ANALYSE DE LA COVARIANCE (ANCOVA)

Christelle NGUENANG
Garmy SAMB
Daouda SISSOKO
Moussa FALL
Hamissou Alaji BOUAHARI
Jean MOYENGA

Eleves Ingénieurs statisticiens Economistes
ENSAE

11 février 2023

Tables des matières

Tables des matières

Introduction

L'analyse de la covariance en abrégé ANCOVA (analysis of covariance), permet d'expliquer une variable quantitative par plusieurs variables à la fois qualitatives et quantitatives. Elle peut être vu comme une extension de l'ANOVA qui incorpore une ou plusieurs variables quantitatives. Par ailleurs, l'ANCOVA peut être aussi vu comme un mélange de l'ANOVA et du modèle de régression linéaire. Car la variable est quantitative, le modèle est linéaire et ils ont les mêmes hypothèses. L'ANCOVA est une méthode statistique visant à tester, par un modèle linéaire général, l'effet sur une variable dépendante continue sur une ou plusieurs variables indépendantes catégorielles, indépendamment de l'effet des autres facteurs quantitatif continus, dits covariables. Elle teste si la variable indépendante influence toujours la variable dépendante après l'introduction d'une ou de plusieurs covariables.

Présentation de l'ANCOVA

- **Définition des concepts**

On appelle :

Variable dépendante : la variable quantitative qui étudiée.

Variables indépendantes : les variables utilisées pour expliquer les variations de la variable dépendantes.

Facteurs : les variables indépendantes catégorielles

Covariables : les variables indépendantes quantitatives. La prise en compte d'une ou plusieurs covariables permet de contrôler une partie de la variation de la variable indépendante d'intérêt.

Présentation de l'ANCOVA

Distinction ANOVA et ANCOVA

L'ANOVA et l'ANCOVA sont deux méthodes statistiques utilisées pour comparer les moyennes entre deux ou plusieurs groupes. La différence entre ces méthodes réside au niveau des covariables et l'hypothèse de linéarité.

Ces différences sont consignées dans le tableau ci-dessous :

Paramètres de comparaison	**ANOVA**	**ANCOVA**
Stands pour	Analyse de variance (ANOVA)	Analyse de covariance
Les usages	Il existe un mélange de modèles linéaires et non linéaires.	Le modèle linéaire est utilisé seul.
Inclus	Variable catégorielle	Des variables à la fois catégorielles et numériques

Présentation de l'ANCOVA

Hypothèses

Une condition préalable à la réalisation d'une ANCOVA est la vérification des hypothèses par les données. Ces hypothèses sont les suivantes :

- **Normalité** : Les distributions de la variable dépendante doivent être normale.
- **Linéarité** : entre la covariable et la variable dépendante.
- **Homogénéité** : la pente de la droite formée par la variable dépendante et la covariable doit être la même pour chaque groupe. Autrement dit, les lignes de régression entre les différents groupes doivent être parallèles.
- **Indépendance** : des observations.

Présentation de l'ANCOVA

Exemple introductif

On cherche à expliquer la variable quantitative Pds final à partir de

la variable quantitative Pds initial

la variable qualitative Traitement

On dispose de ce fait de 20 observations

Traitements	Pds initial	Pds final
T1	27,2 ;32,0 ;33,0 ;26,8	32,6 ;36,6 ;37,7 ;31,0
T2	28,6 ;26,8 ;26,5 ;26,8	33,8 ;31,7 ;30,7 ;30,4
T3	28,6 ;22,4 ;23,2 ;24,4	35,2 ;29,1 ;28,9 ;30,2
T4	29,3 ;21,8 ;30,3 ;24,3	35,0 ;27,0 ;36,4 ;30,5
T5	20,4 ;19,6 ;25,1 ;18,1	24,6 ;23,4 ;30,3 ;21,8

Tables des matières

Modélisation

Notation

On note Y la variable quantitative ici le Poids final des huitres.

- Y_{ij} est la valeur de Y pour l'observation j pour le traitement i.
- X_{ij} est la valeur du poids initial de l'observation j pour le traitement i.

Chaque individu de l'échantillon est repéré par un double indice (i ; j), i représentant le niveau du facteur auquel appartient l'individu, et j correspondant à l'indice de l'individu dans le niveau i (j = 1,...,n_i). n = $\sum_i n_i$ est le nombre d'observations.

Nous supposons ici que les n_i sont les mêmes pour chaque groupe de traitement.

Modélisation

Ecriture des moyennes

- Moyenne globale des poids finals :

$$\bar{Y} = \frac{1}{n_i I} \sum_i \sum_j Y_{ij} = \frac{1}{I} \sum_i \bar{Y}_{i.}$$

- Moyenne des poids finals pour le traitement i :

$$\bar{Y}_{i.} = \frac{1}{n_i} \sum_j Y_{ij}$$

- Moyenne global du poids initial :

$$\bar{X} = \frac{1}{n_i I} \sum_i \sum_j X_{ij}$$

- Moyenne des poids initiaux pour le traitement i :

$$\bar{X}_{i.} = \frac{1}{n_i} \sum_j X_{ij}$$

Ecriture des variances ou somme des carrés

1. Somme des carrés totals de X

$$(SCT)_X = \sum_i \sum_j (X_{ij} - \bar{X})^2$$

2. Somme des carrés des traitements pour X

$$(SCF)_X = \sum_i \sum_j (\bar{X}_{i.} - \bar{X})^2$$

3. Somme des carrés totals de Y

$$(SCT)_Y = \sum_i \sum_j (Y_{ij} - \bar{Y})^2$$

4. Somme des carrés des traitements pour Y

$$(STF)_Y = \sum_i \sum_j (Y_{i.} - \bar{Y})^2$$

Modélisation

5. Somme des produits totale de X et Y

$$SPT = \sum_i \sum_j (X_{ij} - \bar{X})(Y_{ij} - \bar{Y})$$

6. somme des produits des traitements de X et Y

$$SPF = \sum_i \sum_j (\bar{X}_{i.} - \bar{X})(\bar{Y}_{i.} - \bar{Y})$$

7. Somme des carrés des erreurs pour X

$$(SCE)_X = \sum_i \sum_j (X_{ij} - \bar{X})^2$$

8. Somme des carrés des erreurs pour Y

$$(SCE)_Y = \sum_i \sum_j (Y_{ij} - \bar{Y})^2$$

Modélisation

9. Somme des produits des erreurs pour X et Y

$$SPE = \sum_i \sum_j (X_{ij} - \bar{X})(Y_{ij} - \bar{Y})$$

Modélisation

9. Somme des produits des erreurs pour X et Y

$$SPE = \sum_i \sum_j (X_{ij} - \bar{X})(Y_{ij} - \bar{Y})$$

Spécification du modèle

Le modèle d'analyse de la covariance s'écrit comme suit :

$$Y_{ij} = \mu_i + \beta_i X_{ij} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2)$$

où,

- μ_i est la valeur du poids final pour un sac de poids initial nul pour le traitement i (ordonnée à l'origine) ;
- β_i est la pente de régression pour le traitement i ;
- σ^2 est la variance résiduelle (identique pour tous les traitements).

Modélisation

Comme dans l'ANOVA, nous allons introduire les termes différentielles en décomposant le modèle pour tout i comme suit :

- $\mu_i = \mu + \alpha_i$ avec $\sum \alpha_i = 0$
  - μ représente l'effet global du traitement
  - α_i réprésente l'effet spécifique du niveau du traitement i.
- $\beta_i = \beta + \gamma_i$ avec $\sum \gamma_i = 0$
  - β représente un effet commun à chaque groupe
  - γ_i représente l'effet spécifique du niveau du traitement i.

Le modèle s'écrit donc

$$Y_{ij} = \mu + \alpha_i + \beta X_{ij} + \gamma_i X_{ij} + \varepsilon_{ij}$$

Le terme $\gamma_i X_{ij}$ représente l'interaction entre le facteur Traitement et le Poids initial.

Modélisation

Par la suite, nous allons considérer que l'interaction est nulle ie $\gamma_i = 0$ pour tout i.

Le modèle sera donc un modèle sans interaction et sera formuler comme suit :

$$Y_{ij} = \mu + \alpha_i + \beta X_{ij} + \varepsilon_{ij}$$

où,

$$Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{X}) + \varepsilon_{ij}$$

Estimation des paramètres

Les estimateurs des paramètres du modèle sont donnés par les formules suivantes

$$\hat{\mu} = \bar{Y}$$

$$\hat{\alpha}_i = \bar{Y}_{i.} - \hat{\beta}(X_{ij} - \bar{X}) - \hat{\mu}$$

$$\hat{\beta} = \frac{\sum_i \sum_j (X_{ij} - \bar{X}_{i.})(Y_{ij} - \bar{Y}_{i.})}{\sum_i \sum_j (X_{ij} - \bar{X}_{i.})}$$

$$\hat{\varepsilon_{ij}} = Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}(X_{ij} - \bar{X})$$

Modélisation

Critère d'influence du facteur

Après avoir estimer les paramètres du modèle, il s'agira de tester l'effet spécifique du facteur sur la variable Y. On teste

$H_0 : \alpha_i = 0, \forall i \in \{1, ..., I\}$ contre $H_1 : \exists$ i' tel que $\alpha_{i'} \neq 0$.

La statistique de test est donnée par

$$F = \frac{CMF}{CME} \text{ avec } CMF = \frac{SCF}{p-1} \text{ et } CME = \frac{SCE}{n_i - p - 1}$$

Sous $H_0 \rightsquigarrow F(p-1; n_i - p - 1)$

Règle de décision :

Si $F > F_{lu}$, on accepte H_0

Si $F \leq F_{lu}$, on rejette H_0

Modélisation

Pour tester l'absence d'effet de la covariable, on pose :

$H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$.

La statistique de test est donnée par

$$F = \frac{\frac{(SPE)^2}{(SCE)_X}}{CME} \text{ sous } H_0 \; F \rightsquigarrow F(1; n_i - p - 1)$$

Règle de décision :
Si $F > F_{lu}$, on accepte H_0
Si $F \leq F_{lu}$, on rejette H_0

Tableau recapitulatif

Sources de variation	ddl	Somme des carrés	Carré moyen	F
Facteur	P - 1	SCF	$CMF = \frac{SCF}{P-1}$	$F = \frac{CMF}{CME}$
Erreurs	n_i - p - 1	SCE	$CME = \frac{SCE}{n_i - p - 1}$	
Total	n_i - p - 2	SCT		

Tables des matières

Cas pratique

Suivons ensemble
un cas pratique

Tables des matières

Conclusion

Merci
de votre attention !