Université Paris Diderot **Économétrie**
Licence 3, 2012-2013 **Vianney Perchet - Aurélie Fischer**

T.D. 6 : Le modèle ANOVA

Dans cette séance, on étudie un cas particulier de modèle linéaire : le modèle ANOVA (ANalysis Of VAriance) : il s'agit du cas où toutes les variables explicatives sont qualitatives. On profite de cette occasion pour faire des révisions sur le modèle linéaire en général (Exercice 1), et pour s'offrir une petite croisière avec Kate Winslet et Leonardo DiCaprio (Exercice 2). L'exercice 3 est un exercice d'application de l'analyse de la variance à 1 facteur.

Exercice 1. On souhaite modéliser une variable quantitative Y à l'aide d'une seule variable qualitative x pouvant prendre p modalités différentes. Le modèle s'écrit

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

où

- $i = 1, \dots, p$ est l'indice du groupe auquel appartient l'individu.
- $j = 1, \dots, n_i$ est le numéro de l'individu au sein du groupe d'indice i.
- Les ε_{ij} sont des variables aléatoires indépendantes de même loi $\mathcal{N}(0, \sigma^2)$.

1) Déterminer les estimateurs des moindres carrés des μ_i.
2) Ecrire l'équation d'analyse de la variance. Interpréter.
3) Comment tester l'hypothèse $\mathbf{H}_0$: la variable x n'a aucune influence sur la variable Y ? Donner la statistique de test et sa loi.

Exercice 2. On traite ici une application sur une base de données célébrissime, disponible à l'adresse suivante : *http ://www.amstat.org/publications/jse/datasets/titanic.txt*. Dans cette base est reporté pour chaque individu présent sur le Titanic lors de son naufrage si c'est un enfant ou un adulte, sa classe en tant que passager (1^{re}, 2^e ou 3^e classe, ou équipage), son sexe, et la variable d'intérêt Y_i, égale à 1 si la personne a pu survivre au naufrage (en obtenant une place dans un canot de sauvetage) et à 0 sinon.

On précise les informations suivantes : $n = 2801$, $n_{\text{femme}} = 425$, $n_{\text{fille}} = 45$, $n_{\text{garçon}} = 64$, $n_{\text{homme}} = 2267$.

On applique tout d'abord un modèle ANOVA tenant compte de l'âge et du sexe :

$$Y_i = b_0 + b_1 \mathbf{1}_{i=\text{ garçon}} + b_2 \mathbf{1}_{i=\text{ fille}} + b_3 \mathbf{1}_{i=\text{ femme}} + \varepsilon_i.$$

1) Les hypothèses usuelles peuvent-elles être satisfaites ? Penser en particulier au fait que $Y_i \in \{0, 1\}$. Doit-on pour autant rejeter cette modélisation ?
2) On estime le modèle et on obtient :

$$\hat{y}_i = 0,22 + 0,24 \mathbf{1}_{i=\text{ garçon}} + 0,41 \mathbf{1}_{i=\text{ fille}} + 0,53 \mathbf{1}_{i=\text{ femme}}.$$

Quelle était la probabilité pour un homme de survivre ? Pour un garçon ? Une fille ? Une femme ?

3) On souhaite passer à un modèle plus simple, par exemple qui ne différencie que les enfants des adultes, sans tenir compte du sexe :

$$Y_i = c_0 + c_1 \mathbf{1}_{i=\text{ enfant}} + \varepsilon_i.$$

Comment estimer les coefficients c_0 et c_1 à partir des résultats précédents ?

Exercice 3. Dans une société de service, la direction cherche à étudier le montant des ventes produites par ses personnels. Elle s'intéresse à étudier l'influence sur le montant des ventes de la catégorie de personnel de l'individu ayant réalisé la vente, et du département dans lequel il travaille au sein de la société. Deux départements ont été retenus pour cette étude, le département Finance et le département Développement, et quatre catégories de personnels sont étudiées : employé(e), responsable junior, responsable senior, VP. Pour mener à bien cette étude, on relève le montant des ventes à la fin du mois pour 12 individus dans chacun des 2 départements : 3 employés, 3 responsables junior, 3 responsables seniors et 3 VP. Les résultats sont les suivants :

	employé			responsable junior			responsable senior			VP		
Dép. développement	13	16.5	19	20.5	21	22	35	46	48.5	12	23	24
Dép. finance	13	19.5	21	23	27	32	47	52	58	16	24	29

Table 1 – Montant des ventes (en milliers d'euros) en fonction de la catégorie de personnel et du département d'affectation.

On pourra noter y_{ijk} le montant des ventes du k^e individu travaillant dans le i^e département et de catégorie de personnel j.

On souhaite interpréter ce tableau à l'aide d'une analyse de la variance en étudiant l'effet des deux facteurs *département d'affectation* et *catégorie de personnel* sur le montant des ventes.
Tous les tests se feront au niveau 5%**.**

1) **On analyse tout d'abord l'influence de la** *catégorie de personnel***.**
   (a) Compléter le tableau de données suivant.

Catégorie de personnel	employé	responsable junior	responsable senior	VP	Global
Nombre d'observations					
Montant moyen des ventes					

Table 2 – Montant moyen des ventes par catégorie de personnel.

   (b) Décrire avec soin le modèle envisagé avec toutes ses hypothèses pour étudier l'effet du facteur *catégorie de personnel* sur le montant des ventes. On précisera quels sont les paramètres du modèle et leurs contraintes, et on donnera la dimension du modèle.
   (c) En vous servant du tableau précédent et de la table de l'analyse de la variance (Tableau 3) ci-dessous, donner une estimation des paramètres de ce modèle (paramètres d'espérance et de la variance résiduelle). On donnera directement l'expression des estimations de ces paramètres, puis leurs valeurs sur les données.

Source	ddl	Sommes des Carrés	Carrés Moyens	F
Modèle		3413.2		
Résidu				
Total		4051.8		

Table 3 – Table de l'ANOVA 1 du montant des ventes en fonction de la catégorie de personnel

(d) En analysant la table de l'ANOVA 1 (Tableau 3), peut-on conclure à l'existence d'une différence qualitative entre les catégories de personnel ?

On commencera par préciser les hypothèses H_0 et H_1 en termes de valeurs sur les paramètres ainsi qu'en termes de comparaison de modèles (on donnera alors la dimension de chacun des deux modèles que l'on compare), puis on donnera la statistique de test, sa loi sous H_0 et la règle de décision.

2) **On étudie maintenant l'influence du département d'affectation.**

(a) Compléter le tableau de données suivant.

Département d'affectation	développement	finance	Global
Nombre d'observations			
Montant moyen des ventes			

TABLE 4 – Montant moyen des ventes par département.

(b) Décrire avec soin le modèle envisagé avec toutes ses hypothèses pour étudier l'effet du facteur *département* sur le montant des ventes. On précisera quels sont les paramètres du modèle et leurs contraintes, et on donnera la dimension du modèle.

(c) En vous servant du tableau précédent et de la table de l'ANOVA 1 (Tableau 5) que vous aurez préalablement complétée, donner une estimation des paramètres de ce modèle (paramètres d'espérance et de la variance résiduelle). On donnera directement l'expression des estimations de ces paramètres, puis leurs valeurs sur les données.

Source	ddl	Sommes des Carrés	Carrés Moyens	F
Modèle		155.04		
Résidu			177.13	
Total				

TABLE 5 – Table de l'ANOVA 1 du montant des ventes en fonction du département

(d) En analysant la table de l'analyse de la variance (Tableau 5) ci-dessus, peut-on conclure que les départements ont des montants de vente différents ?

On commencera par préciser les hypothèses H_0 et H_1 en termes de valeurs sur les paramètres ainsi qu'en termes de comparaison de modèles (on donnera alors la dimension de chacun des deux modèles que l'on compare), puis on donnera la statistique de test, sa loi sous H_0 et la règle de décision.