both species, CS and N, separately, as samples. In Figure 3.1, the histograms of raw samples are displayed. The graphs are extremely skewed, which implies that the noise effect would be strong and outstanding if we use raw data to do further analysis.

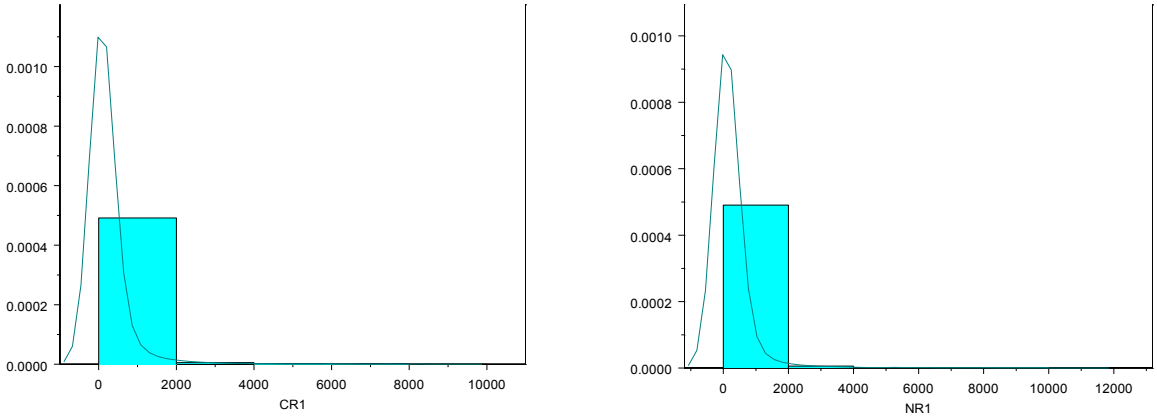

Figure 3.1 Histograms of signals in CR1 and NR1 observed at 0 hour

Figure 3.2 shows the histograms of sample data after doing log2-transformation. Even thoughthey are not perfectly bell-shaped, these two histograms are approximately normal and less skewed as what they were in Figure 3.1. By comparing Figure 3.1 and 3.2, the benefit of log2-transformation is obvious. On the other hand, noise contributes the residual portion, statically. Normal QQ-plot is usually used to identify if the transformed variable is approximately normal. The general rule is that if the graph is approximately a straight li ne, the data set can be approximately described as normally distributed; otherwise, the residual effect of this specific data set may be noisy. In the graph on the left, the data is heavy-tailed normally distributed, and right graph shows the data is light-tailed normally distributed, which implied the log2-transformation is actually

improve the data quantity for further analysis. Even log2-transformed data has already normalized the data and made them light-tailed normally distributed, it is not standard enough. We need second step here.

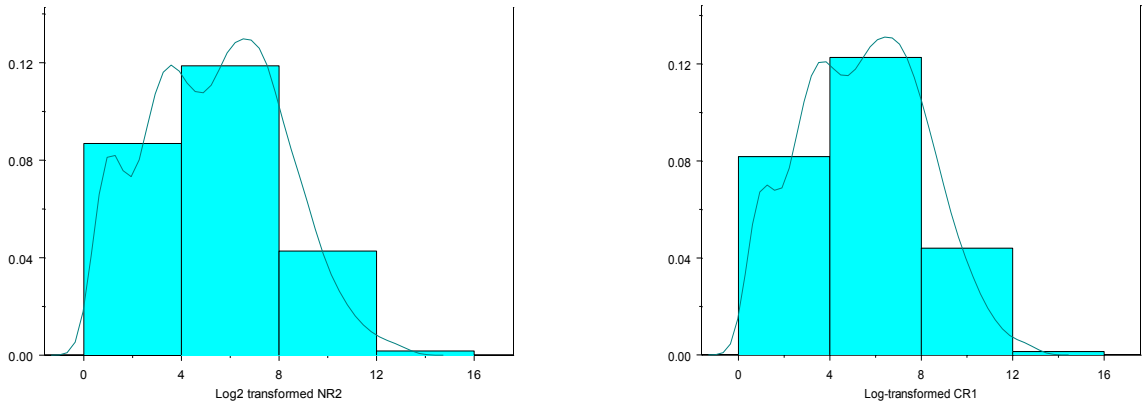

Figure 3.2 Histograms of signals in log2-transformed NR1 and in log2-transformed NR1 measured at 0 hour

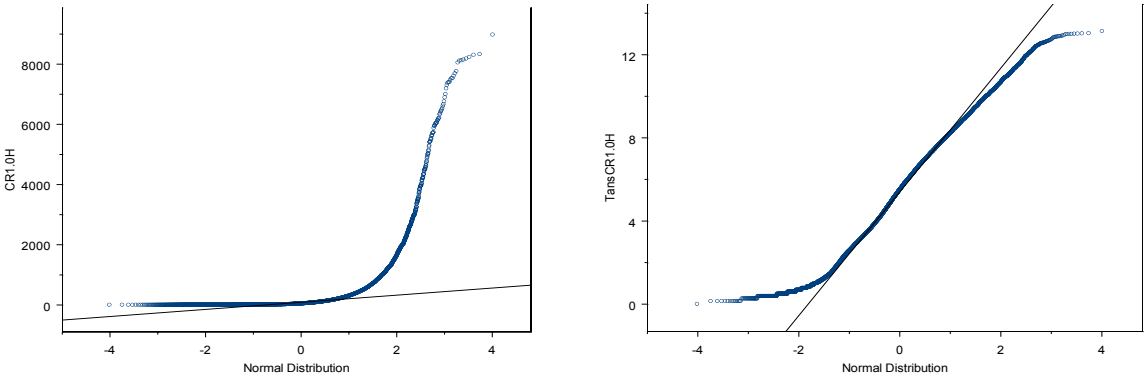

Figure 3.3 Before and after QQ-plots for CR1 at 0 hour with log2-transformation

Secondly, We can try to standardize this dataset again by calculating z-score, $\frac{transformed\ data - sample\ mean}{sample\ deviation}$, sample by sample for all 72 samples. The importance of this step is to make data more standard to do analysis and avoid useless noisy effects. Similarly, we can show the effects of this step by histograms and Normal Q-Q plots, as

what they are for the first step. With the help of this two-step method, the normalized data is marked as THESIS in Dataset section.

After normalizing thesis data, we can make a scatter matrix to check the correlation among six time points on same subject. Taking CR1 as an example, the results are shown in Figure 3.4. For any two time points, they are strongly correlated and there is not outstanding pattern in terms of time lags showing on the scatter matrix in Figure 3.4. This fact supports our general experiment modelling in Chapter 1 and the basic knowledge of correlation structure for split-plot design in Chapter 2.

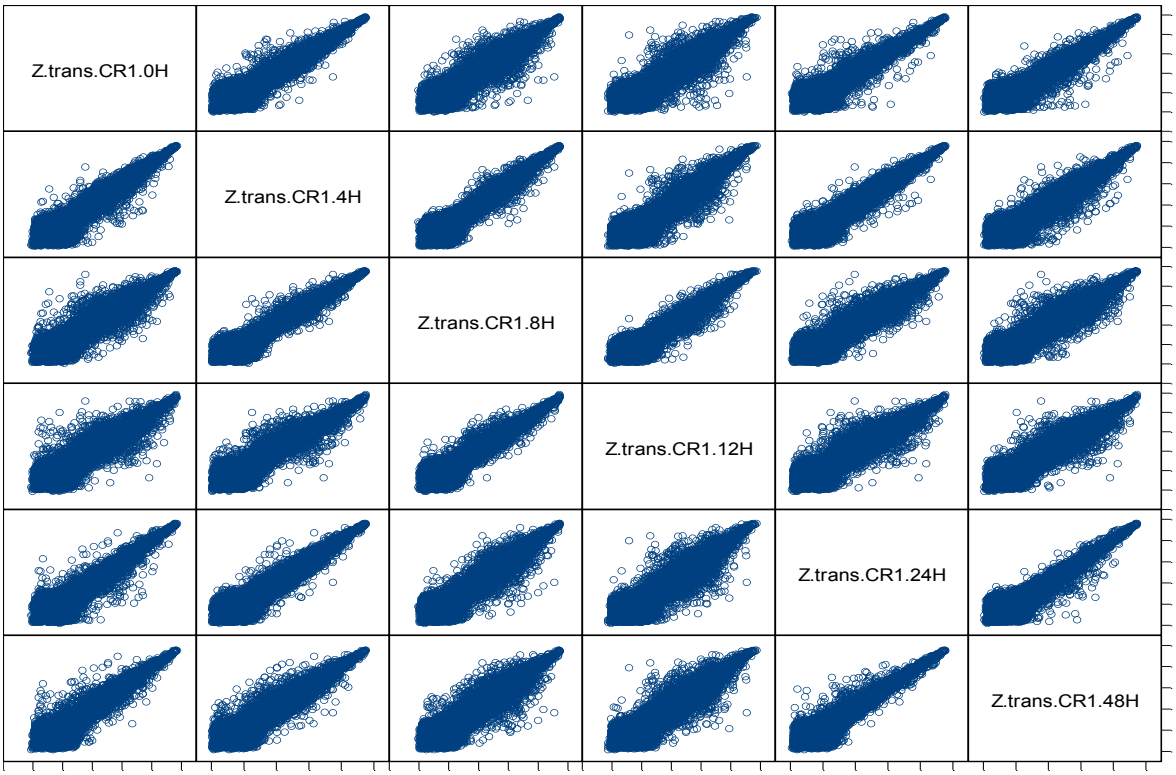

Figure 3.4 Scatter matrix for CR1 for all six time points

Reshaping. Notice that in NORMDATA, the data set follows a multivariate structure, which is not very convenient in many analyses. To minimize trouble in further

analysis, we will reshape the data set into univariate structure. Using PROC IMPORT statement, we can import our raw data into SAS Software and reshape it. In this procedure, some simple statements with their options may be used here, such as PROC TRANSPOSE, DATA, etc. The details and commands of statements will show up in Appendix V. The reshaped data will replace the previous dataset named as NORMDATA.

Section 3.3 General Modelling Procedure

This experiment includes factors SPECIES, TREATMENTS, TIMES, and REPLICATES. One thing we need to notice is that the replicates here are not real replicates as we knew for standard designs. These three replicates are sampled from the same biological subjects, which is called biological replicates. And again, the REPLICATES factor is nested within TIMES factor.

Considering the example in Chapter 2, we can apply similar philosophy in our experiment. All possible effects are main effects of SPECIES, TREATMENTS, and TIMES, and all their interactions, and also the nested effect of REPLICATES nested within TIMES effect. Mathematically, this model can be written as following

Model 3: $y_{ijtk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \theta_t + (\alpha\theta)_{it} + (\beta\theta)_{jt} + (\alpha\beta\theta)_{ijt} + \gamma_{k(ijt)} + e_{ijkt}$

where y_{ijtk} denote the signal for k^{th} measurement at t^{th} time point of i^{th} species with j^{th} treatment for any $i = 1, 2, j = 1,2, t = 1, \dots ,6, k = 1,2,3$; μ is the grand mean; α_i, β_j, $(\alpha\beta)_{ij}$, θ_t, $(\alpha\theta)_{it}$, $(\beta\theta)_{jt}$, $(\alpha\beta\theta)_{ijt}$ are main and interaction effects in terms of SPECIES, TREATMENTS and TIMES factors; $\gamma_{k(ijt)}$ represents the nested kth

REPLICATES effect; and e_{ijkt} is the random error terms following *i.i.d.* $N(0, \sigma^2)$. Other basic assumptions are the similar as those for the general model stated in Chapter 2.

Recall split-plot design we discussed in Chapter 2, we always identify whole-plot factor, split-plot factor, and error terms in each stage, respectively. It is reasonable to believe that, in this case, SPECIES and TREATMENTS are set up as a 2 factorial design, and that they are whole-plot factors, at the same time; and TIMES can be regarded as split-plot factor. The biological replicates are basically nothing but repeated measures here. We will discuss the reason later after we check Model 3.

Notice that, for each GeneChip, the degree of freedom on left and right sides of model are not equal, we can show this in Table 3.3.1. There is no doubt that the degree of freedom in right hand side is 71 since we have 72 signals for each GeneChip. On the right hand side, total degree of freedom except that for error term is 71, as well. Table 3.3.1 suggests that error term has been partitioned into those error terms in stages of split-plot design. Therefore, we can consider to delete the general error term and modify our assumption above. The modified model is

Model 4: $y_{ijtk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \theta_t + (\alpha\theta)_{it} + (\beta\theta)_{jt} + (\alpha\beta\theta)_{ijt} + \gamma_{k(ijt)}$

Now, let us import THESIS into SAS to check rest terms. In SAS, we can use both PROC GLM and PROC MIXED to deal with a certain split-plot design. Once correct codes are written down, results from both procedure are the same.

In this case, by using PROC GLM, we can write following statements:

```
Proc glm data=NORMDATA;
class ID SPE TR T REP;
model Y = SPE TR SPE*TR T SPE*T T*TR SPE*TR*T REP(SPE*TR*T)/SS3;
test h=TR E=SPE*TR;
random SPE*TR REP(SPE*TR*T)/test;
run;
```

Table 3.3.1 Degree of freedom for Model 3

Source in Left Side	Degree of Freedom	Sources in Right side	Degree of Freedom
		α_i	2-1
		β_j	2-1
		$(\alpha\beta)_{ij}$	(2-1)(2-1)
		θ_t	6-1
y_{ijtk}	72-1	$(\alpha\theta)_{it}$	(2-1)(5-1)
		$(\beta\theta)_{jt}$	(2-1)(5-1)
		$(\alpha\beta\theta)_{ijt}$	(2-1)(2-1)(5-1)
		$\gamma_{k(ijt)}$	2×2×6×(3-1)
		e_{ijkt}	?
Total	71	Total	71+?

The first output getting from these statements is the EMS table (Table 3.3.2) for each term in model, which is exactly same as what we did in Chapter 2 and in Appendix II. Noting that we only used the data of the first gene. We discussed about how to use EMS table to determine the stage errors. See the results in EMS column of Table 3.3.2, it is clear now that SPE*TR (interaction effect of SPECIES and TREATMENTS) is the whole plot error and REP (SPE*TR*T), REPLICATES effect nested within SPECIES, TREATMENTS and TIMES factors, is the split-plot error. One thing we need to pay special attention here is if we take all genes' observations into consideration the coefficients would be huge since SAS takes gene ID as a variable in default. This is not reasonable in experiment. If only those observations of one certain gene (such as GENE

00001) were concerned, the results would exactly the same as what we calculated in Chapter 2. And our assumption is whole-plot error follows *i.i.d.* $N(0, {\sigma_w}^2)$ and split-plot error follows *i.i.d.* $N(0, {\sigma_s}^2)$, and they are mutually independent.

Table 3.3.2 SAS output of EMS for model in Model 4

Source	Type III Expected Mean Square
SPE	Var(Error) + Var(REP(SPE*TR*T)) + 18 Var(SPE*TR) + Q(SPE,SPE*T SPE*TR*T,)
TR	Var(Error) + Var(REP(SPE*TR*T)) + 18 Var(SPE*TR) + Q(TR,TR*T,SPE*TR*T)
SPE*TR	Var(Error) + Var(REP(SPE*TR*T)) + 18 Var(SPE*TR) + Q(SPE*TR*T)
T	Var(Error) + Var(REP(SPE*TR*T)) + Q(T,SPE*T,TR*T,SPE*TR*T)
SPE*T	Var(Error) + Var(REP(SPE*TR*T)) + Q(SPE*T,SPE*TR*T)
TR*T	Var(Error) + Var(REP(SPE*TR*T)) + Q(TR*T,SPE*TR*T)
REP(SPE*TR*T)	Var(Error) + Var(REP(SPE*TR*T))

Another output is the ANOVA table (Table 3.3.3), which directly implies that all terms of model 4 are significant, since all p-values are extremely small.

When using PROC GLM for split-plot design, we have noticed that only one error term (usually, the error in the very last stage) is used to calculate all F values in the default, so only those F values of effects in very last stage is correct in default results, for instance, T, SPE*T, and TR*T in Table 3.3.3. We need to tell SAS which error term we

want to use when calculating F value for specific effect. TEST option here is add to specify which one to do hypothesis testing and which stage error to use.

Table 3.3.3 ANOVA for model in model 4 by using PROC GLM

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SPE	1	16517.86240	16517.86240	1868.59	<.0001
TR	1	350.21838	350.21838	39.62	<.0001
SPE*TR	1	585.31561	585.31561	66.21	<.0001
T	5	991.62840	198.32568	22.44	<.0001
SPE*T	5	171.16638	34.23328	3.87	0.0016
TR*T	5	218.52642	43.70528	4.94	0.0002
REP(SPE*TR*T)	53	6225.59124	117.46399	13.29	<.0001

Noting that random term, REP (SPE*TR*T) here, shows neither in MODEL statement nor in RANDOM statement. In this experiment, we can write code as below in SAS and output shows in Table 3.3.4.

```
Proc mixed data=NORMDATA method=ml;
class ID SPE TR T REP;
model Y = SPE TR SPE*TR T SPE*T T*TR;
random SPE*TR;
run;
```

As we said above, the results should be the same in both procedures. Even though some p-values in PROC MIXED procedure don't show up appropriately and some p-values in PROC GLM is not even correct in default, we can still see the results for T, SPE*T, and TR*T are the same, and we have ensured these values are correct in default.

That is enough to prove both procedure function well and the model in model 4 is appropriate tested by our dataset.

Table 3.3.4 ANOVA table for model in model 4 by using PROC MIXED

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
SPE	1	0	1867.46	.
TR	1	0	39.60	.
SPE*TR	1	0	66.18	.
T	5	12E5	22.42	<.0001
SPE*T	5	12E5	3.87	0.0016
TR*T	5	12E5	4.94	0.0002

Section 3.4 Computation of Gene-Specific Significance Models

After finishing general modelling, we need to focus on the main purpose of our experiment, which is to test which gene(s) expression is/are significantly different from others in terms of TREATMENTS. In order to reach this goal, we need to apply our model in model 4 to each GeneChip and observe the results one by one. First of all, we need to sort our data NORMDATA in SAS to perform mixed model ANOVA for individual genes. In SAS, SORT procedure can be applied here.

```
Proc sort data=NORMDATA;
by ID SPE TR T REP;
run;
```

Under this procedure, our data NORMDATA are sorted by ID, then by SPECIES, then by TREATMENTS, etc. For our purpose, only ID one really matters, as this ensures that all of the values for each clone are grouped into contiguous block (Gibson and Wolfinger 2004). And then we can perform gene-specific ANOVA. Typing the following text into the editor pane of SAS and submit it:

```
ods exclude all;
ods noresults;
proc mixed data = NORMDATA;
by ID;
class ID SPE TR T REP;
model Y=SPE TR SPE*TR T SPE*T T*TR SPE*TR*T/ outp= NormData1;
random REP(SPE*TR*T);
lsmeans SPE TR SPE*TR T SPE*T T*TR SPE*TR*T/diff;
ods output covparms=COVparms tests3=Tests3 lsmeans=lsms diffs=Diffs;
run;
ods exclude none;
ods results;
```

By running this set of statements, SAS would automatically apply model to 16436 genes one by one and provide ANOVA table for each Gene ID in Test3 results. Test3 results are shown in Table 3.4.1. In this set of ANOVA tables, only items we concern about are the p-values for TREATMENTS effect, because it implies the significance of TREATMENTS effect. Instead of showing the huge amount of results here, we pick the results of 10 genes in terms of the p-values of TREATMENTS effect to explain.

In column ProbF, we can notice that values for all genes except first and fourth ones are relatively large, over 0.1. For first gene, p-value = 0.0201, which implies this gene expresses differently under treatments if we take 0.05 as level of significance. Similarly, for the fourth gene, it p-value is 0.0677. It is a little bit higher than 0.05, but it is possible to consider it is differently expressed if a higher biological significant level is chosen. On the other hand, for the rest of eight genes, their p-value is considered to be

large in most cases, statistically. We can conclude that their TREATMENTS effects are not significant, and furthermore, their gene expressions don't change much under different treatments in this experiment. Depending on the choice of significance levels, gene expression can be easily determined from the results as Table 3.3.1. This analysis can be generalized to all 16436 genes to check if their TREATMENTS effects are significant or not. If yes, we can pay special attention to it in further research to explore their special properties.

Table 3.4.1 Brief ANOVA for first 10 genes in terms of TREATMENTS effect

Obs	ID	Effect	NumDF	DenDF	FValue	ProbF
1	GENE00001	TR	1	48	5.78	0.0201
2	GENE00002	TR	1	48	0.09	0.7659
3	GENE00003	TR	1	48	1.61	0.2107
4	GENE00004	TR	1	48	3.50	0.0677
5	GENE00005	TR	1	48	0.21	0.6461
6	GENE00006	TR	1	48	0.32	0.5733
7	GENE00007	TR	1	48	0.92	0.3418
8	GENE00008	TR	1	48	2.49	0.1211
9	GENE00009	TR	1	48	0.10	0.7537
10	GENE00010	TR	1	48	2.49	0.1213

Section 3.5 Further Concern

In previous sections, we did analysis by modelling data as split-plot. The results are quite clear in procedures; however, we also noticed that there are many potential properties ignored when using split-plot. For example, the split-plot design never considers the possible reason why a researcher did measurements at six different time

points and the possible biological pattern implied by measurements. Actually, we can do some brief analysis based on time. Recall that in Figure 3.4, it is no problem that correlation between any two time points are extremely high; however, we may have already noticed that the plot is slightly spread out as the time difference increases at first four time points, and that the shape of scatter plots are similar among 0 hour, 24 hour, and 48 hours. Those can be concluded as an obscure pattern saying that the correlation decreases slightly with the increasing of time lag among the first three time points and that the correlation among 0 hour, 24 hour, and 48 hours are relatively stable. This pattern can be clearly displayed by covariance correlation table. Randomly pick all observations of CS with MOC applied, the correlation table is given by S-plus in Table 3.5.1.

Table 3.5.1 Correlation table for CS under treatment MOC at six time points

	0 hour	4 hours	8 hours	12 hours	24 hours	48 hours
0 hour	1.0000000					
4 hours	0.9787216	1.0000000				
8 hours	0.9397674	0.9720687	1.0000000			
12 hours	0.9308847	0.9418207	0.9682751	1.0000000		
24 hours	0.9667132	0.9626963	0.9230444	0.9029759	1.0000000	
48 hours	0.9462432	0.9430237	0.9010002	0.8883521	0.9822352	1.0000000

Due to the property of correlation between two variables, values in cells are symmetric. For instance, correlation between 4 hours and 24 hours is exactly same as what it is for 24 hours and 4 hours. In the second column (0 hour column) of Table 3.4.1, we can see the correlation declines in time interval [0, 12], from 1.0000000 to 0.9308847;

at the same time, corr(0 hour, 24 hours) raised up to 0.9667132 and corr(0 hour, 48 hours) = 0.9462432. This fact may imply a cycling biological pattern, such as the diurnal cycle.

In split-plot design, this special pattern is totally ignored, since split-plot design has a symmetric component assumption about measurements made from same subjects. To make full use of this new potential information, more precise models, such as repeated-measures design, can be considered if current results could not satisfy the researchers' requirements. Briefly, the repeated measurement design provides more flexibility on component assumption on the measurements made from the same experiment units. Next, we will spend a chapter to discuss repeated measurement experiment design.

CHAPTER 4: REPEATED-MEASURES DESIGN

In the previous chapter, we discussed Split-plot Design. There is another commonly used experimental design with some special property and advantages. It is the Repeated Measurements Design. In this chapter, we will introduce knowledge related to Repeated Measurements Design and compare it with Split-plot Design.

Section 4.1: Background

In recent decades, researches have been expanding from natural sciences to social and behavioral sciences. People are frequently chosen to be subjects of research. A good example is doing survey by questionnaire.

The researchers design different questions for different group of people based on their different occupations, regions, levels of education, and so on. And it is rarely that they pick a single person in each group to finish their survey because it would waste the effort of researcher developer. Multiple people would be randomly picked. This kind of research design implies their potential concerns, which are people in same group share some common parts and also the difference among people in same group can make the research more precise.

The example above demonstrates that selecting multiple people may be useful to improve research somehow. But how does this kind of design help? Doing survey can be looked as a kind of experiment. And then, we can review this survey statistically. Different versions of questionnaires are similar to different kinds of treatments applied in experiment. We will show details in following sections in this chapter.

Section 4.2 Statistical Introduction

In the example of last section, the group we described in last section are typically called as experimental units, or subjects in statistics. Researchers apply each treatment to each subject and get multiple observations (different results from different people in same group) for each subject. Such a design is called a Repeated Measures Design (Montgomery, 1997). Repeated measures refers to multiple measures on the same experimental unit. Usually, repeated measures are made over time, but they can be over space. A very common situation is for treatments to be applied to experimental units in a completely randomized design. Then measurements are made at several different times (Littell, et al. 2002). Let's consider the measurements made in a sequence of time point. Those observations get from same subjects should be related and dependent since they share some common properties from the subject. And also, observations which are measured closer in time should have a stronger impact than those get measured far away in time. These abstract thinking can be concluded as the variance and covariance structure, which is exactly the special property for Repeated measures design.

Section 4.3 Model

We mentioned variance and covariance structure in previous section, here we will explain it more specifically with example and its model. For instance, a medical institution plans to examine the effects of n drugs for insomnia patients. *n* drugs are assigned to n groups each randomly. Researchers then measure the time taking to fall asleep daily in following *t* successive days for patients with the help of each kind of drugs.

The experiment units, or subjects, are patients here. Drugs are considered to be treatments. It's clear that the repeated measures are taking by times. The effects we can conclude here are DRUG, TIME, DRUG*TIME. The model for this example is

$$y_{ijk} = \mu + \alpha_i + \beta_k + (\alpha\beta)_{ik} + e_{ijk}$$

where

μ is the overall mean; α_i, β_j, $(\alpha\beta)_{ij}$ are i^{th} drug effect, k^{th} time effect and their interaction effect, respectively; e_{ijk}'s are random error for j^{th} patient taking i^{th} drug at k^{th} day.

Directly from the form of model, two questions are raised:

1. Different patients should make difference; however, no main or even interaction effect involving patient effect. How could this model explain the variability of different patient?
2. This model is exactly the same with certain kind of 2 factorial design. Why do we say this model works for Repeated measures design?

These two question can be answered by the special properties of error term e_{ijk}. These properties are called variance and covariance structure of e_{ijk}. Since time factor is not randomly assigned, we cannot assume error terms are independent generally as what we do for other standard design, or even not as Split-plot design.

Section 4.4 Variance and Covariance Structure

As what we said in section 2 of this chapter, measurements made on same subjects share common property from subject and measurements influent others stronger when they are closer in time, in space, or in condition. Formally speaking, those two conclusions are two aspects of covariance structure in errors:

1. Measurements made from same subject are stronger, positively in most cases, correlated than those made from different subjects. (Sometimes, this aspect is called as between-subject variation)

2. Measurements made closer in time, in space, or under conditions are highly correlated than those made far apart. Correlation can follow various structure, it is not always linear.

Due to the importance of covariance structure for repeated measures design, this structure has to be considered and specified in doing statistical analysis technique. Many methods have been developed to deal with covariance structure for repeated measures design. We introduce three general methods.

Univariate ANOVA. This approach is also known as split-plot in time ANOVA. Why so? Univariate ANOVA treats repeated measures data same as split-plot data. Subjects and time factors are matching with whole plot units and split plot factors, which means only the first aspect of covariance structure is considered in analysis. It's not always appropriate for repeated measures data, but if correlations between measures on the same subject are the same regardless of time proximity, then this approach is a perfectly good method of analysis (Littell, et al. 2006).

Analysis of Contrasts. Instead of doing analysis on data directly, analysis of contrasts analyzes linear combinations of data on each subject. In this approach, we introduce regression concept by considering measurements on same subject are regressed on time. The slopes of these regression line are treated as the effects of time on each subject. By doing analysis of these slopes, we can analyze the effects of treatments on these time effects. Since the covariance structure is not considered here, this approach is not optimal.

Mixed-model Methodology. This approach contains two steps. The first step is estimating covariance structure. Second is substituting estimated covariance structure into

the model. And then we can use (generalized) least-squared method to assess treatment and time effects. However, estimation of covariance structure is not an easy job to do. With the development of statistical programs, this complex estimation can be done by computer software, for instance, PROC MIXED in SAS.

For the convenience of writing down, matrix forms are introduced in almost all aspects of statistics. Matrix form is applied to statistical models. Details are shown in Appendix III. In terms of matrix form, modelling covariance structure is finding the form of **G** and $\boldsymbol{R_{ij}}$, portions of **R** that corresponds to an individual subject. There are many candidate covariance structures to choose from.

Simple covariance structure is as the name implies, which is the simplest structure. It assumes all measurements (observations) are independent no matter they are made on same subjects or not, and variance of measurements are homogenous. Then the correlation function is 0 in this case. Backing to example in section 3, simple covariance structure as

$$cov(y_{ijk}, y_{ijl}) = 0 \; if \; k \neq l \; and \; var(y_{ijk}) = {\sigma_{SIM}}^2$$

Recall the matrix form in Appendix III,$\mathbf{G} = \mathbf{0}$ and $\boldsymbol{R_{ij}} = {\sigma_{SIM}}^2 \boldsymbol{I}$, where **I** is an identity matrix.

Notice that, under simple structure, there is an independence assumption, which leads this structure not realistic for repeated measures data most of time.

Compound symmetric structure specifies repeated measurements on same subject share same covariance and each has same variance, representing by ${\sigma_{CS,b}}^2$ and ${\sigma_{CS,b}}^2 + {\sigma_{CS,w}}^2$, respectively. Because of this kind of representation, compound symmetric structure is also called variance components structure. The correlation function

is$\frac{{\sigma_{CS,b}}^2}{{\sigma_{CS,b}}^2+{\sigma_{CS,w}}^2}$. Correlation here is free of time difference and it is nonnegative, in other words, any pair of measurements on same subjects should have same nonnegative correlation. In matrix form, it can be shown as

$$\mathbf{G} = {\sigma_{CS,b}}^2 \boldsymbol{I} \; and \; \boldsymbol{R_{ij}} = {\sigma_{CS,w}}^2 \boldsymbol{I}, or$$

$$\mathbf{G} = \mathbf{0} \; and \; \boldsymbol{R_{ij}} = {\sigma_{CS,w}}^2 \boldsymbol{I} + {\sigma_{CS,b}}^2 \boldsymbol{J}$$

Depend on certain cases, people can choose why to define their matrices.

Similar to the definition of AR(1) in time series, this covariance structure assumes all measurements have the same variance and the covariance between measurements on same subjects has a regressive relation; specifically speaking, covariance decreases with the increasing of time difference.

$$cov(y_{ijk}, y_{ijl}) = {\sigma_{AR(1)}}^2 {\rho_{AR(1)}}^{|k-l|}, and$$

$$\mathbf{G} = \mathbf{0} \; and \; \boldsymbol{R_{ij}} = {\sigma_{AR(1)}}^2 {\rho_{AR(1)}}^{|k-l|}$$

Consider the second aspect of the covariance structure for repeated measures design, AR(1) meets the requirement much better than first two, and it, perhaps, is the most commonly used structure.

The only difference between this structure and last one is variance partitions into within-subject and between-subjects due to the random effect adding in to subjects. So the variance and covariance now look as following

$$var(y_{ijk}) = {\sigma_{\mathrm{AR(1)+RE,b}}}^2 + {\sigma_{\mathrm{AR(1)+RE,w}}}^2 \; and$$

$$corr_{\mathrm{AR(1)+RE}} = \frac{{\sigma_{\mathrm{AR(1)+RE,b}}}^2 + {\sigma_{\mathrm{AR(1)+RE,w}}}^2 {\rho_{AR(1)+RE}}^{|k-l|}}{{\sigma_{\mathrm{AR(1)+RE,b}}}^2 + {\sigma_{\mathrm{AR(1)+RE,w}}}^2}$$

In terms of matrix, $\mathbf{G} = {\sigma_{\mathrm{AR(1)+RE,b}}}^2 \boldsymbol{I} \; and \; \boldsymbol{R_{ij}} = {\sigma_{AR(1)}}^2 {\rho_{AR(1)+RE}}^{|k-l|}$

Toeplitz structure, sometimes called 'banded', specifies that covariance depends only on time difference, but not as a mathematical function with a smaller number of parameters. The correlation function is $\frac{\sigma_{TOEP,|diff|}}{{\sigma_{TOEP}}^2}$ and $\mathbf{G} = \mathbf{0}$ and $\mathbf{R}$ is a matrix with ${\sigma_{TOEP}}^2$ on the main diagonal and $\sigma_{TOEP,|diff|}$ on a sub-diagonal $|k - l|$, where k is the row number and l is the column number. (Littell, et al., 2000)

The unstructured structure specifies no patterns generally. It needs a very large amount of parameter and requires a large data set to make result precisely.

AR (1) and TOEP are both widely used handing covariance structure issues for repeated measures data. However, they have an outstanding limit; that is, both those structures make sense when time points (or spaces) are equally spaced.

In reality, time points are not always evenly assigned in experiment. We need covariance structure for this general case. Ante-dependence model with order 1 can handle this problem. In this case, R is a matrix with ${\sigma_k}^2$in main diagonal and $\sigma_1\sigma_k\prod_{i=1}^{k-1}\rho_i$ on off-diagonal, where k is the number of measurements.

There are many other covariance structures useful in unequal spaced cases. The most common of these are SP (POW) (spatial power law), SP (GAU) (gaussian), and SP (SPH) (spherical). (Littell, et al. 2000)

For example, SP (POW) provides a generalization of the AR (1) structure for unequal spaced data. It produces $\mathrm{cov}(y_{ijk} - y_{ijl}) = {\sigma_{SP(POW)}}^2\rho^{|k-l|}$ and ρ is an autoregressive parameter with $|\rho| < 1$.

Section 4.5 Advantages and Disadvantages of Repeated-measures Design

A repeated measures design is using the same participants for all of your experimental conditions (Field, 2011). This is in contrast to other standard designs. A benefit of using repeated measures (using the same participants for both manipulations) is it allows the researcher to exclude the effects of individual differences that could occur if two different people were used instead (Howitt & Cramer, 2011). Each subject serves as own control so that between-subject effects get isolated, which makes the analysis focus on treatment effects more precisely.

Recall what we said in section 1 of this chapter, repeated measures designs ask fewer subjects than other related independent groups designs; this fact makes repeated measures designs more efficient and convenient. Apart from all advantages above, repeated measures design can detect the effects of all independent variable, no matter they are significant or relatively small, so we can say repeated measures is more sensitive than other standard designs.

Repeated measures designs also have their limits. Generally, these limits can be conclude as following three aspects:

1. Carry-over Effect. It happens when a treatment is applied before the effects of previous one has worn off. So researchers need sufficient long time between treatments.

2. Latent Effect. If two treatments were relatively related, the later one may activate the effects of previous treatment. This can mislead the results and conclusions of experiment.

3. Learning Effect. Since this kind of design use same subject for all conditions (treatments), subject can improve, or get worse, at each time of measurements, especially when subjects are people. This effect can change the conclusions made for treatments.

Section 4.6 Repeated Measures Designs and Split-plot Designs

In section 4, we notice that the first aspect of covariance structure is the similarity to the assumption made for measurements on same whole plot in Split-plot design, which implies Split-plot design and repeated measures design are comparable. The only feature to make them different is the second aspect of covariance. In repeated measures design, we take repeated measurements over time for each subject. These subjects can be looked on as whole plots. So if all these repeated measurements are equally correlated over time, then time effect is nothing but a sub-plot effect. Actually, this case is following the Univariate ANOVA approach in Section 4.

The ANOVA table for a split-plot design can be used as an approximation ANOVA to its related repeated measurement. The thing we need to do is we need to modify the degrees of freedom of repeated factors and interactions involving them. This new degree of freedoms are called Conservative degree of freedom. The specific method is replace the degrees of freedom of repeated factors by 1's. For instance, consider Table 2 in Chapter 1. If we use it as the approximate ANOVA table for related repeated measures design. Here β_k can be considered as the repeated factor effect for its related repeated measures design. W can see conservative d.f.'s as what are shown in Table 4.5.1.

And then we can use Conservative d.f. column in Table 4.5.1 to do tests for effects. If these two designs give different results, modifying degree of freedom, using REPEATED statement, or testing assumptions can be used to solve some cases, but not all. This issue is beyond the scope here. However, if repeated measurements are significant when conservative d.f. are used, conclusions from both designs are sound.

Table 4.5.1 Approximate ANOVA of repeated measurements analysis

	Sources	*d.f.*	*Conservative d.f.*
	α_i	m-1	m-1
Whole plot	θ_j	n-1	n-1
	$(\alpha\theta)_{ij}$	(m-1)(n-1)	(m-1)(n-1)
	β_k	b-1	1
	$(\alpha\beta)_{ik}$	(m-1)(b-1)	m-1
Split plot	$(\theta\beta)_{jk}$	(n-1)(b-1)	n-1
	$(\alpha\theta\beta)_{ijk}$	(m-1)(n-1)(b-1)	(m-1)(n-1)
	ε_{ijk}		

CHAPTER 5: SUMMARY

There is no simple answer to say which design fits the experiment best. Based on different biological requirements, different designs would be chosen. Split-plot design is convenient in having to grow only a limited number of plants and saves time and funds, and it can satisfy most research goals. However, repeated-measures design provides flexibility of component structure, which gives us a chance to taking covariance between time points into account and supports our cyclic pattern assumption. There must be more designs can fit the data and experiment better, so the research based on this set of data will last long.

REFERENCE

Adam-Blondon, A.F., S. Aubourg, et al. (2005) Updated of Our Knowledge on the Structure of the Grapevine Genome. *International Grape Genomics Symposium*. Saint Louis, Missouri, USA, Missouri State University.

Aradhya, M.K., G.S. Dangl, el al. (2003). Genetic Structure and Differentiation in Cultivated Grape *Vitis vinifera L*. *Genetical Research*. **81(3)**: 179-192.

Alleweldt, G., P. Spiegel-Roy, et al. (1990). Grapes (Vitis). *Genetic Resources of Temperate Fruit and Nut Crops*. J.N.Moore and J.R.Ballington, ISHS. 290:291-337.

Bisson, L.F., A.L. Waterhouse, et al. (2002). The Present and Future of the International Wine Industry. Nature. **418**: 696-699.

Field, A. (2011). *Discovering Statistics Using SPSS, Third Edition*. (pp. 15-18). Thousand Oaks, CA: SAGE Publication.

Fisher, R. A. (1966). *The Design of Experiments, Eighth Edition*. Hafner Publishing Company. New York.

Gibson, G., Russell D. Wolfinger. (2004). Gene Expression Profiling using Mixed Models. A. Saxton, Ed. *Quantitative Genetic Analysis with SAS Software*. SAS Institute.

Hicks, Charles R., Kenneth V. Turner, Jr. (1999). *Fundamental Concepts in the Design of Experiments, Fifth Edition*. New York, NY: Oxford University Press, Inc.

Howitt, D., Cramer, D. (2011). *Introduction to Research Methods in Psychology*, *Third Edition*. (pp. 164, 179-181). Harlow, Essex: Pearson Education Limited.

Littell, Ramon C., George A. Milliken, Walter W. Stroup, Russell D. Wolfinger, Oliver Schabenberger. (2006). *SAS® for Mixed Models, Second Edition*. Cary, NC: SAS Institute Inc.

Littell, Ramon C., Jane Pendergast, Ranjini Natarajan. (2000). TUTORIAL IN BIOSTATISTICAS: Modelling covariance structure in the analysis of repeated measures data. *Statist. Med.* **19**:1793-1819.

Littell, Ramon C., Walter W. Stroup, Rudolf J. Freund. (2002). *SAS® for Linear Models, Fourth Edition*. Cary, NC: SAS Institute Inc.

Montgomery, Douglas C. (1997). *Design and Analysis of Experiments, Eighth Edition*. Hoboken, NJ: John Wiley & Sons, Inc.

Reisch, B. J., C. Pratt. Grapes. Fruit Breeding. (1996). J. Janick and J. N. Moore. New York. Wiley and Sons. **2**: 297-370.

Salmaso, M., G. Faes, et al. (2004). Genome Divisity and Gene Haplotypes in the Grapevine (Vitis vinifera L.), as reserved by single nucleotide polymorphisms. *Molecular Breeding*. **14**: 385-395.

Wu, W., Dave, N., Tseng, George C., Richards, T., Xing, Eric P., Kaminski, N. (2005). Comparison of Normalization Methods for CodeLink Bioarray. *BMC Bioinformatics*. **6**: 309.

GLOSSARY

Analysis of variance (ANOVA): a procedure for constructing statistical tests by partitioning the total variance into different sources.

Biological replicates: biological samples obtained in replicate form independent sources representing the same condition.

Decomposition: separation of a complex variance term in an ANOVA model into its components, which is attributable to all effects and their interactions.

Degree of freedom: the number of levels that can vary freely in a term of an ANOVA model. It is typically one less than the number of levels in the factor.

Error variation: the variation associated with an estimated quantity. It is the square of the standard error and in commonly used to assess the accuracy of estimation.

Fixed effect: a term in an ANOVA model for which the levels are going to be repeated exactly if the experiment is repeated. We are generally interested in the mean values associated with levels of a fixed effect.

Gene expression profiling: the monitoring of differences in the level of expression of thousands of individual genes across a series of treatments.

Mixed-model ANOVA: an ANOVA model in which some terms are treated as random effects and others as fixed effects. In a mixed model there may be multiple sources of random variation.

Normalization: the process of removing certain systematic biased form microarray data.

Null hypothesis: a hypothesis for which the effects of interest are assumed to be absent. Commonly used as a basis for constructing statistical tests.

Power: the probability that a real effect can be identified by a statistical test. It is one minus the type II error probability.

P-value: a measure of the evidence against the null hypothesis in a statistical test. It is the probability of the occurrence of a test statistic equal to, or more extreme than, the observed value under the assumption that the null hypothesis is true.

Random effect: a term in ANOVA model for which the levels represent a sample from a population of levels. In a replicated experiment the same values will not repeat. We are generally interested in the variability associated with a random effect.

Residual: the difference between an observed data value and its expectation as predicted by a model. It is the lowest-level term in ANOVA model, denoted as e_i.

Residual sums of squares: the sum of all the residuals squared. It is a measure of the total discrepancy between a model and the observed data.

Restricted maximum likelihood: a numerical method for estimating variance components in a mixed ANOVA model.

Significance level: the size of p-value that is regarded as providing sufficient evidence against a null hypothesis. If the p-value falls below the significance level, the null hypothesis is rejected.

Technical replicates: multiple RNA samples obtained from the same biological source.

Type I error: the event of rejecting a null hypothesis when it is true.

Type II error: the event of failing to reject a null hypothesis when it is false.

APPENDICES

Appendix I Deriving Sum of Squares

Step 1: Decomposition

$$y_{ijk} - \overline{y_{...}} = (\overline{y_{i..}} - \overline{y_{...}}) + (\overline{y_{.j.}} - \overline{y_{...}}) + (\overline{y_{ij.}} - \overline{y_{i..}} - \overline{y_{.j.}} + \overline{y_{...}}) + (\overline{y_{..k}} - \overline{y_{...}})$$
$$+ (\overline{y_{i.k}} - \overline{y_{i..}} - \overline{y_{..k}} + \overline{y_{...}}) + (\overline{y_{.jk}} - \overline{y_{.j.}} - \overline{y_{..k}} + \overline{y_{...}}) + (y_{ijk} - \overline{y_{ij.}} - \overline{y_{i.k}}$$
$$- \overline{y_{.jk}} + \overline{y_{i..}} + \overline{y_{.j.}} + \overline{y_{..k}} + \overline{y_{...}})$$

When doing decomposition for each term in model, we need to check if left and right sides are equal after cancellation. Left and right sides are equal for our case. At the same time, left side is the estimate of μ, and portions in right side are estimates for α_i , θ_j , $(\alpha\theta)_{ij}$, β_k, $(\alpha\beta)_{ik}$, $(\theta\beta)_{jk}$, $(\alpha\theta\beta)_{ijk}$, respectively.

Step 2: Adding Summation Symbol for Both Sides

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(y_{ijk} - \overline{y_{...}})^2$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}[(\overline{y_{i..}} - \overline{y_{...}}) + (\overline{y_{.j.}} - \overline{y_{...}}) + (\overline{y_{ij.}} - \overline{y_{i..}} - \overline{y_{.j.}} + \overline{y_{...}})$$
$$+ (\overline{y_{..k}} - \overline{y_{...}}) + (\overline{y_{i.k}} - \overline{y_{i..}} - \overline{y_{..k}} + \overline{y_{...}}) + (\overline{y_{.jk}} - \overline{y_{.j.}} - \overline{y_{..k}} + \overline{y_{...}})$$
$$+ (y_{ijk} - \overline{y_{ij.}} - \overline{y_{i.k}} - \overline{y_{.jk}} + \overline{y_{i..}} + \overline{y_{.j.}} + \overline{y_{..k}} + \overline{y_{...}})]^2$$

Sum of squares on both sides are equal naturally since roots for both sides equals from step 1.

Step 3: Simplify right side

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}\Big[(\overline{y_{i..}}-\overline{y_{...}})+(\overline{y_{.j.}}-\overline{y_{...}})+(\overline{y_{ij.}}-\overline{y_{i..}}-\overline{y_{.j.}}+\overline{y_{...}})+(\overline{y_{..k}}-\overline{y_{...}})$$

$$+(\overline{y_{i.k}}-\overline{y_{i..}}-\overline{y_{..k}}+\overline{y_{...}})+(\overline{y_{.jk}}-\overline{y_{.j.}}-\overline{y_{..k}}+\overline{y_{...}})$$

$$+(y_{ijk}-\overline{y_{ij.}}-\overline{y_{i.k}}-\overline{y_{.jk}}+\overline{y_{i..}}+\overline{y_{.j.}}+\overline{y_{..k}}+\overline{y_{...}})\Big]^2$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{i..}}-\overline{y_{...}})^2+\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{.j.}}-\overline{y_{...}})^2+\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{ij.}}-\overline{y_{i..}}-\overline{y_{.j.}}+\overline{y_{...}})^2$$

$$+\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{..k}}-\overline{y_{...}})^2+\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{i.k}}-\overline{y_{i..}}-\overline{y_{..k}}+\overline{y_{...}})^2$$

$$+\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{.jk}}-\overline{y_{.j.}}-\overline{y_{..k}}+\overline{y_{...}})^2$$

$$+\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(y_{ijk}-\overline{y_{ij.}}-\overline{y_{i.k}}-\overline{y_{.jk}}+\overline{y_{i..}}+\overline{y_{.j.}}+\overline{y_{..k}}+\overline{y_{...}})^2$$

All interaction terms are equal to zero based on our assumption for model. Taking first two term of right side in step 1 as an example.

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}\Big((\overline{y_{i..}}-\overline{y_{...}})+(\overline{y_{.j.}}-\overline{y_{...}})\Big)^2$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{i..}}-\overline{y_{...}})^2+\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{i..}}-\overline{y_{...}})(\overline{y_{.j.}}-\overline{y_{...}})$$

$$+\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{.j.}}-\overline{y_{...}})^2$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{i..}}-\overline{y_{...}})^2+\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\alpha_i)(\theta_j)+\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{.j.}}-\overline{y_{...}})^2$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{i..}} - \overline{y_{...}})^2 + b\sum_{i=1}^{m}\alpha_i\sum_{j=1}^{n}\theta_j + \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}\left(\overline{y_{.j.}} - \overline{y_{...}}\right)^2$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{i..}} - \overline{y_{...}})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}\left(\overline{y_{.j.}} - \overline{y_{...}}\right)^2$$

$$= nb\sum_{i=1}^{m}(\overline{y_{i..}} - \overline{y_{...}})^2 + mb\sum_{j=1}^{n}(\overline{y_{.j.}} - \overline{y_{...}})^2$$

For rest terms on right side, similar way and assumption can be applied to. After simplification, we can get following result for right side of equation.

Right side

$$= nb\sum_{i=1}^{m}(\overline{y_{i..}} - \overline{y_{...}})^2 + mb\sum_{j=1}^{n}(\overline{y_{.j.}} - \overline{y_{...}})^2 + b\sum_{i=1}^{m}\sum_{j=1}^{n}(\overline{y_{ij.}} - \overline{y_{i..}} - \overline{y_{.j.}} + \overline{y_{...}})^2$$

$$+ mn\sum_{k=1}^{b}(\overline{y_{..k}} - \overline{y_{...}})^2 + n\sum_{i=1}^{m}\sum_{k=1}^{b}(\overline{y_{i.k}} - \overline{y_{i..}} - \overline{y_{..k}} + \overline{y_{...}})^2$$

$$+ m\sum_{j=1}^{n}\sum_{k=1}^{b}(\overline{y_{.jk}} - \overline{y_{.j.}} - \overline{y_{..k}} + \overline{y_{...}})^2$$

$$+ \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{b}\left(y_{ijk} - \overline{y_{ij.}} - \overline{y_{i.k}} - \overline{y_{.jk}} + \overline{y_{i..}} + \overline{y_{.j.}} + \overline{y_{..k}} + \overline{y_{...}}\right)^2$$

Step 4: Checking Degree of Freedom

Listing out degree of freedom for each corresponding term in the same order as what step 1 shows. With basic algebra, it is easy to check left and right side have same degree of freedom

mnb-1 = (m-1) + (n-1) + (m-1) (n-1) + (b-1) + (m-1) (b-1) + (n-1) (b-1) + (m-1) (n-1) (b-1)

Then we can figure out formulas for all sum of squares and estimates for each parameter.

Appendix II Method and Rules for EMS Calculation (Taking Model 1 as an Example)

1. Setting down a table. Marking row headings (Sources) as variable terms in model and writing down column headings as subscriptions with numbers of observations. Here, you need to make sure to add F for fixed levels of factor and R for random levels.

2. Filling the table. For each column, if fixed subscription appears in corresponding row heading, then writing down 0; otherwise, filling by number of observations for that fixed subscription. If random subscription shows up in corresponding row heading, then writing down 1; otherwise, filling by number of observations for that random subscription. Note that all cells are filled by 1 in error row, anyway.

3. Calculating degree of freedom for each random variable. Degree of freedom is, as usual, computed as number of observation for corresponding subscription subtracting 1, or product of them. Total degree of freedom is N-1, where N is the production of all numbers of observations.

After first three steps, we can get Appendix IIA for Model 1. Notice that degree of freedom for random error term is not computable in this case.

4. Calculation EMS. We prefer to take two variable terms, α_i, $(\theta\beta)_{jk}$, in Table A as examples for explain the abstract calculation procedure. Since α_i is a fixed variable and $(\theta\beta)_{jk}$ is random, we can see the slightly difference between fixed and random cases.

For α_i, we need to list out all variances related to it, firstly. These variance contains variance of general random error and all variances with subscriptions containing at least the notation of the variance we concern, so these variances are σ_ε^2, $\sigma_{\alpha\theta}^2$, and σ_α^2. No β term is involved since factor B effect (β's) haven't added in yet, in this experiment. Now, we add coefficients for each term. Coefficients for σ_ε^2 are always 1 for any term. To find coefficient for $\sigma_{\alpha\theta}^2$, we can cover up i and j columns in FRF sector in Table A, we

get b, which is the coefficient for $\sigma_{\alpha\theta}^2$. Similar procedure can be applied to find the coefficient for σ_α^2 by covering up i column, and nb is the coefficient for σ_α^2. Adding them together, we get $\sigma_\varepsilon^2 + b\sigma_{\alpha\theta}^2 + nb\sigma_\alpha^2$. (A tip here is that it would benefit later by writing σ_ε^2 as the first term and the term using symbol representing corresponding effect as the last one.)

Appendix IIA Degree of freedom for terms in Model 1

Sources	***d.f.***	***m*** ***F*** ***i***	***n*** ***R*** ***j***	***b*** ***F*** ***k***
α_i	m-1	0	n	b
θ_j	n-1	m	1	b
$(\alpha\theta)_{ij}$	(m-1)(n-1)	0	1	b
β_k	b-1	0	n	0
$(\alpha\beta)_{ik}$	(m-1)(b-1)	0	n	0
$(\theta\beta)_{jk}$	(n-1)(b-1)	0	1	0
$(\alpha\theta\beta)_{ijk}$	(m-1)(n-1)(b-1)	0	1	0
ε_{ijk}		1	1	1
Total	mnb-1			

Rules start to apply to adjust terms.

1. First and last terms are always remaining.
2. For middle variance component of EMS, covering up the symbol representing corresponding effect in subscription, α here. If remaining subscription(s) is fixed, then dropping out this variance term; otherwise, keeping it. Here, remaining subscription is θ, which is random effect, so we keep this term.

Then we get $EMS\ (\alpha_i) = \sigma_\varepsilon^2 + b\sigma_{\alpha\theta}^2 + nb\sigma_\alpha^2$.

For $(\theta\beta)_{jk}$, related variances are σ_ε^2, $\sigma_{\alpha\theta\beta}^2$, $\sigma_{\theta\beta}^2$, and coefficients are 1, 1, and m, respectively. After apply rule 1 and 2, we get $EMS\ ((\theta\beta)_{jk}) = \sigma_\varepsilon^2 + m\sigma_{\theta\beta}^2$. Now rule 3 need to be introduced.

3. If the effect is random, then EMS keeps the same as what we have gotten. If it is fixed, we have to replace the variance component of that fixed effect by Sum of its squares dividing its degree of freedom.

   So, $EMS\ (\alpha_i)$ turns to be $\sigma_\varepsilon^2 + b\sigma_{\alpha\theta}^2 + nb\frac{\sum_i \alpha_i^2}{m-1}$. However, $EMS\ ((\theta\beta)_{jk})$ keeps the same. Together with all other effect terms, we can get Appendix IIB.

Appendix IIB EMS for all terms in model 1

Sources	EMS
α_i	$\sigma_\varepsilon^2 + b\sigma_{\alpha\theta}^2 + nb\frac{\sum_i \alpha_i^2}{m-1}$
θ_j	$\sigma_\varepsilon^2 + mb\sigma_\theta^2$
$(\alpha\theta)_{ij}$	$\sigma_\varepsilon^2 + b\sigma_{\alpha\theta}^2$
β_k	$\sigma_\varepsilon^2 + \sigma_{\alpha\theta\beta}^2 + m\sigma_{\theta\beta}^2 + mn\frac{\sum_k \beta_k^2}{b-1}$
$(\alpha\beta)_{ik}$	$\sigma_\varepsilon^2 + \sigma_{\alpha\theta\beta}^2 + n\frac{\sum_i\sum_j (\alpha\beta)_{ij}^2}{(m-1)(b-1)}$
$(\theta\beta)_{jk}$	$\sigma_\varepsilon^2 + m\sigma_{\theta\beta}^2$
$(\alpha\theta\beta)_{ijk}$	$\sigma_\varepsilon^2 + \sigma_{\alpha\theta\beta}^2$
ε_{ijk}	σ_ε^2

Appendix III Matrix Forms for Models

Matrix Form for General Linear Models. Suppose n data $y_1, \dots, y_n$ are observed and explained by p explanatory variables with n values for each, $x_{11}, \dots, x_{1p}, \dots, x_{n1}, \dots, x_{np}$. The model for it is

$$y_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + e_i$$

Here β*'s* are unknown fixed effects parameters and $e_i's$ are i.i.d. $N(0, \sigma^2)$distributed random variables/ errors. However, we need to write n equations to represent all observations. In order to make the process simpler, we can use a single equation by using matrix notations.

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{e}$$

where

Y represents the vector of all response variables, and $\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$;

X is the known matrix of all explanatory variables, and X is defined as $\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$;

β is the unknown fixed effects parameter vector, and $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$; and

e is the random error vector with $\mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$ and **e** follows $N(\mathbf{0}, \boldsymbol{V})$ where $\boldsymbol{V} = \sigma^2 \boldsymbol{I}$.

Matrix Form for General Linear Mixed Models

For mixed model, both fixed and random effects are involved, such as repeated-measures designs. The random effects would allow elements of **Y** to be correlated. There are two approaches to modify general linear model.

First, we can change the assumptions made for random error matrix **e** follows $N(\mathbf{0}, \boldsymbol{R})$. In this case, matrix **e** is called as covariance matrix. Notice that elements in **e** are not i.i.d. any longer.

Second, we can add random effects and coefficients in model. Based on general linear model, we can add **Zu** terms, where **Z** matrix is a conditional model and **u** follows $N(\mathbf{0}, \boldsymbol{G})$. At the same time, **e** follows $N(\mathbf{0}, \boldsymbol{R})$. Models without a **Z** matrix that capture complex covariance structure directly through the variance matrix of the errors **e** are called marginal models (Littell 2007). In summary, the second approach concludes as following

$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$

u follows $N(\mathbf{0}, \boldsymbol{G})$

e follows $N(\mathbf{0}, \boldsymbol{R})$

$\text{cov}[\mathbf{u}, \mathbf{e}] = \mathbf{0}$

It's important to understand that assumption $\text{cov}[\mathbf{u}, \mathbf{e}] = \mathbf{0}$ is the key feature to avoid parameter confounding. To be specific, $\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}$ denote the vector of response variables, the known matrix of all explanatory variables, and the unknown fixed effects parameter vector, respectively. New term **Z** is a matrix filled with 1's and 0's depending on specific model you use. Vector **u** contains all random effects, so it is not parameter since they are not fixed.

The conditional distribution of $\mathbf{Y}|\mathbf{u}$ and the marginal distribution of **Y** are

$\mathbf{Y}|\mathbf{u}$ follows $N(\boldsymbol{X\beta} + \boldsymbol{Zu}, \boldsymbol{R})$

Y follows $N(\boldsymbol{X\beta}, \boldsymbol{V})$

$\mathbf{V} = \boldsymbol{var}(\boldsymbol{Y}) = \boldsymbol{ZGZ'} + \boldsymbol{R}$

Appendix IV Related SAS Codes

```
proc import out=signals datafile='C:\Users\yxz195\Desktop\signals.xlsx'
DBMS=XLSX REPLACE;
```

```
SHEET="Sheet2";
GETNAMES=YES;
RUN;
proc transpose data=signals
out=signals_trans;
by ID;   /*NOTICE:in order to make SAS recognize ID is sorted with ascending order,
ID's should be labeled as GENE00001--GENE16436*/
VAR _1-_72;
run;
data Reduced_names(drop=_NAME_ _LABEL_ );
set signals_trans;
run;
data Signal (rename=(COL1=Signals));
set Reduced_names;
run;
/* proc print data=Signal (OBS=100); run; */

proc import out=Design_matrix datafile='C:\Users\yxz195\Desktop\design.txt'
DBMS=dlm REPLACE;
delimiter='09'x;   /*read space-delimited text file into sas*/
GETNAMES=no;
RUN;
data Design (rename=(VAR1=SPE VAR2=TR VAR3=REP VAR4=T));
set Design_matrix;
RUN;
/* PROC PRINT DATA=Design (OBS=200); RUN; */

data Thesis;
set Design ;
set Signal;
run;
/* Proc print data= Thesis (OBS=500); run; */

data NormData; /*normalized data to make them ready to be analyzed*/
set Thesis;
Y=log2(Signals+1); /*not log2(Signals) since there are some signals are 0 in this
dataset*/
run;
/* Proc print data=NormData (OBS=720); RUN; */

/* SPLIT-PLOT DESIGN-GENERAL*/

Proc glm data=NormData;
class ID SPE TR T REP;
model Y = SPE TR SPE*TR T SPE*T T*TR REP(SPE*TR*T)/SS3;
test h=TR E=SPE*TR;
```

```
random SPE*TR REP(SPE*TR*T)/test;
RUN;

Proc mixed data=NormData method=ml;
class ID SPE TR T REP;
model Y = SPE TR SPE*TR T SPE*T T*TR;
random SPE*TR;
RUN;

These two small programs produce same F-ratios.*/

/* Gene-Specific Significant Models */

proc sort data=NormData;
by ID SPE TR T REP;
RUN;
/*Proc print data=NormData (OBS=720); RUN;*/

/*title 'Scatterplot - Two Variables';
proc gplot data= NormData(N=72);
plot Y*T;
RUN;*/

proc import out=genes datafile='C:\Users\yxz195\Desktop\GENES.xlsx'
DBMS=XLSX REPLACE;
SHEET="Sheet1";
GETNAMES=YES;
RUN;
proc print data= gene1;
run;

ods graphics on;
title 'Genes Data';
proc corr data=gene1 nomiss plots=matrix(histogram);
var T0 T4 T8 T12 T24 T48;
run;
ods graphics off;

/*ods exclude all;
ods noresults;
proc mixed data = NormData;
by ID;
class ID SPE TR T REP;
model Y=SPE TR SPE*TR T SPE*T T*TR SPE*TR*T/ outp= NormData1;
random REP(SPE*TR*T);
lsmeans SPE TR SPE*TR T SPE*T T*TR SPE*TR*T/diff;
```

```
ods output covparms=COVparms tests3=Tests3 lsmeans=lsms diffs=Diffs;
run;
ods exclude none;
ods results;
```