L'économétrie pour les nuls : La régression linéaire

Après la première partie la semaine dernière "L'économétrie pour les nuls : Introduction" (si vous ne comprenez rien à l'économétrie, commencez donc par cet article), le Captain' continue donc aujourd'hui ce dossier en vous expliquant plus en détail le principe de la régression linéaire à plusieurs variables, en introduisant en plus la notion de variable indicatrice (dummy variable in english).

Nous allons continuer avec notre modèle qui essaye d'expliquer le prix d'une voiture (notre variable dépendante "Y") en fonction de variables explicatives (X1, X2, D1). La variable X1 correspond comme dans l'article précédent au poids d'une voiture. La variable X2 représente la consommation en carburant de la voiture et la variable D1 est une variable indicatrice, prenant la valeur 0 si la voiture est une voiture domestique (donc américaine dans notre exemple) ou 1 si la voiture est une voiture étrangère. Notre modèle ressemble donc à cela:

Le dernier terme ? correspond au terme d'erreur, qui représente la déviation entre ce que le modèle prédit et la réalité. Comme précédemment notre but ici va être de déterminer (1) les variables significatives, c'est à dire voir si les différents coefficients sont différents de 0, (2) la valeur de la constante alpha et des différents coefficients "beta" qui permettent de minimiser l'erreur entre notre droite de régression linéaire estimée et les valeurs réelles de Y et enfin (3) la précision de notre modèle, en utilisant, entre autre, le "R-squared".

Le raisonnement est le même qu'avec seulement une variable, sauf qu'il est difficile de travailler graphiquement avec plusieurs variables. En effet, une régression linéaire à une variable explicative peut s'expliquer dans un graphique en 2D (avec en abscisse la variable X et en ordonnée la variable Y). Lorsque l'on passe à 3 variables explicatives, il faudrait montrer cela dans un environnement en 4 dimensions (une dimension par variable explicative + une dimension pour la variable dépendante).

L'introduction d'une variable indicatrice "D1" ne doit pas vous perturber. Cela permet de travailler avec des variables qualitatives à la place de variables quantitatives, en codant la variable qualitative sous la forme binaire (par exemple "homme = 0 et femme = 1). Il s'agit simplement d'une variable qui permet de voir l'impact de la provenance d'une voiture sur son prix. Par exemple, et étant donné que D1 = 0 si la voiture est américaine et D1 = 1 si la voiture est étrangère, une valeur positive du coefficient ?3 signifierait simplement que les voitures étrangères coûtent en moyenne plus cher que les voitures américaines.

Allez, lançons Stata, et exécutons cette régression.

Etape 1: Tester la significativité des variables. Pour cela, il suffit de regarder le "t-stat" (t) ou bien la P-value (P>?t?), et comparer ces valeurs à des "valeurs seuils". Pour faire simple, une variable est significative avec un intervalle de confiance de 95% si son t-stat est supérieur à 1,96 en valeur absolue, ou bien si sa P-value est inférieure à 0,05. Dans notre exemple, on voit que la variable "mpg", qui correspond à la consommation en essence de la voiture n'est pas significative (t-stat trop faible en valeur absolue et P-value trop forte). De plus, l'intervalle de confiance à 95%, allant de -126.17 à 169.99

comprend la valeur 0. Il est donc impossible de rejeter l'hypothèse ?2 = 0.

Les deux autres variables "weight" et "foreign" sont significatives (t-stat de 5,49 et 5,37 donc supérieur à la valeur seuil de 1,96). De plus, l'intervalle de confiance ne comprend pas la valeur 0. Pour ?1 par exemple, l'intervalle de confiance permet de dire "je suis sûr à 95% que la valeur de ?1 se trouve entre 2,20 et 4,72. Le coefficient (=3.467 pour ?1 par exemple) correspond exactement au milieu de l'intervalle de confiance de la variable.

Mais on fait quoi maintenant qu'on a trouvé que la variable "consommation de la voiture" n'est pas significative? Et bien on relance la régression, mais en supprimant la variable. En effet, les résultats de la régression peuvent être modifié par l'inclusion de variables non significatives, et il est donc préférable d'analyser le résultat d'une régression finale contenant uniquement des variables significatives. Voici donc le résultat de notre nouvelle régression.

Etape 1 (nouvelle régression): C'est bon, nos deux variables sont significatives (t-stat > 1,96 en valeur absolue).

Etape 2: Étude des coefficients. La valeur estimée de ?1 est égale à 3,32 et celle de ?3 à 3637. Comment lire cela? Cela signifie que "toutes choses égales par ailleurs", une voiture pesant une livre (unité de masse américaine) de plus, coûtera en moyenne 3,32 $ de plus. Même raisonnement en ce qui concerne l'analyse du coefficient de notre variable indicatrice; "toutes choses égales par ailleurs", une voiture étrangère coûte en moyenne 3637 dollars de plus qu'une voiture américaine.

Etape 3: Mais quelle est la précision de notre modèle ? Pour cela, il est possible de regarder le "R-squared", qui mesure la proportion de la variance de Y (variable dépendante) qui est expliquée par la variation des toutes les variables explicatives. Le R-squared est par construction compris entre 0 et 1 ; plus on se rapproche de 1, plus le modèle est précis. Dans notre exemple, 49% de la variation de Y peut-être expliquée par les variations de X1 et D1. En gros, c'est pas mal mais pas terrible terrible non plus. Il manque en effet de nombreuses variables à notre modèle pour que celui ci permettent d'estimer avec précision le prix d'une voiture en fonction de ses caractéristiques.

Il n'existe pas de valeur du R-squared à partir de laquelle le modèle peut-être considéré comme bon ou mauvais (cela dépend du modèle). Pour donner un ordre d'idée dans cette situation (ne mettez surtout pas ça dans vos exams d'économétrie), un R-squared proche de 0,8 est signe d'un bon modèle, tandis que si votre R-square est proche de 0,2 , c'est pas la folie (peut-être pas mal de variables omises).

Conclusion: Et voilà, en seulement deux articles, vous êtes désormais capable de faire une analyse économétrique "en coupe instantanée", c'est à dire avec un échantillon à un moment donné "t" unique. Nous allons voir au prochain article comment procéder dans le cas d'une analyse en série temporelle, pour étudier les variations d'une variable dans le temps (par exemple la relation entre le cours du pétrole, le CAC40 et la croissance mondiale entre 2000 et 2012). Et là, ça va se compliquer un peu..

Dossier l'économétrie pour les nuls par le Captain' :
- Chapitre 1 : Introduction
- Chapitre 2 : La régression linéaire
- Chapitre 3 : (en cours de rédaction)