

DM 3 - TISD - Master 1 Pro  
A me remettre dans mon casier (RdC Bat M2) pour le 15/12

**Tran Viet Chi**, [chi.tran@math.univ-lille1.fr](mailto:chi.tran@math.univ-lille1.fr), bureau 316 (bâtiment M3).

*Le nombre de décimales pour les résultats numériques est laissé à votre appréciation.*

**La programmation est à réaliser sous SAS.**

*Dans tous les tests demandés, le seuil sera 5%. Pour les tests lus dans les sorties SAS, on donnera la valeur numérique de la statistique de test et la p-valeur associée.*

*La base de données est accessible sur <http://math.univ-lille1.fr/~tran/enseignements.html>.*

**Exercice 1 (Déterminants du salaire et discrimination salariale)**  
(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.

**Partie A : statistiques descriptives**

1. Pour chacune des variables `NONWH`, `HISP`, `FE`, `UNION`, dessiner le diagramme secteur et donner les fréquences empiriques de leurs différentes modalités.
2. Croiser les variables `NONWH` et `UNION` avec la `proc freq`. Ces variables sont-elles indépendantes ?
3. 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 B : impact de l'expérience et de l'éducation sur le salaire**

Nous considérons le modèle linéaire suivant, où  $\varepsilon$  est un bruit :

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

4. 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  $\hat{\varepsilon}$ .

5. Donner le coefficient de détermination  $R^2$  de la régression. Commentez.
6. On étudie les résidus  $\varepsilon$  à partir des résidus estimés  $\hat{\varepsilon}$ .
  - 6.1. Comment montrer avec **SAS** que les résidus  $\varepsilon$  sont centrés ?
  - 6.2. Faire, avec la `proc univariate` et la commande `normaltest`, un test de normalité des résidus de la régression. Commenter. Dessiner un histogramme et un QQ-plot pour corroborer ce résultat.
  - 6.3. Pourquoi tester la normalité de la variable `epsilon` est-il important ?
7. Calculer avec **SAS** les intervalles de confiance à 95% des estimateurs MCO de  $\beta_1$ ,  $\beta_2$  et  $\beta_3$ . On pourra utiliser la `proc reg` et l'option `clb`.
8. A partir des sorties **SAS**, dire si les coefficients  $\beta_1$ ,  $\beta_2$  et  $\beta_3$  sont significatifs. Commenter les signes de ces coefficients, lorsque cela est pertinent.
9. Pour quel niveau d'expérience le salaire est-il maximisé ?
10. Calculer le salaire horaire maximum prédit par le modèle, en dollars, pour un homme Blanc non syndiqué et ayant 14 années d'étude.

La théorie économique du Capital Humain nous dit que si les capacités individuelles sont corrélées avec le niveau d'étude d'une part et avec la rémunération d'autre part, alors le salaire devrait croître plus vite avec l'expérience pour les plus diplômés. Une manière d'incorporer un tel effet est de rajouter au modèle (1) la variable `EDEX` dont on appellera  $\beta_4$  le coefficient. L'effet de l'expérience sur le log du salaire dépend alors de l'expérience *et* du niveau d'étude :

$$\partial \text{LNWAGE} / \partial \text{EX} = \beta_2 + 2\beta_3 \text{EX} + \beta_4 \text{ED}.$$

11. Réaliser la régression avec la variable supplémentaire `EDEX`.
12. Quel test peut-on faire pour tester la validité de cette théorie économique ? Quelle est la conclusion ?

### Partie C : discrimination salariale Hommes/Femmes

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

13. Interpréter le coefficient  $\alpha_F$  qu'on lit dans les sorties **SAS** et son exponentielle.
14. Tester l'hypothèse selon laquelle la différence de salaire entre un homme et une femme est nulle, toutes les autres variables étant égales.
15. 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.