# Regression Analysis on Timber Data Using R

ARIJIT NASKAR

August 12, 2021

## Introduction

Gmelina arborea is an important timber yielding tree species growing in tropical and subtropical regions of South and South East Asia. It is a popular species because of its rapid growth and promising wood characteristics. The species shows large variability in tree characteristics related to adaptation, productivity and quality. An experiment was conducted to provide genetic basis to the observed variation across different provenances in India, Myanmar and Thailand. It is of interest to know what the genetic component of this variability is and how much can be attributed to environmental factors. Seedlings were collected from 8 localities. These were grown in vitro. Seedling nodal segments were used to initiate in vitro cultures. The axillary shoots were excised at the base and nodal segments from these shoots were sub cultured. The process was repeated for 6 subculture generations. It is expected that for subsequent subcultures, variability in the response should decrease. Three characteristics were recorded as 'response': Elongation, induction of multiple shoots and Rooting. In addition to this, data on seed length, seed width and germination was also recorded for seeds from each locality.

### Objective:

Analysis of genetic and environmental components of tree characteristics.

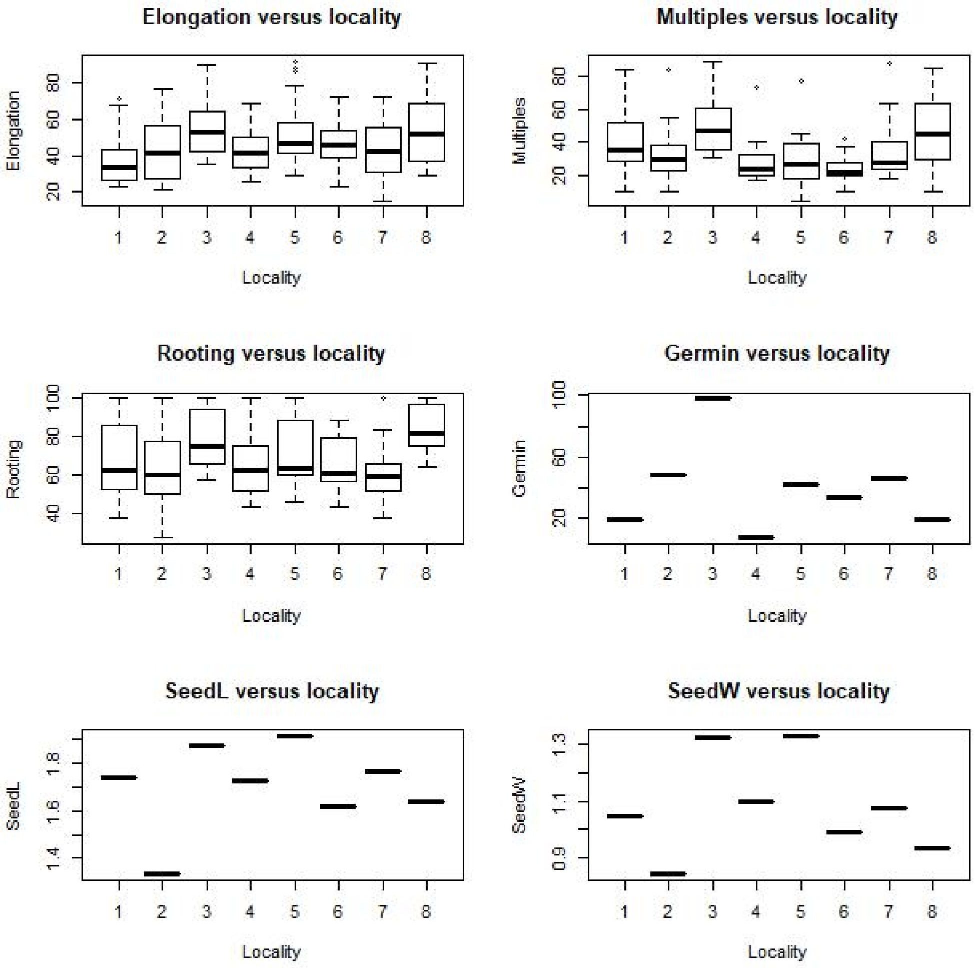
## Exploratory Data Analysis

The response variables are integer type data. "Locality", "Subculture" and "year" are categorical data or factors.

Now we see the nature of changes of different response variables over locality,subcultures and years,

## (a).(over Locality)

From this boxplots we can conclude the followings



**(1).**

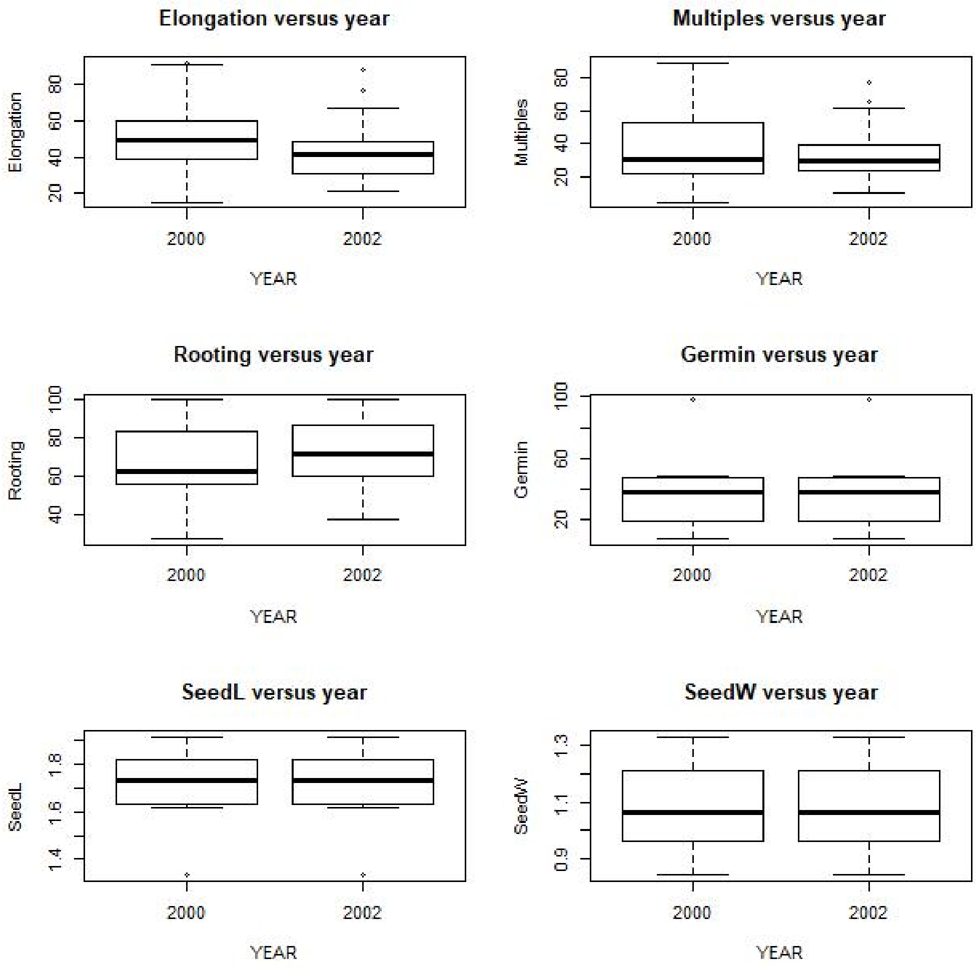
The length and width of seeds,Germination is fixed in every localities.

**(2).**

All the response variables fluctuates over different localities.

## (b).(over year)

From this boxplots we can conclude the followings,



**(1).**

The rootings increase over years.

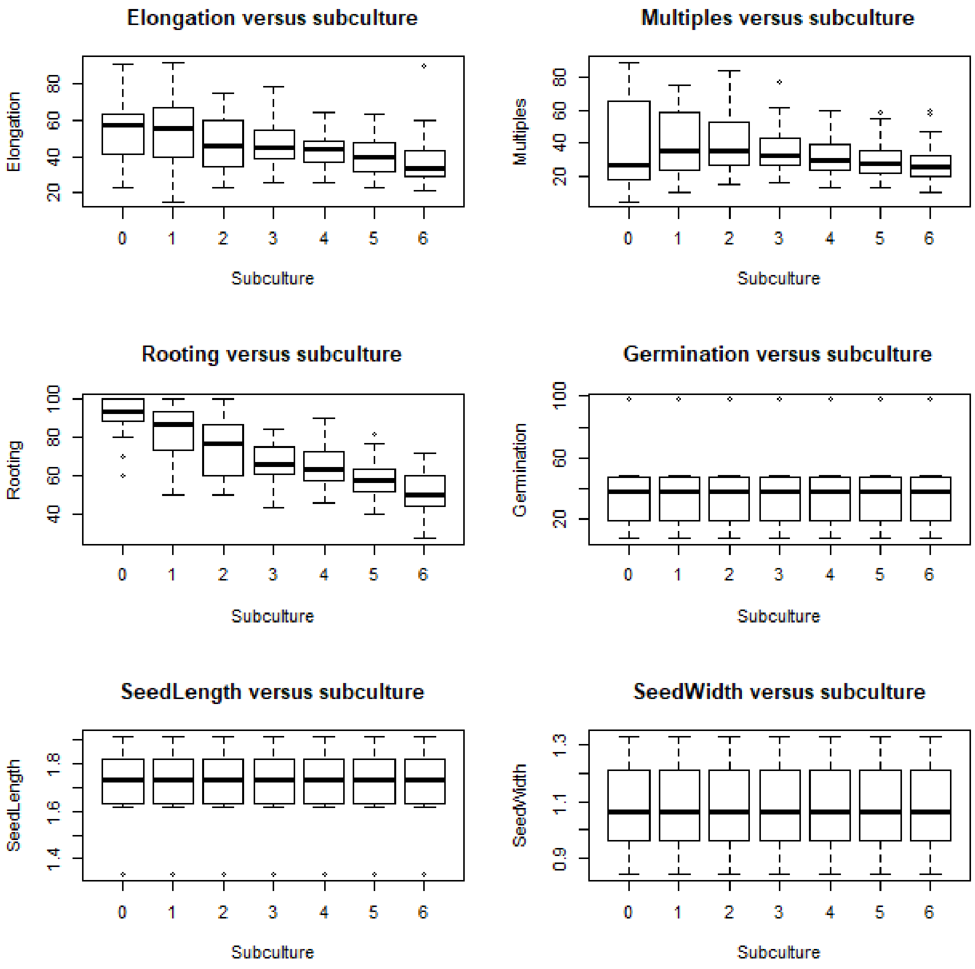
**(2).**

The elongation of the trees decreases over the years.

**(3).**

The other response variables(like induction of multiple shoots,length and width of seeds,Germination) remain same over years.

## (c).(over subculture)



From these boxplots we can conclude the followings,

**(1).**

The rootings decreases over subcultures.

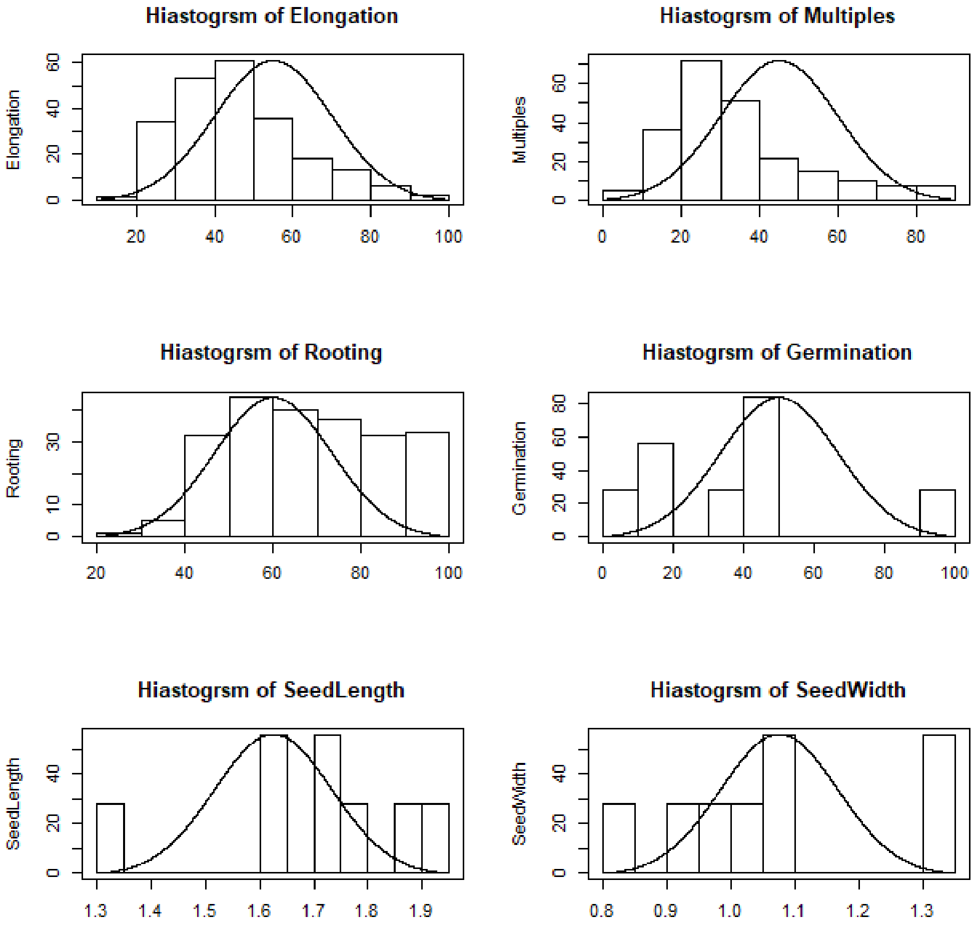
**(2).**

The elongation of the trees decreases over the subcultures.

**(3).**

The other response variables(like induction of multiple shoots,length and width of seeds,Germination) approximately remain same over different subcultures.

From the correlation matrix,the correlation of seed width and seed width is 0.9081 which is the highest correlation.


Hiastogrsm of Elongation

Hiastogrsm of Multiples

Hiastogrsm of Rooting

Hiastogrsm of Germination

Hiastogrsm of SeedLength

Hiastogrsm of SeedWidth

After fitting normal probability density curve with histograms of the response variables we can say the distribution of Rooting is approximately normal distribution.The distribution,seed width of Elongation and Multiples are positively skewed.The distribution of seed length is negatively skewed.

# Inferential Data Analysis

lets check the hypothesis t.hat correlation coefficient between Seed width and Seed length are statistically significant or not which is,

$H_0 : \rho = 0$ vs $H_1 : \rho > 0$ with level of significance 5%.

The value of the statistic is 32.277 and the p-value is $2\times10^{-16}$.As the p-value is less than .05 then in the light of given data we can reject the null hypothesis.

So the correlation co-efficient of Seed width and Seed length is statistically significant and positive.

We can check the difference between mean elongation and rooting of 2000 and 2002 respectively, is statistically significant or not.

For elongation,first we have to check the hypothesis of homoscedasticity elongation of 2000 and 2002 respectively, i.e.

$H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$ with level of significance 5%.

The test statistic or the F value is 1.5207 and the p-value is .0281 which is greater than .025.So in the light of given data we accept the null hypothesis i.e. the elongation 2000 and 2002 respectively is homoscedastic.

the hypothesis be,(under the assumption of homoscedasticity)

$H_0 : \mu_1 - \mu_2 = 0$ vs $H_1 : \mu_1 - \mu_2 \neq 0$ with level of significance 5%.

The value of the test statistic or the t value is 4.569 and the corresponding p-value is $8.3\times10^{-6}$which is less than .025.So in the light of given data we reject the null hypothesis i.e the the difference between mean elongation of 2000 and 2002 respectively, is statistically significant.

Similarly,for Rooting,

The F-value is 1.0596 and the corresponding p-value is 0.7609($>$0.025).So in the light of given data we accept the null hypothesis i.e. the rooting of 2000 and 2002 respectively is homoscedastic.

The value of the test statistic or the t value is -1.7471 and the corresponding p-value is 0.082 which is greater than 0.025.So in the light of given data we accept the null hypothesis i.e the the difference between mean elongation of 2000 and 2002 respectively, is not statistically significant.

Now,we should find the impact of the qualitative variables on the quantitative variables of the trees.

We obtain the effect of number of bedrooms on the price of the house.

The **ANOVA** model is,

$y_{ij} = \alpha + \mu_i + e_{ij}$ i=1,2,..,$n_i$,j=1,2,.12

Here,$y_{ij}$ = observation on the qualitative variable of $j$th tree at i th level of qualitative variable

$\mu_i$ =additional effect of i th qualitative variable on the quantitative varaible.

$n_i$ =number of observations at i level of qualitative varaible.

We set the null hypothesis,

$H_0 : \mu_1 = \mu_1 = \mu_2 = .. = \mu_{12} = 0$ vs $H_1 : H_0$is not true

Let we To determine the effect of locality on the response variables we use the **ANOVA method,**

All the p values are less than 0.05.So all the response variables differs significantly over Locality.

To determine the effect of replicate on the response variables we use the same procedure.

All the p values are less than 0.05.So all the response variables differs significantly over replicate.

To determine the effect of replicate on the response variables we use the same procedure.

All the p values are less than 0.05.So all the response variables differs significantly over replicate.

To determine the effect of subculture on the response variables we use the same procedure.

All the p values are less than 0.05.So all the response variables ef differs significantly over subculture.

# Conclusion and Discussion

The seed width and seed length is fixed in each locality ,subculture and year and also it is remain same over years and subcultures but fluctuates over different locality.The seed width increase if the seed length increases.The elongation of trees changes over years.Germination doesn't change over time.Rooting decreases over subcultures.Germination fluctuates over different localities but remain same over different subcultures.The distribution of Rooting is approximately normal.The distribution of Elongation,Seed width and Multiples are positively skewed.The distribution of seed length is negatively skewed.The response variables such as Elongation,Seed length,Seed width and Seed length differs significantly over locality,subculture and Replication.

# Appendix:

## Theoretical Formula

$\alpha$=level of significance,$t_{\alpha;n}$=Upper $\alpha$ point of t distribution with degrees of freedom n.

$F_{\alpha;n,m}$=Upper $\alpha$ point of F distribution with degrees of freedom n and m.

$\tau_\alpha$=Upper $\alpha$ point of standard normal distribution.

For $H_0 : \rho = 0$ vs $H_1 : \rho > 0$ ,

Test statistic is $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ and critical region is $\{\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} > t_{\alpha;n-2}\}$ where r=correlation co-efficient ,n is no of observation.

For $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$, Test statistic is $\frac{S_1^2}{S_2^2}$and critical region is $\{|\ \frac{S_1^2}{S_2^2}\ |< F_{\alpha/2;n-1,m-1}\}$ where $S_1$=sample sd of X ,$S_2$=sample sd of Y , n=no. of observation of X and m=no. of observation of Y.

For $H_0 : \mu_1 - \mu_2 = 0$ vs $H_1 : \mu_1 - \mu_2 \neq 0$,

Test statistic is $\frac{\bar{y}-\bar{x}}{S\sqrt{\frac{1}{n}+\frac{1}{m}}}$and critical region is $\{|\ \frac{\bar{y}-\bar{x}}{S\sqrt{\frac{1}{n}+\frac{1}{m}}}\ |< t_{\alpha/2;n+m-1}\}$ where $\bar{x}$= sample mean of X and $\bar{y}$=Sample mean of Y.

The **ANOVA** model is,

$y_{ij} = \alpha + \mu_i + e_{ij}$ i=1,2,..,$n_i$,j=1,2,.12

We set the null hypothesis,

$H_0 : \mu_1 = \mu_1 = \mu_2 = .. = \mu_{12} = 0$ vs $H_1 : H_0$is not true

The test statistics is $F = \frac{\sum n_i(\bar{y}_i-\bar{y})}{\sum\sum(\bar{y}_{ij}-\bar{y}_i)}$ and the critical region is $\{F > F_{p-1,n-p}\}$,p is the number of levels of the qualitative variables and $n = \sum n_i$.

## R codes

```
X=read.table("C:\\Users\\User\\Desktop\\Timber.csv",header=T,sep=",")
attach(X)
par(mfrow=c(3,2))
for(j in 5:10)
{
boxplot(X[,j]~Locality,ylab=colnames(X)[j], main=paste(colnames(X)[j],
"versus locality"))
}
par(mfrow=c(3,2))
for(j in 5:10)
{
boxplot(X[,j]~YEAR,ylab=colnames(X)[j], main=paste(colnames(X)[j],"versus year"))
}
par(mfrow=c(3,2))
for(j in 5:10)
{
boxplot(X[,j]~Subculture,ylab=colnames(X)[j], main=paste(colnames(X)[j],
"versus subculture"))
}
par(mfrow=c(3,2))
for(j in 5:10)
{
hist(X[,j],ylab=colnames(X)[j],xlab="", main=paste("Hiastogrsm
of",colnames(X)[j]))
x=seq(-3,3,by=.001)
par(new=T)
plot(dnorm(x),type="l",xaxt="n",yaxt="n",xlab="",ylab="")
}
```

```
cor(X)
x=cor(SeedWidth,SeedLength)
y=(x/sqrt(1-x^2))*sqrt(length(SeedWidth)-1)
p=1-pt(y,df=length(SeedWidth)-1))
a=Germination[YEAR==2000]
b=Germination[YEAR==2002]
c=Rooting[YEAR==2000]
d=Rooting[YEAR==2002]
var.test(a,b)
var.test(c,d)
t.test(a,b)
t.test(c,d)
summary(aov(Elongation~as.factor(Locality)))
summary(aov(Multiples~as.factor(Locality)))
summary(aov(Rooting~as.factor(Locality)))
summary(aov(Germin~as.factor(Locality)))
summary(aov(SeedL~as.factor(Locality)))
summary(aov(Multiples~as.factor(REPLICATE)))
summary(aov(Rooting~as.factor(REPLICATE)))
summary(aov(Germin~as.factor(REPLICATE)))
summary(aov(SeedL~as.factor(REPLICATE)))
summary(aov(Multiples~as.factor(Subculture)))

summary(aov(Rooting~as.factor(Subculture)))
summary(aov(Germin~as.factorSubculture)))
summary(aov(SeedL~as.factor(Subculture)))
```

## Bibliography

I have read some reference like

1.Fundamentals of Statistics,Goon,Gupta,Dasgupta.

2.Outline of statistics,Goon,Gupta,Dasgupta.

I also take some help from some articles related to statistical analysis.