

L'économétrie pour les nuls : Introduction

Ah l'économétrie ! La hantise de plusieurs milliers d'étudiants en école de commerce et en fac chaque année. Et pourtant, comprendre la base de l'économétrie et être capable d'analyser les résultats d'une étude simple ne relève pas de l'impossible. Petit geek de l'économétrie à ses heures perdues, le Captain' va essayer de démystifier pour vous cette matière, dans un dossier spécial "L'économétrie pour les nuls", dont voici le premier opus.

"L'économétrie est un outil d'analyse quantitative, permettant de vérifier l'existence de certaines relations entre des phénomènes économiques et de mesurer concrètement ces relations sur la base d'observations de faits réels" (Eric Dor, "Econométrie"). L'économétrie se base donc sur les mathématiques et les statistiques afin d'identifier des relations entre différentes variables. Mais tout comme vous n'avez pas besoin d'être un pro en microprocesseur pour utiliser un ordinateur, il n'est pas foncièrement nécessaire d'être un crack en math pour pouvoir se servir de l'économétrie comme outil d'analyse.

Commençons donc avec un modèle simple à deux variables. Vous souhaitez tester s'il existe une relation linéaire entre les variables y et x . Pour le poser sous la forme d'une équation, votre but est d'estimer les coefficients ? de l'équation ci-dessous (avec un terme d'erreur mais que je ne présente pas dans cette intro pour plus de simplicité).

Je vous ai dit qu'il ne fallait pas forcément être un crack en math, mais il faut bien en faire un minimum tout de même. Le modèle ci-dessus est un modèle du niveau mathématique de classe de seconde, le fameux $y = f(x)$.

Pour prendre un exemple, supposons que vous ayez envie de voir le lien entre le poids d'une voiture et son prix de vente. Avant de faire votre étude, vous avez donc créé une base de données avec de nombreuses infos sur les voitures, qui pourrait ressembler à celle-ci (si vous avez envie de vous amuser avec cette base, elle est disponible au format Stata ici "auto.dta").

En réalisant un graphique avec en abscisse le poids d'une voiture (weight) et son prix (price) pour chacune des 74 voitures de notre base de données, cela donne ça (chaque point correspondant à une voiture):

Il semble bien qu'il y ait une relation croissante entre le poids d'une voiture et son prix. Graphiquement on remarque en effet une tendance: plus une voiture est lourde, plus son prix est élevé (en moyenne). C'est bien beau tout cela, mais si vous êtes chargé d'étude dans l'industrie automobile et que vous allez voir votre patron en lui disant cela, il va vous dire "tu es bien sympa mon petit, mais il me faut des chiffres précis et être sur que cette relation est significative et n'est pas le fruit du hasard".

Et là, deux solutions s'offre à vous. (1) La technique de l'autruche: vous laissez traîner le problème et inventez des chiffres au hasard en faisant des calculs à l'arrache sur Excel. (2) Lire le Captain', ouvrir un

logiciel économétrique et faire une étude simple de régression linéaire.

Et aussi incroyable que cela puisse paraître, la seconde solution est beaucoup plus simple et rapide que la première. Il suffit d'un seul calcul dans un logiciel spécialisé (type Stata, Eviews, Scilab, SAS...) pour avoir l'ensemble des réponses aux questions de votre patron. C'est là qu'intervient la magie de la "régression linéaire et de la méthode des moindres carrés ordinaire". Sans rentrer dans les détails, le but de cette technique est de tracer une droite sur le graphique précédent, tel que l'écart (au carré) entre les points du nuage et votre droite de régression soit le plus faible possible.

Vous pouvez le faire avec une règle et un crayon de bois, et tâtonner en calculant pour chaque point l'écart entre les points et votre droite. Ou bien utiliser une fonction automatique sur un logiciel économétrique. Dans Stata (le logiciel utilisé par le Captain'), la commande "regress price weight" indique que vous souhaitez faire une régression linéaire entre votre variable y (price) et votre variable x (weight). Et là pas de panique, le Captain' va vous apprendre à lire les résultats de ces tableaux blindés de chiffres.

Votre but était donc d'estimer quelle est la relation entre le poids d'une voiture et son prix, en déterminant les coefficients β_0 et β_1 de votre équation initiale. Pour rappel des mathématiques de classe de seconde, votre coefficient β_0 va correspondre à l'ordonnée à l'origine et votre coefficient β_1 à la pente de la droite de régression. Sur le tableau ci-dessus, vous avez ces deux informations; β_0 correspondant au coefficient de la constante (_cons) et bleu, soit -6,70 et β_1 correspondant au coefficient devant le poids (weight) en rouge, soit 2,05. Votre équation initiale peut donc s'écrire: " price = -6,70 + 2,05 * weight "

En moyenne donc, et en se basant sur les données de votre échantillon, une voiture de 2000 livres (lbs) coûtera environ "-6,70 + 2,05 * 2000 = 4000 dollars). Si l'on reprend notre graphique précédent, on peut alors tracer la droite de régression linéaire, qui est la droite la plus proche possible des différents points. On voit bien que si l'on se base sur une voiture de 2000 lbs, son prix estimé via la droite rouge est proche de 4000 dollars.

Mais cette relation est-elle significative? Pour cela, il convient de regarder le "t-stat" de notre tableau de régression (en vert), et de comparer la valeur de ce "t-stat" (dans notre cas 5,42) à des valeurs déterminées statistiquement. Il est possible de ne pas rentrer dans les calculs, et uniquement de se baser sur cette règle "si le t-stat est supérieur à 1,96 en valeur absolue, alors la variable est significative". Ici c'est le cas, donc vous pouvez aller voir votre patron en lui disant "eh chef, une voiture qui pèse 1 livre de plus qu'une autre coûtera en moyenne 2,05 dollars de plus, et la relation est significative". Si vous voulez même faire le mariole, vous pouvez lui dire "[...] et cette relation est significative avec un intervalle de confiance de 95%", la valeur de t-stat de 1,96 correspondant à cet intervalle de confiance.

Bien sûr ici, on voit bien qu'il y a de gros écarts entre la droite de régression en rouge et de nombreux points. Ceci s'explique car le modèle est ultra-simplifié (une seule variable explicative). Il est d'ailleurs possible d'estimer le pourcentage de la variation de prix expliqué par la variation de poids, en regardant le R-squared et l'adjusted R-squared (en jaune dans le tableau). Dans notre exemple, on trouve que 29% de la variation de prix peut s'expliquer par une variation de poids (et donc que 71% de la variation de prix provient de variables omises dans notre modèle ou bien de facteurs aléatoires).

Conclusion: Notre exemple peut mettre en avant un premier piège de l'économétrie: la différence entre causalité et de corrélation. Par exemple, si vous rajoutez discrètement des parpaings dans le coffre de votre voiture, le prix devrait selon notre modèle augmenté (ce qui n'a aucun sens). La relation estimée

entre le poids d'une voiture et son prix est davantage une relation de corrélation que de causalité. Ce n'est pas le poids à proprement parlé qui entraîne une hausse du prix, c'est le fait que le poids est souvent relié à d'autres variables (taille du moteur, nombre de portes, confort...) qui elles ont réellement une relation de causalité avec le prix d'une voiture. Rendez-vous la semaine prochaine pour la suite de notre dossier "L'économétrie pour les nuls", où nous verrons un modèle à plusieurs variables explicatives permettant de tester cela et de faire des analyses plus complexes.

Dossier l'économétrie pour les nuls par le Captain':

- Chapitre 1 : Introduction
- Chapitre 2 : La régression linéaire
- Chapitre 3 : (en cours de rédaction)