DM3 TISD - M1 Pro
à rendre pour le Vendredi 18 décembre 2009

L'exercice 2 est théorique. Pour l'exercice 1, la table **SAS** est téléchargeable depuis
`http://math.univ-lille1.fr/~tran/enseignements.html`
Les codes sont à envoyer au plus tard le 17 décembre soir à : `chi.tran@univ-lille1.fr`.
Merci de rendre une copie et un code par binôme.

Exercice 1 (Déterminants du salaire (d'après l'examen de janvier 2008))
Nous analysons les données d'une enquête auprès des ménages effectuée par le U.S. Census Bureau en mai 1985. Le fichier étudié, `cps85.sas7bdat`, contient 532 individus, pour lesquels sont renseignées les variables suivantes : le **gain** qui est le log du salaire horaire en dollars (`LNWAGE`), l'âge (`AGE`), le nombre d'années d'étude (`ED`), le nombre d'années d'expérience (`EX=AGE-(ED+6)`, qui est le nombre d'années à partir de la fin des études en admettant que celles-ci commencent à 6 ans), le carré du nombre d'années d'expérience (`EXSQ=EX*EX`), le produit du nombre d'années d'expérience et du nombre d'années d'études (`EDEX=EX*ED`), les indicatrices `FE`, `NONWH`, `HISP` et `UNION` qui valent 1 si l'individu est respectivement une femme, ni Blanc ni Hispanique, Hispanique, syndiqué.

1. Faire quelques statistiques descriptives pour les variables `LNWAGE`, `AGE`, `ED`, `EX` (traitées séparément et calcul de corrélations) et résumer en une dizaine de ligne *au plus*.

Partie A : impact de l'expérience et de l'éducation sur le salaire

Nous considérons le modèle linéaire suivant, où ε est un bruit :

$$\texttt{LNWAGE} = \alpha + \alpha_F\,\texttt{FE} + \alpha_U\,\texttt{UNION} + \alpha_N\,\texttt{NONWH} + \alpha_H\,\texttt{HISP} + \beta_1\,\texttt{ED} + \beta_2\,\texttt{EX} + \beta_3\,\texttt{EXSQ} + \varepsilon \quad (1)$$

2. Estimer les paramètres de ce modèle par MCO sous **SAS** avec une `proc reg`. Récupérer avec la commande `output` dans une table de sortie les valeurs prédites de `LNWAGE` et les résidus estimés $\widehat{\varepsilon}$.

3. Donner le coefficient de détermination R^2 de la régression. Commentez.

4. On étudie les résidus ε à partir des résidus estimés $\widehat{\varepsilon}$. Faire, avec la `proc univariate` et la commande `normaltest`, un test de normalité des résidus de la régression. Dessiner un histogramme et un QQ-plot pour corroborer ce résultat. Pourquoi tester la normalité de la variable `epsilon` est-il important ?

5. Calculer avec **SAS** les intervalles de confiance à 95% des estimateurs MCO de β_1, β_2 et β_3. On pourra utiliser la `proc reg` et l'option `clb`.

6. A partir des sorties **SAS**, dire si les coefficients β_1, β_2 et β_3 sont significatifs. Commenter les signes de ces coefficients, lorsque cela est pertinent.

7. Pour quel niveau d'expérience le salaire est-il maximisé ?

8. Calculer le salaire horaire maximum prédit par le modèle, en dollars, pour un homme Blanc non syndiqué et ayant 14 années d'étude.

Partie B : discrimination salariale Hommes/Femmes

Nous revenons au modèle (1) et nous intéressons à la différence de salaires entre hommes et femmes.

9. Interpréter le coefficient α_F qu'on lit dans les sorties **SAS** et son exponentielle. Tester l'hypothèse selon laquelle la différence de salaire entre un homme et une femme est nulle, toutes les autres variables étant égales.

10. Faire un test de Chow pour tester H_0 *: les paramètres du modèle (1) pour la sous-population des hommes et sur la sous-population des femmes sont les mêmes.* Vérifier la validité du test de Chow. Conclure.

Exercice 2 (EMV et régression logistique (d'après le sujet de janvier 2009))
On considère un échantillon de couples i.i.d. $(Y_i, X_i)_{i\in[\![1,n]\!]}$, où $Y_i \in \{0,1\}$ et $X_i \in \mathbb{R}$ est une variable explicative. On suppose que :

$$Y_i = \mathbf{1}_{Y_i^*>0} \quad \text{avec} \quad Y_i^* = \alpha + \beta X_i + \varepsilon_i, \tag{2}$$

où les $(\varepsilon_i)_{i\in[\![1,n]\!]}$ sont des résidus i.i.d. , indépendants des X_i, centrés et à densité. On notera f leur densité (supposée continue) et F leur fonction de répartition. On supposera de plus que leur distribution est symétrique, ce qui se traduit par : f paire ou de façon équivalente $F(u) = 1 - F(-u)$.

1. Quelle est la loi de Y_i conditionnellement à $X_i = x_i$? Montrer que $\mathbb{P}(Y_i = 1 \,|\, X_i = x_i) = F(\alpha + \beta\, x_i)$. En déduire l'espérance et la variance conditionnelles de Y_i.

2. On pose $p_i = F(\alpha + \beta x_i)$. Ecrire la vraisemblance des $Y_1, \cdots Y_n$ conditionnellement aux $X_1, \cdots X_n$.

3. Montrer que :

$$\frac{\partial \ln\, p_i}{\partial \alpha} = \frac{f(\alpha+\beta x_i)}{F(\alpha+\beta x_i)}$$

faire de même pour la dérivée partielle par rapport à β et pour les dérivées partielles de $\ln(1-p_i)$.

4. En déduire que l'estimateur du maximum de vraisemblance $(\widehat{\alpha}, \widehat{\beta})$ satisfait les équations suivantes :

$$\begin{aligned}
&\sum_{i=1}^{n} Y_i \frac{f(\widehat{\alpha}+\widehat{\beta}X_i)}{F(\widehat{\alpha}+\widehat{\beta}X_i)} - (1-Y_i)\frac{f(\widehat{\alpha}+\widehat{\beta}X_i)}{1-F(\widehat{\alpha}+\widehat{\beta}X_i)} = 0 \\
&\sum_{i=1}^{n} X_i Y_i \frac{f(\widehat{\alpha}+\widehat{\beta}X_i)}{F(\widehat{\alpha}+\widehat{\beta}X_i)} - X_i(1-Y_i)\frac{f(\widehat{\alpha}+\widehat{\beta}X_i)}{1-F(\widehat{\alpha}+\widehat{\beta}X_i)} = 0.
\end{aligned} \tag{3}$$

Dans la suite du problème, on s'intéresse à la régression logistique, obtenue pour des résidus ε_i tels que

$$F(x) = \frac{1}{1+e^{-x}}. \tag{4}$$

5. Montrer que cette distribution est symétrique.

6. Réécrire dans le cas de la régression logistique les dérivées partielles de la question **3** en fonction de F la fonction de répartition des ε_i.

7. Citer un algorithme qu'on pourrait utiliser pour maximiser numériquement la log-vraisemblance. Pourquoi cet algorithme fonctionnerait-il bien ici ?

8. Montrer que dans le cas de la régression logistique, le système (3) se ré-écrit :

$$\begin{aligned} &\sum_{i=1}^{n} \big(Y_i - F(\widehat{\alpha} + \widehat{\beta} X_i)\big) = 0 \\ &\sum_{i=1}^{n} \big(Y_i - F(\widehat{\alpha} + \widehat{\beta} X_i)\big) X_i = 0. \end{aligned} \tag{5}$$

9. Donner l'information de Fisher du modèle sous forme d'une espérance et en faisant intervenir la densité f des résidus.

Jusqu'à la fin du problème, on se place dans le cas d'un modèle logistique où X est également une variable dichotomique, c'est-à-dire qui prend les valeurs 0 ou 1. On se donne le tableau de contingence suivant :

		Y	
		0	1
X	0	$n_{00} = 26$	$n_{01} = 26$
	1	$n_{10} = 16$	$n_{11} = 32$

TAB. 1 – Tableau de contingence de X et Y

10. Réécrire les équations (5) avec uniquement n_{00}, n_{01}, n_{10}, n_{11} (sans les X_i ou Y_i).

11. Résoudre les équations obtenues à la question **10** pour obtenir l'estimateur du maximum de vraisemblance (Indication : on pourra poser $a = F(\widehat{\alpha})$ et $b = F(\widehat{\alpha} + \widehat{\beta})$).