Analyse de la covariance (ANCOVA)

Frédéric Bertrand

2016-2017

Sources

Ce cours s'appuie sur :

- le livre de Kutner, Nachtsheim, Neter, Li **Applied Linear Statistical Models**, Fifth Edition, aux éditions McGraw-Hill Irwin, 2004

Généralités

Introduction

L'analyse de la covariance (ANCOVA) est une technique qui combine certaines des caractéristiques de l'analyse de la variance et de la régression linéaire. Elle peut servir aussi bien pour des études planifiées (plan de type II) ou non (plan de type I).

L'idée à la base de l'analyse de la covariance est d'ajouter à un modèle d'analyse de la variance, associé à une ou plusieurs variables qualitatives, une ou plusieurs variables quantitatives qui pourraient être liées à la réponse étudiée.

Introduction (suite)

En réalisant cet ajout, nous cherchons à réduire la variance du terme d'erreur ε présent dans le modèle et rendre ainsi l'analyse plus précise.

D'un point de vue mathématique, les modèles d'analyse de la covariance sont en fait simplement un type particulier de modèle de régression linéaire.

Réduction de la variance résiduelle

Incitation au voyage

Considérons une étude sur l'effet de trois films incitant au voyage dans un même pays étranger P. Son déroulement est le suivant :

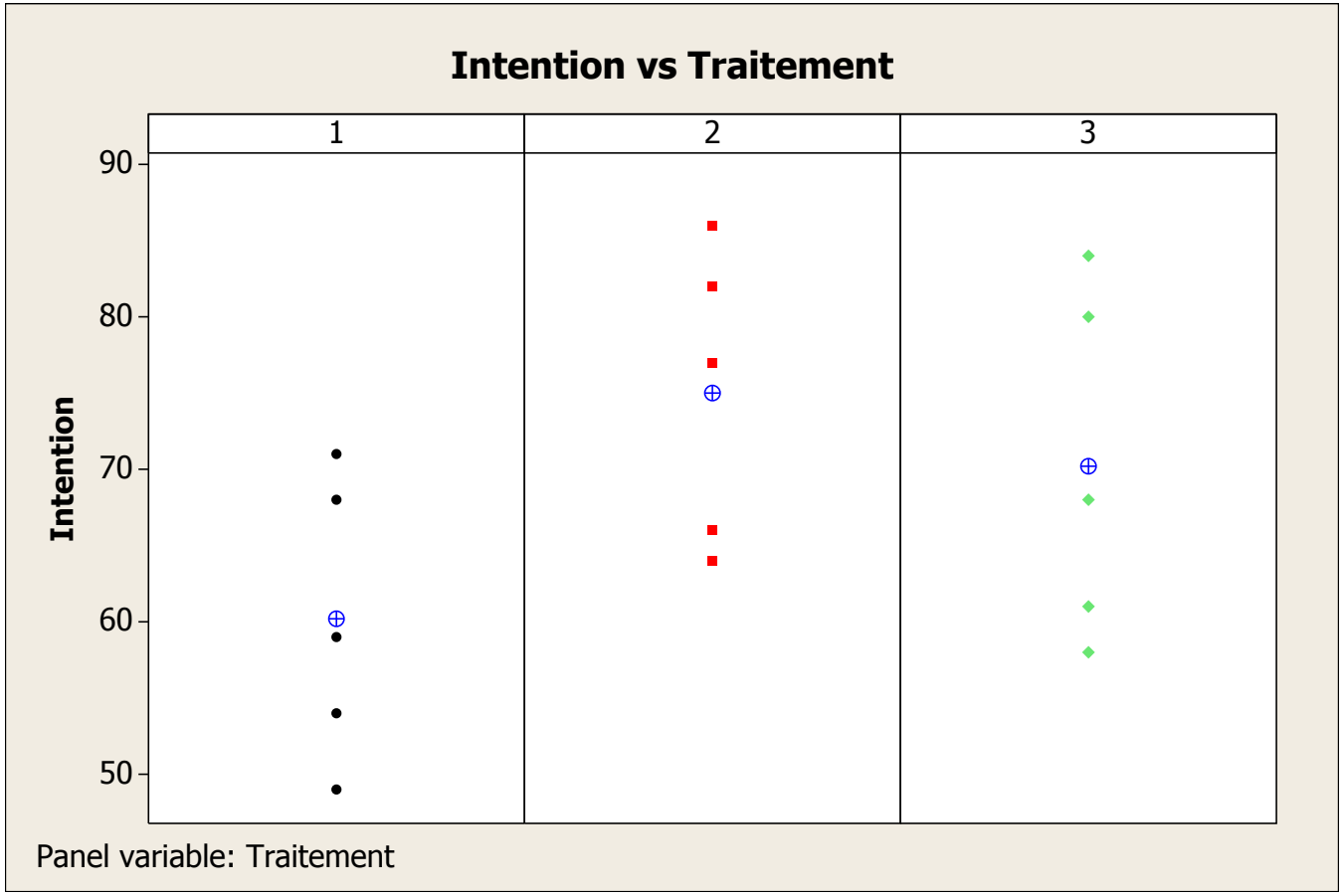
1. chaque sujet a reçoit un questionnaire *avant la projection* pour évaluer la manière dont il perçoit le pays P et établir un score de sympathie du sujet pour ce pays ;
2. chaque sujet voit l'un des trois films de cinq minutes ;
3. chaque sujet reçoit à nouveau un questionnaire sur le contenu du film et son désir de voyager dans le pays P.

Incitation au voyage (suite)

Dans ce type de situation, il est possible de se servir de l'analyse de la covariance. Afin d'avoir une idée intuitive de l'intérêt vraisemblablement majeur qu'il y aurait à la faire, nous représentons sur les deux figures suivantes les scores d'intention de voyage récoltés après que chacun des trois films promotionnels ait été présenté à un groupe de cinq sujets. Les symboles différents servent à distinguer chacun des traitements (films) utilisés.

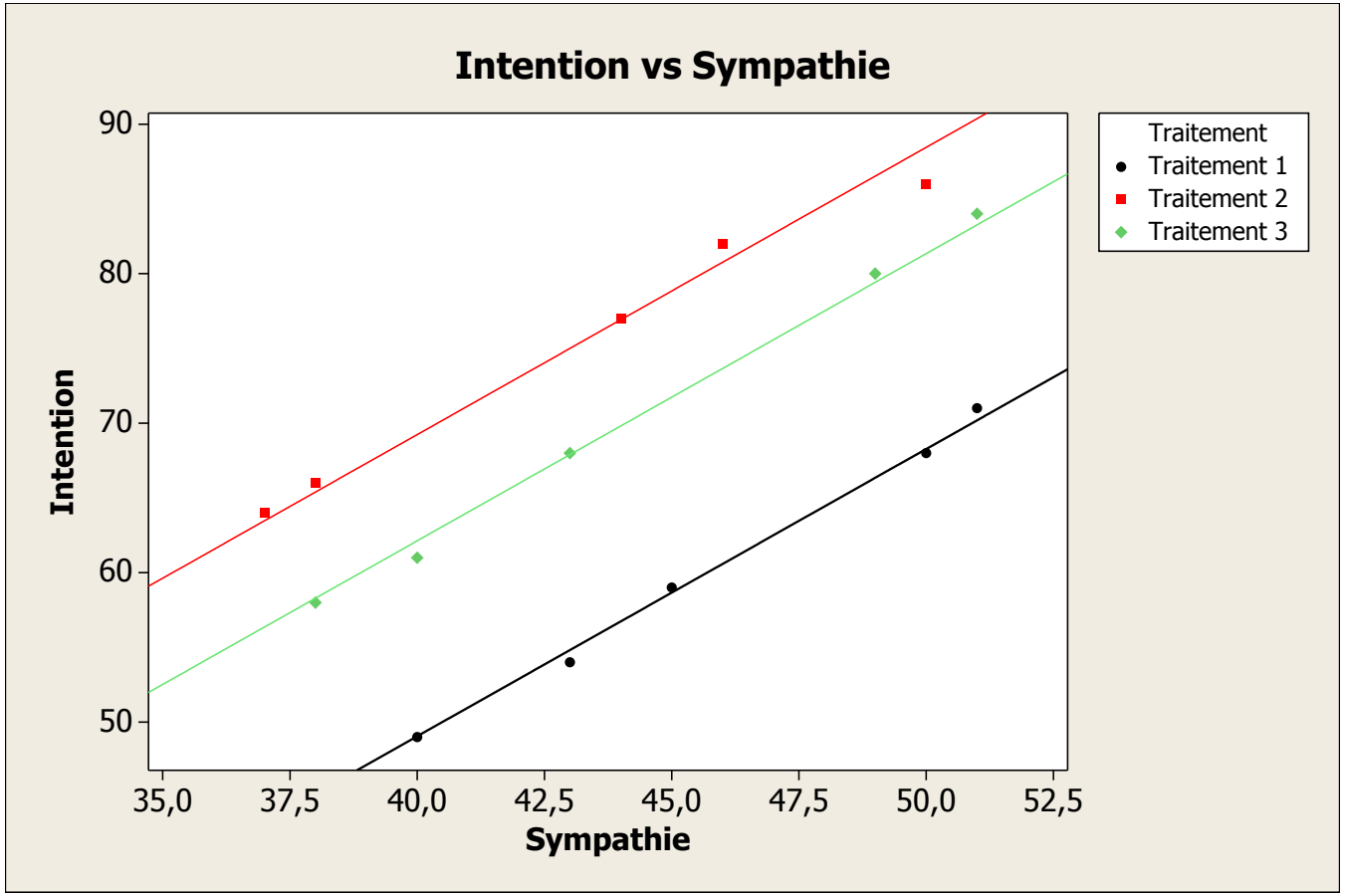
Incitation au voyage (suite)

Sur la première d'entre elles, nous constatons que les variances des termes d'erreur au sein de chaque groupe, qui s'estiment visuellement par la dispersion des points autour des estimations des moyennes $\overline{Y_{i\bullet\bullet}}$ de chacun des groupes, sont importantes ce qui est caractéristique d'une variance élevée pour le terme d'erreur du modèle d'analyse de la variance à un facteur associé.


Intention vs Traitement
1
2
3
90
80
70
60
50
Intention
Panel variable: Traitement

Incitation au voyage (suite)

Utilisons désormais l'information auxiliaire que nous avons récoltée : les valeurs initiales du score de sympathie pour le pays P avant la projection. La figure suivante représente les scores d'intention de voyage récoltés après la projection en fonction des quinze scores de sympathie initiaux, les symboles différents servant encore à distinguer chacun des traitements (films) utilisés.


Intention vs Sympathie
Intention
90
80
70
60
50
35,0
37,5
40,0
42,5
45,0
47,5
50,0
52,5
Sympathie
Traitement
Traitement 1
Traitement 2
Traitement 3

Incitation au voyage (suite)

- Bien que cela ne soit pas nécessaire pour utiliser l'analyse de la covariance, il s'avère que dans cet exemple la relation entre les scores d'intention de voyage récoltés après la projection et les quinze scores de sympathie initiaux semble être approchée de manière satisfaisante par une relation linéaire.
- La dispersion des points autour des droites de régression est bien moins élevée que celle observée sur la première figure, ce qui est caractéristique d'une variance plus faible pour le terme d'erreur du modèle.

Conclusion

L'analyse de la covariance se sert de la relation entre la réponse (scores d'intention de voyage récoltés après la projection dans notre exemple) et une ou plusieurs variables quantitatives pour lesquelles nous disposons d'observations (étude préprojection dans notre exemple) afin de diminuer la variance du terme d'erreur du modèle et rendre l'analyse portant sur la comparaison des moyennes des traitements plus puissante.

Covariables

Définition

Nous appelons covariable, en anglais concomitant variable, toute variable quantitative qui est ajoutée à un modèle d'ANOVA.

Choix des covariables

Le choix des covariables est un processus très important. S'il s'avère que les variables retenues n'ont aucun lien avec la réponse étudiée, le gain du modèle d'ANCOVA par rapport à celui du modèle d'ANOVA sera inexistant et nous retiendrons vraisemblablement au final ce modèle plus simple.

Choix des covariables

- Les covariables fréquemment utilisées lors d'études portant sur des êtres humains sont l'âge, la catégorie socio-professionnelle, l'aptitude, des données recueillies lors d'études antérieures, ...
- Les covariables fréquemment utilisées lors d'études portant sur des magasins sont le nombre d'employé, le volume des ventes pendant la dernière période précedant l'étude, ...

Recueil des observations des covariables

Afin de pouvoir interpréter sans ambigüité les résultats obtenus, les covariables doivent être observées soit avant l'étude, soit pendant l'étude à condition les traitements appliqués aux sujets lors de celle-ci ne puisse en aucune manière modifiée l'observation de celles-ci.

Une préétude d'opinion satisfait à cette condition. Dans la plupart des situations, l'observation de l'âge du sujet pendant l'étude également.

Recueil des observations des covariables

Une entreprise réalise un stage intensif pour des ingénieurs afin de leur apprendre des compétences en gestion. Deux méthodes d'enseignement ont été utilisée et les ingénieurs ont été affectés à l'une ou l'autre au hasard. À la fin de la formation, chaque stagiaire a été évalué à l'aide d'un score quantifiant ce qu'il a retenu de celle-ci. Lors du dépouillement, la personne chargée de l'analyse a décidé d'utiliser le temps de travail personnel, que chaque ingénieur devait noter, comme covariable et n'a pas pu mettre en évidence d'effet de la méthode d'apprentissage.

Recueil des observations des covariables

Ce résultat était surprenant et une étude de la répartition des temps de travail a montré que ceux-ci dépendaient de la méthode utilisée. En effet, l'une des deux méthodes était basé sur du e-learning, ce qui a particulièrement intéressée les stagiaires de ce groupe et les a incités à travailler plus. Ainsi, aussi bien la réponse que la covariable étaient lié au traitement. Ce phénomène, associé à la forte corrélation entre temps de travail personnel et score d'apprentissage, a masqué l'effet traitement.

Recueil des observations des covariables

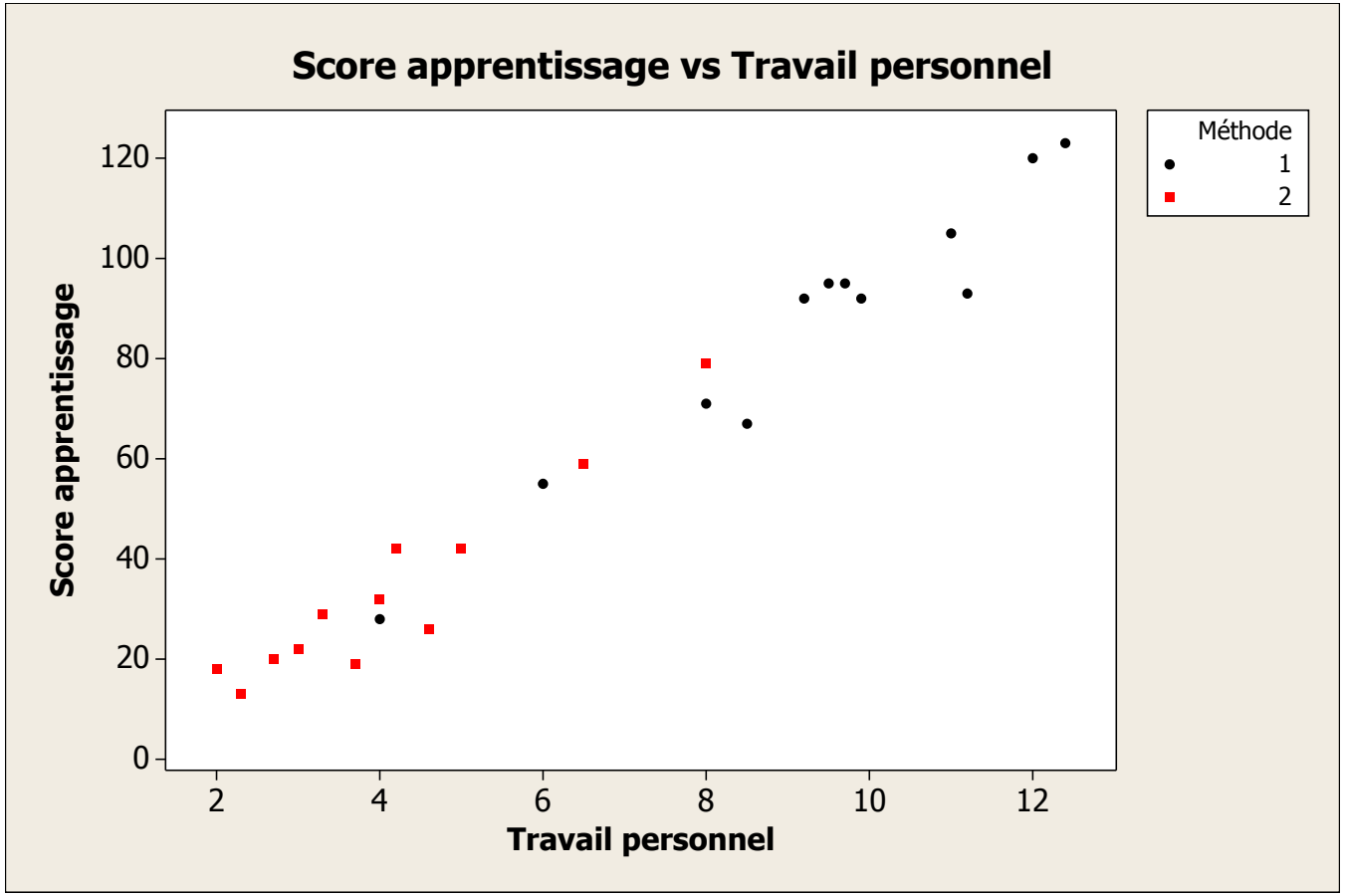
Lorsqu'une covariable dépend des traitements étudiés, le modèle d'analyse de la covariance risque de ne pas pouvoir évaluer correctement les effets de certains (ou la plupart) des traitements sur la réponse et fournir ainsi des résultats trompeurs.

Remarque

Une figure peut être utile pour déterminer si une covariable dépend des traitements ou non.

Recueil des observations des covariables

Considérons la figure suivante qui représente le score d'apprentissage en fonction du temps de travail personnel pour l'exemple précédent. Le traitement 1 correspond au e-learning.


Score apprentissage vs Travail personnel
Score apprentissage
Travail personnel
0
20
40
60
80
100
120
2
4
6
8
10
12
Méthode
1
2

Recueil des observations des covariables

Nous constatons que la plupart des personnes pour lesquels le travail personnel a été élevé sont associées au traitement 1. Réciproquement, la plupart des ingénieurs du groupe traitement 2 se sont peu investis dans la formation.

Par conséquent, les observations pour chacun des traitements ont tendance à se concentrer dans des intervalles disjoints de l'axe des abscisses.

Il est intéressant de contraster cette situation avec celle qui a été présentée dans le premier exemple où un tel regroupement n'apparaissait pas.

Analyse de la covariance à un facteur

Un exemple intégralement traité

Introduction

L'exemple suivant porte sur l'**évaluation de la capacité d'une plante à repousser** et à **produire des graines** lorsqu'elle a été broutée. Nous disposons des informations suivantes :

- La **réponse** observée est le **poids de graines produites**.
- Une information indirecte sur la **taille initiale** de la plante (le **diamètre** supérieur de son **porte greffe**) avant l'intervention éventuelle d'un animal a été relevée.
- Nous savons si la **plante** a été **broutée ou non**.

Introduction (suite)

Comme les plantes plus grandes produisent vraisemblablement plus de graines que des plantes plus petites, la **prise en compte de la taille de la plante** est nécessaire pour analyser ces données.

Réglage des contrastes et récupération des données

```
options(contrasts=c("contr.sum","contr.poly"))
compensation<-read.csv(file.choose(),header=T)
attach(compensation)
names(compensation)
```

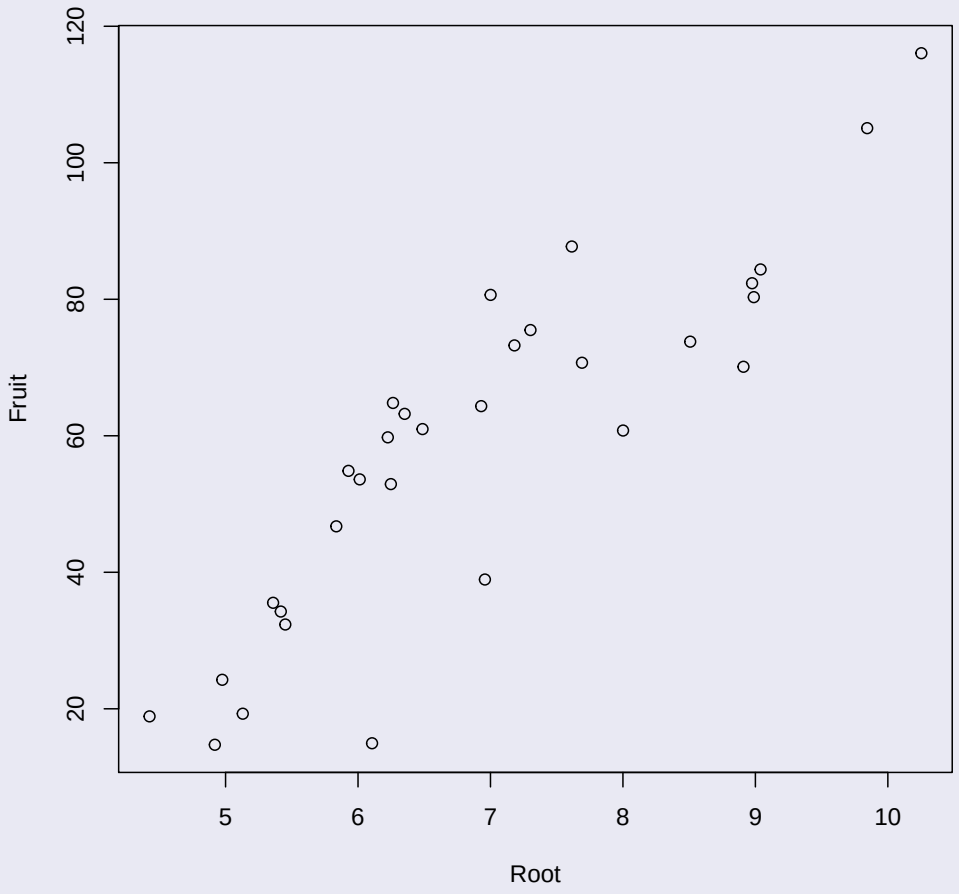
```
## [1] "Root"    "Fruit"    "Grazing"
```

Aperçu des données

```
head(compensation)

##    Root Fruit  Grazing
## 1 6.225 59.77 Ungrazed
## 2 6.487 60.98 Ungrazed
## 3 4.919 14.73 Ungrazed
## 4 5.130 19.28 Ungrazed
## 5 5.417 34.25 Ungrazed
## 6 5.359 35.53 Ungrazed
```

Représentation graphique des données

```
plot(Root,Fruit)
```

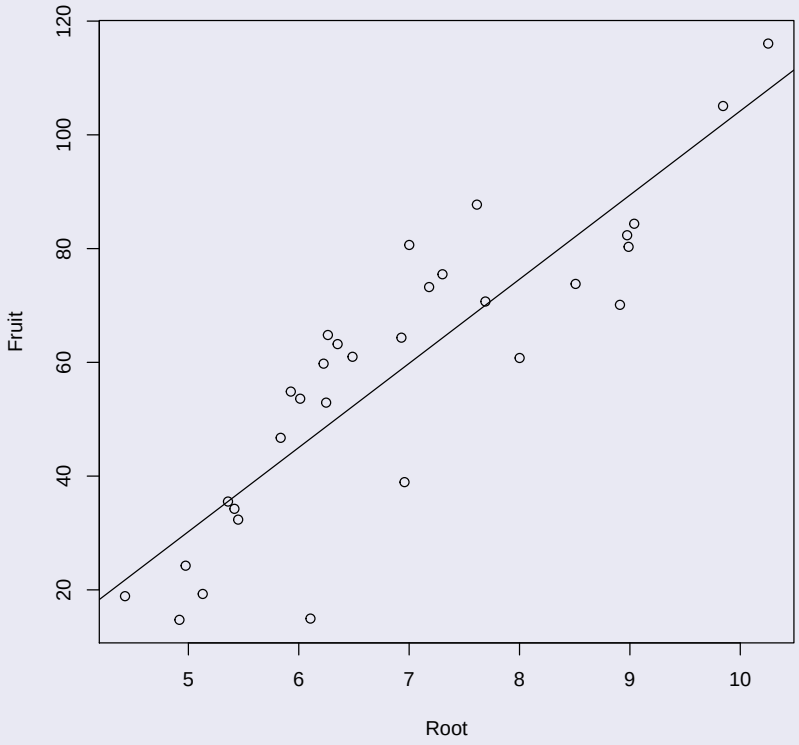


Ajustement linéaire

```
lm(Fruit~Root)

##
## Call:
## lm(formula = Fruit ~ Root)
##
## Coefficients:
## (Intercept)          Root
##     -43.74         14.79
```

Représentation graphique avec ajustement linéaire

```
plot(Root,Fruit)
abline(lm(Fruit~Root))
```



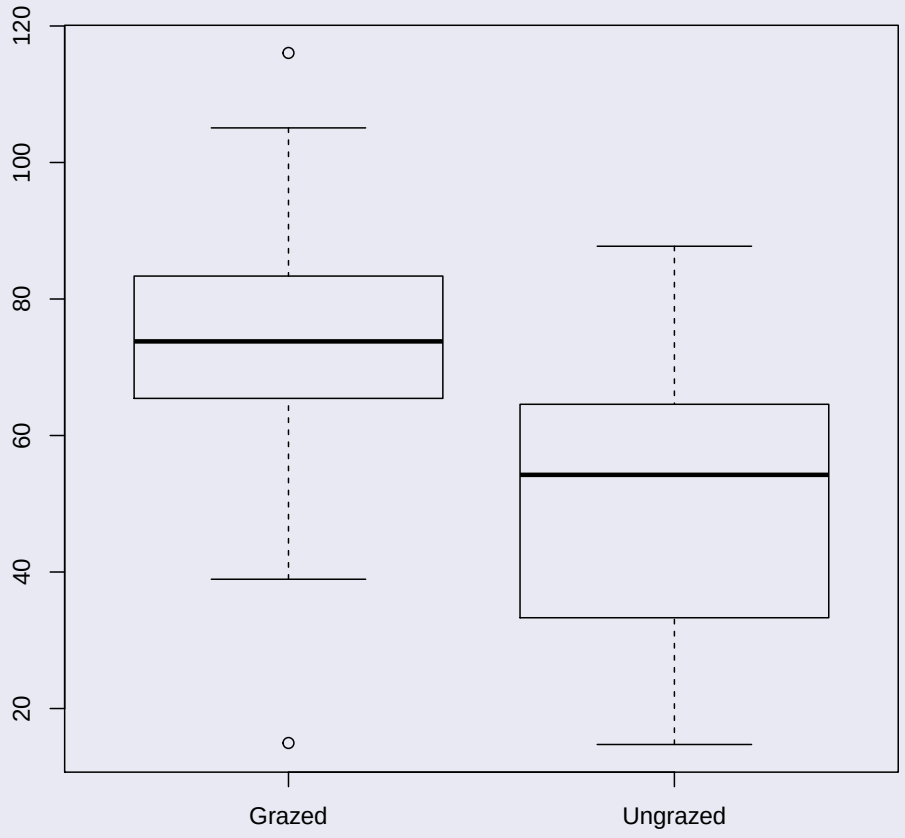
Moyennes par groupe

```
tapply(Fruit,Grazing, mean)

##   Grazed Ungrazed
## 72.49182 50.88050
```

Surprenant, car les plantes qui ont été broutées produisent plus de graines. Confirmé par le graphique suivant mais est-ce vraiment le cas ?

Moyennes par groupe

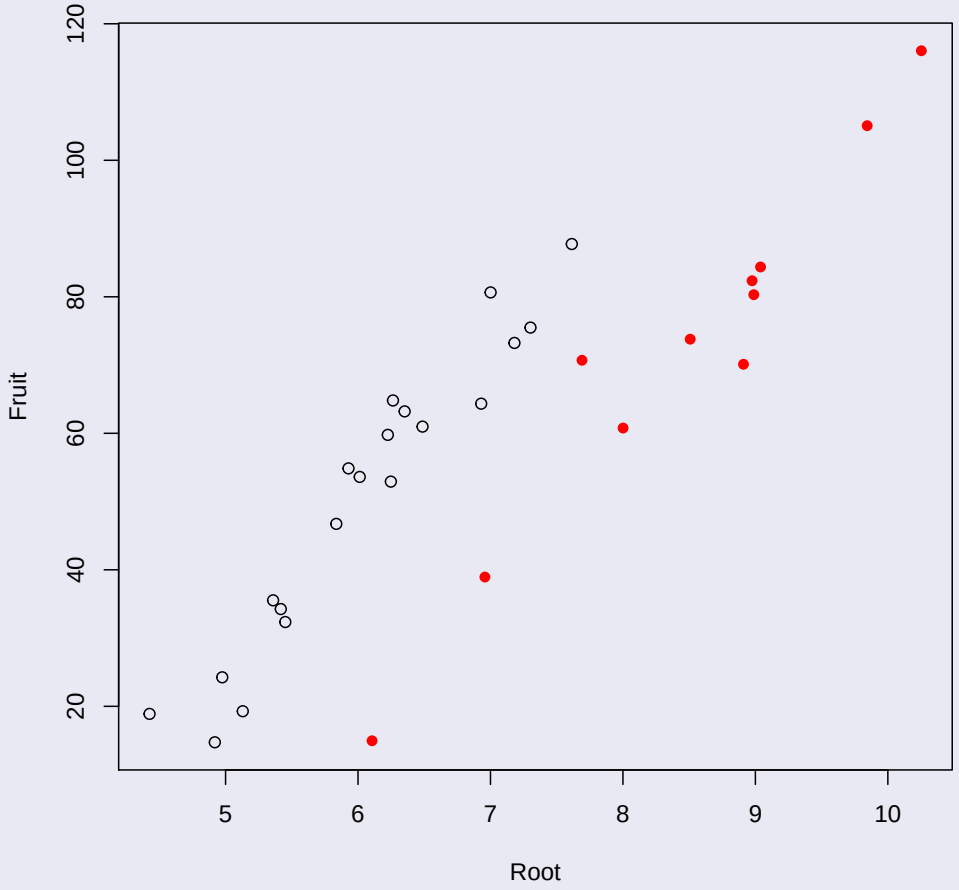
```
plot(Grazing,Fruit)
```



Représentation graphique avec groupe d'appartenance

```
plot(Root,Fruit,type="n")
points(Root[Grazing=="Ungrazed"],
       Fruit[Grazing=="Ungrazed"])
points(Root[Grazing=="Grazed"],
       Fruit[Grazing=="Grazed"],pch=16,
       col="red")
```

Représentation graphique avec groupe d'appartenance



Modèle d'ANCOVA

```
ancova<-lm(Fruit~Root*Grazing)
summary(ancova)


##
## Call:
## lm(formula = Fruit ~ Root * Grazing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0028  -3.8163  -0.6113   3.8098  15.8213
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -105.6583     8.5308 -12.386  1.2e-12 ***
## Root            23.1689     1.1485  20.173  < 2e-16 ***
## Grazing1       -11.2909     8.5308  -1.324    0.197
## Root:Grazing1   -0.8275     1.1485  -0.720    0.477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.227 on 27 degrees of freedom
## Multiple R-squared:  0.9481,Adjusted R-squared:  0.9423
## F-statistic: 164.3 on 3 and 27 DF,  p-value: < 2.2e-16
```

Tableau d'ANOVA de l'ANCOVA

```
anova(ancova)
```

```
## Analysis of Variance Table
##
## Response: Fruit
##              Df  Sum Sq Mean Sq  F value    Pr(>F)
## Root          1 15578.8 15578.8 401.7856 < 2.2e-16 ***
## Grazing       1  3507.1  3507.1  90.4494 4.121e-10 ***
## Root:Grazing  1    20.1    20.1   0.5191    0.4774
## Residuals    27  1046.9    38.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.aov(ancova)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Root          1  15579   15579 401.786  < 2e-16 ***
## Grazing       1   3507    3507  90.449 4.12e-10 ***
## Root:Grazing  1     20      20   0.519    0.477
## Residuals    27   1047      39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Une pente par groupe est-elle nécessaire ?

```
ancova2<-lm(Fruit~Grazing+Root)
anova(ancova2,ancova)

## Analysis of Variance Table
##
## Model 1: Fruit ~ Grazing + Root
## Model 2: Fruit ~ Root * Grazing
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 1067.0
## 2     27 1046.9  1    20.128 0.5191 0.4774
```

Le test de Fisher partiel n'est pas significatif au seuil de $\alpha = 5\%$.

Est-il possible de simplifer encore plus le modèle ?

```
ancova3<-lm(Fruit~Root)
anova(ancova3,ancova2)

## Analysis of Variance Table
##
## Model 1: Fruit ~ Root
## Model 2: Fruit ~ Grazing + Root
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     29 4574.1
## 2     28 1067.0  1    3507.1 92.03 2.388e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```

Le fait de savoir si la plante a été broutée ou non est une information significative au seuil de $\alpha = 5\%$.

Est-il possible de simplifer encore plus le modèle ?

```
ancova3<-lm(Fruit~Grazing)
anova(ancova3,ancova2)

## Analysis of Variance Table
##
## Model 1: Fruit ~ Grazing
## Model 2: Fruit ~ Grazing + Root
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     29 16838
## 2     28  1067  1     15771 413.86 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```

Le fait de connaître le diamètre du porte-greffe est une information significative au seuil de $\alpha = 5\%$.

Résumé du modèle d'ANCOVA

```
summary.lm(ancova2)

##
## Call:
## lm(formula = Fruit ~ Grazing + Root)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3565  -3.1694   0.0446   3.0649  16.4688
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -106.617      8.354 -12.763 3.43e-13 ***
## Grazing1     -17.296      1.803  -9.593 2.39e-10 ***
## Root          23.163      1.139  20.344  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.173 on 28 degrees of freedom
## Multiple R-squared:  0.9471,Adjusted R-squared:  0.9433
## F-statistic: 250.4 on 2 and 28 DF,  p-value: < 2.2e-16
```

Tableau d'ANOVA du modèle d'ANCOVA

```
anova(ancova2)

## Analysis of Variance Table
##
## Response: Fruit
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Grazing    1  3314.5  3314.5  86.977 4.387e-10 ***
## Root       1 15771.4 15771.4 413.859 < 2.2e-16 ***
## Residuals 28  1067.0    38.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```

Choix de modèle automatique

```
step(ancova)


## Start:  AIC=117.11
## Fruit ~ Root * Grazing
##
##                Df Sum of Sq    RSS    AIC
## - Root:Grazing  1    20.128 1067.0 115.70
## <none>                      1046.9 117.11
##
## Step:  AIC=115.7
## Fruit ~ Root + Grazing
##
##           Df Sum of Sq     RSS    AIC
## <none>                  1067.0 115.70
## - Grazing  1    3507.1  4574.1 158.82
## - Root     1   15771.4 16838.4 199.22
##
## Call:
## lm(formula = Fruit ~ Root + Grazing)
##
## Coefficients:
## (Intercept)         Root     Grazing1
##     -106.62        23.16       -17.30
```

Modèle retenu automatiquement

```
summary(ancova2)


##
## Call:
## lm(formula = Fruit ~ Grazing + Root)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3565  -3.1694   0.0446   3.0649  16.4688
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -106.617      8.354 -12.763 3.43e-13 ***
## Grazing1     -17.296      1.803  -9.593 2.39e-10 ***
## Root          23.163      1.139  20.344  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.173 on 28 degrees of freedom
## Multiple R-squared:  0.9471,Adjusted R-squared:  0.9433
## F-statistic: 250.4 on 2 and 28 DF,  p-value: < 2.2e-16
```

Coefficients du modèle

Voici les estimations des coefficients du modèle d'ANCOVA à droites parallèles.

```
coef(ancova2)

## (Intercept)    Grazing1        Root
##  -106.61665   -17.29600    23.16264
```

Coefficients du modèle

Ordonnée à l'origine pour la droite ajustée au groupe des plantes non broutées

```
coef(ancova2)[1]-coef(ancova2)[2]

## (Intercept)
##   -89.32066
```

Ordonnée à l'origine pour la droite ajustée au groupe des plantes broutées

```
coef(ancova2)[1]+coef(ancova2)[2]

## (Intercept)
##   -123.9126
```

Coefficients du modèle

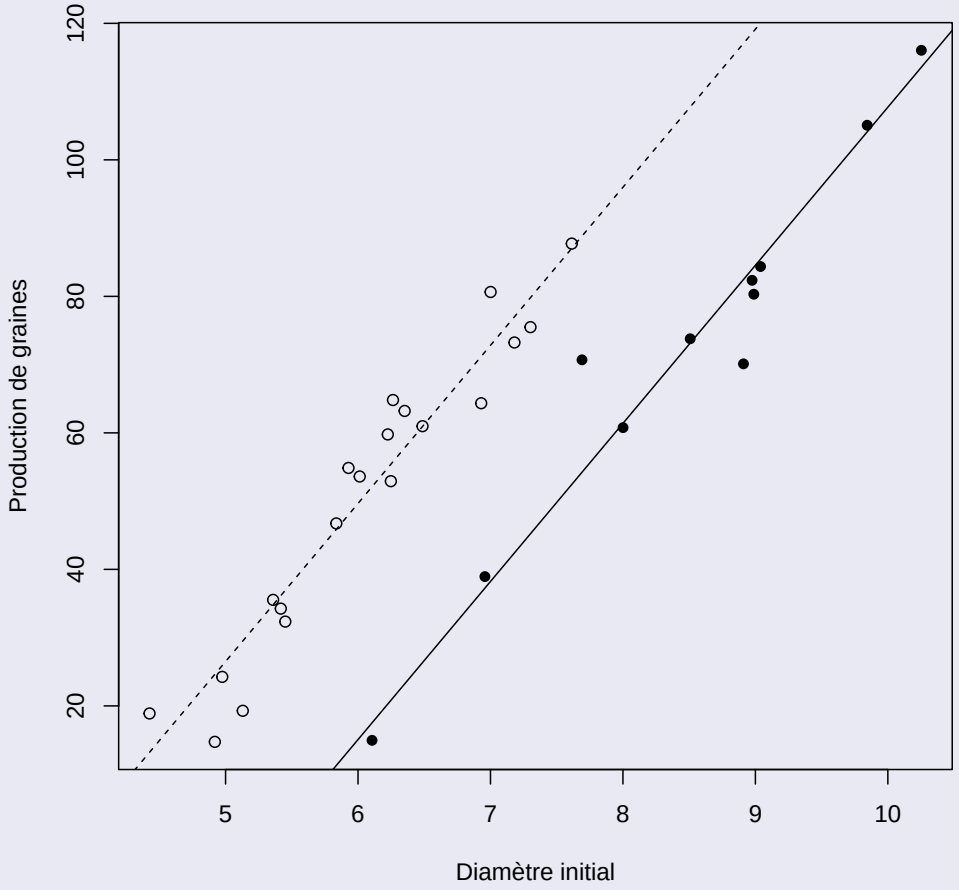
Pente commune aux deux droites (ajustées aux deux groupes)

```
coef(ancova2)[3]

##      Root
## 23.16264
```

Graphique du modèle retenu

```
sf<-split(Fruit,Grazing)
sr<-split(Root,Grazing)
plot(Root,Fruit,type="n",
     ylab="Production de graines",
     xlab="Diamètre initial")
points(sr[[1]],sf[[1]],pch=16)
points(sr[[2]],sf[[2]])

abline(coef(ancova2)[1]+coef(ancova2)[2],
       coef(ancova2)[3])
abline(coef(ancova2)[1]-coef(ancova2)[2],
       coef(ancova2)[3],lty=2)
```

Graphiques


Production de graines
20
40
60
80
100
120
5
6
7
8
9
10
Diamètre initial

Attention

Que se serait-il passé si nous n'avions pas pris en compte la taille de la plante ?

```
tapply(Fruit,Grazing,mean)

##   Grazed Ungrazed
## 72.49182 50.88050
```

Attention

Que se serait-il passé si nous n'avions pas pris en compte la taille de la plante ?

```
summary(aov(Fruit~Grazing))

##            Df Sum Sq Mean Sq F value Pr(>F)
## Grazing     1   3315    3315   5.708 0.0236 *
## Residuals  29  16838     581
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

Attention

Que se serait-il passé si nous n'avions pas pris en compte la taille de la plante ?

```
coef(aov(Fruit~Grazing))

## (Intercept)     Grazing1
##    61.68616     10.80566
```

Production moyenne du groupe non-brouté

Attention

Production moyenne du groupe non-brouté

```
coef(aov(Fruit~Grazing))[1]-
  coef(aov(Fruit~Grazing))[2]

## (Intercept)
##     50.8805
```

Production moyenne du groupe brouté

```
coef(aov(Fruit~Grazing))[1]+
  coef(aov(Fruit~Grazing))[2]

## (Intercept)
##    72.49182
```

Valeur ajustées des moyennes des groupes

Calculons maintenant les moyennes des deux groupes qui ont été ajustées par le modèle d'ANCOVA pour tenir compte de la différence des distributions des tailles entre les deux groupes.

Production moyenne du groupe non-brouté

```
coef(ancova2)[1]-coef(ancova2)[2]+
  coef(ancova2)[3]*mean(Root)

## (Intercept)
##    70.82361
```

Valeur ajustées des moyennes des groupes

Production moyenne du groupe brouté

```
coef(ancova2)[1]+coef(ancova2)[2]+
  coef(ancova2)[3]*mean(Root)

## (Intercept)
##    36.23162
```

En utilisant les moyennes ajustées pour tenir compte de la variable taille, les conclusions s'inversent !

Conclusion

Intérêt du modèle d'ANCOVA

Prendre en compte la taille initiale du porte greffe change complètement le résultat et corrige les biais liés à l'inhomogénéité des groupes.