UNIVERSITE PARIS 1 – PANTHEON SORBONNE
LICENCE DE SCIENCES ECONOMIQUES

STATISTIQUE APPLIQUEE
P. Sevestre

Fiche N°6

(avec corrigé)

L'objet de ce TD est de vous initier à la démarche et à quelques premiers rudiments de l'économétrie. Comme la statistique, l'économétrie recouvre un ensemble de méthodes dont l'objet est d'aider à répondre à toutes sortes de questions en analysant des données statistiques.

Par exemple, considérons la question suivante, déjà envisagée dans les TD précédents: Quel salaire pouvez-vous espérer avoir à l'embauche après avoir obtenu un master en économie à Paris 1? Pour y répondre, plusieurs options sont envisageables. La première consisterait à construire un échantillon ne comportant que des étudiants ayant obtenu un diplôme d'économie de niveau Bac+4 ou Bac+5 à Paris 1 et à calculer la moyenne des salaires obtenus (après avoir tenu compte convenablement des effets de l'inflation comme cela a été expliqué dans un TD précédent). Cette façon de procéder présente plusieurs inconvénients. Dans la pratique, constituer un tel échantillon peut se révéler coûteux, sauf pour l'Administration de l'Université qui dispose du fichier des anciens diplômés. Sinon, trouver plusieurs centaines de ces anciens étudiants au hasard dans la population supposerait d'interroger un nombre extraordinaire de personnes et coûterait donc très cher. D'autre part, cet échantillon serait inutilisable pour répondre au même type de questions pour des étudiants d'autres disciplines et/ou d'autres Universités. Enfin, faire une telle moyenne pour des diplômés étant entrés dans la vie active à des dates différentes et donc ayant été confrontés à des conjonctures différentes risque de conduire à une sur-estimation ou à une sous-estimation de votre salaire espéré. L'autre option consiste à utiliser un échantillon de personnes dont une partie a les caractéristiques recherchées, totalement ou en partie, et à recourir à un modèle économétrique pour prévoir le salaire des individus ayant les caractéristiques considérées.

Comme vous l'avez déjà vu, le salaire d'une personne au moment de son entrée dans la vie active dépend potentiellement de très nombreux éléments: notamment, ses caractéristiques personnelles (son diplôme, son Université, s'il est dynamique dans sa recherche d'emploi, s'il peut se faire aider par des relations...), celles de l'entreprise qui l'embauche (secteur d'activité, taille,...) et aussi de la situation sur le marché du travail. Appelons y le logarithme du salaire

et $z_1, z_2, ..., z_p$ l'ensemble des facteurs que vous avez identifiés comme susceptibles d'avoir une influence sur le salaire d'embauche d'un individu. On écrirait alors le modèle économétrique comme:

$$y_i = b_0 + b_1\ z_{1i} + b_2\ z_{2i} + ... + b_p\ z_{pi} + v_i \tag{1}$$

1) Pourquoi le modèle comporte-t-il une perturbation, notée ici v_i, ajoutée aux facteurs potentiels identifiés $z_1, z_2, ..., z_p$?

L'objet de cette perturbation est de rendre compte de toutes les variables qui peuvent avoir eu une influence sur le salaire d'embauche obtenu par l'individu i et auxquelles on peut ne pas penser quand, en tant qu'économètre, on écrit (on "spécifie") le modèle ; par exemple, le fait que, la veille de l'entretien d'embauche de l'individu i, l'entreprise ait signé un gros contrat et ait eu des besoins de main d'oeuvre supplémentaire met l'individu en position favorable pour négocier son salaire. C'est ce qu'on appelle un "choc favorable". A l'inverse, si l'entreprise est en difficulté et doit réduire ses coûts, cela affectera négativement le salaire qu'il peut obtenir. De même, si cet individu a déjà une offre d'une autre entreprise, il n'acceptera pas, à conditions de travail équivalentes, un salaire inférieur à celui qu'on lui a offert par ailleurs. On peut penser ainsi à une quasi-infinité d'événements qui peuvent, plus ou moins directement, avoir une influence sur le salaire d'embauche. Il est évidemment impossible d'en établir la liste exhaustive et ces facteurs explicatifs "oubliés" sont donc "intégrés" dans leur globalité dans la perturbation v_i.

2) Interprétez en termes littéraires les coefficients b_0, $b_1, ..., b_p$. Quel changement d'interprétation le fait que y soit le salaire et non plus son logarithme entraînerait-il? Comment s'interpréteraient les coefficients si dans le modèle $\ln(y)$ était expliqué par $\ln(x_1), \ln(x_2), ..., \ln(x_k)$?

Chaque coefficient b_j est égal à $\partial y/\partial x_j$. Par conséquent, le coefficient b_j mesure l'impact sur y d'une petite variation de x_j. Par exemple, si y est le salaire (en euros) et x_j l'ancienneté, mesurée en années, $\partial y/\partial x_j$ représente le supplément de salaire, en euros, que donne une année d'ancienneté supplémentaire.

Si la variable x_j est une variable indicatrice qui correspond à une variable qualitative prenant deux modalités (par exemple une indicatrice "Femme" pour la variable "Genre" dont les deux modalités sont "Homme" et "Femme"), et si le modèle comporte une constante, le coefficient mesure l'écart entre la moyenne de y pour la modalité représentée par cette variable x_j et la moyenne de y pour l'autre modalité. Par exemple, soit la variable "Femme" valant 1 si l'individu i est une femme et 0 sinon, le coefficient de cette variable "Femme" dans une

régression avec constante représente la différence entre le salaire d'une femme et celui d'un homme, toutes choses égales par ailleurs. S'il y a discrimination à l'encontre des femmes, on s'attend donc à avoir un coefficient négatif.

Plus généralement, si la variable indicatrice comprend n modalités (par exemple, le niveau de diplôme) dont l'une doit être supprimée pour permettre l'estimation du modèle (cf. ci-dessous), le coefficient de cette indicatrice dans une régression avec constante représente la différence entre le salaire des individus ayant le diplôme indiqué et celui des individus ayant le diplôme correspondant à la catégorie "absente" du modèle (cf. ce qui a été fait en cours).

Ici, puisque y est le logarithme du salaire, $b_j = \partial \ln(w)/\partial x_j = (\partial w/w)/\partial x_j$ représente l'effet, en termes relatifs, sur y d'une petite variation de x_j. C'est ce qu'on appelle une semi-élasticité. Par exemple, si y est le log du salaire et x_j l'ancienneté, mesurée en années, $\partial y/\partial x_j$ représente le supplément de salaire, en pourcentage, que donne une année d'ancienneté supplémentaire.

Enfin, si dans le modèle $\ln(w)$ est expliqué par $\ln(x_1), \ln(x_2), ..., \ln(x_k)$, les coefficients représentent l'élasticité de w par rapport à $x_1, x_2, ..., x_k$. Autrement dit, ces coefficients mesurent l'impact, en pourcentage, sur y d'une variation de x de 1%.

3) En fait, parmi tous les facteurs explicatifs $z_1, z_2, ..., z_p$ auxquels on peut penser, certains ne sont pas observables pour tous les individus, voire ne le sont pas du tout. Il n'est donc pas possible de les prendre en compte dans le modèle estimé si l'on ne dispose pas des données statistiques correspondantes. On est donc conduit finalement à estimer un modèle différent de (1) :

$$y_i = b_0 + b_1\ x_{1i} + b_2\ x_{2i} + ... + b_k\ x_{ki} + u_i \qquad (2)$$

où $x_{1i}, x_{2i}, ..., x_{ki}$ sont les facteurs effectivement observables (i.e. ceux pour lesquels on a des données) et où u_i inclut maintenant non seulement l'effet des facteurs potentiels auxquels vous n'avez pas pensé (v_i) mais aussi ceux dont vous savez qu'ils ont un impact sur le salaire mais pour lesquels vous n'avez pas les données statistiques nécessaires (en totalité ou en partie). On considère que l'ensemble de ces facteurs non explicitement pris en compte par le modèle peut être représenté par une variable aléatoire dont on doit préciser les caractéristiques.

Par exemple, le modèle de régression linéaire multiple est défini par l'équation (2) avec les hypothèses suivantes:

$H0$: La variable y est reliée aux variables $x_1, x_2, ..., x_k$ par le modèle $y_i = b_0 + b_1\ x_{1i} + b_2\ x_{2i} + ... + b_k\ x_{ki} + u_i,\ \forall i.$

$H1$: $E(u_i\ /\ x_{1i}, x_{2i}, ..., x_{ki}) = 0,\ \forall i.$

$H2a$: $V\ (u_i\ /\ x_{1i}, x_{2i}, ..., x_{ki}) = \sigma^2,\ \forall i.$

$H2b$: $(u_i, u_j \ / \ x_{1i}, x_{2i}, ..., x_{ki}, x_{1j}, x_{2j}, ..., x_{kj}) = 0, \ \forall i \neq j.$
$H3$: Les variables explicatives x_{jt} ne sont pas colinéaires.

Rappelez ce que signifient ces hypothèses.

L'hypothèse $H0$ stipule que le modèle linéaire que l'on souhaite estimer correspond bien à la forme de la relation (inconnue) qui existe entre la variable y et les variables $x_1, x_2, ..., x_k$. Supposons que la relation soit en réalité non linéaire; on aura alors une estimation biaisée de l'impact des variables x_j sur y.

L'hypothèse H1 stipule que les facteurs omis dans le modèle (u_i) sont sans corrélation avec ceux qui y sont pris en compte. Ceci permet notamment d'en déduire que

$$E(y_i \ / \ x_{1i}, x_{2i}, ..., x_{ki}) = b_0 + b_1 \ x_{1i} + b_2 \ x_{2i} + ... + b_p \ x_{ki} + E(u_i / x_{1i}, x_{2i}, ..., x_{ki})$$
$$= b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_p x_{ki}, \ \forall i..$$

Cette équation permet de dire que si on a des estimations $\widehat{b}_0, \widehat{b}_1, \widehat{b}_2, ...\widehat{b}_k$ sans biais des paramètres $b_0, b_1, b_2, ...b_k$, une prévision sans biais de $E(y)$ conditionnellement aux caractéristiques observables $x_{i1}, x_{i2}, ..., x_{ik}$, est donnée par $\widehat{y}_i = \widehat{b}_0 + \widehat{b}_1 x_{1i} + \widehat{b}_2 x_{2i} + ... + \widehat{b}_p x_{ki}$.

L'hypothèse H2a stipule que toutes les perturbations u_i ont la même variance. En fait, cette hypothèse, avec H1, revient à considérer que les caractéristiques des distributions de probabilité des perturbations sont identiques. L'hypothèse H2b stipule en outre que les perturbations u_i ne sont pas corrélées entre elles. Une hypothèse un peu plus forte qui induit cela est l'hypothèse "i.i.d." (indépendamment et identiquement distribuées): Les perturbations sont supposées tirées dans une même loi et sont indépendantes les unes des autres, ce qui permet notamment d'appliquer la loi des grands nombres et le théorème central-limite.

L'hypothèse H3 suppose que l'information apportée par chaque variable explicative du modèle n'est pas complètement redondante. Un contre-exemple "classique" est celui des variables indicatrices associées au classement des individus en groupe. Par exemple, si on crée une variable indicatrice "Homme" pour les individus qui sont des hommes (=1 si homme, 0 sinon) et "Femme" pour les individus qui sont des femmes (=1 si femme, 0 sinon), on voit bien que la somme de ces deux variables "Homme"+"Femme" vaut toujours 1 et est donc parfaitement colinéaire avec la variable constante du modèle (associée au coefficient b_0). En d'autres termes, on sait qu'un individu qui n'est pas un homme est une femme, et réciproquement. Il est donc redondant de mettre ces deux variables simultanément dans le modèle, dès lors que celui-ci comporte une constante.

4) Ecrire le modèle et ses hypothèses sous forme matricielle

Le modèle s'écrit, sous forme matricielle:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ 1 & x_{13} & & x_{k3} \\ \vdots & \vdots & \cdots & \\ 1 & x_{1N} & & x_{kN} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \\ b_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_N \end{pmatrix}.$$

Soit,

$$y = Xb + u$$

où:

- y est le vecteur $(N, 1)$ des observations sur la variable expliquée,
- X est la matrice $(N, k+1)$ des observations sur les variables explicatives (y compris la variable constante),
- b est le vecteur $(k+1, 1)$ des coefficients à estimer,
- u est le vecteur des perturbations aléatoires.

La transcription, sous forme matricielle, des hypothèses $H0$ à $H3$ est donnée par :

$H0$: Le modèle est donné par $y = Xb + u$.
$H1$: $E(u/X) = 0$.
$H2$: $V(u/X) = \sigma^2 I$.
$H3$: La matrice X est de rang $k + 1 < N$.

5) Montrer que l'estimateur des Moindres Carrés Ordinaires de b s'écrit:

$$\hat{b} = (X'X)^{-1}X'y$$

La minimisation de la somme des carrés des résidus

$$\underset{b_0, b_1, ..., b_k,}{Min} \sum_{i=1}^{N} (y_i - b_o - b_1 x_{1i} - ... - b_k x_{ki})^2 .$$

s'écrit matriciellement,

$$\underset{b}{Min} (y - Xb)'(y - Xb)$$

$$\Longleftrightarrow \underset{b}{Min} (y'y - 2b'X'y + b'X'Xb)$$

En notant S la quantité à minimiser, les conditions du premier ordre associées à ce programme s'écrivent:

$$\left.\frac{\partial S}{\partial b}\right|_{b=\hat{b}} = -2X'y + 2X'X\hat{b} = 0\,. \tag{3}$$

Ces conditions constituent un système de $k+1$ équations à $k+1$ inconnues $(\hat{b}_o, \hat{b}_1, ..., \hat{b}_k)$, appelées "équations normales". Ce système d'équations admet une solution unique qui correspond à un minimum si la condition du second ordre:

$$\left.\frac{\partial^2 S}{\partial b \partial b'}\right|_{b=\hat{b}} = 2X'X\,, \quad \text{matrice définie positive}$$

est satisfaite.

Cette condition est satisfaite puisque d'après l'hypothèse H3, la matrice X est de rang $k+1$. Par conséquent, la matrice $X'X$, de dimension $(k+1, k+1)$ est également de rang $k+1$ et est définie positive. En pré-multipliant les équations (3) par $(X'X)^{-1}$, on trouve $\hat{b}$ qui minimise la somme des carrés des résidus:

$$\hat{b} = (X'X)^{-1}X'y$$

Cet estimateur est appelé estimateur des Moindres Carrés Ordinaires de b dans le modèle (2).

6) Montrer que, compte tenu des hypothèses posées, $\hat{b}$ est un estimateur sans biais de b et que sa variance est égale à $\sigma^2\ E(X'X)^{-1}$.

On sait que

$$E(\hat{b}) = E_X[E(\hat{b}/X)]$$

Or,

$$\begin{aligned} E(\hat{b}/X) &= E[(X'X)^{-1}X'y\ /\ X] \\ &= (X'X)^{-1}X'E(y/X) \\ &= (X'X)^{-1}X'E(Xb + u\ /X) \\ &= b + (X'X)^{-1}X'E(u/X)] \\ &= E_X[b] \\ &= b \end{aligned}$$

D'autre part, $\quad V(\hat{b}) = \sigma^2 E(X'X)^{-1}$

En effet, on a

$$V(\hat{b}) = E_X[V(\hat{b}/X)] + V_X[E(\hat{b}/X)]$$

Or, $V_X[E(\hat{b}/X)] = V_X b = 0$ puisque b n'est pas aléatoire.

D'autre part,

$$E_X[V(\hat{b}/X)] = E_X[\sigma^2(XX)^{-1}] = \sigma^2 E(XX)^{-1}$$

En effet,

$$\begin{aligned} V(\hat{b}/X) &= E\{[(\hat{b} - E(\hat{b})][(\hat{b} - E(\hat{b})]'/X\} \\ &= E\{[(X'X)^{-1}X'u][(X'X)^{-1}X'u]'/X\} \\ &= E\{[(X'X)^{-1}X'uu'X(X'X)^{-1}/X\} \\ &= (X'X)^{-1}X'E(uu'/X)X(XX)^{-1} \\ &= (X'X)^{-1}X'(\sigma^2 I)X(XX)^{-1} \\ &= \sigma^2(XX)^{-1} \end{aligned}$$

Ainsi,

$$V(\hat{b}) = \sigma^2 E(X'X)^{-1}$$

7) Dans le tableau suivant, on a donné les estimations obtenues avec deux spécifications différentes du modèle. Qu'en concluez-vous?

La première observation à faire dans ces résultats, c'est que l'effet d'avoir un diplôme en économie est très différent selon le modèle que l'on estime. On va voir pourquoi dans la question suivante.

Par ailleurs, on peut utiliser les connaissances que vous avez déjà en statistique pour répondre à certaines questions. Par exemple, le fait d'être une femme est-il défavorable en termes de salaire à l'embauche?

De façon générale, une variable explicative est pertinente si son coefficient est significativement différent de zéro, i.e. si une variation de la variable induit effectivement une variation de y. Pour vérifier cela, on procède à des tests de significativité, i.e. on teste

$$H_0 : b_j = 0 \quad \text{contre} \quad H_1 : b_j \neq 0.$$

Le principe du test est exactement le même que celui vu pour la moyenne d'une loi normale avec variance inconnue. Ici, on suppose que les perturbations u_i sont i.i.d. et compte tenu du nombre élevé d'observations, on peut appliquer le théorème central -limite. Les coefficients estimés suivent (asymptotiquement) une loi normale. Par conséquent, la région critique s'écrit:

Variable	Modalité	Coefficient	Ecart-type	Coefficient	Ecart-type
Intercept		9,73	0,06	9,95	0,08
discipline	autres	0,02	0,10	0,03	0,08
discipline	eco_gestion	0,23	0,07	0,14	0,06
discipline	lettres	-0,04	0,13	-0,04	0,12
discipline	sciences	0,23	0,08	0,07	0,07
discipline	shs	Ref,	-	Ref.	-
niveau_bacplus2				-0,38	-5,87
niveau_maitrise				-0,15	-2,93
niveau_master2				0,00	-
diplome_univ				-0,16	0,00
origine_etrangère				-0,14	0,08
femme				0,08	0,06
diplome_paris1				-0,11	0,01
année_embauche	1990			0,84	0,30
année_embauche	1991			0,00	0,30
année_embauche	1992			0,33	0,30
année_embauche	1993			-0,19	0,21
année_embauche	1994			-0,56	0,30
année_embauche	1995			0,12	0,17
année_embauche	1996			-0,27	0,15
année_embauche	1997			-0,46	0,18
année_embauche	1998			0,19	0,11
année_embauche	1999			-0,06	0,11
année_embauche	2000			0,06	0,07
année_embauche	2001			-0,09	0,09
année_embauche	2002			0,01	0,08
année_embauche	2003			0,00	0,08
année_embauche	2004			0,03	0,06
année_embauche	2005			Ref.	-

$$\omega = \left\{ (y_t, X_t)_{i=1,...,N} / \frac{\left|\hat{b}_j - 0\right|}{\hat{\sigma}_{\hat{b}_j}} \geq N_{1-\alpha/2} \right\} .$$

Si l'inégalité qui figure dans cette expression est vérifiée par l'estimation $\hat{b}_j$, on considère que $\hat{b}_j$ est trop différent de la valeur considérée b_j^o =(ici, 0) pour pouvoir accepter l'hypothèse H_o. Dans le cas contraire, on accepte cette hypothèse.

Par conséquent, puisque $N_{1-\alpha/2}$, le fractile de la loi normale vaut 1.96 si l'on fait un test à 5¨%, tout coefficient tel que le rapport $\hat{b}_j/\hat{\sigma}_{\hat{b}_j}$ est supérieur à 1.96 en valeur absolue est significativement différent de zéro et la variable qui lui est associée a alors une influence significative sur y.

Dans le cas qui nous intéresse ici, on peut dire que, d'après ces estimations, le fait d'être une femme ne semble pas avoir d'influence significative sur le salaire à l'embauche. En effet, la statistique de test vaut $t = 0.08/0.06 = 1.33$ qui est inférieur à 1.96. D'après ces estimations, il n'y aurait pas de discrimination à l'encontre des femmes en matière de salaire à l'embauche. Mais les résultats vus en cours montrent qu'il y en a en termes d'accès à l'emploi.

8) Montrez que si on estime le modèle

$$y_i = b_0 + b_1\ x_{1i} + b_2\ x_{2i} + ... + b_p\ x_{ki} + u_i \tag{4}$$

alors que les facteurs explicatifs de y comportent également des variables $w_{i1}, w_{i2}, ..., w_{im}$, corrélées avec $x_{i1}, x_{i2}, ..., x_{ik}$, les estimations obtenues de b sont biaisées.

Le modèle estimé s'écrit matriciellement

$$y = Xb + u$$

et l'estimateur des MCO est donné par

$$\hat{b} = (X'X)^{-1}X'y$$

Mais puisque les facteurs explicatifs sont X et Z, on a, pour la relation entre y et ses variables explicatives:

$$H0 : y = X\ b + Z\ c + u$$

Par conséquent,

$$\begin{aligned} E(\hat{b}) &= E[b + (X'X)^{-1}X'(Zc + u)] \\ &= E_{X,Z}[b + E[(X'X)^{-1}X'(Zc + u)/X, Z]] \\ &= E_X[b + (X'X)^{-1}X'(Zc + E(u/X, Z))] \\ &= E_X[b + (X'X)^{-1}X'(Zc)] \\ &= b + E_X[(X'X)^{-1}X'(Zc)] \\ &\neq b \end{aligned}$$

Le fait de ne pas tenir compte de facteurs explicatifs qui sont corrélés avec ceux pris en compte (X)conduit à biaiser les estimations.

9) En supposant que les estimations fournies en colonne 3 sont sans biais, quel salaire ce modèle prévoit-il que vous obtiendrez avec votre master en économie obtenu à Paris 1?

Pour répondre à cette question, il faut considérer que la conjoncture quand vous serez recruté(e) en 2008 ou 2009 ne sera pas très différente de ce qu'elle est actuellement, ce qui semble acceptable (l'effet de la conjoncture sur les salaires est à peu près stable depuis 2003). Il faut ensuite associer une valeur 1 aux variables qui correspondent à vos caractéristiques et en déduire votre salaire "prévu". Comme je l'ai dit en cours, certains coefficients estimés sont un peu trop "bizarres" pour croire vraiment au résultat obtenu.