Midterm Exam

Haoyu Li

11/2/2020

Instruction

This is your midterm exam that you are expected to work on it alone. You may NOT discuss any of the content of your exam with anyone except your instructor. This includes text, chat, email and other online forums. We expect you to respect and follow the GRS Academic and Professional Conduct Code.

Although you may NOT ask anyone directly, you are allowed to use external resources such as R codes on the Internet. If you do use someone's code, please make sure you clearly cite the origin of the code.

When you finish, please compile and submit the PDF file and the link to the GitHub repository that contains the entire analysis.

Introduction

In this exam, you will act as both the client and the consultant for the data that you collected in the data collection exercise (20pts). Please note that you are not allowed to change the data. The goal of this exam is to demonstrate your ability to perform the statistical analysis that you learned in this class so far. It is important to note that significance of the analysis is not the main goal of this exam but the focus is on the appropriateness of your approaches.

Data Description (10pts)

Please explain what your data is about and what the comparison of interest is. In the process, please make sure to demonstrate that you can load your data properly into R.

```
#load data from my github
leaf_size<-read.csv("https://raw.githubusercontent.com/jasonhaoyuli/MA-678-Midterm-exam/main/Leaf%20siz
print(leaf_size)
```

```
##    leaf_id tree_id length width number_of_veins number_of_corners
## 1        1       1    7.9   7.0              5                 1
## 2        2       1    8.2   7.5              7                 1
## 3        3       1    9.5   7.7              8                 1
## 4        4       1    8.5   6.9              6                 1
## 5        5       1    6.8   6.7              5                 1
## 6        6       2   11.5  12.0              6                 5
## 7        7       2   12.5  14.5              6                 5
## 8        8       2   13.5  15.5              9                 5
## 9        9       2   13.0  15.0              8                 5
## 10      10       2   13.2  15.9              4                 5
```

```
## 11      11       3   12.2  13.4              8                5
## 12      12       3   11.2  13.0              7                5
## 13      13       3   13.4  13.8              6                5
## 14      14       3   14.0  16.1              6                5
## 15      15       3   14.2  15.2              5                5
## 16      16       4    7.5   7.0              6                1
## 17      17       4    6.3   7.1              7                1
## 18      18       4    7.2   8.2              5                1
## 19      19       4    6.1   7.5              6                1
## 20      20       4    7.1   5.2              8                1
## 21      21       5    6.2   5.8              3                5
## 22      22       5    4.8   5.3              3                5
## 23      23       5    5.1   4.3              4                5
## 24      24       5    5.4   4.6              5                5
## 25      25       5    6.0   5.2              3                5
```

Brief description of the leaf size dataset: The dataset that I collected is about the size of the leaves by measuring leaves from five different trees, and I also collect couples factors that may affect the size of leaves. The comparison of interest is that to whether different trees have different effects on the size of leaves, and also how individual factors affect the size of leaves.

EDA (10pts)

Please create one (maybe two) figure(s) that highlights the contrast of interest. Make sure you think ahead and match your figure with the analysis. For example, if your model requires you to take a log, make sure you take log in the figure as well.

```
#create a new column that calculate the size of the leaves by multiple the length and width of leaves
leaf_size$size<-leaf_size$length*leaf_size$width
#visualize size of leaves grouped by different tree
library(tidyverse)
```

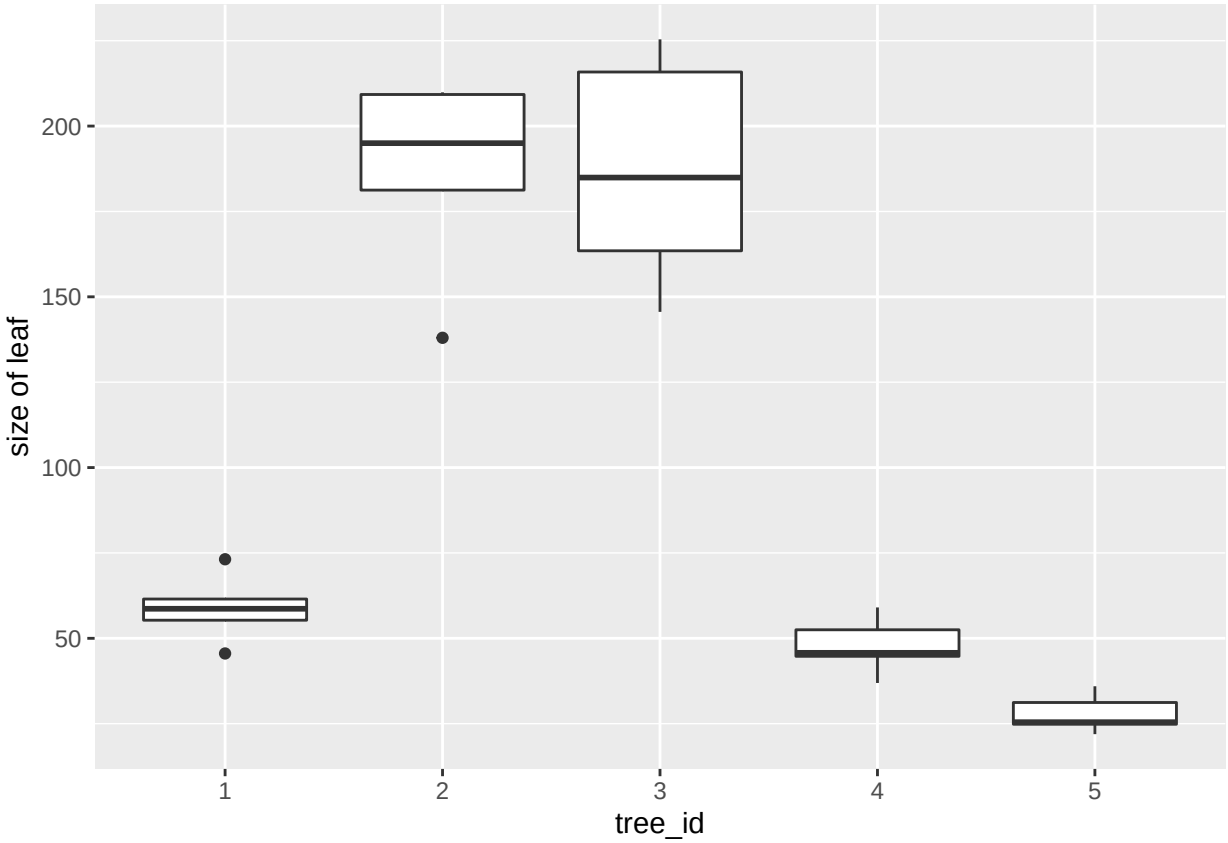
```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --
```

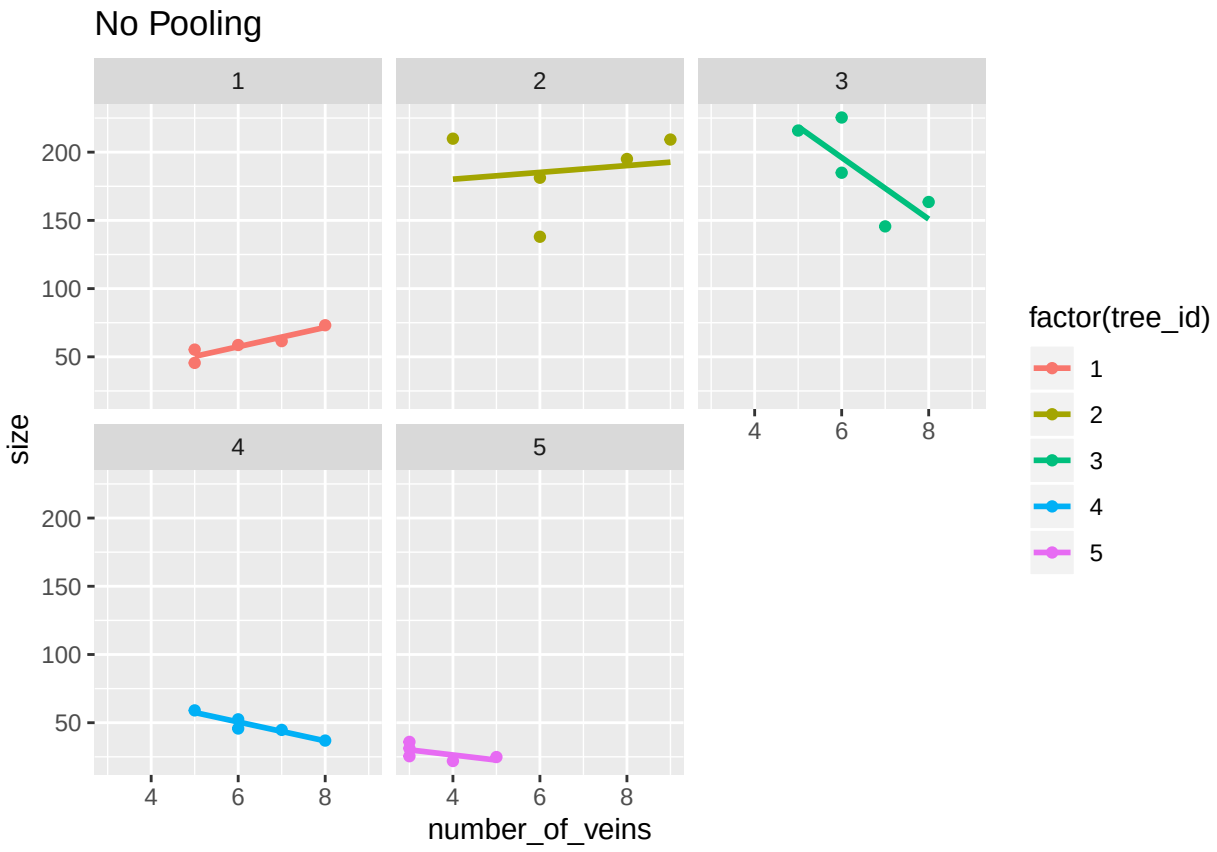
```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
ggplot(data=leaf_size)+
  geom_boxplot(mapping=aes(x=factor(tree_id),y=size))+
  labs(x="tree_id",y="size of leaf")
```




```
#plot fitted model for each group(tree),which is no pooling model
ggplot(data=leaf_size)+
  geom_point(aes(x=number_of_veins, y=size, color = factor(tree_id)))+
  geom_smooth(aes(x=number_of_veins, y=size, color = factor(tree_id)), formula = "y~x", method="lm", se=
  facet_wrap(~factor(tree_id))+
  labs(title = "No Pooling", xlab="number_of_veins")
```



Power Analysis (10pts)

Please perform power analysis on the project. Use 80% power, the sample size you used and infer the level of effect size you will be able to detect. Discuss whether your sample size was enough for the problem at hand. Please note that method of power analysis should match the analysis. Also, please clearly state why you should NOT use the effect size from the fitted model.

```
#install.packages("pwr")
library(pwr)
```

```
## Warning: package 'pwr' was built under R version 3.6.3
```

```
#doing a power analysis
#n=25 since we do a two sample,two side power analysis
#calculate the standard deviation of size of leaves
sd<-sd(leaf_size$size)
sd
```

```
## [1] 74.27431
```

```
pwr.t.test(n=25,sig.level=0.05,power=0.8,type = "two.sample",alternative="two.sided")
```

```
##
```

```
##      Two-sample t test power calculation
##
##               n = 25
##               d = 0.8087121
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The result shows the effect size we will detect is 1.17, under the hypothesis that we assume the mean population difference between different trees/group is 0, we have 80% chance to detect a effect size of 0.8 given the sample size is 50(25 for no difference group,25 for there is difference group).

Reason why we should not use this effect size: from the power analysis, we know we have 80% chance to detect that variances between groups, which the difference between populations u1-u2 =dsd=*0.8*74.27=59.416. That difference is much large than we expected since we know size of leaves between similar kind of tree can be close, so we want to detect that small difference.

Modeling (10pts)

Please pick a regression model that best fits your data and fit your model. Please make sure you describe why you decide to choose the model. Also, if you are using GLM, make sure you explain your choice of link function as well.

```
library(rstanarm)
```

```
## Warning: package 'rstanarm' was built under R version 3.6.3
```

```
## Loading required package: Rcpp
```

```
## This is rstanarm version 2.21.1
```

```
## - See https://mc-stan.org/rstanarm/articles/priors for changes to default priors!
```

```
## - Default priors may change, so it's safest to specify priors, even if equivalent to the defaults.
```

```
## - For execution on a local, multicore CPU with excess RAM we recommend calling
```

```
##   options(mc.cores = parallel::detectCores())
```

```
library(arm)
```

```
## Warning: package 'arm' was built under R version 3.6.3
```

```
## Loading required package: MASS
```

```
## Warning: package 'MASS' was built under R version 3.6.3
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loading required package: lme4

## Warning: package 'lme4' was built under R version 3.6.3

##
## arm (Version 1.11-2, built: 2020-7-27)

## Working directory is C:/Users/Jasonli/Desktop

##
## Attaching package: 'arm'

## The following objects are masked from 'package:rstanarm':
##
##     invlogit, logit
```

```
leaf_size$treeid<-factor(leaf_size$tree_id)
fit<-stan_lmer(size~number_of_veins+number_of_corners+(1+number_of_veins+number_of_corners|treeid),data=
```

```
## Warning: There were 15 divergent transitions after warmup. See
## http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
## to find out why this is a problem and how to eliminate them.

## Warning: Examine the pairs() plot to diagnose sampling problems
```

```
coef(fit)
```

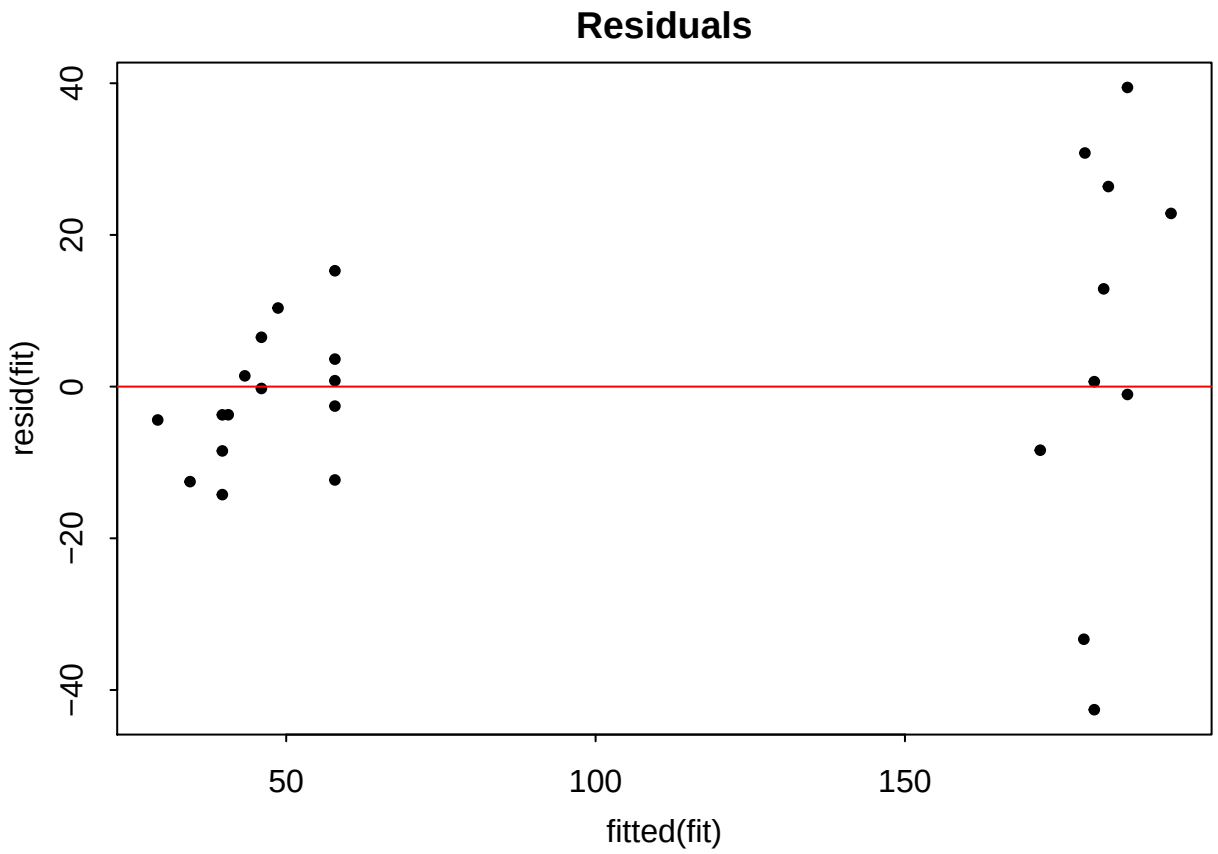
```
## $treeid
##   (Intercept) number_of_veins number_of_corners
## 1    36.25827     0.001679035         21.602506
## 2    37.61753     0.758412700         27.685802
## 3    39.78683    -7.047009648         37.690519
## 4    37.92263    -2.682583863         24.167688
## 5    33.70012    -5.221912160          4.329205
##
## attr(,"class")
## [1] "coef.mer"
```

Explaination of the chosen model: By doing EDA, we see leaf sizes are very different across different groups. Thus, we want to fit a multilevel linear model to account these differences. The varying intercept reflect the difference between groups without considering the variance within groups. The varying slopes reflects the variances within the groups.

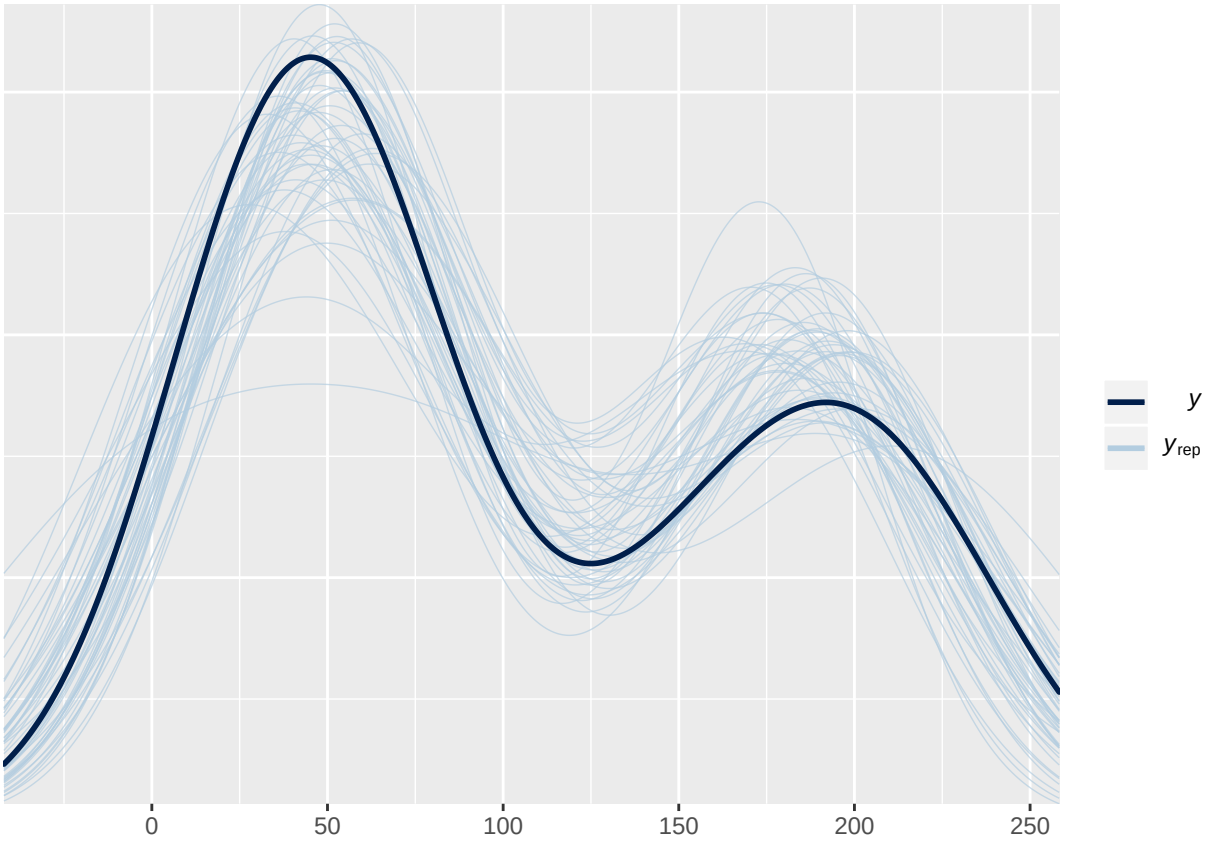
Validation (10pts)

Please perform a necessary validation and argue why your choice of the model is appropriate.

```
#redisual plot
par(mar=c(3,3,2,1), mgp=c(2,.7,0), tck=-.01)
plot(fitted(fit),resid(fit),pch=20,main="Residuals")
abline(0,0,col="red")
```




```
#posterior predictive check
pp_check(fit)
```



The residual shows the unequal spread, so the errors do not have equal variance. This might not be a good model, but it might be due to small sample size as we can see for lower and higer fitted value still have equal variance. But when we look at the posterior predictive check, the observed value fall within predictive value, so I still stick to this model since it describes multilevel data at its best.

Inference (10pts)

Based on the result so far please perform statistical inference to compare the comparison of interest.

```
#Bayesian posterior uncertainty intervals
posterior_interval(fit)
```

```
##                                                 5%         95%
## (Intercept)                            -39.625499  112.085675
## number_of_veins                        -13.759069    7.055069
## number_of_corners                       -1.428328   50.168285
## b[(Intercept) treeid:1]                -33.884965   25.529642
## b[number_of_veins treeid:1]             -5.502628   16.304562
## b[number_of_corners treeid:1]          -31.763740   22.769834
## b[(Intercept) treeid:2]                -29.640383   30.484289
## b[number_of_veins treeid:2]             -5.529212   16.309881
## b[number_of_corners treeid:2]          -14.125733   23.141953
## b[(Intercept) treeid:3]                -23.991878   44.034041
## b[number_of_veins treeid:3]            -20.631039    6.367116
## b[number_of_corners treeid:3]           -2.075266   37.961860
## b[(Intercept) treeid:4]                -24.120574   31.704229
```

```
## b[number_of_veins treeid:4]                          -9.567499   11.802703
## b[number_of_corners treeid:4]                       -23.658680   29.786234
## b[(Intercept) treeid:5]                             -44.722548   25.838194
## b[number_of_veins treeid:5]                         -16.983061   10.752493
## b[number_of_corners treeid:5]                       -39.214353   -1.692541
## sigma                                                15.886023   28.416722
## Sigma[treeid:(Intercept),(Intercept)]                 9.408166 1879.069598
## Sigma[treeid:number_of_veins,(Intercept)]          -253.557677  176.025422
## Sigma[treeid:number_of_corners,(Intercept)]        -181.789658  223.418919
## Sigma[treeid:number_of_veins,number_of_veins]         5.695935  480.868447
## Sigma[treeid:number_of_corners,number_of_veins]   -175.990445   96.348898
## Sigma[treeid:number_of_corners,number_of_corners]   58.316557 1045.319321
```

```
#visualize
library(bayesplot)
```

```
## Warning: package 'bayesplot' was built under R version 3.6.3
```

```
## This is bayesplot version 1.7.2
```

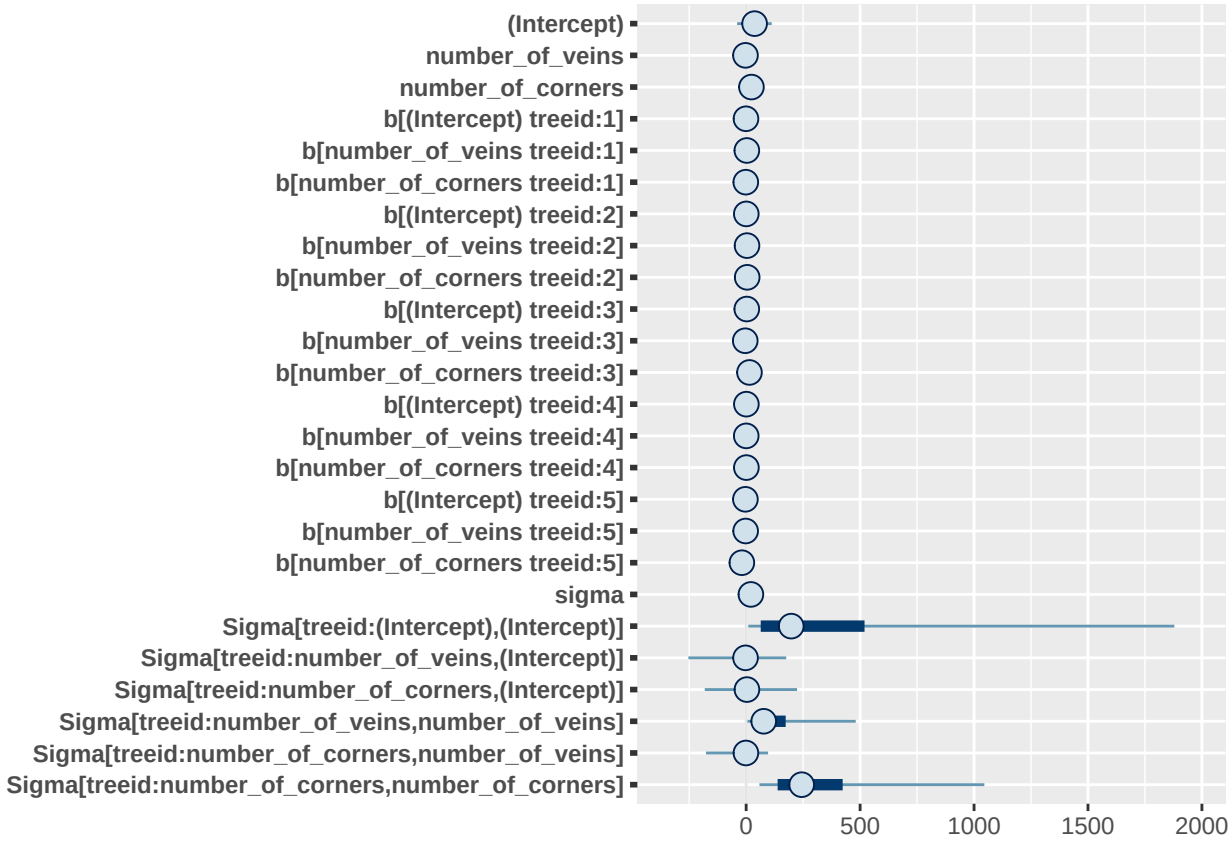
```
## - Online documentation and vignettes at mc-stan.org/bayesplot
```

```
## - bayesplot theme set to bayesplot::theme_default()
```

```
##    * Does _not_ affect other ggplot2 plots
```

```
##    * See ?bayesplot_theme_set for details on theme setting
```

```
sim<-as.matrix(fit)
mcmc_intervals(sim)
```



Discussion (10pts)

Please clearly state your conclusion and the implication of the result.

conclusion: From the results of our model, we can conclude that there are factors such as number of veins indeed have effect on the size of the leaves, and size of leaves have large variances across different trees.

Limitations and future opportunity. (10pts)

Please list concerns about your analysis. Also, please state how you might go about fixing the problem in your future study.

1.Sample size: since the sample sizes from the data collection exercise are very limited, we can see large variances across different tree, thus the statistical error from this dataset is very unequal as we'll expected.

2.How to improve model: If we have more time, we can collect more data from different trees, and the species of tree is also important, if we collect the data from the same species but just different trees, there's also a lack of independence problem. ### Comments or questions If you have any comments or questions, please write them here.