LE MODÈLE LINÉAIRE SIMPLE

DONNÉES, NUAGES DE POINTS

Il est fréquent, dans une étude élémentaire, de chercher à préciser une liaison éventuelle entre deux grandeurs pour lesquelles on dispose d'une série d'observations jointes. Ainsi :

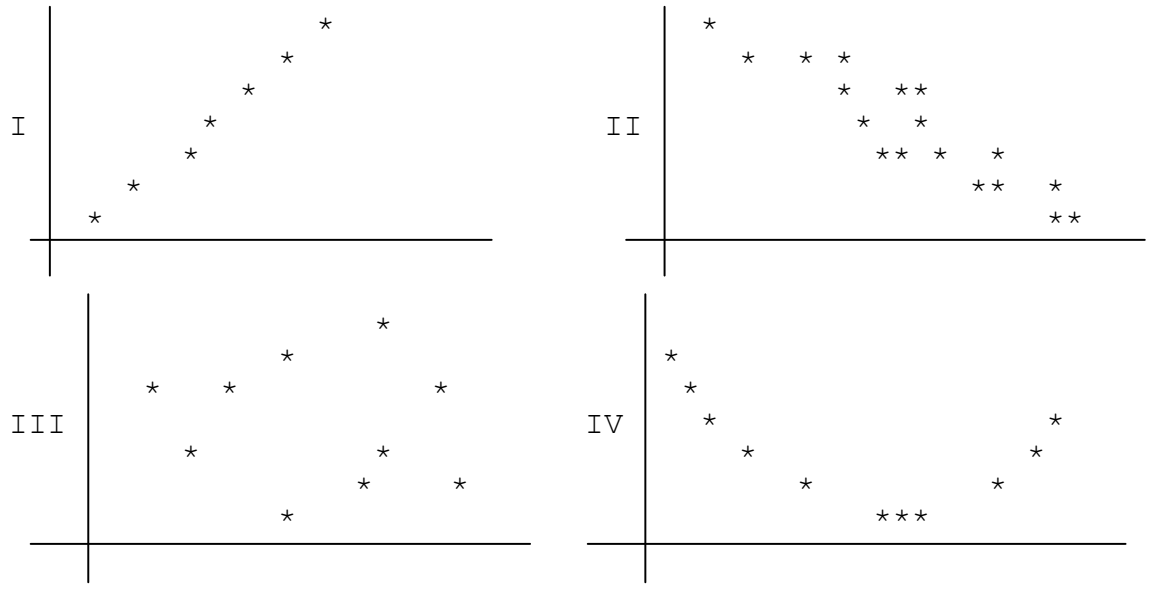
- la taille et le poids des éléments d'un groupe d'étudiants ;
- le salaire et le solde bancaire moyen des clients d'une banque ;
- le revenu annuel et le nombre moyen de voiture par habitant en France depuis 1948 ;
- le revenu par habitant et le taux d'analphabétisme en 1960 des pays d'Afrique ;
- le taux de criminalité et le nombre de policier par habitant en France depuis 1958 ;
- la production de beaujolais et le nombre de jours de soleil annuels depuis 1968.

Les données peuvent être des **séries temporelles**, ou **chronologiques** (exemples 3, 5 et 6), ou en **coupe transversale** (tels les exemples 2 et 4). On peut également considérer des **données croisées**, dites encore de **panel**, présentant les deux dimensions, tels le salaire annuel net, la consommation alimentaire, l'épargne et l'impôt sur le revenu relevés pour un groupe de 250 salariés de l'industrie suivis pendant 15 ans.

On note : (x_1,y_1), (x_2,y_2), ... , (x_N,y_N), la série de N observations des deux grandeurs : X et Y.

Des unités étant convenablement choisies, on peut représenter l'observation i de la série précédente par le point : (x_i,y_i), du plan X_xY. En figurant ainsi les N observations, on obtient le **nuage des points** associé au couple de séries statistiques.

Les nuages de points associés à des séries statistiques à deux caractères peuvent présenter différentes formes.



Les points du nuage I sont presque alignés, tandis que le nuage II laisse simplement apparaître une direction d'allongement privilégiée. Dans ces deux cas on dit que le nuage présente un caractère **linéaire**. Le nuage III ne manifeste pas de structure particulière, le nuage IV enfin semble se placer approximativement selon une courbe régulière qui n'est visiblement pas une droite.

La méthode présentée postule une liaison de nature linéaire, elle n'a donc d'intérêt que si le nuage est de l'un des deux premiers types (le dernier cas pouvant toutefois probablement s'y ramener, moyennant un changement de variable, ou une respécification du modèle...)

L'AJUSTEMENT LINÉAIRE, POINT DE VUE GÉOMÉTRIQUE

L'**ajustement linéaire** est la recherche de la « meilleure » droite résumant la structure du nuage.

Critère des moindres carrés, droite de régression de y en x

On cherche une droite, d'équation :

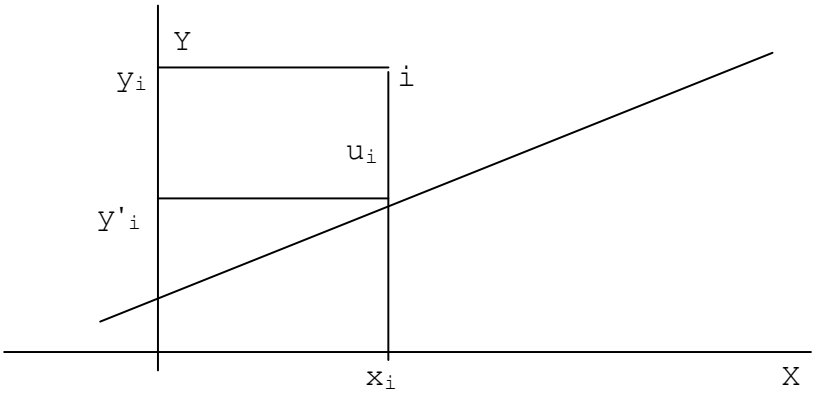
$$y = A.x + B$$

qui approche « au mieux » les données.

On considère en général que l'une des variables dépend de l'autre (par exemple la consommation dépend du revenu, ou l'investissement du taux d'intérêt); on suppose que y est la première, ou **variable à expliquer**, et x la seconde, ou **variable explicative**.

Notant donc : $y = A.x + B$, la droite retenue, on donne les définitions suivantes. Pour chaque observation i :

- x_i est la **valeur observée** de la variable explicative x ;
- y_i est la valeur observée de la variable à expliquer y ;
- $y'_i = A.x_i + B$ est la **valeur théorique**, ou **ajustée**, de la variable à expliquer, associée à la valeur observée x_i ;
- $u_i = y_i - y'_i$ est l'**erreur d'ajustement** (ou **résidu**), c'est-à-dire l'écart entre la valeur observée et la valeur théorique calculée de la variable à expliquer.



La « meilleure » droite retenue est en fait celle qui rend minimale la somme des carrés des erreurs d'ajustement : Σu_i^2. On l'appelle **droite des moindres carrés** de y en x ou **droite de régression** de y en x.

On montre que les coefficients: A et B, de la droite de régression de y en x s'expriment en fonction des données par :

$$\mathbf{A} = \frac{\Sigma\,(x_i - \bar{x}).(y_i - \bar{y})}{\Sigma\,(x_i - \bar{x})^2}$$

$$\bar{y} = \mathbf{A}.\bar{x} + \mathbf{B}$$

la seconde relation, qui permet de calculer B après A, signifie que la droite passe par le point moyen du nuage : ($\bar{x}$, $\bar{y}$).

Exemple: y : 5 4 6 5 9 9
x : 1 3 4 7 7 9

On figure le nuage des points, puis on calcule successivement :

$\bar{x} \approx 5{,}17$ et $\bar{y} = 6{,}5$
$A \approx 0{,}569$ et $B \approx 3{,}56$

et on obtient la droite de régression :

$y \approx 0{,}569.x + 3{,}56$

que l'on trace sur le graphique ci-dessous.

Droite de régression de x en y

La régression de y en x donne des rôles différents aux deux variables, on peut renverser le problème et régresser la variable x sur y.

On obtient une droite : $x = C.y + D$

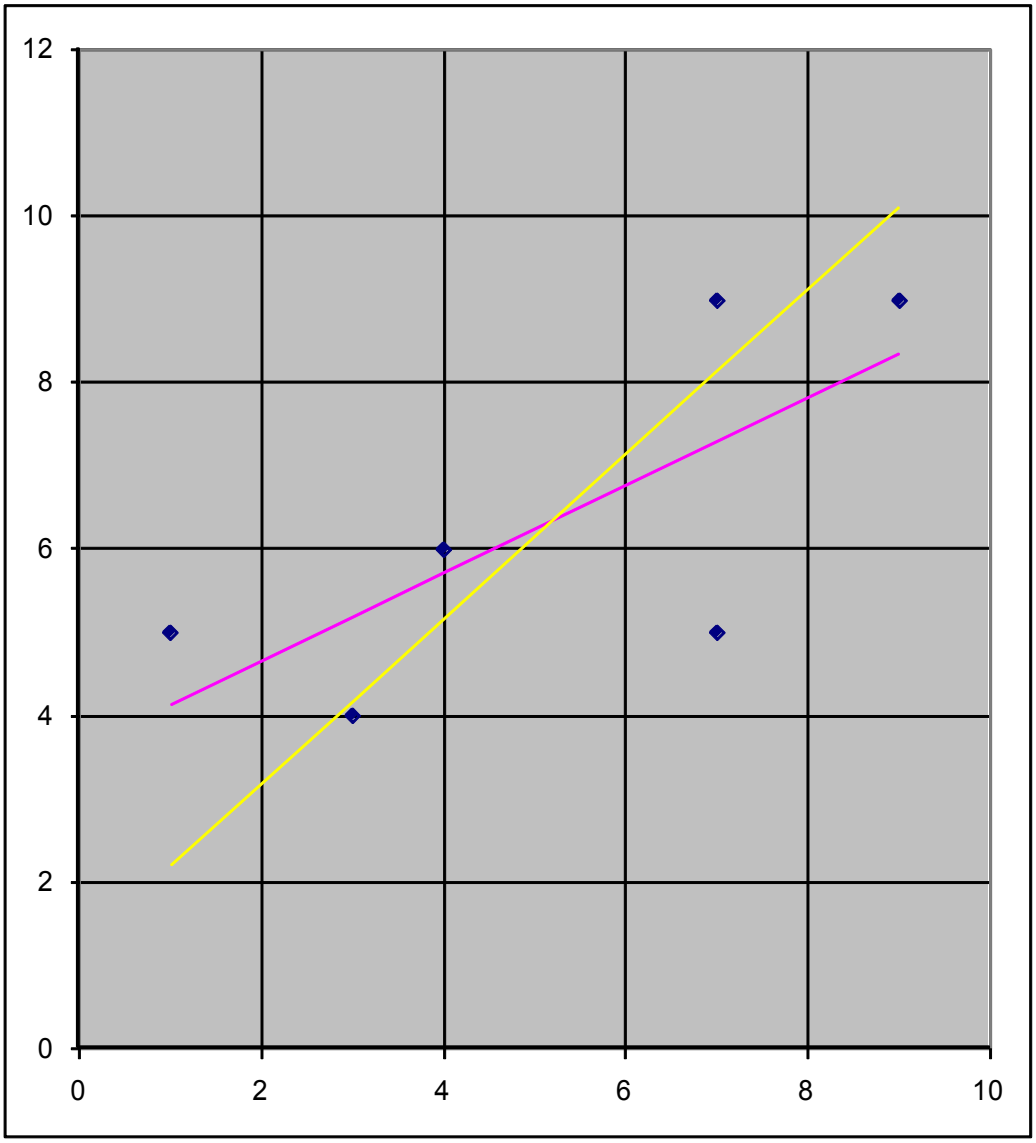
de coefficients : C et D, donnés par :

$$\mathbf{C} = \frac{\Sigma\,(x_i - \bar{x}).(y_i - \bar{y})}{\Sigma\,(y_i - \bar{y})^2}$$

$$\bar{x} = \mathbf{C}.\bar{y} + \mathbf{D}$$

Les deux droites de régression différent en général, mais sont peu différentes si le nuage est proche de l'alignement.

Exemple : en reprenant l'illustration numérique précédente, on trouve : $C \approx 1{,}186$ et $D \approx -2{,}54$, soit la droite : $x \approx 1{,}186.y - 2{,}54$ ou encore $y \approx 0{,}843.x + 2{,}14$ figurée sur le graphique.



Coefficient de corrélation

Il est des cas ou aucune des deux variables ne paraît devoir expliquer l'autre (par exemple le taux d'équipement en réfrigérateurs et celui en magnétoscopes). On s'intéresse alors davantage à mesurer l'intensité de la liaison linéaire éventuelle, qu'à régresser l'une des variables sur l'autre. Pour ce faire, on calcule le **coefficient de corrélation**, noté : **R**, entre les deux variables :

$$R = \frac{\Sigma (x_i - \bar{x}).(y_i - \bar{y})}{[\Sigma (x_i - \bar{x})^2]^{0,5}.[\Sigma (y_i - \bar{y})^2]^{0,5}}$$

On peut montrer que ce coefficient R est toujours compris entre -1 et 1. Il mesure comme suit l'intensité de la liaison linéaire :

- si R est proche de 1 : il y a une liaison linéaire marquée, et les deux variables varient dans le même sens ;
- si R est proche de 0 : il n'y a pas de liaison linéaire ;
- si R est proche de –1 : il y a une liaison linéaire marquée, et les deux variables varient en sens contraire.

Dans le premier et le dernier cas le nuage montre un bon alignement, et les deux droites de régression sont proches. Dans le second cas, le nuage n'a pas de caractère linéaire, et la régression n'est guère justifiée.

Exemple : dans le cas traité, on trouve : $R \approx 0{,}821$ ce qui traduit numériquement l'aspect linéaire observé.

Analyse de la variance

Revenant à la droite de régression de y en x, on montre que les résidus : u_i, sont de somme et de moyenne nulles, tandis que les valeurs ajustées de la variable à expliquer : y'_i, ont la même moyenne : $\bar{y}$, que les valeurs observées : y_i.

On a en outre la relation :

$$\Sigma (y_i - \bar{y})^2 = \Sigma (y'_i - \bar{y})^2 + \Sigma u_i^2$$

ou **équation d'analyse de la variance**. On dit que la **variation totale** (autour de la moyenne) est égale à la **variation expliquée** augmentée de la **variation résiduelle**.

On montre que le carré du coefficient de corrélation défini précédemment s'obtient encore par :

$$R^2 = \frac{\Sigma (y'_i - \bar{y})^2}{\Sigma (y_i - \bar{y})^2} = \frac{\text{variation expliquée}}{\text{variation totale}}$$

La régression telle qu'elle vient d'être présentée reste une méthode descriptive; il importe en particulier de ne pas croire que l'observation d'une corrélation implique nécessairement une liaison de **causalité**.

Ainsi la série annuelle du nombre des divorces depuis la guerre peut-elle être bien corrélée avec celle donnant la longueur totale des autoroutes, sans qu'on doive supposer qu'une relation de causalité relie les deux phénomènes. Une forte corrélation entre deux séries économiques peut, par exemple, être simplement l'effet mécanique d'une tendance temporelle commune.

Remarque : origine du terme régression

On peut s'étonner de l'emploi pour désigner l'ajustement linéaire du terme régression, dont le sens usuel paraît a priori sans rapport avec la méthode exposée. Il fait suite à une étude, publiée en 1886, du statisticien anglais Francis Galton, expliquant la taille des enfants par celle des parents : il observait une pente inférieure à 1, c'est dire un retour vers la moyenne ou une régression. Il s'agit en fait d'un artefact résultant de la disymétrisation, car l'explication de la taille des parents par celle des enfants fait apparaître le même phénomène inversé !

POINT DE VUE ÉCONOMÉTRIQUE

Le modèle

Guidé par ses idées économiques, et l'observation du nuage, l'économètre suppose une liaison effective entre les variables (par exemple la consommation et le revenu), ou **équation économétrique**, de la forme :

$$\mathbf{Y = a.X + b + \varepsilon}$$

où **a** et **b** sont les coefficients inconnus du modèle, et ε, une **perturbation aléatoire**, appelée simplement **aléa**.

La présence de cet aléa traduit le caractère approché des « lois » en économie, à la différence par exemple des lois physiques, qui se veulent des lois exactes (quoique les données réelles soient généralement entachées d'erreurs dues à l'imperfection des mesures).

On peut considérer que le terme aléatoire : ε, rassemble toutes les influences autres que celle de la variable explicative X d'incidences secondaires sur la variable à expliquer Y et non explicitement prises en compte dans le modèle.

Pour dire les choses autrement, on suppose l'existence d'une « vraie » droite associée à loi du phénomène modélisé, et chaque couple d'observations dans une relation :

$$\mathbf{y_i = a.x_i + b + \varepsilon_i}$$

incluant la perturbation, mais la position exacte de cette droite est inconnue, l'économètre ne connaît ni les valeurs de a et b, ni celles des ε_i.

C'est à partir des seuls N points ou couples données observées : $(\mathbf{x_i,y_i})$, qu'il faut estimer les différentes quantités, et juger de la pertinence du modèle. Ces données observées sont l'équivalent pour l'économètre de l'échantillon du statisticien classique, à ceci près qu'il a généralement beaucoup moins de liberté pour la constitution de celui-ci…

Propriétés des estimations des moindres carrés

À défaut de connaître la vraie droite, on retient la droite des moindres carrés, les valeurs : **A** et **B**, calculées comme précédemment par la méthode des moindres carrés, ne sont plus simplement les coefficients d'une droite géométriquement satisfaisante, mais des estimations statistiques des coefficients : a et b, du modèle.

La relation $\mathbf{Y = A.X + B}$ est l'**équation estimée**., tandis que chaque relation $\mathbf{y_i = A.X_i + B + u_i}$ (à ne pas confondre avec la relation « vraie » du paragraphe précédent) fournit le **résidu** $\mathbf{u_i}$ correspondant.

Les propriétés des **estimateurs des moindres carrés**, dépendent des caractéristiques de l'aléa ε.

On suppose que les aléas : ε_i, suivent la même loi normale: $\mathbf{N}(0,\sigma)$, centrée, et d'écart-type : σ (inconnu), et sont **indépendants**, ces présupposés sont généralement appelés les **hypothèses des moindres carrés**. L'hypothèse de normalité n'est pas déraisonnable si l'on admet que ces aléas résultent de l'addition de nombreux effets secondaires mineurs indépendants. Sous ces conditions :

- les résidus calculés: ui, approchent les aléas inconnus: ε_i, et la quantité $\mathbf{[(\Sigma\, u_i^2)/(N-2)]^{0,5}}$ liée à la somme des carrés des résidus, est une bonne estimation de l'écart-type: σ, de l'aléa. Elle est appelée: **écart-type résiduel** ;

- les estimateurs : A et B, sont les « meilleurs possibles » (en un sens mathématique qu'on ne précisera pas davantage pour l'instant) ;

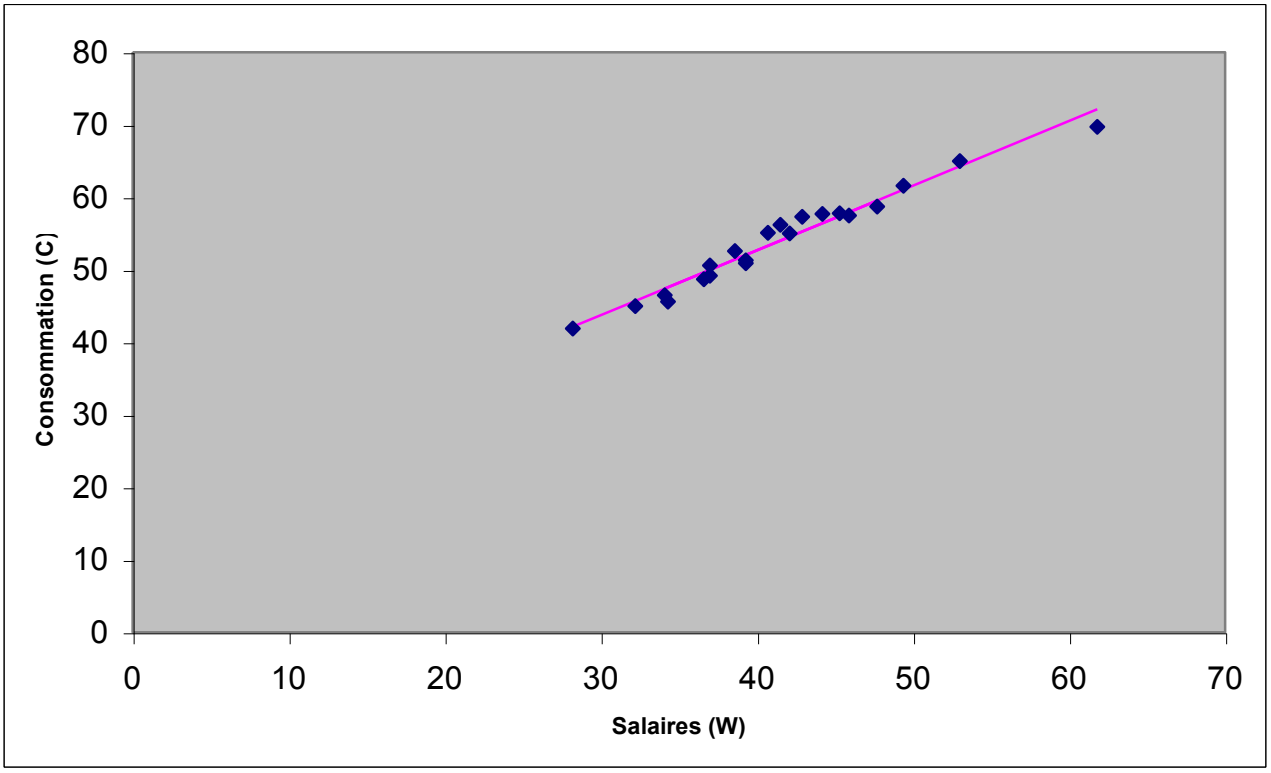
- les estimateurs : A et B, suivent des lois normales : **N**(a,s_A) et **N**(b,s_B), dont les espérances : a et b, sont les quantités estimées ; ces estimateurs sont sans biais ;

- les écart-type : s_A et s_B, des estimateurs : A et B, peuvent également être estimés.

Pour une précision minimale des estimations, on demande généralement que le nombre : N, d'observations utilisées approche au moins la quinzaine.

Les logiciels économétriques calculent et éditent les valeurs estimées: A et B, ainsi que les différentes quantités précédentes, le carré du coefficient de corrélation et d'autres indicateurs d'usage plus technique qui seront présentés dans la suite du cours.

Un exemple



Régression de la consommation annuelle des ménages : C, sur le total des salaires : W, aux États-Unis de 1921 à 1941 (sorties du logiciel économétrique Ystat)

```
                  OLS  -- DEPENDENT VARIABLE: C

 SAMPLE SIZE(  1 to   21) =            21 (DF=19)
 SUM OF SQUARED RESIDUALS =            32.285574
           VARIANCE (MSE) =             1.699241
STANDARD ERROR (ROOT MSE) =             1.303549
               R-SQUARED =              0.965706
      ADJUSTED R-SQUARED =              0.963901
  F-STATISTIC(  1,    19) =           535.029521 (p=0.0009)
        SUM OF RESIDUALS =              0.000000
  DURBIN-WATSON STATISTIC =              0.927815
```

RIGHT-HAND VARIABLE	ESTIMATED COEFFICIENT	STANDARD ERROR	T-STATISTIC	PROB.	STANDARDIZED COEFFICIENT
1 W	0.892316559	(0.03858)	23.13071	0.000	0.9827033
2 Constant	16.981097387	(1.62530)	10.44796	0.000	0.0000000

Test de significativité d'un coefficient (test de Student)

Bien qu'il soit possible de construire des intervalles de confiance pour les coefficients du modèle, on préfère en économétrie donner les estimations ponctuelles en écrivant l'équation estimée, et tester simplement la **significativité** des coefficients.

Il s'agit de tester si, pour un niveau de confiance donné (en général 95%), l'intervalle de confiance peut ou non contenir la valeur 0. En effet si la valeur véritable du coefficient peut être 0, il n'est même pas certain que la variable explicative (ou le terme constant) intervienne réellement dans le modèle.

Sachant que pour un risque α, l'intervalle de confiance pour a est :

$$[A - t_\alpha.s_A ; A + t_\alpha.s_A]$$

le test revient à examiner si le rapport :

$$\frac{|\mathbf{A}|}{s_A} = \frac{|\textbf{coefficient estimé}|}{\textbf{écart-type estimé}}$$

dépasse ou non t_α.

On fait en général ce test au risque α = 5%, ce qui donne, en utilisant la valeur approchée $t_{0,05} \approx 1,96 \approx 2$:

- $|$coefficient estimé$|$/écart-type estimé < 2 : coefficient non significatif au risque 5% ;
- $|$coefficient estimé$|$/écart-type estimé > 2 : coefficient significatif au risque 5%.

Pour ces raisons, il est d'usage dans les articles, de présenter les équations estimées en plaçant sous chaque coefficient soit son écart-type estimé, soit directement le ratio précédent (l'usage n'est malheureusement pas unifié, le second choix paraissant toutefois préférable). Le lecteur quelque peu versé en économétrie peut ainsi effectuer aisément ce test pour chaque coefficient.

Ce test est généralement appelé **test de Student**, car, strictement, lorsque l'échantillon utilisé est de petite taille (N < 30), il conviendrait d'employer une loi de Student, voisine de la loi normale mais plus dispersée, pour tenir compte du fait que l'écart-type est lui-même estimé.

Lors d'une étude économétrique, le test de Student sur chacun des coefficients est beaucoup plus important que l'examen du coefficient de corrélation.

Un « bon » test de Student doit toutefois être regardé avec une certaine modestie, ce test suppose en effet la pertinence du modèle, mais il n'a pas vocation à la confirmer; en fait, il sert essentiellement à mettre en doute ou à écarter les variables d'influence incertaine.

Exemple : la régression précédente s'écrit :

$$C = 0{,}892.W + 16{,}98$$
$$\quad (23{,}13) \qquad (10{,}45)$$

(on a noté entre parenthèses les **ratios de Student**)
Les rapports sont de l'ordre de 20 et 10, ce qui montre la forte significativité des deux coefficients estimés - à supposer, comme il vient d'être dit, le modèle correct et les hypothèses des mco satisfaites...

Prévision

L'équation estimée permet de faire des **prévisions** (s'il s'agit de séries temporelles), ou plus généralement de calculer la valeur :

$$Y_{N+1} = A.x_{N+1} + B$$

de la variable expliquée par le modèle, correspondant à une nouvelle valeur : x_{N+1}, de l'explicative.

En supposant le modèle toujours valable, la valeur véritable serait en fait :

$$y_{N+1} = a.x_{N+1} + b + \varepsilon_{N+1}$$

On voit que la prévision de y_{N+1} par Y_{N+1} n'est qu'une estimation, d'une part parce que A et B ne sont que des estimations d a et b, et d'autre part parce que la nouvelle perturbation aléatoire : ε_{N+1}, est également inconnue.

On peut montrer que la variance de l'erreur de prévision : $Y_{N+1} - y_{N+1}$ vaut :

$$\text{var}(Y_{N+1} - y_{N+1}) = \sigma^2.[1 + 1/N + (x_{N+1} - \overline{x})^2/\Sigma(x_i - \overline{x})^2]$$

c'est à dire, que, outre une incertitude incompressible liée à la présence d'un aléa, la prévision est d'autant moins précise que la nouvelle valeur : x_{N+1}, de l'explicative est éloignée du domaine de celles ayant servi à l'estimation du modèle, mais d'autant plus précise que la taille de l'échantillon utilisé est grande.

Il est enfin possible que le modèle ne soit plus valable ! (par exemple un modèle de consommation d'énergie après le choc pétrolier de 1973).

Exemple : avec l'équation de consommation américaine estimée précédemment, si on suppose pour 1942 un total des salaires : W, égal à 65, on déduit la consommation :

$$C = 0{,}892.65 + 16{,}98 = 74{,}96$$

en fait, il s'agit plutôt ici d'une **simulation** que d'une prévision.

LE MODÈLE LINÉAIRE MULTIPLE

Une grandeur économique dépend rarement d'une seule variable, et les économètres considèrent en général des modèles à plusieurs variables explicatives (par exemple, l'investissement des entreprises pourra dépendre à la fois des profits, du taux d'intérêt et du niveau de capital fixe précédent).

LE MODÈLE

Le modèle linéaire multiple postule que la grandeur à expliquer est une expression linéaire des variables explicatives retenues, perturbée par un aléa.

Ainsi le modèle de variable à expliquer : Y, et à trois variables explicatives : X, Z et T (ou plutôt quatre en comptant explicitement la constante), s'écrit :

$$\mathbf{Y = a.X + b.Z + c.T + d + \varepsilon}$$

où **a**, **b**, **c** et **d** sont les coefficients, inconnus, du modèle, et ε la perturbation aléatoire ou aléa.

Le problème est à nouveau d'estimer et d'éprouver le modèle à partir des seules N observations des variables: $\mathbf{y_i}$, $\mathbf{x_i}$, $\mathbf{z_i}$ et $\mathbf{t_i}$.

Terminologie

On a coutume d'appeler indifféremment Y la variable à expliquer, la variable **dépendante** (des autres) ou la variable **endogène** (du modèle), et X, Z et T (et la constante) les variables explicatives, **indépendantes** ou **exogènes**.

ESTIMATIONS DES MOINDRES CARRÉS, PROPRIÉTÉS

Principe et premières propriétés

Si la représentation dans un plan du nuage des points n'est plus possible, la méthode des **mco** (pour « moindres carrés ordinaires », ou olsq, pour "ordinary least squares", chez les Anglo-Saxons) s'emploie encore; elle revient à chercher l'expression :

$$\mathbf{Y = A.X + B.Z + C.T + D}$$

qui minimise la somme des carrés des erreurs d'ajustement : $u_i = y_i - (A.x_i + B.z_i + C.t_i + D)$, ou résidus, pour les valeurs observées.

Les valeurs : A, B, C et D, correspondantes sont les estimations des moindres carrés des coefficients : a, b, c et d, du modèle (les logiciels économétriques les calculent par des méthodes matricielles appropriées indiquées en appendice).

Comme dans le modèle simple, la méthode permet de calculer les valeurs ajustées ou théoriques de la variable expliquée, les résidus, ainsi que les diverses quantités apparaissant dans la relation *d'analyse de la variance* et le coefficient R^2, appelé à présent **coefficient de corrélation multiple**.

Les propriétés simplement algébriques, ou géométriques, sont conservées.

Si constante figure parmi les explicatives, les résidus sont de somme et de moyenne nulle, tandis que la variable ajustée a même moyenne que la variable observée, la relation d'analyse de la variance est satisfaite. Le coefficient R^2 est compris entre 0 et 1, la proximité de 1 traduisant toujours une forte liaison linéaire multiple, mais d'un seul point de vue descriptif, l'introduction de variables explicatives quelconques le faisant croître mécaniquement…

Propriétés des estimateurs des mco

Sous les mêmes hypothèses (dites des moindres carrés) sur l'aléa ε, les estimateurs vérifient encore les propriétés indiquées pour le modèle simple.

On suppose donc que les aléas : ε_i, suivent la même loi normale : $\mathbf{N}(0,\sigma)$, centrée, et d'écart-type : σ (inconnu), et sont indépendants. On demande en outre à nouveau que le nombre: N, d'observations dépasse largement le nombre de variables explicatives, si possible d'au moins une quinzaine. Sous ces conditions :

- les résidus u_i approchent les aléas ε_i, et la quantité $[(\Sigma\, u_i^2)/(N-4)]^{0,5}$ est une bonne estimation de l'écart-type : σ, de l'aléa ; elle est appelée : écart-type résiduel. Dans le cas général, le dénominateur est : N-k, où k est le nombre de variables explicatives, constante comprise si elle est présente. Cette quantité : N-k, est le nombre de **degrés de liberté** de la régression ;

- les estimateurs : A, B, C et D, sont sans biais ;

- ils sont de variance minimale ou efficaces, c'est à dire les plus précis, parmi ceux linéaires par rapport aux y_i et sans biais ;

- ils suivent des lois normales : $\mathbf{N}(a,s_A)$, $\mathbf{N}(b,s_B)$, $\mathbf{N}(c,s_C)$ et $\mathbf{N}(d,s_D)$, dont les espérances : a, b, c et d, sont donc les quantités qu'ils estiment ;

- les écart-type : s_A, s_B, s_C et s_D, de ces estimateurs : A, B, C et D, peuvent être également estimés ;

- sous certaines hypothèses complémentaires concernant la suite des données (c'est à dire la suite d'échantillons croissants utilisés), les différents estimateurs sont consistants et convergents.

Ces estimateurs suivant donc des lois normales, et leurs écart-type pouvant être estimés, on peut opérer le test de Student de significativité pour chaque coefficient.

Par exemple, pour le coefficient b, le test au niveau de risque α revient à examiner si le rapport :

$$\frac{|\mathbf{B}|}{\mathbf{s_B}} = \frac{|\textbf{coefficient estimé}|}{\textbf{écart-type estimé}}$$

dépasse ou non t_α.

On fait en général le test au risque $\alpha = 5\%$, ce qui, avec la valeur approchée : $t_{0,05} \approx 2$, revient à examiner la position du ratio précédent par rapport à 2 :

- si le ratio dépasse 2, le coefficient est significativement différent de 0 au risque 5% ;

- si le ratio est inférieur à deux, le coefficient n'est pas significativement différent de 0 au risque 5%.

PRATIQUE ÉLÉMENTAIRE DE LA RÉGRESSION

Indiquons comment lire les résultats d'une régression linéaire multiple.

On peut commencer par jeter un rapide coup d'œil sur le coefficient de corrélation multiple R^2 mais sans y attacher trop d'importance : une valeur élevée – le cas le plus fréquent lorsqu'on traite des séries chronologiques brutes – peut n'être qu'un artefact (résultant par exemple d'un trend

commun) sans réelle valeur prédictive, tandis qu'à l'inverse un modèle peut être pertinent mais affecté d'une forte imprécision due à une perturbation aléatoire importante…

On examine ensuite les ratios de Student attachés à chaque coefficient - à ne pas confondre avec les coefficients estimés eux-mêmes ! Il n'y a pas grand chose à dire des coefficients pour lesquels ces tests sont très mauvais : même à supposer que la variable concernée intervienne réellement, la valeur estimée est trop incertaine pour être commentée valablement. Si on en a la possibilité, on peut reprendre les calculs en ôtant des explicatives tout ou partie des variables présentant de mauvais tests de Student.

Les coefficients les plus significatifs sont en revanche à examiner avec attention: on peut commencer par regarder si leur signe correspond à ce que l'on aurait attendu, puis étudier les valeurs numériques elles-mêmes, les rapporter à leur signification économique (consommation marginale, élasticité, etc.) et les comparer à celles proposées par les travaux antérieurs ou la théorie...

Il convient pour les examens précédents de ne pas faire de fixation déraisonnable sur le seuil de 5% (ou toute autre valeur fixée à l'avance), ni de perdre de vue que, comme en toute question d'estimation, l'imprécision comme la possibilité de conclusions erronées sont consubstantielles à la méthode.

On gardera par ailleurs à l'esprit que l'échelle des coefficients dépend des unités en lesquelles sont exprimées les données, du moins en cas de variables hétérogènes (euros, billions de dollars, espérance de vie, taux bancaire, nombre de voitures, d'enfants ou de téléviseurs, ratio per capita, etc.). Les **coefficients standardisés**, calculés sur les séries centrées-réduites et parfois édités permettent éventuellement de s'affranchir de ces effets d'échelle pour comparer l'influence relative des différentes variables explicatives.

L'examen des résidus (supposés approcher les aléas si le modèle est correct et convenablement estimé), dont le diagramme doit être édité, permet enfin de repérer éventuellement certaines observations **atypiques**, et surtout de tester a posteriori les hypothèses faites sur ces aléas, notamment leur indépendance temporelle. Ces points importants seront examinés dans les chapitres suivants.

Si la lecture, et l'appréciation, d'une équation estimée peut ainsi paraître relativement aisée, le travail de l'économètre pour tenter de trouver une équation convenable expliquant une variable donnée, l'est beaucoup moins. Il doit en effet déterminer la forme du modèle, chercher les véritables explicatives parmi celles pour lesquelles des données fiables lui sont accessibles, etc.

Il est naturellement possible que la grandeur étudiée ne puisse pas être expliquée par un modèle économétrique (tels les gains hebdomadaires d'un particulier au PMU).

Exemple : régression de la consommation annuelle des ménages : C, sur le profit des entreprises : P, le profit décalé : P_{-1}, et le total des salaires : W, aux États-Unis de 1921 à 1941.

On remarque que si ce modèle est correct, le modèle linéaire simple considéré précédemment ne l'était pas, alors même que les estimations et indicateurs obtenus avaient pu paraître satisfaisants...

OLS - DEPENDENT VARIABLE: C

```
SAMPLE SIZE(  1 to   21) =           21 (DF=17)
SUM OF SQUARED RESIDUALS =           17.879449
```

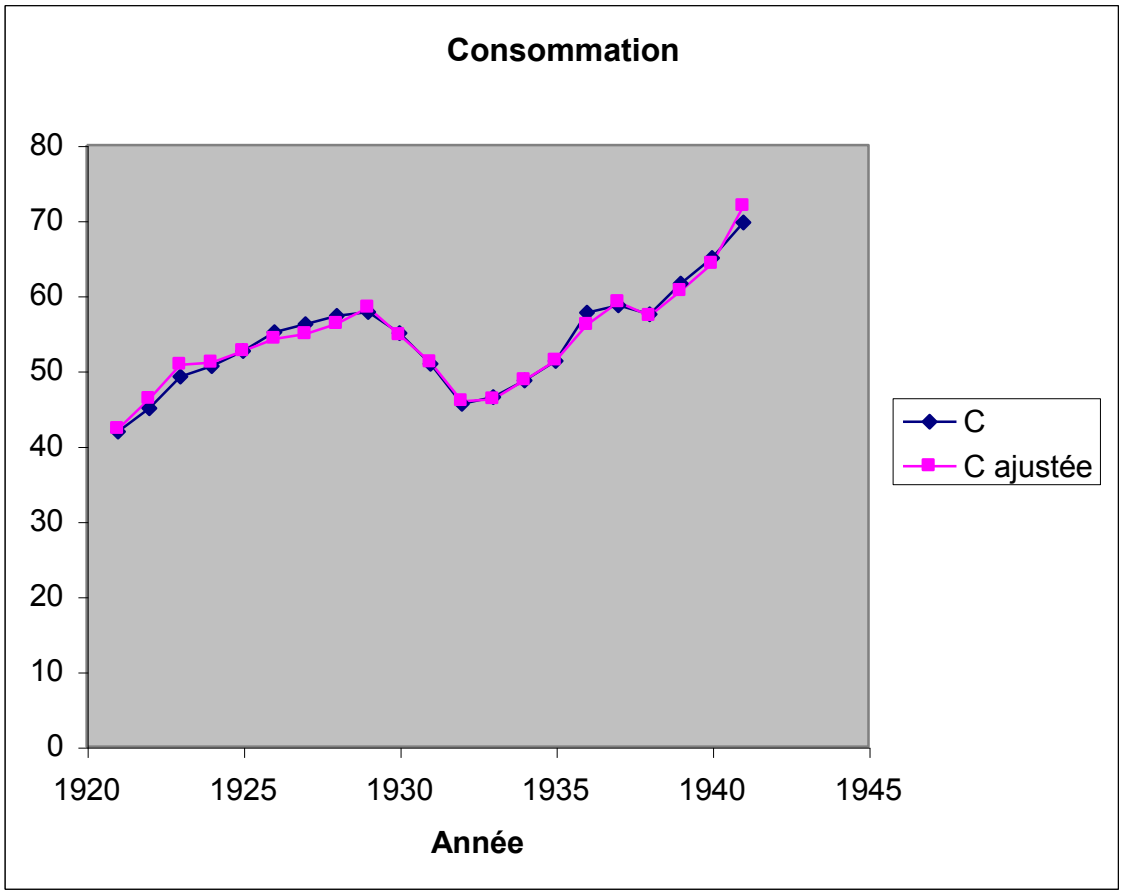
```
           VARIANCE (MSE) =            1.051732
STANDARD ERROR (ROOT MSE) =            1.025540
               R-SQUARED =             0.981008
      ADJUSTED R-SQUARED =             0.977657
  F-STATISTIC(  3,    17) =           292.707595 (p=0.0004)
        SUM OF RESIDUALS =             0.000000
  DURBIN-WATSON STATISTIC =             1.367474
```

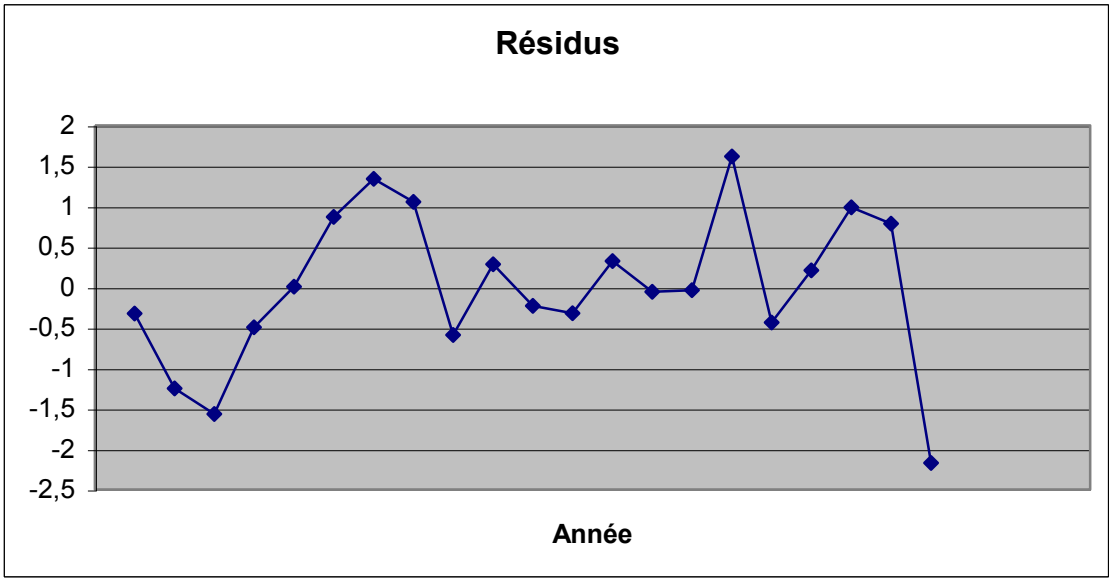
	RIGHT-HAND VARIABLE	ESTIMATED COEFFICIENT	STANDARD ERROR	T_STATISTIC	PROB.	STANDARDIZED COEFFICIENT
1	P	0.192934381	(0.09121)	2.11527	0.049	0.11867562
2	P-1	0.089884898	(0.09065)	0.99158	0.335	0.05277306
3	W	0.796218750	(0.03994)	19.93342	0.000	0.87687132
4	Constant	16.236600272	(1.30270)	12.46382	0.000	0.00000000

La régression obtenue peut ainsi s'écrire :

$$C = 0{,}193.P + 0{,}090.P_{-1} + 0{,}796.W + 16{,}24$$
$$\quad (2{,}12) \qquad (0{,}99) \qquad (19{,}9) \qquad (12{,}5)$$

tandis qu'on peut figurer les séries chronologiques de consommation brutes et ajustées ainsi que le diagramme des résidus :





On remarque tout d'abord un coefficient R^2 très élevé, mais cet indicateur, comme on l'a dit, n'est pas très significatif pour juger de l'intérêt économétrique d'une régression ni de la pertinence du modèle.

On note ensuite que les signes des coefficients estimés paraissent conformes à ce que l'on pouvait attendre; mais on observe cependant que le coefficient du profit retardé : P_{-1}, n'est pas significatif au risque 5%. Au vu de cela, et à moins que des raisons théoriques fortes ne nous en dissuadent, on serait tenté d'abandonner cette variable et de reprendre la régression avec seulement P et W pour explicatives...

Les valeurs estimées des deux autres coefficients montrent qu'un accroissement du total des salaires est répercuté à près de 80% sur la consommation, et un accroissement des profits des entreprises pour environ 20%, ce dernier coefficient, quoique significatif au risque 0,05, étant estimé avec moins de précision.

Le diagramme des résidus fait apparaître un écart important pour la dernière période, conduisant à s'interroger sur la permanence et la stabilité du modèle pendant la guerre. La structure en arches assez régulière met également en doute l'hypothèse d'indépendance des aléas…

APPENDICE (MATHÉMATIQUES DE LA RÉGRESSION)

Présentation matricielle de la régression multiple

Soit une variable y à régresser sur les k variables explicatives x, z, t, ..., w à l'aide de N observations de chacune d'entre elles.

On note encore Y le vecteur des observations de la variable y en colonne et X la matrice N_xk portant de même les explicatives en colonnes, y compris éventuellement une colonne de 1 si le modèle prévoit la constante, comme c'est le plus souvent le cas.

Le modèle s'écrit :

$$Y = X.a + \varepsilon$$

où a est le vecteur à k composantes des coefficients, et ε un vecteur aléatoire de dimension N figurant les aléas.

Les hypothèses des mco se traduisent en disant que le vecteur ε des aléas, suit une loi normale multidimensionnelle centrée et de matrice de variances-covariance $\sigma^2.I$ (où I note la matrice unité de dimension NxN) :

$$\varepsilon \sim \mathbf{N}(0, \sigma^2.I)$$

D'un point de vue simplement géométrique, la minimisation de la somme des carrés des résidus revient à projeter orthogonalement le vecteur Y sur le sous-espace engendré par les colonnes de X. Cela se traduit par la relation matricielle :

$$X'.(Y - X.A) = 0 \quad \text{ou} \quad X'.Y = (X'.X).A \quad (\textbf{équations normales})$$

où X' désigne la matrice transposée de X et A le vecteur des coefficients estimés. En supposant les colonnes de X linéairement indépendantes, X'.X est inversible et A vaut donc :

$$A = (X'.X)^{-1}.X'.Y = a + (X'.X)^{-1}.X'.\varepsilon$$

La première expression indique comment calculer le vecteur A, et la seconde montre qu'il suit une loi normale multidimensionnelle d'espérance a et de matrice de variance-covariance :

$$V(A) = \sigma^2.(X'.X)^{-1}$$

Le vecteur des y ajustés est le vecteur normal :

$$Y_a = X.A$$

tandis que le vecteur des résidus est le vecteur normal orthogonal au précédent :

$$u = Y - Y_a = Y - X.A = [I - X.(X'.X)^{-1}.X'].Y$$

et le carré scalaire $u'.u/\sigma^2$ suit une loi de chi-deux à N-k degrés de liberté, ce qui justifie l'estimation de σ et les tests de Student qui en dérivent.

Soit une nouvelle observation des explicatives: x_{N+1} , rangée en un vecteur ligne de dimension k, la prévision, ou valeur ajustée correspondante, de y: Y_{N+1} , est donnée par :

$$Y_{N+1} = x_{N+1}.A$$

tandis que la valeur vraie, encore inconnue, sera :

$$y_{N+1} = x_{N+1}.a + \varepsilon_{N+1}$$

et la variance de l'erreur de prévision : $Y_{N+1} - y_{n+1}$, est :

$$var(Y_{N+1} - y_{n+1}) = x_{N+1}.V(A).x'_{N+1} + var(\varepsilon_{N+1}) = \sigma^2.[x_{N+1}.(X'.X)^{-1}.x'_{N+1} + 1]$$

<u>Retour sur le maximum de vraisemblance</u>

On peut encore présenter les choses dans le cadre d'une estimation par le maximum de vraisemblance.

En conservant l'écriture matricielle, le modèle peut aussi s'écrire :

$$\varepsilon = Y - X.a$$

ou pour l'observation i :

$$\varepsilon_i = y_i - (a_1 + x_i.a_2 + z_i.a_3 + ... + w_i.a_k)$$

Sachant qu'on a supposé les aléas normaux, centrés et de même variance σ^2, la densité est :

$$f_a = (2.\pi.\ \sigma^2)^{-1/2}.\exp\{\frac{-[y_i - (a_1 + x_i.a_2 + z_i.a_3 + ... + w_i.a_k)]^2}{2.\ \sigma^2}\}$$

et celle de l'ensemble des observations (correspondant à un échantillon de taille N), c'est à dire la vraisemblance, est donc :

$$L = K.\sigma^{-N}.\exp\{\frac{-\Sigma\ [y_i - (a_1 + x_i.a_2 + z_i.a_3 + ... + w_i.a_k)]^2}{2.\sigma^2}\}$$

quantité que l'on doit maximiser par rapport aux coefficients $a_1, a_2, a_3, \ldots, a_k$ et à σ, les inconnues du modèle.

On voit aisément que le choix de A, c'est à dire des coefficients estimés $a_1, a_2, a_3, ..., a_k$ qui maximise cette vraisemblance est précisément celui qui minimise la somme des carrés des résidus. Les estimateurs des mco sont donc également ceux du maximum de vraisemblance.

----÷≈≈≡≡oo0OΩO0oo≡≡≈≈÷----

(19.02.2009)