Econométrie et analyse de données

Cours

L'économétrie permet de comprendre la relation qui existe (ou non) entre un phénomène et les variables qui peuvent l'influencer.
x est la variable **explicative** ou **indépendante**.
y est la variable **expliquée** ou **dépendante**.

On recueille les données de nature qualitative et on les représente sous forme de nuage de points ➔ l'objectif est de trouver la droite qui représente le mieux la **direction** ou la **relation** entre les différents points.

Définitions :

→ **Moyenne** : mesure la tendance centrale la plus importante : $\bar{x} = \frac{\sum xi}{n}$

→ **Médiane** : on classe les données par ordre croissant, la médiane correspond à la valeur centrale. (Si nombre de données impair, la médiane est la valeur centrale, si nombre de données pair, on prend la moyenne des deux valeurs centrales)

→ **Mode** : valeur de l'observation qui a la plus grande fréquence

→ **Percentile** : p^1 est la valeur telle que au moins P% des observations ont une valeur inférieure ou égale à cette valeur, et au moins (100-P)% ont une valeur supérieur ou égale à cette valeur. Ex : quartile.

→ **Variance** : C'est la mesure de dispersion. Basée sur la différence entre la valeur de chaque opération xi et la moyenne $\bar{x}$. La différence entre chaque observation xi et la moyenne $\bar{x}$ est appelée écart par rapport à la moyenne.
Variance de la population : $\sigma^2 = \frac{\Sigma\, x_i - \bar{x}}{N}$

Variance de l'échantillon : $s^2 = \frac{\Sigma\, x_{i-\bar{x}^2}}{n-1}$

→ **Ecart type** : racine carrée de la variance.
Pour l'échantillon : $s = \sqrt{s}^2$
Pour la population : $\sigma = \sqrt{\sigma}^2$

Régression linéaire :

$y_i = \beta_0 + \beta_1 x_{i,1}$

Régression multiple :

$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_k + u_i$
$i = 1, \ldots, I$

- ➔ y variable dépendante ou à expliquer
- ➔ x_1, x_2 variable indépendante ou à expliquer

} Variable endogène

- ➔ β_1, β_2 sont les paramètres à estimer
- ➔ u_i mesure la différence entre les valeurs réellement observées de la variable dépendante et les valeurs qui auraient été observées si la relation spécifiée avait été rigoureusement exacte → marge d'erreur de spécification (la seule variable explicative n'est pas suffisante pour rendre compte de la totalité du phénomène expliqué), erreur de mesure (données ne représentent pas exactement le phénomène), erreur de fluctuation d'échantillonnage (d'un échantillon à l'autre les observations et donc les estimations sont légèrement différentes).
- ➔ β_0 est la constante

Estimateur de Gauss-Markov : sous les hypothèses, $\hat{\beta}_0$ et $\hat{\beta}_1$ sont les meilleurs estimateurs linéaires, aucun autre estimateur linéaire sans bien n'aura de variance inférieure. Propriété : estimateurs sans biais et variance minimale.

Econométrie et analyse de données

L'absence de biais : s'il était possible de répéter l'échantillon un nombre infini de fois, on obtiendrait la vraie estimation en moyenne.
Variance minimale : sans biais : si la distribution de $\hat{\beta}$ a une petite variance, l'estimation a plus de chance d'être proche de la moyenne. Variance de la distribution peut être réduite si on augmente la taille de l'échantillon.

La droite recherchée qui serait la plus pertinente serait celle qui passerait le plus près des observations, c-à-d celle qui minimiserait l'écart entre les observations et les points de la droite recherchée. → C'est l'objet de la **méthode des moindres carrées** (MCO).

MCO :

Minimiser le carré de la différence : $(y_i - (\hat{\beta}_1 x_1 + \hat{\beta}_0))^2$

Estimateur du coefficient : $\hat{\beta} = \frac{\Sigma_{i=1}^{I} (x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^{I} (x_i - \bar{x})^2}$

Estimateur des MCO est un estimateur de Gauss-Markov :

Estimateur MCO est **BLUE** (Best (de variance minimale) Linear Unbiased Estimator). Estimateurs sans biais, convergents, etc, relatif à la fonction de densité des estimateurs.
Estimateur non biaisé : estimateur donc la fonction de répartition est centrée sur la vraie valeur du paramètre que l'on cherche à estimer.

Propriétés :
Distribution de la moyenne statistique : fonction de densité centrée sur la moyenne de la population mais dont la variance devient plus petite lorsque l'échantillon devient plus grand. Alors que l'échantillon devient plus grand, la fonction de répartition se déplace en rapprochant ainsi la moyenne de la vraie valeur du paramètre estimé → **distribution asymptotique**.
Les meilleurs estimateurs sont ceux qui sont **efficaces** → estimateur dit efficace s'il est celui qui a la + petite variance. Car la variance de l'estimateur est également celle de l'erreur d'estimation. Cela revient à minimiser la variance de l'erreur.

Convergence des estimateurs : condition supplémentaire qui garantit que les erreurs d'estimations tendent vers 0 lorsque l'on augmente indéfiniment le nombre d'observations.

Types de variables à analyser :

Variables **continues** : âge, chiffres d'affaires…
Variables **discrètes** : genre d'un individu (homme ou femme)… ➔ très utilisées car elles paramètres de considérer les valeurs anormales, ou intégrer la saisonnalité, facteurs qualificatifs ou la caractérisation d'un individu.
Variable **qualitative** : couleurs des cheveux…

Types de données :

Séries temporelles : il s'agit de variables observées à intervalles de temps réguliers. *Exemple* : données financières comme le cours des actions.
Coupe instantanée : on observe un phénomène à un instant donné.
Panel : la variable représente les valeurs prises par un échantillon d'individus à intervalles réguliers, donc il y a deux dimensions (temporelle et en coupe)
Cohorte : comme le panel mais les individus ne sont pas les mêmes dans le temps

Corrélation :

Coefficient de corrélation : mesure la relation entre 2 variables (mais par la causalité entre des variables/phénomènes)

$$\rho_{xy} = \frac{\sum_{i=1}^{I} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{I} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{I} (y_i - \bar{y})^2}}$$

Econométrie et analyse de données

La corrélation entre 2 variables mesure leur intensité de liaison.

- ➔ Corrélation **positive** : si y augmente, x augmente aussi. Nuage de points très regroupés : forte corrélation, proche de 1.
- ➔ Corrélation **négative** : si y augmente, x diminue. *Exemple* : nbre d'enfants d'une femme, effet sur son salaire.
- ➔ **Pas de corrélation** : égale à 0, les deux variables n'ont quasiment pas de liens entre elles, x n'influence pas y.

Représente la relation entre deux variables mais pas la causalité entre des variables/phénomènes.

Test Student et test de Fisher :

Test Student : habituellement utilisé pour tester les hypothèses qui portent sur les coefficients de la régression pris individuellement. Nécessite la connaissance de la moyenne et de l'écart type.
Test du Fisher : on l'utilise pour les tests sur un ensemble de coefficients

Approche empirique, comment procéder :

Question de recherche, analyse des données, statistiques descriptives, corrélation et test de chi-2, régression, test d'autocorrélation. Attention à la multicollinéarité.

Test d'hypothèses :

Objectif : montrer qu'un échantillon bien particulier se confirme à certaines hypothèses.

- ➔ Spécification de l'hypothèses à tester
- ➔ Règle de décision à utiliser pour rejeter ou non l'hypothèses
- ➔ Type d'erreur qui peut être rencontré en appliquant cette règle de décision

Erreur de type I : on rejette à tort l'hypothèse nulle
Erreur de type II : on accepte à tort l'hypothèse nulle

Hypothèse nulle : égale aux valeurs que le paramètre doit prendre si l'hypothèse n'est pas correcte
Hypothèse alternative : spécification du type de valeur qui devrait prendre le paramètre si la théorie était correcte

Pour le Test Student :
Hypothèse testée, au seuil $\alpha\%$:
$H_0 : \beta_1 = 0$
$H_1 : \beta_0 \neq 0$
Si la valeur absolue du t calculé est supérieur à la valeur critique, on rejette l'hypothèse nulle. Dans le cas contraire, on accepte l'hypothèse nulle.

Règle de décision :
Si Test Student inférieur ou supérieur au seuil, on rejette l'hypothèse H_0 (donc on accepte H_1).

Econométrie et analyse de données

TD

Méthode calcul Test de Student :

- Calcul de la somme des x et la somme des y
- Calcul de la moyenne des x ($\bar{x}$) et de la moyenne des y
- Calcul de la moyenne au carré des x et des y
- Noter le minimum et le maximum des x et des y
- Calcul du coefficient de corrélation

=COEFFICIENT.CORRELATION(XX:XX;YY:YY)

$$\rho_{xy} = \frac{\sum_{i=1}^{I}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{I}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{I}(y_i - \bar{y})^2}}$$

- Calcul de $x - \bar{x}$ et de $(x - \bar{x})^2$ pour chaque donnée
- Calcul de $y - \bar{y}$
- Calcul de $(x - \bar{x})(y - \bar{y})$ pour chaque donnée
- Calcul de $\hat{\beta}$: somme des $(x - \bar{x})(y - \bar{y})$ divisée par la somme des $(x - \bar{x})^2$ (pour chaque donnée)
- Calcul de la constante : $\bar{y} - (\hat{\beta} \times \bar{x})$
- Calcul de $\hat{y}$ (pour chaque donnée) : constante+$\hat{\beta} \times x$
- Calcul du résidu pour chaque donnée : $y - \hat{y}$
- Calcul du résidu au carré pour chaque valeur
- Calcul de la somme des résidus au carré
- Calcul de la variance de l'erreur : $\frac{résidu\ au\ carré}{10-2}$ (car 10 valeurs dans l'exercice)
- Calcul de la variance estimée : $\frac{variance\ erreur}{somme\ (x-\bar{x})^2}$
- Calcul de la racine carrée de la variance estimée
- Calcul du test de student : $\frac{\hat{\beta}}{racine\ carrée\ variance\ estimée}$

Bon courage ! 💖