Statistiques appliquées - L3 d'Economie
Devoir - Groupes 8, 9 & 12 - Corrigé

Marc SANGNIER - marc.sangnier@ens-cachan.fr

7 décembre 2007

Lien entre les années d'études et le salaire d'embauche

Section 1 - Modèle économique

$$y_i = b + ax_i \tag{1}$$

Dans l'équation (1), y_i représente le salaire d'embauche de l'individu $i$et x_i le nombre d'années d'études de l'individu i. a et b sont des nombres réels.

Question 1.1

La variable exogène (ou variable explicative) de ce modèle est x, le nombre d'années d'études.

Question 1.2

La variable endogène (ou variable à expliquer) de ce modèle est y, le salaire d'embauche.

Question 1.3

A priori, on peut penser que le réel a sera positif. En effet, on s'attend à ce que le nombre d'années d'études influence de façon positive le salaire auquel l'individu peut prétendre. On peut éventuellement interprèter b comme un seuil minimum en dessous duquel le salaire ne peut pas descendre (par exemple, comme le salaire mimimum légal).

Question 1.4

La théorie du capital humain part de l'idée que toute formation est un investissement qui permet au travailleur d'augmenter ses capacités, donc sa productivité et ainsi, ses revenus futurs. Une année de formation supplémentaire donne lieu, selon cette théorie, à un arbitrage entre le coût de la formation (les frais supportés, mais aussi le coût d'opportunité de ne pas travailler) et l'accroissement des revenus futurs attendu du fait de la formation. Dans le modèle présenté ici, on fait l'hypothèse implicite que le niveau de formation accroît le salaire d'embauche, un indicateur de l'ensemble des revenus futurs. C'est cette hypothèse que l'estimation économètrique va nous permettre de mesurer (et éventuellement de tester, mais pas dans ce devoir).

Section 2 - Les données

On dispose de $n = 40$ observations (fictives) du couple $(X;Y)$. X (respectivement Y) est la variable aléatoire représentant les années d'études (le salaire de l'individu). Les années sont comptabilisées en années complètées (les redoublements ne sont donc pas pris en compte). Les salaires sont donnés sous la forme d'indices, le plus faible ayant été choisi comme base de calcul, ils sont exprimés dans la même unité monétaire, ils sont donc comparables entre eux. Pour plus de réalité, on peut supposer que ces salaires représentent pour tous les individus la même fraction du salaire mensuel. Ces observations sont rassemblées dans les tableaux présentés en annexe.

Question 2.1

Un nombre d'années d'études égal à 12 représente le niveau du baccalauréat dans le système éducatif français (5 années d'école primaire + 4 années de collège + 3 années de lycée). De la même façon, un nombre d'années d'études égal à 17 représente un niveau d'étude équivalent à Bac+5.

Question 2.2

La moyenne empirique de la distribution des salaires est $\bar{Y} = \sum_{i=1}^{n} \frac{y_i}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{40}5710 = 142,75$

Question 2.3

La moyenne empirique de la la distribution des années d'études est $\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{40}613 = 15,325.$

Question 2.4

$cov(X;Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{X}\bar{Y} = \frac{1}{40}89395 - 142,75 * 15,325 = 47,23$

Question 2.5

Pour calculer le coeficient de corrélation linéaire du couple $(X;Y)$ il faut tout d'abord calculer les écarts-types des deux distributions.
La variance empirique des années d'étude est $\sigma_X^2 = \frac{1}{40}\sum_{i=1}^{n}(x_i - \bar{X})^2 = \frac{1}{40}216,775 = 5,419$
D'où l'écart-type $\sigma_X = \sqrt{\sigma_X^2} = \sqrt{5,419} = 2,328$
La variance empirique des salaires est $\sigma_Y^2 = \frac{1}{40}\sum_{i=1}^{n}(y_i - \bar{Y})^2 = \frac{1}{40}24947,5 = 623,687$
D'où l'écart-type $\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{623,687} = 24,973$
Soit $\rho(X;Y)$le coefficient de corrélation linéaire du couple :
$\rho(X;Y) = \frac{cov(X;Y)}{\sigma_X \sigma_Y} = \frac{47,23}{2,328*24,973} = 0,8124$

Section 3 - Modèle économètrique

$$y_i = b + ax_i + \varepsilon_i \tag{2}$$

Dans l'équation (2), y_i représente le salaire d'embauche de l'individu i, x_i le nombre d'années d'études de l'individu i et ε_i des caractèristiques inobservées (ou non explicitées) de l'individu i ayant un impact sur son salaire d'embauche. a et b sont des nombres réels.

Question 3.1

Le terme ε_i regroupe trois types d'erreurs :
- une erreur de spécification : la variable explicative n'est sans doute pas la seule à influencer la variable expliquée ;
- une erreur de mesure : les données peuvent être imparfaitement mesurées ;
- une erreur d'échantillonnage : les observations dépendent de l'échantillon étudié.

Question 3.2

Les hypothèses faites pour estimer ce modèle par la méthode des moindres carrés sont les suivantes :
- le modèle est linéaire en x_i ;
- les valeurs obsrevés de x_i le sont sans erreurs (la variable explicative n'est pas aléatoire) ;
- l'espérance mathématique de l'erreur est nulle : $\forall i, E(\varepsilon_i) = 0$ (en moyenne le modèle est bien spécifié) ;
- la variance de l'erreur est constante : $\forall i, V(\varepsilon_i) = \sigma^2$ (hypothèse d'homoscédasticité) ;
- les erreurs sont indépendantes entre elles, elles ne sont pas corrélées ;
- l'erreur est indépendante de la variable explicative.

Section 4 - Méthode des moindres carrés ordinaires

Vous allez maintenant procèder pas à pas à l'estimation de ce modèle. La méthode des moindres carrés ordinaires consiste à minimiser l'écart entre la valeur observée de y_i et sa valeur prédite le modèle. Pour une observation donnée de x_i, la valeur prédite est $\hat{y_i} = ax_i + b$. On désire donc déterminer a et b de façon à minimiser pour toutes les observations le terme $y_i - \hat{y_i}$. Pour ce faire, on définit une fonction de perte quadratique, c'est à dire la somme des écarts au carré. Soit S cette fonction que nous allons chercher à minimiser :

$$S = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{3}$$

Question 4.1

$$S = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 \iff S = \sum_{i=1}^{n}(b + ax_i + \varepsilon_i - ax_i - b)^2$$

$$\iff S = \sum_{i=1}^{n}\varepsilon_i^2$$

On remarque au passage que :

$$S = \sum_{i=1}^{n}(y_i - b - ax_i)^2$$

Question 4.2

Les conditions du premier ordre sont :

$$\frac{\partial S}{\partial a} = 0 \tag{4}$$

$$\frac{\partial S}{\partial b} = 0 \tag{5}$$

Question 4.3

$$\frac{\partial S}{\partial b} = 0 \iff -1\sum_{i=1}^{n}(y_i - b - ax_i) = 0 \tag{6}$$

$$\iff \sum_{i=1}^{n} y_i - nb - a\sum_{i=1}^{n} x_i = 0 \iff b = \frac{1}{n}\sum_{i=1}^{n} y_i - a\frac{1}{n}\sum_{i=1}^{n} x_i \iff b = \bar{Y} - a\bar{X}$$

$$\frac{\partial S}{\partial a} = 0 \iff -2\sum_{i=1}^{n} x_i(y_i - b - ax_i) = 0 \tag{7}$$

$$\iff \sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} x_i - a\sum_{i=1}^{n} x_i^2 = 0 \iff \sum_{i=1}^{n} x_i y_i - (\bar{Y} - a\bar{X})\sum_{i=1}^{n} x_i - a\sum_{i=1}^{n} x_i^2 = 0$$

$$\iff \frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\bar{Y} - a\bar{X})\frac{1}{n}\sum_{i=1}^{n} x_i - a\frac{1}{n}\sum_{i=1}^{n} x_i^2 = 0 \iff \frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\bar{Y} - a\bar{X})\bar{X} = a\frac{1}{n}\sum_{i=1}^{n} x_i^2$$

$$\iff \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{Y}\bar{X} = a\frac{1}{n}\sum_{i=1}^{n} x_i^2 - a\bar{X}\bar{X} \iff \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{Y}\bar{X} = a\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{X}\bar{X}\right)$$

On reconnaît à droite l'écriture de la variance : $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2$

$$\iff \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{Y}\bar{X} = a\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2 \iff \sum_{i=1}^{n} x_i y_i - n\bar{Y}\bar{X} = a\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$$\iff \sum_{i=1}^{n} x_i y_i - n\bar{Y}\bar{X} + n\bar{Y}\bar{X} - n\bar{Y}\bar{X} = a\sum_{i=1}^{n}(x_i - \bar{X})^2 \iff \sum_{i=1}^{n} x_i y_i - n\bar{Y}(\frac{1}{n}\sum_{i=1}^{n} x_i) + n\bar{Y}\bar{X} - n\bar{X}(\frac{1}{n}\sum_{i=1}^{n} y_i) = a\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$$\iff \sum_{i=1}^{n} x_i y_i - \bar{Y}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{Y}\bar{X} - \bar{X}\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n}(x_i - \bar{X})^2 \iff \sum_{i=1}^{n} x_i y_i - x_i\bar{Y} + \bar{Y}\bar{X} - y_i\bar{X} = a\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$$\iff \sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}) = a\sum_{i=1}^{n}(x_i - \bar{X})^2$$

Les solutions sont $\hat{a}$ et $\hat{b}$ tels que :

$$\hat{a} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2} \tag{8}$$

$$\hat{b} = \bar{Y} - \hat{a}\bar{X} \tag{9}$$

Question 4.4

On en déduit : $\hat{a} = 8,71$ et $\hat{b} = 9,18$

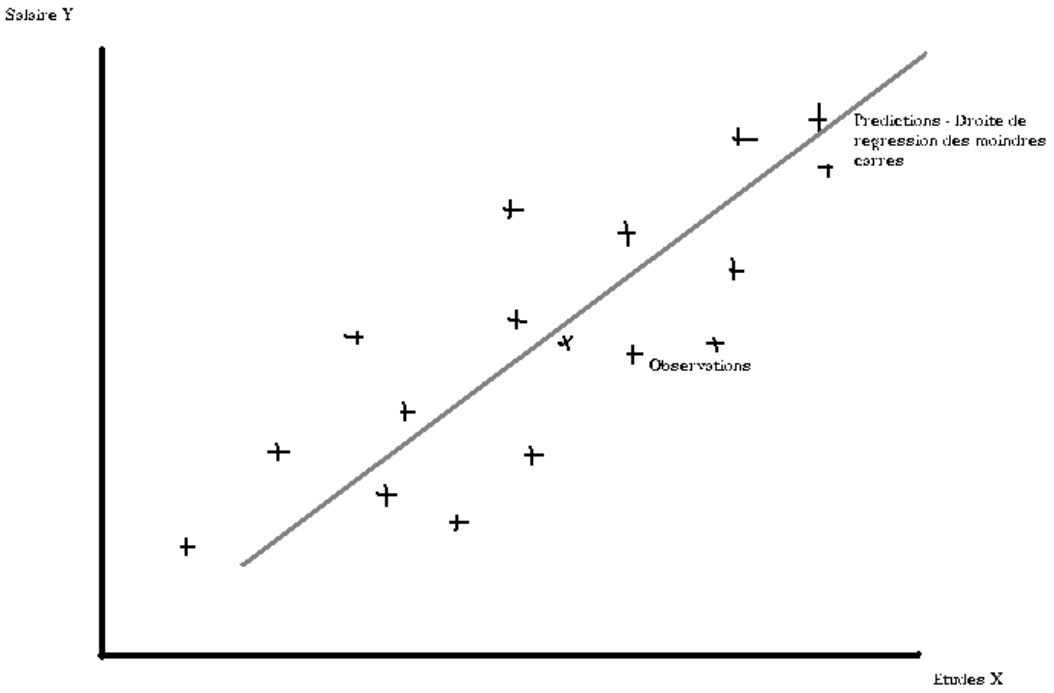
Question 4.5

Expression du salaire d'embauche théorique $y = \hat{a}x + \hat{b} \iff y = 8,71x + 9,18$

Question 4.6

$\bar{Y} = 142,75$ et $\bar{X} = 15,325$
$8,71 * 15,325 + 9,18 = 142,66 \approx 142,75 = \bar{Y}$ (erreurs d'arrondis)
La droite définie à la question précédente passe donc bien par le point $(\bar{X}; \bar{Y})$.

Question 4.7



Question 4.8

$$\frac{\partial y}{\partial x} = \hat{a} = 8,71$$

Question 4.9

De l'équation précèdente, on peut déduire que l'augmentation de la durée d'études d'une année entraîne une hausse de 8,71 points (compte tenu de la mesure des salaires choisis pour notre échantillon) du salaire d'embauche.

Section 5 - Analyse de la variance

L'équation fondamentale d'analyse de la variance est :

$$\sum_{i=1}^{n}(y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{y_i} - \hat{\bar{Y}})^2 + \sum_{i=1}^{n}\varepsilon_i^2 \tag{10}$$

De l'équation (10) on peut déduire l'expression du coefficient de détermination du modèle, R^2 :

$$R^2 = \frac{\sum(\hat{y_i} - \bar{Y})^2}{\sum(y_i - \bar{Y})^2} \tag{11}$$

Question 5.1

Le terme de gauche $\sum_{i=1}^{n}(y_i - \bar{Y})^2$ représente la variance des observations (la variance des salaires observés).
Le terme $\sum_{i=1}^{n}(\hat{y_i} - \hat{\bar{Y}})^2$ représente la variance du modèle (la variance des salaires prédits).
Le terme $\sum_{i=1}^{n}\varepsilon_i^2$ représente la variance non expliquée par le modèle (la variance des erreurs).
L'équation fondamentale de la variance exprime l'égalité entre la variance des observations et la somme de la variance des prédiction du modèle et de celle des perturbations.

Question 5.2

Compte tenu des équation précédentes, on peut écrire :

$$\sum_{i=1}^{n}(\hat{y_i} - \hat{\bar{Y}})^2 = \sum_{i=1}^{n}(y_i - \bar{Y})^2 - \sum_{i=1}^{n}\varepsilon_i^2$$

D'où, en injectant cette écriture dans celle du coefficient de détermination :

$$R^2 = \frac{\sum(y_i - \bar{Y})^2 - \sum\varepsilon_i^2}{\sum(y_i - \bar{Y})^2} = 1 - \frac{\sum\varepsilon_i^2}{\sum(y_i - \bar{Y})^2}$$

Le coefficient de détermination mesure donc la part de la variance des obsrevations expliquée par le modèle.

Question 5.3

$$R^2 = 1 - \frac{\sum\varepsilon_i^2}{\sum(y_i - \bar{Y})^2} = 1 - \frac{8482,199}{24947,5} \approx 0,66$$

Cela signifie que le modèle présenté ici explique 66% de la variance des observations. C'est à dire qu'on peut imputer les deux tiers de la dispersion des salaires d'embauche aux différences dans la durée des études.