A Model Suggestion for Cellulose Synthase Complex Data

STAT 550 Report

Tom Hyeongcheol Park

1.Summary

A biologist group have conducted an experiment on Arabidopsis thaliana seedlings to figure out which factors affect the growth of the plant. The velocity of Cellulose Synthase Complex particle(CSC) is known as a factor of the growth. The group wants to know if a genetic type of the plant or a location of CSC can explain the velocity difference. They collected four variables: genetic type of plants(Type), velocity of particle(Velocity), the location of the particle(MT), and which seedling the particle comes from(Seedling).

The goal of this report is to suggest a proper statistical model for the data. It is found that linear mixed effect model with nested random effect is the best for the analysis. Advantage and disadvantage of the model are explained comparing the model to another candidate model. It is seen that the suggested model has higher chance of making correct decision when there is no difference of velocities. On the other hand, the model has lower chance of making correct decision when there exists a difference of velocities. However, collecting more seedlings can cure the disadvantage. Two simulations are done to prove the arguments.

2. Introduction

Cellulose Synthase Complex(CSC) plays a major role in the growth of Arabidopsis thaliana. The CSCs move within microtubule domains producing cellulose into the cell. Hence, the CSCs velocity is an important factor of growth. The genetically mutated plant, mor1-1 has been known to disorganize microtubule networks which may lead to slower CSCs velocity. A group of biologists wants to see if the CSCs velocity of normal type Arabidopsis thaliana is different from the CSCs velocity of the mutated one. They also want to check if there is a velocity difference between CSCs going outside of microtubule domains and CSCs staying inside of the domains. To check the velocity difference, the biologists have collected sample data from each type of plants. The goal of this report is to suggest a proper statistical model to analyze the data. The rest of the report is arranged as follows. The data are explained along with how the data was collected. Based on the context of the data, an adequate model is suggested comparing the model with another candidate model. Following simulation results support the suggested model. Finally, we summarize our discussion and brief advantage and disadvantage of the model.

3. Proposed Statistical Method

3.1 Data description

Four cloned seedlings are randomly chosen from wild type plants and three cloned seedlings are randomly chosen from mor1-1 type plants. Different numbers of CSC velocities are measured from each seedling. Also, for each seedling, numbers of CSCs inside the domain is different from numbers of CSCs outside the domain. The sample size of CSC velocities for wild type is 491 and the sample size of CSC velocities for mor1-1 type is 497. Type, MT, and Seedling are categorical variables and Velocity is continuous variable.

Variable	Scale	Explanation
Type	Categorical	Genetic type of plant: ‘wt’ (Wild Type) and ‘mor1-1’ (mor1-1)
MT	Binary	The location of particle. ‘Y’(Inside of the microtubule domain), ‘N’(Outside of the domain
Seedling	Categorical	Index of the seedlings. ‘1’,’2’,’3’,’4’ for Wild Type, ‘1’,’2’,’3’ for mor1-1 plants.
Velocity	Continuous	The velocity of CSC. Response variable

3.2 Proposed analysis: linear mixed effect model with nested random effect

Linear mixed effect model with nested random effect(LMEM) is recommended for the two purposes: the CSCs velocity difference between the two plant types, the velocity difference between CSCs inside of the domain and CSCs outside of the domains. The model has three factors. The variables, Type and MT are fixed effects and Seedling is a nested random effect. There are some reasons why we call the Type and MT as fixed effects. First, the variables are under control and fixed by the biologists. Secondly, they are always our interest even we replicate the same experiment several times. On the other hand, we call the seedlings as random effect because they are randomly drawn. Another explanation of their randomness is that we do not know what distinct traits the seedlings have. To be more specific, the biologists cannot control the different characteristics of the seedlings. Seedlings can only have one genetic type so that we say the random effect(Seedling) is nested within the fixed effect(Type). In other words, the fact that a couple of randomly chosen seedlings are selected from each type implies the types nest the seedlings.

The main reason of choosing the model is that the model can catch some differences(variances) between the same type of seedlings. Many observations of CSCs are from a single seedling, implying that the observations from one seedling share same traits together. Besides, there may exists some interactions between the CSCs within a seedling. Hence, the relation between two CSCs within a seedling is not the same as the relation between two CSCs each from different seedlings. This means that we cannot regard every single CSC as independent observation. We should make a group of CSCs for each seedling and differentiate them with other seedling groups. The differences between seedlings may affect our result unless we eliminate those.

One may argue that we should use linear model or analysis of variance(ANOVA) model. However, the models have a serious shortage. They ignore the differences between seedlings. They assume that CSCs are all independent and do not allow grouping. The wrong assumption easily leads us to a wrong conclusion. That is, we have higher chance to conclude that there is no difference of CSCs velocities between the two types even though it is not true. The probability of wrong decision is called Type 1 error. The type 1 error gets higher when there is more variance(difference) between seedlings and less variance within each seedling. In contrast, The LMEN has lower and consistent type 1 error comparing to the other two models.

However, he LMEM has a weak point. The drawback is that we may not be able to detect the velocity difference between the two types, even though there exists a true difference. Same situation is likely to happen to the velocity difference between the position of CSCs(variable MT). The probability that a model catches the true difference when there exists the difference is called power. That is to say, our model may not have enough power. This is because the model suffers from the lack of enough samples(seedlings). The model uses only 4 samples for wild type plants and 3 samples for mur1-1 type plants. We are using too small number of examples for each type, which endangers our conclusion.

Therefore, to collect more seedlings for each type is highly recommended. However, we do not have to collect many CSCs velocities for each seedling when we collect more data. This is because the number of seedlings is more important than the number of CSC velocities for each seedling[i]. Many CSCs velocities for each seedling only elaborate the representative CSC velocities for each seedling.

3.2.2 Simulation

Two simulations are conducted. Both simulations ignore the variable MT for simplicity. The numbers of particles for each seedling are all set equal. The numbers of seedlings for each type are also set equal. Figure 1-1 shows that one of the improper model, linear regression has higher type 1 error rate. Especially the error rate goes worse when the seedlings in a type are more different to one another and the CSCs in a seedling are more similar to one another. On the other hand, our model's type 1 error rate is ideally low and consistent.

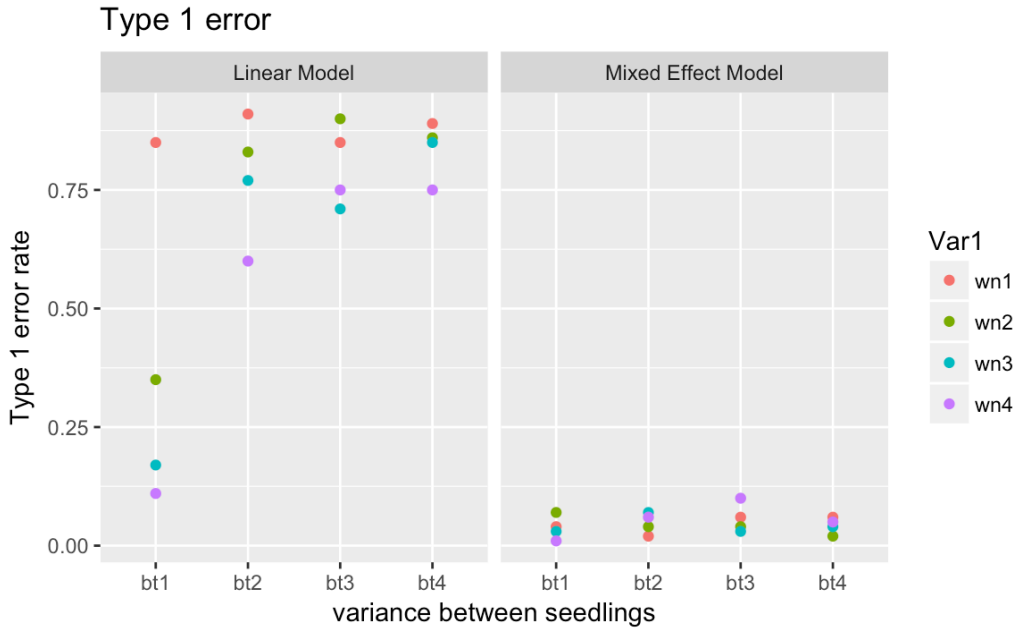

Figure 1

In this plot, Var1(wn1, wn2, wn3, wn4) means variance within seedling. Bigger suffix number means bigger variance(wn1=0.1, wn2=0.6,wn3=1.1, wn4=1.6). Likewise, bt1, bt2, bt3, bt4 means variance between seedlings(bt1=0.1, bt2=0.6,bt3=1.1, bt4=1.6). The simulation assumes that the seedlings and the CSC particles follow Normal distribution with mean zero. The number of simulation is 100.

Figure 2 is to verify how large number of seedlings increases the power of our model. The number of CSCs within a seedling is set to five, which is regarded as a small number. It clearly shows that number

of seedlings cure the lower power problem even the number of CSCs for each seedling is relatively small.

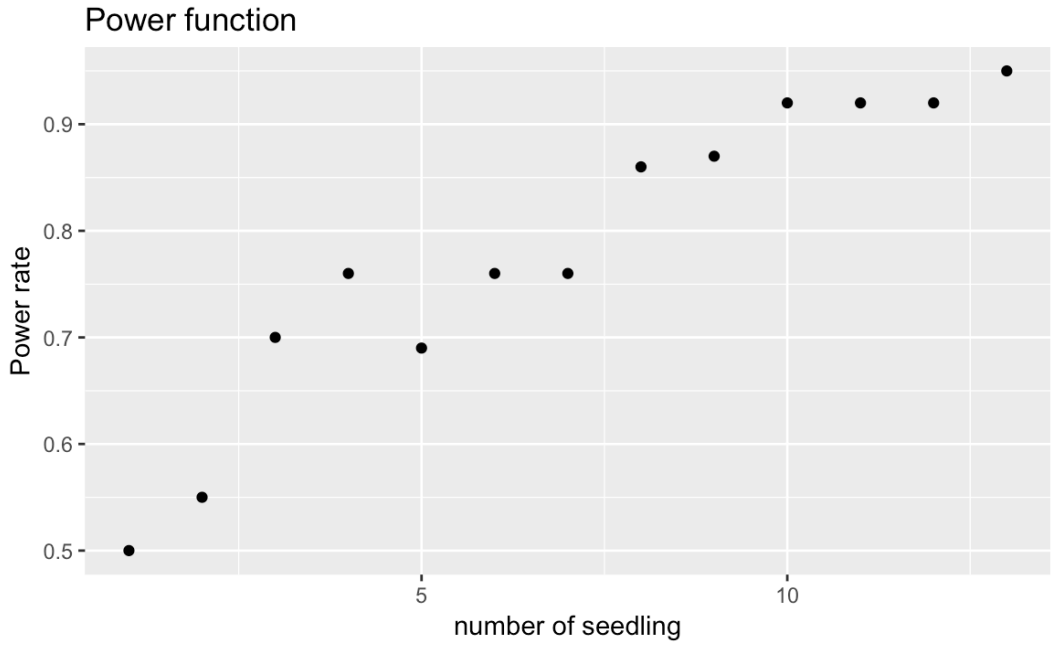

Figure 1

The simulation assumes that the true mean difference of velocities between the two types is 1. The variance between the seedlings is set to 1. For each seedling, CSC particles are set to follow the standard normal distribution. The number of simulation is 100.

4.conclusion

It is found that linear mixed effect model with nested random effect is appropriate for the analysis. The model is free from the concern about inflated type 1 error, whereas it may suffer from lower power problem. However, the problem can be cured by adding more samples(seedlings) to our model.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(reshape2)
library(tidyverse)
library(nlme)

MEM_power <- function(mean_difference,between_var,within_var,n_of_simulation) {
  n=120
  count_p <- rep(NA, n_of_simulation)
  for (i in 1:n_of_simulation) {
    mean_x1 <- rnorm(3,0,between_var)
    mean_x2<- rnorm(3,mean_difference,between_var)
    x11 <- rnorm(n,mean_x1[1],within_var)
    x12 <- rnorm(n,mean_x1[2],within_var)
    x13 <- rnorm(n,mean_x1[3],within_var)
    x21 <- rnorm(n,mean_x2[1],within_var)
    x22 <- rnorm(n,mean_x2[2],within_var)
    x23 <- rnorm(n,mean_x2[3],within_var)
    mydata <- data.frame(x11,x12,x13,x21,x22,x23)
    mydata <- mydata %>% gather(colnames(mydata),key='seedling',value='velocity') %>%
mutate(type=c(rep("wt",3*n),rep("m",3*n)))
    lm1 <- lme(velocity ~ type, random=~1|seedling,data=mydata)
    p_value <- summary(lm1)$tTable[2,5]
    count_p[i] <- p_value < 0.05
  }
  pwr <- mean(count_p)
  return(pwr)
}

LR_power <- function(mean_difference,between_var,within_var,n_of_simulation) {
  n=120
  count_p <- rep(NA, n_of_simulation)
  for (i in 1:n_of_simulation) {
    mean_x1 <- rnorm(3,0,between_var)
    mean_x2<- rnorm(3,mean_difference,between_var)
    x11 <- rnorm(n,mean_x1[1],within_var)
    x12 <- rnorm(n,mean_x1[2],within_var)
    x13 <- rnorm(n,mean_x1[3],within_var)
    x21 <- rnorm(n,mean_x2[1],within_var)
    x22 <- rnorm(n,mean_x2[2],within_var)
    x23 <- rnorm(n,mean_x2[3],within_var)
    x1 <- c(x11,x12,x13)
    x2 <- c(x21,x22,x23)
    mydata <- data.frame(x1,x2)
    mydata <- mydata %>% gather(colnames(mydata),key='seedling',value='velocity') %>%
      mutate(type=c(rep("wt",3*n),rep("m",3*n)))
    tt1 <- t.test(velocity ~ type,data=mydata)
    p_value <- tt1$p.value
    count_p[i] <- p_value < 0.05
```

```
  }
  pwr <- mean(count_p)
  return(pwr)
}

within_var <- seq(0.1, 1.6, 0.5)
between_var <- seq(0.1, 1.6, 0.5)
type1_matrix <- matrix(NA, 4, 4)
for (i in 1:length(within_var)) {
  for (j in 1:length(between_var)){
    type1_matrix[i,j] <- LR_power(0,between_var[j],within_var[i],100)
  }
}

colnames(type1_matrix) <- c('bt1','bt2','bt3','bt4')
rownames(type1_matrix) <- c('wn1','wn2','wn3','wn4')
type1_matrix
type1_d.f.LR <- as.data.frame(as.table(type1_matrix))
type1_d.f.LR["LR"] <- NA
type1_d.f.LR$LR <- rep(1,16)
type1_d.f
ggplot(type1_d.f)+geom_point(aes(Var2,Freq,color=Var1))+labs(title = "Type 1 error")+xlab("variance
between seedlings")+ylab("Type 1 error rate")

within_var <- seq(0.1, 1.6, 0.5)
between_var <- seq(0.1, 1.6, 0.5)
type1_matrix <- matrix(NA, 4, 4)
for (i in 1:length(within_var)) {
  for (j in 1:length(between_var)){
    type1_matrix[i,j] <- LR_power(0,between_var[j],within_var[i],100)
  }
}

colnames(type1_matrix) <- c('bt1','bt2','bt3','bt4')
rownames(type1_matrix) <- c('wn1','wn2','wn3','wn4')

type1_d.f.LR <- as.data.frame(as.table(type1_matrix))
type1_d.f.LR["LR"]<- NA
type1_d.f.LR$LR <- rep("Linear Model",16)

type1_matrix <- matrix(NA, 4, 4)
for (i in 1:length(within_var)) {
  for (j in 1:length(between_var)){
    type1_matrix[i,j] <- MEM_power(0,between_var[j],within_var[i],100)
  }
}

colnames(type1_matrix) <- c('bt1','bt2','bt3','bt4')
rownames(type1_matrix) <- c('wn1','wn2','wn3','wn4')
type1_d.f.ME <- as.data.frame(as.table(type1_matrix))
type1_d.f.ME["LR"]<- NA
```

```
type1_d.f.ME$LR <- rep("Mixed Effect Model",16)
type_1 <- rbind(type1_d.f.ME,type1_d.f.LR)

ggplot(type_1)+geom_point(aes(Var2,Freq,color=Var1))+facet_wrap(~LR)+labs(title = "Type 1
error")+xlab("variance between seedlings")+ylab("Type 1 error rate")

pwr_MEM <- function(seedling){
  n_of_simulation=100
  mean_diff=1
  btw_var=1
  wtn_var=1
  pwr_vct <- rep(NA,seedling-2)
  for (i in 1:(seedling-2)) {
    count_p <- rep(NA, n_of_simulation)
    for (j in 1:n_of_simulation){
      seed_means.1 <- rnorm(i+2,0,1)
      seed_means.2 <- rnorm(i+2,1,1)
      data.f <-
data_frame(type=c(rep('wt',5*(i+2)),rep('mt',5*(i+2))),seedling=rep(NA,10*(i+2)),velocity=rep(NA,10*(i
+2)))
      seedling.vec.1 <-seedling.vec.2 <-  velocity.vec.2  <- velocity.vec.1 <- c()

      for (k in 1:(i+2)) {
        velocity.vec.1 <- c(velocity.vec.1, rnorm(5, seed_means.1[k],1))
        velocity.vec.2 <- c(velocity.vec.2, rnorm(5, seed_means.2[k],1))
        seedling.vec.1 <- c(seedling.vec.1,rep(k, 5))
        seedling.vec.2 <- c(seedling.vec.2,rep(k, 5))
      }
      data.f$velocity <- c(velocity.vec.1,velocity.vec.2)
      data.f$seedling <- c(seedling.vec.1,seedling.vec.2)
    lm1 <- lme(velocity ~ type, random=~1|seedling,data=data.f)
    p_value <- summary(lm1)$tTable[2,5]
    count_p[j] <- p_value < 0.05
    }
    pwr_vct[i] <- sum(count_p)/n_of_simulation
  }
  return (pwr_vct)
}

y <- pwr_MEM(15)
x <- seq(3:15)
df.final <- data.frame(y,x)
ggplot(df.final)+geom_point(aes(x=x,y=y))+labs(title = "Power function")+xlab("number of
seedling")+ylab("Power rate")
```

---

[i] Heather Woltman, Andrea Feldstain, J. Christine MacKay, Meredith Rocchi An introduction to hierarchical linear modeling Tutorials in Quantitative Methods for Psychology 2012, Vol. 8(1), p. 52-69