Université Paris 7 **Économétrie**
Licence 3, 2009-2010 **Pierre Alquier - Joseph Salmon**

TD n° 6 : Le modèle ANOVA

Dans ce TD, on étudie un cas particulier de modèle linéaire : le modèle ANOVA (*ANalysis Of VAriance*) : il s'agit en fait du cas où toutes les variables explicatives sont qualitatives. On profite de cette occasion pour revoir tout un tas de choses sur le modèle linéaire en général (Exercice 1), et pour s'offrir une petite croisière avec Leonardo DiCaprio et Kate Winslet (Exercice 2).

Exercice 1. On se place dans le cas où on souhaite modéliser une variable quantitative y à l'aide d'une seule variable qualitative x pouvant prendre les modalités m_1, ..., m_p. On introduit donc le modèle, pour $1 \leq i \leq n$,

$$y_i = b_0 + \sum_{j=1}^{p-1} b_j \mathbb{1}_{x_i = m_j} + \varepsilon_i$$

avec les hypothèses **(H0)-(H5)** et les notations matricielles habituelles.

1) Ecrire le modèle sous forme matricielle.
2) En déduire une façon très simple de calculer les estimateurs des moindres carrés $\hat{b}_0$, ..., $\hat{b}_{p-1}$. On pourra introduire la notation

$$n_j = \text{ "le nombre d'individus } i \text{ tels que } x_i = m_j\text{".}$$

3) Réécrire l'équation d'analyse de la variance dans ce cas. Quelle nouvelle interprétation peut-on donner aux différents termes de cette équation ?
4) Comment tester l'hypothèse $\mathbf{H}_0$: la variable x n'a aucune influence sur la variable y ? On écrira complètement la statistique de test utilisée 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*. Cette base donne pour chaque individu présent sur le Titanic lors de son naufrage son âge (enfant ou adulte), sa classe en tant que passager (1ère, 2ème ou 3ème classe, ou équipage), son sexe (homme ou femme) et la variable d'intérêt, qui est $y_i = 1$ si la personne a pu survivre au naufrage (en obtenant une place dans un canot de sauvetage), ou $y_i = 0$ sinon. On va donc appliquer le modèle ANOVA, pour le moment en ne tenant compte que de l'âge et du sexe :

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

On précise les information suivantes $n = 2801$: $n_{\text{femme}} = 425$, $n_{\text{fille}} = 45$, $n_{\text{garçon}} = 64$ dont on peut déduire $n_{\text{homme}} = n - n_{\text{femme}} - n_{\text{fille}} - n_{\text{garçon}} = 2267$.

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

$$\hat{y}_i = 0.22 + 0.24 \mathbb{1}_{i=\text{ garçon}} + 0.41 \mathbb{1}_{i=\text{ fille}} + 0.53 \mathbb{1}_{i=\text{ femme}}.$$

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

3) Celà est-il cohérent avec le film ?

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

$$y_i = c_0 + c_1 \mathbb{1}_{i=\text{ enfant}} + \varepsilon_i,$$

comment peut-on estimer les coefficients c_0 et c_1 à partir des résultats des questions précédentes ?