STATISTIQUE

Script et Test statistique

1/ Les tests paramétriques

- Test du Chi2
- Test de Student

2/ Transformations de variable

- Test de normalité
- Test d'homogénéité des variances
- Linéarisation

3/ Les regressions linéaires (Variable continue)

- Linéaire simple
- Multiple
- Avec intercation

4/ ANOVA (Variable catégorielle)

- 1 Facteur
- 2 Facteur
- Avec intercation

5/ Covariance

- Covariance
- Covariance avec interaction

6/ Modèle liéaire généralisé (GLM)

- Modèle
- Représentation graphique

7/ Analyse des résidus

- Normalité
- Homogénéité

Script important

Importation des données .txt

tab <-**read.table**("*chemin d'accès*", header = *TRUE / FALSE* , dec = ", / .")

Dans propriété : avec / ou \\

Decimal utilisé

En-tête

Importation des données .csv

tab <-**read.csv**("*chemin d'accès*", sep = " ; ", header = *TRUE / FALSE* , dec = ", / .")

Séparateur entre les données

Creation de vecteur

V <- **c**(*1* , *2* , *3*, ...)

Formation de list

Creation de matrix

tab <-**matrix**(**data**=c(1, 2, 3, ...), **ncol**= *nbc*, **nrow**=*nbl*, **dimnames**= **list** (c("*nml1*", "*nml2*", ...), c("*nmc1*","*nmc2*", ...)))

Creation d'un tableau

Changer l'affichage des graphiques

par (**mfrow** = **c**(*1* , *2*))

nombre de ligne

nombre de colonne

Creation de graphique

plot(*y* ~ *x* , **data** = *tab*)

nuage de point

boxplot(*y* ~ *x* , **data** = *tab*)

boite

Test Paramétrique

TEST CHI2

Comparaison de 2 échantillon dont les 2 variables sont catégorielles

1/ Creation du tableau de contingence

tab <-**matrix**(**data**=c(1, 2, 3, ...), **ncol**= *nombre de colonne*, **nrow**=*nombre de ligne*, **dimnames**= **list** (c("*nom ligne 1*", "*nom ligne 2*", ...), c("*nom colonne 1*","*nom colonne 2*", ...)))

	A	B	C
a	1	3	5
b	2	4	6

c( ) = Creation d'un vecteur

Attention : Les valeurs doivent être ajouté dans le bonne ordre (comme dans le tableau)

2/ Test du Chi2

chic.test (x= *tab*)

Donne la valeur de **chi2**, **p-value** et le **ddl**

```
> #Tableau de Contingence
> tab1<-matrix(data=c(14,4,3,7,6,8,1,8,0,5),ncol=5,nrow=2,
+       dimnames=list(c("a","b"),c("R","M","L","0","<0")))
> tab1
   R M L 0 <0
a 14 3 6 1  0
b  4 7 8 8  5
> #Test du Chi2
> chisq.test(x=tab)

        Pearson's Chi-squared test

data:  tab
X-squared = 17.092, df = 4, p-value = 0.001855
```

H0 : Pas de difference significative de la croissance entre les 2 espèces
H1 : INVERSE

p>0,01	p<0,01
Acceptation de H0	Rejet de H0

P-value = 0,0019 < 0,01
=> Rejet de H0
Il y a une difference significative de croissance entre les 2 espèces

TEST DE STUDENT

Comparaison de 2 moyennes de variable continue

1/ Création de 2 vecteurs à comparer

V1 <- **c** (1, 2, 3)
V2 <- **c** (a, b, c)

V1	V2
1	a
2	b
3	c

2/ Calcul des moyennes

mean (*V2*)
mean (*V1*)

3/ Test de Student

t.test (*V1*, *V2*, **paired** (*TRUE*/*FALSE*), **var.equal** = *TRUE*/*FALSE*, **conf.level** = *0,01*)

Echantillons appariés ?

Variances égales ?

intervalle de confiance

```
> #Creation des vecteurs
> MS<-c(5.7,8.2,6.9,6,3.8,3.9,4.3,2.7)
> MH<-c(4,5.8,4.9,4.8,3.6,3.5,2.9,1.2)
> #Calcul des moyennes
> mean(MS)
[1] 5.1875
> mean(MH)
[1] 3.8375
> #Test de Student
> t.test(MS,MH,paired=TRUE,var.equal=FALSE,conf.level=0.90)

        Paired t-test

data:  MS and MH
t = 5.1025, df = 7, p-value = 0.001396
alternative hypothesis: true mean difference is not equal to 0
90 percent confidence interval:
 0.8487416 1.8512584
sample estimates:
mean difference
           1.35
```

p>0,01	p<0,01
Acceptation de H0	Rejet de H0

H0 : Pas de difference significative entre les 2 moyennes (methode)
H1 : INVERSE

P-value = 0,0014 < 0,01
=> Rejet de H0
Il y a une difference significative entre les 2 methode

Transformations de variable

TEST DE NORMALITÉ

Comparaison d'une distribution à une distribution théorique continue normale : Test de Kolmogorov-Smirnov

H0 : Pas de difference entre la distribution observée de xi et la distribution théorique de xi attendue sou l'hypothèse que xi suit une loi normal

H1 : INVERSE

1/ Création de la distribution théorique

Xth <- **rnorm** (1000, mean (*tab*), sd(*tab*))

Si D tend vers 0 - p>0,01	Si D tend vers 1 - p<0,01
Acceptation de H0	Rejet de H0

2/ Test de normalité

observée ← **ks.test**(*tab*, *Xth*) → théorique

Exemple:

```
> # test de normalité
> tab1<-c(3.62, 3.48, 3.45, 3.42, 3.80, 3.71, 3.51, 3.70, 3.40, 3.55, 3.73, 3.48,
+       3.53, 3.39, 3.29, 3.67, 3.62, 3.63, 3.67, 3.42, 3.47, 3.43, 3.75, 3.72,
+       3.56, 3.39, 3.39, 3.67, 3.38, 3.70, 3.69, 3.48)
> X1<-rnorm(1000,mean(tab1),sd(tab1))
> ks.test(tab1,X1)

        Asymptotic two-sample Kolmogorov-Smirnov test

data:  tab1 and X1
D = 0.13975, p-value = 0.58
alternative hypothesis: two-sided
```

Ici, **D tend vers 0** donc on ne peux pas rejeter **H0**

P-value = 0,58 >0,01 => Grande

Il n'y a pas de difference significative entre les 2 distributions

TEST D'HOMOGÉNÉITÉ

Comparaison de 2 variances de 2 échantillons observés indépendants

H0 : Pas de difference entre les variances observés des 2 échantillon

H1 : Difference significative entre les 2 variances

1/ Calcul des variances

V1 <- **var**(*tab1*)
V2 <- **var**(*tab2*)

2/ Test de Fisher

var.test(*tab1*, *tab2*)

Variance la + grande

Variance la + petite

Si F tend vers 1 - p>0,01	Si F s'éloigne de 1 - p<0,01
Acceptation de H0	Rejet de H0

Exemple:

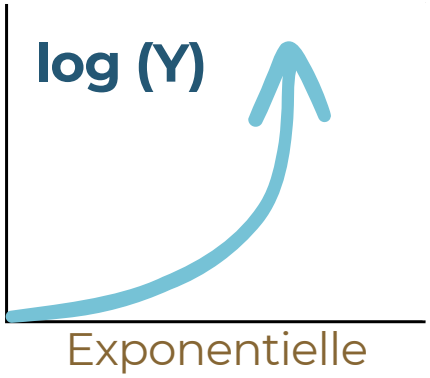
```
> PN<-c(110, 115, 80, 75, 120, 96, 73, 105)
> PS<-c(100, 89, 76, 121, 68, 75, 112, 94)
> #Calcule des variances
> V1<-var(PN)
> V2<-var(PS)
> #Test de Fisher
> var.test(PS,PN)

        F test to compare two variances

data:  PS and PN
F = 1.0014, num df = 7, denom df = 7, p-value = 0.9986
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2004813 5.0018309
sample estimates:
ratio of variances
          1.001386
```

Ici, **F tend vers 1** donc on peut accepter H0

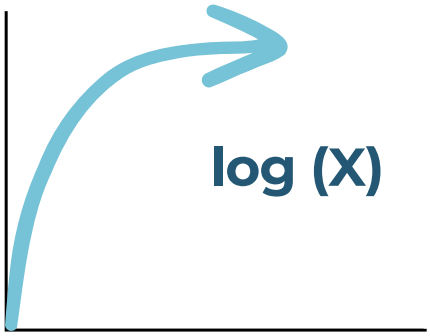
p-value =0,998>0,01, donc il n'y a pas de difference significative entre les variances
On accepte **HO**

LINÉARISATION

Utilisé lorsqu'une courbe d'echantillon est exponentielle ou logarithmique



Exponentielle



Logarithmique

1/ Definir un tableau avec X et Y

tab **<- data.frame** (*nom* = **c** (*donnée*, ...), *nom* = **c**(*donnée*, ...))

2/ Création du graphique avant modification

plot(*y* ~ *x*, **data**= *tab*)

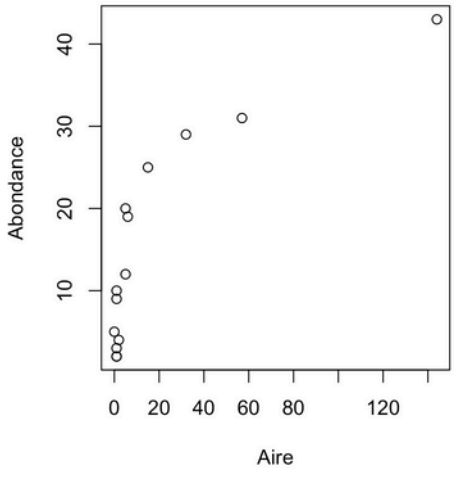
3/ Linéarisation

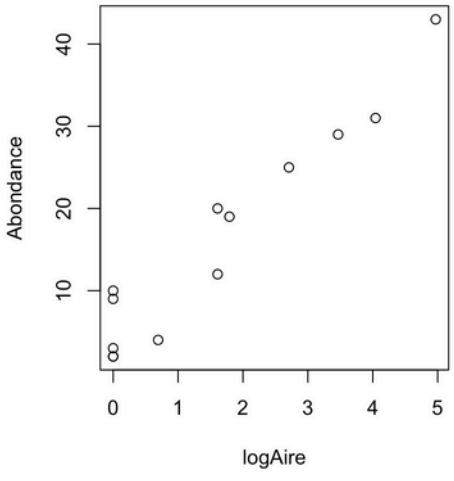
log<- **log** (*tab* $ *xy*)

4/ Création du grapgique après linéarisation

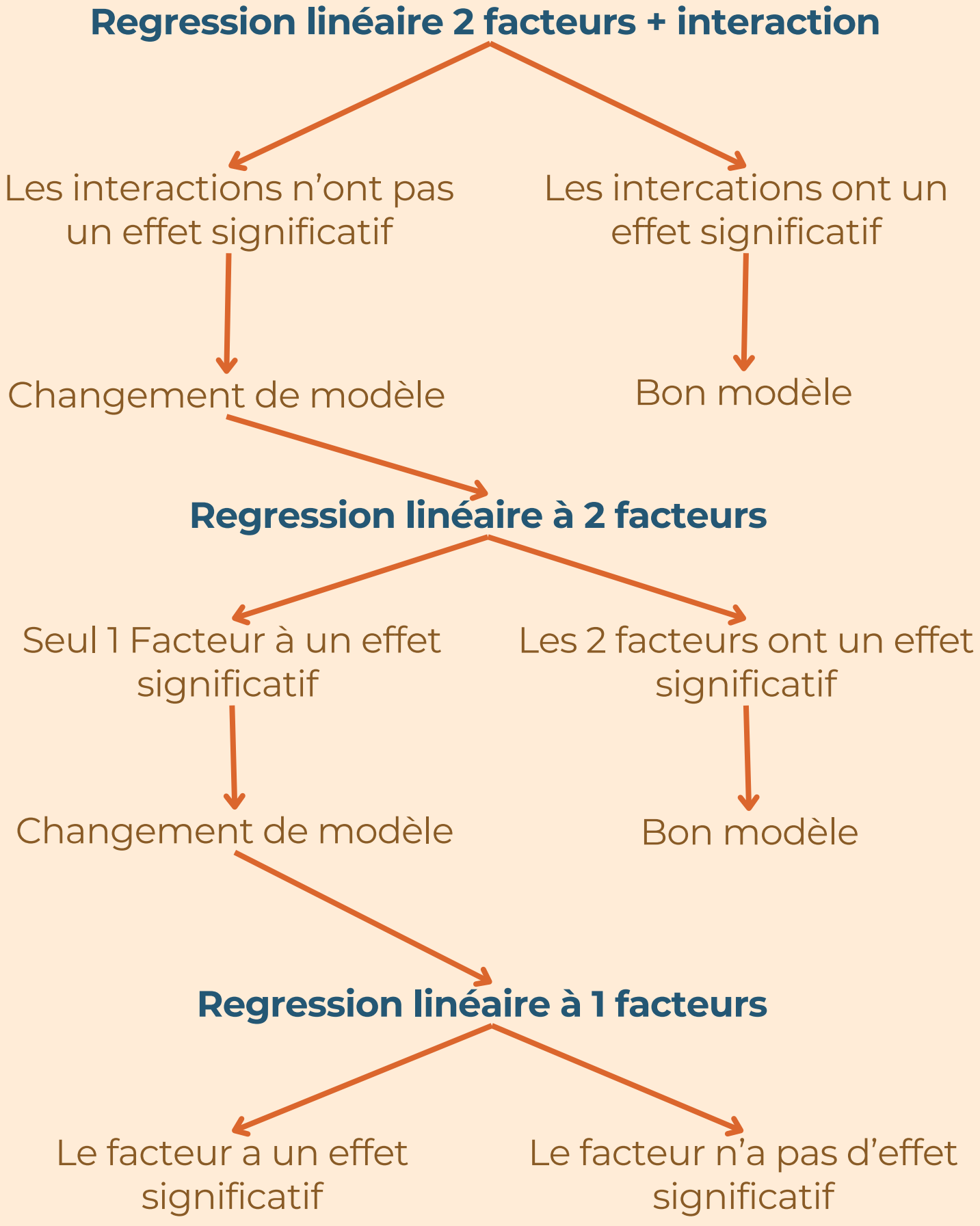
plot(*y* ~ *log*, **data**= *tab*)

```
> tab<-data.frame(ID=c(1:14),
+   Abondance=c(5,2,2,9,3,10,4,12,20,19,25,29,31,43),
+   Aire=c(0,1,1,1,1,1,2,5,5,6,15,32,57,144))
> par(mfrow=c(1,2))
> #Création du graphique
> plot(Abondance~Aire,data=tab)
> #Graphique de type log
> logAire<-log(tab$Aire)
> #Graphique
> plot(Abondance~logAire,data=tab)
```





REGRESSION LINÉAIRE



REGRESSION SIMPLE

Condition à respecter pour faire une regression linéaire :

- **Normalité** de Y
- **Homogénéité**
- **Linéarité**

H0 : X n'influence pas Y
H1 : X influence Y

1/ Création du nuage de point

plot (*tab$X*, *tab$Y*)

2/ Création de la droite

M<-**lm** (*y* ~ *x*, **data** = *tab*)

3/ Test de comparaison

anova(*M*)

Si F tend vers 1 - p>0,01	Si F s'éloigne de 1 - p<0,01
Acceptation de H0	Rejet de H0

4/ Determibation de l'equation de droite

summary(*M*)

Intercetp = **b**
X = **a**
R-squared = **[1 ; 0]**

Les "error" ne doivent pas être supérieur aux "estimate"

```
tab2<-read.table("/Users/eleane/Desktop/BEE/Outils analytique /Statistique/Atelier/dataAtelier4.txt",
                 header=TRUE,dec=".")
plot(tab2$LT,tab$Racines)
M1<-lm(Racines~LT,data=tab2)
anova(M1)
summary(M1)
```

```
Analysis of Variance Table

Response: Racines
          Df Sum Sq Mean Sq F value    Pr(>F)
LT         1 37.983  37.983    32.7 3.458e-06 ***
Residuals 29 33.685   1.162
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Call:
lm(formula = Racines ~ LT, data = tab2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0505 -0.7821  0.0590  0.4369  2.6127

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.28742    0.66548  -1.935   0.0629 .
LT           0.05821    0.01018   5.718 3.46e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.078 on 29 degrees of freedom
Multiple R-squared:   0.53,     Adjusted R-squared:  0.5138
F-statistic:  32.7 on 1 and 29 DF,  p-value: 3.458e-06
```

F= 32,7 -> il est eloigné de 1
P= 3,458e-06< 0,01
Donc **on rejette H0** et **accepte H1**

Racine = -1,28742 (**b**) +0,05821 (**a**) **LT**

R-squared = 0,5138
Donc il y a **51%** de la dispersion total des racines expliqué par la longueur des feuilles (LT)

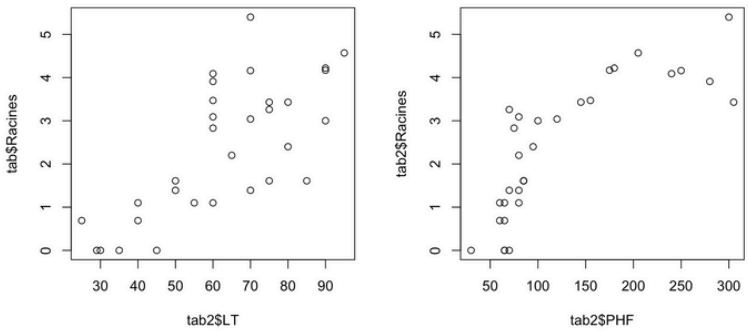
REGRESSION MULTIPLE

Regression où l'on prend en compte 2 facteurs : On test des relations multiples en ajoutant une dimension supplémentaire

Modèle ADDITIF

Création du modèle
(comme pour une simple mais en ajoutant un facteur)

M<-**lm**(*y* ~ *x1* + *x2*, **data**= *tab*)
anova(*M*)
summary(*M*)



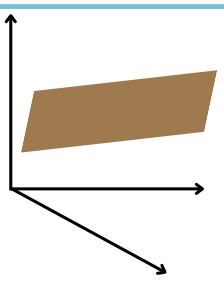

```
tab2<-read.table("/Users/eleane/Desktop/BEE/Outils analytique /Statistique/Atelier/dataAtelier4.txt",
                 header=TRUE,dec=".")
par(mfrow=c(1,2))
plot(tab2$LT,tab$Racines)
plot(tab2$PHF,tab2$Racines)
M<-lm(Racines~LT+PHF,data=tab2)
anova(M)
summary(M)
```

```
Analysis of Variance Table

Response: Racines
          Df Sum Sq Mean Sq F value    Pr(>F)
LT         1 37.983  37.983  79.852 1.085e-09 ***
PHF        1 20.366  20.366  42.816 4.296e-07 ***
Residuals 28 13.319   0.476
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Call:
lm(formula = Racines ~ LT + PHF, data = tab2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.50533 -0.43046  0.06791  0.35581  1.33523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.385798   0.426129  -3.252  0.00298 **
LT           0.036763   0.007292   5.041 2.48e-05 ***
PHF          0.011685   0.001786   6.543 4.30e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6897 on 28 degrees of freedom
Multiple R-squared:  0.8142,    Adjusted R-squared:  0.8009
F-statistic: 61.33 on 2 and 28 DF,  p-value: 5.86e-11
```

Si F tend vers 1 - p>0,01	Si F s'éloigne de 1 - p<0,01
Acceptation de H0	Rejet de H0

F sont éloignés de 1
P < 0,01
Donc **on rejette H0** et **accepte H1**

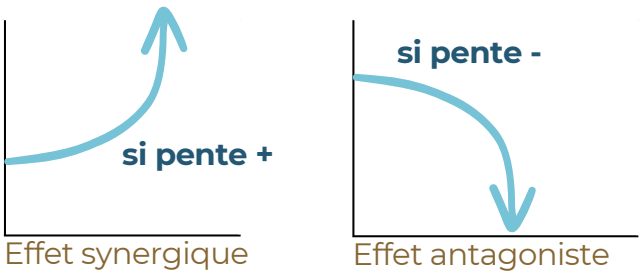


Racine = -1,3858 (c) +0,0117 (b) **PHF** + 0,05821 **LT**
(a)

R-squared = 0,8009
Donc il y a **80%** de la dispersion total des racines expliqué par la longueur des feuilles (LT) et le potentiel hydrique foliaire (PHF)

REGRESSION AVEC INTERCATION

Regression où l'on prend en compte 2 facteurs et leur interaction



Effet synergique Effet antagoniste

Création du modèle

M<-**lm**(*y ~ x1 + x2 + x1*x2* , **data**= *tab*)

Si F tend vers 1 - p>0,01	Si F s'éloigne de 1 - p<0,01
Acceptation de H0	Rejet de H0

anova(*M*)

```
Analysis of Variance Table

Response: Racines
          Df Sum Sq Mean Sq F value    Pr(>F)
LT         1 37.983  37.983 78.2961 1.830e-09 ***
PHF        1 20.366  20.366 41.9818 6.045e-07 ***
LT:PHF     1  0.220   0.220  0.4544     0.506
Residuals 27 13.098   0.485
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F tend vers 1
P=0,506 > 0,01
Donc **on accepte H0**
Pas d'effet de l'intercation, pas le bon modèle

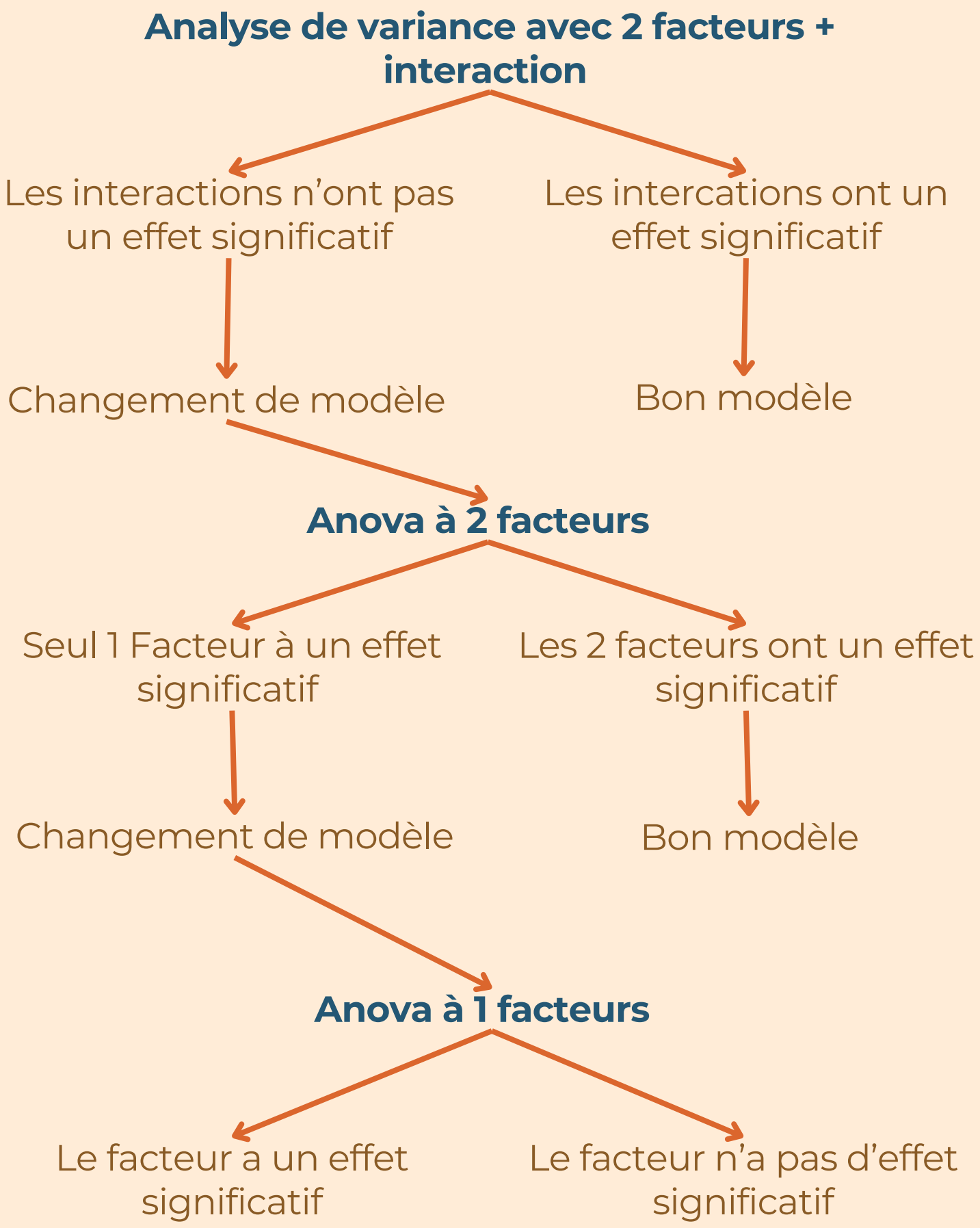
summary(*M*)

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.4022 -0.4172  0.1090  0.3237  1.2960

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.912e+00  8.917e-01  -2.145  0.04114 *
LT           4.533e-02  1.469e-02   3.086  0.00464 **
PHF          1.779e-02  9.231e-03   1.927  0.06457 .
LT:PHF      -9.095e-05  1.349e-04  -0.674  0.50596
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6965 on 27 degrees of freedom
Multiple R-squared:  0.8172,    Adjusted R-squared:  0.7969
F-statistic: 40.24 on 3 and 27 DF,  p-value: 4.211e-10
```

$$\widehat{\text{Racine}} = \underset{d}{-1{,}912} + \underset{c}{1{,}779e\text{-}02}\ \text{PHF} + \underset{b}{4{,}533e\text{-}02}\ \text{LT} - \underset{d}{9{,}095e\text{-}05}\ \text{*LT*PHF}$$

R-squared = 0,7969
=> Donc il y a **79%** de la dispersion total des racines expliqué par la longueur des feuilles (LT), le potentiel hydrique foliaire (PHF) et l'interaction de ces 2 paramètres

Anova (variable catégorielle)



ANOVA À 1 FACTEUR

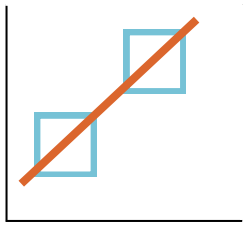
Analyse de variance avec un facteur categoriel

1/Rendre des variables catégorielles

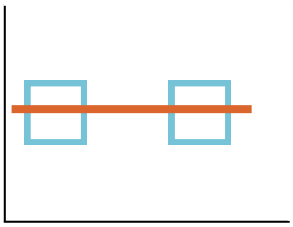
tab$x1*<-**as.factor**(*tab$x1)

2/Representation graphique

boxplot(*y* ~ *x1* , **data** = *tab*)



Effet de X sur Y



Pas d'effet de X sur Y

Si F tend vers 0 - p>0,01	Si F tend vers 0 l'infini - p<0,01
Acceptation de H0	Rejet de H0

3/ Création et analyse du modèle

M <- **lm** (*y* ~ *x1*, **data** = *tab*)

H0 : X n'influence pas Y
H1 : X influence Y

anova(*M*)

```
Response: Racines
          Df Sum Sq Mean Sq F value    Pr(>F)
Gpe        2 52.913 26.4566  39.497 7.067e-09 ***
Residuals 28 18.756  0.6698
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F (=39,5) tend vers l'infini
P=7,067e-09 < 0,01
Donc **on rejette H0**
Effet des groupes sur les racines

Nominateur

18,75 >> 0,6698

Dénominateur

Dénominateur supérieur au Nominateur
=> effet de Gpe fort

summary(*M*)

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.45667 -0.41344 -0.06667  0.38333  1.86812

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1719     0.2046   5.727 3.82e-06 ***
GpeB          1.8948     0.3410   5.556 6.08e-06 ***
GpeC          3.2631     0.3918   8.329 4.62e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8184 on 28 degrees of freedom
Multiple R-squared:  0.7383,    Adjusted R-squared:  0.7196
F-statistic:  39.5 on 2 and 28 DF,  p-value: 7.067e-09
```

= moyenne des racines de GpeA

= écart-type, marge d'erreur de la moyenne

Moyenne GpeB = Moyenne Gpe A + Ecart avec B
= 1,17+ 1,89 = 3,06

Changer de referentiel pour avoir l'eccart entre Bet C

tab$x1 <- **relevel**(*tab$x1*, "*b*")

ANOVA À 2 FACTEURS

Analyse de variance avec deux facteur catégoriel

Même script

```
tab$x1<-as.factor(tab$x1)
boxplot( y ~ x1 + x2, data = tab)
M <- lm ( y ~ x1 + x2, data = tab)
anova(M)
summary(M)
tab$x1 <- relevel(tab$x1, "b")
```

H0 : X1+X2 n'influence pas Y
H1 : X1+X2 influence Y

Si F tend vers 0 - p>0,01	Si F tend vers 0 l'infini - p<0,01
Acceptation de H0	Rejet de H0

```
Response: Racines
          Df Sum Sq Mean Sq F value    Pr(>F)
Gpe        2 52.913 26.4566 41.9162 5.253e-09 ***
Origine    1  1.714  1.7138  2.7153     0.111
Residuals 27 17.042  0.6312
```

F (=41,9) tend vers l'infini
P=5,253e-09 < 0,01
Donc **on rejette H0**
Effet des groupes sur les racines

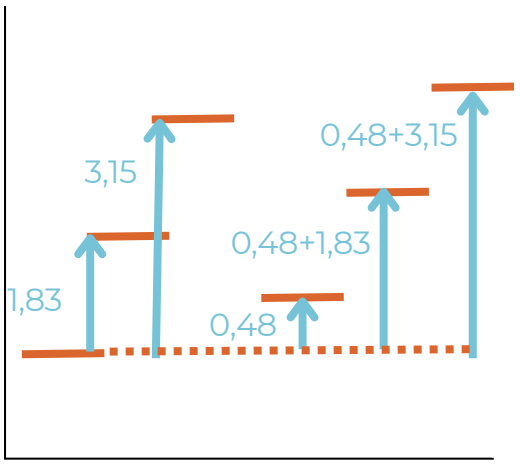
```
Call:
lm(formula = Racines ~ Gpe + Origine, data = tab2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.66922 -0.41269 -0.02617  0.32643  1.59911

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9626     0.2357   4.084 0.000355 ***
GpeB          1.8383     0.3328   5.524 7.48e-06 ***
GpeC          3.1535     0.3861   8.168 9.00e-09 ***
OrigineUSA    0.4782     0.2902   1.648 0.110982
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7945 on 27 degrees of freedom
Multiple R-squared:  0.7622,    Adjusted R-squared:  0.7358
F-statistic: 28.85 on 3 and 27 DF,  p-value: 1.424e-08
```

= moyenne des racines de GpeA

= difference de moyenne des racines entre EurA et EurB/EurC

= difference de moyenne des racines entre EurA et USA A

On peut additionner les ecart car il s'agit d'un modèle sans intercation et donc **additif**



ANOVA À 2 FACTEURS+INTERACTION

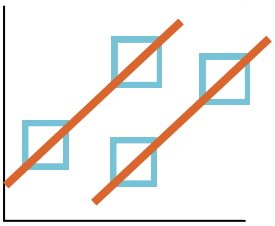
Analyse de variance avec deux facteur catégoriel + interaction de ces 2 facteurs

Même script

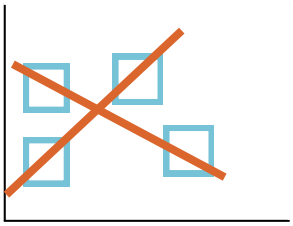
```
tab$x1<-as.factor(tab$x1)
boxplot( y ~ x1 + x2, data = tab)
M <- lm ( y ~ x1 + x2, data = tab)
anova(M)
summary(M)
tab$x1 <- relevel(tab$x1, "b")
```

H0 : X1+X2+X2*X1 n'influence pas Y
H1 : X1+X2+X1*X2 influence Y

Si F tend vers 0 - p>0,01	Si F tend vers 0 l'infini - p<0,01
Acceptation de H0	Rejet de H0



Pas d'interaction



Interaction

```
Response: Racines
            Df Sum Sq Mean Sq F value    Pr(>F)
Gpe          2 52.913 26.4566 44.4996 5.794e-09 ***
Origine      1  1.714  1.7138  2.8826    0.1020
Gpe:Origine  2  2.178  1.0892  1.8320    0.1809
Residuals   25 14.863  0.5945
```

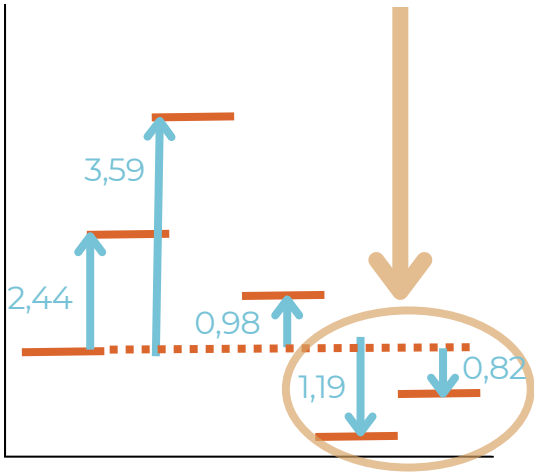
F (=44,5) tend vers l'infini
P=5,794e-09 < 0,01
Donc **on rejette H0**
Différence significative des effets de Gpe

Changment de modèle

MAIS : Interaction pas significativement différente (p>0,01)
Donc pas d'effet de l'interaction des 2 facteurs

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.3620 -0.4764 -0.1114  0.4068  1.4556

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.7444     0.2570   2.896  0.00773 **
GpeB              2.4406     0.4633   5.267 1.87e-05 ***
GpeC              3.5856     0.6028   5.949 3.29e-06 ***
OrigineUSA        0.9770     0.3886   2.514  0.01873 *
GpeB:OrigineUSA  -1.1900     0.6469  -1.839  0.07776 .
GpeC:OrigineUSA  -0.8195     0.7726  -1.061  0.29897
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7711 on 25 degrees of freedom
Multiple R-squared:  0.7926,    Adjusted R-squared:  0.7511
F-statistic: 19.11 on 5 and 25 DF,  p-value: 8.028e-08
```

Valeur aberrante car interaction non significative



Covariance
(continue et catégorielle)

Analyse de variance avec 1 variable catégorielle, 1 variable continue et interaction

- Les interactions n'ont pas un effet significatif → Changement de modèle
- Les intercations ont un effet significatif → Bon modèle

Changement de modèle → **Analyse de variance avec 1 variable catégorielle et 1 variable continue**

- Seul 1 Facteur à un effet significatif → Changement de modèle
- Les 2 facteurs ont un effet significatif → Bon modèle

Changement de modèle → **Analyse de variance avec 1 variable**

- Le Facteur n'à pas un effet significatif
- Le facteur a un effet significatif

COVARIANCE À 2 FACTEURS

Analyse de variance avec un facteur catégoriel + un facteur continu

H0 : X1+X2+n'influence pas Y
H1 : X1+X2 influence Y

1/ Representation graphique

Plot(*tab*$*x1* , *tab*$*y*)

→ Variable continue

boxplot(*y* ~ *x2*, **data** = *tab*)

→ Variable catégorielle

2/ Rendre la variable catégorielle

tab$*x2*<-**as.factor**(*tab*$*x2*)

3/ Création et analyse du modèle

M<- **lm**(*y* ~*x1* + *x2*, **data** = *tab*)
anova(*M*)
summary(*M*)
tab$*x1* <- **relevel**(*tab*$*x2*, "*b*")

Si F tend vers 0 - p>0,001	Si F tend vers 0 l'infini - p<0,001
Acceptation de H0	Rejet de H0

anova(*M*)

```
Response: Racines
          Df Sum Sq Mean Sq F value    Pr(>F)
Gpe        2 52.913 26.4566 47.4286 1.46e-09 ***
LT         1  3.694  3.6945  6.6231   0.01588 *
Residuals 27 15.061  0.5578
```

F (=47,4) tend vers l'infini
P=1,46e-09 < 0,001
Donc **on rejette H0**
Différence significative des effets de Gpe et LT

summary(*M*)

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.51043 -0.48020 -0.05323  0.32102  1.42981

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.051362   0.510673  -0.101  0.92063
GpeB         1.357347   0.374774   3.622  0.00119 **
GpeC         2.571116   0.447367   5.747 4.13e-06 ***
LT           0.024193   0.009401   2.574  0.01588 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7469 on 27 degrees of freedom
Multiple R-squared:  0.7899,    Adjusted R-squared:  0.7665
F-statistic: 33.83 on 3 and 27 DF,  p-value: 2.73e-09
```

= ordonnée a l'origine de GpeA

= pente de LT par rapport à GpeA

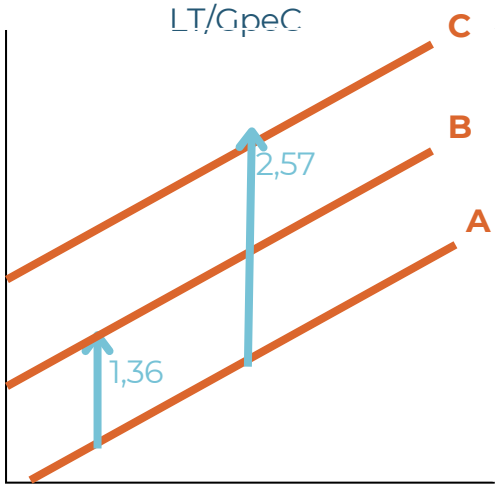
= difference entre LT/GpeA et LT/GpeB ou LT/GpeC

$\widehat{Racines}$ = -0,051 + 0,024 **LT** A

$\widehat{Racines}$ = 1,36-0,051 + 0,024 **LT** B

$\widehat{Racines}$ = 2,57-0,051 + 0,024 **LT** C

ADDITIVITÉ



COVARIANCE À 2 FACTEURS + INTERACTION

Analyse de variance avec un facteur catégoriel + un facteur continu + l'interaction des 2 facteurs

Même script

```
tab$x1<-as.factor(tab$x1)
M <- lm ( y ~ x1 + x2 + x1 * x2 , data = tab)
anova(M)
summary(M)
tab$x1 <- relevel(tab$x1, "b")
```

H0 : X1+X2+ X1*X2 n'influence pas Y
H1 : X1+X2+X1*X2 influence Y

Si F tend vers 0 - p>0,01	Si F tend vers 0 l'infini - p<0,01
Acceptation de H0	Rejet de H0

```
Response: Racines
          Df Sum Sq Mean Sq F value    Pr(>F)
Gpe        2 52.913 26.4566 59.6563 3.041e-10 ***
LT         1  3.694  3.6945  8.3306  0.007922 **
Gpe:LT     2  3.974  1.9870  4.4805  0.021727 *
Residuals 25 11.087  0.4435
```

F (=59,7) tend vers l'infini
P=3,041e-10 < 0,01
Donc **on rejette H0**
Interaction significative

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.40374 -0.36945 -0.05548  0.43078  1.27715

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.86925    0.53500  -1.625 0.116754
GpeB         5.66919    1.71156   3.312 0.002818 **
GpeC         5.51945    1.75581   3.144 0.004266 **
LT           0.04037    0.01006   4.015 0.000477 ***
GpeB:LT     -0.06418    0.02431  -2.641 0.014059 *
GpeC:LT     -0.04309    0.02314  -1.862 0.074420 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6659 on 25 degrees of freedom
Multiple R-squared:  0.8453,    Adjusted R-squared:  0.8144
F-statistic: 27.32 on 5 and 25 DF,  p-value: 2.247e-09
```

= ordonnée à l'origine / pente de GpeA

= Ecart de B / C avec l'ordonnée a l'origine de A

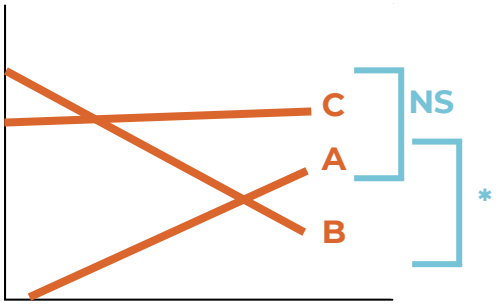
= Ecart de B / C avec la pente de A

R-squared = 0,81
81 % de Racines s'explique par les 2 facteurs et leur interaction

$\widehat{Racines}$ = -0,87 + 0,04 **LT** A

$\widehat{Racines}$ = (5,66-0,87) + (-0,064+0,04) **LT** B

$\widehat{Racines}$ = (5,52-0,87) + (-0,043-0,04) **LT** C



Modèle linéaire Généralisé

MODÈLE LINÉAIRE GENERALISÉ

Analyse pour toutes etudes où la distribution ne suit pas une loi normale : De type binomial

1/ Representation graphique

Plot(*tab*$*x* , *tab*$*y*)

Type :

- normal
- inverse gaussian
- gamma
- poisson
- binomial

2/ Création et analyse du modèle

M<- **glm**(*y* ~*x*, **family** = *binomial*, **data** = *tab*)

anova(*M*, **test** = "*Chisq*")

summary(*M*)

anova(*M*, **test** = "*Chisq*")

```
Model: binomial, link: logit

Response: Presence

Terms added sequentially (first to last)
```

Modèle sans l'effet de la distance => Horizontale

```
         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        75    104.039
Distance  1   28.748        74     75.291 8.243e-08 ***
```

Modèle avec l'effet de la distance

p>0,01	p<0,01
Acceptation de H0	Rejet de H0

= Valeur du Chi2 = ecrat entre les 2 résidus

P=8,243e-08 < 0,01
Donc **on rejette H0**
=> Effet de la distance (X) sur la presence (Y)

summary(*M*)

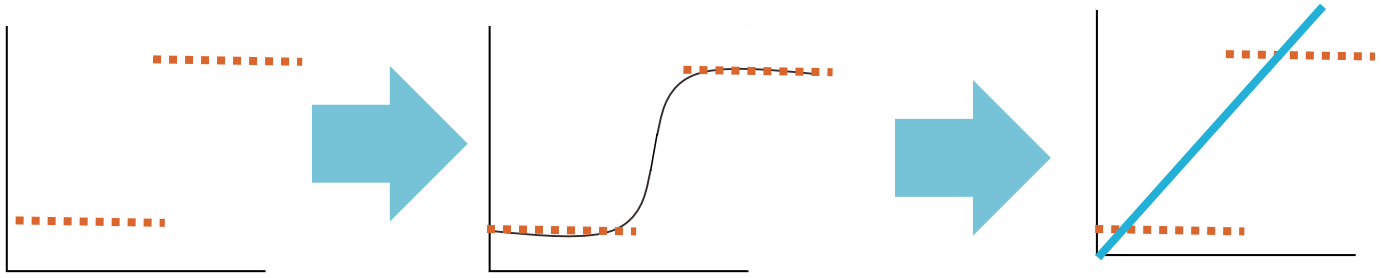
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.732062   0.652313   4.188 2.81e-05 ***
Distance    -0.037577   0.009583  -3.921 8.81e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 104.039  on 75  degrees of freedom
Residual deviance:  75.291  on 74  degrees of freedom
AIC: 79.291

Number of Fisher Scoring iterations: 5
```

Effet positif ou négatif ?? => Summary

= Ordonnée à l'origine

= pente

$\widehat{\textbf{logit(présence)}} = 2{,}73 + -0{,}038\ \textbf{LT}$



GRAPHIQUE GLM

1/ Création d'un nouveau tableau

Valeurs max et min

tab2 <- **data.frame**(**X** = **seq** (**from** = *1*, **to** = *201*, **by** = *1*))

Même nom que dans le fichier d'origine

Pas de 1

2/ Ajout d'une colonne de prediction sur le modèle dans tab2

pred <- **predict**(*M* , **newdata** = *tab2*, **type** = "response")

Résultat de la sygmoïde de logis(Y)

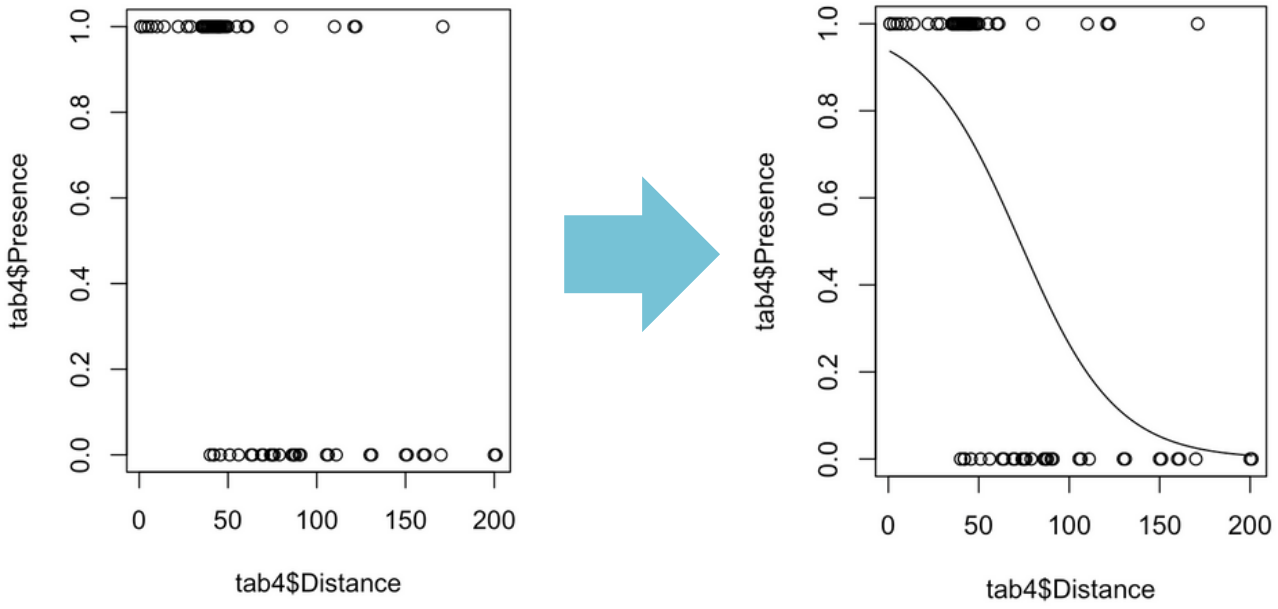
3/ Graphique

par(**new** = *TRUE*)

Permet d'ajouter un nouveau plot au dessus d'un deja existant

lines(*tab2*$*X*, *pred*)

X Y



Analyse des résidus

NORMALITÉ DES RESIDUS

Analyse de la distribustion des résidus en comparaison avec une loi Normal

1/ Extraction des residus du modèle

R <- **resid**(*M*)

2/ Representation graphique

hist (*R*)

3/ Test de normalité des résidus

Xth <- **rnorm** (1000, mean (*R*), *sd*(*R*))

ks.test (*R* , *Xth*)

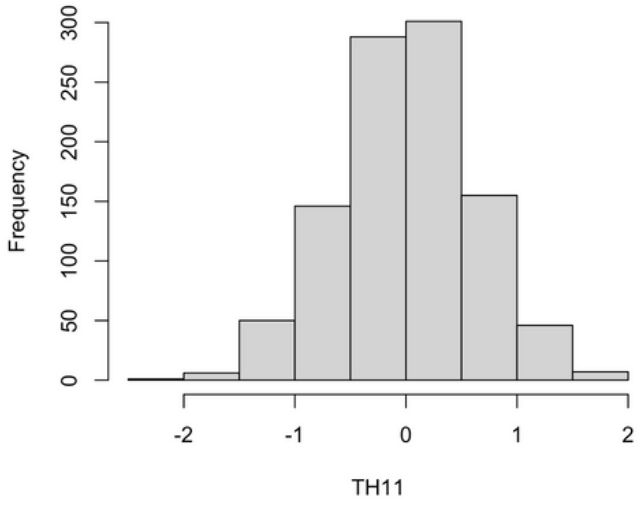
Si D tend vers 0 - p>0,01	Si D tend vers 1 - p<0,01
Acceptation de H0	Rejet de H0

H0 : Les résidus suivent une loi normale

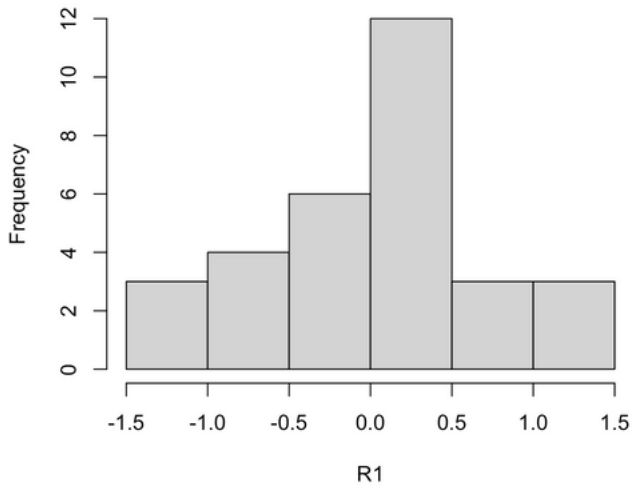
H1 : Les résidus ne suivent pas de loi normale

```
        Asymptotic two-sample Kolmogorov-Smirnov test

data:  R1 and TH11
D = 0.095387, p-value = 0.9473
alternative hypothesis: two-sided
```

P=0,94 > 0,001 => On ne peut pas rejeter H0
Il n'y a pas de difference significative entre la loi normale et la distribution des residus



Residus Theorique



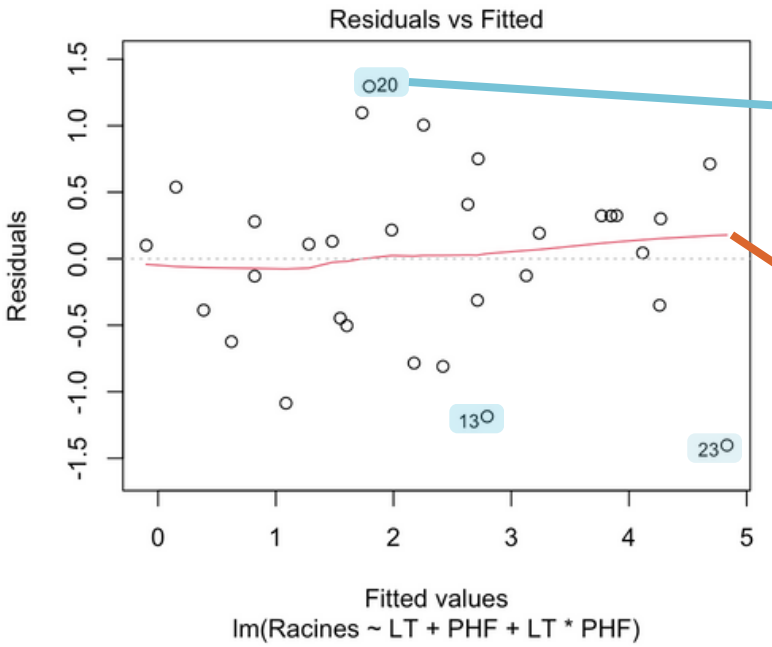
Residus

HOMOGÉNÉITÉ DES RESIDUS

Analyse de l'homogénéité des variances résiduelles

Trouver la representation graphique qui nous intéresse

plot (*M11*, **which**= c (*1*)) → **Donne un diagnostic du modèle**



Residus qui sortent de l'homogénéité des variances résiduelles du modèle

Moyenne mobile = Barycentre

Plus elle est proche de 0 plus le modèle est correct

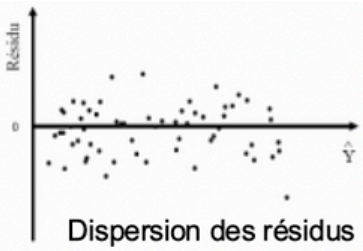
Plus il y a de point non conforme plus le modèle est mauvais

Heterogénéité peut etre du a :

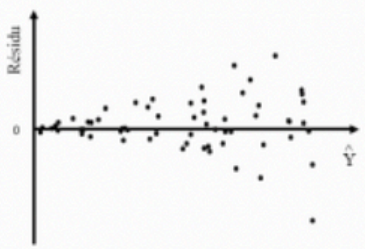
- Point non indépendant
- obsevateur different
- dependance des points
- spatial
- temporel

Dependance

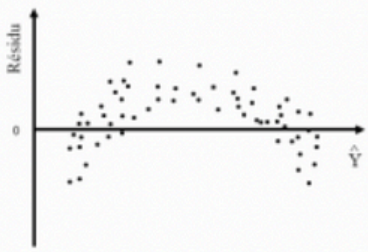
- L'homoscédasticité (homogénéité des variances)



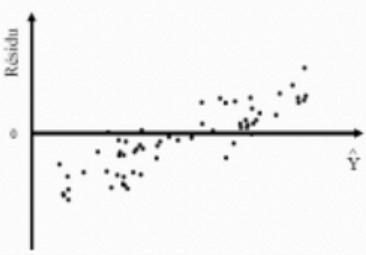
Dispersion des résidus autour de 0 semble constante



La variance des résidus augmente avec Ŷ



Surestimation des faibles et fortes valeurs de Y et sous estimation des valeurs moyennes (régression linéaire simple pas suffisante)



Surestimation des faibles valeurs de Y et sous estimation les valeurs fortes valeurs de Y (régression linéaire simple pas suffisante)