Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 37-45
Statistique Canada

Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey

JOHN L. ELTINGE et IBRAHIM S. YANSANEH[1]

RÉSUMÉ

Les auteurs décrivent certaines méthodes diagnostiques simples utilisées pour guider la construction de cellules de correction pour la non-réponse. S'inspirant des travaux de Little (1986), ils étudient la construction de cellules de correction par regroupement d'unités d'échantillonnage selon la probabilité estimée de réponse ou selon la réponse estimée aux questions de l'enquête. Ils examinent plus particulièrement l'évaluation de la sensibilité des estimations corrigées de la moyenne à la variation de k, c'est-à-dire le nombre de cellules utilisées, le dépistage de cellules particulières qui nécessitent une mise au point supplémentaire, la comparaison des estimations corrigées et non corrigées de la moyenne et la comparaison des estimations obtenues au moyen des cellules fondées sur la probabilité estimée de réponse, d'une part, et sur la réponse estimée aux questions, d'autre part. Les auteurs justifient les méthodes proposées et les illustrent par une application à l'estimation du revenu moyen des unités de la U.S. Consumer Expenditure Survey.

MOTS CLÉS: Données incomplètes; données manquantes; quasi-randomisation; propension à répondre; analyse de sensibilité; correction par pondération.

1. INTRODUCTION

1.1 Énoncé du problème

Les analystes d'enquête recourent souvent à la construction de cellules de correction pour tenir compte de la non-réponse. L'idée générale consiste à définir d'abord des groupes, ou «cellules», d'unités d'échantillonnage qu'on estime présenter à peu près la même probabilité de réponse ou produire à peu près la même valeur pour une question particulière, comme celle sur le revenu, puis, à corriger par pondération ou à effectuer une simple imputation «hot-deck» dans chaque cellule de correction. L'estimateur corrigé obtenu d'une moyenne ou d'un total de la population est alors entaché d'un biais dû à la non-réponse approximativement nul, à condition que les covariances intracellulaires entre les réponses aux questions et les probabilités de réponse soient approximativement nulles.

Certains travaux antérieurs sur la correction pour la non-réponse consistaient à créer des cellules de correction en groupant des variables de classification démographiques ou géographiques simples. Cependant, Little (1986) et d'autres chercheurs ont étudié la construction de cellules par groupement direct d'unités d'échantillonnage selon la probabilité estimée de réponse ou selon la valeur estimée des réponses. Dans le présent article, nous examinons certaines méthodes diagnostiques simples qui facilitent l'application de ces idées à la création de cellules. Nous nous concentrons surtout sur la sensibilité des résultats au nombre de cellules utilisées, au dépistage de cellules particulières qui nécessitent une mise au point supplémentaire, à la comparaison des estimations corrigées et non corrigées de la moyenne et à la comparaison des estimations obtenues d'après les cellules fondées sur les probabilités estimées, d'une part, et sur les réponses estimées aux questions, d'autre part. Nous illustrons ces méthodes diagnostiques grâce aux données sur le revenu collectées dans le cadre de la U.S. Consumer Expenditure Survey.

1.2 Notation, biais dû à la non-réponse et cellules de correction

Représentons par U une population donnée de taille N et les questions d'enquête par Y_i, $i \in U$; et considérons l'estimation de la moyenne de la population $\bar{Y} = N^{-1}\sum_{i \in U} Y_i$. Tirons un échantillon s de taille n de la population U et représentons par π_i la probabilité que l'unité i soit incluse dans l'échantillon.

Supposons que la non-réponse satisfait le modèle de quasi-randomisation qui suit (Oh et Scheuren 1983). Posons que R_i est une variable indicatrice égale à 1 si l'unité d'échantillonnage i choisie est un répondant et égale à 0, autrement. Enfin, supposons que les R_i sont des variables aléatoires de Bernoulli (η_i) mutuellement indépendantes, pour lesquelles les probabilités fixes de réponse η_i peuvent varier d'une unité à l'autre. En outre, définissons les poids de sondage $\lambda_i = \pi_i^{-1}$ et la réponse moyenne non corrigée pondérée selon le plan de sondage

$$\hat{\bar{Y}}_1 \overset{\text{def}}{=} \left(\sum_{i \in s} \lambda_i R_i\right)^{-1} \sum_{i \in s} \lambda_i R_i Y_i. \tag{1.1}$$

À cause d'écarts entre les η_i, l'estimateur non corrigé $\hat{\bar{Y}}_1$ est entaché d'un biais dû à la non-réponse approximativement égal à $N^{-1}\bar{\eta}^{-1}\sum_{i \in U}\eta_i(Y_i - \bar{Y})$, où $\bar{\eta} = N^{-1}\sum_{i \in U}\eta_i$ et où les espérances mathématiques sont déterminées à la fois sur le plan de sondage original et sur le modèle de quasi-randomisation. Pour réduire ce biais, on partage souvent la population

[1] John L. Eltinge, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, U.S.A.; Ibrahim S. Yansaneh, Westat, 1650 Research Blvd., Rockville, MD 20850-3195, U.S.A.

en k «cellules de correction» U_h, et l'échantillon s, en groupes correspondants s_h, puis on utilise l'estimateur corrigé

$$\hat{\bar{Y}}_k \overset{\text{def}}{=} \sum_{h=1}^{k} w_h \bar{Y}_{hR}, \tag{1.2}$$

où $w_h = (\sum_{i\in s}\lambda_i)^{-1}\sum_{i\in s_h}\lambda_i$ et $\bar{Y}_{hR} = (\sum_{i\in s_h}\lambda_i R_i)^{-1}\sum_{i\in s_h}\lambda_i R_i Y_i$. Soulignons que, si $k = 1$, alors les estimateurs (1.1) et (1.2) sont identiques. Pour une discussion générale des méthodes basées sur les cellules de correction, consulter, par exemple, Cassel, Särndal et Wretman (1983), Oh et Scheuren (1983), et Kalton et Maligalig (1991).

L'estimateur corrigé $\hat{\bar{Y}}_k$ est entaché d'un biais résiduel dû à la non-réponse approximativement égal à

$$N^{-1}\sum_{h=1}^{k} \bar{\eta}_h^{-1} \sum_{i\in U_h} (\eta_i - \bar{\eta}_h)(Y_i - \bar{Y}_h), \tag{1.3}$$

où N_h représente le nombre d'unités dans U_h et où $(\bar{\eta}_h, \bar{Y}_h) = N_h^{-1}\sum_{i\in U_h}(\eta_i, Y_i)$. Par conséquent, on préfère construire des cellules telles que la covariance entre η_i et Y_i est approximativement nulle dans chaque cellule. En pratique, on s'efforce de le faire en construisant des cellules qui sont approximativement homogènes en regard des probabilités de réponse η_i ou des réponses aux questions Y_i, ou des deux. Dans certains cas, on définit *a priori* des ensembles «naturels» de cellules grâce à des combinaisons de variables de classification connues tant pour les répondants que pour les non-répondants. Par exemple, Ezzati et Khare (1992) utilisent 72 cellules définies selon l'âge, la race, la religion, la situation d'urbanisation et la taille du ménage pour apporter des corrections pour la non-réponse à une partie des données de la National Health and Nutrition Examination Survey. Toutefois, fort souvent, en pratique, la liste des variables susceptibles d'être utilisées pour construire les cellules est assez longue, ce qui peut produire un nombre considérable de cellules contenant peu de répondants, voire aucun. Par conséquent, plusieurs auteurs ont mis au point des méthodes permettant de déterminer les variables de classification les moins importantes et à regrouper les cellules de correction peu peuplées de façon à ce que chaque cellule retenue soit raisonnablement homogène. Consulter, par exemple, Tremblay (1986), Lepkowski, Kalton et Kasprzyk (1989), Kalton et Maligalig (1991), Goskel, Judkins et Mosher (1991) et la discussion connexe sur le regroupement des strates a posteriori de Little (1993). De surcroît, les méthodes axées sur la création de cellules de correction sont apparentées à d'autres, comme les méthodes de correction fondées sur la régression [consulter Rao (1996, section 2.4) et les auteurs qu'il cite] et la méthode itérative généralisée (Deville, Särndal et Sautory 1993).

1.3 Cellules de correction fondées sur l'estimation de la propension à répondre ou sur les réponses prévues

Comme, en principe, les cellules de correction sont approximativement homogènes, on pourrait argumenter que de telles cellules définissent implicitement un modèle pour les valeurs de η_i ou de Y_i, ou des deux. Une modélisation plus explicite mène à deux méthodes connexes de création des cellules. En premier lieu, représentons par X_i un vecteur de variables auxiliaires observées pour les unités d'échantillonnage i répondantes ainsi que non répondantes et servons-nous des valeurs observées dans l'échantillon (R_i, X_i) pour ajuster un modèle pour $\eta_i = \eta(X_i)$ par régression linéaire, logistique ou probit. Puis, construisons les cellules de l'échantillon s_h en groupant les unités d'échantillonnage selon la probabilité estimée de réponse $\hat{\eta}_i$. Comme deuxième solution, considérons la régression des réponses Y_i en fonction d'un vecteur auxiliaire X_i pour produire les réponses estimées $\hat{Y}_i$ pour les unités d'échantillonnage répondantes ainsi que non répondantes. Puis, construisons les cellules de l'échantillon s_h en groupant les unités selon les valeurs de $\hat{Y}_i$.

Ces deux méthodes ont été proposées par Little (1986) comme extension des travaux de Rosenbaum et de Rubin (1983, 1984) sur les scores de propension calculés d'après des données d'observation. Consulter aussi David, Little, Samuhel et Triest (1983). Au départ, les méthodes ont été élaborées dans le contexte d'un modèle, mais elles s'étendent directement au cadre de référence actuel. Selon Little (1986), l'utilisation de cellules fondées soit sur les valeurs de $\hat{\eta}_i$, soit sur celles de $\hat{Y}_i$, pourrait réduire le biais dû à la non-réponse et celle des cellules fondées sur $\hat{Y}_i$ permettrait aussi de contrôler la variance. En outre, dans certains cas, la méthode des cellules fondées sur $\hat{\eta}_i$ et sur $\hat{Y}_i$ est plus souple que celle des cellules définies *a priori*. De surcroît, les cellules de correction fondées sur $\hat{Y}_i$ sont reliées conceptuellement aux notions de stratification optimale (consulter, par exemple, Cochran 1977, sections 5A.7 et 5A.8).

Little (1986) ne propose pas de règle particulière pour déterminer la construction des cellules. Cependant, en s'inspirant des travaux connexes sur les données d'observation effectuées par Cochran (1968) et par Rosenbaum et Rubin (1984), on pourrait envisager de répartir les unités en cellules définies en se fondant sur les quantiles $k^{-1}j$ estimés des populations de $\hat{\eta}_i$ ou de $\hat{Y}_i$ où $j = 1, 2, ..., k-1$. Cette méthode des quantiles égaux permet de contrôler dans une certaine mesure le nombre prévu de répondants dans chaque cellule. En outre, la lecture des deux références susmentionnées donne à penser que, pour un ensemble donné de prédicteurs X_i, on peut réaliser la plupart de la réduction faisable du biais grâce à un nombre relativement faible de cellules, disons $k = 5$. Une étude de cas effectuée par Czajka, Hirabayashi, Little et Rubin (1992) comprend la construction de $k = 6$ cellules de correction fondées sur les $\hat{\eta}_i$ dans chacune de plusieurs strates en appliquant des règles un peu plus complexes que la règle des quantiles égaux considérée ici. Cependant, il ne faut pas surinterpréter le fait qu'un petit nombre de cellules pourrait être adéquat. Par exemple, si on omettait un régresseur important, les estimateurs corrigés d'après les cellules pourraient être entachés d'un biais résiduel important, quel que soit le nombre utilisé de cellules fondées sur la probabilité de réponse ou sur la réponse estimée.

Enfin, on peut remplacer la correction par pondération par l'imputation. Par exemple, la simple imputation «hot-deck» permet de remplacer une valeur manquante dans une cellule de correction donnée en sélectionnant au hasard des répondants repères dans la même cellule. Parallèlement à (1.1) et à (1.2), l'estimateur résultant de la moyenne est $\hat{\bar{Y}}_{\text{imp}} = (\sum_{i \in s} \lambda_i)^{-1} \sum_{i \in s} \lambda_i Y_i^*$, où Y_i^* représente une valeur observée ou imputée, selon le cas. En pratique, on s'appuie souvent sur la correction par pondération pour tenir compte de la non-réponse des unités et sur l'imputation, pour tenir compte de la non-réponse à une question. Cependant, pour un ensemble donné de cellules, l'estimateur ponctuel corrigé par pondération (1.2) et l'estimateur par imputation $\hat{\bar{Y}}_{\text{imp}}$ sont entachés du même biais approximatif (1.3). Par souci de simplicité, la suite du présent article portera avant tout sur la correction par pondération, mais il ne faut pas perdre de vue que, pour un ensemble donné de cellules, le problème de réduction du biais est le même qu'on se serve de ces cellules pour effectuer la correction par pondération ou par simple imputation «hot-deck».

1.4 Plan de l'article

Nous examinons dans le présent article certains détails de l'application des méthodes de construction de cellules fondées sur la probabilité estimée de réponse et sur la réponse estimée. Nous accordons une attention particulière aux méthodes diagnostiques permettant de déceler les problèmes que pose un ensemble particulier de cellules, et nous justifions et illustrons ces méthodes en décrivant en détail leur application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey. À la section 2, nous donnons certains renseignements généraux sur le problème de la non-réponse concernant le revenu. À la section 3, nous décrivons et appliquons plusieurs méthodes diagnostiques, y compris la comparaison des estimations $\hat{\bar{Y}}_k$ et des erreurs-types pour plusieurs valeurs de k (section 3.1), l'évaluation partielle du biais intracellulaire (section 3.2.1), l'évaluation de la largeur des cellules comparativement à la précision des estimations $\hat{\eta}_i$ (section 3.2.2) et la comparaison des estimations corrigées et non corrigées de la moyenne $\hat{\bar{Y}}_k$ et $\hat{\bar{Y}}_1$ (section 3.3). À la section 4, nous montrons qu'on peut appliquer des méthodes diagnostiques semblables à la correction des cellules fondées sur les chiffres prévus de revenu $\hat{Y}_i$, et nous comparons les estimations du revenu moyen calculé d'après les cellules fondées sur les probabilités estimées, d'une part, et sur le revenu estimé, d'autre part. À la section 5, nous résumons les idées principales qui sous-tendent le présent article et mentionnons certains domaines dans lesquels il conviendrait de poursuivre les travaux.

2. NON-RÉPONSE CONCERNANT LE REVENU DANS LE CAS DE LA U.S. CONSUMER EXPENDITURE SURVEY

2.1 Consumer Expenditure Survey, méthodes de pondération et estimation de la variance

La U.S. Consumer Expenditure Survey (CE) est une enquête à plan d'échantillonnage stratifié à plusieurs degrés avec renouvellement effectuée par le Census Bureau pour le Bureau of Labor Statistics (BLS). Les éléments de l'échantillon sont des «unités de consommation», grossièrement équivalentes aux ménages. Durant l'enquête, on demande à chaque unité d'échantillonnage sélectionnée de participer à cinq interviews. La méthode actuelle de pondération de la CE tient compte des probabilités de sélection initiale, de la correction pour la non-interview, de la stratification a posteriori fondée sur plusieurs variables démographiques et de mises au point supplémentaires; consulter Zieschang (1990) et le United States Bureau of Labor Statistics (1992). En raison de la complexité des travaux de pondération de la CE, le BLS a décidé d'utiliser des estimateurs de la variance fondés sur des méthodes de pseudorépétition à 44 échantillons répétés. Cette pseudorépétition est approximativement équivalente à la répétition compensée type (Wolter 1985, chapitre 3). Toutes les erreurs-types mentionnées ici sont fondées sur la méthode de pseudorépétition, toutes les étapes supplémentaires d'estimation des paramètres et de correction de la pondération étant exécutées séparément pour chaque répétition.

2.2 Non-réponse concernant le revenu

En général, on estime que la correction pour la non-interview incluse dans la méthode de pondération de la CE tient compte comme il convient de la non-réponse d'une unité, comme l'absence de contact ou le refus de participer à une interview particulière. Donc, nous ne nous pencherons plus sur la question de la non-réponse d'une unité ici. Cependant, le BLS craint que les estimations du revenu moyen soient entachées d'un biais dû à la non-réponse partielle aux questions sur le revenu de la CE. Suivent certains renseignements généraux.

Les données détaillées sur le revenu sont collectées durant les deuxième et cinquième interviews de la CE et sont utilisées pour produire des estimations du revenu moyen des unités de consommation (U.S. Bureau of Labor Statistics, 1991) et d'autres paramètres. Les données sur le revenu de la CE sont collectées grâce à un ensemble complexe de questions et le taux de non-réponse à ces questions est relativement élevé. Pour donner une indication sommaire de la réponse ou de la non-réponse à l'ensemble complet de questions sur le revenu, le BLS classifie chaque unité de consommation qui participe à la seconde ou à la cinquième interview comme déclarant complètement ou incomplètement leur revenu. La définition officielle de l'«unité déclarant complètement son revenu» est relativement compliquée; pour une discussion détaillée, consulter Garner et Blanciforti (1994). La méthode appliquée à l'heure actuelle par le BLS pour estimer le revenu moyen consiste à utiliser la réponse moyenne non corrigée $\hat{\bar{Y}}_1$ définie par (1.1), où les R_i représentent les indicateurs de déclaration complète du revenu, Y_i représente le revenu et les poids λ_i correspondent à ceux décrits à la section 2.1. La moyenne pondérée $\hat{\bar{Y}}_1$ est calculée d'après les données de la deuxième ainsi que de la cinquième interview pour une période de référence précise, mais ne s'appuie pas directement sur la structure par panel des données de la CE. Conformément à cette approche, nous ne ferons la distinction entre les données de la deuxième et de la cinquième interviews dans le présent article que pour construire les modèles de $\hat{\eta}_i$ et $\hat{Y}_i$.

Ici, nous avons utilisé les données des rapports de la deuxième et de la cinquième interviews pour toutes les unités de consommation pour lesquelles une deuxième interview était prévue en 1990. Les données de la deuxième interview se rapportent à 5,125 unités et celles de la cinquième, à 5,093 unités. Pour chaque unité interviewée (ayant déclaré complètement ou incomplètement son revenu), nous avons tiré des enregistrements du BLS des données sur un grand nombre de variables démographiques et de variables de dépenses que nous avons utilisées comme variables auxiliaires dans les travaux de modélisation décrits aux sections 3 et 4 ci-après. À la deuxième ainsi qu'à la cinquième interview, environ 14% des unités de consommation interviewées ont déclaré incomplètement leur revenu.

3. CELLULES FONDÉES SUR LES PROBABILITÉS ESTIMÉES DE RÉPONSE

Nous examinons pour commencer la construction de cellules de correction fondées sur les probabilités estimées de réponse. Nous avons ajusté séparément les modèles de régression logistique utilisés pour calculer la probabilité qu'une unité déclare complètement son revenu $\eta_i = \eta(X_i)$ aux données de la deuxième et de la cinquième interviews décrites à la section 2. Les détails de l'ajustement des modèles, y compris l'estimation des paramètres et le calcul des erreurs-types, sont décrits dans Yansaneh et Eltinge (1993). Nous avons calculé toutes les estimations de la variance par la méthode de pseudorépétition décrite à la section 2. Nous nous sommes servis des modèles résultant des ajustements finals pour estimer, pour chaque unité ayant participé à la deuxième et à la cinquième interviews, les probabilités de déclarer complètement le revenu $\hat{\eta}_i$. Conformément à la stratégie décrite à la section 1.3, nous avons regroupé les unités selon la valeur de $\hat{\eta}_i$ en k cellules dont nous avons défini les limites par la méthode des quantiles égaux.

3.1 Analyse initiale de la sensibilité au nombre choisi de cellules

Les trois premières colonnes du tableau 1 donnent les estimations ponctuelles corrigées $\hat{\bar{Y}}_k$ du revenu moyen et les erreurs-types associées pour plusieurs valeurs de k. La comparaison de ces estimations ponctuelles indique dans quelle mesure les estimations corrigées sont sensibles au choix d'une valeur particulière de k. Pour $k \geq 5$, les estimations ponctuelles présentées sont relativement stables, variant de \$32,630 à \$32,664. Cette observation concorde avec l'idée énoncée à la section 1.3 selon laquelle $k = 5$ cellules pourrait fournir la plupart de la réduction effective du biais produite par une méthode donnée de construction des cellules; consulter Rosenbaum et Rubin (1984, section 1 et appendice A) pour certains renseignements mathématiques connexes.

En outre, soulignons que pour $k \geq 3$, l'erreur-type qui entache $\hat{\bar{Y}}_k$ est également assez stable, variant de \$508 à \$530. Cette observation contredit en partie l'idée générale selon laquelle le choix du nombre approprié de cellules s'appuie sur un compromis entre le biais et la variance. En ce qui concerne l'ensemble de données examinées ici, il semble que la réduction effective du biais se produise assez rapidement (disons, pour $k = 5$,), alors qu'une augmentation considérable de la variance ne survient qu'après qu'on ait dépassé la valeur $k = 20$. Ce résultat n'est pas irréaliste, puisque, même pour $k = 20$, le nombre de réponses sur le revenu par cellule demeure assez grand (variant de 461 à 569), donc ne donne pas lieu au problème général de l'estimateur instable associé à un nombre croissant de cellules peu peuplées. Par contre, le compromis entre le biais et la variance pourrait poser des problèmes plus graves pour des valeurs assez faibles de k dans le cas d'applications où la taille effective de l'échantillon est plus petite, comme l'estimation de petites sous-populations.

Tableau 1
Estimations corrigées du revenu moyen quand les limites des cellules sont déterminées d'après les quantiles des probabilités estimées de réponse

Nombre de cellules	Estimation ponctuelle	Erreur-type	ET$(\hat{\bar{Y}}_k - \hat{\bar{Y}}_1)$	Ratio EQM$(\hat{\gamma}_k)$
Non corrigée ($k = 1$)	32,967	569	S/O	S/O
$k = 3$ cellules	32,736	530	112	1.30
$k = 4$ cellules	32,779	518	122	1.28
$k = 5$ cellules	32,630	523	138	1.53
$k = 6$ cellules	32,664	515	122	1.51
$k = 10$ cellules	32,640	514	116	1.58
$k = 15$ cellules	32,638	515	118	1.58
$k = 20$ cellules	32,634	508	118	1.63

3.2 Deux méthodes diagnostiques simples applicables aux cellules

Pour compléter l'analyse de sensibilité qui précède, il est utile d'examiner certains ensembles de cellules de correction plus en détail. Représentons par $C_1 = \{s_1, ..., s_k\}$ un ensemble donné de cellules de correction à examiner, comme les cellules créées par la division en quantiles égaux pour $k = 3$ ou $k = 5$ décrites à la section 3.1. Nous pouvons perfectionner les cellules contenues dans l'ensemble C_1 en effectuant une division en quantiles égaux pour une valeur plus grande de k ou en divisant directement une ou plusieurs cellules de l'ensemble C_1. Ce perfectionnement pourrait être utile quand les observations empiriques indiquent que 1) l'estimateur de la moyenne de la cellule $\bar{Y}_{hR}$ risque d'être fortement biaisé ou que 2) une cellule est large comparativement à la précision avec laquelle les valeurs η_i sont estimées. Nous décrivons aux sous-sections 3.2.1 et 3.2.2 deux méthodes diagnostiques simples qui permettent de résoudre les problèmes (1) et (2), respectivement. Dans chaque sous-section, la méthode diagnostique proposée mène au dépistage des «cellules problématiques» éventuelles et à la construction d'un ensemble plus perfectionné de cellules de correction que nous appelons C_2. La comparaison des estimations de $\bar{Y}$ fondées sur C_1 et C_2 permet alors de décider quel est l'ensemble de cellules de correction fondées sur $\hat{\eta}_i$ qui convient le mieux.

3.2.1 Évaluation du biais à l'intérieur des cellules

Comme nous l'avons fait remarquer à la section 1.2, un estimateur corrigé $\hat{\bar{Y}}_k$ donné réduit, mais n'élimine pas complètement, le biais dû à la non-réponse et le biais résiduel de $\hat{\bar{Y}}_k$ dépend du biais qui entache les estimations de la moyenne intracellulaire $\bar{Y}_{hR}$. Considérons l'autre estimateur de la moyenne intracellulaire

$$\bar{Y}_{h\eta} = \left(\sum_{i \in s_h} \hat{\eta}_i^{-1} \lambda_i R_i \right)^{-1} \sum_{i \in s_h} \hat{\eta}_i^{-1} \lambda_i R_i Y_i. \tag{3.1}$$

Si les estimations $\hat{\eta}_i$ étaient égales aux probabilités réelles de réponse η_i, alors (3.1) serait un estimateur approximativement non biaisé de la moyenne réelle de la sous-population $\bar{Y}_h$. Le cas échéant, un estimateur du biais intracellulaire $E(\bar{Y}_{hR} - \bar{Y}_h)$ serait $\hat{B}_h = \bar{Y}_{hR} - \bar{Y}_{h\eta}$, et l'estimateur correspondant du biais global $E(\hat{\bar{Y}}_k - \bar{Y})$ serait $\hat{B} = (\sum_{h=1}^{k} \sum_{j \in s_h} \lambda_j)^{-1} \sum_{h=1}^{k} (\sum_{j \in s_h} \lambda_j) \hat{B}_h$.

Puisque les valeurs de $\hat{\eta}_i$ sont sujettes à des erreurs d'estimation, les termes $\hat{B}_h$ et $\hat{B}$ ne donnent qu'une indication partielle des problèmes éventuels de biais. Par exemple, une grande valeur de $\hat{B}_h$ pourrait être le reflet d'un biais important entachant $\bar{Y}_{hR}$ ou de biais entachant l'autre estimateur $\bar{Y}_{h\eta}$, à cause des erreurs $\hat{\eta}_i - \eta_i$; lire les commentaires de mise en garde de Little (1986, p. 146) concernant l'utilisation directe des poids $\hat{\eta}_i^{-1}$ pour établir l'estimation corrigée de $\bar{Y}$. Donc, si on observe une grande valeur de $\hat{B}_h$, cela vaut la peine d'envisager la mise au point de la cellule h, mais la décision finale quant à l'utilisation de l'ensemble perfectionné de cellules ainsi obtenu dépendra du fait que cet ensemble produit ou non une estimation nettement différente de la moyenne globale $\bar{Y}$.

Nous présentons aux tableaux 2 et 3 les valeurs de $\hat{B}_h$, les erreurs-types associées et les valeurs de la statistique t pour les cellules formées par division en quantiles égaux pour $k = 3$ et $k = 5$, respectivement. Soulignons que, dans le cas où $k = 3$, le test diagnostic s'appliquant à $\hat{B}_h$ indique une contribution éventuelle au biais pour la cellule de rang le plus bas. Cette observation concorde avec l'idée énoncée à la section 3.1 selon laquelle $k = 3$ cellules pourrait ne pas donner une correction satisfaisante pour la non-réponse. En outre, la valeur correspondante de $\hat{B}$ est 111, avec une erreur-type de 75; cette valeur de $\hat{B}$ est très proche de la différence $\hat{\bar{Y}}_3 - \hat{\bar{Y}}_5 = 106$ des estimations $\hat{\bar{Y}}_3$ et $\hat{\bar{Y}}_5$ du tableau 1.

À la lumière des résultats qui précèdent, nous avons divisé en deux la cellule à faible $\hat{\eta}_i$ du cas où $k = 3$. Nous avons déterminé les limites supérieures des deux nouvelles cellules (soit, $h = 1'$ et $h = 1''$) grâce aux quantiles estimés 0.167 et 0.333 de la population de $\hat{\eta}_i$. Les valeurs résultantes de $\hat{B}_h$ et les erreurs-types sont 90 et 197 pour la cellule $1'$, et -42 et 79, pour la cellule $1''$. En outre, l'ensemble ainsi perfectionné de quatre cellules donne $\hat{B} = 30$, avec une erreur-type de 75, et l'estimation corrigée de $\bar{Y}$, égale à \$32,652, ainsi que l'erreur-type, égale à \$518, sont proches des valeurs obtenues par la méthode de division en quantiles égaux pour $k = 5$.

Tableau 2
Statistiques $\hat{B}_h$ cellulaires pour les cellules fondées sur les probabilités, $k = 3$

h	$\hat{B}_h$	$\text{et}(\hat{B}_h)$	$t = \hat{B}_h/\text{et}(\hat{B}_h)$
1	269	136	1.98
2	-19	43	-0.44
3	84	45	1.87

Tableau 3
Statistiques $\hat{B}_h$ cellulaires pour les cellules fondées sur les probabilités, $k = 5$

h	$\hat{B}_h$	$\text{et}(\hat{B}_h)$	$t = \hat{B}_h/\text{et}(\hat{B}_h)$
1	96	217	0.44
2	-72	116	-0.62
3	-52	56	-0.93
4	-16	27	-0.59
5	98	50	1.96

Contrairement aux résultats obtenus pour $k = 3$, les valeurs de $\hat{B}_h$ obtenues pour $k = 5$ posent assez peu de problèmes, sauf, éventuellement, dans le cas de la cellule $h = 5$, pour laquelle la valeur de la statistique t est égale à 1.96. Pour $k = 5$, la valeur de $\hat{B}$ est 11, avec une erreur-type de 93. Une division supplémentaire de la cellule $h = 5$ n'a pas modifié notablement l'estimation de $\bar{Y}$ ou de l'erreur-type associée. Les valeurs de $\hat{B}_h$ résultant de la division des cellules correspondant à une valeur de k plus grande par la méthode des quantiles égaux présentent encore moins de signes de l'existence d'un biais intracellulaire. Par exemple, pour $k = 6$, la statistique t est inférieure ou égale à 1.65 pour les valeurs de $\hat{B}_h$ de chacune des six cellules et, pour $k = 10$, la statistique t est inférieure ou égale à 1.54 pour les valeurs de $\hat{B}_h$ de chacune des dix cellules.

3.2.2 Comparaison entre la largeur des cellules et la précision des estimations η_i

La comparaison de la largeur des cellules de correction à la largeur des intervalles de confiance associés aux probabilités de réponse η_i fournit une autre méthode diagnostique pour repérer les cellules problématiques éventuelles. Premièrement, représentons par $a_h = (\sum_{i \in s_h} \lambda_i R_i)^{-1} \sum_{i \in s_h} \lambda_i$, le facteur de correction pour la non-réponse appliqué aux unités répondantes de la cellule h. Deuxièmement, conformément aux résultats types de la régression logistique, notons qu'un intervalle de confiance d'environ 95% pour η_i est

$$(LB_i, UB_i) = ([1 + \exp\{-X_i'\hat{\theta} + 1.96 D_i^{1/2}\}]^{-1}$$
$$[1 + \exp\{-X_i'\hat{\theta} - 1.96 D_i^{1/2}\}]^{-1}),$$

où $\hat{\theta}$ est le vecteur des estimations des paramètres de régression logique, où $D_i = X_i' \hat{V}_\theta X_i$, et où $\hat{V}_\theta$ est la matrice de covariance estimée d'après la pseudorépétition pour $\hat{\theta}$. Représentons par $\bar{d}_h$ la moyenne de l'échantillon pondérée en fonction de λ_i des largeurs des intervalles de confiance $UB_i - LB_i$ pour les unités i de la cellule h, et comparons $\bar{d}_h$ à la largeur de la cellule h. Si la cellule h est relativement large, tant en valeur absolue que comparativement à $\bar{d}_h$, alors, la division de cette cellule peut produire de nouvelles cellules ayant des facteurs de pondération a_h nettement différents. Inversement, si $\bar{d}_h$ est beaucoup plus grand que la largeur de la cellule h, alors, les écarts entre les $\hat{\eta}_i$ dans cette cellule peuvent résulter davantage de l'erreur d'estimation que d'écarts entre les η_i réels. Le cas échéant, une division supplémentaire de la cellule modifiera vraisemblablement peu les facteurs de pondération a_h; et, par conséquent, l'estimateur de $\bar{Y}$ corrigé pour la non-réponse obtenu variera assez peu.

Nous présentons aux tableaux 4 et 5 les limites des cellules, les largeurs des cellules, $\bar{d}_h$, ainsi que les valeurs de a_h pour $k = 5$ et $k = 10$, respectivement. Pour $k = 5$, la largeur des cellules 2 à 5 n'est pas grande comparativement aux valeurs de $\bar{d}_h$. Essentiellement, chacune de ces cellules est divisée en deux pour produire le cas où $k = 10$ cellules. Les paires résultants de a_h pour $k = 10$ sont assez proches des valeurs correspondantes de a_h dans les cellules 2 à 5 pour $k = 5$.

Tableau 4

Limites des cellules fondées sur la probabilité estimée de réponse, largeur moyenne des cellules, largeur des intervalles de confiance et facteur de correction pour la non-réponse, $k = 5$

h	Limite inférieure	Limite Supérieure	Largeur de la cellule	$\bar{d}_h$	a_h
1	0.384	0.810	0.426	0.197	1.35
2	0.810	0.861	0.051	0.139	1.20
3	0.861	0.894	0.033	0.110	1.13
4	0.894	0.924	0.030	0.088	1.08
5	0.924	0.994	0.070	0.067	1.07

En revanche, pour $k = 5$, la cellule 1 est plus de deux fois plus large que $\bar{d}_1$. Quand $k = 10$, cette cellule est divisée en cellules plus petites ayant des facteurs de pondération corrigés pour la non-réponse a_h légèrement différents, soit 1.45 et 1.27, respectivement. Cependant, les estimations correspondantes de la moyenne cellulaire sont assez proches, à savoir $\bar{Y}_{1R}$ = \$24,045 et $\bar{Y}_{2R}$ = \$24,582 pour $k = 10$. Donc, dans cet exemple, les estimations corrigées pour la non-réponse $\hat{\bar{Y}}_5$ et $\hat{\bar{Y}}_{10}$ sont relativement proches, parce que quatre des cinq divisions cellulaires entraînent une variation assez faible des poids et que la cinquième produit des cellules dont la moyenne est similaire.

Tableau 5

Limites des cellules fondées sur la probabilité estimée de réponse, largeur des cellules, largeur moyenne des intervalles de confiance et facteur de correction pour la non-réponse, $k = 10$

h	Limite inférieure	Limite supérieure	Largeur de la cellule	$\bar{d}_h$	a_h
1	0.384	0.762	0.378	0.220	1.45
2	0.762	0.810	0.048	0.174	1.27
3	0.810	0.840	0.030	0.146	1.21
4	0.840	0.861	0.021	0.132	1.19
5	0.861	0.878	0.017	0.111	1.14
6	0.878	0.894	0.016	0.108	1.11
7	0.894	0.908	0.014	0.093	1.09
8	0.908	0.924	0.016	0.083	1.08
9	0.924	0.944	0.020	0.072	1.08
10	0.944	0.994	0.050	0.062	1.06

Enfin, les facteurs a_h du tableau 5 indiquent que les taux moyens de réponse dans le cas de $k = 10$ cellules se situent dans une fourchette raisonnable, allant de $(1.45)^{-1} = 0.69$ à $(1.06)^{-1} = 0.94$. Certains ensembles de données sur la non-réponse sont caractérisés par une fourchette plus large, donc plus susceptibles de produire des écarts plus prononcés après la division des cellules. Inversement, d'autres sont caractérisés par une distribution plus serrée des probabilités de réponse, donc moins susceptibles d'être affectés notablement par la division des cellules.

3.3 Comparaison des estimations corrigées d'après les cellules aux estimations non corrigées

Pour conclure l'évaluation des cellules fondées sur les $\hat{\eta}_i$, nous comparons les estimations corrigées $\hat{\bar{Y}}_k$ aux estimations non corrigées $\hat{\bar{Y}}_1$. Premièrement, le tableau 1 indique que, pour les valeurs déclarées de $k \geq 5$, les différences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ sont supérieures ou égales à \$303. Deuxièmement, pour $k \geq 5$, les erreurs-types estimées des différences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ sont toutes inférieures ou égales à \$138, et les valeurs correspondantes de la statistique t sont toutes supérieures à 2.44. Donc, pour $k = 5$ par exemple, un test formel de l'hypothèse $H_0: \mathrm{E}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_5) = 0$ mènerait au rejet de cette dernière pour les seuils de signification types, ce qui signifie que la méthode des cellules de correction produit une modification importante de l'estimation du revenu moyen.

De surcroît, une comparaison grossière de l'efficacité de $\hat{\bar{Y}}_1$ et $\hat{\bar{Y}}_k$ se dégage du ratio des erreurs quadratiques moyennes estimées

$$\hat{\gamma}_k = \{\hat{V}(\hat{\bar{Y}}_k)\}^{-1}[\hat{V}(\hat{\bar{Y}}_1) + \max\{0, (\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)^2 - \hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)\}]$$

où $\hat{V}(\hat{\bar{Y}}_1)$, $\hat{V}(\hat{\bar{Y}}_k)$, et où $\hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$ sont les estimations de la variance basée sur la pseudorépétition pour les moyennes indiquées. Pour interpréter ce ratio, supposons pour le moment que $\hat{\bar{Y}}_k$ est un estimateur approximativement non biaisé de $\bar{Y}$. Alors, $\hat{\gamma}_k$ est un estimateur de l'erreur quadratique moyenne de l'estimateur non corrigé $\hat{\bar{Y}}_1$, comparativement à l'erreur quadratique moyenne de $\hat{\bar{Y}}_k$.

Conséquemment, $\hat{\gamma}_k$ reflète la perte d'efficacité résultant de l'utilisation de l'estimateur biaisé, non corrigé $\hat{\bar{Y}}_1$ au lieu de l'estimateur corrigé, non biaisé $\hat{\bar{Y}}_k$. Cependant, cette interprétation doit être considérée avec prudence, puisqu'elle dépend de la supposition que $\hat{\bar{Y}}_k$ est un estimateur approximativement non biaisé de $\bar{Y}$, et, puisque les $\hat{\gamma}_k$ sont des fonctions des termes aléatoires $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$, $\hat{V}(\hat{\bar{Y}}_1)$, $\hat{V}(\hat{\bar{Y}}_k)$, et $\hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$.

Comme l'a suggéré un arbitre, on pourrait aussi considérer le ratio des erreurs quadratiques moyennes

$$\{\hat{V}(\hat{\bar{Y}}_\eta)\}^{-1}[\hat{V}(\hat{\bar{Y}}_k) + \max\{0, (\hat{\bar{Y}}_k - \hat{\bar{Y}}_\eta)^2 - \hat{V}(\hat{\bar{Y}}_k - \hat{\bar{Y}}_\eta)\}]$$

où $\hat{\bar{Y}}_\eta$ est égal à l'expression (1.1) avec λ_i remplacé par $(\hat{\eta}_i)^{-1}\lambda_i$. Cette approche équivaudrait à comparer chaque estimation $\hat{\bar{Y}}_k$ fondée sur les cellules à $\hat{\bar{Y}}_\eta$. Cette démarche est appropriée si $\hat{\bar{Y}}_\eta$ est approximativement non biaisé, mais l'absence de biais peut être problématique dans certains cas; à cet égard, consulter Little (1986, p. 146).

La dernière colonne du tableau 1 donne les ratios estimés $\hat{\gamma}_k$ pour des valeurs particulières de k. Pour $k \geq 5$, chaque $\hat{\gamma}_k$ présenté est plus grand que 1.5. Enfin, soulignons que chaque estimation corrigée $\hat{\bar{Y}}_k$ est inférieure à l'estimation non corrigée $\hat{\bar{Y}}_1$. Cette situation tient au fait que, pour une valeur donnée de k, les cellules associées aux probabilités de réponse les plus grandes ont tendance à produire une estimation plus grande de la moyenne $\bar{Y}_{hR}$. Par exemple, pour $k = 5$, les valeurs de $\bar{Y}_{hR}$ sont \$24,333, \$33,729, \$33,398, \$34,620 et \$37,057 pour $h = 1$ (cellule à $\hat{\eta}_i$ la plus faible) à $h = 5$ (cellule à $\hat{\eta}_i$ la plus élevée), respectivement.

4. CELLULES FONDÉES SUR LES VALEURS ESTIMÉES DU REVENU

Les idées diagnostiques générales décrites à la section 3 s'appliquent également aux cellules fondées sur $\hat{Y}_i$. Pour illustrer ce point, nous ajustons séparément des équations de régression pondérée où Y_i = représente le revenu déclaré par les répondants à la deuxième et à la cinquième interviews. Yansaneh et Eltinge (1993) décrivent en détail les calculs, y compris l'estimation des paramètres et des erreurs-types. Nous nous sommes servi des modèles de régression résultants pour calculer les estimations du revenu $\hat{Y}_i$ pour les unités qui ont déclaré complètement et incomplètement leur revenu, puis nous avons regroupé les unités en cellules d'après leur valeur de $\hat{Y}_i$, les limites des cellules étant déterminées par la méthode des quantiles égaux.

Nous présentons au tableau 6 les résultats de l'analyse fondamentale de sensibilité et de la mesure de l'efficacité pour les cellules fondées sur les $\hat{Y}_i$; sa présentation est la même que celle du tableau 1. Les résultats de l'analyse de sensibilité sont qualitativement similaires, mais non identiques, à ceux présentés pour les cellules fondées sur les $\hat{\eta}_i$. Au cours de travaux supplémentaires non décrits en détail ici, nous avons examiné la division de chaque cellule fondée sur les $\hat{Y}_i$ en quantiles égaux. Pour $k \geq 4$, les estimations résultantes de la moyenne et les erreurs-types connexes ne diffèrent pas notablement de celles présentées au tableau 6.

Tableau 6
Estimations corrigées du revenu moyen quand les limites des cellules sont déterminées d'après les quantiles du revenu estimé

Méthode de correction	Estimation ponctuelle	Erreur-type	ET$(\hat{\bar{Y}}_k - \hat{\bar{Y}}_1)$	Ratio EQM
Non corrigée ($k = 1$)	32,967	569	S/O	S/O
$k = 3$ cellules	32,512	509	106	2.01
$k = 4$ cellules	32,468	512	108	2.14
$k = 5$ cellules	32,473	511	115	2.12
$k = 6$ cellules	32,492	508	117	2.08
$k = 10$ cellules	32,488	510	119	2.07
$k = 15$ cellules	32,478	504	124	2.16
$k = 20$ cellules	32,495	513	124	2.02

Les deux dernières colonnes du tableau 6 permettent de comparer $\hat{\bar{Y}}_k$ à l'estimation non corrigée $\hat{\bar{Y}}_1$. Pour $k \geq 4$, les différences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ sont supérieures ou égales à \$472, et les erreurs-types estimées, inférieures ou égales à \$124. Les valeurs de la statistique t associée sont toutes supérieures à 3.80. En outre, les ratios des erreurs quadratiques moyennes estimées $\hat{\gamma}_k$ sont tous plus grand que 2.0.

Qui plus est, les cellules fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$ produisent des estimations corrigées légèrement différente du revenu moyen, mais les écarts observés ne sont pas statistiquement significatifs pour les seuils de signification α ordinaires. Par exemple, pour $k = 5$, l'écart entre les estimations fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$ est égal à 32,630\$ - \$32,473 = \$157, l'erreur-type est égale à \$122 et la statistique t est égale à 1.29. Pareillement, pour $k = 10$, la différence entre les estimations fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$ est de \$152, avec une erreur-type de \$104. Donc, les données permettent peu de faire la distinction entre les résultats obtenus par les deux méthodes générales de construction de cellules.

Enfin, soulignons qu'un ensemble donné de cellules fondées sur les $\hat{Y}_i$ est fondamentalement lié à une variable Y particulière, comme le revenu de l'unité de consommation. Par conséquent, cet ensemble de cellules ne donnera pas nécessairement de bons résultats pour l'estimation de la moyenne d'une variable Y différente.

5. DISCUSSION

5.1 Résumé des méthodes

Nous examinons dans le présent article certaines méthodes diagnostiques simples permettant de construire des cellules de correction pour la non-réponse. Nous pouvons résumer comme suit la méthodologie.

1. D'après des travaux de modélisation préliminaires et les variables auxiliaires étudiées X_i, calculer la probabilité estimée de réponse $\hat{\eta}_i$ pour chaque unité d'échantillonnage (répondants et non-répondants).

2. Construire k cellules de correction dont les limites sont déterminées par les quantiles $k^{-1}j$ estimés de la population de $\hat{\eta}_i$, $j = 1, 2, ..., k - 1$. Calculer l'estimation de la moyenne corrigée résultante, $\hat{\bar{Y}}_k$.
3. Répéter (2) pour plusieurs nombre entiers $k > 1$. À mesure que k augmente, repérer le point où $\hat{\bar{Y}}_k$ devient à peu près constante. Compte tenu des résultats de Rosenbaum et Rubin (1984) et des résultats empiriques exposés ici, les valeurs de k proches de 5 pourraient présenter un intérêt particulier.
4. Utiliser des méthodes diagnostiques simples (par exemple $\hat{B}_h$ et $\bar{d}_h$ décrits à la section 3.2) pour repérer les cellules de corrections obtenues par division en quantiles égaux posant éventuellement des problèmes. Si l'application de la méthode diagnostique indique que certaines cellules sont problématiques, essayer d'effectuer une mise au point supplémentaire de ces cellules. Calculer les estimations de $\bar{Y}$ d'après les ensembles perfectionnés de cellules et comparer les nouvelles estimations aux valeurs de $\hat{\bar{Y}}_k$ obtenue au point 3.
5. Évaluer l'effet global de la correction en comparant les différences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ aux erreurs-types se$(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$ et en calculant les ratios des erreurs quadratiques moyennes estimées $\hat{\gamma}_k$.
6. Répéter les étapes (1) à (5), au besoin, pour les cellules de correction fondées sur les $\hat{Y}_i$. Comparer les estimations finales de $\bar{Y}$ obtenues par les méthodes fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$.

5.2 Domaines dans lesquels il faut poursuivre les travaux

Les résultats de la présente étude donnent à penser qu'il pourrait être utile de poursuivre les travaux dans deux domaines. Premièrement, le problème que pose la non-réponse aux questions sur le revenu dans le cas de la CE est similaire à celui que pose la non-réponse dans le cas de plusieurs autres enquêtes à grande échelle. Cependant, comme pour toute étude de cas, on devrait éviter de surgénéraliser les résultats empiriques présentés ici. Il serait utile d'appliquer les méthodes diagnostiques que nous décrivons aux problèmes que posent d'autres estimations (par exemple, moyennes croisées) ou à des ensembles de données sur la non-réponse présentant des caractéristiques légèrement différentes (par exemple, taille effective de l'échantillon plus grande ou plus petite, ou distribution plus étalée ou plus serrée des estimations des $\hat{\eta}_i$). Cet exercice apporterait des éclaircissements sur les caractéristiques opératoires des méthodes de construction de cellules fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$ dans la pratique. Deuxièmement, il serait intéressant d'étendre l'étude à des problèmes à plusieurs variables (par exemple, rapport entre les données sur le revenu de la deuxième et de la cinquième interview de la CE).

REMERCIEMENTS

Les auteurs remercient Richard Dietz, Thesia Garner, Paul Hsen, Eva Jacobs, Geoffrey Paulin, Stuart Scott et Stephanie Shipp pour les nombreuses discussions fructueuses sur la Consumer Expenditure Survey, et Wayne Fuller, Steve Miller, Geoff Paulin, Stuart Scott, trois arbitres, et le rédacteur, pour leurs commentaires précieux sur des versions antérieures de l'article. Les présents travaux ont été effectués pendant un séjour des auteurs au Bureau of Labor Statistics organisé dans le cadre du ASA/NSF/BLS Research Fellow Program et financé grâce à une bourse de la National Science Foundation (SES-9022443). Les travaux de recherche de Eltinge ont également été financés en partie par une bourse du National Institutes of Health (CA 57030-04). Les opinions exprimées dans le présent article sont celles des auteurs et ne représentent pas nécessairement les politiques du Bureau of Labor Statistics.

BIBLIOGRAPHIE

CASSEL, C.-M., SÄRNDAL, C.-E., et WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. Dans *Incomplete Data in Sample Surveys*, (Vol. 3), (Éds. W.G. Madow, I. Olkin, et D. Rubin). New York: Academic Press, 143-160.

COCHRAN, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 205-213.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley.

CZAJKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., et RUBIN, D.B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business and Economic Statistics*, 10, 117-131.

DAVID, M.H., LITTLE, R.J.A., SAMUHEL, M., et TRIEST, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 168-173.

DEVILLE, J.-C., SÄRNDAL, C.-E., et SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

EZZATI, T., et KHARE, M. (1992). Nonresponse adjustments in a national health survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 339-344.

GARNER, T.I., et BLANCIFORTI, L.A. (1994). Household income reporting: An analysis of U.S. Consumer Expenditure Survey data. *Journal of Official Statistics* 10, 69-91.

GOKSEL, H., JUDKINS, D.R., et MOSHER, W.D. (1991). Nonresponse adjustments for a telephone follow-up to a national in-person survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 581-586.

KALTON, G., et MALIGALIG, D.S. (1991). A comparison of methods of weighting adjustment for nonresponse. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 409-428.

LEPKOWSKI, J., KALTON, G., et KASPRZYK, D. (1989). Weighting adjustments for partial nonresponse in the 1984 SIPP panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 296-301.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 139-157.

LITTLE, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

OH, H.L., et SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. Dans *Incomplete Data in Sample Surveys*, (Vol. 2), (Éds. W.G. Madow, I. Olkin et D.B. Rubin). New York: Academic Press, 143-184.

RAO, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

ROSENBAUM, P.R., et RUBIN, D.B. (1983). The central role of the propensity score in observational studies for causal effects *Biometrika*, 70, 41-55.

ROSENBAUM, P.R., et RUBIN, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-524.

TREMBLAY, V. (1986). Critères pratiques pour la définitions des classes de pondération. *Techniques d'enquête*, 12, 91-103.

UNITED STATES BUREAU OF LABOR STATISTICS (1991). News: Consumer Expenditures in 1990. Publication USDL 91-607, United States Department of Labor, Washington, DC.

UNITED STATES BUREAU OF LABOR STATISTICS (1992). BLS Handbook of Methods. Bulletin 2414, United States Department of Labor, Washington, DC.

WOLTER, K.M. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag.

YANSANEH, I.S., et ELTINGE, J.L. (1993). Construction of adjustment cells based on surrogate items or estimated response propensities. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 538-543.

ZIESCHANG, K.D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.