INTRODUCTION A L'ETUDE DES VARIABLES QUALITATIVES

Plan

- Introduction
- Définition
- Catégories de variables qualitatives
- Modèles pour Données avec Troncature
- Les Modèles pour Données Censurées
- Définition de Troncature et Censure
- Les Modèles à variables endogène Dichotomique
- Les Modèles de Régression Discrète
- Conclusion

Introduction

- L'objet de ce cours est de présenter des méthodes économétriques couramment utilisées dans les études portant sur des données d'enquête en coupe instantanée (cross-section).
- Les variables que l'on souhaite expliquer prennent parfois des valeurs discrètes en nombre fini (par exemple, zéro ou un). Dans ce cas, les méthodes économétriques utilisées sont connues sous les noms de logit et probit.
- Parfois encore, les variables à expliquer ne sont que partiellement observables, c'est à dire qu'elles sont tronquées ou censurées. Dans ce cas, les méthodes économétriques à utiliser sont connues sous le nom modèle Tobit.
- Enfin, les variables que l'on souhaite expliquer prennent parfois un nombre infini de valeurs discrètes, ce qui est notamment le cas lorsqu'elles constituent des données de comptage. Il existe des méthodes économétriques appropriées à ce type de données, dont la plus connue est fondée sur la loi de Poisson.

Définition

- La variable qualitative est une Variable qui ne peut être numériquement mesurée que par une échelle nominale ou une échelle ordinale. Cette appellation vient du fait que les nombres désignant les modalités de la variable sont le résultat d'un procédé de codage numérique arbitraire et non pas d'interprétation physique ou arithmétique concrète. Des opérations arithmétique courante comme calcule d'une somme, moyenne,..., n'ont pas de sens. Exemple du "statut matrimonial".
- Les modèles à variables dépendantes qualitatives sont très largement utilisés en micro-économie appliquée (modèle d'offre de travail, modèles de sélection endogène d'échantillon, expériences naturelles, credit scoring, ...) et le nombre d'applications en macro-économie appliquée ne cesse d'augmenter (modèle de déséquilibre...)

Catégories de variables qualitatives

Les modèles économétriques contenant des variables endogènes qualitatives peuvent être classés en trois catégories à savoir :

- Les Modèles pour Données avec Troncature
- Les Modèles pour Données Censurées
- Les Modèles à variables endogène Dichotomique

Modèles pour Données avec Troncature

- L’idée de ces modèles est l’existence d’un seuil au-delà duquel la variable dépendante n’est plus observable, ou n’a pas de signification économique en d’autres termes les variables dépendantes sont tronquées à un certain point, les observations de Y au dessus du seuil ne sont pas incluses.

Soit Y le revenu exprimé en fonction des années de scolarité X

La régression est la suivante : $Yi = B\,Xi + Ui$ (1)

Avec Ui est le résidu.

On observe Yi si $Yi \leq C$ où C est une constante donnée.

L’équation (1) devient $B\,Xi - Ui \leq C$

- On remarque que le résidu dépend de la variable explicative et par la suite l’application des MCO comme méthode d’estimation donne des estimateurs biaisés et non convergents d’où le recours à d’autres méthodes d’estimation,

Les Modèles pour Données Censurées

- Le modèle de régression le plus simple qui comprend une variable dépendante censurée est le modèle Tobit, une forme simple du modèle est la suivante :

Y*= XtB + Ut

Yt = Yt* si Yt*>0, Yt= 0 si non

Ici Y* est une variable latente observée seulement quand elle est positive.

- Quand la variable latente est négative, la variable dépendante prend une valeur nulle.
- Il est aisé de modifier le modèle Tobit de telle sorte qu'une censure se produise pour d'autres valeurs que zéro, de telle sorte que la censure s'applique `a des valeurs supérieures plutôt qu'inférieures, ou de telle sorte que la valeur sur laquelle se produit la censure change (d'une manière non stochastique) sur l'échantillon

Définition de Troncature et Censure

- Les modèles à variable dépendante limitée sont conçus pour traiter des échantillons tronqués ou censurés d'une certaine manière. Ces deux termes sont facilement confondus.
- Un échantillon a été tronqué si certaines de ses observations qui devaient y être ont été systématiquement exclues de l'´echantillon. Par exemple, un échantillon de ménages avec des revenus inférieurs à $100,000 exclut nécessairement tous les ménages ayant des revenus supérieurs `a ce niveau. Il ne s'agit pas d'un échantillon aléatoire de tous les ménages. Si la variable dépendante est le revenu, ou une série corrélée avec le revenu, les résultats qui utilisent l'´echantillon tronqué pourraient être potentiellement très fallacieux.

Définition de Troncature et Censure

- De l’autre côté, un échantillon a été censuré si aucune observation n’a été systématiquement exclue, mais si une certaine information contenue par ces observations a été supprimée. Songeons au “censeur” qui lit le courrier des gens et occulte certaines parties de celui-ci. Les destinataires reçoivent encore leur courrier, mais des passages de celui-ci sont illisibles. Pour continuer sur ce premier exemple, supposons que les ménages avec tous les niveaux de revenu soient inclus dans l’´echantillon, mais que pour ceux dont les revenus excèdent $100,000, le montant report´e est toujours exactement $100,000.
- Dans ce cas, l’´echantillon censuré est encore un échantillon aléatoire de tous les ménages, mais les valeurs reportées pour les ménages à hauts revenus ne sont pas les véritables valeurs. Nous pouvons assimiler les variables dépendantes discrètes à un type de censure encore plus prononcé. Par exemple, si nous nous contentions de classer les revenus des ménages dans des intervalles en dollars, la variable dépendante serait des réponses qualitatives ordonnées. Cependant, censurer à ce point n’est pas habituellement considéré comme une censure.

Les Modèles à variables endogène Dichotomique

- Une variable dépendante qualitative procède d'une dichotomie indiquant que tel événement est survenu ou non, que telle condition est en jeu ou non, c'est une variable indicatrice expliquée, exemple : individu appartient ou non à la population active. On assigne la valeur 1 à l'existence de l'événement et la valeur 0 à son inexistence.

- Le caractère dichotomique de la variable dépendante n'interdit pas le recours à la méthode des MCO pour estimer une équation de régression. Néanmoins plusieurs problèmes se posent.

Les Modèles à variables endogène Dichotomique

Les problèmes rencontrés sont:

- L’hypothèse selon laquelle le terme d’erreur suit une loi normale est violée, mais on peut utiliser le théorème central limite pour les grands échantillons.
- L’hypothèse de la non corrélation entre le terme d’erreur et la variable explicative est également violée. Toutefois, on peut surmonter cette difficulté.
- Enfin, il peut arriver que les valeurs estimées de la variable dépendante se trouvent hors de l’intervalle [0 1] mais la solution sera la concentration dans cet intervalle les probabilités estimées à l’aide de la fonction cumulative normale (modèle probit) ou de la fonction logistique (modèle logit).

Les Modèles de Régression Discrète

- Les modèles de régression discrète sont des modèles dont la variable dépendante prend des valeurs discrètes
- Le cas le plus simple de ces modèles est celui où la variable dépendante Y est binaire, c'est-à-dire elle prend deux valeurs seulement qu'on note 0 si l'événement ne se produit pas et 1 si l'événement se produit .

Exemple :

Y= 1 si l'individu travaille

Y=0 si l'individu ne travaille pas

- Et le cas le plus complexe est ce lui où la variable dépendante prend plus de deux valeurs. Ce qui va nous emmener à une classification des cas :

1-variables catégoriques

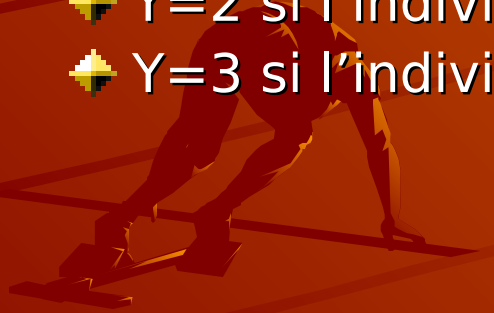
2-variables non catégoriques

Les Modèles de Régression Discrète (variables catégoriques)

- Une variable catégorique est une variable où chaque réponse peut être classée dans une catégorie particulière. Ces catégories peuvent être mutuellement exclusives ou mutuellement exhaustives. Une catégorie mutuellement exclusive signifie que toutes les réponses d'enquête possibles doivent faire partie d'une seule catégorie, alors qu'une catégorie mutuellement exhaustive signifie qu'une catégorie doit tenir compte de toutes les réponses possibles. Une variable catégorique peut être ordonnée, désordonnée ou séquentielle.

Les Modèles de Régression Discrète (variables catégoriques)

- Variable catégorique ordonnée : est une variable dite catégorique dans laquelle les catégories possibles peuvent être classées dans un ordre spécifique ou dans un ordre naturel quelconque.
- Exemple :
- Y=1 si l'individu gagne moins de 10 000dh
- Y=2 si l'individu gagne entre 10 000dh et 30000dh
- Y=3 si l'individu gagne plus que 30 000dh



Les Modèles de Régression Discrète (variables catégoriques)

- Variable catégorique désordonnée : est une variable où les catégories possibles ne suivent pas un ordre naturel. Le sexe et le genre de logement en sont des exemples.
- Exemple :
- Y=1 si le mode de transport est la voiture
- Y=2 si le mode de transport est le bus
- Y= 3 si le mode de transport est le train
- Dans cet exemple on peut définir la variable dépendante selon l'ordre qu'on désire, c'est pourquoi on l'appelle variable catégorique désordonnée.

Les Modèles de Régression Discrète (variables catégoriques)

- variable catégorique séquentielle : Ce type de variables peut être illustré à travers l'exemple suivant :
- Y=1 si l'individu n'a pas continué ses études secondaires (ES)
- Y=2 si l'individu a continué ses (ES) et non universitaires(EU)
- Y=3 si l'individu a continué ses (EU) et non de degré professionnel.
- Y=4 si l'individu a continué ses études de degré professionnel

Les Modèles de Régression Discrète (variables non catégoriques)

- L’exemple de variables non catégoriques est rencontré au cas où la variable Y indique le nombre de brevets sortis de la compagnie durant l’année. Ici Y prend des valeurs 0, 1, 2,3. .. Mais Y n’est pas une variable catégorique, pourtant c’est une variable discrète.

Conclusion

L’objet de ce cours est de présenter les différentes méthodes d’analyse et d’estimation des modèles avec des variables qualitatives, il s’agit principalement de traiter les thèmes suivants :

- **Thème 1** : les méthodes d’estimation des modèles de régression discrète
- **Thème 2** : Les modèles à variable qualitative binaire
- **Thème 3** : Les modèles multinomiaux
- **Thème 4** : les modèles à variable dépendante limitée
- **Thème 5** : L’analyse discriminante
- **Thème 6** : les modèles à équations simultanées avec variables tronquées et censurées et les méthodes d’estimation en deux étapes