Ecole Normale Supérieure
FIMFA – Statistique 2012-2013

Feuille d'exercices n°6.

Exercice 1. Soit R le revenu d'un individu. On veut analyser la différence de revenu selon le sexe. Pour ceci, on considère le modèle

$$R = \mu + \beta s + \varepsilon$$

où s est une variable qui vaut 1 si l'individu est un homme et 0 si c'est une femme, et ε est une variable gaussienne centrée de variance connue σ^2. Les paramètres (μ, β) sont inconnus. On observe des réalisations $(R_i, s_i)_{1\leq i\leq n}$ i.i.d. suivant le modèle précédent. Dans la suite, $(s_1, ..., s_n)$ est considérée comme déterministe.

1. Comment s'interprète le coefficient β ? Que se passe-t-il si $\beta = 0$, $\beta < 0$?
2. Donner une condition sur $(s_1, ..., s_n)$ pour que l'estimateur du maximum de vraisemblance de (μ, β) soit bien défini et préciser alors son expression et sa loi.
3. Proposer un intervalle de confiance pour β. En déduire un test $H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$ de niveau α.
4. On veut tester $H_0 : \beta \leq 0$ contre $H_1 : \beta > 0$. Proposer un test de niveau α et le comparer au test de la question précédente.

Exercice 2. On considère le modèle linéaire

$$y_i = \alpha w_i + \beta z_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

où $\underline{w} = (w_1, \ldots, w_n)^T$ et $\underline{z} = (z_1, \ldots, z_n)^T$ sont deux vecteurs **déterministes** de $\mathbb{R}^n$ et $\underline{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ est un vecteur aléatoire gaussien à composantes i.i.d. de loi normale $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$. On note $\underline{y} = (y_1, \ldots, y_n)^T$.

Les paramètres α, $\beta \in \mathbb{R}$ et $\sigma^2 > 0$ sont **inconnus**. Le but de l'exercice consiste à comparer la qualité de l'estimation de ces paramètres lorsque les vecteurs $\underline{w}$ et $\underline{z}$ sont orthogonaux et lorsque ils ne le sont pas.

Dans tout l'exercice, on suppose que

$$\langle \underline{w}, \underline{z}\rangle^2 \neq \|\underline{w}\|^2 \|\underline{z}\|^2.$$

1. Montrer que l'estimateur des moindres carrés $\hat{\underline{b}}_n = (\hat{\alpha}_n, \hat{\beta}_n)^T$ du paramètre $\underline{b} = (\alpha, \beta)^T$ est donné par

$$\hat{\underline{b}}_n = \frac{1}{\|\underline{w}\|^2 \|\underline{z}\|^2 - \langle \underline{w}, \underline{z}\rangle^2} \begin{pmatrix} \|\underline{z}\|^2 \langle \underline{w}, \underline{y}\rangle - \langle \underline{w}, \underline{z}\rangle\langle \underline{z}, \underline{y}\rangle \\ \|\underline{w}\|^2 \langle \underline{z}, \underline{y}\rangle - \langle \underline{w}, \underline{z}\rangle\langle \underline{w}, \underline{y}\rangle \end{pmatrix}$$

et préciser sa loi.

2. Montrer que si $\underline{w}$ et $\underline{z}$ sont orthogonaux, alors les deux estimateurs $\hat{\alpha}_n$ et $\hat{\beta}_n$ de α et β sont indépendants.

   A partir de maintenant, on abandonne l'hypothèse d'orthogonalité de $\underline{w}$ et $\underline{z}$, et on note $\theta \in]0, \pi[$ l'angle formé par ces deux vecteurs. **On suppose, pour simplifier les calculs, que $\|\underline{w}\| = \|\underline{z}\| = 1$.**
3. Préciser la loi de la variable aléatoire
$$\frac{|\sin\theta|(\hat{\beta}_n - \beta)}{\sqrt{\hat{\sigma}_n^2}},$$
   où $\hat{\sigma}_n^2$ est l'estimateur habituel (sans biais) de σ^2 dans le modèle de régression.
4. Construire un intervalle de confiance (exact et symétrique) de niveau $1 - \alpha$ pour β.
5. Soit $L^2(\theta)$ la variable aléatoire égale au carré de la longueur de l'intervalle de confiance trouvé dans la question précédente. Etudier la fonction
$$g(\theta) = \mathbb{E}\left[L^2(\theta)\right], \quad \theta \in]0, \pi[$$
   et interpréter les résultats obtenus à la lumière du comportement de g à la frontière de son domaine de définition.

Exercice 3. Nous disposons pour n entreprises, $i = 1, ..., n$, des valeurs du capital K_i, de la valeur ajoutée VA_i et de la quantité de travail L_i. Nous supposons que la fonction de production F de ces entreprises est du type Cobb-Douglas, c-à-d
$$VA_i = F(L_i, K_i) = e^{\alpha} L_i^{\lambda} K_i^{\gamma}.$$
Les paramètres α, λ, γ sont inconnus.

1. Ecrire un modèle de régression linéaire associé lorsque les erreurs sont supposées gaussiennes indépendantes, centrées et de variance commune σ^2 (inconnue). On pourra écrire ce modèle sous la forme
$$Y = X\beta + E.$$
   Rappeler alors l'expression matricielle de l'estimateur des moindre carrés ordinaires $\hat{\beta}$ de β (on suppose X injective, i.e. de rang 3) et préciser sa matrice de variance-covariance.

   Dans la suite, on utilisera les données suivantes : pour 1658 entreprises, on a obtenu à l'aide de l'estimateur MCO de β, la formule de régression
$$\log VA_i = 3.136 + 0.738 \log L_i + 0.282 \log K_i$$
   et SCR(=somme des carrés des résidus)= 148.27. Nous donnons également
$$(X'X)^{-1} = \begin{pmatrix} 0.0288 & 0.0012 & -0.0034 \\ 0.0012 & 0.0016 & -0.0010 \\ -0.0034 & -0.0010 & 0.0009 \end{pmatrix}$$

2. Donner un estimateur $\hat{\sigma}^2$ sans biais de σ^2 et un estimateur sans biais de $\mathrm{Var}(\hat{\beta})$.
3. Donner un intervalle de confiance de niveau 95% pour les paramètres α et λ.
4. Tester l'hypothèse $H_0 : \gamma = 0$ contre $H_1 : \gamma > 0$.
5. On souhaite à présent tester l'hypothèse selon laquelle la fonction de production F est à rendements d'échelle constants, c-à-d

$$F(\mu L, \mu K) = \mu F(L, K), \quad \forall \mu \in \mathbb{R}_+^*.$$

   Proposer un test de niveau 5% pour H_0 : "rendements d'échelle constants" contre H_1 : "rendements d'échelle croissants."
6. Nous nous intéressons à la fonction de production dans trois secteurs

   T07 : minerais et métaux ferreux, première transformation de l'acier
   T15B : fabrication de biens d'équipement ménager
   T18 : industries textiles et habillement.

   Ces trois secteurs sont-ils régis par la même fonction de production ? Pour le savoir, nous avons estimé par la méthode des moindres carrés ordinaires les paramètres de l'équation $\log VA = \alpha + \lambda \log L + \gamma \log K$ séparément sur chacun des trois secteurs, puis sur l'ensemble des entreprises. Nous avons obtenus les résultats suivants :

   T07 : $n = 40$, SCR= 2.57
   T15B : $n = 24$, SCR= 1.17
   T18 : $n = 238$, SCR= 17.8
   Ensemble : $n = 302$, SCR= 41.12

   Pour un niveau de 5%, tester l'hypothèse selon laquelle la fonction de production est la même dans les trois secteurs.