

Attrition et non-réponse dans l'Enquête sur
la dynamique du travail et du revenu

Brahim Boudarbat
(Université de Montréal)

Lee Grenon
(British Columbia Inter-University Research Data Centre)

Séminaire méthodologique du CIQSS
21 novembre 2008

Motivation et objectif :

- Les études longitudinales sont devenues très populaires mais font, inévitablement, face au problème de l'érosion de l'échantillon.
- Faut-il se soucier de ce problème? Les déterminants? Les conséquences? Les solutions?
- "The increased availability of panel data ... has been one of the most important developments in applied social science research in the last thirty years. Panel data have permitted social scientists to examine a wide range of issues that could not be addressed with cross-sectional data ... Nevertheless, the most potentially damaging and frequently-mentioned threat to the value of panel data is the presence of biasing attrition- that is, attrition that is selectively related to outcome variables of interest" (Fitzgerald et al., 1998, p.252).
- "A major concern for any panel survey is the possible effect of non-response (attrition) on analytical results. For example, systematic response biases may occur because people with particular characteristics are more traceable and/or co-operative than those with other characteristics" (Gray et al., 1996, p.163).

- “The increased availability of longitudinal surveys is a major advance in facilitating the empirical analysis of individual behavior over time. [The analysis of attrition] is important because nonrandom attrition results in a data set that is no longer representative of the population from which the survey was sampled” (Zabel, 1998, p.479).
- “Differences in characteristics between two samples at initial and subsequent waves is a threat to external validity, and differences in the relationships between variables in the samples is a threat to internal validity” (Miller et Wright, 1995, p.921)
- Qu’en est-il de l’Enquête sur la dynamique du travail et du revenu?
 - Un taux combiné d’attrition et de non-réponse de 22% au deuxième panel (1996-2001)

Liste des enquêtes longitudinales (accessibles au CIQSS) :

Enquête sur la dynamique du travail et du revenu (EDTR) : permet de comprendre le bien-être économique des Canadiens : à travers quels changements économiques doivent passer les personnes et les familles? Et quel rôle jouent à cet égard les changements touchant le travail rémunéré, la composition de la famille, la réception de paiements de transfert gouvernementaux, ou d'autres facteurs?

Enquête auprès des jeunes en transition (EJET) : vise à étudier les transitions école- travail chez les jeunes (adolescents et jeunes adultes) et les facteurs qui influent ces transitions. L'enquête inclut quasiment toutes les expériences d'études formelles ainsi que celles reliées au marché du travail, et facteurs d'influence tels que le contexte familial, les expériences scolaires, les réalisations, les aspirations et attentes et expériences sur le marché du travail.

Enquête longitudinale auprès des immigrants (ELIC) : a pour objet d'examiner le processus d'intégration d'un immigrant au cours des quatre années qui suivent son arrivée au pays, période cruciale durant laquelle il noue des liens économiques, sociaux et culturels avec la société canadienne.

Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ) : a été conçue pour recueillir des renseignements sur les facteurs qui influent sur le développement social et émotionnel ainsi que sur le comportement des enfants et des jeunes.

Enquête sur le milieu de travail et les employés (EMTE) : a pour but général d'examiner de quelle manière les employeurs et leurs employés réagissent et s'adaptent au changement dans un environnement concurrentiel axé sur la technologie.

Enquête nationale auprès des diplômés (END) : vise à mesurer la situation à court et à moyen terme sur le marché du travail des diplômés des programmes publics canadiens d'enseignement universitaire, collégial et de formation professionnelle et technique.

Enquête nationale sur la santé de la population (ENSP) : vise à recueillir des renseignements sur la santé de la population canadienne ainsi que des renseignements sociodémographiques connexes.

Étude longitudinale du développement des enfants au Québec : suivi d'une cohorte d'enfants nés entre octobre 1997 et juillet 98 (naissances simples) et leur famille.

Ampleur de l'attrition / non-réponse dans certains panels :

Enquête sur l'Établissement des nouveaux immigrants (ÉNI) :

- Suivi d'une cohorte d'immigrants âgés de 18 ans et plus arrivés au Québec entre la mi-juin et novembre 1989, et résidant dans la grande région de Montréal au moment de la première entrevue un an plus tard.
- Quatre passages d'observation ont été réalisés :
 - Un an après l'arrivée : 1000 entrevues
 - Deux ans : 729 entrevues (*« les pertes sont dues pour partie à des refus de répondre ou des absences prolongées et pour partie à des migrations hors de la grande région de Montréal ailleurs au Canada ou dans le monde »*).
 - Trois ans : 508 entrevues complétées.
 - Dix ans (n'était pas prévu initialement) : 429 entrevues (88 individus étaient perdus de vue après la 1^{ère} année et 83 perdus de vue après la 2^e année).

Enquête longitudinale auprès des immigrants du Canada (ÉLIC) :

- Vise à étudier la façon dont les nouveaux immigrants s'adaptent au mode de vie du Canada au fil du temps. *Il s'agit d'une enquête à participation volontaire.*

- Trois vagues :

- Avril 2001 - mai 2002 : environ 12 000 nouveaux immigrants âgés de 15 ans et plus, au pays depuis environ six mois.
- 2003 : environ 9 300 immigrants du premier cycle ont été interviewés.
- 2005 : 7716 répondants (815 non-répondants, 163 hors du champ de l'enquête, 628 cas non résolus).

Immigrant hors du champ de l'enquête : ne satisfaisait pas aux critères définissant la population d'intérêt. Exemples : immigrants décédés, ceux qui vivaient en établissement et ceux qui avaient quitté le Canada.

Cas non résolus ou non dépistés : sont ceux identifiés à l'étape de la collecte pour lesquels il n'y a eu aucun contact avec l'immigrant sélectionné. Aucun renseignement n'a été recueilli permettant de le repérer.

Non-répondants : repéré et présent au Canada mais, pour une raison donnée, n'a pu répondre à l'interview. (Guide de l'utilisateur des microdonnées de l'ÉLIC)

Enquête nationale auprès des diplômés (END) :

- Mène des interviews des diplômés à deux moments différents, soit deux ans et cinq ans après l'obtention du diplôme des établissements postsecondaires du Canada.
- L'interview de suivi (5 ans) ne concerne que les diplômés ayant répondu à l'interview initiale (approche « en entonnoir »).
- Cohortes enquêtées jusqu'à présent : 1982, 1986, 1990, 1995 et 2000
- Le taux de réponse global à l'Enquête de suivi de 2000 est de 68,5 %.

Longitudinal Health and Life Style Survey (HALS), Grande-Bretagne (Gray et al., 1996) :

- 2 vagues : 1984-85 et 1991-92
- 3 passages à chaque vague

TABLE 2
Response at each stage of HALS1 and HALS2

<i>Stage</i>	<i>n</i>	<i>%</i>
HALS1 interview	9003†	100
HALS1 measurement	7414	83
HALS1 self-completion	6572	73
HALS2 interview	5352	59
HALS2 measurement	4480	50
HALS2 self-completion	3817	42

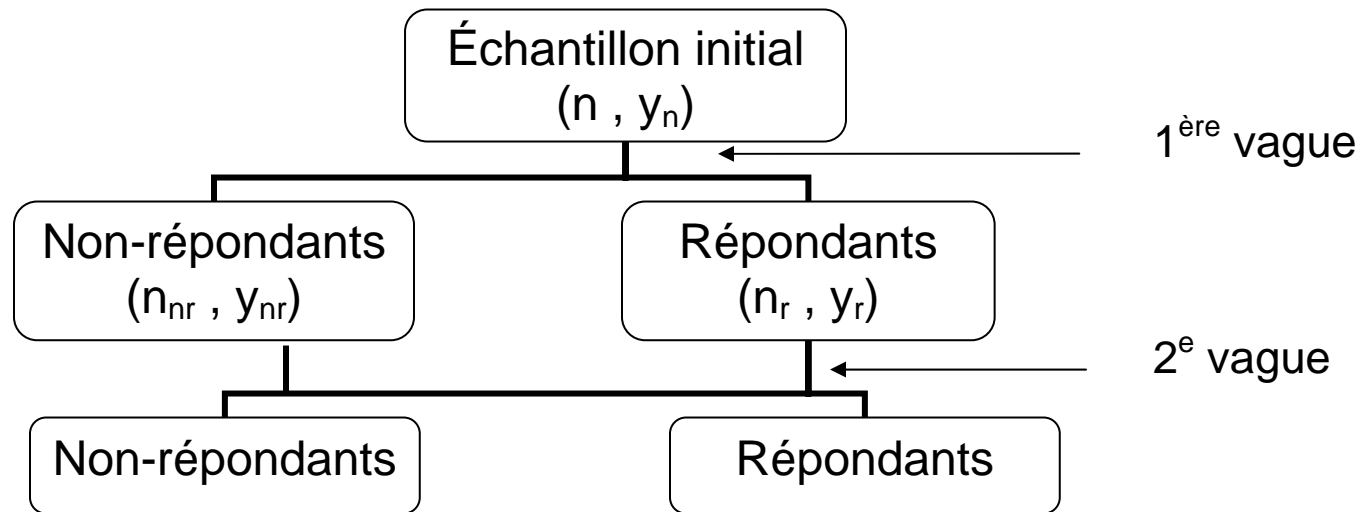
†This represents a response rate at HALS1 of 73.5%.

European Community Household Panel (ECHP); Behr et al. (2005) :

- 14 pays de l'Union européenne, 5 vagues (1994-1998)
- Objectif principal : faire des comparaisons à travers les pays.
- 255 926 répondants à la première vague.
- Taux de non participation à la dernière vague : 35,6% (27,4% inéligibles, 8,2% non répondants)
- Grande variabilité du taux de non-réponse à travers les pays : Ireland 46% vs. Portugal \approx 20%

Cohort Contraception survey-in 2000 (COCON); Razafindratsima et al. (2004) :

- Observer les pratiques de contraception chez les femmes sur une période de 5 ans.
- Perte d'un tiers de l'échantillon après 2 ans (2863 femmes en 2000; 2218 en 2001, et 1912 en 2002).
- Le taux d'attrition est plus élevé la première année (22,5% vs. 13,8% la 2^e année).



L'effet des non-réponses dépend de plusieurs facteurs (Groves, 1989) :

$$y_r = y_n + \underbrace{(n_{nr}/n)(y_r - y_{nr})}_{\text{biais}}$$

Déterminants et conséquences de l'attrition/non-réponse dans les études existantes :

Déterminants :

- Caractéristiques individuelles : statut socioéconomique, caractéristiques personnelles
- Caractéristiques de l'enquête : sujet, durée du panel, nombre de visites, mode d'entrevue, longueur du questionnaire/ durée de l'entrevue, interviewer (le même?), maintien du contact, méthodes de localisation des individus, etc.

Conséquences (biais) :

- Généralement faibles

Gray et al. (1996) :

À partir de la compilation de plusieurs études existantes :

- Niveau de scolarité : faible
- Statut d'emploi : chômage
- Revenu : faible

À partir de "the Longitudinal Health and life style survey – HALS" (Grande-Bretagne)

- Existence de différents sous-groupes selon le nombre de non-réponses aux 5 vagues de l'enquête.
- Ceux susceptibles d'avoir un comportement persistant (i.e., 5 non-réponses sur 5) : les jeunes (18-34 ans), les personnes âgées (65 ans +) et ceux dont le revenu n'est pas connu.
- Ceux avec 3 à 4 non-réponses sur 5 : non-blancs, locataires, célibataires, familles monoparentales, résidents du milieu urbain, personnes à faible revenu, ou avec un faible niveau d'instruction, fumeurs réguliers.

Behr, Bellgardt et Rendtel (2005) : European Community Household Panel

Considèrent trois types de variables:

- variables liées à l'enquête sur terrain : déménagement ou changement de l'interviewer
- variables liées au comportement vis-à-vis de l'enquête : âge, sexe, situation matrimoniale, niveau de scolarité.
- autres variables d'analyse importantes : revenu du ménage, activité sur le marché du travail

Trouvent que :

- le premier groupe de variables est le plus significativement associé avec l'attrition. Les jeunes ont également tendance à ne pas répondre.
- l'attrition a peu d'impact sur la mobilité du revenu (et donc sur le classement des pays européens).

Razafindratsima et al. (2004) : Cohort Contraception survey - 2000

Trouvent :

- une faible association entre le refus de participer et les caractéristiques des interviewers, la durée de l'entrevue ou le fait de poser des questions sensibles (ex : un enquêteur homme qui interroge une femme sur les méthodes de contraception).
- L'attrition est l'affaire essentiellement des femmes plus jeunes, plus âgées, celles ayant de faibles qualifications, femmes étrangères, et celles qui vivent seules ou en dehors d'un couple.
- Néanmoins, il y a un faible biais de sélection qui découle de l'attrition → supporte l'utilisation des données de panel en dépit du problème d'attrition.

Hill et Willis (2001) : Health and Retirement Study (HRS)

- aucune évidence que la durée de l'entrevue affecte la participation aux vagues subséquentes (effet négatif seulement quand la durée de l'entrevue dépasse 94 mn)
- garder le même interviewer à travers les cohortes a un impact positif sur les taux de réponse.

De leur côté, O'Muircheartaigh et Campanelli (1999) trouvent qu'une forte variance des taux de réponses à The British Household Panel Study, est liée aux différences entre les interviewers → suggèrent de focaliser sur la formation de ces derniers.

Fitzgerald et al. (1998) : The Michigan Panel Study of Income Dynamics (1968-89, perte d'environ 50% de l'échantillon initial)

- Les non-répondants ont tendance à avoir de faibles revenus et niveau d'instruction, de l'instabilité dans le mariage, et semblent, généralement, provenir de la partie inférieure de la "distribution socioéconomique".
- En dépit du taux de réponse élevé, les auteurs trouvent que ceci n'a pas sérieusement affecté la représentativité du PSID à travers le temps.

Lee et Panis (1998) aboutissent la même conclusion pour le PSID :

- "Although we find evidence of significant selectivity in attrition behavior, the biases that are introduced by ignoring selective attrition are very mild".

van den Berg et Maarten Lindeboom (1998) : The Netherlands Labour Supply Panel Survey (1985–1990)

- Analysent la relation entre la durée dans le chômage et l'emploi et la durée de participation à l'enquête.
- Quelques sources de biais de sélection :
 - un individu qui n'aime pas les formalités serait peu disposé à coopérer avec les agences de recrutement, et seraient également peu disposé à participer à des entrevues de longues durées dans une enquête longitudinale.
 - un individu avec des habilités de communication limitées performerait moins bien dans les entrevues pour l'emploi ou l'enquête (il essaierait d'éviter ces entrevues).
 - un individu qui passe beaucoup de temps à la recherche d'opportunités sur le marché du travail aurait moins de temps pour participer aux enquêtes.
- Conclusion :
 - "... even though we formally find significant dependence between labor market durations and attrition, it does not really matter whether we take account of this or not."

Zabel (1998) : Panel Study of Income Dynamics and the Survey of Income and Program Participation

- Le taux d'attrition augmente avec le nombre de vagues (passages).
- L'attrition est réduite si on garde le même enquêteur.
- Le biais dû à l'attrition est faible dans l'estimation des équations de participation au marché du travail et des salaires.
- Mais le comportement sur le marché du travail diffère entre les non-répondants et les répondants.

Robins et West (1986) : Seattle and Denver Income Maintenance Experiments

- Évaluent l'importance du biais dû à l'attrition dans les données sur l'offre de travail

- Conclusion:

- "Although not conclusive, the analysis suggests that attrition bias is probably not a serious enough problem in the SIME/DIME data to warrant extensive correction procedures".

Enquête sur la dynamique du travail et du revenu

Tableau 1 : Taux d'attrition et de non-réponse au 2^e panel de l'EDTR

Vague	Hors du champ d'observation		Non-répondants (dans le champ d'observation)		Total		
	Nombre (1)	% échant. (2)	Nombre (3)	% échant. (4)	(1)+(3)	(2)+(4)	
1996		(23 598 répondants, 16 - 64 ans)					0,00
1997	647	2,74	1 319	5,59	1 966	8,33	
1998	1 001	4,24	1 589	6,73	2 590	10,98	
1999	1 629	6,90	2 047	8,67	3 676	15,58	
2000	2 451	10,39	2 780	11,78	5 231	22,17	
2001	2 949	12,50	2 211	9,37	5 160	21,87	

Notes : 1/3 des individus de l'échantillon initial étaient au moins une fois hors champ ou non-répondants entre 1996 et 2001

Tableau 2 : Caractéristiques des non-répondants / hors champ

	Moyenne pour l'année :			
	1996		2000	
	Répondants en 1997	Non- répondants/ hors champs en 1997	Répondants en 2001	Non- répondants/ hors champs en 2001
Âge	38,3	37,2	42,6	40,0
Femme	0,504	0,499	0,508	0,483
Marié	0,628	0,526	0,672	0,528
Un enfant à la maison	0,179	0,130	0,149	0,112
Milieu urbain	0,821	0,875	0,805	0,842
A déménagé au cours de l'année	0,150	0,189	0,145	0,228
Immigrant	0,182	0,242	0,171	0,210
Étudiant	0,200	0,215	0,117	0,150
A travaillé durant l'année	0,770	0,739	0,786	0,740
Salaire horaire (\$)	14,68	13,81	17,69	15,64
Revenu familial (\$)	58 572	54 172	69 317	65 693

Modélisation (spécification économétrique) :

Quoi?

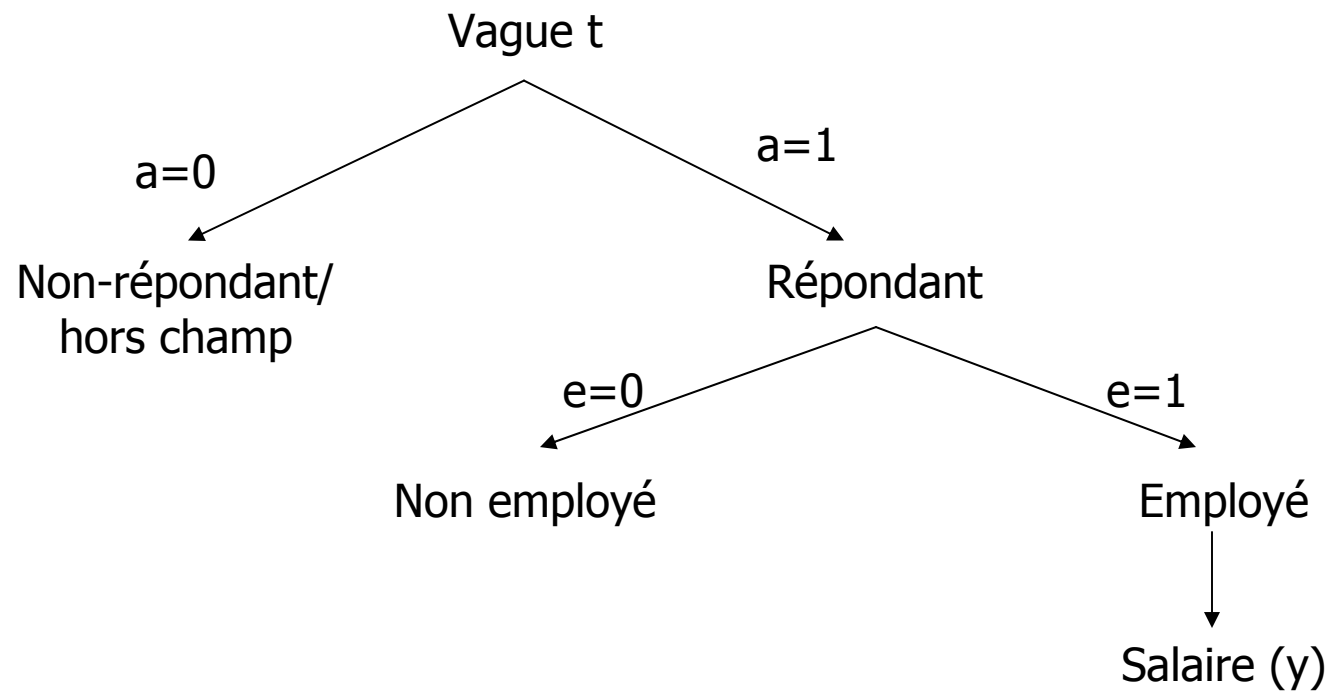
- Attrition ou non-réponse? Les deux?
- Nombre de non-réponses/attritions?
- Moment où l'attrition/non-réponse intervient?
- Non-réponse : complète ou partielle?
- Variable(s) d'intérêt? Offre de travail, salaire par exemple.

Comment?

- Modèles?
- A-t-on toutes les variables nécessaires pour modéliser correctement l'attrition/non-réponse et les variables d'intérêt?

- Considérons un panel qui comprend T vagues.
- La non-réponse (et attrition) survient à partir de la vague $t \geq 2$
- Pour des raisons de simplicité, nous considérons que la non-réponse est état absorbent (pas de retour dans l'échantillon après une non-réponse).
- Variables d'intérêt :
 - Être employé
 - Salaire

Modèle séquentiel :



➤ Critère de participation à l'enquête :

$$a_{it}^* = Z_{i,t-1}\theta + \varepsilon_{1it}, \quad i=1,\dots,n; \quad t=2,\dots,T_i \quad (1)$$

où « i » représente l'individu et « t » représente la vague.

L'individu i reste dans l'échantillon et participe à la vague t ($a_{it} = 1$) si $a_{it}^* \geq 0$. Il devient un non-répondant ($a_{it} = 0$) si $a_{it}^* < 0$.

➤ Critère d'emploi:

$$e_{it}^* = X_{it}\alpha + \varepsilon_{2it}, \quad i=1,\dots,n; \quad t=2,\dots,T_i \quad (2)$$

Le répondant i est employé ($e_{it} = 1$) si $e_{it}^* \geq 0$. Il n'est pas employé ($e_{it} = 0$) si $e_{it}^* < 0$.

➤ Équation de revenu:

$$y_{it} = W_{it}\beta + \varepsilon_{3it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T_i \quad (3)$$

où y_{it} est le log du revenu annuel d'emploi

Z_{it} , X_{it} et W_{it} sont des vecteurs de caractéristiques observables

ε_{1it} , ε_{2it} et ε_{3it} sont des termes aléatoires qui capturent l'hétérogénéité et les variables non observées.

Pour tenir compte de la sélection :

- Modèle structurel où les termes aléatoires des trois équations (ε_{1it} , ε_{2it} et ε_{3it}) sont librement corrélés.
- Problème : la vraisemblance implique des distributions $3 \times T_i$ -variées (!!!).

Simplifications :

- Modèle avec effet aléatoire ($\varepsilon_{1it} = u_{1i} + v_{1it}$, $\varepsilon_{2it} = u_{2i} + v_{2it}$ et $\varepsilon_{3it} = u_{3i} + v_{3it}$), v_{1it} , v_{2it} et v_{3it} sont mutuellement indépendants.
- Estimation en deux étapes :
 - Étape 1 : on estime les équations de participation et d'emploi à l'aide d'un modèle bivarié censuré avec données de panel et par maximum de vraisemblance simulée (Gourieroux et Monfort, 1996; et Greene, 2002)
 - Étape 2 : on estime l'équation de salaires en incluant deux termes de corrections estimés à partir des résultats de la première étape (Ham, 1982).

Résultats empiriques :

Tableau 3: Estimation du modèle bivarié censuré (modèle de sélection)

	Équation de participation à l'enquête		Équation d'emploi	
	Coefficient	Err. type	Coefficient	Err. type
Constante	0,8029***	0,1473	-1,1922***	0,3678
Âge / 10	-0,0021	0,0489	3,015***	0,1407
Âge ² / 100	0,0068	0,0056	-0,4465***	0,0167
Femme	0,0346**	0,0159	-1,007***	0,0601
Immigrant	-0,1059***	0,0301	-0,1436	0,1101
Immigrant x minorité visible	-0,1285***	0,0413	-0,461***	0,1512
Étudiant	0,0221	0,0273	-0,7122***	0,0499
Marié	0,1751***	0,0273	0,4068***	0,0717
Célibataire	-0,0442	0,0328	-0,1853*	0,1044
Femme x enfant âge préscolaire	-	-	-1,1864***	0,0729
Diplôme études secondaires	0,043*	0,0226	0,8015***	0,0598
Diplôme postsecondaire	0,0903***	0,0235	1,3324***	0,0741
Diplôme universitaire	0,0968***	0,0292	1,6728***	0,098

(*), (**) et (***) : significatif au niveau 10, 5 et 1 % respectivement.

Tableau 3: (Suite)

	Équation de participation à l'enquête		Équation d'emploi	
	Coefficient	Err. type	Coefficient	Err. type
Taille du ménage	0,0098	0,0068	-	-
Zone urbaine	0,1724***	0,052	-0,297*	0,1749
log taille population urbaine	-0,0579***	0,0098	0,0396	0,0338
Propriétaire du logement	-0,0145***	0,0038	-	-
A déménagé durant l'année de référence	-0,1658***	0,0224	-	-
log revenu familial	0,0805***	0,0224	-	-
log revenu hors travail	-	-	-0,3469***	0,0335
Taux de chômage local	-	-	-0,0638***	0,0084
État de santé ¹	0,0295***	0,0084	0,2341***	0,0154
Autres contrôles :				
Province de résidence, vague				
σ_ε	0,0287**	0,0128	2,2024***	0,0445
$\text{corr}(\varepsilon_1, \varepsilon_2)$	0,0260**	0,0113		

(*), (**) et (***) : significatif au niveau 10, 5 et 1 % respectivement.

¹ Cette variable comprend cinq catégories allant de 1=faible à 5=excellente.

Tableau 4: Équation de salaire ajustée du biais de sélection

	Coefficient	Err. type
Terme de correction de l'attrition/non réponse	-1,0115***	0,0388
Terme de correction de l'emploi	-0,2167***	0,0195

(*), (**) et (***) : significatif au niveau 10, 5 et 1 % respectivement.

Tableau 4: Équation de salaire ajustée du biais de sélection (suite)

	Non-corrigée		Corrigée	
	Coef.	Err. type	Coef.	Err. type
Constant	1,7000***	0,0124	1,9446***	0,0333
Expérience	0,0429***	0,0005	0,0308***	0,0008
Expérience^2	-0,0007***	0,0000	-0,0004***	0,0000
Femme	-0,2315***	0,0038	-0,1741***	0,0048
Immigrant	-0,0464***	0,0079	-0,0225**	0,0090
Immigrant x minorité visible	-0,1598***	0,0117	-0,0984***	0,0145
Diplôme études secondaires	0,1476***	0,0059	0,1070***	0,0074
Diplôme postsecondaire	0,2754***	0,0060	0,1999***	0,0095
Diplôme universitaire	0,5866***	0,0072	0,4935***	0,0041
Zone urbaine	-0,1433***	0,0116	-0,1861***	0,0057
log taille population urbaine	0,0144***	0,0010	0,0231	0,0241
Année : 1997	0,0247***	0,0061	0,1734***	0,0109
1998	0,0459***	0,0063	0,1924***	0,0071
1999	0,0822***	0,0063	0,1970***	0,0066
2000	0,1284***	0,0067	0,2620***	0,0125
2001	0,1583***	0,0069	0,3354***	0,0022
R ² ajusté	0,3790		0,3897	

Conclusion :

- Les enquêtes longitudinales font inévitablement face au problème de l'érosion de l'échantillon
- Il est important d'évaluer l'ampleur et les conséquences potentielles de ce problème dans le cas des enquêtes dont l'analyse des données a un impact sur les politiques publiques et les connaissances académiques.
- Dans le cas de l'EDTR, Statistique Canada évalue continuellement la qualité des données et développe des méthodologies à même d'en améliorer la fiabilité.
- Parallèlement à ces efforts, nous pensons qu'une évaluation de la sélectivité potentielle due à l'attrition/non-réponse est nécessaire. Cet exercice a été fait pour plusieurs panels aux États-Unis et en Europe.
- Les résultats de nos analyses indiquent que l'attrition/non-réponse dans l'EDTR n'est pas aléatoire.
- Il y a une corrélation positive et statistiquement significative (mais petite) entre les termes aléatoires des équations de participation et d'emploi, et entre les termes aléatoires des équations de participation et de salaire.

- Les facteurs qui sont positivement associés avec la participation à l'enquête : sexe (femme), situation matrimoniale (marié), niveau de scolarité, milieu de résidence (urbain), revenu familial, statut d'immigrant (natif ou immigrant n'appartenant à une minorité visible), mobilité géographique (non).
- Certains de ces facteurs affectent aussi la probabilité d'être employé et le salaire.
- Pour les recherches futures :
 - Distinguer « non-réponse » de « hors champ »
 - Tenir compte du retour des non-répondants dans l'échantillon
 - Tenir compte des caractéristiques de l'enquête (enquêteurs, durée des entrevues, types de questions, etc.)
 - Penser à d'autres caractéristiques socioéconomiques qui pourraient expliquer le comportement des individus vis-à-vis de l'enquête (les inclure dans le questionnaire des enquêtes futures)
 - Élaborer des modèles plus fiables pour évaluer l'ampleur du biais de sélection et ses conséquences.