Licence AES 3ème année

Université de Lille II

Notes du cours d'introduction à l'économétrie

P. De Vreyer

Partie III: Le modèle de régression simple.

1. Introduction.

Nous sommes maintenant en mesure de présenter le modèle de régression simple. Celui-ci relie deux variables aléatoires: une variable expliquée (encore appelée dépendante ou endogène), que nous noterons Y et une variable explicative (encore appelée indépendante ou exogène), que nous noterons X. La relation entre Y et X dépend de paramètres qui sont l'objet de l'estimation:

$$Y = \alpha + \beta . X$$

En économétrie, cette relation découle en principe du raisonnement économique. Par exemple, la théorie keynésienne de la consommation prédit que la consommation (ici Y) est une fonction linéaire du revenu courant (ici X). Cependant, même si la théorie est vérifiée, la relation est rarement observée exactement, parce que des perturbations de tous ordres interviennent: Y et X peuvent être mesurés avec erreur ou des évènements imprévus peuvent être intervenus qui perturbent temporairement la relation. Pour cette raison le modèle économétrique diffère du modèle économique en ce que l'on ajoute un terme aléatoire à la relation précédente pour obtenir:

$$Y = \alpha + \beta . X + \epsilon$$

Avant de progresser il est bon de faire le point sur ce que l'économètre observe et sur ce qu'il doit estimer. Ce qui est observé, ce sont les valeurs de Y et de X pour un ensemble d'observations rangées dans un échantillon. Ce qui est inconnu ce sont les valeurs des paramètres, α et β, et la distribution du terme d'erreur ϵ. Ce que l'économètre doit estimer ce sont les paramètres et la moyenne et l'écart-type de la distribution de ϵ.

Les observations à disposition de l'économètre sont rangées dans les vecteurs $y = y(y_1, ..., y_n)'$ et $x = (x_1, ..., x_n)'$ où n est la taille supposée de l'échantillon. Il est habituel d'écrire les modèles économétriques à partir des observations dont on dispose et non pas des variables aléatoires dont la réalisation est à leur origine. Deux écritures peuvent être adoptées. Le modèle peut être écrit au niveau de l'ensemble des observations:

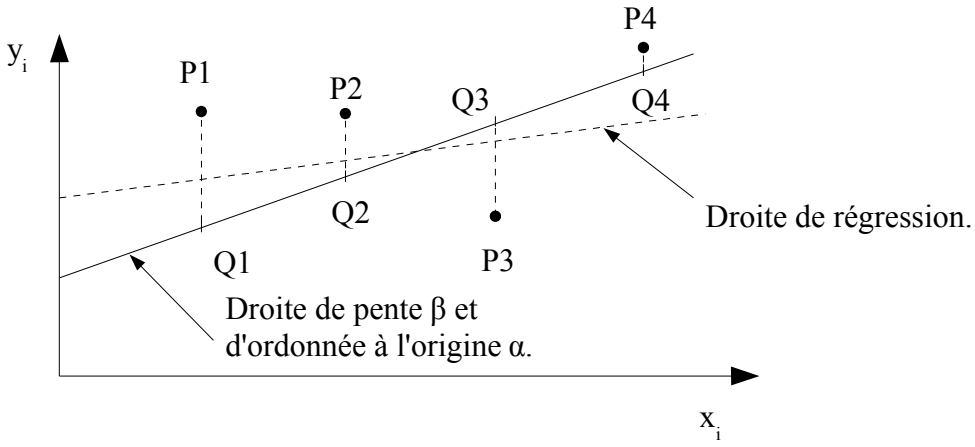
$$y = \alpha + \beta . x + \epsilon$$

Il est important de bien noter que dans cette écriture y, x et ϵ sont des vecteurs colonnes et que les paramètres α et β sont des nombres réels constants. Le modèle peut alternativement être écrit au niveau des observations individuelles:

$$y_i = \alpha + \beta . x_i + \epsilon_i$$

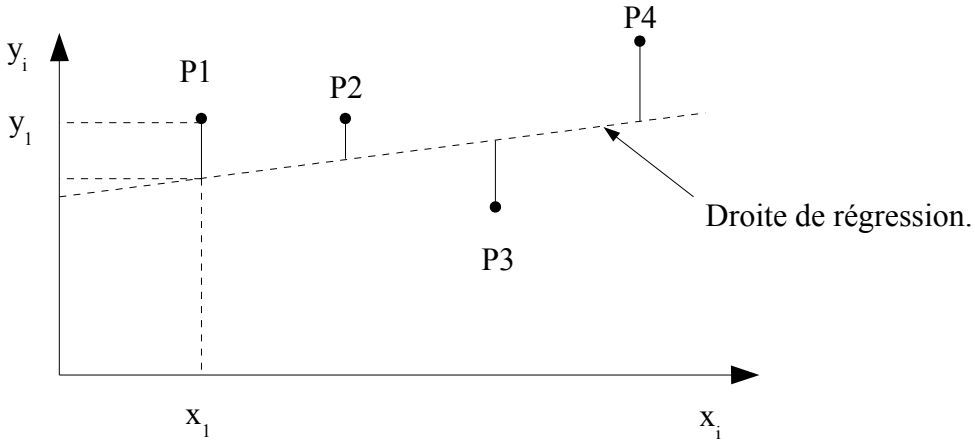
Enfin, il devrait être clair que $\epsilon = (\epsilon_1, ..., \epsilon_n)$.

Supposons que l'on dispose de 4 observations de la variable explicative, x_1, x_2, x_3, x_4. Si la relation théorique (économique) n'était pas perturbée, on observerait également 4 valeurs de la variable expliquée, notées $\tilde{y}_1, \tilde{y}_2, \tilde{y}_3, \tilde{y}_4$. Les quatre couples de points $(x_i, \tilde{y}_i)$ pourraient alors être représentés sur une droite de pente β et d'ordonnée à l'origine α. Cependant la perturbation ne permet pas d'observer les valeurs $\tilde{y}_i$. A la place, ce que l'on observe ce sont des valeurs y_i qui diffèrent des valeurs $\tilde{y}_i$, de façon telle que les couples (x_i, y_i) ont peu de chances d'être alignés.



Les points Q1 à Q4 ne sont pas observés. Ce qui l'est ce sont les points P1 à P4. Pour cette raison la valeur exacte des paramètres à estimer est inconnue et le restera. Le travail de l'économètre dans ce modèle simple est de tracer une droite qui, tout en ne connaissant pas la localisation de la véritable droite, doit être aussi proche que possible de celle-ci. Cette droite est appelée droite de régression.

Comment procéder ? Pour le comprendre nous devons tout d'abord introduire la notion de résidu. Dès lors qu'il y a plus de deux observations, ce n'est que par coincidence qu'il est possible de trouver une droite qui passe par tous les couples d'observations.



Par exemple, dans le graphique ci-dessus, pour la valeur de X observée x_1 l'observation de Y est égale à y_1, mais cette valeur n'est pas sur la droite de régression. Cette droite conduit à une valeur différente de Y, appelée la prédiction de Y, égale à $\hat{y}_1$. Le résidu est l'écart entre la valeur observée de Y et sa valeur prédite. C'est en fait une estimation de la perturbation. On le note donc $\hat{\epsilon}_i$. Dans le cas présent, pour la première observation: $\hat{\epsilon_1} = y_1 - \hat{y}_1$.

Il est clair que la meilleure adéquation de la droite de régression à l'ensemble des couples d'observations est obtenue lorsque les résidus de l'estimation sont les plus faibles possibles. Il est également clair que si l'on positionne la droite de régression de façon à ce que certains résidus soient nuls, les autres résidus seront probablement grands en valeur absolue. Il faut donc trouver une façon de positionner la droite qui tienne compte de la taille de l'ensemble des résidus simultanément. La première idée est de chercher à minimiser la somme des résidus. Cependant ce n'est pas une bonne idée parce que certains résidus sont positifs et d'autres négatifs, de sorte que leur somme peut s'annuler alors même que l'adéquation de la droite aux observations est mauvaise. En fait il est facile de montrer que la somme des résidus est nulle lorsque l'estimation de $\alpha, \hat{\alpha},$ est égale à $\bar{y}$ et l'estimation de $\beta, \hat{\beta},$ est nulle. La bonne façon d'aborder le problème est de minimiser la somme des carrés des résidus: $SCR = \hat{\epsilon_1} + \hat{\epsilon_2} + \hat{\epsilon_3} + \hat{\epsilon_4}$.

2. Un exemple simple avec deux puis avec trois points.

Supposons d'abord que quand x=1, y=3 et quand x=2, y=5. Notre objectif est d'estimer les coefficients (paramètres) α et β. Selon le modèle économétrique, quand:

- x = 1, $\hat{y} = \hat{\alpha} + \hat{\beta}$ et $\hat{\epsilon} = y - \hat{y} = 3 - \hat{\alpha} - \hat{\beta}$
- x = 2, $\hat{y} = \hat{\alpha} + 2\ \hat{\beta}$ et $\hat{\epsilon} = 5 - \hat{\alpha} - 2\ \hat{\beta}$

La somme des carrés des résidus est alors égale à:

$$SCR = (3 - \hat{\alpha} - \hat{\beta})^2 + (5 - \hat{\alpha} - 2\ \hat{\beta})^2$$

Il faut trouver $\hat{\alpha}$ et $\hat{\beta}$ qui minimisent cette somme. Deux conditions (dites conditions du premier ordre) doivent être vérifiées:

- $\partial SCR / \partial \hat{\alpha} = 4\ \hat{\alpha} + 6\ \hat{\beta} - 16 = 0$
- $\partial SCR / \partial \hat{\beta} = 6\ \hat{\alpha} + 10\ \hat{\beta} - 26 = 0$

Ces deux conditions sont non seulement nécessaires, mais aussi suffisantes pour assurer que les

coefficients $\hat{\alpha}$ et $\hat{\beta}$ qui les vérifient conduisent effectivement au minimum de la SCR. Le système conduit ici à: $\hat{\alpha}=1$ et $\hat{\beta}=2$. Comme il y a forcément une droite qui passe par deux points distincts, dans le cas présent les résidus estimés sont nuls, de sorte que la SCR est égale à 0, au minimum. Dès lors qu'il y a plus de deux points cette situation devient tout à fait exceptionnelle. Supposons donc maintenant qu'un troisième point est ajouté à nos deux observations: quand x=3, y=6. Le résidu de cette troisième observation s'écrit: $\hat{\epsilon}=6-\alpha-3\ \hat{\beta}$. Le problème est maintenant de trouver $\hat{\alpha}$ et $\hat{\beta}$ qui minimisent:

$$SCR=(3-\hat{\alpha}-\hat{\beta})^2+(5-\hat{\alpha}-2\ \hat{\beta})^2+(6-\hat{\alpha}-3\ \hat{\beta})^2$$

Les conditions du premier ordre s'écrivent:

- $\partial SCR/\partial\hat{\alpha}=6\ \hat{\alpha}+12\ \hat{\beta}-28=0$
- $\partial SCR/\partial\hat{\beta}=12\ \hat{\alpha}+28\ \hat{\beta}-62=0$

La résolution du système conduit à: $\hat{\alpha}=1{,}67$ et $\hat{\beta}=1{,}50$.

3. Le cas général.

Considérons maintenant le cas général où le nombre d'observations est égal à n. Dans ce cas la somme des carrés des résidus est:

$$\begin{aligned} SCR&=\hat{\epsilon}'\hat{\epsilon}\\ &=\sum_{i=1}^{n}\hat{\epsilon_i}^2\\ &=\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2 \end{aligned}$$

où la première écriture résulte des règles de multiplication des vecteurs $(\hat{\epsilon}=\begin{pmatrix}\hat{\epsilon_1}\\ \vdots\\ \hat{\epsilon_n}\end{pmatrix})$. Les conditions du premier ordre s'écrivent:

- $\partial SCR/\partial\hat{\alpha}=-2\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)=-2\sum_{i=1}^{n}\hat{\epsilon}_i=0$
- $\partial SCR/\partial\hat{\beta}=-2\sum_{i=1}^{n}x_i(y_i-\hat{\alpha}-\hat{\beta}x_i)=-2\sum_{i=1}^{n}x_i\hat{\epsilon}_i=0$

Ce système se simplifie et on peut y faire apparaître facilement des expressions connues d'estimateurs empiriques:

- $\hat{\alpha}+\bar{x}\hat{\beta}-\bar{y}=0$
- $n\bar{x}\hat{\alpha}+(\sum_{i=1}^{n}x_i^2)\hat{\beta}-\sum_{i=1}^{n}x_iy_i=0$

Ce qui conduit finalement à:

$$\hat{\alpha} = \bar{y} - \bar{x}\,\hat{\beta}$$

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \bar{x}^2} = \frac{cov(x, y)}{V(x)}$$

Les estimateurs $\hat{\alpha}$ et $\hat{\beta}$ sont appelés estimateurs des moindres carrés ordinaires.

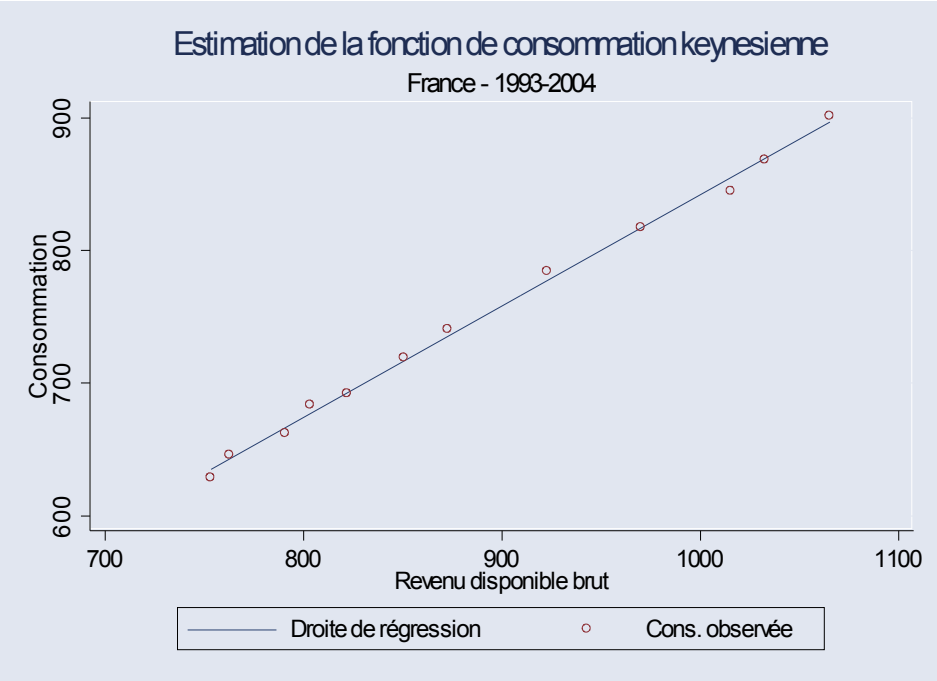
4. Un exemple illustratif.

Les données ci-dessous représentent les dépenses de consommation et le revenu disponible des ménages français entre 1993 et 2004:

Année	***Revenu disponible brut***	***Consommation***
1993	753.4	628.1
1994	762.9	645.3
1995	790.9	661.5
1996	803.5	682.8
1997	822.3	691.5
1998	850.8	719.1
1999	872.8	739.9
2000	923	783.9
2001	970.4	817.4
2002	1015.5	844.4
2003	1032.9	868
2004	1065.6	901.2

La régression de la consommation sur le revenu disponible brut conduit aux résultats suivants:

$$Consommation\ prédite = 2{,}89 + 0{,}839\ RDB$$

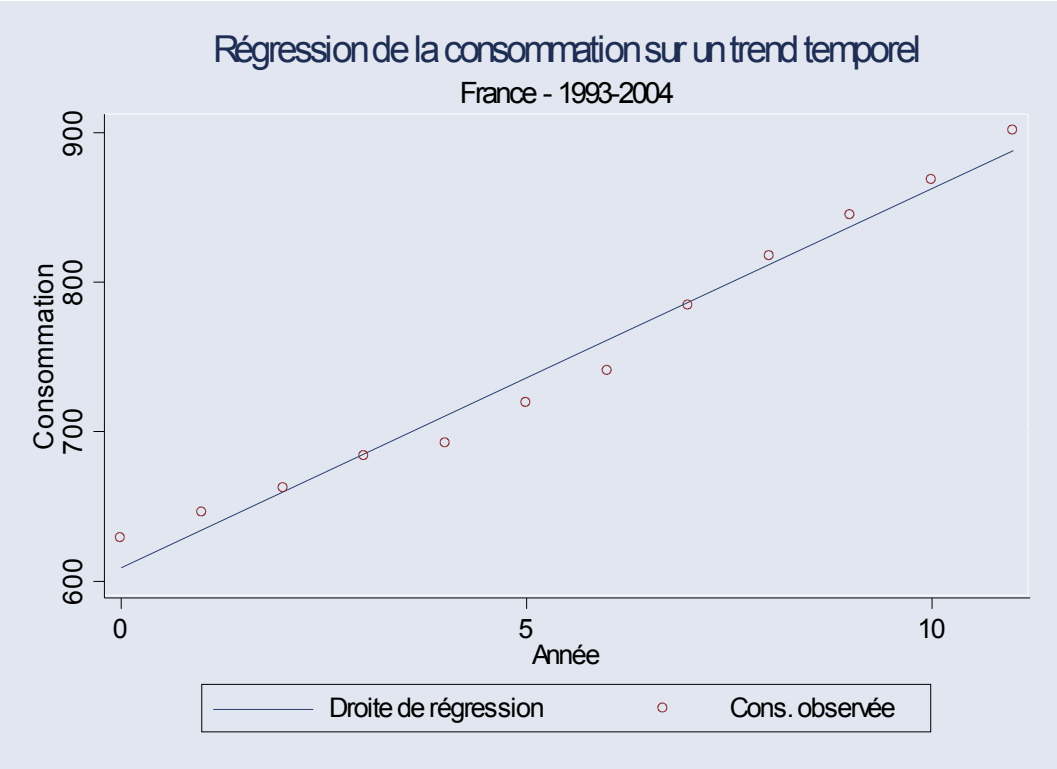


La figure ci-dessus montre les observations et la droite de regression. Comment interpréter les résultats ? Le coefficient $\hat{\beta}$ indique que lorsque le revenu disponible brut (RDB) des ménages français augmente de 1 euro, ceux-ci, en moyenne, augmentent leur consommation d'un peu moins de 84 centimes. Quelle interprétation donner au coefficient $\hat{\alpha}$ (que l'on appelle la constante du modèle) ? Au sens strict, ce coefficient indique la valeur de la consommation lorsque le revenu disponible brut des ménages est nul. Dans le cas présent celle-ci vaudrait 2,89 milliards d'euros, un montant très faible. Cependant cette interprétation n'a pas ici beaucoup de sens car le revenu disponible brut des ménages n'est jamais nul. Lorsque, à partir de cette estimation, on tire la conclusion que si le revenu disponible brut des ménages était nul, on observerait ce niveau de consommation, on fait l'hypothèse que si l'on avait à notre disposition des observations avec de très faibles valeurs du RDB le comportement de consommation des ménages resterait le même, ce qui est très loin d'être acquis. Par exemple on peut imaginer que dans ce cas les ménages consommeraient la totalité de leur revenu. En fait, dans le cas présent, la constante n'a pas d'interprétation directe, elle sert juste à placer la droite de régression plus ou moins haut dans le graphique.

Supposons maintenant que l'on effectue la régression de la consommation non plus sur le RDB, mais sur un « trend » temporel, c'est à dire une variable t qui vaut 0 en 1993, 1 en 1994, 2 en 1995 etc. Ceci conduit aux résultats suivants:

$$\textit{Consommation prédite} = 609 + 25{,}4 * t$$

Cette fois le coefficient de t indique que chaque année, en moyenne, la consommation des ménages augmente de 25,4 milliards d'euros. La constante ici admet une interprétation: quand t = 0 cela signifie que l'on est en 1993. La valeur de $\hat{\alpha}$ est donc simplement la valeur prédite de la consommation en 1993 (voir graphique ci-dessous).



L'adéquation de la seconde droite de régression aux données semble moins bonne que la première. Pour en être sûr il faut disposer d'une mesure de cette adéquation. C'est l'objet du prochain paragraphe.

5. Mesure du pouvoir explicatif du modèle: le R^2.

Souvenez-vous que dans notre modèle la variable à gauche du signe « = », Y, est appelée variable expliquée et que la variable à droite, X, reçoit le nom de variable explicative. La raison en est que lorsque l'on régresse Y sur X, ce que l'on cherche à faire est d'expliquer comment les variations de X déterminent celles de Y. Il est très important de bien réaliser que si X prenait toujours la même valeur, X ne pourrait pas contribuer à comprendre pourquoi Y varie. De même si Y prenait toujours la même valeur, les variations de X ne pourraient pas aider à comprendre pourquoi Y reste constant (et d'ailleurs y aurait-il encore quelque chose à expliquer ?)

La question qui se pose est donc la suivante: quelle proportion de la variation de Y est expliquée par celle de X, selon notre modèle ? Ce que le modèle doit expliquer est la variance empirique de y observée dans l'échantillon, à savoir: $V(y)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2=\frac{1}{n}SCT$ où SCT est la somme des carrés totaux. Pour voir ce qu'explique le modèle une fois estimé, remarquons que:

$$\begin{aligned} y_i &= \hat{y}_i+\hat{\epsilon}_i \\ &= \hat{\alpha}+\hat{\beta}\,x_i+\hat{\epsilon}_i \\ &= \bar{y}-\hat{\beta}\,\bar{x}+\hat{\beta}\,x_i+\hat{\epsilon}_i \end{aligned}$$

si l'on soustrait $\bar{y}$ des deux côtés de l'équation on obtient:

$$y_i-\bar{y}=\hat{\beta}(x_i-\bar{x})+\hat{\epsilon}_i=\hat{y}_i-\bar{y}+\hat{\epsilon}_i$$

En faisant la somme pour obtenir la SCT à gauche de l'équation on trouve:

$$
\begin{aligned}
SCT &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{\epsilon}_i^{\,2} \\
&= \hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} \hat{\epsilon}_i^{\,2} \\
&= SCE + SCR
\end{aligned}
$$

L'obtention de ce résultat n'est pas immédiate et résulte des conditions du premier ordre dont la résolution conduit aux estimateurs des MCO. En effet:

$$
\begin{aligned}
\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} (\hat{\beta}(x_i - \bar{x}) + \hat{\epsilon}_i)^2 \\
&= \sum_{i=1}^{n} (\hat{\beta}^2 (x_i - \bar{x})^2 + \hat{\epsilon}_i^{\,2} + 2\ \hat{\beta}(x_i - \bar{x})\hat{\epsilon}_i) \\
&= \sum_{i=1}^{n} \hat{\beta}^2 (x_i - \bar{x})^2 + \sum_{i=1}^{n} \hat{\epsilon}^2 + 2\ \hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})\hat{\epsilon}_i
\end{aligned}
$$

Le dernier terme du membre de droite est nul lorsque les conditions du premier ordre sont vérifiées (revoyez ces conditions pour vous en convaincre).

La somme $\hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$ est appelée somme des carrés expliquée et notée SCE. L'explication de cette appelation réside dans l'égalité: $\hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ de sorte que la SCE apparaît comme (n fois) la variance empirique de la prédiction de y (il est facile de vérifier que $\bar{\hat{y}} = \bar{y}$).

Nous sommes maintenant en mesure d'introduire la mesure du pouvoir explicatif de la régression.

Définition: R² ou coefficient de détermination.

Le R² est égal à: $R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$

Ce coefficient est toujours compris entre 0 et 1. Plus il est proche de 1 et plus le pouvoir explicatif de la régression est élevé.

Illustration: reprenons maintenant les deux régressions montrées en exemple. Laquelle est la meilleure ? C'est la première: son R² vaut 0,9967 (ce qui est très élevé, mais guère surprenant étant donné la nature des observations), alors que celui de l'autre ne vaut « que » 0,9801. L'impression visuelle était donc la bonne.