Modeling Correlated Binary Outcomes in the Morning Glory Family

Ruiyu Yang, Elizabeth Housworth

April 2017

1 Background and Introduction

1.1 What are comparative phylogenetic methods?

Phylogenetic comparative methods are used for statistical inference on species taking into consideration their phylogenetic information. A commonly-asked question is whether two or more traits of a number of species from one family are associated. For instance, whether the fruits of angiosperm plants evolved in adaptation to the size of animals that are in charge of their seed dispersal [6]; or whether the gregarious social behavior of insect larvae evolved in response to their alarming coloration [10]. Some other questions are asked as well, such as what the rate of phenotypic evolution is or what the ancestral states of two extant species are. But here we focus on the association study of the traits.

Let's look at the following fictitious example a family of four species. Suppose that we have two traits of this family of species: the seed mass and their fruit size. The phylogenetic tree is given. People might ask if there is a correlation between seed mass and the size of the fruit. Or has evolutionary changes in seed mass caused their fruits to grow bigger?

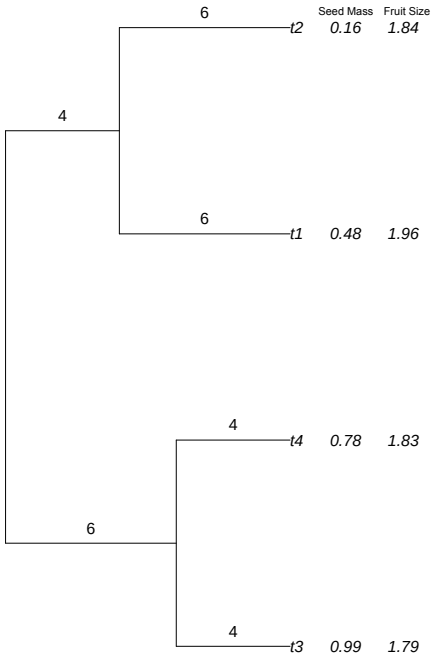

Figure 1: A fictitious example of a phylogeny on a family of four species with two traits listed beside the tips. It is of interest whether the evolution of seed mass and the size of fruits are independent or one happened in response to the other. This example is made up solely for the purpose of illustration.

Example 1.1. Suppose we have a phylogenetic tree with four species and two traits shown in Figure 1. It is of our interest whether the two traits have evolved independently, or that the fruit size has evolved in response to different seed mass. Here we assume that the traits in this fictitious phylogeny evolved according to Brownian motion to make one of the most common phylogenetic comparative method – phylogenetically independent contrast (PIC) applicable. In this approach the values for the traits at all ancestral nodes are first approximated [4] and then contrasts are taken for all sister species for all traits respectively and the relevant regression is formulated as

$$\nabla y_{ij} = \beta_i \nabla x_{ij} + \sqrt{v_{ij}}\epsilon_{ij},$$

in which $\nabla y_{ij} = y_i - y_j$ and $\nabla x_{ij} = x_i - x_j$ for sister species i and j (they are on adjacent nodes/tips) with ϵ_{ij} follows normal distribution with mean 0 and variance σ^2; v_{ij} is the variance involved in Brownian motion and is taken to be the sum of corrected branch lengths under nodes/tips i and j in [4]. Since Brownian motion proceeds independently on each branch the contrasts x_{ij}'s are independent and so are y_{ij}'s.

A more general method, similar to PIC, is a generalized least square (GLS) approach that incorporates the correlation information among the covariates directly in the error structure as below, where

$$\textit{fruit Size} = \textit{Seed Mass} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon},$$

with ϵ_i following normal distribution with mean zero and variance σ, and $\mathbb{E}(\epsilon_i \epsilon_j) = \rho_{ij}\sigma^2$ where R_{ij} is the phylogenetic correlation between species i and j. In our application, we take ρ_{ij} to be the sum of branch lengths over the segments where species i and j shared evolution over the total evolution time for species j.

$$R = \begin{array}{c} \begin{array}{cccc} \text{t1} & \text{t2} & \text{t3} & \text{t4} \end{array} \\ \left[\begin{array}{cccc} 1 & 0.4 & 0 & 0 \\ 0.4 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.6 \\ 0 & 0 & 0.6 & 1 \end{array}\right] \end{array} \tag{1}$$

In general, to uncover whether the variance in one trait can be explained by other traits, we can model them using a regression model such as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ represents the response variables or transformed responses (e.g. using logit function) while $\mathbf{X}$ stores observations for some trait(s), $\boldsymbol{\beta}$ is the regression parameter and $\boldsymbol{\epsilon}$ is the error structure. In a simple regression model the error $\boldsymbol{\epsilon}$ is often assumed to follow a Gaussian distribution with an identity matrix as its variance-covariance matrix. However, in that scenario there lies an hidden assumption that the observations/responses are independent of each other which is not the case for phylogenetic data.

1.2 Some common methods for comparative phylogenetic studies

One of the most common methods is PIC proposed by [3] as mentioned in Example 1.1. This approach is devised under the assumption that the evolution of the species follow a Brownian motion model. In this method the trait divergence are obtained for pairs of sister species to deal with the dependence among data. However this method is not generalized enough to be applied to binary traits as Brownian motion evolution assumption is constrained to continuous variables. Another method applicable only to continuous response variables using generalized least squares (GLS) was later developed by [5]. The dependence among data is incorporated in the error structure with a variance-covariance matrix. The method was later extended by [9] to incorporate within-species variation and measurement error to more precisely determine the error structure.

1.3 What are our data, goal, and approach in this project?

In biology, hereditary symbiosis is when the symbiont is vertically transmitted from mother to offspring through seeds or eggs. It is to biologists' greatest interest whether these interactions are mutually beneficial or just random events. Previous studies on grasses infected by fungal endophytes that are vertically transmitted from mother to offspring have shown that symbiotic plants are better equipped to defend against herbivory

than non-symbiotic ones due to the production of toxic alkaloids by the endophyte [1, 2]. Thus over time the symbiotic plants tend to be dominant in the community. Similar phenomenon is observed in the morning glory family where many species are infected and contain highly-concentrated alkaloids. In particular we are going to investigate whether there is hidden relationships between presence of the symbiont and the vigor of the species.

In our case, we have three traits for each species in the morning glories: the number of samples, the number of samples found to have ergot alkaloid, and its average seed mass. Our interest is to find out whether the presence of ergot alkaloid is correlated with the seed mass, taking into consideration the phylogenetic dependence. Note that due to the binary nature of our response variable, we cannot adopt PIC nor GLS for their requirement of continuity of the response variables.

add a table

In this project we adopted the approach of generalized estimating equations (GEE) proposed by [8]. It is an extension of generalized linear models (GLM) in which the mean of the response variable is assumed to be in the exponential family, such as Gaussian, Exponential, Poisson, and Binomial distributions, so the models can be applied to both continuous and categorical response variables. In essence, GLM are regression models that incorporates the correlation among the covariates. Suppose we have N repeated measurements at times $t = 1, 2, \ldots, N$ for the i-th subject. (We think of the family as one subject and the species traits as the repeated observations of the same family at different times $t = 1, 2, \ldots, N$.) At each measurement, we have a response variable $y_{i,t} \in \mathbb{R}^{1\times 1}$ and some covariates $\mathbf{x_{i,t}} = (1, x_{i,t,1}, x_{i,t,2}, \ldots, x_{i,t,m}) \in \mathbb{R}^{(m+1)\times 1}$, with m being the number of covariates. Then the regression model is that

$$g(\mathbb{E}(y_{i,t})) = \mathbf{x}_{i,t}^T \boldsymbol{\beta}_i,$$

where $\boldsymbol{\beta}_i = (\beta_{i,0}, \beta_{i,1}, \ldots, \beta_{i,m})$ is the regression parameter and g is a link function, such as probit and logit functions. In GLM the variance for in the response variable is modeled by

$$\mathrm{Var}(y_{i,t}) = \phi \mathscr{V}(\mathbb{E}(y_{i,t}))$$

where ϕ is the dispersion parameter and $\mathscr{V}$ is the variance function, both defined with respect to the distribution for $y_{i,t}$ (in the exponential family). However the variance only holds when all responses $y_{i,1}, y_{i,2}, \ldots, y_{i,N}$ are independent, which is not the case when the response variables are phenotypic traits of a phylogeny. Instead the variance for all responses is formulated as

$$V_i = \phi A_i^{1/2} R_i A_i^{1/2}$$

in GEE, where A_i is the diagonal response variance matrix assuming that the response variables are independent, i.e. the (t, t)-th element of A_i is $\mathrm{Var}(y_{i,t})$; R_i is the correlation matrix among the responses $y_{i,1}, \ldots, y_{i,N}$. If the responses are assumed to be independent then R_i is the identity matrix. Thus the estimating equations become

$$\sum_i \frac{\partial g^{-1}(\mathbf{X}_i^T \boldsymbol{\beta_i})}{\partial \boldsymbol{\beta_i}} V_i^{-1} (\mathbf{y_i} - g^{-1}(\mathbf{X}_i^T \boldsymbol{\beta})) = 0,$$

where $\mathbf{X}_i = (\mathbf{x}_{i,1}^T, \ldots, \mathbf{x}_{i,N}^T)^T \in \mathbb{R}^{N\times(m+1)}$.

There are several reasons that lead to us choosing a GEE model for this phylogenetic comparative study. Firstly, for each species in this family (each observation of one subject at different times), the response variable is taken as the proportion of the sample that possess a special characteristic over the total sample size and the sample sizes are varying and quite small. To account for this sampling error, we assume that the error is related to a binomial distribution dependent on its sample size. Therefore we need to adopt GLM since the data non-normal. In addition, our response variable is the estimated expectation of the outcome, so we choose a marginal regression model. Secondly, since there is time dependence among the covariates due to their phylogenetic relationship (longitudinal correlation), we adopt a model that incorporates this information. Since in this project we only focus on one family, we have only one subject. Our response variable is the presence of a syndrome which is of value in $[0, 1]$ and thus we choose logit link function to model this dichotomous outcome.

2 How to validate our approach?

To test the proposed models Monte Carlo simulation is often carried out, which could be carried out under the null hypothesis (or the alternative hypothesis). In each case, the proposed methods are credited only if most of the times they give a matching conclusion as the simulation assumption, that is the type-I error (or the Type-II error) is very small. For instance, in Example 1.1 we can simulate the evolution of the fruit size along the given tree irrespective of the seed mass; in this simulation we would like to see the testing approach gives us the conclusion that the two traits evolved independently. Alternatively, we could also generate the fruit size along the phylogeny dependent on the seed mass; then the method being tested is expected to show some correlation in the evolution of the two traits.

3 Model

The authors in [7] gave a general description of approaches modelling logistic regression models depending on whether the binary response is subject response or population average, whether the covariates are time-dependent. For illustrative purposes, we tweak the example in [7] a little. Suppose the data set stores the information on 1,000 patients and the response variable is an indicator of rehospitalization within 30 days after their discharge for the same condition for which they were first hospitalized, and the covariates are the repeated measurements of the blood pressure taken over different times. Then clearly the covariates are time dependent and the response variable is a subject response. In our data set the morning glories, there is only one subject – the morning glory family, and the traits for the species are measurements taken at different times. Note that we treat our response variable as *population average* as we take the proportion of sample having ergot alkaloid as an estimate for the real probability of having ergot alkaloid. In comparison, population in [7] means a collection of subjects whereas here means a collection of measurements at the same time.

We developed and implemented a statistical method for assessing the association between seed mass and the presence of ergot alkaloids in morning glory while accounting for the dependency in the data that occurs due to phylogeny. The method implements the technique described in [7] where there is one individual or replicate (Morning Glory) and a time dependency described by the phylogeny on 82 extant species rather than by autocorrelation among a sequence of timed events. We used simulations (parametric bootstrapping) to estimate the statistical p-value of association measured on the real data. For each species, some number of accessions, N_t, were assessed for the presence of ergot alkaloids, so the data contains for each species the number of accessions assessed and the number containing ergot alkaloids. This may be as few as one accession or as many as 101. Sometimes all or no accession contains ergot alkaloids, but, sometimes, only some of them contain ergot alkaloids. Thus, each species can be viewed as having some probability of containing alkaloids, rather than has having a strict presence/absence of ergot alkaloids in all individuals from the species.

The model for the association was a logistic regression model:

$$\text{logit}(P_t) = \beta_0 + \beta_1 x_t + \epsilon_t$$

where P_t is the probability that species t in the Morning Glory family has ergot alkaloids present, β_0 is the intercept interpreted as the average logit of the probability of the presence of ergot alkaloids over all the species, and β_1 is the slope where the difference from zero is interpreted as the strength of the association between seed mass and ergot alkaloid presence, x_t is the seed mass for species t, and ϵ_t is the usual normal residual error for the model.

We estimated β_0 and β_1 using the method of generalized estimating equations (GEE) described in section 3.1 in [7] adapting the method for one individual and a time dependency structure described by Brownian motion along the phylogeny rather than by an autocorrelation process for longitudinal data. The method provides estimate for the intercept and for the slope given in the usual statistical notation with hats: $\hat{\beta}_0 = -0.91$ and $\hat{\beta}_1 = 0.0015$.

To estimate the p-value for assessing the significance $\hat{\beta}_1$, we simulated values of $\hat{\beta}_1$ obtained under the null hypothesis that $\beta_1 = 0$ using the morning glory phylogeny and the estimate of $\hat{\beta}_0^{H_0} = -0.83$ obtained from the null hypothesis model:

$$\text{logit}(P_t) = \beta_0 + \epsilon_t$$

We will denote the simulated values for $\hat{\beta}_1$ under the null hypothesis as $\hat{\beta}_1^{sim:H_0}$.

This leads to a $\hat{\beta}_0^{H_0} = -0.83$, and thus constant probability of having ergot alkaloids $P_0 = 0.30$ that does not depend on the species seed mass. We also kept the number, N_t, of accessions assessed for ergot alkaloids in species t as fixed as in the original data.

The simulation was conducted by repeating these steps

1. Simulate a multivariate normal error vector, $\boldsymbol{\epsilon}$ with zero mean and a variance structure given by Brownian motion along the phylogeny standardized as a correlation matrix so that the variance along the diagonal is one.

2. Add this multivariate normal vector, $\boldsymbol{\epsilon}$ as a noise to the $\hat{\beta}_0^{H_0}$ obtained under the null hypothesis, and then we apply the inverse logit function on the sum to get the probability of having ergot alkaloids for each species $\mathbf{P}^{H_0} = \text{inv.logit}(\hat{\beta}_0^{H_0} + \boldsymbol{\epsilon})$.

3. Since each species has varying small sample sizes, we try to incorporate sampling bias by simulating $N_{alk,t}^{H_0}$ the number of sample with ergot alkaloids following a binomial distribution $Binomial(N_t, P_t^{H_0})$ for each species t, where $P_t^{H_0}$ is the t-th entry of $\mathbf{P}^{H_0}$.

4. For each species t, get the simulated proportion of ergot alkaloids through $P_{alk,t}^{H_0} = \frac{N_{alk,t}^{H_0}}{N_t}$ and we let $\mathbf{P}_{alk}^{H_0} = (P_{alk,1}^{H_0}, P_{alk,2}^{H_0}, \ldots, P_{alk,N}^{H_0})$.

5. Apply GEE to $\mathbf{P}_{alk}^{H_0}$ using the original seed mass to obtain $\hat{\beta}_1^{sim:H_0}$, a simulated value for $\hat{\beta}_1$ under the null hypothesis.

After repeating the simulation 1000 times we obtain a distribution of $\hat{\beta}_1^{sim:H_0}$, treated as the empirical null distribution of $\hat{\beta}_1$ and we determined that the two-sided p-value for our original estimate $\hat{\beta}_1 = 0.0015$ is approximated to be greater than 0.7. We conclude that there is no statistical association between the chance of having ergot alkaloids and the seed mass after accounting for phylogeny.

References

[1] Wesley T Beaulieu, Daniel G Panaccione, Corey S Hazekamp, Michelle C Mckee, Katy L Ryan, and Keith Clay. Differential allocation of seed-borne ergot alkaloids during early ontogeny of morning glories (convolvulaceae). *Journal of chemical ecology*, 39(7):919–930, 2013.

[2] Daniel Cook, Wesley T Beaulieu, Ivan W Mott, Franklin Riet-Correa, Dale R Gardner, Daniel Grum, James A Pfister, Keith Clay, and Clairton Marcolongo-Pereira. Production of the alkaloid swainsonine by a fungal endosymbiont of the ascomycete order chaetothyriales in the host ipomoea carnea. *Journal of agricultural and food chemistry*, 61(16):3797–3803, 2013.

[3] Joseph Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15, 1985.

[4] Theodore Garland, Jr and Anthony R Ives. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *The American Naturalist*, 155(3):346–364, 2000.

[5] Alan Grafen. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 326(1233):119–157, 1989.

[6] Pedro Jordano. Angiosperm fleshy fruits and seed dispersers: a comparative analysis of adaptation and constraints in plant-animal interactions. *The American Naturalist*, 145(2):163–191, 1995.

[7] Trent L Lalonde, Anh Q Nguyen, Jianqiong Yin, Kyle Irimata, and Jeffrey R Wilson. Modeling correlated binary outcomes with time-dependent covariates. *Journal of Data Science*, 11(4), 2013.

[8] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[9] Emilia P Martins and Thomas F Hansen. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149(4):646–667, 1997.

[10] Birgitta Sillén-Tullberg. Evolution of gregariousness in aposematic butterfly larvae: a phylogenetic analysis. *Evolution*, 42(2):293–305, 1988.

A appendix

A.1 Validation of Our Approach Through Simulated Correlated Data

The simulated data under the null hypothesis, that is $\beta_1 = 0$ while taking into consideration the phylogenetic relationship among the Morning Glory family species, is generated in the following manner.

1. Generate a sample of seed mass according to the normal distribution $N(\bar{\mathbf{x}}, V)$, where V is the response marginal variance, $\mathbf{x} = (x_1, \ldots, x_T)$ and $\bar{\mathbf{x}}$ is the average seed mass.
2. Consider the given phylogenetic tree with tip values of the probability of having Ergot. The tip values are rounded up to one if it's positive and zero otherwise.
3. We use the method of maximum parsimony to reconstruct the ancestral nodes. To decide on the ancestral nodes that could have value of either zero or one, we follow the assumption that if the children have Ergot then their parents are more likely to have Ergot than not to.
4. For each species, in the phylogenetic tree from the root to that species tip, whenever there is a change from 0 to 1, we increase its simulated seed mass by a constant (in our experiment it's chosen to be 60).

After the seed mass are simulated, we carry out the same hypothesis testing process on the simulated data. The p-values we get for the simulation is all zero, hence our approach for modeling the correlated binary output – the chance of having Ergot for the Morning Glory family – is justified.

A.2 R code for hypothesis tesing using GEE

```
# we add more specifications in gee(), such as link function = "logit" or family = "
    ↪ binomialbinomial"
library(MASS)

#library for estimating equations
library("gee")

#library for importing a newick tree
library("ape")
MyTree <- read.tree("~/Google␣Drive/correlated␣association/ITS_w_outgrps_cured.newick")

#library for comparative phylogeny
library("phytools")

#library for read_excel
library("readxl")
mydata <- read_excel("~/Google␣Drive/correlated␣association/Alkaloid␣Seed␣Mass␣Table␣for␣
    ↪ Doug␣Aug␣31␣2016_clean.xlsx",
                                                                    na = "null")
```

```
#lib for inv.logit
library("boot")
#################################################################
# clean data: since there more data in excel than
# in newick, we need to remove the extra data in excel
#################################################################

I = match(MyTree$tip.label,mydata$Species)
I_remain_data = I[which(!is.na(I))]
#remove extra data
mydata <-mydata[I_remain_data,]

#################################################################
# pruning MyTree
#################################################################
Index_extra_tree = which(is.na(I))

# Species_extra = MyTree$tip.label[Index_extra]
pruned.tree<-drop.tip(MyTree,MyTree$tip.label[Index_extra_tree])

# verify
match(pruned.tree$tip.label,mydata$Species)

colnames(mydata) = c("Species" , "Num_Acc","Num_Alk","SeedMass")

#add a new feature: the chance of having Ergot
mydata["Alk_P"] <- mydata["Num_Alk"]/mydata["Num_Acc"]

#sort mydata
mydata<-mydata[match(pruned.tree$tip.label,mydata$Species),]

#################################################################
# generalized linear models
#################################################################

library(nlme)

#phylogenetic correlation assuming the continuous phenotypes evolved according to
    ↪ Brownian motion
sigma<- vcv.phylo(pruned.tree, cor=TRUE)

gee_model1 = gee(Alk_P ~ SeedMass,id = c(rep(0,82)), data = mydata, corstr = "fixed",
    ↪ family = binomial, R = sigma )
beta0 = gee_model1$coefficients[1]
beta1 = gee_model1$coefficients[2]

# assume beta_1 is zero
gee_model1_null = gee(Alk_P ~ 1,id = c(rep(0,82)), data = mydata, corstr = "fixed",family
    ↪  = binomial, R = sigma )
# get the chance of having Ergot for morning glory family under the null hypothesis that
    ↪ beta_1 is zero
beta0_null <- gee_model1_null$coefficients[1]
p_null <- as.numeric(inv.logit(beta0_null))
```

```
############ Simulation to get distribution of beta under the null hypothesis
    ↪ ###############
data_sim <- mydata
mu <- c(rep(0,82))
# sigma<- vcv.phylo(pruned.tree, cor=TRUE)
Nsim = 2000

# the normalized numer of Ergot for each taxa follows normal distribution (0, sigma)
#DOUBLE CHECK, right now the error has unit variances for each species
noise <- mvrnorm(Nsim, mu = mu, Sigma = sigma )
# sim_Alk is the simulated number of Ergot each taxa should have under the null
    ↪ hypothesis
sim_Alk_P = matrix( c(rep(0,Nsim*82)),nrow=Nsim,ncol=length(mydata$Species))
sim_Alk_num = matrix( c(rep(0,Nsim*82)),nrow=Nsim,ncol=length(mydata$Species))
# data_sim$Alk = matrix( c(rep(0,Nsim*82)),nrow=Nsim,ncol=length(mydata$Species))
for(i in 1:82){
#### add noise to beta_null then use inv.logit link function to transform that into
    ↪ simulated presence of alkaloids
  #simulated chance of having alkaloids
  sim_Alk_P[,i] = inv.logit(noise[,i]+beta0_null)
  #simulate the number of sample that have alkaloids according to binomial distribution
  sim_Alk_num[,i] = rbinom(Nsim,mydata$Num_Acc[i],sim_Alk_P[,i])
}

# generate beta1 distribution under the null hypothesis
beta_sim<-c(rep(0,Nsim))
count_exception = 0
for(j in 1:Nsim){
  # When divergences happen in gee(), that loop is skipped
  error_beta_sim<-tryCatch({
  #get the simulated chance of having Ergot for each species
  data_sim$Alk_P = sim_Alk_num[j,]/data_sim$Num_Acc
  # get the simulated beta_1 from the simulated data
  beta_sim[j] = gee(Alk_P ~ SeedMass,id = c(rep(0,82)), data = data_sim, family =
      ↪ binomial, corstr = "fixed", R = sigma )$coefficient[2]},
  error=function(e){e}
)
  if(inherits(error_beta_sim,"error")){
    count_exception = count_exception+1
  }
}

#get rid of the values of beta_sim that diverged
# beta_sim = beta_sim[beta_sim!=0]
# get p-value
mean(abs(beta_sim) > abs(beta1))
```

A.3 R code for simulating seed mass under null hypothesis

```
#changed gee parameters: family = binomial and made sure sim_alk are nonnegative

#library for estimating equations
library("gee")
```

```
#for generating random multivariate normal r.v.'s
library(mvtnorm)

#library for importing a newick tree
library("ape")
MyTree <- read.tree("~/Google␣Drive/correlated␣association/ITS_w_outgrps_cured.newick")

#library for comparative phylogeny
library("phytools")

#library for read_excel
library("readxl")

#library for inv.logit
library("boot")

#library for nonlinear models
library(nlme)

#lib for excel
library(readxl)
mydata <- read_excel("~/Google␣Drive/correlated␣association/Alkaloid␣Seed␣Mass␣Table␣for␣
    ↪ Doug␣Aug␣31␣2016_clean.xlsx",
                     na = "null")
#################################################################
# clean data: more data in excel than in newick
#################################################################
I = match(MyTree$tip.label,mydata$Species)
I_remain_data = I[which(!is.na(I))]
mydata <-mydata[I_remain_data,]

# mydata <- mydata[match(MyTree$tip.label,mydata$Species),]
colnames(mydata) = c("Species" , "Num_Acc","Num_Alk","SeedMass")
mydata["Alk_P"] <- mydata["Num_Alk"]/mydata["Num_Acc"]

#################################################################
# pruning MyTree
#################################################################
Index_extra_tree = which(is.na(I))
pruned.tree<-drop.tip(MyTree,MyTree$tip.label[Index_extra_tree])

#sort mydata so the species match the MyTree tips' labels
mydata<-mydata[match(pruned.tree$tip.label,mydata$Species),]
# get maximum phylogeny for the given topology using Alk tips

MP.tree<-pruned.tree
MP.tree$Alk_P<- mydata$Alk_P[match(pruned.tree$tip.label,mydata$Species)]
MP.tree$Alk_P[which(MP.tree$Alk_P>0)] = 1

#approximate ancestral nodes with maximum parsimony
library("mvSLOUCH")
MP.ouchtree<- ape2ouch(MP.tree)
#correct the names of the internal nodes
MP.ouchtree@nodelabels[1:(MP.ouchtree@nnodes-MP.ouchtree@nterm)]<-as.character(
```

```
  1:(MP.ouchtree@nnodes-MP.ouchtree@nterm))

# root = 2 gives us root 1 ; 2nd state in level (0 1) is 1
# acctran accelerates transformation along evolution, thus realizing our assumption that
    ↪ if a child has Ergot, their parent are more likely to have Ergot as well
# The experiment is different from our assumption that most ancestral nodes will be
    ↪ 1.....
# The reality: most ancestral nodes are 0 and setting root = 1 does not affect the
    ↪ ancestral nodes close to root being 0
# NOTE: acctran or deltran does not affect how we calculate count_flip
# root = 2, set jump to a negative number
# root = 1, set jump to a positive number
MP.fitch<- fitch.mvsl(MP.ouchtree, MP.tree$Alk_P, deltran = FALSE, acctran = TRUE, root =
    ↪  1)
MP1.ouchtree<-MP.ouchtree
plot(MP.ouchtree)
for(i in 1:length(MP1.ouchtree@nodelabels)){
  MP1.ouchtree@nodelabels[i] = as.character(MP.fitch[i])
}
MP1.ouchtree@nodelabels[1] = "1"

#simulate the SeedMass
library(MASS)
mu_SM = c(rep(mean(mydata$SeedMass),82))
#should I add sig_SM in the simulation
sig_SM = var(mydata$SeedMass)
sigma<- vcv.phylo(pruned.tree, cor=TRUE) #phylogenetic correlation according to brownian
    ↪ motion
Nsim_SM = 100
#assumed that each species has the same variance
sim_SM <- mvrnorm(Nsim_SM, mu = mu_SM, Sigma = sigma*sig_SM )

#calculate the number of jumps for each species
count_flip = c(rep(0,163)) #only uses the latter 82 entries so the index is consistent
for(k in 82:163){ #1:81 are for internal nodes
  ancestor_curr = k
  while(ancestor_curr!=1){
    ancestor_prev = ancestor_curr
    ancestor_curr = as.numeric(MP.ouchtree@ancestors[ancestor_prev])
    fitch_curr = as.numeric(MP.fitch[ancestor_curr])
    fitch_prev = as.numeric(MP.fitch[ancestor_prev])
    # ergot is gained during evolution: 1 if value = 0; 2 if value = 1.
    if((fitch_curr==1) & (fitch_prev==2) ){
      count_flip[k] = count_flip[k]+1;
    }
    # ergot is lost during evolution
    if((fitch_curr==2) & (fitch_prev==1) ){
      count_flip[k] = count_flip[k]-1;
    }
    }}

# the constant we add to the seed mass if that species gain Ergot in their evolution
jump = 60
```

```
# if there are even number of jumps make it zero jump; odd then one jump
sim_SM[,which(count_flip[82:163]==1)] = sim_SM[,which(count_flip[82:163]==1)]+jump #
    ↪ adding jumps to the species
#Exception handling
if(length(which((sim_SM<0) == TRUE)) > 0 ){
  cat("WRONG simulation")
  Sys.sleep(5)
}
############################################################
# generalized linear models
############################################################
#initialization
pvalue = rep(0,Nsim_SM) #stores pvalues for each round of simulation
length_sim = rep(0,Nsim_SM) #stores the number of simulated beta1 under H0 in each
    ↪ simulation
count_exception_beta1 = NULL #stores the round index when divergences of gee() for beta1
    ↪ happens

#repeat simulations Nsim_SM times
for(round_SM in 1:Nsim_SM){
    #stores the loop index when divergences of gee() for beta_sim happens
    count_exception_beta_sim = NULL
    #ERROR HANDLING for getting beta1 from gee; if gee() diverges that loop is skipped
    Error_beta1 <- tryCatch(
      { mydata["sim_SM"]<-sim_SM[round_SM,]
      beta1 = gee(Alk_P ~ sim_SM,id = c(rep(0,82)), data = mydata, corstr = "fixed",family
          ↪ =binomial, R = sigma )$coefficients[2]
      beta_null = gee(Alk_P ~ 1,id = c(rep(0,82)), data = mydata, corstr = "fixed",family=
          ↪ binomial, R = sigma )$coefficients[1]
      p_null <- inv.logit(beta_null)
      p_null = as.numeric(p_null)
      },
      error=function(e) e)

  # stores the round index for beta1 when gee() diverges
  if(inherits(Error_beta1, "error")){
    count_exception_beta1 = c(count_exception_beta1,round_SM)
    next}

  #continue simulation for beta1 under null hypothesis
  ############ Simulation to get distribution of beta ##############
  data_sim <- mydata
  mu <- c(rep(0,82))
  Nsim = 500
  noise <- mvrnorm(Nsim, mu = mu, Sigma = sigma )
  sim_Alk_num = matrix( c(rep(0,Nsim*82)),nrow=Nsim,ncol=length(mydata$Species))
  sim_Alk_p = matrix( c(rep(0,Nsim*82)),nrow=Nsim,ncol=length(mydata$Species))

  #simulate number of sample with ergot alkaloids for each species
  for(i in 1:82){
    sim_Alk_p[,i] = inv.logit(noise[,i]+beta_null)
    sim_Alk_num[,i] = rbinom(Nsim,data_sim$Num_Acc[i],sim_Alk_p[,i])
  }
```

```
  # generate beta distribution under the null hypothesis
  beta_sim<-c(rep(0,Nsim))
  for(j in 1:Nsim){
    # if gee() diverges that loop is skipped
    Error_beta_sim <- tryCatch(
    {data_sim$Alk_P = sim_Alk_num[j,]/data_sim$Num_Acc
    beta_sim[j] = gee(Alk_P ~ sim_SM,id = c(rep(0,82)), data = data_sim, family=binomial,
        ↪ corstr = "fixed", R = sigma )$coefficient[2]},
    error=function(e) e
    )

    #stores the loop index for beta_sim when gee() diverges
    if (inherits(Error_beta_sim, "error")){
      count_exception_beta_sim = c(count_exception_beta_sim,j)
      next}
  }#end for loop for j

  #clean beta_sim to get rid of ones diverged
  if (!is.null(count_exception_beta_sim)){
     beta_sim = beta_sim[-count_exception_beta_sim]
  }

  # get p-value
  pvalue[round_SM] = mean(abs(beta_sim) > abs(beta1))
  length_sim[round_SM] = length(beta_sim)
}#end round_SM

 pvalue_effective = pvalue[-count_exception_beta1]
 save(pvalue_effective,count_exception_beta1,pvalue,file = "pvalue.RData")
```