Voyage à travers le longitudinal à plan complexe

Par

Johanne Boisjoly Ph.D.

Professeure titulaire

Département des Sciences Humaines

Université du Québec à Rimouski

Professeure associée

Département de Sociologie

Université de Montréal

Johanne_boisjoly@uqar.qc.ca

Boisjoly@northwestern.edu

L ’évolution des pratiques des chercheurs

Les deux premières conférences devraient porter sur les enquêtes longitudinales à plan complexe et les difficultés liées à leur usage. Nous souhaiterions que ces conférences soient données par des représentants de Statistique Canada. Les deux dernières conférences devraient aborder les problèmes liés à l'utilisation des données recueillies dans des enquêtes longitudinales à plan complexe du point de vue des chercheurs en sciences sociales.

Plan de la conférence

- Introduction
- Les enquêtes à plan complexes
  - définitions
- Présentation de deux enquêtes à plan complexes
- Évolution des pratiques des chercheurs
- Techniques de « réplications » permettant le calcul d ’erreurs-types qui tiennent compte du plan d ’échantillonnage de l ’enquête utilisée.
- Conclusion

Introduction

- Expérience d ’analyse de données longitudinales
  - données américaines seulement
  - acquise au cours des 10 dernières années:
  - d ’abord lors d ’un séjour de deux ans au Survey Reseacrh Center, Institute for Social Research, University of Michigan 1993-1995
    - Le PSID: « Panel Study of Income Dynamics »
  - ensuite par l ’analyse des données de l ’Enquête « Add Health », National Longitudinal Study of Adolescent Health (Carolina Population Center, University of North Carolina at Chapel Hill), depuis 1994.

Grands principes

Nous sommes dans l ’univers de l *’inférence statistique*

- la grande majorité des chercheurs qui utilisent les données du PSID ou de ADD Health, tout comme ceux de toutes les autres enquêtes de ce type, veulent être en mesure de généraliser les résultats de leurs travaux à l ’ensemble de la population.

Sciences sociales et statistiques

- Problème de décalage entre les concepts statistiques tels qu ’élaborés par les statisticiens, d ’une part;
- et leur intégration dans les pratiques de recherches des chercheurs en sciences sociales, d ’autre part. Ceci est dû, entre autres:
  - au caractère parfois rebutant des « équations mathématiques » qui fondent les concepts statistiques;
  - à la disponibilité « à retardement » dans les logiciels statistiques des procédures adéquates d ’analyse de données.

Les plans d ’échantillonnage complexes

- Définition:
  - les enquêtes utilisées par les chercheurs en sciences sociales sont généralement faites à partir d ’un échantillonnage à plan complexe:
    - l ’échantillon est stratifié
    - mode de sélection par grappes (« clusters »)
    - probabilités de sélection inégales d ’une unité d ’analyse à l ’autre
    - pré-requis: tous les éléments de la population avait une probabilité connue et non nulle d ’être sélectionnés.

Rappel des notions de base de l ’échantillonnage

- 1- la stratification
- 2- la sélection par grappes

La stratification et ses conséquences sur les estimations

- Définition:
  - dans un échantillonnage par strate, les éléments de la population sont classifiés par strate et la sélection de l ’échantillon se fait séparément dans chacune des strates;
  - l ’identification des strates requiert de l ’information détaillée sur les variables définissant la stratification. Si ces informations ne sont pas disponibles on procède alors par des techniques dites de « post-stratification ».

- On stratifie pour plusieurs raisons: (Lee et all)
  - réduction de la variance totale si le strates sont davantage homogènes que l ’ensemble de la population;
  - des estimations peuvent être obtenues pour chacune des strates avec une précision prédéterminée (chaque strate peut avoir une taille spécifique compte tenu de la précision souhaitée);
  - les strates (géographiques entres autres) peuvent être utilisées pour organiser la collecte de données;
  - différents objectifs d ’échantillonnage peuvent être atteints pour chacune des strates.

- Conséquences de la stratification sur les estimations:
  - il est possible d ’obtenir des estimations séparées pour chaque strate;
  - l ’application de poids appropriés (qui redonnent à chaque individu sa place dans la population) permet d ’obtenir des estimations pour l ’ensemble de l ’échantillon ;
  - dans ce cas, si on ne tient pas compte de la stratification, on SUR-ESTIME les erreurs-types. (tous les tests sont trop sévères - ce qui n ’est pas en soi un problème). (Nous y reviendrons plus loin, à l ’aide d ’un exemple détaillé).

La sélection par grappes et ses conséquences sur les estimations

- Définition:
  - dans un échantillonnage par grappes, ce sont des groupes entiers d ’éléments de population qui sont sélectionnés plutôt que les éléments eux-mêmes;
  - on procède généralement de cette manière pour réduire les coûts de « terrain »;
  - on utilise le plus souvent une hiérarchie de niveaux géographiques et de structures imbriquées pour définir les grappes.

Références:

- Lee, Eun Sul, Ronald, N. Forthofer, Ronald J. Lorimor. Analyzing Complex Survey Data. Sage: Quantitative Applications in the Social Sciences, No. 71, 1989.
- Mosteller, Frederick, John W. Tukey. Data Analysis and Regression. A second course in statistics. Addison Wesley, 1977.

Conséquences sur les résultats d ’analyse

- Tous les logiciels statistiques tels que SPSS, SAS, STATA, pour ne citer que les plus répandus, font le calcul des intervalles de confiance, des erreurs-types, etc., en prenant pour acquis que l ’échantillon est de type aléatoire avec remplacement (SRSWR: simple random sampling with replacement).

Conséquences du plan d ’échantillonnage complexes sur l ’analyse des données

- Les plans d ’échantillonnage complexes ne signifient pas automatiquement que l ’on doive procéder à une analyse complexe:
  - Les **techniques d ’analyse** sont les mêmes.
  - Les estimations doivent cependant être ajustées en regard du plan d ’échantillonnage.

Présentation des deux enquêtes à plan complexes

Raisons distinctes qui expliquent cette complexité

- Raisons de coûts:
  - Le PSID est une enquête dont le plan d ’échantillonnage visait d ’abord et avant tout à réduire les coûts de recueil de données d ’une enquête par questionnaire administré en face-à-face.
- Raisons de méthodologie:
  - L ’enquête Add Health a été conçue pour tirer avantages des avancées méthodologiques, en particulier au plan des approches multi-niveaux.

Panel Study of Income Dynamics (PSID)

- adresse: http://www.isr.umich.edu/src/psid/
- Échantillon représentatif de la population américaine tant au plan des individus que des familles qui la composent (assuré par les règles de suivi).
- But: mesurer les aspects dynamiques des comportements économiques et démographiques, tout en incluant des dimensions sociologiques et psychologiques.

- Durée et étendue:
  - L’enquête démarrée en 1968 auprès de 4,800 familles, comprend maintenant plus de 7,000 familles et plus de 62,000 individus.
  - Pour certains individus le PSID a permis de recueillir à ce jour des données sur les 34 dernières années de leur vie.
  - L’enquête s’est déroulée chaque année depuis 1968, mais depuis 1997 elle est devenue bi-annuelle.

Plan d'échantillonnage du PSID

- Résulte au départ de la fusion de deux échantillons probabilistes indépendants:
  - L'échantillon SRC: 3,000 familles
  - L'échantillon SEO: « Survey of Economic Opportunity »: 2,000 familles à faibles revenus, dont le « chef » avait moins de 60 ans
  - Échantillon stratifié et par grappes
  - Non proportionnel

Accès aux données du PSID:

- Accès: toutes les données de base du PSID sont disponibles gratuitement et peuvent être téléchargées à partir du site WEB de l ’enquête.
  - Certains fichiers comprenant des informations médicales, ainsi que des informations géographiques ne sont cependant disponibles que suite à la signature d ’un contrat de confidentialité garanti par un dépôt en argent.

National Longitudinal Study of Adolescent Health (Add Health)

- adresse: http://www.cpc.unc.edu/projects/addhealth/
- Population: adolescents américains de la 7ième à la 12ième année.
- Visait à permettre de mieux comprendre les causes de l'état de santé des adolescents et des comportements affectant cet état de santé.

Plan d ’échantillonnage

- 80 écoles secondaire ont été sélectionnées à partir d’une liste de 26,666 [les grappes] avec une probabilité de sélection proportionnelle à leur taille;
- L’échantillon a été stratifié sur la base de la région, de l’état, du niveau d’urbanisation, du type d’école (privée ou publique), et du pourcentage d’étudiants Blancs.
- On a ensuite administré le questionnaire à tous les élèves de ces écoles.

- En 1994-1995 l’enquête à l’école (« in-school ») a rejoint 90,118 répondants dans 132 écoles.
- En 1995, on a tiré parmi les répondants de l’année précédente un échantillon principal (« core ») : 200 élèves sélectionnés dans des écoles choisies avec une probabilité proportionnelle à leur taille.
- A cela s’est ajouté un échantillon « saturé »: tous les élèves de 16 écoles.

- Ce sont ensuite ajoutés des échantillons spéciaux:
  - Échantillon d'élèves handicapés
  - Échantillon d'élèves Noirs, Cubains, Porto-Ricains, Chinois provenant de milieux favorisés.
  - Échantillons génétiques: jumeaux identiques et fraternels, frères et sœurs, demi-frères et demi-sœurs, cousins et adolescents non-apparentés vivant dans un même ménage (familles reconstituées).
- En 1995, l'enquête à la maison a été réalisée auprès des 20,745 répondants sélectionnés.

- En 1996, une nouvelle vague d'enquête « wave-II » a été réalisée auprès des répondants de la vague précédente (n=14,738).
- Cinq ans plus tard, en 2001-2002, a été réalisée une nouvelle enquête auprès des répondants de la première vague à la maison (« wave-I ») pour mesurer la transition des répondants vers l'âge adulte. Le terrain vient tout juste de se terminer et plus de 16,000 personnes y ont participé.

- Cette enquête a pour particularité d'avoir été explicitement conçue pour tenir compte des divers contextes dans lesquels évoluent les adolescents: famille, réseaux d'amis, école, partenaires amoureux, voisinage.
- L'enquête se distingue en ce que les éléments de contexte ne proviennent pas des répondants eux-mêmes, mais de tous les individus qui les entourent.

Particularités de Add Health

- Les informations sur la famille proviennent d'entrevues réalisées auprès des parents des répondants et de chacun de leurs frères et sœurs (le cas échéant)
- Les données sur l'école proviennent de tous les autres élèves de l'école, de même que des administrateurs des écoles.
- Les données sur les amis et les partenaires amoureux proviennent directement de questionnaires qui leur ont été administrés.
- Les données sur les voisinages où habitent les répondants proviennent des données du recensement, agrégées au niveau du secteur de recensement.

Accès aux données de Add Health

- données publiques vendues pour 225.00$ USD
  - il s ’agit d ’un sous-échantillon
- données complètes disponibles par contrat et dépôt d ’un montant d ’argent.

Pré-requis à la correction des erreurs-types

- Connaissance **étendue** du plan d ’échantillonnage de l ’enquête utilisée
- Présence parmi les variables disponibles de l ’enquête des variables nécessaires à la correction des erreurs-types:
  - Ces données sont disponibles pour les deux enquêtes décrites ici.
  - Elles ne sont pas disponibles généralement dans les banques de données publiques (exemple IDD-Sherlock).

Évolution des pratiques des chercheurs

- Il y a eu des données longitudinales avant que les méthodes d ’analyses ne soient disponibles:
  - Ce fut le cas du PSID qui a commencé à recueillir des données plus de dix ans avant que les méthodes d ’analyses longitudinales ne soient disponibles.
  - Dans la première décennie les résultats de chaque année d ’enquête sont analysés comme s ’ils provenaient d ’une série d ’enquête transversales successives.
- L’on était bien loin en 1968, au début du PSID, des procédures de correction des erreurs-types.

Les années 80

- Développement des méthodes d'analyse longitudinales
  - Ex. RATE (Nancy Tuma et al.), TDA

Les années 90

- Développements majeurs:
  - 1- On a réalité que de disposer de plusieurs enfants de la même famille, de plusieurs membres d'un même ménage, de plusieurs personnes d'un même voisinage, etc., constituait un AVANTAGE au plan de l'analyse:
    - Permet la prise en compte de variables non-mesurées (modèles de « fixed effects »).

- 2- L’analyse multi-niveaux a permis de formaliser davantage toute cette nouvelle manière de hiérarchiser la structures des effets individuels et collectifs.
  - Encore beaucoup limitée aux logiciels spécialisés, tels HLM, MLN, etc.
  - Présente dans SAS: Proc Mixed (livre de Paul Alison en préparation à paraître chez SAS).
  - Nouvelle approche: Latent Growth Models
    - LISREL
    - À venir livre sur le sujet par Kenneth Bollen et Patrick Currant.

- 3- Les procédures de correction des erreurs-types en regard du plan d'échantillonnage sont graduellement apparues, principalement dans SUDAAN et dans STATA (elles sont maintenant disponible de manière très limitée dans SAS).

Les étapes à suivre pour la prise en compte des effets de plan d’échantillonnage

Préalables

Qui est membre de l ’échantillon?

- Qui fait partie de l ’enquête sans en être membre?
- Comment l ’établir?

Quel est le plan d ’échantillonnage?

- base de stratification (SRATA).
- identification des grappes (PSU: « primary sampling unit »).
- faudra-t-il corriger toutes les estimations en regard du plan d ’échantillonnage?
  - Si l’échantillon est uniquement stratifié, ce n’est pas nécessaire sauf si on désire obtenir les plus petites erreurs-types « honnêtes » possibles.
  - Si l’échantillon comporte des grappes, c’est ABSOLUMENT NÉCESSAIRE.

Quelles sont les variables de poids et leur signification

- les poids sont-ils standardisés (lors de leur application l ’échantillon sera-t-il gonflé artificiellement?)
- le poids tient-il compte de l ’attrition différenciée de l ’échantillon?
  - Si oui, il faudra prendre garde d ’utiliser les poids les plus récents compte tenu de l ’échantillon sélectionné.

La pondération

- Ne fait pas l ’unanimité, en particulier chez les économistes, mais aussi chez les praticiens de certaines disciplines (notamment chez ceux qui ont l ’habitude de travailler sur des échantillons restreints; i.e. psychologues).
- Certains économistes vont affirmer qu ’il suffit d ’introduire les variables de stratification au cœur du processus de sélection des membres de l ’échantillon comme variables de contrôle dans les équations de régression par exemple pour contrer l ’effet de plan
  - Or, ceci n ’est pas suffisant lorsque l ’on a des enquêtes longitudinales à plan complexe.

Outils disponibles pour la correction des erreurs-types

- SUDDAN
  - pionniers dans l ’estimation d ’erreurs-type.
- STATA
  - disponible pour une grande variété de modèles
  - différence et similarités entre _robust et svy
- SAS
  - est-ce disponible dans la version 8? Oui mais limité
    - http://www.sas.com/rnd/app/papers/survey.pdf

PROCÉDURES DISPONIBLES DANS SUDAAN ET DANS STATA

	SUDAAN	STATA
Moyennes, totaux, proportions	DESCRIPT	SVYMEAN
Tableaux de contingence	CROSSTAB	SVYTAB
Régression linéaire	REGRESS	SVYREG
Régression logistique	LOGISTIC	SVYLOGIT
Régression logistique multinomiale	MULTILOG	SVYMLOG
« Proportional Hazards »	SURVIVAL	Non disponible
Modèles log-linéaires	CATAN	Non disponible
Probit	Non disponible	SVYPROB
Analyse d'une sous-population	Énoncé SUBPOPN	Option SUBPOP

Étapes

- Procéder d'abord à l'analyse sans tenir compte des effets de plans.
- Établir les modèles non pas sur cette base, mais sur les bases théoriques de la recherche effectuée.
- Ne procéder à la correction pour les effets de plans qu'à la toute fin de l'analyse.

Exemple

- Voici un exemple des effets de la prise en considération du plan d'échantillonnage de Add Health dans une analyse de régression linéaire.
- robustadh1.xls

Que faire si aucune procédure de correction n’est disponible?

Méthode SIMPLE bien qu'imprécise

- Si possible, mesurer l 'effet de plan à partir de moyennes de variables-clés sur le même échantillon.
- Faire une mise-en-garde dans la présentation des résultats
  - la règle de stat can! (cf. IDD Sherlock)

Procédures maison:

- Nécessaires dans les cas où aucun module de correction n ’est disponible pour la technique utilisée:
  - modèles multi-niveaux
  - modèles TOBIT
  - modèles d ’équations structurelles
  - etc.
- jackknife
- bootstrap

Jackknife

- Définition: technique de réplication des résultats qui consiste à refaire les estimations n fois en enlevant une unité à la fois d'un échantillon de taille n.
- La variation de ces estimations constitue l'erreur-type.

- Application à Add Health
  - Comme le PSU est l ’école, on retire un école à la fois et on refait la régression 132 fois.
  - On fait ensuite la moyenne des estimations (qui est égale aux estimations originales).
  - La variation de ces estimations autour de la moyenne constitue l ’erreur-type désirée.

Boothstrap

- Définition: la technique du boothstrap consiste à tirer n échantillons de taille n sans remplacement, un très grand nombre de fois (1,000) et procéder à l'estimation désirée. L'écart-type des estimations autour de l'estimation originale constitue alors l'erreur-type de la statistique.
- Exemple des corrélations polychoriques pour introduire dans des modèles d 'équations structurelles
  - PRELIS2 ne permet de prendre les poids en considération

Inconvénient de ces méthodes

- Exigent de la « programmation » intensive et du temps d'ordinateur intensif, surtout si les échantillons sont grands.

Problème

- Parfois contesté lorsque l ’on soumet des articles
- Exemple du modèle TOBIT
- Exemple: jkksvy1.xls

Nouveau sujet

- l ’assignation des valeurs manquantes
  - http://www.sas.com/rnd/app/papers/mianalyze.pdf

Conclusion

- Faut-il conclure que les milliers d ’articles publiés sans prise en considération des effets de plans basent leurs conclusions sur des résultats erronés?
- Ce serait trop sévère.