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

TD N° 4 : Analyse de la variance et Variables qualitatives

Ce TD a pour but de réviser l'équation d'analyse de la variance (Exercice 1), et d'utiliser son application principale, à savoir la définition du coefficient R^2, qui est un indicateur de la qualité du modèle. Dans l'Exercice 2, on utilise le R^2 pour choisir un modèle reliant le prix de la communication téléphonique au temps de communication. Enfin, dans l'Exercice 3 on donne un exemple d'utilisation de variables qualitatives.

Exercice 1. (Tiré du livre de R. Bourbonnais) Un économiste, Oscar, étudie deux variables x et y, il propose le modèle suivant :

$$y_i = ax_i + b + \varepsilon_i,$$

$i \in \{1, ..., n = 20\}$, avec les ε_i i.i.d. $\mathcal{N}(0, \sigma^2)$. Il estime les paramètres a et b, et obtient

$$\hat{y}_i = 1,251x_i - 32,95.$$

Son logiciel l'informe aussi que $R^2 = 0.23$ et $\hat{\sigma}^2 \simeq 10.66$.

1) Ecrire l'équation d'analyse de la variance, en précisant la signification et la définition de toutes les quantités utilisées.
2) Retrouver, à partir des résultats numériques d'Oscar, la valeur de chacune de ces quantités.

Exercice 2. On souhaite modéliser la consommation d'un produit en fonction de son prix. En suivant la théorie microéconomique classique, si les consommateurs ont le choix entre k produits substituables, $j = 1, ..., j = k$, la demande en chacun de ces biens sera une fonction de chacun des prix $p_1, ..., p_k$. Ici, pour simplifier les choses, on se place dans le cas d'un produit non substituable (par exemple, un abonnement de téléphone portable). On suppose donc que la consommation totale C dans un pays est fonction du prix moyen de l'heure de commmunication p, $C = f(p)$. On observe C_i et p_i dans 20 pays européens, $i = 1, ..., n$ avec $n = 20$.

1) Une première économètre, Alice, part de l'idée qu'une variation du prix Δp se traduira toujours par la même variation de la consommation ΔC, autrement dit :

$$\frac{\Delta p}{\Delta C} \simeq \text{cste}.$$

Proposer un modèle compatible avec cette hypothèse.

2) Un autre économètre, Bob, propose l'idée que c'est l'elasticité de la demande aux prix qui est constante (i.e. le pourcentage de la variation de C est proportionnel au pourcentage de la variation de p), autrement dit,

$$\frac{C}{\Delta C}\frac{\Delta p}{p} \simeq \text{cste}.$$

Proposer un modèle compatible avec cette hypothèse.

3) On se décide finalement à estimer 4 modèles différents, on résume ici les résultats obtenus :

Modèle A	$\widehat{C}_i$	$= 120$	$-40p_i$	$R^2 = 0.234$
Modèle B	$\widehat{\log C}_i$	$= 8$	$-2.3p_i$	$R^2 = 0.401$
Modèle C	$\widehat{C}_i$	$= 132$	$-87 \log p_i$	$R^2 = 0.229$
Modèle D	$\widehat{\log C}_i$	$= 9.1$	$-3.4 \log p_i$	$R^2 = 0.312$

Quel modèle (A, B, C ou D) correspond à la proposition d'Alice ? A celle de Bob ?

4) Au vu de ces résultats, quel modèle semble être le plus adapté au données ? Quel lien entre ΔC et Δp celà représente-t'il ?

Exercice 3. Les trois économètres, Alice, Bob et Oscar souhaitent modéliser le salaire d'un individu en fonction de sa formation et de son ancienneté. Ils observent donc, sur un échantillon de n individus i, les variables sal_i (salaire mensuel brut en euros), anc_i (l'ancienneté de l'individu dans son entreprise actuelle, en mois), et diplome_i, le dernier diplôme obtenu par l'individu, variable pouvant prendre trois modalités dans cette enquête : $\text{diplome}_i = \text{infbac}$ si l'individu n'a pas obtenu le bac, $\text{diplome}_i = \text{bac}$ si l'individu a obtenu le bac et aucun diplôme ensuite, et enfin $\text{diplome}_i = \text{sup}$ si l'individu est diplômé de l'enseignement supérieur.

1) Oscar propose le modèle suivant :

$$\text{sal}_i = a_0 + a_1\text{anc}_i + a_2\mathbb{1}(\text{diplome}_i = \text{infbac}) + a_3\mathbb{1}(\text{diplome}_i = \text{bac}) + a_4\mathbb{1}(\text{diplome}_i = \text{sup}) + \varepsilon_i$$

avec les hypothèses usuelles sur ε_i. Que pensez-vous de ce modèle ?

2) Alice propose le modèle suivant :

$$\text{sal}_i = b_0 + b_1\text{anc}_i + b_2\mathbb{1}(\text{diplome}_i = \text{bac}) + b_3\mathbb{1}(\text{diplome}_i = \text{sup}) + \varepsilon_i.$$

Elle l'estime et obtient :

$$\widehat{\text{sal}}_i = 1024 + 62\text{anc}_i + 389\mathbb{1}(\text{diplome}_i = \text{bac}) + 875\mathbb{1}(\text{diplome}_i = \text{sup}).$$

Commenter ces résultats.

3) Bob propose le modèle suivant :

$$\text{sal}_i = c_0 + c_1\text{anc}_i + c_2\mathbb{1}(\text{diplome}_i = \text{infbac}) + c_3\mathbb{1}(\text{diplome}_i = \text{sup}) + \varepsilon_i.$$

Ce modèle est-il réellement différent de celui d'Alice ? Si oui, pourquoi ? Sinon, peut-on retrouver les coefficients de ce modèle à partir de ceux du modèle d'Alice ?