INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tento projekt je spolufinancován Evropským sociálním fondem a Státním rozpočtem ČR
InoBio – CZ.1.07/2.2.00/28.0018

Statistical Analysis in Ecology using R

ANOVA, Linear Regression, ANCOVA

Juliette Chamagne
Institute of Evolutionary Biology and Environmental Sciences
University of Zurich

November 12th, 2013

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
INVESTMENTS IN EDUCATION DEVELOPMENT

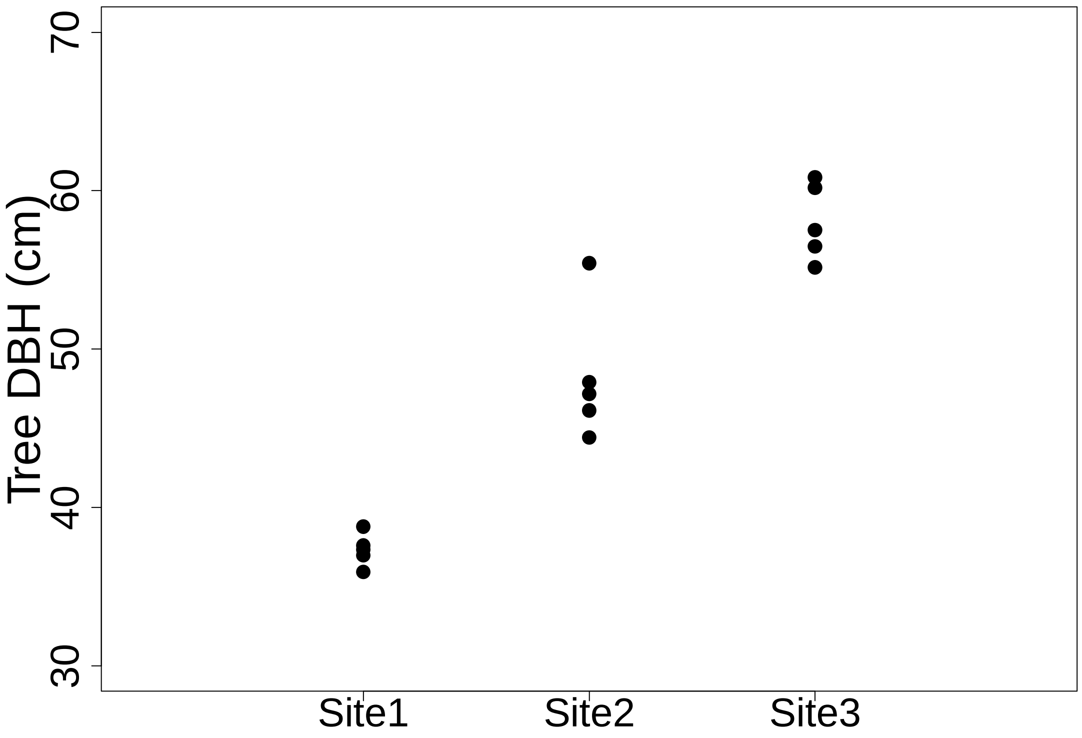
Recap

- Type of data
- We want to **explain Y**
  - Central tendency
  - Dispersion around the central tendency
- Correlations
- Visuals (graphs)

ANalysis Of VAriance (ANOVA)

- Used to analyze how **mean values** of continuous variable (Y) **differ among groups** of elements.
- Y (size)
- Groups (sites)
- Elements (trees)

- k = 3 sites
- n = 15 trees



ANOVA: Assumptions

- **Independence:** The elements have been sampled randomly and independently from each group.
- **Normality:** For each group the values of Y follow a normal distribution.
- **Homoscedasticity:** the variance should be the same for each group.

ANOVA: Hypotheses

- **H_0:** The **null hypothesis**. The mean value of Y is the same for all groups. The observed variability is only within groups and is due to random sampling. **Nothing interesting is happening.**

- **H_A:** The **alternative hypothesis**. The mean value of Y differs among populations. There is variability within groups AND among groups. **Something might be of interest.**

- **We test H_0**

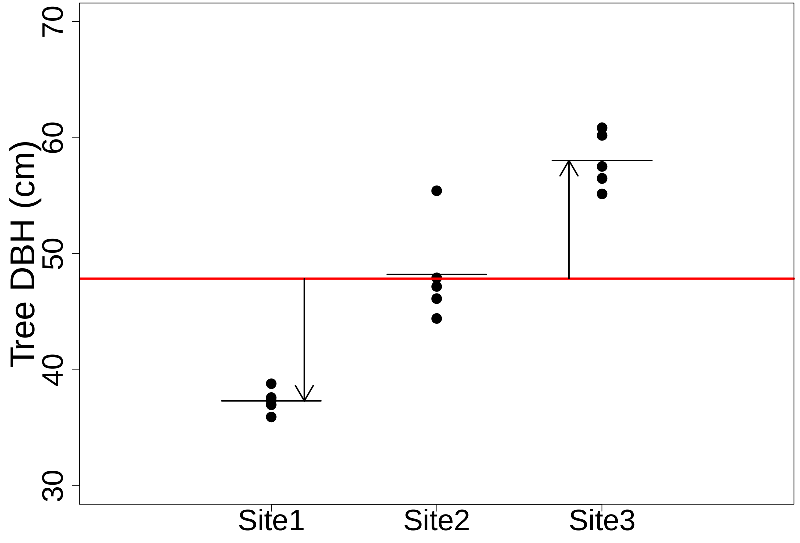
ANOVA: F ratio:

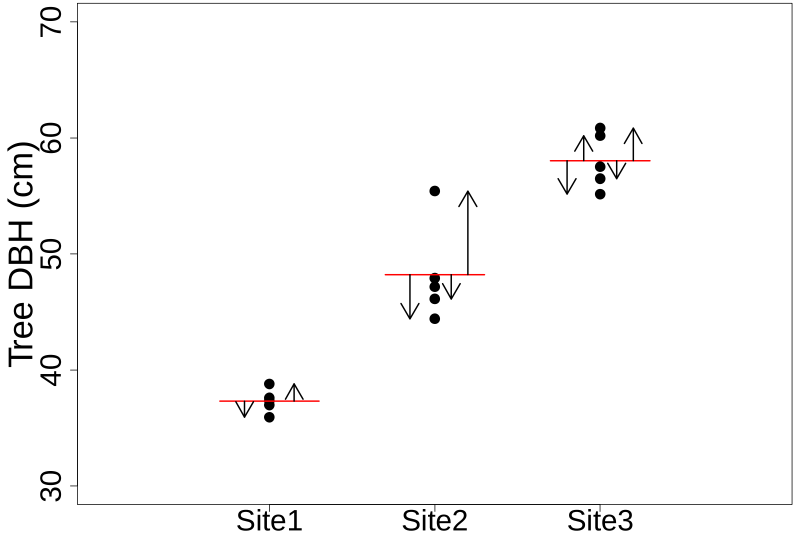
Variation among groups/Variation within groups

$$SS_{among} = \square_{ij} (\bar{y}_j - \bar{y})^2 = \square_{j} n_j (\bar{y}_j - \bar{y})^2$$

$$SS_{within} = \square_{ij} (y_{ij} - \bar{y}_j)^2$$

- **“Group effects”**
- **“Residuals”**





SS = Sum of Squares

n_j = number of elements in group j

y_{ij} = value of element i in group j

$\bar{y}_j$ = mean of yij for group y

$\bar{y}$ = overall mean

ANOVA: F ratio:

Variation among groups/Variation within groups

- **F-ratio = MS among groups/MS within groups**
- MS = SS/df
- MS = Mean Square

- df = degree of freedom
- df total = n-1
- df among = k-1
- df within = n-k
- **Total = residual + group**

k = number of groups (j = 1 to k)
n = total sample size (sum of n_j)

$MS_{total} = MS_{residual} + MS_{group}$
$SS_{total} = SS_{residual} + SS_{group}$
$df_{total} = df_{residual} + df_{group}$

ANOVA: F ratio:

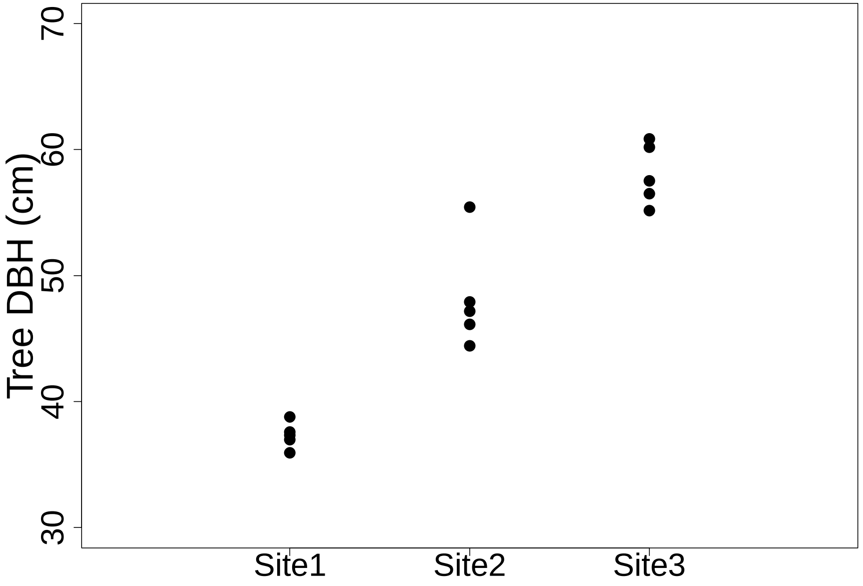
Variation among groups/Variation within groups

- If F-ratio = 1: $MS_{residual} = MS_{group}$
- H_0 is true: no difference between groups
- The bigger the F-ratio is, the more likely H_0 to be rejected

- R squared indicates the fraction of the total variation that is due to differences among groups.
- **R squared = SS of group effects / SS total**
- 0 < R squared < 1

ANOVA in R: aov()

```
> model.aov <- aov(size ~ site, data = data)
> summary(model.aov)
            Df    Sum Sq  Mean Sq   F value    Pr(>F)
site        2     1073.4     536.7     64.54     3.79e-07 ***
Residuals   12    99.8       8.3
```



ANOVA in R: aov()

```
> model.aov <- aov(size ~ site, data = data)
> summary(model.aov)
              Df       Sum Sq  Mean Sq     F value       Pr(>F)
site          2        1073.4       536.7        64.54        3.79e-07 ***
Residuals     12       99.8         8.3
```

n = 15 trees	-> df_{total}	= n-1	= 14
k = 3 sites	-> df_{groups}	= k-1	= 2
	-> $df_{residuals}$	= 14 – 2	= 12

ANOVA in R: aov()

```
> model.aov <- aov(size ~ site, data = data)
> summary(model.aov)
             Df     Sum Sq    Mean Sq    F value     Pr(>F)
site         2      1073.4      536.7      64.54       3.79e-07 ***
Residuals    12     99.8        8.3
```

$MS_{among} = MS_{groups} = SS_{groups} / df_{groups} = 1073.4/2$

$MS_{within} = Ms_{residuals} = SS_{residuals} / df_{residuals} = 99.8/12$

ANOVA in R: aov()

```
> model.aov <- aov(size ~ site, data = data)
> summary(model.aov)
            Df     Sum Sq  Mean Sq    F value    Pr(>F)
site        2      1073.4      536.7      64.54       3.79e-07 ***
Residuals   12     99.8        8.3
```

F value = $MS_{groups}/MS_{residuals}$=536.7/8.3

Much bigger than 1

ANOVA in R: aov()

```
> model.aov <- aov(size ~ site, data = data)
> summary(model.aov)
            Df    Sum Sq  Mean Sq   F value    Pr(>F)
site        2     1073.4    536.7     64.54      3.79e-07 ***
Residuals   12    99.8      8.3
```

P value = Probability of observing data as extreme as this if the null hypothesis were true.

Here, the P value is almost 0, so it is very unlikely that H_0 is true.

Therefore, we accept H_A : ”There is a difference among groups”.

ANOVA

H_A:

Tree size = Mean size + site effect + effects of age, genes, species…

Y = Fitted value + residual (error)

H_0:

Tree size = Mean size + effects of age, genes, species…

Data = Fitted value + residual (error)

SS site effect = SS residual (model without site effect)

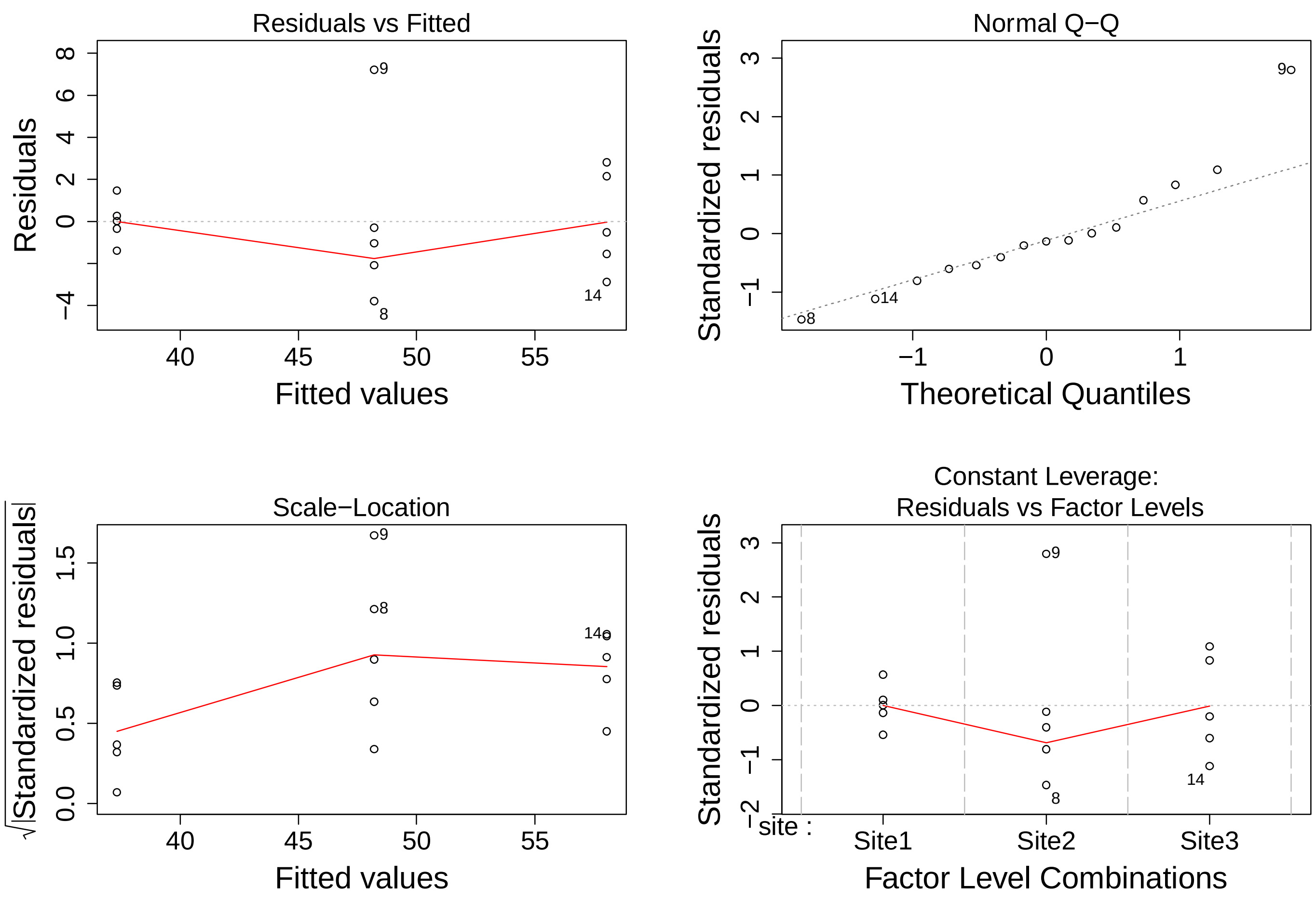
- SS residual (model with site effect)

ANOVA

- Answers 3 questions:
  - Do mean values of Y differ among the sites from which samples were taken?
  - Is the variability of group means larger than expected from the variability within groups?
  - Does a model with site effects describe the data better than a model without site effects?

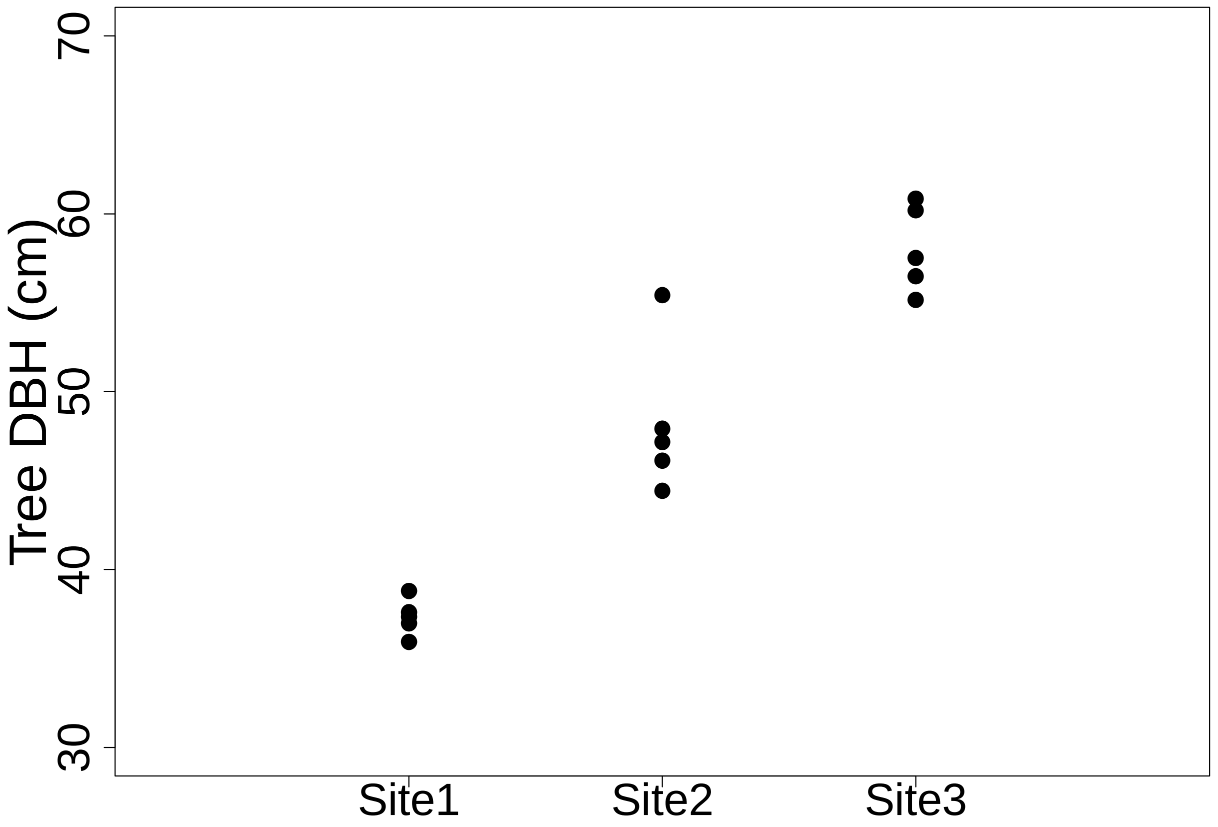
ANOVA in R: checking assumptions

```
> par(mfrow=c(2,2))
> plot(model)
```



One Way-ANOVA: one factor

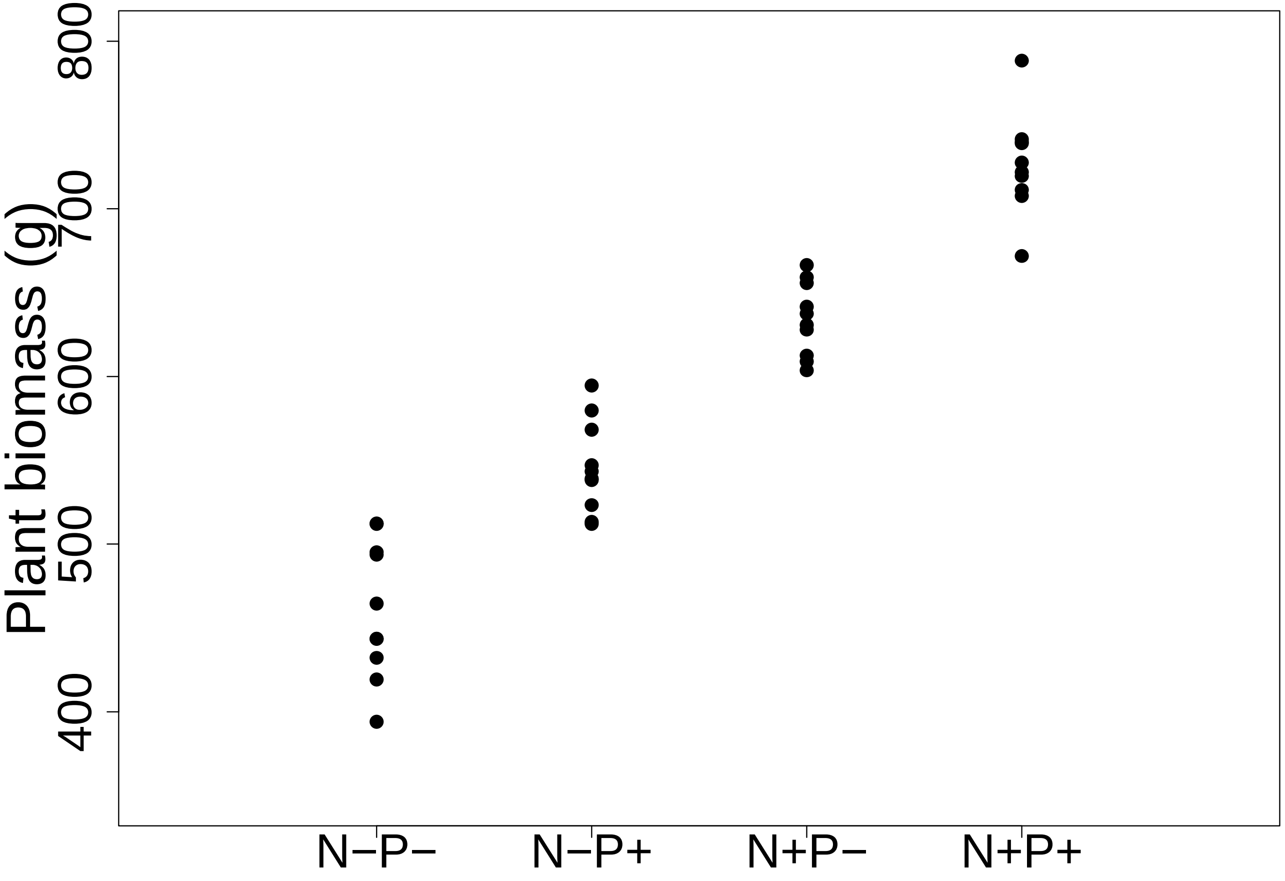
```
model <- aov(Y~X, data=data)
```



Two Way-ANOVA: two factors

- Y ~ X1 * X2
- Y = Plant biomass (g)
- N+/N- : Nitrogen addition / control
- P+/P- : Phosphate addition / control

Two Way-ANOVA: two factors

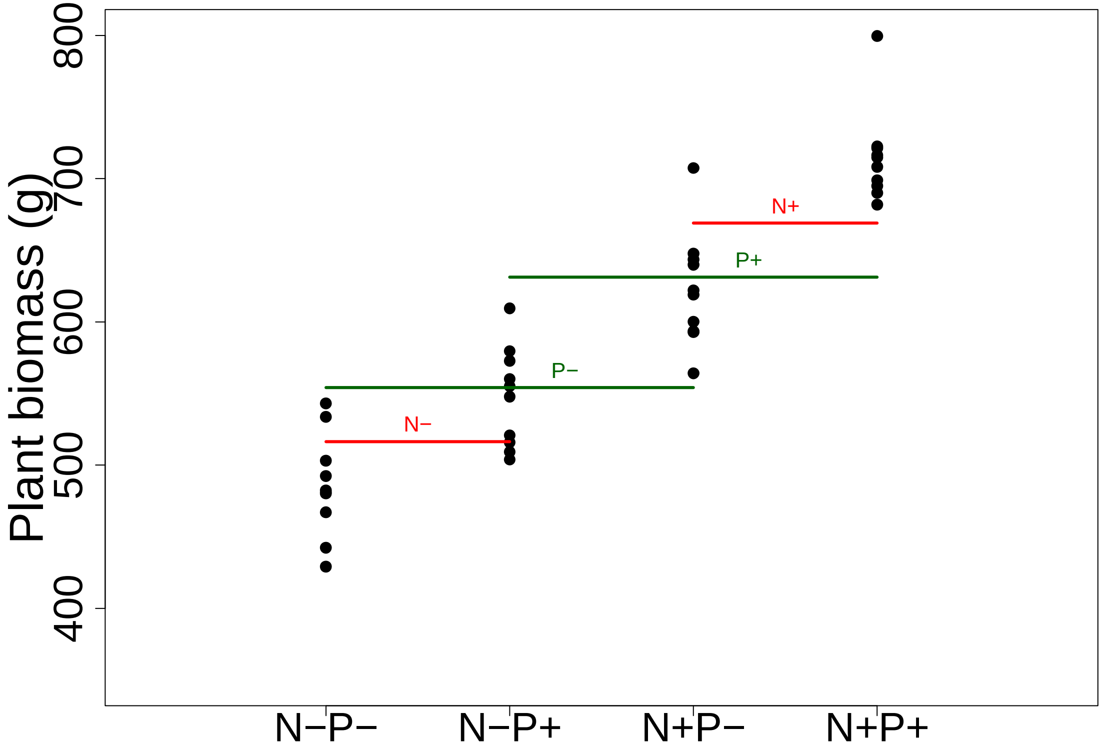


Two Way-ANOVA: two factors

```
> model.aov <- aov(biomass ~ nitrogen * phosphate, data =
fert)
> summary(model.aov)
                   Df Sum Sq Mean Sq F value   Pr(>F)
nitrogen            1 191576  191576 232.983  < 2e-16 ***
phosphate           1  78295   78295  95.217 1.19e-11 ***
nitrogen:phosphate  1      1       1   0.001    0.976
Residuals          36  29602     822
```

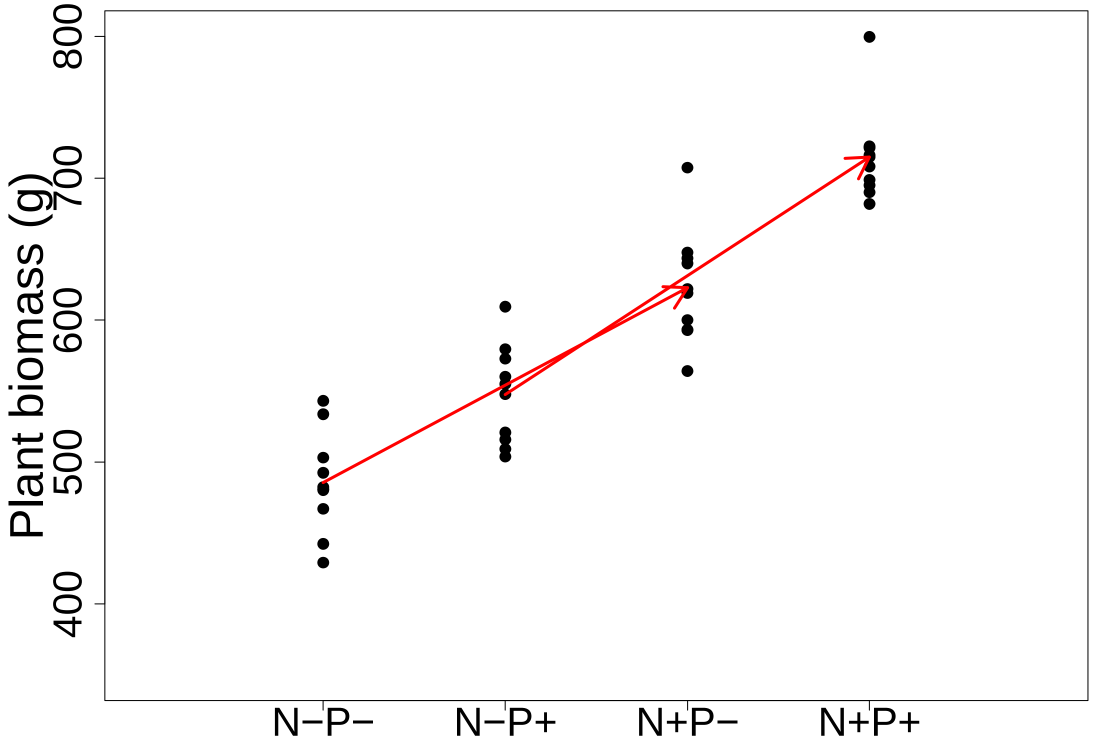
Two Way-ANOVA: two factors

```
> summary(model.aov)
                   Df Sum Sq Mean Sq F value   Pr(>F)    
nitrogen            1 191576  191576 232.983  < 2e-16 ***
phosphate           1  78295   78295  95.217 1.19e-11 ***
nitrogen:phosphate  1      1       1   0.001    0.976    
Residuals          36  29602     822                     
```

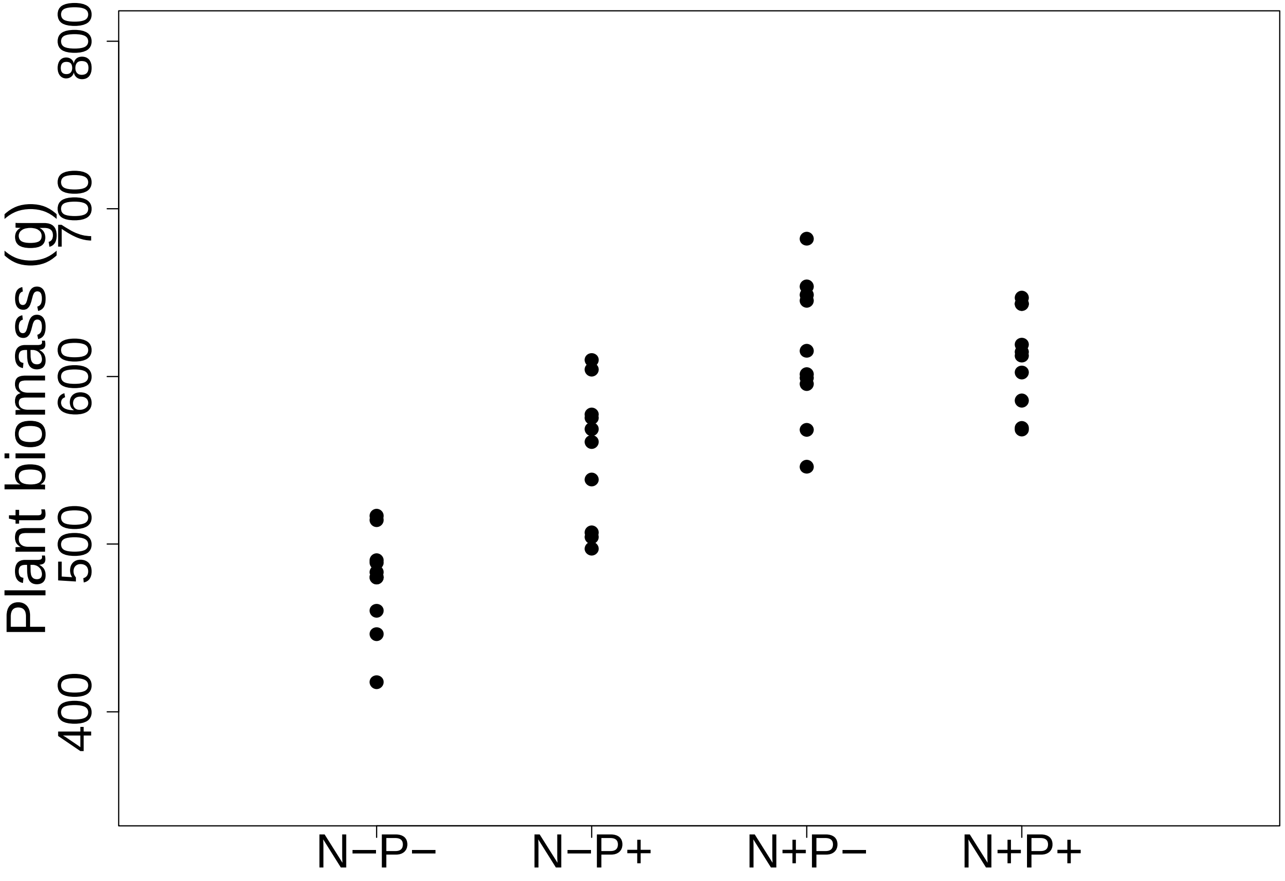


Two Way-ANOVA: two factors

```
> summary(model.aov)
                   Df Sum Sq Mean Sq F value   Pr(>F)    
nitrogen            1 191576  191576 232.983  < 2e-16 ***
phosphate           1  78295   78295  95.217 1.19e-11 ***
nitrogen:phosphate  1      1       1   0.001    0.976    
Residuals          36  29602     822                     
```



Two Way-ANOVA: two factors

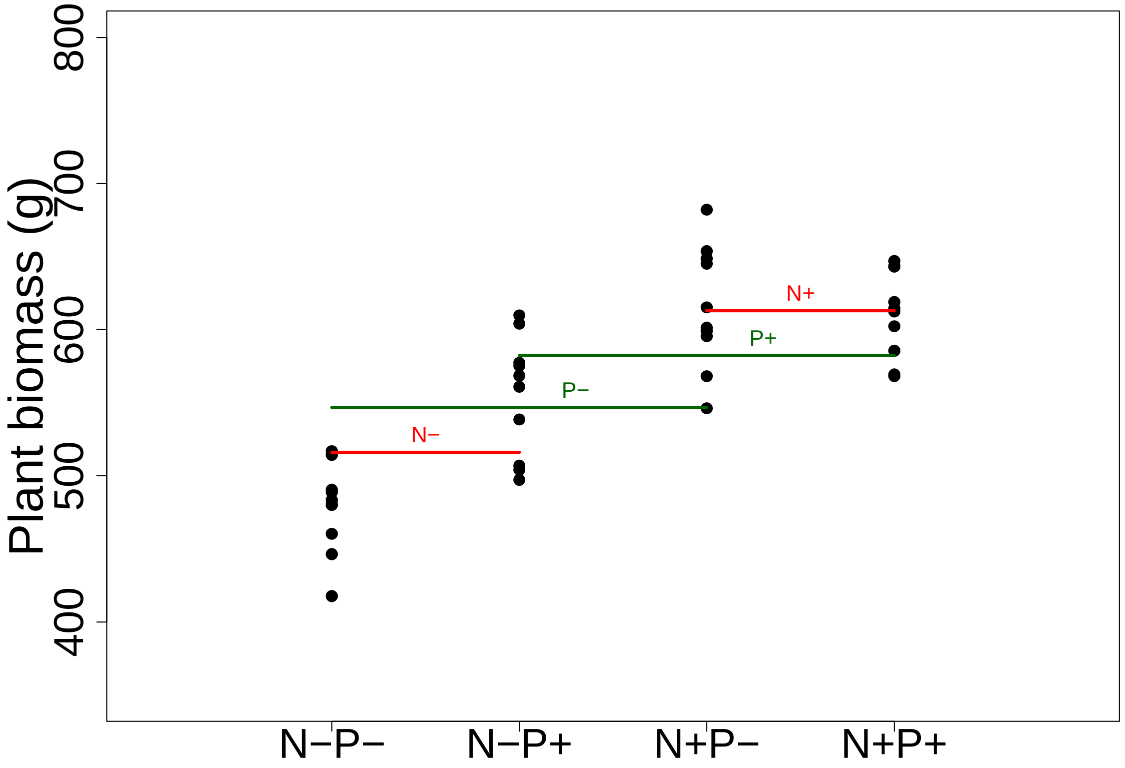


Two Way-ANOVA: two factors

```
> model.aov <- aov(biomass ~ nitrogen * phosphate, data =
fert)
> summary(model.aov)
                   Df Sum Sq Mean Sq F value   Pr(>F)
nitrogen            1  93956   93956  72.499 3.79e-10 ***
phosphate           1  12747   12747   9.836  0.00340 **
nitrogen:phosphate  1  16586   16586  12.798  0.00101 **
Residuals          36  46655    1296
```

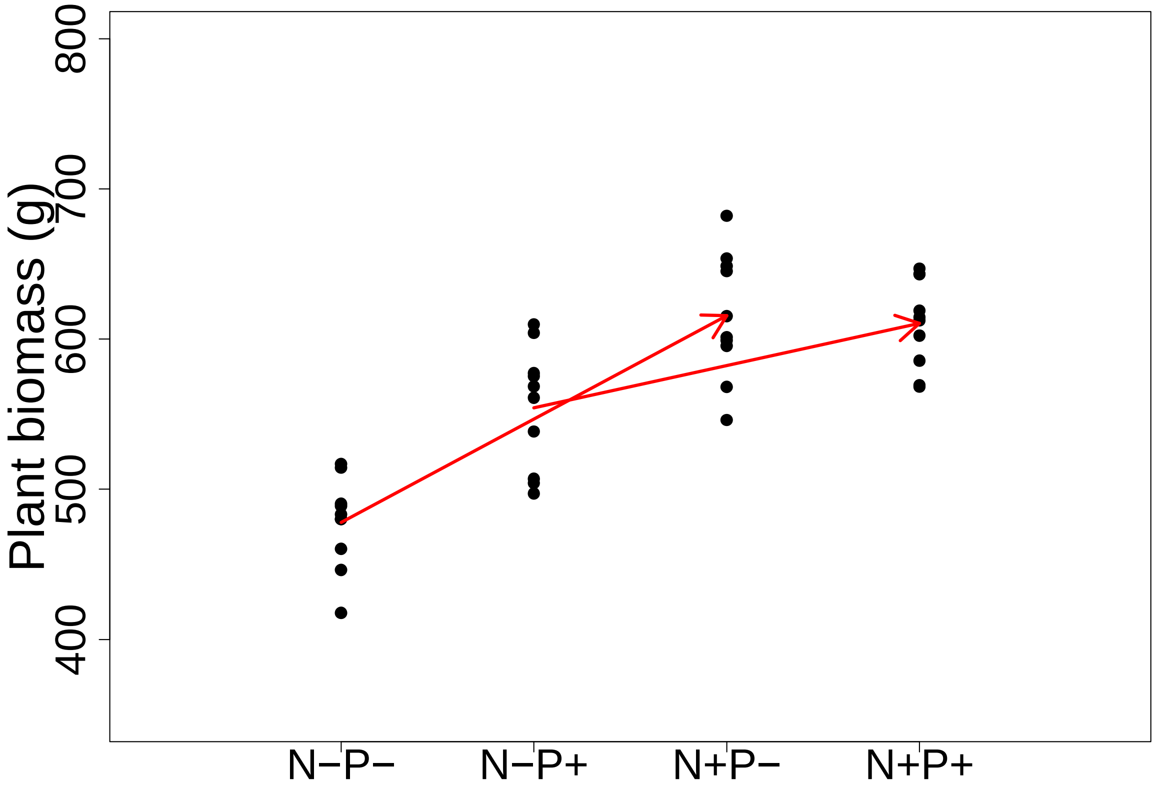
Two Way-ANOVA: two factors

```
> summary(model.aov)
                   Df Sum Sq Mean Sq F value   Pr(>F)    
nitrogen            1  93956   93956  72.499 3.79e-10 ***
phosphate           1  12747   12747   9.836  0.00340 ** 
nitrogen:phosphate  1  16586   16586  12.798  0.00101 ** 
Residuals          36  46655    1296                     
```



Two Way-ANOVA: two factors

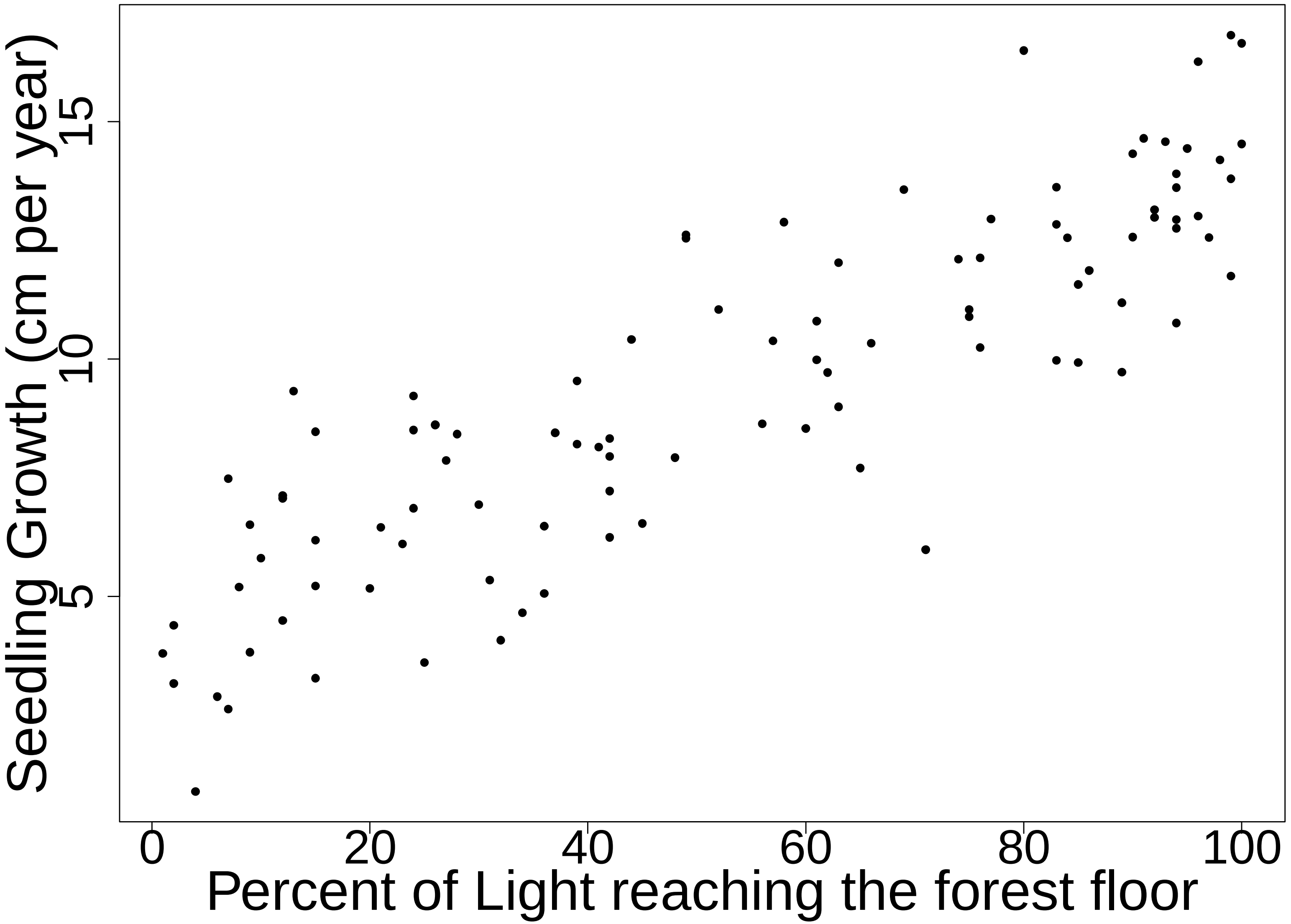
```
> summary(model.aov)
                   Df  Sum Sq  Mean Sq F value    Pr(>F)    
nitrogen            1   93956    93956  72.499 3.79e-10 ***
phosphate           1   12747    12747   9.836  0.00340 ** 
nitrogen:phosphate  1   16586    16586  12.798  0.00101 ** 
Residuals          36   46655     1296                      
```



Linear regression

- Same assumptions as ANOVA:
    - **Independence**
    - **Normality**
    - **Homoscedasticity**
- Here the explanatory variable **X is continuous**
- Instead of difference among groups, we want to model the **intercept** (a= value of Y when X=0) and the **slope** (b) of the regression.
- Y ~ **a** + **b***X

Linear regression

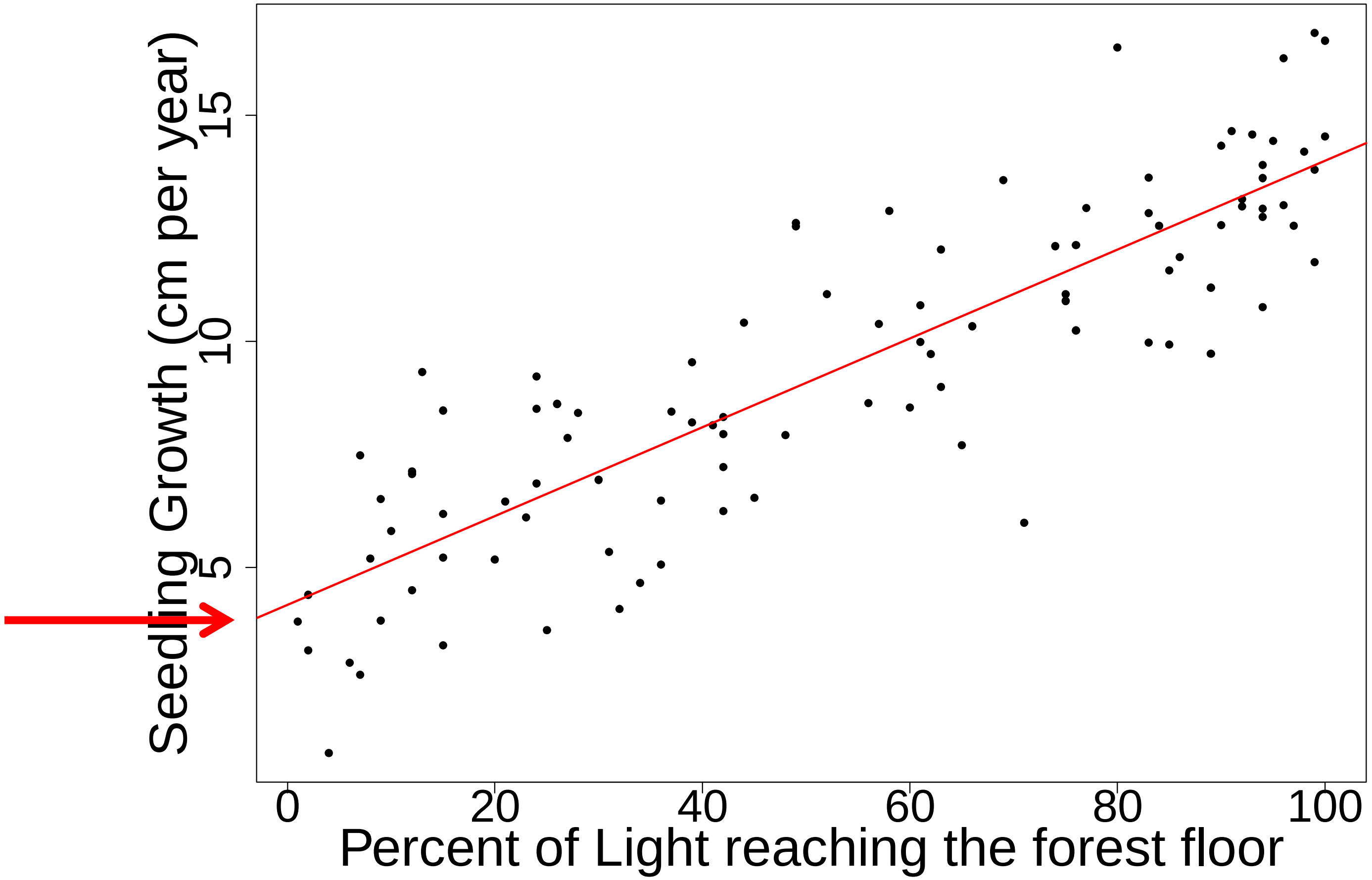


Linear regression in R: lm()

```
> model.lm <- lm(growth ~ light, data = data)
> summary(model.lm)

Call:
lm(formula = growth ~ light, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.1620 -1.1587 -0.0605  1.2966  4.4653 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.168827   0.362514   11.50   <2e-16 ***
light       0.098287   0.005811   16.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.854 on 98 degrees of freedom
Multiple R-squared: 0.7449,	Adjusted R-squared: 0.7423 
F-statistic: 286.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

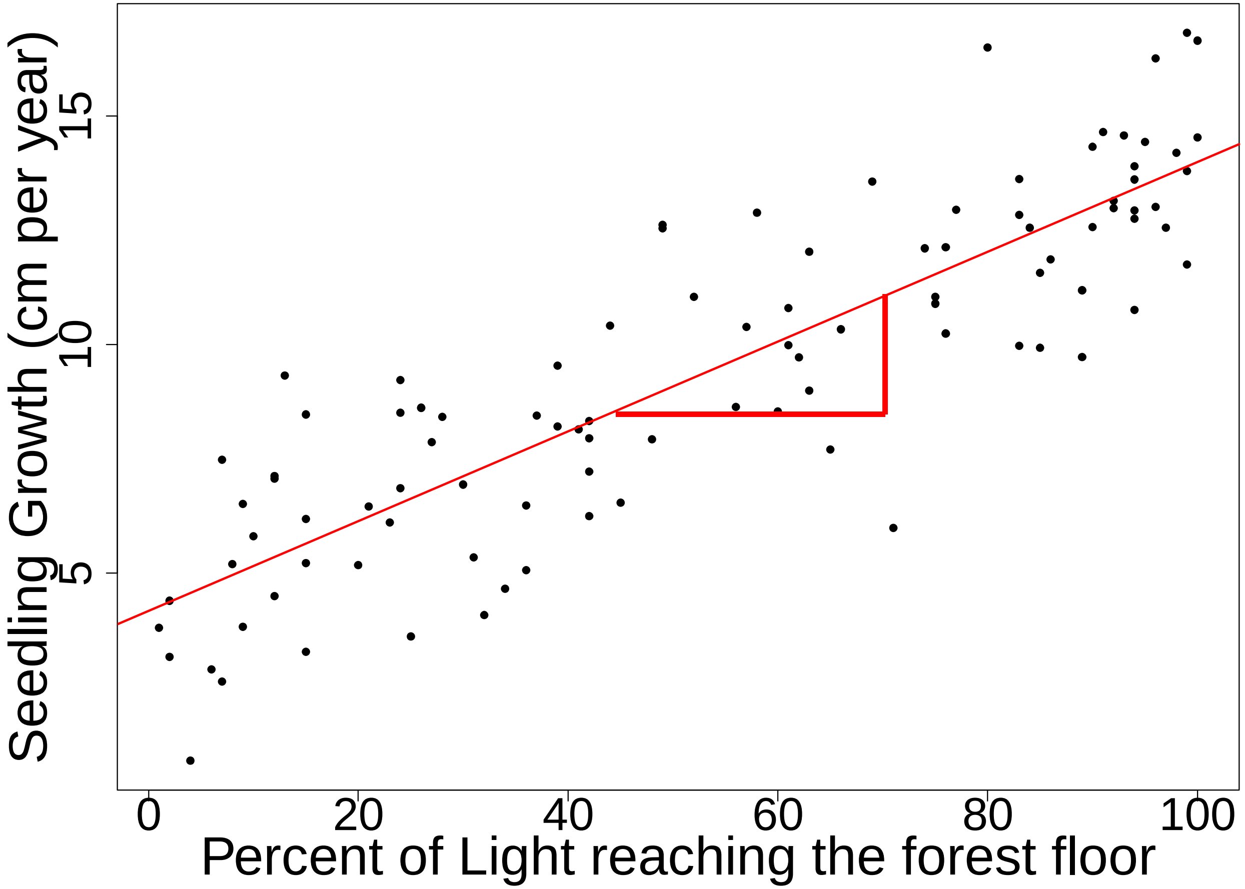
Linear regression in R: lm()

```
> summary(model.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.168827   0.362514   11.50   <2e-16 ***
light       0.098287   0.005811   16.91   <2e-16 ***
```



Linear regression in R: lm()

```
> summary(model.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.168827   0.362514   11.50   <2e-16 ***
light       0.098287   0.005811   16.91   <2e-16 ***
```



Linear regression in R: lm()

```
> model.lm <- lm(growth ~ light, data = data)
> summary(model.lm)

Call:
lm(formula = growth ~ light, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1620 -1.1587 -0.0605  1.2966  4.4653

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.168827   0.362514   11.50   <2e-16 ***
light       0.098287   0.005811   16.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.854 on 98 degrees of freedom
Multiple R-squared: 0.7449,   Adjusted R-squared: 0.7423
F-statistic: 286.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

SE of the mean:

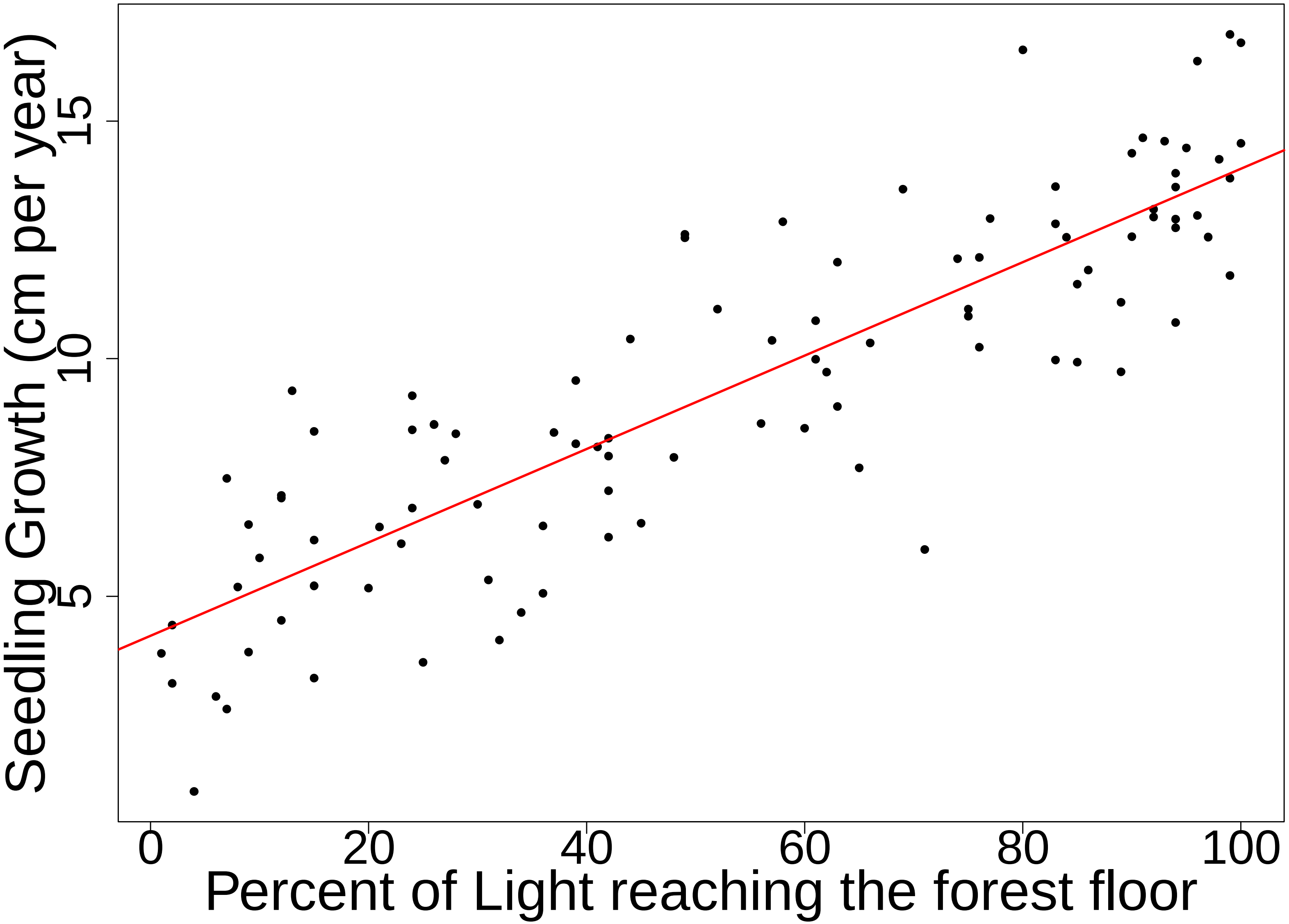
SD/√n

SE: Standard Error

SD: Standard Deviation

n: population size

Linear regression in R: lm()



Linear regression in R: lm()

```
> model.lm <- lm(growth ~ light, data = data)
> summary(model.lm)

Call:
lm(formula = growth ~ light, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1620 -1.1587 -0.0605  1.2966  4.4653

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.168827   0.362514   11.50   <2e-16 ***
light       0.098287   0.005811   16.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.854 on 98 degrees of freedom
Multiple R-squared: 0.7449,   Adjusted R-squared: 0.7423
F-statistic: 286.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

Linear regression in R: lm()

```
> anova(model.lm)
Analysis of Variance Table

Response: growth
          Df Sum Sq Mean Sq F value    Pr(>F)
light      1 983.50  983.50  286.11 < 2.2e-16 ***
Residuals 98 336.88    3.44
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Gives the same output as summary(aov(model.aov))

ANOVA is a special case of LM

- Confusing because we can use the function anova() to get a table of analysis of variance whatever the model was made with (aov() or lm())
- ANOVA = special case of LM when X:factors
- lm() in R works for anova, but the way it’s shown is a bit different.
- Same assumptions

ANOVA is a special case of LM

- In the end, you can always use lm() :
  - As long as your Y is continuous, and all the assumptions are met;
  - You can use the function anova(model) to get the degrees of freedom, SS, MS, F-test and p-value of every X variable in your model
  - You can use the function summary(model) to get the estimated means and SE of each levels, and the differences between levels

Exercise: Forest: ANCOVA (ANalysis of COVAriance)

- Productivity: continuous Y variable that we want to explain
- Forest type: factorial X variable
- Species Diversity: continuous X variable

- ANCOVA: Y ~ fact.X * cont.X