Mixed Model Analysis for Repeated Measures of Lettuce Growth

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF UNIVERSITY OF MINNESOTA
BY

Levi Dawson Pederson

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

YANG LI

May, 2015

Acknowledgements

I would like to give special thanks to Mike Mageau for putting up with the work of weighing all the plants every week. And a thank you to Yang Li for helping me out of all the problems that I couldn't get anywhere with.

Abstract

We conducted an experiment to compare the growth of lettuce using three different treatments. Each treatment had different spacing between each lettuce plant. Mixed model analysis was used to analyze the growth of the lettuce over time. SAS was used for fitting an appropriate covariance structure to the data in order to represent the correlation between time points using PROC MIXED. We also tested higher order polynomial terms in the model and used ESTIMATE statements to compare the treatments at each day along with comparing weights between days within treatments.

"Because mixed models are more complex and more flexible than the general linear model, the potential for confusion and errors is higher." –Homer & Simpson (2005)

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The data for this particular statistical analysis came from an agricultural study in a greenhouse in Silver Bay, MN. Mike Mageau of the Center of Sustainable Community Development came up with the experiment in order to increase yield and profit from growing lettuces. Mostly lettuce is grown in the greenhouse, with a variety of other vegetables, and roughly 400 heads of lettuce are harvested every week. The lettuce grows in a hydroponic system where cups of dirt, perlite, and seedlings are placed into rafts (or floats). These rafts float in pools of water and allow the roots to grow underneath the raft into the water, while the lettuce grows above. All rafts have the same dimensions; however, there are two types of rafts: 18-hole and 36-hole rafts. The 36-hole rafts generally are used when the lettuce is still relatively small and after a certain age, when they are large enough, are moved to the 18-hole raft. Current production uses only nine of the eighteen holes with the lettuce equally spaced apart, so no two plants will impede another's growth process.

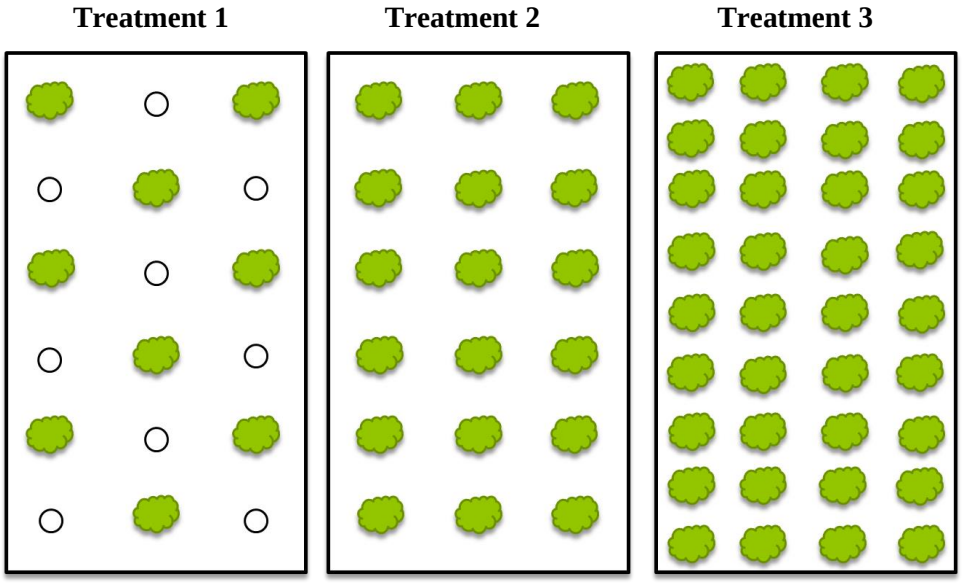

Figure 1.1: Illustration of the three treatments (different spacing between lettuces)

It was to be tested whether or not the lettuce can be grown with less spacing between the plants. Three treatments are designed to assess the effects of three different spacing settings. Treatment 1 is the control, or current process with nine plants in eighteen holes. Treatment 2 had all eighteen holes filled with plants, and treatment 3 had all 36 holes in the 36-hole raft filled with plants (See **Figure 1.1**). Each plant was randomly assigned a treatment, each float was in the same pool of water, and all plants received the same amount of lighting, nutrients, etc. In a pilot experiment, it was found that at least 15 plants are needed to achieve the desired precision. Therefore, it was determined to have 18 plants from each treatment setting. Thus, two floats of treatment 1, one float of treatment 2, and one float of treatment 3 were used in the experiment. For treatment 1 and 2, all plants were weighted, while 18 plants were randomly selected to be weighed from the single raft of treatment 3. This gives a total of 54 plants being studied over 11 unequally spaced time points for a total of 594 observations. The lettuces were measured in grams on December 1st, 5th, 8th, 12th, 15th, 18th, 21st and 26th, as well as January 2nd, 5th, and 9th. The 9th being the day of harvest. See the Appendix for the structure of the data set.

We hope to answer the following question: Do certain treatments tend to have a higher yield? Is it possible to harvest a week earlier and get similar weights? Does the lettuce reach an optimum yield quicker than other treatments? These questions can help significantly improve the yield and, thus, profit. If the results show that Treatment 2 is significantly better or no different than the first treatment, then the yield can be doubled since we only use half of the holes in Treatment 1. The profit, however, is more complicated since it is related to other factors including the labor cost, the cost of floats, the cost of lighting, etc. In this study, we only focus on the statistical analysis of the growth and leave profit maximization for future studies.

In the subsequent chapters, we will go over the basic models generally used with repeated measures and show the procedure how to estimate fixed effects, random effects, and covariance parameters in the model. Options on selecting appropriate covariance structures, and how to fit them is also covered. Finally, we will show the related SAS procedures for data analysis and results.

Figure 1.2 and **1.3** help visualize the growth of the lettuce. The first figure is the profile plots of each individual plant in each treatment. Looking closely, one may notice some lines go down over time. Each time point for each plant gets a measurement of weight in grams that include the weight of the plant, roots, soil, cup, and any water in the soil at the time. Thus, there can be a certain measurement error each day of weighing depending on the scale and how much water is in the soil. When there is an excess of water in the soil on one day and less water during the next day's measurement, there will be a negative slope in the graphs below. This can be seen during the first couple weeks when the growth of the plant is minimal. The second figure shows a point at every time point representing the average weight for the eighteen plants in the corresponding treatment. The line going through the mean shows the normal error, or deviation, of the plants above and below the mean, and one can see that the plants tend to deviate from the mean more as time goes on.

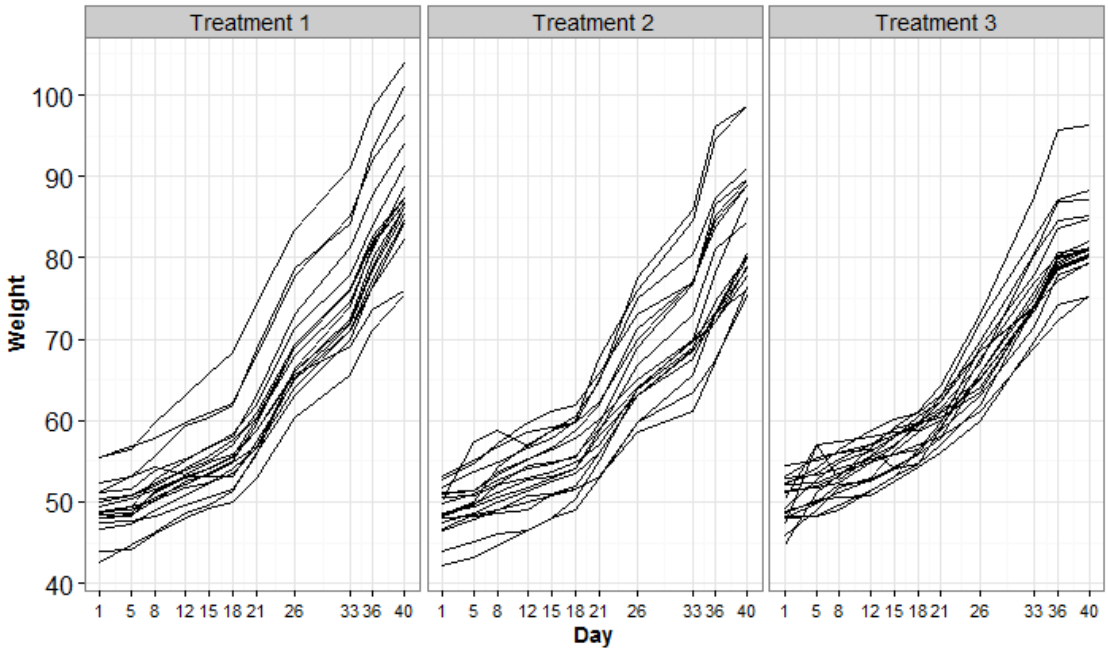

Figure 1.2: Profile plots of the three treatments.

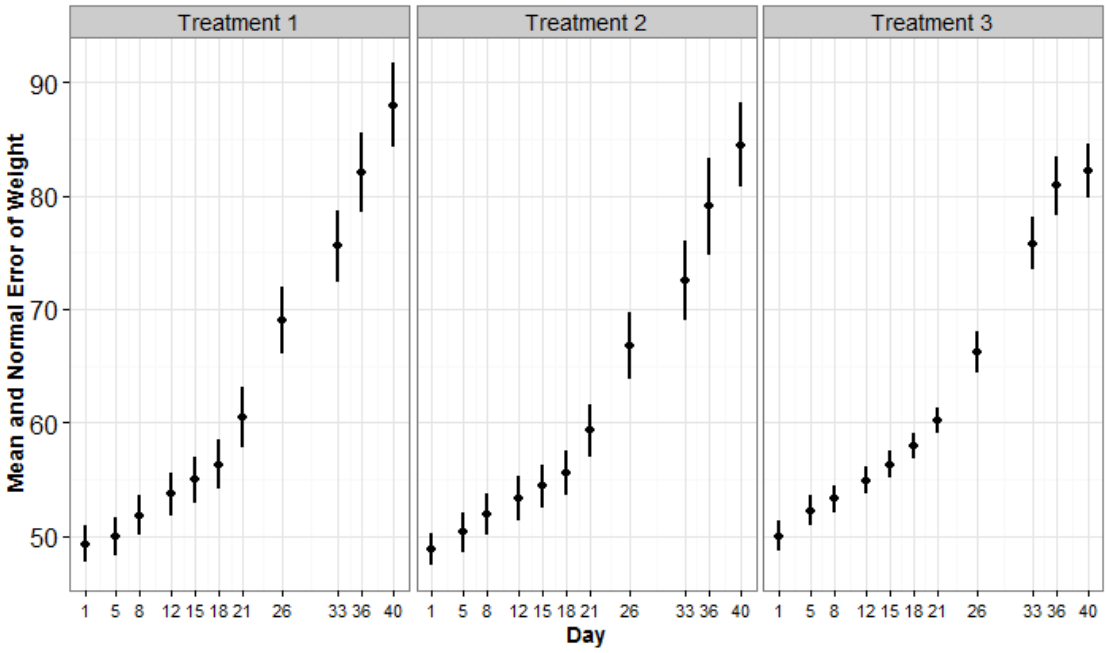

Figure 1.3: Plots of the three treatments with the mean and standard error of the measurements at each day.

Chapter 2: Model Specification

A typical model for repeated measures data is the following linear model:

$$Y_{ijk} = \mu + \alpha_i + \tau_k + (\alpha\tau)_{ik} + e_{ijk}$$

$$i = 1,2,..q \quad j = 1,2,..n \quad k = 1,2,..t$$

where μ is a constant intercept parameter, α_i is the fixed effect corresponding to treatment *i*, τ_k is the fixed effect corresponding to day k, $(\alpha\tau)_{ik}$ is the interaction effect due to treatment i on day k, and e_{ijk} is the random error of an individual plant *j* in treatment *i* on day *k.*

Even though the plants were randomly assigned to the treatments, time is *not* randomly assigned to the plants. Therefore we cannot assume the random errors for the *same plant* are independent. Measurements from the same subject are correlated. Therefore, we can assume that the errors for different plants are independent:

$$Cov[e_{ijk}, e_{i'j'l}] = 0, \text{ if either } i \neq i' \text{ or } j \neq j'.$$

Also, since measurements on the same plant are taken over time, they may have different variances at each time point and correlation between pairs of measurements may depend on the time interval between the two time points, known as the lag. In the most general setting, we have:

$$Var[e_{ijk}] = \sigma_k^2 , \qquad Cov[e_{ijk}, e_{ijl}] = \sigma_{k,l}$$

where σ_k^2 is the heterogeneous variances at time point t. This general structure has a symmetric variance-covariance matrix with σ_k^2 on the diagonal and $\sigma_{k,l}$ on the off-diagonal corresponding to row *k* and column *l*. This gives $t(t+1)/2$ parameters that need to be estimated and in this particular experiment, with eleven time points, we will have 66 parameters to estimate. In most cases, if we can assume some covariance structure in the model, then the number of covariance parameters can be greatly reduced. Details will be discussed in Chapter 4.

We can express the vector of observations on plant *j* in treatment *i* as $\vec{Y}_{ij} = [Y_{ij1}, Y_{ij2}, \cdots, Y_{ijt}]'$ and the corresponding variance-covariance structure is

$$Var\left[\vec{Y}_{ij}\right] = \Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1,t} \\ \vdots & \ddots & \vdots \\ \sigma_{t,1} & \cdots & \sigma_t^2 \end{bmatrix}.$$

This assumes that the covariance structure Σ is the same for each plant. Furthermore, if we stack the vectors on top of each other, we can represent the vector of all plant observations as $\boldsymbol{Y} = \left[\vec{Y}'_{11}, \vec{Y}'_{12}, \cdots, \vec{Y}'_{1n}, \vec{Y}'_{21}, \cdots, \vec{Y}'_{tn}\right]'$ then $Var[\boldsymbol{Y}] = \boldsymbol{V} = diag\{\Sigma\}$ which means the variance structure of $\boldsymbol{Y}$ is block diagonal with Σ on the diagonals. Sometimes, as in this experiment, it is advantageous to include a between-subjects random effect since the plants used in the experiment were randomly selected from a population of lettuces. Our model then becomes

$$Y_{ijk} = \mu + \alpha_i + b_{ij} + \tau_k + (\alpha\tau)_{ik} + e_{ijk} \tag{2.1}$$

where $b_{ij} \sim iid\ N(0, \sigma_b^2)$ is the random effect for plant *j* assigned to treatment *i*. In matrix notation, we can write this model as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e} \tag{2.2}$$

where $\boldsymbol{X}$ is the matrix of known constants, $\boldsymbol{\beta}$ is the vector of fixed, unknown parameters, $\boldsymbol{u}$ is the vector of random effects, $\boldsymbol{Z}$ is the corresponding matrix of known constants, and $\boldsymbol{e}$ is the vector of errors, e_{ijk}. The random effects vectors $\boldsymbol{u}$ and $\boldsymbol{e}$ are assumed to be independent, where $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G})$ and $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{R})$ by letting $\boldsymbol{G} = Var(\boldsymbol{u})$ and $\boldsymbol{R} = Var(\boldsymbol{e})$. The mean and variance of $\boldsymbol{Y}$ are $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$ and $Var(\boldsymbol{Y}) = \boldsymbol{ZGZ}' + \boldsymbol{R}$, respectively. Here, the between-plant error is represented by $\boldsymbol{ZGZ}'$ and $\boldsymbol{R}$ represents the within-plant error (Littell et. al 2006).

Chapter 3: Estimation of Parameters

With the normal assumption for the responses, we can proceed with estimation in the Gaussian mixed model. So far we have that the response variable $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$, where $\boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}$. This gives the following joint pdf,

$$f(\boldsymbol{Y}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{V}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{V^{-1}}(\boldsymbol{Y} - \boldsymbol{X\beta})\}$$

where n is the dimension of $\boldsymbol{Y}$. From this we get the log-likelihood function

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) = c - \tfrac{1}{2}\log(|\boldsymbol{V}|) - \tfrac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{V^{-1}}(\boldsymbol{Y} - \boldsymbol{X\beta}),$$

where c is a constant and $\boldsymbol{\theta}$ represents the variance components vector involved in $\boldsymbol{V}$, that is, $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$. By differentiating the log-likelihood function with respect to the parameter vectors, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, we obtain the following equations (see Appendix (A)),

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \boldsymbol{X'V^{-1}Y} - \boldsymbol{X'V^{-1}X\beta}, \tag{3.1}$$

$$\frac{\partial l}{\partial \theta_r} = \frac{1}{2}\left\{(\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{V^{-1}}\frac{\partial \boldsymbol{V}}{\partial \theta_r}\boldsymbol{V^{-1}}(\boldsymbol{Y} - \boldsymbol{X\beta}) - tr\left(\boldsymbol{V^{-1}}\frac{\partial \boldsymbol{V}}{\partial \theta_r}\right)\right\}, \tag{3.2}$$

$$r = 1, \ldots, q,$$

where θ_r is the rth component of $\boldsymbol{\theta}$, which has dimension q. Setting the above equations equal to zero and solving them are the standard procedure in finding the maximum likelihood estimators (MLE) for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Now let p be the dimension of $\boldsymbol{\beta}$ and for simplicity, assume that $\text{rank}(\boldsymbol{X}) = p$; that is, $\boldsymbol{X}$ is of full column rank. Letting $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ represent the MLE's, from (3.1) one obtains

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'}\widehat{\boldsymbol{V}}^{\boldsymbol{-1}}\boldsymbol{X})^{-1}\boldsymbol{X'}\widehat{\boldsymbol{V}}^{\boldsymbol{-1}}\boldsymbol{Y}, \tag{3.3}$$

where the matrix $\boldsymbol{V}$ with the variance components involved is replaced by their MLE, that is, $\widehat{\boldsymbol{V}} = \boldsymbol{V}(\widehat{\boldsymbol{\theta}})$. Thus, the MLE of $\boldsymbol{\beta}$ can be calculated after the MLE of $\boldsymbol{\theta}$ is found. Notice, this closed form MLE of $\boldsymbol{\beta}$ is the same as the ordinary least squares estimate of $\boldsymbol{\beta}$. Also, if X is not of full rank, one can replace $(\boldsymbol{X'}\widehat{\boldsymbol{V}}^{\boldsymbol{-1}}\boldsymbol{X})^{-1}$ with its generalized inverse $(\boldsymbol{X'}\widehat{\boldsymbol{V}}^{\boldsymbol{-1}}\boldsymbol{X})^{-}$ in order to estimate $\boldsymbol{\beta}$.

MLEs of the variance components generally tend to be biased, whereas REML estimates are at least approximately unbiased. When one is trying to estimate $\boldsymbol{\theta}$, the fixed

effects are called nuisance parameters and the REML is a method that finds an estimate of $\boldsymbol{\theta}$ without having to deal with those nuisance parameters. In order to get around this, one has to transform the data, which we will now illustrate under a general setting.

Assume without loss of generality that $\text{rank}(\boldsymbol{X}) = p$. Let $\boldsymbol{A}$ be an $N \times (N-p)$ matrix such that

$$rank(\boldsymbol{A}) = N - p, \quad \boldsymbol{A}'\boldsymbol{X} = 0.$$

Define $\boldsymbol{z} = \boldsymbol{A}'\boldsymbol{Y}$, and it is straightforward to show that $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{A}'\boldsymbol{V}\boldsymbol{A})$. The joint pdf of $\boldsymbol{z}$ is given by

$$f_R(\boldsymbol{z}) = \frac{1}{(2\pi)^{(N-p)/2}\,|\boldsymbol{A}'\boldsymbol{V}\boldsymbol{A}|^{1/2}} \exp\left\{-\frac{1}{2}\boldsymbol{z}'(\boldsymbol{A}'\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{z}\right\},$$

where symbol R, in $f_R(\boldsymbol{z})$, corresponds to "restricted". By taking the log-likelihood of $\boldsymbol{z}$, which is called the restricted log-likelihood, we have the following equation

$$l_R(\boldsymbol{\theta}) = c - \frac{1}{2}\log(|\boldsymbol{A}'\boldsymbol{V}\boldsymbol{A}|) - \frac{1}{2}\boldsymbol{z}'(\boldsymbol{A}'\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{z}, \tag{3.5}$$

where, again, c is a constant. Notice that there is no fixed effects parameter $\boldsymbol{\beta}$ in the equation. Now differentiating the restricted log-likelihood function and expressing in terms of $\boldsymbol{Y}$, we obtain,

$$\frac{\partial l_R}{\partial \theta_i} = \frac{1}{2}\left\{\boldsymbol{Y}'\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \theta_i}\boldsymbol{P}\boldsymbol{Y} - tr\left(P\frac{\partial \boldsymbol{V}}{\partial \theta_i}\right)\right\}, \quad i = 1, \ldots, q, \tag{3.6}$$

where

$$\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{A}'.$$

The REML estimator of $\boldsymbol{\theta}$ is thus defined to be the maximizer of (3.5), and such a maximization satisfies the REML equation $\frac{\partial l_R}{\partial \boldsymbol{\theta}} = 0$ (Jiang 2007).

Now what is left is to find the best linear unbiased predictor (BLUP) of $\boldsymbol{u}$. Henderson (1950) gave the following mixed model equations:

$$\begin{aligned} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\widehat{\boldsymbol{u}} &= \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\widehat{\boldsymbol{u}} &= \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Y} \end{aligned} \tag{3.7}$$

He described these estimates as being "joint maximum likelihood estimates," and later he explained that his derivation came from maximizing the joint density of $\boldsymbol{Y}$ and $\boldsymbol{u}$ with

respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$, while assuming that $\boldsymbol{u}$ and $\boldsymbol{e}$ are normally distributed (Henderson 1973). The joint density is

$$f(\boldsymbol{y},\boldsymbol{u}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}+\frac{q}{2}}}\left(\det\begin{bmatrix}\boldsymbol{G} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix}\right)^{-\frac{1}{2}}$$
$$\times \exp\left\{-\frac{1}{2\sigma^2}\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu}\end{pmatrix}'\begin{bmatrix}\boldsymbol{G} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix}^{-1}\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu}\end{pmatrix}\right\}. \qquad (3.8)$$

Maximizing $f(\boldsymbol{y},\boldsymbol{u})$ with respect to the parameters, $\boldsymbol{\beta}$ and $\boldsymbol{u}$, requires minimizing

$$\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu}\end{pmatrix}'\begin{bmatrix}\boldsymbol{G} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix}^{-1}\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu}\end{pmatrix}$$
$$= \boldsymbol{u}'\boldsymbol{G}^{-1}\boldsymbol{u} + (\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})'\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu}).$$

Differentiating this with respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$ and equating each to zero gives Henderson's mixed model equations (3.7) (Robinson 1991).

Solving (3.7) for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ (See Appendix (B)), we get the same estimate for $\boldsymbol{\beta}$ as in the maximum likelihood estimate (3.3) and the resulting predictor of $\boldsymbol{u}$ is

$$\hat{\boldsymbol{u}} = \left(\hat{\boldsymbol{G}}^{-1} + \boldsymbol{Z}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{Z}\right)\boldsymbol{Z}'\hat{\boldsymbol{R}}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{G}}\boldsymbol{Z}'\hat{\boldsymbol{V}}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Chapter 4: Modeling Covariance Structure

One of the challenges in repeated measures analysis is to find the right covariance structure that fits the data. As seen from above, the estimate for $\boldsymbol{\beta}$ depends on the variance parameters estimate of $\boldsymbol{\theta}$. This variance-covariance matrix, $\boldsymbol{V}(\boldsymbol{\theta})$, can have many different variations and it is important to have the appropriate one for the within-subject correlation in order to arrive at an accurate conclusion. If we ignored this correlation by using a model that is too simple, then the Type I error rate for fixed effect tests will increase. On the other hand, if the model is too complicated, then it can lead to a sacrifice in test power and the efficiency of tests for the fixed effects. Even though the true covariance structure is seldom known, an approximately accurate covariance model must be specified for a valid analysis.

It is a good idea to start with an unstructured (UN) covariance model to look for any patterns that might suggest a certain structure. For example, if you notice that there is heterogeneous variances over time, then you should rule out homogeneous covariance structures like the AR(1) model.

UNSTRUCTURED: The unstructured covariance structure is the most complex because it allows unequal covariances for each combination of time as well as unequal variances at each time point, giving a total of $t(t+1)/2$ parameters. This can be expressed by:

$$\boldsymbol{V}_{ij} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \dots & \sigma_{1t} \\ & \sigma_2^2 & \sigma_{23} & \dots & \sigma_{2t} \\ & & \sigma_3^2 & \vdots & \sigma_{3t} \\ & & & \ddots & \vdots \\ & & & & \sigma_t^2 \end{bmatrix}$$

This is generally more difficult to fit due to over-parameterization and it may result in less powerful tests. Since covariance structures are symmetric, we refrain from filling in the lower triangle of the matrix here, and for the rest of the paper.

SIMPLE: On the opposite end of the spectrum, the simple covariance structure assumes that there is no correlation between any pair of observations, even on the same subject. Therefore, it assumes that all observations are independent of each other. This gives a structure with 0 for the off-diagonal elements and σ^2 on the main diagonal and is

expressed as $\boldsymbol{V}_{ij} = \sigma^2 \boldsymbol{I}_{\boldsymbol{t}}$, where $\boldsymbol{I}_{\boldsymbol{t}}$ is the $t \times t$ identity matrix. It gets the name "simple" because there is only one parameter that needs to be estimated. Typically, this structure is seldom true with repeated measures data since measurements taken from the same subject are in general correlated.

COMPOUND SYMMETRY (CS): A step up from simple structure, is the compound symmetry covariance structure that assumes equal variances on the main diagonal and equal covariances on the off-diagonal. This is the simplest model for repeated measures data for it assumes a constant correlation between observations no matter the distance of time between the observations (Wang & Goonewardene 2004). The two parameters used in CS are the between subject variance (σ_b^2) and the within subject variance (σ_e^2) where $\sigma^2 = \sigma_b^2 + \sigma_e^2$ and $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_e^2)$ and the matrix can be expressed as (Littell et. al 2006):

$$\boldsymbol{V}_{ij} = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho \\ & 1 & \ddots & \rho \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix}$$

This seldom comes into play because of the time-dependent correlations that most likely exist with repeated measures data.

AUTOREGRESSIVE OF ORDER 1 (AR(1)): The first-order autoregressive covariance structure assumes that the correlation between adjacent observations is ρ as well as assuming the correlation between observations n units apart have correlation ρ^n. Therefore, the correlation between observations is a function of distance in time. Another assumption is that of equal variances along the diagonal, giving only two parameters to estimate, and the structure looks like the following:

$$\boldsymbol{V}_{ij} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{t-1} \\ & 1 & \rho & \cdots & \rho^{t-2} \\ & & 1 & \ddots & \vdots \\ & & & \ddots & \rho \\ & & & & 1 \end{bmatrix}$$

One thing to note is that AR(1) structure requires equally spaced time points. If the data has unequally spaced time points, as in our case, the corresponding covariance structure is the spatial power law covariance structure.

ANTE(1): The first-order antedependence covariance structure allows heteroscedasticity (property of unequal variances) over time, covariance among different pairs of measurements, and unequal correlations. Because of this, there are $t + (t-1)$ parameters to be estimated, and the matrix is expressed as:

$$\boldsymbol{V}_{ij} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_1\rho_2 & \cdots & \sigma_1\sigma_t\rho_1\rho_2 \dots \rho_{t-1} \\ & \sigma_2^2 & \sigma_2\sigma_3\rho_2 & \cdots & \sigma_1\sigma_t\rho_2\rho_3 \dots \rho_{t-1} \\ & & \sigma_3^2 & \ddots & \vdots \\ & & & \ddots & \sigma_{t-1}\sigma_t\rho_{t-1} \\ & & & & \sigma_t^2 \end{bmatrix}$$

Unlike the AR(1) model, this structure allows unequal time spacing (Wang & Goonewardene 2004). If one has unequal time spacing, spatial covariance structures can be tested as well. Two common structures are SP(POW) and SP(SPH) and the connection is that the unequally spaced data can be viewed as a spatial process in one dimension (Littell et. al 2006).

SPATIAL POWER LAW (SP(POW)): This structure is a direct association to the AR(1) model using correlation as a function of distance in time, except instead of the constant d, we now have d_{ij}, where d_{ij} is the distance between time points i and j. SP(POW) is represented by

$$\boldsymbol{V}_{ij} = \sigma^2 \begin{bmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} & \cdots & \rho^{d_{1t}} \\ & 1 & \rho^{d_{23}} & \cdots & \rho^{d_{2t}} \\ & & 1 & \ddots & \vdots \\ & & & \ddots & \rho^{d_{(t-1)t}} \\ & & & & 1 \end{bmatrix}$$

SPATIAL SPHERICAL (SP(SPH)): This structure is a little tedious to express in matrix format, but the (i,j)th element can be expressed as $\sigma^2 \left[1 - \frac{3d_{ij}}{2\rho} + \frac{d_{ij}^3}{2\rho^3}\right] I(d_{ij} \leq \rho)$. Here, $I(d_{ij} \leq \rho)$ is the indicator function and d_{ij} is the same as defined in SP(POW). If two measurements are separated by a time interval greater than ρ, a quantity called range in spatial statistics, they will be uncorrelated (SAS/STAT 9.2 User's Guide 2008).

Chapter 5: Choosing Covariance Structure using Proc Mixed in SAS

We now get into detail of using SAS for the data analysis. First we fit linear mixed models to the data, and then compare information criteria to select the best covariance structure.

There are multiple information criteria that can be used to measure goodness of fit, including Akaike Infromation Criteria (AIC), AICC, which is an adjusted AIC for small sample sizes, and Bayesian Information Criteria (BIC), also known as Schwarz's Bayesian Criteria (BIC). These criteria are computed using the Residual Log Likelihood that SAS computes for a given model. Letting L_R denote the residual likelihood, the criteria can be defined as follows:

$$\text{AIC} = -2lnL_R + 2k,$$

$$\text{AICC} = AIC + \frac{2k(k+1)}{n-k-1},$$

$$\text{BIC} = -2lnL_R + k\ln(n).$$

Here, *k* is the number of parameters to be estimated and *n* is the total number of observations. When comparing two different models with goodness of fit statistics, the model with a smaller number has a better fit.

If one model is a special case of another (nested), one may also compare the two models using the likelihood ratio test (LRT). For example, if one wants to compare the CS model with the UN model, the null hypothesis would be

$$H_0\text{: The restriced model (CS) is true.}$$

The alternative hypothesis would then be

$$H_A\text{: The restricted model (CS) is not true.}$$

The test statistic is

$$\chi^2 = 2[(\text{Log likelihood of full model}) - (\text{Log likelihood of restricted model})]$$

which can also be written as

$$\chi^2 = 2\ln\left(\frac{\text{Likelihood of full model}}{\text{Likelihood of restricted model}}\right)$$

hence, the name "likelihood ratio test." If the restricted model is true, the test statistic has an asymptotically chi-square distribution with degrees freedom $d_F - d_R$, where d_F and d_R are the number of parameters estimated in the full and restricted models, respectively. One then rejects the null hypothesis if the test statistic is greater than $\chi^2_{d_F - d_R}$. When models are *nested* (one is a special case of another), any of the above criteria is applicable, and when they are *non-nested* models, typically AIC and BIC criteria are used (Archontoulis & Miguez 2015).

PROC MIXED is the procedure in SAS that can be used to fit a general linear mixed model. When using the MIXED procedure, the general linear mixed model from (1.2) can be fit using the MODEL, CLASS, RANDOM, and REPEATED statements. The CLASS statement is used to indicate which variables are the classification or categorical variables. The MODEL statement is used to define your model with the response on the left of the equal sign and the fixed effects on the right of the equal sign. The RANDOM statement is a way to incorporate the random effects $\boldsymbol{u}$ into the model and the REPEATED statement is used to specify covariance structures for repeated measurements on subjects (SAS/STAT 9.2 User's Guide 2008).

From the data set in the Appendix (C), we can see that SPACING, PLANT, and DAY are all classification variables. The variable DAY could also be treated as continuous by leaving it out of the CLASS statement, but for the first stage of implementing the covariance structure, one should leave it as a classification variable, otherwise you may be underspecifying the mean structure and result in biased estimates for the variance-covariance parameters and, thus, lead to an incorrect assessment of the covariance structure (Littell et. al 2006).

The basic structure of code is as follows:

```
proc mixed data=Project.Spacingdata;
      class spacing day plant;
      model weight=spacing day spacing*day;
run;
```
(5.1)

SAS would then provide an ordinary least squares fit of the model (2.1). In order to fit a covariance structure to the model, we must add RANDOM and/or REPEATED

statements to (5.1). These statements cause SAS to compute REML estimates, by default, of the covariance parameters for the specified structure.

Options are available for printout with the RANDOM and REPEATED statements by adding the option following a slash (/). A few of the common options are TYPE=, R, RCORR, G, GCORR, V, VCORR, and SUBJECT=. TYPE= allows you to specify which covariance structure you would like to fit. R and RCORR can be used in the REPEATED statement to printout the $\boldsymbol{R}$ matrix in covariance or correlation form. G, GCORR, V, and VCORR can similarly be used to find the $\boldsymbol{G}$ matrix or $\boldsymbol{V}$ matrix, but in the RANDOM statement. SUBJECT= option specifies the variables whose levels are used to identify the block diagonal structure in $\boldsymbol{G}$ or $\boldsymbol{R}$.

I will now present the SAS code for identifying each of the covariance structures listed in Chapter 3 and make comparisons. Notice that in the REPEATED statements, we have 'SUBJECT=PLANT'. This specifies the R matrix to be block-diagonal with a sub-matrix for each plant. If the data had the plants numbered 1-18 for each spacing, then one would need to write 'SUBJECT=PLANT(SPACING)', but since there are no common numberings in our data, it is sufficient to designate only 'SUBJECT=PLANT'.

UNSTRUCTURED:

```
proc mixed data=Project.Spacingdata;
      class spacing plant day;
      model weight=spacing day spacing*day;
      repeated day / subject=plant type=un r rcorr;
run;
```

The repeated statement above is used to define the covariance matrix of $\boldsymbol{Y}$ conditional on the random effects $\boldsymbol{u}$, $Var(\boldsymbol{Y}|\boldsymbol{u}) = \boldsymbol{R}$. Since there is a REPEATED statement, but no RANDOM statement, PROC MIXED models the covariance matrix of $\boldsymbol{Y}$ directly (Littell et. al 2006). In **Table 5.1** we can see on the diagonal that the variances increase with time and, thus, would not choose a structure that has homoscedasticity. Also, looking at **Table 5.2**, SP(POW) would be a good covariance structure to test for the lettuce data since the correlation decreases slowly with time. Otherwise a heterogeneous AR(1) model would also be good to test.

UN Estimated R Matrix										
8.6787	7.8474	7.9725	8.2171	7.915	7.7273	8.8836	9.6828	10.0738	10.5188	9.4936
7.8474	10.3991	10.0365	9.8413	9.5887	9.268	10.3535	10.6739	10.8766	11.7195	10.7823
7.9725	10.0365	10.9109	10.9188	10.8299	10.5281	11.7351	12.6364	12.9184	13.9607	12.7812
8.2171	9.8413	10.9188	11.8409	11.881	11.7953	13.2956	14.6472	15.1625	16.7507	15.4065
7.915	9.5887	10.8299	11.881	12.5717	12.8	14.2522	15.7735	16.5795	18.4753	16.9319
7.7273	9.268	10.5281	11.7953	12.8	13.4518	15.0218	16.8563	18.0041	20.1578	18.5186
8.8836	10.3535	11.7351	13.2956	14.2522	15.0218	18.2056	21.2154	23.0583	26.2561	24.1708
9.6828	10.6739	12.6364	14.6472	15.7735	16.8563	21.2154	27.5478	30.5638	35.5473	32.4669
10.0738	10.8766	12.9184	15.1625	16.5795	18.0041	23.0583	30.5638	36.5554	42.4436	38.7668
10.5188	11.7195	13.9607	16.7507	18.4753	20.1578	26.2561	35.5473	42.4436	51.0928	46.6873
9.4936	10.7823	12.7812	15.4065	16.9319	18.5186	24.1708	32.4669	38.7668	46.6873	44.6185

Table 5.1: SAS Output for the estimated variance-covariance R Matrix for Unstructured.

UN Estimated R Correlation Matrix										
1	0.826	0.8193	0.8106	0.7578	0.7152	0.7067	0.6262	0.5656	0.4995	0.4824
0.826	1	0.9422	0.8869	0.8386	0.7836	0.7525	0.6306	0.5579	0.5084	0.5006
0.8193	0.9422	1	0.9606	0.9247	0.869	0.8326	0.7289	0.6468	0.5913	0.5793
0.8106	0.8869	0.9606	1	0.9738	0.9346	0.9055	0.811	0.7288	0.681	0.6703
0.7578	0.8386	0.9247	0.9738	1	0.9843	0.9421	0.8476	0.7734	0.729	0.7149
0.7152	0.7836	0.869	0.9346	0.9843	1	0.9599	0.8756	0.8119	0.7689	0.7559
0.7067	0.7525	0.8326	0.9055	0.9421	0.9599	1	0.9473	0.8938	0.8609	0.8481
0.6262	0.6306	0.7289	0.811	0.8476	0.8756	0.9473	1	0.9631	0.9475	0.9261
0.5656	0.5579	0.6468	0.7288	0.7734	0.8119	0.8938	0.9631	1	0.9821	0.9599
0.4995	0.5084	0.5913	0.681	0.729	0.7689	0.8609	0.9475	0.9821	1	0.9778
0.4824	0.5006	0.5793	0.6703	0.7149	0.7559	0.8481	0.9261	0.9599	0.9778	1

Table 5.2: SAS Output for the estimated correlation matrix R for Unstructured.

COMPOUND SYMMETRIC:

```
proc mixed data=Project.Spacingdata;
        class spacing plant day;
        model weight=spacing day spacing*day;
        repeated day / subject=plant type=cs r rcorr;
run;
```

When looking at the UN model, if the variances tend to increase (or decrease) with time, then one may try a heterogeneous compound symmetric covariance structure by replacing CS with CSH in the TYPE= option.

ANTE(1):

```
proc mixed data=Project.Spacingdata;
      class spacing plant day;
      model weight=spacing|day / s;
      repeated day/ sub=plant type=ante(1);
run;
```

Spacing|day is a short hand form of typing the same model statement in CS above. Adding 's' after the slash prints out the fixed effects estimates for the model.

SP(POW):

```
proc mixed  data=Project.Spacingdata2 order=data;
      class spacing  day  plant;
      model weight=spacing|day;
      repeated day / type=sp(pow) (day1) sub=plant r;
      random int / subject=plant v;
run;
```

'Spacingdata2' is a new data set where we add the variable DAY1=DAY. By not placing DAY1 into the class statement, it is treated as a continuous variable, thus for the repeated statement (DAY1) is used to indicate the time levels in the SP(POW) covariance structure. The ORDER=DATA option in the PROC MIXED statement, tells SAS to preserve the ordering of the levels of the class variable DAY.

SP(SPH):

```
proc mixed data=Project.Spacingdata order=data;
      class spacing day plant;
      model weight=spacing|day;
      repeated day / type=sp(sph) (day1) sub=plant;
      random int / sub=plant;
run;
```

Each of the above outputs produce a table labeled "Fit Statistics" where the -2 Res Log Likelihood, AIC, AICC, and BIC numbers are located. **Table 5.3** summarizes these

statistics. Recall that we are looking for the best fit, hence, the smallest goodness of fit statistic in each column. These numbers are in bold. Even though the unstructured model gives the best log likelihood, with so many parameters (recall 66 of them to estimate) the model will have problems of over-fitting and tougher computation of the estimates of fixed effects. Thus, we look at the information criterion that punishes model complexity and choose the smallest number from those columns. For our lettuce data, ANTE(1) covariance structure provides the best fit. For more options in covariance structures, see SAS 9.2 User's Guide for the MIXED procedure.

Structure type	-2ResLogLik.	AIC	AICC	BIC
UN	**1898.6**	2030.9	2048.8	2162.2
CS	2894	2898	2898	2902
CSH	2628.3	2652.3	2652.9	2676.1
ANTE(1)	1983.5	**2025.5**	**2027.2**	**2067.3**
SP(POW)	2161.8	2165.8	2165.9	2169.8
SP(SPH)	2153.4	2157.4	2157.5	2161.4

Table 5.3: Summary of goodness of fit statistics for select covariance structures.

Chapter 6: Results

Now that we have ANTE(1) as the best fit covariance structure for the model, we want to study the properties of the growth curve as a function of time. We are fitting a polynomial trend with order three using

```
proc mixed data=Project.Spacingreg;
      class spacing plant;
      model weight=spacing spacing*d spacing*d2 spacing*d3 / s
                  noint ddfm=kr;
      repeated / sub=plant type=ante(1);
run;
```
(6.1)

'Spacingreg' is a new data set where we define D=DAY, D2=DAY*DAY, and D3=DAY*DAY*DAY. Thus, we now have our squared and cubic term to put into the model. Also, DAY (or D now) is not in the CLASS statement. This is because D must be treated as a continuous variable in order for PROC MIXED to run effectively. In the MODEL statement, S produces the solution of the fixed effects, NOINT suppresses the intercept from being in the output (leaving this out will give the same results, but different format), and DDFM=KR specifies the Kenward-Roger method for calculating the denominator degrees of freedom. The standard errors for the treatment effects will be biased downward. Kenward-Rogers is a more general degree-of-freedom procedure which can correct for the downward bias, and the Satterthwaite's approximation procedure is a special case of the Kenward-Rogers. For designs with missing data, the Kenward-Rogers is recommended (SAS/STAT 9.2 User's Guide 2008). In the REPEATED statement, we no longer have DAY in front of the '/' since it is no longer a CLASS variable.

Type 1 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
spacing	2	83.9	5.68	0.0049
spacing*d	3	142	630.88	<.0001
spacing*d2	3	268	217.91	<.0001
spacing*d3	3	258	8.89	<.0001

Table 6.1: SAS output for testing significance of fixed effects. Type 1 implies the sequential ANOVA testing for fixed effects.

Solution for Fixed Effects						
Effect	spacing	Estimate	Standard Error	DF	t Value	Pr > |t|
spacing	1	45.2797	0.7259	68.7	62.38	<.0001
spacing	2	44.9343	0.7259	68.7	61.90	<.0001
spacing	3	48.3006	0.7259	68.7	66.54	<.0001
spacing*d	1	0.02063	0.1088	191	0.19	0.8497
spacing*d	2	0.1511	0.1088	191	1.39	0.1663
spacing*d	3	0.04055	0.1088	191	0.37	0.7097
spacing*d2	1	0.009423	0.006159	241	1.53	0.1273
spacing*d2	2	-0.00191	0.006159	241	-0.31	0.7573
spacing*d2	3	0.01839	0.006159	241	2.99	0.0031
spacing*d3	1	0.000249	0.000094	258	2.66	0.0084
spacing*d3	2	0.000399	0.000094	258	4.26	<.0001
spacing*d3	3	-0.00012	0.000094	258	-1.23	0.2204

Table 6.2: SAS output for solutions of the fixed effects estimates.

Table 6.1 shows the significance of the fixed effects in the model. Since each effect has a p-value < 0.05, we conclude that all of these terms must be included in the model. Higher order polynomial trends are not considered because they tend to pick some spurious oscillation patterns. **Table 6.2** shows the 'Solution for Fixed Effects' coming from /S in the model statement. This table provides coefficients of the third degree polynomials for each treatment of plant spacing. From this table, the fitted polynomial equations for each treatment/spacing are

Treatment 1: $\text{weight} = 45.2797 + 0.02063\,D + 0.009423\,D^2 + 0.000249\,D^3$

Treatment 2: $\text{weight} = 44.9343 + 0.1511\,D - 0.00191\,D^2 + 0.000399\,D^3$

Treatment 3: $\text{weight} = 48.3006 + 0.04055\,D + 0.01839\,D^2 - 0.00012\,D^3$

Recall, that we wanted to know if certain treatments give a higher yield than the others. The ESTIMATE statements below can be added to (5.1) to get estimates for differences between each treatment at each day (Littell, Pendergast, & Natarajan 2000). This will help show how much of a difference there is between each treatment on each day of measurement and how significant that difference is. To see where the numbers come from in the statements, let's look at the second ESTIMATE statement where we want to estimate the difference between the average weight of Treatment 1 (9hole) and Treatment 2 (18hole) on the 5th day. Thus, we have the following estimate:

$$\text{Treatment 1} - \text{Treatment 2 on day 5} =$$
$$[45.2797(1) + 0.02063(5) + 0.009423\,(5^2) + 0.000249\,(5^3)] - [44.9343(1) + 0.1511\,(5) - 0.00191\,(5^2) + 0.000399\,(5^3)]$$

The second estimate statement below reflects this. For example, 1 -1 0 after SPACING is for taking the difference of the intercepts of the first two treatments and giving the third treatment's intercept a coefficient of 0. Therefore, we place a 0 in the third position after each effect so we don't include the third treatment in the estimate.

```
estimate '9hole-18hole day1' spacing 1 -1 0 spacing*d 1 -1 0
        spacing*d2 1 -1 0 spacing*d3 1 -1 0;
estimate '9hole-18hole day5' spacing 1 -1 0 spacing*d 5 -5 0
       spacing*d2 25 -25 0 spacing*d3 125 -125 0;
estimate '9hole-18hole day8' spacing 1 -1 0 spacing*d 8 -8 0
```

```
      spacing*d2 64 -64 0 spacing*d3 512 -512 0;
estimate '9hole-18hole day12' spacing 1 -1 0 spacing*d 12 -12 0
      spacing*d2 144 -144 0 spacing*d3 1728 -1728 0;
estimate '9hole-18hole day15' spacing 1 -1 0 spacing*d 15 -15 0
      spacing*d2 225 -225 0 spacing*d3 3375 -3375 0;
estimate '9hole-18hole day18' spacing 1 -1 0 spacing*d 18 -18 0
      spacing*d2 324 -324 0 spacing*d3 5832 -5832 0;
estimate '9hole-18hole day21' spacing 1 -1 0 spacing*d 21 -21 0
      spacing*d2 441 -441 0 spacing*d3 9261 -9261 0;
estimate '9hole-18hole day26' spacing 1 -1 0 spacing*d 26 -26 0
      spacing*d2 676 -676 0 spacing*d3 17576 -17576 0;
estimate '9hole-18hole day33' spacing 1 -1 0 spacing*d 33 -33 0
      spacing*d2 1089 -1089 0 spacing*d3 35937 -35937 0;
estimate '9hole-18hole day36' spacing 1 -1 0 spacing*d 36 -36 0
      spacing*d2 1296 -1296 0 spacing*d3 46656 -46656 0;
estimate '9hole-18hole day40' spacing 1 -1 0 spacing*d 40 -40 0
      spacing*d2 1600 -1600 0 spacing*d3 64000 -64000 0;
```

Similar coding can be written for Treatment 1 – Treatment 3 and Treatment 2 – Treatment 3. E.g. if we switched to estimation of '9hole-36hole' for each day, we would switch the numbers in the second and third position for each effect, i.e. SPACING 1 0 -1. The output for the above is shown in **Table 6.3**. It shows that there is no significant difference between Treatment 1 and Treatment 2 up until day 33, 36, and day 40. For these last few days, the difference is still only 3-4 grams. The difference between Treatment 1 and 3 was significantly different, except on day 33 and 36. Similarly, the difference between Treatment 2 and 3 was significantly different at each day, except on day 36 and 40.

Since we are interested in the last few time points of measurements, we can also estimate differences of weights between certain days within each treatment to answer the question on whether or not lettuces can be harvested sooner with similar yields.

```
estimate 'day33-day36 of 9hole' spacing*d -3 0 0 spacing*d2 -207 0 0
      spacing*d3 -10719 0 0;
estimate 'day33-day40 of 9hole' spacing*d -7 0 0 spacing*d2 -511 0 0
      spacing*d3 -28063 0 0;
```

```
estimate 'day36-day40 of 9hole' spacing*d -4 0 0 spacing*d2 -304 0 0
      spacing*d3 -17344 0 0;
estimate 'day33-day36 of 18hole' spacing*d 0 -3 0 spacing*d2 0 -207 0
      spacing*d3 0 -10719 0;
estimate 'day33-day40 of 18hole' spacing*d 0 -7 0 spacing*d2 0 -511 0
      spacing*d3 0 -28063 0;
estimate 'day36-day40 of 18hole' spacing*d 0 -4 0 spacing*d2 0 -304 0
      spacing*d3 0 -17344 0;
estimate 'day33-day36 of 36hole' spacing*d 0 0 -3 spacing*d2 0 0 -207
      spacing*d3 0 0 -10719;
estimate 'day33-day40 of 36hole' spacing*d 0 0 -7 spacing*d2 0 0 -511
      spacing*d3 0 0 -28063;
estimate 'day36-day40 of 36hole' spacing*d 0 0 -4 spacing*d2 0 0 -304
      spacing*d3 0 0 -17344;
```

The results show that there is significant difference between each pair of days within each treatment. This ranges from 2-12 grams for the difference of the weight between pairs of days. **Table 6.4** shows the output.

Estimates				
Label	Estimate	Standard Error	t Value	Pr > |t|
9hole-18hole day1	0.2261	0.9652	0.23	0.8155
9hole-18hole day5	-0.04255	0.9004	-0.05	0.9624
9hole-18hole day8	-0.05012	0.9355	-0.05	0.9574
9hole-18hole day12	0.1522	0.9853	0.15	0.8776
9hole-18hole day15	0.4319	1.0132	0.43	0.6711
9hole-18hole day18	0.7942	1.0435	0.76	0.4488
9hole-18hole day21	1.2149	1.0878	1.12	0.2674
9hole-18hole day26	1.9801	1.2024	1.65	0.1034
9hole-18hole day33	2.9971	1.4050	2.13	0.0356
9hole-18hole day36	3.3463	1.4975	2.23	0.0280
9hole-18hole day40	3.6722	1.6618	2.21	0.0302

Table 6.3: Partial SAS output for the estimates of the difference of Treatment 1 and 2 at each day.

Estimates				
Label	Estimate	Standard Error	t Value	Pr > |t|
day33-day36 of 9hole	-4.6822	0.1626	-28.79	<.0001
day33-day40 of 9hole	-11.9492	0.4553	-26.24	<.0001
day36-day40 of 9hole	-7.2670	0.2995	-24.26	<.0001
day33-day36 of 18hole	-4.3331	0.1626	-26.64	<.0001
day33-day40 of 18hole	-11.2742	0.4553	-24.76	<.0001
day36-day40 of 18hole	-6.9411	0.2995	-23.18	<.0001
day33-day36 of 36hole	-2.6941	0.1626	-16.56	<.0001
day33-day40 of 36hole	-6.4499	0.4553	-14.17	<.0001
day36-day40 of 36hole	-3.7557	0.2995	-12.54	<.0001

Table 6.4: SAS output for the estimates of differences between days within each treatment.

Conclusion

Starting with the harvest, we can see that Treatment 1 is significantly better than Treatment 2 by roughly 3.7 ± 1.7 grams. It is also, significantly better than Treatment 3 by 5.1 ± 1.7 grams. Thus, we conclude that on harvest day, Treatment 1 gives a higher yield. With an average of 88 grams per plant, we could harvest roughly 792 grams per raft. On the other hand, Treatment 2 would give a total of 1521 grams per raft and Treatment 3 would give roughly 2959.2 grams per raft. So even though the first treatment gives a significantly better yield, it may be more profitable to decide if losing a few grams per plant is worth keeping the original growing method of nine plants per raft during the winter months.

If the lettuces are harvested a week early, on day 33, Treatment 1 is still significantly better than Treatment 2, but there is not enough evidence that it is significantly better than Treatment 3. Also, Treatment 3 is significantly better than Treatment 2 by roughly 3.4 ± 1.4 grams. With an average of 679.5, 1305, and 2728.8 grams per raft for treatments 1, 2, and 3 respectively, it would be worth it to have 36 plants per raft during December and January.

Suppose it was decided to go with Treatment 1 with nine holes per raft. Is it worth harvesting earlier? From **Table 6.4** we can see that the difference from the weight on the 36^{th} day compared to the weight on the 40^{th} day is 7.3 ± 0.3 grams. Thus, we could roughly gain seven grams per plant by waiting another four days to harvest on day 40. If it was harvested a week earlier, we would lose almost 12 grams per plant. Suppose there is five rafts that get harvested each week. With a total of 45 plants, we would get roughly 3397.5 grams with an average of 75.5 grams per plant harvesting on day 33. Waiting until day 40 would give us about 3960 grams of lettuce with an average of 88 grams per plant. Thus, waiting one week, we could gain almost 600 more grams worth of lettuce and since it was harvested every week, it may be worth the extra week of growth during the time of December through January.

Similarly, if he went with Treatment 2, **Table 6.4** tells us that the difference of a week is about 11.3 ± 0.5 grams per plant. Again, suppose there are five rafts that get

harvested each week. This will give roughly 6525 grams for harvesting on day 33 and 7605 grams for harvesting on day 40 with an average of 72.5 and 84.5 grams per plant respectively. Which means we could gain almost 1100 grams waiting until day 40.

Lastly, looking at Treatment 3 from **Table 6.4**, we see that the difference from day 33 to day 40 is 6.4 ± 0.5 grams. Continuing with our example of five rafts per week, on day 33 we would get roughly 13,641 grams of lettuce with an average of 75.8 grams per plant. Also, on day 40 we might see a harvest of 14,796 grams with an average of 82.2 grams per plant. That is a difference of again, 1150 grams of yield by just waiting a week.

In conclusion, I would suggest setting up the greenhouse to harvest every week using Treatment 3 with 36 plants per raft. It would be worth the wait as well to let the lettuce grow until day 40, to get the extra 1150 grams. Even though the estimates are small in **Table 6.3** and **6.4**, it adds up when you multiply how many rafts and plants are being harvested in the long run. This can only be said for the growth months of December and January. The summer may produce similar or much different results, especially if the lettuce doubles in size and needs the space to grow. But as for the winter, it is safe to say that we can grow 36 of them together on one raft and gain $14{,}796 - 3960 = 10{,}836$ grams of lettuce! This would significantly increase profits comparing to the current process of nine plants per raft.

References

Archontoulis, S. V. & Miguez, F. E. (2015). Nonlinear Regression Models and Applications in Agricultural Research. *Agron. J. 105*: 1-13.

Henderson, C.R. (1950). Estimation of genetic parameters. *Ann. Math. Stat.* 21, 309.

Henderson, C.R. (1963). Selection index and expected genetic advance. In: Statistical Genetics and Plant Breeding. Hanson, W. D. & Robinson, H. F. (Eds.), pp. 141-63. National Academy of Sciences-Nationial Research Council, Washington. Publication 982.

Henderson, C.R. (1973). Sire evaluations and genetic trends. Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush, Champaign, Illinois: American Society of Animal Science, 10.

Jiang, J. (2007). Linear and Generalized Linear Mixed Models and Their Applications. New York, NY. Springer.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Tutorial in biostatistics: modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine* 19, 1793-1918.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). SAS for Mixed Models. Cary, NC. SAS Institute Inc.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Sciences.* 6, 15.

SAS/STAT 9.2 User's Guide (Book Excerpt): The MIXED Procedure. (2008). Cary, NC. SAS Institute Inc.

Wang, Z. & Goonewardene, L. A. (2004). The use of MIXED models in the analysis of animal experiments with repeated measures data. *Can. J. Anim. Sci.* 84, 1.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. New York, NY. Springer.

Appendix

Appendix (A):

$$l(\boldsymbol{\beta},\boldsymbol{\theta}) = c - \frac{1}{2}\log(|\boldsymbol{V}|) - \frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X\beta})'\boldsymbol{V^{-1}}(\boldsymbol{Y}-\boldsymbol{X\beta})$$

We can rewrite this equation as

$$l(\boldsymbol{\beta},\boldsymbol{\theta}) = c - \frac{1}{2}\log(|\boldsymbol{V}|) - \frac{1}{2}(\boldsymbol{Y'V^{-1}Y} - \boldsymbol{Y'V^{-1}X\beta} - \boldsymbol{\beta'X'V^{-1}Y} + \boldsymbol{\beta'X'V^{-1}X\beta}).$$

Now differentiating with respect to $\boldsymbol{\beta}$ we get that

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = -\frac{1}{2}\{-\boldsymbol{Y'V^{-1}X} - (\boldsymbol{X'V^{-1}Y})' + [\boldsymbol{\beta'X'V^{-1}X} + (\boldsymbol{X'V^{-1}X\beta})']\}$$

$$= -\frac{1}{2}(-2\boldsymbol{Y'V^{-1}X} + 2\boldsymbol{\beta'X'V^{-1}X})$$

$$= -\boldsymbol{Y'V^{-1}X} + \boldsymbol{\beta'X'V^{-1}X}\,.$$

Equating this to zero and taking the transpose of both sides gives

$$\boldsymbol{X'V^{-1}X\beta} = \boldsymbol{X'V^{-1}Y}.$$

Now solving for $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = (\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}Y}.$$

Differentiating $l(\boldsymbol{\beta},\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ gives the following:

$$\frac{\partial l}{\partial \theta_r} = -\frac{1}{2}tr\left(\boldsymbol{V^{-1}}\frac{\partial \boldsymbol{V}}{\partial \theta_r}\right) - \frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X\beta})'\left[-\boldsymbol{V^{-1}}\frac{\partial \boldsymbol{V}}{\partial \theta_r}\boldsymbol{V^{-1}}\right](\boldsymbol{Y}-\boldsymbol{X\beta})$$

$$= -\frac{1}{2}\left\{(\boldsymbol{Y}-\boldsymbol{X\beta})'\left[\boldsymbol{V^{-1}}\frac{\partial \boldsymbol{V}}{\partial \theta_r}\boldsymbol{V^{-1}}\right](\boldsymbol{Y}-\boldsymbol{X\beta}) - tr\left(\boldsymbol{V^{-1}}\frac{\partial \boldsymbol{V}}{\partial \theta_r}\right)\right\}.$$

Appendix (B)

Now in order to get Henderson's mixed model equations in (3.7) we start with the following:

$f(Y,u) = f(Y|u)f(u)$ where $Y|u = Y - X\beta - Zu = e \sim N(0,R)$ and $u \sim N(0,G)$ thus,

$$f(Y|u) = \frac{1}{(2\pi)^{n/2}|R|^{1/2}}\exp\left\{-\frac{1}{2}(Y-X\beta-Zu)'R^{-1}(Y-X\beta-Zu)\right\}$$

$$f(u) = \frac{1}{(2\pi)^{q/2}|G|^{1/2}}\exp\left\{-\frac{1}{2}u'G^{-1}u\right\}$$

combining gives equation (3.8).

Maximizing (3.8) requires minimizing what is in the exponential:

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{Y - X\beta - Zu} \end{pmatrix}' \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix}^{-1} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{Y - X\beta - Zu} \end{pmatrix}$$

$$= \boldsymbol{u'G^{-1}u} + (\boldsymbol{Y - X\beta - Zu})'\boldsymbol{R^{-1}}(\boldsymbol{Y - X\beta - Zu})$$

$$= \boldsymbol{u'G^{-1}u} + \boldsymbol{Y'R^{-1}Y} - \boldsymbol{Y'R^{-1}X\beta} - \boldsymbol{Y'R^{-1}Zu} - \boldsymbol{\beta'X'R^{-1}Y} + \boldsymbol{\beta'X'R^{-1}X\beta}$$

$$+ \boldsymbol{\beta'X'R^{-1}Zu} - \boldsymbol{u'Z'R^{-1}Y} + \boldsymbol{u'Z'R^{-1}X\beta} + \boldsymbol{u'Z'R^{-1}Zu}.$$

Differentiating with respect to $\boldsymbol{\beta}$ we have

$$\frac{\partial f(\boldsymbol{Y},\boldsymbol{u})}{\partial \boldsymbol{\beta}} = -\boldsymbol{Y'R^{-1}X} - (\boldsymbol{X'R^{-1}Y})' + [\boldsymbol{\beta'X'R^{-1}X} + (\boldsymbol{X'R^{-1}X\beta})'] + (\boldsymbol{X'R^{-1}Zu})'$$

$$+ \boldsymbol{u'Z'R^{-1}X} = -2\boldsymbol{Y'R^{-1}X} + 2\boldsymbol{\beta'X'R^{-1}X} + 2\boldsymbol{u'Z'R^{-1}X}$$

and setting equal to zero gives the transpose of the first equation in (3.7):

$$\boldsymbol{\beta'X'R^{-1}X} + \boldsymbol{u'Z'R^{-1}X} = \boldsymbol{Y'R^{-1}X}.$$

Now differentiating with respect to $\boldsymbol{u}$ we have

$$\frac{\partial f(\boldsymbol{Y},\boldsymbol{u})}{\partial \boldsymbol{u}} = [\boldsymbol{u'G^{-1}} + (\boldsymbol{u'G^{-1}})'] - \boldsymbol{Y'R^{-1}Z} + \boldsymbol{\beta'X'R^{-1}Z} - (\boldsymbol{Z'R^{-1}Y})' + (\boldsymbol{Z'R^{-1}X\beta})'$$

$$+ [\boldsymbol{u'Z'R^{-1}Z} + (\boldsymbol{Z'R^{-1}Zu})']$$

$$= 2\boldsymbol{u'G^{-1}} - 2\boldsymbol{Y'G^{-1}Z} + 2\boldsymbol{\beta'X'R^{-1}Z} + 2\boldsymbol{u'Z'R^{-1}Z}$$

and equating to zero gives the transpose of the second equation in (3.7):

$$\boldsymbol{\beta'X'R^{-1}Z} + \boldsymbol{u'}(\boldsymbol{G^{-1}} + \boldsymbol{Z'R^{-1}Z}) = \boldsymbol{Y'R^{-1}Z}.$$

Lastly, by taking the transpose of both sides of the equations above gives Henderson's mixed model equations in (3.7).

Appendix (C)

Data set:

Trt.1	Day										
plant	**1**	**5**	**8**	**12**	**15**	**18**	**21**	**26**	**33**	**36**	**40**
1	44	44.1	46	48.1	49.2	50	53	60.4	65.7	71	75.3
2	48.4	48.6	51.1	53	54.1	55.5	58.7	65.7	69.2	73.7	76
3	55.5	56.5	59.8	63.1	65.9	68.3	74.5	83.4	91	98.5	103.9
4	48.6	49.5	52.8	55.1	56.8	58	63	73	81.2	87.8	94
5	48.1	48.3	51.4	53.1	53.2	53.2	57.5	65	71	77.4	84.6
6	42.6	44.8	46.2	48.6	49.7	51.3	56	63.1	69.9	76.5	84.1
7	55.4	56.8	57.9	59.7	61	62.2	68	77.8	85.2	92	97.4
8	47.5	47.6	48.3	49.6	50.5	51.5	55.9	66.5	74	81.5	86.7
9	48.8	49.5	50.2	51.7	52.4	53.5	56.7	64.1	71	76.3	82.3
10	51.2	53.1	54.3	53.4	54.9	55.8	58.8	66	72.1	79	86.2
11	49.4	50.3	51.5	53.1	54	55.7	60.1	68	74.7	81	88.7
12	50.4	50.8	52.3	53.9	55.1	57	61	69.4	76.1	82.6	87.4
13	50	50.7	52	54.1	55.6	57.6	60.3	68.9	76	81.9	86.5
14	48	48.5	50.1	52.6	53.4	54.8	56.9	65.3	71.8	78.5	85.3
15	52.3	53.1	55.7	59.3	60.4	62	68.7	78.8	84.2	93.4	101
16	51.1	51.6	53.9	55.2	56.7	58.3	62	71.2	77.9	84	91.2
17	46.7	47.2	49.1	50.8	52.3	54.1	56.7	66	72.5	79.8	86.1
18	48.8	49.1	50.5	52.1	53.6	55	59.6	68.9	76	82.1	87.3

Trt.2	Day										
plant	**1**	**5**	**8**	**12**	**15**	**18**	**21**	**26**	**33**	**36**	**40**
1	48.1	48.3	48.6	49	51	52	56	63	67.5	73.2	75.9
2	48.2	49.8	51.1	52.7	53.1	54	58.7	66.7	73	81	84.4
3	51	51.3	53.7	55.2	56.7	58.8	62	71.2	76.9	85.1	89.5
4	47.4	48.6	50	51.3	52.5	53.6	56.1	63.1	68.5	73	78.7
5	51.2	51.4	53.4	55.3	56.4	57.8	60.1	65	70	73.3	80.4
6	50.6	50.7	52.1	54.5	54.9	55.5	59.8	68.9	76.7	86.5	89.6
7	50.1	57.4	58.7	56.8	58.7	59.8	62.4	70	77.1	83.8	88.9
8	44	45.1	46	46.4	48.1	49	53	59.7	65.7	73	79.8
9	52.8	54.7	57.2	59.8	61.1	62	65.9	73.1	76.9	84.6	88.9

10	53.1	55.1	56.8	58.6	59.1	59.7	65.1	75	80.5	87.3	91.1
11	48.4	50	54.2	57	58.8	60.5	64.8	77.6	86	96.1	98.6
12	48.2	49.7	52.1	53	53.8	54.8	57	63.9	69	74.6	80.1
13	49.8	50.9	52.8	54.1	54.7	55.7	58.9	64	69.7	72.1	78.9
14	46.4	47.8	49	50	50.8	51.6	53.2	59.7	63.4	67.3	75.4
15	51.8	53.7	54.9	56.6	57.8	59.9	67.8	76	84.5	94.6	98.7
16	42.1	43.1	44.7	46.5	48.1	50.3	55	63.8	70	78.1	87.3
17	46.7	48.4	49.1	50.7	51	51.5	53.1	58.5	61.2	67.1	76.4
18	48.6	49.4	50.5	51.8	52.9	54	58.2	63	68.7	72.1	77.8

Trt.3	**Day**										
plant	**1**	**5**	**8**	**12**	**15**	**18**	**21**	**26**	**33**	**36**	**40**
1	53.1	55.4	56.1	56.4	54	54.6	59	67.5	77	80.3	82.1
2	51.4	51.7	52.3	54.8	56.9	59.7	61.5	67	80.2	84.5	85.1
3	49.3	52.8	54.8	56.2	58.1	59.6	62.3	68.6	73.9	79.5	81.3
4	48.1	49.8	51.1	52.8	54.2	55.8	57.9	65.8	75	80.1	81
5	48.3	48.2	49.1	51.3	53.1	56.5	59.8	68.9	77.9	83.6	84.7
6	45.9	49	51.7	53	56.2	58.5	59.9	65.1	74	78.7	80.2
7	47.3	57	57.4	58.1	58.6	58.8	61	64	73.9	79	80.4
8	44.8	51.2	53.3	55.6	58	60.5	62.4	69.8	80.4	86.9	87.1
9	52.1	52.4	53.6	54.9	55.9	57.1	58	61.2	69	72.3	75.4
10	54.4	55.1	56	57.1	59	59.8	62.8	72.1	82.5	87.2	88.2
11	48.1	48.2	49.7	51.4	53.1	54.7	57	62	74.1	77.2	79.5
12	52.3	53.5	55.2	57.1	58.2	59.8	60.3	63.5	74.3	78.5	80
13	48.7	50.1	50.5	50.8	52.5	54.2	56.1	59.9	69.7	74.3	75.2
14	52.1	54	56.4	58.8	60.1	61	63	68.6	74	79.8	80.9
15	48.9	50	52.1	52.6	54	56.1	59.2	65.9	75.7	80.6	81
16	51.1	51.9	53.1	56.7	59.1	61.1	64.2	73	87.6	95.7	96.3
17	50.4	57.1	52.8	55.4	57	59.4	61.2	65.3	73.4	79.2	81.1
18	53	53.3	54.1	55.3	55.8	56.5	58.7	63.1	71.5	78	79.3