Analyses de Variance à un ou plusieurs facteurs

Régressions

Analyse de Covariance

Modèles Linéaires Généralisés

Professeur Patrice Francour

francour@unice.fr

Une grande partie des illustrations viennent du site Internet de l'Université d'Ottawa
(cours de Biostatistiques appliquées; © Antoine Morin et Scott Findlay)

Quand utiliser l'ANOVA

- Pour tester l'effet d'une variable indépendante "discrète"
- Chaque variable indépendante est appelée un facteur et chaque facteur peut avoir deux ou plusieurs niveaux ou traitements (ex: niveau d'irrigation; température d'élevage; région géographique, *etc*)
- Une ANOVA teste si toutes les moyennes sont égales, donc H_0: égalité et H_1: au moins une différence
- Si H_0 est rejetée pour un seuil ", l'ANOVA ne dit pas où sont les différences
- A utiliser quand le nombre de niveaux est *supérieur à deux*

Pourquoi ne pas utiliser plusieurs tests de t?

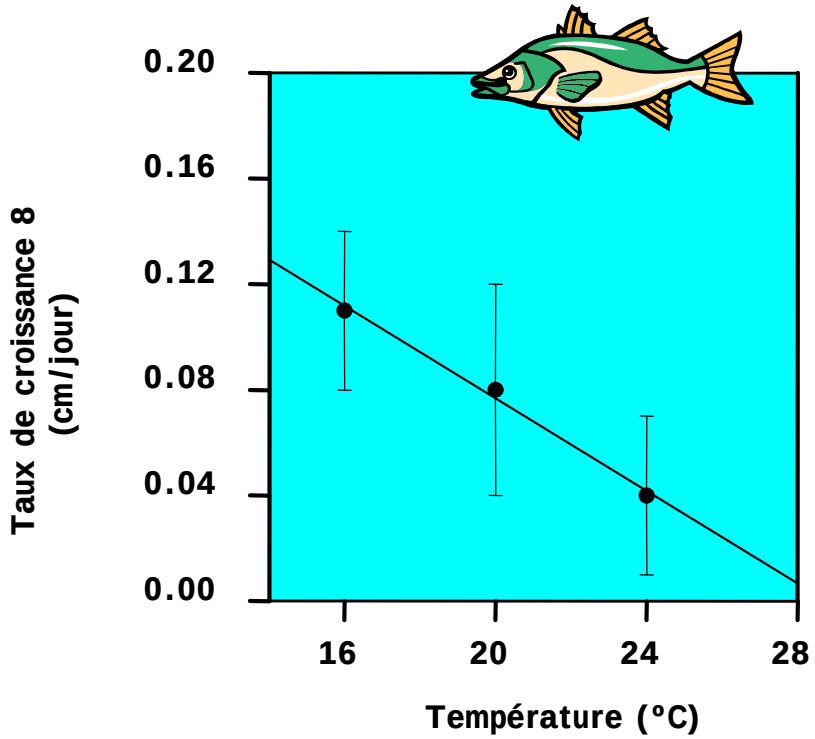
- Pour un nombre de comparaisons *k* , si H_0 est vraie, la probabilité de l'accepter pour tous les *k* est $(1 - ")^k$
- ex: pour 4 moyennes, $(1 - ")^k = (0.95)^6 = .735$; alors, " (pour toutes les comparaisons) = 0.265
- En comparant les moyennes des 4 échantillons provenant de la même population on s'attend à détecter des différences significatives pour une paire dans 27% des cas

Les différents types d'ANOVA

- **Type I ("effets fixes")** : les traitements sont déterminés par le chercheur

ANOVA Type I: effet de la température sur le taux de croissance de la truite

- 3 traitements (Température) déterminés par le chercheur
- la variable dépendante est le taux de croissance (**8**), et le facteur (**T**) est la température
- **T** étant contrôlé, on peut estimer l'effet de l'augmentation d'une unité de T (température) sur 8 (le taux de croissance)…
- …et prédire 8 pour d'autres températures

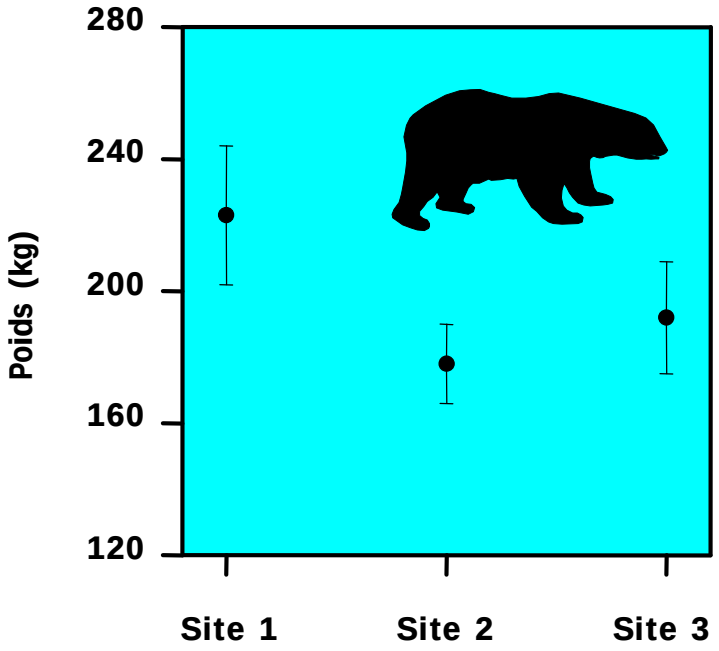


Les différents types d'ANOVA

- **Type I ("effets fixes")** : les traitements sont déterminés par le chercheur
- **Type II ("effets aléatoires")** : les traitements ne sont pas sous le contrôle de l'expérimentateur

ANOVA Type II: poids de l'ours noir et dispersion géographique

- 3 sites (groupes) échantillonnés
- La variable dépendante est le poids et le site est le facteur
- Pour des sites différents les facteurs contrôlant la variabilité sont inconnus…
- …alors, on ne peut prédire le poids pour d'autres sites



Les différents types d'ANOVA

- **Type I ("effets fixes")** : les traitements sont déterminés par le chercheur
- **Type II ("effets aléatoires")** : les traitements ne sont pas sous le contrôle de l'expérimentateur
- **Type III ("modèle mixte")** : au moins un facteur du Type I et au moins un du Type II

Différences entre les modèles

- Pour le Type I, les facteurs peuvent être manipulés par l'expérimentateur, pas dans le Type II
- Le Type I nous permet d'estimer l'effet du traitement, de faire des prédictions, pas le Type II
- Les calculs pour les deux types sont identiques mais seulement pour l'ANOVA à un critère de classification !

Pourquoi le nom ANOVA?

- Dans une ANOVA, la variance totale est répartie en deux composantes:
  - *intergroupe* : variance des moyennes des différents groupes (traitements)
  - *intragroupe (erreur)* : variance des observations autour de la moyenne du groupe

Procédure	*Variable dépendante*	*Variable(s) indépendante(s)*
ANOVA 1 facteur	**1 continue**	**1 discontinue***

* peuvent être discontinues ou traitées comme discontinues (=discrètes)

Deuxième phase de l'ANOVA

- Si la première phase de l'ANOVA (comparaison des variances inter et intragroupes) rejette H_0, alors il faut faire des comparaisons multiples de moyennes.

- Les comparaisons multiples peuvent être planifiées (*a priori*) ou non planifiées (*a posteriori*).

- Une comparaison planifiée est indépendante des résultats de l'ANOVA; la théorie prédit quels traitements devraient être différents.

La croissance d'un poisson est comparée pour différentes températures. Si la théorie prévoit qu'au-dessous de 10° la croissance devient très faible, voire nulle, les comparaisons se feront donc au-dessus et en dessous de cette valeur seuil (critique).

Deuxième phase de l'ANOVA

- Si la première phase de l'ANOVA (comparaison des variances inter et intragroupes) rejette H_0, alors il faut faire des comparaisons multiples de moyennes.

- Les comparaisons multiples peuvent être planifiées (*a priori*) ou non planifiées (*a posteriori*).

- Une comparaison planifiée est indépendante des résultats de l'ANOVA; la théorie prédit quels traitements devraient être différents.

- Une comparaison non planifiée est dépendante des résultats de l'ANOVA.

La croissance d'un poisson est comparée pour différentes températures. Si la théorie prévoit seulement que la croissance baisse quand la température baisse, les comparaisons se feront donc entre tous les échantillons.

Deuxième phase de l'ANOVA

- Si la première phase de l'ANOVA (comparaison des variances inter et intragroupes) rejette H_0, alors il faut faire des comparaisons multiples de moyennes.

- Les comparaisons multiples peuvent être planifiées (*a priori*) ou non planifiées (*a posteriori*).

- Une comparaison planifiée est indépendante des résultats de l'ANOVA; la théorie prédit quels traitements devraient être différents.

- Une comparaison non planifiée est dépendante des résultats de l'ANOVA.

Attention : l'ANOVA est plus fiable et plus robuste que les comparaisons multiples. Une CM ne doit pas être faite si H_0 (1° phase ANOVA) est acceptée ! Elle pourrait éventuellement voir des différences là où il n'y en a pas !!

ANOVA à plusieurs facteurs

- Ce qui a été dit précédemment concernait 1 seul facteur
- Si plusieurs facteurs **indépendant** peuvent agir, il faut utiliser une ANOVA à plusieurs facteurs (MANOVA)
- Contrairement à ANOVA à 1 facteur, il faut proposer plusieurs H_0
- Une ANOVA à plusieurs facteurs évite de recourir à plusieurs ANOVA à 1 facteur pour tester la même chose.
- En plus, une ANOVA à plusieurs facteurs permet de tester les interactions entre facteurs.

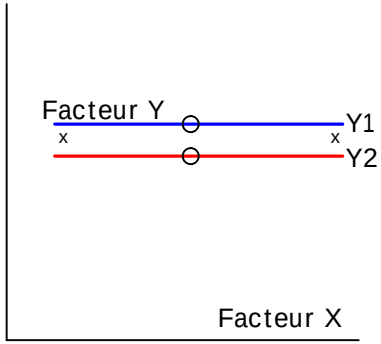
ANOVA à plusieurs facteurs

Exemple : La croissance d'une plante est comparée en fonction de la quantité d'engrais (E1, E2 et E3) fournie et du niveau d'irrigation (I1, I2 et I3).

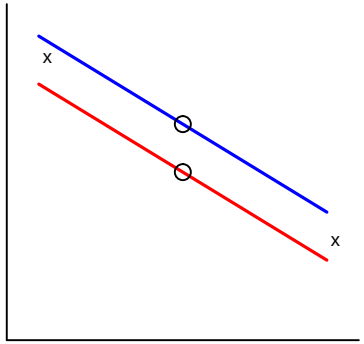
Il est possible de proposer 3 ANOVA à 1 facteur (Irrigation) pour chacune des quantités d'engrais testée. Il faut donc 3 expériences pour répondre à la même question.

La probabilité d'accepter H_0 pour toutes les expériences est de $(0.95)^3 = 0.86$. Donc la probabilité de rejeter au moins une H_0 qui est vraie est " = 0.14.

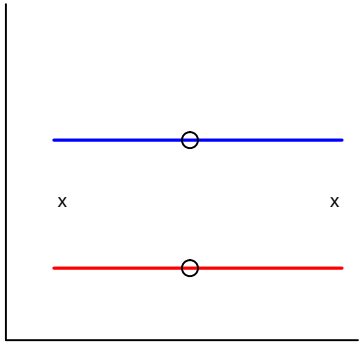
En plus les interactions, éventuelles, entre engrais et irrigation ne sont pas testées.



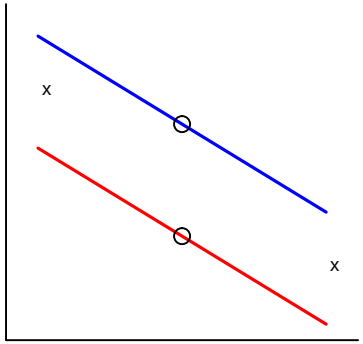
A X: pas d'effet; Y: faible effet (ou rien si même ligne); pas d'interaction



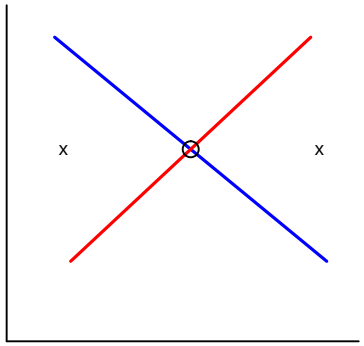
B X: effet important; Y: faible effet; pas d'interaction



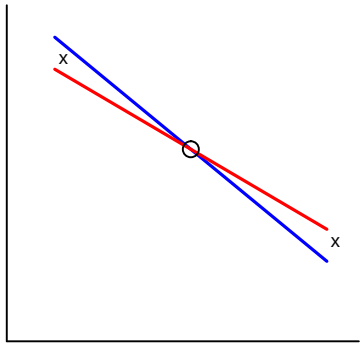
C X: pas d'effet; Y: effet important; pas d'interaction



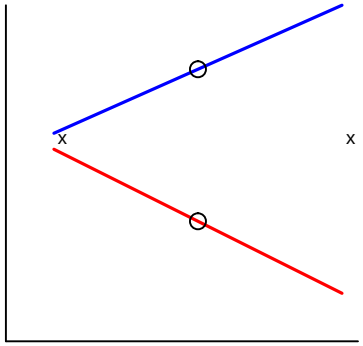
D X: effet important; Y: effet important; pas d'interaction



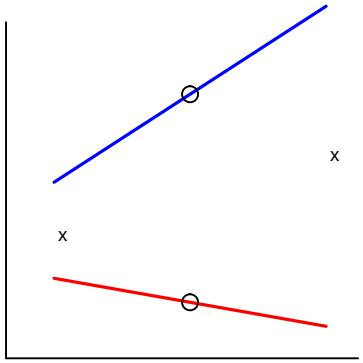
E X: pas d'effet; Y: pas d'effet; interaction forte



F X: effet important; Y: pas d'effet; légère interaction



G X: pas d'effet; Y: effet important; interaction importante



H X: effet modéré; Y: effet important; interaction importante

Les différents types d'ANOVA à plusieurs facteurs

- **Type I ("effets fixes")** : les traitements sont déterminés par le chercheur

Exemple : la croissance en taille d'un poisson en fonction du pH et de la température de l'eau, **tous deux fixés par l'expérimentateur**.

La variable dépendante est le taux de croissance et les deux facteurs sont le pH et la température.

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 truites.

Les différents types d'ANOVA à plusieurs facteurs

- **Type I ("effets fixes")** : les traitements sont déterminés par le chercheur

- **Type II ("effets aléatoires")** : les traitements ne sont pas sous le contrôle de l'expérimentateur

Exemple : la taille d'un lézard en fonction de la région et de l'altitude, **tous deux aléatoires (non fixés par l'expérimentateur)**.

La variable dépendante est la taille et les deux facteurs sont la région et l'altitude.

Même si la taille diffère en fonction de la région ou de l'altitude, on ne savoir quel facteur est responsable de cette variabilité et **prédire** la taille pour une autre région ou une autre altitude.

Les différents types d'ANOVA à plusieurs facteurs

- **Type I ("effets fixes")** : les traitements sont déterminés par le chercheur
- **Type II ("effets aléatoires")** : les traitements ne sont pas sous le contrôle de l'expérimentateur
- **Type III ("modèle mixte")** : au moins un facteur du Type I et au moins un du Type II

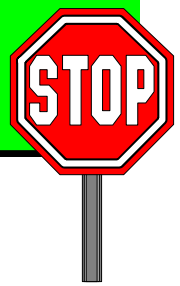
Exemple : la taille d'un ours en fonction de la région (**variable aléatoire**) et du sexe (**variable fixée**).

La variable dépendante est la taille et les deux facteurs sont la région et le sexe.

Même si la taille diffère en fonction de la région ou du sexe, on ne savoir quel facteur est responsable de cette variabilité et **prédire** la taille des ours de chaque sexe pour une autre région. Par contre, on peut prédire (peut-être) la différence entre les sexes.

Les facteurs fixes versus les facteurs aléatoires pour l'ANOVA

	Facteur fixe	Facteur aléatoire
• Manipulation par l'expérimentateur?	Oui	Non
• Estimation de l'effet des traitements?	Oui	Non
• Prédiction?	Oui	Non
• Calcul de l'ANOVA à un critère de classification • Calcul de l'ANOVA à plusieurs critères de classification	Identique Différent (très !)	



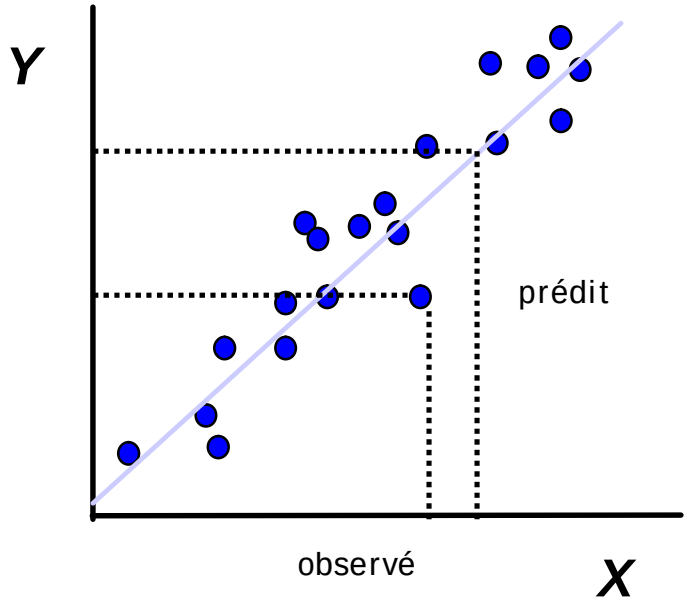
Il faut donc renseigner soigneusement le modèle dans le logiciel utilisé pour faire les calculs !

Procédure	*Variable dépendante*	*Variable(s) indépendante(s)*
ANOVA 1 facteur	**1 continue**	**1 discontinue***
ANOVA n facteurs	**1 continue**	**2 ou plus discontinues ***

* peuvent être discontinues ou traitées comme discontinues (=discrètes)

Régression simple

- Ajustement d'une ligne droite à travers un nuage de points
- Test et quantification de l'effet d'une variable indépendante X sur la variable dépendante Y
- L'intensité de l'effet est donnée par la pente (b) de la régression
- L'importance de l'effet est donné par le coefficient de détermination (r^2)



Régression simple : coefficients de corrélation et de régression

- La pente est obtenue par:

$$b = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{N}(X_i - \overline{X})^2} = \frac{Cov(X,Y)}{s_X^2}$$

- Le coefficient de corrélation *r:*

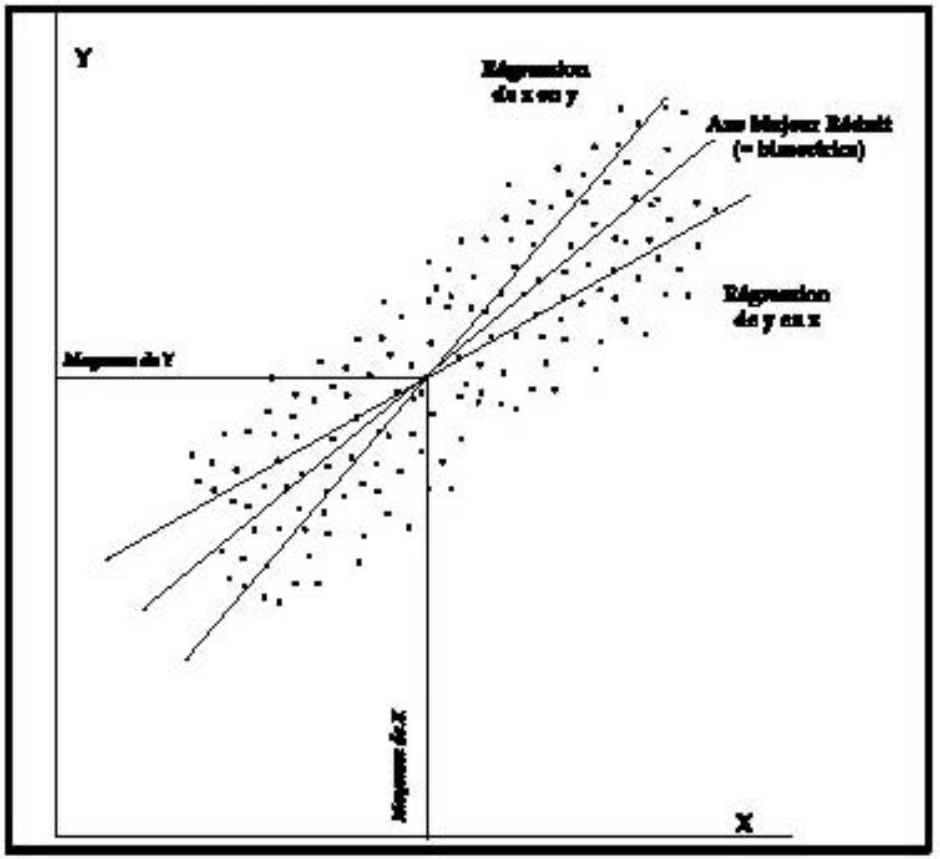
$$r = \frac{Cov(X,Y)}{s_X s_Y}$$

- Alors

$$r = b\frac{s_X}{s_Y}$$

AXE MAJEUR RÉDUIT ou AXE PRINCIPAL RÉDUIT

Droite de Tessier, Relation d'Allométrie de Tessier



Régression de y en x

$$y = a.x + b$$

Coeff. Corrélation = **r**

Régression de x en y

$$x = c.y + d$$

Coeff. Corrélation = **r**

$$\mathbf{c = r^2/a}$$

La pente de l'Axe Majeur Réduit sera :

$$\sqrt{a / c} = \frac{a}{r}$$

Connaissant la pente, on trouve l'ordonnée à l'origine en sachant que la droite de Tessier passe forcément par les moyennes; r ne change pas.

Remarques : - ce n'est rien d'autre que la **moyenne géométrique** des deux pentes, **a et 1/c**, exprimées dans le même repère
- plusieurs méthodes ont été proposées pour calculer un modèle II (aucune dans les packages classiques !) : (i) la méthode du maximum de vraissemblance, (ii) axe majeur (moindres rectangles - projections perpendiculaires et non verticales), (iii) axe majeur réduit (Tessier, 1948)

Procédure	*Variable dépendante*	*Variable(s) indépendante(s)*
ANOVA 1 facteur	**1 continue**	**1 discontinue***
ANOVA n facteurs	**1 continue**	**2 ou plus discontinues ***
Régression simple	**1 continue**	**1 continue**
Régression multiple	**1 continue**	**2 ou plus continues**

* peuvent être discontinues ou traitées comme discontinues (=discrètes)

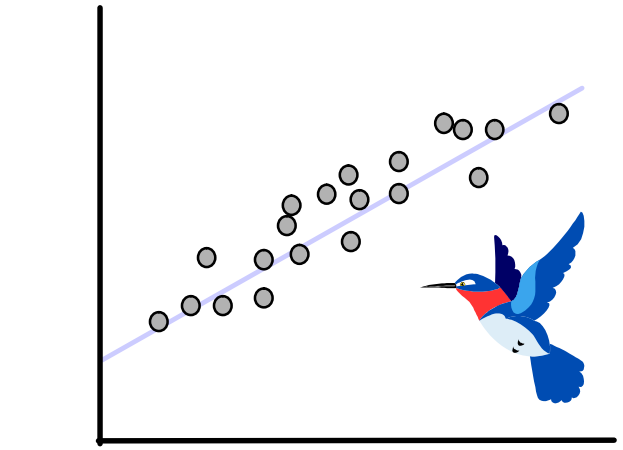
Régression et ANOVA

Comparaison Taille-Poids chez différents groupes de Vertébrés

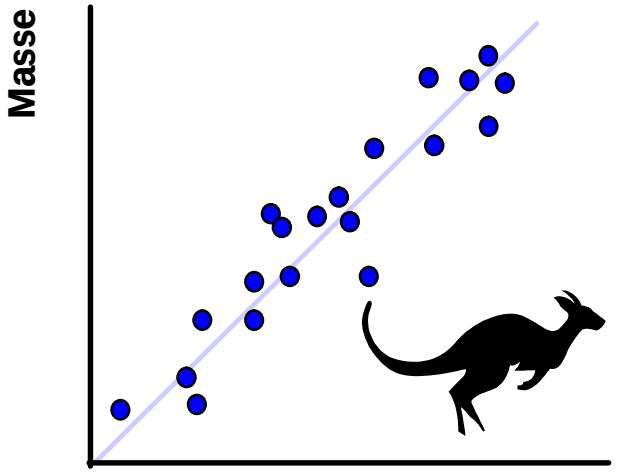
Pour une taille donnée, il est normal (on s'attend à …) que le poids d'un mammifère soit plus important que celui d'un oiseau.

Deux régressions différentes s'imposent.

Maintenant imaginons que l'on cherche à comparer des tailles et des poids sans tenir compte du groupe taxinomique : R^2 serait probablement très faible (pas de corrélation et donc pas de régression) !



Taille



Taille

Régression et ANOVA

Comparaison du Poids d'un animal en fonction de différents régime alimentaire

Si le régime alimentaire est riche, il est normal (on s'attend à …) que le poids de l'animal soit plus élevé.

Si plus de 2 régimes alimentaires sont comparés, une ANOVA à 1 facteur (le régime) s'impose.

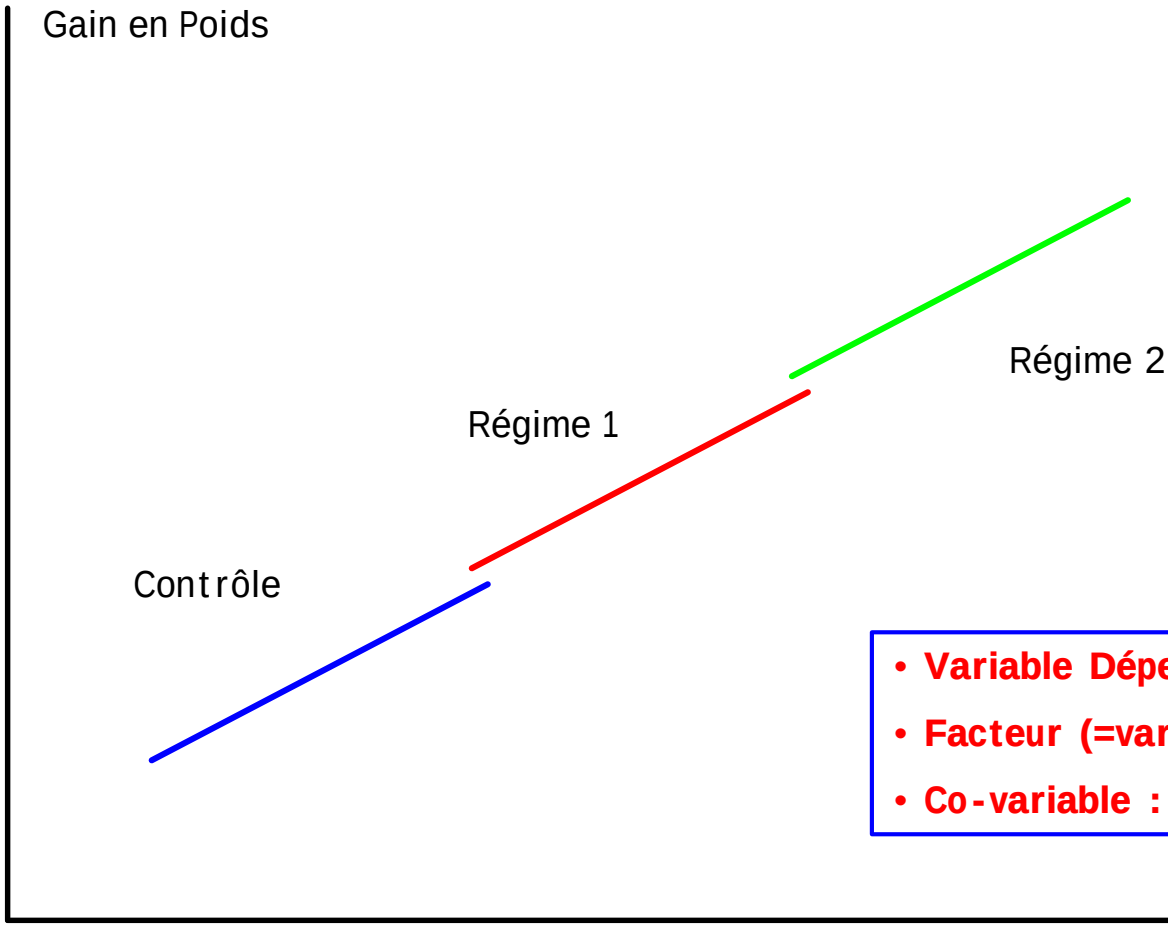
Mais quelle est la condition à respecter ?

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

Si cette condition n'est pas réalisée, il est possible d'introduire dans le modèle la variabilité que l'on connaît déjà : l'effet de la taille.

C'est donc une ANOVA (1 facteur) avec une **co-variable** (taille). On parle d'**ANCOVA**

Régression et ANOVA



Une ANOVA classique mettra en évidence une différence significative entre les régimes alimentaires.

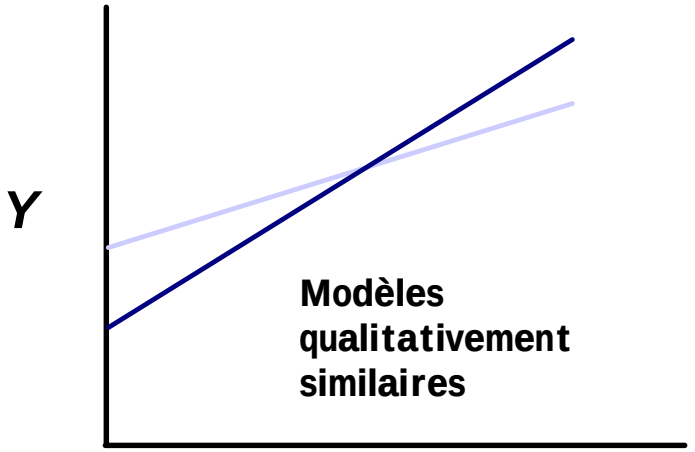
Par contre, une simple observation du graphique montre que les gains en poids ne sont liés qu'aux gammes différentes de taille utilisées.

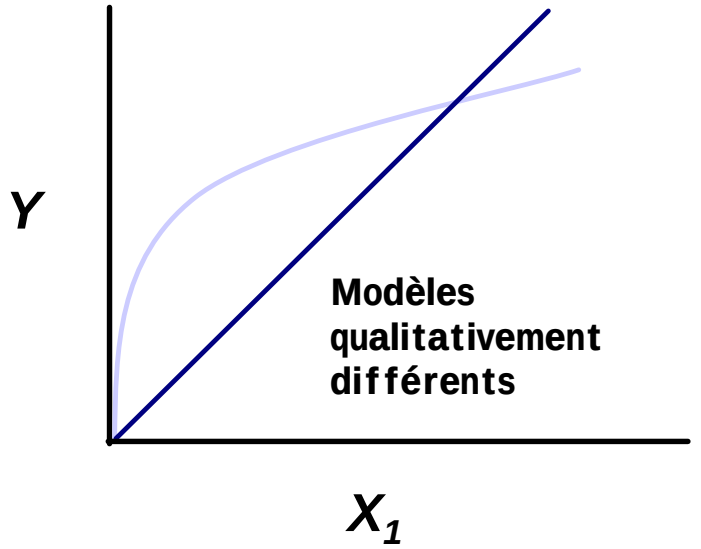
L'introduction de la taille en co-variable ne mettra plus en évidence de différence significative entre les régimes alimentaires !

- **Variable Dépendante : continue**
- **Facteur (=var. indép.) : discret (discontinu)**
- **Co-variable : continue**

Utilisation de l'ANCOVA

- Lorsque l'on fait ces comparaisons, on suppose que les modèles sont qualitativement similaires pour tous les niveaux de la variable discontinue (la co-variable) …
- …autrement ce serait comme comparer des pommes et des oranges !
- ANCOVA est utilisée afin de comparer des modèles *linéaires* généralement.





Les Procédures sont homogènes

Procédure	*Variable dépendante*	*Variable(s) indépendante(s)*
ANOVA 1 facteur	**1 continue**	**1 discontinue***
ANOVA n facteurs	**1 continue**	**2 ou plus discontinues ***
Régression simple	**1 continue**	**1 continue**
Régression multiple	**1 continue**	**2 ou plus continues**
ANCOVA	**1 continue**	**Au moins 1 discontinue* et au moins une 1 continue**

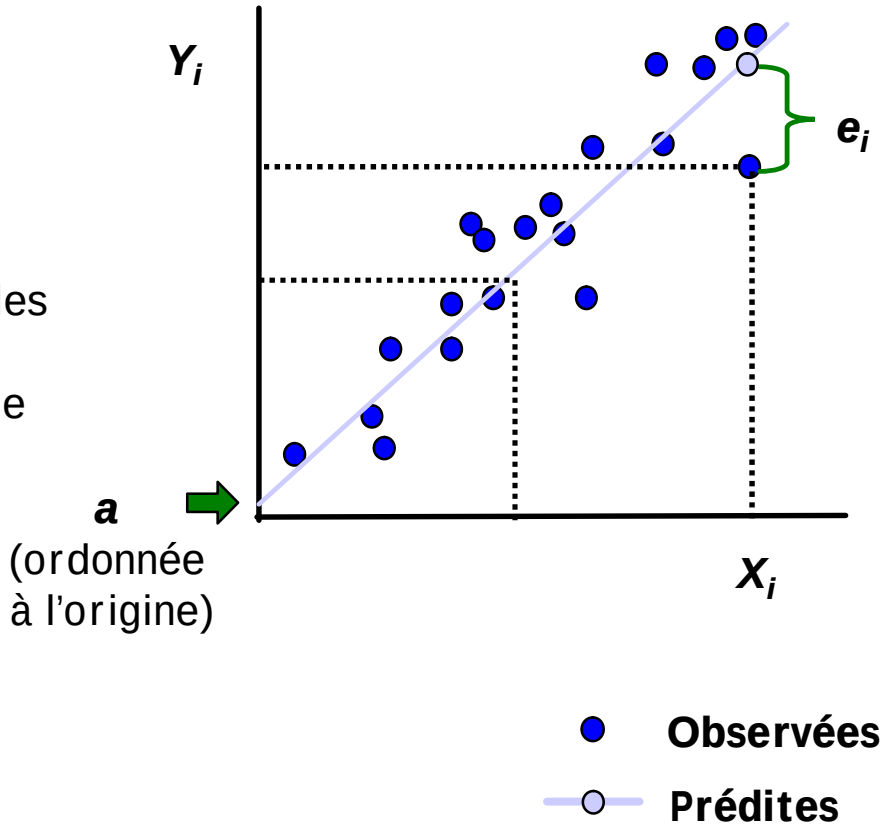
* peuvent être discontinues ou traitées comme discontinues (=discrètes)

Le modèle de la régression simple

- Le modèle de la régression:

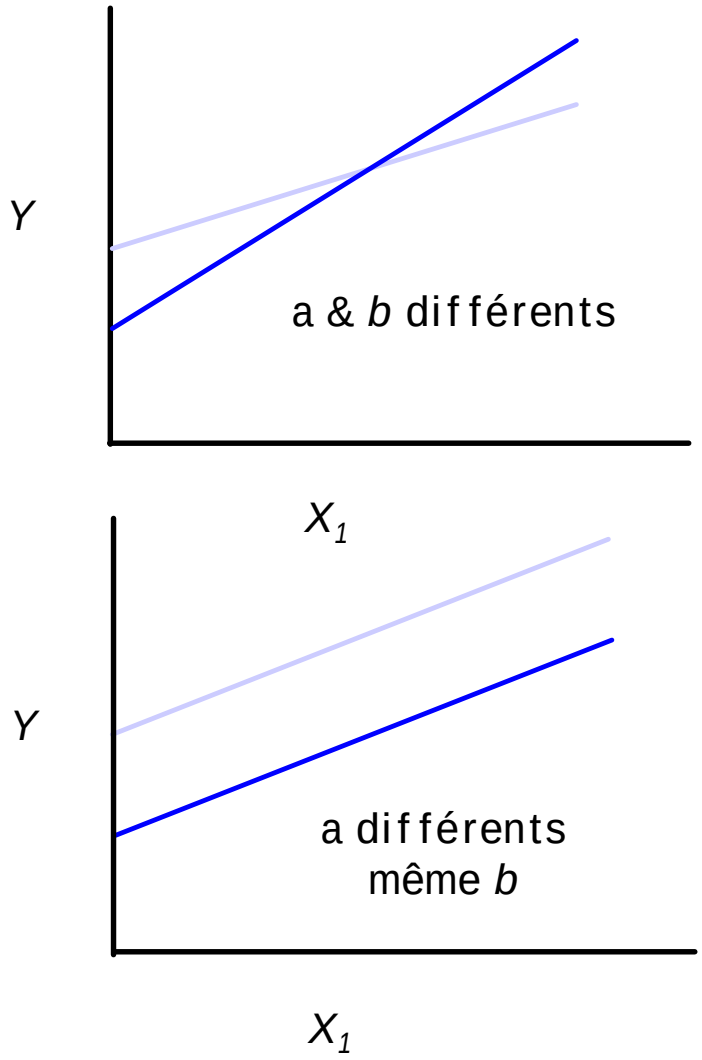
$$Y_i = a + bX_i + e_i$$

- alors, toutes les régressions simples sont décrites par 2 paramètres: l'ordonnée à l'origine (a) et la pente (*b*)



Modèles GLM simples

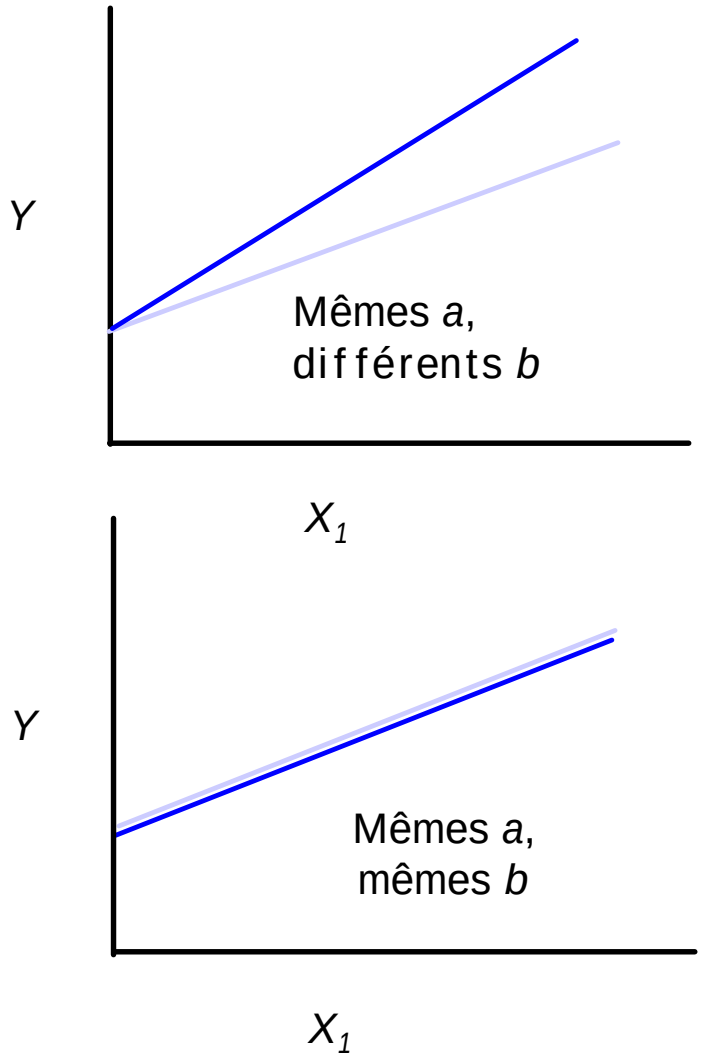
- Les ordonnées à l'origine (*a*) et les pentes (*b*) sont différentes.
- Les ordonnées à l'origine sont différents mais les pentes sont les mêmes.



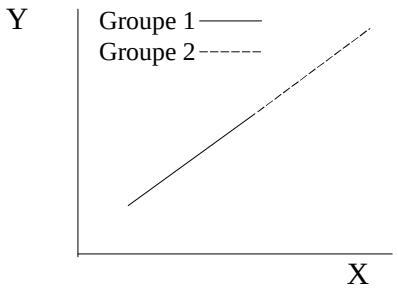
Modèles GLM simples

- Mêmes ordonnée à l'origine (*a*) mais les pentes (*b*) sont différentes.

- Mêmes pentes et mêmes ordonnées à l'origine .

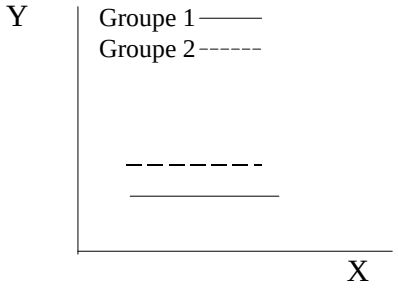


Y: variable dépendante; X: co-facteur (Prédicteur Continu); G: variable indépendante (Prédicteur Catégoriel; discret). On teste les effets de X, G et X*G (interaction) sur la variable Y



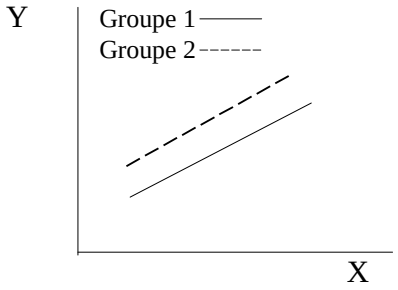
Cas 1 : X est significatif, G et X*G ne le sont pas.

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



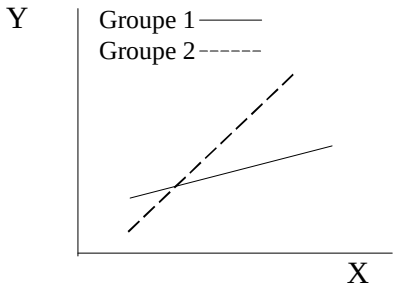
Cas 2 : G est significatif, X et X*G ne le sont pas.

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



Cas 3 : G et X sont significatifs, X*G ne l'est pas.

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



Cas 4 : G, X et X*G sont significatifs.

Y change en changeant X, alors X affecte Y. Les points d'intersection des deux groupes sont différents, alors G affecte Y également. En plus, les deux pentes sont différentes (les lignes ne sont pas parallèles) donc l'effet de Y sur X dépend de la valeur de G (c'est-à-dire, dépend du groupe). Alors X*G est significatif.

Modèles GLM simples

Ils peuvent donc être utilisés pour comparer des droites de régression.

Par exemple, pour comparer les droites de régression entre la taille et le poids pour différentes espèces :

- Poids = variable dépendante
- Taille = variable indépendante = prédicteur continu = co-variable
- Espèce = facteur ou catégorie

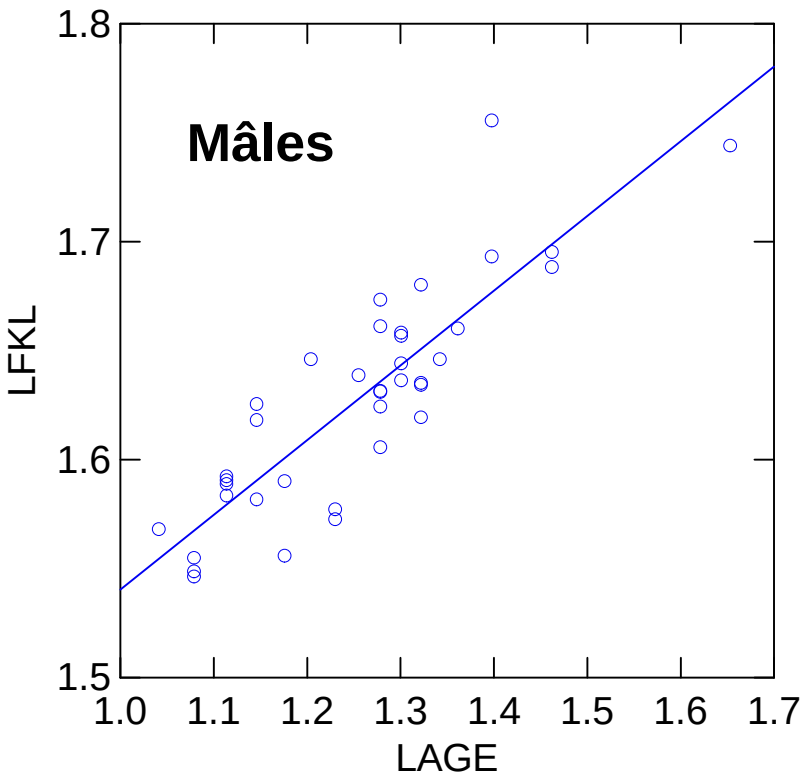
Analyse :

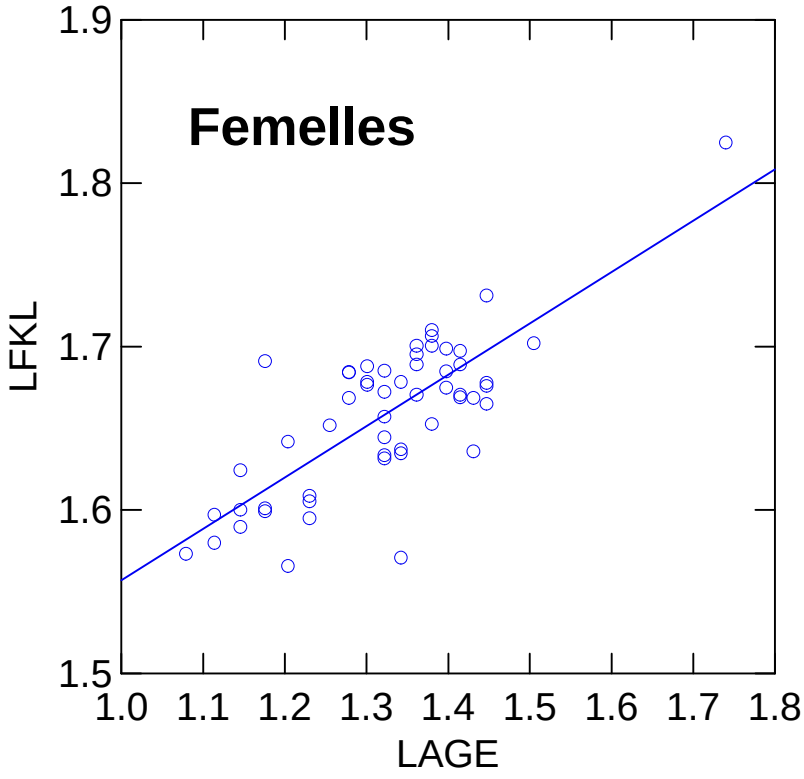
- Comparer les pentes
- Si les pentes ne sont pas statistiquement différentes, comparaison des ordonnées à l'origine
- Si les pentes sont statistiquement différentes, la comparaison des ordonnées à l'origine ne s'impose pas.

Tester les pentes revient à tester les interactions

Tester les ordonnées à l'origine revient à tester le prédicteur catégoriel

Effets du sexe et de l'âge sur les esturgeons

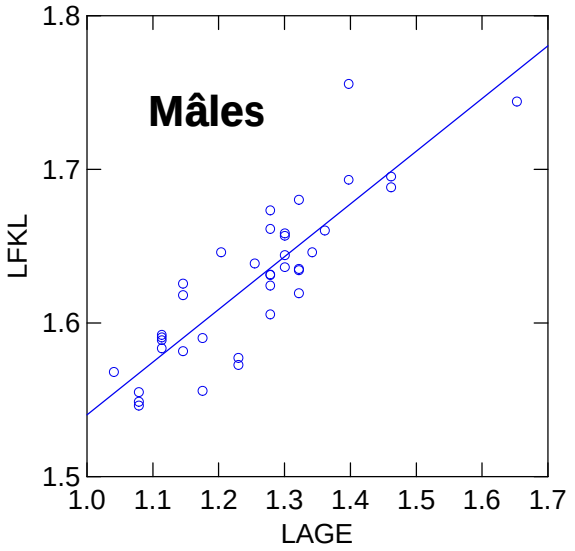


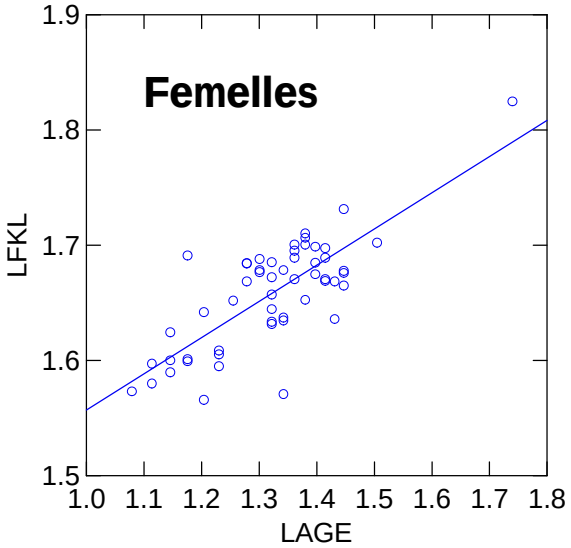


Analyse

- Log(forklength)(LFKL) est la variable dépendante, log(age) (LAGE) est la variable indépendante continue, et sex (SEX$) est la variable discontinue (2 niveaux)

- ***Q1***: la pente de la régression de LFKL sur LAGE est la même pour les deux sexes?





Effets du sexe et de l'âge sur les esturgeons

Dep Var: LFKL N: 92 Multiple R: 0.835 Squared multiple R: 0.697

Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
LAGE	0.143	1	0.143	176.650	0.000
SEX$	0.000	1	0.000	0.504	0.479
SEX$*LAGE	0.000	1	0.000	0.337	0.563
Error	0.071	88	0.001		

Conclusion 1 : la pente est la même pour les deux sexes - p(SEX$*LAGE) > .05

Q2 : l'ordonnée à l'origine est-elle la même?

Effets du sexe et de l'âge sur les esturgeons

Dep Var: LFKL N: 92 Multiple R: 0.834 Squared multiple R: 0.696

Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
SEX$	0.001	1	0.001	1.851	0.177
LAGE	0.143	1	0.143	178.163	0.000
Error	0.072	89	0.001		

Conclusion 2 : Ordonnée à l'origine est la même pour les deux sexes - p(SEX$ > .05)
Le meilleur modèle est donc la régression commune.

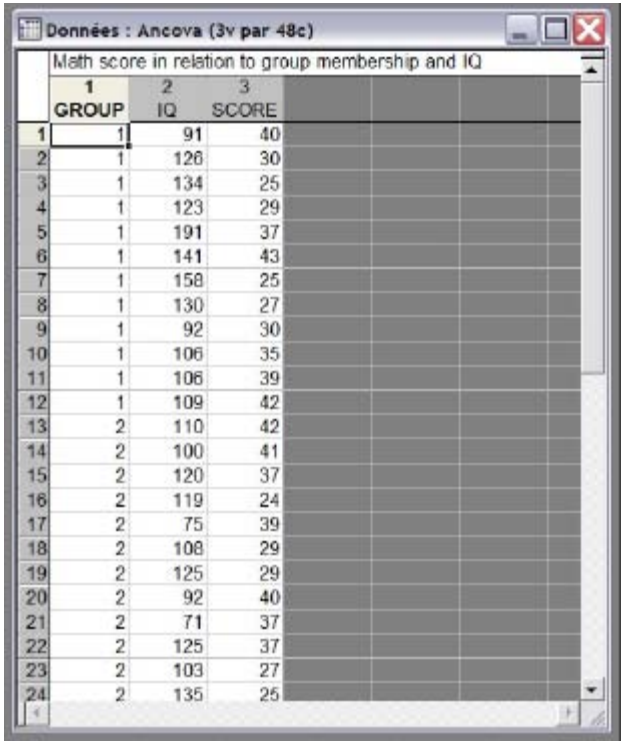
Effets du sexe et de l'âge sur les esturgeons

Dep Var: LFKL N: 92 Multiple R: 0.830 Squared multiple R: 0.690

Adjusted squared multiple R: 0.686 Standard error of estimate: 0.029

Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	1.211	0.031	0.0	.	39.191	0.000
LAGE	0.336	0.024	0.830	1.000	14.144	0.000

Exemple d'Analyse de Covariance (Statistica)



Données : Ancova (3v par 48c)

Math score in relation to group membership and IQ

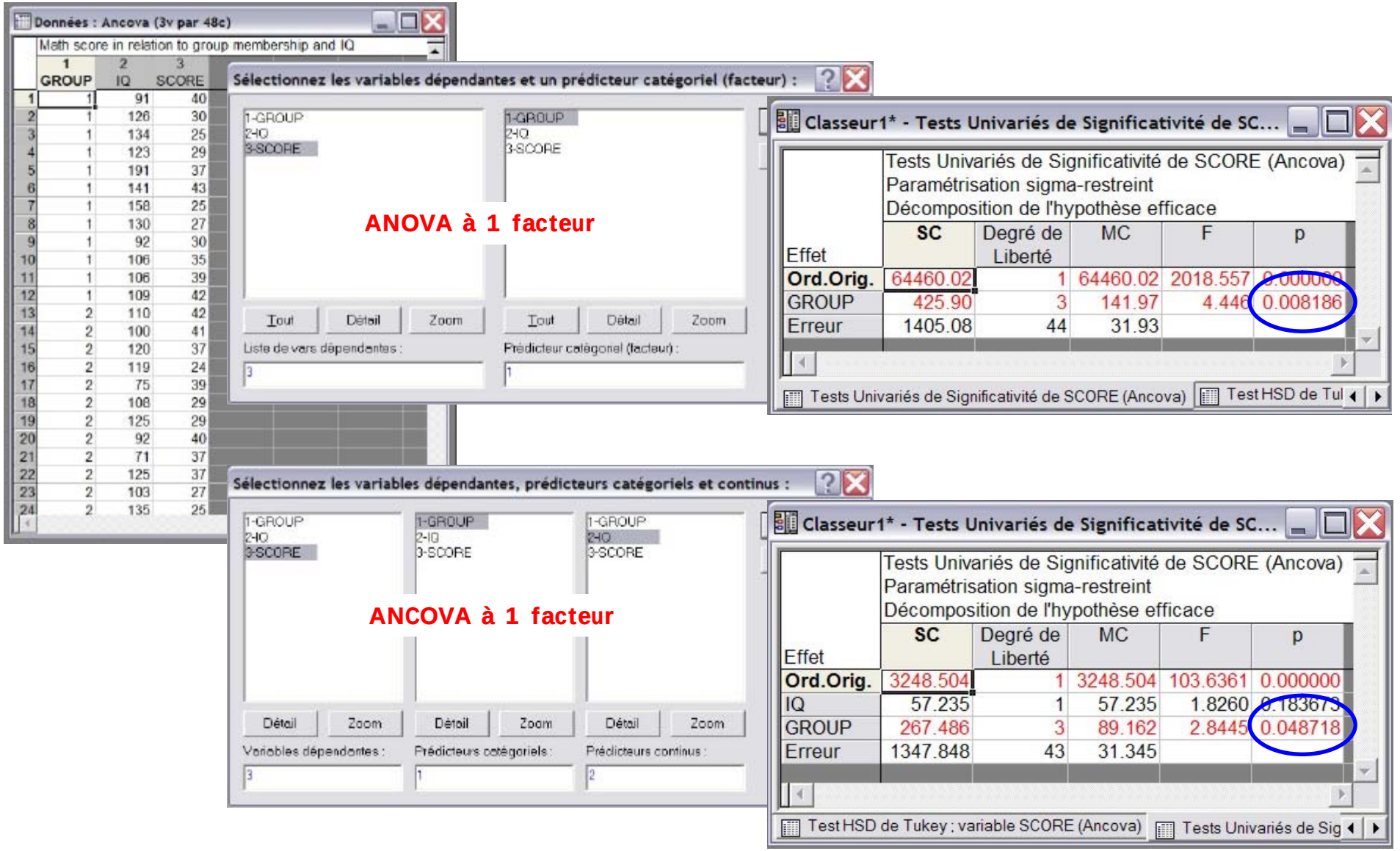
	1 GROUP	2 IQ	3 SCORE
1	1	91	40
2	1	126	30
3	1	134	25
4	1	123	29
5	1	191	37
6	1	141	43
7	1	158	25
8	1	130	27
9	1	92	30
10	1	106	35
11	1	106	39
12	1	109	42
13	2	110	42
14	2	100	41
15	2	120	37
16	2	119	24
17	2	75	39
18	2	108	29
19	2	125	29
20	2	92	40
21	2	71	37
22	2	125	37
23	2	103	27
24	2	135	25

Les résultats en mathématiques (*Score*) sont comparés entre différents groupes d'étudiants recevant chacun des méthodes d'enseignement différentes (*Group*). Le quotient intellectuel (*IQ*) est pris comme co-variable.

Dans cet exemple, **il n'y a pas d'interaction** entre le prédicteur catégoriel (*Group*) et le prédicteur continu (*IQ*).

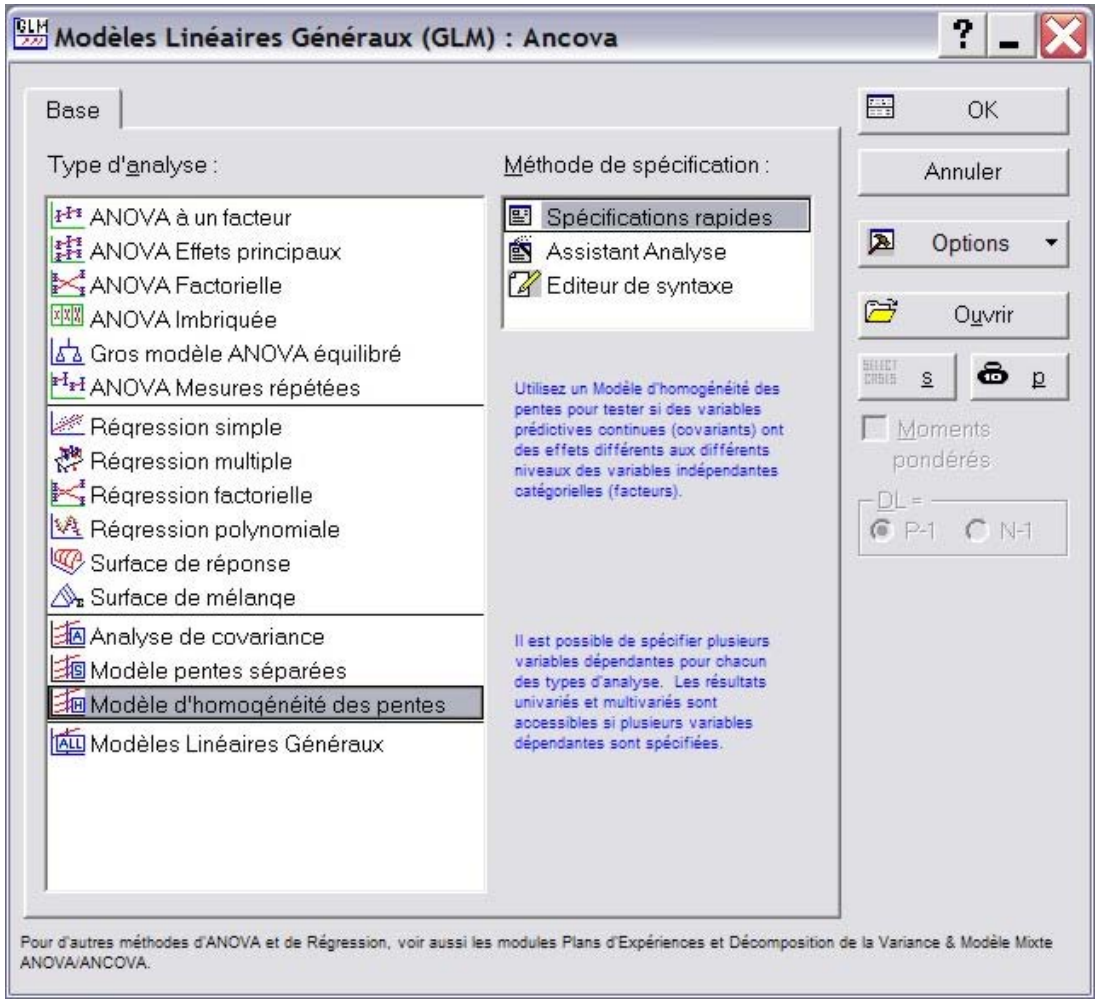
La méthode d'enseignement dispensée aux différents groupes (*Group*) est supposée indépendante du niveau intellectuel des étudiants (*IQ*).

Exemple d'Analyse de Covariance (Statistica)



Exemple d'Analyse de Covariance (Statistica)

Dans Statistica, la démarche est la suivante :



- prendre l'option générale Homogeneity of Slopes : permet de tester si oui ou non les pentes diffèrent (**pas d'a priori**)
- si les pentes diffèrent réellement [p(interaction) < 0.05], passer au modèle de co-variance à pentes séparées (Separate-slope model)
- si les pentes ne diffèrent pas [p(interaction) > 0.05], passer au modèle traditionnel (Analysis of covariance)