

DM 4 - M1 Maths-fi

A me remettre dans mon casier (RdC Bat M2) pour le
03/01/2010

Tran Viet Chi, chi.tran@math.univ-lille1.fr, bureau 316 (bâtiment M3).

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)

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.

Préliminaire. Explorer la base de données avec quelques statistiques descriptives simples. Faire un résumé en quelques lignes.

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

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

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

1. 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 $\hat{\varepsilon}$.
2. Donner le coefficient de détermination R^2 de la régression. Commentez.
3. On étudie les résidus ε à partir des résidus estimés $\hat{\varepsilon}$.
 - 3.1. Comment montrer avec **SAS** que les résidus ε sont centrés ?
 - 4.2. 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.
 - 4.3. 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.
9. Nous considérons (1) et nous intéressons à la différence de salaires entre hommes et femmes. Interpréter le coefficient α_F qu'on lit dans les sorties **SAS** et son exponentielle.
10. Tester l'hypothèse selon laquelle la différence de salaire entre un homme et une femme est nulle, toutes les autres variables étant égales.