Analyse de la covariance : ANCOVA

Support de cours (3)

Année Universitaire 2022-2023

R. Abdesselam

Courriel : rafik.abdesselam@univ-lyon2.fr
Web : http://perso.univ-lyon2.fr/~rabdesselam/Fr/
Support pédagogique : http://perso.univ-lyon2.fr/~rabdesselam/Documents/

M2 – SISE

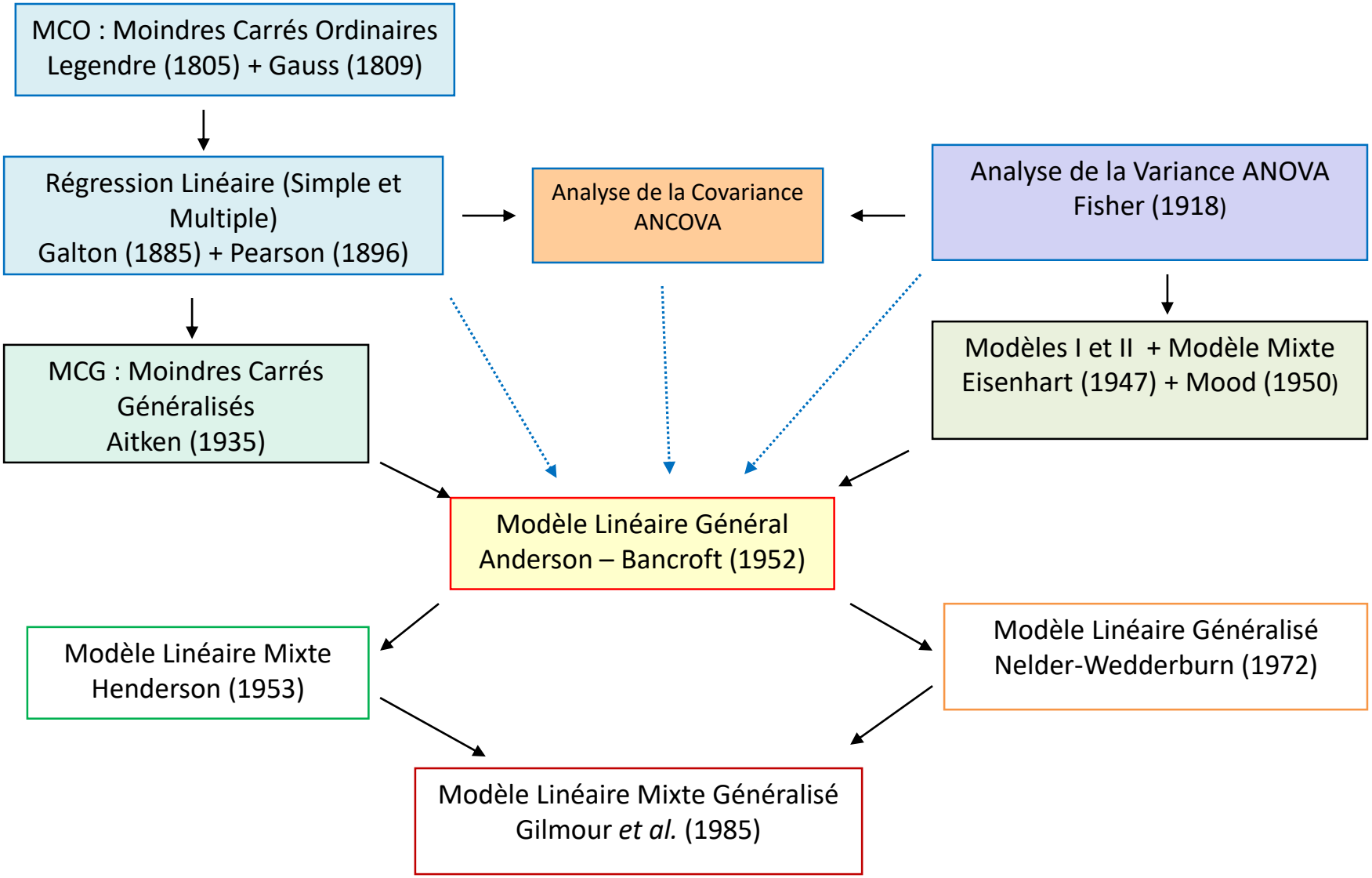


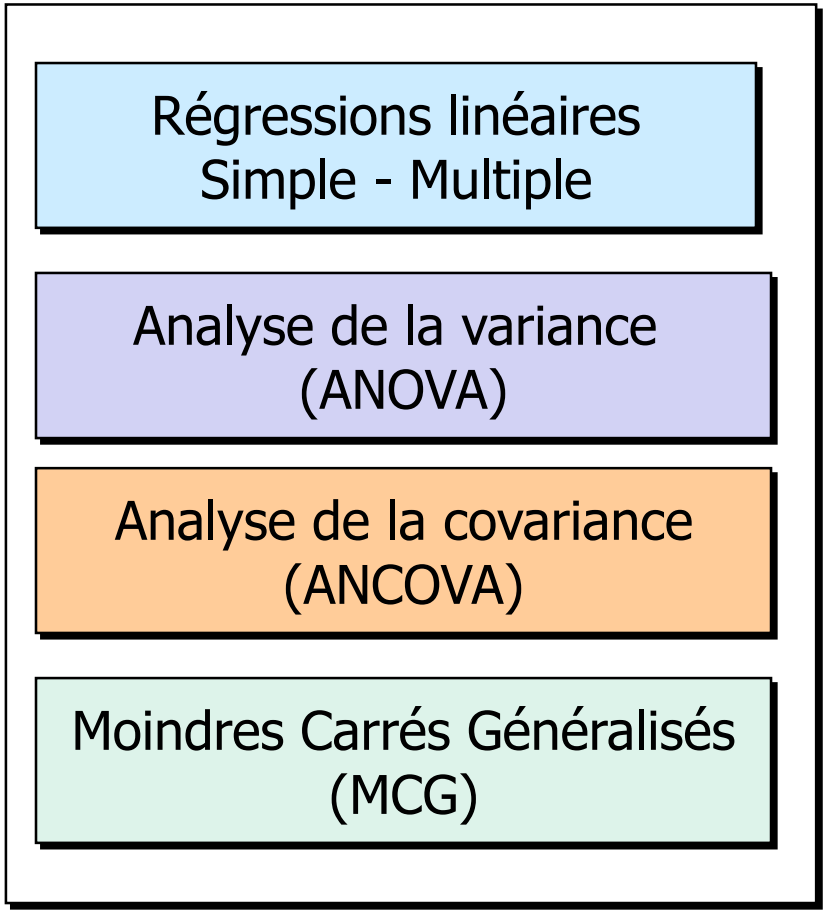
Schéma récapitulatif (Source P. Dagnélie - modifiée)

Introduction

- Le MLG est un modèle de la forme suivante:

$$Y = b\,X + e$$

- *Y* est un vecteur variable à expliquer,
- *b* est le vecteur des coefficients estimés,
- *X* est une matrice de vecteurs variables explicatives,
- *e* représente le terme d'erreur.



Les principaux modèles MLG

Modèle	*Variable à expliquer (endogène)*	*Variable(s) explicative(s) (exogène(s))*
Régression simple	**1 continue**	**1 continue**
ANOVA à un critère	**1 continue**	**1 nominale***
ANOVA à critères multiples	**1 continue**	**2 ou plus nominales***
ANCOVA	**1 continue**	**Au moins 1 nominale* et au moins une 1 continue**
Régression multiple	**1 continue**	**2 ou plus continues**

* Discrète ou discontinue

Analyse de la variance : ANOVA (suite)

- ANOVA est une régression multiple sur variables explicatives nominales (facteurs).
- Dans une ANOVA, la variance totale est répartie en deux sources de variations :
  - ◇ Inter-groupes : variance des moyennes des différents groupes (niveaux du facteur)
  - ◇ Intra-groupe (erreur) : variance des observations autour de la moyenne du groupe.

Les différents types d'ANOVA

- ANOVA Type I (effets fixes) : *les traitements sont fixés ou contrôlés par l'expérimentateur ou le chercheur,*
- ANOVA Type II (effets aléatoires) : *les traitements ne sont pas sous le contrôle de l'expérimentateur ou du chercheur,*
- ANOVA Type III (modèle mixte) : *on est en présence d'au moins un facteur de type I et d'au moins un facteur du type II.*

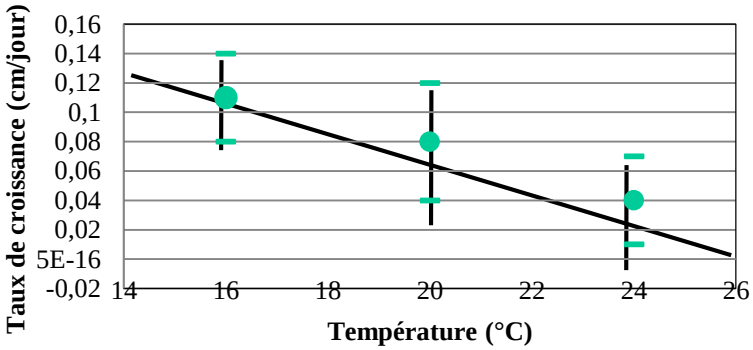
Les différents types d'ANOVA à un facteur

ANOVA Type I - Effets fixes

" *les traitements sont déterminés ou contrôlés par l'expérimentateur* "

Exemple - Pisciculture : effet de la température sur le taux de croissance du poisson.

- ***A*** *est le facteur contrôlé : 3 niveaux de température (°C) déterminés (fixés) par l'utilisateur – chercheur,*
- ***Y*** *est le taux de croissance (cm/jour) , la variable continue à expliquer,*
- *on peut estimer l'effet de l'augmentation d'une unité de A (température) sur Y (taux de croissance)*
- *… on peut alors prédire Y pour d'autres températures .*



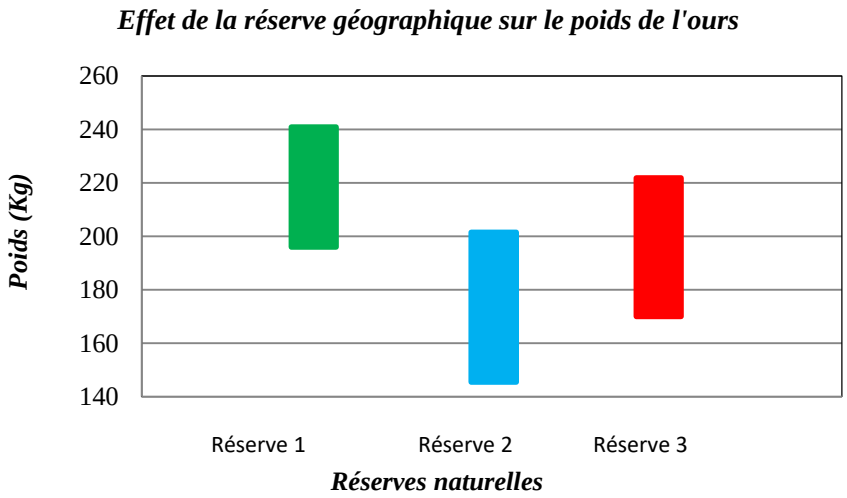
Les différents types d'ANOVA à un facteur

ANOVA Type II - Effets aléatoires

"les traitements ne sont pas sous le contrôle de l'expérimentateur"

Exemple : poids de l'ours et la dispersion géographique.

- ***A*** *est le facteur non contrôlé : 3 niveaux (aléatoires - réserves géographiques ou groupes) échantillonnés par l'utilisateur – chercheur,*
- *Y* est le poids de l'ours (kg) , la variable continue à expliquer,
- *Pour des réserves différentes, les facteurs contrôlant la variabilité sont inconnus,*
- *… alors, on ne peut pas prédire Y le poids de l'ours pour d'autres réserves.*



Les différences entre les modèles ANOVA

- Pour le type I, les facteurs peuvent être manipulés par l'utilisateur, pas dans le type II,
- Le type I, permet d'estimer l'effet du facteur et faire ainsi des prédictions, mais pas le type II,
- Pour l'ANOVA à 1 facteur, les calculs Type I et Type II sont identiques.

- Sorties SAS : Type I (procédure ANOVA).
  Type I et Type III (procédure GLM).

  Sorties SPAD : Type III.

ANOVA à plusieurs facteurs

- *A utiliser lorsque plusieurs facteurs indépendants peuvent agir,*
- *Contrairement à l'ANOVA à 1 facteur, il faut proposer plusieurs hypothèses nulles H_0,*
- *Elle évite de recourir à plusieurs ANOVA à 1 facteur pour tester la même chose,*
- *En plus, elle permet de tester les interactions entre facteurs.*

Exemple : La croissance d'une céréale en fonction de la quantité d'Engrais (Q1, Q2, Q3) déversée et du volume d'irrigation (V1, V2, V3).

- *On peut effectuer 3 ANOVA à 1 facteur (Quantité d'engrais) pour chacun des volumes d'eau testé. Il faut 3 expériences pour répondre à la même question.*
- *La probabilité d'accepter H_0 pour toutes les expériences est de $(0.95)^3$ = 86%, donc rejeter au moins une fois H_0 qui est vraie avec une probabilité de 14%,*
- *… En plus, les éventuelles interactions entre engrais et irrigation ne sont pas testées.*

Les différents types d'ANOVA à plusieurs facteurs

ANOVA Type I - Effets fixes

" les traitements sont déterminés et sous contrôle de l'expérimentateur "

Exemple : croissance en taille d'un poisson en fonction de la température et du pH de l'eau.

- *Les deux facteurs sont les variables explicatives **A** : température et **B** : pH . Les deux facteurs sont fixés par l'utilisateur,*
- ***Y** est le taux de croissance (cm/jour) , la variable continue à expliquer (dépendante),*
- *on peut estimer l'effet de l'augmentation d'une unité de **A** (température) sur **Y** (taux de croissance)*
- *… comme les facteurs sont contrôlés, on peut estimer l'effet de l'accroissement d'une unité de température ou de pH sur le taux de croissance et le prédire pour d'autres poissons.*

Les différents types d'ANOVA à plusieurs facteurs

ANOVA Type II - Effets aléatoires

"les traitements ne sont pas sous le contrôle de l'expérimentateur"

Exemple : la taille d'un ours en fonction de la région et de l'altitude.

- *Les deux facteurs sont les variables nominales explicatives **A** : région et **B** : l'altitude. Les deux facteurs (aléatoires) ne sont pas fixés par l'utilisateur,*
- *Y est la taille, variable continue à expliquer (dépendante),*
- *… comme les facteurs ne sont pas contrôlés, même si la taille diffère en fonction de la région ou de l'altitude, on ne peut pas savoir quel facteur est responsable de cette variabilité, on ne peut donc pas prédire la taille pour une autre région ou une autre altitude.*

Les différents types d'ANOVA à plusieurs facteurs

ANOVA Type III - Modèle mixte

"Au moins un facteur de type I et au moins un facteur du type II"

Exemple : la taille d'un ours en fonction de la région et du sexe.

- *Les deux facteurs sont les variables explicatives **A** : région (variable aléatoire) et **B** : sexe (variable fixée),*
- ***Y** est la taille, variable continue à expliquer (dépendante),*
- *… même si la taille diffère en fonction de la région ou du sexe, on ne peut pas savoir quel facteur est responsable de cette variabilité, on ne peut donc pas prédire la taille des ours de chaque sexe pour une autre région. Par contre, on peut éventuellement prédire la différence entre les sexes.*

Les différents facteurs pour l'ANOVA

	Facteur fixe	*Facteur aléatoire*
Manipulation par l'expérimentateur ?	**OUI**	**NON**
Estimation de l'effet des niveaux du facteur	**OUI**	**NON**
Prédiction ?	**OUI**	**NON**
Calculs de l'ANOVA à un facteur	**Identiques**	
Calculs de l'ANOVA à plusieurs facteurs	**Différents**	

Attention, pour faire les calculs, il faut bien renseigner le modèle selon le logiciel utilisé .

Analyse de la covariance : ANCOVA

Introduction

• L'analyse de la covariance (ANCOVA) est une technique qui utilise les caractéristiques de l'ANOVA 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'ANCOVA est d'ajouter à un modèle d'ANOVA, associé à une ou plusieurs variables qualitatives (facteurs contrôlés), une ou plusieurs variables quantitatives qui pourraient être liées à l'étude. Cet ajout va chercher à réduire la variance du terme d'erreur et rendre ainsi l'analyse plus précise.

• ANCOVA est une régression multiple sur variables explicatives mixtes : nominales (facteurs) et continues (co-variables).

• Dans un modèle ANOVA, la valeur de la variable à expliquer est déterminée, à l'aléas ε près, par les classes dans lesquelles sont faites les mesures ou observations.

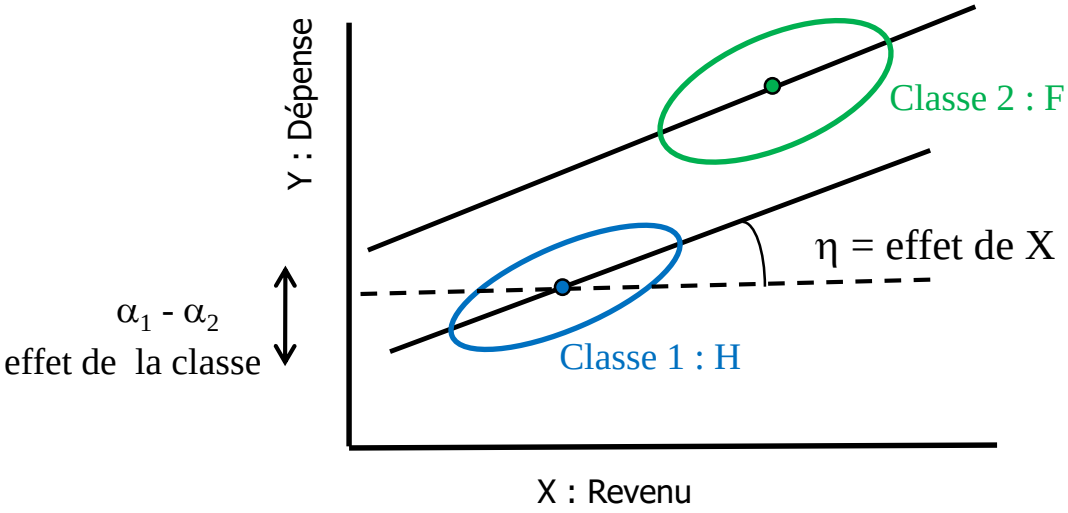
• On peut cependant imaginer un modèle où cette valeur est à l'intérieur de chaque classe k, fonction également d'une ou plusieurs variables explicatives continues.

• D'un point de vue mathématique, le modèle d'ANCOVA est un type particulier de modèle de régression linéaire.

Analyse de la covariance : ANCOVA

Exemple 1 : dépense individuelle en habillement en fonction du sexe (facteur) et pour chaque sexe fonction du revenu (continue) de l'individu.

Modèle d'ANCOVA : facteur sexe sans effet sur la pente de la régression Y : Dépense fonction de X : Revenu



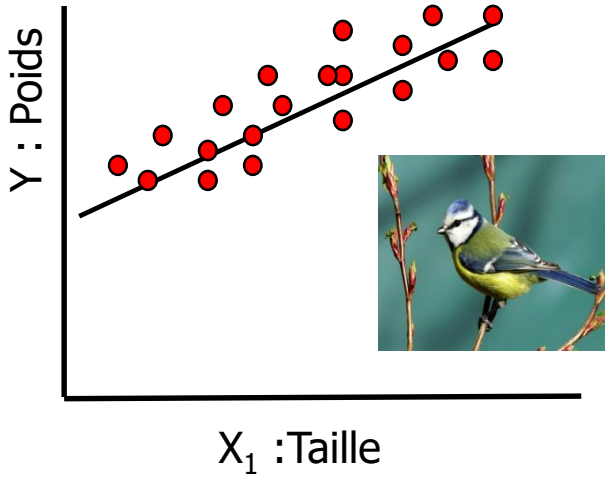
Modèle : $Y_{ik} = \mu + \alpha_k + \eta\, x_{ik} + \varepsilon_{ik}$ observation i dans la classe k

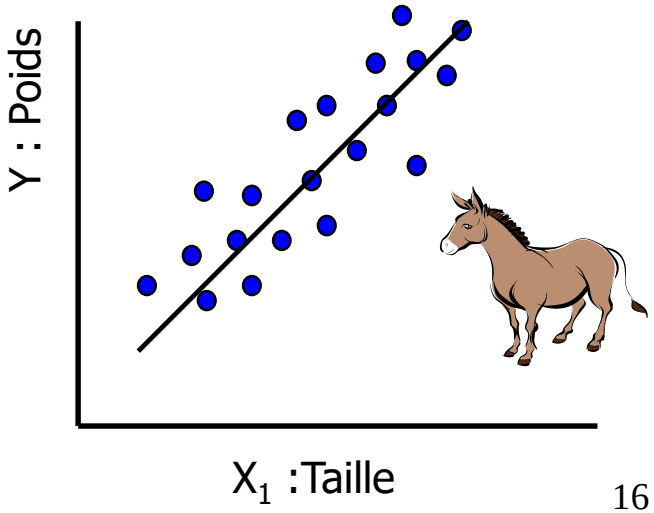
- *En donnant la même pente η aux 2 droites passant par les centres de classe, on suppose ici que le revenu a le même effet quel que soit le sexe.*
- *l'écart (α_1 - α_2) entre les 2 droites mesure l'effet du facteur sexe.*
- *On aurait pu supposer un effet du revenu différencié suivant le sexe en traçant des droites non parallèles.*

Utilité de l'ANCOVA
Régression & ANOVA

Exemple 1 : Comparaison Taille – Poids chez différents groupes de Vertébrés

- Pour une taille donnée, il est normal que le poids d'un mammifère soit plus important que celui d'un oiseau.
- Deux régressions différentes s'imposent.
- Si l'on cherche à comparer des tailles et des poids sans tenir compte du groupe taxinomique : le coefficient de détermination R^2 serait probablement très faible. Pas de corrélation et donc pas de régression linéaire !





Utilité de l'ANCOVA : Régression & ANOVA

Exemple 2 : Effets de différents régimes alimentaires sur le Poids

- Si le régime alimentaire est riche, il est normal que le poids soit plus élevé.
- Si plus de 2 régimes alimentaires sont comparés, une ANOVA à un facteur (Régime) s'impose.

Mais quelle est la condition à respecter ?

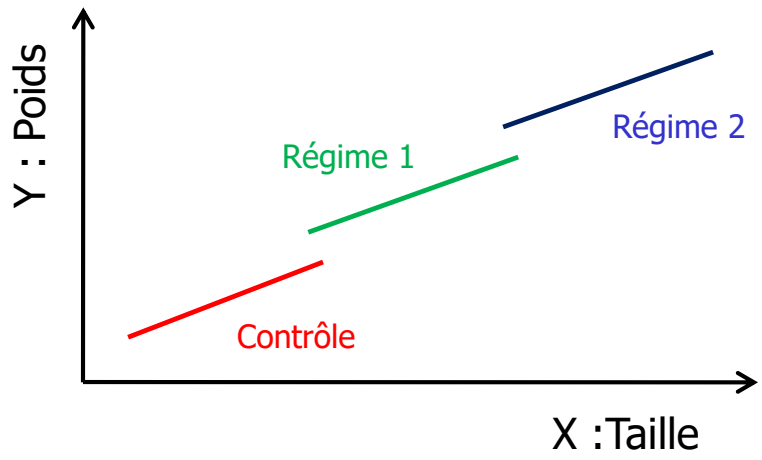
- Le poids dépend de la taille; il faut donc qu'au début de l'expérience, avant l'application du régime alimentaire, le poids soit identique. Si cette condition n'est pas respectée, l'expérience est biaisée.

- Si cette condition n'est pas vérifiée, il faut introduire dans le modèle la variabilité due à la taille : effet taille.

…. C'est une ANOVA (1 facteur : régime) avec une variable continue (taille : co-variable) pour expliquer la variable continue (poids) …. Il s'agit là d'une ANCOVA.

Utilité de l'ANCOVA : Régression & ANOVA

Exemple 2 : Comparaison du Poids en fonction de différents régimes alimentaires.



- Un modèle simple d'ANOVA mettra en évidence une différence significative entre les régimes alimentaires.
- Par contre, en visualisant le graphique, on voit que les gains en poids ne sont liés qu'aux gammes différentes de la taille.
- L'introduction dans le modèle de la variable taille (co-variable) ne mettra plus en évidence une différence significative entre les régimes alimentaires.

ANCOVA : Conditions d'application

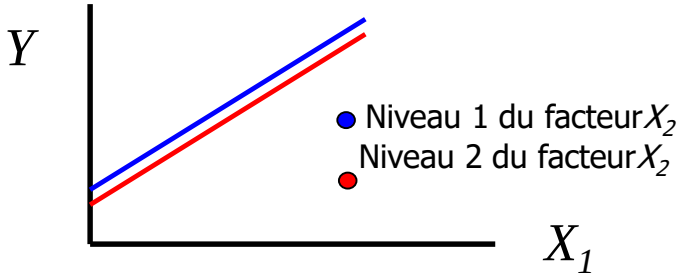
- Les résidus sont indépendants et distribués normalement
- La variance des résidus est égale pour toutes les valeurs de X et indépendantes des valeurs de la variable discontinue (homoscédasticité)
- pas d'erreur sur les variables indépendantes
- linéarité

Modèle ANCOVA : hypothèses nulles

Y : variable à expliquer, X_1 : variable explicative (continue) et X_2 : facteur (nominale) à 2 niveaux.
*Significativité des effets de X_1, X_2 et X_1*X_2 (interaction) sur Y.*

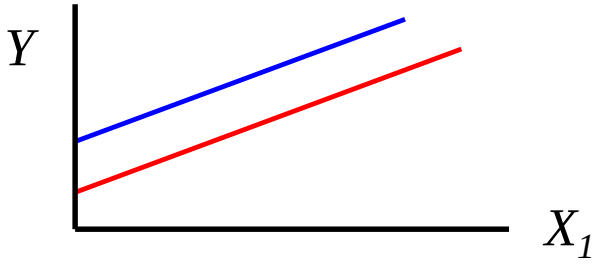
1) X_1 est significatif, X_2 et X_1*X_2 ne le sont pas.

Y change en changeant X_1 , alors X_1 a un effet significatif sur Y. Par contre, les 2 points d'intersection et les 2 pentes sont les mêmes.



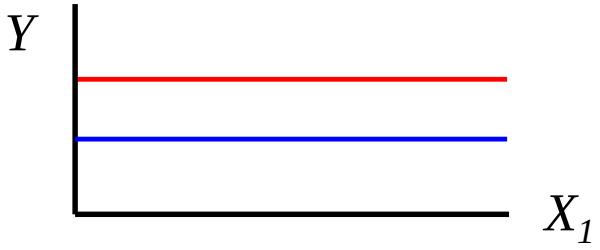
2) X_1 et X_2 sont significatifs, X_1*X_2 ne l'est pas.

*Y change en changeant X_1 , alors X_1 affecte Y. Les points d'intersection des 2 groupes sont différents, alors X_1 affecte Y également. Par contre les 2 pentes sont égales (parallèles) donc l'effet de Y sur X_1 ne varie pas en fonction de la valeur de X_2 (dépendant du groupe). Alors X_1 *X_2 n'est pas significatif.*



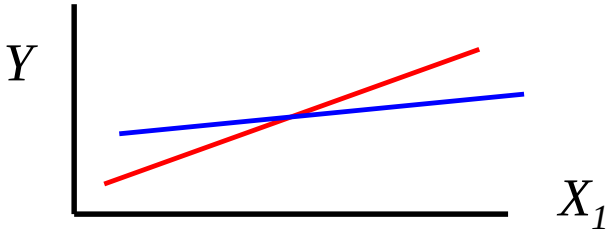
3) X_2 est significatif, X_1 et X_1*X_2 ne le sont pas.

*Y ne change pas en changeant X_1 , alors X_1 n'a pas d'effet sur Y. Les points d'intersection des 2 groupes sont différents, alors X_2 a un effet significatif sur Y. Par contre, les 2 pentes sont égales (zéro) donc Alors X_1 *X_2 n'a pas d'effet sur Y.*



4) X_1 , X_2 et X_1*X_2 sont significatifs.

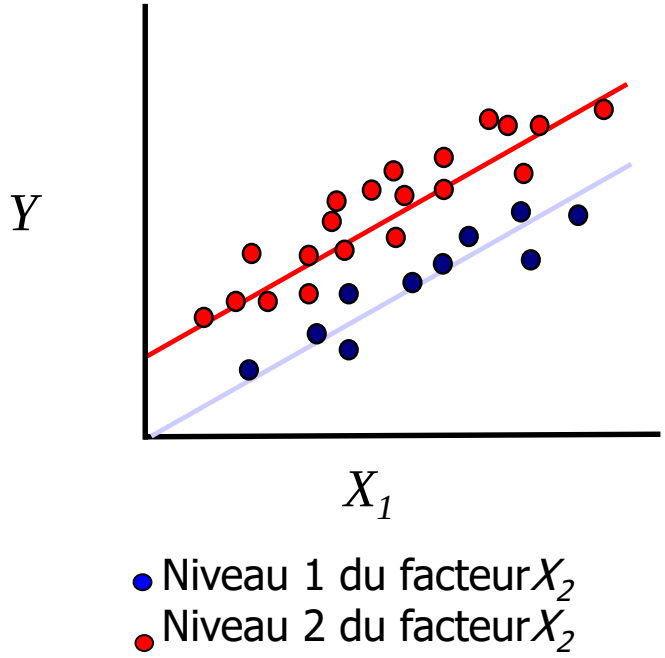
*Y change en changeant X_1 , alors X_1 affecte Y. Les points d'intersection des 2 groupes sont différents, alors X_2 affecte Y également. En plus, les 2 pentes sont différentes (non parallèles) donc l'effet de Y sur X_1 dépend de la valeur de X_2 (dépend du groupe). Alors X_1 *X_2 est significatif.*



Comment procéder ?

- Ajuster le modèle d'ANCOVA, tester pour les différences entre les pentes.

$H_0 : \alpha_i = \text{constante}$



- Si H_0 est rejetée, séparer les régressions pour chaque niveau de la variable discontinue
- Si H_0 est acceptée, ajuster une régression commune.

→ Régressions séparées

→ Régression commune

Exemple : Effets des résultats du test d'aptitude et du sexe de l'employé sur le niveau de production

- Niveau de productivité (NPRO) est la variable dépendante,

  Résultats au test d'aptitude (RTAP) est la variable indépendante continue,

  Sexe de l'employé (SEXE) est la variable qualitative (2 niveaux)
- La pente de la régression de NPRO sur RTAP est la même pour les deux sexes ?

Effets des résultats du test d'aptitude et du sexe de l'employé sur le niveau de production

SAS : Modèle ANCOVA

Number of Observations Read 27
Number of Observations Used 27

The GLM Procedure
Dependent Variable: NPRO Niveau de productivité

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	114.1647390	38.0549130	86.39	<.0001
Error	23	10.1315573	0.4405025		
Corrected Total	26	124.2962963			

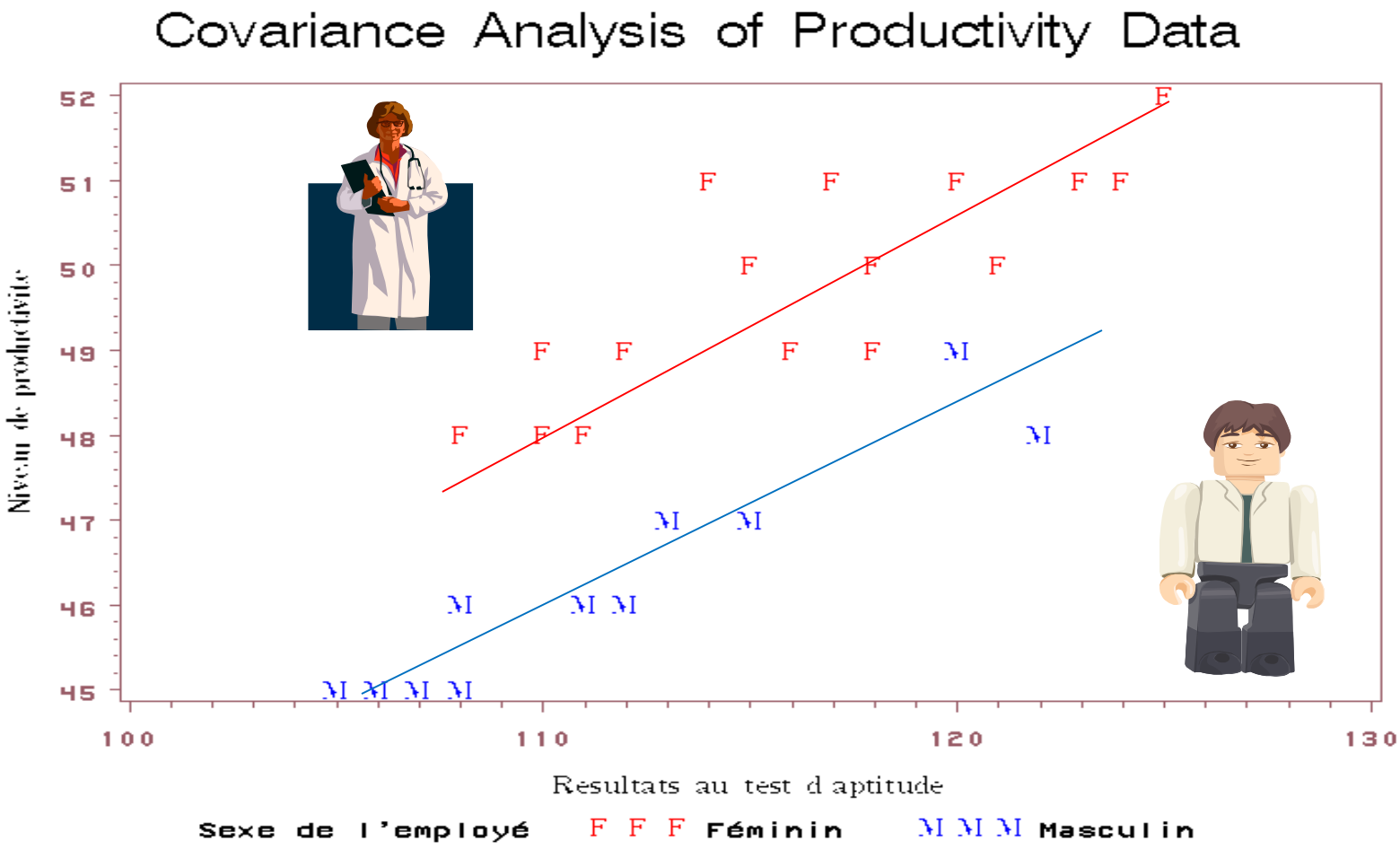
R-Square	Coeff Var	Root MSE	NPRO Mean
0.918489	1.372128	0.663704	48.37037

Source	DF	Type III SS	Mean Square	F Value	Pr > F
RTAP	1	32.30044345	32.30044345	73.33	<.0001
SEXE	1	0.50308096	0.50308096	1.14	0.2963
RTAP*SEXE	1	0.16890743	0.16890743	0.38	0.5419

Effets des résultats du test d'aptitude et du sexe de l'employé sur le niveau de production SPAD : Modèle ANCOVA

```
IDENTIFICATION OF THE ADJUSTMENT COEFFICIENTS
ENDOGENOUS VARIABLE (Y) ... Niveau de productivité
          FACTOR   3 ... Genre de l'employé
        VARIABLE   2 ... Résultat du test d'aptitude
     INTERACTION   2  3
                         Résultat du test d'aptitude
                         Genre de l'employé
ESTIMATION / COEFFICIENTS
LEAST SQUARES ADJUSTMENT (WITH CONSTANT TERM)
   27 CASES,  4 PARAMETERS (CONSTANT IN QUEUE).
  IDEN   LABEL                          COEFFICIENT  STAND.DEV.  STUDENT  PROBA.  T.VALUE
                                                 23
FACTOR(S)
  FEMI - Féminin                           3.0025       2.809    1.069   0.296    1.04
* MASC - Masculin                         -3.0025       2.809    1.069   0.296   -1.04
  TAPT - Résultat du test d'aptitude        0.2115       0.025    8.565   0.000    5.69
2nd ORDER INTERACTION(S)
  TAPT - Résultat du test d'aptitude
  FEMI - Féminin                          -0.0153       0.025    0.619   0.542   -0.61
* TAPT - Résultat du test d'aptitude
* MASC - Masculin                          0.0153       0.025    0.619   0.542    0.61
         CONSTANT                        23.9818       2.808    8.539   0.000    5.68
GLOBAL ADJUSTMENT TEST
RESIDUAL SUM OF SQUARES ..............  SCE =          10.1274
MULTIPLE CORRELATION COEFFICIENT ......  R   =           0.9584   R2 =       0.9185
ESTIMATED RESIDUAL VARIANCE ...........  S2  =           0.4403   S  =       0.6636
   FISHER  =    86.428          DEG. OF FREEDOM =  3   23
   P-VALUE =    0.0001          TEST VALUE      =    7.01
SOURCE                    SUM OF     FISHER  DEG. OF      P-VALUE  TEST
                          SQUARES            FREEDOM              VALUE
RESIDUAL                  10.127                23
2nd ORDER INTERACTION(S)
| Résultat du test d'aptitude
| Genre de l'employé
+------------------>          0.169     0.384   1  23     0.5417     -0.61
FACTOR(S)
| Genre de l'employé
+------------------>          0.503     1.143   1  23     0.2962      1.04
```

Effets des résultats du test d'aptitude et du sexe de l'employé sur le niveau de production (Modèle ANCOVA)



Effets des résultats du test d'aptitude et du sexe de l'employé sur le niveau de production

Modèle Régression multiple

The REG Procedure

Dependent Variable: NPRO Niveau de productivité

Number of Observations Used 27

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	113.99583	56.99792	132.80	<.0001
Error	24	10.30046	0.42919		
Corrected Total	26	124.29630			

Root MSE	0.65512	R-Square	0.9171	Dependent Mean	48.37037
Adj R-Sq	0.9102	Coeff Var	1.35439		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	22.93754	2.69634	8.51	<.0001
RTAP	Résultats	1	0.20920	0.02411	8.68	<.0001
CSEXE		1	2.52944	0.28177	8.98	<.0001